Traffic & Transportation

# Pilot Workload During Walk-around Inspection Based on HRV Features

Jiajun YUAN[1], Ligang WANG[2], Lu TIAN[3], Haotian QIAO[4], Qin DONG[5]

Original Scientific Paper
Submitted: 28 May 2025
Accepted: 10 Oct 2025
Published: 27 May 2026

[1] Corresponding author, 1071716402@qq.com, Sichuan Provincial Engineering Research Centre of Domestic Civil Aircraft Flight and Operation Support, Flight Technology College, Civil Aviation Flight University of China, Guanghan, China
[2] wangligang810@126.com, Guanghan Branch, Civil Aviation Flight University of China, Guanghan, China
[3] 984149946@qq.com, Aircraft Repair & Overhaul Plant, Civil Aviation Flight University of China, Guanghan, China
[4] 2622738049@qq.com, School of Transportation & Logistics, Southwest Jiaotong University, Chengdu, China
[5] frances_dq@qq.com, Guanghan Branch, Civil Aviation Flight University of China, Guanghan, China



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

**ABSTRACT**
The pre-flight walk-around inspection represents a critical safety procedure in civil aviation; however, its cognitive and physiological workload mechanisms remain underexplored. This study systematically investigated pilot workload during standardised SR20 aircraft inspections by integrating heart rate variability (HRV) feature analysis, expert evaluation and task performance assessment. Forty-one flight students, stratified by training and experience levels, participated under controlled real-aircraft conditions. HRV signals were collected using wearable optical sensors, and multiple time-, frequency- and nonlinear-domain features were extracted. Feature selection was conducted via statistical testing and random forest ranking, identifying mean HR, maximum HR, standard deviation of HR, mean NNI and median NNI as the most workload-sensitive indicators. Six deep learning architectures were developed and compared under two feature-processing strategies: complete HRV features and PCA-reduced features. Results revealed a clear gradient of workload across experience groups, with experienced pilots exhibiting the lowest load and highest task performance, whereas untrained participants demonstrated the highest load. In terms of classification, the Depthwise CNN achieved the best overall performance (accuracy = 0.9372) with full HRV features, while the CNN-GRU hybrid was most effective (accuracy = 0.9186) after PCA reduction. These findings highlight the importance of aligning feature dimensionality with model architecture and provide an empirical foundation for establishing objective workload monitoring frameworks in aviation safety management.

**KEYWORDS**
workload assessment; HRV; aviation safety; deep learning; walk-around inspection.

## 1. INTRODUCTION

Pre-flight walk-around inspection is a mandatory procedure stipulated in Annexe 6 of the International Civil Aviation Organisation (ICAO) Standards and Recommended Practices, playing a crucial role in aviation safety assurance systems [1]. This procedure requires flight crews to conduct comprehensive inspections of aircraft external structures, critical components and related systems through visual inspection, tactile assessment and instrument-assisted methods prior to takeoff. This inspection process aims to identify and eliminate mechanical failures or abnormal conditions that could potentially lead to flight safety incidents through proactive risk management [2] [3]. However, significant safety risks persist in current walk-around

inspection operations. On one hand, some aviation operators reduce inspection time in pursuit of operational efficiency, resulting in incomplete inspections; on the other hand, the lack of effective supervision mechanisms may lead to perfunctory inspection procedures [4]. The 2024 Alaska Airlines Boeing 737 MAX 9 incident fully demonstrates the severity of this issue, where door plug detachment resulted in injuries to seven passengers. The accident investigation revealed that the failure to identify missing bolts during pre-flight inspection was a critical contributing factor [5]. Walk-around inspection, as a complex multidimensional task involving over 100 parallel visual inspection items [6], poses significant workload challenges for operators. From a physiological perspective, inspection personnel must continuously move, frequently use elevated platforms, and adopt various postures to complete inspection actions, demanding high levels of physical fitness and flexibility. From a cognitive perspective, the inspection process involves the simultaneous execution of multiple tasks, including procedural memory, anomaly identification and risk assessment, imposing high requirements on memory maintenance, attention allocation and professional judgment capabilities. Furthermore, commercial aviation's stringent punctuality requirements add additional time pressure, exacerbating workload intensity.

Currently, pilot workload assessment methods are primarily categorised into three types: subjective, performance and objective assessments. Subjective assessments, represented by NASA Task Load Index (NASA-TLX) [7] and Subjective Workload Assessment Technique (SWAT) [8], while operationally convenient, are susceptible to cognitive bias and emotional factors [9]. Performance assessments indirectly reflect workload by analysing task completion status, but lack sensitivity to physiological and psychological changes during operations. Objective assessments achieve precise workload quantification through monitoring physiological signals such as heart rate (HR), electrocardiography (ECG) and electroencephalography (EEG) [10]. Heart rate variability (HRV), as a core physiological indicator reflecting autonomic nervous system (ANS) activity, represents the antagonistic effects of sympathetic and parasympathetic nerves in cardiac rhythm regulation and has become an important quantitative tool for assessing stress levels and psychological load [11].

In recent years, multiple studies have confirmed the effectiveness of HRV in pilot workload assessment. Wang et al. [12] systematically reviewed the application of HRV in detecting pilot mental workload (MWL), finding significant correlations between HRV indicators and pilot mental workload levels, though inconsistencies exist across different studies. Machine learning models were employed for mental workload classification based on HRV features, with some models achieving accuracies exceeding 90% in binary classification tasks, but lower accuracies in multi-class classification tasks. Alaimo et al. [13] experimentally investigated pilot workload during flight tasks using HRV as an objective indicator and NASA-TLX questionnaire as a subjective assessment tool, finding that pilot workload was significantly higher during approach and landing phases compared to takeoff and climb phases, with HRV indicators (such as low frequency/high frequency ratio and standard deviation of normal-to-normal intervals) showing significant correlations with workload levels. Mohanavelu et al. [14] studied cognitive load in fighter pilots within flight simulator environments, discovering that HRV features (such as very low frequency and total power) exhibited statistical significance across different flight phases and task load conditions, particularly under low visibility and additional cognitive task conditions, where low frequency normalised units and high frequency normalised units features could significantly distinguish task load variations. Mansikka et al. [15] investigated the relationship between HR and HRV and task performance in fighter pilots during simulated flight tasks, finding that HR and HRV indicators could effectively distinguish pilot performance levels under different task demands, with the mean of RR intervals indicator particularly effective in distinguishing between high-performance and sub-standard flight tasks. Koskelo et al. [16] collected continuous ECG data from flight students during simulated flight tasks, observing significant decreases in most HRV indicators with increasing cognitive workload. Cao et al. [17] recruited 30 active commercial airline pilots to perform flight tasks under different carbon dioxide concentration environments, finding that HRV reduction was associated with age increase, obesity status and high-difficulty operations, with carbon dioxide exposure and HRV both independently affecting pilot performance. Park et al. [18] collected multimodal data from 28 pilots performing various vertical takeoff and landing (VTOL) flight tasks, analysing and inferring behavioural patterns related to pilot performance and perceived workload, finding decreased HRV in high workload tasks. Zhu et al. [19] collected photoplethysmography (PPG) signals from three helicopter pilots across three flight phases during actual flights, extracting HRV features for mental workload identification, with most HRV features showing significant differences among the three mental workload levels. The highest classification accuracy was

achieved using the k-nearest neighbour (KNN) algorithm, while the largest area under the curve (AUC) was obtained using the random forest (RF) algorithm.

Although existing research has confirmed the effectiveness of HRV as an objective assessment indicator for pilot workload, these studies have primarily focused on flight task phases, with insufficient in-depth exploration of workload characteristics and physiological mechanisms during the critical pre-flight walk-around inspection phase. Walk-around inspection, as an important preliminary phase for ensuring flight safety, possesses unique characteristics including parallel multitasking, strict time constraints and complex information integration, resulting in workload patterns significantly different from flight tasks, necessitating targeted research. Therefore, this study aims to systematically explore pilot workload characteristics during walk-around inspection tasks through integrating HRV feature analysis, professional assessment and performance quantification, providing a theoretical foundation and practical guidance for establishing a scientific and objective walk-around inspection workload assessment system.

## 2. MATERIALS AND METHODS

### 2.1 Participants

This study was approved by the Ethics Review Committee of Civil Aviation Flight University of China (CAFUC) (No: 2024-7), strictly adhered to the relevant provisions of the Declaration of Helsinki, and obtained written informed consent from all participants. Forty-one male flight students were recruited from CAFUC as research subjects, with ages ranging from 19 to 24 years. All participants possessed normal colour vision (no red-green colour blindness or colour weakness), intact auditory and visual perception functions, held Class I medical certificates compliant with the Civil Aviation Administration of China (CAAC) "Regulations on Management of Civil Aviation Personnel Medical Certificates", and had no alcohol consumption or caffeine intake within 24 hours prior to the experiment. Among all participants, 16 were students with approximately 250 hours of actual flight experience, while 25 were students who had only received theoretical training without actual flight or walk-around inspection experience. From the latter group, 12 participants were randomly selected to receive 60 minutes of theoretical learning and procedural practice for walk-around inspection. The standardised 60-minute training program was divided into a 30-minute theoretical session and a 30-minute practical session. In the theoretical component, 15 minutes were allocated to instruction on 62 specific items and corresponding acceptance criteria of the SR20 aircraft preflight inspection, 10 minutes to the demonstration of abnormality recognition methods, and 5 minutes to the introduction of time management strategies. During the practical component, a certified flight instructor first performed a complete preflight inspection on an actual SR20 aircraft while explaining key operational points. Subsequently, trainees independently conducted two inspection rounds using the standardised checklist.

### 2.2 Experimental equipment

This study was conducted in the hangar of the Guanghan Branch of CAFUC, utilising a real aircraft environment to ensure the ecological validity of experimental results. The Cirrus SR20 general aviation aircraft was selected as the walk-around inspection subject (as shown in *Figure 1*). This aircraft model is widely used for pilot training due to its advanced safety systems, modern avionics equipment and excellent fuel efficiency, providing a standardised inspection environment for this study. The hangar environment consisted of an enclosed indoor space with appropriate temperature and humidity conditions, eliminating the influence of weather factors and environmental noise on experimental results while ensuring stable operation of data collection equipment. The experimental site layout was strictly designed according to standard walk-around inspection procedures, ensuring that participants could complete the entire walk-around inspection protocol. This setup aimed to maximally simulate the cognitive load and physiological responses of pilots performing walk-around inspection tasks under actual operational conditions, providing effective task-induced conditions for subsequent HRV analysis.

Physiological data collection was performed using the Polar Verity Sense optical HR monitoring armband, as shown in *Figure 2*. This device is equipped with a second-generation multispectral PPG sensor array, utilising a spectral combination of 530/660/880 nm wavelengths to achieve non-invasive continuous monitoring of HRV. The device maintains an internal sampling frequency of 135 Hz [20], ensuring high-precision raw signal acquisition. Processed data are transmitted to external recording devices via Bluetooth protocol, with PPG data transmission frequency at approximately 1 Hz, while pulse-to-pulse interval (PPI) data transmission frequency

is dynamically adjusted based on HR, typically around 1 Hz (corresponding to approximately 60 bpm HR) [21][22]. The monitoring device was worn on the participant's left forearm, positioned 5–10 cm from the wrist, ensuring close contact between the sensor and skin without impeding blood circulation. The wearing tightness was adjusted to allow insertion of one finger.

*Figure 1 – SR20 aircraft*


*Figure 2 – Polar Verity Sense sensor*

## 2.3 Experimental task

The standardised pre-flight walk-around inspection procedure for the Cirrus SR-20 aircraft was employed, requiring participants to execute tasks strictly according to the predetermined sequence of inspection points, as shown in *Figure 3*. The inspection process commenced from the left fuselage and proceeded clockwise in the following order: left fuselage, empennage, right fuselage, right wing trailing edge, right wingtip, right side of nose, nose landing gear and propeller, left side of nose, left landing gear and left wing, left wingtip and left wing trailing edge. The entire inspection protocol encompassed 62 specific inspection items, each with clearly defined inspection standards and acceptance criteria, ensuring systematicity, completeness and standardisation of the inspection process, thereby providing a unified task benchmark for subsequent cognitive load and physiological response analysis.

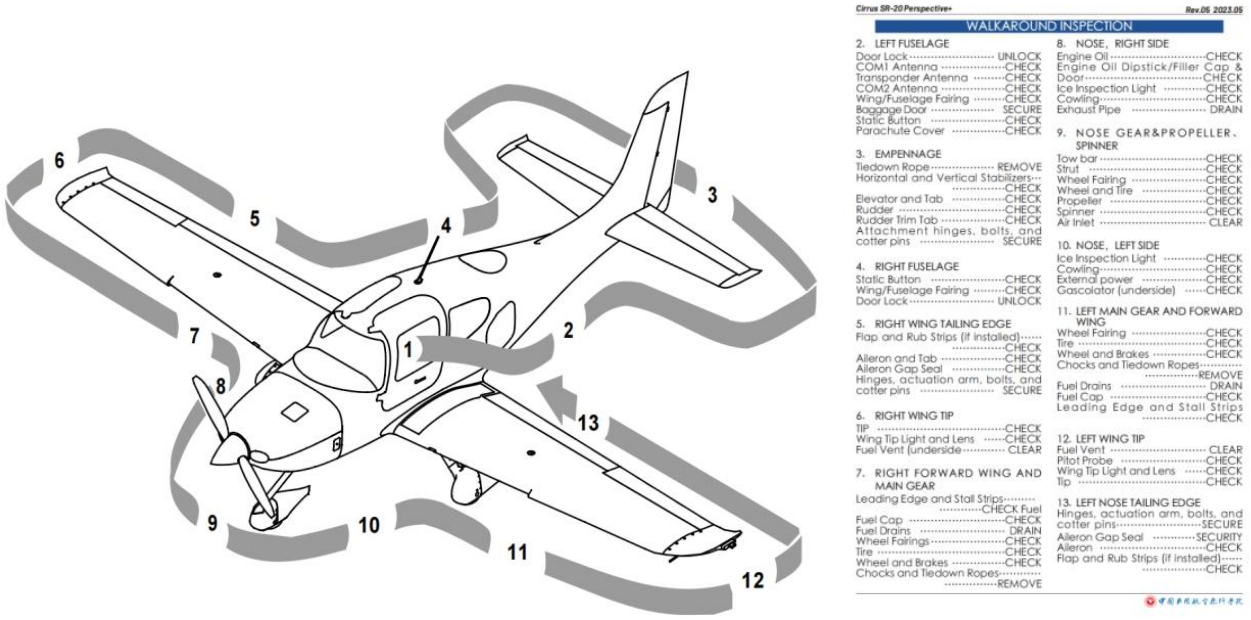


*Figure 3 – Walk-around inspection route and checklist*

### 2.4 Experimental procedure

All participants received standardised pre-flight walk-around inspection training prior to formal experimentation, including inspection procedures, technical standards and safety requirements. All participants underwent a 5-minute adaptive wearing period before the formal experiment to ensure equipment stability and participant comfort, with experimental personnel confirming normal data transmission. Based on participants' flight experience and training background, a stratified randomisation method was employed to evenly distribute them across three SR20 aircraft. Participants strictly followed the standard inspection route illustrated in *Figure 3* for the walk-around inspection, while an assessor with extensive flight instruction experience scored participants' inspection performance in real-time according to the scoring criteria specified in *Figure 4*. High-definition video equipment was used to record participants' inspection behaviours and movement trajectories throughout the entire process. Upon completion of each experiment, raw physiological data from HR monitoring devices, video recording files and assessor scoring sheets were collected, establishing a data foundation for subsequent comprehensive analysis.

**Civil Aviation Flight University of China**

| SR20G6 Aircraft Walk-Around Inspection Scoring Sheet | | | | |
|---|---|---|---|---|
| **Pilot Name** | | **Flight Experience: Yes/No** | | |
| **Start Time - End Time:** | | **Duration:** | | |
| **Category** | **Scoring Criteria** | | **Score** | **Points Earned** |
| Knowledge (K) | Inquire whether the pilot is clear on the specific names and functions of the inspection points during the walk-around inspection, such as: airspeed tubes, static pressure ports, and inspection notes for the landing gear, etc. | | 10 | |
| Skills (S) | Did the pilot understand that there is a specific route for the walk-around inspection? | | 10 | |
| | Did the pilot check that the cabin door locks and functions are operational? | | 10 | |
| | Did the pilot bend down and get on the aircraft to inspect the lower and upper fuselage antennas? | | 10 | |
| | Did the pilot check that the static pressure ports are unobstructed? | | 10 | |
| | Did the pilot check that the condition of the landing gear is good and free of visible lines? | | 10 | |
| | Did the pilot check the oil level of the engines? | | 10 | |
| | Did the pilot check the fuel levels in the left and right tanks and ensure that the tank covers are securely closed? | | 10 | |
| | Did the pilot inspect according to the prescribed route, ensuring that there are no omissions and that any identified issues have been addressed? | | 10 | |
| Attitude (A) | Assess the pilot's understanding of the importance of the walk-around inspection by briefly describing its purpose. | | 10 | |
| | | Total: | 100 | |

*Figure 4 – Walk-around inspection route and checklist*

## 3. DATA PROCESSING

### 3.1 Date preprocessing

Raw HR interval data obtained from the Polar Verity Sense optical HR monitoring device were exported in text format and imported into Excel worksheets for preliminary processing. Data preprocessing included the following steps: (1) identification and removal of missing values and outliers; (2) application of linear interpolation methods to fill data gaps, ensuring completeness of time series and consistency of sampling frequency; (3) temporal window segmentation based on precisely recorded experimental start and end time points to construct complete HRV time series for each participant during the walk-around inspection task.

### 3.2 Feature extraction

HRV feature extraction was performed using the open-source HRV-analysis library [23] in a Python environment. Considering the dynamic characteristics of the walk-around inspection task, a 30-second sliding time window [24] was established with 40% overlap between windows to capture temporal changes in physiological states during task execution. The extracted HRV features encompassed three categories: time-domain, frequency-domain and nonlinear features, as shown in *Table 1*. The extracted multidimensional HRV features comprehensively reflect autonomic nervous system activity states, providing a reliable physiological basis for quantitative assessment of cognitive load and physiological stress levels during walk-around inspection tasks.

*Table 1 – Features and definitions of HRV*

| Features | Definition | Features | Definition |
|---|---|---|---|
| mean_nni | Mean normal-to-normal interval | std_hr | Standard deviation of heart rate |
| sdnn | Standard deviation of normal-to-normal intervals | lf | Low-frequency power |
| sdsd | Standard deviation of differences between adjacent normal-to-normal intervals | hf | High-frequency power |
| pnni_20 | Percentage where the absolute difference between consecutive normal-to-normal intervals > 20 ms | lf_hf_ratio | Ratio of low-frequency power to high-frequency power |
| pnni_50 | Percentage where the absolute difference between consecutive normal-to-normal intervals > 50 ms | lfnu | Normalised low-frequency power |
| nni_20 | Number where the absolute difference between consecutive normal-to-normal intervals > 20 ms | hfnu | Normalised high-frequency power |
| nni_50 | Number where the absolute difference between consecutive normal-to-normal intervals > 50 ms | total_power | Total spectral power |
| ressd | Root mean square difference between adjacent normal-to-normal intervals | vlf | Very low frequency power |
| median_nni | Median normal-to-normal interval | sd1 | Standard deviation 1 from Poincaré plot (short-term variability) |
| range_nni | Range normal-to-normal interval (maximum minus minimum) | sd2 | Standard deviation 2 from Poincaré plot (short-term variability) |
| cvsd | Coefficient of variation for successive differences | ratio_sd2_sd1 | Ratio of standard deviation 2 to standard deviation 1 |
| cvnni | Coefficient of variation for normal-to-normal intervals | csi | Sympathetic nervous index |
| mean_hr | Mean heart rate | cvi | Parasympathetic nervous index |
| max_hr | Maximum heart rate | modified_csi | Modified sympathetic nervous index |
| min_hr | Minimum heart rate | triangular_index | Triangular index |

### 3.3 Feature selection

Considering the relatively limited sample size in this study (n<50), the Shapiro-Wilk test for normal distribution was conducted on extracted HRV features using SPSS 26.0 statistical software. Results revealed that p-values for all HRV features were less than 0.05, indicating a non-normal distribution of the data. Therefore, the Kruskal-Wallis H test was employed to evaluate statistical differences in HRV features among participant groups with different experience levels. To identify HRV features most closely associated with cognitive load, an RF algorithm [23] was utilised for feature importance ranking. As an ensemble learning method, random forest constructs multiple decision trees and aggregates their prediction results, effectively handling small sample data while providing feature importance assessment. Stratified five-fold cross-validation was employed to ensure model generalisability and robustness, with the feature_importances_attribute utilised to quantify the contribution of each HRV feature to cognitive load prediction. The number of decision trees was set to 100, maximum depth to 10 and minimum samples split to 5, balancing model complexity with prediction performance. *Table 2* summarises the statistical significance test results and importance rankings of each HRV feature, providing a quantitative basis for feature selection and interpretation in subsequent cognitive load assessment models.

*Table 2 – Significance and importance of HRV features*

| Features | Significance | Importance | Features | Significance | Importance |
|---|---|---|---|---|---|
| mean_nni | 0.043* | 0.048 | std_hr | 0.000** | 0.048 |
| sdnn | 0.851 | 0.027 | lf | 0.607 | 0.030 |
| sdsd | 0.440 | 0.025 | hf | 0.200 | 0.027 |
| pnni_20 | 0.184 | 0.025 | lf_hf_ratio | 0.120 | 0.024 |
| pnni_50 | 0.095 | 0.030 | lfnu | 0.120 | 0.026 |
| nni_20 | 0.021* | 0.024 | hfnu | 0.120 | 0.026 |
| nni_50 | 0.071 | 0.027 | total_power | 0.854 | 0.029 |
| rmssd | 0.432 | 0.024 | vlf | 0.271 | 0.031 |
| median_nni | 0.002* | 0.053 | sd1 | 0.438 | 0.026 |
| range_nni | 0.314 | 0.030 | sd2 | 0.867 | 0.029 |
| cvsd | 0.065 | 0.028 | ratio_sd2_sd1 | 0.321 | 0.026 |
| cvnni | 0.426 | 0.030 | csi | 0.009* | 0.028 |
| mean_hr | 0.000** | 0.053 | cvi | 0.321 | 0.025 |
| max_hr | 0.000** | 0.078 | modified_csi | 0.761 | 0.028 |
| min_hr | 0.765 | 0.038 | triangular_index | 0.762 | 0.015 |

*$*p < 0.05$; $**p < 0.001$*

### 3.4 Deep learning algorithms

Convolutional neural networks (CNN) is a deep learning model specifically designed for processing data with a grid structure, with core components including convolutional layers, pooling layers and fully connected layers. Convolutional layers extract local features through convolution operations, pooling layers perform dimensionality reduction and feature compression, and fully connected layers complete final classification or regression tasks. The key advantages of CNN include: (1) parameter sharing mechanism reduces model complexity; (2) local connectivity reflects spatial correlation; (3) translation invariance enhances model robustness. CNN processes HRV time-series data through one-dimensional convolution operations, where convolutional kernels slide across the sequence to extract local temporal features, and pooling operations are employed for dimensionality reduction, thereby adapting effectively to physiological signals such as HRV that exhibit local correlations. CNN and its derivatives demonstrate superior performance in automatic feature extraction, long-sequence processing, multi-scale information integration and eliminating the need for handcrafted features, leading to enhanced classification and prediction capabilities, temporal and nonlinear patterns, and thereby enhancing classification and prediction performance.

1) Large kernel CNN

The basic CNN captures local HRV features through fixed-size convolutional kernels, while the large kernel CNN is a direct extension whose core improvement lies in employing larger 1D convolutional kernels to explicitly extract long-term dependencies of HRV. Compared with stacking small kernels, the large-kernel design reduces feature redundancy and enables more efficient modelling of long-period HRV variations. Furthermore, batch normalisation is applied to stabilise feature distributions, and pooling operations are integrated to balance feature preservation with computational efficiency. The fundamental formulation is:

$$Y(c',t)=MaxPool(BN(ReLU(\sum_{c=0}^{C-1}\sum_{k=0}^{K-1}X(c,t-k)\cdot W(c,c',k)+b(c')))) \quad (1)$$

Here, $K$ denotes the convolution kernel size (with $K$=5 or 7 in the large-kernel model), $X(c,t)$ represents the HRV input feature, where $c$ is the channel index and $t$ is the temporal index. $BN(\cdot)$ denotes the batch normalisation operation, and $MaxPool()$ denotes the max pooling operation.

2) Depthwise CNN

As a lightweight variant of conventional CNN, depthwise CNN optimises feature extraction through a two-stage process of depthwise convolution and pointwise convolution. In the first stage, each channel is convolved independently to capture intra-channel local temporal dependencies of HRV. In the second stage, a 1×1 convolution is employed to fuse multi-channel features and establish cross-dimensional correlations. This architecture significantly reduces the number of parameters while preserving critical temporal information, making it well-suited for efficient processing of high-resolution HRV data. The processing pipeline consists of three main steps:

— Depthwise convolution, where each channel is convolved independently to capture intra-channel local temporal dependencies.

$$Y_{dw}(c,t)=ReLU(BN(\sum_{k=0}^{K-1}X(c,t-k)\cdot W_{dw}(c,k)+b_{dw}(c))) \tag{2}$$

— Pointwise convolution, where 1×1 kernels are applied to fuse multi-channel features and establish cross-dimensional correlations.

$$Y_{pw}(c',t)=ReLU(\sum_{c=0}^{C-1}Y_{dw}(c,t)\cdot W_{pw}(c,c')+b_{pw}(c')) \tag{3}$$

— Downsampling, where pooling operations are employed to reduce feature dimensionality and computational complexity while retaining critical information.

$$Y_{pool}(c',t)=\mathrm{m}_{i=0}^{pool-1}Y_{pw}(c',2t+i) \tag{4}$$

3) Residual CNN

To address the vanishing gradient problem in deep CNNs, the residual CNN introduces residual connections, representing a key extension of CNNs in the depth dimension. Its core component is the residual block: the main branch employs consecutive convolution and batch normalisation layers to learn higher-order HRV features, while the residual branch directly reuses the input through an identity mapping to preserve lower-order features. The two branches are fused by element-wise addition, which simultaneously retains fundamental temporal information and captures complex associations, thereby enabling the construction of deeper networks for extracting fine-grained HRV features. The key equation is:

$$ResBlock(X)=MaxPool(Activation(BN(F(X))+X)) \tag{5}$$

where the main branch transformation is:

$$F(X)=BN(Conv1D(K_2,BN(Conv1D(K_1,X)))) \tag{6}$$

4) Multi-scale CNN

For characterising HRV features across different temporal scales, the multi-scale CNN employs parallel multi-scale convolutional branches to achieve comprehensive feature capture, representing a CNN extension with scale adaptability. Specifically, different branches utilise convolution kernels spanning 3, 5 and 7 heartbeat intervals to extract HRV features corresponding to 2–4 s, 4–8 s and 8–12 s, respectively. The multi-scale information is then fused through channel concatenation, followed by a secondary convolution to establish cross-scale correlations, thereby providing a holistic representation of the multi-temporal characteristics of HRV. The key formulation involves three main steps:

— Multi-scale parallel convolution, where convolutional kernels of different receptive fields are applied simultaneously to extract HRV features at multiple temporal scales.

$$Y_K(c',t)=ReLU(\sum_{k=0}^{K-1}X(0,t-k)\cdot W_K(0,c',k)+b_K(c')) \tag{7}$$

— Feature concatenation, where outputs from different branches are combined along the channel dimension to integrate multi-scale information.

$$Y_{merge}(c',t)=\begin{cases} Y_3(c',t) & 0\le c'<32 \\ Y_5(c'-32,t) & 32\le c'<64 \\ Y_7(c'-64,t) & 64\le c'<96 \end{cases} \quad (8)$$

— Secondary fusion, where an additional convolution is performed to establish cross-scale correlations and refine the integrated representation.

$$Y_{final}(c'',t)=BN(ReLU(\sum_{k=0}^{2} Y_{pool}(c',t-k)\cdot W_{merge}(c',c'',k)+b_{merge}(c''))) \quad (9)$$

5) CNN-LSTM hybrid

The proposed model integrates the strengths of CNN and LSTM, where CNN is employed to extract local features and LSTM is utilised to capture long-term dependencies, thereby addressing the dual requirements of local fluctuations and global temporal dynamics in HRV. In this framework, the front-end CNN with a small 3-point kernel extracts short-term variations, while the back-end LSTM leverages its forget, input and output gates to model long-term trends, with the gating mechanism ensuring effective retention of critical temporal information.

— CNN-based local feature extraction

$$F(c',t)=BN(ReLU(\sum_{k=0}^{2} X(0,t-k)\cdot W_{cnn}(0,c',k)+b_{cnn}(c'))) \quad (10)$$

— Temporal modelling with LSTM

LSTM captures long-term dependencies through its gating mechanism, with the core principle being the dynamic update of the cell state. It integrates the roles of the forget gate, the input gate and the output gate. The most critical formulation is the cell state update equation:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (11)$$

where, $f_t = \sigma(W_f \cdot [h_{t-1}, F_t] + b_f)$ is the forget gate, $i_t = \sigma(W_i \cdot [h_{t-1}, F_t] + b_i)$ is the input gate, $\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, F_t] + b_c)$ is the candidate cell state and $h_t = o_t \odot tanh(C_t)$ is the output gate, which represents the final hidden state. Together, these components enable effective memory and updating of long-term temporal features in HRV data.

6) CNN-GRU hybrid

As a lightweight alternative to CNN-LSTM, the CNN-GRU hybrid model combines the local feature extraction capability of CNN with the temporal modelling efficiency of GRU. Specifically, the front-end CNN captures short-term HRV fluctuations, whereas the back-end GRU, through its dual-gating mechanism, processes feature sequences to learn long-term dependencies. With fewer parameters compared to LSTM, the CNN-GRU hybrid is better suited for small-sample HRV scenarios, effectively balancing local detail preservation and global temporal analysis.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot tanh(W_h \cdot [r_t \odot h_{t-1}, F_t] + b_h) \quad (12)$$

## 4. RESULTS

### 4.1 Analysis of evaluator scoring

To ensure the objectivity and reliability of the assessment outcomes, the potential influence of evaluator subjectivity on trainee scores was first examined. Three evaluators (designated as evaluator 1, evaluator 2 and evaluator 3) independently assessed trainees across the Knowledge (K), Skills (S) and Attitude (A) dimensions, as well as the Total score (T). The Shapiro-Wilk test was applied to examine the normality of score distributions, given the sample size (n<50). Concurrently, internal consistency reliability was evaluated using

Cronbach's alpha (α). Pre-revision results indicated that evaluator 1 achieved an α of 0.683, evaluator 2 an α of 0.889 and evaluator 3 an α of 0.841. These findings suggested that the internal consistency of evaluator 1 was only acceptable and notably lower than that of the other two evaluators.

Based on the results of normality testing, Pearson product–moment correlation coefficients were computed for normally distributed data, while Spearman rank correlations were used for non-normally distributed data. The correlations between evaluator identity and the K, S, A dimensions, as well as the Total score, were 0.382, 0.512, 0.422 and 0.503, respectively, indicating moderate associations between evaluator scoring patterns and trainee performance.

To address evaluator subjectivity and insufficient internal consistency, the assessment criteria were systematically revised. Subjective items (e.g. "Does the pilot clearly understand the walk-around inspection route?") were replaced with observable and quantifiable standards (e.g. "Did the trainee inspect the integrity of fuselage skin rivets?"). The revised rubric was subsequently validated through inter-evaluator reliability analysis, video-based verification and repeated internal consistency testing.

Post-revision analysis demonstrated marked improvements. Cronbach's α increased to 0.930 for evaluator 1, 0.907 for evaluator 2 and 0.884 for evaluator 3. All three evaluators thus achieved excellent internal consistency, confirming stronger dimensional coherence under the revised criteria. Re-analysis of evaluator correlations revealed that the coefficients between evaluator identity and the dimensions decreased to 0.409 (K), 0.246 (S), 0.187 (A) and 0.235 (T). Furthermore, the Knowledge, Skills and Attitude dimensions each demonstrated significant positive correlations with the Total score, with the Skills dimension exerting the greatest influence. Compared with pre-revision results, these findings confirm that the revised criteria substantially improved inter-evaluator reliability, reduced evaluator bias, and enhanced both objectivity and scientific rigour.

As illustrated in *Figure 5*, post-revision scoring distributions exhibited distinct characteristics. Evaluator 3 assigned the highest median and mean scores, with a more concentrated distribution, suggesting superior overall performance among the trainees assessed by this evaluator. Collectively, these results demonstrate that the implementation of scenario-based simulation training [26], combined with cognitive load management strategies, can effectively improve training effectiveness and elevate the quality of performance assessment.

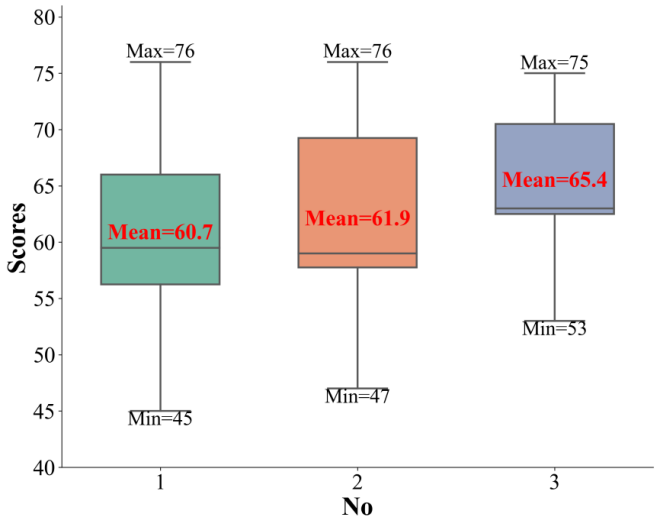


*Figure 5 – Scores by different evaluators*

## 4.2 Analysis of participant workload

Based on flight experience and training background, participants were divided into three groups: Group c comprised participants with no flight experience and no relevant training; Group b included participants who had received training but lacked actual flight experience; and Group a consisted of participants with flight experience. Due to limitations in familiarity with professional task domains, Group c participants exhibited the highest cognitive load levels. In contrast, Group b participants displayed moderate cognitive load levels.

*Figure 6* presents the performance score distributions for each participant group after re-evaluation using revised scoring criteria combined with video recordings. Results indicate a negative correlation between task performance and cognitive load levels across different experience levels: Group a achieved the highest

comprehensive performance scores, Group b showed intermediate performance, and Group c received the lowest scores. This gradient distribution pattern further validates the positive impact of flight experience and structured training on task performance quality.

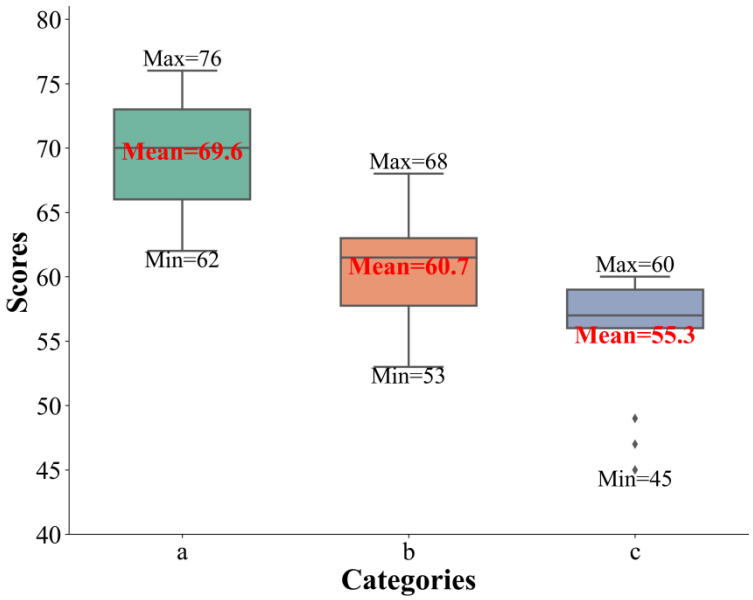


*Figure 6 – Scores by different participants*

### 4.3 Analysis of important HRV features

Based on statistical significance ($p<0.05$) and importance ranking, the top five features were selected: mean_hr, max_hr, std_hr, mean_nni and median_nni. The mean values of these features were calculated for participants in Groups a, b and c, as presented in *Table 3*. Group a participants, who exhibited the lowest workload, demonstrated the lowest mean_hr, max_hr and std_hr, along with the highest mean_nni and median_nni. Conversely, Group c participants, who showed the highest workload, displayed the highest mean_hr, max_hr and std_hr, coupled with the lowest mean_nni and median_nni.

*Table 3 – Average values of significant and important features for different participant groups*

| Features | a | b | c |
|---|---|---|---|
| mean_hr (bpm) | 74.41 | 77.99 | 80.23 |
| max_hr (bpm) | 102.84 | 109.55 | 113.14 |
| std_hr (bpm) | 16.15 | 17.69 | 18.33 |
| mean_nni (ms) | 903.01 | 901.76 | 895.84 |
| median_nni (ms) | 872.88 | 854.42 | 852.04 |

### 4.4 Comparison of classifier results

To comprehensively evaluate the predictive capability of HRV features in workload classification, six CNN-based deep learning models (depthwise CNN, large kernel CNN, CNN-LSTM hybrid, residual CNN, multi-scale CNN and CNN-GRU hybrid) were employed for classifier construction. Two feature selection strategies were designed for all HRV features (complete feature set) and HRV features after principal component analysis (PCA) dimensionality reduction, where PCA retained principal components explaining 95% of the cumulative variance to mitigate feature redundancy and overfitting risks.

Feature standardisation was performed using RobustScaler, which scales data based on the median and interquartile range to enhance robustness against extreme values. Labels were integer-encoded, mapping high, medium and low workload levels to numerical labels, and were further converted to one-hot encoding during deep learning model training to adapt to Softmax outputs. The dataset was split into training and testing sets

at an 8:2 ratio using stratified sampling to preserve class distribution. All random processes were fixed with a seed of 42 to ensure experimental reproducibility.

Six temporal data augmentation methods were designed, including time warping, amplitude scaling, jittering, window slicing, Gaussian filtering and segment permutation. These methods perturb the time axis, amplitude or temporal structure, thereby introducing measurement noise and individual variability to enhance model robustness and generalisation capability. To prevent overfitting, augmentation ratios were adaptively set according to model type: twofold for machine learning models and threefold for deep learning models, with a rotation strategy employed to ensure sample diversity. Dropout and batch normalisation were incorporated into all models to mitigate overfitting, while strategies such as global average pooling were adopted to reduce parameter counts.

To address class imbalance, three oversampling techniques, SMOTE, ADASYN and Borderline-SMOTE, were applied, with the optimal method for each model determined via 5-fold cross-validation. This strategy accounted for model-specific sensitivity to sampling approaches, aiming to maximise the recognition rate of minority classes while controlling dataset expansion. Classifier performance was evaluated using the following key metrics: mean accuracy, precision, recall, Cohen's kappa coefficient and F1-score. As shown in *Table 4*, substantial performance differences were observed among classifiers in HRV-based workload classification under the two feature processing strategies (all features and PCA-reduced features). Results show that depthwise CNN achieved the best overall performance on the full feature set, while CNN-GRU hybrid demonstrated superior accuracy after dimensionality reduction. These findings suggest that model effectiveness is strongly influenced by feature dimensionality, with depthwise and large-kernel CNNs benefiting from richer input representations, whereas GRU-based hybrids show greater robustness under compact feature spaces.

*Table 4 – Prediction results of different classification models*

| Performance | Models | Accuracy | Precision | Recall | Cohen's Kappa coefficient | F1-score |
|---|---|---|---|---|---|---|
| All HRV features | Depthwise CNN | 0.9372±0.008 | 0.9377±0.008 | 0.9372±0.008 | 0.9038±0.012 | 0.9373±0.008 |
| | Large kernel CNN | 0.9346±0.011 | 0.9350±0.011 | 0.9346±0.011 | 0.8994±0.018 | 0.9344±0.011 |
| | CNN-LSTM hybrid | 0.9324±0.006 | 0.9330±0.006 | 0.9324±0.006 | 0.8964±0.009 | 0.9324±0.006 |
| | Multi-scale CNN | 0.9053±0.004 | 0.9060±0.004 | 0.9053±0.004 | 0.8543±0.007 | 0.9049±0.004 |
| | Residual CNN | 0.9282±0.014 | 0.9286±0.014 | 0.9282±0.014 | 0.8896±0.021 | 0.9279±0.014 |
| | CNN-GRU hybrid | 0.8968±0.017 | 0.8967±0.017 | 0.8968±0.017 | 0.8413±0.027 | 0.8964±0.017 |
| PCA-reduced HRV features | Depthwise CNN | 0.9029±0.027 | 0.9047±0.025 | 0.9029±0.027 | 0.8518±0.041 | 0.9032±0.027 |
| | Large kernel CNN | 0.8806±0.017 | 0.8810±0.017 | 0.8806±0.017 | 0.8165±0.027 | 0.8801±0.018 |
| | CNN-LSTM hybrid | 0.9176±0.013 | 0.9182±0.013 | 0.9176±0.013 | 0.8736±0.020 | 0.9175±0.013 |
| | Multi-scale CNN | 0.8907±0.024 | 0.8913±0.024 | 0.8907±0.024 | 0.8321±0.036 | 0.8904±0.024 |
| | Residual CNN | 0.9162±0.018 | 0.9171±0.017 | 0.9162±0.018 | 0.8716±0.026 | 0.9162±0.018 |
| | CNN-GRU hybrid | 0.9186±0.007 | 0.9187±0.007 | 0.9186±0.007 | 0.8751±0.010 | 0.9185±0.007 |

## 5. DISCUSSIONS

### 5.1 Mechanisms of workload differences

Considering that participants were randomly assigned to groups, this suggests that original scores were subject to certain subjective biases due to individual evaluator prejudices, experience levels and differences in understanding assessment criteria. Compared to initial results, this demonstrates that the revised scoring criteria significantly reduced evaluator subjective bias and enhanced assessment objectivity and scientific rigour. Evaluator 2 demonstrated a wider score distribution range, potentially reflecting individual differences in participant skill levels or inconsistencies in assessment standard implementation. Evaluator 1 exhibited the lowest median and mean values, suggesting a relatively weaker overall performance of this participant group. These findings indicate that implementing scenario-based simulation training [26] and cognitive load

management strategies to improve existing training systems could enhance training effectiveness and assessment quality.

When confronted with complex inspection protocols, operational standards and emergency response procedures, they lacked pre-established cognitive schemas and mental representation frameworks, necessitating heavy reliance on working memory for information encoding, analytical processing and decision-making, resulting in significantly increased extraneous cognitive load. This result aligns with expectations from expert-novice cognitive load theory [27], which posits that novices typically demonstrate higher cognitive load due to limited task familiarity and a lack of well-developed cognitive schemas.

Standardised training provided them with essential theoretical knowledge foundations and basic operational skills, significantly reducing task unfamiliarity compared to Group c and improving task execution efficiency [28]. Group a participants showed the lowest cognitive load levels, with their proficient mastery of walk-around inspection protocols and extensive practical experience enabling rapid adaptation to task requirements, accurate decision-making under pressure conditions, and minimisation of operational errors and corrective behaviours. Participants with extensive experience were able to execute tasks by leveraging established procedural knowledge and automated skills, reducing demands on limited cognitive resources. While training cannot completely substitute for practical experience, it effectively reduces novice cognitive load by establishing foundational cognitive structures.

### 5.2 Correlation between HRV features and workload

HRV regulation is a complex process involving both the sympathetic and parasympathetic nervous systems. Sympathetic nervous system activation accelerates HR, while parasympathetic nervous system activation decelerates it. HR and rhythm are primarily regulated by the ANS. Increased workload has been demonstrated to elevate psychological and physiological stress, activating the sympathetic nervous system to release catecholamines (such as epinephrine and norepinephrine), thereby increasing HR and myocardial contractility, resulting in elevated mean_hr and max_hr [29][30]. Simultaneously, increased workload suppresses parasympathetic activity, leading to reduced HRV [31]. For instance, acute psychological stress under time-constrained task conditions has been shown to induce transient catecholamine release, which correlates with elevated std_hr [32].

Previous research [33] has demonstrated that increased workload shortens NN intervals, thereby reducing mean_nni and median_nni. This reduction in NNI is attributed to heightened workload triggering sympathetic nervous system activation, resulting in catecholamine release and increased myocardial contractility, while simultaneously suppressing parasympathetic activity. During walk-around inspections, participants must continuously conduct detailed examinations of specific aircraft components, and this concentrated investment of visual attention leads to more focused gaze patterns and lower NNI [34].

### 5.3 Classifier performance differences

When using the full HRV feature set, the depthwise CNN achieved the highest accuracy ($0.9372 \pm 0.008$) and F1-score, followed closely by the large kernel CNN and CNN-LSTM hybrid. The superior performance of depthwise CNN likely stems from the parameter- and compute-efficiency of depthwise-separable convolutions. By decoupling spatial filtering (depthwise) and channel mixing (pointwise), the architecture greatly reduces parameter count and computational cost while retaining representational flexibility, thereby ameliorating overfitting risks in moderate-sized datasets [35][36]. Indeed, depthwise separable convolutions have been widely adopted in lightweight models for mobile and embedded applications owing to such efficiency gains.

The large kernel CNN also performed strongly on the full feature set. The rationale is that larger convolutional kernels effectively expand the receptive field, allowing the network to capture longer-range temporal dependencies without requiring deep stacking. Recent work [37] shows that replacing many small kernels with a few large ones can approach or even match the representational power of transformer-style architectures, especially when combined with re-parameterisation techniques. In the HRV/physiological signal domain, where discriminative patterns may span multiple time scales, this capacity to “see farther” is advantageous.

The CNN-LSTM hybrid also showed robust performance under full-feature settings, evidencing the benefit of combining CNN’s local feature extraction with LSTM’s long-range memory capabilities. CNN layers extract temporal motifs and local feature maps, while the LSTM integrates sequential dependencies across

time; such CNN-LSTM hybrids have been successfully applied to ECG/physiological signal classification [38]. The gating mechanisms in LSTM (forget, input, output gates) are well-suited to manage noise and temporal context in time series data.

In contrast, the multi-scale CNN and CNN-GRU hybrid lagged somewhat when given the full HRV feature set. Multi-scale convolutional architectures often include multiple branches (e.g. different kernel sizes or dilation rates), which may introduce redundancy or parameter inefficiency if not carefully regularised or fused; when the input features already encode multi-scale statistics, the architectural overhead may not yield net gains. The GRU-based hybrid, while parameter-light, may not fully exploit the richness of the unreduced feature set, especially if finer temporal patterns require more expressive gating than GRU provides.

After PCA dimensionality reduction, the performance ordering shifted: CNN-GRU became the top-performing architecture, marginally outperforming CNN-LSTM and residual CNN. This suggests that in compressed feature spaces with reduced redundancy, simpler recurrent architectures (GRU) may generalise better, likely due to fewer parameters, lower overfitting risk and more stable training dynamics. In effect, the GRU's lighter gating architecture appears more robust when feature inputs are compact and less noisy. Empirical comparisons of recurrent units [39] show that GRU can match or exceed LSTM performance in many sequence modelling tasks while using fewer parameters, making it particularly suitable in constrained / compressed feature regimes.

The residual CNN offered stable performance across both paradigms, illustrating that skip connections help in training deeper architectures without vanishing gradients. However, while residual connections foster deeper representation learning, they do not by themselves guarantee the best performance if the architecture is mismatched to the input feature space [40].

## 6. CONCLUSIONS

This study is among the first to systematically investigate pilot workload during pre-flight walk-around inspections, integrating physiological indicators, expert scoring and task performance metrics. Results showed clear workload gradients by experience: untrained participants exhibited the highest cognitive load, trained but inexperienced students demonstrated intermediate levels, while experienced pilots achieved the lowest workload and superior inspection performance. HRV analysis revealed that heart rate indices (mean HR, maximum HR, HR standard deviation) and interval-based measures (mean NNI, median NNI) were the most sensitive workload indicators.

From a modelling perspective, six CNN-based architectures were benchmarked. Findings indicated that depthwise and large-kernel CNNs excel when using the complete HRV feature set, whereas GRU-based hybrids generalise more robustly under PCA-based dimensionality reduction. Residual connections contributed to stable training but required alignment with the feature space for optimal results.

From a practical standpoint, several design lessons emerge: (1) depthwise separable and large-kernel CNNs are particularly effective in rich, high-dimensional feature spaces; (2) lighter recurrent architectures such as GRU are preferable when dimensionality reduction is applied; and (3) residual connections ensure stability but must be matched with data complexity to maximise benefit.

Limitations include the use of a single PCA retention threshold; future work should explore multiple thresholds (e.g. 90%, 95%, 99%) to examine ranking robustness. Additionally, statistical tests (e.g. Friedman with post-hoc or McNemar's paired comparisons) and confidence intervals should accompany performance comparisons to establish significance. Finally, ablation experiments, such as removing recurrent modules, varying kernel sizes, or pruning multi-scale branches, would clarify how each architectural component contributes to performance.

## ACKNOWLEDGEMENT

We would like to thank the foundations of the Fundamental Research Funds for the Central Universities (Grant No. ZJ2023-009).

## REFERENCES

[1] International Civil Aviation Organization. Annex 6: Operation of Aircraft. Montreal: ICAO;2024.

[2] Civil Aviation Administration of China (CAAC). Regulations on the Operational Certification of Public Air Transport Carriers Using Large Aircraft (CCAR-121-R5). Beijing: CAAC;2020.

[3] Federal Aviation Administration. FAR Part 91: General Operating and Flight Rules. Washington, DC: U.S. Government Publishing Office.2023.

[4] Shappell S A, Wiegmann D A. The human factors analysis and classification system-HFACS. (Report No. DOT/FAA/AM-00/7). Washington, DC: Federal Aviation Administration, Office of Aerospace Medicine;2000.

[5] National Transportation Safety Board. Aircraft accident report: Alaska Airlines Flight 1282. Washington, DC: NTSB;2024.

[6] Skybrary. Maintenance workload [Internet]. Available from: https://skybrary.aero/articles/maintenance-workload

[7] Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research//*Advances in psychology*. North-Holland, 1988;52:139-183. DOI: 10.1016/s0166-4115(08)62386-9.

[8] Vidulich MA, Tsang PS. Techniques of subjective workload assessment: a comparison of SWAT and the NASA-Bipolar methods. *Ergonomics*, 1986;29(11),1385-1398. DOI: 10.1080/00140138608967253.

[9] Masi G, et al. Stress and workload assessment in aviation—a narrative review. *Sensors,* 2023;23(7):3556. DOI: 10.3390/s23073556.

[10] Luzzani G, et al. A review of physiological measures for mental workload assessment in aviation. *Aeronautical Journal*, 2024;128(1323):928-949. DOI: 10.1017/aer.2023.101.

[11] Malik M. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: task force of the European society of cardiology and the North American society for pacing and electrophysiology. *Annals of Noninvasive Electrocardiology*, 1996;1(2),151-181. DOI: 10.1111/j.1542-474x.1996.tb00275.x.

[12] Wang P, Houghton R, Majumdar A. Detecting and predicting pilot mental workload using heart rate variability: a systematic review. *Sensors*, 2024;24(12):3723. DOI: 10.3390/s24123723.

[13] Alaimo A, et al. Aircraft pilots workload analysis: heart rate variability objective measures and NASA-task load index subjective evaluation. *Aerospace*, 2020;7(9):137. DOI: 10.3390/aerospace7090137.

[14] Mohanavelu K, et al. Cognitive workload analysis of fighter aircraft pilots in flight simulator environment. *Defence Science Journal*, 2020;70(2). DOI: 10.14429/dsj.70.14539.

[15] Mansikka H, et al. Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics*, 2016;59(10):1344-1352. DOI: 10.1080/00140139.2015.1136699.

[16] Koskelo J, et al. Cardiac autonomic responses in relation to cognitive workload during simulated military flight. *Applied Ergonomics*, 2024;121:104370. DOI: 10.1016/j.apergo.2024.104370.

[17] Cao X, et al. Heart rate variability and performance of commercial airline pilots during flight simulations. *International Journal of Environmental Research and Public health*, 2019;16(2):237. DOI: 10.3390/ijerph16020237.

[18] Park JH, et al. How is the pilot doing: VTOL pilot workload estimation by multimodal machine learning on psycho-physiological signals.//2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). IEEE, 2024;2311-2318. DOI: 10.1109/ro-man60168.2024.10731202.

[19] Zhu W, et al. Assessment of pilot mental workload based on physiological signals: A real helicopter cross-country flight study//2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT). IEEE, 2023;638-643. DOI: 10.1109/iccasit58768.2023.10351548.

[20] Navalta JW, et al. Heart rate processing algorithms and exercise duration on reliability and validity decisions in biceps-worn Polar Verity Sense and OH1 wearables. *Scientific Reports*, 2023;13(1),11736. DOI: 10.1038/s41598-023-38329-w.

[21] Takahashi M, et al. Cardiac parasympathetic outflow during dynamic exercise in humans estimated from power spectral analysis of P–P interval variability. *Experimental Physiology*,2016;101(3):397-409. DOI: 10.1113/ep085420.

[22] Marathon Handbook. Heart rate variability [Internet]. Available from: https://marathonhandbook.com/heart-rate-variability/

[23] Champseix R, Ribiere L, Le Couedic C. A python package for heart rate variability analysis and signal preprocessing. *Journal of Open Research Software*, 2021;9(1). DOI: 10.5334/jors.305.

[24] Lipponen JA, Tarvainen MP. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. *Journal of Medical Engineering and Technology*, 2019;43(3):173-181. DOI: 10.1080/03091902.2019.1640306.

[25] Yang S, et al. The impacts of temporal variation and individual differences in driver cognitive workload on ECG-based detection. *Human Factors*, 2021;63(5):772-787. DOI: 10.1177/0018720821990484.

[26] Hurd KD, et al. Effectiveness of simulation-based training for obstetric internal medicine: Impact of cognitive load and emotions on knowledge acquisition and retention. *Obstetric Medicine*, 2021;14(4):242-247. DOI: 10.1177/1753495x211011915.

[27] Paris F, et al. Differences between experts and novices in the use of aircraft maintenance documentation: evidence from eye tracking. *Applied Sciences*, 2024;14(3):1251. DOI: 10.3390/app14031251.

[28] FasterCapital. Aviation Training Research: Cognitive Load Management in Flight Training—Strategies and Challenges [Internet]. Available from: https://www.fastercapital.com/content/Aviation-Training-Research--Cognitive-Load-Management-in-Flight-Training--Strategies-and-Challenges.html

[29] Altmann A, et al. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 2010;26(10):1340-1347. DOI: 10.1093/bioinformatics/btq134.

[30] Mahdavi N, et al. Unraveling the interplay between mental workload, occupational fatigue, physiological responses and cognitive performance in office workers. *Scientific Reports*, 2024;14(1):17866. DOI: 10.1038/s41598-024-68889-4.

[31] Mund D, Schulte A. Experimental evaluation of heart-based workload measures as related to their suitability for real-time applications//International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2020;372-382. DOI: 10.1007/978-3-030-50788-6_27.

[32] Sammito S, et al. Guideline for the application of heart rate and heart rate variability in occupational medicine and occupational health science. *Journal of Occupational Medicine & Toxicology*, 2024;19(1):15. DOI: 10.1186/s12995-024-00414-9.

[33] Kim HG, et al. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investigation*, 2018;15(3):235-245.

[34] Maggi P, Di Nocera F. Sensitivity of the spatial distribution of fixations to variations in the type of task demand and its relation to visual entropy. *Frontiers in Human Neuroscience*, 2021;15:642535. DOI: 10.3389/fnhum.2021.642535.

[35] Howard AG, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv*:1704.04861:2017. DOI: 10.1201/9781351003827-3.

[36] Haase D, Amthor M. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020;14600-14609. DOI: 10.1109/cvpr42600.2020.01461.

[37] Ding X, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022;11963-11975. DOI: 10.1109/cvpr52688.2022.01166.

[38] Zihlmann M, Perekrestenko D, Tschannen M. Convolutional recurrent neural networks for electrocardiogram classification//2017 Computing in Cardiology (CinC). IEEE, 2017;1-4. DOI: 10.22489/cinc.2017.070-060.

[39] Chung J, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv*:1412.3555,2014.

[40] Xu G, et al. Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey. *arXiv preprint arXiv*:2405.01725,2024.