

Article

Wind Shear and Aircraft Aborted Landings: A Deep Learning Perspective for Prediction and Analysis

Afaq Khattak ^{1,*}, Jianping Zhang ², Pak-Wai Chan ³, Feng Chen ^{1,*}, Arshad Hussain ⁴ and Hamad Almujiabah ⁵

¹ The Key Laboratory of Infrastructure Durability and Operation Safety in Airfield of CAAC, Tongji University, 4800 Cao'an Road, Jiading, Shanghai 201804, China

² Civil Unmanned Aircraft Traffic Management Key Laboratory of Sichuan Province, The Second Research Institute of Civil Aviation Administration of China, Chengdu 610041, China; zhangjp@caacsri.com

³ The Hong Kong Observatory, 134A Nathan Road, Kowloon, Hong Kong, China

⁴ NUST Institute of Civil Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan; drarshad@nice.nust.edu.pk

⁵ Department of Civil Engineering, College of Engineering, Taif University, P.O. Box 11099, Taif City 21974, Saudi Arabia; hmujiabah@tu.edu.sa

* Correspondence: khattak@tongji.edu.cn (A.K.); fengchen@tongji.edu.cn (F.C.)

Abstract: In civil aviation, severe weather conditions such as strong wind shear, crosswinds, and thunderstorms near airport runways often compel pilots to abort landings to ensure flight safety. While aborted landings due to wind shear are not common, they occur under specific environmental and situational circumstances. This research aims to accurately predict aircraft aborted landings using three advanced deep learning techniques: the conventional deep neural network (DNN), the deep and cross network (DCN), and the wide and deep network (WDN). These models are supplemented by various data augmentation methods, including the Synthetic Minority Over-Sampling Technique (SMOTE), KMeans-SMOTE, and Borderline-SMOTE, to correct the imbalance in pilot report data. Bayesian optimization was utilized to fine-tune the models for optimal predictive accuracy. The effectiveness of these models was assessed through metrics including sensitivity, precision, F1-score, and the Matthew Correlation Coefficient. The Shapley Additive Explanations (SHAP) algorithm was then applied to the most effective models to interpret their results and identify key factors, revealing that the intensity of wind shear, specific runways like 07R, and the vertical distance of wind shear from the runway (within 700 feet above runway level) were significant factors. The results of this research provide valuable insights to civil aviation experts, potentially revolutionizing safety protocols for managing aborted landings under adverse weather conditions, thereby improving overall airport efficiency and safety.

Keywords: civil aviation safety; aborted landings; deep learning; SHAP



Citation: Khattak, A.; Zhang, J.; Chan, P.-W.; Chen, F.; Hussain, A.; Almujiabah, H. Wind Shear and Aircraft Aborted Landings: A Deep Learning Perspective for Prediction and Analysis. *Atmosphere* **2024**, *15*, 545. <https://doi.org/10.3390/atmos15050545>

Academic Editor: Stephan Havemann

Received: 16 March 2024

Revised: 21 April 2024

Accepted: 26 April 2024

Published: 29 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An aborted landing, often termed as a “go-around”, is a situation typically encountered during the final approach phase of a flight. This occurs when landing safely is not possible due to various reasons, including extreme weather, an occupied runway, poor visibility, or an unstable approach [1]. In such cases, it is crucial to halt the landing attempt, ascend again, and either prepare for a different landing approach or head to an alternate airport, as depicted in Figure 1. During this phase, the aircrew follows specific procedures and rigorously checks the relevant checklists. Aborted landings can negatively impact airport efficiency, airline timeliness, and increase the workload for air traffic controllers [2,3].

The accurate evaluation of aborted landings is important for enhancing safety measures and developing effective strategies to reduce their frequency. Aborted landings, especially those occurring during wind shear, are rare and complex events influenced by a multitude of interconnected factors, making it challenging to fully understand every aspect

of these incidents. By understanding the detailed interactions between these factors, we can develop sophisticated safety protocols that not only prevent aborted landings but also minimize their impact on airport throughput and scheduling.

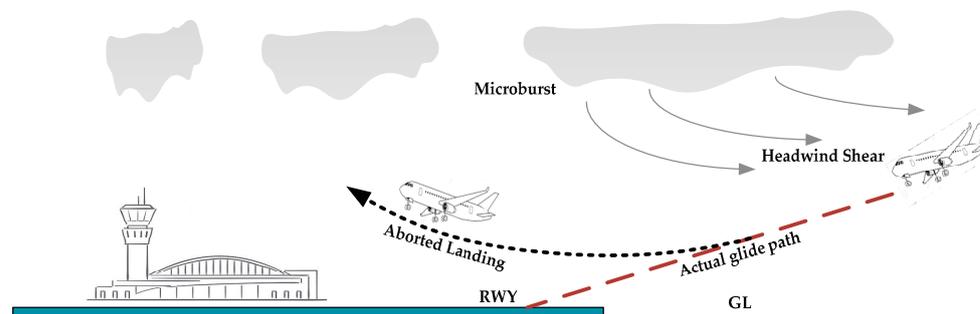


Figure 1. Aborted landing due to wind shear near airport runway.

2. Related Work

Investigating the diverse factors and standards that affect aborted landings has been a significant area of ongoing research. Studies have explored various elements, such as weather and environmental conditions, the psychological states of pilots and air traffic controllers, and unstable aircraft landings. One study focusing on environmental influences on aircraft aborted landings identified key factors like runway visibility, wind speed near the runway, and localizer deviation as significant [4]. Another research pointed out that atmospheric pressure, wind speed, and visibility play major roles in causing aborted landings [5]. Further research noted that severe thunderstorms and wind speeds over 29 mph near runways are critical factors, although visibility levels showed less significance [6]. Additionally, poor weather conditions, especially convective storms affecting the runway's glide path, have been linked to the occurrence of aborted landings [7].

Unstable or non-stabilized approaches are significant contributors to commercial aviation accidents during landing [8–10]. A stable approach requires adherence to specific criteria related to configuration and speed. Approaches that do not meet these criteria are deemed unstable, substantially increasing the risk of incidents or aborted landings. Research analyzing aborted landings found that factors like flight separation, approach stability, departing aircraft, and the aircraft's altitude above the runway significantly affect the probability of an aborted landing [11]. Various studies on aircraft aborted landings have mainly focused on the efficiency and attitudes of pilots and air traffic controllers. It was uncovered that aborted landings could be linked to a temporary impairment in rational decision-making due to negative emotional impacts [12]. An investigation into pilot performance and visual scanning behavior during aborted landings revealed that most pilots, about two-thirds, committed errors, including significant deviations in flight path, during an aborted landing [13]. According to [3], a lack of situational awareness among air traffic controllers is a key factor in aborted landings. Another study indicated that the decisions regarding aborted landings are heavily influenced by the experience and age of air traffic controllers [14].

This research represents a novel attempt to apply deep learning models to analyze aircraft aborted landings. It employs three advanced deep learning techniques, including the conventional deep neural network (DNN) [15], the deep and cross network (DCN) [16], and the wide and deep network (WDN) [17] to predict aircraft aborted landings due to wind shear, using pilot reports (PIREPs) as the primary data source. Data augmentation techniques like Synthetic Minority Over-Sampling Technique (SMOTE) [18], KMeans-SMOTE [19], and Borderline-SMOTE [20] were used to address the imbalance in PIREP data. Bayesian optimization [21] was implemented for selecting and refining the hyperparameters of these models and their learning processes. Since deep learning models are not inherently interpretable, the post hoc interpretation strategy of the Shapley Additive Explanations (SHAP) algorithm [22] was utilized to assess the impact of various factors. The combination

of deep learning algorithms with SHAP can aid in developing targeted interventions to reduce aborted landings during wind shear events.

The paper is structured as follows: Section 3 provides a detailed overview of the study location, PIREP data, and the proposed deep learning architectures. Section 4 discusses the performance evaluation and interpretation of the optimal deep learning models via SHAP analysis. Section 5 elucidates the findings of the study.

3. Materials and Methods

3.1. Study Location

Hong Kong International Airport (HKIA), a major aviation center in the Hong Kong region, is located on the man-made island of Lantau, positioned along the subtropical coastline of mainland China (Figure 2). The typical weather patterns in Hong Kong are characterized by tropical cyclones and the southwest monsoon, which frequently lead to severe thunderstorms and intense rainfall in the area [23,24]. This airport is known to be more vulnerable to wind shear compared to many other airports globally [25,26].



Figure 2. Complex Lantau island to the south of HKIA.

3.2. Data Description

The incidence of aborted landings due to wind shear is relatively rare, leading to a skewed dataset in pilot reports (PIREPs) that predominantly reflects successful landings. PIREPs are formal accounts given by pilots about weather phenomena experienced during a flight [27], aimed at providing warnings to other pilots and helping air traffic control be aware of potential dangers. This ensures pilot safety by enabling them to steer clear of such hazards [28]. These reports include details like aircraft type, flight number, time, temperature, precipitation, and current weather conditions, such as intense thunderstorms and wind shear. As mentioned earlier, HKIA is particularly prone to wind shear, ranking it among the world's most vulnerable airports in this respect. Thus, PIREPs from HKIA predominantly contain data on wind shear events, including information on altitude, intensity, and the position of the wind shear relative to the runway threshold [29]. They also describe factors leading to wind shear, like gust fronts or sea breezes, and incidents of aborted landings.

In aviation, the classification of wind shear encounters in relation to the runway is typically labeled as RWY, MD, or MF. Figure 3 depicts the runway as a gray rectangle marked RWY. The rectangles to the right of the runway indicate distance in miles to the final approach, with each 1MF rectangle representing one nautical mile to the final approach. The green rectangles on the left side show the distance from the end of the runway used for departures. For instance, a two-mile final (2MF) is a spatial measurement of two nautical miles from the runway's edge at the arrival threshold, shown as a blue

circle in the diagram. Figure 4 presents two examples of urgent and non-urgent PIREPs. Table 1 outlines various factors derived from PIREPs that could influence the occurrence of aborted landings during wind shear conditions. The following section will discuss the study location and the proposed deep learning architectures in more detail.

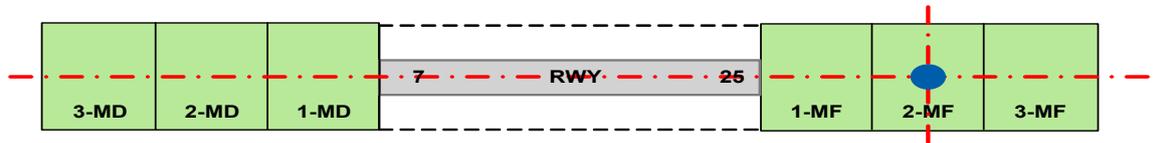


Figure 3. Designation of wind shear horizontal encounter locations from runway.

PIREP C172	Urgent PIREP PA28
Obs Time: 2022-01-19T02:52:00Z	Obs Time: 2021-02-13T21:05:00Z
Flight level type: DURD	Flight level: 090
Cloud Cover: BKN	Flight level type: OTHER
Cloud Top: 18	Turb intensity: SEV
Cloud Base: 12	Urgent PIREP: IPL UUA /OV IPL260030/TM
PIREP: SMO UA /OV SMO030004/TM 0252/FLDURGD/TP	2105/FL090/TP PA28/TB SEV/RM DOWN DRAFT LOSS
C172/SK BKN012-TOP018	2000 FT
(a)	(b)

Figure 4. Examples of onboard PIREPs: (a) non-urgent PIREP; (b) urgent PIREP.

Table 1. Description and coding of different factors extracted from HKIA-based PIREPs.

Factors from PIREPs	Data Type	Description and Coding
Type of aircraft	Discrete	0: narrow-body aircraft, 1: wide-body aircraft
Assigned approach runway	Discrete	0: 07R, 1: 07C, 2: 07L; 3:25R, 4: 25C, 5: 25L
Intensity of wind shear	Continuous	Negative (−): tail wind, positive (+): headwind
Wind shear horizontal distance from the runway	Discrete	0: RWY, 1:1MF, 2; 2MF, 3: 3MF
Wind shear vertical distance from the runway	Continuous	-
Cause of the wind shear	Discrete	0: gust front; 1: sea breeze
Precipitation	Discrete	0: No, 1: yes
Season of the year	Discrete	0: winter, 1: spring, 2: summer; 3: autumn

3.3. Deep Neural Network Architectures

This section provides an overview of the different architectures employed in this study for the prediction of aircraft aborted landing. Additionally, it emphasizes the fundamental attributes of the aforementioned architectures.

3.3.1. Conventional Deep Neural Network (DNN)

The conventional DNN refers to an artificial neural network architecture that incorporates numerous hidden layers positioned between the input and output layers. In essence, a conventional DNN can be described as an advanced computational model that encompasses multiple layers, facilitating the processing of multiple inputs and generating multiple outputs in a feed-forward manner, as shown in Figure 5 [30]. The layers referred to as hidden layers are composed of nodes that receive input from the preceding layer. These nodes execute a mathematical operation, commonly referred to as an activation function, on the input and produce an output that serves as the input for the subsequent layer. As illustrated in Figure 5, the input layer consists of three nodes, also referred to as neurons, while the output layer comprises two nodes. The hidden layers consist of intermediary layers, each containing neurons. One of the main perks of a DNN lies in its ability to accurately approximate a nonlinear function to any desired level of precision. This feat is made possible through a meticulous choice of appropriate activation functions. Notably, the Rectified Linear Unit (RLU), Hyperbolic Tangent, and sigmoid activation functions are

widely employed in different fields. The conventional DNN also has the capability to be employed for tasks that involve regression and classification [31].

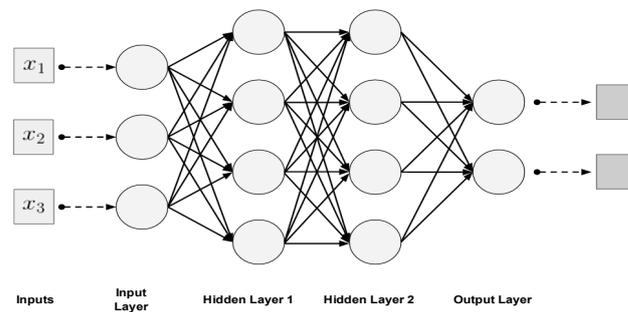


Figure 5. Example of DNN architecture.

3.3.2. Deep and Cross Network (DCN)

The DCN model architecture is an improved version of the conventional DNN that commences with an embedding and stacking layer, followed by a cross network and a deep network operating in parallel. Subsequently, a final layer is implemented to combine the outcomes from both of the preceding networks. A comprehensive DCN model is illustrated in Figure 6. The DCN model retains the advantages of a DNN model while also introducing a novel cross network that exhibits improved efficiency in learning bounded-degree feature interactions. Specifically, the DCN employs feature crossing at every layer, eliminating the need for manual feature engineering and introducing minimal additional complexity to the DNN model. For the details of the working principle of the DCN, refer to the paper [16].

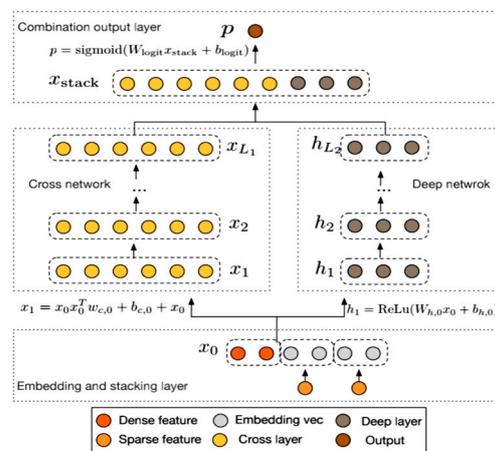


Figure 6. Example of DCN architecture (adapted from [16]).

3.3.3. Wide and Deep Network (WDN)

The WDN approach involves the simultaneous training of wide linear models (WLM) and a DNN in order to leverage the advantages of both memorization and generalization in different fields of application. The wide component can be described as a generalized linear model. The deep component refers to a type of neural network known as a feed-forward neural network. The prediction is made by combining the output log odds of the deep and wide components using a weighted sum. The input is subsequently passed through a logistic loss function for the purpose of joint training. This process involves back-propagating the gradients from the outcome to both the wide and deep components of the model concurrently, utilizing an optimization strategy. The WDN architecture is shown in Figure 7. For details regarding the WDN model, refer to [17].

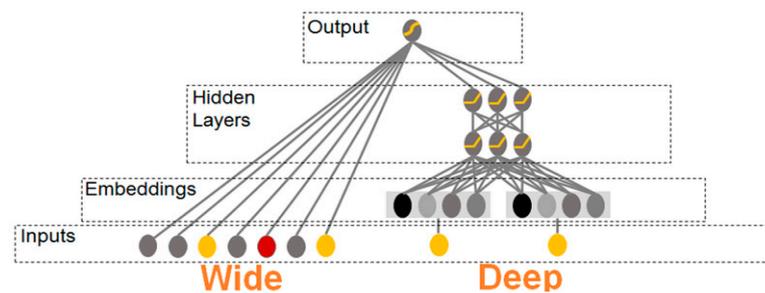


Figure 7. Architecture of WDN (adapted from [17]).

3.4. Deep Learning Model Interpretation

Certain machine learning algorithms, such as the random forest algorithm and extreme gradient boosting algorithm, possess an intrinsic capability to provide insights on the importance of factors. Nevertheless, deep learning models are capable of estimating other models, although they do not inherently possess this characteristic. Therefore, in order to develop interpretable deep learning models, SHAP analysis is coupled with an optimal deep learning model. SHAP is a mathematical approach that possesses the capability to elucidate the predictions generated by both machine learning [22,32–34] and deep learning [30,35,36] models. This approach has its roots in the principles of game theory and can be employed to explain predictions by quantifying the individual contributions of each factor to the prediction. It determines the most significant factor and quantifies its impact on the model's prediction.

3.5. Performance Indicators

The metrics for a binary classification problem are characterized by true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Positive and negative are standard terms used to describe the classes in a binary classification problem. Accordingly, a true positive occurs when both the actual and estimated classes are positive. When the actual class is negative but the estimated class is positive, this is a false positive. Based on these, different metrics can be evaluated, including sensitivity; also being known as recall or the true-positive rate, sensitivity is a quantitative metric that assesses the ability of a model to accurately identify positive instances within a given dataset. Precision is a metric that quantifies the ratio of accurately predicted positive instances. The F1-score is a metric that incorporates both precision and sensitivity. It accounts for imbalanced class distribution in the data and is commonly regarded as one of the most reliable indicators of a model's effectiveness. The Matthews Correlation Coefficient (MCC) is another metric utilized to assess the efficacy of binary classification models. It considers both true and false positives and negatives, making it widely recognized as a balanced measure that remains applicable even when the classes exhibit significant disparities in size. It yields a numerical value ranging from -1 to $+1$. A coefficient with a value of $+1$ signifies an ideal prediction, while a value of 0 indicates no improvement and -1 denotes complete discordance between the prediction and the observed outcome. The mathematical expressions of these indicators are shown by Equations (1)–(4).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{F1-Score} = \frac{TP}{TP + \frac{1}{2}(TN + FN)} \quad (3)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + TN) \times (TP + FN) + (TN + FP) \times (TN + FN)}} \quad (4)$$

4. Results and Discussion

This study focused on examining the instances of aborted landings due to wind shear at Hong Kong International Airport (HKIA). Utilizing PIREP data from 1 January 2015 to 23 July 2023 from HKIA, a thorough analysis revealed a total of 3585 wind shear events affecting both arriving and departing flights. Of these, our research specifically examined the 2024 cases reported by flights arriving at HKIA. In these 2024 wind shear incidents, there were 476 aborted landings and 1552 successful landings. Standard protocols for data preparation and pre-processing were followed in this study [37]. The data were randomly divided, with 70% allocated to training and the remaining 30% for testing, with all deep learning models being evaluated on this split. This division was carried out using a randomly chosen seed.

A binary classification problem was established, designating successful landings during wind as the majority class and aborted landings as the minority class. In order to tackle the issue of imbalanced data, three data augmentation techniques including SMOTE, KMeans-SMOTE, and Borderline-SMOTE were used to balance the training data. Post-treatment, the dataset consisted of 1093 successful and 1093 aborted landing instances, as depicted in Figure 8. The research was conducted in a Jupyter Notebook environment, using custom-written Python code that leveraged libraries such as Pandas, Numpy, sklearn, TensorFlow, and DeepTables.

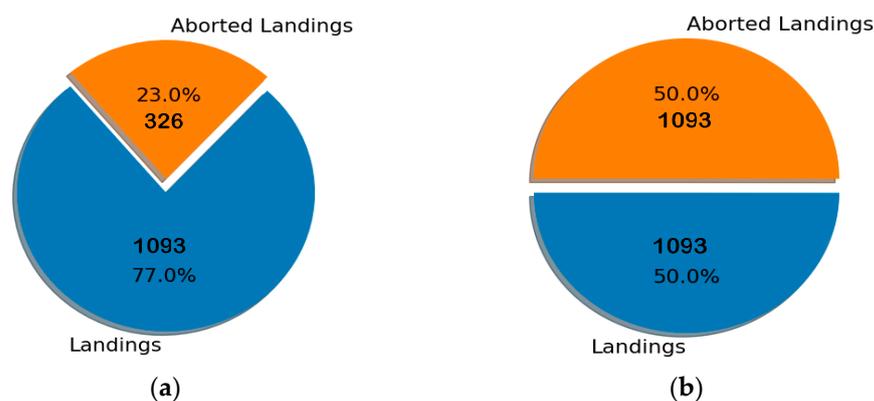


Figure 8. Instances of training data: (a) original data instances; (b) instances resampled via SMOTE, KMeans-SMOTE, and Borderline-SMOTE.

Following the initial setup, the hyperparameters for each deep learning architecture were identified. Bayesian optimization was then employed to refine these hyperparameters, with the primary aim of enhancing the F1-score. This method is preferred for hyperparameter tuning due to its efficacy in producing more accurate hyperparameter estimates. It has been shown to yield better values for hyperparameters with considerably fewer iterations compared to grid and random search methods, as indicated in previous research [38,39].

For the conventional DNN and WDN models, key hyperparameters included the number of hidden layers, the neurons in each layer, the type of activation function, the training algorithm, and the learning rate. The activation function and optimizer, being non-numeric, were conventionally transformed into numerical proxies. Additionally, a uniform dropout rate of 0.1 was applied across all hidden layers in the networks. In the case of the DCN model, an extra hyperparameter, the number of cross layers, was also considered. Tables A1–A3 in Appendix A display the ranges and optimal values for these hyperparameters as determined by Bayesian optimization under various data treatment strategies.

4.1. Performance Analysis and Comparison

Following the determination of the optimal hyperparameter combination, the deep learning models were subsequently retrained on the established training dataset. Monitoring of

the validation loss was performed in order to mitigate the risk of over-fitting, and the strategy of early stopping was employed. A confusion matrix was developed for the proposed deep learning models using both the original untreated data and resampled data. Initially, it was observed that deep learning models exhibited a poor performance when applied to imbalanced datasets, as shown in Figure 9. Among the 150 aborted landing instances in the testing dataset, the DNN correctly classified 8 instances, the DCN correctly classified 30 instances, and the WDN correctly classified 7 instances of aborted landings. The performance metrics of the proposed deep learning models, when applied to untreated data, are presented in Table 2. The experimental results indicate that the utilization of untreated PIREP data in DNN, DCN, and WDN models led to a decrease in the F1-score, with values of 9.88%, 29.27%, and 8.33%, respectively. Additionally, the MCC values were also lower, measuring 0.138, 0.218, and 0.057 for DNN, DCN, and WDN, respectively.

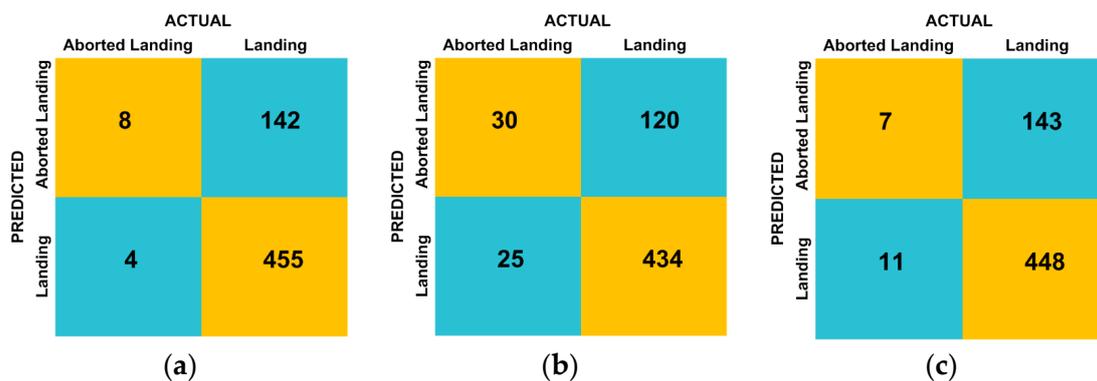


Figure 9. Confusion matrix of proposed deep learning model using untreated data: (a) the DNN model; (b) the DCN model; (c) the WDN model.

Table 2. Performance assessment of proposed deep learning models using untreated data.

Deep Learning Models	Performance Measures				Run Time (Seconds)
	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC	
DNN	66.67	76.21	9.88	0.138	37
DCN	54.55	78.34	29.27	0.218	46
WDN	38.89	75.80	8.33	0.057	47

A confusion matrix, as shown in Figure 10, was subsequently constructed for the conventional DNN model using data that had undergone pre-processing techniques including SMOTE, KMeans-SMOTE, and Borderline-SMOTE. The utilization of data processing techniques led to significant improvements in the precise categorization of aborted landings. In the context of the DNN + SMOTE scenario, it was observed that 116 out of 150 instances accurately classified aborted landings. In the context of DNN + KMeans-SMOTE, the accurate classification was attained in 118 instances out of a total of 150. DNN + Borderline-SMOTE resulted in the efficient classification of 128 instances of aborted landings out of 150. The findings shown in Table 3 indicate that the DNN + SMOTE, DNN + KMeans-SMOTE, and DNN + Borderline-SMOTE models led to a higher F1-score, with values of 69.05%, 71.73%, and 77.34%, respectively. The MCC values were also higher compared to untreated data, measuring 0.583, 0.618, and 0.701, respectively.

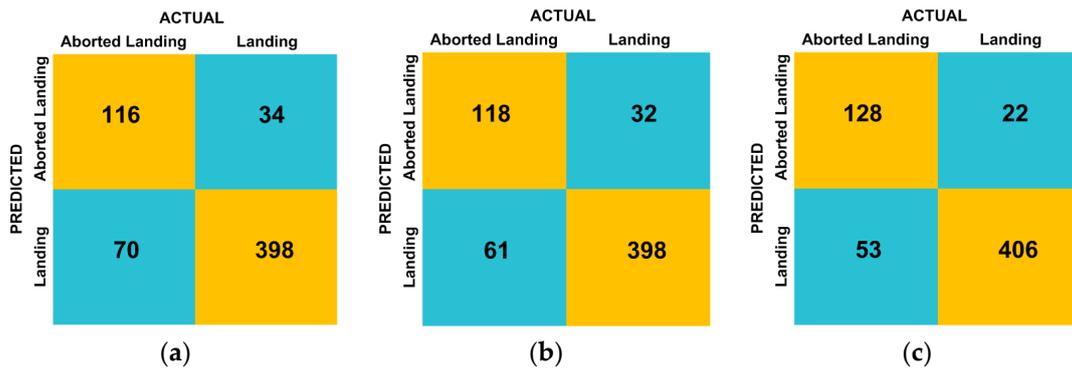


Figure 10. Confusion matrix and ROC plot for DNN model (a) with SMOTE-treated data; (b) with KMeans-SMOTE-treated data; (c) and with Borderline-SMOTE-treated data.

Table 3. Performance assessment of DNN model using different data augmentation strategies.

Data Augmentation Techniques	Performance Measures				
	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC	Run Time (Seconds)
SMOTE	62.37	91.96	69.05	0.582	38
KMeans-SMOTE	65.92	92.56	71.73	0.618	40
Borderline-SMOTE	70.15	94.56	77.34	0.701	35

The development of a confusion matrix for the DCN model also involved the utilization of SMOTE, KMeans-SMOTE, and Borderline-SMOTE, as shown in Figure 11. In the case of DCN, data treatment techniques also resulted in notable advancements in the precise categorization of aborted landings. In the specific scenario involving DCN + SMOTE, it was noted that 127 instances out of a total of 150 were successfully classified as aborted landings. In the context of the DCN + KMeans-SMOTE approach, a total of 150 instances were evaluated, resulting in accurate classification in 134 instances. The implementation of the DCN + Borderline-SMOTE technique yielded a successful classification rate of 142 out of 150 instances of aborted landings. The results presented in Table 4 demonstrate that the DCN + SMOTE, DCN + KMeans-SMOTE, and DCN + Borderline-SMOTE models yielded superior F1-scores, achieving 73.41%, 76.15%, and 82.56%, respectively. The MCC values also exhibited higher magnitudes, with respective measurements of 0.642, 0.686, and 0.773.

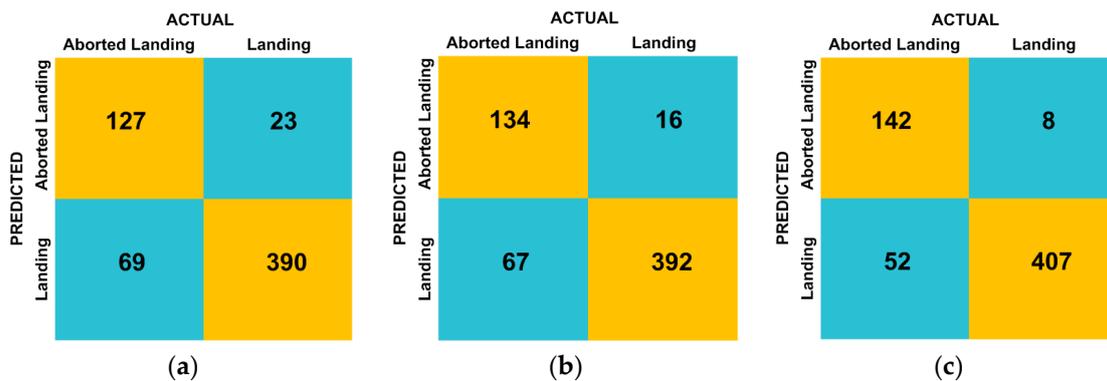


Figure 11. Confusion matrix and ROC plot for DCN model (a) with SMOTE-treated data; (b) with KMeans-SMOTE-treated data; (c) and with Borderline-SMOTE-treated data.

Table 4. Performance assessment of DCN model using different data augmentation strategies.

Data Augmentation Techniques	Performance Measures				
	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC	Run Time (Seconds)
SMOTE	64.62	94.95	73.41	0.642	49
KMeans-SMOTE	66.85	96.67	76.15	0.686	46
Borderline-SMOTE	73.59	98.62	82.56	0.773	51

A confusion matrix was also developed for the WDN models, as shown in Figure 12, and the findings are displayed in Table 5, which indicate that the WDN + SMOTE, WDN + KMeans-SMOTE, and WDN + Borderline-SMOTE models exhibited F1-scores of 68.48%, 74.18%, and 78.75%, respectively, and the MCC metric was measured at 0.576, 0.657, and 0.719, respectively.

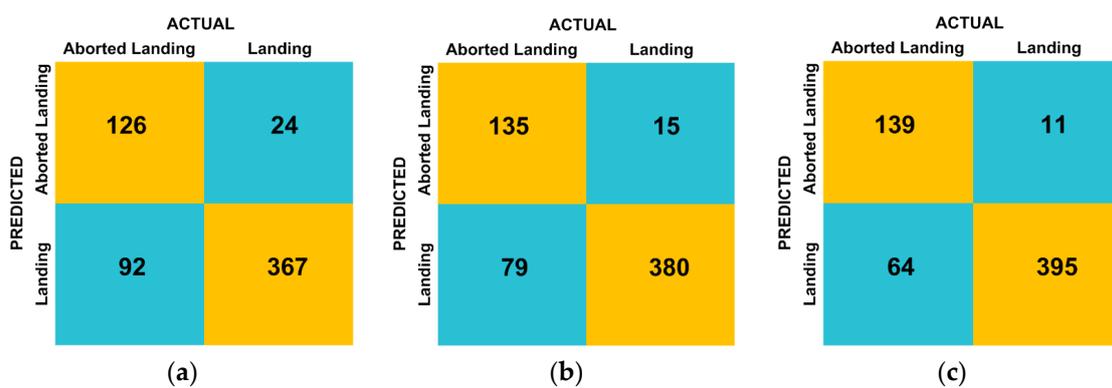


Figure 12. Confusion matrix and ROC plot for WDN model (a) with SMOTE-treated data; (b) with KMeans-SMOTE-treated data; (c) and with Borderline-SMOTE-treated data.

Table 5. Performance assessment of WDN model using different data augmentation strategies.

Data Augmentation Techniques	Performance Measures				
	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC	Run Time (Seconds)
SMOTE	57.80	93.86	68.48	0.576	40
KMeans-SMOTE	63.44	96.25	74.18	0.657	52
Borderline-SMOTE	68.82	97.73	78.75	0.719	54

Based on above findings, it can be concluded that the DCN + Borderline-SMOTE and WDN + Borderline-SMOTE techniques yielded superior F1-score values of 82.56% and 78.75%, respectively. These techniques also demonstrated higher MCC scores of 0.773 and 0.719, respectively. In addition, the optimal deep learning models were compared to binary logistic regression (BLR) using both untreated and treated data. The findings indicate that both the F1-score and MCC obtained from BLR, in the case of both untreated and treated data, were much lower than those obtained for optimal deep learning models, as shown in Table 6. The closeness of the results for the optimal deep learning models necessitated the utilization of SHAP analysis for the interpretation of these models, as detailed in the following section.

Table 6. Performance comparison of optimal deep learning models and BLR.

Deep Learning Models + Data Augmentation Techniques	Performance Measures			
	Sensitivity (%)	Precision (%)	F1-Score (%)	MCC
DCN + Borderline-SMOTE	73.59	98.62	82.56	0.773
WDN + Borderline-SMOTE	68.82	97.73	78.75	0.719
BLR + Untreated data	50.03	76.67	39.73	0.060
BLR + SMOTE	43.36	82.41	45.56	0.266
BLR + KMeans-SMOTE	51.24	84.61	52.43	0.368
BLR + Borderline-SMOTE	57.42	86.56	58.36	0.445

4.2. Interpretation of Optimal Deep Learning Models

The development of an accurate deep learning model for aborted landings is of great importance, as an optimized deep learning model has the potential to provide a deeper comprehension of the relationship between aborted landings and the various factors that contribute to them. Following predictive analysis by deep learning models, SHAP bee swarm plots [40] were generated for both optimal DCN + Borderline-SMOTE and WDN + Borderline-SMOTE in order to evaluate the significance and contribution of various factors. As depicted in Figure 13, the input factors are arranged on the vertical axis in descending order of ascending influence, commencing with the factor exerting the greatest influence. The plot illustrates the contribution of these factors, with the SHAP value represented on the horizontal axis and a color scale ranging from blue (indicating low significance) to red (indicating high significance).

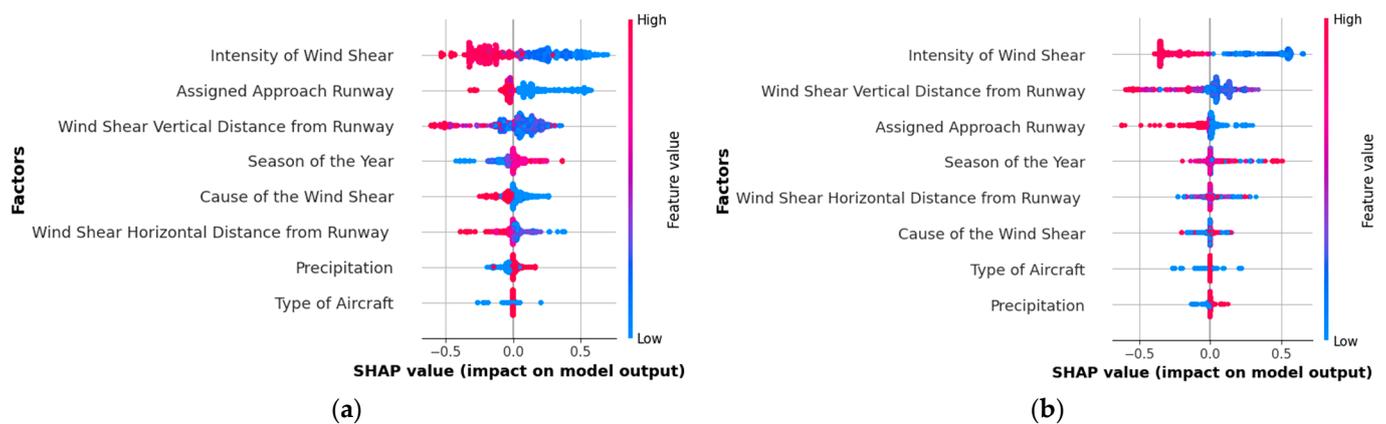


Figure 13. SHAP bee swarm plot of deep learning models: (a) DCN + Borderline-SMOTE; (b) WDN + Borderline-SMOTE.

For both optimal deep learning models, the three primary factors that exhibited significance were the intensity of wind shear, the assigned approach runway, and the vertical distance of wind shear from the runway. These findings indicate that although there may be slight variations in the performance of these deep learning models, each one may possess distinct advantages in different scenarios. When considering the intensity of wind shear, it can be observed that the blue dots are positioned to the right of the vertical reference line on the SHAP bee swarm plot. This positioning suggests a significant impact of negative wind shear magnitude, indicating the impact of tail wind shear and its influence on aborted landings during wind shear events. In a similar vein, it can be observed that runways at HKIA that are assigned lower codes are indicative of a greater impact on the occurrence of aborted landings. The occurrence of southerly or southeasterly gusts of wind at HKIA increases the probability of wind shear, potentially resulting in notable aborted landings at runway 07R. The aborted landings were additionally impacted by the lower altitude of wind shear events. The results of the factor importance and contribution analysis

presented in this study were found to be consistent with previous research conducted by others [41–43].

Furthermore, SHAP interaction plots were developed to analyze the top three factors that are considered significant. Figure 14a illustrates the correlation between the intensity of wind shear and the vertical distance of wind shear from the runway. The presence of red and blue dots positioned above a horizontal green reference dashed line signifies a significant level of effect exerted by the respective factor. It can be observed that the combined influence of tail wind shear, as indicated by negative values, and the low altitude of wind shear, as indicated by blue dots, often results in aborted landings. Nevertheless, the observed head wind shear does not have any substantial influence, and the presence of wind shear at high altitudes does not yield any noteworthy effects on aborted landings. The information presented in Figure 14b indicates there is a higher probability of aborted landings due to trail wind shear at runway 07R. Nevertheless, there were no noteworthy instances of aborted landings recorded on other runways. Figure 14c shows a notable concentration of purple dots, symbolizing runway 07R, in the region situated above the horizontal green reference dashed line and below an altitude of 700 ft, indicating a higher occurrence of aborted landings. Based on the findings, it may be inferred that there is a higher probability of aborted landings for aircraft at runway 07R when wind shear phenomena manifest at altitudes below 700 ft.

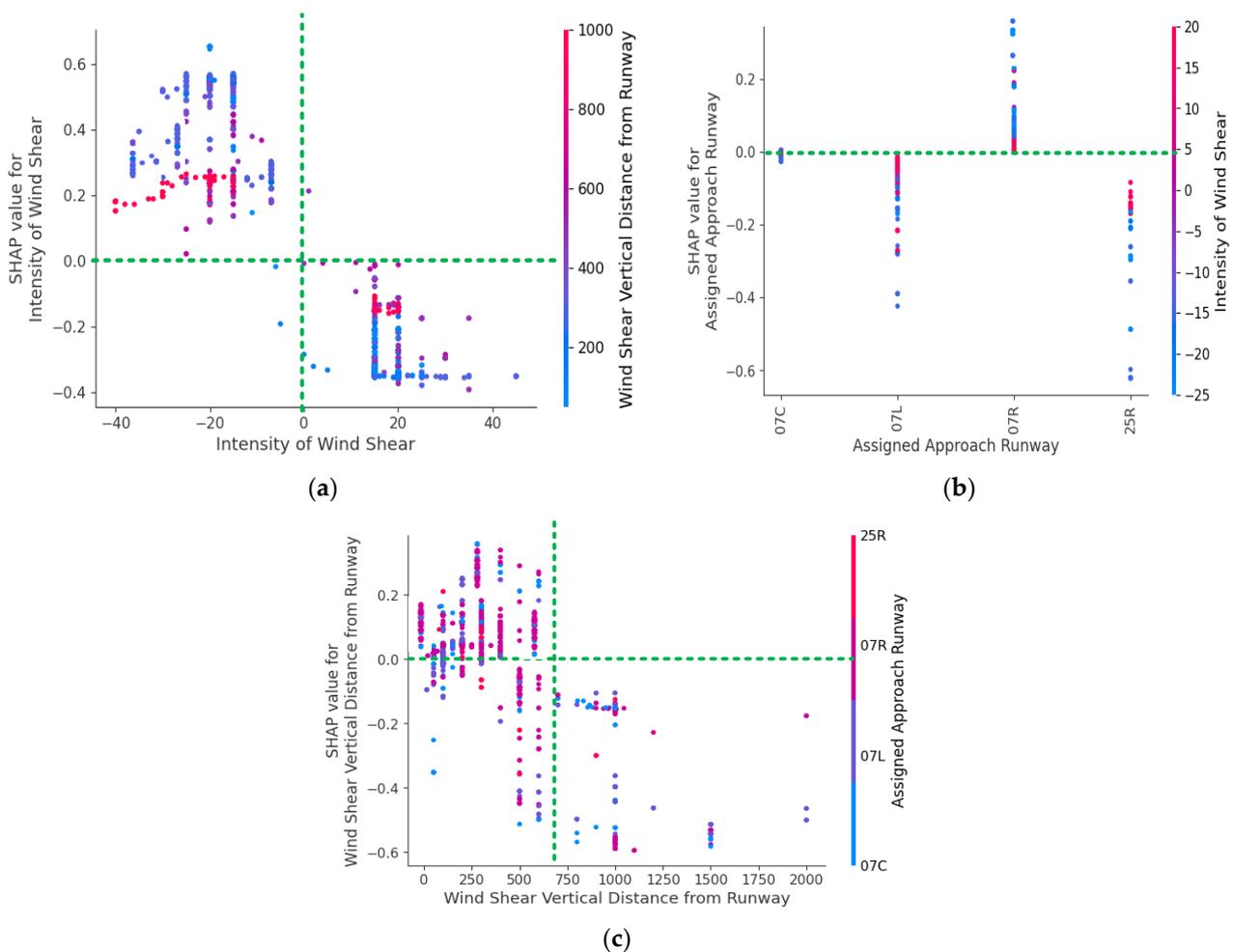


Figure 14. SHAP interaction plots; (a) interaction of intensity of wind shear and wind shear vertical distance from runway; (b) interaction of assigned approach runway and intensity of wind shear; (c) interaction of wind shear vertical distance from runway and assigned approach runway.

5. Conclusions and Future Direction

This study employed three advanced deep learning models, including conventional deep neural networks (DNNs), a deep and cross network (DCN), and wide and deep networks (WDN), to predict instances of aircraft aborted landings due to wind shear at HKIA. Utilizing PIREP data from 1 January 2015 to 23 July 2023, our analysis identified a total of 3585 wind shear events affecting both arriving and departing flights. Of these events, we specifically examined the 2024 cases reported by flights arriving at HKIA, which encompassed 476 aborted landings and 1552 successful landings. To address the imbalance in the data, augmentation techniques such as SMOTE, KMeans-SMOTE, and Borderline-SMOTE were employed. Additionally, the SHAP algorithm was utilized to provide deeper insights into how various factors influence the model's predictions.

Initial results with untreated PIREP data showed lower F1-scores and MCC values, demonstrating the challenges posed by imbalanced data. After applying data treatment, the DCN model, enhanced with Borderline-SMOTE, achieved superior results, with an F1-score of 82.56% and an MCC of 0.773. Following closely was the WDN model, also treated with Borderline-SMOTE, which demonstrated an F1-score of 78.75% and an MCC of 0.719. These outcomes demonstrate the significant potential of targeted data treatments for enhancing model performance. The SHAP analysis identified the intensity of wind shear, the assigned approach runway, and the vertical distance of wind shear from the runway as critical factors influencing aborted landings. Notably, tail wind shear at runway 07R and a lower altitude of wind shear occurrence (below 700 feet above runway level) significantly increased the likelihood of aborted landings.

The findings from this study provide invaluable insights for aircrew, aviation safety researchers, policymakers, and air traffic controllers. They highlight the importance of considering specific environmental and operational conditions when developing strategies to mitigate the risk of aborted landings. Future studies should include a broader array of data sources, such as international flight operations and flights over varied terrain or near large structures, to better understand the diverse factors contributing to aborted landings. There is also a need to explore additional deep learning models to compare effectiveness across different architectures and potentially discover more robust or efficient solutions. Investigating the temporal variability of wind shear effects and its long-term impact on aborted landings could provide deeper insights for training and operational protocols. Assessing the integration of these deep learning models into real-time systems could revolutionize how air traffic controllers manage incoming flights during adverse weather conditions, enhancing both safety and efficiency. Based on the identified key factors, revised aircraft operational protocols and updated flight training programs focusing on handling wind shear conditions could be developed to significantly reduce the probability of aborted landings.

Author Contributions: Conceptualization, A.K.; Data curation, P.-W.C.; Formal analysis, A.K.; Funding acquisition, F.C.; Methodology, A.K. and P.-W.C.; Project administration, F.C. and A.H.; Resources, J.Z. and A.H.; Software, J.Z. and H.A.; Supervision, J.Z., P.-W.C. and F.C.; Writing—original draft, A.K.; Writing—review and editing, A.H. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding: The present study received financial support from the National Natural Science Foundation of China (Grant No. 52250410351), the National Foreign Expert Project (Grant No. QN2022133001L), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), and Xiaomi Young Talent Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are unavailable due to privacy. The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would also like to express our gratitude to our colleagues at the Hong Kong Observatory of Hong Kong International Airport for their guidance.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Hyperparameters of DNN models with different data augmentation strategies.

Parameters	Range	Substitute Number	Optimal Values		
			SMOTE	KMeans-SMOTE	Borderline-SMOTE
Number of hidden layers	[2, 5]		2	2	2
Number of neurons	[2, 100]		[92, 89]	[90, 86]	[90, 94]
Activation function	[Relu, sigmoid, softplus, softsign, tanh, selu]	[1,2,3,4,5,6]	sigmoid	softsign	sigmoid
Learning rate	[0.01, 2]		0.15	0.11	0.19
Optimizer	[SGD, Adm, Adagrad, Adadelta]	[1,2,3,4]	SGD	SGD	Adm

Table A2. Hyperparameters of DCN models with different data augmentation strategies.

Parameters	Range	Substitute Number	Optimal Values		
			SMOTE	KMeans-SMOTE	Borderline-SMOTE
Number of hidden layers	[2, 5]		3	2	3
Number of neurons	[2, 100]		[89, 88, 73]	[85, 86]	[94, 91, 83]
Number of cross layers	[1, 6]		3	4	4
Activation function	[Relu, sigmoid, softplus, softsign, tanh, selu]	[1,2,3,4,5,6]	sigmoid	sigmoid	sigmoid
Learning rate	[0.01, 2]		0.21	0.16	0.17
Optimizer	[SGD, Adm, Adagrad, Adadelta]	[1,2,3,4]	Adm	Adadelta	Adm

Table A3. Hyperparameters of WDN models with different data augmentation strategies.

Parameters	Range	Substitute Number	Optimal Values		
			SMOTE	KMeans-SMOTE	Borderline-SMOTE
Number of hidden layers	[1, 3]		2	2	3
Number of neurons	[2, 100]		[87, 75]	[84, 80]	[96, 93, 85]
Activation function	[sigmoid, softplus, softsign, tanh, selu]	[1,2,3,4,5,6]	sigmoid	tanh	softplus
Learning rate	[0.01, 2]		0.09	0.15	0.13
Optimizer	[SGD, Adm, Adadelta, Adagrad]	[1,2,3,4]	Adagrad	Adm	Adagrad

References

1. Limor, Y.; Borowsky, A. Exploring the Type and Number of Flight Crews' Errors During Reported Incidents of Unsafe Missed Approach Maneuvers. Ph.D. Thesis, Ben-Gurion University of the Negev, Faculty of Engineering Sciences, Be'er Sheva, Israel, 2016.
2. Blajev, T.; Curtis, W. Go-around decision-making and execution project. *Final. Rep. Flight Saf. Found.* **2017**.
3. Jou, R.-C.; Kuo, C.-W.; Tang, M.-L. A study of job stress and turnover tendency among air traffic controllers: The mediating effects of job satisfaction. *Transp. Res. Part E: Logist. Transp. Rev.* **2013**, *57*, 95–104. [[CrossRef](#)]
4. Zaal, P.; Campbell, A.; Schroeder, J.A.; Shah, S. Validation of Proposed Go-Around Criteria Under Various Environmental Conditions. In Proceedings of the AIAA Aviation 2019 Forum, Dallas, TX, USA, 17–21 June 2019; p. 2993.
5. Chou, C.-S.; Tien, A.; Bateman, H. A machine learning application for predicting and alerting missed approaches for airport management. In Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 3–7 October 2021; pp. 1–9.
6. Donavalli, B.; Mattingly, S.P.; Massidda, A. *Impact of Weather Factors on Go-Around Frequency*; National Academies of Sciences, Engineering, and Medicine: Washington, DC, USA, 2017.

7. Proud, S.R. Analysis of aircraft flights near convective weather over Europe. *Weather* **2015**, *70*, 292–296. [[CrossRef](#)]
8. Jiao, Y.; Sun, H.; Wang, C.; Han, J. Research on unstable approach detection of civil aviation aircraft. *Procedia Comput. Sci.* **2018**, *131*, 525–530. [[CrossRef](#)]
9. Lai, H.-Y.; Chen, C.-H.; Khoo, L.-P.; Zheng, P. Unstable approach in aviation: Mental model disconnects between pilots and air traffic controllers and interaction conflicts. *Reliab. Eng. Syst. Saf.* **2019**, *185*, 383–391. [[CrossRef](#)]
10. Moriarty, D.; Jarvis, S. A systems perspective on the unstable approach in commercial aviation. *Reliab. Eng. Syst. Saf.* **2014**, *131*, 197–202. [[CrossRef](#)]
11. Dai, L.; Liu, Y.; Hansen, M. Modeling go-around occurrence using principal component logistic regression. *Transp. Res. Part C Emerg. Technol.* **2021**, *129*, 103262. [[CrossRef](#)]
12. Causse, M.; Dehais, F.; Péran, P.; Sabatini, U.; Pastor, J. The effects of emotion on pilot decision-making: A neuroergonomic approach to aviation safety. *Transp. Res. Part C Emerg. Technol.* **2013**, *33*, 272–281. [[CrossRef](#)]
13. Dehais, F.; Behrend, J.; Peysakhovich, V.; Causse, M.; Wickens, C.D. Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: A behavioral and eye tracking study. *Int. J. Aerosp. Psychol.* **2017**, *27*, 15–28. [[CrossRef](#)]
14. Kennedy, Q.; Taylor, J.L.; Reade, G.; Yesavage, J.A. Age and expertise effects in aviation decision making and flight control in a flight simulator. *Aviat. Space Environ. Med.* **2010**, *81*, 489–497. [[CrossRef](#)]
15. Sze, V.; Chen, Y.-H.; Yang, T.-J.; Emer, J.S. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [[CrossRef](#)]
16. Wang, R.; Fu, B.; Fu, G.; Wang, M. Deep & cross network for ad click predictions. In Proceedings of the ADKDD'17, Halifax, NC, Canada, 14 August 2017; pp. 1–7.
17. Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M. Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016; pp. 7–10.
18. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
19. Zheng, X. *SMOTE Variants for Imbalanced Binary Classification: Heart Disease Prediction*; University of California, Los Angeles: Los Angeles, CA, USA, 2020.
20. Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 878–887.
21. Wu, J.; Chen, X.-Y.; Zhang, H.; Xiong, L.-D.; Lei, H.; Deng, S.-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40.
22. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
23. Hon, K.-K. Predicting low-level wind shear using 200-m-resolution NWP at the Hong Kong International Airport. *J. Appl. Meteorol. Climatol.* **2020**, *59*, 193–206. [[CrossRef](#)]
24. Chen, F.; Peng, H.; Chan, P.w.; Ma, X.; Zeng, X. Assessing the risk of windshear occurrence at HKIA using rare-event logistic regression. *Meteorol. Appl.* **2020**, *27*, e1962. [[CrossRef](#)]
25. Khattak, A.; Chan, P.-w.; Chen, F.; Peng, H. Estimating turbulence intensity along the glide path using wind tunnel experiments combined with interpretable tree-based machine learning algorithms. *Build. Environ.* **2023**, *239*, 110385. [[CrossRef](#)]
26. Khattak, A.; Chan, P.-w.; Chen, F.; Peng, H. Assessing wind field characteristics along the airport runway glide slope: An explainable boosting machine-assisted wind tunnel study. *Sci. Rep.* **2023**, *13*, 10939. [[CrossRef](#)]
27. Chen, S.; Kopald, H.; Avjian, B.; Fronzak, M. Automatic pilot report extraction from radio communications. In Proceedings of the 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 18–22 September 2022; pp. 1–8.
28. Schwartz, B. The quantitative use of PIREPs in developing aviation weather guidance products. *Weather Forecast.* **1996**, *11*, 372–384. [[CrossRef](#)]
29. Hon, K.K.; Chan, P.W. Historical analysis (2001–2019) of low-level wind shear at the Hong Kong International Airport. *Meteorol. Appl.* **2022**, *29*, e2063. [[CrossRef](#)]
30. Pradhan, B.; Ibrahim Sameen, M.; Pradhan, B.; Ibrahim Sameen, M. Predicting injury severity of road traffic accidents using a hybrid extreme gradient boosting and deep neural network approach. *Laser Scanning Syst. Highw. Saf. Assess. Anal. Highw. Geom. Saf. Using LiDAR* **2020**, 119–127.
31. Schrupf, F.; Serdack, P.R.; Fuchs, M. Regression or Classification? Reflection on BP prediction from PPG data using Deep Neural Networks in the scope of practical applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2172–2181.
32. Wang, D.; Thunell, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **2022**, *301*, 113941. [[CrossRef](#)]
33. Khattak, A.; Chan, P.-W.; Chen, F.; Peng, H. Time-Series Prediction of Intense Wind Shear Using Machine Learning Algorithms: A Case Study of Hong Kong International Airport. *Atmosphere* **2023**, *14*, 268. [[CrossRef](#)]
34. Feng, D.-C.; Wang, W.-J.; Mangalathu, S.; Taciroglu, E. Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. *J. Struct. Eng.* **2021**, *147*, 04021173. [[CrossRef](#)]

35. Wang, C.; Feng, L.; Qi, Y. Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China. *Environ. Res.* **2021**, *202*, 111740. [[CrossRef](#)]
36. Lai, Y.; Sun, W.; Schmöcker, J.-D.; Fukuda, K.; Axhausen, K.W. Explaining a century of Swiss regional development by deep learning and SHAP values. *Environ. Plan. B Urban Anal. City Sci.* **2022**, *50*, 23998083221116895. [[CrossRef](#)]
37. Muraina, I. Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts. In Proceedings of the 7th International Mardin Artuklu Scientific Research Conference, Mardin, Turkiye, 10–12 December 2021.
38. Boelrijk, J.; Pirok, B.; Ensing, B.; Forré, P. Bayesian optimization of comprehensive two-dimensional liquid chromatography separations. *J. Chromatogr. A* **2021**, *1659*, 462628. [[CrossRef](#)]
39. Eggenberger, K.; Feurer, M.; Hutter, F.; Bergstra, J.; Snoek, J.; Hoos, H.; Leyton-Brown, K. Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In Proceedings of the NIPS workshop on Bayesian Optimization in Theory and Practice, Lake Tahoe, NV, USA, 10 December 2013.
40. Nguyen, M.H.; Mai, H.-V.T.; Trinh, S.H.; Ly, H.-B. A comparative assessment of tree-based predictive models to estimate geopolymer concrete compressive strength. *Neural Comput. Appl.* **2023**, *35*, 6569–6588. [[CrossRef](#)]
41. Lei, L.; Chan, P.; Li-Jie, Z.; Hui, M. Numerical simulation of terrain-induced vortex/wave shedding at the Hong Kong International Airport. *Meteorol. Z.* **2013**, *22*, 317–327. [[CrossRef](#)]
42. Chen, F.; Peng, H.; Chan, P.-W.; Zeng, X. Wind tunnel testing of the effect of terrain on the wind characteristics of airport glide paths. *J. Wind. Eng. Ind. Aerodyn.* **2020**, *203*, 104253. [[CrossRef](#)]
43. Chan, P. A significant wind shear event leading to aircraft diversion at the Hong Kong international airport. *Meteorol. Appl.* **2012**, *19*, 10–16. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.