Towards Trustworthy Aircraft Safety: Explainable AI for Accurate Incident and Accident Predictions

Maryam Amin[1], Umara Noor[1], Manahil Fatima[1], Zahid Rashid [2], Jörn Altmann [2, 3, 4]

[1] Department of Software Engineering, Faculty of Computing and Information Technology, International Islamic University, Islamabad 44000, Pakistan
[2] Technology Management Economics and Policy Program, College of Engineering, Seoul National University, Seoul, South Korea
[3] Institute of Engineering Research (IOER), College of Engineering, Seoul National University Seoul, South Korea
[4] Integrated Major in Smart City Global Convergence, Seoul National University, Seoul, South Korea

maryam.phdcs190@iiu.edu.pk, umara.zahid@iiu.edu.pk, manahil.mscs1111@iiu.edu.pk, rashidzahid@snu.ac.kr, jorn.altmann@acm.org

Abstract. Despite technological advancements, ensuring aircraft safety remains a challenge, however, Machine learning (ML)-based approaches for predicting future incidents play a crucial role in addressing flight safety. As ML models increase in complexity, their decision-making process becomes less transparent, posing significant challenges to trustworthiness. While simpler models demonstrate lower accuracy, more intricate models such as deep neural networks achieve higher accuracy but sacrifice interpretability. In this study, we enhance trustworthiness in aircraft safety prediction by leveraging a dataset of past accidents and incidents to prevent similar accidents from occurring in the future. To achieve this, we apply Random Forest and Extreme Gradient Boosting models to classify different categories of aircraft incidents. Additionally, we apply two powerful explainable artificial intelligence (XAI) techniques: Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) to provide insights into both local and global predictions made by the models. Notably, our results reveal high accuracy in these predictions while maintaining trustworthiness. This research contributes to the advancement of XAI and offers valuable insights for safety-critical applications and decision support systems.

Keywords: Aircraft Safety, Random Forest, Extreme Gradient Boosting (XGBoost), Explainable Artificial Intelligence, Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive exPlanations (SHAP).

1 Introduction

Safety of aircraft is critical as it poses a significant risk of accidents and incidents, threatening aircraft operations and imposing substantial economic costs on the aviation industry. Aircraft incidents and accidents can have a profound impact, resulting in loss of life, severe injuries, and significant economic losses, as well as

damaging the reputation of airlines and aircraft manufacturers and undermining public trust in the aviation industry. The annual safety report by the International Civil Aviation Organization revealed a global accident rate of 2.05 accidents per million departures in 2022, representing a 6.3% increase from the previous year's statistics [23].

As the fastest means of transportation, the aircraft industry is poised for significant expansion, with global demand expected to triple by 2050, thereby increasing the demand for aviation safety to meet the escalating requirements [25]. However, the dominant safety frameworks currently utilized by air traffic controllers are largely reactive, focusing on minimizing the impact of safety incidents after they happen. However, such systems are often criticized as a basic form of risk management, and during emergencies, they become less efficient and resource-intensive. Pilots and air traffic controllers rely on real-time data to make their safety-critical decisions, ensuring timely and effective responses to evolving situations in the flight.

There exist complex non-linear interactions and interdependencies between various factors such as mechanical, weather, human, and communication, coupled with their dynamic evolution over time, pose significant challenges in developing precise physical models that can accurately capture the complex relationships governing aircraft safety. Machine learning (ML) has demonstrated its ability to accurately model and predict intricate physical phenomena, leading to the widespread application of this technology especially for predicting the safety of complex systems. In the aviation industry, ML-based predictive safety approaches are important which prioritize risk prevention, anticipate hazards, and mitigate them before incidents occur. However, due to black-box nature of ML models, their decision-making process is difficult to interpret by humans, leading to a lack of transparency and subsequent trust issues.

In literature ML-based aircraft's safety prediction demonstrates impressive predictive capabilities but their opacity of decision-making processes undermines trust among airline stakeholders, thereby severely limiting their widespread adoption in real-world applications, where safety and trustworthiness are paramount. The safety prediction made by simpler models (e.g., linear regression, decision trees) demonstrates limited predictive accuracy while exhibiting strong power of prediction's interpretability, whereas more complex models (e.g., deep neural networks) achieve superior accuracy but offer low reasoning of the decision-making process [1]. Although some surveys and reviews have provided general guidelines for explanations of predictions made by ML models there exists a significant gap between theoretical advancements and practical implementation of integrating trustworthiness in aircraft safety predictions as only two research endeavors on it, one for aircraft failure diagnosis [15] and other for runway surface contamination [16]. Also, the relative importance of each feature and their impact in the decision-making process remain unexplored which poses a significant challenge in the pursuit of trustworthy ML models.

The main objective of this paper is to investigate the application of ML methods in the prediction of aircraft safety with higher accuracy and to demonstrate reasoning for the model's prediction to ensure trustworthiness. This is done for predicting different

types of aircraft accidents such as accidents, incidents, criminal occurrence, hijacking, ground fire, sabotage, and other unknown occurrences, through developing two autoregressive-inspired time series ensemble approaches of Random Forest (RF) and Extreme Gradient Boosting (XGBoost) classification models which are known for their efficiency, speed, and accuracy on large datasets [22]. The models are trained on a vast collection of historical accident and incident records from around the world, providing a rich source of information that enables the models to learn from past experiences and improve their predictive capabilities. Similar to other ensemble methods, RF and XGBoost are inherently not interpretable, therefore, we utilize two powerful techniques: Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to develop simplified models that provide both global and local explanations, enabling the understanding of the model's predictions and the contribution of individual features to the output. The performance of our ML models is evaluated and compared with other similar accident/incident prediction approaches. The results of this study reveal remarkable predictive accuracy while maintaining transparency and ensuring trustworthiness. Our findings contribute to advancing the field of XAI and provide valuable insights for safety-critical applications and decision support systems.

Following are main contributions of this research.

- The study's results exhibit outstanding predictive accuracy of RF (Random Forest) and XGBoost models in predicting aviation safety incidents, while maintaining a high degree of transparency and ensuring the trustworthiness of the models.
- XAI techniques like LIME and SHAP are demonstrated to provide clear local and global explanations of model predictions in aviation safety systems.
- The study's findings identified potential issues in aviation systems before they resulted in critical failures, fostering trust in AI systems, which is crucial for their adoption in safety-critical applications.

This research paper is organized into the following sections. Section 2 covers the literature review, Section 3 provides details of Methodology while the results and discussions are described in Section 4. Finally, the paper ends with concluding remarks in Section 5.

2 Literature Review

Aircraft safety prediction by using ML algorithms is a highly focused research area and many researchers have been contributing regarding different dimensions. The research in [2] applies data-mining and sequential deep-learning techniques to accident investigation textual reports published by the National Transportation Safety Board (NTSB) to get predictions regarding adverse events. Zeng et. al. [3] introduce an innovative method combining the least absolute shrinkage and selection operator (LASSO) with long short-term memory (LSTM) for aviation safety prediction which demonstrates improved efficiency and robustness while maintaining excellent generalization ability. The study in [4] presents a novel deep learning technique based

on auto-encoders and bidirectional gated recurrent unit networks to handle extremely rare failure predictions in aircraft predictive maintenance modeling. The authors of [5] introduce an analytical methodology which combines data cleaning, correlation analysis, classification-based supervised learning, and data visualization to identify critical parameters and remove extraneous factors. The research in [6] develops a methodology to identify and classify human factor categories from textual aviation incident reports by using semi-supervised Label Spreading and supervised Support Vector Machine (SVM). Silagyi in [7] applies SVM models to predict the severity of aircraft damage and personal injury during approach and landing accidents. The study in [8] investigates cognitive workload in aviation by applying a stacking ensemble machine learning algorithm (support vector machine, random forest, and logistic regression) on electroencephalogram (EEG) data collected from ten collegiate aviation students during live-flight operations in a single-engine aircraft.

Focusing on aircraft safety prediction, the research in [12] explores the value and necessity of XAI when using DNNs (Deep Neural Networks) for Predictive Maintenance in Aerospace Integrated Vehicle Health Management. Saraf et. al. in [13] investigate the intersection of AI and aviation safety by exploring implications, possibilities, innovation capacity, skills development, and ethical regulation. The authors in [14] conduct a comprehensive literature review to explore the applications of AI in safety-critical domains by identifying Themes and Techniques, Future Research Directions, and Practical Implications. The research in [18] analyzes AI's usefulness within the aviation domain and synthesizes findings into a conceptual framework called the Descriptive, Predictive, and Prescriptive model.

Hernandez et. al. in [17] focus trustworthiness of AI-based automated solutions in air traffic management and propose a novel framework which encompasses technical robustness, transparency, security, and safety. The practical challenges related to need of transparency and explainability, are also presented. The study in [15] addresses the challenge in aviation maintenance and proposes an XAI methodology, called Failure Diagnosis Explainability (FDE) which enhances transparency and enables checking whether a new failure aligns with expected diagnosis values. The research in [16] combines XGBoost models with the XAI technique SHAP to address the challenge of runway surface contamination (e.g., snow, ice, slush) during winter seasons, which reduces tire-pavement friction and poses safety risks for aviation.

In the existing literature, there are some research gaps. First, trustworthiness is often overlooked in the context of aircraft incident/accident predictions. Although ML models demonstrate strong predictive power but reliable, transparent, and safety are crucial for building trust among aviation professionals and passengers. Second, some models may sacrifice accuracy for interpretability striking the right balance between accuracy and interpretability remains a challenge. Third, there remains a gap between theoretical advancements and practical implementation of integrating trustworthiness in aircraft safety predictions as only two research endeavors on it for aircraft failure diagnosis [15] and runway surface contamination [16].

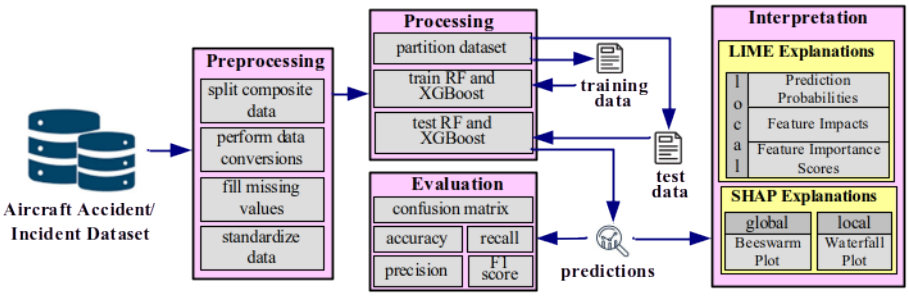

Fig. 1. The Proposed Methodology

3 Methodology

The methodology comprises of steps as shown in Figure 1, each step is discussed in detail in the following subsections.

3.1 Dataset

The dataset selected for this research focuses on aviation safety and comprises a comprehensive collection of worldwide accidents, failures, and hijackings involving airliners, corporate jets, and military transport aircraft [24]. The dataset is selected because examining past accidents, researchers can determine the underlying causes and contributing factors, which can inform strategies to prevent similar accidents in the future. With 23519 data points and 23 features, this extensive dataset is contained in a single CSV file, covering incidents from 1919 to November 2, 2022, providing a valuable resource for analysis and insight into aviation safety trends and patterns. The dataset is obtained from kaggle which is a vast repository of publicly accessible datasets across various domains.

3.2 Preprocessing

Preprocessing plays a crucial role in ML as it ensures data correctness and consistency, and suitability for analysis. The selected dataset contains missing information, composite values and inconsistent data format which need completeness, splitting and standardization in order to improve its quality. A two-step preprocessing is performed; one by using MS Excel and other by using Python. In Microsoft Excel the data is split based on delimiter characters such as : '|' , '/' , ':' for instance the column Onboard_Crew contains composite data containing the number of 'Fatalities' and 'occupants' of an incident separated by '/' which is splitted into two columns 'Onboard_Crew_ Fatalities' and 'Onboard_Crew_Occupants'. The data of date column such as Incident_Date is converted into timestamp. By using Python Null values in Object column are replaced by forward fill method (ffil) and question marks used as missing values in different columns were first replaced by NAN which are

then replaced by ‘unknown’. The Null values in integer columns are replaced with mean values and the dates in incident_date column are converted into dd-mm-yyyy format and the categorical data is converted into numerical data.

3.3 Feature Selection

RF extracts features in a recursive manner, selecting the most informative features at each node of the decision tree. The process is repeated multiple times, resulting in a collection of decision trees, each with their own set of extracted features. XGBoost extracts features in a greedy manner, selecting the most informative features at each node of the decision tree. The process is repeated multiple times, resulting in a collection of decision trees, each with their own set of extracted features. Therefore, both RF and XGBoost ensemble approaches reduce overfitting and improve generalization. Finally, columns containing text narratives, such as Incident_cause(es) and Incident_subcategory, from the dataset used in this study while these narratives could indeed provide valuable insights for explaining aircraft incident and accident predictions, we plan to explore these aspects in future work to enhance the comprehensiveness of our analysis.

3.4 Random Forest

This study builds prediction models that classify aviation events using the resilient Random Forest (RF) model by using dataset of past incidents. As an advancement of the bagging (Bootstrap Aggregating) technique, RF was created in 2001 and combines several decision trees to increase the model's robustness and accuracy [21]. It is renowned for its effectiveness, speed, and accuracy on big datasets with lots of variables and is utilized for both regression and classification problems. In a variety of industries, including finance, healthcare, and e-commerce, RF is used to predict risks.

Random Forest is chosen for its exceptional performance, offering a rare combination of speed, accuracy, and scalability. It efficiently processes large datasets with numerous features, minimizing bias and robustly handling missing values and outliers, making it an ideal algorithm for our analysis. To predict a new point x:

Classification: Let $\widehat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then

$$\widehat{C}_{rf}^{B}(x) = majority\ vote\left\{\widehat{C}_b(x)\right\}_1^B \tag{1}$$

3.5 eXtreme Gradient Boosting

Using a large dataset of historical incidents, this research builds prediction models that categorize aircraft incidents using the cutting-edge XGBoost algorithm. Since its release in 2014, XGBoost, a highly scalable and effective implementation of gradient boosting decision trees has gained a great deal of attention and praise. It has proven

successful in machine learning competitions and has been used in a variety of transportation risk assessment applications across a range of industries, including road traffic, aviation, and shipping.

Because of XGBoost's exceptional performance, handling of big datasets, and speed of computation, it was chosen to train the airplane safety predictor. Furthermore, multicollinearity, a common problem in our data is successfully reduced using XGBoost's decision tree ensemble technique, guaranteeing reliable and accurate predictions. In real life, the model has to be trained on the data, which are often represented as an n-dimensional vector of outcomes (y) and a n times m matrix of input variables (X). A decision tree fk(x) is obtained at each iteration by minimizing an objective function.

$$obj\left(f_k(x)\right)=\sum_{i=1}^{n} L\left(y_i,\hat{f}\left(x_i\right)^{\lfloor k-1\rfloor}+f_k(x_i)\right)+\Omega\left(f_k(x)\right) \tag{2}$$

where (x_i,y_i) is the i-th observation, $\sum_{i=1}^{n} L\left(y_i,\hat{f}\left(x_i\right)^{\lfloor k-1\rfloor}+f_k(x_i)\right)$ is the empirical estimate of the loss, $\hat{f}\left(x_i\right)^{\lfloor k-1\rfloor}$ is the current estimate of the model(i.e., the model computed at the previous iteration k-1), and $\Omega\left(f_k(x)\right)$ is a penalty term that penalized the tree complexity.

3.6 Experimental Setup

The experiment was conducted on a laptop equipped with a 12th Gen Intel® Core™ i5-1235U 1.30 GHz processor, 8.00 GB RAM, and a 64-bit Windows 10 operating system with an x64-based processor. The Python code was developed and executed within Jupyter Notebook to perform tasks such as data analysis and scientific exploration. This setup provided a robust environment for executing computational tasks efficiently, ensuring that the data analysis processes were both reliable and reproducible. The choice of Jupyter Notebook facilitated an interactive coding experience, allowing for real-time visualization and iterative development, which are crucial for thorough scientific investigation.

3.7 Models Training

Our goal is to predict aircraft safety based on previous accident/incident dataset and the features of the dataset are used to predict incident category. The types of incidents are labeled into six classes; ***Accident*** class with 19543 records (Label 0), ***Criminal occurrence*** (sabotage, shoot down) having 1256 entries (Label 1), ***Hijacking*** with 1092 (Label 2), ***Incident*** having 12 records (Label 3), ***occurrence unknown*** with 570 entries (Label 4) and ***other occurrence*** (ground fire, sabotage) with 1046 records (Label 5). The dataset is partitioned into eighty percent training (18815 records) and twenty percent test (4704 records) data frames. Since the

categorical data is converted into numerical data therefore all six classes are labeled with numbers from 0 to 5. Both RF and XGBoost are trained on training data.

4 Results and Discussions

4.1 Performance of Models

On test dataset both models demonstrated an accuracy of 90.11% and 82.91% respectively. Considering the substantial imbalance in the dataset, where only 5.1% of cases belong to the incident class, relying solely on accuracy as a performance metric for classification in this research is inadequate. Accuracy may not accurately reflect the model's performance on the minority class. Therefore, the performance of the RF classification model is assessed using confusion matrices, which provide a detailed breakdown of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) predictions. Table 1 displays the confusion matrix for RF model's predictions, with the columns representing predicted classes and the rows representing actual classes, offering a clear visualization of the model's performance. The high values of TP for class 0, 2 and 5 gives confidence in model's performance whereas the marginal scores of other classes and few zeros in case of class 3 are due to imbalance distribution of classes in dataset. The results of confusion matrix provide essential information about model's performance and help analyze misclassifications.

For evaluation we also apply precision, recall and F1 score for each classification class as shown in Table 2 which makes a weighted average precision, recall and F1 score as 0.89, 0.90 and 0.88 respectively. These promising scores reveal model's powerful predictive ability on unseen data and its performance beyond training data.

In Table 3 the effectiveness of our models is evaluated by comparing their performance to similar research endeavors, notably [15], which tackled aircraft failure diagnosis prediction, and [16], which addressed runway surface contamination prediction, providing a framework for evaluating our approach's efficacy. The results show that our research employs Random Forest and XGBoost machine learning models to forecast aircraft accidents and incidents worldwide, yielding high accuracy scores of 90.11% and 82.91%, respectively, demonstrating the effectiveness of our approach in predicting aviation safety risks.

Table 1. Confusion Matrix

	0	1	2	3	4	5
0	3815	9	13	0	16	18
1	159	73	4	0	2	29
2	74	0	165	0	0	1
3	1	0	0	0	0	0
4	97	0	0	0	24	2
5	34	5	1	0	0	162

Table 1. Evaluation Metrics

class	precision	recall	f1-score
0	0.91	0.99	0.95
1	0.84	0.27	0.41
2	0.90	0.69	0.78
3	0	0	0
4	0.57	0.20	0.29
5	0.76	0.80	0.78

Table 1. Comparison with Existing Work

Ref	dataset	Model Used	Accuracy	Explanation
15	Netherland	RF	81%	FDE
16	Norway	XGBoost	NA	SHAP
This work	global	RF XGBoost	90.11% 82.91%	SHAP and LIME

4.2 Interpretation

The complexity of RF and XGBoost models, which aggregate scores from numerous decision trees (between 50 and 250) renders them challenging to interpret and comprehend. This opacity has contributed to the growing interest in Explainable Artificial Intelligence (XAI), as the increasing reliance on sophisticated black-box algorithms like XGBoost and deep neural networks necessitates a better understanding of their decision-making processes [16]. XAI refers to a set of processes and methods designed to enhance human understanding and trust in machine learning algorithms [27]. As AI models grow in complexity, their decision-making processes become increasingly opaque, posing challenges for interpretability. XAI techniques aim to illuminate these "black-box" models, making their predictions more transparent and reliable. XAI encompasses a range of techniques and methodologies aimed at demystifying the complex decision-making processes of black-box ML models, thereby rendering their predictions more comprehensible, trustworthy, and accountable, thereby fostering greater human understanding and confidence in AI-driven decision-making. XAI is actively used in diverse fields such as agriculture, games, information systems, smart cities, social media, sports, [19].

SHapley Additive exPlanations

SHAP is a powerful framework for explaining the predictions of ML models [20]. SHAP (SHapley Additive exPlanations) is based on Shapley values, which have their roots in cooperative game theory. SHAP provides global as well as local explanations for predictions and can be used for tabular, text, image, and genomic data. It helps us understand why a specific instance received a particular prediction. SHAP treats any supervised learning model as a black box and calculates Shapley values for each feature by evaluating all potential feature combinations and their respective contributions. This method assesses the impact of individual features on a model's predictions. It connects optimal credit allocation (determining how much each feature contributes) with local explanations.

For SHAP explanations the predictions made by XGBoost model is utilized. The goal of using shapley values is to distribute the prediction among variables. This makes Shapley values part of the additive feature attribution methods, which means they have an explanation model that is a linear function of binary variables:

$$g(z) = \varnothing_0 + \sum_{j=1}^{m} \varnothing_j z_j \tag{3}$$

where $z \in \{0,1\}m$ is a coalition vector giving the absence/presence of input variables in x and m is the number of variables in the original model. Methods with this explanation model assign an importance effect $\varnothing_j$ to each variable and summing the effects of all variables approximates the output of the original model.

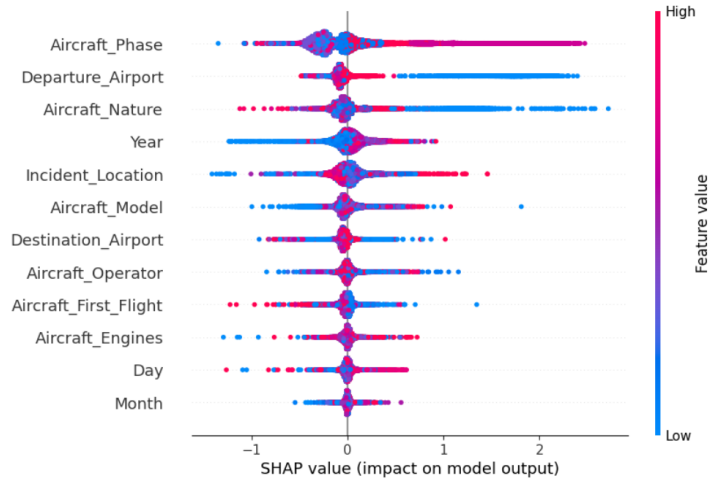

Fig. 2. SHAP Beeswarm Plot

SHAP values were intended for localized explanations, providing insights into individual predictions. However, Tree SHAP's high-speed estimations enable the generation of local explanations for entire datasets, facilitating a more extensive understanding of the model's overall performance. By plotting local explanations for a complete test set, we can amalgamate individual insights into a comprehensive global understanding of the model's behavior and decision-making processes.

Figure 2 shows a beeswarm plot of local SHAP values for each test sample, aggregated to form a global explanation of the classification model's overall performance, revealing how the model generates predictions for all instances in the test set. The plot displays the variables in decreasing order of importance, with increasing SHAP values (moving right on the x-axis) indicating a higher likelihood of accident class and negative values indicating a lower likelihood, with point density and color representing individual variable values.

First important observation from Figure 2 is that globally the most impactful features in prediction of accident/incident are ‘Aircraft-phase’, ‘departure_Airport’ and ‘Destination_Airport’ which reveals that landing and takeoff flights at airport are most critical stages of a flight. Second, the ‘Aircraft_Nature’ and ‘Aircraft_Model’ are also impactful as they reveal the poor mechanical aspect of an aircraft. are placed

at the top. Third, since we have split ‘date’, ‘day’ and ‘year’ the season and weather at an instance also play a significant role in the safety prediction.

The waterfall plot in Figure 3 focuses on explaining a single prediction (local) made by the model. It starts from the expected value of the model output (usually the average prediction) and shows how each feature’s contribution (positive or negative) moves the prediction from the expected value to the actual model output for that specific instance. Each row in the waterfall plot represents a feature. The SHAP value of a feature reflects how much that feature’s evidence influences the model’s output. The plot uses color-coding: red for positive and blue for negative contributions that helps to understand which features are driving model’s decision for specific prediction.

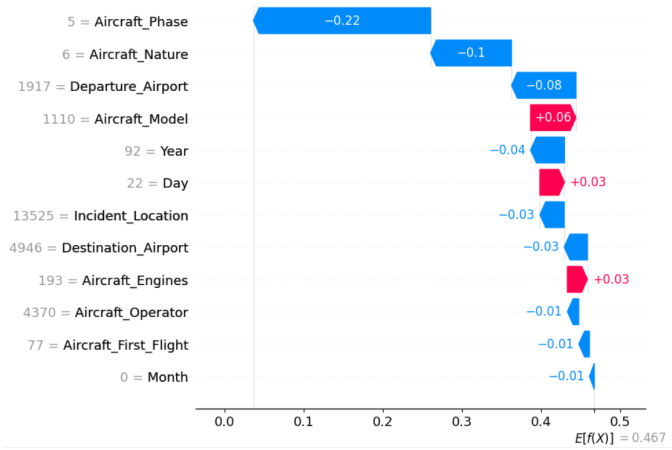

Fig. 3. SHAP Waterfall Plot

First important observation in Figure 3 is that for a given prediction, positive SHAP values such as ‘Aircraft_Model’, ‘Day’ and ‘Aircraft_Engines’. Second, negative values such as ‘Aircraft_Phase’, ‘Aircraft_Nature’ and ‘Departure_Airport’ contribute negatively to reach a prediction. However, the global explanation in Figure 2 ranks these features contradictory as compared to Figure 3. This contradiction can be explained that each prediction, SHAP determines how much each feature contributes to that specific prediction, known as local SHAP values. To derive global explanations, SHAP takes the average of the absolute local SHAP values for each feature across all data instances. So as a result this average indicates the overall significance of each feature in the model’s predictions globally.

Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) provides local, interpretable explanations for individual predictions made by any machine learning model [26]. It treats any supervised learning model as a black box, can be applied to various types of models. LIME focuses on explaining predictions within the vicinity

of a specific data point. It samples data points around the instance being explained, creates a simpler surrogate model, and approximates the original model's behavior. For LIME explanations the predictions made by RF model is utilized.

Figure 4 displays a bar chart depicting prediction probabilities for six different classes; 0 indicates Accident, 1 represents Criminal occurrence, 2 shows Hijacking, 3 indicates Incident, 4 represents occurrence unknown and 5 shows other occurrence as described in section 3.7. These probabilities represent the likelihood of different outcomes. The highest probability (0.75) of class 0 corresponds that the model classifies the given instance as 'Accident' whereas the other probabilities are 0.13, 0.10, and the lowest (0.01) for other classes. In Figure 4 and 5 the horizontal bars use different colors to represent values of prediction probabilities, making it easy to visually compare the values.

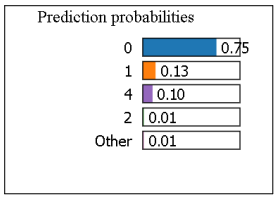

Fig. 4. LIME Prediction Probabilities

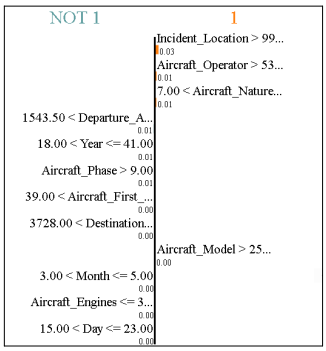

Fig. 5. LIME Feature Impacts

Fig. 6. LIME Feature Importance Scores

In Figure 5, there's a list of features with corresponding weights. 'Incident_Location' has a strong positive weight (99%), meaning it significantly influences the prediction. Conversely, 'Departure_Aircraft' has a negative weight (-25%), reducing the likelihood of this outcome. The Figure 6 lists several features on the right side. These features are likely input variables used by ML model such as "Incident Location," "Aircraft Operator," "Departure Airport," "Year," "Aircraft Phase," "Aircraft First Flight," "Aircraft Model," "Month," and "Day". Each feature has a corresponding value next to it. This value represents the importance or impact of that feature on the model's prediction. For instance, a high value 4881 of 'Destination_Airport' indicates that changing that feature significantly affects the model's output. In Figure 6, the orange color indicates features that positively influence reaching a prediction label, while the gray color represents features that negatively impact achieving a prediction.

This is particularly significant when comparing these results with findings of prior research efforts in interpreting predictions for aircraft safety domain those have just focused on two aspects one for aircraft failure diagnosis [15] and other for runway surface contamination [16]. The study's findings focus on aircraft incident and accident predictions which reveal exceptional predictive accuracy for RF and XGBoost models, coupled with a high level of transparency and trustworthiness.

Additionally, XAI techniques such as LIME and SHAP effectively offer clear local and global explanations for the model's predictions.

5 Conclusion

The research successfully accomplished its objectives, which centered on predicting aircraft accidents and incidents for safety-critical systems. By leveraging historical data, the study forecasted future accidents. The AI models RF and XGBoost both achieved remarkable accuracy in these predictions. Moreover, the research demonstrated the effective application of XAI techniques, specifically LIME and SHAP, to provide comprehensive explanations for both local and global predictions in order to enhance trustworthiness.

The study exhibits several limitations. Firstly, the results heavily rely on a publicly available dataset, which may introduce biases or inaccuracies. Additionally, the dataset suffers from imbalanced class distribution, missing values, and inconsistencies in data format. Secondly, the predictive features used are limited; incorporating environmental factors such as temperature, air pressure, and humidity could enhance accuracy. Thirdly, the research explored only two ML models, RF and XGBoost, warranting further investigation into more complex techniques like deep learning. Also currently we have provided both local and global explanations for predictions of XGBoost using SHAP, and interpretations of local predictions for RF using LIME. The application of both XAI frameworks to RF and XGBoost, could provide valuable insights for explaining aircraft incident and accident predictions to allow a more interesting direct comparison on the performance and level of explainability of the two frameworks. Lastly, to improve accuracy and interpretability, additional XAI methods should be considered.

While the airline industry frequently displays hesitance toward adopting novel technologies since safety is the top priority and new technologies are held to extremely high standards before they can be adopted. AI holds significant promise for enhancing safety and ensuring trustworthiness. In forthcoming studies, emphasizing feature engineering and enhancing model accuracy will be pivotal. Furthermore, ensuring the interpretability of predictions is essential for their effective adoption within the airline industry particularly in decision support system.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT), with project No. NRF-2022R1A2C1092077, NRF-RS-2023-00302083 (as part of the EC-funded Swarmchestrate Project), and BK21 FOUR (Fostering Outstanding Universities for Research) funded by Korea's Ministry of Education (MOE) and the NRF of Korea.

References

1. Myo, T., Ahmed, M. R., Al Hadidi, H., and Al Baroomi, B. Trends and Challenges of Machine Learning-Based Predictive Maintenance in Aviation Industry. In: *International*

Conference on Aeronautical Sciences, Engineering and Technology, pp. 362-368. Springer Nature, Singapore (2023).
2. Zhang, X., Srinivasan, P., and Mahadevan, S.: Sequential deep learning from NTSB reports for aviation safety prognosis. *Safety science*, pp. *142*, (2021).
3. Zeng, H., Guo, J., Zhang, H., Ren, B., and Wu, J.: Research on aviation safety prediction based on variable selection and LSTM. *Sensors*, *23*(1), pp. 41, (2022).
4. Dangut, M. D., Jennions, I. K., King, S., and Skaf, Z.: A rare failure detection model for aircraft predictive maintenance using a deep hybrid learning approach. *Neural Computing and Applications 35*(4), pp. 2991-3009, (2023).
5. Lee, H., Madar, S., Sairam, S., Puranik, T. G., Payan, A. P., Kirby, M., ... and Mavris, D. N.: Critical parameter identification for safety events in commercial aviation using machine learning. *Aerospace*, *7*(6), pp. 73, (2020).
6. Madeira, T., Melício, R., Valério, D., and Santos, L.: Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace 8*(2), pp. 47, (2021).
7. Silagyi II, D. V., and Liu, D.: Prediction of severity of aviation landing accidents using support vector machine models. *Accident Analysis & Prevention*, pp. *187*, (2023).
8. Taheri Gorji, H., Wilson, N., VanBree, J., Hoffmann, B., Petros, T., and Tavakolian, K.: Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Scientific Reports 13*(1), (2023).
9. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... and Hussain, A.: Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* 16(1), pp. 45-74, (2024).
10. Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B.: A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access,* (2023).
11. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... and Herrera, F.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* 99, (2023).
12. Shukla, B., Fan, I. S., and Jennions, I. Opportunities for explainable artificial intelligence in aerospace predictive maintenance. *PHM Society European Conference,* V.5,pp.11, 2020
13. Saraf, A. P., Chan, K., Popish, M., Browder, J., and Schade, J. Explainable artificial intelligence for aviation safety applications. In: *AIAA Aviation 2020 Forum*, (2020).
14. Sutthithatip, S., Perinpanayagam, S., & Aslam, S.: (Explainable) Artificial Intelligence in Aerospace Safety-Critical Systems. In:*IEEE Aerospace Conference 2022,* pp.1-12, (2022).
15. Zeldam, S. G. Automated failure diagnosis in aviation maintenance using explainable artificial intelligence (XAI) (Master's thesis, University of Twente), (2018).
16. Midtfjord, A. D., De Bin, R., & Huseby, A. B.: A decision support system for safer airplane landings: Predicting runway conditions using XGBoost and explainable AI. *Cold Regions Science and Technology*, pp. *199*, (2022).
17. Hernandez, C. S., Ayo, S., and Panagiotakopoulos, D.: An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. In: *IEEE/AIAA 40th Digital Avionics Systems Conference (DASC),* pp. 1-10, (2021).
18. Weber, P., Carl, K. V., and Hinz, O. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, *74*(2), pp.867-907, (2024).
19. Degas, A., Islam, M. R., Hurter, C., Barua, S., Rahman, H., Poudel, M., ... and Arico, P.: A survey on artificial intelligence (ai) and explainable ai in air traffic management: Current trends and development with future research trajectory. *Applied Sciences*, *12*(3), (2022).

20. Messalas, A., Kanellopoulos, Y., and Makris, C. Model-agnostic interpretability with shapley values. In: *10th International Conference on Information, Intelligence, Systems and Applications,* pp.1-7. IEEE, (2019).
21. Liu, Y., Wang, Y., and Zhang, J. New machine learning algorithm: Random forest. In *: 3rd International Conference on Information Computing and Applicatizons,* pp. 246-252. Springer Berlin Heidelberg, *Chengde, China,* (2012).
22. Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In: *22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. (2016).
23. ICAO safety Report, http://www.icao.int/safety/Documents/ICAO_SR_2023_20230823.pdf, last accessed 2024/6/25
24. Aircraft Accidents, Failures & Hijacks Dataset, https://www.kaggle.com/datasets/deepcontractor/aircraft-accidents-failures-hijacks-dataset, last accessed 2024/6/25
25. Gössling, S., and Humpe, A. The global scale, distribution and growth of aviation: Implications for climate change. *Global Environmental Change*, *65*, 102194, (2020).
26. Zafar, M. R., & Khan, N.(2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning & Knowledge Extraction*, 3(3),525-541.
27. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 11(5), e1424.