Data-Driven Long-Landing Event Detection and Interpretability Analysis in Civil Aviation

Article *in* IEEE Access · June 2022

DOI: 10.1109/ACCESS.2022.3182796

5 authors, including:

Xiong Yang
Civil Aviation University of China
4 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Jin Ren
Shenzhen Polytechnic University
11 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Haigang Zhang
Shenzhen Polytech
37 PUBLICATIONS **481** CITATIONS
SEE PROFILE

Jinfeng Yang
Shenzhen Polytech
94 PUBLICATIONS **1,807** CITATIONS
SEE PROFILE

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 16, 2022, accepted June 4, 2022, date of publication June 13, 2022, date of current version June 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182796

Data-Driven Long-Landing Event Detection and Interpretability Analysis in Civil Aviation

XIONG YANG[1], JIN REN[1,2], JUNCHEN LI[3], HAIGANG ZHANG[1], AND JINFENG YANG[1]

[1]Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic, Shenzhen 518055, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[3]College of Engineering, Southern University of Science and Technology, Shenzhen 518055, China

Corresponding author: Jinfeng Yang (jfyang@szpt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62076166, in part by the China Postdoctoral Science Foundation under Grant 2021M703371, in part by the Shenzhen Science and Technology Program under Grant RCBS20200714114940262, in part by the Postdoctoral Foundation Project of Shenzhen Polytechnic under Grant 6021330002K, and in part by the General Higher Education Project of Guangdong Provincial Education Department under Grant 2020ZDZX3082 and Grant 2020ZDZX3085.

ABSTRACT Long-landing Events (LLEs) can occur as a result of pilot's improper operation, resulting in shorter available runways and higher operating costs. The LLE can be effectively pinpointed by analyzing data from the Quick Access Recorder (QAR), which records all of the pilot's operations during takeoff and landing. Traditionally, domain experts inspect LLEs by manually setting thresholds on uni-dimensional data. However, they cannot detect effectively the defects caused by the pilot's maneuvering technique because the potential mutual information between different features in the large amount of data is not considered. This paper proposes a data-driven LLE detection and causation analysis workflow, which can automatically mine and analyze the mutual information, to overcome the existing problems. Firstly, a dataset is established based on the extracted QAR data from 2002 flights, considering the landing phase of the aircraft. Subsequently, this paper proposes a Hybrid Feature Selection (HFS) method for selecting features that are highly correlated with LLEs in both supervised and unsupervised ways. A categorical Light Gradient Boosting Machine (LGBM) with a Bayesian optimization (LGBMBO) model is used to determine the performance improvement. Furthermore, the model is visualized to analyze the marginal effect of key parameters for the LLEs by using SHapley Additive exPlanations (SHAP). The experimental results demonstrate that our model reduces computational cost and achieves better performance. Additionally, this paper demonstrates that LLEs can be avoided during the landing phase by maintaining the appropriate descent speed, aircraft altitude, and descent angle.

INDEX TERMS Flight safety, long-landing event, anomaly detection, shapley additive explanations, feature engineering, machine learning.

I. INTRODUCTION

The aircraft landing phase is the period with the highest accident rate, since it is closely associated with the geographical environment of the airport, weather conditions, pilot's operational expertise, and level of maintenance of the aircraft parts [1]. The accident data analysis indicates that the Runway Excursion (RE) category is the most common type of accident reported each year, which occurs during the landing phase [2]. The RE accounted for 14.9% of all accidents and included 50% of all fatal accidents, with 44 fatalities in 2019 [3]. The long-landing event (LLE) is one of the corresponding RE accident indicators. LLEs occur when the aircraft's landing distance lies beyond the normal range; it shortens the runway's available length, posing a significant risk. Even worse, the aircraft may run off if the pilots lack the necessary proficiency, leading to unpredictable consequences [4], [5]. Therefore, the detection of LLEs and their causes can help in preventing RE accidents, benefit airlines management and pilot training, and help in the development of aviation safety regulations.

The Quick Access Recorder (QAR) is an onboard data recording device which collects the entire flight information via sensors placed throughout the aircraft during the flight

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Luo.

phases. Research on the QAR has gained considerable attention from civil aviation safety experts for the evaluation of flight quality and detection of risk events due to its characteristics of low cost and portability [6]. The Exceedance Detection (ED) method is one of the most efficient methods to monitor the QAR data; it includes various trigger logic expressions based on operational manuals, training programs, and risk assessment procedures on uni-dimensional data [7]. However, it depends on the experience of domain experts, making it difficult to effectively and efficiently utilize the potential mutual information between different features in the huge amount of data; it is thus difficult to effectively identify the defects caused by the pilot's maneuvering technique [8]–[11]. Consequently, how to unify the ability to detect anomalous events and the inference of their causality presents considerable challenges in the analysis of large amounts of aviation data.

The paper attempts to break down the challenges into three sub-problems as follows: selecting the features most relevant to LLEs from a large number of features; determining whether the proposed anomaly detection model is optimal when compared with other machine learning algorithms; lastly, excavating the correlation between key parameters and LLEs. This study solves these problems by employing a data-driven workflow based on an interpretable machine learning approach. The main contributions of this paper are as follows:

- The QAR data from 2002 flights is collected to build a dataset of LLE anomalies that occur during the landing phase.
- A fast and accurate Hybrid Feature Selection (HFS) algorithm, which presents the advantages of both unsupervised and supervised algorithms, is proposed for the QAR data to reduce data redundancy and determine the most relevant features.
- This study adopts Light Gradient Boosting Machine (LGBM) by incorporating Bayesian Optimization (BO) to overcome the problem of sample imbalance in the abnormal data.
- Employing the interpretable machine learning framework, i.e., SHapley Additive exPlanations (SHAP), to excavate the correlations between the parameter within the anomaly detection model.

The rest of the paper is structured as follows. Section 2 presents a description of the related works. Section 3 depicts the proposed system architecture, including dataset construction, feature engineering, anomaly detection model, and interpretability analysis. Section 4 presents the experimental results, and explains and analyzes the models using SHAP. Lastly, Section 5 presents the conclusion.

II. LITERATURE REVIEW

A. ANOMALY DETECTION APPROACHES FOR AVIATION DATA

Extensive research has been conducted to detect aviation abnormal event. The anomaly detection models in aviation data can be classified into three categories based on the findings of previous studies: clustering-based, regression-based, and classification-based models. Clustering-based models employ an unsupervised learning algorithm which analyzes the similarity between aviation data based on different rules. The Multiple Kernel-based Anomaly Detection (MKAD) algorithm, developed by Das *et al.* [12], is one of the first unsupervised methods designed for anomaly detection in aviation. The MKAD can detect potential safety anomalies in very large databases of discrete and continuous data based on kernel functions and One Class SVM (OC-SVM) [13]. Li *et al.* [14] applied a cluster-based Anomaly Detection (ClusterAD) method on multidimensional flight data, which is based on the DBSCAN algorithm. The algorithm can automatically identify multiple types of flight operation patterns unlike the MKAD. Puranik *et al.* [15] used the nearest Local Outlier Factor (LOF)-based neighborhood method to quantify the anomaly degree of flight data clustering results.

A regression-based models feed multiple independent variable parameters as inputs to fit a dependent variable, and outliers are identified by comparing the difference between the model's predictive output and the ground truth of the dependent variable. The Semi-Markov Switching Vector AutoRegressive (SMS-VAR) model proposed by Melnyk *et al.* [16], performs regression fitting of multidimensional flight data sequences, to detect anomalies based on the dissimilarities between the model predictions and the data observations. Tong *et al.* [17] developed a deep prediction model based on Long Short-Term Memory (LSTM), and then used Mean Square Error (MSE) as an evaluation criterion to identify hard landing events. Puranik *et al.* [18] used Random Forest (RF) models to establish predictive model and identified anomalies using quartiles for the MSE between the output of the model and the true value.

The classification-based models vary from the clustering or regression models, which require anomaly labels that are assigned by experts in the aviation industry. However, the classification models are advantageous as they can objectively compare the performance of each model. Consequently, they present considerable potential in the anomaly detection of aviation data. Memarzadeh *et al.* [19] developed a Convolutional Variational Auto-Encoder (CVAE), which is an unsupervised deep generative model for anomaly detection in aviation high-dimensional time-series data. Rey *et al.* [20] implemented an XGBoost model in multiple airports to detect runway anomalies; the model was sufficiently accurate to select the contributing factors and present an extensive analysis.

B. ANOMALY ANALYSIS APPROACHES FOR AVIATION DATA

The aim of anomaly analysis approaches for aviation data are to find the factors which contribute to aviation accidents. Wang *et al.* [4] identified the key flight parameter characteristics of LLEs by analyzing the QAR data while proposing

TABLE 1. The different kinds of anomaly detection approaches for aviation data.

Category	Characteristics	Method	Author	Year
Clustering-based	Unsupervised learning approach, anomalies are distinguished based on similarity between samples	MKAD	S.Das [12], [13]	2010
		ClusterAD	L.Li [14]	2015
		LOF	T. G. Puranik [15]	2017
Regression-based	Continuous values are taken as supervision, samples beyond a certain range are considered as outliers	SMS-VAR	I. Melnyk [16]	2016
		LSTM	C. Tong [17]	2018
		RF	T. G. Puranik [18]	2020
Classification-based	Discrete anomaly labels are taken as supervision	CVAE	M. Memarzade [19]	2020
		XGBoost	M. Rey [20]	2021

preventive measures from the pilot operation perspective. They developed the logistic regression and the linear regression models to further analyze the correlation between touchdown distance and these flight parameter variables. In 2018, Wang *et al.* [21] analyzed the potential correlations between the flare operation and the landing performance on long and hard landing events by employing two regression models. Barry [22] implemented Bayesian networks to the airborne recorded flight data, to effectively perform a quantitative risk assessment of aviation events by selecting features directly related to the known the risk causal factors.

The approaches presented in the previous studies are difficult to implement uniformly in a single model. The anomaly detection approach can accurately detect anomalies, but cannot interpret them. Conversely, the anomaly analysis approach can explain anomalies, but presents limitations such as small data samples and non-uniform evaluation indexes. Consequently, evaluating the effectiveness of the model for large-scale data applications is difficult. This paper combines the advantages of these two approaches to integrate risk event detection and cause analysis, to address this problem.

III. METHODOLOGY

A. SYSTEM ARCHITECTURE

In this study, it is assumed that the aircraft is in good operating condition, that the pilots are able to maneuver the aircraft normally, and that the data recorded by the QAR are consistent with the pilot's maneuvers, so as to investigate the relationship between LLEs and pilot operation characteristics. In addition, the influence of external factors, such as runway conditions and weather conditions, are not the focus in this paper. The proposed anomalous event detection and analysis scheme consists of four steps: dataset construction, HFS-based feature selection, anomaly detection model, and analysis of experimental results, including performance evaluation and interpretability analysis, as shown in the Figure 1.

B. DATASET CONSTRUCTION

1) PARAMETERS

The parameters are selected based on two aspects: the types of parameters in the QAR data and the parameters related to the landing phase. The QAR records four types of flight parameters: alarm, environment, status, and operation parameters. Among these, the status and operation parameters are the most important indicators, which represent the pilot's operational expertise. Subsequently, the landing phase of the aircraft can be divided into three phases: the glide slope, the flare-out, and the deceleration taxiing phase [23]. During the glide slope phase, the aircraft transitions from gliding to an approximate flying state, generally employing the small throttle sliding method. During the flare-out phase, the aircraft mainly adjusts its speed, attitude, and angle for landing. Lastly, during the deceleration taxiing phase, the aircraft's drag must be increased to reduce the speed and the landing glide distance as much as possible. The glide slope phase is affected by the vertical height of the aircraft as it passes the runway threshold, the gliding angle, and other factors. The deceleration taxiing phase after grounding, is affected by the grounding speed of the aircraft, the deceleration device, the slope of the airport runway, the wind speed on the ground, and other factors. Consequently, the parameters which characterize the flight dynamics of the aircraft during the landing phase are selected from the aircraft status and the operation parameters of the QAR data. These parameters can be classified into the velocity-type, position-type, status-type, engine-type, and deceleration-type. Subsequently, the relevant parameters are selected from more than 2000 parameters in the QAR data, with 151 columns; the samples of the selected parameter types and examples are presented in the Table 2.

TABLE 2. the samples of the landing-related parameters.

Category	Parameters	Unit
Velocity	Ground Speed (_GS)	KNOTS
	Computed Speed (_CAS)	KNOTS
	Instantaneous Vertical Velocity (_IVV)	FT/MIN
Position	Radio Altitude (_ALT_RADIO)	FEET
	Glide Angle Minus A/C Dev (_GA_M_3D)	DEGS
	Glide Devitation in dots (_GLIDE)	DOTS
Status	Flap Angle (_FLAP)	DEG
	Elevator Angle (_ELEV)	DEG
	Pitch Angle (_PITCH)	DEG
Engine	N1, N2	%RPM
Deceleration	Spd Brake Handle Posn (aSPDBRKPOS)	DEG

2) DATA EXTRACTION

The LLEs typically occur during the landing phase, which is the most common phase for aviation accidents. The aim of the data extraction process is to select the landing phase from the entire flight data frame. The process is performed according to the same rules for each flight in this study: each flight must have the same dimensions and a unified sampling point. During the landing phase, the unified point is defined as the maximum aircraft braking value, which is usually considered as the landing point of the aircraft. For each flight, 49 time-stamps are taken before the landing point, and ten

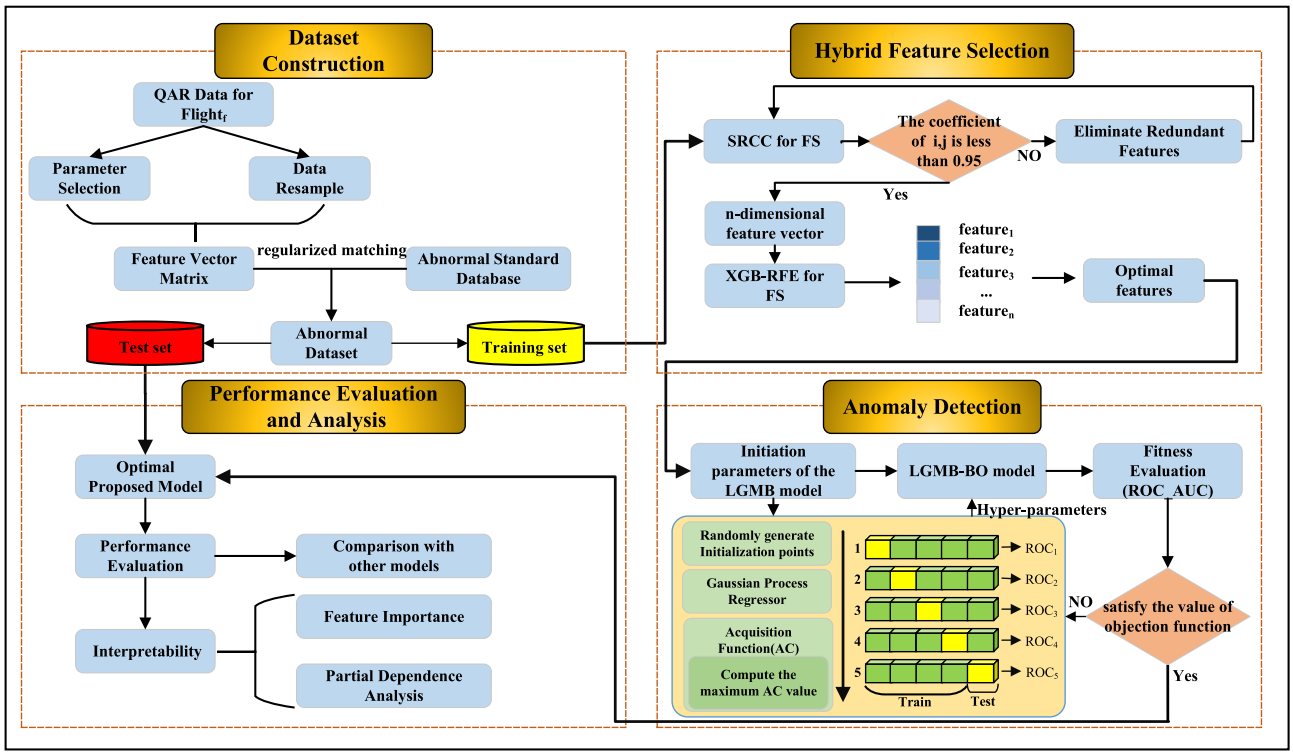


FIGURE 1. System architecture.

time-stamps are taken after the landing point to establish a feature vector matrix for the flight. Therefore, a total of 60-time stamps are extracted per flight. The range of the values taken include the aircraft's altitude, which are within 600 feet of the runway height. For each flight, f, a vector at the time $t(t \in (0, 60))$, is defined in the (1).

$$k_f^t = [k_1^t, k_2^t, \ldots, k_p^t], \tag{1}$$

where k_p^t means the flight value of the p^{th} flight parameter at time step t.

3) LABELLING OF ANOMALOUS EVENT THROUGH REGULARIZED MATCHING

The QAR and the abnormal events databases are obtained using the airline's decoding software, AirFase. The QAR database includes multi-dimensional time-series data, such as flight registration numbers, airline registration numbers, aircraft type numbers, flight execution dates, parameters records, and flight operation time. The AirFase software creates the abnormal events database by analyzing the documents of the Flight Standards Department of the Civil Aviation Administration of China (CAAC) [24] and the airline QAR monitoring standards, which contains the flight registration number, aircraft type number, abnormal phase, abnormal events occurrence events and abnormal level, and so on. In the abnormal events database, an LLE is defined as the aircraft landing which occurs when the landing point is more than 700 meters from the runway threshold. Lastly, the QAR and abnormal events databases are matched to obtain the abnormal datasets through regularization filtering, matching

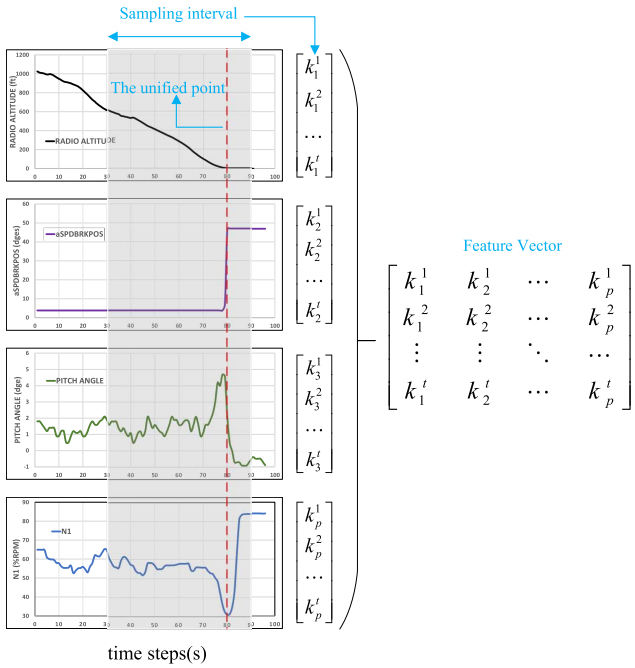


FIGURE 2. Data extraction process.

flight numbers, determining abnormal time, labeling abnormalities, and so on.

C. HYBRID FEATURE SELECTION

Several sensors perform the same function in an aircraft. These sensors ensure that when one fails, the others remain in a working condition. The redundant mechanism ensures the

safety of the aircraft. However, it adds unnecessary complexity in the data analysis. Therefore, this paper combines the Unsupervised Feature Selection (UFS) and Supervised Feature Selection (SFS) algorithms to design an HFS algorithm for high-dimensional aviation data.

1) UNSUPERVISED FEATURE SELECTION FOR ELIMINATING REDUNDANT FEATURES

The UFS evaluates the relevance of the features by analyzing the internal structure between the features [25]. The unsupervised method of the Spearman rank correlation coefficient (SRCC) feature selection filters redundant features without using label information. The assumptions of Spearman's correlation are that the data must be at least ordinal and that one variable must be monotonically correlated with another variable [26]. The QAR data exactly satisfy the assumptions of SRCC. Therefore, the variables with high monotonicity can be filtered by SRCC as a way to remove the redundant features from the QAR data. The SRCC includes values ranging from -1 to 1. The values close to 1 represent a strong monotonic increasing relationship between two variables, while those close to -1 represent a strong monotonic decreasing relationship. When the correlation values close to 0, it indicates that there is no correlation between two variables [27]. The absolute value of the correlation matrix is taken because both monotonic increasing and monotonic decreasing correlations are relevant in the current data set. The variables with high monotonicity are filtered out by SRCC as a way to remove redundant features from the QAR data. The SRCC is defined as follows,

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{2}$$

where d represents the corresponding subtraction of the elements in the two variables, a and b to obtain a ranking difference set, and $d_i = A_i - B_i$. N represents the total number of individual elements, which in the instance, is the number of rows of the QAR data. If the correlation coefficient of the two variables is closer to 1, it indicates they have a strong correlation, which can be considered as redundancy. The correlation threshold is set to 0.95 based on the experiments conducted in this study.

2) SUPERVISED FEATURE SELECTION FOR SELECTING THE MOST RELEVANT FEATURES

The SFS selects more relevant features based on the correlation between the features and class labels, resulting in better performance in the subset [28]. This study proposes a high-performance automatic supervised feature selection method, which is obtained by cross-validation using the eXtreme Gradient Boosting (XGBoost) and Recursive Feature Elimination (RFE) methods. XGBoost is an ensemble machine learning algorithm based on a gradient boosting decision trees (GBDT) framework [29]. GBDT is a type of integrated learning boosting, where the boosting method trains a bunch of individual models continuously, each model learning from the error of the previous model [30]. Assuming that the GBDT model input sample is

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{3}$$

where F represents the set of all weak learners.

Compared with GBDT, XGBoost introduces a regular term in the cost function for controlling the complexity of the model as follows,

$$\Omega(f_t) = \mu T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2. \tag{4}$$

The objective function of XGBoost is composed of two parts: training loss and regularization, as represented in the (5).

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{5}$$

Then a Taylor second-order expansion of the loss function is performed to estimate the loss function. By approximating the objective function, the whole model has a faster convergence rate. Then, the optimal tree structure corresponding to the objective function is defined as follows,

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T, \tag{6}$$

where G_j and H_j represent the cumulative sum of the first-order and second-order partial derivatives concerning the approximate loss function, respectively.

XGBoost then calculates the importance of the features by summing the number of splits of the features in each tree to select the important features.

RFE is a wrapper feature selection method, which aims to select the best or worst features based on the coefficients, remove the selected features, and repeat the process in the remaining set of features until all the features are traversed [31]. Firstly, the estimator is trained on the initial set of features to determine the importance of each feature. The least important features are then removed from the current set of features. The process is repeated recursively on the pruned set until the final number of features to be selected are obtained. The XGBoost and RFE feature selection methods form a set of embedding methods for feature selection. The procedure of the combined XGBoost and RFE feature selection algorithms are as follows:

Step1. The XGBoost model is trained on the original features, and each feature is assigned a weight, i.e., feature importance.

Step2. The features with the smallest absolute value weights are moved out of the feature set space.

Step3. The process is repeated recursively until the number of remaining features reach a set feature count threshold.

Step4. The performance of the selected feature subset is evaluated and lastly, the feature subset with the best performance is selected.

D. ANOMALY DETECTION MODEL

1) LGBM MODEL FOR ANOMALY DETECTION

The LGBM algorithm is an integrated algorithm, which is based on the gradient boosting framework proposed by Microsoft Research. The LGBM presents a higher accuracy and faster model training time when compared with XGBoost and is better optimized for the construction of decision trees [32]. Thus, the main computational principles of both algorithms are identical. The improvements made by the LGBM in the construction of decision trees are reflected in the proposed GOSS algorithm, EFB algorithm, Histogram algorithm, and leaf-wise leaf node generation method. The GOSS algorithm is used to reduce the training sample size, the EFB algorithm is used to reduce the sample dimension or reduce the sample features, and the Histogram algorithm is used to identify the best splitting point. Particularly, the leaf-wise growth approach is computationally less expensive and avoids overfitting when compared with the XGBoost approach use level-wise leaf node growth.

2) BAYESIAN OPTIMIZATION ALGORITHM FOR HYPERPARAMETER TUNING

The LGBM is tedious and complicated since the parameters can be only adjusted manually, which deteriorates the performance of the algorithm. Therefore, in this study, the parameters are automatically adjusted by using the Bayesian optimization algorithm to ensure that the LGBM algorithm achieves the best results. The BO works by constructing a posterior distribution of functions (Gaussian process) that best describes the function that wants to optimize [33]. As the number of observations grows, the posterior distribution improves, then the model will focu on the region of the parameter space with the most promising validation scores.

The BO contains two core parts: the probabilistic surrogate model (PSM) and the acquisition function (AC). The PSM can be divided into parametric and non-parametric models based on the model parameters. The Gaussian process (GP), which is a non-parametric model with a variable number of parameters when compared with the parametric one, can better describe the unknown objective function and strongly fit the function due to its greater flexibility and scalability.

The acquisition function (AC) is an important basis to determine the next evaluation point using the Gaussian Process-Upper Confidence Bound (GP-UCB), as defined in the (7),

$$x_{t+1} = \arg\max \mu_t(x) + \sqrt{\beta_t}\delta_t(x), \tag{7}$$

where the parameter β_t represents a constant for equilibrium exploration and exploitation, $\mu_t(x)$ represents the mean, and $\delta_t(x)$ represents the standard deviation.

TABLE 3. Confusion matrix description.

	Actual anomaly	Actual normal
Predicted anomaly	TP	FP
Predicted normal	FN	TN

E. MODEL INTERPRETATION

This paper applies the machine learning interpretability tool, SHAP, to analyze the LGBMBO model. SHAP consists of several machine learning interpretable strategies proposed by Lundberg and Lee in 2016 [34]. It is derived from a game-theoretic approach called Shapley values, which fundamentally interpret the output of any machine learning model in terms of whether it is parameterized or not. SHAP can explain the working concept of machine learning by reverse-engineering the output of any predictive model. SHAP interprets the model's predicted value as the sum of the imputed values of each input feature, which is a linear model expressed as an additive feature imputation method. The additive feature imputation method is defined as follows,

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j, \tag{8}$$

where $f(x)$ represents the original model, $g'(x)$ represents the explanatory model, x' represents the simplified input for variable x, and M represents the number of input features. In the SHAP interpretable model, ϕ_0 represents the mean value of the mapped SHAP for all training samples, ϕ_j represents the SHAP value corresponding to variable j, and x_j' indicates the value of 1 when variable j is selected and 0 otherwise.

The SHAP value ϕ_j is defined as follows,

$$\phi_j = \sum_{s \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\} - f_x\{S\}], \tag{9}$$

where N represents the set of all features, S represents a sequential subset of N, and M represents the number of input features in the (8).

The SHAP value is used to measure the degree of influence of a feature on the predicted result. A positive SHAP value indicates that the variable has a promoting effect on the prediction value in the sample. Whereas, a negative SHAP value indicates that the variable has a restraining effect on the prediction value in the sample. A larger absolute value of the SHAP value indicates that the variable has a greater influence on the predicted value in the sample.

IV. EXPERIMENTS AND ANALYSIS

A. EVALUATION MEASURES

In this paper, the confusion matrix description is shown in Table 3.

For unbalanced data, a single accuracy metric cannot sufficiently reflect the performance of the model because the accuracy (Acc) of the majority and minority sample classification must be considered together. Therefore, the F1 score (F1) and the Matthews correlation coefficient (MCC) are used to analyze the confusion matrix, and the receiver

operating characteristic (ROC) curve and the precision-recall (PR) curve, are used to determine the performance of the classifier.

The following definitions give the evaluation indicators of the classifier:

1) Acc is defined as:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP}. \tag{10}$$

2) F1 is defined as:

$$F_1 = \frac{2TP}{2TP + FP + FN}. \tag{11}$$

3) MCC is defined as:

$$MCC = \frac{TP/N - S \times P}{\sqrt{P \times S(1-S)(1-P)}}, \tag{12}$$

where N, S, P can be defined as follows,

$$N = TN + TP + FN + FP, \tag{13}$$

$$S = \frac{TP + FN}{N}, \tag{14}$$

$$P = \frac{TP + FP}{N}. \tag{15}$$

4) The ROC curve presents a curve with FPR as the y-axis and TPR as the x-axis, PR curve presents a curve with precision rate as the y-axis and recall rate as the x-axis, which can be defined as follows,

$$FPR = \frac{FP}{FP + TN}, \tag{16}$$

$$TPR = \frac{TP}{TP + FN}, \tag{17}$$

$$precision = \frac{TP}{TP + FP}, \tag{18}$$

$$recall = \frac{TP}{TP + FN}. \tag{19}$$

The area enclosed by ROC curve and PR curve are ROC_AUC and PR_AUC respectively.

Acc, F1, ROC_AUC and PR_AUC are in the range of [0,1], and MCC is in the range of [-1,1], with closer to 1 indicating better algorithm performance and vice versa.

B. EXPERIMENTAL SETTINGS

In this experiment, a total of 20 GigaByte (GB) QAR data elements were collected from 2002 flights of an airline during actual operation. Each flight obtained 60 time stamps during the landing stage through resampling, for a total of 120120 sample-points. In the constructed anomaly dataset, the ratio of long landing data to normal data is 1:17, indicating that the dataset is unbalanced. For the dataset processing, the original data set is first divided into the training set and test set through stratified sampling in the ratio of 8:2. Subsequently, the HFS algorithm is used to select the appropriate features in the training set, while the test set uses the feature matrix after the hybrid feature selection to ensure the objectivity of the test results. The experiment is conducted using a computer with Windows 10, 2.9 GHz, Inteli7 processor, RTX3080 GPU, python 3.8.10, and compiler as VScode environment using the jupyter notebook.

C. FEATURE SELECTION ANALYSIS

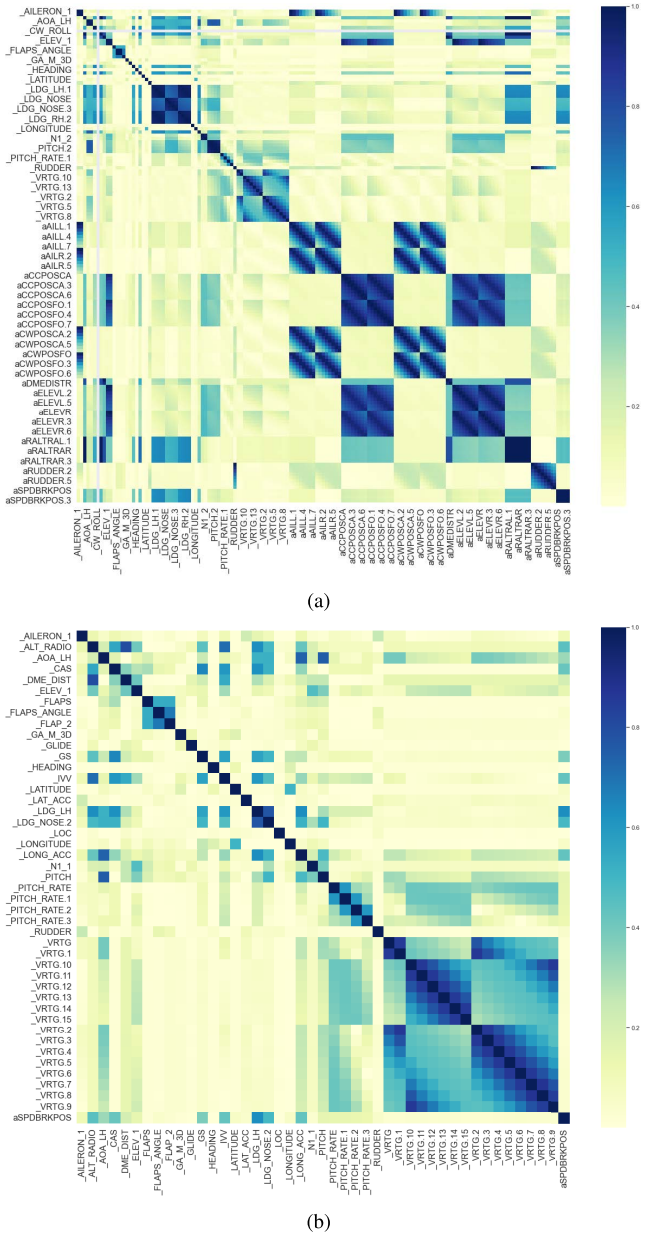
1) DEPICTION OF THE RESULT OF FEATURE SELECTION

The feature number is reduced from 151 dimensions to 62 dimensions after the SRCC deletes the redundant features from the raw data. The following figure presents the comparison between the raw data and the SRCC features. In the heat map, the darker the color, the greater the correlation between the features. It can be observed from Figure 3 (a), that there are more dark areas in the heat map of the raw data, indicating that there are more redundant values in the raw data. Subsequently, it can be observed from Figure 3 (b) that the dark areas become significantly reduced after the SRCC processing, and the color becomes significantly lighter, indicating that the redundant values can be effectively eliminated after the raw data is subjected to the SRCC processing.

The XGB+RFE, a supervised feature selection method, is used to select features associated with anomalous events after removing redundant values using the UFS approach. This study adopts a cross-validation approach which can automatically select the best subset of features to ensure that the proposed algorithm automatically selects the appropriate features. Furthermore, the ROC_AUC score with 5-fold cross-validation is used as the objective function. Figure 4 presents the selected key feature parameters. It shows the number of different feature subsets and cross-validation scores; the point where the vertical line crosses the curve is the point with the largest cross score. And the horizontal one coordinate corresponding to the vertical line is the best feature subset selected. It can be observed from Tabel 4 that the algorithm selects the best feature subset number of 16. Figure 5 presents the hot map obtained after the HFS processing; it can be observed that there are fewer data dimensions and a lower correlation between features after the XGB+RFE processing when compared with the SRCC processing.

2) FEATURE SELECTION PERFORMANCE COMPARISON

This study compares different feature selection methods separately in terms of performance and computation time to demonstrate the necessity of feature selection. Classical feature selection algorithms such as XGBoost, Random Forest (RF) [35], and Principal Component Analysis (PCA) [36] are selected and then compared with the algorithm proposed in this study. The features with a feature importance is greater than 0.01 are selected when using RF in the feature selection. The decomposition vector with cumulative explained Variance greater than 95% is selected, while *n_components* is equal to 12, when using the PCA in feature dimensionality reduction. The classifier model used, is the LGBM algorithm without hyperparameter optimization, while the parameter values of LGBM are the original model values. Table 5 and Figure 6 present the comparison results.



(a)

(b)

FIGURE 3. Comparison of the hot maps of raw data and after the SRCC processing: (a) The hot map of the raw data; (b) The hot map after the SRCC processing.

The two aspects of performance and feature selection time can be compared by using the information presented in Table 5 and Figure 6. Firstly, in terms of performance, it can be observed by comparing the original data with the SRCC processed data, that Acc, F1, MCC, ROC_AUC, and PR_AUC do not significantly vary. Furthermore, the comparison of HFS with other feature selection algorithms demonstrates that the HFS method has the best performance with 98.04%, 78.20%, 0.7863%, 0.9955%, and 0.9408% for ACC, F1, MCC, ROC_AUC, and PR_AUC respectively. Particularly, the HFS effect is most pronounced for F1 scores and MCC values. The F1 scores of HFS are 4.59%, 3.037%, 2.84%, 1.13%, 60.35%, and 0.4% higher than the raw data, SRCC_FS, XGB_FS, RF_FS, PCA, and

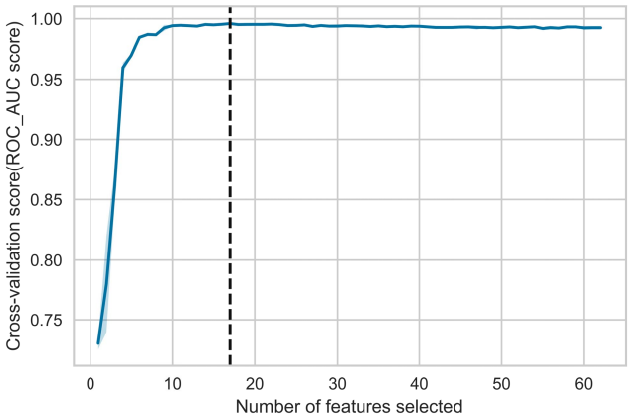


FIGURE 4. Number of key features selected by the XGB-RFE.

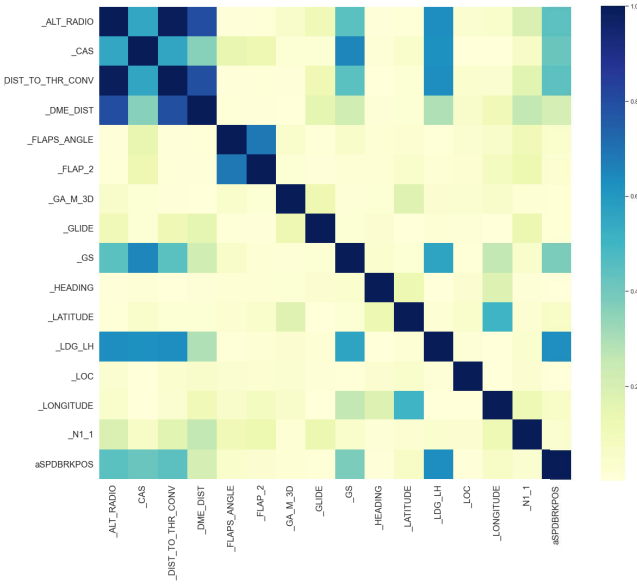


FIGURE 5. The hot map after the HFS processing.

XGB+RFE_FS. The MCC values of HFS are 0.04, 0.0268, 0.0252, 0.0098,0.5357, and 0.0032 higher than the raw data, SRCC_FS, XGB_FS, RF_FS, PCA, XGB+RFE_FS. Subsequently, the comparison of the feature selection time demonstrates that the PCA has the shortest feature selection time of 1.4 s, while XGB+RFE_FS is the algorithm with the longest feature selection time of 4301.8 s when compared with the time of XGB_FS, XGB+RFE_FS, and HFS, all three of which use the XGBoost feature selection algorithm, with 40.5 s, 4301.8 s, and 760.4 s, respectively. Therefore, the following conclusions can be drawn from the above data comparison:

1. For the data processed by using the SRCC when compared with the original data, the data dimension is significantly reduced, but the anomaly detection performance does not deteriorate;
2. The HFS method significantly improves the anomaly detection performance when compared with the SRCC algorithm. And it reduces the feature selection time by 82.3% when compared with the XGB+RFE algorithm;

TABLE 4. Best feature subset after the HFS processing.

Number	Feature name	Meaning
1	_ALT_RADIO	Radio altitude
2	_CAS	Computerd air speed
3	_DIST_TO_THR_CONV	Distance from the runway threshold
4	_DME_DIST	Distance from the DME
5	_FLAPS	Flaps angle
6	_FLAPS_2	Flaps angle_2
7	_GA_M_3D	Glide angle
8	_GS	Ground speed
9	_HEADING	Aircraft heading
10	_LATITUDE	Aircraft latitude
11	_LDG_LH	Landing gear status
12	_LOC	Localizer deviation
13	_LONGITUDE	Aircraft longitude
14	_N1	Rotor speed of engine
15	aSPDBRKPOS	Brake handle position
16	_GLIDE	Glide slope deviation

TABLE 5. Feature selection performance comparison.

FS methods	Number	Acc(%)	F1(%)	MCC	FS time(s)
raw data	151	97.77	73.61	0.7463	0
SRCC_FS	62	97.82	75.17	0.7595	40.5
XGB_FS	46	97.83	75.36	0.7611	9.4
RF_FS	19	97.96	77.07	0.7765	48.9
PCA	12	95.98	17.85	0.2506	1.4
XGB+RFE_FS	20	98.01	77.80	0.7831	4301.8
HFS	16	98.04	78.20	0.7863	760.4

3. The HFS method has the least number of features and presents better performance than other feature selection algorithms.

D. ANOMALY DETECTION MODEL ANALYSIS

1) BO HYPERPARAMETER

The LGBM is first hyperparameterized using a BO algorithm when using the LGBM algorithm for LLE detection. The BO obtains the optimal hyperparameters of the LGBM model through iterative calculation by presetting the hyperparameter space, as shown in the following Table 6.

2) ANOMALY DETECTION MODEL PERFORMANCE COMPARISON AND ANALYSIS

The idea of selecting an best anomaly detection model is to choose the best among the best, which could be divided into two steps. At the beginning, some baseline models are chosen without optimization and the best model is selected through experiment. Then the best model is optimized with hyperparameters. In this study, ten groups of basic machine learning algorithms without optimization are selected for comparison experiments to compare the effectiveness of those models. The parameters of the ten machine learning classifiers are as follows: (1) The n_estimators of XGBoost (XGB) [20] is 100; (2)The min_samples_leaf of the decision tree (Tree) is 100; (3) RF [37] sets the number of n_estimators to 50; (4) The closest neighbor of KNN is 10; (5) SVM uses the poly kernel function; (6) The number of GBDT n_estimators is 500; and (7) The solver of LR is lbfgs, and the class_weight is balanced; (8) NB and MLP use default parameters. Table 7 and Figure 7 present the comparison results.

It can be observed from the above data that these tree-based model reach better performance than other machine learning

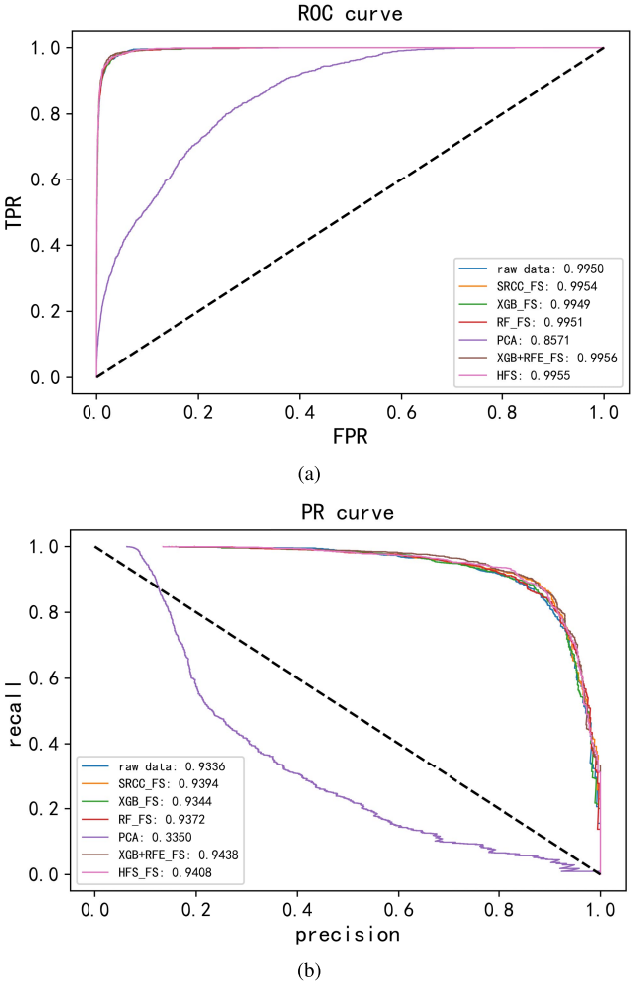


FIGURE 6. Comparison of ROC curve and PR curve of feature selection algorithms: (a) ROC curve; (b) PR curve.

models, indicating that these tree-based models are more suitable for aviation data anomaly detection; they include Tree< GBDT< RF< XGB< LGB. Among the ten groups of basic models, LGB has the best performance, i.e., its Acc,F1,MCC,ROC_AUC and PR_AUC are 98.04%, 78.20, 0.7862, 0.9951% and 0.9408% respectively. Therefore, LGB is selected for further hyperparameter optimization. The F1, MCC, and PR_AUC values are significantly improved, increased by 20.35%, 19.47%, and 4.38% respectively by using BO for enhancing the LGB algorithm.

E. MODEL INTERPRETABILITY ANALYSIS

The SHAP tool is used to visualize the model feature importance and local dependency maps after using HFS and LGBMBO algorithm model on the training set. Feature importance is a global interpretive tool; it is defined as the magnitude of the SHAP value of all the feature value mapping, which indicates the degree of influence on LLEs.

In Figure 8, the vertical axis represents the name of each feature, and the horizontal represents the magnitude of the SHAP value. The feature names from top to bottom are obtained by ranking the importance of the predicted results.

TABLE 6. BO hyperparameter and meaning.

Hyperparameters	Initial value	Parameter space	BO	Meaning
num_leaves	31	(25,4000)	1223	max number of leaves in one tree
max_depth	-1	(5,63)	38	Maximum depth of the tree
lambda_l2	0	(0.0,0.05)	0.0307	L1 regularization
lambda_l1	0	(0.0,0.05)	0.0489	L2 regularization
min_data_in_leaf	20	(100,2000)	109	minimal number of data in one leaf
scale_pos_weight	1	(1,100)	19	weight of labels with positive class

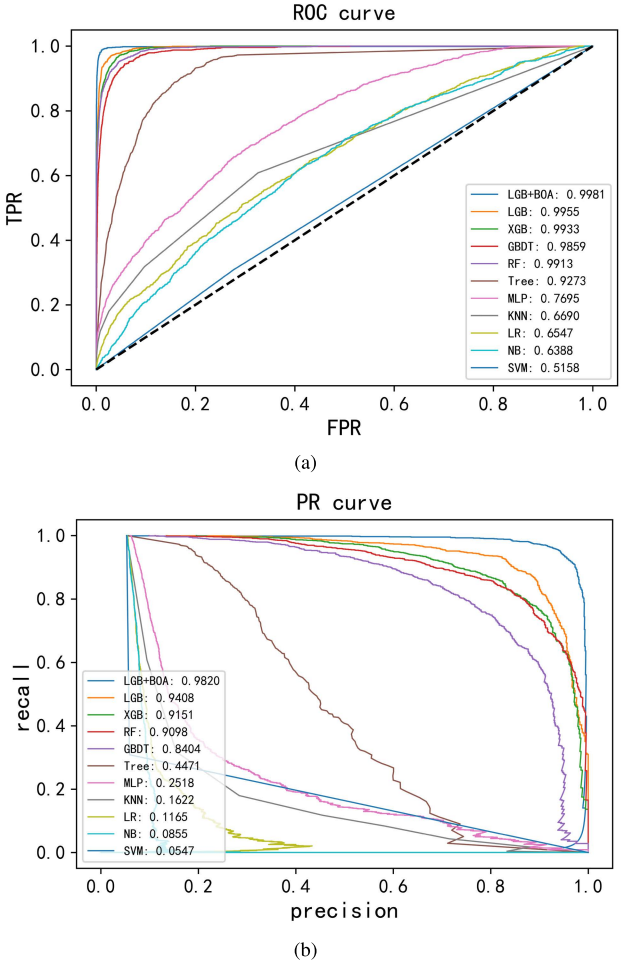


FIGURE 7. Comparison of ROC curve and PR curve of anomaly detection algorithms: (a) ROC curve; (b) PR curve.

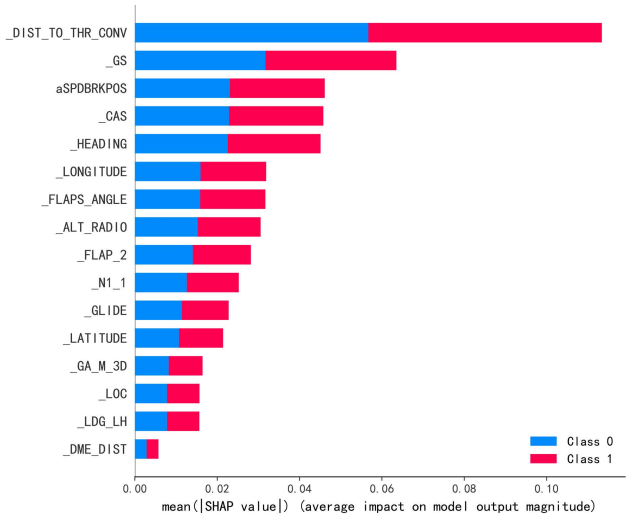


FIGURE 8. Importance ranking of SHAP value.

TABLE 7. Anomaly detection model performance comparison.

Methods	Acc(%)	F1(%)	MCC	ROC_AUC	PR_AUC
LGB+BO	99.40	94.11	0.9394	0.9981	0.9820
LGB	98.04	78.20	0.7863	0.9955	0.9408
XGB	97.85	75.35	0.7620	0.9933	0.9151
Tree	95.14	37.43	0.3807	0.9273	0.4471
RF	96.34	72.93	0.7304	0.9913	0.9098
KNN	94.85	10.82	0.1874	0.6690	0.1622
SVM	70.09	9.86	0.1576	0.5158	0.0547
GBDT	97.06	63.50	0.6572	0.9859	0.8404
LR	62.98	14.24	0.0982	0.6547	0.1165
NB	39.54	12.43	0.0849	0.6388	0.0855
MLP	94.87	20.31	0.2499	0.7695	0.2518

It can be observed that the distance from the runway threshold, ground speed, braking, compute airspeed, and heading angle are the five features that most affect the LLE.

The partial dependence plot (PDP) is a local interpretability tool which can be used to view each parameter's marginal contribution to the predicted outcome. On the PDP, each point represents a row of data; the vertical axis represents the feature value, and the vertical coordinate indicates the effect of the feature value on the final landing result. A value greater than 0 on the vertical coordinate indicates that it promotes the LLE, and a value less than 0 indicates that the point in time is in a normal event. The horizontal coordinates represent the range of values of the feature. The whole graph shows the mapping correlation between the feature distribution and SHAP values.

In this study, the trend of each parameter is observed in two dimensions: altitude and runway plane, to conveniently observe the flight status of the aircraft during landing. The altitude is referenced by the radar altitude, considering the airport plane as the altitude origin. Furthermore, the runway plane considers the runway threshold as its origin. A distance greater than 0 indicates that the aircraft flies to the runway, while a distance less than 0 indicates that the aircraft flies from the runway threshold to the endpoint. These two dimensions are analyzed based on several key parameters which affect the flight state: ground speed, brake, glide slope deviation, and glide angle.

The yellow line in the Figure 9, Figure 10 and Figure 11 represents the regression line of the parameter distribution. The yellow line represents a trend throughout the data, and the black dashed line is the horizontal line where SHAP value is 0. In these figures, if the yellow line is below the black dashed line, it indicates that the LLE will

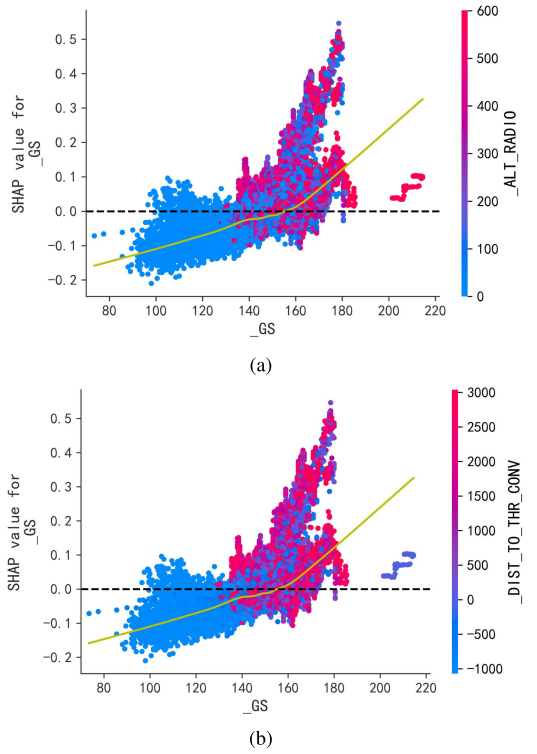


FIGURE 9. The trend of ground speed during landing: (a) from altitude viewpoint; (b) from runway plane viewpoint.

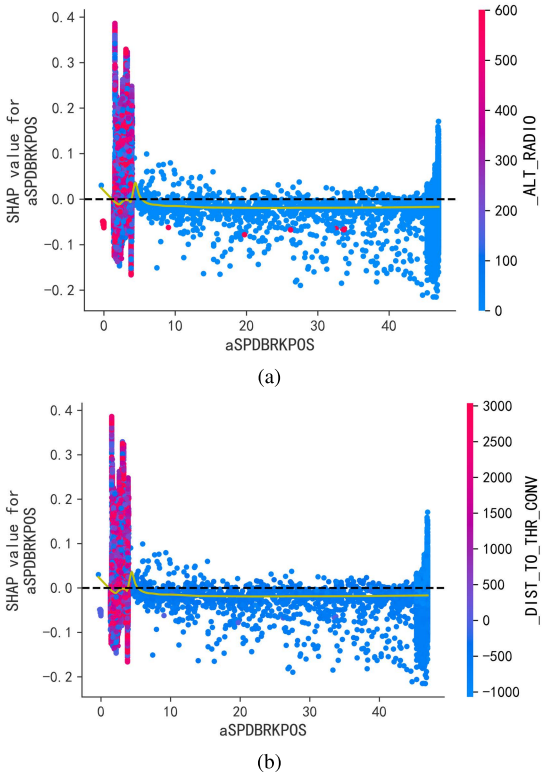


FIGURE 10. The trend of brake during landing: (a) from altitude viewpoint; (b) from runway plane viewpoint.

be suppressed at this time; otherwise, the occurrence of the LLE will be promoted. It can be observed from Figure 9 that the slope of the regression line is greater when the aircraft is at high altitude than at low altitude, and the higher the aircraft speed, the more likely is occurrence of an LLE. Therefore, maintaining the appropriate airspeed during the landing phase is crucial.

Brake value is an angle variable, and the greater the brake value, the greater the braking force. It can be observed from Figure 10 that the regression lines of the brake value distributions are typically lower than black dashed line, indicating that braking has an inhibitory effect on the LLE, which is consistent with physical logic. In fact, the brakes serve to mitigate the effects of the LLEs, i.e., to increase the actual length of the aircraft available on the runway and reduce the scheduling risk.

Figure 11 depicts the effect of glide slope deviation on the LLE. In fact, glide slope deviation and glide slope are the difference relation of vertical direction. Despite the multi-path and near-field effects, the general trend of glide slope deviation can be observed. It can be observed from Figure 11 that, if glide slope deviation is less than 0 during the flare-out phase, i.e., the yellow line is above the black dashed line, it indicates that the aircraft is below the ILS glide path, which suppresses LLEs, and vice versa. The actual glide angle of the aircraft is related to the altitude of the aircraft and the distance from the ILS glide-path localizer, so the pilot should be pay attention to the aircraft altitude and the distance from glide-path localizer.

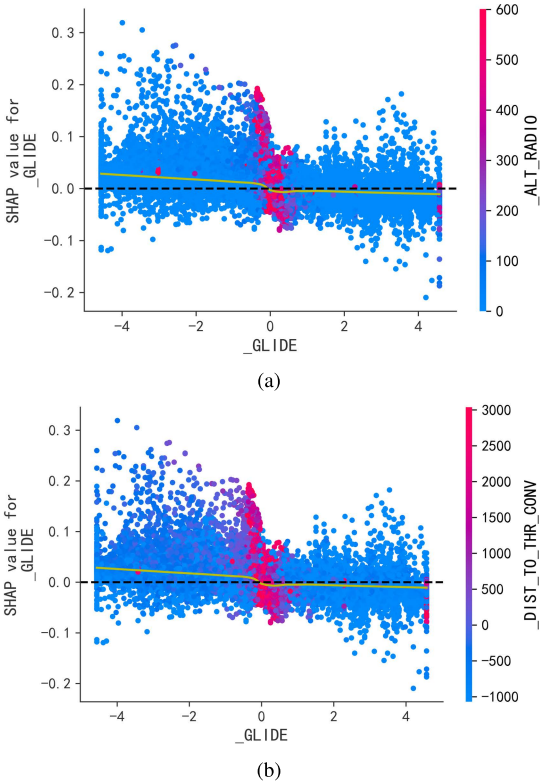


FIGURE 11. The trend of glide slope deviation during landing: (a) from altitude viewpoint; (b) from runway plane viewpoint.

Figure 12 illustrates the glide angle of the aircraft, from which it can be observed that most aircraft land along a

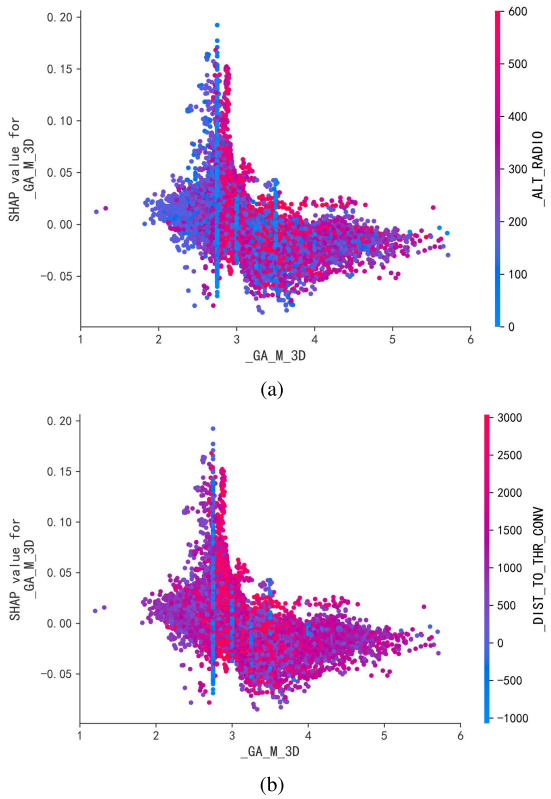


FIGURE 12. The trend of glide angle during landing: (a) from altitude viewpoint; (b) from runway plane viewpoint.

2.8 degree angle of glide. It is consistent with the rule in the flight manual that the angle of glide of the aircraft typically lies between 2.5 and 4 degrees. Therefore, the pilot must ensure that the glide angle lies within the appropriate range during the approach phase. Essentially, a low glide angle may result in an LLE.

V. CONCLUSION

Risk event detection and cause analysis are necessary and sufficient for flight safety and accident prevention. Risk event detection is a prerequisite for risk cause explanation, and a model can only convincingly explain the causes of risk with a high detection rate. Conversely, risk cause explanation is a supplement to risk event detection and can reflect the practical value of detection. In this paper, the proposed workflow unifies the ability to detect anomalous events and the inference of their causality.

Our model reduces computational cost and achieves better performance. Besides, with the help of SHAP, it proves that LLEs can be avoided during the landing phase by maintaining the appropriate descent speed, aircraft altitude, and descent angle, which is consistent with the requirements of aviation regulations and pilot training manuals.

Despite the strengths of proposed workflow in the study, some limitations should be noted. The current work does not allow for some external environmental factors, such as weather and runway conditions, which may pose a risk of misjudgment. Furthermore, the lack of temporal correlation among events makes it difficult to depict a causal chain of accidents. In future work, the aviation daily monitoring data will be used to explore more challenging scenarios, such as the assessment of the pilot's flight quality and the construction of aviation abnormal event knowledge graphs for causal inference of events.

REFERENCES

[1] B. C. Airplanes, "Statistical summary of commercial jet airplane accidents, worldwide operations 1959–2020," Boeing Crop., Chicago, IL, USA, Tech. Rep., 2021. [Online]. Available: https://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf

[2] AIRBUS. (2019). *A Statistical Analysis of Commercial Aviation Accidents 1958–2019*. [Online]. Available: https://accidentstats.airbus.com/statistics/accident-categories

[3] *Safety Report (2020)*, ICAO, Montreal, QC, Canada, 2020.

[4] L. Wang, C. Wu, and R. Sun, "An analysis of flight quick access recorder (QAR) data and its applications in preventing landing incidents," *Rel. Eng. Syst. Saf.*, vol. 127, pp. 86–96, Jul. 2014.

[5] S. Zhou, Y. Zhou, Z. Xu, W. Chang, and Y. Cheng, "The landing safety prediction model by integrating pattern recognition and Markov chain with flight data," *Neural Comput. Appl.*, vol. 31, no. S1, pp. 147–159, Jan. 2019.

[6] *Manual on Flight Data Analysis Programmes (FDAP)*, Standard 10000AN/501, ICAO, 2014.

[7] *Flight Operational Quality Assurance*, Standard AC 120-82, FAA, Apr. 2004.

[8] L. Li, R. Hansman, R. Palacios, and R. Welsch, "Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring," *Transp. Res. C, Emerg. Technol.*, vol. 64, pp. 45–57, Mar. 2016, doi: 10.1016/j.trc.2016.01.007.

[9] J. Oehling and D. J. Barry, "Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data," *Saf. Sci.*, vol. 114, pp. 89–104, Apr. 2019, doi: 10.1016/j.ssci.2018.12.018.

[10] K. Sheridan, T. G. Puranik, E. Mangortey, O. J. Pinon-Fischer, M. Kirby, and D. N. Mavris, "An application of DBSCAN clustering for flight anomaly detection during the approach phase," in *Proc. AIAA Scitech Forum*, Jan. 2020, p. 1851.

[11] F. Mannering, C. R. Bhat, V. Shankar, and M. Abdel-Aty, "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis," *Analytic Methods Accident Res.*, vol. 25, Mar. 2020, Art. no. 100113.

[12] S. Das, B. L. Matthews, A. N. Srivastava, and N. C. Oza, "Multiple kernel learning for heterogeneous anomaly detection: Algorithm and aviation safety case study," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 47–56.

[13] S. Das, B. L. Matthews, and R. Lawrence, "Fleet level anomaly detection of aviation safety data," in *Proc. IEEE Conf. Prognostics Health Manage.*, Jun. 2011, pp. 1–10.

[14] 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.

[15] T. G. Puranik, H. Jimenez, and D. N. Mavris, "Utilizing energy metrics and clustering techniques to identify anomalous general aviation operations," in *Proc. AIAA Inf. Syst. Infotech Aerosp.*, Jan. 2017, p. 789.

[16] I. Melnyk, A. Banerjee, B. Matthews, and N. Oza, "Semi-Markov switching vector autoregressive model-based anomaly detection in aviation systems," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1065–1074.

[17] C. Tong, X. Yin, S. Wang, and Z. Zheng, "A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data," *Future Gener. Comput. Syst.*, vol. 88, pp. 552–558, Nov. 2018.

[18] T. G. Puranik, N. Rodriguez, and D. N. Mavris, "Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning," *Transp. Res. C, Emerg. Technol.*, vol. 120, Nov. 2020, Art. no. 102819.

[19] M. Memarzadeh, B. Matthews, and I. Avrekh, "Unsupervised anomaly detection in flight data using convolutional variational auto-encoder," *Aerospace*, vol. 7, no. 8, p. 115, Aug. 2020.

[20] M. Rey, D. Aloise, F. Soumis, and R. Pieugueu, "A data-driven model for safety risk identification from flight data analysis," *Transp. Eng.*, vol. 5, Sep. 2021, Art. no. 100087.
[21] L. Wang, Y. Ren, and C. Wu, "Effects of flare operation on landing safety: A study based on ANOVA of real flight data," *Saf. Sci.*, vol. 102, pp. 14–25, Feb. 2018.
[22] D. J. Barry, "Estimating runway veer-off risk using a Bayesian network with flight data," *Transp. Res. C, Emerg. Technol.*, vol. 128, Jul. 2021, Art. no. 103180.
[23] *Airplane Flying Handbook (FAA-H-8083-3A)*," Skyhorse Publishing, New York, NY, USA, 2011.
[24] *The Implementation and Management Flight Operational Quality Assurance (FOQA) [S]*, Standard AC-121/135-FS-2012-45R1, CAAC, Jun. 2015.
[25] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 907–948, Feb. 2020.
[26] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018.
[27] R. A. Armstrong, "Should Pearson's correlation coefficient be avoided?" *Ophthalmic Physiol. Opt.*, vol. 39, no. 5, pp. 316–327, Sep. 2019.
[28] S. H. Huang, "Supervised feature selection: A tutorial," *Artif. Intell. Res.*, vol. 4, no. 2, pp. 22–37, Apr. 2015.
[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.
[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
[33] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz, "BOA: The Bayesian optimization algorithm," in *Proc. Genetic Evol. Comput. Conf. (GECCO)*, vol. 1, 1999, pp. 525–532.
[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.
[35] J. Rogers and S. Gunn, "Identifying feature relevance using a random forest," in *International Statistical and Optimization Perspectives Workshop-Subspace, Latent Structure and Feature Selection*. Cham, Switzerland: Springer, 2005, pp. 173–184.
[36] H. Abdi and L. J. Williams, "Principal component analysis, Wiley interdisciplinary reviews," *Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[37] J. Li, H. Zhang, and J. Yang, "Discrimination model of QAR high-severity events using machine learning," in *Proc. Int. Conf. Intell. Sci. Big Data Eng.* Cham, Switzerland: Springer, Oct. 2019, pp. 430–441.

XIONG YANG was born in 1995. He received the B.S. degree in electronic information engineering and the M.Sc. degree from the Civil Aviation University of China, in 2018 and 2021, respectively. His research interests include machine learning, data mining, and flight safety.

JIN REN was born in 1990. He received the B.S. and the Ph.D. degrees in mechanical engineering and from Central South University, in 2013 and 2020, respectively. He is currently working with the Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic. His research interests include speech and language processing, multi-modal understanding (speech-text-image), semantic perception, knowledge-enhanced language representation, and human–robot interaction.

JUNCHEN LI was born in 1993. He received the B.S. degree in electronic information engineering and the M.Sc. degree from Civil Aviation University of China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the College of Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include machine learning, data mining, and smart city.

HAIGANG ZHANG was born in 1989. He received the B.S. degree from the University of Science and Technology Liaoning, in 2012, and the Ph.D. degree from the School of Automation and Electrical Engineering, University of Science and Technology Beijing, in 2018. He is currently an Associate Researcher with the Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic. His research interests include machine learning, image process, and biometrics.

JINFENG YANG as born in 1971. He received the B.S. degree from the Zhengzhou University of Light Industry, China, in 1994, the M.Sc. degree from Zhengzhou University, in 2001, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2005. He is currently a Professor with the Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic. His major research interests include machine learning, pattern recognition, computer vision, multimedia processing, and data mining. He was a PC Member/ a Referee of international journals and top conferences, including the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), *Innovative Veterinary Care Journal* (IVC), *Pattern Recognition* (PR), *Physical Review Letters* (PRL), International Conference on Computer Vision (ICCV), *Computer Vision and Pattern Recognition* (CVPR), and European Conference on Computer Vision (ECCV). He is a Syndic of the Chinese Society of Image and Graphics and the Chinese Association for Artificial Intelligence.