International Journal of Advances in Intelligent Informatics
Vol. 5, No. 3, November 2019, pp. 285-296
ISSN 2442-6571

Anomaly detection on flight route using similarity and grouping approach based-on automatic dependent surveillance-broadcast

Mohammad Yazdi Pusadan [a,b,1,*], Joko Lianto Buliali [a,2], Raden Venantius Hari Ginardi [a,3]

[a] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b] Department of Informatics, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia
[1] yazdi.diyanara@untad.ac.id; [2] joko@cs.its.ac.id; [3] hari@its.ac.id
* corresponding author

ARTICLE INFO

Article history
Received June 26, 2018
Revised July 28, 2018
Accepted December 20, 2018
Available online November 30, 2019

Keywords
Segment
Log-likelihood ratio
Grouping similarity
Accuracy
Anomaly detection

ABSTRACT

Flight anomaly detection is used to determine the abnormal state data on the flight route. This study focused on two groups: general aviation habits (C1) and anomalies (C2). Groups C1 and C2 are obtained through similarity test with references. The methods used are: 1) normalizing the training data form, 2) forming the training segment 3) calculating the log-likelihood value and determining the maximum log-likelihood (C1) and minimum log-likelihood (C2) values, 4) determining the percentage of data based on criteria C1 and C2 by grouping SVM, KNN, and K-means and 5) Testing with log-likelihood ratio. The results achieved in each segment are Log-likelihood value in C1Latitude is -15.97 and C1Longitude is -16.97. On the other hand, Log-likelihood value in C2Latitude is -19.3 (maximum) and -20.3 (minimum), and log-likelihood value in C2Longitude is -21.2 (maximum) and -24.8 (minimum). The largest percentage value in C1 is 96%, while the largest in C2 is 10%. Thus, the highest potential anomaly data is 10%, and the smallest is 3%. Also, there are performance tests based on F-measure to get accuracy and precision.

1. Introduction

Li and Hansman [1] investigated the anomaly detection with the cluster method to classify flights under a general pattern, found that flights outside the standard pattern were anomalous. Another advance development is an abnormal flight detection based on cluster analysis without knowing the norm standard [2]. The difference from the previous study lies in the type and airline that are not specified, and flight problems should not be known. These detection processes need experts to validate the results of anomaly detection analysis.

The anomaly detection required computational analysis. There are some data training in a specific month period as a modeling reference. Models formed based on custom analysis of flight patterns on an airline route. The clustering method K-Means cluster analysis in earthquake epicenter clustering by categorizing data based on properties is exact because the resulting model does not require ground truth as a referral or a justification from an expert. In this study, the number of k specified is two: a big group of standardized data and the small one as anomalous data [3]. They perform data analysis that has the same call sign of routes and airlines. The detector recognizes specific segmented locations as anomalies.

This study uses flight data based on Automatic Dependent Surveillance-Broadcasting (ADS-B) as data sources, is carefully observed in 1998. It consists of broadcast information about the position of the

aircraft (latitude, longitude), height (altitude), as well as related information about the aviation guidance recommended by the International Civil Aviation Organization (ICAO) [4]. In 2017, ADS-B studied aircraft monitoring from outer space. The result is a worldwide air traffic control and compares the accuracy of ADS-B information from the ground with those in space [5]. Furthermore, ADS-B utilization for an aviation anomaly detection study generated an algorithm on traffic and warning on aviation traffic [6] and the determination of areas for conflict detection on aviation traffic routes [7].

Several recent studies continue to be developed on classification methods using support vector machines (SVM). This method is more appropriate to solve binary classification problems with two classes in features that are linear at an interval [8]. Here, forming binary classes is a necessity for multi-class classification. The binary grouping techniques in SVM begin by looking for an optimized hyper hypothesis that distinguishes positive and negative samples [8][9]. Also, the SVM classification method has a robust training process. An implementation of a regression method is essential for better SVM performance[10]. The other classification method is k-nearest neighbors (k-NN). This method depends on a set of previously determined labels and classifiers. In the training data that is the closest distance to the classifier, several different groups will be formed from one another [11]. In the latest study, the k-NN classification can be applied to discrete and real data and produces better accuracy values. However, this lazy algorithm requires precise accuracy, one of which is in determining the closest (regional) distance [12] using the Euclidean distance [13].

Another grouping method is carried out as a comparison (i.e., K-means clustering). The developed model is better than the baseline if the results of the comparison between the classification model (supervised) and the clustering model (unsupervised) have minimal precision. For optimal results, the cluster method used is different from the classic cluster algorithm, which is without random centroid initialization [14], is done to reduce the iteration process. The k-means cluster initialization uses maximum and minimum values to obtain the optimal cluster results [15]. Another model used is the similarity model, used as a determinant of data with similar characteristics. The latest research similarity method occurred internally and externally [16]. The internal formed from pre-processed training data, while the external is a reference data set. The similarity method is log-likelihood, compare the log-likelihood value to obtain a measure of similarity (coefficient) [17]. The grouping method and classification can be done based on this coefficient. The next step is the similarity model testing, which divides the log-likelihood value of the test results with the log-likelihood value of the model. This method is called the log-likelihood ratio or exact log-likelihood ratio method [18].

This study determines the anomaly detection area on a flight route without having to know flight anomaly criteria based on the formed segment. This research also proposes a new testing method based on maximum and minimum similarity models of 30 days of trained data reference.

2. Method

Fig. 1 shows the framework in this study. This research consists of several stages, such as data collection, preprocessing, segment, training, and testing. The following sections describe these stages in more detail.

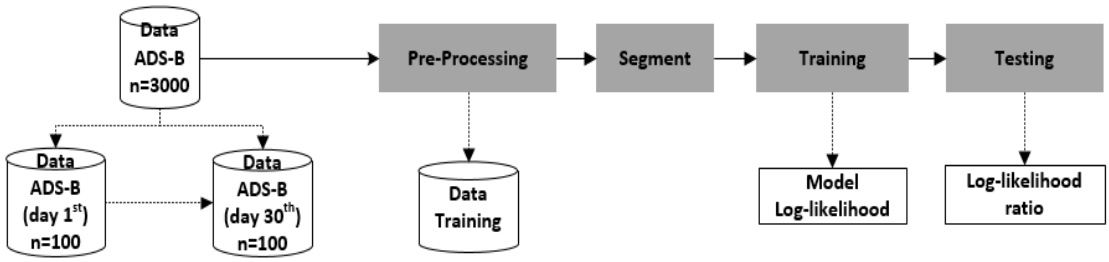

Fig. 1. Research Framework

2.1. Dataset

The data source used in this study is ADS-B, collected from the RTL 1090 radio device [19]. The data period is one month with 100 records in MySQL DBMS, used in the computational process. The parameters are date, latitude, longitude, and altitude. Table 1 shows a sample of ADS-B data for callsign LNI860 / SUB-PLW.

Table 1. ADS-B data

Date	Latitude	Longitude	Altitude
180414	-7.2823	112.2996	35000
180414	-7.2823	112.2996	35000
180414	-7.2823	112.2996	35000

The data with the DATE attribute contains a date, month, and year information — for example, the attribute DATE: 180414, the flight data on April 18, 2014. Furthermore, the attribute indicates the position of LATITUDE and LONGITUDE is a Cartesian coordinate (latitude, longitude). In general, the latitude value is negative (-), which denotes the longitude coordinates. Meanwhile, the value of longitude is positive (+) which denotes the latitude coordinates. The ALTITUDE attribute is data about the height of the plane as measured from the ground. For altitude units are feet or nautical miles, then on ALTITUDE = 35000 means 35000 feet = 5.76 nautical miles.

To get the data set in tabular format, first, the ADS-B data obtained is done by the feature extraction process. Data obtained from the process of broadcasting information from Register Transfer Level Software Defined Radio ADSB-B (RTL-SDR ADS-B) system. RTL-SDR ADS-B receives information from the ADS-B signal of the aircraft in the data format, as in Fig. 2.

```
##180414##225958.338015##
##180414##225951.639000##
8A03A2:A-:SJY580:0:5:0:2:2412:-7.2823:112.2996:8:0:F350:35000:
0:0:2425::0:0:0::80:77:0:257:464:467:792:0:0:0:370::1013:0::0:
:::0:0:0::0:12:0:7:17:1:1:1:0:0:0:341:25:-47:34:2:0:0:7774:4:0:34:
217148::2:33804::3166:10:4:32:64:30:22:16:72:1400680791:58:T-WARR6
```

Fig. 2. information from the ADS-B signal of the aircraft

Some bold numbers indicate the main attributes extracted to the DBMS (MySQL) for computational analysis.

2.2. Preprocessing

Pre-processing generated training data from in the form of training data derived from the mean of each record (rec-1 until rec-n, n = 100) ADS-B data (for 30 days). Fig. 3, shows the stages of pre-processing to form training data.

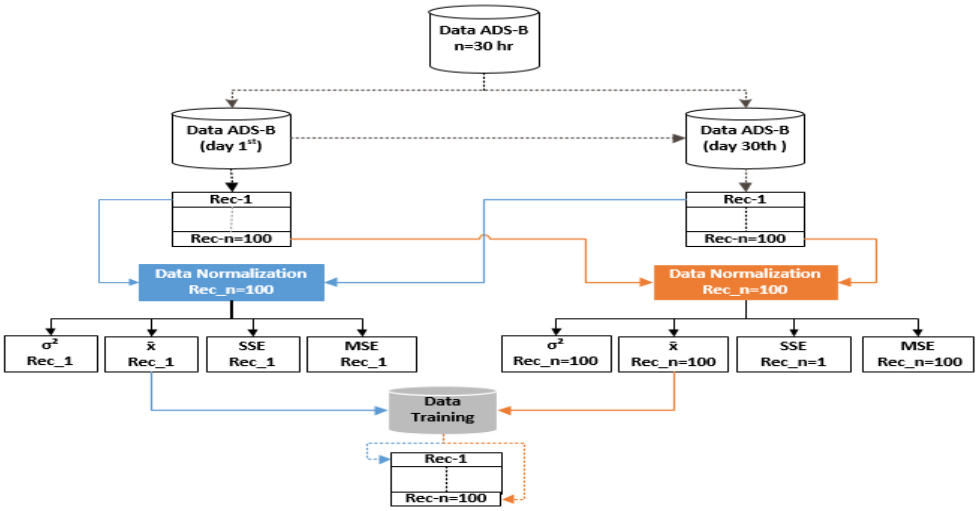

Fig. 3. Preprocessing

2.3. Segment

The segment forms data into groups based on the nature or habit of data records. Based on the experiment, the number of segments is 5. Each segment contains parameters of latitude and longitude with 20 records. Fig. 4 shows the process of training data segmentation. Table 2 illustrates five segments of the dataset with 20 records in each segment.

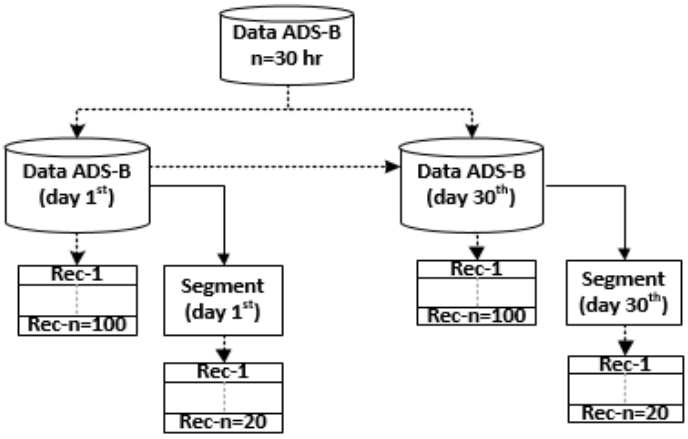

Fig. 4. The segmentation stages

Table 2. Five segments of the dataset

Date	Latitude	Longitude	Altitude	Record	Segment
180414	-7.2823	112.29964	35000	1	1
180414	-7.2823	112.29964	35000	2	
...	...	...	...	...	
180414	-7.28696	112.32369	35000	20	
180414	-7.28696	112.32369	35000	21	2
180414	-7.28696	112.32369	35000	22	
...	...	...	...	...	
180414	-7.28696	112.32369	35000	40	
180414	-7.28696	112.32369	35000	41	3
180414	-7.28696	112.32369	35000	42	
...	...	...	...	...	
180414	-7.30316	112.40718	35000	60	
180414	-7.30316	112.40718	35000	61	4
180414	-7.30316	112.40718	35000	62	
...	...	...	...	...	
180414	-7.30316	112.40718	35000	80	
180414	-7.30707	112.42721	35000	81	5
180414	-7.30316	112.40718	35000	82	
...	...	...	...	...	
180414	-7.30316	112.40718	35000	100	

2.4. Training and Models

Training determines a model. A similarity test (log-likelihood) for each segment determine the max and min values. These values are used as a classifier and early centroid for the grouping process. Fig. 5 shows the stages of modeling based on similarity test.

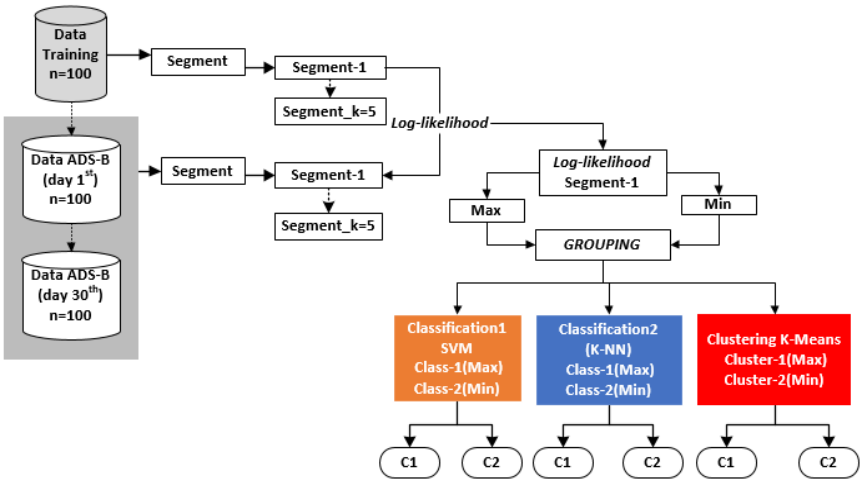

Fig. 5. Model creation through the similarity test

2.5. Testing

This stage tested the generated model. The similarity test compares the obtained log-likelihood value with the model log-likelihood value: C1 and C2 respectively. There are two log-likelihood ratios (LLR) values: LLR_C1 and LLR_C2. If LLR_C1 is less than LLR_C2 (LLR_C1 <LLR_C2), then the data test data is C1.Otherwise, it should be C2 Fig. 6 shows the testing stages between models.

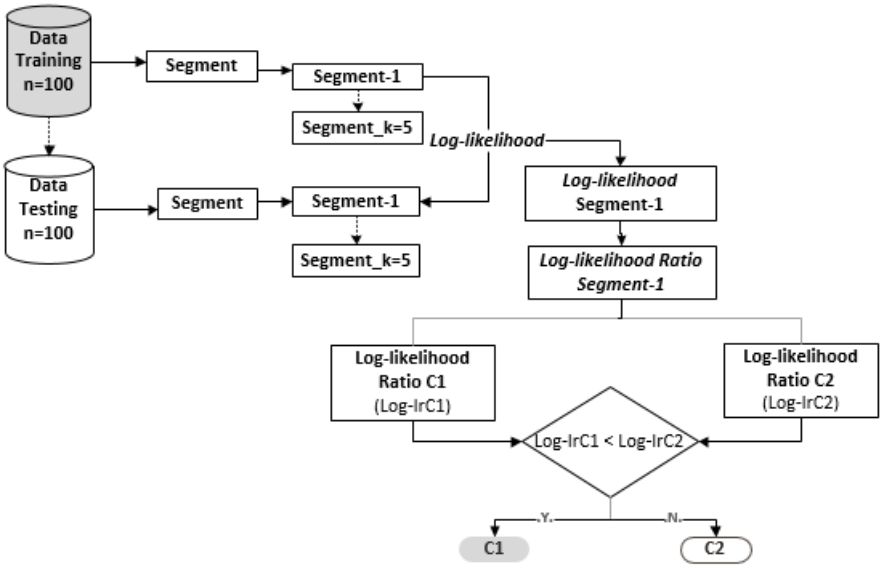

Fig. 6. Testing stages per segment

2.6. Performance Testing

The next stage uses a confusion matrix for computational accuracy measurements. The resulting indicator is the value of accuracy and precision. This measure is generated from several parameters, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). There are formulas (1) and (2) used to measure system performance based on precision and precision values.

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

$$precision = \frac{TP}{(TP + FP)} \quad (2)$$

2.7. Mean (μ), Standard Deviation (σ^2), Mean Square Error (MSE), and Sum Square Error (SSE)

This study used several methods: data normalization (pre-processing), grouping, and similarity test. The following equations express the data normalization [20].

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, k = 1,2,3,\ldots,l \quad (3)$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \mu_k)^2 \quad (4)$$

$$MSE = \frac{SSE}{n} \quad (5)$$

$$SSE = \sum_{i=1}^{N}(x_{ik} - \mu_k)^2 \quad (6)$$

The purpose of data in normalization is to obtain reference data for similarity modeling. The reference data is the average of some obtained training data in each record. Here, MSE and SSE are determined based on latitude and longitude. On the other hand, grouping divided data into two classifications and clustering. The classification used two methods: Support Vector Machine (SVM) and K-Nearest Neighborhood (K-NN). This research used the K-Means clustering method.

2.8. Support Vector Machine (SVM)

In some cases, the Support Vector Machine (SVM) [21] is suitable to handle the classification with an imbalanced dataset. Although the resulting sensitivity is less useful, the resulting accuracy is better. For example, Prahara *et al.* [22] proposed the use of SVM for the introduction of motor vehicle license plates based on edge detection. The process takes place to form an area that signifies the shape of the vehicle plate. A combination of Histogram of Oriented Gradients (HOG) + SVM is used to localize the number plate from a specific region with the candidate number plate extracted using the vertical edge density method to achieve reasonable accuracy. Latah and Toker [23] use of SVM for classifying the type Daniel of Service (DoS) attack, so that the user cannot access the computer network system. The resulting accuracy is 96.25% detected DoS attacks of type flooding.

SVM application of a data set is Z{(x_1,y_1),(x_2,y_2),...,(X_n,y_n)}. For $x_i \in R^n$ and $y_i \in \pm 1$, which is the label of the group and i=1, 2,...,n. The following equations formulate a set of SVM hyperplanes [23].

$$w \cdot x_i + b \geq +1; y_i = +1 \quad (7)$$

$$w \cdot x_i + b \leq 1; y_i = -1 \quad (8)$$

2.9. K-Nearest Neighborhood (K-NN)

K-Nearest Neighborhood is a supervised classification method characterized by a labeled class [24]. The application of K-NN is to classify signal speech in a noisy environment. The result is that there are two signal classifications in the speech feature. There are two groups, namely, the speech signal and the non-speech signal. Then, testing is done to determine which test data are in which group.

In this study, the accuracy was 80%. Bhattacharya *et al.* [25] proposed modifications of k-NN algorithm applied to fifteen numerical data sets from UCI data repository machine learning. Based on the 5-fold and 10-fold cross-validation, the average accuracy is better than using a typical cluster algorithm.

2.10. K-Means

K-Means clustering is used to define a set of data that resides in one group, where the distance between group members and the centroid is minimal [26]–[28]. The following formula describes the calculation centroid in a cluster.

$$C_{k,j} = \frac{x_{ikj} + x_{2kj} + ... + x_{akj}}{a}, j = 1,2,..., p \quad (9)$$

where $C_{k,j}$ is the centroid of group-k, variable-j, and a is the number of members in group k.

The general method is described in the following steps [3]: 1) determine the data be clustered, 2) apply K-means cluster analysis to the earthquake data, 3) compute the Krzanowski and Lai criteria for optimum K number of clusters, and 4) determine the optimum K number of cluster. In this study, we use the KL index to determine the optimum K. KL index is used to determine the optimum K number of clusters. Based on the KL index of the formed cluster, the most extensive KL index indicated that the amount of the cluster is the optimal number of clusters. There are 7 clusters with attributes latitude mean, longitude means, magnitude means, frequency means, and SSE.

2.11. Log-likelihood and Log-likelihood Ratio

The similarity approach is determined by the log-likelihood method for model similarity and log-likelihood ratio for testing. The likelihood principle is that if two experiments involve a model with θ parameter, give the same likelihood value, then the inference to θ must be the same, as explained in the following equations [29].

$$L(X;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \quad (10)$$

X is an observation matrix $x_1^T,...,x_n^T$ in every line

$$\ell(X;\theta) = \log L(X;\theta) \quad (11)$$

$$L(y,X;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\{-\frac{1}{2\sigma^2}(y_i - \beta^T x_i)\} \quad (12)$$

If it is in the model of $p(x_i,\theta)$, negative log-likelihood (LL) [30] is

$$\ln L(\theta) = -\sum_i y_i \ln(p(x_i,\theta)) + (1-y_i)\ln(1-p(x_i,\theta)) \quad (13)$$

For testing, the log-likelihood ratio test [31] is

$$L^2 = \max_{\substack{0 \le k_1 \le k_2 \le n \\ [\alpha n] \le k_2 \le [(1-\beta)n]}} -2\log \Lambda_{k_1,k_2} \quad (14)$$

where a likelihood ratio is

$$\lambda(X) = \frac{L_0^*}{L_1^*} \quad (13)$$

and Log-likelihood ratio is

$$-2\log\lambda = 2(\ell_1^* - \ell_0^*) \quad (14)$$

3. Results and Discussion

The achieved result of the pre-processing stage is forming training data. Based on reference data from ADS-B for 30 days accompanied by the process of data normalization, it generated training data. Table 3 shows the reference data ADS-B for 30 days.

Table 3. Data Source ADS-B for 30 days

Date	Time	Latitude	Longitude	Altitude	Record
180414	123844.20	-7.282300	112.299640	35000	1
180414	123844.41	-7.282300	112.299640	35000	2
...	...	...	...	...	...
150514	90055.55	-7.443790	112.728960	35000	3000

In the ADS-B data for 30 days, the number of records is 100 per day. Table 4 showed training data along with normalization values, namely MSE (Mean Square Error) and SSE (Sum Square Error).

Table 4. Data training for normalization results

Latitude	Longitude	MSE Latitude	MSE Longitude	SSE Latitude	SSE Longitude
-7.283655	112.465800	0.02	0.03	0.52	1.03
-7.285892	112.470210	0.02	0.04	0.53	1.07
...	...	...	...	...	...
-7.341225	112.636857	0.02	0.03	0.54	0.94

The normalization process eliminates duplicate data from existing ADS-B reference data. In this study, the parameters used in training data were latitude and longitude.

The segment aims to determine an area for the anomaly detection process in a real-time period. This study has five determined segments. The data per segment based on training data were 100 records. Then, the similarity of each segment data was tested by reference data ADS-B per day for 30 days of data.

Model is something built on similarity test results (log-likelihood). Several performed steps are segment, calculation of log-likelihood value, and grouping. The classification used two methods (SVM and KNN), while clustering implemented the K-Means method.

Table 5 shows an anomaly detection model based on a similarity test (log-likelihood). The test similarity with training data obtained data, containing the log-likelihood value of 30 records. The log-likelihood value grouping based on max log-likelihood (C1) and min log-likelihood (C2) values. Log-likelihood latitude and log-likelihood longitude values determine C1 and C2.

In the classification and clustering section, we use C1 and C2 values as classification and centroid initialization. So it was obtained several log-likelihood data (30 records) located on C1 and C2 with the percentage. Here is the maximum log-likelihood (C1) and minimum (C2) values obtained for the latitude and longitude parameters in each segment.

Table 5. Max and min log-likelihood per segment

C1 & C2	Latitude					Longitude				
	Segment					*Segment*				
	I	*II*	*III*	*IV*	*V*	*I*	*II*	*III*	*IV*	*V*
max log-likelihood (C1)	-15.97	-15.97	-15.97	-15.97	-15.97	-15.98	-15.97	-15.98	-15.97	-16
min log-likelihood (C2)	-20.30	-19.46	-19.30	-19.48	-19.50	-21.31	-21.20	-22.54	-23.72	-24.8

For the number of data based on classification and clustering (SVM, K-NN, and K-Means) each segment has the same number and percentage. **Segment 1** was the number of data (n) and percentage (%) in SVM, KNN, and K-Means, i.e.: n_C1 = 27 (90%) and n_C2 = 3 (10%). **Segment 2** on SVM, was: n_C1 = 27 (90%) and n_C2 = 3 (10%). While in KNN and K-Mean, it was obtained the same number of data and percentages, namely: n_C1 = 28 (93%) and n_C2 = 2 (6%). **Segment 3** on SVM, KNN, and K-Means obtained the same number of data and percentage, was: n_C1 = 28 (93%) and n_C2 = 2 (6%). **Segment 4** and **segment 5** obtained the same number of data and percentages, namely: on SVM with n_C1 = 28 (93%) and n_C2 = 2 (6%). For KNN and K-Means it was obtained n_C1 = 29 (96%) and n_C2 = 1 (3%).

The distribution of data in each class C1 and C2 is showed in the graphical representation. For grouping with KNN and K-Means, it is always pair in each segment. As a result, the percentage and the distribution of data in C1 and C2 were equal. Fig. 7 and Fig. 8 show the data distribution with SVM grouping. Fig. 9, Fig. 10, and Fig. 11 show the data distribution using **KNN** and **K-Means**

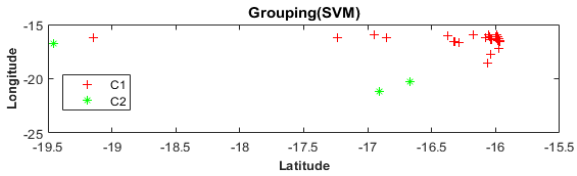

Fig. 7. SVM with n_C1=27 (90%) and n_C2=3 (10%) in segment 1 and segment 2

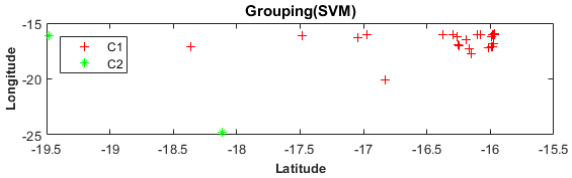

Fig. 8. SVM with n_C1=28 (93%) and n_C2=2 (6%) in segment 3, segment 4, and segment 5

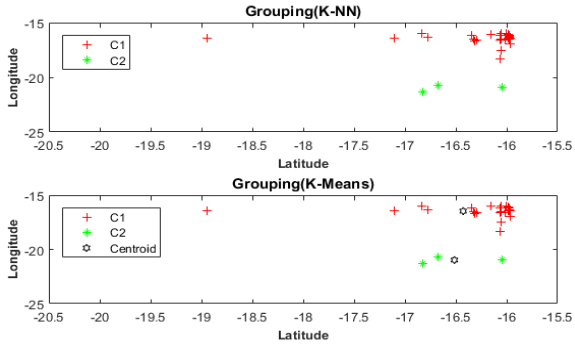

Fig. 9. K-NN and K-Means with n_C1=27 (90%) and n_C2=3 (10%) in segment 1

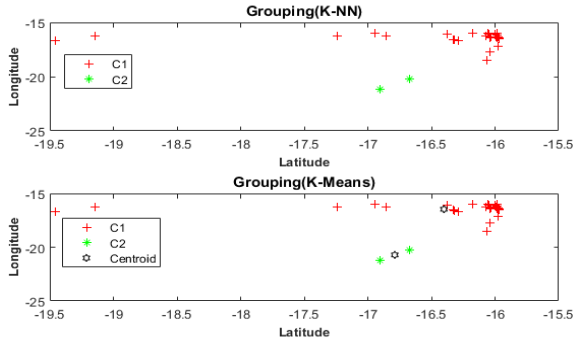

Fig. 10. K-NN and K-Means with n_C1=28 (93%) and n_C2=2 (6%) in segment 2 and segment 3

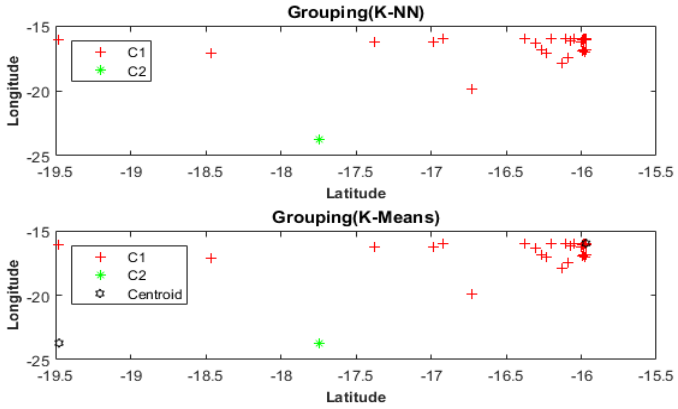

Fig. 11. K-NN and K-Means with n_C1=29 (96%) and n_C2=1 (3%) in segment 4 and segment 5

The testing process is performed by using ADS-B test data that is not ADS-B (30 days) reference data for training data. The stages are similar to the modeling process. Test data is determined in several

segments (segment 1 to 5). Furthermore, a similarity test (log-likelihood) of the test data was conducted with training data. Based on the log-likelihood value, we compared the log-likelihood value in the training data (C1 and C2). The result is a log-likelihood ratio (LLR) calculation. Small LLR values, then the test data is a particular class (C1 or C2).

The following test results are log-likelihood latitude and longitude values between test data and sequential training data per segment. **Segment 1** was: -16.002 (log-likelihood latitude), -16.9111 (log-likelihood longitude); **segment 2** was: -15.9891 (log-likelihood latitude), -17.0575 (log-likelihood longitude); **segment 3** was: -15.9695 (log-likelihood latitude), -16.8107 (log-likelihood longitude); **segment 4** was: -15.9673 (log-likelihood latitude), -17.0178 (log-likelihood longitude); and **segment 5** was: -15.9751 (log-likelihood latitude), -16.997 (log-likelihood longitude).

Test results determined based on LLR calculations along with the determination of the following sequential classes per segment. The result is detailed as follows: **Segment 1**, obtained LLR_C1 latitude, LLR_C1 longitude is (79.23, 83.42), and LLR_C2 latitude, LLR_C2 longitude is (79.63, 84.30) and test result is in class C1. **Segment 2**, obtained LLR_C1 latitude, LLR_C1 longitude is (79.17, 84.10), and LLR_C2 latitude, LLR_C2 longitude is (79.57, 84.97) and the test result is in class C1. **Segment 3** obtained LLR_C1 latitude, LLR_C1 longitude is (79.08, 82.96), and LLR_C2 latitude, LLR_C2 longitude is (79.48, 83.84), and the test result is in class C1. **Segment 4**, obtained LLR_C1 latitude, LLR_C1 longitude is (79.07, 83.91) and LLR_C2 latitude, LLR_C2 longitude is (79.47, 84.79) and test result is in class C1. **Segment 5**, obtained LLR_C1 latitude, LLR_C1 longitude is (79.11, 83.82), and LLR_C2 latitude, LLR_C2 longitude is (79.51, 84.69) and the test result is in class C1.

For accuracy and precision measurement is shown in Table 6 and Table 7. Table 6 shows the accuracy (%) of each grouping that occurs in the segment, while in Table 7 shows the precision (%) on each group on the segment.

Table 6. Accuracy (%) in each group that occurred in the segment

Grouping	Segment Accuracy				
	I	*II*	*III*	*IV*	*V*
SVM	100%	97%	97%	97%	100%
K-NN	97%	100%	100%	97%	97%
K-Means	97%	100%	100%	97%	97%

Table 7. Precision (%) in each group that occurred in the segment

Grouping	Segment Precision				
	I	*II*	*III*	*IV*	*V*
SVM	100%	96%	96%	96%	100%
K-NN	100%	100%	100%	96%	96%
K-Means	100%	100%	100%	96%	96%

Based on F-measure, Table 8 shows FPR and TPR per segment based on SVM, and K-NN & K-Means.

Table 8. FPR and TPR

	Segment									
	1		*2*		*3*		*4*		*5*	
	SVM	*K-NN & K-Means*	*SVM*	*K-NN & K-Means*	*SVM*	*K-NN & K-Means*	*SVM*	*K-NN & K-Means*	*SVM*	*K-NN & K-Means*
FPR	0%	0%	25%	0%	33%	0%	33%	50%	0%	50%
TPR	100%	96%	100%	100%	100%	100%	100%	100%	100%	100%

Percentage of 100% in ROC is due to the rate of TPR in each grouping (SVM, K-NN, and K-Means) is 100%. There is only one percentage worth 96% (i.e., in segment 1 in the K-NN and K-Means groupings). FPR percentage has a different rate. The maximum is 50%, while the minimum is 0%.

4. Conclusion

The result is a log-likelihood (LL) model in aviation anomaly detection based on ADS-B. The subsequent studies are determined based on the formation of segments and testing processes, and the segment is formed by distance-based clustering. In addition, the testing process is done by forming a new model again in testing data (re-modeling) without calculating the value of the Log-likelihood Ratio. Based on the clustering approach, the highest percentage of the most data on C1 is in the fourth segment and fifth segment. The percentage of K-NN and K-Means is 96%, and SVM is 93%. While the highest percentage of C2 is in the first and second segments with a 10% percentage of SVM, K-NN, and K-Means.

References

[1] L. Li, M. Gariel, R. J. Hansman, and R. Palacios, "Anomaly detection in onboard-recorded flight data using cluster analysis," in *Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th*, 2011, p. 4A4--1, doi: 10.1109/DASC.2011.6096068.

[2] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, "Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations," *J. Aerosp. Inf. Syst.*, vol. 12, no. 9, pp. 587–598, Sep. 2015, doi: 10.2514/1.I010329.

[3] P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means cluster analysis in earthquake epicenter clustering," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 2, pp. 81–89, Jul. 2017, doi: 10.26555/ijain.v3i2.100.

[4] D. S. Hicok and D. Lee, "Application of ADS-B for airport surface surveillance," in *17th DASC. AIAA/IEEE/SAE. Digital Avionics Systems Conference. Proceedings (Cat. No. 98CH36267)*, 1998, vol. 2, pp. F34--1, doi: 10.1109/DASC.1998.739823.

[5] X. Zhang, J. Zhang, S. Wu, Q. Cheng, and R. Zhu, "Aircraft monitoring by the fusion of satellite and ground ADS-B data," *Acta Astronaut.*, vol. 143, pp. 398–405, 2018, doi: 10.1016/j.actaastro.2017.11.026.

[6] M. Gariel, F. Kunzi, and R. J. Hansman, "An algorithm for conflict detection in dense traffic using ADS-B," *AIAA/IEEE Digit. Avion. Syst. Conf. - Proc.*, pp. 1–12, 2011, doi: 10.1109/DASC.2011.6095916.

[7] M. Orefice, V. Di Vito, F. Corraro, G. Fasano, and D. Accardo, "Aircraft conflict detection based on ADS-B surveillance data," *2014 IEEE Int. Work. Metrol. Aerospace, Metroaerosp. 2014 - Proc.*, pp. 277–282, 2014, doi: 10.1109/MetroAeroSpace.2014.6865934.

[8] G. Cheng and X. Tong, "Fuzzy Clustering Multiple Kernel Support Vector Machine," *2018 Int. Conf. Wavelet Anal. Pattern Recognit.*, pp. 7–12, doi: 10.1109/ICWAPR.2018.8521307.

[9] R. Wang, W. Li, R. Li, and L. Zhang, "Signal Processing : Image Communication Automatic blur type classification via ensemble SVM," *Signal Process. Image Commun.*, vol. 71, no. 37, pp. 24–35, 2019, doi: 10.1016/j.image.2018.08.003.

[10] L. V Utkin, "An imprecise extension of SVM-based machine learning models," *Neurocomputing*, 2018, doi: 10.1016/j.neucom.2018.11.053.

[11] Z. Liu, Z. Zhang, Y. Liu, J. Dezert, and Q. Pan, "Knowledge-Based Systems A new pattern classification improvement method with local quality matrix based on K-NN," *Knowledge-Based Syst.*, 2018, doi: 10.1016/j.knosys.2018.11.001.

[12] S. S. Aung, I. Nagayama, and S. Tamaki, "Regional Distance-based k-NN Classification," pp. 56–62, 2017, doi: 10.1109/ICIIBMS.2017.8279719.

[13] K. Shankar and M. Ilayaraja, "Secure Optimal k -NN on Encrypted Cloud Data using Homomorphic Encryption with Query Users," *2018 Int. Conf. Comput. Commun. Informatics*, pp. 1–7, 2018, doi: 10.1109/ICCCI.2018.8441290.

[14] S. F. Hussain and M. Haris, "A k-means based Co-clustering (kCC) Algorithm for Sparse, High Dimensional Data," *Expert Syst. Appl.*, 2018, doi: 10.1016/j.eswa.2018.09.006.

[15] G. Tzortzis and A. Likas, "The MinMax k-Means clustering algorithm," *Pattern Recognit.*, vol. 47, no. 7, pp. 2505–2516, Jul. 2014, doi: 10.1016/j.patcog.2014.01.015.

[16] W. He, D. Zhao, Y. Zheng, and J. Xie, "An Expected Patch Log Likelihood Denoising Method Based on Internal and External Image Similarity," in *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI)*, 2018, pp. 1–4, doi: 10.1109/ISSI.2018.8538103.

[17] V. M. Suhila and B. C. Kovoor, "Optimized H ybrid Approach for T opic S earch using Log Likelihood and RV Coefficient," *2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput.*, pp. 338–341, 2017, doi: 10.1109/ICECDS.2017.8390062.

[18] J. Lee and H. Chung, "Exact and approximate log-likelihood ratio of M-ary QAM with two-time dimensions," *ICT Express*, vol. 5, no. 3, pp. 173–177, 2019, doi: 10.1016/j.icte.2018.08.004.

[19] C. Laufer, *The Hobbyist's Guide to the RTL-SDR: Really Cheap Software Defined Radio*. 2015, available at: Google Scholar.

[20] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to pattern recognition: a matlab approach*. Academic Press, 2010, available at: Google Scholar.

[21] H. Hartono, O. S. Sitompul, T. Tulus, and E. B. Nababan, "Biased support vector machine and weighted-smote in handling class imbalance problem," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, p. 21, Mar. 2018, doi: 10.26555/ijain.v4i1.146.

[22] A. Prahara, A. Pranolo, and R. Drezewski, "GPU Accelerated Number Plate Localization in Crowded Situation," *Int. J. Adv. Intell. Informatics*, vol. 1, no. 3, pp. 150–157, 2015, doi: 10.26555/ijain.v1i3.46.

[23] M. Latah and L. Toker, "A novel intelligent approach for detecting DoS flooding attacks in software-defined networks," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 11–20, Mar. 2018, doi: 10.26555/ijain.v4i1.138.

[24] T. Lakshmi Priya, N. R. Raajan, N. Raju, P. Preethi, and S. Mathini, "Speech and non-speech identification and classification using KNN algorithm," *Procedia Eng.*, vol. 38, pp. 952–958, 2012, doi: 10.1016/j.proeng.2012.06.120.

[25] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "An affinity-based new local distance function and similarity measure for kNN algorithm," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 356–363, 2012, doi: 10.1016/j.patrec.2011.10.021.

[26] Y. Ding, Y. Zhao, X. Shen, M. Musuvathi, and T. Mytkowicz, "Yinyang k-means: A drop-in replacement of the classic k-means with consistent speedup," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, available at: Google Scholar.

[27] S. Ahmadian, A. Norouzi-Fard, O. Svensson, and J. Ward, "Better Guarantees for k-Means and Euclidean k-Median by Primal-Dual Algorithms," *SIAM J. Comput.*, pp. FOCS17-97-FOCS17-156, Oct. 2019, doi: 10.1137/18M1171321.

[28] X. Liu *et al.*, "Multiple Kernel k-means with Incomplete Kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019, doi: 10.1109/TPAMI.2019.2892416.

[29] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 2015, no. 4, doi: 10.1198/tech.2005.s319.

[30] I. Kulikovskikh and S. Prokhorov, "Minimizing the effects of floor and ceiling to improve the convergence of log-likelihood," *Procedia Eng.*, vol. 201, pp. 779–788, 2017, doi: 10.1016/j.proeng.2017.09.627.

[31] D. Jarušková and V. I. Piterbarg, "Log-likelihood ratio test for detecting transient change," *Stat. Probab. Lett.*, vol. 81, no. 5, pp. 552–559, 2011, doi: 10.1016/j.spl.2011.01.006.