Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio

doi: 10.3850/978-981-18-8071-1_P519-cd

Data Preparation for Precursor Identification in Unstable Approach Events in flight data

Jie Yang, Diyin Tang, Jinsong Yu*, Yue Song, Lingkun Kong

School of Automation Science and Electrical Engineering, Beihang University.
E-mail: megan_yj@buaa.edu.cn, tangdiyin@buaa.edu.cn, yujs@buaa.edu.cn (Corresponding), songyue@buaa.edu.cn, konglingkun@buaa.edu.cn.

Unstable approaches have been identified as the main factor in most aviation accidents, making the identification of precursors to achieve such event prediction critical for ensuring the safety and reliability of flights. However, data preparation before precursor identification is challenging due to high-dimensional variable-length time series in a specific flight phase. In this study, we propose a pipeline for flight data preparation that offers standardized inputs for the precursor mining phase and labeled outputs for the unstable approach identification phase. The raw inputs are processed by an automatic feature selection based on correlation analysis. Additionally, a uniform dynamic time warping method is proposed to transform inputs with variable lengths into equal lengths for modeling, addressing the challenge of input variability caused by different tasks and weather conditions. The effectiveness of the preparation method in flight data is validated using flight data collected from regional aircraft. It is also possible to be extended to other adverse events occurring in flight phases in terms of precursor identification.

Keywords: Data Preparation, Unstable Approach Events, Automatic Feature Selection Strategy, Uniform Dynamic Time Warping, Precursor Mining

1. Introduction

Unstable approach events have been identified as common and potentially dangerous events in many aviation accidents. It can result in a hard landing, loss of aircraft control, runway excursion, and collision with terrain or infrastructure [10]. So, it stimulates the demand for precursor identification for event prediction, which provides a proactive way to ensure safety in aviation. However, flight data always tend to be high dimensional and variable-length. In this paper, we propose a pipeline for data preparation of precursor-based event prediction.

Precursor-based event prediction is a proactive way to find safety events by identifying precursors in advance. Precursors are defined as key anomalous behaviors triggering safety events, also with a high likelihood of the events occurring in the future. It not only offers one effective way to predict safety events but also provides automated explanations. The unstable approach prediction can be achieved by building a model according to standard inputs and labeled outputs. Fig. 1 shows a general descending and approach phases of an aircraft before touchdown. The key annotated phases, precursor mining phase (PMP) and unstable approach detection phase (UADP) are the raw data sources for generating standard inputs and labeled outputs.

Flight data always tend to be high-dimensional, which needs to be prepossessed without altering the original feature representation before modeling. Dimension reduction offers a way to transfer high-dimensional data to low-dimensional data. It could be executed by two main methods, feature selection, and feature extraction. Compared to feature extraction, feature selection that outputs a feature subset from the original feature set so as to preserve the original feature representation outperforms feature extraction which alters the original representation [13]. Feature selection methods include filter, wrapper [11], embedded [1], and hybrid [4]. The filter methods, like Pearson correlation, are applied in our automatic feature selection process. It is mainly because they have a good performance and own high-efficiency computing, especially for high dimensional flight data [14].

Moreover, the length of each flight in a specific

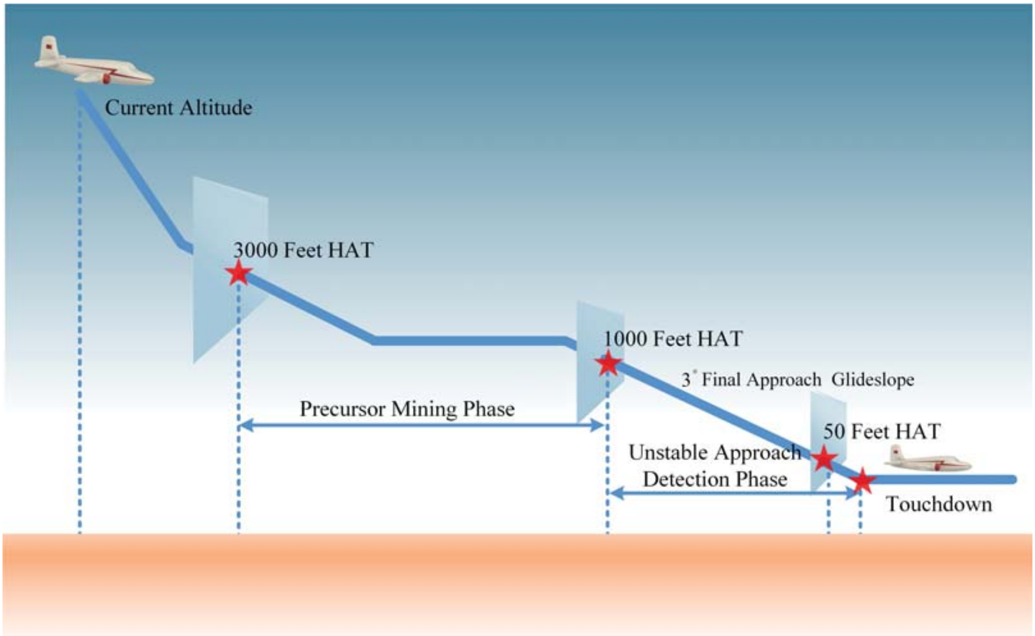

Fig. 1. Descending and approach phases of aircraft

phase is varied, which is caused by its flight plan, weather condition, and so on. Also, before reaching a stable approach different pilots may need uncertainty adjustments in a descending phase. It brings difficulty when defining a fixed length of precursor mining phase in unstable approach event prediction. Janakiraman [9] re-sampled time series at every quarter nautical mile starting approximately 25 nautical miles away from the runway. Ackley et al. [2] selected every 250-feet altitude from 3000 feet to landing. However, they are unable to characterize flight paths, either in terms of lateral distance or vertical height. In this case, dynamic time warping (DTW) could be used to rearrange variable-length time series in the precursor mining phase to find a warped query with the shortest distance compared to a flight path template. However, when each warped query meets the shortest distance, each may not be at the same length though it considers the flight path to change query arrangement.

This paper proposes a pipeline to prepare standard flight data for precursor-based unstable approach prediction. It applies automatic feature selection based on the Pearson coefficient to lower flight data dimension and raises a unified DTW to later each query's arrangement to represent flight path, also keeping each query's length the same as a reference.

2. Identifying Unstable Approach Events

In this paper, data preparation consists of two steps, that is generating standard low-dimensional inputs with equal-length precursor mining phase (PMP) and labeling events in the unstable approach detection phase (UADP). So, unstable approach prediction can be achieved by building a model according to standard inputs and labeled outputs. In Section 2, the labeled outputs of the unstable approach can be obtained by the following process.

Suppose multi-dimensional time series as inputs $X^{(N,L,D)}$ where N is the number of samples, L is the maximum length of time series and D is the number of sensor variables, standard inputs X_{input} in PMP and labeled outputs Y_{output} in

UADP can be expressed as,

$$X_{input} = X_{pm}^{(N,l,d)} \tag{1}$$

$$Y_{output} = RULE(X_{ua}) \tag{2}$$

where X_{pm} and X_{ua} are time series in PMP and UADP alternatively. l is the number of selected sensor variables ($l \ll L$) and d is the fixed length of the time series. $RULE(\cdot)$ is exceedance rules to define unstable approach events and $Y_{output} \in [0, 1]$ is labels of unstable approach events where 1 states unstable approach happens and vice versa.

In terms of unstable approach events time series in PMP and UADP are required to be identified in an aircraft descending and approach phases. In order to make the use of time series from the flight recorder, an appropriate altitude need to be identified and selected carefully. According to the distribution of the altitude at which unstable approach events occur, time series in UADP X_{ua} is defined where the altitude is below 1000 feet before landing. Furthermore, the time series between 3000 feet and 1000 feet altitude in a descending phase is time series in PMP X_{pm}, which captures around 2 minutes sufficient to seek precursors happening prior to unstable approach events. In summary, PMP aims to mine precursors triggering unstable approaches, while UAIP needs to detect unstable approaches.

To determine which flight contains an unstable approach, the exceedance rules $RULE(\cdot)$ is applied to detect unstable approach, mainly referring to several reports from Flight Safety Foundation [7] [6], International Civil Aviation Organization [8], European Union Aviation Safety Agency [5] and Boeing [3]. According to specific flight data, sensor variables and their threshold requires to be adjusted in reality considering that the flight data always doesn't contain all the information used to verify the requirements in a specific standard above. The exceedance rules in the paper are shown in Table 1, including Glideslope Deviation, Localizer Deviation, Sink Rate, Drift Angle Rate, Late Flap Extension, Speed Brake Deployment, N1 Thrust, Final Approach Airspeed Reference. The final approach airspeed reference V_{app} states whether an airplane is descending too slowly or too high in the final approach phase. Due to the lack of V_{app} in flight recorders, this paper estimates it by its definition which is the mean airspeed under 50 feet high above touchdown in a statistical way.

An unstable approach is detected when at least one of the rules is met. In this case, we could obtain the labeled outputs Y_{output} of unstable approach events in UADP via the built exceedance rules in Table 1.

3. Automated Feature Selection Using Statistical Methods

In the following Sections, given time series in PMP X_{pm}, standard low-dimensional and equal-length input X_{input} can be get by automatic feature selection based on correlation coefficient and uniform dynamic time warping, alternatively.

First, a correlation analysis is exhibited by Pearson's correlation coefficient to reduce feature space, as well as to enhance predictive performance. Highly correlated variables would present similar feature importance scores so as to skew feature importance ranking and increase the model's complexity and computation. Given a parameter pair x and y, the Pearson's correlation coefficient $\rho_{x,y}$ is represented as,

$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{3}$$

where $cov(x, y)$ is the covariance of a parameter pair (x, y), and σ_x and σ_y are the standard deviation of x and y separately. When a parameter pair's $|\rho_{x,y}| > \delta$, then one of them will be deleted. So, given the datasets $X \in \mathbb{R}^{n\times m}$, its' numerical variable vector $Var \in \mathbb{R}^{m\times 1}$, the variable vector to be saved Var_{save} and to be deleted Var_{del} and visited feature vector F_{vi}, the new selected variables Var_{sel} can be obtained by Feature selection based on correlation analysis algorithm, shown in below:

In the above algorithm, outliers remover must be done before normalization in order to improve the effectiveness of normalization. Also, when discrete variables exist, an extra criterion is introduced by whether a variable has limited constants to identify discrete variables, in case of acciden-

Table 1. Exceedance Rules for Unstable Approach.

Variables	Specific Parameters	Type	Rules	Duration
Glideslope Deviation	Glideslope Deviation (GSD)	Analogue	$\lvert$GSD$\rvert$ $>$ 1 dot	5
Localizer Deviation	Localizer Deviation (LD)	Analogue	$\lvert$LD$\rvert$ $>$ 1/4 dot	5
Sink Rate	Sink Rate (SR)	Analogue	SR $>$ 1000 feet/min	3
Drift Angle Rate	Standard Deviation of Drift Angle Rate (SDYR)	Analogue	SDYR $>$ 1.25	3
Late Flap Extension	The Delta Flap Level Position (DFLP)	Discrete	$\lvert$DFLP$\rvert$ $>$ 0	1
Speed Brake Deployment	The Delta Brake Position (DBP)	Discrete	$\lvert$DBP$\rvert$ $>$ 0	1
N1 Thrust	Engine N1 Thrust (ENT)	Analogue	ENT/ENT_{ref} $<$ 35%	5
Airspeed	Calibrated Airspeed (CAS)	Analogue	CAS $<V_{app}$ -5 kts or CAS $>V_{app}$ + 10 kts	5

Note: ENT_{ref}: Engine N1 Thrust Reference; V_{app}: 140 kts.

Algorithm 1 Automated Feature Selection Using Statistical Methods

Input: $X \in \mathbb{R}^{n\times m}$, $Var \in \mathbb{R}^{m\times 1}$, $F_{vi} = [\,]$

Remove outliers in X, Normalize X, and delete variables with low standard deviation which is lower than 10^{-5} in X

Calculate the correlation coefficient matrix $\rho \in \mathbb{R}^{m\times m}$ of each pair (i, j) in Var, where $i, j \in [1, 2, ..., m]$

Update $\rho = \rho[\, |\rho| > \delta \;\&\; \rho! = 1\,]$

repeat

  Obtain the indices $corr^{i}_{del}$ of the non-NAN elements in i^{th} column in ρ

  Save the variables corresponding to $corr^{i}_{del}$ into Var_{del}

  Update F_{vi} by $Var_{del} \cup F_{vi}$

until $|F_{vi}| = m$ {Stop when all the variables have been visited}

Obtain Var_{new} by deleting Var_{del} in Var

Calculate $\{FDR_1, FDR_2, ..., FDR_k, ...\}$ for each variable in Var_{new}

Determine the number of selected features l by $l = min\{k \mid \sum_{i=1}^{k} FDR_i >= \zeta\}$

Return Var_{sel} with l highest FDR in a descending order

Output: Var_{sel}

tally removing those variables with outliers.

After variables reduction by Pearson's correlation coefficient, the next step is to employ Fisher Discriminant Ratio (FDR) to select the optimal feature subset in Var_{new} to improve the discriminatory power of selected features between two classes. The FDR is presented as:

$$FDR = \frac{(m_1 - m_2)^2}{(\sigma_1 + \sigma_2)^2} \tag{4}$$

where m_1 and m_2 are the mean values and σ_1 and σ_2 is the standard deviation of a feature in two classes. Then all variables will be ranked by their FDR values $FDR_1, FDR_2, ..., FDR_s$ in descending order. The number of selected features l are determined by,

$$FDR_i = \frac{FDR_i}{\sum_{i=1}^{s} FDR_i} \tag{5}$$

$$l = \{min\{k\} \mid \sum_{i=1}^{k} FDR_i >= \zeta\} \tag{6}$$

where ζ is the cumulative FDR threshold for feature selection. Therefore, l sensor variables are carefully selected in Section 3.

4. Uniform Dynamic Time Warping

The precursor mining phase aims to identify precursors that lead to unstable approach events, but due to differences in tasks and weather conditions during the descending phase, the resulting time series may have variable lengths, which is not suitable for modeling that relies on equal-length inputs. The standard dynamic time warping (DTW) [12] can find the shortest path to align variable-length time series, but it cannot ensure that all warped time series are comparable because they may be aligned to different indices of the reference signal. Therefore, this paper proposes an extension of DTW called uniform dynamic time warping (UDTW), which can transform variable-length time series into equally-length ones by ensuring that every element in the reference signal is matched to an index in the query signal.

In the specific context of PMP, UDTW defines an alignment match between a reference signal and a query signal inferred from the ratio altitude (RA) high above touchdown (HAT) with a fixed length. This alignment matrix is then applied to the remaining variables. RA is chosen as the reference variable because pilots must strictly follow standard tasks during flight phases, and the tasks are always relative to a specific RA that offers direct instructions. For instance, a stable approach must be completed at around 1000 feet HAT in instrument meteorological conditions or 500 feet HAT in visual meteorological conditions. Therefore, it is reasonable to use RA HAT as a reference to rearrange other variables and ensure that all warped time series have equal lengths.

Suppose Reference signal $R \in \mathbb{R}^{k\times 1}$, Query signal $Q \in \mathbb{R}^{g\times 1}$, and the length of R is p, the UDTW algorithm can provide a reference alignment matrix $\{I, q', R, Q'\}$, X_{input}, then applied into other variables'. Therefore, X_{input} in all flight recordings could be represented as equal-length time series as same as the RA HAT reference. The specific algorithm can be viewed in Algorithm 2 below.

Algorithm 2 Uniform Dynamic Time Warping

Input: Reference signal R, query signal Q, and new reference indices $I = \{1, 2, \ldots, p\}$, $i = 0$

1: Apply standard DTW to align R and Q, and obtain their respective indices r and q
2: Obtain new indices $a = I \cap r^c$
3: **repeat**
4: Search forward to find the nearest reference index r_s to a_i and its corresponding query index q_s to r_s
5: Update $i = i + 1$
6: **until** $i = |a|$
7: Obtain warped query indices q' by merging q and q_s, and order q' according to the corresponding indices in I.

Output: Warped query signal $Q' = Q[q']$ and its respective indices q'

Note: r^c denotes the complement of set r.

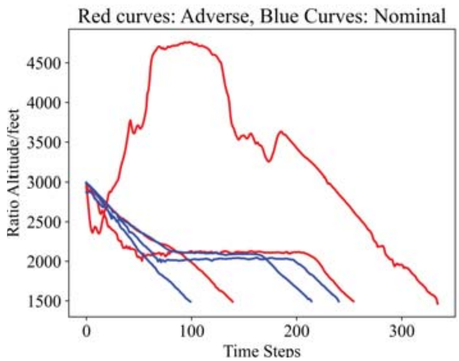

Fig. 2. Variable-length Time Series in Precursor Mining Phase

5. Experiment and Discussion

In this study, the datasets are collected from enhanced aircraft flight recorders (EAFR) of regional aircraft. The datasets contain 84 available commercial regional flights and each flight has 3056 variables sampled at 1 Hz. These variables include the states, position, and orientation of airplanes, data from Inertial Measurement Units (IMUs), configurations related to flaps, gears, and brakes, system status, engine parameters, and weather conditions. Fig. 2 represents 6 cases in PMP in terms of ratio altitude as an example, 3 for adverse events of unstable approach in red colors and 3 for nominal events of stable approach in blue colors. Flight data always have variable-length time series, which need to be processed before modeling.

Automated feature selection using statistical methods consists of two steps, feature selection based on Pearson's correlation and FDR. As shown in Algorithm 1, the number of variables is cut into 1018 due to low standard deviations. These variables may not be very informative or discriminating in distinguishing between different classes, which contribute less information to modeling and should be removed ahead. Then, feature selection based on correlation analysis helps to reduce the variables' dimension to 693. The correlation matrix heatmaps in Var and Var_{new} are shown in Fig. 3, separately. In the heatmaps, col-

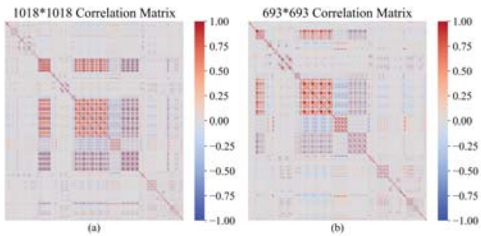

Fig. 3. Heatmaps of correlation matrix (a) before feature selection ; (b) after feature selection

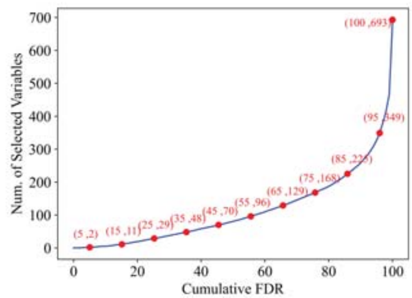

Fig. 4. Num. of Selected Features VS Cumulative FDR

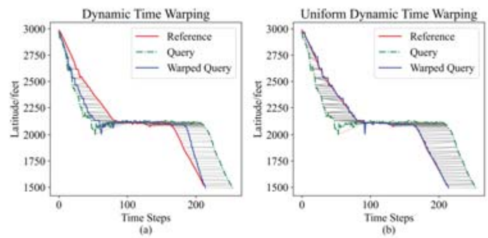

Fig. 5. Results for (a) DTW and (b) UDTW

ors approaching red indicate a positive correlation between two variables, while colors approaching blue indicate a negative correlation. The intensity of the color reflects the strength of the correlation, with darker colors indicating a stronger correlation. As seen from Fig. 3(a), there exist many highly correlated variables whose colors are close to red and blue. After applying a correlation-based method for dimension reduction, the number of variables was reduced by approximately 32 %, from 1018 to 693. It also can be roughly judged from the lower areas of red and blue colors in Fig. 3(b), compared to Fig. 3(a). Algorithm 1 achieves a good performance in an initial dimension reduction by removing redundant variables with highly correlated to others.

In the following, we calculate each variable's FDR and obtain cumulative FDR to select a feature subset according to a specific cumulative FDR threshold. Fig. 4 shows the number of selected features among different cumulative FDR. When the threshold was set to 95 %, almost half of the variables were removed, leaving 349 variables in the selected feature subset, as shown in Fig. 4. The selected feature subset is chosen based on their FDR values, with those above the specified cumulative FDR threshold retained in the subset. Applying the FDR method in feature selection is a simple and efficient way to greatly reduce dimension by removing less informative features.

Furthermore, as previously mentioned, time series in PMP are always of variable lengths, as depicted in Fig. 2. To address this issue, we propose a UDTW algorithm that converts variable-length time series into equal-length time series. Results in Fig. 5 state that the UDTW offers equal-length time series in the reference variable ratio altitude. Moreover, the alignment matrix obtained from UDTW, represented as dash curves, offers a more concise matching approach with the warped query being closer to the reference.

What's more, Fig. 6 displays the local views of the above alignment matrix between reference indices and query indices for DTW and UDTW in specific points: the first 10 and last 10 points. UDTW ensures that each element in the reference signal is matched with an index in the query signal. For instance, in Fig. 6(a), the fourth point in reference indices is present in UDTW but missing in DTW. Also, Fig. 7 indicates that UDTW consistently exhibits smaller distances between the reference and query signal in terms of ratio altitude compared to DTW. In addition, the red points in Fig. 7 indicate the distances for the adverse samples, revealing that the distances in adverse samples are relatively larger than those in nominal

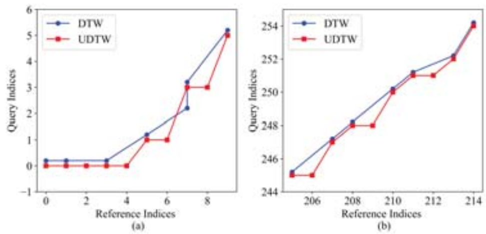

Fig. 6. Local views of alignment matrix between reference indices and query indices for (a) the first 10 points (b) the last 10 points

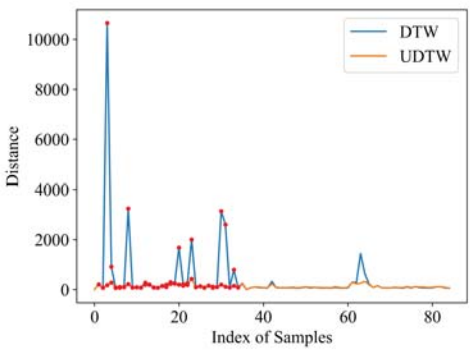

Fig. 7. Distance in all samples for (a) DTW and (b) UDTW

samples, as anticipated. In summary, this idea in UDTW not only achieves equal-length time series but also results in smaller distances between their reference and query signals.

6. Conclusion

This paper proposed a pipeline for preparing data in precursor-based unstable approach events. The pipeline generated lower-dimensional and equal-length time series as inputs in PMP, along with labeled outputs of unstable approach in UAIP. The results obtained from the real flight recorder demonstrated that the proposed data preparation method effectively performed in providing standardized inputs and outputs for modeling. Furthermore, it could be applied to data preparation for other adverse events occurring during take-off and climb phases, where pilots must adhere to strict standard tasks, similar to the context discussed in this paper.

However, although the number of variables in a flight is significantly reduced by automatic feature selection and FDR analysis, this reduction is only a rough statistical process and remains independent of further modeling. Thus, it is necessary to fine-tune the variables in conjunction with the modeling process to achieve a lower-dimensional and more comprehensive feature representation.

References

1. Abd-Alsabour, N. (2018). On the role of dimensionality reduction. *J. Comput. 13*(5), 571–579.
2. Ackley, J. L., T. G. Puranik, and D. Mavris (2020). A supervised learning approach for safety event precursor identification in commercial aviation. In *AIAA Aviation 2020 Forum*, pp. 2880.
3. Boeing (2014). Asias unstable approach criteria thresholds. `https://www.icao.int/SAM/Documents/2014-UNSTAPPCH/BOEING%20ASIAS%20UA%20Criteria%20Threshold%20Chart%20Runway%20Safety%20Seminar.pdf`. Accessed: March 26, 2023.
4. Eesa, A. S., Z. Orman, and A. M. A. Brifcani (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. *Expert systems with applications 42*(5), 2670–2679.
5. European Union Aviation Safety Agency (2019). Guideline for the implementation of flight data monitoring precursor- unstable approach. `https://www.easa.europa.eu/sites/default/files/dfu/Study_WGB_Precursors_20191021.pdf`. Accessed: March 26, 2023.
6. Foundation, F. S. (2000). Flight safety foundation approach-and-landing accident reduction task force 2000.11-fsf alar briefing note 7.1-stabilized approach. Technical Report ALAR Briefing Note 7.1, Flight Safety Foundation. Accessed: March 26, 2023.
7. Foundation, F. S. (2010). Flight safety foundation approach-and-landing accident reduction

toolkit update 2010.5. Technical Report ALAR-Toolkit 2010.5, Flight Safety Foundation. Accessed: March 26, 2023.

8. International Civil Aviation Organization (2017). Go-around decision-making and execution project 2017-new stabilized approach and go-around guidelines. Technical Report Doc 10184, International Civil Aviation Organization, Montreal. Accessed: March 26, 2023.
9. Janakiraman, V. M. (2018). Explaining aviation safety incidents using deep temporal multiple instance learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 406–415.
10. Juarez Garcia, E., C. Stephens, and N. J. Napoli (2022). Detecting high-risk anomalies in aircraft dynamics through entropic analysis of time series data. In *AIAA AVIATION 2022 Forum*, pp. 3257.
11. Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artificial intelligence 97*(1-2), 273–324.
12. Mizutani, E. (2006). The dynamic time warping algorithms. In *Mechanical Engineering Seminar, Tokyo Metropolitan University*, pp. III.
13. Yu, J., J. Yang, D. Tang, and J. Dai (2018). An optimal burn-in policy for cellular phone lithium-ion batteries using a feature selection strategy and relevance vector machine. *Energies 11*(11), 3021.
14. Zebari, R., A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J. Appl. Sci. Technol. Trends 1*(2), 56–70.