Mechatronics and Intelligent Transportation Systems
https://www.acadlore.com/journals/MITS

Assessing Automatic Dependent Surveillance-Broadcast Signal Quality for Airplane Departure Using Random Forest Algorithm

Rani Silvani Yousnaidi[1], Rossi Passarella[2*], Rizki Kurniati[1], Osvari Arsalan[1], Aditya[1], Indra Gifari Afriansyah[1], Muhammad Rifqi Fathan[1], Marsella Vindriani[2]

[1] Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya, 30129 Palembang, Indonesia
[2] Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, 30129 Palembang, Indonesia

[*] Correspondence: Rossi Passarella (passarella.rossi@unsri.ac.id)

Received: 04-04-2023 **Revised:** 05-19-2023 **Accepted:** 05-24-2023

Citation: R. S. Yousnaidi, R. Passarella, R. Kurniati, O. Arsalan, Aditya, I. G. Afriansyah, M. R. Fathan, M. Vindriani, "Assessing Automatic Dependent Surveillance-Broadcast signal quality for airplane departure using random forest algorithm," *Mechatron. Intell Transp. Syst.*, vol. 2, no. 2, pp. 64–71, 2023. https://doi.org/10.56578/mits020202.

Abstract: This study aims to assess the safety level of the Automatic Dependent Surveillance-Broadcast (ADS-B) signal quality during airplane departures at Sultan Mahmud Badaruddin II Airport. The Aero-track application was utilized to monitor commercial aircraft departures and collect observation data. The collected data underwent processing using data analysis algorithms and labeling processes, resulting in a comprehensive dataset for evaluating ADS-B signal quality. Signal quality was categorized into four levels, and a model was built using the Random Forest algorithm, achieving an accuracy of 99%. Comparative analysis with SVM and Naive Bayes algorithms showed accuracy values of 93% and 97% respectively. Consequently, the Random Forest Model was chosen for estimating ADS-B signal quality during commercial aircraft takeoff and landing.

Keywords: Automatic Dependent Surveillance-Broadcast (ADS-B); Signal quality; Classification; Random forest

1 Introduction

The three main causes of aviation accidents during takeoff and landing—technical issues, weather, and human error—generally account for 63% of all accidents involving aircraft [1]. Accident analysis during the take-off or landing phase now depends solely on data from the black box, whereas, as information technology advances, data from the Automatic Dependent Surveillance-Broadcast (ADS-B) should be a concern in order to see alternative points of view in the accident analysis process [2]. Yet, the ADS-B signal quality is the most important component in convincing the analytical results utilizing ADS-B data. ADS-B allows aircraft to broadcast their identification and current position, determined by the global navigation satellite system, to ground stations or other aircraft above 1090 MHz [3]. The ADS-B ground station equipment on the ground receives the signal from ADS-B aircraft and sends it to the ATC (Air Traffic Controller). The ATC display then uses the data to track the aircraft [4]. ADS-B is particularly significant, especially for airports with a high operating level, since it is utilized to assist operational activities with the goal of boosting the safety, capacity, and efficiency of national airspace system operations, lowering radar installation, and providing off-radar coverage. One of the Object studies in this research is Sultan Mahmud Badaruddin II Airport ($02°54'01''$S $104°42'00''$E), which sees an average of 16 departures each day from both local and foreign destinations. This airport's busy schedule ought to be complemented by a strong ADS-B signal.

The purpose of this study was to evaluate Sultan Mahmud Badaruddin II Airport's ADS-B signal quality for measuring its level of safety. To delve deeper, a mechanism for classifying the quality of ADS-B data at Sultan Mahmud Badaruddin II Airport is required. This research has its own level of difficulty because there is no reference for analyzing ADS-B signals at Indonesian airports, so this is a novelty for this research, which will ultimately provide an overview of data quality around the airport area. The Random Forest classification method was picked as one of the options. Based on prior research, the Random Forest method produced an accuracy of 86.14% and an f1-score of 86.93% when used to classify maritime transportation, specifically ships [5].

We organize this study as follows: The first section is the introduction, which contains the study's aims; the second section is the methodology, which covers the technique adopted to reach the objectives, as well as the methods of data collecting, data processing, and random forest algorithm implementation. The third section provides the results of the methodology's execution, as well as talks about the process of analyzing the results. The conclusions are revealed in the last section.

2 Methodology

The Research Methodology in this research describes the procedures or techniques used to identify the problem or objective, select the data, and process and analyze the information. The general flow of methodology is illustrated in Figure 1.

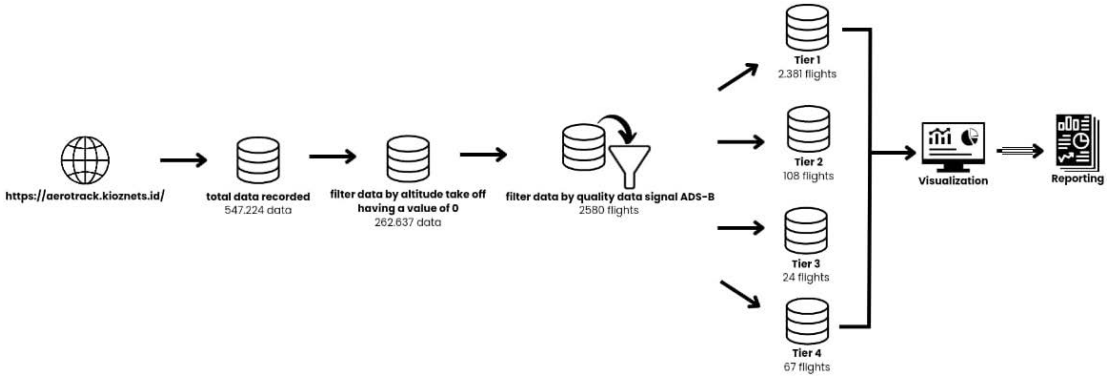


Figure 1. Research flow diagram

The main data source in this study is an application named AERO-TRACK [6], which is used to record flight departure data at Sultan Mahmud Badaruddin II Airport, after which the data is downloaded in database format and translated into CSV format with the following data information as seen in Table 1.

Table 1. Metadata of the database from Aero-track

Data information	
Recording period	1 February 2022 – 1 August 2022
Download date	15 August 2022
Format	.csv
Data size	68,441 MB
Number of rows	547.224
Number of columns	21

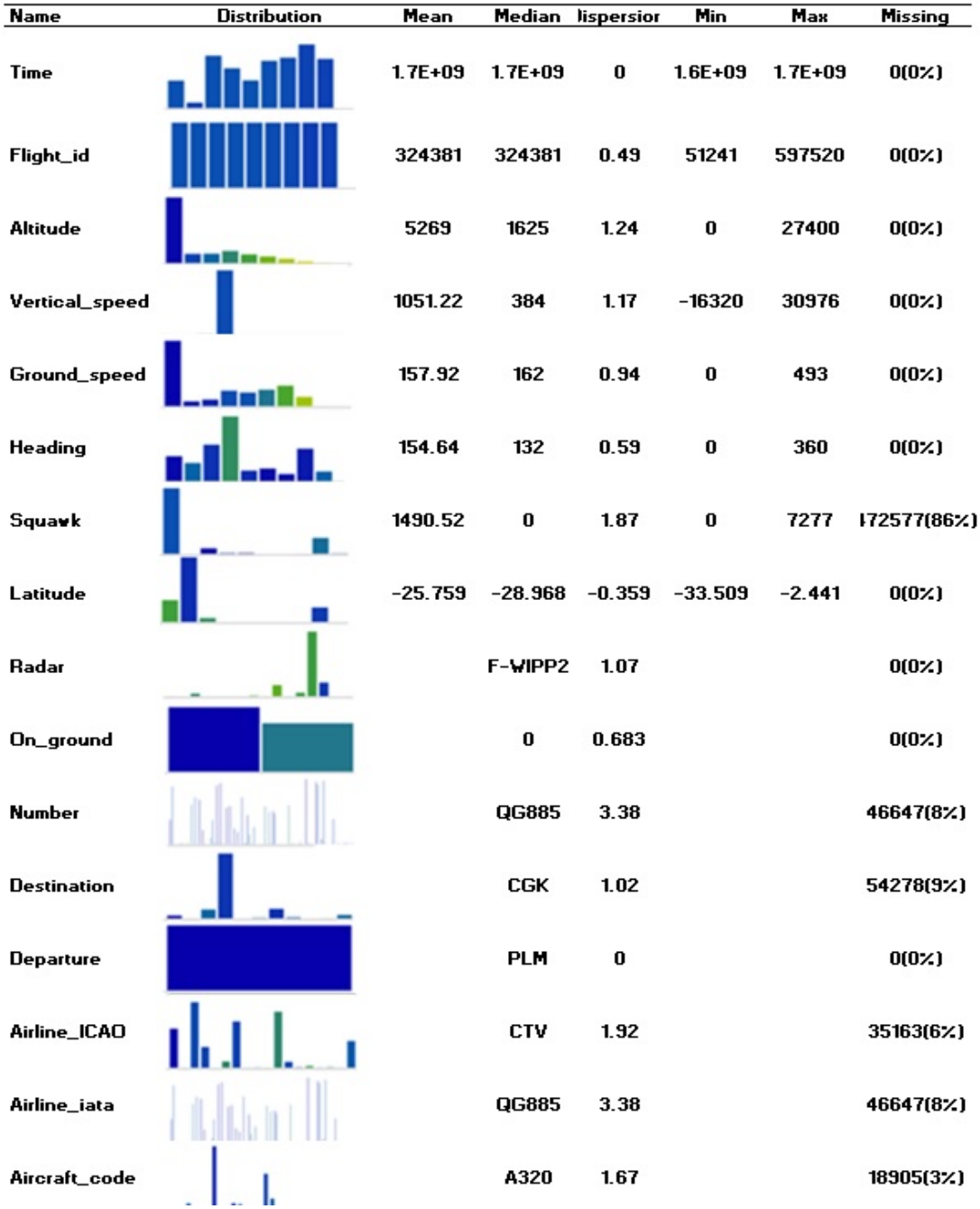
2.1 Sultan Mahmud Badaruddin II Airport

Sultan Mahmud Badaruddin II International Airport (IATA: PLM, ICAO: WIPP) is a commercial airport in Talang Betutu, South Sumatra, owned by the government of Indonesia. The airport has one runway with two landing or takeoff directions, 11 and 29, with a length of 9,834 feet (3000 meters) with an asphalt surface. According to statistics, the total number of passengers who used this airport in 2018 was 5,126,298.

2.2 Data Preprocessing

Preprocessing is where the data from Table 1 is processed. Reduced data size, association discovery, data normalization, irregularity removal, and feature extraction are the main goals of preprocessing. The strategies used in this process include data cleaning, integration, transformation, and reduction [7]. The data cleaning phase is the initial step in preprocessing procedures used to identify anomalies, clean out noise in the data, detect missing values, and fix incorrect data [8]. Figure 2 shows a variable in a dataset that is missing a value. A missing value can reveal bias and lower the quality of a flight analysis' results [9]. Resolving missing values can be done in several ways, including (a) eliminating objects that contain missing values, (b) manually resolving missing values, (c) employing global or object-to-object consistency, and (d) finding the most likely answer to the problem [10].

We choose the data that will be used in the visualization process and discard the remainder at the preprocessing stage. Preprocessing is carried out using Jupyter Notebook, the Python programming language, and the panda's library. The following are the phases of preprocessing:



Name	Distribution	Mean	Median	lispersior	Min	Max	Missing
Time		1.7E+09	1.7E+09	0	1.6E+09	1.7E+09	0(0%)
Flight_id		324381	324381	0.49	51241	597520	0(0%)
Altitude		5269	1625	1.24	0	27400	0(0%)
Vertical_speed		1051.22	384	1.17	-16320	30976	0(0%)
Ground_speed		157.92	162	0.94	0	493	0(0%)
Heading		154.64	132	0.59	0	360	0(0%)
Squawk		1490.52	0	1.87	0	7277	172577(86%)
Latitude		-25.759	-28.968	-0.359	-33.509	-2.441	0(0%)
Radar			F-WIPP2	1.07			0(0%)
On_ground			0	0.683			0(0%)
Number			QG885	3.38			46647(8%)
Destination			CGK	1.02			54278(9%)
Departure			PLM	0			0(0%)
Airline_ICAO			CTV	1.92			35163(6%)
Airline_iata			QG885	3.38			46647(8%)
Aircraft_code			A320	1.67			18905(3%)

Figure 2. Statistic feature

1. Date and icao24 filtering. This step is important so that the data may be sorted according to a flight with a specific date schedule.

2. Removing any duplicate information discovered in time variables. Duplicate data in rows has an impact on how well the data is visualized.

3. Don't keep altitude rows that aren't datasets. The following criteria apply to altitude deletion:

a. Deleting altitude 0 and using the latest line altitude recorded for each trip as its final altitude.

b. If the flying altitude does not start at 0, then the altitude is still taken.

c. The altitude line is eliminated if the final altitude is zero.

4. New columns next to the time column: difference, average, and tier. These columns are utilized in the processing and evaluation of data quality.

5. Using the formula second time minus first time, followed by other iterations, to determine the difference. The data needed to process data quality includes the difference.

6. Determining the average by dividing the total number of differences by the sum of all differences.

7. The criteria should be used to classify the signal quality as viewed from the average as seen in Table 2. Description:

Tier 1: Data standards with high-performance traffic separation services are included in this class.

Tier 2: Data standards with traffic situational awareness services and procedural separation can be found in this class.

Tier 3: This group of data standards includes those with traffic advisory functions (flight information services).

Tier 4: Unacceptable data standards are used in this class.

8. Using the date, icao24, aircraft code, registration, airline icao, average, and category columns as the dataset for the classification method as well as additional research material.

Table 2. Signal class ADS-B [11]

Parameter	Tier 1	Tier 2	Tier 3	Tier 4
Aircraft updates	$0.5\text{ s} < \text{x} < 10\text{ s}$	$10\text{ s} \leq \text{x} < 20\text{ s}$	$20\text{ s} \leq \text{x} < 60\text{ s}$	$x \geq 60\text{ s}$

Note: x = interval (second)

2.3 Random Forest Implementation

One classification and regression-based method using a decision tree aggregation procedure is the Random Forest. This technique is employed because it results in fewer errors and has a decent classification accuracy [12]. When the majority of the data is absent, Random Forest provides an efficient approach for predicting the missing data, maintains accuracy, and has a mechanism for balancing class mistakes in the data set [13]. With Random Forest, the stages of compilation and estimation are:

a. The bootstrap stage, where random samples with recoveries of size n are drawn from the cluster of training data.

b. The random sub-setting stage, where a tree is constructed from the data but m ¡ d explanatory variables are randomly selected; the best splits are produced in this stage.

c. Repeating steps, a – b as much as k times in order to produce k random trees.

d. Doing a joint estimation using k-trees (e.g., using the majority vote for classification cases or the average for regression cases).

Using Jupyter Notebook, the Python programming language, and the sklearn package, this classification technique is implemented. The dataset is initially split into three categories: 20% of it is unseen data or data whose labels have been purposefully removed; The training data is drawn from the remaining dataset (80%), which is divided into 70%, which is used to teach the computer to recognize patterns; and the remaining 30% is test data, which is used to evaluate the effectiveness of the machine's training. This method of splitting the data is called the "holdout" method [14–19]. There are 1,444 rows of training data after being separated, as can be seen. The confusion matrix is used to test the data and assess the algorithm's effectiveness. Table 3 shows the multi-class confusion matrix's shape.

Table 3. Confusion matrix multi-class

	Predicted classification				
	Classes	Tier 1	Tier 2	Tier 3	Tier 4
	Tier 1	TN	FP	TN	TN
Actual classification	Tier 2	FN	TP	FN	FN
	Tier 3	TN	FP	TN	TN
	Tier 4	TN	FP	TN	TN

Accuracy is a formula to determine the comparison of true (negative and positive) predictions with the overall data, as seen in Eq. (1), while precision is a formula to determine the ratio of data that is predicted positive to the overall data, as seen in Eq. (2). Next is referred to as "recall." Recall is a formula for knowing the true positive predicted value compared to the true positive value of the overall data, as seen in Eq. (3) and the last is the F1 score, which is the sum of recall and precision averaged, with a formula that can be written as in Eq. (4) [20].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

3 Results

The Random Forest classification algorithm will be used to train and evaluate the data after preprocessing in an effort to categorize it. Figure 3 displays the performance of the training data, showing that Random Forest can accurately categorize all categories because there is enough data scattered throughout each class for the model to do so.

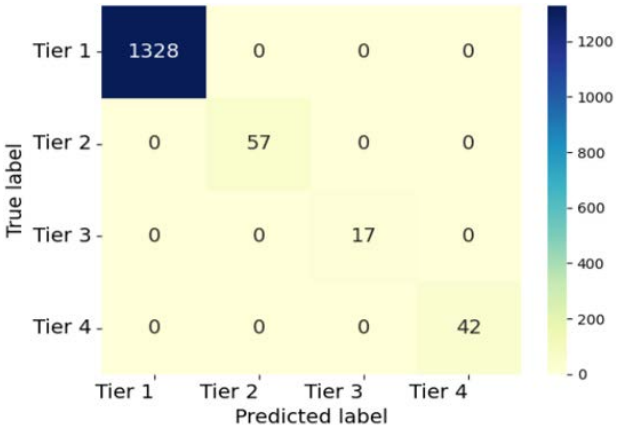


Figure 3. Training data performance

The 620 rows of test data were used to test the trained model. Figure 4 displays the test data's performance; there are 0–2 misclassifications in each category. This is because each class only has a tiny amount of test data, which leads to errors in the model.

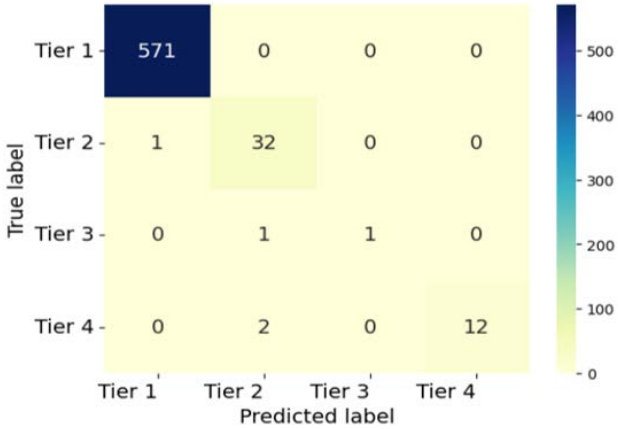


Figure 4. Test data performance

This study examines performance based on the f1-score caused by imbalanced datasets. The test data performance, which results in an f1-score value of 100% in Tier 1, 94% in Tier 2, 67% in Tier 3, and 92% in Tier 4, giving an overall f1-score value of 88%, is shown in Figure 5. The fact that there is a sizable disparity in the amount of data for each category has an impact on the f1-score's value.

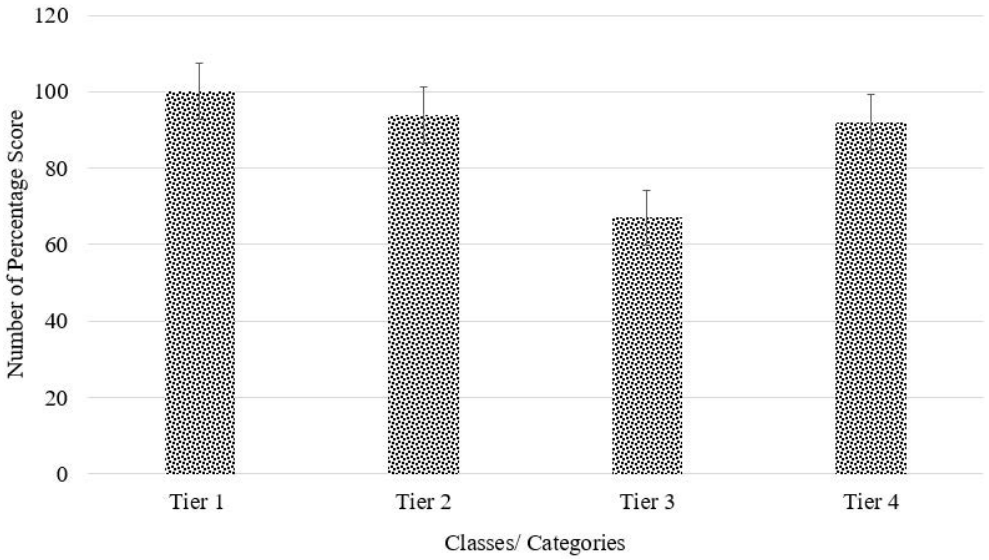


Figure 5. Random forest F1-score percentage

4 Discussion

Tier 1 has the best performance because there is enough training data and test data for the model to accurately classify, whereas Tier 3 has the worst performance because there is not enough training data and test data for the model to perform at its best. Because there are 2,381 flights in Tier 1, 108 in Tier 2, 24 in Tier 3, and 67 in Tier 4, the data categorization findings, shown in Figure 6, show that the Sultan Mahmud Badaruddin II Airport's ADS-B signal quality is in the good category.

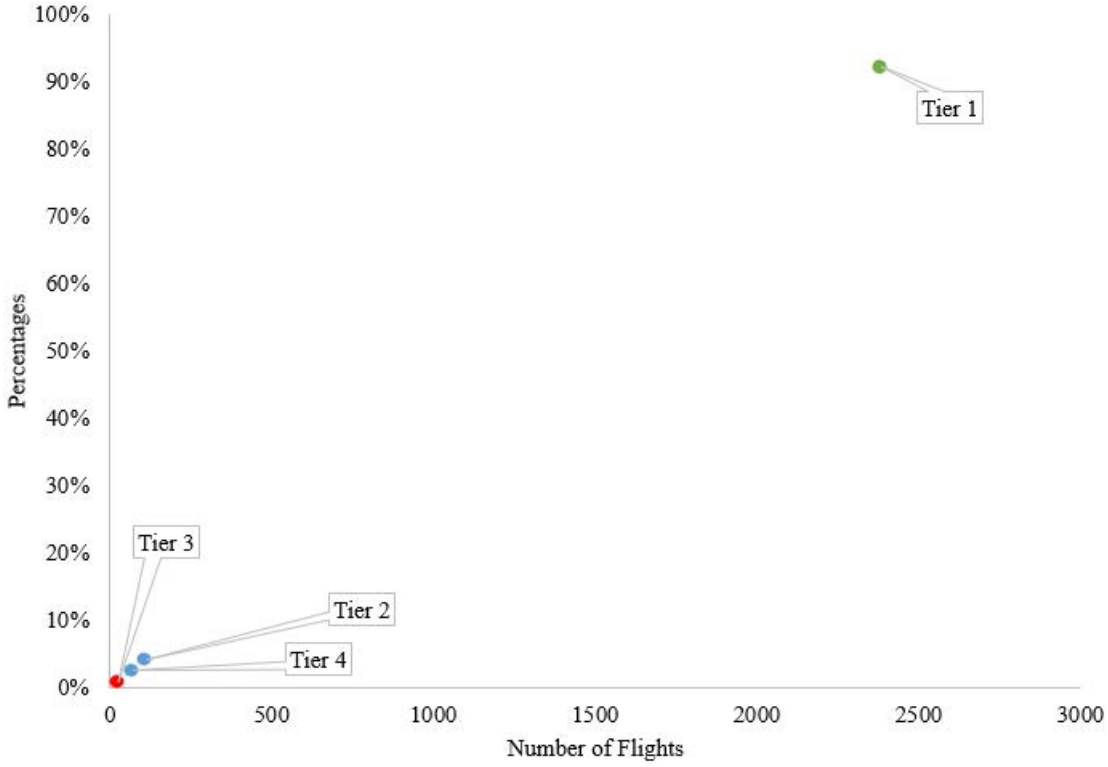


Figure 6. ADS-B signal quality

As most aircraft at Sultan Mahmud Badaruddin II Airport are categorized as Tier 1, it can be inferred from the trend in Figure 6 that signal quality will be high between February 1 and August 1 of 2022.

The findings of the ADS-B signal quality classification in this study revealed that the Random Forest algorithm produced an accuracy of 99% with this imbalanced data, which was significantly better than the accuracy of multi-

class classification with imbalanced data in earlier studies. From the results obtained, it is necessary to test again with other algorithms, which in this study used the SVM and Naïve Bayes algorithms with the same procedure as the procedure in Random Forest. The accuracy obtained using the SVM algorithm is 93%, while the Naïve Bayes algorithm produces an accuracy of 97%. However, random forest also has drawbacks such as a high probability of overfitting, making it very difficult to interpret the prediction results, taking a long training time, and using a lot of memory. All of these are influenced by the number of trees in the forest. However, these shortcomings are not so impactful if we still use a small number of datasets and classes [21–23]. As proven in this study.

5 Conclusions

As a result of Sultan Mahmud Badaruddin II Airport's dominance in the Tier 1 category, the data analysis results revealed that the ADS-B signal quality on airplane departure data is generally good. It follows from this that most active pilots adhere to the ADS-B use regulations. This study performs well even with data that is skewed and accurately classifies existing groups of classes, especially Tier 1. The quantity of training and test data in that category supports this. With the test results using the generative classification type, namely the Naïve Bayes algorithm, and the discriminative type, namely SVM and Random Forest, further testing is recommended to attempt utilizing a distribution-free classification type.

Data Availability

The data used are available from the corresponding author upon request.

Conflicts of Interest

We hereby declare that we have no pecuniary or other personal interest, direct or indirect, in any matter that raises or may raise a conflict with this manuscript.

References

[1] S. Angriyana, "Data boeing: 63% kecelakaan pesawat itu saat take off dan landing," 2018. https://travel.detik.com/travel-news/d-4280766/data-boeing-63-kecelakaan-pesawat-itu-saat-take-off-dan-landing

[2] R. Passarella and S. Nurmaini, "Behavioral evidence of public aircraft with historical data: The case of boeing 737 MAX 8 PK-LQP," *J. Appl. Eng. Sci.*, vol. 20, no. 4, pp. 1254–1262, 2022. https://doi.org/10.5937/jaes0-38696

[3] X. Ying, J. Mazer, G. Bernieri, M. Conti, L. Bushnell, and R. Poovendran, "Detecting ADS-B spoofing attacks using deep neural networks," *In 2019 IEEE Conference on Communications and Network Security (CNS), Washington, DC, USA*, pp. 187–195, 2019. https://doi.org/10.1109/CNS.2019.8802732

[4] M. Sohibi, D. Dermawan, and Lasmadi, "Rancang bangun receiver menggunakan antena 1090 MHz dan low noise amplifier untuk menambah jarak jangkau penerimaan sinyal dan data parameter target ads-b berbasis rtl820t2," *AVITEC*, vol. 2, no. 2, pp. 129–144, 2020. https://dx.doi.org/10.28989/avitec.v2i2.765

[5] Y. Wang, L. Yang, X. Song, and X. Li, "Ship classification based on random forest using static information from AIS data," *J. Phys. Conf. Ser.*, vol. 2113, no. 1, p. 012072, 2021. https://dx.doi.org/10.1088/1742-6596/2113/1/012072

[6] M. R. Fathan, A. Aditya, I. G. Afriansyah, R. S. Yousnaidi, R. Passarela, O. Arsalan, R. Kurniati, and M. Vindriani, "Aero-track: Perangkat lunak perekam data penerbangan aeronautika," *Generic*, vol. 15, no. 1, pp. 15–19, 2023.

[7] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC*, vol. 132, p. 116045, 2020. https://doi.org/10.1016/j.trac.2020.116045

[8] H. M. Marin-Castro and E. Tello-Leal, "Event log preprocessing for process mining: A review," *Appl. Sci.*, vol. 11, no. 22, p. 10556, 2021. https://doi.org/10.3390/app112210556

[9] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," *Appl. Intell.*, vol. 27, no. 1, p. 79, 2007. https://doi.org/10.1007/s10489-006-0032-0

[10] R. Somasundaram and R. Nedunchezhian, "Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values," *Int. J. Comput. Appl.*, vol. 21, no. 10, pp. 14–19, 2011. https://doi.org/10.5120/2619-3544

[11] International Civil Aviation Organization Asia and Pacific Office, "ADS-B implementation and operations guidance document," 2015. https://www.icao.int/APAC/Documents/edocs/APX-H%20-%20AIGD%20-%20Edition%208%200%20-%20Sep%202015%20to%20be%20adopted%20by%20APANPIRG26.pdf#search=Search%2E%2E%2EADS%2DB%20Implementation%20and%20Operations%20Guidance%20Document

[12] S. Lee, A. Abdullah, N. Jhanjhi, and S. Kok, "Classification of botnet attacks in iot smart factory using honeypot combined with machine learning," *PeerJ*, vol. 7, p. e350, 2021. http://dx.doi.org/10.7717/peerj-cs.350

[13] V. Y. Kulkarni and P. K. Sinha, "Effective learning and classification using random forest algorithm," *Int. J. Eng. Innov. Tech.*, vol. 3, no. 11, pp. 267–273, 2014.

[14] A. James and V. Tripathi, "Time series data analysis and ARIMA modeling to forecast the short-term trajectory of the acceleration of fatalities in Brazil caused by the corona virus (covid-19)," *PeerJ*, vol. 9, p. e11748, 2021. https://doi.org/10.7717/peerj.11748

[15] M. P. Keane and K. I. Wolpin, "Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior," *Int. Econ. Rev.*, vol. 48, no. 4, pp. 1351–1378, 2007. https://doi.org/10.1111/j.1468-2354.2007.00465.x

[16] A. Singh, "Cross-validation techniques," *Analytics Vidhya*, 2020.

[17] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40–79, 2010. https://doi.org/10.1214/09-SS054

[18] L. Devroye and T. Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 601–604, 1979. https://doi.org/10.1109/TIT.1979.1056087

[19] A. Kumar, "Hold-out method for training machine learning models," 2023. https://vitalflux.com/hold-out-method-for-training-machine-learning-model/

[20] K. Ramasubramanian and A. Singh, "Machine learning model evaluation," *Machine Learning Using R*, pp. 425–464, 2017. https://doi.org/10.1007/978-1-4842-2334-5_7

[21] R. Xu, "Improvements to random forest methodology," Ph.D. dissertation, Iowa State University, 2013.

[22] M. M. Cleaver, "Using random forest modeling to predict earthworm distribution in the ottawa national forest," Ph.D. dissertation, Michigan Technological University, 2018.

[23] C. J. Mantas, J. G. Castellano, S. Moral-García, and J. Abellán, "A comparison of random forest based algorithms: random credal random forest versus oblique random forest," *Soft Comput.*, vol. 23, pp. 10 739–10 754, 2019. https://doi.org/10.1007/s00500-018-3628-5