Hindawi
International Journal of Aerospace Engineering
Volume 2023, Article ID 6452129, 16 pages
https://doi.org/10.1155/2023/6452129

Research Article

Pilot Maneuvering Performance Analysis and Evaluation with Deep Learning

Shiwen Zhang, Zhimei Huo, Yanjin Sun, Fujuan Li, and Bo Jia

China Eastern Technology Application R&D Center, China Eastern Airlines, China

Correspondence should be addressed to Bo Jia; icesea137@163.com

Received 15 July 2022; Revised 4 December 2022; Accepted 8 March 2023; Published 15 March 2023

Academic Editor: Jinchao Chen

Quick access recorder (QAR) data have been used to evaluate pilot performance for decades. However, traditional evaluation methods suffer from the inability to consider multiple parameters simultaneously, and most of them need to select features manually in advance. To study the relationship between QAR data and pilot performance, this paper puts forward one-dimensional convolutional neural networks (1-D CNN) which consider QAR metrics in an integrated manner. This paper obtained indicators describing the operational status of an aircraft first. Then, the correlation between indicators and pilot performance (skill levels) was studied. Inspired by the fact that CNN can extract both local and global features, this paper has developed an approach to achieve the state-of-the-art result in pilot performance evaluation, which was tested on operating data of Boeing 737. The results prove that other methods do not work well, while the 1-D CNN improves the prediction accuracy of 5 pilot skill levels. Besides, when it is used on a binary classification problem, the result improves to 78.18%. Finally, the indicators were grouped into 5 common factors by factor analysis and fed into 1-D CNN in different combinations. Each common factor plays a different role in pilot performance evaluation, which can provide advice for the future.

1. Introduction

Although the number of accidents and fatalities per passenger mile traveled of aviation is much lower than those of other transportation modes [1], the risk of aviation security is still an important issue that results in injuries and deaths and also causes negative impacts on the global aviation industry and economy. Traditionally, experts learn from accidents to prompt and deploy countermeasures to avoid accidents and injuries in the future [2]. However, this approach always learns lessons afterwards and is not effective in preventing accidents from happening in the first place. The accident prevention approaches in airline industry have gradually shifted to positive ways, which detect hazards and symptoms before accidents occur. Studies that hold this idea can be divided into reliability analysis of aircraft equipment, prediction of environmental threats, and assessment of pilot performance. Because aircrafts and their systems are becoming more reliable, the majority of accidents are caused by human factors [3]. That is why the International Civil Aviation Organization (ICAO) and International Air Transport Association (IATA) all suggest airlines to evaluate pilot performances rationally for advanced training and management. Although there are pros and cons of pilot operation and his/her skill maturity has been studied for decades, most widespread methods of assessing pilot performances are based on the subjective evaluations from instructors [4]. This may lead to an inconsistent evaluation because of different standards across instructors.

Civil airlines generate massive amounts of flight data, recorded and saved by QAR, as a result of routine monitoring [5]. For the objective study of pilot operational performance, a great amount of QAR data are utilized by airlines and their engineers. However, extracting safety-related information from QAR data still focuses on detecting exceedances, which are set to analyze one-dimensional data separately [6]. Numerous cases have been conducted based on such exceedance detection. Worldwide, the most popularly used flight data monitoring tools (Spirent GRAF, Teledyne AirFASE, and Sagem AGS) rely mainly on it, which characterize events that fall outside operator-determined standards [7]. Even the evaluation criteria of some novel pilot

performance evaluation systems were still the risk of exceedances [8]. Since the flight data is multidimensional and complex, such a method of exceedance detection does not provide a comprehensive evaluation of the pilot performance. The morning report of atypical flights (developed by NASA), combining with clustering tools (such as K-means), was proposed to discover anomalous events using QAR data [9]. Similarly, the Sequence Miner used the multiple kernel anomaly detection method to analyze both discrete and continuous QAR data [10]. However, these approaches can only provide risk warning for a single flight and do not effectively discriminate pilot operational performances.

Since the QAR data contains aircrafts' kinetic and maneuvering information, we attempt to use it to evaluate pilot operational performance objectively. Although some airlines and studies have assessed pilot performance using QAR data, these exceedance-based approaches always require experts to set thresholds in advance. In 2019, Wang et al. developed a method to evaluate landing operation performance by utilizing the distributions of QAR data and incident data, but this approach only targeted three unsafe factors in the landing phase, rather than comprehensive features of multidimensional flight data [11]. This paper is aimed at developing a new data-driven method to identify pilot capability without any expert-defined criterion. Thus, a 1-D CNN is proposed to detect the pilot performance free of initial assumption. For using supervised machine learning algorithms, a reliable labeled dataset is required as a training set.

Therefore, the proposed approach in this paper has the following advantages: First, the evaluation of pilot performance is done objectively and automatically through the data-driven method and can be continuously updated with the latest QAR data. Second, the fusion capability of CNN for features is utilized to avoid the limitation caused by considering each exceedance separately and the misleading due to subjective assumption. Third, this paper collects large-scale QAR data and pilot skill levels as the training set for the first time, which ensures the reliability of the machine learning algorithm.

The rest of the article is organized as follows. Section 2 reviews the relevant literatures. Section 3 puts forward the data collection, 1-D CNN classifier, and questionnaire for pilot performance evaluation. Section 4 presents the results of data correlation analysis and performance prediction using 1-D CNN. Section 5 discusses the prediction results obtained from different methods. Section 6 concludes the study with some suggestions for future research.

2. Literature Review

Aviation accidents not only result in casualties but also lead to great psychological harm, material damage, and public confidence loss [12]. Reports from AirSafe, ICAO, and Jet Airliner Crash Data Evaluation Center contain a large number of aviation safety incidents, and most of them are more or less caused by pilot factors [13]. Low and Yang explored the effects of human, technical, and operating factors on the insecurity records of 50 airlines from 2004 to 2015 [14]. Results showed that the human factor plays the most influential role in ensuring aviation safety. The operations of pilot will affect aviation safety to a large extent, so research on aviation safety assisted by pilot performances is the most popular area in the aviation field.

During a flight, hundreds to thousands of flight parameters are recorded in QAR data, such as altitude, airspeed, pitch angle, roll angle, engine parameters, and control surface position. Lots of aviation safety experts have devoted to studying factors of pilot maneuvering by using QAR data [15], especially on specific events. In 2014, Wang et al. discovered that the period of 200 feet to touchdown and the flare action are key features of long landing events [16]. Then, they put forward prevention measures from the perspective of pilot operation. There have been several such studies since then. In 2018, Lv et al. proposed a method to evaluate the overrun risk of Airbus 320 [17]. Chung and Kim aimed to study characteristics of hard landing by utilizing the QAR data of 24 recorded hard landing incidents of Boeing 777 in 2021 [18]. The results showed that the main causal factors were derived from low vertical path and late flare. In the same year, Kang et al. also proposed an innovative deep sequence-to-sequence model to estimate landing distance [19]. Experiments on 44,176 Airbus 321 flights showed that the error between the real and predicted landing distance is 26 meters. All the above are studies based on QAR data, but they only reflect pilot performances for specific events.

The flight operation quality assurance (FOQA) program, also known as flight data monitoring in Europe, strives to improve airline safety by utilizing flight data [15]. However, because the exploration of QAR data is still focalized on detecting exceedances, following approaches can only expose a minority of the information buried in the data [20]. To obtain more useful information, several machine learning (ML) methods need to be used.

Algorithms and statistical models that enable computers to learn without explicit programming are referred to as ML, which has been applied in various scientific fields [21]. ML can be divided into supervised and unsupervised learning. Because unsupervised learning is focusing on discovering hidden patterns in data, it is typically used to discriminate items from different categories [22]. Common methods, such as K-means and K-medoids, have a disadvantage which requires the number of clusters in advance. Some nature-inspired metaheuristic algorithms are offered to address the problem [23, 24]. Recently, Chen et al. proposed a density-based clustering analysis, which used ant colony system-based algorithm to effectively improve the unsupervised learning method. The method achieved remarkable results in path optimization of unmanned aerial vehicles [25, 26]. In studies using unsupervised learning methods to analyze QAR data, one of the first practices was the Morning Report software to detect anomalies from ordinary flight data [27]. In 2022, Zeng et al. proposed a DBSCAN (density-based spatial clustering of applications with noise) clustering analysis method for aircrafts' outlier data detection [28]. However, evaluating pilot performance is not only to discriminate pilots with different abilities but also to achieve

how well a pilot performs. Therefore, it is necessary to use supervised learning, which needs specific labels for training. Classification trees, random forests, and support vector machines have achieved good results in many fields such as image recognition [29] and disease diagnosis [30]. In 2018, Lv et al. used random forest, support vector machine (SVM), and logistic regression models in their study to predict the probability of overrun and achieved reasonable results [17].

An artificial neural network (ANN) can replicate how human brain neurons process information on a computer, according to Haykin [31]. ANN has been a popular and useful branch for ML as computer techniques developed. Because of its great accuracy, processing speed, fault tolerance, scalability, and convergence, ANN is recognized as being superior in the field of data analysis [32]. Recently, as more efficient ANN, deep learning methods with complex multilayers have been used widely due to better outcomes [33]. One of the common deep learning methods is recurrent neural network (RNN), which is used in time series problems. Long short-term memory is an RNN upgrade, employed in Google, Apple, and Amazon's voice recognition platforms [34]. Another deep learning architecture is deep neural network, which can be classified as multilayer perceptron (MLP), stacked autoencoder, and deep belief network (DBN) based on their layer types and learning methods [35]. The convolutional neural network (CNN) is a biologically inspired deep learning method, which was originally used to accomplish image classification and face identification tasks [36, 37]. Using a filter (kernel), convolution techniques can extract complicated characteristics from an image automatically. Deep features retrieved by CNN offer stronger discriminative and robust representation capacity, allowing features to represent images well. Inspired by this, Amrani and Jiang proposed a method to extract more discriminative characteristics of synthetic aperture radar images, obtaining 99.7% accuracy on the moving and stationary target acquisition and recognition benchmark [38]. They also proposed a very deep CNN to differentiate electrocardiogram (ECG) of heartbeats. Results showed that the method can effectively distinguish heartbeats between normal and three common types of arrhythmia with an error rate less than 10% [39]. Aside from the high performance and the resistance to noise, another significant advantage of CNN can combine feature extraction and classification tasks into a single body.

In order to take advantage of the benefits of CNN, some conversion algorithms have been utilized to represent 1-D vibration signals [40, 41]. However, using the method of CNN has some downsides and restrictions. 2-D CNN are not ideal for real-time applications on mobile and moderate devices, and the networks require a massive size of dataset for training. The first compact and adaptive 1-D CNN was proposed to operate directly on patient-specific ECG signals in 2015 and addressed some flaws by Kiranyaz et al. [3]. 1-D CNN has become popular with state-of-the-art performance in a variety of signal processing applications, such as structural health monitoring and damage identification, in a relatively short period [42–44]. Therefore, the 1-D CNN is proposed to combine a large number of indicators from QAR, which could reflect the maneuvering performance of pilots.

3. Methodology

3.1. Data Acquisition and Processing

3.1.1. Flight Data Acquisition. All 54,893 flights for this research were obtained from an airline in China between January 1, 2021, and March 21, 2021. Only the data of Boeing 737-700 or 737-800 were taken into account. In each flight, thousands of factors were sampled at several different rates from 0.25 to 4 Hz. Excluding the pilot information not recorded, the remaining flights contained 2809 pilots performed as pilot flying (PF) in one or several flights. Of these, 500 were instructors, 666 were captains, 108 were cruise captains, 1057 were first officers, and 418 were second officers.

The indicators were filtered through raw data collected by sensors. For example, the touchdown moment was filtered using records of the air-ground switch installed on landing gears. In total, 79 indicators were selected covering all phases of a flight from taxiing, takeoff, to landing (Table 1). The specific filter of these indicators was set according to airlines' security management experiences and company policies [6]. 17 indicators are from the final approach, 15 from landing, and 17 from takeoff and initial climb, accounting for 62% of the total. According to the statistics of the phases where unsafe events took place, the probabilities of events occurring in four phases above are also higher than in other phases [45]. In addition, 18 indicators are applied to all flight phases.

3.1.2. Data Collection and Cleaning. When the aircraft encounters complex weather such as turbulence in the air, the sensors could induce errors or interruptions. Besides, QAR data might be incorrectly recorded due to temporary hardware failures during transmission and storage.

Usually, there are several types of flight data errors. Firstly, the data exceeds the limit, which is possible to achieve. For example, the common operating empty weight of the Boeing 737-800 excluding passengers and fuel is nearly 41 tons, so flights recorded less than 20 tons of the gross weight of landing are unlikely to happen. Secondly, some filter conditions are set inappropriately. Since the indicators are not directly recorded by the sensors, incorrect results could be produced by engineers due to insufficient consideration. For example, if the time duration from the runway entrance to the touchdown point was set as the flare time, it would be calculated too big when encountered a go-around, including the whole go-around after entering the runway for the first time. Thirdly, the fixed data jumps. For example, the approach speed (V_{app}) of an aircraft is fixed when the landing states are determined, but it is not recorded at every sampling point. Instead of recording "does not exist," the QAR sometimes records a jump pattern, in which the parameter is recorded as a quick succession of its minimum or maximum value. These common errors, as well as others occurring randomly, can remain undetected

Table 1: Pilot performance indicators.

Pilot performance indicator	Flight phase
Touchdown speed (-VREF)	Landing
Speed at 50 ft of landing (-VREF)	
Runway departure speed	
Reverse thrust overuse	
Retard height before touchdown	
Pitch rate at landing	
Pitch at touchdown	
Nose gear touchdown first	
Load at touchdown	
Late reverse thrust use	
Gross weight of landing	
Flare time	
Change of heading during landing MAX	
Bounced landing	
Approach speed at 50 ft (-VAPP)	
Takeoff autopilot turn-on altitude	Initial climb
Roll from 50 to 500 ft MAX	
Roll from 500 to 1000 ft MAX	
Initial climb speed (-V2)	
Speed at landing gear is released	Final approach
Roll from 400 to 50 ft MAX	
Roll from 1000 to 400 ft MAX	
MAX deviation of localizer from 500 to 150 ft	
MAX deviation of localizer from 1000 to 500 ft	
MAX deviation of glide slope from 500 to 150 ft	
MAX deviation of glide slope from 1000 to 500 ft	
Landing gear down height in landing	
Landing configuration setting height	
GPWS warning glideslope	
Final approach using speed spoiler	
Decline rate from 500 to 50 ft MAX	
Decline rate from 1000 to 500 ft MAX	
Approach speed at 500 ft (-VREF)	
Approach speed at 500 ft (-VAPP)	
Approach speed at 1000 ft (-VREF)	
Approach speed at 1000 ft (-VAPP)	
Straight-line taxiing speed MAX	Taxiing
MAX longitudinal load on the ground	Takeoff, landing, taxiing
Roll swing below 100 ft	Takeoff, landing
Exceeding tire limit speed	
Speed below 10,000 ft MAX	Descend
Decline rate above 10,000 ft MAX	
Passenger cabin height warning	Cruise
Mach MIN	
Exceed maximum limiting speed (MAX-VMO)	
Exceed altitude restriction of using flaps	
Altitude overrun	

Table 1: Continued.

Pilot performance indicator	Flight phase
Roll above 1000 ft MAX	Approach
Decline rate from 3000 to 2000 ft MAX	
Decline rate from 2000 to 1000 ft MAX	
Too low landing gear audio warning	All
Too low flaps audio warning	
Stick shaker warning	
Speed with landing gear down MAX	
Speed (MIN-VLS)	
Overspeed of configuration 7 (MAX-VFE)	
Overspeed of configuration 6 (MAX-VFE)	
Overspeed of configuration 5 (MAX-VFE)	
Overspeed of configuration 4 (MAX-VFE)	
Overspeed of configuration 3 (MAX-VFE)	
Overspeed of configuration 2 (MAX-VFE)	
Overspeed of configuration 1 (MAX-VFE)	
Maximum lateral load	
High thrust using speed spoiler	
GPWS warning terrain too low	
GPWS warning terrain pull up	
GPWS warning sink rate	
GPWS warning pull up	
Takeoff pitch rate MAX	Takeoff
Takeoff pitch rate MIN	
Takeoff EGT over temperature	
Takeoff change of configuration height	
Speed at landing gear retracting	
Roll below 50 ft MAX	
MAX deviation of heading from 100 knots to off-ground	
Load at lifting wheel	
Landing gear up height in takeoff	
Rotation speed	
Meteorological radar still on until engines shutdown	Shut off

during the postprocessing. Data cleaning is generally a trade-off between thoroughness and minimizing the loss of valid data. To avoid deleting valid data, the following criteria were intentionally defined.

In the preliminary stage, the main indicators were selected and cleaned as below. Because of aircraft's properties, the highest operating speed of Boeing 737-800 is 340 knots. A 50-knot margin was added and any airspeed above 390 knots was considered a data error [46]. The lowest speed of an aircraft is its stall speed, which is normally increased by 30% to get the landing reference speed (V_{ref}). The lowest (V_{ref}) is 114 knots of an empty B737-700. Therefore, we set the lowest real airspeed to 88 knots, and any speed below it was discarded in our research. The maximum pitch may occur in takeoff and its value is usually between 15 and 20 degrees, while the minimum pitch is -1 degree in descent. Setting a margin of 5 degrees, the data were considered incorrect when the pitch exceeded the range of -6°~25°. Also, civil aviation aircraft's roll angle is no more than 35° generally. Increasing the margin of 5 degrees, the trustworthy range was set to -40°~40°. Other flight indicators were also set following the flight manual to limit the range and increase margins as appropriate.

In the study, 79 indicators were extracted from the raw data of each flight. For each indicator, the mean value of the same PF was calculated, representing his/her average performance on this specific maneuvering.

3.2. 1-D Convolutional Neural Networks. The method of ANN in a variety of ways makes life easier for human beings [47]. CNN is a sort of feed forward ANN inspired by the human visual cortex [37]. Related models are mostly employed to process 2-D information such as pictures and videos [48].

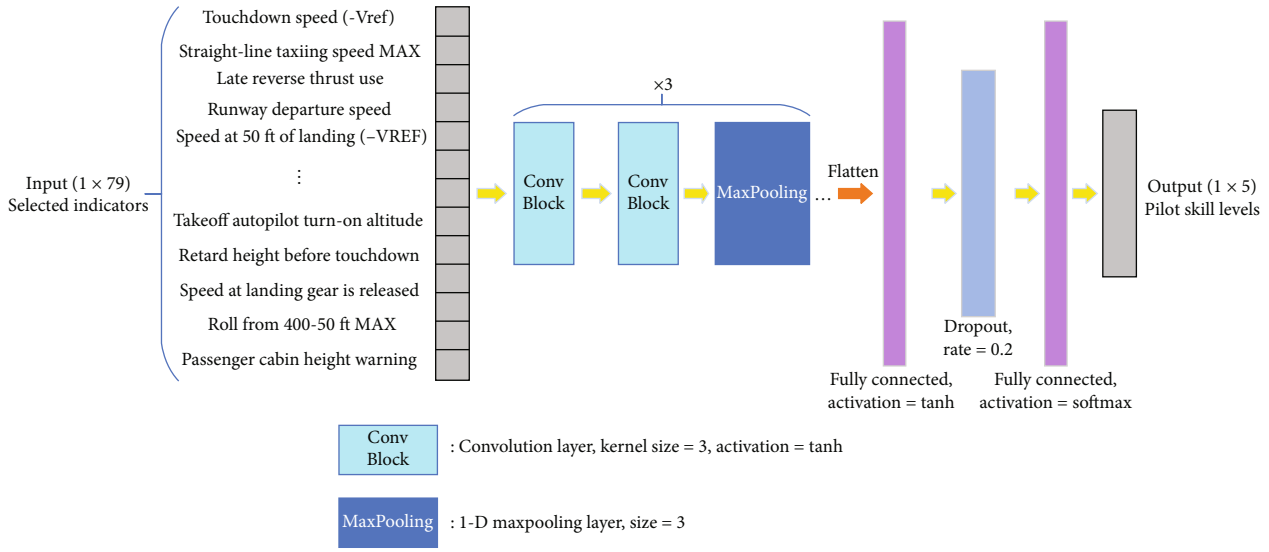

Figure 1: The 1-D CNN architecture.

1-D CNN, working on 1-D signals, is an accommodative version of 2-D CNN. Its convolution kernel and feature map utilize an 1-D vector instead of 2-D matrices. Besides, the computational cost is less [49]. Extraction of features and classification are done in a single body of 1-D CNN which removes the manual handling requirement for extraction of features. Because CNN performs a set of multiscale subband decompositions in each hidden convolution-pooling layer, 1-D CNN has the ability to recognize and isolate patterns from numerous flight indicators.

Indicators, representing the average performance of pilots, are used as the input of 1-D CNN. The architecture of the proposed 1-D CNN is shown in Figure 1. The forward propagation of the l-th convolutional layer is described by

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} \text{conv1D}\left(w_{ik}^{l-1}, s_i^{l-1}\right), \tag{1}$$

where x_k^l is defined as the input, b_k^l is defined as the bias of the k-th neuron at layer l, s_i^{l-1} is the output of the i-th neuron at layer $l-1$, and w_{ik}^{l-1} is the kernel from the i-th neuron at layer $l-1$ to the k-th neuron at layer l. conv1D$(\cdot, \cdot)$ is used to perform 1-D convolution. In order to obtain the network sparsity and improve the robustness of the features, the MaxPooling operator is used and it extracts local maximum values of feature maps as

$$\hat{s}_k^l = \sigma\left(\alpha_k^l \max\left(s_k^l\right) + \beta_k^l\right), \tag{2}$$

where α_k^l, β_k^l, and max $(\cdot)$ denote the weight, deviation, and the MaxPooling function of the k-th pooling unit at l-th MaxPooling layer, respectively, and σ represents the activate function.

In principle, the increase of convolutional and MaxPooling layers implies that more feature information could be learned. Meanwhile, too many convolutional layers will lead to the vanishing gradient problem and increase the computational cost of model training. The MaxPooling is adopted in the network for a downsampled output, which reduces the spatial size of the output and decreases the number of features and the computational complexity of the network. We made use of the block that contained two convolutional layers following with one MaxPooling layer. To get the efficient architecture, one block was added after each trial of the experiment. The structure with 3 blocks converges with fastest speed, obtaining over 80% accuracy on the training set after 17 epochs (Figure 2). If the model has 2 blocks, the training accuracy exceeds 80% after about 35 epochs. If the model has only 1 block, the accuracy of the model on the training set after 40 epochs still cannot exceed 70%, and the trained network is considered not to have a stable performance. Since the data dimension had dropped to 1 after 3 blocks, no more convolutional nor MaxPooling layer could be added. In addition, comparing the final results of the network with 2 blocks and 3 blocks on the test set, the prediction accuracy of 3 blocks was higher. Therefore, the proposed architecture contains 6 convolutional layers and 3 MaxPooling layers. The number of filters of the first two convolutional layers was set to 16, while the others were 64. The activation function was tanh, and the kernel size of MaxPooling layers was 3 [50].

The output data from the last MaxPooling layer was flattened to combine the features extracted from the previous layers and then fed into the fully connected layer. The first fully connected layer outputs 128 channels, using tanh as the activation function. The dropout layer was added to improve the robustness of the network, dropping out a percentage of outputs from the previous layer [51]. The dropout rate was 0.2. The last fully connected layer performed the classification utilizing softmax as the activation function and output 5 channels, corresponding to the skill levels of pilots.

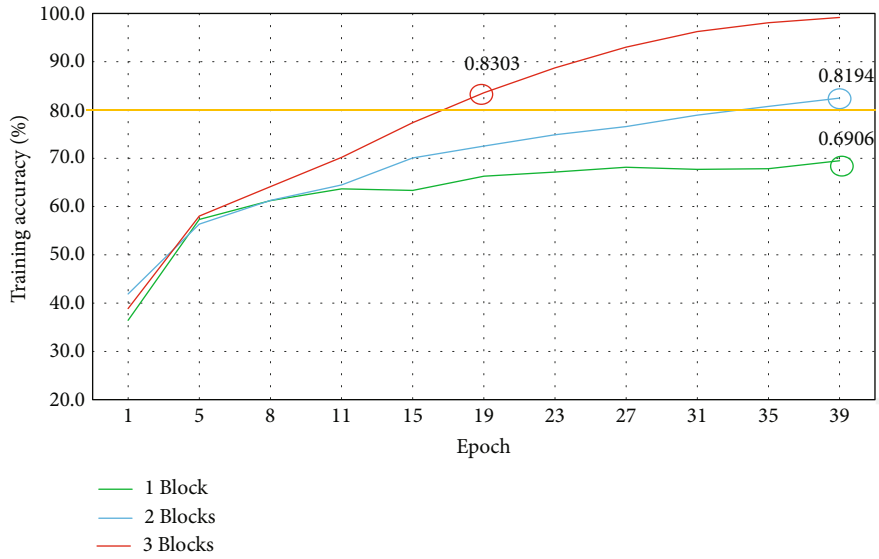

Figure 2: Effect of model architecture on training accuracy.

Assume the input layer $l = 1$ and the output layer $l = L$. The error in the output layer is described as

$$E = E\left(y_1^L, \cdots, y_{N_L}^L\right) = \sum_{i=1}^{N_L} \left(y_i^L - t_i\right)^2. \quad (3)$$

Corresponding to the input value, the output vector is $[y_1^L, \cdots, y_{N_L}^L]$. The purpose of backpropagation is to find the derivative of error according to an individual weight, w_{ik}^{l-1}, and bias of the neuron k, b_k^l, so that the gradient descent approach can be performed to a minimum error [3].

3.3. Analysis of the Questionnaires. The pilot performance in airlines is generally evaluated as a composite ability, but the actual performance is not an indivisible entity. In the ICAO description of pilot competencies, there are 8 core competencies of a pilot, including the application of procedures, communication, aircraft flight path management (automation), aircraft flight path management (manual), leadership and teamwork, problem-solving and decision-making, situation awareness, and workload management [52]. In [53], the abilities reflected by flight data were also classified into 4 categories such as integrity, smoothness, accuracy, and punctuality.

In this paper, the opinions of expert pilots on the importance of indicators were collected by questionnaires, and then, specific indicators would be downscaled by factor analysis. By grouping variables with high correlation, multiple indicators could be converted into fewer integrated indicators that are independent of each other and contain key information to determine pilot performance.

During 79 indicators presented in this paper, some are more sensitive to the pilot performance, while the rest have weaker associations with the performance. Therefore, removing the indicators with weaker associations before the prediction could improve the efficiency of observation and evaluation. Besides, some indicators reflect the same ability to determine performance. It is necessary to know whether the indicator could improve the results of 1-D CNN.

3.4. Questionnaire Design. The questionnaire is divided into 2 parts: the first part consists of basic information, including age, skill level, flight hours, and graduate school of pilots. The second part is the main section. Experts are collected to evaluate the indicator importance to pilot performance. Importance is noted from 1 to 5. The score of 5 means that it is much more important, while the score of 1 means the lowest importance.

Expert pilots who participated in the questionnaire were from different airlines in China, with flight hours ranging from 22 to 26,000 hours. The questionnaire was designed in a multiple-choice format with 60 questions. It summarized the flight performance of 79 indicators listed in Table 1. In order to ensure the typicality and validity of the survey results, the questionnaires with complete basic information were selected. A total of 348 valid questionnaires were obtained.

Factors were extracted from the questionnaire responses using factor analysis. The factor analysis is a technique which is used to reduce a large number of variables into fewer factors [54].

4. Results

4.1. Data Analysis. All the data were collected, cleaned, and processed. When calculating the correlation, the ranking of "instructor, captain, cruise captain, first officer, second officer" was replaced with "5, 4, 3, 2, 1" in descending order. (According to the company's policy, after graduating from flight school and joining the airline, one becomes a second officer first. After approximately one year, if the technical assessment is satisfactory, he/she will be promoted to first officer. In about 4 to 5 years, if the safety record is good

Table 2: Correlations of indicators with flight skill levels.

Indicator	Correlation coefficient	Significance (two-tailed)
Takeoff autopilot turn-on altitude	-.387**	0.000
Runway departure speed	-.358**	0.000
Late reverse thrust use	-.311**	0.000
Approach speed at 1000 ft (-VAPP)	-.307**	0.000
Retard height before touchdown	.277**	0.000
Landing configuration setting height	.264**	0.000
Change of heading during landing MAX	-.254**	0.000
Flare time	-.243**	0.000
Landing gear up height in takeoff	-.231**	0.000
Landing gear down height in landing	.222**	0.000
Exceed altitude restriction of using flaps	.195**	0.000
Approach speed at 500 ft (-VAPP)	-.165**	0.000
Takeoff pitch rate	-.157**	0.000
Straight-line taxiing speed MAX	-.155**	0.000
MAX deviation of localizer from 1000 to 500 ft	-.144**	0.000
Roll above 1000 ft MAX	-.134**	0.000
MAX deviation of heading from 100 knots to off-ground	-.112**	0.000
Stick shaker warning	.108**	0.000
MAX deviation of glide slope from 1000 to 500 ft	-.106**	0.000
Overspeed of configuration 2 (MAX-VFE)	.105**	0.000

**$p < 0.01$.

and the technical assessment is passed, the pilot will be promoted to cruise captain. About one year afterwards, the pilot can be promoted to captain. After 3 years of experience as a captain, one can be promoted to instructor.) The correlation among indicators is shown using the Spearman correlation analysis (Table 2). The table shows indicators with higher correlations under significance at 0.01.

There are 20 out of 79 indicators whose absolute correlations are greater than 0.1, and 10 of them are greater than 0.2. It is denoted that there are positive or negative weak correlations between 10 indicators and flight skill levels. The four most strongly correlated indicators are all negatively correlated. It is indicated that pilots with higher skill levels turn on the autopilot at a lower altitude after takeoff ($r = -0.387^{**}$), exit the runway with less speed ($r = -0.358^{**}$), use reverse thrust earlier after landing ($r = -0.311^{**}$), and hold less approach speed at 1000 ft ($r = -0.307^{**}$), respectively. The two most positively correlated indicators are the retard height before touchdown ($r = 0.277^{**}$) and the landing configuration setting height ($r = 0.264^{**}$).

4.2. Pilot Performance Prediction Using 1-D CNN. First, the indicators were normalized to remove the effect of units. Then, each pilot's average performance and his/her skill level were put into CNN networks, which consists of a 7-layer structure mentioned above. The learning rate, epochs, batch size, and dropout rate of the data were set to 0.0001, 40, 1, and 0.2. 80% of the data was selected as training dataset randomly and the rest 20% as the test data. The accuracy of CNN networks on the training dataset converged to 98% after 30 epochs. Since our training data was labeled with the pilot's skill levels, we cannot absolutely say that skill levels equal to the pilot performance. Too much convergence in the training phase will cause over fitting, while less convergence leads to an incompletely trained model. During experiments, when the accuracy on the training set was in 60-80% interval, the result on the test set was better (around 60%). Nevertheless, we made the model completely trained here, and the prediction results on the test dataset were $55.90\% \pm 1.42$.

One prediction with 57.45% accuracy is shown in Figure 3. The highest accuracy of prediction for the first officer is 0.70, while the wrong prediction of accuracy with 0.13 and 0.17 being the first officer is classified as the second officer and other categories. The accuracy of prediction is 0.56 for those whose true category is instructor, while the prediction of accuracy with 0.23 is wrong assigned to captains. Those whose true category is captain are classified as captain and instructor with 0.55 and 0.17, respectively, and 0.29 is classified as lower skill levels. The accuracy of classified to the second officer is 0.38, while it is classified to the first officer with wrong accuracy of 0.49, belonging to the second officer. The prediction accuracy of 0.10 is classified as cruise captain only, the rest is classified in other categories, and 0.71 is classified as the first officer wrongly. The proper reason is that the number of cruise captain is less.

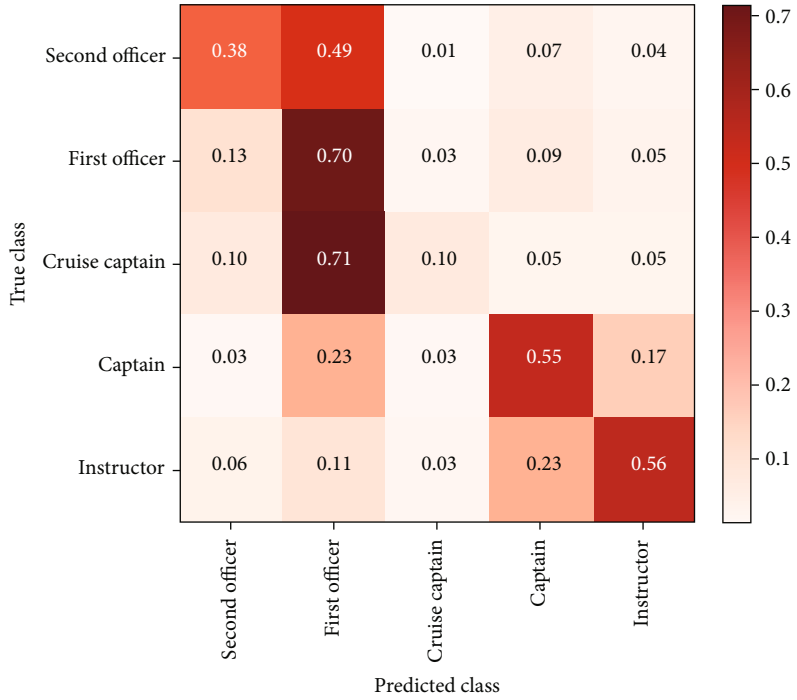

Figure 3: The prediction of individual performances in five levels.

4.3. Factor Analysis of Questionnaires. The answers collected for each question in the questionnaire constituted the set of important ranks of an indicator. Cronbach's α coefficient for indicators in the questionnaire was 0.989, denoting a good consistency of questions in the questionnaire. Then, based on the information obtained from 348 valid responses, a factor analysis of flight performance was conducted. Validity testing of the sample data was required before conducting factor analysis. The Kaiser-Meyer-Olkin (KMO) test and Bartlett's test [55] were used to measure the adequacy and independence assumptions of samples, respectively. The result of KMO is 0.976. It is higher than 0.7 meaning that it is suitable for using factor analysis for the existence of a potential factor structure among the indicators. Also, the significance of Bartlett's test is 0.000, less than 0.01. The assumption of independence of the items is not valid and the correlation matrix is not an identity matrix. The validity of the sample data is confirmed and the sample is suitable for factor analysis.

The common factors were selected with eigenvalues greater than 1, and then, 5 common factors were obtained. The value of 73.32% of the total variance could be explained. The communality is the sum of the squared component loadings up to the number of components extracted. For each indicator, the maximum communality is 0.861 of "Exceed altitude restriction of using flaps." The minimum communality of "Takeoff EGT over temperature" is 0.445, which is the only one less than 0.5. There are only 16 communalities of indicators less than 0.7, showing that the rest of the indicators could be well explained by the common factors.

The indicators were classified using the rotated factor matrix, and the results are shown in Table 3. There are 30 indicators with a larger correlation coefficient in the first common factor than the other factors, so they could be classified into one category. According to the meaning of indicators in this category, we named manipulation accuracy as the first common factor. Similarly, the categories of guidance tracking, exceeding recognition, general manipulation habits, and risk prevention are identified. After corresponding indicators in the questionnaire to the indicators of flight performance in Table 1, there are 34 indicators of manipulation accuracy, 18 indicators of exceeding recognition, 11 indicators of risk prevention, 9 indicators of guidance tracking, and 7 indicators of general manipulation habits in each category.

5. Discussion

5.1. Comparison with Traditional Methods. The traditional approach in the area of using QAR data to evaluate pilot performance needs experts to set limits in advance. When an indicator exceeds those limits, the flight is considered at risk and the pilot will be penalized after that. This approach usually focuses on one single indicator at a time and assumes a significant association between indicators and flight performance.

After analyzing the relationships between flight performance indicators and pilots' skill levels with the Spearman correlation, only 10 indicators' absolute correlation coefficients are greater than 0.2, which means weakly correlated. The absolute correlation coefficients of the rest are less than 0.2, meaning there is no linear relationship. Therefore, if these indicators are used in isolation, they would not reflect pilot performance precisely. Using such results to reward or punish pilots is criticized, which is why airlines and pilots

Table 3: The rotated factor matrix.

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Principal component
Roll above 1000 ft MAX	0.843	0.140	0.214	0.221	0.090	
Decline rate above 10,000 ft MAX	0.827	0.176	0.156	0.197	0.147	
Roll from 500 to 1000 ft MAX	0.779	0.253	0.285	0.241	0.101	
Decline rate from 3000 to 2000 ft MAX	0.768	0.248	0.171	0.285	0.169	
Takeoff autopilot turn-on altitude	0.755	0.184	0.234	0.196	0.150	
Takeoff change of configuration height	0.712	0.184	0.311	0.209	0.087	
Late reverse thrust use	0.659	0.216	0.146	0.417	0.245	
Speed below 10,000 ft MAX	0.656	0.152	0.174	0.404	0.353	
Initial climb speed (-V2)	0.655	0.398	0.303	0.241	0.112	
Straight-line taxiing speed MAX	0.647	0.211	0.319	0.128	0.162	
MAX deviation of glide slope from 1000 to 500 ft	0.642	0.379	0.391	0.083	0.244	
Roll from 1000 to 400 ft MAX	0.641	0.393	0.252	0.290	0.239	
Roll from 50 to 500 ft MAX	0.623	0.442	0.319	0.151	0.146	
MAX deviation of localizer from 1000 to 500 ft	0.618	0.360	0.417	0.132	0.237	
High thrust using speed spoiler	0.612	0.261	0.281	0.371	0.188	Manipulation accuracy
Approach speed at 1000 ft (-VAPP)	0.608	0.366	0.251	0.279	0.303	
Pitch at touchdown	0.601	0.507	0.254	0.097	0.167	
Retard height before touchdown	0.593	0.488	0.139	0.234	0.134	
Approach speed at 500 ft (-VREF)	0.578	0.455	0.239	0.316	0.287	
Flare time	0.577	0.399	0.157	0.344	0.278	
Decline rate from 2000 to 1000 ft MAX	0.574	0.377	0.189	0.451	0.262	
Speed with landing gear down MAX	0.571	0.335	0.236	0.503	0.138	
Landing gear up height in takeoff	0.556	0.191	0.263	0.510	0.328	
Load at touchdown	0.556	0.332	0.341	0.236	0.105	
Takeoff change of configuration height	0.547	0.356	0.135	0.531	0.208	
Final approach using speed spoiler	0.544	0.231	0.131	0.355	0.480	
Approach speed	0.542	0.423	0.241	0.395	0.267	
Takeoff pitch rate	0.534	0.466	0.338	0.241	0.117	
Pitch rate at landing	0.520	0.386	0.335	0.040	0.219	
Meteorological radar still on until engines shutdown	0.439	0.191	0.248	0.346	0.412	
Roll below 50 ft MAX	0.217	0.678	0.311	0.228	0.267	
Roll from 400 to 50 ft MAX	0.376	0.632	0.270	0.281	0.301	
Touchdown speed (-VREF)	0.509	0.579	0.143	0.359	0.179	
Speed at 50 ft of landing (-VREF)	0.467	0.572	0.249	0.181	0.238	
Pitch rate at landing	0.461	0.566	0.166	0.348	0.250	Guidance tracking
MAX deviation of glide slope from 500 to 150 ft	0.382	0.563	0.321	0.328	0.305	
Roll swing below 100 ft	0.443	0.554	0.375	0.150	0.258	
MAX deviation of localizer from 500 to 150 ft	0.416	0.535	0.304	0.287	0.324	
Decline rate from 500 to 50 ft MAX	0.251	0.527	0.363	0.464	0.240	
Decline rate from 1000 to 500 ft MAX	0.489	0.504	0.298	0.288	0.178	
Exceed altitude restriction of using flaps	0.254	0.183	0.780	0.303	0.249	
Gross weight of landing	0.301	0.046	0.738	0.089	0.170	
Overspeed of configuration	0.287	0.251	0.720	0.309	0.227	
Exceeding tire limit speed	0.275	0.239	0.707	0.351	0.099	Exceeding recognition
Exceed maximum limiting speed (MAX-VMO)	0.227	0.327	0.612	0.285	0.356	
Nose gear touchdown first	0.157	0.288	0.592	-0.015	0.350	
Change of heading during landing MAX	0.287	0.513	0.572	0.130	0.168	
MAX deviation of heading from 100 knots to off-ground	0.144	0.486	0.528	0.169	0.307	

Table 3: Continued.

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Principal component
Takeoff EGT over temperature	0.329	0.351	0.430	-0.071	0.152	
Speed at landing gear retracting	0.410	0.190	0.363	0.685	0.130	General manipulation habits
Speed at landing gear is released	0.408	0.183	0.415	0.632	0.159	
Landing configuration setting height	0.481	0.320	0.239	0.570	0.323	
Maximum lateral load	0.410	0.501	0.248	0.538	0.091	
Runway departure speed	0.468	0.329	0.243	0.501	0.308	
MAX longitudinal load on the ground	0.480	0.497	0.224	0.499	0.158	
GPWS warning	0.118	0.243	0.372	0.079	0.766	Risk prevention
Bounced landing	0.380	0.266	0.179	0.199	0.649	
Stick shaker warning	0.140	0.216	0.481	0.142	0.625	
Passenger cabin height warning	0.181	0.223	0.419	0.312	0.586	
Speed (MIN-VLS)	0.357	0.404	0.317	0.210	0.454	

resist the extensive use of QAR, also known as "overuse of flight data."

Among the 10 indicators with higher correlation coefficients, 6 of them belong to the common factors of manipulation accuracy. Three indicators of them belong to general manipulation habits. There is only one indicator that belongs to exceeding recognition, and none of the indicators with high correlation appears in the risk prevention and guidance tracking categories. The relationships between manipulation accuracy or general manipulation habits and flight performance are stronger, so the indicators belonging to these categories are easier to reflect performances by comparing them with predefined thresholds. In contrast, the other three categories of indicators are difficult to evaluate performance in this way.

The method using 1-D CNN effectively solves the problem of not considering multiple indicators at the same time. The features among different indicators were extracted by the method of 1-D CNN and formed higher-level features, which eventually resulted in good prediction capability at the pilot skill levels.

5.2. The Impact of Skill Levels. Although the overall prediction accuracy is 55.90% ± 1.42, there are two situations need to be considered. First, the prediction accuracy is lowest when the pilots are cruise captains, and most of them are classified in the category of first officer, as in Figure 3. This is because the cruise captain is considered as a transitional level in the airline companies and the number of individuals in this category is less. In this paper, the percentage of cruise captain only account for 3.84%. In addition, most cruise captains only have a short period after their first officer time, indicating the two levels have similar characteristics. Second, most errors are caused by classifying pilots to the neighboring levels of their own, with 49% of second officers being considered as the first officer and 23% of instructors being considered as the captain. Because there is no rigorous flight performance criteria that could be used as training labels, the promotion of pilots can partly reflect their improvement of flying ability. However, there are some counterexamples with higher levels and lower performances. Especially, instructors cannot usually perform better than captains, because the age increase might cause a decline in physical capability and maneuvering performance. Therefore, although the overall accuracy is not high, we still believe that the 1-D CNN method in this paper achieves desirable results in practice.

To relieve the second situation above, the classification problem was converted into a binary one. The instructor and captain were treated as the "high performance" category, while the cruise captain, first officer, and second officer were treated as the "low performance" category. After the processing of 1-D CNN, its prediction accuracy reached 78.18% (Figure 4).

5.3. Comparison with Other Machine Learning Methods

5.3.1. Compared Methods. In this paper, some traditional ML methods (SVM, logistics regression, K-nearest neighbor, and decision trees) were used to classify pilot performances. The SVM proposed by Cortes and Vapnik [56] is a supervised learning model commonly used for pattern recognition, classification, and regression analysis. A data point is viewed as a p-dimensional vector, and the aim of SVM is to separate such points with a $(p-1)$-dimensional hypersurface. The hypersurface that has the greatest margin between different classes is the feasible option [17].

Since around 1970, the logistic model has been frequently utilized for binary regression [57]. The logistic regression is estimating the parameters of a logistic model (Equation (4)) by maximum likelihood.

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}, \tag{4}$$

where μ is a location parameter and s is a scale parameter.

The principle of K-nearest neighbor is that when predicting a new value, it is judged which category it belongs to according to the category of the points closest to it [58]. There is no need to estimate parameters and train. However,

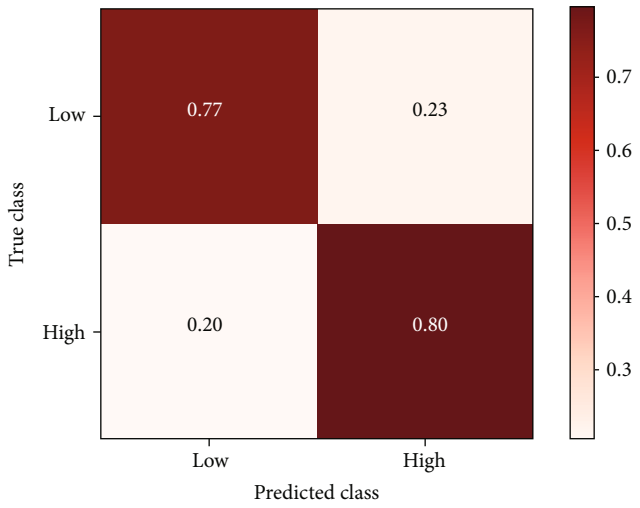

Figure 4: A 2-class prediction with individual performances.

the K value is not fixed, and it has a great influence on the classification results. Prediction results are susceptible to noisy data. When the samples are unbalanced, the prediction of new samples will be biased towards the party with the dominant number.

The decision tree algorithm was proposed early [59]. Each internal node of the tree-like structure represents a "test" on an attribute. It starts at the tree root and split the data on the feature that results in the largest information gain:

$$\text{Information Gains}(s) = H(t) - H(s,t), \tag{5}$$

where $H(t)$ represents entropy of a node of the decision tree and $H(s,t)$ is the entropy of a candidate split at node t of a decision tree. The main disadvantage of decision trees is that, even with prepruning, they often overfit and generalize poorly. Thus, in most applications, the ensemble method introduced by the subsequent random forest is often used instead of a single decision tree.

Three relatively recent methods have been compared here. Fawagreh et al. created the random forest, an ensemble learning technique for classification that builds a multitude of decision trees during training [60]. It compensates for decision trees' tendency to overfit to their training set. AdaBoost is a machine learning approach that is based on the idea of combining many relatively weak and inaccurate prediction rules to create a highly accurate prediction rule. It is often referred to as one of the best out-of-the-box classifier [61]. XGBoost is proposed by Chen et al. in 2015, gaining much popularity and attention recently as the algorithm of choice for many winning teams of machine learning competitions [62].

Two main deep learning methods (MLP and DBN) were also used to compare. MLP is a feed-forward network which consists of a number of neurons connected by linking weights. It is composed of nonlinear layers, which are stacked and trained in a purely supervised manner.

DBN is a generative model that provides joint probability distribution of visible data and class labels. To learn DBN, there are two phases called pretraining and fine-tuning. Pretraining is unsupervised learning with stacked restricted Boltzmann machines (RBMs). In the fine-tuning phase, the network is trained using labeled data based on backpropagation [62].

5.3.2. Results on Pilot Performance Prediction. During traditional methods, the logistic regression gets the highest accuracy, with a percentage of 52.55%. The prediction results of K-neighborhood and decision tree are lower, only reaching 44.91% and 40.73%. Although their accuracies are poor, the results of them are consistent with the actual situation.

The prediction accuracy of SVM, 37.09%, is the worst. The prediction accuracies of each category are 0%, except that of first officer reaching 100%. In other words, all pilots are classified to first officer by SVM. Therefore, for the problem of classifying pilot performance, it is not applicable (shown as dashed bar in Figure 5).

The random forest, AdaBoost, and XGBoost achieve good results. The overall accuracy of the random forest is higher, especially it reaches 91% in category of first officer, but the accuracies are lower in other categories. This leads to the method being heavily influenced by data unbalance. The accuracies of AdaBoost and XGBoost are 58.9% and 63.82%, respectively, which are slightly lower than the method proposed in this paper.

In the comparison deep learning methods, the MLP was comprised of two hidden layers, and each had 64 neurons. Every hidden layer was followed by a dropout layer with the rate of 0.1. The activation function was chosen as ReLU. The DBN had two hidden layers with 128 neurons each. Set the same dropout rate and activation function for this network. The iteration of backpropagation was set to 100 to make the model converge. The accuracy of MLP on the training dataset after 100 epochs was only 0.84. In order to make the 1-D CNN comparable with it, so the epochs of 1-D CNN was chosen here as 20, having an accuracy of 0.81 after the training phase.

The deep learning methods all obtain better prediction results than the traditional ML methods with the current parameter settings. The DBN is 60.25%, the MLP is 62.79%, and the 1-D CNN proposed in this paper is 64.97%. Here, these deep learning methods achieve higher prediction results, because no excessive accuracy (e.g., 98%) was required in the training dataset. This helps to avoid the degradation of prediction accuracy on the test dataset caused by overfitting.

The 1-D CNN method proposed in this paper achieves the best prediction accuracies of 49% and 70% in categories of second officer and captain, respectively (Figure 5). In the category of instructor, the method in this paper is only slightly lower than MLP by 1%. Its prediction accuracy also reaches 63%, which is relatively satisfactory. In the first officer category, the method proposed in this paper is significantly lower than DBN and the random forest, but it is more stable in other categories. In the category of cruise captain, the method in this paper achieves the highest of 19%. This indicates that the 1-D CNN is relatively good for complex problems with multicategories and imbalanced datasets.

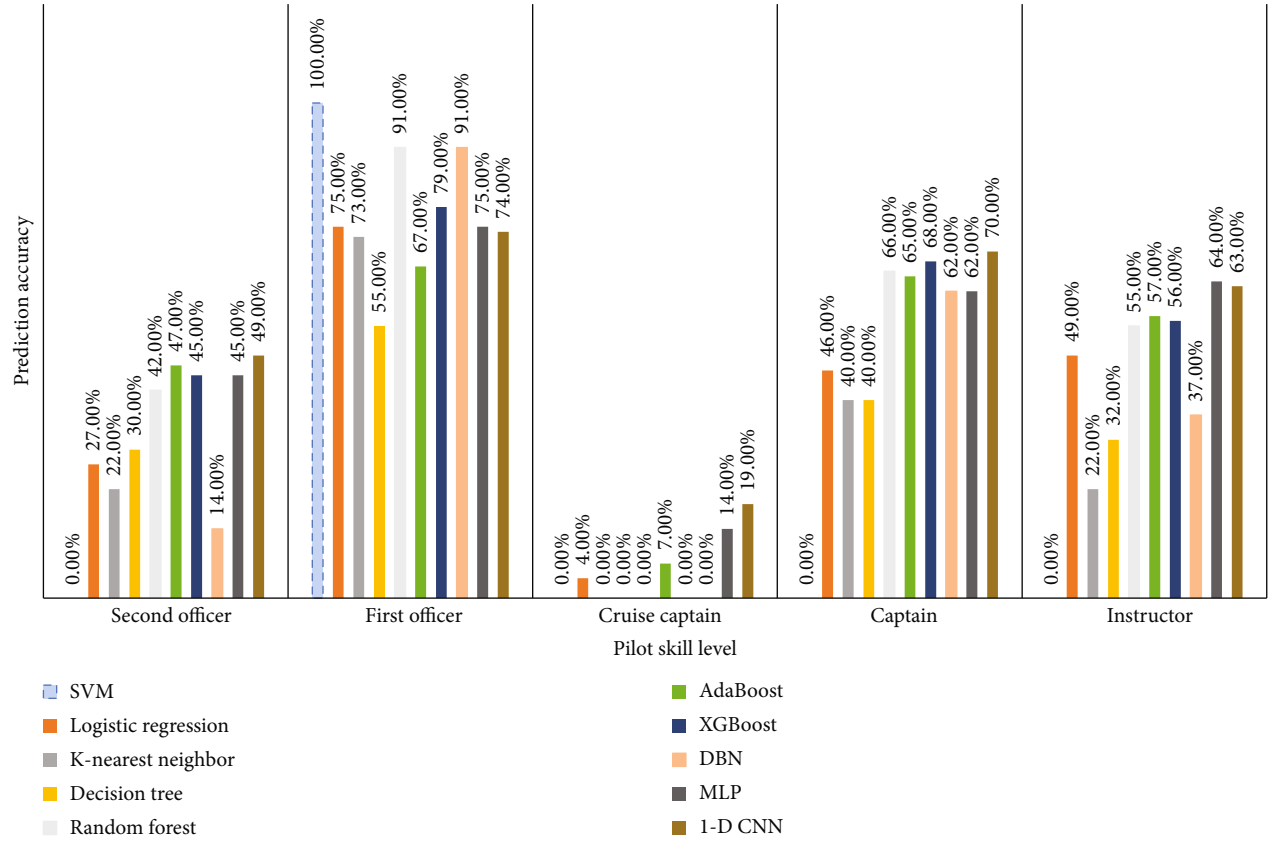

Figure 5: The prediction results using machine learning methods.

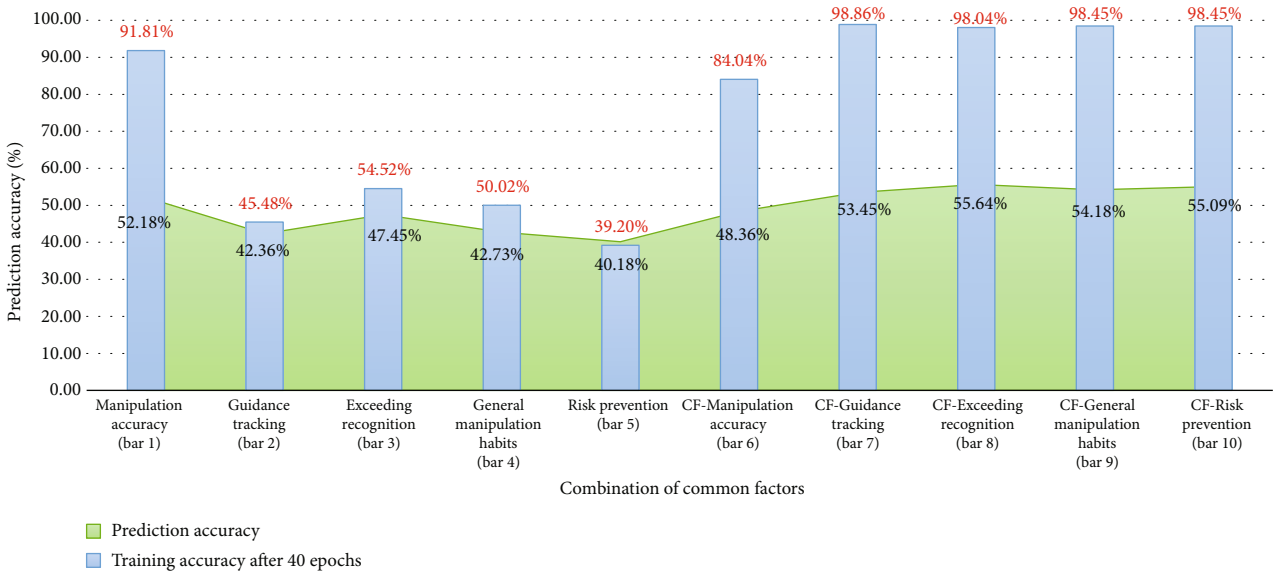

Figure 6: The prediction results with different common factors.

Since airlines tend to classify pilot performance to reflect more minor differences, this leads to more classification categories and imbalanced datasets. Therefore, the approach in this paper is more applicable to practical needs.

5.4. Effect of Common Factors. Different combinations of indicators were used as inputs according to the common factors obtained above. The prediction results of 1-D CNN are shown as percentages in black in Figure 6. The abbreviation CF in the figure indicates all 79 common factors. When the inputs contain the factors of manipulation accuracy, the prediction accuracies are all above 50% (bars 1, 7, 8, 9, and 10). The results are still higher than 53% after excluding one other common factor separately (the 7th, 8th, 9th, and

10th bars), which show minor differences from the prediction using all indicators. It is indicated that the manipulation accuracy has the greatest influence on pilot performance. However, if only the indicators of manipulation accuracy were used as inputs (the 1st bar), the result is 52.18%, slightly worse than previous cases.

When indicators of another common factor are used as inputs individually, all the prediction accuracies are below 50% (bars 2, 3, 4, and 5). Although the values are greater than 40%, most pilots are assigned to the category of first officer in these cases. The proposed method is unable to distinguish the pilots anymore. Indicators of single common factor except manipulation accuracy cannot predict well, but the combination of four common factors together yields a slightly better result (the 10th bar) with an accuracy of 48.36%.

The convergence rates of the 1-D CNN models are shown as the percentages in red in Figure 6. In the training process, the indicators of manipulation accuracy combined with other common factors make the method converge faster than using manipulation accuracy alone. Using the four other common factors together has a faster convergence speed than using them separately and has a higher accuracy on the training set reaching 84.04% after 40 epochs. Here, the accuracies around 50% indicate that the training did not yield fairly fixed models.

It is implied that the performance of pilots needs to be evaluated through diverse indicators, but the main concern is the manipulation accuracy. Attention and development of such abilities need to be focused on during routine flight and training.

6. Conclusions

Airlines have an access to a large amount of flight data recorded in the form of QAR data during their daily operations. However, analyzing and extracting useful conclusions from such a large amount of data is still a challenge, especially for evaluating pilots. A new method of 1-D CNN is proposed in this paper. The routine flight data collected is used automatically to distinguish pilot performances via deep learning.

Based on operational experience, this paper obtained 79 indicators for evaluating pilot performance by setting filter conditions. Routine data of 54,893 flights were used, all of which were operated by Boeing 737-700 or 737-800. The correlation between indicators and pilot skill levels was examined and it is found that there is little correlation between them. It illustrates that the traditional exceedance detection of expert-defined criteria is somehow not feasible. That is why experts are skeptical of using QAR data in this way. After general data cleaning, the indicators are put into the 1-D CNN model proposed in this paper, and pilot skill levels are used as training labels. On the pilot performance estimation with five categories, the total prediction accuracy reached 55.90% ± 1.42. Compared to other ML methods and some deep learning methods, the proposed approach in this paper achieved the best estimation results. In addition, the results of 1-D CNN method improved to 78.18% on the binary classification issue. It shows that the method in this paper could evaluate pilot performance by using the information contained in airline QAR data effectively and give reasonable hints to airlines for subsequent measures.

The questionnaire is designed to obtain the importance of indicators further. The results of the 348 responses were collected for analysis by using factor analysis. The 79 indicators are classified into 5 common factors, which are manipulation accuracy, guidance tracking, exceeding recognition, general manipulation habits, and risk prevention. In order to study the influence of various common factors on the evaluation of pilot performance, the indicators were fed into 1-D CNN in different combinations, respectively. The results show that the pilot performance is mainly influenced by the manipulation accuracy. Besides, other categories also contain some deeply buried information and are beneficial to improve the learning speed of the model.

Currently, the data analysis is limited to Boeing 737-700 and 737-800. A future enhancement of this study is to extend the proposed method to other aircraft types. Since our training data was labeled with the pilot's skill levels, we cannot absolutely say that skill levels equal to the pilot performance. On one hand, we need to adjust the convergence degree of the model appropriately during the training. Too much convergence will cause overfitting, while less convergence obtains the model which is not completely trained. On the other hand, finding reference objects other than the pilot skill levels which are used as training labels may also improve the final evaluation of the model, but there is still a long way to go in this field.

Data Availability

The flight data (QAR data) used to support the findings of this study were supplied by China Eastern Airlines under license and so cannot be made freely available. Requests for access to these data should be made to Bo Jia (icesea137@163.com).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Liu and N. Moini, "Benchmarking transportation safety performance via shift-share approaches," *Journal of Transportation Safety & Security*, vol. 7, no. 2, pp. 124–137, 2015.

[2] T. J. Logan, "Error prevention as developed in airlines," *Physics*, vol. 71, no. 1, pp. S178–S181, 2008.

[3] S. Kiranyaz, A. Gastli, L. Ben-Brahim, N. Al-Emadi, and M. Gabbouj, "Real-time fault detection and identification for mmc using 1-D convolutional neural networks," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8760–8771, 2019.

[4] J. Klinect, P. Murray, A. Merritt, and R. Helmreich, "Line operations safety audit (LOSA): definition and operating characteristics," in *Proceedings of the 12th International Symposium on Aviation Psychology*, pp. 663–668, Ohio State University Dayton, OH, 2003.

[5] A. Circular, "Flight operational quality assurance," Federal Aviation Administration, US Department of Transportation, AC (120-82), 2004.

[6] C. A. Authority, "Cap 739: flight data monitoring".

[7] N. Maille, "On the use of data-mining algorithms to improve FOQA tools for airlines," in *2013 IEEE Aerospace Conference*, pp. 1–8, Big Sky, MT, USA, 2013.

[8] S. Liu, Y. Zhang, and J. Chen, "A system for evaluating pilot performance based on flight data," in *International Conference on Engineering Psychology and Cognitive Ergonomics*, pp. 605–614, Las Vegas, NV, USA, 2018.

[9] N. P. Maille and I. C. Statler, "Comparative analyses of operational flights with AirFASE and the morning report tools," Tech. rep., Technical Report NASA/TM-2009-215379, 2009.

[10] S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 1, pp. 101–113, 2009.

[11] L. Wang, J. Zhang, C. Dong, H. Sun, and Y. Ren, "A method of applying flight data to evaluate landing operation performance," *Ergonomics*, vol. 62, no. 2, pp. 171–180, 2019.

[12] J. Squalli, "Mutual forbearance, the representativeness heuristic and airline safety," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 13, no. 3, pp. 143–152, 2010.

[13] H. Kharoufah, J. Murray, G. Baxter, and G. Wild, "A review of human factors causations in commercial air transport accidents and incidents: from to 2000-2016," *Progress in Aerospace Sciences*, vol. 99, pp. 1–13, 2018.

[14] J. M. Low and K. K. Yang, "An exploratory study on the effects of human, technical and operating factors on aviation safety," *Journal of Transportation Safety & Security*, vol. 11, no. 6, pp. 595–628, 2019.

[15] K. Mitchell, B. Sholy, and A. Stolzer, "General aviation aircraft flight operations quality assurance: overcoming the obstacles," *IEEE Aerospace and Electronic Systems Magazine*, vol. 22, no. 6, pp. 9–15, 2007.

[16] L. Wang, C. Wu, and R. Sun, "An analysis of flight quick access recorder (QAR) data and its applications in preventing landing incidents," *Reliability Engineering & System Safety*, vol. 127, pp. 86–96, 2014.

[17] H. Lv, J. Yu, and T. Zhu, "A novel method of overrun risk measurement and assessment using large scale QAR data," in *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (Big-DataService)*, pp. 213–220, Bamberg, Germany, 2018.

[18] S. S. Chung and H. D. Kim, "B777 hard landing trend analysis based on quick access recorder (QAR) data," *Journal of Advanced Navigation Technology*, vol. 25, no. 2, pp. 169–176, 2021.

[19] Z. Kang, J. Shang, Y. Feng et al., "A deep sequence-to-sequence method for accurate long landing prediction based on flight data," *IET Intelligent Transport Systems*, vol. 15, no. 8, pp. 1028–1042, 2021.

[20] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, "Analysis of flight data using clustering techniques for detecting abnormal operations," *Journal of Aerospace Information Systems*, vol. 12, no. 9, pp. 587–598, 2015.

[21] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: a review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.

[22] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A review of medical image segmentation algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 27, pp. e6–e6, 2021.

[23] B. Alatas, "ACROA: artificial chemical reaction optimization algorithm for global optimization," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13170–13180, 2011.

[24] B. Niu and H. Wang, "Bacterial colony optimization," *Discrete Dynamics in Nature and Society*, vol. 2012, Article ID 698057, 28 pages, 2012.

[25] J. Chen, C. Du, Y. Zhang, P. Han, and W. Wei, "A clustering-based coverage path planning method for autonomous heterogeneous UAVs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, 2022.

[26] J. Chen, F. Ling, Y. Zhang, T. You, Y. Liu, and X. Du, "Coverage path planning of heterogeneous unmanned aerial vehicles based on ant colony system," *Swarm and Evolutionary Computation*, vol. 69, article 101005, 2022.

[27] B. G. Amidan and T. A. Ferryman, "Atypical event and typical pattern detection within complex systems," in *2005 IEEE Aerospace Conference*, pp. 3620–3631, Big Sky, MT, USA, 2005.

[28] C. Zeng, R. Wang, and Q. Zuo, "Analysis of abnormal flight and controllers data based on DBSCAN method," *Security and Communication Networks*, vol. 2022, Article ID 7474270, 8 pages, 2022.

[29] S. Dhivya, J. Sangeetha, and B. Sudhakar, "Copy-move forgery detection using surf feature extraction and SVM supervised learning technique," *Soft Computing*, vol. 24, no. 19, pp. 14429–14440, 2020.

[30] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th international conference on electronic devices, systems and applications (ICEDSA)*, pp. 1–4, Ras Al Khaimah, United Arab Emirates, 2016.

[31] S. Haykin, *Neural Networks and Learning Machines, 3/E*, Pearson Education India, 2009.

[32] A. Mozaffari, M. Emami, and A. Fathi, "A comprehensive investigation into the performance, robustness, scalability and convergence of chaos-enhanced evolutionary algorithms with boundary constraints," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2319–2380, 2019.

[33] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International conference on engineering and technology (ICET)*, pp. 1–6, Antalya, Turkey, 2017.

[34] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[35] H. Patel, A. Thakkar, M. Pandya, and K. Makwana, "Neural network with deep learning architectures," *Journal of Information and Optimization Sciences*, vol. 39, no. 1, pp. 31–38, 2018.

[36] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, Zurich, Switzerland, 2014.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[38] M. Amrani and F. Jiang, "Deep feature extraction and combination for synthetic aperture radar target classification," *Journal of Applied Remote Sensing*, vol. 11, no. 4, article 042616, 2017.

[39] M. Amrani, M. Hammad, F. Jiang, K. Wang, and A. Amrani, "Very deep feature extraction and fusion for arrhythmias detection," *Neural Computing and Applications*, vol. 30, no. 7, pp. 2047–2057, 2018.

[40] W. Zhang, G. Peng, and C. Li, "Bearings fault diagnosis based on convolutional neural networks with 2-D representation of vibration signals as input," in *MATEC Web of Conferences*, vol. 95, p. 13001, EDP Sciences, 2017.

[41] O. Janssens, V. Slavkovikj, B. Vervisch et al., "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.

[42] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, and D. J. Inman, "Wireless and real-time structural damage detection: a novel decentralized method for wireless sensor networks," *Journal of Sound and Vibration*, vol. 424, pp. 158–172, 2018.

[43] O. Abdeljaber, O. Avci, M. S. Kiranyaz, B. Boashash, H. Sodano, and D. J. Inman, "1-D CNNs for structural damage detection: verification on a structural health monitoring benchmark data," *Neurocomputing*, vol. 275, pp. 1308–1317, 2018.

[44] Z. H. Khattak, J. Rios-Torres, M. D. Fontaine, and A. J. Khattak, "Inferring safety critical events from vehicle kinematics in naturalistic driving environment: application of deep learning algorithms," *Journal of Intelligent Transportation Systems*, pp. 1–18, 2022.

[45] S. Shappell, C. Detwiler, K. Holcomb, C. Hackworth, A. Boquet, and D. A. Wiegmann, "Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system," in *Human error in aviation*, pp. 73–88, Routledge, 2017.

[46] J. Oehling and D. J. Barry, "Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data," *Safety Science*, vol. 114, pp. 89–104, 2019.

[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[48] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[49] P. Sharma, S. Chandan, and B. Agrawal, "Vibration signal-based diagnosis of defect embedded in outer race of ball bearing using 1-D CNN," in *2020 International Conference on Computational Performance Evaluation (ComPE)*, pp. 531–536, Shillong, India, 2020.

[50] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: mixed, gated, and tree," in *Artificial Intelligence and Statistics, PMLR*, pp. 464–472, Cadiz, Spain, 2016.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[52] H. Mansikka, D. Harris, and K. Virtanen, "An input–process–output model of pilot core competencies," *Aviation Psychology and Applied Human Factors*, vol. 7, no. 2, pp. 78–85, 2017.

[53] S. Ruishan and X. Yabing, "Research on indicating stucture for operation characteristic of civil aviation pilots based on QAR data," *Journal of Safety Science and Technology*, vol. 8, no. 11, p. 6, 2012.

[54] D. Goretzko, T. T. H. Pham, and M. Bühner, "Exploratory factor analysis: current use, methodological developments and recommendations for good practice," *Current Psychology*, vol. 40, no. 7, pp. 3510–3521, 2021.

[55] N. Shrestha, "Factor analysis as a tool for survey analysis," *American Journal of Applied Mathematics and Statistics*, vol. 9, no. 1, pp. 4–11, 2021.

[56] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[57] J. S. Cramer, "The origins of logistic regression".

[58] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers-a tutorial," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–25, 2022.

[59] B. Kamin'ski, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Central European Journal of Operations Research*, vol. 26, no. 1, pp. 135–159, 2018.

[60] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering*, vol. 2, no. 1, pp. 602–609, 2014.

[61] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference*, pp. 37–52, Springer, 2013.

[62] T. Chen, T. He, M. Benesty et al., "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.