Traffic & Transportation

Pilot Workload Identification During Engine Failure Landings

Jiajun YUAN[1], Huaxian LIN[2], Chenyang ZHANG[3], Xijun KE[4], Panpan YU[5], Qingfeng ZHANG[6]

Original Scientific Paper
Submitted: 26 Dec 2024
Accepted: 4 Apr 2025
Published: 24 Feb 2026

[1] 1071716402@qq.com, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China
[2] 185551013@qq.com, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China
[3] 18523507789@163.com, School of Transportation & Logistics, Southwest Jiaotong University, Chengdu, China
[4] 2622738049@qq.com, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China
[5] 3224589424@qq.com, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China
[6] Corresponding author, yjj@cafuc.edu.cn, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
Pilot workload is a critical factor affecting flight safety, particularly in emergency situations such as engine failure during forced landings. This study aims to assess pilot workload during engine failure-induced forced landings compared to normal landings using EEG signals. EEG data were recorded from 21 pilots using a flight simulator and EEG acquisition equipment during both forced and normal landings. Subjective workload scores were obtained using a subjective workload scale. EEG features were extracted, and nonparametric tests were applied to assess the significance of these features at different workload levels. Subsequently, machine learning algorithms (SVC, KNN, RF and LightGBM) were employed to develop models for workload evaluation. The LightGBM model achieved a peak accuracy of 99.5% using the top 80% of all features (time-domain, frequency-domain and nonlinear features) as inputs. This study provides an effective quantitative approach for assessing pilot workload during emergency situations and offers valuable insights for improving flight safety and optimising pilot training programs.

KEYWORDS
aviation safety; engine failure; workload; machine-learning; EEG.

1. INTRODUCTION

Flight safety is the foremost pursuit and eternal theme of civil aviation, and flight accidents are usually not caused by a single factor [1], including bad weather, operational errors, engine failure, etc. Among these, engine failure is one of the specific circumstances that seriously threaten flight safety. Statistics from the Civil Aviation Administration of China (CAAC) Aviation Safety Report indicate that engine failure is one of the top three events that cause accident symptoms [2]. The ability of pilots to access and utilise pertinent information swiftly is vital for effectively managing such incidents. In this unique situation, pilots face the challenge of simultaneously processing multiple items of information. This increased workload can strain resources and potentially leave limited capacity for additional concurrent tasks. Consequently, this condition could lead to operational mistakes by pilots, thus endangering flight safety. Therefore, there is a direct correlation between pilot workload and flight safety, and evaluating pilot workload becomes more critical in the field of aviation. However, research on pilot workload specifically during engine failure situations remains limited, highlighting the need for further exploration in this area.

The current methods of evaluating pilot workload fall into three main types: subjective evaluation, performance assessment and physiological measurements [3]. The subjective measurement method asked the pilot to describe their workload during flight, categorise and rank flight tasks according to their self-perception, and then quantitatively evaluate them. This approach, which is widely utilised in workload research, quantifies elements such as effort, mood and fatigue. Subjective assessment methods seek to quantify people's perceptions and evaluations of the task demands that they experience [4]. Among the many subjective assessment methods, the most representative are the NASA task load index (NASA-TLX) scale, the subjective workload assessment technique (SWAT) and the modified Cooper-Harper scale (MCH). Flight performance assessment is mainly used to assess the pilot workload by analysing pilot performance during mission execution. Third, the physiological data were collected. When the workload of the human body changes, the physiological parameters of the human body also change to a certain extent. Therefore, it is possible to reflect the degree of pilot workload through changes in physiological signals. Physiological parameters identify the workload level of pilots primarily by monitoring changes in their specific physiological indicators. This physiological parameter-based analysis measures and reflects the mental load of the on-the-job pilot. Physiological signals that are often collected and analysed in this process include electroencephalogram (EEG), electrocardiogram (ECG) and electrooculogram (EOG) to provide a comprehensive assessment of workload [5]. The physiological signal detection method is used to study the workload by analysing the physiological signal data. Compared with other methods, this method offers a quantitative and objective analysis of pilot workload.

As a contemporary physiological measuring instrument, EEG presents superior temporal resolution that can accurately measure up to millisecond or even microsecond intervals compared to other physiological signals. The EEG signal contains abundant real-time biological data, making it an integral method for examining workload in the current research phase [6]. Concurrently, it is gradually being utilised to evaluate pilot workload in emergencies. Noel et al. [7] used an artificial neural network to classify feature combinations of physiological signals, namely ECG, EEG and eye movements, from pilots at different flight times, which improved the data collection efficiency. Jiahao Fu et al. [8] investigated the changes in EEG load under different task combinations and proved that both computational difficulty and sub-tasks affect the cerebral workload, and the EEG signals change with the workload. Mohanavelu et al. [9] constructed a classifier using machine learning algorithms to efficiently pilot workloads during the take-off, cruise and landing phases, based on ECG and EEG recognition. Guo Z. et al. [10] extracted the entropy value based on the performance of EEG signals in different channels, constructed the corresponding algorithm, and were able to recognise the workload of car drivers better. Huynh et al. [11] verified the feasibility of using deep learning algorithms to train a model that classifies the brain states using the EEG data. Ghaderi et al. [12] investigated the pilot in-flight simulation attention levels. They found that pupil dilation and EEG can be used together to examine pilot behaviour and that both are effective workload measures. Wang et al. [13] designed an airport traffic pattern flight simulation experiment. They used the ECG physiological data and NASA-TLX psychological data obtained from this experiment, combined with RNN and LSTM, to construct a workload recognition model for pilots, which provides a useful reference for the prevention and reduction of human errors caused by workload in flight missions. Asgher et al. [14] showed that LSTM performed well in multi-task load classification of MWL-fNIRS brain signals, with classification accuracies ranging from 83.11% to 95.51%. Hogervorst et al. [15] measured the workload of different levels of N-back tasks using four types of measurements, namely EEG, ECG and eye movement, and proved that the closest relationship between EEG signals and workload was found. Current research commonly gathers different physiological data under normal flight conditions to probe the relationship between these data and workload. Some studies artificially inflated the workload by introducing unlikely tasks in actual flight situations. However, few studies have examined high-stakes and real-life situations, such as engine failure prompting a forced landing.

Most current studies in this field recruit participants without real flight experience, which may lead to deviations between the experimental results and actual flight scenarios [16]. Whether participants have real flight experience is critical for simulation flight research, as pilots with real aircraft experience demonstrate more accurate situational awareness and emergency response capabilities, and their operational behaviours align more closely with real flight logic [17]. Therefore, this study recruited pilots with real aircraft flight experience to participate in the experiment.

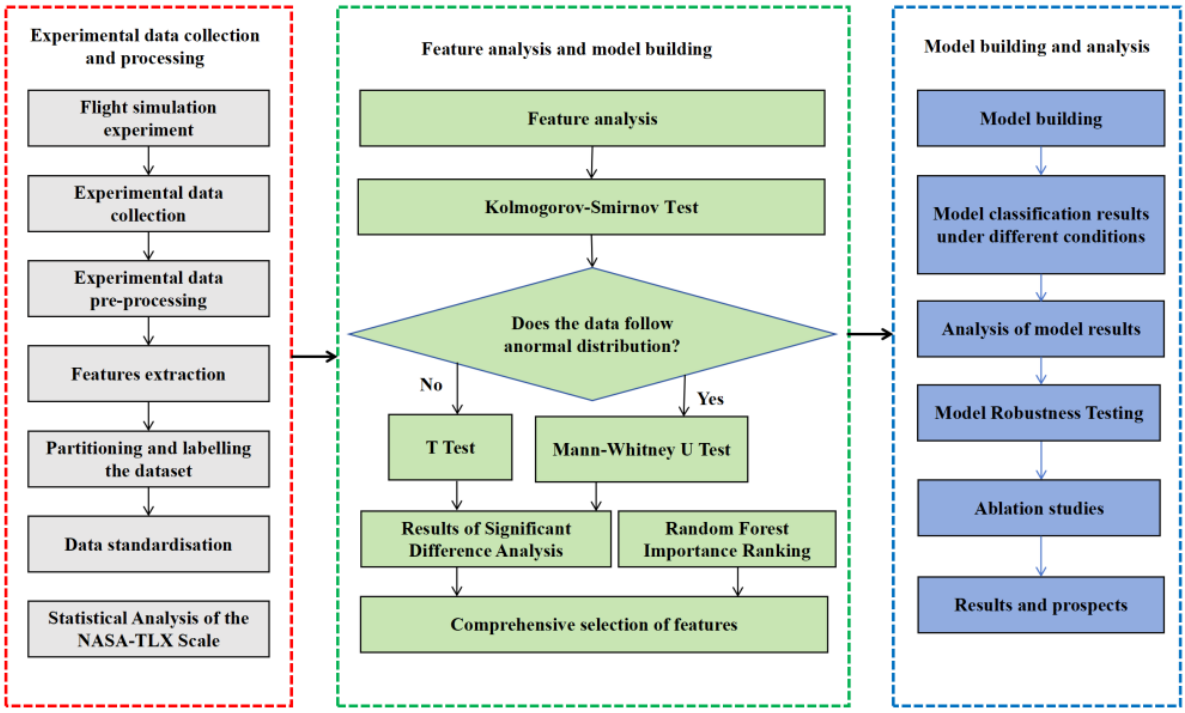


Figure 1 – Research flow chart

Figure 1 shows a flowchart of the study. The objective of this research is to employ EEG as a tool for assessing the workload of pilots during emergency landings due to engine failure compared to routine landing procedures. To get closer to the real data, we chose to conduct the experiment using the X-PLANE simulation and operation software, which is recognised as the world's leading flight simulation and operation software. It is closer to actual failure simulation and has been widely used in aircraft design, performance testing and flight training [18]. In this study, simulated flight experiments were designed to simulate both engine failure forced landings and normal training landings, collected EEG physiological data and NASA-TLX subjective scale data from these experiments, which then facilitated the categorisation of workload into high and low, further extracting the time and frequency-domains along with nonlinear and multidimensional correlation features from the EEG data. The mentioned feature values were evaluated and selected by combining the results of random forest importance ranking and statistical analysis of significance. Finally, based on the selected EEG physiological feature metrics, machine learning algorithms were constructed to build a workload assessment model and compare the performance of different models. Subsequently, the model was tested for robustness and ablation learning was performed based on different brain region feature divisions to explore the contribution of different brain regions to pilot workload recognition during engine failure and forced landing. This preliminary study allows for a more intuitive identification of the physiological conditions of flight cadets during engine-failure-forced landings, thus providing an important reference for enhancing training programs for flight cadets and future aviation safety.

2. MATERIALS AND METHODS

In the introduction, we emphasised the critical importance of pilot workload for flight safety, especially during emergency forced landings due to engine failure. To better understand this issue, this study employed experimental methods to collect and analyse relevant data. The following section details the experiment's design, participants, equipment and procedures, establishing a foundation for the subsequent data analysis.

2.1 Participants

Twenty-one healthy male flight cadets from the Civil Aviation Flight Academy of China participated in this experiment, and their average age was 22.5±3.5 years. They have flown an average of more than 230 hours in the simulators and real aeroplanes. The Edinburgh Sharpshooter Survey showed that all the participants were right-handed. The subjects had normal or corrected-to-normal vision and hearing. None of the

participants took drugs, consumed coffee or caffeinated beverages, or consumed alcohol or alcoholic beverages 24 hours prior to the start of the experiment. All participants agreed to the use of a physiological signal collection device before the experiment, had read and signed a consent form before the experiment, and were subsequently given a participation fee. This study complied with the principles of the Declaration of Helsinki and was approved by the Ethical Review Board of Southwest Jiaotong University (SWJTU-2109-001-QT). All participants read and signed an informed consent form prior to the start of the experiment.

2.2 Experimental equipment

X-PLANE simulation software was chosen for the experiment, and the model of the simulator was a Cessna172SP Skyhawk. The simulator implemented the corresponding internal instrument functions, which allowed the pilot to accomplish more realistic flight tasks. The display system simulated the real field of view seen by the pilot during flight. It can realise the real-time dynamic interactive effect between the pilot and virtual scenes in flight.

2.3 EEG signal acquisition device

The EEG signal acquisition equipment used in this experiment was a 5-channel portable EEG device, Emotive Insight, manufactured by Emotive Bioinformatics Inc. of the United States of America, which has the advantages of high safety, durability and ease of use, with a sampling rate of 128 Hz. The five transducers were distributed in accordance with the international 10-20 EEG system and were placed in the temporal lobe (T7, T8), frontal lobe (AF3, AF4) and occipital lobe (Pz). The experimental setup is illustrated in *Figure 2*.

Figure 2 – Flight simulator and the Emotive Insight

2.4 Experimental procedures

The subjects were randomly lined up, and the experiments were conducted one by one. A subjective questionnaire was administered at the end of the experiment. All subjects were required to wear the Polar Verity Sense heart rate measurement device and Emotive Insight EEG data acquisition device throughout the experiment. Each subject was asked to complete the NASA-TLX scale during all the experiments based on actual subjective situations. The civil aviation simulation experiments in which the subjects participated mainly included the silent state experiment before the formal experiment, the flight simulation experiment under normal conditions of landing and take-off routes, and the experiment of forced landing under the special condition of engine failure. The silent state experiment required subjects to sit still and rest in a simulated cockpit to eliminate the effects of emotions such as nervousness or excitement. To ensure the completeness and reliability of the collected physiological data, all subjects started the experiment at 9:00 a.m. and 2:00 p.m. The experimental procedure is illustrated in *Figure 3*.

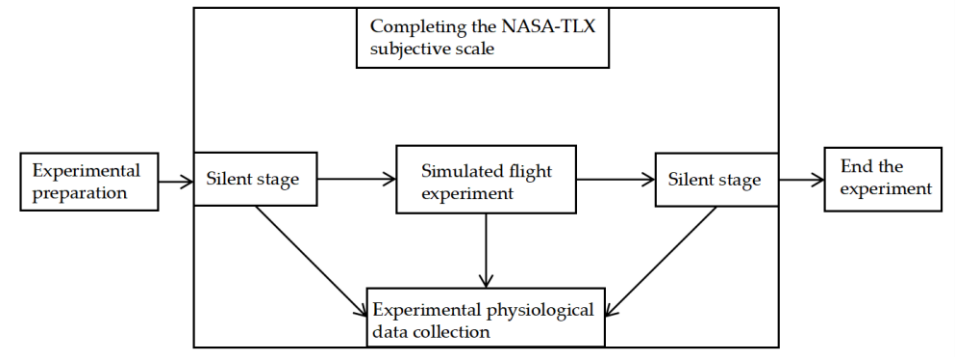


Figure 3 – Schematic diagram of the experimental procedure

The airport of this simulation experiment was runway 19 at Beijing Capital International Airport. The environment was set as clear sky, increasing the side wind from the ground to 3,000 feet, with a wind speed of 15 knots, and the wind direction was specified. We set the normal take-off and landing route flight simulation experiment as in Experiment 1, and the engine failure unpowered forced landing experiment as in Experiment 2.

Participants were required to pilot the Cessna 172SP using the cockpit view from Runway 19 at Capital International Airport. The flight consists of visualising a left take-off and landing circuit, with ultimate landing on Runway 19 and no autopilot used. The initial parameters included a maximum flight slope of 30° and a take-off of 180°. A stable speed of 90 knots and a climb rate of 500 feet per minute should be established, continuing until 1,000 feet in altitude. A left turn was initiated with a 20° bank, not exceeding a maximum of 30°. After the first turn, pilots should maintain a track of 90° for the second leg and start the second turn after reaching an altitude of 1,500 ft. The third leg required flying at 360° and maintaining an altitude of 1,500 ft. The third turn should be initiated when the runway entrance is at a 45° angle behind the fuselage on the left, permitting pilots to start descending and decide the timing for the fourth turn while keeping the maximum bank within 30°. The approach requires a descent rate within 500 ft/min, with runway entry more than 50 ft above the runway elevation. With a descent rate of approximately 40 ft/min, grounding should occur with a ground overload of approximately 1.2, with strictly prohibited cross-grounding inhibition.

Experiment 2 required the pilot to fly a Cessna C172SP from Runway 19 at the Capital International Airport (CIA) using the cockpit view. After reaching 1,000 ft, the pilot can start turning to the left and execute a left landing route (east of Runway 19) with a width of approximately 1.3 nautical miles from the runway, an altitude of 1,500 ft at the trim level, a minimum speed of 90 knots, and a stopping procedure triggered by the system within 5 to 30 s after the aircraft passes over the entrance of the runway. The system triggers a stop procedure within 30 seconds, and the aircraft is controlled to land on Runway 19. An autopilot should not be used throughout the flight, and the flight slope should not exceed 30 °. The requirement for the return from an unpowered forced landing is to deal with the engine failure as soon as possible, and the airspeed should be strictly maintained at 65 knots until passing the Runway 19 entrance. During the return, the subjects could choose the appropriate return route and downward trajectory by themselves, and flaps and levelling could be used. The altitude of the runway approach must not be less than 50 feet or greater than 200 feet. At the end of each flight mission, the subjects filled out the NASA-TLX workload scale at the end of the silent state. The simulated flight path diagram is shown in *Figure 4*.

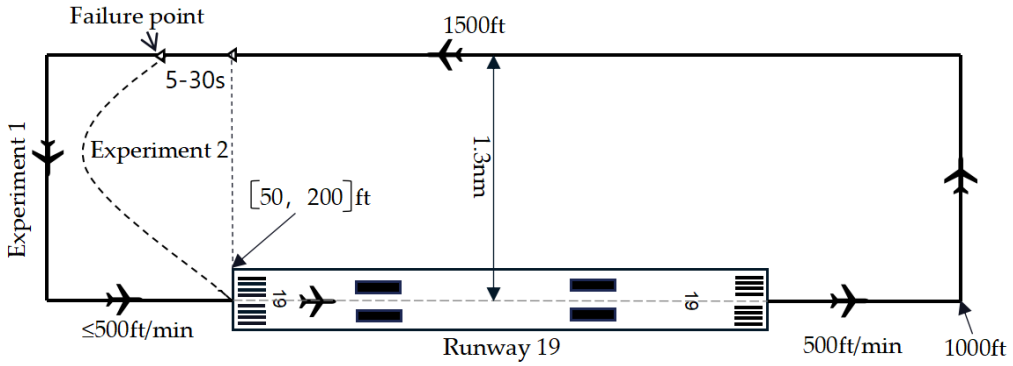


Figure 4 – Simulated flight path diagram

3. DATA ANALYSIS

The experiment collected EEG data and subjective workload ratings from pilots across different flight conditions. This dataset forms a robust foundation for subsequent analyses, enabling an in-depth exploration of pilot workload changes during engine failure-induced forced landings. Detailed data analysis will uncover workload patterns across flight phases and evaluate the machine learning models' effectiveness in workload identification.

3.1 Subjective assessment of workload

In this study, the NASA-TLX scale was used to subjectively assess the workload of pilots, with larger final scores on the scale indicating greater workload [19]. The scale is a multidimensional, comprehensive assessment scale comprising six dimensions: mental demands, physical demands, time demands, effort, performance and frustration [20]. Statistical analysis, including normality testing and significance analysis, was performed using the Statistical Product and Service Solution Software (SPSS). The NASA-TLX scores for the two experiments are shown in *Figure 5*.

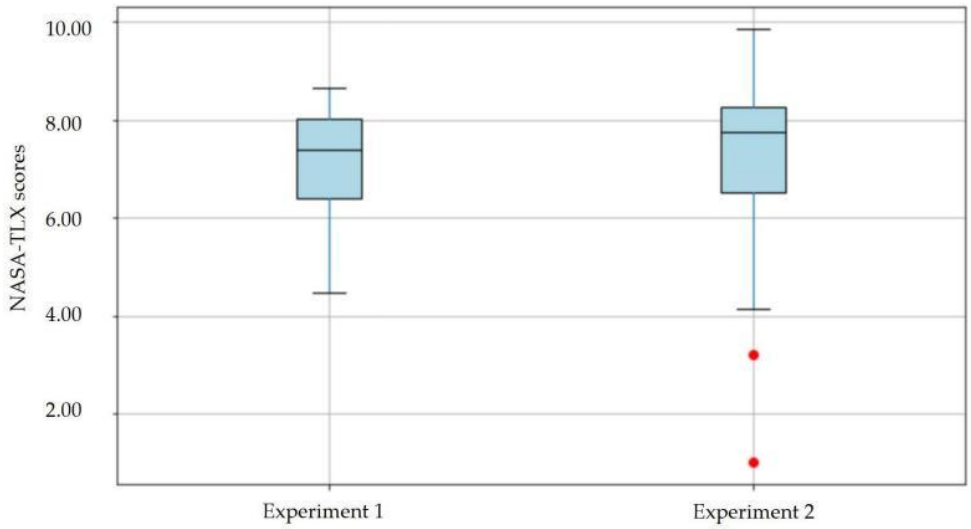


Figure 5 – Boxplot of NASA-TLX scores

Upon confirming engine failure, pilots must immediately execute engine failure handling procedures. Pilots must troubleshoot the engine issue, attempt an engine restart, search for a suitable emergency landing site and maintain aircraft control simultaneously. Furthermore, the urgency and complexity of the tasks demand rapid decision-making and precise execution increases the pilot workload. Therefore, the workload during emergency procedures is significantly higher than that during normal operations.

3.2 Pre-processing of EEG signals

It is critical to process raw EEG signals captured in experiments to obtain clean signals. First, the channels were localised according to the International 10-20 EEG System. Next, the EEG signals were re-referenced with the average of all signals to exclude spatial bias or artefacts that may have resulted from the position of the reference electrodes used during recording. The bad electrodes were then interpolated. The interpolation of bad electrodes involves calculating the potential data points of faulty or missing electrodes based on the values acquired from neighbouring electrodes. Subsequently, the EEG signals were filtered using an FIR filter to obtain 1-40 HZ signals. Finally, an independent component analysis algorithm was used to eliminate artefacts. The EEG signal pre-processing programs were based on the Python open-source package MNE 1.3.0.

3.3 Feature extraction and selection

Continuous EEG data were segmented at 2 s, with a 1-second overlap set. Then, the time-domain, frequency-domain and nonlinear EEG features were extracted for the five channels, as shown in *Table 1*. The time-domain features were extracted as mean, variance, standard deviation, peak-to-peak amplitude, skewness, kurtosis, root-mean-square, zero crossing, Hjorth mobility and Hjorth complexity. The frequency-domain features were extracted as power spectral densities in bands δ (1-3 Hz), θ (4-7 Hz), α (8-11 Hz), β1 (12-20 Hz), β2 (21-29 Hz) and γ1 (30-40 Hz). The nonlinear features extracted in this study are the sample and approximate entropy, which are sensitive to workload.

Table 1 – List of time-domain and nonlinear features

Domain	Feature	Description
Time	Mean	Represents the overall signal level over a specific time window
	Variance (vari)	Shows the dispersion of the signal
	Standard deviation (std)	The square root of the variance indicates the spread of the signal
	Peak to peak amplitude (ptp-amp)	The difference between the maximum and minimum of the signal
	Skewness	Indicates the asymmetry of the signal data distribution
	Kurtosis	Measures the sharpness of the signal data distribution
	Root mean square (rms)	The root-mean-square of the signal amplitude represents the signal energy
	Zero crossings (zc)	Used to count the number of times the signal crosses zero
	Hjorth mobility (hm)	An index of signal activity level is used to evaluate the activity level of the signal [21]
	Hjorth complexity (hc)	An index of signal complexity is used to evaluate the complexity level of the signal
NonLinear	Approximate entropy (apen)	Used to assess the complexity and irregularity of the signal [22]
	Sample entropy (sampen)	Similar to Approximate Entropy, it is used to evaluate the complexity and irregularity of the signal

To improve the accuracy of the assessment of workload classification, this study utilised feature selection methods, including significance analysis and random forests. The obtained features were first tested for normality, and there are two commonly used methods for normality testing: the Kolmogorov-Smirnov test (K-S test) and the Shapiro-Wilk test (S-W test). When analysing small sample data with less than 50 sample sizes, the tendency is to look at the normality test results obtained by the S-W test; when analysing large sample data with more than 50 sample sizes, the tendency is to look at the normality test results obtained by the K-S test [23]. A normality test was performed on the extracted characteristic data using the K-S test with a confidence level set at $P > 0.05$. In addition, the t-test and Mann-Whitney U test are usually used for significant difference analysis. Based on the results of the K-S test, the obtained features were categorised into those that obeyed and those that did not obey the normal distribution. For the features obeying a normal distribution, the t-test was used to select the features that were significantly different under the two experiments, and for the features not obeying a normal distribution, the Mann-Whitney U-test was used to select the features that were significantly different under the two experiments, with the level of significance set at $P<0.05$. All statistical analyses were performed using SPSS software. Subsequently, according to the P-value of the statistical analysis of the significance of the results of the corresponding P-value, topographic maps were drawn with the Matrix Laboratory MATLAB, which can visually observe whether the features of different brain regions have significant differences. *Figure 6* shows the nonlinear features as an example. It can be observed that only the sample entropy of the T8 channel in the nonlinear features is not significantly different.

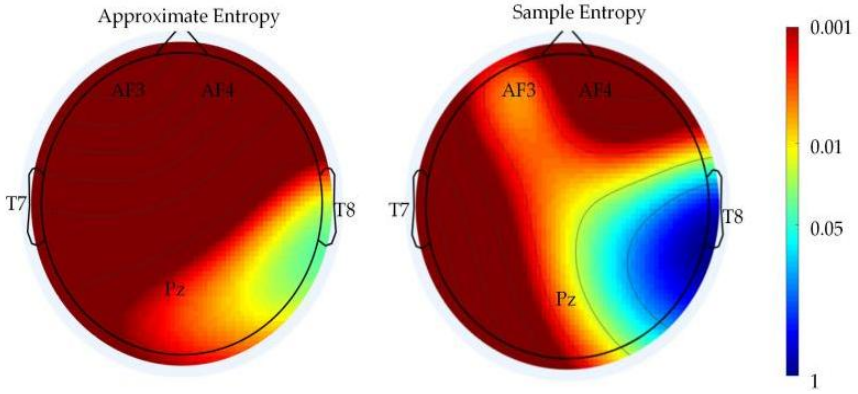


Figure 6 – P-value topographies of nonlinear features

To assess the pilot workload more accurately, the features with significant differences were selected for the study, and they were ranked in order of importance using the random forest algorithm. The results are shown in *Tables 2* and *3*, where the numbers represent the feature importance. The change in the number of features before and after feature selection is shown in *Figure 7*, where the number of features before and after screening varies more in both cases of selecting time-domain features alone and selecting overall features.

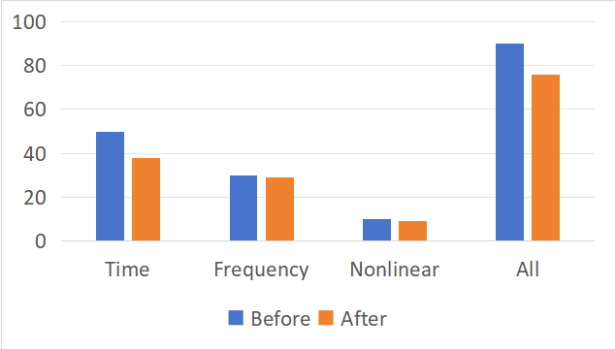


Figure 7 – Quantities before and after feature selection

Table 2 – Results of random forest importance ranking for time-domain, frequency-domain and nonlinear features

	Feature	Importance	Feature	Importance	Feature	Importance
Time-domain	T7-mean	0.063622	AF4-hc	0.023181	AF4-zc	0.012615
	T7-rms	0.053198	AF4-hm	0.022725	T7-kurtosis	0.012087
	T7-ptp-amp	0.049247	AF3-ptp-amp	0.022483	Pz-kurtosis	0.011355
	T7-hm	0.043444	Pz-mean	0.022145	AF4-kurtosis	0.010425
	T8-ptp-amp	0.040979	T8-hm	0.020648	AF3-zc	0.010190
	T7-hc	0.039854	AF3-mean	0.020268	T8-kurtosis	0.009769
	T8-std	0.039337	T8-hc	0.019843	T7-zc	0.007574
	AF3-rms	0.034860	Pz-hc	0.019117	T8-zc	0.006546
	T7-std	0.031283	Pz-hm	0.017928	T7-vari	0.000356
	Pz-ptp-amp	0.030002	AF4-mean	0.017613	T8-vari	0.000097
	AF4-rms	0.026583	AF3-std	0.017189	AF3-vari	0.000089
	AF4-ptp-amp	0.023997	T8-rms	0.013809	AF4-vari	0.000004
	AF4-std	0.023386	AF3-skewness	0.012990		
Frequency-domain	T7-β1	0.066899	T8-γ1	0.037645	T8-δ	0.026038
	T7-γ1	0.057836	AF3-β1	0.034427	T8-α	0.024854
	T7-β2	0.054356	AF3-β2	0.034053	AF4-δ	0.023145
	T7-α	0.048349	AF4-β1	0.033316	T8-θ	0.023107
	AF4-β2	0.045673	T7-θ	0.031643	AF3-α	0.022905
	Pz-β1	0.042498	Pz-γ1	0.030374	AF3-θ	0.021900
	AF4-γ1	0.041666	T7-δ	0.029047	AF3-δ	0.019994
	Pz-β2	0.040916	AF3-γ1	0.028126	AF4-α	0.011593
	T8-β1	0.040796	Pz-θ	0.026283	AF4-θ	0.010929
	T8-β2	0.040566	Pz-α	0.026223		
Nonlinear	AF3-apen	0.11676	T7-apen	0.100563	Pz-sampen	0.094501
	AF3-sampen	0.113601	AF4-sampen	0.096083	T8-apen	0.09136
	T7-sampen	0.101494	Pz-apen	0.095674	AF4-apen	0.090558

Table 3 – Results of random forest importance ranking for all features

Feature	Importance	Feature	Importance	Feature	Importance
T7-β1	0.042199	T7-hc	0.014379	AF4-hm	0.005736
T7-γ1	0.036178	T8-α	0.013887	T8-rms	0.005533
T7-mean	0.035455	T8-δ	0.013439	AF4-zc	0.005318
T7-rms	0.035322	T8-θ	0.012821	T8-hm	0.005263
T7-β2	0.034173	T8-std	0.012477	AF3-std	0.005256
T7-α	0.029636	T8-ptp-amp	0.012385	T7-sampen	0.005122
Pz-β1	0.028574	T7-std	0.010696	T8-hc	0.005073
AF4-β2	0.027861	AF3-α	0.010145	AF4-sampen	0.004741
T8-β2	0.026075	Pz-ptp-amp	0.009593	AF3-zc	0.004643
T8-β1	0.025817	AF3-θ	0.009215	AF4-α	0.004435
Pz-β2	0.024781	AF4-std	0.008924	T8-apen	0.004013
T8-γ1	0.024343	Pz-hc	0.008271	AF4-apen	0.00386
AF4-γ1	0.023434	AF3-mean	0.008116	AF4-θ	0.0041
T7-θ	0.020705	Pz-hm	0.007856	T7-kurtosis	0.003792
AF4-β1	0.019426	AF4-ptp-amp	0.007805	AF3-skewness	0.003686
AF3-β1	0.018441	AF3-apen	0.007792	T7-zc	0.003346
AF3-β2	0.017864	AF3-sampen	0.00772	Pz-kurtosis	0.003244
Pz-γ1	0.017071	Pz-mean	0.007564	T8-kurtosis	0.003122
AF3-rms	0.016826	AF4-mean	0.007169	AF4-kurtosis	0.003024
Pz-α	0.015448	AF3-δ	0.006736	T8-zc	0.001722
AF3-γ1	0.015048	AF3-ptp-amp	0.006678	T7-vari	0.000054
T7-hm	0.014971	Pz-apen	0.00666	AF3-vari	0.000041
T7-ptp-amp	0.014757	AF4-δ	0.006609	T8-vari	0.000035
Pz-θ	0.014672	AF4-hc	0.00649	AF4-vari	0.000003
T7-δ	0.014572	T7-apen	0.006046		
AF4-rms	0.014512	Pz-sampen	0.005991		

3.4 Model building

In this study, we used four different machine-learning algorithms, SVC, KNN, RF and LightGBM, to construct workload assessment models.

1) SVM is a mighty and flexible machine-learning algorithm commonly used to analyse pilot workloads [24]. SVC is a classification algorithm that implements SVM. If the data are not linearly divisible, the kernel function can be used to map the data to the high-dimensional feature space and then search for the hype-plane, which is well adapted to high-dimensional data and smaller training sets. The radial basis function RBF kernel used in this study can be expressed as *Equation 1*, where g is the kernel parameter, which represents the width of the kernel function.

$$K(x_i, x_j) = exp(\frac{Px_i - x_jP^2}{g^2}) \quad (1)$$

2) KNN is an instance-based learning algorithm that performs classification or regression by calculating the distance between the samples to be predicted and all samples in the training set. For the classification task, KNN selects the K training samples closest to the samples to be predicted and determines the class of the samples to be predicted by voting [25]. Where the generalised distance metric is called Minkowski distance, and the defining expression is

$$\boldsymbol{d} = (\sum_{(\boldsymbol{n=1})}^{\boldsymbol{n}} |\boldsymbol{x_i} - \boldsymbol{y_i}|^{\boldsymbol{p}})^{(\boldsymbol{1/p})} \tag{2}$$

where P=1 is the Manhattan distance, and P=2 is the most commonly used standard Euclidean distance. KNN is a parameter-free algorithm that can be adapted to various complex data distributions, but has a high computational overhead for large datasets and high-dimensional data.

3) RF is an integrated learning algorithm that builds a more robust classifier or regression by constructing multiple decision trees [26]. RF uses self-sampling and the random selection of feature subsets to generate multiple decision trees. In a classification problem, each tree votes to select the most categories for the final prediction. In the regression problem, the average of the outputs of each tree was used as the final prediction. Random forest is more robust to noise and outliers, and it can also handle high-dimensional data and large datasets with a low risk of overfitting. In addition, RF can be used to evaluate the importance of features for feature selection.

4) LightGBM is an implementation of the gradient boosting tree framework, which deals with classification and regression problems using efficient decision tree learning algorithms [27]. LightGBM is designed to have high efficiency, low memory footprint and high accuracy, and it employs feature-based discretisation and sample-based sparse feature representation techniques to reduce the memory footprint, increase the training speed and use a leaf-wise growth with depth restriction (leaf-wise) algorithm on top of the regular histogram algorithm.

$$(p_m, f_m, v_m) = arg \min_{(p,f,v)} L(T_{m-1}(X).split(p, f, v), Y) \tag{3}$$

$$T_m(X) = T_{m-1}(X).split(p, f, v) \tag{4}$$

This results in an accelerated ranking of feature values and data segmentation, leading to a faster construction of decision trees.

3.5 Results and discussion

This study statistically evaluated the performance of four machine learning classifiers (SVC, KNN, RF and LightGBM) based on different feature selections from the significance analysis and random forest importance rankings. These are visually represented in *Figures 8–11*. Here, the feature selection span was set from 20% to 100% and then increased by 10%.

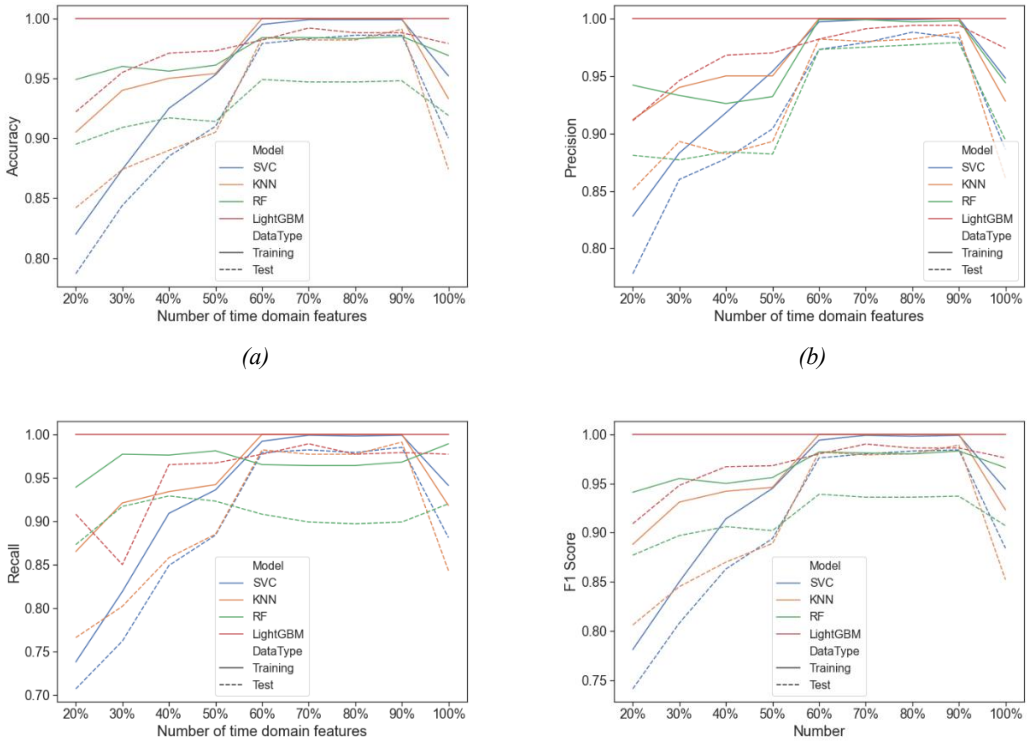


Figure 8 – Results of SVC, KNN, RF and LightGBM when different percentages of features are selected as inputs after ranking the time-domain features with significance in order of random forest importance

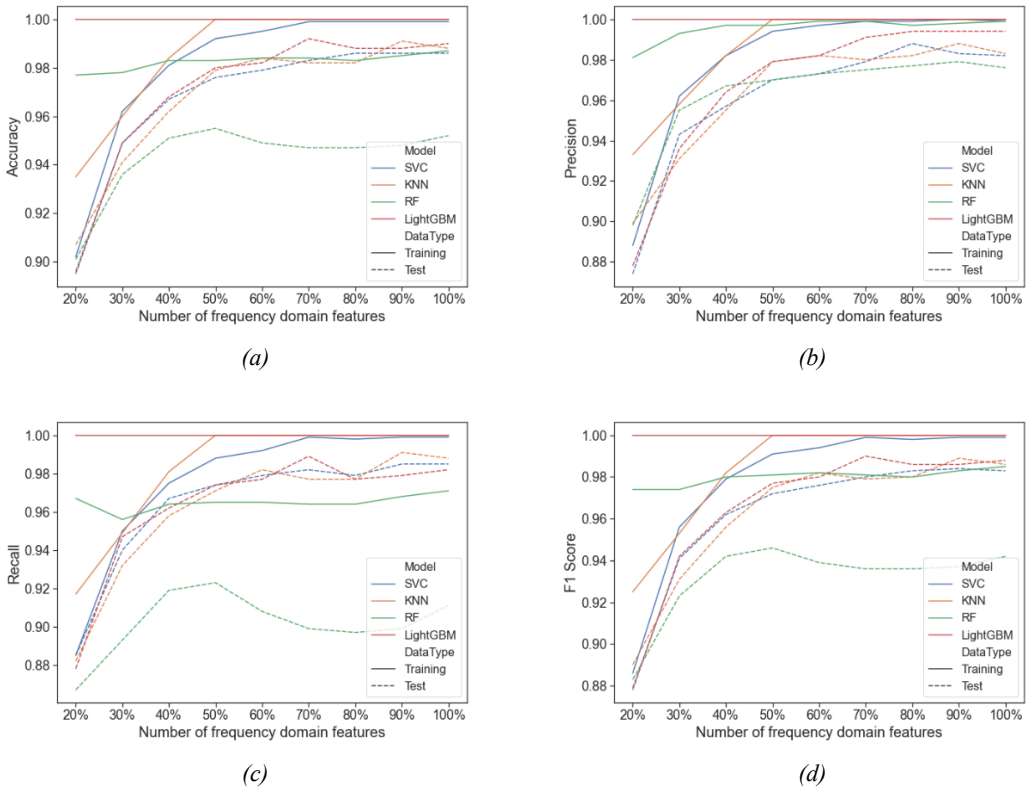
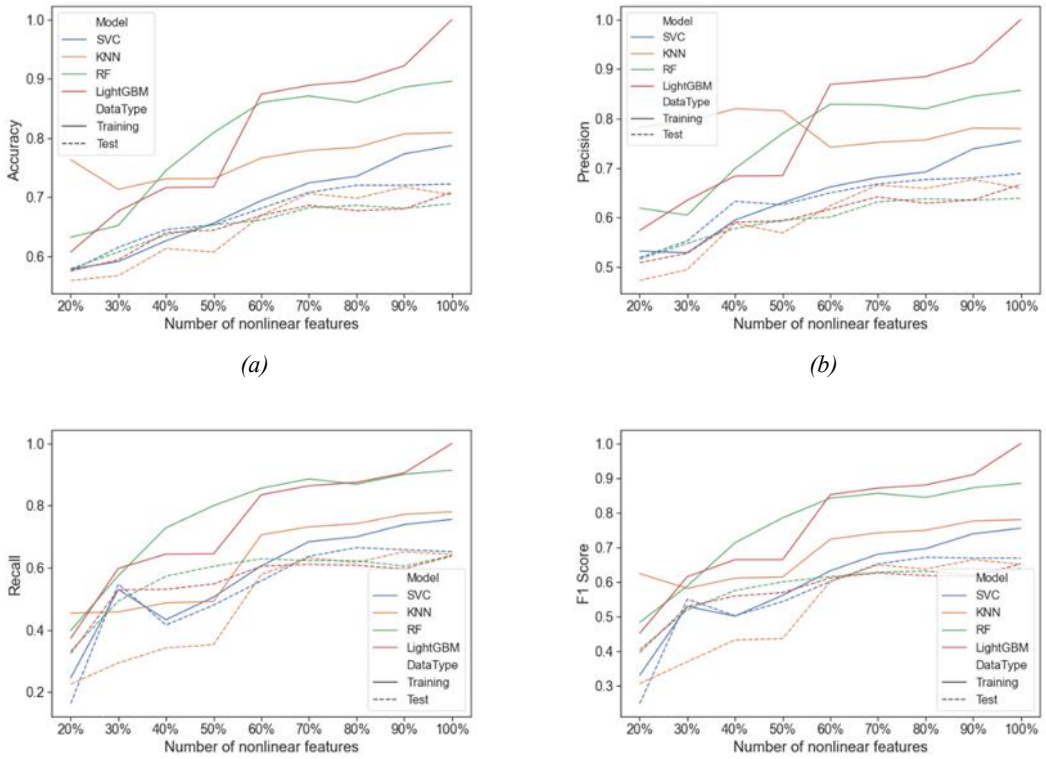


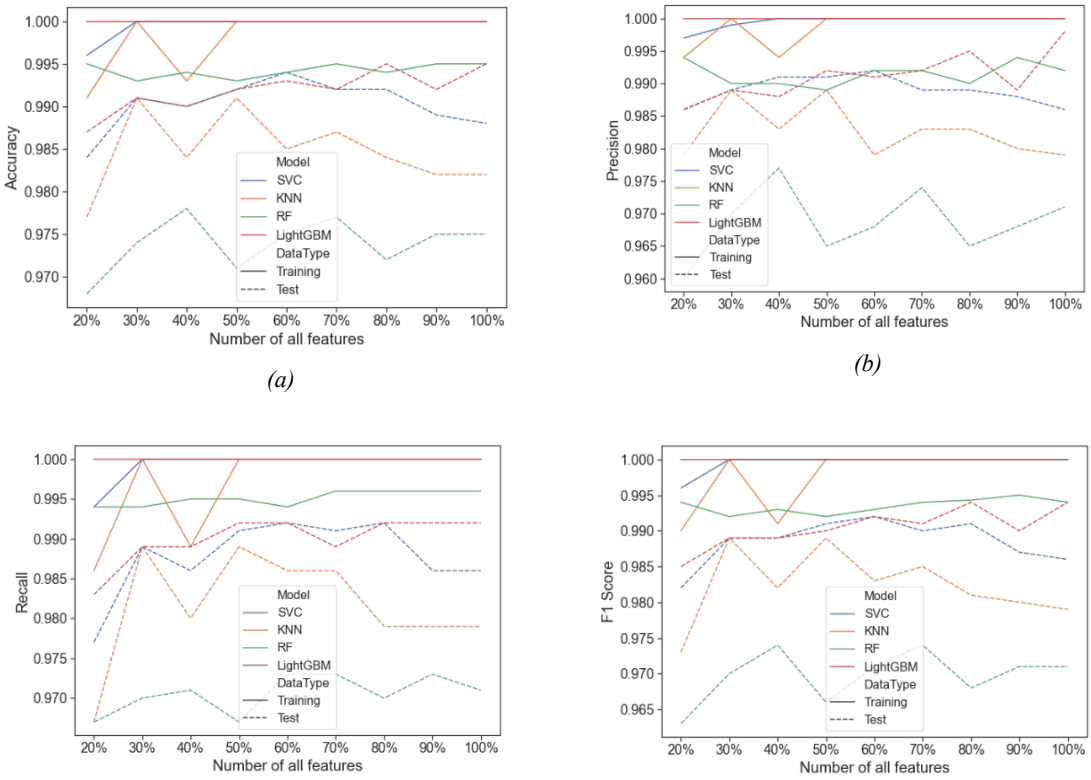
Figure 9 – Results of SVC, KNN, RF and LightGBM when different percentages of features are selected as inputs after ranking the frequency-domain features with significance in order of random forest importance



(a) *(b)*

(c) *(d)*

Figure 10 – Results of SVC, KNN, RF and LightGBM when different percentages of features are selected as inputs after ranking the nonlinear features with significance in order of random forest importance



(a) *(b)*

(c) *(d)*

Figure 11 – Results of SVC, KNN, RF and LightGBM when different percentages of features are selected as inputs after ranking all features with significance in order of random forest importance

In this study, the NASA-TLX score of Experiment 2 was slightly higher than that of Experiment 1, which indicates that the pilot workload is greater in engine failure forced landing than in normal landing. However, after the significance analysis, it was found that there was no significant difference between the scale scores of these two experiments ($p>0.05$), which also reflects the limitations of using subjective assessment methods to identify assessment workloads that are susceptible to the impression of objective factors, such as individual differences.

When distinct time-domain or frequency-domain features serve as data samples, LightGBM outperforms other constructed machine learning models in terms of workload identification accuracy. By selecting the top 70% of features from either the time-domain or the frequency-domain, the computational time decreases, and high accuracy rates are maintained. Specifically, the model achieved an accuracy of 99.2% in the time and frequency domains. The feature distributions are: 7 T7, 5 T8, 5 AF3, 7 AF4, 4 PZ in the time domain and 6 T7, 4 T8, 3 AF3, 3 AF4 and 5 PZ in the frequency domain. By using all features within the time domain, an accuracy of 97.9% can be achieved, whereas using all frequency domain features yields a 99% accuracy rate. However, utilising all the features increases the computational time required. Using significantly different nonlinear features, the SVC model achieved a peak accuracy of 72.2% using the initial full feature set. This comprised three features from T7 and T8, four from AF3 and AF4, and two from the PZ. Furthermore, the model performs optimally when all three feature classes are assessed collectively; the test set results show minimal differences compared with the training set. Using the top 80% of all features (time domain, frequency domain and nonlinear) as inputs allowed the LightGBM model to achieve an apex accuracy of 99.5%.

Doyon et al. [28] showed that the region where the temporal lobe is located also plays a crucial role in the whole visual processing, and the temporal lobe is involved in the high-level processing and integration of visual information, which is one of the critical structures for understanding and recognising the visual world. At the same time, the pilots mainly scan the cockpit displays visually to obtain the information they need during the experiments, so different visual processing situations will affect the workload of the pilots when they face different experiments, which directly affects pilot decision-making and thus flight safety [29][30]. A recent study by Chikhi et al. [31] found that the assessment of workload based on EEG analysis of information in the θ, α and β bands was the most sensitive. Based on the results of this study, it can be found that various models show a strong ability to assess workload with high accuracy, precision, recall and F1 scores for both time and frequency domain features and that the accuracy of model recognition to assess workload is the highest when frequency domain features are used as the data samples. However, for nonlinear features, the performance of the model is relatively weaker, possibly because the processing of nonlinear features is usually more complex than that of time and frequency domain features. This poses a greater challenge for the model and may also be caused by the small number of nonlinear features selected in this study. This also suggests that the change in pilot workload during an unpowered forced landing is most closely related to the frequency domain features. Similar results were obtained by Zhang et al. [32]. In addition, it can also be found here that by using more features as inputs, the model is not necessarily better. However, on the contrary, it may also have a negative impact on the computational speed and accuracy of the model. This validates the findings of Ayinde et al. [33].

In addition, this study demonstrated the feasibility of using EEG data captured with Emotive Insight to identify different levels of workload assessed in pilots. This was confirmed by Jafet et al. [34] and Li et al. [35]. In addition, according to the statistical results of the NASA-TLX scale, pilots have a greater workload during engine failure landings than during normal landings because they have to perform more complex tasks and maintain a high level of concentration at all times, which is consistent with the earlier findings of Bernhardt et al. [36]. This result also helps pilots rationally allocate their attention during training and improve their stress resistance when facing special situations to adjust their state and handle them correctly scientifically. Meanwhile, from the results of the subjective scale, the difference between the workload during forced landing and normal landing was not apparent. However, it has a high accuracy rate when distinguishing the workload of these two situations through physiological data, EEG signals, which indicates that there is a large difference between the workload during forced landing and normal landing. Therefore, this study also has some limitations when using the NASA-TLX for subjective workload assessment.

4. CONCLUSIONS

The participants in this study were flight cadets with practical flight experience, enabling them to demonstrate situational awareness and emergency response capabilities more accurately, which allowed for a

more objective assessment of pilot workload. The study employed 5-channel EEG equipment to collect brain data during both simulated engine failure forced landings and normal landings. The objective was to understand the relationship between pilot workload and EEG signals during engine failure. Models assessing workload were constructed using machine learning, with performance metrics including accuracy, precision, recall and F1 scores.

Additionally, the model performance was assessed using ROC combined with AUC. It was found that the changes in different brain regions under different workload states were differentiated, and the input feature samples from the temporal lobe were the most numerous when the model reached the highest accuracy. When the three types of time domain and frequency domain nonlinear features are considered individually, the best result is based on the frequency domain features to assess the pilot workload during forced landing under special circumstances of engine failure, which is very close to that obtained by the combined consideration of the three types of features. This study also verifies the feasibility of using EEG data to assess the workload of pilots in simulated flight scenarios, which can help us better study the changes in pilot workload during real flight.

While this study achieved certain results, it also has certain limitations. First, this experiment was completed based on a flight simulation, which is still some way off from actual engine failures and forced landings. Second, in the subjective assessment of workload, individual differences were ignored, and only the results of the obtained scales were analysed uniformly. In the future, we aim to study the data of more diverse pilots facing special situations in real flight situations to analyse the workload. We will try to validate some of the findings of this study by combining the knowledge of neurology, and then continuously improve the model system. Ultimately, through the integration of comprehensive flight data and training modules, we can improve the pilot's ability in special situations and enhance aviation safety.

ACKNOWLEDGEMENT

We would like to thank the two foundations of the Key Laboratory of Flight Techniques and Flight Safety, CAAC (Grant No. F2024KF10C), China Eastern Airlines Technology Application Research and Development Centre Co., Ltd (Grant No. KC-202309-0041).

REFERENCES

[1] Ng CBR, Bil C, O'Bree T. An expert system framework to support aircraft accident and incident investigations. *The Aeronautical Journal*. 2021;125(1289):1131-1156. DOI: 10.1017/aer.2021.11

[2] Aviation Safety Office of CAAC. CAAC aviation safety report 2017. *Beijing: Civil Aviation Administration of China*; 2018.

[3] Charles RL, Nixon J. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*. 2019;74:221-232. DOI: 10.1016/j.apergo.2018.08.028

[4] Wang L, Wang X, He X. Research on the system of factors influencing the workload of civil aviation pilots. *Ergonomics*. 2016;22:45-48. DOI: 10.13837/j.issn.1006-8309.2016.03.0009

[5] Anders C, Arnrich B. Wearable electroencephalography and multi-modal mental state classification: A systematic literature review. *Computers in Biology and Medicine*. 2022;150:106088. DOI: 10.1016/j.compbiomed.2022.106088

[6] Gjoreski M, et al. Cognitive load monitoring with wearables–lessons learned from a machine learning challenge. *IEEE Access*. 2021;9:103325-103336. DOI: 10.1109/access.2021.3093216

[7] Noel JB, Bauer Jr KW, Lanning JW. Improving pilot mental workload classification through feature exploitation and combination: A feasibility study. *Computers & Operations Research*. 2005;32(10):2713-2730. DOI: 10.1016/j.cor.2004.03.022

[8] Fu J, et al. EEG-based brain load study of multi factorial cognitive tasks. *Aerospace Medicine and Medical Engineering*. 2020;33(1):35-44. DOI: 10.16289/j.cnki.1002-0837.2020.01.006

[9] Mohanavelu K, et al. Machine learning-based approach for identifying mental workload of pilots. *Biomedical Signal Processing and Control*. 2022;75:103623. DOI: 10.1016/j.bspc.2022.103623

[10] Guo Z, et al. A method for recognizing drivers' brain load level based on EEG entropy value. *Journal of Southeast University (Natural Science Edition)*. 2015;45(05):980-984. DOI: 10.3969/j.issn.1001-0505.2015.05.028

[11] Huynh P, Warner G, Lin H. Effects of EMD and feature extraction on EEG analysis. *Journal of Advances in Information Technology*. 2020;11(1). DOI: 10.12720/jait.11.1.26-34

[12] Ghaderi M, et al. Estimation of piloting attention level based on the correlation of pupil dilation and EEG. *International Conference on Intelligent Tutoring Systems. Cham: Springer Nature Switzerland*; 2023:381-390.

[13] Wang H, et al. Cognitive load identification of pilots based on physiological-psychological characteristics in complex environments. *Journal of Advanced Transportation*. 2020;(1):1-16. DOI: 10.1155/2020/5640784

[14] Asgher U, et al. Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Frontiers in Neuroscience*. 2020;14:584. DOI: 10.3389/fnins.2020.00584

[15] Hogervorst MA, Brouwer AM, Van Erp JBF. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience*. 2014;8:322. DOI: 10.3389/fnins.2014.00322

[16] Zhang C, et al. Variation of pilots' mental workload under emergency flight conditions induced by different equipment failures: a flight simulator study. *Transportation Research Record*. 2024;2678(4):365-377. DOI: 10.1177/03611981231184188

[17] Santos S, et al. The Effect of expertise during simulated flight emergencies on the autonomic response and operative performance in military pilots. *International Journal of Environmental Research and Public Health*. 2022;19(15):9141. DOI: 10.3390/ijerph19159141

[18] Kusumoputro B. Data acquisition for rocket model identification with aircraft simulator X-Plane and MATLAB// *2019 16th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering. IEEE*; 2019:1-4.

[19] Hart SG. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload/Elsevier*. 1988. DOI: 10.1016/S0166-4115(08)62386-9

[20] von Janczewski N, et al. A subjective one-item measure based on NASA-TLX to assess cognitive workload in driver-vehicle interaction. *Transportation Research Part F: Traffic Psychology and Behaviour*. 2022;86:210-225. DOI: 10.1016/j.trf.2022.02.012

[21] Sasidharan D, Venugopal G, Swaminathan R. Complexity analysis of surface electromyography signals under fatigue using hjorth parameters and bubble entropy. *Journal of Mechanics in Medicine and Biology*. 2023;23(06):2340051. DOI: 10.1142/S0219519423400511

[22] Zhang M, et al. Aerodynamic system instability identification with sample entropy algorithm based on feature extraction. *Propulsion and Power Research*. 2023;12(1):138-152. DOI: 10.1016/j.jppr.2022.02.004

[23] Tabesh H, Heidari A, Saki A. Normality assessment, a substantial appraisal in medical studies: A simulation study for power comparison of various types of normality tests. *Current Journal of Applied Science and Technology*. 2014;4:2646-2660.

[24] Feng C, et al. A comprehensive prediction and evaluation method of pilot workload. *Technol Health Care*. 2018;26(S1):65-78.

[25] Duru AD. Determination of increased mental workload condition from EEG by the use of classification techniques. *International Journal of Advances in Engineering and Pure Sciences*. 2019;31(1):47-52.

[26] Kose MR, Ahirwal MK, Atulkar M. Dynamic characterization of functional brain connectivity network for mental workload condition using an effective network identifier. *International Journal of Information Technology*. 2023;15(1):229-238. DOI: 10.1007/s41870-022-01151-0

[27] Lei S, et al. Net load segmented forecasting method for data center based on GS-LightGBM model// *2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT)*. IEEE; 2023:1-6.

[28] Doyon J, Milner B. Right temporal-lobe contribution to global visual processing. *Neuropsychologia*. 1991;29(5):343-360. DOI: 10.1016/0028-3932(91)90024-3

[29] Yu C, et al. Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, Space, and Environmental Medicine*. 2014;85(7):708-714. DOI: 10.3357/asem.3847.2014

[30] Hsu CK, Lin SC, Li WC. Visual movement and mental-workload for pilot performance assessment// *Engineering Psychology and Cognitive Ergonomics: 12th International Conference, EPCE 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings 12. Springer International Publishing*; 2015:356-364.

[31] Chikhi S, Matton N, Blanchet S. EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology*. 2022;59(6):e14009. DOI: 10.1111/psyp.14009

[32] Zhang Y, et al. Recognising drivers? mental fatigue based on EEG multi-dimensional feature selection and fusion. *Biomedical Signal Processing and Control*. 2023;79:104237. DOI: 10.1016/j.bspc.2022.104237

[33] Ayinde BO, Inanc T, Zurada JM. Redundant feature pruning for accelerated inference in deep neural networks. *Neural Networks*. 2019;118:148-158. DOI: 10.1016/j.neunet.2019.04.021

[34] Rodriguez J, Del-Valle-Soto C, Gonzalez-Sanchez J. Affective states and virtual reality to improve gait rehabilitation: a preliminary study. *International Journal of Environmental Research and Public Health*. 2022;19(15):9523. DOI: 10.3390/ijerph19159523

[35] Ji L, et al. A comprehensive experimental framework based on analysis of the pilots' EEG and NASA-TLX questionnaire in a VR environment. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. 2023;237(7):869-878. DOI: 10.1177/09544119231179842

[36] Bernhardt KA, Poltavski D, Petros T, et al. The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment. *Applied Ergonomics*. 2019;77:83-91. DOI: 10.1016/j.apergo.2019.01.008