applied sciences

MDPI

Article

A Real-Time Detection of Pilot Workload Using Low-Interference Devices

Yihan Liu [1,†], Yijing Gao [1,†], Lishengsa Yue [1,*], Hua Zhang [2], Jiahang Sun [1] and Xuerui Wu [1]

[1] Key Laboratory of Road and Traffic Engineering of Ministry of Education, Department of Transportation Engineering, Tongji University, Shanghai 201804, China; liuyihan1118@tongji.edu.cn (Y.L.); 2153330@tongji.edu.cn (Y.G.); 2153337@tongji.edu.cn (J.S.); 2152764@tongji.edu.cn (X.W.)
[2] Urban Mobility Institute, Tongji University, Shanghai 201804, China; xiaohai_hua@tongji.edu.cn
* Correspondence: 2014yuelishengsa@tongji.edu.cn
† These authors contributed equally to this work.

Abstract: Excessive pilot workload is one of the significant causes of flight accidents. The detection of flight workload can help optimize aircraft crew operation procedures, improve cockpit human–machine interface (HMIs) design, and ultimately reduce the risk of flight accidents. However, traditional detection methods often employ invasive or patch-based devices that can interfere with the pilot's control. In addition, they generally lack real-time capabilities, while the workload of pilots actually varies continuously. Moreover, most models do not take individual physiological differences into account, leading to the poor performance of new pilots. To address these issues, this study developed a real-time pilot workload detection model based on low-interference devices, including telemetry eye trackers and a pressure-sensing seat cushion. Specifically, the Adaptive KNN-Ensemble Pilot Workload Detection (AKE-PWD) model is proposed, combining KNN in the outer layer for identifying the physiological feature cluster with the ensemble classifier corresponding to this cluster in the inner layer. The ensemble model employs random forest, gradient boosting trees, and FCN–Transformer as base learners. It utilizes soft voting for predictions, integrating the strengths of various networks and effectively extracting the sequential features from complex data. Results show that the model achieves a detection accuracy of 82.6% on the cross-pilot testing set, with a runtime of 0.1 s, surpassing most studies that use invasive or patch-based detection devices. Additionally, the model demonstrates high accuracy across different individuals, indicating good generalization. The results are expected to improve flight safety.

Keywords: pilot workload; low-interference detection; ensemble learning; soft voting

check for updates

Citation: Liu, Y.; Gao, Y.; Yue, L.; Zhang, H.; Sun, J.; Wu, X. A Real-Time Detection of Pilot Workload Using Low-Interference Devices. *Appl. Sci.* **2024**, *14*, 6521. https://doi.org/10.3390/app14156521

Academic Editors: António Correia, Daniel Schneider, Benjamim Fonseca, Tommi Kärkkäinen and João M. F. Rodrigues

Received: 14 May 2024
Revised: 17 July 2024
Accepted: 23 July 2024
Published: 26 July 2024

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

1. Introduction

A flight accident results in significant casualties and substantial financial losses [1]. Human factors contribute to approximately 75% of all accident causes [2]. Among these, aviation maintenance errors account for about 12–15% [3] since maintenance workers sometimes perform a variety of tasks within a limited timeframe with stress and fatigue [4]. However, a larger proportion is due to pilot operational issues [2].

The pilot operation issues that cause accidents mainly fall into two categories: One is the failure to complete standard operating procedures (SOPs) within their allowable operational time window (AOTW) [5]. Sometimes, the AOTW is simply too short for the pilot to complete the SOPs. In other cases, high workload impedes the pilots' ability to complete them within the AOTW. The second category is because of pilot operation errors [6]. A significant cause of these two types of situations is the high workload of pilots, which can significantly slow a pilot's reaction, hampering their ability to finish the procedures within the prescribed time limits [7]. It can also increase the probability of pilots making errors [8]. However, it is also worth noting that despite human errors often being

blamed for many accidents, pilots frequently intervene in a timely manner to ensure flight safety and do so far more often than they cause accidents under high workload [9].

If real-time detection of pilot workload is achievable, the data can be analyzed to identify trends and patterns in pilot workload under various flight conditions [10]. By examining data from periods of high workload, specific conditions and key causes of elevated workload can be identified. These insights can then be leveraged to redesign aircraft crew and maintenance operation procedures [11]. For example, this process can help in avoiding cascading high workloads to reduce accidents. Moreover, the human–machine interfaces (HMIs) in modern aircraft systems are increasingly digital and complex [12], which increases the pilot workload and poses potential safety risks. Based on the real-time detection of pilot workload, we can support modifications to cockpit HMIs to improve their designs [13]. Additionally, we can provide comprehensive metrics to help aviation regulatory bodies establish standards and guidelines, such as setting the requirements for cockpit configuration and the minimum crew [14].

At present, the main methods for detecting workload can be categorized into three types: subjective assessment, task performance measurement, and physiological measurement. Both subjective assessment and task performance measurement are post-evaluation methods [15] which cannot meet the real-time detection requirements for pilot workload. Physiological measurement utilizes physiological parameters to assess the workload of individuals while performing tasks. Studies [16–18] have shown that this method can effectively monitor pilots' mental health by tracking various physiological indicators, making it relatively objective and capable of real-time detection.

However, the physiological devices commonly used in previous studies [17], such as electrode-type electroencephalography (EEG), electrocardiography (ECG), and galvanic skin response (GSR), are often equipped with invasive or patch-based components that may interfere with participants' performance. In contrast to these devices that cause interference, there exists a category of detection devices, such as telemetry eye trackers [19], in-ear EEG [20], and textile-based sensors [21]. These are characterized by low detection interference with pilot operations, thus making the workload data collection more authentic.

In addition, in studies using physiological measurements for workload detection, most experiments evaluate the average workload throughout the entire process based on the overall task difficulty after the task is completed [22,23]. Actually, during flight, the workload of pilots varies continuously, showing significant differences across different flight phases [24]. Applying the same workload level label to the whole task overlooks the continuous changes in workload during task performance, making the models developed in these studies less suitable for actual flight scenarios.

Moreover, these studies often overlook the different relationship between workload and human physiological characteristics among individuals. Physiological responses to workload specifically denote the body's automatic, measurable reactions to the demands placed on an individual during tasks [25]. However, these responses vary not only across different tasks but also across individuals due to factors such as genetics, physical condition, and experience [26]. For instance, when different pilots face the same high-workload task, some may not show noticeable differences in heart rate, while others may experience a significant increase. Individual differences in how indicators like EEG, eye movement, and heart rate variability respond to workload can be observed in existing studies [27–29]. These individual differences make many current models less effective and reliable when applied to a broad number of pilots in practice.

To address these challenges, this study developed a novel real-time pilot workload detection method using low-interference devices, called the Adaptive KNN-Ensemble Pilot Workload Detection, abbreviated as AKE-PWD. The main contributions of this study are as follows:

1. The proposed method utilizes low-interference detection metrics that hardly interfere with flight operations, including the eye movement indicators from telemetry eye trackers and the shifting rate center of gravity in the pressure-sensing seat cushion.

2. The proposed method analyzes the relationship between multi-source data and workload by utilizing a 10 s time window, allowing it to detect continuous workload levels within a single task in real time.
3. The proposed model integrates the strengths of three base learners and employs a soft voting mechanism, achieving high prediction accuracy through refined structural design. It incorporates a KNN module before the ensemble classifier to identify the clusters of individual's physiological feature–workload mapping pattern. By differentiating the processing of each cluster of physiological features, it enhances the model's generalizability across different pilots.

The remainder of this paper is organized as follows: Section 2 reviews the relevant studies; Section 3 introduces the proposed AKE-PWD methodology; Section 4 details the experimental setup and evaluates the proposed methodology; Section 5 presents the results and discussion; and Section 6 concludes the study.

2. Literature Review

2.1. Causes of Flight Accidents

2.1.1. Incomplete SOPs within the AOTW

The AOTW is a critical time window for completing essential procedures during flight, and its length varies with the complexity of the flight environment and the tasks involved [30]. When SOPs cannot be completed within the AOTW, it can lead to flight accidents [5]. In some cases, the AOTW itself is too short to complete these procedures. Another scenario is when the high-stress or complex flight environment increases the pilot's workload, affecting their performance and preventing them from completing the required tasks within the AOTW [5]. High workload can lead to delays in procedure execution as pilots are overwhelmed with multiple tasks simultaneously, stretching their cognitive and physical resources [31].

Moreover, simulation studies [32] have found that the time distribution for pilots to complete SOPs under different workload shows significant variance, further supporting that AOTW settings should adapt to different workload levels and flight scenarios. Improving SOP design and adjusting AOTW settings can help pilots complete their tasks within the prescribed time, reducing the risk of accidents due to insufficient time [33].

2.1.2. Pilot Operation Errors Due to Excessive Workload

High workload can increase the probability of pilots making errors [8]. Glaser et al. [34] indicated that stress acts as an intervening variable between workload and performance. Specifically, high workloads are associated with increased stress, which, in turn, can negatively impact performance and cause pilots to make mistakes. Dehais [35] pointed out that high workload can decrease visual and auditory attention, memory, and execution ability, greatly increasing the probability of pilots making errors during flight tasks. According to MacDonald [36], workload was proven to be a key determinant of stress and fatigue levels among operators performing repetitive tasks.

2.2. Main Methods for Detecting Workload

2.2.1. Subjective Measurement of Workload

Subjective measurement techniques, initially introduced by Pass [37], provide a holistic measure of workload. These measures typically assess the psychological workload induced by tasks and their behavioral performance through structured self-report scales. This method has become widely used by scholars in many studies [38,39], demonstrating good psychometric properties. Popular subjective workload scales include the Cooper–Harper single-factor rating scale [40], the NASA task load index (NASA-TLX) [41], and the subjective workload assessment technique (SWAT) [42], all known for their reliability. NASA-TLX was originally designed to offer a comprehensive assessment of workload in tasks, especially in the aviation and aerospace sectors. The subsequent extensive research has optimized the selection of its subscales and the weighted averaging approach. Over the

past two decades, it has demonstrated both ease of use and consistent responsiveness to key experimental manipulations [43].

The main advantages of subjective measurement techniques are their sensitivity and simplicity. However, due to the simplicity and limited number of direct questions, they only provide an overall measure of workload and cannot distinguish between different types of workloads. This method is not a real-time measurement; rather it is conducted after the operator's task is completed. This subjective and offline approach, asynchronous with the subject's actual perception processes, can potentially lead to discrepancies between the assessment results and the subject's true condition [15].

2.2.2. Task Performance Measurement of Workload

Task performance measurement assesses workload based on the performance of subjects in completing predetermined tasks. For instance, F. Paas and others [44] have evaluated psychological effort and cognitive load levels through subjects' reading and vocabulary scores; and R. Brünken and others [45] found that the accuracy and response times in tasks effectively reflect workload levels. However, current task performance measurements are mainly used for reading or memory tasks and are not very suitable for assessing pilots' workload during flight tasks. Furthermore, according to the theory of workload-task performance relationship proposed by De Waard [46], there is no absolute negative correlation between them. Similarly, this type of performance evaluation is also a post-evaluation, which cannot detect real-time workload during the process.

2.2.3. Physiological Measurement of Workload

Physiological measurement records the physiological signals of subjects to evaluate their workload [37]. Due to its objectivity and capacity for immediate feedback, this method effectively supports the ongoing assessment of pilot workload. However, most physiological measurement devices, such as EEG, ECG, and GSR, are often equipped with components that may interfere with participants' performance. For instance, the patch-based dry electrodes, although they do not use gel, could restrict participants' movements or distract their attention because of their tight attachment to the skin [47,48].

In recent years, several low-interference devices have been developed. For instance, remote eye trackers can measure a subject's eye movements from a distance [19]. The advantage of this device is its non-intrusiveness and freedom. It can achieve high accuracy and vast tracking robustness in eye movement detection while allowing users to barely feel the presence of the device [49]. E-Textile is a kind of textile-based sensor, which can measure various parameters [50,51]. Unlike sensors based on silicon or piezoelectric materials, e-Textile is similar to conventional wearable fabrics in both cost and texture. As a result, these sensors can be seamlessly incorporated into various textiles for unobtrusive and low-interference detection [52,53]. Among them, textile-based pressure sensors in seat cushions can provide valuable information for the recognition of sitting postures, which, in turn, can be used to identify workload through body gesture recognition [54]. moreover, Looney et al. [55] introduced a kind of in-ear EEG device that makes EEG measurement portable and comfortable. Jeong et al. [56] use an echo state network (ESN) to discriminate attention states on the basis of in-ear EEGs. Kuatsjah et al. [57] used features collected by a two-channel in-ear EEG system and achieved an accuracy of 79.30% in classifying workload during a visuomotor tracking task. This device relies on custom-made hearing aid earplugs, which require a wax impression of the ears to ensure a tight fit with the ear canal [20]. However, pilots already need to wear headsets for communication purposes, such as maintaining contact with the ground. Although these in-ear devices cause low interference to pilots, they may conflict with the communication headsets the pilots wear. So, it is believed that the in-ear EEG sensor and associated signal processing methods should be incorporated into the aviation headset when in practical use [58]. Since this study does not have access to such integrated devices, these in-ear devices will not be considered.

2.3. Physiological Indicators

EEG data record the brain's electrical activity, which results from ion movement in neurons during cognitive processes [59]. When neurons activate, the voltage across the neuronal cells changes, leading to fluctuations [60]. The signal is typically divided into several frequency bands based on the range of brainwave power, i.e., delta band, theta band, alpha band, beta band, and gamma band, with each band representing different aspects of brain activity [61]. Dolce and Waldeier [62] showed that delta wave power escalates during complex endeavors. However, due to the low frequency of the delta band, the clarity and reliability of the signal within this range can be easily affected by various types of interference and artifacts. Pope et al. [63] used a task engagement index to develop the first EEG-based adaptive system and identified the ratio beta/(alpha + theta) as the most sensitive. Freeman et al. [64] later confirmed the effectiveness of this index. W. Lang [65] and A. Mecklinger [66] found that alpha band power increases with increasing workload, most prominently at the temporal and occipital lobes, but also at the frontal lobe during complex tasks [67]. Naveen Kumar et al. [61] used the average power of the alpha band measured from four EEG channels on the frontal lobes (AF3, AF4) and temporal lobes (T7, T8) as an indicator of task workload. Their research validated that with an increase in workload, the average power of the alpha band also increased. This method is less challenging in terms of EEG signal processing and is more feasible.

ECG signals are sensitive to workload, especially within safety-critical dynamic systems like transportation [68]. Several studies have demonstrated a correlation between ECG responses and workload levels [69–71]. The heart rate (HR) and heart rate variability (HRV) of ECG signals represents the natural fluctuations in the intervals between heartbeats. The key time-domain metrics of HRV encompass the standard deviation of NN intervals (SDNN), the percentage of successive RR intervals that vary by over 50 ms (pNN50), heart rate (HR), and the root mean square of successive differences between RR intervals (RMSSD) [72]. The frequency-domain includes the absolute power of the ultra-low-frequency band (ULF), the very-low-frequency band (VLF), the low-frequency band (LF), the high-frequency band (HF), and the LF-to-HF power ratio (LF/HF) [73]. Amir Tjolleng et al. [70] collected six ECG measures in both the time and frequency domains to classify drivers' workload levels. P.G.A.M. Jorna [71] indicated that cardiovascular measures are well suited to indexing the different mental states of pilots, as well as their dynamic responses to variations in workload.

Existing research has demonstrated some significant correlation between certain eye movement metrics and subjects' workload. Velichkovsky et al.'s study [74] suggests that upon immediate hazard detection, there is an increase in fixation duration and subsequent adjustments in fixation frequency to accommodate changes in visual demand. Marquart [42] and Van Orden et al. [75] also found that both fixation duration and frequency increase under high workload. Moresi et al. [76] and Wierda et al. [77] demonstrated a close relationship between pupil diameter and workload.

In addition, ergonomic and psychological studies have demonstrated that body posture can serve as one of the indicators for workload assessment [53]. Giakoumis et al. [78] showed that behavioral features enhanced the performance of automatic stress detection systems using physiological features. Valentina Nino et al. [79] demonstrated that body postures assumed to perform an activity are negatively affected by workload perception. Jonathan Aigrain [80] used body posture to build an SVM model to recognize workload during task processes, where the amount of body movement was positively correlated with the level of stress. Therefore, indicators related to body posture could potentially be used to assess the workload of pilots during tasks.

Machine Learning Workload Detection Models

Halverson et al. [81] and W. L. Lim et al. [82] utilized SVM for workload modeling. The former employs eye-tracking and the latter uses EEG data, both achieving satisfactory predictive outcomes. Zhu et al. [83] developed an ANN model based on functional near-

infrared spectroscopy (fNIRS) data and reached an accuracy of 70.02% in classifying the workload of processing wayfinding information into three classes. Hamidur Rahman [84] extracted heart rate variability from facial videos, combined with vehicle dynamics parameters, and studied four machine learning algorithms, namely, logistic regression (LR), support vector machine (SVM), linear discriminant analysis (LDA), and neural network (NN), for classifying driver workload; the results show that LR achieved the highest accuracy, reaching 92% in binary classification.

Taheri Gorji, H. et al. [85] utilized recursive feature elimination (RFE), along with a stacking ensemble machine learning algorithm composed of SVM, RF, and LR, achieving 91.67% accuracy in a three-class setting. Niall McGuire et al. [86] applied bagging and stacking ensemble techniques to the STEW dataset, including a stacking BLSTM consisting of eight learners, which achieved a classification accuracy of 97% in a three-class setting. The results from this study underscore that ensemble networks can enhance the accuracy beyond that of individual learners while reducing prediction variance.

However, these models did not account for individual physiological differences, and the accuracy of the trained models was not validated on new individuals. To address the issue of poor model generalization, some studies have optimized machine learning methods to mine shared features among subjects. They have constructed cross-subject workload detection models to mitigate the impact of inter-subject variability and non-stationary characteristics on workload recognition results. Wang et al. [87] proposed a cross-operator cognitive workload recognition method using frequency domain features combined with an SVM classifier. Appriou et al. [88] compared the cross-operator performance of different classifiers, demonstrating the superiority of CNNs in extracting shared features. Yang et al. [89] demonstrated that normalization approaches relying on the choice of appropriate baselines help compensate for the effects of temporal variation and individual differences. Zhou Y. et al. [90] developed a workload classification model based on CNN and adversarial domain generalization, incorporating an adversarial domain generalization module to enhance model robustness and generalization by extracting shared features. Another line of research addresses individual differences by clustering individual physiological characteristics to distinguish different physiological patterns. Chen et al. [26] used K-means to cluster the data before classifying it with SVM. Yuna Noh [91] used the fuzzy c-means clustering algorithm to differentiate the workload characteristics of different drivers and created personalized driver workload profiles (PDWP) to reflect the unique physiological responses of drivers, achieving good accuracy in workload classification.

3. Methods

The Adaptive KNN-Ensemble Pilot Workload Detection (AKE-PWD) model is proposed. It adopts a strategy of soft voting ensemble learning, with random forest, gradient boosting trees, and an FCN–Transformer model as the base learners. In order to solve the problem of individual differences regarding physiological features, the K-nearest neighbors (KNN) algorithm is introduced in the outer layer of the model.

During data preprocessing, all features were clustered into four distinct categories using the spectral clustering technique. For the model construction, first, the KNN algorithm identifies the physiological feature cluster to which a sample data belong. After that, the corresponding ensemble model is then selected to perform feature extraction and classification prediction for the identified cluster, thereby increasing the robustness and cross-pilot generalization of the model. The model structure is shown in Figure 1.

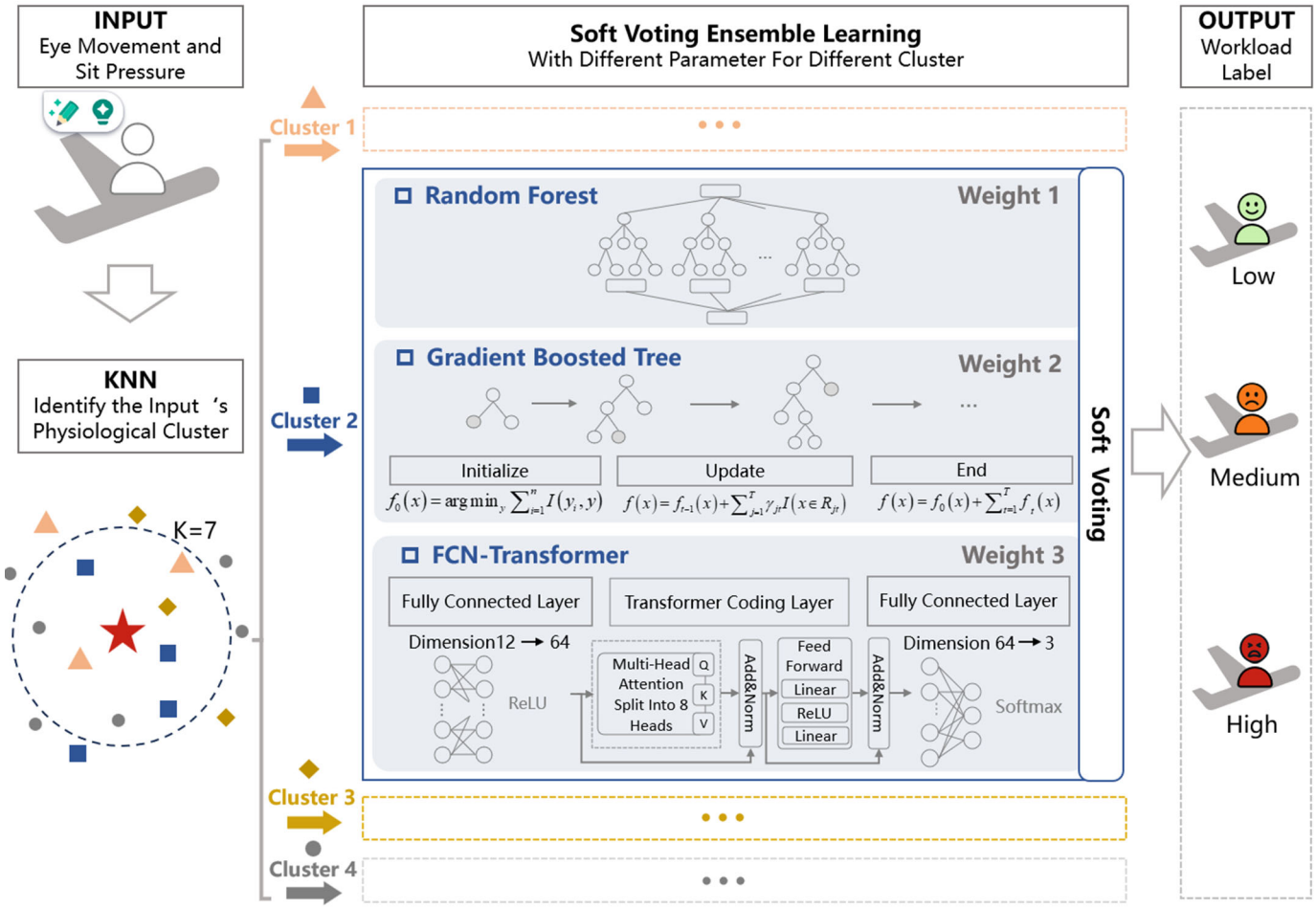


Figure 1. The architecture of the entire model.

3.1. Identify the Physiological Feature Cluster by KNN

To address the issue of individual physiological differences affecting the cross-pilot accuracy of model, we employed a differentiated processing method for various patterns of physiological features, processing different categories of features with models tuned to specific parameters. All collected feature data are clustered into four classes using spectral clustering. The KNN algorithm is then used to determine the cluster of the input features. Based on model performance, K = 7 is set, meaning that the 7 nearest feature points to the input data are selected. This nearest neighbor selection is based on Euclidean distance calculations. Subsequently, the most frequent feature category (1–4) among these 7 points is chosen as the cluster for the input data, and an appropriate ensemble learning model is selected based on this category for further processing.

3.2. Ensemble Learning

This study tested more than ten common models for workload classification and adopted an ensemble learning strategy to derive the final classifier. The application of ensemble learning through a soft voting mechanism can mitigate the issues of high variance and overfitting that often plague individual models [92]. While a single model may exhibit overfitting on distinct data subsets, the ensemble approach using soft voting systematically diminishes these effects, thereby enhancing the stability and accuracy of the overall model on unseen data.

In selecting the base learners for the ensemble, the choice was made by considering both the enhancement of overall model accuracy and the management of complexity, as well as the characteristics of the base learners themselves. Our dataset contains sequential data from physiological signals, which means that capturing temporal dependencies and

intricate patterns is crucial. Additionally, the data exhibit non-linear relationships and high dimensionality. Based on these characteristics, three models—random forest (RF), gradient boosting trees (GBT), and FCN–Transformer—were selected as base learners.

3.2.1. Random Forest

Random forest (RF) improves the accuracy and stability of predictions by constructing multiple decision trees and using average prediction values. It has advantages in processing non-linear data and automatic feature selection and performs well in handling high-dimensional data and preventing overfitting [93]. It is particularly suitable for processing datasets with complex structures in this study.

First, RF performs self-sampling by randomly selecting multiple samples from the original dataset to construct multiple sub-datasets, with duplicate sampling allowed. The number of samples extracted is a hyperparameter, usually set to the square root of the size of the original dataset. Next, it constructs a decision tree and trains one decision tree for each sub-dataset. During the node splitting process, a subset of features is randomly selected to determine the optimal split point. The number of features selected is also a hyperparameter, typically set to the logarithm base 2 of the total number of features. Finally, the prediction results are obtained using a voting method, with the specific formula as follows:

$$Y = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

where N is the number of decision trees, and y_i represents the result of the ith tree.

3.2.2. Gradient Boosting Trees

Gradient boosting trees (GBT) enhance prediction accuracy by incrementally adding decision trees to correct errors from the previous model. Each new tree is constructed to reduce residuals, transforming weak learners into a strong learner over the course of training [94]. GBT is particularly effective at handling datasets with strong and complex interactions between variables, offering detailed optimization suitable for capturing subtle patterns within the data.

First, the model is initialized to a constant value:

$$f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} l(y_i, \gamma).$$

For iterations t = 1 to T (the total number of trees), first, we compute the residuals r_{it}; the, we construct a tree to predict r_{it}, resulting in leaf node regions R_{jt} and corresponding scores γ_{jt}. Finally, the model is updated:

$$f(x) = f_0(x) + \sum_{J=1}^{T} \gamma_{jt} I\left(x \in R_{jt}\right).$$

Finally, the following model is obtained:

$$f(x) = f_0(x) + \sum_{t=1}^{T} f_t(x).$$

3.2.3. FCN–Transformer

The FCN–Transformer model serves as a pivotal base learner, incorporating a self-attention mechanism to conduct an in-depth analysis of inputs. It combines the advantages of fully connected networks (FCN) and Transformers, effectively extracting sequential features from complex data and particularly excelling at capturing long-range dependencies and intricate patterns within sequence data. In practice, feature data are continuously input in 10 s time windows, embodying certain time-series characteristics, which leverages the strengths of the Transformer model to enhance prediction accuracy.

- Input Linear Transformation Layer

The input first passes through a linear transformation layer, denoted as fc1. The purpose of this layer is to transform the dimensions of the input data from the original feature space to an intermediate representation space. The linear transformation is defined by the following formula:

$$x' = ReLU(W_1 x + b_1),$$

where x' represents the input feature vector, and W_1 and b_1 are the weights and biases of the linear layer, respectively. ReLU is applied to introduce non-linearity, aiding the model in learning more complex features.

- Transformer Encoder Layer

A multi-head self-attention mechanism is at the core of the Transformer architecture, aimed at enabling the model to capture information from different representation subspaces simultaneously. The input sequence X first undergoes three sets of linear transformations to generate the query (Q), key (K), and value (V) matrices:

$$Q = XW^Q, K = XW^K, V = XW^V,$$

where W^Q, W^K, and W^V are learnable weight matrices.

We utilize the dot-product attention function to calculate the similarity between queries and all keys, followed by normalization through the softmax function to compute the attention scores:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where d_k denotes the dimensionality of the key vectors, which serves to stabilize gradients. The softmax function ensures that all attention scores are positive and add up to 1.

In order to enhance the expressive power of the model, the aforementioned process is executed in parallel across multiple heads, with each head corresponding to a different weight matrix. The output is then integrated through a linear transformation after concatenating the results from all heads.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_n)W^O,$$

where each $head_i$ is an independent self-attention operation result. Through iterative parameter tuning, this study employs an 8-head self-attention mechanism. W^O represents another learnable weight matrix.

The Transformer model adopts residual connections to help alleviate the gradient vanishing problem in deep networks. The output of each sublayer, whether it be a self-attention layer or a feed-forward network, is added to the input of that sublayer and then normalized within the layer:

$$LayerNorm(x + Sublayer(x)),$$

where Sublayer(x) represents the output of the sublayer itself. This design enhances the stability and efficiency of the training process.

Each encoder layer also incorporates a feed-forward network, which independently applies the same operation to the output at each position. This network consists of two linear transformations and a ReLU activation function:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2.$$

Here, W_1, W_2, b_1, and b_2 are network parameters. The feed-forward network further enhances the model's ability to handle non-linear problems.

- Output Linear Transformation Layer

After being processed by the Transformer encoder layer, the output y is then passed through another linear transformation layer. The purpose of this layer is to map the output of the encoder layer to the final class space. The formula for this step is as follows:

$$z = W_2 y + b_2,$$

where W_2 and b_2 denote the weights and biases of this layer. The final output z represents scores for different categories, which can be utilized for classification tasks.

3.2.4. Soft Voting Ensemble

The model employs a soft voting method to integrate the above three base learners. By averaging their predictions to make the final decision, this approach effectively integrates the predictive capabilities of multiple models. The soft voting ensemble strategy is defined as follows: assuming there are M base learners, each model provides a probability distribution $p_m(i)$ as the predicted output for sample i, representing the probabilities of the sample belonging to each class. The prediction $p(i)$ of the ensemble model is calculated as the arithmetic average of predictions from all individual models:

$$p(i) = \frac{1}{M}\sum_{m=1}^{M} p_m(i),$$

where $p_k(i)$ represents the probability distribution predicted by model m for sample i, and M is the total number of models (M = 3).

The final class selection is made based on the averaged probability distribution, whereby each sample is classified into the category with the highest average probability, denoted as follows:

$$c(i) = \arg\max_k p_k(i),$$

where $p_k(i)$ denotes the average probability of sample i belonging to category k.

The ensemble integrates the unique capabilities of random forest (RF), gradient boosting trees (GBT), and FCN–Transformer models. The RF model contributes through its multi-tree structure, offering robust classification stability and effectively handling the complexity and high dimensionality of the data. The GBT progressively corrects errors from previous predictions and optimizes its accuracy over iterations, and it is particularly good at capturing non-linear relationships. Meanwhile, the FCN–Transformer excels in processing time-series data, capturing long-term dependencies and intricate patterns inherent in the dataset. This synergy allows the ensemble model to outperform any single model by leveraging their combined strengths to effectively manage the complex interactions and non-linear patterns present in the dataset.

4. Experimental Setup

4.1. Experimental Design

To elicit different levels of pilot workload among participants, this study designed three categories of simulated flight tasks, labeled A, B, and C, with increasing levels of difficulty. Task A is to pilot a single-engine light aircraft to take off and ascend to an altitude of 1500 m under the guidance of the tutorial. Task B is to pilot a Boeing 737 aircraft to approach from 3 nautical miles away and 3 km in the air under favorable weather conditions and land on a designated runway. Task C is similar to Task B but is conducted under adverse weather conditions. Weather conditions not only affect visual circumstances but also affect the piloting through the settings of dynamic parameters, thereby significantly increasing the pilot's workload. The specific details are shown in Table 1.

Table 1. The specific settings for three types of tasks.

Task	Weather and Dynamics Simulation	Screen Display	Experimental Difficulty	Overall Workload Level Expectation
A	Noon; Cloudless; No Turbulence		Easy	Low
B	Noon; Scattered Cumulus Clouds; Some Precipitation and Storms; Some Turbulence		Moderate	Medium
C	Dusk; Thick Cumulus Clouds at 1500 m; Heavy Precipitation and Storms; Severe Turbulence		Difficult	High

4.2. Experiment Procedure

This study utilized the joysticks, rudder pedals, and throttle controls jointly developed by Thrustmaster and Boeing to emulate the flight dynamics controlling experience of real pilots. Flight scenarios were constructed using the X-Plane11 software. It incorporates 28,000 aerodynamic models to deliver realistic physical simulations and has been certified by the FAA, placing it among the most advanced flight simulation software available [95]. This experimental setup is widely used in scientific research and engineering experiments in the field of simulated flight [96,97].

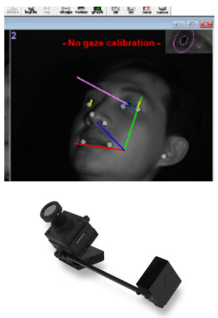
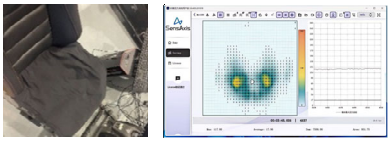
Before starting the experiment, participants were required to undergo unified training to ensure they had mastered the basic skills necessary for simulated flight. During this process, their individual flight ability was also assessed.

The specific experimental steps are as follows:

First, we introduce the participants to the experimental procedures, connect the participants to the ECG and ECG electrodes, and adjust the eye-tracking equipment with the participant's cooperation. Then, we start the flight simulation program, as well as the data recording of the eye-tracking device, ECG, EEG, and pressure-sensing seat cushion simultaneously. We the sequentially guide the participant to complete tasks A, B, and C. Throughout the formal experiment, various physiological data of the participants were synchronously collected, and the detection equipment used is shown in Table 2. After completing each type of task, participants need to fill out the NASA-TLX scale to subjectively evaluate their workload.

The experimental scenario is shown in Figure 2. Throughout the experiment, the environmental variables, such as noise and temperature, were consistently controlled at the same level.

Table 2. Experimental device.

Detection Object	Equipment	Introduction	Appearance
Eye Movement	SmarteyePro Telemetry Eye Tracker	Equipped with systems including SmarteyePro (a remote eye-tracking system, version 9.3), pylon viewer (Camera Setup Software, version 7.3.0), and HRT (human factors research system); achieves telemetry through computer image processing technology and is completely non-interfering	
Seat Pressure	Pressure-Sensing Seat Cushion	Equipped with SenAxis visualization monitoring software	
EEG&ECG	EEG&ECG Signal Detector	Including 32-channel EEG cap and 4-channel EMG electrodes, equipped with TMSI processing system	

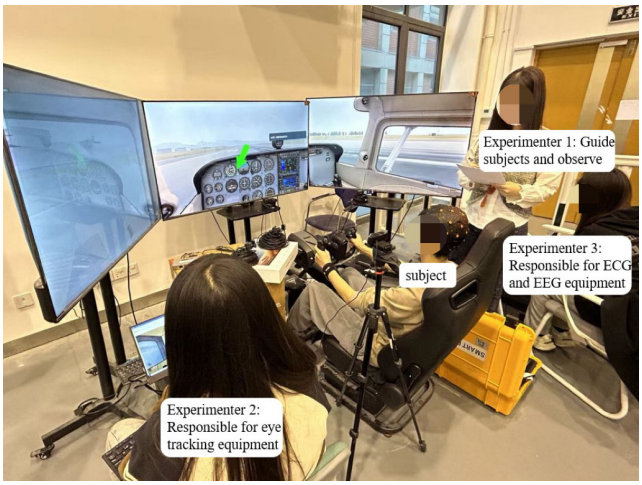


Figure 2. The experimental scenario.

4.3. Participants

A total of 20 participants from the local community and university were recruited for this experiment, and their average age was 22.3. Their heights and weights met the criteria for civil flight pilots in China, with heights ranging from 164 to 185 cm and weights ranging between 80% and 130% of their standard weight (standard weight = height − 110). All participants had visual acuity (including corrected vision) of 0.8 or higher. The experimental protocol for this study was approved by the Institutional Ethics Committee of Tongji University, Shanghai, China. All participants provided informed consent in accordance with human research ethics guidelines.

4.4. Indicator Extraction

4.4.1. NASA-TLX

The NASA task load index (NASA-TLX) exhibits low inter-subject variability and is one of the most widely used subjective workload assessment tools [41]. It consists of six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. The total workload score for each participant is determined by the weighted average of the scores across these dimensions, with higher scores indicating greater workload.

The specific calculation formula is as follows:

$$F = \sum_{i=1}^{6} M_i \times \frac{P_i}{15},$$

where F indicates the total score for workload assessment; M_i is the score selected by the participant of the ith dimension; and P_i is the number of times the ith item was selected in 15 pairs.

4.4.2. Seat Pressure

Considering that the original pressure magnitude is significantly influenced by individual weight and aircraft ascent and descent, the indicator ultimately extracted from the pressure sensor is the center of gravity shifting rate, which reduces the impact of irrelevant factors.

The raw data provided by the pressure-sensing seat cushion is a 32 × 32 matrix, representing the pressure distribution for each frame, with a sampling rate of 20 Hz. From the pressure distribution matrix, the center of gravity position at each moment can be determined.

$$x = \frac{\sum_{i=1}^{32}\sum_{j=1}^{32} j \cdot P_{ij}}{\sum_{i=1}^{32}\sum_{j=1}^{32} P_{ij}}, y = \frac{\sum_{i=1}^{32}\sum_{j=1}^{32} i \cdot P_{ij}}{\sum_{i=1}^{32}\sum_{j=1}^{32} P_{ij}} \quad (1)$$

$$v_{\arg} = \frac{1}{10 f_s}\sum_{t=1}^{N} f_s \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \quad (2)$$

Through Formula (1), the center of gravity shifting rate can be calculated, and the average value of this rate over every 10 s is computed using Formula (2). This indicator can measure the tension level of participants' posture changes. Statistical results demonstrate that it has a certain correlation with the subjective scales calculated through NASA-TLX ($r > 0.75$).

4.4.3. Eye Tracking Data

The device's sampling rate is 60 Hz. It can output information such as timestamps, blinks, eyelid opening sizes, pupil diameters, and gaze coordinates. Over a time interval of more than 100 ms, if the Euclidean distance between gaze coordinates at adjacent timestamps is less than 50, it is considered one fixation. Since there is no browsing behavior in this experiment, eye movements during non-fixation periods are considered saccades. The saccade amplitude is the Euclidean distance between the centers of two fixation points.

The video screen from the rear camera is divided into area of interest A (AOI A), area of interest B (AOI B), and non-interest area, as shown in Figure 3, to obtain eye movement indicators within specific areas of interest. Eye movement indicators are calculated in 10 s windows. Table 3 shows the definitions of each indicator.

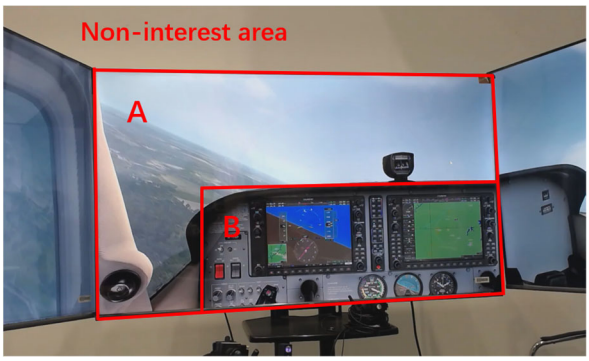


Figure 3. The division of areas of interest.

Table 3. Eye movement indicators and their explanations.

Eye Movement Indicators	Explanation
Average Fixation Duration	The mean duration of each fixation within a time window (seconds).
Number of Fixations/Saccades/Blinks	The number of fixations, saccades, and blinks within a time window.
Average Saccade Amplitude	The mean Euclidean distance between the start and end coordinates of each saccade within a time window.
Left (Right) Pupil Diameter	The average diameter of the left (right) pupil at each timestamp within a time window (meters).
Fixation Duration in Area of Interest A (B)	The average duration of each fixation within area of interest A (B) in a time window (seconds).
Number of Fixations in Area of Interest A (B)	The number of fixations within area of interest A (B) in a time window.

4.4.4. ECG Data

The ECG signal processing workflow is illustrated in Figure 4. After preprocessing to remove noise, the process begins with a first-order difference using Formula I to identify all inflection points of the ECG signal. Then, a second-order difference is applied using Formula II to calculate all the peaks of the ECG signal. Finally, the peak threshold is determined using Formula III to identify the peak points of the R-waves. Thus, the RR intervals are obtained, allowing for the calculation of the average heart rate every 10 s. Statistical results indicate that the average heart rate (denoted as HR) of each group has a strong positive correlation with the scores of subjective scales ($r > 0.8$), suggesting that it can effectively indicate the magnitude of workload in flight tasks.

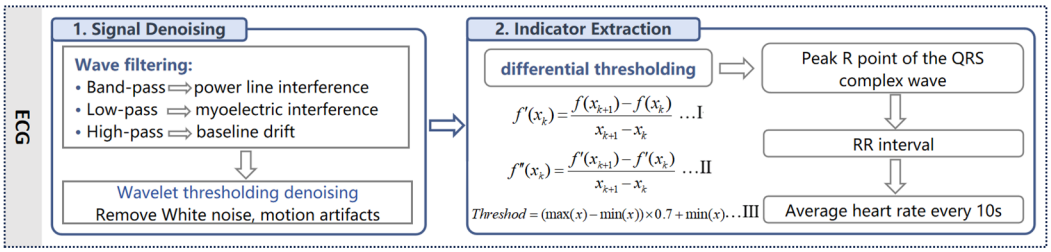


Figure 4. The preprocessing of ECG data.

4.4.5. EEG Data

The raw EEG signals were preprocessed using the EEGLAB toolbox in MATLAB. The specific workflow is illustrated in Figure 5.

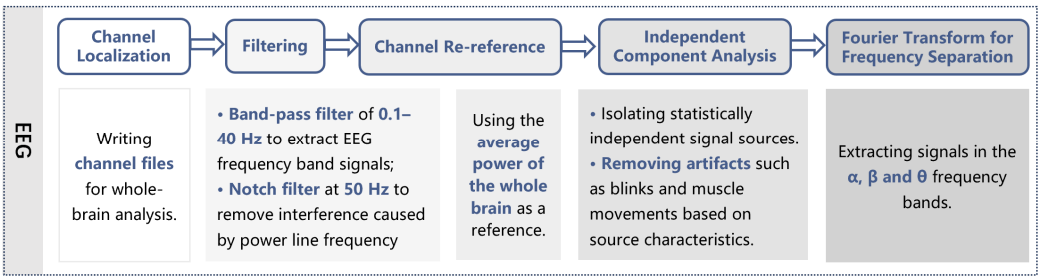


Figure 5. The preprocessing of EEG data.

Following the approach of Kumar et al. [61], this study used the average power of the alpha band every 10 s from four EEG channels—FP1 and FP2 in the frontal lobe and T7 and T8 in the temporal lobe—as the characteristic indicators of workload. To eliminate variations between individuals, these values were normalized within each participant. The EEG workload characteristic value for the ith participant during the jth task is denoted as E_{ij}. Statistical tests demonstrated that E_{ij} can significantly distinguish between the levels of workload across different flight task difficulties ($p < 0.05$), and it showed a Pearson correlation coefficient greater than 0.85 with subjective scales of NASA-TLX, indicating a strong positive correlation. Therefore, it may be concluded that E_{ij} can represent the magnitude of workload.

4.4.6. Workload Labels

From the above analysis, it is evident that both ECG and EEG indicators can effectively represent the magnitude of workload in flight tasks. The EEG indicator is denoted as E, and heart rate as HR. We then draw an E-HR scatter plot, as shown in Figure 6a, which shows a clear positive correlation between them within each participant. Principal component analysis (PCA) performed on the dataset composed of E and HR shows that the first principal component accounts for the vast majority of the variance, as seen in Figure 6b. Calculating its loadings, the results show that both E and HR have the same loading of 0.7071 on the first principal component, indicating that the main features of the dataset are distributed along the y = x direction. Therefore, using K-means, the data points are clustered into three categories along the characteristic direction based on the values of E and HR, as illustrated in Figure 6c. It can be observed in Figure that both E and HR increase with the increment of workload labels, as shown in Figure 6d,e. Thus, each sample point is labeled with one of the three categories: low, medium, and high.

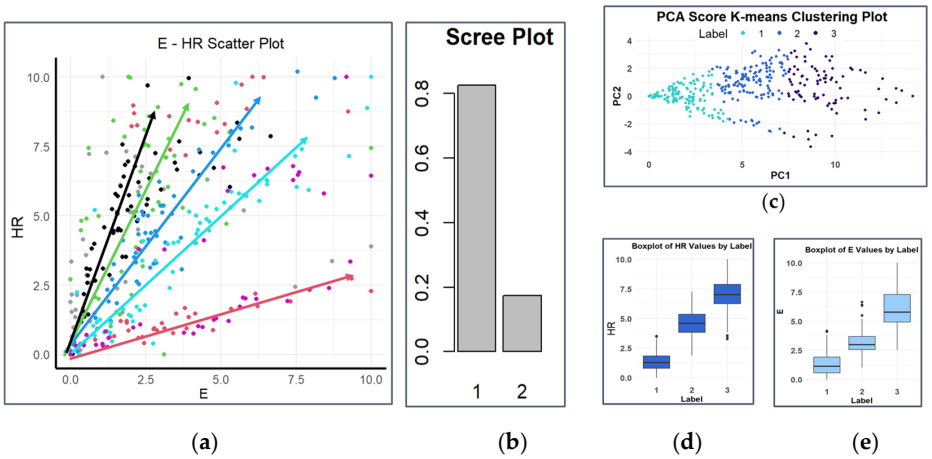


Figure 6. The process of statistical analysis for a dataset composed of the indicators HR and E: (**a**) an E−HR scatter plot (only partial data are included to make the image clearer); (**b**) the PCA of the dataset; (**c**) clustering results along the main feature direction; (**d**) the distribution of HR in three types of workload labels; (**e**) the distribution of E in three types of workload labels.

Ultimately, the dataset consists of 11 low-interference detection metrics, i.e., 10 remote eye-tracking indicators, 1 seat cushion pressure indicator, and the workload label.

4.5. Spectral Clustering of Physiological Features

Individuals exhibit differences in their physiological responses even when faced with the same situation [91,98,99]. This study employs spectral clustering to separate features with significant differences, laying the groundwork for subsequent differentiated model training.

The spectral clustering module is characterized by constructing a similarity graph and utilizing its eigendecomposition to identify and differentiate individual physiological variations effectively, which provides a foundation for personalized driver workload assessment. The basic framework of spectral clustering can be described as follows:

First, a weighted undirected graph is established as follows:

$$G = (v, \varepsilon, \omega),$$

where

v is the node set, representing the set of data points for individual physiological characteristics;

ε is the edge set, representing the similarity between data points;

ω is the similarity matrix, where each element ωij denotes the similarity between nodes i and j.

The similarity matrix ω is constructed using the radial basis function (RBF) kernel to measure the similarity between two nodes, with an expression as follows:

$$\omega_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

involving σ, a predetermined scale parameter that controls the rate of similarity decay.

Based on the similarity matrix W, we further construct the graph Laplacian matrix L. Dimensionality reduction is achieved by solving for the eigenvalues and eigenvectors of L. Subsequently, clustering is performed using the K-means algorithm in the reduced-dimensional space. Utilizing the aforementioned methods, spectral clustering algorithms can effectively cluster the physiological characteristics data of pilots.

The number of clusters is determined based on the Calinski–Harabasz index, with the optimal ratio of inter-cluster separation to intra-cluster compactness achieved with four categories, as shown in Figure 7. In this way, all physiological characteristic data are clustered into four categories. Subsequently, distinct soft voting ensemble learning classifiers will be trained based on these four categories of features.

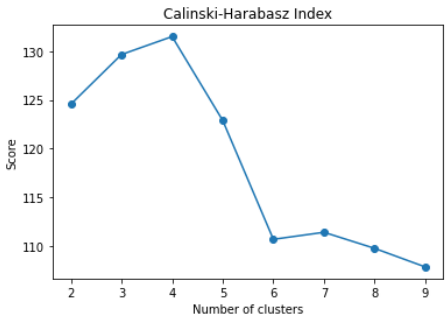


Figure 7. The Calinski–Harabasz index of different numbers of clusters.

4.6. Data Augmentation Using CTGAN

As the FCN–Transformer model is a type of deep learning architecture, there exists a risk of overfitting when applied to small-sample datasets, especially with complex neural

networks. To mitigate this problem, the study employs the conditional generative adversarial network for tabular data (CTGAN) algorithm to augment the original dataset of 1050 entries to a dataset containing 6000 entries.

The operational principle of CTGAN is based on the framework of generative adversarial networks (GANs), which includes two main components: a generator (G) and a discriminator (D). The goal of the generator is to produce data realistic enough to deceive the discriminator, while the discriminator aims to differentiate between generated and real data. Throughout the training process, these two components compete against each other, with the generator eventually learning to produce data indistinguishable from real data. The interaction between the generator G and the discriminator D can be described by the following optimization problem:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data(x)}}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))].$$

Here, x represents real data; z is a noise vector drawn from a prior distribution; $G(z)$ denotes the data generated by the generator; $D(x)$ is the discriminator's assessment of the real data; and $D(G(z))$ is the discriminator's assessment of the generated data. CTGAN modifies this foundational architecture to handle tabular data more efficiently. For categorical variables, CTGAN implements a training-by-sampling strategy, introducing conditional vectors and a generator loss function to address issues of category imbalance. In CTGAN, the generator is conditionally trained, meaning that for each category, it generates data based on a conditional vector that guides the generation process [100]. The conditional vectors specify the generation conditions, while categorical variables are represented as multi-dimensional one-hot vectors. During training, the conditional generator can produce any combination of one-hot vectors. For numerical variables, CTGAN utilizes a variational Gaussian mixture model (VGM) and employs the Wasserstein GAN loss function for gradient penalty, enhancing the stability and effectiveness of the training process. Overall, CTGAN, with these technological enhancements, is better equipped to handle the imbalance of categories in tabular data and generate synthetic data that more closely mirror the distribution of real data.

The Jensen–Shannon divergence (JSD) is a common method of measuring the similarity between two probability distributions [101]. The definition is as follows:

$$JSD = H(M) - \frac{1}{2}(H(P) + H(Q)),$$

where:

$M = \frac{1}{2}(P + Q)$ is a mixture distribution of P and Q;

$H(P)$ is the Shannon entropy for distribution P, $H(P) = -\sum_x P(x)\log P(x)$.

The similarity between the distributions is greater when the Jensen–Shannon distance is closer to zero [102]. The JS divergence of the data before and after augmentation using CTGAN is shown in Table 4, where the left column lists the features, and the right column displays the JSD of each feature distribution before and after augmentation. Results demonstrate the effective data augmentation performance of CTGAN.

Table 4. The JS divergence of features.

Feature	JS Divergence	Feature	JS Divergence
Fixation duration	0.023	Right pupil diameter	0.079
Fixation frequency	0.017	Fixation duration in AOI A	0.055
Saccade frequency	0.010	Fixation frequency in AOI A	0.005
Blink frequency	0.018	Fixation duration in AOI B	0.014
Saccade distance	0.008	Fixation frequency in AOI B	0.042
Left pupil diameter	0.121	Workload label	0.019

4.7. Model Training and Testing

After data cleaning and indicator extraction, the dataset contains a total of 1073 entries. By applying data augmentation based on CTGAN, it is effectively expanded to 6000 entries structured into 12 dimensions, including 10 eye movement indicators, 1 seat cushion indicator, and 1 workload-level label.

To assess the reliability of the model, the leave-one-subject-out cross-validation method is used to divide the training and test sets. For the data of 20 subjects, the data from 19 subjects are used as the training set, and the remaining one is used as the testing set each time. In this way, the model's accuracy on new individuals is tested, ensuring the model's generalizability.

In the training process, the RF model utilizes 110 decision trees, and the GBT model employs 120 weak learners and uses a multi-class error rate as the evaluation metric. The FCN–Transformer model is set with a learning rate of 0.002, runs for 20 epochs, and processes data in batches of 32, utilizing the Adam optimizer for adjustments during training.

4.8. Ablation Experiment

To further demonstrate the structural efficacy of the proposed AKE-PWD model, three ablation experiments were conducted:

1. Three classifier models (RF/GBT/FCN–Transformer): This experiment evaluated each base learner independently, testing their performance when predicting separately.
2. Soft Voting Ensemble Learning model (RF + GBT + FCN–Transformer): The ensemble learning strategy employed a soft voting mechanism, integrating RF, GBT, and FCN–Transformer. This experiment analyzed the improvements in prediction accuracy offered by the ensemble approach compared to the individual models, i.e., RF/GBT/FCN–Transformer.
3. Adaptive KNN-Ensemble model (KNN + Soft Voting Ensemble (RF + GBT + FCN–Transformer)): Utilizing the KNN algorithm, this model identified the cluster of physiological features relevant to the input data. Based on these clusters, corresponding ensemble models were selected for feature processing. This approach was tested to evaluate its effectiveness in enhancing model generalizability and accuracy by considering individual physiological differences.

5. Results and Discussion

5.1. Model Performance

By employing KNN to identify the physiological feature clusters for appropriate ensemble learning methods, the proposed model demonstrates superior accuracy, real-time performance, and cross-driver generalizability in workload recognition. The model achieves an accuracy of 82.6% on the testing set, with a runtime of only 0.1 s, and maintains a standard deviation of less than 2% across validation sets, indicating high stability.

From Figure 8, it is shown that the AKE-PLD achieves higher prediction accuracy on new individuals compared to other models. The introduction of the ensemble strategy has improved the model accuracy by about 11% compared to the single-model approach. Moreover, utilizing KNN to identify physiological features and selecting the appropriate ensemble model has further enhanced model accuracy by 21% compared to the original ensemble learner. This demonstrates the accuracy and generalizability of the model, proving the effectiveness of the proposed model's structure.

The classification performance was evaluated using accuracy, precision, recall, and F1 score, which are defined by the following formulae:

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}.$$

Among them,
TP (true positives): the number of correctly predicted positive samples;
TN (true negatives): the number of correctly predicted negative samples;
FP (false positives): the number of negative samples incorrectly predicted as positive;
FN (false negatives): the number of positive samples incorrectly predicted as negative.

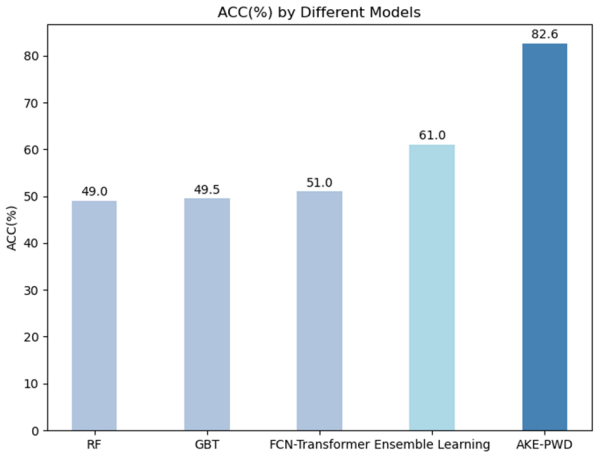


Figure 8. The results of ablation experiments.

Table 5 details the model performance on three workload levels (Level 1 = low; Level 2 = moderate; and Level 3 = high). Level 1 exhibits excellent performance across all three metrics. Metrics of Level 2 show consistency, but are slightly lower compared to Level 1, suggesting that the model's recognition ability for moderate workload is somewhat inferior. Level 3 has high precision but a lower recall rate, meaning that some of the samples with high workload were not identified by the model.

Table 5. The performance of the three labels in terms of precision, recall, and F1 score.

Workload Level	Precision	Recall	F1-Score
Level 1	0.8557	0.9431	0.8973
Level 2	0.7863	0.7931	0.7897
Level 3	0.8493	0.7470	0.7949
Macro Average	0.8304	0.8278	0.8273
Weighted Average	0.8258	0.8258	0.8242

Figure 9 displays the model's confusion matrix. Label 1 shows excellent predictive accuracy at 94%. Labels 2 and 3 have lower recall rates of 79% and 75%, each exhibiting some degree of confusion with adjacent labels, suggesting potential areas for improvement.

Figure 10 shows the ROC curves generated individually for the three workload levels by treating each one as the positive class while considering the remaining classes as negatives. The overall curves are close to the upper left corner, with the areas under the curve (AUC) for the three classes being 0.94, 0.82, and 0.85, respectively. Level 1 still demonstrates excellent discriminative ability, while Levels 2 and 3 are slightly inferior, with Level 3 performing marginally better than Level 2.

Based on the above results, it can be seen that the predictive performance of the model across the three categories of workload is generally excellent. There is still room for further improvement at higher levels of workload. This may be due to the increased noise in signal collection caused by more intense driving maneuvers under high workload, as well as the greater variability in physiological responses among different individuals at high workload levels.

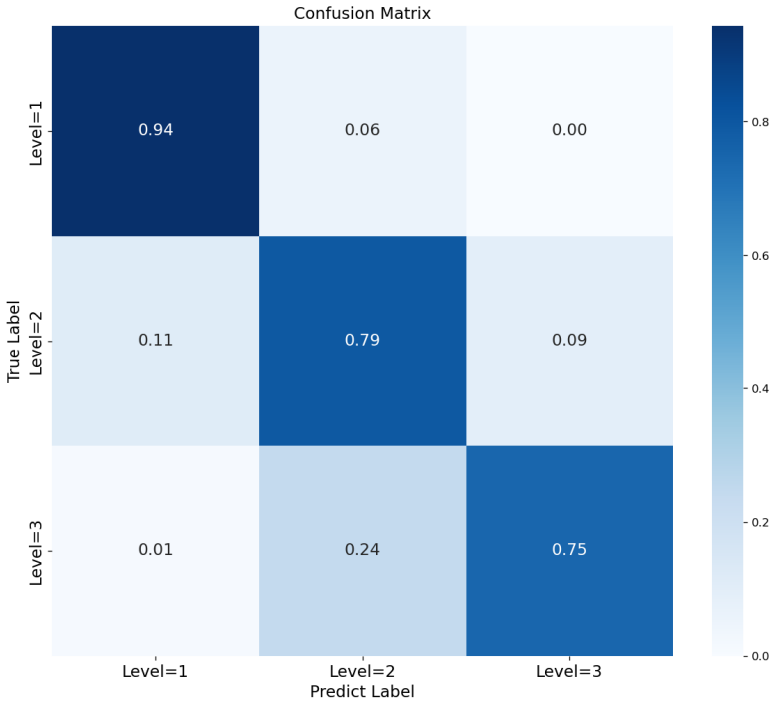


Figure 9. The confusion matrix of the model results.

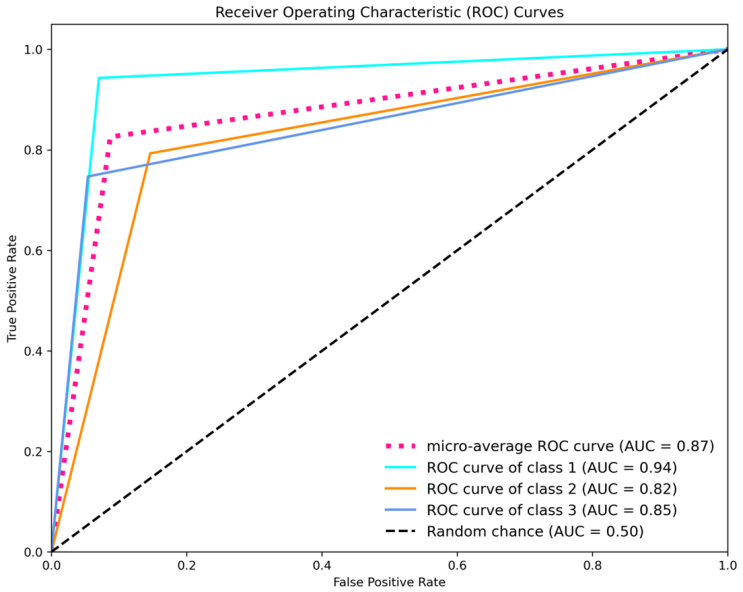


Figure 10. The receiver operating characteristic (ROC) curves.

5.2. Comparison with Other Studies

In addressing physiological differences between individuals, some studies [88–90] have optimized machine learning to uncover shared features to mitigate the impact of inter-subject variability on workload recognition outcomes. Inspired by their algorithmic ideas, this study also experimented with a workload detection model based on a fully connected network and adversarial domain generalization (FCN-DG), incorporating an adversarial domain generalization module to enhance robustness and generalizability by extracting common features, as shown in Figure 11.

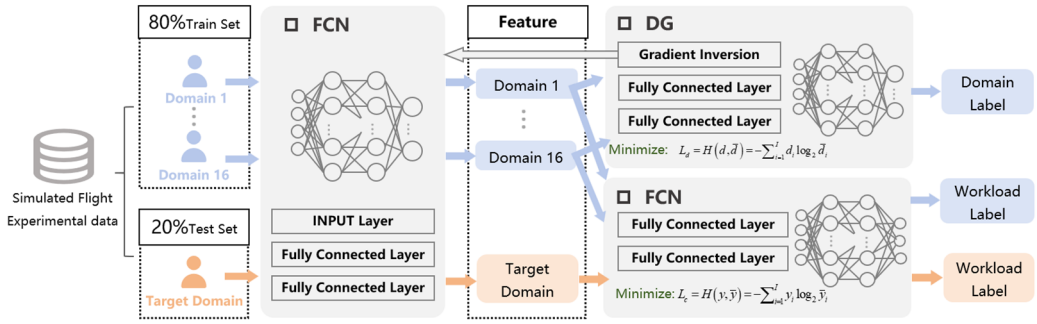


Figure 11. The architecture of the FCN-DG model.

However, the shared features extracted through confrontation were suboptimal for classification purposes, achieving only 65% accuracy via this model, while it costs more time training the adversarial generalization models. And the highest accuracy of other scholars' adversarial generalization models only reached 69%. This may be due to the following factors: (1) the association between the shared physiological features identified and workload is derived from group-level analyses, representing the commonality of all individuals, which are not strong enough to assess an individual's workload at a specific moment accurately; (2) physiological responses that are consistent within some individuals may vary across single individuals such that some variables that are non-responsive to pilot workload at the group level may be valuable for assessing workload at the individual level [26].

Therefore, the proposed model proves more advantageous in terms of both efficiency, accuracy, and generalizability.

Table 6 lists several other studies from recent years that have also used machine learning models for workload detection. It is evident that although categorizing into the same three classes, the proposed model surpasses most models built on traditional physiological measurement techniques in terms of accuracy through its low-interference detection approach.

Table 6. Comparison between the method proposed in this study and existing methods.

Authors (or Model Name)	Year	Physiological Signal	Feature Selection	Classification Model	Number of Categories	Whether Device Interferes with Operation	Can Be Achieved in Real Time	ACC
D. He et al. [103]	2022	ECG, GSR	None	K-NN	3	Yes	No	72%
S. Barua et al. [104]	2017	ECG, GSR	SFFS, MDA	K-NN	3	Yes	No	78%
L. Zheng et al. [105]	2021	HRV, GSR	Pearson-r	ANN	3	Yes	No	73%
Li et al. [106]	2023	EEG, Eye tracking	LDA	LR	2	Yes	No	82.7%
Q. Zhu et al. [83]	2021	fNIRS, ECG	None	LDA	3	Yes	No	68%
Lim et al. [82]	2018	EEG	None	SVM	3	Yes	No	69%
FCN-DG	2024	Eye tracking, Seat pressure	DG-FCN	FCN	3	No	Yes	65%
AKE-PWD	2024	Eye tracking, Seat pressure	None	ensemble learning	3	No	Yes	82.6%

5.3. The Role of Human Factors in Predicting Workload

This study mainly focuses on low-interference detection metrics for workload assessment. However, other metrics, such as the ones related to human factors, may also contribute to the model prediction. To explore and validate this idea, we incorporated age, gender, and flight ability assessment scores from the pre-experiment for each subject into the dataset.

The results showed that adding age and gender did not improve the model's performance. This may be due to the relatively concentrated age distribution of the recruited pilots and the predominantly male composition (to reflect the gender ratio in the pilot population). As a result, these variables did not provide sufficient variability to enhance the model's predictive performance; thus, they were not included in the model.

However, incorporating the flight ability assessment scores increased the model's prediction accuracy by approximately 2%. Specifically, as shown in Table 7, for the prediction of high workload level, the inclusion of the flight ability metric resulted in improvements in precision, recall, and F1-score. This finding indicates that human factor data related to the subject's baseline abilities can indeed enhance the model's predictive performance of pilot workload. This flight ability is subjectively judged through simulated flight tests. However, given that there might be various ways to evaluate the flight ability, its reference value for broader application can be demonstrated further.

Table 7. Model performance of level 3 workload before and after incorporating flight ability.

	Precision	Recall	F1-Score
Before	0.8493	0.7470	0.7949
After	0.8497	0.7936	0.8172

5.4. Limitations and Future Works

This study has several limitations that need to be addressed:

The relationship between workload and human physiological characteristics varies significantly between individuals, yet the model categorizes them into just four classes. Expanding the dataset with more participants would improve the spectral clustering algorithm, allowing the model to detect more-precise physiological patterns and refine its predictive accuracy. Additionally, conducting experiments with different aircraft and simulators under various flight conditions can help collect diverse data and enhance the model's generalizability across different flying environments.

This study did not consider the potential impact of environmental factors such as temperature, humidity, and noise on pilot workload as these variables were controlled at consistent levels throughout the experiment. Future work will include further investigation into environmental factors on workload to enhance the applicability of the model.

Moreover, the experiments in this study were conducted using flight simulation devices. Although these advanced devices are officially certified, they cannot fully replicate all the detailed behaviors of real flight. For example, real aircraft maneuvers involve movements of rolling and yawing, which causes the gravitational force and the pressure exerted on the pilot to not be aligned. This may affect the measurement of pressure distribution.

6. Conclusions

High pilot workload is a major contributing factor to aviation accidents, and workload detection can help reduce the incidence of such accidents. To address the issues in existing workload detection models, such as interference with pilot operations and the lack of real-time applicability, and to enhance the model's prediction accuracy and generalizability, this paper proposed the Adaptive KNN-Ensemble Pilot Workload Detection (AKE-PWD) model. By harnessing low-interference detection metrics, i.e., using the telemetry-based eye-tracking metrics and seat pressure metrics as model inputs, the approach achieves a streamlined detection process without being perceived by the subjects at all. By gathering continuous data from flight simulation experiments and employing 10 s time windows as sampling points, it ensures real-time detection capabilities. By constructing the Adaptive KNN-Ensemble model, which includes an outer KNN and an inner ensemble classifier composed of RF, GBT, and FCN–Transformer, we can identify the cluster of physiological features relevant to the input data and subsequently select the corresponding ensemble model which integrates the strengths of the three base learners. In this way, detection generalizability across different pilots is facilitated. The real-time workload detection across

different pilots achieved an accuracy of 82.6%, surpassing baseline workload detection methods. Additionally, it operates with a running time of less than 0.1 s and can detect the workload every 10 s, ensuring real-time capability. The research findings have broad application prospects. For instance, specific conditions and causes of elevated workload can be identified from the real-time detection data. The findings can guide the design and improvement of the cockpit HMIs and the aircraft crew operation procedures. Additionally, it can provide comprehensive metrics to help aviation regulatory bodies establish standards and guidelines. This provides new insights into the workload of human–machine interaction and has significant potential for the safe, intelligent, and efficient development of future aviation traffic systems.

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

Funding: This research is funded by the National Natural Science Foundation Project under Grant 52302442.

Institutional Review Board Statement: The study was approved by the Institutional Ethics Committee of Tongji University for studies involving humans with the approval number tjdxsro82 on 15 April 2024.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request. Due to confidentiality agreements, access to the raw data may be restricted.

Acknowledgments: The College of Transportation and Engineering of Tongji University provided all the equipment we need. Our sincere gratitude goes to our college and all the volunteers who actively participated in this research.

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

References

1. Sobieralski, J.B. The Cost of General Aviation Accidents in the United States. *Transp. Res. Part A Policy Pract.* **2013**, *47*, 19–27. [CrossRef]
2. Kharoufah, H.; Murray, J.; Baxter, G.; Wild, G. A review of human factors causations in commercial air transport accidents and incidents: From to 2000–2016. *Prog. Aerosp. Sci.* **2018**, *99*, 1–13. [CrossRef]
3. Rashid, H.S.J.; Place, C.S.; Braithwaite, G.R. Investigating the investigations: A retrospective study in the aviation maintenance error causation. *Cogn. Technol. Work* **2013**, *15*, 171–188. [CrossRef]
4. Dalkilic, S. Improving aircraft safety and reliability by aircraft maintenance technician training. *Eng. Fail. Anal.* **2017**, *82*, 687–694. [CrossRef]
5. Bashatah, J.; Sherry, L. Model-Based Analysis of Standard Operating Procedures' Role in Abnormal and Emergency Events. In Proceedings of the INCOSE International Symposium, Detroit, MI, USA, 25–30 June 2022; Volume 32, pp. 1220–1246.
6. Amalberti, R.; Wioland, L.I.E.N. Human error in aviation. In *Aviation Safety, Human Factors-System Engineering-Flight Operations-Economics-Strategies-Management*; CRC Press: Boca Raton, FL, USA, 2020; pp. 91–108.
7. Kantowitz, B.H.; Campbell, J.L. Pilot workload and flightdeck automation. In *Automation and Human Performance*; CRC Press: Boca Raton, FL, USA, 2018; pp. 117–136.
8. Dismukes, R.K.; Kochan, J.A.; Goldsmith, T.E. Flight crew errors in challenging and stressful situations. *Aviat. Psychol. Appl. Hum. Factors* **2018**, *8*, 35–46. [CrossRef]
9. Holbrook, J.; Barshi, I. Why Learning From All Operations Is Imperative. In *Aviation Safety InfoShare*; NASA: Dallas, TX, USA, 2020.
10. Kantowitz, B.H.; Casper, P.A. Human workload in aviation. In *Human Error in Aviation*; Routledge: London, UK, 2017; pp. 123–153.
11. Cahill, J.; Losa, G. Flight crew task performance and the design of cockpit task support tools. In Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore! London, UK, 28–31 August 2007; pp. 83–87.

12. Thomas, P.; Biswas, P.; Langdon, P. State-of-the-art and future concepts for interaction in aircraft cockpits. In *Universal Access in Human-Computer Interaction. Access to Interaction: 9th International Conference, UAHCI 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings, Part II 9*; Springer International Publishing: Cham, Switzerland, 2015; pp. 538–549.
13. Piechulla, W.; Mayser, C.; Gehrke, H.; König, W. Reducing drivers' mental workload by means of an adaptive man–machine interface. *Transp. Res. Part F Traffic Psychol. Behav.* **2003**, *6*, 233–248. [CrossRef]
14. Speyer, J.J.; Fort, A.; Fouillot, J.P.; Blomberg, R.D. Assessing workload for minimum crew certification. In Proceedings of the AGARD Conference on Methods to Assess Workload, (AGARD–CPP–282), Stuttgart, Germany, 27 September–1 October 1987; pp. 90–115.
15. Zhou, Y.; Huang, S.; Xu, Z.; Wang, P.; Wu, X.; Zhang, D. Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 799–818. [CrossRef]
16. Charles, R.L.; Nixon, J. Measuring mental workload using physiological measures: A systematic review. *Appl. Ergon.* **2019**, *74*, 221–232. [CrossRef]
17. Tao, D.; Tan, H.; Wang, H.; Zhang, X.; Qu, X.; Zhang, T. A Systematic Review of Physiological Measures of Mental Workload. *Int. J. Environ. Res. Public Health* **2019**, *16*, 2716. [CrossRef]
18. Orlandi, L.; Brooks, B. Measuring mental workload and physiological reactions in marine pilots: Building bridges towards redlines of performance. *Appl. Ergon.* **2018**, *69*, 74–92. [CrossRef]
19. Klingner, J.; Kumar, R.; Hanrahan, P. Measuring the task-evoked pupillary response with a remote eye tracker. In Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, Savannah, GA, USA, 26–28 March 2008; pp. 69–72.
20. Belkhiria, C.; Peysakhovich, V. Electro-encephalography and electro-oculography in aeronautics: A review over the last decade (2010–2020). *Front. Neuroergon.* **2020**, *1*, 606719. [CrossRef]
21. Meyer, J.; Lukowicz, P.; Troster, G. Textile pressure sensor for muscle activity and motion detection. In Proceedings of the 2006 10th IEEE International Symposium on Wearable Computers, Montreux, Switzerland, 11–14 October 2006; pp. 69–72.
22. Mohanavelu, K.; Poonguzhali, S.; Janani, A.; Vinutha, S. Machine learning-based approach for identifying mental workload of pilots. *Biomed. Signal Process. Control* **2022**, *75*, 103623. [CrossRef]
23. Xi, P.; Law, A.; Goubran, R.; Shu, C. Pilot workload prediction from ECG using deep convolutional neural networks. In Proceedings of the 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Istanbul, Turkey, 26–28 June 2019; pp. 1–6.
24. Lee, Y.H.; Liu, B.S. Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviat. Space Environ. Med.* **2003**, *74*, 1078–1084.
25. Hancock, P.A.; Matthews, G. Workload and performance: Associations, insensitivities, and dissociations. *Hum. Factors* **2019**, *61*, 374–392. [CrossRef]
26. Chen, J.; Zhang, Q.; Cheng, L.; Gao, X.; Ding, L. A Cognitive Load Assessment Method Considering Individual Differences in Eye Movement Data. In Proceedings of the 2019 IEEE 15th International Conference on Control and Automation (ICCA), Edinburgh, UK, 16–19 July 2019; IEEE: San Diego, CA, USA, 2019; pp. 295–300.
27. Bargary, G.; Bosten, J.M.; Goodbourn, P.T.; Lawrance-Owen, A.J.; Hogg, R.E.; Mollon, J.D. Individual differences in human eye movements: An oculomotor signature? *Vis. Res.* **2017**, *141*, 157–169. [CrossRef]
28. Miyake, S.; Yamada, S.; Shoji, T.; Takae, Y.; Kuge, N.; Yamamura, T. Physiological responses to workload change. A test/retest examination. *Appl. Ergon.* **2009**, *40*, 987–996. [CrossRef]
29. Riding, R.J.; Glass, A.; Butler, S.R.; Pleydell-Pearce, C.W. Cognitive style and individual differences in EEG alpha during information processing. *Educ. Psychol.* **1997**, *17*, 219–234. [CrossRef]
30. Kourdali, H.K.; Sherry, L. Available operational time window: A method for evaluating and monitoring airline procedures. *J. Cogn. Eng. Decis. Mak.* **2017**, *11*, 371–381. [CrossRef]
31. Hart, S.G.; Bortolussi, M.R. Pilot errors as a source of workload. *Hum. Factors* **1984**, *26*, 545–556. [CrossRef]
32. Stimpson, A.J.; Ryan, J.C.; Cummings, M.L. Assessing pilot workload in single-pilot operations with advanced autonomy. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Washington, DC, USA, 19–23 September 2016; SAGE Publications: Los Angeles, CA, USA, 2016; Volume 60, pp. 675–679.
33. Commercial Aviation Safety Team. Improving Aviation Safety Report. 2020. Available online: https://www.cast-safety.org/pdf/2020-AVS-002_Improving_Aviation_Safety_Report_sd25_web.pdf (accessed on 16 July 2024).
34. Glaser, D.N.; Tatum, B.C.; Nebeker, D.M.; Sorenson, R.C.; Aiello, J.R. Workload and social support: Effects on performance and stress. *Hum. Perform.* **1999**, *12*, 155–176. [CrossRef]
35. Dehais, F.; Causse, M.; Vachon, F.; Régis, N.; Menant, E.; Tremblay, S. Failure to detect critical auditory alerts in the cockpit: Evidence for inattentional deafness. *Hum. Factors* **2014**, *56*, 631–644. [CrossRef]
36. MacDonald, W. The impact of job demands and workload on stress and fatigue. *Aust. Psychol.* **2003**, *38*, 102–117. [CrossRef]
37. Paas, F. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *J. Educ. Psychol.* **1992**, *84*, 429. [CrossRef]
38. Paas, F.; Tuovinen, J.E.; Tabbers, H.; Van Gerven, P.W. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. In *Cognitive Load Theory*; Routledge: New York, NY, USA, 2016; pp. 63–71.

39. Paas, F.; Ayres, P.; Pachman, M. Assessment of Cognitive Load in Multimedia Learning. In *Recent Innovations in Educational Technology that Facilitate Student Learning*; Robinson, D.H., Schraw, G., Eds.; Information Age Publishing: Charlotte, NC, USA, 2008; pp. 11–35.
40. Mitchell, D.G. Fifty Years of the Cooper-Harper Scale. In Proceedings of the AIAA Scitech 2019 Forum, San Diego, CA, USA, 7–11 January 2019; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2019; p. 0563.
41. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Hancock, P.A., Meshkati, N., Eds.; North-Holland: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183. [CrossRef]
42. Marquart, G.; Cabrall, C.; Winter, J. Review of Eye-Related Measures of Drivers' Mental Workload. *Procedia Manuf.* **2015**, *3*, 2854–2861. [CrossRef]
43. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 Years Later. In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco, CA, USA, 16–20 October 2006; Sage Publications: Los Angeles, CA, USA, 2006; Volume 50, pp. 904–908.
44. Paas, F.; Renkl, A.; Sweller, J. Cognitive Load Theory and Instructional Design: Recent Developments. *Educ. Psychol.* **2003**, *38*, 1–4. [CrossRef]
45. Brünken, R.; Plass, J.L.; Leutner, D. Assessment of Cognitive Load in Multimedia Learning with Dual-Task Methodology: Auditory Load and Modality Effects. *Instr. Sci.* **2004**, *32*, 115–132. [CrossRef]
46. De Waard, D. The Measurement of Drivers' Mental Workload. Ph.D. Thesis, University of Groningen, Groningen, The Netherlands, 1996.
47. Fu, Y.; Zhao, J.; Dong, Y.; Wang, X. Dry electrodes for human bioelectrical signal monitoring. *Sensors* **2020**, *20*, 3651. [CrossRef]
48. Chi, Y.M.; Jung, T.P.; Cauwenberghs, G. Dry-contact and noncontact biopotential electrodes: Methodological review. *IEEE Rev. Biomed. Eng.* **2010**, *3*, 106–119. [CrossRef]
49. Niehorster, D.C.; Cornelissen, T.H.; Holmqvist, K.; Hooge, I.T.; Hessels, R.S. What to expect from your remote eye-tracker when participants are unrestrained. *Behav. Res. Methods* **2018**, *50*, 213–227. [CrossRef]
50. Farringdon, J.; Moore, A.J.; Tilbury, N.; Church, J.; Biemond, P.D. Wearable sensor badge and sensor jacket for context awareness. Digest of Papers. In Proceedings of the Third International Symposium on Wearable Computers, San Francisco, CA, USA, 18–19 October 1999; pp. 107–113.
51. Marschall, M.; Harrington, A.C.; Steele, J.R. Effect of work station design on sitting posture in young children. *Ergonomics* **1995**, *38*, 1932–1940. [CrossRef]
52. Xu, W.; Li, Z.; Huang, M.C.; Amini, N.; Sarrafzadeh, M. ecushion: An etextile device for sitting posture monitoring. In Proceedings of the 2011 International Conference on Body Sensor Networks, Dallas, TX, USA, 23–25 May 2011; pp. 194–199.
53. Liang, G.; Cao, J.; Liu, X. Smart cushion: A practical system for fine-grained sitting posture recognition. In Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kona, HI, USA, 13–17 March 2017; pp. 419–424.
54. Qiu, J.; Helbig, R. Body Posture as an Indicator of Work. *Hum. Fact.* **2012**, *54*, 626–635. [CrossRef]
55. Looney, D.; Kidmose, P.; Park, C.; Ungstrup, M.; Rank, M.L.; Rosenkranz, K.; Mandic, D.P. The in-the-ear recording concept: User-centered and wearable brain monitoring. *IEEE Pulse* **2012**, *3*, 32–42. [CrossRef]
56. Jeong, D.H.; Jeong, J. In-ear EEG based attention state classification using echo state network. *Brain Sci.* **2020**, *10*, 321. [CrossRef]
57. Kuatsjah, E.; Zhang, X.; Khoshnam, M.; Menon, C. Two-channel in-ear EEG system for detection of visuomotor tracking state: A preliminary study. *Med. Eng. Phys.* **2019**, *68*, 25–34. [CrossRef]
58. Wilson, N.; Guragain, B.; Verma, A.; Archer, L.; Tavakolian, K. Blending human and machine: Feasibility of measuring fatigue through the aviation headset. *Hum. Factors* **2020**, *62*, 553–564. [CrossRef]
59. Zhang, J. Basic Neural Units of the Brain: Neurons, Synapses and Action Potential. *arXiv* **2019**, arXiv:1906.01703. [CrossRef]
60. Atkinson, R.C.; Shiffrin, R.M. Human Memory: A Proposed System and Its Control Processes. *Psychol. Learn. Motiv.* **1968**, *2*, 89–195. [CrossRef]
61. Kumar, N.; Kumar, J. Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study. *Procedia Comput. Sci.* **2016**, *84*, 70–78. [CrossRef]
62. Dolce, G.; Waldeier, H. Spectral and Multivariate Analysis of EEG Changes During Mental Activity in Man. *Electroencephalogr. Clin. Neurophysiol.* **1974**, *36*, 577–584. [CrossRef]
63. Pope, A.T.; Bogart, E.H.; Bartolome, D.S. Biocybernetic System Evaluates Indices of Operator Engagement in Automated Task. *Biolog. Psychol.* **1995**, *40*, 187–195. [CrossRef]
64. Freeman, F.G.; Mikulka, P.J.; Prinzel, L.J.; Scerbo, M.W. Evaluation of an Adaptive Automation System Using Three EEG Indices with a Visual Tracking Task. *Biolog. Psychol.* **1999**, *50*, 61–76. [CrossRef]
65. Lang, W.; Lang, M.; Kornhuber, A.; Diekmann, V.; Kornhuber, H.H. Event-Related EEG-Spectra in a Concept Formation Task. *Hum. Neurobiol.* **1988**, *6*, 295–301. [PubMed]
66. Mecklinger, A.; Kramer, A.F.; Strayer, D.L. Event Related Potentials and EEG Components in a Semantic Memory Search Task. *Psychophysiology* **1992**, *29*, 104–119. [CrossRef]
67. Fink, A.; Grabner, R.H.; Neuper, C.; Neubauer, A.C. EEG Alpha Band Dissociation with Increasing Task Demands. *Cogn. Brain Res.* **2005**, *24*, 252–259. [CrossRef]

68. Guo, W.; Tian, X.; Tan, J.; Zhao, L.; Li, L. Driver's Mental Workload Estimation Based on Empirical Physiological Indicators. In Proceedings of the 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 11–13 November 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 344–347. [CrossRef]
69. Heine, T.; Lenis, G.; Reichensperger, P.; Beran, T.; Doessel, O.; Deml, B. Electrocardiographic Features for the Measurement of Drivers' Mental Workload. *Appl. Ergon.* **2017**, *61*, 31–43. [CrossRef]
70. Tjolleng, A.; Jung, K.; Hong, W.; Lee, W.; Lee, B.; You, H.; Son, J.; Park, S. Classification of a Driver's Cognitive Workload Levels Using Artificial Neural Network on ECG Signals. *Appl. Ergon.* **2017**, *59*, 326–332. [CrossRef]
71. Jorna, P.G.A.M. Heart Rate and Workload Variations in Actual and Simulated Flight. *Ergonomics* **1993**, *36*, 1043–1054. [CrossRef]
72. Shao, S.; Wang, T.; Wang, Y.; Su, Y.; Song, C.; Yao, C. Research of HRV as a Measure of Mental Workload in Human and Dual-Arm Robot Interaction. *Electronics* **2020**, *9*, 2174. [CrossRef]
73. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 290215. [CrossRef]
74. Velichkovsky, B.M.; Rothert, A.; Kopf, M.; Dornhöfer, S.M.; Joos, M. Towards an Express-Diagnostics for Level of Processing and Hazard Perception. *Trans. Res. F-Traffic* **2002**, *5*, 145–156. [CrossRef]
75. Van Orden, K.F.; Limbert, W.; Makeig, S.; Jung, T.-P. Eye Activity Correlates of Workload During a Visuospatial Memory Task. *Hum. Fact.* **2001**, *43*, 111–121. [CrossRef]
76. Moresi, S.; Adam, J.J.; Rijcken, J.; Van Gerven, P.W.; Kuipers, H.; Jolles, J. Pupil Dilation in Response Preparation. *Int. J. Psychophysiol.* **2008**, *67*, 124–130. [CrossRef]
77. Wierda, S.M.; van Rijn, H.; Taatgen, N.A.; Martens, S. Pupil Dilation Deconvolution Reveals the Dynamics of Attention at High Temporal Resolution. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 8456–8460. [CrossRef]
78. Giakoumis, D.; Drosou, A.; Cipresso, P.; Tzovaras, D.; Hassapis, G.; Gaggioli, A.; Riva, G. Using Activity-Related Behavioural Features Towards More Effective Automatic Stress Detection. *PLoS ONE* **2012**, *7*, e43571. [CrossRef]
79. Nino, V.; Claudio, D.; Monfort, S.M. Evaluating the Effect of Perceived Mental Workload on Work Body Postures. *Int. J. Ind. Ergon.* **2023**, *93*, 103399. [CrossRef]
80. Aigrain, J.; Dubuisson, S.; Detyniecki, M.; Chetouani, M. Person-Specific Behavioural Features for Automatic Stress Detection. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Ljubljana, Slovenia, 4–8 May 2015; IEEE: New York, NY, USA, 2015; Volume 3, pp. 1–6. [CrossRef]
81. Halverson, T.; Estepp, J.; Christensen, J.; Monnin, J. Classifying Workload with Eye Movements in a Complex Task. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Boston, MA, USA, 22–26 October 2012; Sage Publications: Los Angeles, CA, USA, 2012; Volume 56, pp. 168–172. [CrossRef]
82. Lim, W.L.; Sourina, O.; Wang, L.P. STEW: Simultaneous Task EEG Workload Data Set. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 2106–2114. [CrossRef]
83. Zhu, Q.; Shi, Y.; Du, J. Wayfinding Information Cognitive Load Classification Based on Functional Near-Infrared Spectroscopy. *J. Comput. Civ. Eng.* **2021**, *35*, 04021016. [CrossRef]
84. Rahman, H.; Ahmed, M.U.; Barua, S.; Begum, S. Non-Contact-Based Driver's Cognitive Load Classification Using Physiological and Vehicular Parameters. *Biomed. Signal Process. Control* **2020**, *55*, 101634. [CrossRef]
85. Taheri Gorji, H.; Wilson, N.; VanBree, J.; Hoffmann, B.; Petros, T.; Tavakolian, K. Using Machine Learning Methods and EEG to Discriminate Aircraft Pilot Cognitive Workload During Flight. *Sci. Rep.* **2023**, *13*, 2507. [CrossRef]
86. McGuire, N.; Moshfeghi, Y. On Ensemble Learning for Mental Workload Classification. In Proceedings of the International Conference on Machine Learning, Optimization, and Data Science, Grasmere, Lake District, UK, 22–26 September 2023; Springer Nature: Cham, Switzerland, 2023; pp. 358–372.
87. Wang, S.; Gwizdka, J.; Chaovalitwongse, W.A. Using Wireless EEG Signals to Assess Memory Workload in the *n*-Back Task. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *46*, 424–435. [CrossRef]
88. Appriou, A.; Cichocki, A.; Lotte, F. Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload from EEG Signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2018; pp. 1–6. [CrossRef]
89. Yang, S.; Kuo, J.; Lenné, M.G.; Fitzharris, M.; Horberry, T.; Blay, K.; Wood, D.; Mulvihill, C.; Truche, C. The Impacts of Temporal Variation and Individual Differences in Driver Cognitive Workload on ECG-Based Detection. *Hum. Fact.* **2021**, *63*, 772–787. [CrossRef]
90. Zhou, Y.; Gong, P.; Wang, P.; Wen, X.; Zang, D. Cross-Operator Cognitive Workload Recognition Based on Convolutional Neural Network and Domain Generalization. *J. Electron. Inf. Technol.* **2023**, *45*, 2796–2805. [CrossRef]
91. Noh, Y.; Kim, S.; Jang, Y.J.; Yoon, Y. Modeling Individual Differences in Driver Workload Inference Using Physiological Data. *Int. J. Automot. Technol.* **2021**, *22*, 201–212. [CrossRef]
92. Mienye, I.D.; Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **2022**, *10*, 99129–99149. [CrossRef]
93. Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
94. Natekin, A.; Knoll, A. Gradient Boosting Machines, A Tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef]
95. Junior, J.M.M.; Khamvilai, T.; Sutter, L.; Feron, E. Test Platform for Autopilot System Embedded in a Model of Multi-Core Architecture Using X-Plane Flight Simulator. In Proceedings of the 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, 8–12 September 2019; IEEE: San Diego, CA, USA, 2019; pp. 1–6. [CrossRef]

96. Qin, H.; Zhou, X.; Ou, X.; Liu, Y.; Xue, C. Detection of Mental Fatigue State Using Heart Rate Variability and Eye Metrics During Simulated Flight. *Hum. Fact. Ergon. Manufact. Serv. Ind.* **2021**, *31*, 637–651. [CrossRef]
97. Tang, H.; Lee, B.G.; Towey, D.; Pike, M. The Impact of Various Cockpit Display Interfaces on Novice Pilots' Mental Workload and Situational Awareness: A Comparative Study. *Sensors* **2024**, *24*, 2835. [CrossRef]
98. Castelhano, M.S.; Henderson, J.M. Stable Individual Differences Across Images in Human Saccadic Eye Movements. *Can. J. Exp. Psychol./Rev. Can. Psychol. Exp.* **2008**, *62*, 1. [CrossRef]
99. Day, M.E. An Eye-Movement Indicator of Individual Differences in the Physiological Organization of Attentional Processes and Anxiety. *J. Psychol.* **1967**, *66*, 51–62. [CrossRef]
100. Lee, J.S.; Lee, O. CTGAN VS TGAN? Which One is More Suitable for Generating Synthetic EEG Data. *J. Theor. Appl. Inf. Technol.* **2021**, *99*, 2359–2372. Available online: https://www.researchgate.net/publication/352007070 (accessed on 5 June 2024).
101. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
102. Osterreicher, F.; Vajda, I. A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653. [CrossRef]
103. He, D.; Wang, Z.; Khalil, E.B.; Donmez, B.; Qiao, G.; Kumar, S. Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, *2676*, 670–681. [CrossRef]
104. Barua, S.; Ahmed, M.U.; Begum, S. Classifying Drivers' Cognitive Load Using EEG Signals. *Stud. Health Technol. Inform.* **2017**, *237*, 99–106. [CrossRef]
105. Zheng, L.; Qiao, X.; Ni, T.; Yang, W.; Li, Y. Driver Cognitive Loads Based on Multi-Dimensional Information Feature Analysis. *China J. Highw. Transp.* **2021**, *34*, 240–250.
106. Li, Q.; Ng, K.K.; Yu, S.C.; Yiu, C.Y.; Lyu, M. Recognising Situation Awareness Associated with Different Workloads Using EEG and Eye-Tracking Features in Air Traffic Control Tasks. *Knowl. Based Syst.* **2023**, *260*, 110179. [CrossRef]

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