Jurnal Nasional Teknik Elektro dan Teknologi Informasi
Volume 13 Number 2 May 2024

Translation of article 10.22146/jnteti.v13i2.9579

Optimization of the KNN Algorithm through Outlier Analysis Comparison (Distance, Density, LOF-Based)

Fitri Ayuning Tyas[1], Mahda Nurayuni[1], Hidayatur Rakhmawati[1]

[1] Program Studi Sistem Informasi, STMIK Muhammadiyah Paguyangan Brebes, Brebes, Jawa Tengah 52276, Indonesia

[Submitted: 19 September 2023, Revised: 1 December 2023, Received: 20 March 2024]
Corresponding Author: Fitri Ayuning Tyas (email: tyas_fa@stmikmpb.ac.id)

ABSTRACT — The current data growth affects data analysis in various fields, such as astronomy, business, medicine, education, and finance. The collected and stored data contain extreme values or observation values different from most other observation value results. These extreme values are called outliers. Outliers on some data often hold valuable information, necessitating thorough examination to determine whether to retain or discard them prior to data mining application. Outlier detection can be performed as a part of data preprocessing using outlier analysis techniques. Commonly utilized outlier analysis techniques encompass distance-based methods, density-based methods, and the local outlier factor (LOF) method. k-nearest neighbors (KNN) are a data mining algorithm susceptible to outliers due to its reliance on the value of *k*. Hence, having an appropriate handling mechanism is essential when employing KNN on datasets that contain outliers. The experimental method was selected to apply the proposed approach, aiming to optimize the KNN algorithm through a comparison of outlier analysis methods (KNN-distance, KNN-density, and KNN-LOF). The results revealed that KNN-density outperformed the others significantly: achieving an average accuracy of 99.34% at k=3 and k=5 for Wisconsin Breast Cancer, 85.25% at k=7 for Glass, and 85.45% at k=5 for Lymphography. Moreover, both the Friedman and Nemenyi tests validate a notable distinction between KNN-density and KNN-LOF.

KEYWORDS — K-Nearest Neighbors, Outlier, Density, Distance, LOF, Friedman Test, Nemenyi Test.

I. INTRODUCTION

The amount of data being generated and stored has been consistently increasing, but despite this growth, much of these data still lack significant value as actionable information. Data analysis is crucial for transforming data into usable information across a range of fields, including astronomy, business, medicine, education, and finance [1], [2]. Data mining enables the utilization of data as a tool to extract knowledge from it [1], [3]. Data can yield knowledge in various forms, such as patterns, formulas, decision trees, and more. Data mining is a study of collecting, cleaning, processing, and analyzing existing data to derive valuable insights from it [4], [5]. Thus, data that initially consists only of facts becomes valuable knowledge or reveals patterns once it undergoes data mining.

In data mining and statistical literature, outliers are often described as abnormalities, discordant, deviant, or anomalous [5]. In terms of classification, outliers are commonly seen as unimportant features, absent data points, or instances that are redundant or inconsistent [6]. Outliers can have adverse impacts on the outcomes of data analysis, thus necessitating special attention [7]. Even though outliers exhibit distinct behavior compared to the majority of data and are frequently seen as noise, they frequently carry valuable information [8]. Noise refers to random fluctuations in a measured variable, which may manifest as deviations in attribute values, incorrect or missing values, and are considered outliers [1], [9]. While outliers are distinct from noise [1], noise contributes to the outlier phenomenon. Knowledge extraction from data containing noise or outliers presents a challenging task within the field of data mining [10]. Research on outlier detection issues is still ongoing in various studies.

The detection of outliers plays a crucial role in data preprocessing. When conducting data mining, the presence of outliers can lead to the generation of inaccurate results [11]. Data preprocessing involves various techniques for handling data prior to its processing stage, including tasks like data cleaning, transformation, and standardization [1], [12]. Outlier detection plays a significant role in various tasks like decision-making, grouping, and pattern classification. It helps uncover rare yet crucial phenomena and identify intriguing or unexpected patterns [13]. Different methods for detecting outliers are categorized into statistical, cluster-based, distance-based, and density-based techniques [14], [15]. RapidMiner is a data mining platform utilized for data processing, offering outlier detection features including distance-based detection, density analysis, and local outlier factor (LOF). Therefore, it is crucial to choose the appropriate outlier detection method to identify any outliers in the dataset.

The k-nearest neighbors (KNN) algorithm is a popular lazy learning algorithm extensively employed in data mining for classification purposes. Its simplicity and straightforward implementation not only contribute to its effectiveness but also make it adaptable to a wide range of applications [16]–[18]. KNN operates by determining the nearest distance between multiple k data objects or patterns in both the training and test datasets. It then selects a class based on the highest occurrence among these k patterns [16]. Finding the appropriate k-value holds significant importance in KNN, particularly within the field of outlier detection. If the k-value is too small, the outcome will be highly influenced by outliers. On the other hand, if the k-value is too large, the outcome will be more resistant to outliers [6], [16], [19]. Hence, it is essential to manage KNN effectively when dealing with datasets containing outliers.

This research proposes a KNN algorithm optimization through outlier analysis (based on distance, density, and LOF).

The proposed methods, named KNN-distance, KNN-density, and KNN-LOF, aim to detect outliers within the dataset. Outlier detection aims to determine the value or number of outliers in a dataset, and then remove any values that contain outliers. Additionally, the performance of the KNN-distance, KNN-density, and KNN-LOF method models was assessed using 10-fold cross-validation. The resulting accuracy average was compared based on the level of significance using the Friedman test. The Friedman test was employed to highlight variations between the proposed approaches, while the Nemenyi test was utilized to identify which proposed method exhibited the most notable distinction.

The following section of this paper outlines various methods employed, discusses the research approach, and presents the obtained outcomes. Section II elucidates the theoretical foundations of outlier detection. Section III discusses the research methodology that has been employed, beginning with dataset collection, conducting proposed method experiments, and evaluating the experimental results. Section IV covers the test results, while Section V draws conclusions from these findings.

II. OUTLIER DETECTION

Outlier detection is a crucial step in implementing data mining [20], and it is extensively utilized in research on identifying abnormal cases within databases [21]. In the previous discussion, it was mentioned that among the techniques for outlier detection are methods based on statistics, cluster, distance, and density. The statistical-based method detects outliers by calculating the average value of data points. Examples of statistical-based methods include Gaussian distributions and histograms [2], [13], [22]. The points with low probability generated by the distribution method are considered outliers [22]. The cluster-based method involves grouping similar objects by calculating the distance matrix between clusters. Examples of cluster-based methods include k-means, self-organizing maps (SOM), and one-class support vector machines (SVM) [13], [22], [23]. Data points that lie distant from the cluster or group are typically regarded as outliers. The distance-based method employs a technique that calculates the distance of each data point from its neighbors. The term “outliers” refers to objects that are situated farther away from their neighbors [22]. Distance-based methods include KNN and outlier detection using indegree number (ODIN) [13], [23], [24]. Density-based methods calculate the density of a data point and compare it with its surrounding data points, which is known as the outlier score [22]. Normal data points and neighboring data points should exhibit similar densities. Alternatively, outliers exhibit varying densities [22]. Density-based techniques are proposed to overcome the limitations of distance-based global outlier detection. Some density-based methods include local outlier probabilities (LoOP), local correlation integral (LOCI), and LOF [2], [9], [13], [15], [22]. The LOF method gained widespread popularity as a density-based technique for outlier detection [2]. LOF operates by assessing the local density ratio surrounding an object against the average density achievable from neighboring objects. The object’s surroundings are defined by the user-provided minimum *k*-neighbor parameters and the closest neighbor distance [15], [23]. Given the variations in outlier detection techniques outlined earlier, it is crucial to select the appropriate approach to optimize the performance of a data mining algorithm.

TABLE I
COLLECTION OF EXPERIMENTAL DATASETS

No	Dataset Name	Number of Instances	Attributes and Labels
1.	Wisconsin Breast Cancer	699	9 attributes + 1 label
2.	Glass	214	10 attributes + 1 label
3.	Harbeman	306	3 attributes + 1 label
4.	Lymphography	148	18 attributes + 1 label
5.	Parkinson	195	22 attributes + 1 label

III. METHODOLOGY

The study employed experimental techniques involving stages for collecting datasets, applying the proposed methods (KNN-distance, KNN-density, and KNN-LOF), and assessing the experimental results.

A. DATASET COLLECTION

This study utilized a dataset containing outliers. The dataset is publicly available for download from both the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php) and Kaggle (https://www.kaggle.com/) [3], [13].

Table I illustrates the specifics of the five datasets utilized in the experimental application of the proposed method. The datasets are Wisconsin Breast Cancer, Glass, Haberman, Lymphography, and Parkinson.

B. EXPERIMENTATION OF THE PROPOSED METHOD

The proposed method enhances the KNN algorithm’s performance by optimizing it through outlier analysis, including KNN-distance, KNN-density, and KNN-LOF, conducted with RapidMiner. Figure 1 illustrates the experimental procedures outlined in the proposed method.

1) DATASET INPUT

The initial action in this experiment involved importing datasets acquired beforehand through the data collection process.

2) OUTLIER ANALYSIS

The applied outlier detection techniques included KNN-distance, KNN-density, and KNN-LOF. Outlier detection is included in the process of data preprocessing. In this stage, the goal of outlier detection is to identify datasets containing outliers, thereby enabling the identification of outliers within the dataset.

3) COMPARING THE ACCURACY OF THE PROPOSED METHODS

The outlier analysis method yields modeling results presented as average accuracy. These average accuracy results were then compared to their significance levels using the Friedman and Nemenyi tests. The Friedman test, as proposed by Demsar [25], is a nonparametric analysis for conducting two-way variation analysis based on ratings. In this study, Friedman’s test was utilized to compare the effectiveness of the proposed methods, utilizing either chi-square, F-distribution, or P-value.

Figure 2 illustrates the performance comparison procedure of the proposed method, including KNN-distance, KNN-density, and KNN-LOF. The first step is to prepare observational data from the experimental results and perform a ranking. The hypotheses set out in the Friedman test in this study were as follows.

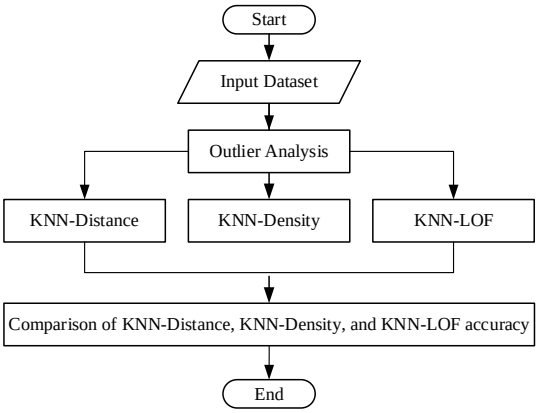

Figure 1. Steps of Experimentation

- H_0 (null hypothesis): the proposed methods employed in this study's experiment show no variation in average accuracy values.
- H_a (alternative hypothesis): the methods employed in the experiments of this study exhibit variations in average accuracy values.

The significance level value (α), also known as the error rate, was established to guide decision-making in hypothesis testing. The significance level values that could be used were 0.05 (5%) and 0.1 (10%). A lower value indicates a higher level of confidence in decision-making.

The next procedure was to calculate Friedman's test statistics. In this study, the Friedman test statistical calculation was conducted utilizing both chi-square and F-distribution.

The chi-square test is a type of nonparametric comparative test used to analyze two variables with nominal data scales [26], [27]. Equation (1) was employed in Friedman's test based on chi-square.

$$X_f^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (1)$$

here, r_i^j represents the jth rank of the k method in the ith dataset out of a total of N datasets with degrees of freedom (DF) (2).

$$DF = (k-1). \quad (2)$$

The hypothesis decision involves comparing the values of chi-square count (X_f^2) with the chi-square table ($X_{\alpha(k-1)}^2$), denoted as follows:

- If $X_f^2 < X_{\alpha(k-1)}^2$, then H_0 was accepted and H_a was rejected.
- If $X_f^2 > X_{\alpha(k-1)}^2$, then H_0 was rejected and H_a was accepted.

Equation (3) was employed in Friedman's test based on F-Distribution (F_f).

$$F_f = \frac{(N-1)X_f^2}{N(k-1)-X_f^2}. \quad (3)$$

Equations (4) and (5) were employed to calculate DF in the Friedman test using the F-distribution(F_f).

$$DF1 = (k-1) \quad (4)$$

$$DF2 = (k-1)(N-1). \quad (5)$$

The hypothesis decision involves comparing the values of F-distribution (F_f) with those in the F-distribution table ($F_{\alpha(k-1),(k-1)(N-1)}$), represented as follows.

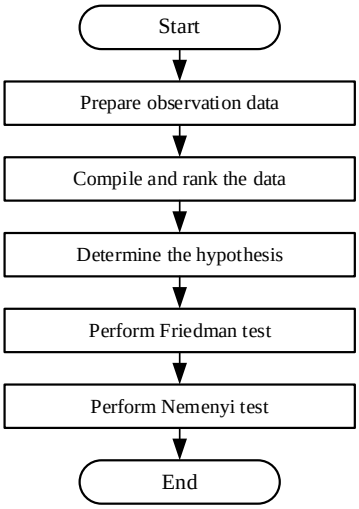

Figure 2. Procedures for comparing the performance of the proposed methods.

- If $F_f > F_{\alpha(k-1),(k-1)(N-1)}$, then H_0 was rejected and H_a was accepted.
- If $F_f < F_{\alpha(k-1),(k-1)(N-1)}$, then H_0 was accepted, and H_a was rejected.

If the H_0 was rejected and the H_a was accepted, the analysis proceeded with the Nemenyi test to identify which methods, when compared in pairs, exhibited significant differences in this study. The Nemenyi test's statistical calculation was performed using the critical difference (CD) value as outlined in (CD) [25] (6). Two or more methods can be considered significantly different if their average rating value exceeds the critical difference (CD).

$$CD = q_\alpha\sqrt{\frac{K(K+1)}{6D}} \quad (6)$$

here, q_α represents the critical value chosen based on the significance level value. K represents the number of methods being compared, and D represents the number of datasets used in each proposed method's experiment.

C. EVALUATION OF EXPERIMENTAL RESULTS

In the evaluation stage, the experimental results were assessed by comparing the accuracy of each proposed method. Conclusions were further drawn from the conducted research.

IV. RESULTS AND DISCUSSION

A. RESULTS

The experiment conducted in this study was to optimize the KNN algorithm based on an outlier analysis comparison. The experiments for the proposed KNN-distance, KNN-density, and KNN-LOF methods were carried out using RapidMiner. The evaluation results of the three proposed methods were in the form of accuracy using a confusion matrix and a 10-fold cross-validation method. The proposed method's average accuracy results across each dataset were compared using both the Friedman and Nemenyi tests to assess their significance levels.

1) KNN-DISTANCE EXPERIMENT

The distance-based outlier detection operator at RapidMiner identified n outliers within the dataset based on the kth distance of its nearest neighbor [28]. The operator searched for outliers using the outlier detection approach recommended in previous studies [29]. The study suggested a method for

TABLE II
NEIGHBOR VALUE (K) AND NUMBER OF OUTLIERS (N)

No.	Dataset Name	Neighbors (*k*)	Outlier (*n*)
1.	Wisconsin Breast Cancer	5	30
2.	Glass	7	12
3.	Haberman	5	30
4.	Lymphography	5	10
5.	Parkinson	3	15

distance-based outlier formulation based on the distance between a point and its *k*th nearest neighbor. Each point was ranked by its distance to its *k*th nearest neighbor. The top *n* points in this rating were identified as outliers [28]. The values of *k* and *n* may be determined depending on the number of neighbors and outlier parameters.

Outlier detection (distances) may be set based on the number of *k* neighbor parameters and the number of *n* outliers by selecting the most appropriate parameters [30], [31]. The selection of each parameter was determined using the trial-and-error method, which involved trying neighboring *k* parameter values of odd numbers (3, 5, and 7) one by one. The value yielding the highest accuracy was identified among the neighboring *k*-values [32]. The number of outliers to be selected should be adjusted based on the number of instances in each dataset. Outlier This implies that the number of selected outliers should not exceed the total number of instances in the dataset during each search. Table II summarizes the number of *k*-neighbors and *n* outliers with high accuracy for each dataset based on trial-and-error results.

Table II shows the selection of *k*-neighbor values and the number of *n* outliers. For instance, when employing KNN-distance with *k*=5 and *n=30* in the Wisconsin Breast Cancer dataset, it achieved the highest performance.

The implementation of the KNN-distance experimental model involved various operations, such as Detect Outlier (distances), Filter Examples, Split Data, Multiply, and Cross-Validation. The Filter Examples operator is designed to remove identified outliers. It accomplished this task by setting the filter parameters that had been added to the false condition. The filter parameters set on this operator were outliers. The Split Data operator generates the exact number of subsets needed from the dataset. The dataset was divided into training and testing data, with a ratio of 90% for training data and 10% for testing data [33]. The model's performance was validated using the cross-validation operator, employing a 10-fold parameter. The way 10-fold cross-validation worked by dividing the dataset into mutually independent 10-fold sets: f_1, f_2, …., f_{10}, with each fold contained one-tenth of the dataset. Furthermore, there were ten sets of datasets: D_1, D_2, …., D_{10} each contained nine folds as practice data and one-fold for testing purposes. Each fold became a one-time test data [34]. During the cross-validation process, there were two pages: the training and the testing pages. The training page was used as an application of the KNN algorithm model. The selection of the *k* value was typically subjective, with a recommended preference for odd values [32], [35]. In this experiment, *k* values of 3, 5, and 7 were utilized due to their higher accuracy compared to other odd *k* values. The test page included the 'apply model' and 'performance (classification)' operators, which assessed the algorithm's performance on each dataset. In the KNN-distance experiment, accuracy was the performance measure selected for evaluation. Table III displays the average accuracy obtained from the KNN-distance experiment, employing a 10-fold cross-validation model for performance evaluation.

TABLE III
AVERAGE ACCURACY RESULTS OF KKN-DISTANCE

No.	Dataset Name	KNN-Distance		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	96.08%	96.14%	95.99%
2.	Glass	82.68%	84.57%	84.79%
3.	Haberman	72.71%	73.94%	75.33%
4.	Lymphography	83.93%	83.17%	84.19%
5.	Parkinson	90.95%	89.67%	88.17%

Table III provides a summary of the average accuracy achieved in the KNN-distance experiment, obtained through 10-fold cross-validation with k values of 3, 5, and 7. Among the datasets, Wisconsin Breast Cancer achieved the highest accuracy at 96.14% with k=5, whereas Haberman yielded the lowest average accuracy of 72.71% with k=7.

2) KNN-DENSITY EXPERIMENT

The density-based outlier detection tool in RapidMiner identifies outliers within the dataset by analyzing data density. Objects situated at a distance farther than *D* and had, at least, a *p* proportion of all objects are classified as outliers [28]. The KNN-density experimental model was applied using various operators, such as Detect Outlier (densities), Filter Example, Split Data, Multiply, and Cross-Validation.

In outlier detection based on densities, suitable parameters for the distance parameter D and the proportion p are determined by searching for appropriate values [31]. The trial-and-error method is used to determine the selection of each parameter, utilizing odd values such as 3, 5, 7, and 9, along with adjusting the proportion of p within the range of 0.1 to 0.9. The *k* values used in this experiment were *k*=3, *k*=5, and *k*=7, as they exhibited higher accuracy compared to other odd *k* values. The results of the parameter values selected in the experiment were obtained based on trial-and-error experimentation with distance and proportion parameters, as illustrated in Table IV. The values of distance and proportion parameters in Table IV are the values of distance and proportion parameters yielding the highest accuracy in each dataset.

The utilization of Filter Examples, Split data, and cross-validation operators does not deviate from the experimental implementation of KNN-distance discussed earlier. Accuracy was the chosen performance metric for the KNN-density experiment. Table V illustrates the average accuracy obtained from this experiment, utilizing a 10-fold cross-validation model performance validation.

Table V showcases the average accuracy achieved by KNN-density through 10-fold cross-validation across all datasets, employing *k* values of 3, 5, and 7. In the case of Wisconsin Breast Cancer, the highest average accuracy of 99.34% was achieved using k values of 3 and 5. Conversely, Haberman demonstrated the lowest average accuracy of 70.79% when employing a *k* value of 3.

3) KNN-LOF EXPERIMENT

The LOF method operates on the principle of local density. The locality is defined by the *k* nearest neighbors whose distances are employed to gauge density. Outliers are determined by comparing an object's local density to that of its neighboring areas, identifying regions with similar densities to their neighbors and points with notably lower densities compared to their surroundings [28], [36].

Jurnal Nasional Teknik Elektro dan Teknologi Informasi
Volume 13 Number 2 May 2024

TABLE IV
DISTANCE AND PROPORTION PARAMETER VALUES

No.	Dataset Name	Distance (*D*)	Proportion (*p*)
1.	Wisconsin Breast Cancer	0.3	0.6
2.	Glass	0.3	0.7
3.	Haberman	0.3	0.8
4.	Lymphography	0.7	0.8
5.	Parkinson	0.9	0.8

TABLE V
AVERAGE ACCURACY RESULTS OF KKN-DENSITY

No.	Dataset Name	KNN-Density		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	99.34%	99.34%	98.92%
2.	Glass	82.02%	85.14%	85.25%
3.	Haberman	70.79%	71.87%	73.08%
4.	Lymphography	83.13%	85.45%	82.18%
5.	Parkinson	89.64%	89.80%	89.85%

Various operators were employed in conducting the KNN-LOF experiment, including Detect Outlier (LOF), Filter Examples, Split Data, Multiply, and Cross-Validation. The lower threshold parameter and upper threshold of the minimum point in RapidMiner's outlier detector operator (LOF) were both set to 10 and 20 [37], respectively and applied uniformly across all datasets. The experiment involving the application of KNN-distance and KNN-density discussed earlier was replicated through the utilization of Filter Examples, Split Data, and Cross-Validation operators. The selected performance measure used was accuracy, consistent with the KNN-distance and KNN-density experiments. Table VI displays the average accuracy value derived from the KNN-LOF experiment employing the 10-fold cross-validation model for performance validation.

A summary of the average results of KNN-LOF accuracy, generated through 10-fold cross-validation with values of k=3, k=5, and k=7, is presented in Table VI. Among all the datasets, Wisconsin Breast Cancer excelled with k=3, achieving an average accuracy of 93.65%, whereas Haberman exhibited the lowest average accuracy with k=3, standing at 67.50%.

D. DISCUSSION

In the experiment, outlier analysis was employed to remove outliers from the dataset. The results obtained after detecting and removing outliers using outlier analysis techniques (distance, density, and LOF) are presented in Table VII.

The number of outliers detected in each dataset is presented in Table VII. Subsequently, any detected outliers were automatically removed. 290 outliers were detected in the Wisconsin Breast Cancer dataset by KNN-density, marking the largest number of outliers identified. No outliers were detected on the Glass dataset by KNN-density, therefore no data deletion occurred. In Wisconsin Breast Cancer and Haberman datasets, the most outliers, totaling 30, were detected by KNN-distance, whereas the least number of outliers, ten in total, was detected by Lymphography among the other datasets. Meanwhile, only lower threshold and upper threshold values were displayed by KNN-LOF. The highest LOF value was regarded as an outlier, meaning that any values surpassing an upper threshold were automatically classified as outliers. In the Wisconsin Breast Cancer dataset, KNN-LOF identified the highest number of outliers, setting the upper threshold value at 22.137, whereas in the Lymphography dataset, it detected the fewest outliers, with an upper threshold value of 1.935.

TABLE VI
AVERAGE ACCURACY RESULT OF KNN-LOF

No.	Dataset Name	KNN-LOF		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	93.65%	93.39%	93.41%
2.	Glass	81.27%	84.03%	82.02%
3.	Haberman	67.50%	69.76%	70.19%
4.	Lymphography	77.22%	74.74%	77.53%
5.	Parkinson	89.54%	88.75%	87.78%

TABLE VII
DETECTED OUTLIERS

No.	Dataset Name	Number of Instances	Detected Outliers			
			Density	*Distance*	*LOF*	
					Min	*Max*
1.	Wisconsin Breast Cancer	699	290	30	0	22.137
2.	Glass	214	0	12	0.926	2.617
3.	Haberman	306	7	30	0.931	3.301
3.	Lymphography	148	7	10	0.916	1.935
4.	Parkinson	195	4	15	0.900	3.255

1) COMPARISON OF THE PERFORMANCE OF THE PROPOSED METHODS

The results of the performance evaluation of the proposed method using 10-fold cross-validation were compared to the Friedman and Nemenyi tests. Below are the steps for comparing the performance of the proposed methods.

First, the preparation of observation data was undertaken. The utilized observation data was the average accuracy result data of each dataset for every proposed method are presented in Tables III, V, and VI. Furthermore, ranking and observation were carried out by sorting the average accuracy values of each dataset from every proposed method. The proposed method's performance, starting with the highest accuracy, was assigned a rating of 1, followed by the method with the second highest accuracy receiving a rating of 2, and so forth. If the same level of accuracy was encountered, the utilized rating was the average rating (RANK.AVG).

In Table VIII, observation data from the Friedman test is presented, showing the ranking of the average accuracy value of each dataset for each proposed method. In Table V, Wisconsin Breast Cancer exhibits the highest accuracy value under the KNN-density method, whether with k=3 or k=5, both yielding the same value of 99.34%. It was assumed that the ranks were 1 and 2, resulting in an average rank of 1.5. For example, the accuracy of the Glass dataset depicted in Table V and Table VI remains consistent, achieving 82.02%, whether utilizing the KNN-density method with k=3 or the KNN-LOF method with k=7. It was assumed that the ranks were 7 and 8, resulting in an average rating of 7.5.

The next step was to determine the significance level value (α). In this study, the chosen α values were 0.05 (5%) and 0.1 (10%). The use of two different significance levels aimed to explore various hypothetical decision-making scenarios.

The hypothesis was subsequently established. The hypotheses tested in this study are as follows:

H_0: KNN-density = KNN-distance = KNN-LOF.

H_a: KNN-density ≠ KNN-distance ≠ KNN-LOF.

TABLE VIII
FRIEDMAN TEST OBSERVATION DATA

No.	Dataset Name	KNN-Density			KNN-Distance			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	1.5	1.5	3	5	4	6	7	9	8
2.	Glass	7.5	2	1	6	4	3	9	5	7.5
3.	Haberman	6	5	3	4	2	1	9	8	7
4.	Lymphography	5	1	6	3	4	2	8	9	7
5.	Parkinson	5	3	2	1	4	8	6	7	9
Total		25	12.5	15	19	18	20	39	38	38.5
Mean of rank		5	2.5	3	3.8	3.6	4	7.8	7.6	7.7
		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9

The null hypothesis, H_0suggested that there was no variance in average accuracy values across the KNN-distance, KNN-density, and KNN-LOF methods. Meanwhile, the H_a hypothesis suggested that there was a divergence in the average accuracy across the KNN-density, KNN-distance, and KNN-LOF methods.

The next step involved calculating the Friedman test statistics after the hypothesis was determined. The calculation of Friedman's test, using chi-square as a basis, began by determining the degrees of freedom (DF) value as described in (2), followed by the calculation of the chi-square count value as outlined in (1), with reference to chi-square tables.

$$DF = (k-1) = (9-1) = 8$$

$$X_f^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

$$X_f^2 = \frac{12*5}{9(9+1)}\left[5^2 + 2.5^2 + 3^2 + 3.8^2 + 3.6^2 + 4^2 + 7.8^2 + 7.6^2 + 7.7^2 - \frac{9(9+1)^2}{4}\right]$$

$$X_f^2 = 0.66667[261.54 - 225]$$

$$X_f^2 = 24.36.$$

Chi-square table ($X^2_{a(k-1)}$) was determined using the CHIINV(α;DF) Excel equation. For the significance level value of 5%, the resulting table chi-squared was CHIINV (0.05;DF) = CHIINV (0.05;8) 15.5073, = 10%, while for the significance level value of (0,1; the resulting table chi-squared was CHIINV(0.1;*DF*) = CHIINV(0.1;8) = 13.36157.

From the calculation results, the chi-squared value of the calculation was greater than the chi-squared value of the table, both in the use of the significance level of 5% and 10%. Therefore, the hypothesis decision is $X_f^2 > X^2_{a(k-1)}$ or H_0 is rejected and H_a is accepted. This implies that variances are present among the KNN-distance, KNN-density, and KNN-LOF methods

Friedman's test based on F-distribution could be calculated by starting with the calculation of DF1 values such as (4) and *DF* as in *(5)*, then looking for F-distribution values (F_f) and F-distribution tables $(F_{\alpha(k-1),(k-1)(N-1)})$.

$$DF1 = (k-1) = (9-1) = 8$$

$$DF2 = (k-1)(N-1) = (9-1)(5-1) = 32$$

$$F_f = \frac{(N-1)X_f^2}{N(k-1) - X_f^2}$$

$$F_f = \frac{(5-1)*24.36}{5(9-1) - 24.36}$$

$$F_f = \frac{97.44}{15.64}$$

$$F_f = 6.23018.$$

The F-distribution table value was derived through the utilization of the Microsoft Excel formula, specifically employing FINV(α;(*DF1*);(*DF*2). Therefore, the F-distribution table for the 5% significance level value was $F_{a(k-1)(N-1)}$ = FINV(0.05;8;32) = 2.2444, while the *F*-distribution table for the 10% significance level value was $F_{a(k-1)(N-1)}$ = FINV(0.1;8;32) = 1.8701.

The results of the calculation show that when both the 5% and 10% significance levels were employed, the F-distribution exceeded the values in the F-distribution table. Therefore, the hypothesis testing result is $F_f > F_{\alpha(k-1),(k-1)(N-1)}$ or H_0 is rejected and H_a is accepted. This implies that distinctions exist among the KNN-distance, KNN-density, and KNN-LOF methods.

The hypothesis results were obtained by using Friedman test statistics employing chi-square and F-distribution, which led to the rejection of H_0 and the acceptance of H_a . Thus, it is concluded that significant differences exist between the proposed methods being compared. Moreover, the Nemenyi test can be conducted to determine the pairs of proposed methods that exhibit the most significant differences. Below, the stages of the Nemenyi test carried out in this study are presented.

The first step was to prepare the observation data that have been presented in Table VIII. The observation data in the table corresponded to those used in the Friedman test. The subsequent procedure involved computing the *CD* value, as indicated in (6), using the chosen critical value (q_a). The critical values utilized in the Nemenyi test are displayed in Table IX.

The utilization of critical values applicable in the Nemenyi test is depicted in Table IX. In this study, the critical values of significance levels (α) 0.05 and 0.1 were utilized in classifier 9, namely 3.102 and 2.855. Classifier 9 was selected due to the comparison of nine accuracy values as shown in Table VII. The CD value according to the obtained significance level is as follows.

$$CD = q_{0.05}\sqrt{\frac{K(K+1)}{6D}} = 3.102\sqrt{\frac{9(9+1)}{6*5}} = 5.3728$$

$$CD = q_{0.1}\sqrt{\frac{K(K+1)}{6D}} = 2.855\sqrt{\frac{9(9+1)}{6*5}} = 4.9450.$$

The next step involved calculating the differences in mean rank between the two methods being compared. In other words, the Nemenyi test displays a pairwise comparison of the proposed method. If the disparity in the mean rank between the two compared methods exceeds the resultant *CD* value, the method is considered significantly different. Table VIII showcases the mean of rank (R_1 to R_9) For example, a difference in mean rank of 2.5 between KNN-density (*k*=3) and

Jurnal Nasional Teknik Elektro dan Teknologi Informasi
Volume 13 Number 2 May 2024

TABLE IX
CRITICAL VALUES FOR NEMENYI TEST

Classifiers	$q_{0,05}$ (5%)	$q_{0,1}$ (10%)
2	1.960	1.645
3	2.343	2.052
4	2.569	2.291
5	2.728	2.459
6	2.850	2.589
7	2.949	2.693
8	3.031	2.780
9	3.102	2.855
10	3.164	2.920

TABLE X
NEMENYI TEST OBSERVATION DATA

		KNN-Density			KNN-Distance			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-density	*K=3*	0	2.5	2	1.2	1.4	1	2.8	2.6	2.7
	K=5	2.5	0	0.5	1.3	1.1	1.5	5.3	5.1	5.2
	K=7	2	0.5	0	0.8	0.6	1	4.8	4.6	4.7
KNN-distance	*K=3*	1.2	1.3	0.8	0	0.2	0.2	4	3.8	3.9
	K=5	1.4	1.1	0.6	0.2	0	0.4	4.2	4	4.1
	K=7	1	1.5	1	0.2	0.4	0	3.8	3.6	3.7
KNN-LOF	*K=3*	2.8	5.3	4.8	4	4.2	3.8	0	0.2	0.1
	K=5	2.6	5.1	4.6	3.8	4	3.6	0.2	0	0.1
	K=7	2.7	5.2	4.7	39	4.1	3.7	0.1	0.1	0
Critical Difference ($q_{0.05}$) = 5.3728 Critical Difference ($q_{0.1}$) = 4.9450										

TABLE XI
NEMENYI TEST RESULTS ($\alpha = 0.05$)

		KNN-Density			KNN-Distance			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-density	*K=3*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-distance	*K=3*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-LOF	*K=3*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*

TABLE XII
NEMENYI TEST RESULTS ($\alpha = 0.1$)

		KNN-Density			KNN-Distance			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-density	*K=3*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-distance	*K=3*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-LOF	*K=3*	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*

KNN-density (k=5) was observed, derived from the mean difference of rank R_1 (5) compared to the mean rank of R_2 (2.5). The overall data from Nemenyi test observations based on the difference in mean of rank are shown in Table X.

The comparison between the methods is illustrated in Table X, with the differences in the mean of rank of each method against the others being highlighted. For instance, when comparing KNN-density with k=3 against itself, the result was 0. When comparing KNN-density with k=3 against k=5, the difference was 2.5. Similarly, comparing KNN-density with k=3 against k=7 yielded a difference of 2, and so forth.

The final step of the Nemenyi test involved comparing the results of the observational data in Table X with the CD obtained in the previous stage, in order to assess the significance of the differences between the proposed methods. The results of comparing the average rating of the observational data with the CD results indicate differences, where the values were recorded as "No" and "Yes." If the CD value exceeded the average rating value, then the conclusion was "No," indicating no significant distinction among the proposed methods (KNN-density = KNN-distance = KNN-LOF). If the CD value fell below the average rating value, the conclusion was "Yes," indicating a significant difference in the proposed method (KNN-density ≠ KNN-distance ≠ KNN-LOF). The results of the Nemenyi test comparison are displayed in Table XI for a significance level of 5% and Table XII for a significance level of 10%.

According to Table XI, it can be inferred that, at a significance level of 5%, no significant difference was observed between the proposed methods in the results of the Nemenyi test. Based on Table XII, it can be concluded that, at a significance level of 10%, significant differences were observed between KNN-density with k=5 and KNN-LOF with k=3, k=5, and k=7, as well as between KNN-LOF with k=3, k=5, and k=7 and KNN-density with k=5. Thus, it has been demonstrated that there exists a significant difference in the proposed method according to the results of the Nemenyi test conducted at a significance level of 10%.

V. CONCLUSION

The results indicate that the KNN-density method consistently achieves high average accuracy across three datasets: it attained an average accuracy of 99.34% for the Wisconsin Breast Cancer dataset at k values of 3 and 5, 85.25% accuracy for the Glass dataset at k = 7, and 85.45% accuracy for the Lymphography dataset at k =5. In the analysis conducted with Friedman's test at significance levels of 5% and 10%, it was observed that H_0 was rejected, and H_a was accepted. This suggests that there are distinctions among KNN-density, KNN-distance, and KNN-LOF. Moreover, according to the Nemenyi test conducted with a significance level of 5%, no significant difference was observed between the proposed methods. When a significance level of 10% was employed, it was demonstrated that a notable distinction existed between KNN-density and KNN-LOF. According to the average accuracy results, it can be inferred that the KNN-density approach effectively enhances the KNN algorithm through the identification and elimination of outliers using density-based outlier analysis. This contribution could serve as a response to the research problem and objectives outlined in this study.

CONFLICTS OF INTEREST

The author assures that during the composition of the scientific article titled "Optimization of the K-Nearest Neighbors Algorithm through Outlier Analysis Comparison (Distance, Density, LOF-Based)," the authors maintain impartiality and have no conflicts of interest with any involved parties.

AUTHORS' CONTRIBUTIONS

Conceptualization, Fitri Ayuning Tyas and Mahda Nurayuni; methodology, Mahda Nurayuni; software, Mahda Nurayuni; validation, Fitri Ayuning Tyas, Mahda Nurayuni, and Hidayatur Rakhmawati; original drafting, Fitri Ayuning Tyas, Mahda Nurayuni, Hidayatur Rakhmawati; reviewing and editing, Fitri Ayuning Tyas; visualization, Fitri Ayuning Tyas; supervision, Fitri Ayuning Tyas; funding, Fitri Ayuning Tyas.

ACKNOWLEDGMENTS

The author extends their gratitude to their fellow colleagues at the Computer Laboratory of STMIK Muhammadiyah Paguyangan Brebes, particularly those associated with the Information Systems Study Program, for their invaluable assistance in completing this research project.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[2] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cogn. Comput.*, vol. 5, no. 1, pp. 1–24, Mar. 2021, doi: 10.3390/bdcc5010001.

[3] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Heidelberg, Germany: Springer, 2011.

[4] I.H. Witten, E. Frank, and M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[5] C.C. Aggarwal, *Data Mining.* New York, NY, USA: Springer, 2015.

[6] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," *J. Syst. Softw.*, vol. 85, no. 5, pp. 1067–1024, May 2012, doi: 10.1016/j.jss.2011.12.019.

[7] D. Armiady, "Analisis Metode DBSCAN (density-based spatial clustering of application with noise) dalam mendeteksi data outlier," *JURIKOM (J. Ris. Komputer)*, vol. 9, no. 6, pp. 2158–2164, Dec. 2022, doi: 10.30865/jurikom.v9i6.5080.

[8] R. Silvi, "Analisis cluster dengan data outlier menggunakan centroid linkage dan k-means clustering untuk pengelompokan indikator HIV/AIDS di Indonesia," *J. Mat. MANTIK*, vol. 4, no. 1, pp. 22–31, May 2018, doi: 10.15642/mantik.2018.4.1.22-31.

[9] M.Y. Pusadan, "Outlier detection pada set data flight recording (pre-processing sumber data ADS-B)," *Seminar Nas. Teknol. Inf. Multimedia 2015*, 2015, pp. 2.1-31–2.1-36.

[10] J. Abellán, J.G. Castellano, and C.J. Mantas, "A new robust classifier on noise domains: Bagging of credal C4.5 trees," *Complexity*, vol. 2017, pp. 1–17, Dec. 2017, Art. no. 9023970, doi: 10.1155/2017/9023970.

[11] A. Duraj and P.S. Szczepaniak, "Outlier detection in data streams — A comparative study of selected methods," *Procedia Comput. Sci.*, vol. 192, pp. 2769–2778, Oct. 2021, doi: 10.1016/j.procs.2021.09.047.

[12] S. Sugidamayatno and D. Lelono, "Outlier detection credit card transactions using local outlier factor algorithm (LOF)," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 13, no. 4, pp. 409–420, Oct. 2019, doi: 10.22146/ijccs.46561.

[13] X. Xu, H. Liu, L. Li, and M. Yao, "A comparison of outlier detection techniques for high-dimensional data," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, pp. 652–662, Jan. 2018, doi: 10.2991/ijcis.11.1.50.

[14] T. Sangeetha and G. Mary A., "A fuzzy proximity relation approach for outlier detection in the mixed dataset by using rough entropy-based weighted density method," *Soft Comput. Lett.*, vol. 3, pp. 1–12, Dec. 2021, doi: 10.1016/j.socl.2021.100027.

[15] H. Xu, L. Zhang, P. Li, and F. Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor," *J. Algorithms Comput. Technol.*, vol. 16, pp. 1–12, Mar. 2022, doi: 10.1177/17483026221078111.

[16] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[17] Z. Deng *et al.*, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, Jun. 2016, doi: 10.1016/j.neucom.2015.08.112.

[18] S. Zhang *et al.*, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[19] J. Ning, L. Chen, C. Zhou, and Y. Wen, "Parameter k search strategy in outlier detection," *Pattern Recognit. Lett.*, vol. 112, pp. 56–62, Sep. 2018, doi: 10.1016/j.patrec.2018.06.007.

[20] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook.* New York, NY, USA: Springer, 2010.

[21] P.A. Ariawan, "Optimasi pengelompokan data pada ketode k-means dengan analisis outlier," *J. Nas. Teknol. Sist. Inf.*, vol. 5, no. 2, pp. 88–95, Aug. 2019, doi: 10.25077/TEKNOSI.v5i2.2019.88-95.

[22] H.C. Mandhare and S.R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," *2017 Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, 2017, pp. 931–935, doi: 10.1109/ICCONS.2017.8250601.

[23] J. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognit.*, vol. 115, pp. 1–11, Jul. 2021, doi: 10.1016/j.patcog.2021.107874.

[24] H. Wang, M.J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, Aug. 2019, doi: 10.1109/ACCESS.2019.2932769.

[25] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

[26] I.C. Negara and A. Prabowo, "Penggunaan uji chi–square untuk mengetahui pengaruh tingkat pendidikan dan umur terhadap pengetahuan penasun mengenai HIV–AIDS di Provinsi DKI Jakarta," *Pros. Senamantra (Seminar Nas. Mat. Terapannya)*, 2018, pp. 1–8.

[27] L.F. Obe, D. Lalang, V. Lakapeni, and D. Fatin, "Pengaruh jumlah anak terhadap pendapatan hasil perkebunan kemiri di Desa Maikang Kecamatan Alor Selatan tahun 2020 menggunakan metode chi kuadrat," *J. Ilm. Wahana Pendidikan*, vol. 7, no. 6, pp. 378–384, Oct. 2021, doi: 10.5281/zenodo.5644452.

[28] F. Akthar and C. Hahne, *RapidMiner 5 Operator Reference*. 2012.

[29] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data*, 2000, pp. 427–438, doi: 10.1145/342009.335437.

[30] Z.A Bakar, R. Mohemad, A. Ahmad, and M.M. Deris, "A comparative study for outlier detection techniques in data mining," *2006 IEEE Conf. Cybern. Intell. Syst.*, 2006, pp. 1–6, doi: 10.1109/ICCIS.2006.252287.

[31] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, Jun. 2017, doi: 10.1016/j.neucom.2017.02.039.

[32] D. Kartini *et al.*, "Perbandingan nilai k pada klasifikasi pneumonia anak balita," *J. Komputasi*, vol. 10, no. 1, pp. 47–53, Apr. 2022, doi: 10.23960%2Fkomputasi.v10i1.2965.

[33] R.M. Candra and A.N Rozana, "Klasifikasi komentar bullying pada Instagram menggunakan metode k-nearest neighbor," *IT J. Res. Dev.*, vol. 5, no. 1, pp. 45–52, Jul. 2020, doi: 10.25299/itjrd.2020.vol5(1).4962.

[34] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[35] M. Rivki and A.M. Bachtiar, "Implementasi algoritma k-nearest neighbor dalam pengklasifikasian follower Twitter yang menggunakan bahasa Indonesia," *J. Sist. Inf. (J. Inf. Syst.)*, vol. 13, no. 1, pp. 31–37, Apr. 2017, doi: 10.21609/jsi.v13i1.500.

[36] A. Mahendra, "Pentapisan dan deteksi data outlier dalam proses sistem akusisi data pada proses sintering," *Arsitron*, vol. 6, no. 1, pp. 1–7, Jun. 2015.

[37] D. Handriyadi, M.A. Bijaksana, and E.B. Setiawan, "Analisis perbandingan clustering-based, distance-based dan density-based dalam mendeteksi outlier," *Seminar Nas. Apl. Teknol. Inf. (SNATI)*, 2009, pp. 101–108.