aerospace

MDPI

Article

Pilot Selection in the Era of Virtual Reality: Algorithms for Accurate and Interpretable Machine Learning Models

Luoma Ke [1], Guangpeng Zhang [2], Jibo He [1,*], Yajing Li [1], Yan Li [1], Xufeng Liu [3] and Peng Fang [3]

[1] Psychology Department, School of Social Sciences, Tsinghua University, Beijing 100084, China
[2] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China
[3] Department of Military Medical Psychology, Air Force Medical University, Xi'an 710032, China
* Correspondence: hejibolaboratory@pku.org.cn

Abstract: With the rapid growth of the aviation industry, there is a need for a large number of flight crew. How to select suitable prospective pilots in a cost-efficient manner has become an important research question. In the current study, 23 pilots were recruited from China Eastern Airlines, and 23 novices were from the community of Tsinghua University. A novel approach incorporating machine learning and virtual reality technology was applied to distinguish features between these participants with different flight skills. Results indicate that SVM with the *MIC* feature selection method consistently achieved the highest prediction performance on all metrics with an accuracy of 0.93, an *AUC* of 0.96, and an *F*1 of 0.93, which outperforms four other classifier algorithms and two other feature selection methods. From the perspective of feature selection methods, the *MIC* method can select features with a nonlinear relationship to sampling labels instead of a simple filter-out. Our new implementation of the SVM + MIC algorithm outperforms all existing pilot selection algorithms and perhaps provides the first implementation based on eye tracking and flight dynamics data. This study's VR simulation platforms and algorithms can be used for pilot selection, training, and personnel selection in other fields (e.g., astronauts).

Keywords: pilot selection; expertise differences; virtual reality; machine learning; eye movement; support vector machine

Citation: Ke, L.; Zhang, G.; He, J.; Li, Y.; Li, Y.; Liu, X.; Fang, P. Pilot Selection in the Era of Virtual Reality: Algorithms for Accurate and Interpretable Machine Learning Models. *Aerospace* **2023**, *10*, 394. https://doi.org/10.3390/aerospace10050394

Academic Editor: Peng Wei

Received: 21 January 2023
Revised: 1 March 2023
Accepted: 9 March 2023
Published: 25 April 2023

1. Introduction

For any country, pilots are an indispensable strategic resource for national safety and civil air transportation. Thus, the pilot selection is crucial, as it can improve the flight performance of the pilot, potentially increase the training success rate, and reduce the cost of training pilots [1]. Furthermore, training cost reduction is an important organizational goal for the aviation industry (i.e., the military and civil aviation), especially during the current down seasons of the COVID pandemic. For example, the cost per person who fails to complete undergraduate pilot training is estimated at $80,000 in the US Air Force [2]. According to a previous study, personnel selection impacts training costs and organizational productivity [3]. According to Boeing's 2015–2034 market outlook, 588,000 new commercial airline pilots will be needed worldwide over the next 20 years [4], which imposes an urgent need for better, more accurate, and lower cost pilot selection methods. In brief, the pilot selection is vital for the aviation industry; an appropriate algorithm and platform can reduce training attrition, enhance flight performance, and improve organizational efficiency.

The current pilot selection method still needs improvement to achieve the above goals for pilot selection in at least three aspects. First, pilot selection is conducted by most civil or military pilot training schools or airlines [5,6]. Traditionally, the pilot selection process uses predictors of performance from cognitive ability tests; however, unfortunately, the moderate effectiveness of these measurements cannot live up to the expectations of pilot selections [7]. A recent study showed only weak correlations between self-report and behavioral measures

of the same construct, and it further suggested that these weak correlations were due to the poor reliability of many behavioral measures and the different response processes involved in the two measure types [8]. Second, the performance rating in a flight simulator can predict pilot selection success with mean validities ranging from 0.24 to 0.35, overweighing the predictability of the cognitive ability test [9]. Unhesitatingly, these traditional flight simulators have unique advantages for pilot selection over personality questionnaires and cognitive task performances. At the same time, their drawbacks are distinct (e.g., high cost, insufficient immersiveness, environment unfriendly, etc.) [10]. Third, machine learning (ML) modeling (or artificial intelligence) has gradually been applied in personnel selection in recent years [11]. ML algorithms can handle a larger number of predictor variables than classical statistical modeling and often reflect complex interactions and nonlinearities, and, thus, ML models typically exhibit high predictive powers [12]. However, some ML methods are not interpretable and cannot incorporate expert or regulatory knowledge. Interpretability for ML algorithms is expected by the National Institute of Standards and Technology (NIST) and many researchers, and it is critical for the fairness, acceptance, and training of the pilot selection [13]. Machine learning algorithms may also help to build pilot selection models of a transparent and interpretable decision-making procedure.

Machine learning features for pilot selection should include eye movement metrics, as piloting is mostly a visual behavior (see detailed reviews [14,15]). Many studies have shown that the eye-scanning mode significantly differs between novice and expert [16–18]. Experts' gaze duration on the instrument is shorter and the frequency of gaze on the instrument is higher than that of novices [19,20]. Although eye movement metrics facilitate aviation research, there is still room for improvement in the method based on traditional eye-tracking tools. For example, the GazePoint GP3 desktop eye tracker was used to evaluate the pilot selection process, and the results showed a positive correlation of eye movements with the usual paper and pencil-based selection tests [21]. However, this traditional desktop eye tracker is time-consuming and unpleasant for participants (e.g., participants were asked to keep their heads still through the whole process), whereas, conversely, Virtual Reality Eye Tracker (e.g., HTC VIVE Pro Eye) can be analyzed with automated analysis scripts without time-consuming manual coding of the areas of interest (AOIs) and is also more pleasant for participants (e.g., participants can move around without restrictions). Furthermore, to our best knowledge, the Virtual Reality Eye Tracker is rarely used in pilot selection, if ever attempted.

Another type of feature for machine learning is flight dynamics, which is often evaluated qualitatively and crudely by peer interviewees, but seldomly quantitively evaluated and used in machine learning. The quick access recorder (QAR) keeps various pilot operating and aircraft parameters as well as environmental and alarming information [22]. Which parameters recorded by QAR could be a candidate indicator for pilot selection? The existing literature provided some information but not the full scope of the contribution of flight dynamics data to pilot selection. For example, studies demonstrated that flight performance could be measured as the deviance from the ideal traffic pattern [23,24] and the pitch angle [25]. Useful as it is, QAR or flight dynamics data are not sufficiently utilized in pilot selection. Algorithmic assessment based on the QAR quantification is more reliable than an instructor looking at the flight path and evaluating flight performance subjectively. No studies use QAR or flight dynamics to select pilots quantitatively by machine learning algorithms after an extensive literature search by our research team.

Algorithms of Pilot Selection

In addition to the necessity of using a VR platform and machine learning to select pilots, two core research questions remain to be answered. Firstly, what input features are available and should be used for machine learning? Secondly, what is the best-performing machine learning algorithm for pilot selection? The following two sections address these two core research questions. (See Table 1 for a summary of input features and algorithms for pilot selection).

First, input features for machine learning can potentially include personality, cognitive task performance, electroencephalography (EEG), heart rate, eye movements, and flight dynamics.

1. More specifically, for personality features, Cattell's 16PF-personality scale using the Support Vector Machine (SVM) generated an accuracy of 64% and 78% in two studies by researchers at The Fourth Military Medical University, China [26,27]. A highly relevant summary whitepaper titled, "The predictive power of assessment for pilot selection", generated by the cut-e Group, a consulting company in Germany specializing in pilot selection, reported that a job success prediction accuracy of 79.3% can be achieved using personality characteristics, flight simulator results, and prior flying experience [28].
2. Cognitive tasks are the commonly-used predictors in pilot selection. These cognitive tasks include General Mental Ability (such as general ability, verbal ability, quantitative ability, or the *g*-factor), spatial ability, gross and fine dexterity, perceptual speed, etc. [29,30]. Cognitive tasks have the advantages of being low cost and easy to implement with paper and pencils or a computer, compared to EEG, eye movement, and flight dynamics. Two studies with cognitive tasks as subcomponents to select pilots achieved a predictability of accuracy in the range of 74% up to nearly 94% [31,32].
3. Only one EEG and machine learning study was identified to select pilots [33]. The rare use of EEG to select pilots is perhaps because of the technical difficulty, intrusiveness, and more than 30 min of EEG preparation time to use traditional EEG. The EEG components used in their SVM machine learning classification were the power spectrum factor of alpha, theta, and delta waves at the O1, Oz, and O2 electrodes, and their relative power of the three EEG waves. As summarized in Table 1, a classification accuracy of 76% was achieved for a combination of EEG, heart rate, and eye movement [33], and each component's predictability accuracy is unknown.
4. Similar to EEG, a quite large amount of research has been performed on using eye movement and machine learning to select pilots. Only one study considered three eye movement parameters in selecting pilots: blink rate, average gaze duration, and pupil diameter [33]. Although they achieved an overall 76% prediction accuracy, no independent contribution of eye movement was provided in their work [33]. Despite little work on the utilization of eye movement in the pilot selection, eye movement was able to distinguish novice and expert vehicle drivers [34], and can be used to predict driver cognitive distraction with a high accuracy of 90% using machine learning algorithms like SVM [35]. Recent advances in VR-based eye-tracker can potentially reduce traditional eye-tracker costs and manual coding efforts, which might make the application of eye-trackers in pilot selection more feasible and practicable.
5. Flight dynamics measured by a flight simulator or QAR is often considered in pilot selection [28]. The seminal work published in *Psychological Bulletin* after reviewing 85 years of research in pilot selection reported that the mean validity of 0.63 can be achieved with a combination of general mental ability and a work sample test [30]. Despite the high predictability of flight dynamics, no published scientific study on pilot selection was identified using flight dynamics and machine learning to our best knowledge. Flight dynamics are often rated crudely via peer ratings, which generates much lower validity than a work sample test (0.49 compared to 0.54, respectively) or flight dynamics [30].

Table 1. Summary of Algorithms and Input Features of Pilot Selection.

Index	Algorithms	Input Feature	Accuracy	Institute, Country	References
1	Support Vector Machine	Cattell's 16PF-personality	78%	The Fourth Military Medical University, China	[26]
2	Support Vector Machine	Cattell's 16PF-personality	64%	The Fourth Military Medical University, China	[27]
3	Extremely randomized tree	Cognitive task performance & personality test etc.	Nearly 94%	United States Air Force Academy, United States	[31]
4	Discriminant analysis	Cognitive task performance	74%	ISPA-Instituto Universitário, Portugal	[32]
5	Logistic regression	Cognitive task performance	77%	ISPA-Instituto Universitário, Portugal	[32]
6	Neural network	Cognitive task performance	76%	ISPA-Instituto Universitário, Portugal	[32]
7	Support Vector Machine	EEG, heart rate measured using ECG, eye movements (blink rate, gaze duration, and pupil diameter)	76%	Beihang University, China	[33]

Second, what is the machine learning algorithm for pilot selection? As summarized in Table 1, prior researchers have explored mostly Support Vector Machine (SVM), extremely randomized trees, discriminant analysis, logistic regression, and neural network etc.

As SVM was used mostly according to the summary in Table 1, we conceived a machine learning framework using an SVM classifier that combined a mutual information coefficient (*MIC*) feature selection method to identify pilot experts from novices, in which the *MIC* method is proposed in the hope of complimenting and improving the performance of existing SVM algorithm further. SVM as a powerful algorithm for classification or regression tasks has been proven superior to many related algorithms [36]. Training SVM is to solve a quadratic programming problem, essentially, given a training dataset defined as $D(x_i, y_i), x \in R^d, y \in \{0, 1\}$. The optimal hyperplane determined by $g(x)$ is as follows:

$$g(x) = \text{sign}(\phi(w) \cdot \phi(x) + b) \tag{1}$$

where $\phi(w) \cdot \phi(x) = K(w, x), K(\cdot)$ denotes the kennel function, $\text{sign}(\cdot)$ is the symbolic function, and b is a bias. Different optimization methods of kernel function will be different. The optimization of SVM is then:

$$\begin{gathered} \phi^*(w) = \underset{\phi(w)}{\text{argmin}} \frac{1}{2}\phi(w)^2 \\ s.t. y_i(\phi(w) \cdot \phi(x_i) + b) \geq 1, i = 1, 2, \ldots n \end{gathered} \tag{2}$$

The Lagrange multiplier method is used to solve quadratic convex optimization:

$$L(\phi(w), b, \alpha) = \frac{1}{2}||\phi(w)||^2 - \sum_i \alpha_i (y_i(\phi(w) \cdot \phi(x_i) + b) - 1) \tag{3}$$

where α represents the Lagrangian multiplier. The *MIC* feature selection ranks the importance of features by estimating the mutual information coefficient between features and classes. It can be described as follows:

$$MIC(X, Y) = H(X) + H(Y) - H(X, Y) \quad (4)$$

H is an entropy, and $H(X, Y)$ is a joint entropy of inputs of X and Y. To prove the superior performance of SVM combining *MIC* rather than the previously ordinary SVM [26,27], experiments on multiple classifiers and feature selection algorithms were carried out and compared.

To summarize, to meet the need for pilot selection and leverage the new VR technology, the current study attempted to use the seldomly utilized eye movement and flight dynamics (QAR-like data) as features for machine learning, and to explore multiple machine learning predictors and feature selection methods to identify the best-performing algorithm, with consideration of the interpretability need of the machine learning algorithms.

2. Methods

2.1. Participants

Forty-six righthanded participants (all males) completed the experiment (age range from 25 to 58 years; expert mean age: 32.52 years, *SD* = 7.28; novice mean age: 29.57 years, *SD* = 5.74). Among all of the participants, twenty-three expert participants were recruited from China Eastern Airlines, and the other 23 novice participants were from the community of Tsinghua University.

The research protocol in the current study has received approval from the Institutional Review Board (IRB) of Tsinghua University (ethical approval code: 2022 Ethic Audit No.17). All participants signed informed consent forms and were informed that they could drop out of the experiment at any time without any form of penalty. They were all paid 80 CNY (about 12 USD, or a present of equal value) for participation.

2.2. Procedure

2.2.1. Flight Task

The flight task is a traditional traffic pattern, which mainly includes six stages: upwind, crosswind, downwind, turning base, base, and final (see Figure 1). Participants were asked to put on the HTC Eye Pro VR and finish the flight task in the VR flight simulator.

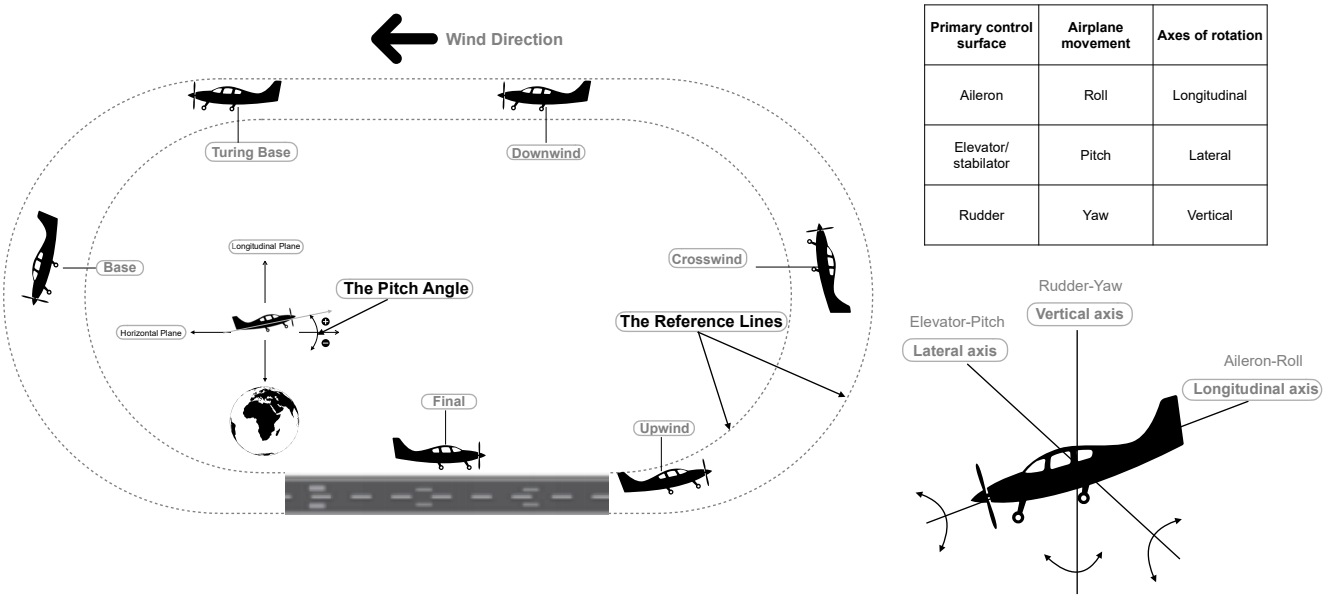

Figure 1. The objective measures of the flight performance related to the traffic pattern.

2.2.2. Experiment Process

The formal experiment was made up of four steps, which included "Information collection", "Flight training", "Formal flight", and "Data collation". The experiment flowchart is shown in Figure 2.

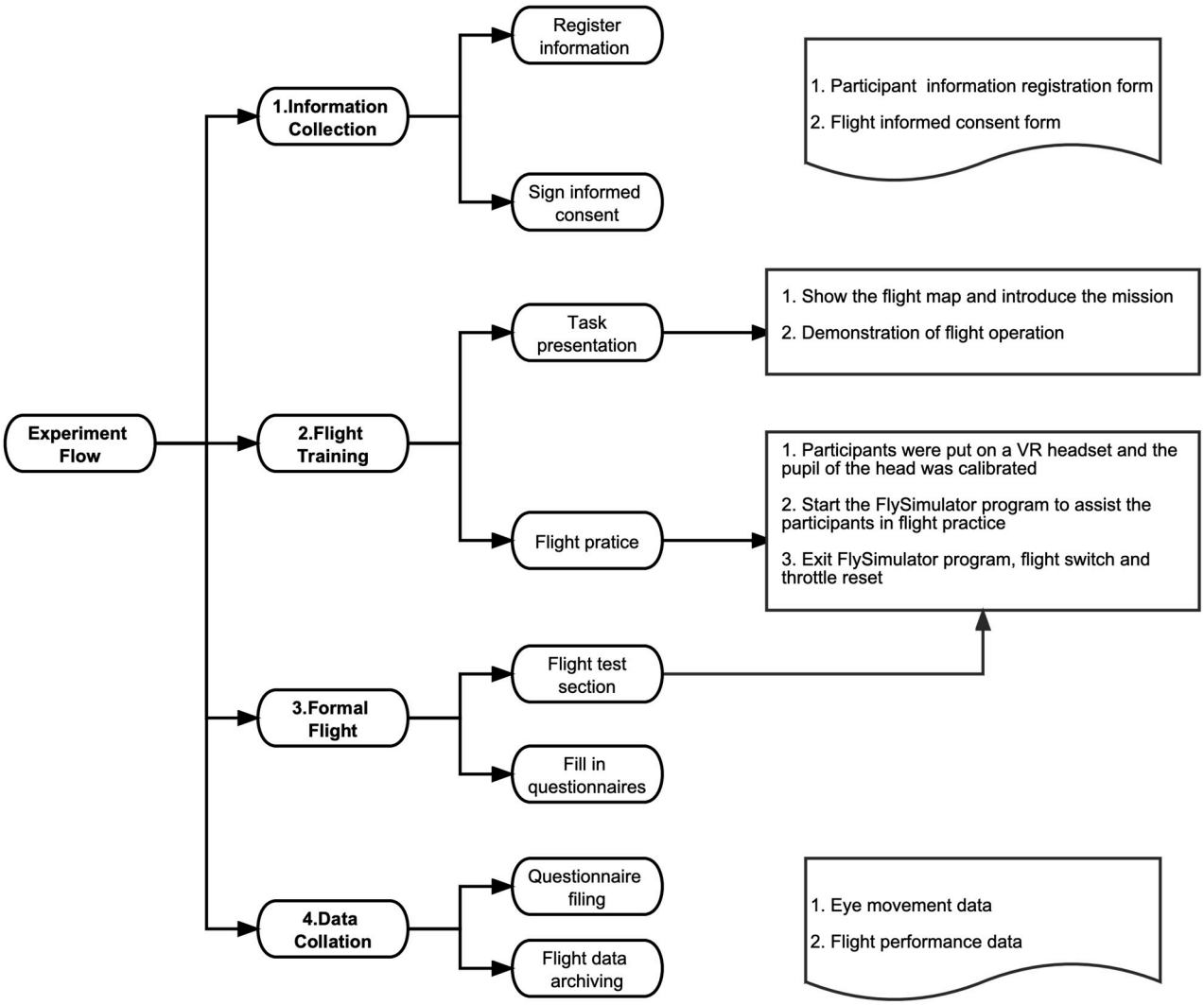

Figure 2. Experiment process flowchart.

The first step is information collection. Before starting the experiment, experimenters ensured that none of the participants had a history of epilepsy, heart or brain diseases, recent endocrine or psychiatric medication, or severe skin allergies. The researchers then used the Snellen eye chart (using the metric system 6/6) to confirm that participants' vision was normal or corrected-to-normal and that they had no visual impairments like color blindness or color weakness. Later, researchers ensured that none of the participants had consumed any alcohol or drugs within the previous 24 h and that the participants had no less than 6 h of sleep and were in a good mental condition before the experiment. This was done to ensure that they had sufficient ability to fly.

The second step was flight training. Participants were shown the flight map and the task mission. They were asked to try to fly within the two reference lines. They were also informed (and it was emphasized twice) that they should come to a complete stop after landing. Next, researchers demonstrated how to use the flight controller, the instruments and how to read the indicators. Afterwards, participants were allowed to practice on

the flight controller until they had the minimum knowledge and confidence to start the formal task.

The third step was the formal experimental flights. The participants were asked to finish the whole task in a maximum of twenty minutes or until their voluntary request to quit the experiment. After the formal flights, participants were asked to complete questionnaires concerning their subjective performance ratings. Finally, participants were acknowledged and reimbursed.

The last step was data collection. Data were generated during the execution of the VR-based flight simulation. The researchers fetched the eye movement and flight data, named according to the participant ID number, and saved it in the experimental computer.

2.3. Feature Selection Method

In pattern recognition, machine learning, and data mining, feature selection is an effective means of data preprocessing [37]. It typically works by reducing the redundancy features of data—decreasing the computation complexity of models, improving the model's robustness, and avoiding overfitting problems [38]. In essence, feature selection is an NP-hard problem because it aims to select a subset from 2^N possible subsets of the dataset, where N represents the number of features of the dataset [39]. Therefore, it is efficient to use feature selection methods to search for the proper combination of features in a polynomial time [40].

Three feature selection algorithms were implemented and compared in this study, namely, mutual information coefficient (*MIC*), support vector machine-based recursive feature elimination (SVM-RFE), and random forest (RF). According to whether feature selection algorithms use classifiers, feature selection algorithms are divided into three categories, which are filter methods (classifier-independent), wrapper methods (classifier-dependent), and embedded methods (classifier-dependent) [41]. *MIC* ranks feature by the estimated mutual information as an instance of filter methods. SVM-RFE, an embedded method, selects the relevant features by their default settings. RF, as a wrapper method, orders a feature-by-feature importance vector. Combining these feature selection methods, selecting a subset of features based on the percentage of ranked features was used.

2.4. Predictors

Pilot selection was a binary classification task in this study. Concretely, the binary predictors were fed as inputting the participant's features and output 1 or 0, where 1 means expert and 0 means a novice. In the present study, we applied feature selection techniques to a dataset and subsequently evaluated the performance of five binary classification algorithms, namely, support vector machine (SVM) [42], k-nearest neighbor (KNN) [43], logistic regression (LR) [44], light gradient boosting machine (LGBM) [45], and decision tree (DTree) [46].

2.5. Cross Validation

Cross validation is a standard method to evaluate model performance in machine learning [47]. It divides the training data into several subsets and uses one subset each time to verify the model obtained from the training of other remaining subsets to reduce the error caused by unreasonable partitioning of the training data. When the number of instances in a dataset or the number of a certain class is too small, it may be beneficial to use a technique called "leave-one-out" cross validation to obtain a more reliable estimate of the accuracy of a classification algorithm [48]. This is a particular condition of k-fold cross validation, where only one instance is left out during each iteration. This allows for using all the data in the estimation process while still providing a way to measure the algorithm's performance on unseen data [49].

2.6. Metrics

This study adopted five commonly used metrics to evaluate our algorithm, i.e., *F*1 score (*F*1), Accuracy (*Acc*), area under the receiver operating characteristic (ROC) curve (*AUC*), *Precision*, and *Recall*. *AUC* calculated by the area under the ROC curve is widely used to assess an unbalanced learning performance [50]. Suppose an *AUC* metric must be labeled as good or bad. In that case, we can reference the following rule-of-thumb from Hosmer and Lemeshow in Applied Logistic Regression (p. 177): 0.5 = No discrimination, 0.5–0.7 = Poor discrimination, 0.7–0.8 = Acceptable discrimination, 0.8–0.9 = Excellent discrimination, and >0.9 = Outstanding discrimination. By these standards, a model with an *AUC* score below 0.7 would be considered poor, and anything higher than 0.7 would be considered acceptable or better. The confusion matrix is a psychophysics and machine learning concept, especially in Signal Detection Theory. The measurement of the confusion matrix includes *FP*, *TP*, *FN*, and *TN*, where *TP* and *FN* are true positive numbers and false negative numbers, respectively; *TN* and *FP* represent the number of true negative and false positive. The formulas of *F*1, *Acc*, *AUC*, *Precision*, and *Recall* are as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$AUC = \frac{\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} 1_{if(p^+ > p^-)}}{n^+ n^-} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TN}{TP + FN} \tag{9}$$

where n^+ is the number of positive samples, n^- is the number of negative samples, $1_{if(p^+ > p^-)}$ represents 1 when $p^+ > p^-$, and p^+ and p^- are the prediction probabilities of positive and negative samples.

2.7. Data Analysis

2.7.1. Flight Performance Data

Six major indicators were applied to evaluate the flight performance in the current study: the total flight time (i.e., time from takeoff to landing, in seconds), pitch angle 1 s before landing (i.e., the pitch angle of the plane 1 s before landing, in degrees), the mean pitch angle 0–10 s before landing, the standard deviation of the pitch angle 0–10 s before landing, mean distance to the center of the two reference lines (in meters), and standard deviation of the distance to the center of the two reference lines (in meters). The total flight time, also named as task completion time, is one of the most commonly-used dependent variables in human factors and aviation psychology. Publications in influential journals in the aviation psychology field rely heavily on task completion time as a performance indicator [51]. The three measurements concerning pitch angles measure the attitude of the aircraft when landing, with a slightly positive angle larger than 0 degree as the optional performance. Large negative values for pitch angle or larger standard deviations of pitch angle indicate potential crash risks. The mean and standard deviation of distance to the center of the two reference lines measure participants' tracking performance to follow a predetermined flight path.

2.7.2. Eye Movement Data

Four key flying instruments (i.e., airspeed indicator, vertical speed indicator, attitude indicator, and altitude indicator) were defined as different areas of interest (AOIs) in the study. Each AOI's percent dwell time (unit: %) was used to evaluate visual attention dispersion. The mean percent dwell time was the cumulative time observed within the AOI divided by the total gaze time and averaged across participants.

2.7.3. Statistic Analysis

An independent sample *t*-test was chosen to compare the means between novices and experts. The test was appropriate for the study to determine whether there was a significant difference between the means of these two groups and whether any differences could be attributed to the expertise differences. Statistical analyses were performed using SPSS Version 24.0 (IBM Corp, Armonk, NY, USA). The significance level was set at $p < 0.05$. The scientific graphs with statistical results were plotted using GraphPad Prism version 9.4.0 for the Mac Operating System (GraphPad Software, San Diego, CA, USA).

2.7.4. Eye Movement Preprocessing & Analysis

The raw data provided by the HTC Eye Pro and Tobii SDK and logged by SRanipal Eye Framework include (a) timestamp in second; (b) gaze origins in millimeters in X, Y, Z axis for left and right eyes (FOL_X, FOL_Y, FOL_Z, FOR_X, FOR_Y, FOR_Z); (c) gaze direction normalized to between −1 and 1 in X, Y, Z axis for left and right eyes (FVL_X, FVL_Y, FVL_Z, FVR_X, FVR_Y, FVR_Z); (d) eye-opening of left and right eyes (EOL and EOR); (e) pupil position normalized to between −1 and 1 in X and Y axis for left and right eyes (PPLX, PPLY, PPRX, PPRY); (f) AOIName logs the 19 predefined areas of interest (AOIs), which include aircraft clocks, airspeed indicator, attitude indicator, vertical speed indicator, radio compass, inlet pressure gauge, altitude indicator, turn-and-slip indicator, aircraft tri-use meter, magnetic course correction calculator, tachometer, cylinder head thermometer, air inlet temperature indicator, current, and voltage meters, tank pressure gauge, spare magnetic compass, left aircraft cockpit glass, front aircraft cockpit glass, and right aircraft cockpit glass. These raw data were converted to features in Table 2.

2.7.5. Flight Dynamics Preprocessing & Analysis

The flight dynamics data were automatically recorded by the VR simulation software. The raw variable included: (a) timestamp in seconds; (b) the gesture of the airplane measured using Roll, Pitch, and Yaw (Heading); (c) the location of the airplane, including longitude, latitude, above ground level (AGL), and above sea level (ASL); (d) flight movement measurements, including velocity and angle of attack (AoA); (e) control device inputs, including rudder, elevator, and roll inputs; and (f) the (longitude, latitude, and height) of the nearest point on the center of the two reference lines.

The above raw values were converted to features in Table 3 according to the QAR (Quick Access Recorder) analysis method inspired by the 35 BAE-146 aircraft QAR data provided by t National Aeronautics and Space (NASA) [52]. It generally takes a Type A aircraft 8 s to land with a beginning height of 730 m and 15 m additional buffer space and a gradient of descent of 4.3% according to the regulation of International Civil Aviation Organization (ICAO) Doc 8168. Thus, 8 s before landing was chosen to calculate the pitch angle and angle of attack [53].

Table 2. Features from eye movements.

Index	Feature/Variable Names	Calculation Method	Performance Indication
1	Standard deviation of fixation in x axis	The standard deviation of FVL_X, e.g., numpy.std(data['FVL_X'])	Indicates the horizontal dispersion of fixations, or how wide participants looked
2	Standard deviation of fixation in y axis	The standard deviation of FVL_Y, e.g., numpy.std(data['FVL_Y'])	Indicates the vertical dispersion of fixations, or whether participants looked up and down. This often relates to whether pilots are able to look forward and near areas to guide their flight path
3	Standard deviation of fixation in z axis	The standard deviation of FVL_Z, e.g., numpy.std(data['FVL_Z'])	Indicates the depth of visual attention
4	Eye opening (%, from 0 to 1)	The mean of EOL or EOR, e.g., numpy.average(data['EOL'])	Indicates how wide the eye opens, which is related to participants' interests and workload
5	Percent dwell time (%) on each AOI	$\frac{\text{AOIName}_i}{\sum_{k=1}^{N} AOIName_k}$ (AOIName_i is a specific AOI, such as the altitude indicator; $k = 1, 2, \ldots$, refer to all the indicators, where $N = 19$, as we have a total of 19 AOIs.)	Indicates the relative attention to a specific AOI, reflecting cognitive processing and understanding of information in that AOI. Note: This is labelled as "AOI" information in machine learning section; all others in this table labelled as "EM" (eye movement)
6	Frequency of AOI transitions (Hz)	The number of sequential pairs (AOIName_i, AOIName_j) where $i! = j$, divided by time	Indicates how actively participants look for information from gauges, which suggests understanding meaning of gauges
7	Fixation duration (ms).	Use i-VT algorithm to detect fixation, with threshold for velocity of 30 o/s	Larger value indicates more time spent on processing visual information
8	Fixation count	The number of fixations	Indicates actively looking for information

Table 3. Features from flight dynamics (QAR data).

Index	Feature/Variable Names	Performance Indication
1	ldg_time (s)	landing time, that is, the time when the airplane lands, which is used as reference time point for the 1 s and 8 s before landing
2	Vert_accel_landing	Vertical acceleration when landing
3	AOA (1 s or 8 s before landing) (°)	Angle of Attack 1 s or 8 s before landing
4	AOA (min & max values)	The minimum and maximum values of Angle of Attack
5	Pitch_angle(1 s or 8 s before landing) (°)	Pitch angle 1 s or 8 s before landing
6	RudderInput (1 s or 8 s before landing)	Rudder input 1 s or 8 s before landing
7	ElevatorInput (1 s or 8 s before landing)	Elevator input 1 s or 8 s before landing
8	RollInput(1 s or 8 s before landing)	Roll input 1 s or 8 s before landing
9	TAS (1 s or 8 s before landing) (m/s)	True air speed (TAS) 1 s or 8 s before landing in unit of m/s
10	GS(1 s or 8 s before landing) (m/s)	Ground speed (GS) 1 s or 8 s before landing in unit of m/s
11	Velocity_Descent_mean (m/s)	Average descent velocity when landing in unit of m/s
12	Longitude_err (*mean* + *SD*) (m)	The mean and standard deviation (SD) of the airplane position in the longitude axis relative to the nearest center of the two reference lines
13	Latitude_err (*mean* + *SD*) (m)	The mean and standard deviation (SD) of the airplane position in the latitude axis relative to the nearest center of the two reference lines
14	Height_err (*mean* + *SD*) (m)	The mean and standard deviation (SD) of the airplane position in the height axis relative to the nearest center of the two reference lines

Table 3. *Cont.*

Index	Feature/Variable Names	Performance Indication
15	dist_err (*mean* + *SD*) (m)	The mean and standard deviation (SD) of the distance of the airplane position to the nearest center of the two reference lines
16	rou (min & max values)	The minimum and maximum values of the turning curvature of the airplane, with larger values indicating possibly unsafe sharp turning
17	acc_h_max (m/s^2)	The maximum value of the vertical acceleration
18	acc_xy_max (m/s^2)	The maximum values of the acceleration in the horizontal plane, with the recommended acc_xy_max value for civil aviation pilots being 1 G
19	Roll (min & max values)	The minimum and maximum values of Roll angle
20	Pitch (min & max values)	The minimum and maximum values of Pitch angle
21	slide_length (m)	The distance the airplane travelled after landing until full stop; the upper limit for this value is often 1800 m at most airports

2.7.6. Machine Learning Modeling

After preprocessing, 19, 7, and 39 features existed in the AOI, EM, and QAR datasets. In Table 3, most rows contain two to three features, such as mean and SD, which adds up to 39 QAR features. This study combined three kinds of data (AOI, EM, and QAR) and then obtained seven combinations of the whole datasets: AOI, EM, QAR, AOI and EM, AOI and QAR, EM and QAR, AOI and EM and QAR. Those datasets all contained 45 participants. One participant was excluded from the machine learning analysis as he was the only instructor pilot whose expertise and age differed from other pilots. To illustrate the predictability of the newly proposed framework based on the SVM + MIC algorithm, comparative experiments were conducted based on the AOI and EM and QAR datasets with other algorithms. The whole machine learning algorithms included three processes: feature selection, training predictors, and evaluating predictors. First, the feature selection methods were used to filter redundant features. Feature selection proportion was set to 15% to 95% with a stride of 10%, which means that a certain proportion of relevant features were selected from the ranked features by the feature selection methods. For example, there are 70 rated features after feature selection in a dataset, then the top 7 features were selected with a feature selection proportion of 10%. Second, training predictors: selected features were fed into multiple predictors, and a leave-one-out cross validation strategy was applied. Lastly, evaluating predictors: all algorithms were implemented by Python 3.9.12 software. The hyper-parameters of predictors were default settings in the Scikit-learn 1.2.0 package and not adjusted. Additionally, ablation experiments were conducted on the seven datasets to explore the importance of different data sources.

3. Results

3.1. Flight Performance Results

An independent-sample *t*-test was performed on the flight QAR indicators. Results are shown in Table 4 and Figure 3. The difference between novice and expert was significant in the pitch angle 1 s before landing ($t(44) = 2.09$, $p < 0.05$, Cohen's $d = 0.62$), the mean pitch angle 0–10 s before landing ($t(44) = 2.31$, $p < 0.05$, Cohen's $d = 0.68$), the mean distance to the center of the two reference lines ($t(44) = 3.96$, $p < 0.01$, Cohen's $d = 1.17$), and the standard deviation of distance to the center of the two reference lines ($t(44) = 3.53$, $p < 0.01$, Cohen's $d = 1.04$). The standard deviation of pitch angle 0–10 s before landing is marginally significant ($t(44) = 1.96$, $p = 0.06$, Cohen's $d = 0.58$). Experts accomplished the whole flight task numerically quicker than novices, as the total flight time is marginally significant ($t(44) = 1.84$, $p = 0.07$, Cohen's $d = 0.54$).

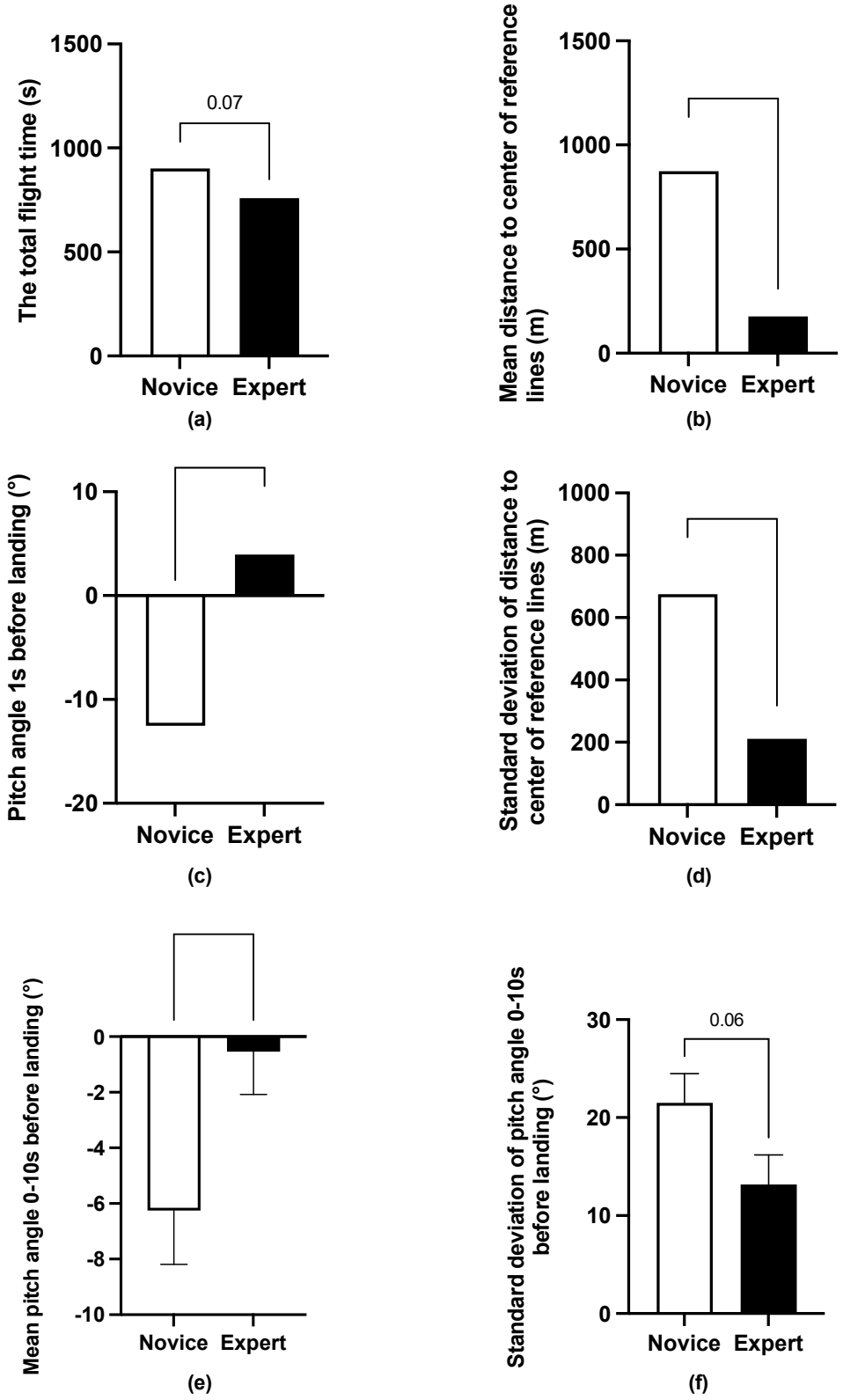

Figure 3. Means and data distribution of flight performance between novices and experts: (**a**) The total flight time (s); (**b**) Mean distance to center of reference lines (m); (**c**) Pitch angle 1 s before landing (°); (**d**) The standard deviation of distance to center of reference lines (m); (**e**) Mean pitch angle 0–10 s before landing (°); (**f**) Standard deviation of pitch angle 0–10 s before landing (°). Note: Error bar represents SEM, * represents $p < 0.05$, and *** represents $p < 0.001$.

Table 4. An independent-sample *t*-test results of the flight performance data.

Indicators	Novice	Expert	*t*	*p*	Cohen's *d*
	Mean (SD)	Mean (SD)			
The total flight time (s)	902.32 (336.73)	759.06 (163.58)	1.84	0.07	0.54
Pitch angle 1 s before landing (°)	−12.54 (29.03)	3.97 (24.43)	2.09	*	0.62
Mean pitch angle 0–10 s before landing (°)	−6.25 (9.28)	−0.54 (7.37)	2.31	*	0.68
Standard deviation of pitch angle 0–10 s before landing (°)	21.50 (14.31)	13.18(14.15)	1.96	0.06	0.58
Mean distance to center of reference lines (m)	873.89 (818.43)	176.67 (205.52)	3.96	***	1.17
Standard deviation of distance to center of reference lines (m)	675.78 (589.07)	211.52 (225.76)	3.53	***	1.04

Note: * represents $p < 0.05$, *** represents $p < 0.001$.

3.2. Eye Movement Analysis Results

As shown in Table 5 and Figure 4, experts spent a significantly higher percentage of time than novices on the key flying instruments (i.e., airspeed indicator, vertical speed indicator, and altitude indicator): the airspeed indicator ($t(44) = 2.11$, $p < 0.05$, Cohen's $d = 0.64$), the vertical speed indicator ($t(44) = 3.83$, $p < 0.001$, Cohen's $d = 1.15$), and the altitude indicator ($t(44) = 2.24$, $p < 0.05$, Cohen's $d = 0.68$). However, the difference among experts and novices was only marginally significant in the attitude indicator ($t(44) = 1.84$, $p = 0.07$, Cohen's $d = 0.56$).

Table 5. An independent-sample *t*-test results of the percent dwell time on each AOI (Unit: %).

Area of Interest (AOI)	Novice	Expert	*t*	*p*	Cohen's *d*
	Mean (SD)	Mean (SD)			
Airspeed indicator	5.32 (4.98)	8.56 (5.45)	2.11	*	0.64
Attitude indicator	31.03 (12.2)	25.15 (9.23)	1.84	0.07	0.56
Vertical speed indicator	5.33 (4.11)	13.11 (8.84)	3.83	***	1.15
Altitude indicator	1.34 (2.22)	2.83 (2.29)	2.24	*	0.68

Note: * represents $p < 0.05$ and *** represents $p < 0.001$.

3.3. Evaluation of Different Proportions of Selected Features

The strategy of selecting features according to the k highest scores was adopted during the feature selection process, where k denotes an integral part of the product of the feature selection proportion and the number of overall features. Due to the continuity of feature proportions, we cannot exhaustively enumerate all proportions. So, the proportion was set to 15% to 95% with a stride of 10% for evaluating how the prediction performance was influenced. The experiments were conducted on the dataset of AOI and EM and QAR by using nine feature proportions (from 15% to 95% with a stride of 10%), five predictors (SVM, KNN, LR, LGBM, and DTree), and three feature selection methods (*MIC*, SVM-RFE, and RF) above. Figure 5 shows each proportion's best *Acc*, *AUC*, *F*1, *Precision*, and *Recall* performance. 65% feature selection proportion outperformed the other eight proportions on *Acc*, *F*1, *AUC*, *Precision*, and *Recall* metrics. Thus, this study adopted the feature proportion of 65% as the best choice.

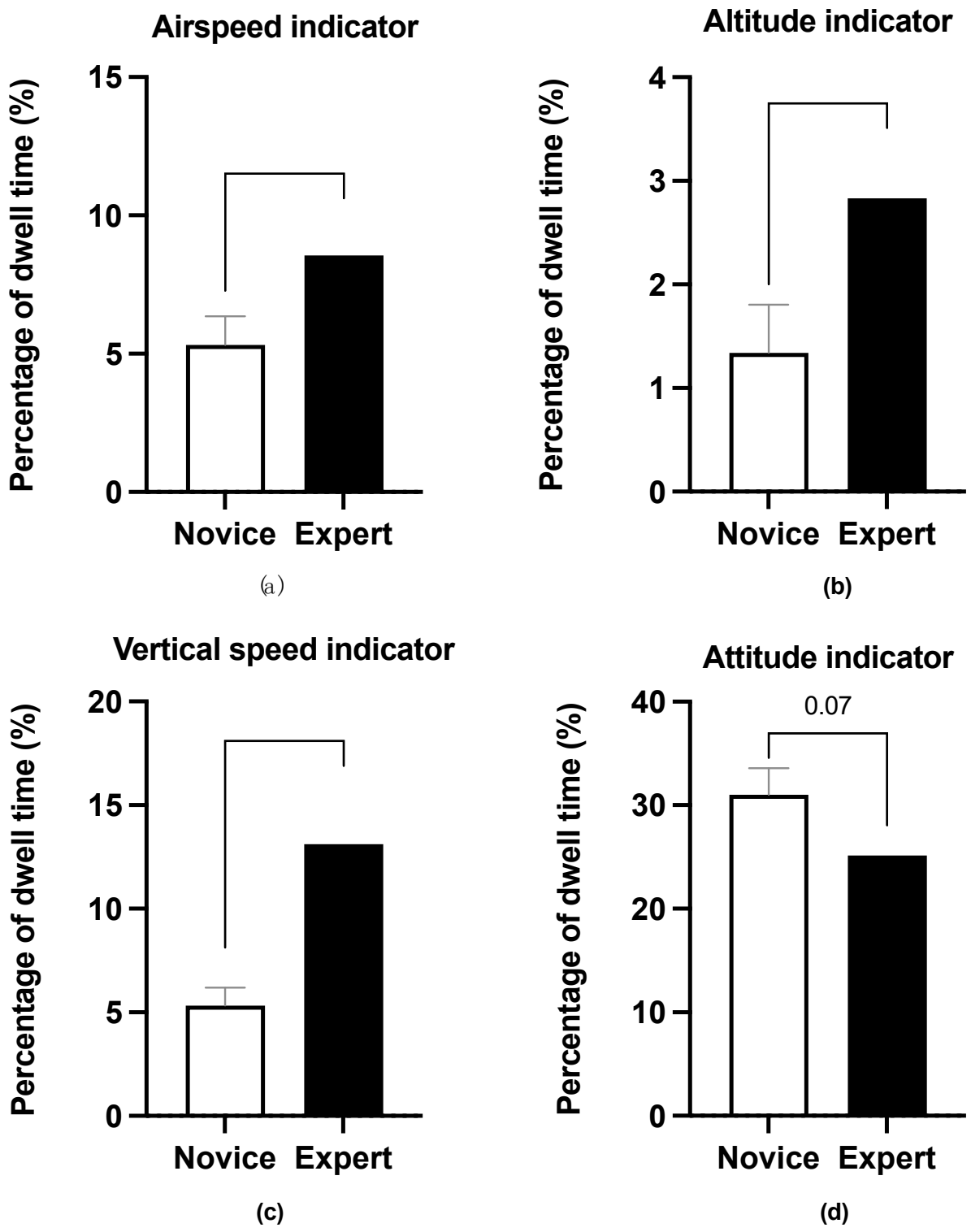

Figure 4. Means and data distribution of percent dwell time between novices and experts: (**a**) Airspeed indicator; (**b**) Altitude indicator; (**c**) Vertical speed indicator; (**d**) Attitude indicator. Note: Error bar represents SEM, * represents $p < 0.05$, and *** represents $p < 0.001$.

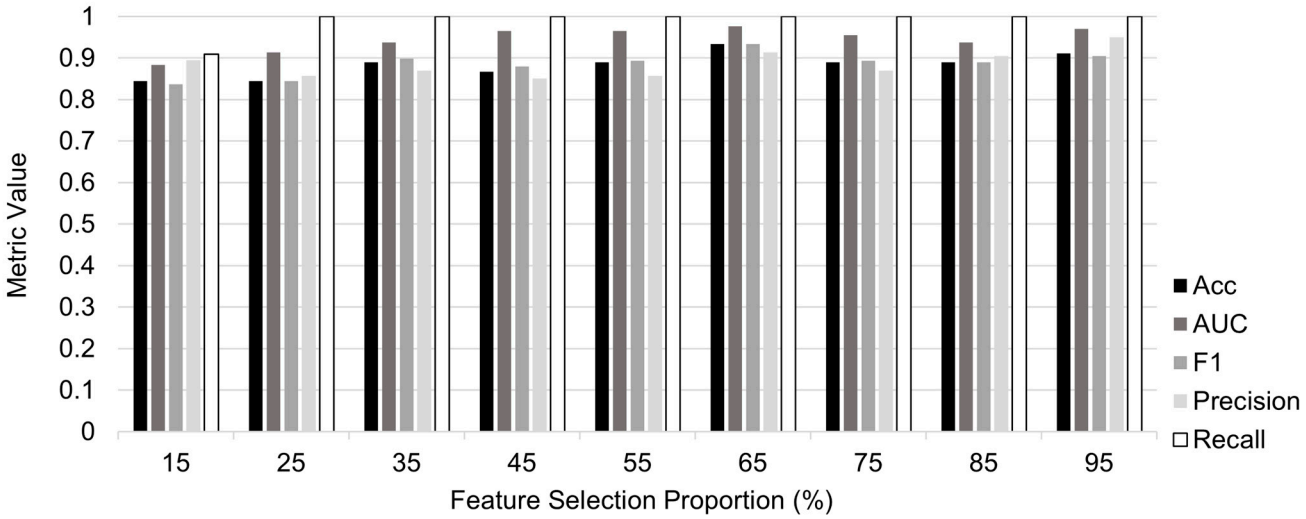

Figure 5. Evaluation of different proportions of the features selected on AOI & EM & QAR datasets.

3.4. Performance Evaluation of Predictors and Feature Selection Methods

This section evaluated 15 models with the combinations of three feature selection methods (*MIC*, SVM-RFE, and RF) and five predictors (SVM, KNN, LR, LGBM, and DTree), as shown in Figure 6. The results indicate that an SVM predictor with the *MIC* feature selection method generally achieved the highest performance on all metrics with an *Acc* of 0.9333, an *AUC* of 0.9644, and an *F*1 of 0.9333. From the perspective of feature selection methods, the *MIC* algorithm is the most suitable feature selection method for the pilot selection prediction task. The *MIC* algorithm can select features with a nonlinear relationship to the sample labels instead of filtering out these features. Additionally, SVM performed better robustness in *Acc*, *AUC*, and *F*1 metrics, and DTree with an *Acc* of 0.8889, an *AUC* of 0.8893, and an *F*1 of 0.8889 when using the *MIC* feature selection method. SVM and DTree both showed outstanding discrimination for pilot selection.

3.5. Ablation Experiments on Datasets

To investigate the impact of different data sources on the performance of pilot selection prediction tasks. Ablation experiments on the seven datasets were conducted using the SVM + MIC algorithm, which has been proven as the best algorithm for pilot selection as described in the above paragraphs. As Figure 7 shows, the EM & QAR and AOI & EM & QAR datasets achieved *AUC* values of 96.64% and 96.44%, respectively, and the AOI dataset got the worst *AUC* value of 82.41%. Table 6 summarizes the prediction performance on the seven datasets. The AOI & EM & QAR dataset obtained the best performance with an *Acc* of 0.9333, an *F*1 of 0.933, a *Precision* of 0.9130, and a *Recall* of 0.9545. The AOI dataset obtained the worst performance on all five metrics.

Table 6. Prediction performance on ablation experiments for the SVM + MIC algorithm.

Dataset	*Acc*	*AUC*	*F1*	*Precision*	*Recall*
AOI	0.7333	0.8597	0.7000	0.7778	0.6364
EM	0.8222	0.8933	0.8400	0.7500	0.9545
QAR	0.8444	0.8874	0.8444	0.8261	0.8636
AOI & EM	0.8667	0.9447	0.8696	0.8333	0.9091
AOI & QAR	0.7556	0.8241	0.7317	0.7895	0.6818
EM & QAR	0.8667	0.9664	0.8696	0.8333	0.9091
AOI & EM & QAR	0.9333	0.9644	0.9333	0.9130	0.9545

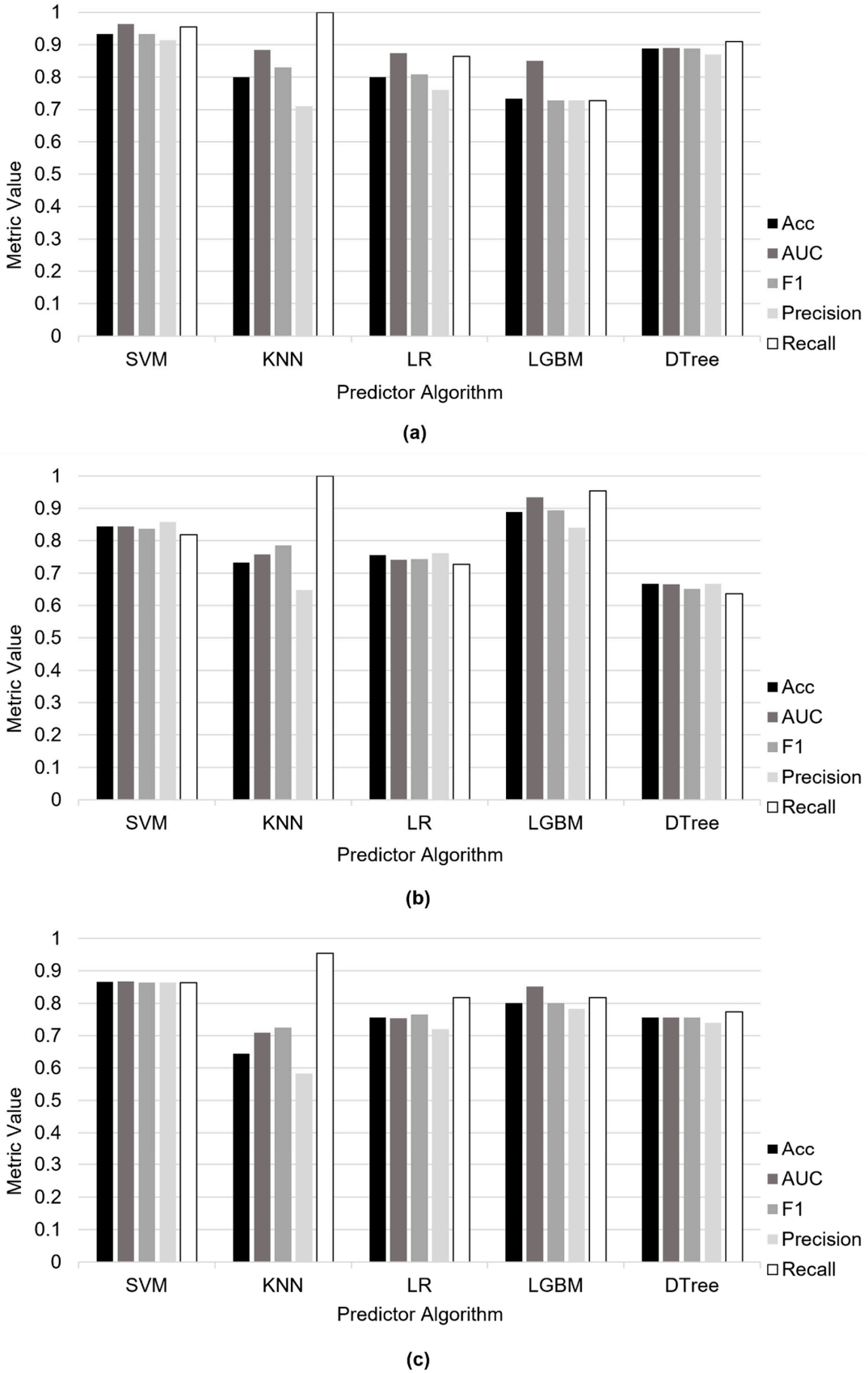

Figure 6. The performance comparison between feature selection methods and predictor algorithm: (**a**) *MIC*, (**b**) SVM-RFE, and (**c**) RF.

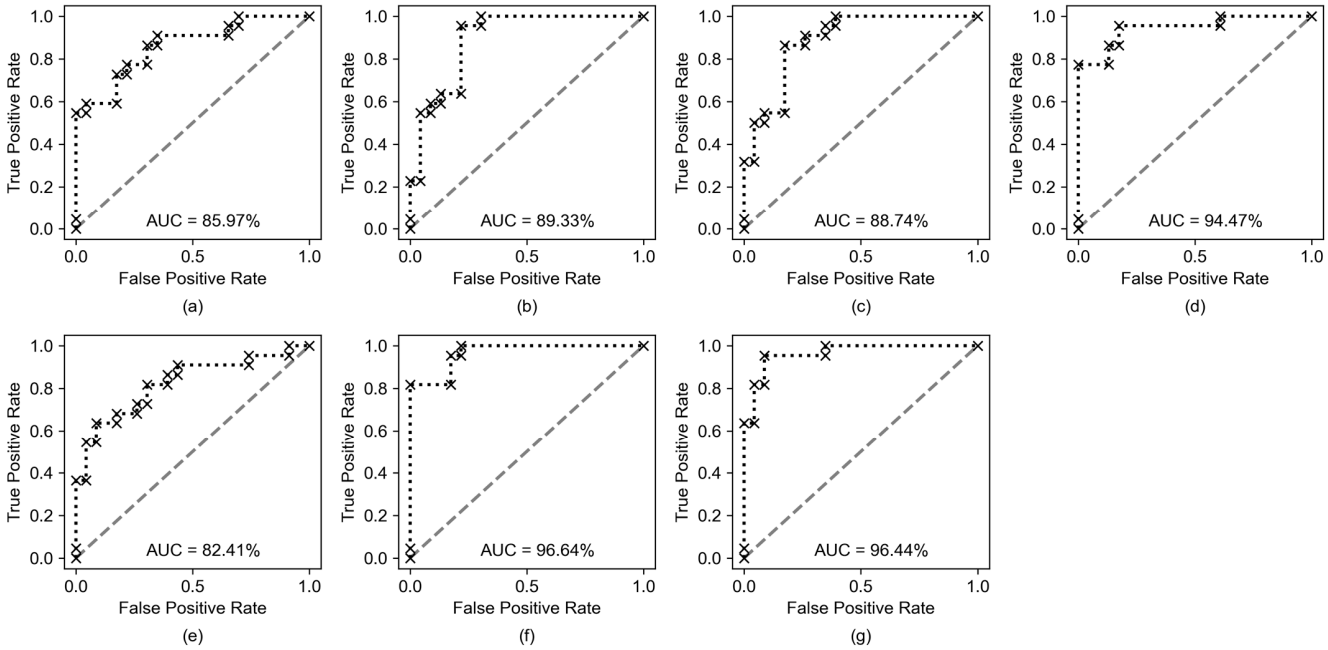

Figure 7. The ROC curves of the SVM + MIC algorithm on the seven datasets is: (**a**) AOI, (**b**) EM, QAR, (**c**) AOI & EM, (**d**) AOI & QAR, (**e**) EM & QAR, (**f**) AOI & EM & QAR, and (**g**) AOI & EM& QAR, where AOI, EM, and QAR represent the area of interest dataset, the eye movement dataset, and the quick access recorder dataset, respectively.

3.6. Interpretable Model Results Based Decision Tree (DTree)

Most of the above-mentioned models (Including SVM, KNN, and LGBM) are not interpretable. Although the SVM + MIC algorithm generated the best accuracy of 0.9333, the DTree model was further analyzed to provide an interpretable model and results for the difference between novice and expert pilots. The DTree had an *Acc* of 0.8889, an *AUC* of 0.8893, and an *F*1 of 0.8889 when using the *MIC* feature selection method. Figure 8 shows the visualization results of DTree. Thus, when the interpretability of models is emphasized, we need to switch from the best performing SVM + MIC algorithm to DTree and trade $0.9333 - 0.8889 = 0.0444$ accuracy for interpretability.

The variables included in the final DTree model include: (1) Percent Dwell Time on Altitude Indicator; (2) Rudder Input 8 s before landing; (3) Ground Speed 1 s before landing; (4) Elevator Input 8 s before landing, and (5) Saccade Count. Figure 9 further depicts the relative contributions of these five variables to the DTree model. Percent Dwell Time on Altitude Indicator contributed most to the DTree model, followed by Ground Speed 1 s before landing, Rudder Input 8 s before landing, next Saccade Count, and Elevator Input 8 s before landing.

An interpretable model is thereby that most expert pilots are the ones who: (1) use the Altitude Indicator frequently (larger than 0.013 of the total time); (2) can maintain Ground Speed 1 s before landing; (3) have smaller than 0.4 elevator inputs 8 s before landing; (4) have more Saccade Counts. Conversely, most novices are the ones who: (1) use the Altitude Indicator less frequently and (2) have a smaller Rudder Input 8 s before landing.

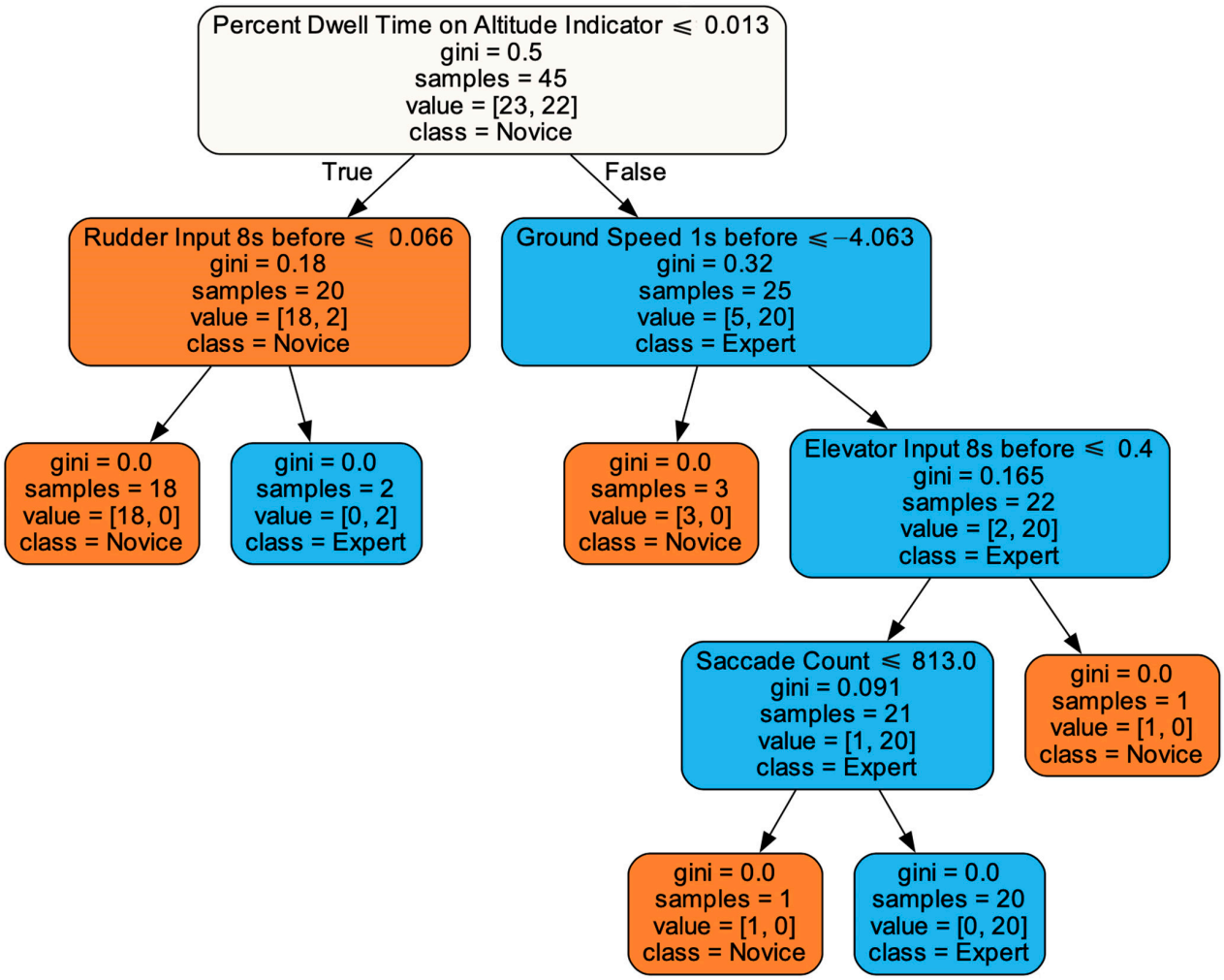

Figure 8. Visualization of Decision Tree classification results. Blue blocks and orange blocks stand for experts and novices, respectively.

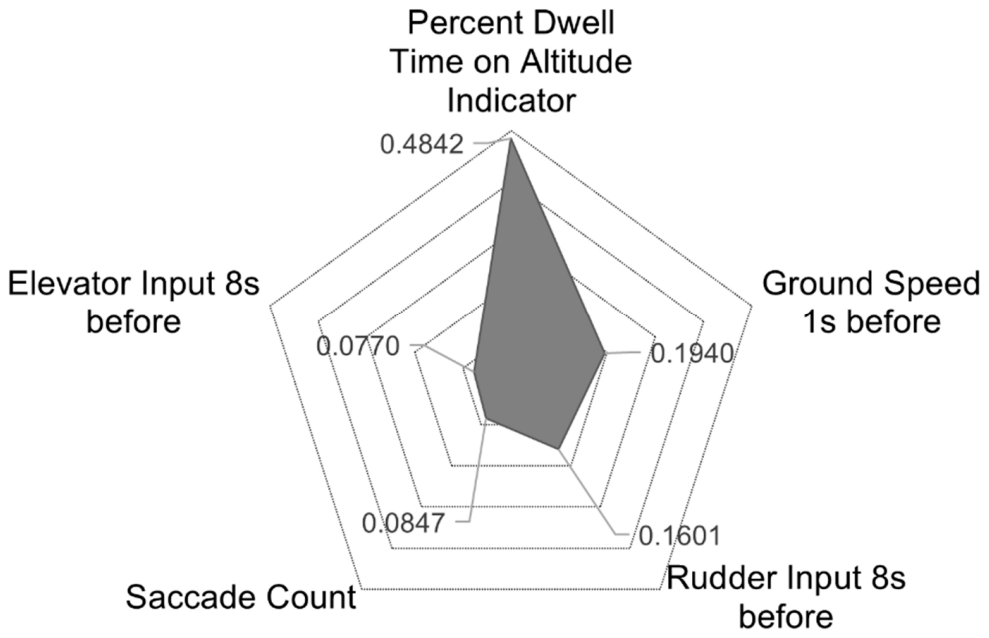

Figure 9. Relative Contribution of Variables in the Decision Tree model.

4. Discussion

To sum up, expert pilots differed from novices in at least three aspects for the behavioral results. Firstly, for flight dynamics, the pilots' actual flight path was nearer to the center of the two reference lines than the novices. Secondly, pilots had a distinctly different eye-moving pattern from novices. During flying, pilots relied heavily on the critical instrumentations of flight, such as the airspeed indicator, the vertical speed indicator, and the altitude indicator. Thirdly, pilots had a longer fixation time on the instruments and exhibited a more structured and efficient eye-scanning pattern. The eye movement results showed that experts had more efficient eye-tracking models [20].

For the machine learning results, our new SVM + MIC algorithm achieved a high classification of 93.33% in pilot selection. The specific training dataset formed decision functions of SVM. Put another way, SVM can maximize the margin between the decision borders from the dataset in a Euclidean space, making SVM more generalizable, obtaining better robustness, and producing less train error, especially when using a small dataset than other predictors [54]. Moreover, SVM can maximally mine the data's latent knowledge, making small sample prediction tasks possible [55]. One of the most significant advantages of SVM is fitting nonlinear and high dimensional data better when using nonlinear kernel functions. Mutual information can measure various relationships between random variables, including linear or nonlinear relationships [56]. In addition, redundant and irrelevant features, such as data noise, decrease the performance of the SVM classifier. These are two main reasons our combined SVM + MIC algorithm outperformed all evaluated models [57].

The contribution of input features to the predictability of pilot selection was explored by feeding each input feature to machine learning algorithms independently or in various combinations. Area of Interest (AOI) dependent analysis was separately analyzed from other measurements of eye movements (labeled as "EM" in Figure 7, including fixation duration, fixation dispersion, and saccade frequency, etc.), as AOI analysis using traditional glass-like eye-tracker and desktop eye tracker often needs time-consuming manual creation and coding of AOIs. Figure 7 and Table 6 show that information on AOI-related measurements contributed to an additional $0.9333 - 0.8667 = 0.0666$ accuracy gained by comparing "QAR & AOI & EM vs. QAR & EM". To retain this accuracy achieved from AOI information, it is essential to use a VR-based eye tracker like HTC Eye Pro instead of a traditional glass-like or desktop eye tracker. In a VR scenario, all objects are in digital forms, and it is thus easy and accurate to know the location and semantic meaning of an object or an AOI. However, traditional eye trackers can only tell researchers where users are looking but cannot tell "what" users are looking at, as current computer vision algorithms cannot recognize all objects' names in videos. In addition, a VR eye tracker (around $1800) costs much less than traditional eye trackers (around $20,000 to $30,000). Thus, considering time and financial costs, many pilot selections should use a VR eye tracker, which was most likely first implemented and investigated in the current study, instead of traditional eye trackers like GazePoint [21] or SMI ETG Spectacle eye tracker [33]. The time cost and the labor to conduct an AOI analysis were perhaps one of the reasons why the two studies undertook only a generic standard eye movement analysis without AOI information [21,33]. Nevertheless, the two studies were the only pioneer former studies we have identified that attempted to use eye movement to select pilots. More research efforts are needed for future studies to explore better algorithms to select pilots using eye movement, especially using AOI information and a VR eye tracker.

Existing flight simulator owners may need other information features as inputs for machine learning, such as QAR flight dynamics, as eye trackers are expensive and, more importantly, impracticable to set a space for a VR eye tracker in their over 100,000,000 USD advanced flight simulator with 6 degrees of freedom (DoF). Our current work provides a good alternative: QAR flight dynamics for enterprise-level users who cannot choose an eye tracker. As shown in Figure 7 and Table 6, QAR outperforms generic eye movements (see "EM" tick label) and AOI information (see "AOI" tick label). QAR alone can predict pilots

with an accuracy of 0.8444, as high as most of the accuracy achievements summarized in Tables 1 and 6.

Although the combination of eye tracking and flight dynamics generates promising discriminability of novices from expert pilots, and these two measurements should be the core of flight behaviors and performances, they are not the complete descriptors of candidate pilots under examination. Other modality variables, such as heart rate, skin conductance, and dry EEG sensors, should be implemented and evaluated in future studies for their predictability of pilot selection success to provide a platform for selection with all-around modalities.

We are fortunate to locate at least six publications that use machine learning to predict pilot selection success, encouraging our work towards the multimodality + machine-learning approach. For example, a recent publication in 2021 by the United States Air Force Academy claimed that an "extremely randomized tree machine learning technique can achieve nearly 94% accuracy in predicting candidate success" based on 8-year data from the historical specialized undergraduate pilot training (SUPT) program [31]. Their long-term work is possibly a milestone that summarizes personnel selection using paper-and-pencil and purely cognitive task performances [31]. It signifies the critical value of machine learning algorithms in pilot selection. The United States Air Force Academy researchers can never be expected and constrained unless the constraint is from historical and technological perspectives. Features with merely cognitive tasks face the limitation of ecological validity and practice effects. For example, the selection of US Air Force (USAF) pilots relied heavily on officer metrics [58] and used the Pilot Candidate Selection Method (PCSM) [59]. Recent advances in Virtual Reality technology with embedded eye-tracking modules can remove the technological constraint for our pioneer researchers [31]. The current work follows their machine learning approach but with the adoption of new technology and achieves similar predictability of nearly 94% using multimodality data of eye tracking and flight dynamics.

However, unfortunately, we cannot find papers using machine learning algorithms to select pilots based on data like heart rate [33] and skin conductance after thorough searches in databases like Web of Science, ProQuest, Google Scholar, Baidu Scholar, etc. Thus, it is highly likely that the machine-learning-based multimodality approach for pilot selection still needs to be explored. No attempts have used machine learning algorithms, heart rate, and skin conductance data to select pilots. Thus, our efforts were in vain to locate other similar papers using the multimodality + machine-learning approach, which may suggest that the current study is one of the pioneering works and an early adopter of HTC Eye Pro Virtual Reality to select pilots using robust machine-learning algorithms and with a more economical and practical platform. Future studies may consider additional machine learning algorithms, especially recent advances in deep learning, such as CNN, RNN, and Transformer algorithms, etc. However, this study has extensively compared as many as five algorithms, including SVM, KNN, LR, LGBM, and Decision Tree. More complex neural network algorithms were not explored in this current study as accuracy alone is not the only indicator goal for our research; interpretability is another, if not the most important, criterion for pilot selection. More complex algorithms can be attempted with our dataset accumulating year by year. In addition, the pilot selection is not a field similar to "hardware with seldom changes" but more like software that warrants a periodic update. With yearly new data for pilot selection, relatively new concepts in machine learning like "Active Learning" [60] should be considered in future algorithms to update the models yearly to reach an overarching goal of "Faster, Higher, Stronger" algorithms for pilot selection.

5. Conclusions

This study contributed to pilot selection in at least three aspects. First, our SVM + MIC algorithm achieves a high predictability of 93.33% accuracy, outperforming most existing SVM and logistic regression models in the literature [32,33]. Second, the Decision Tree model with an 88.89% accuracy shows an interpretable finding: novice pilots who use the altitude indicator less frequently have smaller Rudder Input 8 s before landing. Third,

Virtual Reality with an embedded eye tracker and the possibility of automated analysis of areas of interest can provide a low-cost, portable, and efficient platform for pilot selection. Furthermore, our research focuses on machine learning to identify differences in expertise between novices and experts. Therefore, except for pilot selection, it also has potential application value for pilot training and personnel selection in other fields, such as astronauts.

6. Limitations and Future Study

Our algorithms can distinguish expertise differences between novices and experts, and these differences can be used to improve training. More importantly, expertise differences may also encompass partially stable features that are not trainable or not easily trainable and can be used for pilot selection. According to the research from Dr. Carretta (a well-known researcher in aviation psychology), traditional pilot selection relies heavily on cognitive ability testing via paper-and-pencils or computers [58,59]. This approach is valid, and cognitive ability is a valid predictor of job performance [30,61]. Some countries have also attempted to conduct pilot selection through simulators that have achieved comparable predictive results to traditional cognitive ability tests [62]. However, the simulator-based pilot selection is too costly and complicated to set up to reach a large area for replication [62]. As a result, we developed a virtual reality-based flight simulator to achieve low cost and a simple setup to solve these problems. At the same time, we used machine learning to distinguish expertise differences to improve pilot training and selection. Given our research limitation, future studies must distinguish these expertise differences between easily trainable or training effect, hard to train, and unfeasible to train. Specifically, the latter two features are hard to use and unfeasible for training and should be used for pilot selection.

This study is limited in several ways. First, reference lines were provided to provide flight directions to cater to the inexperience of novices who have no prior knowledge of traffic pattern. The design also allows the calculations of two crucial performance indicators, mean distance to the center of the two reference lines (in meters) and standard deviation of distance to the center of the two reference lines (in meters). However, this design is limited for pilots as the reference lines may constrain or modify pilots' habitual flight traffic pattern. Pilots with good knowledge and tracking performance can still outperform novices with the existence of the two reference lines, but their actual flight performance may be imperfectly measured. Future studies may consider other options to use markers at turning points in the traffic pattern [63] or L shape indicator in the segmented circle to indicate traffic patterns for pilots.

Another limitation of this study is that the participants were all males. Regardless of this limitation, this study is still scientifically sound and representative of most pilots for several reasons. First, there are currently no women pilots flying internation-ally in China Eastern Airlines and Shanghai Airlines from 2021 to 2022. Second, we used the systematic sampling method to recruit participants [64] and ended up with all male participants because of the large proportion of male pilots. Third, women pilots represent only 1.05% of the total pilots in China Eastern Airline Corporation and 1.27% in Shanghai Airline, 3.3% in Turkey [65], 5.4% in Canada and the U.S., and 5.7% in British Airways in 2013 [66]. Future studies should consider recruiting more female participants and pilots in their studies. The aviation industry should also consider attracting and accepting more women pilots to build an inclusive and diversified work culture for both genders.

Author Contributions: Conceptualization, J.H. and X.L.; methodology, L.K. and G.Z.; software, G.Z. and J.H.; validation, G.Z. and Y.L. (Yajing Li); formal analysis, L.K., G.Z. and P.F.; investigation, L.K. and Y.L. (Yan Li); resources, J.H., Y.L. (Yan Li) and X.L.; data curation, L.K. and Y.L. (Yajing Li); writing—original draft preparation, L.K., J.H., G.Z. and Y.L. (Yajing Li); writing—review and editing, X.L., P.F., J.H. and L.K.; visualization, L.K. and G.Z.; supervision, J.H. and X.L.; project administration, J.H. and X.L.; funding acquisition, J.H. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is sponsored by the National Key R & D Program of China (2022YFB4500600), the National Natural Science Foundation of China (Grant No. T2192933), the National Key Laboratory Project of Human Factors Engineering (grant number 6142222210201), Aviation Safety and Security Association (Number ASSA2022234) to J.H.; "The year 2022 Major Projects of Military Logistic Research Grant", and "Key Project of Air Force Equipment Comprehensive Research, Grant Number KJ2022A000415" to X.L. Science and Technology Commission of the Military Commission National Defense Science and Technology Innovation Special Zone project [Grant number: 224-CXCY-M113-07-01-01(04)].

Data Availability Statement: Restrictions apply to the availability of these data. Data were collected under sponsorship from pilots in an airline company with privacy concerns, and thus were not publicly available. The data may be available upon request for academic purposes and with permission from our sponsors.

Acknowledgments: We acknowledge the aviation knowledge from general aviation pilot Jie Liu and civil aviation pilot Lian Duan from China Eastern Airline. We also appreciate the programming help from Xin Meng and Yongtao Du. Research Assistant Wenqing Zhang also contributed to the data collection.

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

References

1. Zheng, Z.; Gao, S.; Su, Y.; Chen, Y.; Wang, X. Cognitive load-induced pupil dilation reflects potential flight ability. *Curr. Psychol.* **2022**, *1*, 11. [CrossRef]
2. Siem, F.M.; Carretta, T.R.; Mercatante, T.A. *Personality, Attitudes, and Pilot Training Performance: Preliminary Analysis*; Air Force Human Resources Lab., Brooks Air Force Base: San Antonio, TX, USA, 1988; Volume 1, p. 32.
3. Byrnes, J.F.; Cascio, W. Costing human resources: The financial impact of behavior in organizations. *Acad. Manag. Rev.* **1984**, *9*, 370–371. [CrossRef]
4. Lutte, R.; Lovelace, K. Airline pilot supply in the US: Factors influencing the collegiate pilot pipeline. *J. Aviat. Technol. Eng.* **2016**, *6*, 8. [CrossRef]
5. Carretta, T.R.; Ree, M.J. Pilot-candidate selection method: Sources of validity. *Int. J. Aviat. Psychol.* **1994**, *4*, 103–117. [CrossRef]
6. Martinussen, M.; Torjussen, T. Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. *Int. J. Aviat. Psychol.* **1998**, *8*, 33–45. [CrossRef]
7. Damos, D.L. Using meta-analysis to compare the predictive validity of single-and multiple-task measures to flight performance. *Hum. Factors* **1993**, *35*, 615–628. [CrossRef]
8. Dang, J.; King, K.M.; Inzlicht, M. Why are self-report and behavioral measures weakly correlated? *Trends Cogn. Sci.* **2020**, *24*, 267–269. [CrossRef]
9. ALMamari, K.; Traynor, A. Multiple test batteries as predictors for pilot performance: A meta-analytic investigation. *Int. J. Sel. Assess.* **2019**, *27*, 337–356. [CrossRef]
10. Robinson, A.; Mania, K.; Perey, P. Flight simulation: Research challenges and user assessments of fidelity. In Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry, Singapore, 16–18 June 2004; pp. 261–268.
11. Goretzko, D.; Israel, L.S.F. Pitfalls of machine learning-based personnel selection: Fairness, transparency, and data quality. *J. Pers. Psychol.* **2022**, *21*, 37–47. [CrossRef]
12. Yarkoni, T.; Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **2017**, *12*, 1100–1122. [CrossRef]
13. Phillips, P.; Hahn, C.; Fontana, P.; Yates, A.; Greene, K.; Broniatowski, D.; Przybocki, M. *Four Principles of Explainable Artificial Intelligence, NIST Interagency/Internal Report (NISTIR)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399 (accessed on 20 January 2023).
14. Peißl, S.; Wickens, C.D.; Baruah, R. Eye-tracking measures in aviation: A selective literature review. *Int. J. Aerosp. Psychol.* **2018**, *28*, 98–112. [CrossRef]
15. Ziv, G. Gaze behavior and visual attention: A review of eye tracking studies in aviation. *Int. J. Aviat. Psychol.* **2016**, *26*, 75–104. [CrossRef]
16. Lai, M.; Tsai, M.; Yang, F.; Hsu, C.; Liu, T.; Lee, S.; Lee, M.; Chiou, G.; Liang, J.; Tsai, C. A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educ. Res. Rev.* **2013**, *10*, 90–115. [CrossRef]
17. Jin, H.; Hu, Z.; Li, K.; Chu, M.; Zou, G.; Yu, G.; Zhang, J. Study on how expert and novice pilots can distribute their visual attention to improve flight performance. *IEEE Access* **2021**, *9*, 44757–44769. [CrossRef]
18. Yu, C.S.; Wang, E.M.Y.; Li, W.C.; Braithwaite, G.; Greaves, M. Pilots' visual scan patterns and attention distribution during the pursuit of a dynamic target. *Aerosp. Med. Hum. Perform.* **2016**, *87*, 40–47. [CrossRef]

19. Bellenkes, A.H.; Wickens, C.; Kramer, A. Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviat. Space Environ Med.* **1997**, *68*, 569–579.
20. Kasarskis, P.; Stehwien, J.; Hickox, J.; Aretz, A.; Wickens, C. Comparison of expert and novice scan behaviors during VFR flight. In Proceedings of the 11th International Symposium on Aviation Psychology, Columbus, OH, USA, 5–8 March 2001; Volume 6, pp. 1–6.
21. Vlačić, S.; Knežević, A.; Rođenkov, S.; Mandal, S.; Vitsas, P.A. Improving the pilot selection process by using eye-tracking tools. *J. Eye Mov. Res.* **2020**, *12*, 10–16910. [CrossRef]
22. Wang, L.; Ren, Y.; Sun, H.; Dong, C. A landing operation performance evaluation system based on flight data. In Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics, Vancouver, BC, Canada, 9–14 July 2017; pp. 297–305.
23. Le Ngoc, L.; Kalawsky, R.S. Visual circuit flying with augmented head-tracking on limited field of view flight training devices. In Proceedings of the AIAA Modeling and Simulation Technologies (MST) Conference, Grapevine, TX, USA, 9–13 January 2013; p. 5226.
24. Oberhauser, M.; Dreyer, D.; Braunstingl, R.; Koglbauer, I. What's real about virtual reality flight simulation? Comparing the fidelity of a virtual reality with a conventional flight simulation environment. *Aviat. Psychol. Appl. Hum. Factors* **2018**, *8*, 22–34. [CrossRef]
25. Wang, L.; Ren, Y.; Wu, C. Effects of flare operation on landing safety: A study based on ANOVA of real flight data. *Saf. Sci.* **2018**, *102*, 14–25. [CrossRef]
26. Sun, J.; Xiao, X.; Cheng, S.; Shen, C.; Ma, J.; Hu, W. Study on pilot personality selection with an SVM-Based classifier. In Proceedings of theInternational Conference on Man-Machine-Environment System Engineering, Nanjing, China, 20–22 October 2018; Springer: Singapore, 2018; pp. 3–10.
27. Xiao, X.; Cheng, S.; Sun, J.; Ma, J.; Hu, W. A Pilot Study of Assessment of Civilian Pilot Personality Based on Support Vector Machine. *Prog. Mod. Biomed.* **2017**, *17*, 111–114.
28. Lochner, K.; Nienhaus, N. The Predictive Power of Assessment for Pilot Selection. Technical Report Published by Cut-e Group. Retrieved 15 January 2023. 2016. Available online: https://www.cute.com/fileadmin/user_upload/Assessing_in_Aviation/White_Paper_Pilot_5238.pdf (accessed on 10 March 2023).
29. Hunter, D.R.; Burke, E.F. Predicting aircraft pilot-training success: A meta-analysis of published research. *Int. J. Aviat. Psychol.* **1994**, *4*, 297–313. [CrossRef]
30. Schmidt, F.L.; Hunter, J.E. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychol. Bull.* **1998**, *124*, 262–274. [CrossRef]
31. Jenkins, P.R.; Caballero, W.N.; Hill, R.R. Predicting success in United States Air Force pilot training using machine learning techniques. *Socio-Econ. Plan. Sci.* **2022**, *79*, 101121. [CrossRef]
32. Maroco, J.; Bártolo-Ribeiro, R. Selection of air force pilot candidates: A case study on the predictive accuracy of discriminant analysis, logistic regression, and four neural network types. *Int. J. Aviat. Psychol.* **2013**, *23*, 130–152. [CrossRef]
33. Wang, X.; Gong, G.; Li, N.; Ding, L.; Ma, Y. Decoding pilot behavior consciousness of EEG, ECG, eye movements via an SVM machine learning model. *Int. J. Model. Simul. Sci. Comput.* **2020**, *11*, 2050028. [CrossRef]
34. Olsen, E.C.B.; Lee, S.E.; Simons-Morton, B.G. Eye movement patterns for novice teen drivers: Does 6 months of driving experience make a difference? *Transp. Res. Rec.* **2007**, *2009*, 8–14. [CrossRef]
35. Liang, Y.; Lee, J.D. A hybrid Bayesian Network approach to detect driver cognitive distraction. *Transp. Res. Part C Emerg. Technol.* **2014**, *38*, 146–155. [CrossRef]
36. Huang, M.W.; Chen, C.W.; Lin, W.C.; Ke, S.W.; Tsai, C.F. SVM and SVM ensembles in breast cancer prediction. *PLoS ONE* **2017**, *12*, e0161501. [CrossRef]
37. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 1–45. [CrossRef]
38. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1060–1073. [CrossRef]
39. Nadimi-Shahraki, M.H.; Zamani, H.; Mirjalili, S. Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study. *Comput. Biol. Med.* **2022**, *148*, 105858. [CrossRef] [PubMed]
40. Sheikhpour, R.; Sarram, M.A.; Gharaghani, S.; Chahooki, M.A.Z. A Survey on semi-supervised feature selection methods. *Pattern Recognit.* **2017**, *100*, 141–158. [CrossRef]
41. Liu, C.; Wang, W.; Zhao, Q.; Shen, X.; Konan, M. A new feature selection method based on a validity index of feature subset. *Pattern Recognit. Lett.* **2017**, *100*, 1–8. [CrossRef]
42. Wang, W.; Han, R.; Zhang, M.; Wang, Y.; Wang, T.; Wang, Y.; Shang, X.; Peng, J. A network-based method for brain disease gene prediction by integrating brain connectome and molecular network. *Brief. Bioinform.* **2022**, *23*, bbab459. [CrossRef]
43. Wolff, J.; Backofen, R.; Gruening, B. Robust and efficient single-cell Hi-C clustering with approximate k-nearest neighbor graphs. *Bioinformatics* **2021**, *37*, 4006–4013. [CrossRef]
44. Dong, C.; Qiao, Y.; Shang, C.; Liao, X.; Yuan, X.; Cheng, Q.; Li, Y.; Zhang, J.; Wang, Y.; Chen, Y.; et al. Non-contact screening system based for COVID-19 on XGBoost and logistic regression. *Comput. Biol. Med.* **2022**, *141*, 105003. [CrossRef]

45. Wei, J.; Li, Z.; Pinker, R.T.; Wang, J.; Sun, L.; Xue, W.; Li, R.; Cribb, M. Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). *Atmos. Chem. Phys.* **2021**, *21*, 7863–7880. [CrossRef]
46. Ghiasi, M.M.; Zendehboudi, S. Application of decision tree-based ensemble learning in the classification of breast cancer. *Comput. Biol. Med.* **2020**, *128*, 104089. [CrossRef]
47. Rafalo, M. Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis. *ICT Express* **2022**, *8*, 183–188. [CrossRef]
48. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846. [CrossRef]
49. Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Theoretical aspects. *J. Chemom.* **2012**, *26*, 361–373. [CrossRef]
50. Liang, X.W.; Jiang, A.P.; Li, T.; Xue, Y.Y.; Wang, G.T. LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM. *Knowl.-Based Syst.* **2020**, *196*, 105845. [CrossRef]
51. Chua, Z.K.; Feigh, K.M. Quantifying pilot performance during landing point redesignation for system design. *J. Aerosp. Comput. Inf. Commun.* **2011**, *8*, 183–196. [CrossRef]
52. Tanveer, A. GitHub-Tanvcodes/Qar_Analytics: Scripts for Working with Publicly Available Quick Access Recorder (QAR) Data from a Fleet of 35 BAE-146 Aircraft. Retrieved on 2 March 2023. 2022. Available online: https://github.com/tanvcodes/qar_analytics (accessed on 10 March 2023).
53. Weber, L. *International Civil Aviation Organization (ICAO)*; Kluwer Law International BV: Alphen aan den Rijn, The Netherlands, 2021.
54. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
55. Yin, Z.; Hou, J. Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. *Neurocomputing* **2016**, *174*, 643–650. [CrossRef]
56. Vergara, J.R.; Estevez, P. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
57. Bolon-Canedo, V.; Sanchez-Marono, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]
58. Carretta, T.U.S. Air Force pilot selection and training methods. *Aviat. Space Environ. Med.* **2000**, *71*, 950–956.
59. Carretta, T.R. Pilot Candidate Selection Method. *Aviat. Psychol. Appl. Hum. Factors* **2011**, *1*, 3–8. [CrossRef]
60. Bhargava, B.C.; Deshmukh, A.; Narasimhadhan, A.V. Modulation and signal class labelling with active learning and classification using machine learning. In Proceedings of the 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 8–10 July 2022; pp. 1–5.
61. Hunter, J.E.; Hunter, R. Validity and utility of alternative predictors of job performance. *Psychol. Bull.* **1984**, *96*, 72–98. [CrossRef]
62. Carretta, T.R.; Ree, M.J. Pilot Selection Methods. In *Principles and Practice of Aviation Psychology*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2003; pp. 357–396.
63. Rafi, M.; Chandrasekaran, B.; Kusmez, B.; Steck, J.E.; He, J. Real-time Google glass heads-up display for rapid air-traffic detection. *J. Aircr.* **2017**, *55*, 263–274. [CrossRef]
64. McCombes, S. Sampling Methods | Types. Techniques & Examples. Scribbr. 1 December 2022. Available online: https://www.scribbr.com/methodology/sampling-methods/ (accessed on 10 March 2023).
65. Yanıkoğlu, Ö.; Kılıç, S.; Küçükönal, H. Gender in the cockpit: Challenges faced by female airline pilots. *J. Air Transp. Manag.* **2020**, *86*, 101823. [CrossRef]
66. Pilot Institute. Women Pilot Statistics: Female Representation in Aviation. 7 October 2022. Available online: https://pilotinstitute.com/women-aviation-statistics/ (accessed on 10 March 2023).