UAA UNIVERSITY AVIATION ASSOCIATION

Collegiate Aviation Review International

Volume 44 | Issue 1 Peer Reviewed Article #9

05-07-2026

Data-Driven Decision-Making in Aviation Safety Management Systems: A Supervised Machine Learning Approach

Vivek Sharma
Florida Institute of Technology

Brooke Wheeler
Florida Institute of Technology

Bhoomin Chauhan
NASA AMES Research Center

Shaun Kelly
Purdue University

One of the objectives of the FAA's safety improvement plan is to continuously collect safety data to identify potential risks. Over the last decade, the application of machine learning (ML) and Artificial Intelligence (AI) models in prediction, classification, and identification has been widely used across both aviation and non-aviation domains. Few studies have explored the use of ML techniques in aviation for predicting safety. Therefore, the purpose of the current study is twofold: (a) to build and compare the classification performance of three supervised machine learning models: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boost (XG Boost), and (b) to apply the Synthetic Minority Oversampling Technique (SMOTE) class imbalance technique to all three models and compare the performance. The total sample size for the current study was N = 17,275, of which 15,870 (91.86%) were classified as accidents and 1,405 (9.14%) were classified as incidents. The NTSB database includes many features, such as event ID, registration number, Federal Aviation Regulation (FAR), flight plan, damage type, event type, crew demographics, and aircraft characteristics. The dependent or outcome variable for the current study was event type (accident or incident). The findings of this study demonstrate that supervised ML models can be effectively used to predict or classify aviation events such as incidents or accidents. Specifically, the Random Forest, SVM, and XG Boost models can be applied to operational data to classify aviation accidents and incidents. The findings of this study offer multiple practical implications for enhancing safety through proactive and predictive data-analytical decision-making, aligning with ICAO's (2018) SMM. A robust SMS is always data-driven, and integrating ML to proactively identify hazards can strengthen its foundation.

Recommended Citation:
Sharma, V., Wheeler, B., Chauhan, B., & Kelly, S. (2026). Data-Driven Decision-Making in Aviation Safety Management Systems: A Supervised Machine Learning Approach. *Collegiate Aviation Review International*, *44*(1), 178–199. Retrieved from https://ojs.library.okstate.edu/osu/index.php/CARI/article/view/10569/9363

Introduction

According to the International Civil Aviation Organization's (ICAO, 2018) Safety Management Manual (SMM), the purpose of safety analysis and safety reporting is to present the organization with the current safety state so that decision-makers can make decisions based on the data presented (ICAO, 2018, p. 6-7). The safety management systems (SMS) framework consists of four pillars: safety policy, safety risk management (SRM), safety assurance, and safety promotion (ICAO, 2018). SRM is considered a fuel for SMS, in which safety-critical hazards are identified, risks are assessed for each hazard, and risk-mitigation strategies are implemented. Hazard identification is a critical component of SRM and depends on proactive data collection and analysis. Data collection and analysis in SMS involve generating meaningful insights from large datasets, including accident/incident reports, flight data logs, maintenance logs, and employee reports (Howell, 2025). The main objective of safety data monitoring and analysis is to help safety decision-makers or top-level management of an organization with the current state of safety. According to ICAO (2018), this can be achieved through a data-driven decision-making (D3M) approach. However, some key challenges safety managers face when implementing the D3M approach include data inconsistency, volume and complexity, limited resources, and a lack of expertise. Machine learning (ML) and Artificial Intelligence (AI) tools are gaining popularity in aviation to help safety managers and personnel make data-driven safety-critical decisions (Demir et al., 2024). ML-based models can be trained on historical safety data to classify risks, predict events, and identify potential hazards (Tafur et al., 2025), making them essential tools for aviation safety. ML modeling opens up possibilities to more proactively address safety.

According to 14 Code of Federal Regulations (CFR) Part 5, all Part 121 U.S. Operators, Part 135 charter operators, Part 91.147 air tour operators, Part 21 aircraft manufacturers, and Part 139 airports are required to have an SMS program to proactively manage safety and risks (National Archive and Records Administration [NARA], 2026). Figure 1 illustrates a conceptual framework for integrating an ML pipeline into the four pillars of the SMS framework (Sharma, 2026). As illustrated in Figure 1, the initial stage of the ML pipeline begins with identifying safety problems and the corresponding data collection measures. These ML steps align with the safety policy pillar, as this is where an organization's safety goals, objectives, and culture are built. In the later stages of the ML pipeline, such as data cleaning and integration, feature selection and engineering, model training, testing, and evaluation, these activities can be aligned with the SRM and Safety Assurance pillars of SMS. The final stages of the ML pipeline, such as model monitoring, deployment to the operational environment, and building safety dashboards, can align with the Safety Assurance of the SMS framework. In the current study, the scope is limited to the sections highlighted in Figure 1, which are associated with Safety Risk Management and Safety Assurance pillars of the SMS.

Figure 1

Conceptual Framework of SMS and ML Pipeline

Note. This figure was adopted from Vivek Sharma's LinkedIn post (Sharma, 2026).

Purpose Statement

According to the Federal Aviation Administration (FAA, n.d.), one of the objectives of the FAA's safety improvement plan is to continuously collect safety data to identify potential risks. Over the last decade, the application of machine learning (ML) and Artificial Intelligence (AI) models in prediction, classification, and identification has been widely used across both aviation and non-aviation domains (Boukerche & Wang, 2020; Black et al., 2023; Nanyoga et al., 2023; Samek et al., 2017; Yang et al., 2022). However, few studies have explored the use of ML techniques in aviation to predict safety (Omrani et al., 2023; Nanyonga et al., 2025; Ramírez et al., 2024). Therefore, the purpose of the current study is twofold: (a) to build and compare the classification performance of three supervised machine learning models: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boost (XG Boost) to classify aviation events: incidents and accidents, and (b) to apply the Synthetic Minority Oversampling Technique (SMOTE) class imbalance technique to all three models and compare the performance before and after SMOTE application.

Research Questions

RQ 1: How do the proposed models, RF, SVM, and XG Boost, compare to each other in terms of overall accuracy, precision, recall, and F1 scores for classifying aviation events: incidents and accidents?

RQ 2: How do the proposed RF, SVM, and XG Boost models perform after applying the SMOTE technique to classify aviation incidents and accidents?

The selection of the proposed ML models in the current study: RF, SVM, and XG Boost, was based on a literature review and theoretical grounding. Various aviation studies have successfully adopted the above-mentioned techniques (Metha et al., 2021; Nanyonga et al., 2025; Omrani et al., 2023). For instance, Nanyonga et al. (2025) conducted a study to classify aviation accidents and incidents using SVM and RF-based ML techniques on the Australian Transportation Safety Board (ATSB) database. Omrani et al. (2023) investigated aviation accidents from the Transportation Safety Board (TSB), NTSB, and ATSB databases to classify injury levels and deployed several ML techniques, including SVM. Similarly, Mehta et al. (2021) built RF, Gradient Boosting, and SVM models to classify aircraft damage. Building on the above literature, the goal of the current study is to classify aviation events and conduct a comprehensive comparative analysis of three robust supervised ML techniques: RF, SVM, and XG Boosting. Although logistic regression is a widely considered machine learning model for binary classification problems, the current study used advanced ML algorithms for their unique advantages in predictive modeling. For instance, SVM is a robust technique that specializes in finding optimal class-separating hyper-plane when using multiple features (Omar et al., 2024). RF is an ensemble technique of decision trees used to make accurate predictions and is often considered to reduce overfitting concerns and improve generalization (Breiman, 2001). XGBoost is an ensemble learning technique that can capture complex nonlinear relationships and improve predictive accuracy (Friedman, 2001).

Study's Significance

A conceptual framework for integrating the ML pipeline into the SMS framework is shown in Figure 1. The objectives of the current study were to train and evaluate supervised ML models to classify aviation events based on historical data. One of the key findings of the proposed study is the development of ML models that can serve as automated hazard-identification tools in the SRM stage. In particular, aviation stakeholders implementing a full SMS Framework, including but not limited to airlines, AMTs, airport operators, and UAS operators, can deploy the proposed ML models in their SMS framework and assist safety analysts by classifying historical safety reports into critical operational hazards. Although the current study uses historical National Transportation Safety Board (NTSB, n.d.) data, the NTSB data capture safety characteristics or safety performance indicators comparable to those used by aviation stakeholders in their safety programs, such as Aviation Safety Action Reports (ASAP), Aviation Safety Information and Analysis Sharing (ASIAS), and National Aeronautical and Space Administration's (NASA's) Aviation Safety Reporting System (ASRS), thereby increasing the generalizability of the current findings. The reader should also be required to note that the practical implications of the current study's findings are to be considered as proof-of-concept and to guide aviation stakeholders in integrating ML practices into the SMS framework.

Literature Review

Use of ML in Aviation Safety

Prior research in other fields and aviation is reviewed to highlight the need for ML in aviation safety. The application of ML techniques in healthcare has witnessed rapid advancements, completely changing the landscape of medical treatment and diagnosis (Ramírez

et al., 2024). Recently, the use of ML-based prediction models to classify driver behaviors, identify high-risk road segments, and forecast road infrastructure failures has gained popularity (Boukerche & Wang, 2020; Yang et al., 2022). The aviation industry has begun adopting data-driven approaches, such as ML, to enhance safety. The application of ML models to autonomously identify aviation operational hazards and assess potential risks has been reported in several studies (Huang et al., 2020; Kuleshov et al., 2023; Borjalilu, 2024). According to Demir et al. (2024), the use of ML models can significantly help aviation organizations explore large quantities of flight operational data. Khalid et al. (2023) reported the effectiveness of ML models in analyzing trends and patterns from historical safety data to help top-level management make informed flight operational safety decisions.

Several studies have examined the application of ML techniques to enhance aviation safety (Karaburun et al., 2024; Nanyonga et al., 2025; New & Wallace, 2025; Omrani et al., 2024; Rahman et al., 2025). Karaburun et al. (2025) explored how the takeoff performance of a B737-300-type aircraft was predicted using different ML approaches, including SVM, regression, RF regression, and XGB. Omrani et al. (2024) examined aviation accident data through various ML approaches, including Artificial Neural Network (ANN), Decision Tree (DT), and SVM, to identify potential contributing factors of accidents. Mehta et al. (2001) investigated several ML models, including SVM, RF, and logistic regression, to predict the severity of airplane crashes. Silagyi and Liu (2023) applied SVMs to predict aircraft damage and personal injury severity using National Transportation Safety Board (NTSB) accident data from 2014 to 2019.

In a recent study, Cankaya et al. (2023) used supervised ML methods, including multinomial logistic regression, SVMs, and deep learning, to predict aircraft damage from the FAA's Aviation Safety Information Analysis and Sharing (ASIAS) accident and incident data from 2020–2024. Nanyonga et al. (2025) conducted a study to monitor the classification performance of four different ML approaches, including SVM, logistic regression, RF, and deep neural networks, on the Australian Transportation Safety Board's (ATSB's) accident data by incorporating a Variational Autoencoder (VAE), a class imbalance technique. These studies demonstrate ML's ability to classify and predict when trained on aviation safety data.

Safety is paramount in aviation, and safety management is a priority for all aviation stakeholders, including airline operators, airport operators, maintenance service providers, airline pilots, and air traffic controllers (ATCs). Airlines and aviation organizations collect operational data through various safety programs, including Flight Operational Quality Assurance (FOQA), Line Operations Safety Audit (LOSA), Aviation Safety Action Program (ASAP), and aviation accident databases, to identify hazards and mitigate potential risks. Although a few studies have focused on using supervised ML techniques to identify aviation-related safety by leveraging the FAA's ASIAS, ATSB, and incident logs from FDM programs (Cankaya et al., 2023; Karaburun et al., 2024; Nanyonga et al., 2025), there is a dearth of research that focuses on building supervised learning models using NTSB to classify event types based on different features. Additionally, most previous studies did not account for class imbalance, a widely observed problem in real-time operational or safety datasets, where accidents occur far less frequently than incidents.

Methods

The current study began by identifying key features (predictors) and the criterion variable, performing data preprocessing and handling missing data, executing model training and evaluation, and finally applying the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance across all models.

Sample and Feature Selection

The population for the current study included all aviation accidents and incidents reported in the NTSB (n.d.) database. The accessible population and sample for the current study included all accidents and incidents reported in the NTSB database between January 2015 and July 2025. The total sample size for the current study was $N = 17{,}275$, of which 15,870 (91.86%) were classified as accidents and 1,405 (9.14%) were classified as incidents. The NTSB database includes many features, such as event ID, registration number, Federal Aviation Regulation (FAR), flight plan, damage type, event type, crew demographics, and aircraft characteristics. The dependent or outcome variable for the current study was event type (accident or incident). The NTSB dataset comprises more than 40 features that can be used as predictors; however, it is necessary to focus only on a subset of variables (features) likely to have a significant impact on aircraft operations. In aviation, accidents or incidents are never caused by a single factor but result from multiple interactions among humans, aircraft systems, environmental conditions, and organizational structures (Reason, 1990; Wiegmann & Shappell, 2003). Therefore, 15 key features were selected for the feature selection process, as shown in Table 1. The selected features were broadly classified into three categories: aircraft, flight crew-related, and environment-related. The selection of features for the current study was also consistent with previous studies that examined ML models to classify aviation events (Cankaya et al., 2023; Omrani et al., 2023; Silyagi & Liu, 2023).

Data Preprocessing

All missing data were identified in Python and handled based on the percentage of missingness. Features with more than 40% missing data were excluded from the model. Missing data for continuous features were imputed with median values, and missing data for categorical features were imputed as *unknown* values. All categorical variables were coded using One-Hot Encoding, which converts them into binary indicators (1 and 0).

Table 1

Handling Missing Data

Features Selected		Missing Data Count	% Missing	Strategy
afm_hrs_last_insp	Number of Hours since last inspection	13051	75.54	Excluded
flight_plan_activated	Type of Flight Plan filed	6907	39.98	*Unknown*
cert_max_gr_wt	Certified Max Gross Weight	5615	32.50	Median

flight_hours	Total number of flight hours	5349	30.96	Median
type_last_insp	Last Inspection Type	4928	28.52	Median
crew_age	Age of Flight Crew in Years	4679	27.08	*Unknown*
eng_type	Engine Type	4612	26.69	*Unknown*
second_pilot	If there is a second pilot	4304	24.91	Unknown
flt_plan_filed	Type of Flight Plan filed	4068	23.54	*Unknown*
far_part	FAA Regulation Part	497	2.87	*Unknown*
acft_category	Aircraft Category	186	1.07	*Unknown*
Occurrence_Code	Type of Occurrence	14	0.08	*Unknown*
wind_vel_kts	Wind Velocity in Knots	0	0.00	-
gust_kts	Gust Velocity in Knots	0	0.00	-
wx_cond_basic	Basic Weather Condition	0	0.00	-

Model Development Hyperparameters and Evaluation Metrics

Three supervised learning models were developed and implemented to evaluate the classification performance of the event types: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XG Boosting). All three models were configured based on the key hyperparameters listed in Table 2. These hyperparameters increase the model's balance by reducing biases, enhancing model robustness, and avoiding overfitting. In the current study, all selected hyperparameters were drawn from prior literature (Nanyoga et al., 2025; Silyagi & Liu, 2023). The use of extensive hyperparameter tuning was avoided to ensure a consistent comparison across all models before and after SMOTE.

Table 2

Handling Missing Data

Model	SMOTE Applied?	Key Hyperparameters
RF	Yes	n_estimators=200; max_depth=2; criterion = 'gini' random_state=7; max_features = 'sqrt'; class_weight not used SMOTE applied
SVM	Yes	kernel='linear'; C=1.0; gamma='scale' probability=False, random_state=7; decision_function_shape = 'ovo'
XGB	Yes	eval_metric='logloss', scale_pos_weight=1, random_state=7, use_label_encoder=False, max_depth=3

As reported in Table 3, the model's performance was evaluated using Accuracy, F1-score, Precision, and Recall. Accuracy is a fundamental indicator of model performance (Powers, 2020).

Table 3

Evaluation Metrics

Metric	Definition	Formula	Interpretation
Accuracy	It is defined as all correctly predicted instances out of the total	$\frac{(\mathbf{TP}+\mathbf{TN})}{(\mathbf{TP}+\mathbf{TN}+\mathbf{FP}+\mathbf{FN})}$	Gives an idea of overall performance of the model.
F1 – Score	It is defined as the weighted average score of precision and recall	$\mathbf{2} * \frac{(\boldsymbol{p} * \boldsymbol{r})}{(\boldsymbol{p}+\boldsymbol{r})}$	Useful when you use a single metric to measure performance with multiple imbalances categories.
Precision (*p*)	It is defined as the proportion of true positive (TP) predictions out of all positive predicted values (TP + FP)	$\frac{\mathbf{TP}}{(\mathbf{TP}+\mathbf{FP})}$	When the model predicts a certain category, how often is it correct?
Recall (*r*)	It is defined as the proportion of true positive (TP) predictions out of all actual positive values (TP + FN)	$\frac{\mathbf{TP}}{(\mathbf{TP}+\mathbf{FN})}$	Out of all actual instances of a specific category, how many did the model correctly detect?

Precision ensures the reliability of a model by reducing false positives (Sokolova & Lapalme, 2009). Recall is a critical metric for identifying the positive class when dealing with an imbalanced dataset, whereas the F1-score is a weighted average of precision and recall and mitigates bias toward a specific metric (He & Garcia, 2009). In addition to the evaluation metrics, the current study used a confusion matrix to assess classification errors. A confusion matrix is a two-dimensional contingency table used to differentiate between "actual" and "predicted" values, and it has been widely used as a performance evaluator in health and aviation studies (Nanyonga et al., 2025; Yang & Berdine, 2024). As reported in Table 4, the errors were analyzed from two different perspectives: False Positives and False Negatives. Along with the errors, the true positive and true negative values were also reported.

Table 4

Confusion Matrix

Actual Values	Predicted Values	
	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Area Under Curve Receiver Operating Curve

In addition to the confusion matrix, a receiver operating characteristic (ROC) curve was used to evaluate the overall performance of the three models: RF, SVM, and XGBoost. An ROC curve was initially developed during World War II to differentiate between signals and noise (Nahm, 2022). ROC curves have been applied in psychology, medicine, and machine learning (Nahm, 2022; Sui et al., 2021; Tanner & Swets, 1954). In machine learning, the ROC plots True Positive Rates against False Positive Rates. The Area Under the Curve (AUC) is a key metric derived from the ROC curve that summarizes the model's overall performance in differentiating between positive and negative values. An AUC score of 1.0 indicates that the model is a perfect

classifier, and a score of.50 indicates that the model is a poor classifier; any prediction is equivalent to random guessing (Chang & Newman, 2024).

Handling Class Imbalance

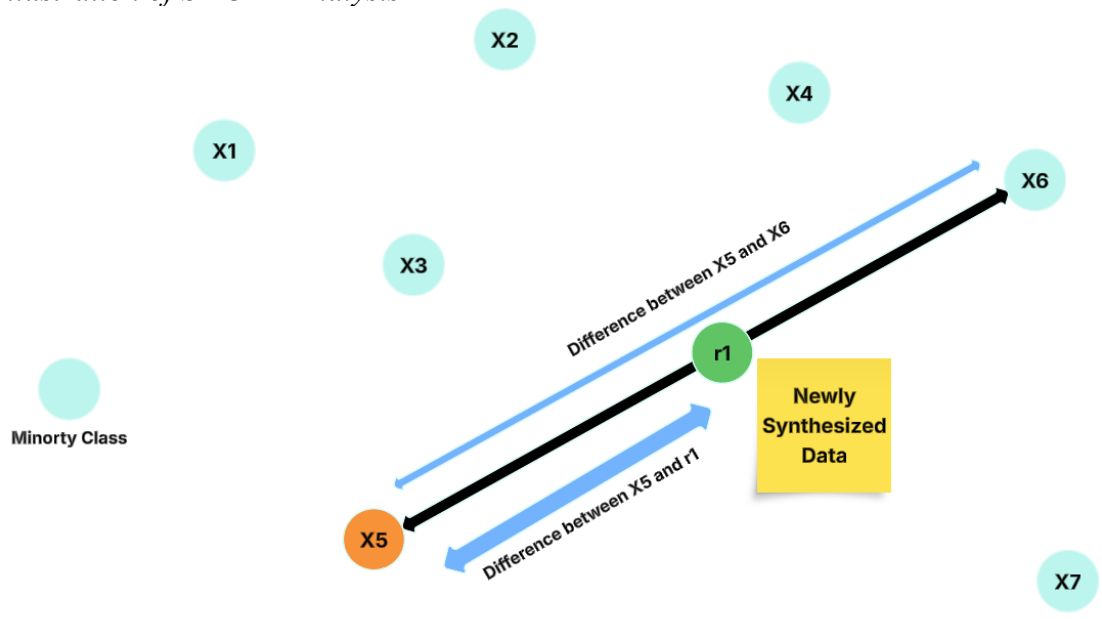
When conducting research on aviation safety data, one key obstacle is class imbalance. This is critical, especially when predicting critical but rare events, such as accidents or incidents. Machine learning performance metrics, such as overall accuracy and receiver operating curve (ROC), tend to perform poorly when the data are imbalanced (Chawla et al., 2002). One of the most widely used techniques for handling class imbalance is the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002; Delgado & Núñez-González, 2022; Riantika et al., 2024). According to Chawla et al. (2002), in SMOTE, minority classes are oversampled by creating new minority class examples rather than by oversampling with a replacement. The SMOTE technique relies on creating new data points for the minority class by selecting samples from the minority class and their nearest neighbors. As noted earlier, the criterion variable of this study was highly imbalanced, with 15,870 (91.86%) instances classified as accidents and 1,405 (9.14%) as incidents. This imbalance could potentially inflate or deflate the performance metrics of the three ML models: RF, SVM, and XGB.

Therefore, the current study employed the SMOTE technique and compared the performance of all models before and after its application. A graphical illustration of SMOTE is shown in Figure 2. In the example below, X5 (orange dot), which is a minority class, is randomly selected, and the nearest neighbors are identified as X1, X2, X3, X4, X6, and X7 (cyan circles). One of these neighbors is randomly selected; in this instance, it is X6. The vector distance between X1 and X6 was computed, and a scalar gap was applied to it. A new synthetic sample (green dots) was generated using the following formula:

$$r1 = X5 + (Diferrence\ beyween\ X1\ and\ r1) * (X6 - X5)$$

Figure 2

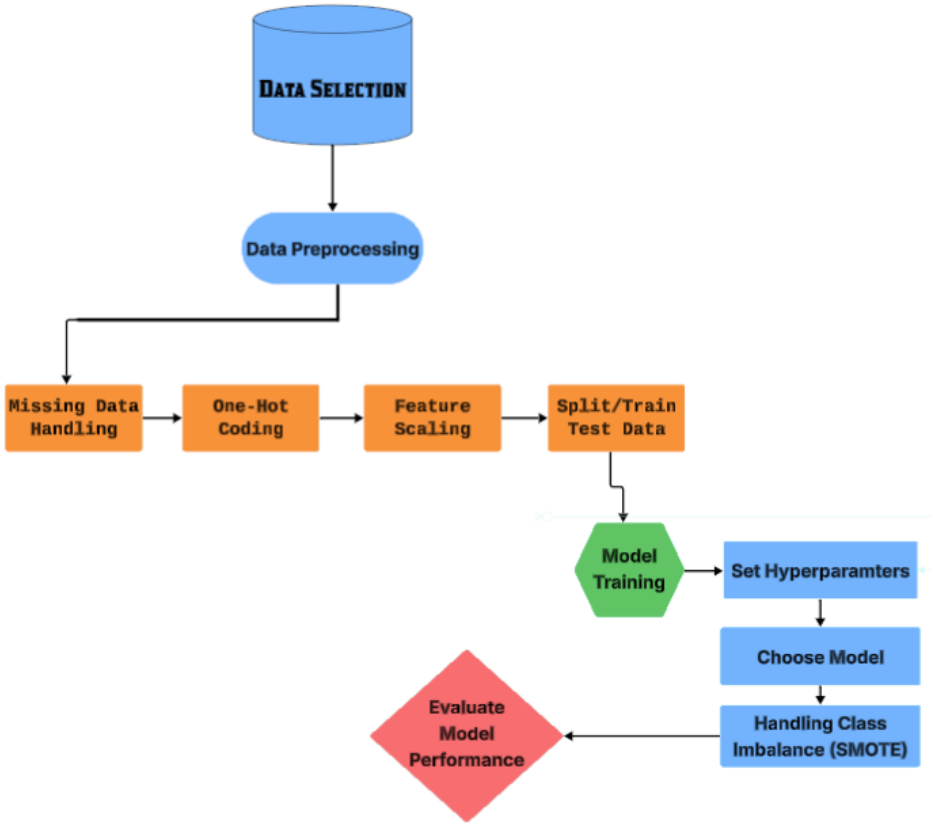
Illustration of SMOTE Analysis



A detailed description of the machine learning pipeline is shown in Figure 3. The ML pipeline adopted in the current study began with data collection and preprocessing, which included handling missing data, encoding categorical variables through a one-hot encoding strategy, and feature scaling. In the next stage, the dataset was divided into two sets: training and testing sets. To address the class imbalance issue, SMOTE was applied only to the training dataset after splitting the data into training and testing datasets. This process ensured that the test data remained unchanged and prevented the model from inflating the results due to the newly created synthetic samples (He & Garcia, 2009).

Figure 3

Machine Learning Pipeline



Results

In this study, three different supervised ML models, RF, SVM, and XGB, were built and compared to differentiate between accidents and incidents based on different features. The Synthetic Minority Oversampling Technique (SMOTE) was employed to mitigate class imbalance. Next, all model performance parameters were compared before and after SMOTE was applied.

Random Forest (RF) Model Performance Before and After SMOTE

The results for the RF model before and after applying SMOTE are presented in Table 5. The model before the SMOTE application had an overall accuracy score of 0.93. However, the recall value for incidents was only 0.27. After applying SMOTE, the recall value for incidents significantly improved to 0.84, whereas the overall accuracy decreased slightly to 0.90. As shown in Figure 4, after applying SMOTE, the true-positive counts for incidents significantly increased, indicating improved incident classification. However, there was a slight increase in the number of false negatives associated with accidents. Additionally, as shown in Figure 5, the receiver operating curve (ROC) achieved a value of 0.93, indicating robust discrimination between accidents and incidents.

Table 5

RF Model Performance Before and After SMOTE

Model	Class	Precision	Recall	F1 – Score	Accuracy
Before SMOTE	Incident	0.68	0.27	0.38	0.93
	Accident	0.94	0.99	0.96	
After SMOTE	Incident	0.43	0.84	0.57	0.90
	Accident	0.98	0.90	0.94	

Figure 4

Confusion Matrix for RF Before and After SMOTE

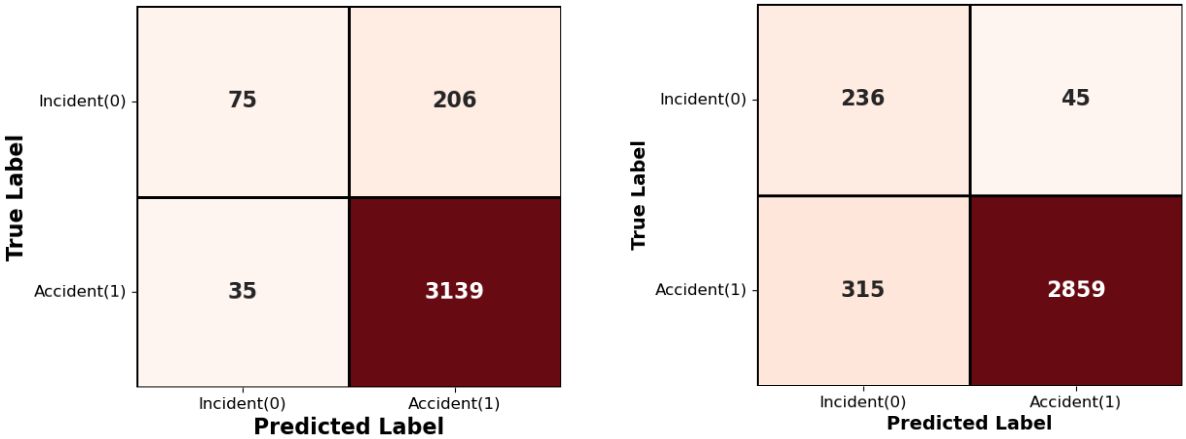
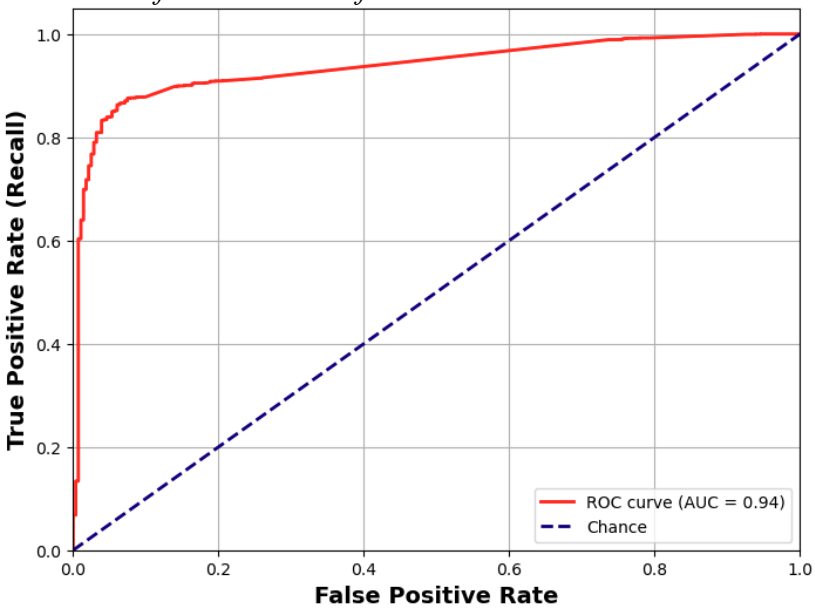


Figure 5

ROC Curve for RF Model After SMOTE



Support Vector Machine (SVM) Model Performance Before and After SMOTE

Similar results were obtained for SVM compared to those for RF. The SVM model performance parameters before and after SMOTE are listed in Table 6. Before applying SMOTE, the overall accuracy score was 0.93, with recall values of 0.26 and 0.88 for incidents and accidents, respectively. After applying SMOTE, the recall for incidents significantly improved to 0.87, while the overall accuracy slightly decreased to 0.88.

Table 6

SVM Model Performance Before and After SMOTE

Model	Class	Precision	Recall	F1 – Score	Accuracy
Before SMOTE	Incident	0.69	0.26	0.37	0.93
	Accident	0.94	0.99	0.96	
After SMOTE	Incident	0.39	0.87	0.54	0.88
	Accident	0.99	0.88	0.93	

As shown in Figure 6, a similar trend was observed in the confusion matrices of the SVM when compared with RF. After applying SMOTE, the true-positive counts for incidents increased from 82 to 245, indicating improved incident classification. However, there was a slight increase in false positives associated with accidents. Additionally, as shown in Figure 7, the receiver operating characteristic (ROC) achieved a value of 0.92, indicating robust discrimination between accidents and incidents in the dataset.

Figure 6

Confusion Matrix for SVM Before and After SMOTE

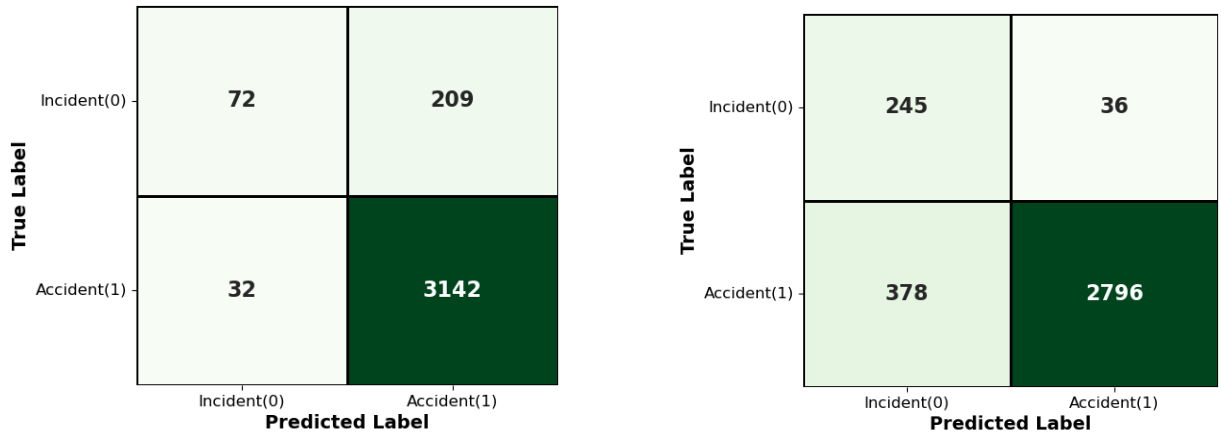
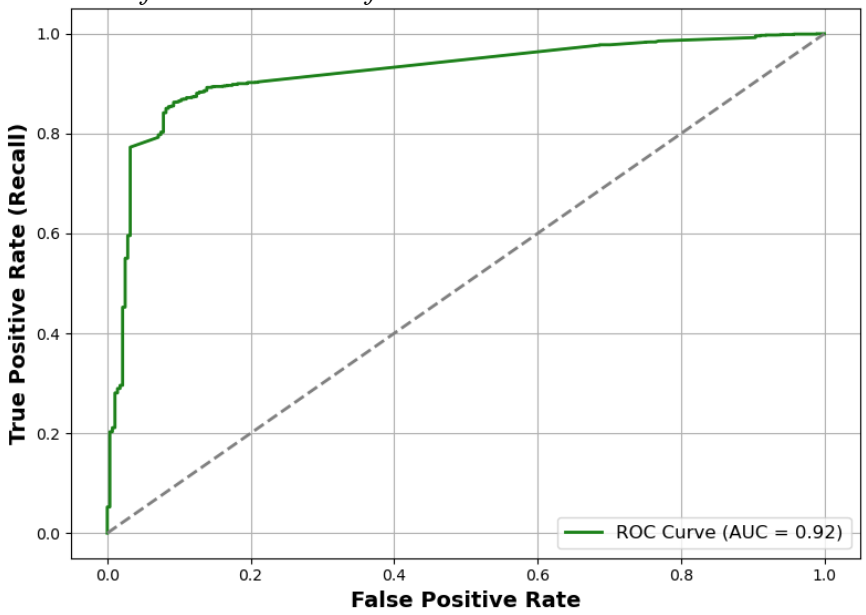


Figure 7

ROC Curve for SVM Model After SMOTE



Extreme Gradient Boost (XG Boost) Model Performance Before and After SMOTE

As reported in Table 7, the performance metrics of the XGBoost models also improved after applying SMOTE, similar to those of the RF and SVM models. The recall values for the incidents improved from 0.29 to 0.88, while the overall accuracy decreased slightly from 0.93 to 0.90.

Table 7

XG Boost Model Performance Before and After SMOTE

Model	Class	Precision	Recall	F1 – Score	Accuracy
Before SMOTE	Incident	0.68	0.29	0.41	0.93
	Accident	0.94	0.99	0.96	
After SMOTE	Incident	0.43	0.88	0.57	0.90
	Accident	0.99	0.90	0.94	

As shown in Figure 8, a similar trend was observed in the confusion matrices of the XG Boost compared to those of the RF and SVM models. After applying SMOTE, the true-positive counts for incidents increased from 82 to 238, indicating improved classification performance. However, there was a slight increase in false positives associated with accidents. Additionally, as shown in Figure 9, the receiver operating curve (ROC) achieved a value of 0.94, indicating robust discrimination between accidents and incidents.

Figure 8

Confusion Matrix for XG Boost Before and After SMOTE

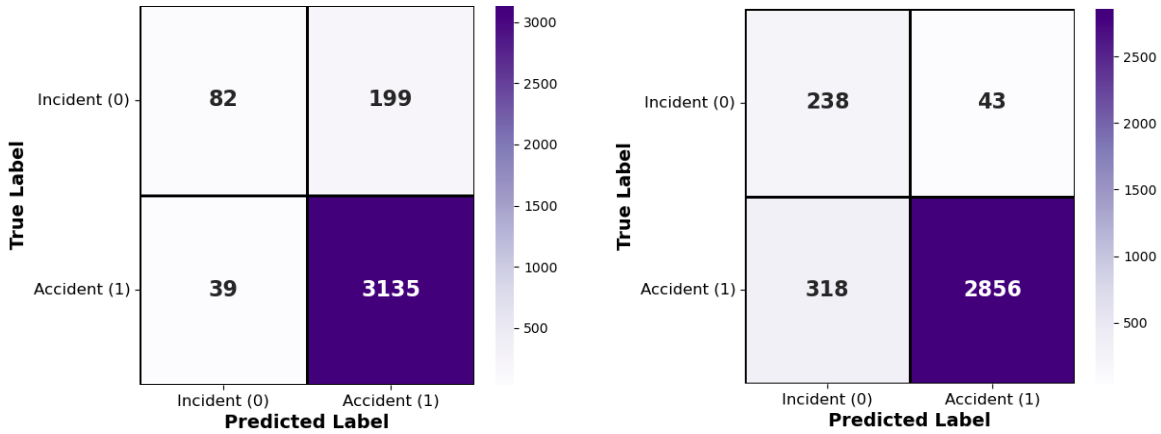
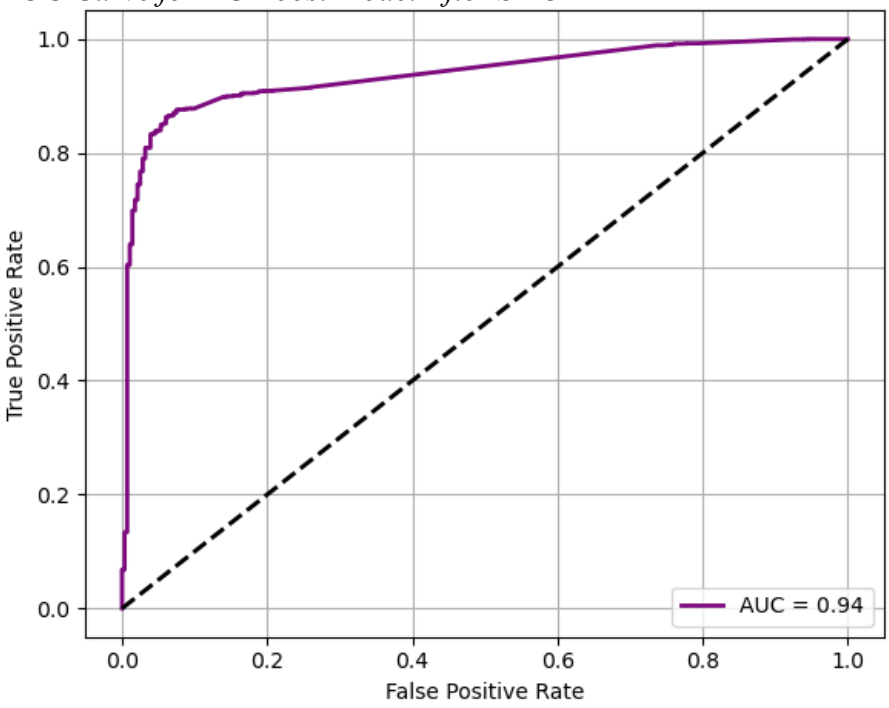


Figure 9

ROC Curve for XG Boost Model After SMOTE



The results of the current study indicate that, before applying SMOTE, all models yielded high overall accuracy but poor recall for the minority class (incidents). After implementing SMOTE, recall for the minority class (incidents) increased consistently across all three models: RF, SVM, and XG Boost. However, the overall accuracy was slightly decreased

Discussion

The primary purpose of the current study was to demonstrate whether supervised machine learning models, including Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XG Boost), can effectively classify aviation accidents and incidents using a focused set of aircraft, crew-related, and environment-related features derived from NTSB records. The findings of this study demonstrate that supervised ML models can effectively classify aviation events, including incidents and accidents. All three models achieved strong overall performance when evaluated using accuracy, precision, recall, and F1 score, indicating that data-driven approaches are well-suited for supporting quantitative safety analysis in aviation. The similarity in the high overall accuracy and performance across all models indicated that the current dataset contained consistent patterns that could have been captured by all algorithms. For instance, the relationship between environmental factors and the likelihood of an aviation event could have been learned similarly across all models. The findings of the current study are consistent with previous studies, indicating that supervised ML models, such as SVM and RF, can be used to classify or predict aviation safety-related events (Nanyoga et al., 2025; Omrani et al., 2023). The findings of the current study are also consistent with those of Silagyi and Liu (20230), who reported the advantages of using SVM modeling to predict the severity of personal injury in aviation accidents. This indicates an application of ML modeling in aviation that can be

developed to further support proactive safety within SMS. Specifically, the Random Forest, SVM, and XG Boost models can be applied to operational data to classify aviation events. The current study's findings indicate that safety-related data collected from routine operations, when combined with ML techniques, can provide fully or partially automated support for safety classifications for airline operators, aircraft manufacturers, and aircraft maintenance technicians.

Moreover, all three models demonstrated good overall accuracy both before and after applying SMOTE. However, RF and XGB achieved slightly higher accuracies after applying SMOTE than SVM did. A plausible explanation is that tree-based models, such as RF and XGB, better handle complex, overlapping patterns in the data when synthetic samples are generated with SMOTE. However, SVM that relies on decision boundaries could be affected by these changes. Before applying class-imbalance techniques, all models performed better in the majority class (accidents). However, after applying SMOTE, the models' ability to correctly identify the minority incident class improved substantially, as evidenced by marked gains in recall and F1 scores. Simultaneously, the overall accuracy decreased slightly, which is a common trade-off when correcting for class imbalance, but remained high at or above 0.88 across all models. A plausible explanation for these findings is that, after implementing SMOTE, the representation of the minority class in the dataset (incidents) was increased by creating synthetic samples. This helped the models learn more patterns associated with the minority class (incidents), thereby improving their detection rates. However, when SMOTE was applied, the dataset was rebalanced by creating synthetic samples of the minority class, which could have expanded the decision regions associated with the minority class. Consequently, there could be an increased tendency for the models to classify a few accidents as incidents, leading to increased false-positive rates and reducing overall accuracy. Aligned with prior work, the findings of the current study also demonstrate that supervised ML models, when combined with class-imbalance techniques, can be effective for classifying aviation events. The results of the current study are consistent with those of previous studies that prioritized the use of class imbalance techniques to handle unequal groups (Dangut et al., 2022; Nanyoga et al., 2025)

In addition, receiver operating characteristic (ROC) curve analyses indicated area under the curve values above 0.90 for the RF, SVM, and XG Boost models following SMOTE, which reflects strong discrimination between accidents and incidents across decision thresholds and underscores the potential value of these models in informing SMS processes. This finding is further supported by a significant improvement in recall for the minority class (incidents). From an aviation safety standpoint, an increase in recall is critical, as it projects a reduced likelihood of not identifying incident events that could potentially be emerging hazards. For example, greater recall values would enable airlines to identify a high proportion of unstable approaches or altitude deviation incidents. This information can help safety managers or analysts develop proactive measures before runway excursions or controlled flight into terrain (CFIT) events occur. The application of SMOTE resulted in higher recall values; however, as reported in Tables 5, 6, and 7, there was a slight reduction in the overall accuracy of all three ML models. Therefore, the reader is reminded that this should not be treated as a reduction in the ML models' operational efficiency. In high-risk domains, such as aviation, false negatives may yield greater safety consequences than false positives. These findings align with the objective of proactive hazard identification under the safety risk management pillar of the SMS (ICAO, 2018). The current findings also imply that organizations implementing ML models within their SMS

framework to classify aviation-related hazards can be successful by prioritizing models with high recall, though this may result in high false positives. Although false-positive events increase the workload of safety analysts, reviewing flagged cases remains critical for identifying hazards early.

These results indicate that the findings of this study offer multiple practical implications for enhancing safety through proactive and predictive data-analytical decision-making, aligning with ICAO's (2018) SMM. A robust SMS is always data-driven, and integrating ML to proactively identify hazards can strengthen its foundation. Safety risk management (SRM) is one of the critical pillars of the SMS framework, and ML models can be augmented to the SRM pillar to proactively identify hazards and prioritize them for a detailed review. Moreover, combining ML techniques as decision-making tools with human resources in the SMS framework will enhance the robustness of safety assurance and support continuous safety improvement in aviation organizations.

Conclusion

The findings of this study demonstrate that supervised ML models can effectively classify aviation events, such as accidents and incidents. All ensemble learning tree-based models, including RF and XGB, and decision boundary-based models, including SVM, achieved strong overall performance. The application of SMOTE improved performance for the minority class and increased true positive rates. These findings emphasize the importance of data-driven practices and the advantages of integrating ML-based approaches for classifying hazards in aviation safety.

Limitations

The current study has several limitations that should be considered when interpreting the results. First, the analysis relied exclusively on historical NTSB (n.d.) accident and incident data, which are subject to reporting practices, coding conventions, and potential underreporting that may introduce bias and limit the generalizability of the findings to other jurisdictions, time periods, or data sources. Second, although 15 key predictors were selected to balance operational relevance with data quality, additional potentially meaningful variables were excluded because of high levels of missing data or preprocessing constraints, which may have limited the models' ability to fully capture the complexity of real-world accident and incident causation. Third, the use of SMOTE to address class imbalance inherently anchors the models to the structure of the existing dataset, improving the detection of typical incident patterns but potentially reducing sensitivity to rare, emerging, or qualitatively different event types that are not well represented in historical data. Finally, model training and evaluation were conducted in a research environment rather than an operational setting; therefore, the study did not assess how safety practitioners would interpret and use model outputs in practice or how false positives and false negatives might influence decision-making and resource allocation.

Future Research

In the context of future research, there is a clear opportunity to extend this work by integrating multiple complementary data sources, such as Flight Operational Quality Assurance,

Line Operations Safety Audit, Aviation Safety Action Program, and Aviation Safety Information Analysis and Sharing system with NTSB records to create richer multi-source datasets that span both the precursors and outcomes of safety events. Such integrated datasets would enable researchers to examine whether the model performance, stability, and transferability improve across different operational environments, operator types, and flight phases. Future studies should also conduct structured comparisons of SMOTE with alternative imbalance-handling strategies, including cost-sensitive learning, ensemble approaches that emphasize the minority class, and anomaly detection techniques explicitly designed for rare events, while incorporating explainable artificial intelligence methods to clarify which features most strongly drive classification decisions. This ML modeling approach could also be applied within an organization (e.g., airline, flight school, or airport) with a sufficiently large internal dataset of operations and safety events. Finally, prospective or near-real-time implementations of these models within SMS, coupled with evaluations of user trust, alert response, and changes in observed safety outcomes, would provide essential evidence of the practical utility of supervised machine learning as a tool for proactive, data-driven SRM in the aviation industry.

Declaration of AI-assisted technologies in the writing process

During the submission check of this work, the authors have used Paper pal for spelling and formatting**.** Paperpal

References

Black, J. E., Kueper, J. K., & Williamson, T. S. (2023). An introduction to machine learning for classification and prediction. *Family Practice, 40*(1), 200–204. https://doi.org/10.1093/fampra/cmac104

Borjalilu, N. (2024). Risk assessment and machine learning models. In *The future of risk management*. IntechOpen. https://doi.org/10.5772/intechopen.1005485

Boukerche, A., & Wang, J. (2020). Machine learning-based traffic prediction models for intelligent transportation systems. *Computer Networks, 181*, 107530. https://doi.org/10.1016/j.comnet.2020.107530

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cankaya, B., Topuz, K., & Glassman, A. M. (2023). Business inferences and risk modeling with machine learning: The case of aviation incidents. In *Proceedings of the 56th Annual Hawaii International Conference on System Sciences* (pp. 1238–1248). https://hdl.handle.net/10125/102783

Chang, P. W., & Newman, T. B. (2024). Receiver operating characteristic curves: The basics and beyond. *Hospital Pediatrics, 14*(7), e330–e334. https://doi.org/10.1542/hpeds.2023-007462

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Dangut, M. D., Skaf, Z., & Jennions, I. K. (2022). Handling imbalanced data for aircraft predictive maintenance using the BACHE algorithm. *Applied Soft Computing, 123*, 108924. https://doi.org/10.1016/j.asoc.2022.108924

Delgado, R., & Núñez-González, J. D. (2022). Bayesian network-based over-sampling method with application to indirect cost-sensitive learning. *Scientific Reports, 12*, Article 9144. https://doi.org/10.1038/s41598-022-12682-8

Demir, G., Moslem, S., & Duleba, S. (2024). Artificial intelligence in aviation safety: Systematic review and biometric analysis. *International Journal of Computational Intelligence Systems, 17*(1), 279. https://doi.org/10.1007/s44196-024-00671-w

Federal Aviation Administration. (n.d.). *Safety: Continuous improvement*. https://www.faa.gov/safety/continuous-improvement

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://www.jstor.org/stable/2699986

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Howell, C. (2025). *6 data mining challenges for aviation safety managers*. Enhancing Aviation Safety: Expert Insights, Tips, and Updates from SMS-Pro. https://aviationsafetyblog.asms-pro.com/blog/aviation-sms-data-mining-challenges-for-safety-managers

Huang, X., Kroening, D., Ruan, W., Sharp, J. J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review, 37*, 100270. https://doi.org/10.1016/j.cosrev.2020.100270

International Civil Aviation Organization. (2018). *Doc 9859: Safety management manual* (4th ed.). https://caainternational.com/wp-content/uploads/2018/05/9859_unedited_en.pdf

Karaburun, N. N., Hatipoğlu, S. A., & Konar, M. (2025). Aircraft takeoff speed prediction with machine learning: Parameter analysis and model development. *The Aeronautical Journal, 129*(1336), 1534–1549. https://doi.org/10.1017/aer.2024.164

Khalid, S., Song, J., Azad, M. M., Elahi, M. U., Lee, J. H., Jo, S., & Kim, H. S. (2023). A comprehensive review of emerging trends in aircraft structural prognostics and health management. *Mathematics, 11*(18), 3837. https://doi.org/10.3390/math11183837

Kuleshov, V., Aygumov, T., Zolkin, A., Tychkov, A. S., & Bityutskiy, A. S. (2023). Use of machine learning for prevention of incidents and reduction of occupational risks at the workplace. In *International Conference on Digital Transformation: Informatics, Economics, and Education (DTIEE2023)*. https://doi.org/10.1117/12.2680682

Mehta, J., Vatsaraj, V., Shah, J., & Godbole, A. (2021, July). Airplane crash severity prediction using machine learning. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT51525.2021.9579711

Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology, 75*(1), 25–36. https://doi.org/10.4097/kja.21209

Nanyonga, A., Wasswa, H., & Wild, G. (2023, December). Aviation safety enhancement via NLP and deep learning: Classifying flight phases in ATSB safety reports. In *2023 Global Conference on Information Technologies and Communications (GCITC)* (pp. 1–5). IEEE. https://doi.org/10.1109/GCITC60406.2023.10426306

Nanyonga, A., Wasswa, H., Joiner, K., Turhan, U., & Wild, G. (2025). Explainable supervised learning models for aviation predictions in Australia. *Aerospace, 12*(3), 223. https://doi.org/10.3390/aerospace12030223

National Archives and Records Administration. (2026). *Safety management systems (14 C.F.R. Part 5)*. Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-14/chapter-I/subchapter-A/part-5

National Transportation Safety Board. (n.d.). *Accident data*. https://www.ntsb.gov/safety/data/Pages/Data_Stats.aspx

New, M. D., & Wallace, R. J. (2025). Classifying aviation safety reports: Using supervised natural language processing in an applied context. *Safety, 11*(1), 7. https://doi.org/10.3390/safety11010007

Omrani, F., Etemadfard, H., & Shad, R. (2024). Assessment of aviation accident datasets in severity prediction through machine learning. *Journal of Air Transport Management, 115*, 102531. https://doi.org/10.1016/j.jairtraman.2023.102531

Omar, E. D., Mat, H., Abd Karim, A. Z., Sanaudi, R., Ibrahim, F. H., Omar, M. A., & Goh, B. L. (2024). Comparative analysis of logistic regression, gradient boosted trees, SVM, and random forest algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. *International Journal of Nephrology and Renovascular Disease, 17*, 197–204. https://doi.org/10.2147/IJNRD.S461028

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*. https://doi.org/10.48550/arXiv.2010.16061

Ramírez, J. G. C., Islam, M. M., & Even, A. I. H. (2024). Machine learning applications in healthcare: Current trends and future prospects. *Journal of Artificial Intelligence General Science, 1*(1). https://doi.org/10.60087/jaigs.v1i1.33

Reason, J. (1990). *Human error*. Cambridge University Press. https://doi.org/10.1017/CBO9781139062367

Riantika, I., Sartono, B., & Notodiputro, K. A. (2024). Effectiveness of SMOTE-ENN to reduce complexity in classification model. *Indonesian Journal of Statistics and Its Applications, 8*(1), 70–82. https://doi.org/10.29244/ijsa.v8i1p70-82

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*. https://doi.org/10.48550/arXiv.1708.08296

Silagyi, D. V., II, & Liu, D. (2023). Prediction of severity of aviation landing accidents using support vector machine models. *Accident Analysis & Prevention, 187*, 107043. https://doi.org/10.1016/j.aap.2023.107043

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sui, Y., Lu, K., & Fu, L. (2021). Prediction and analysis of novel key genes ITGAX, LAPTM5, and SERPINE1 in clear cell renal cell carcinoma through bioinformatics analysis. *PeerJ, 9*, e11272. https://doi.org/10.7717/peerj.11272

Tafur, C. L., Camero, R. G., Rodríguez, D. A., Rincón, J. C. D., & Saenz, E. R. (2025). Applications of artificial intelligence in air operations: A systematic review. *Results in Engineering, 25*, 103742. https://doi.org/10.1016/j.rineng.2024.103742

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401–409. https://doi.org/10.1037/h0058700

Wiegmann, D. A., & Shappell, S. A. (2001). Human error analysis of commercial aviation accidents: Application of the Human Factors Analysis and Classification System. *Aviation, Space, and Environmental Medicine, 72*(11), 1006–1016. https://pubmed.ncbi.nlm.nih.gov/11718505/

Yang, S., & Berdine, G. (2024). Confusion matrix. *The Southwest Respiratory and Critical Care Chronicles, 12*(53), 75–79. https://doi.org/10.12746/swrccc.v12i53.1391

Yang, Y., He, K., Wang, Y. P., Yuan, Z. Z., Yin, Y. H., & Guo, M. Z. (2022). Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods. *Physica A: Statistical Mechanics and Its Applications, 595*, 127083. https://doi.org/10.1016/j.physa.2022.127083