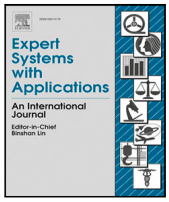
Expert Systems With Applications 242 (2024) 122808

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Analysis of behavioural curves to classify iris images under the influence of alcohol, drugs, and sleepiness conditions

Leonardo Causa [a], Juan E. Tapia [b,*], Andres Valenzuela [c], Daniel Benalcazar [a], Enrique Lopez Droguett [d], Christoph Busch [b]

[a] Electric Engineering Department, University of Chile, Santiago, Chile
[b] da/sec-Biometrics and Internet Security Research Group, Hochschule, Darmstadt, Germany
[c] Informatics Department, Universidad de Santiago de Chile, Chile
[d] Department of Civil and Environmental Engineering, and Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA

ARTICLE INFO

Keywords:
Biometrics
Iris
Fitness for duty

ABSTRACT

This paper proposes a new method to estimate behavioural curves from Near-Infra-Red (NIR) iris images for classifying Fitness for Duty using a biometric capture device. Fitness for Duty (FFD) techniques detect whether a subject is Fit to safely perform a given task, which means no reduced alertness condition and security, or the subject is unfit, that could impact a reduced alertness condition by sleepiness or consumption of alcohol and drugs. The analysis showed essential differences in pupil and iris behaviour to classify the workers in "Fit" or "Unfit" conditions. The best results can distinguish subjects robustly under alcohol, drug consumption, and sleep conditions. The Multi-Layer-Perceptron and Gradient Boosted Machine reached the best results in all groups with an overall accuracy for Fit and Unfit classes of 74.0% and 75.5%, respectively. These results open a new application for iris capture devices.

1. Introduction

In a 24/7 society, many people work day and night for public safety, health services, or economic reasons. An estimated 15%–25% of the workforce works in shifts (Balasubramanian et al., 2020; Gusman, Standlee, Reid, & Wolfe, 2023; Peter, Reindl, Zauter, Hillemacher, & Richter, 2019; Wickwire, Geiger-Brown, Scharf, & Drake, 2017). Working in rotating shifts at night can be a significant risk factor. Also, several studies performed in different regions, such as the U.S., Australia, U.K., Japan and others, show an essential correlation between shift work at night with alcohol and drug consumption, which can be considered as critical triggers for occupational accidents (Borrelli et al., 2023; Wickwire et al., 2017). Workplace alcohol and drug use directly affect an estimated 15% of the U.S. workforce; about 10.9% work under the influence of alcohol or with a hangover (Frone, 2006). The Australian Government alcohol guidelines report shows 13% of shift-workers and 10% of those on standard schedules reported consuming alcohol at risky levels for short-term harm (Dorrian & Skinner, 2012). It is essential to note that these numbers are similar in different parts of the world.

These fitness impairments for work cause decreased performance and an increased likelihood of accidents and are also associated with productivity loss and increased economic costs (Iqbal, Shafiq, Singh, & Afzal, 2023). Alcohol and substance abuse disorders and their consequences on the heart, liver, immune system, and other organs are on the rise worldwide (Martini, Fregna, Bosia, Perrozzi, & Cavallaro, 2022; Pinheiro et al., 2015), which are made worse by the economic and health crisis generated by COVID-19. According to the National Institute of Drug Abuse (NIDA), alcohol, illicit drugs, and tobacco cost between $600 billion and $740 billion per year via lost work productivity, crime, and healthcare (Birnbaum et al., 2011; Florence, Zhou, Luo, & Xu, 2016; NIDA, 2020; Sacks, Gonzales, Bouchery, Tomedi, & Brewer, 2015; Xu, Bishop, Kennedy, Simpson, & Pechacek, 2015). Fatigue is estimated to cost employers another 136 billion dollars (Council, 2017; Rosekind et al., 2010; Xu & Hall, 2021; Yung, 2016). Only in the U.S., the total cost of workplace injuries in 2019 was 171 billion dollars. Sleep deprivation impairs alertness and short-term memory. It produces global decreases in brain activity, especially in networks that control

* Corresponding author.
E-mail addresses: lcausa@ing.uchile.cl (L. Causa), juan.tapia-farias@h-da.de (J.E. Tapia), andres.valenzuela.g@usach.cl (A. Valenzuela), dbenalcazar@ug.uchile.cl (D. Benalcazar), eald@g.ucla.edu (E.L. Droguett), christoph.busch@h-da.de (C. Busch).

https://doi.org/10.1016/j.eswa.2023.122808
Received 17 August 2023; Received in revised form 29 November 2023; Accepted 30 November 2023
Available online 5 December 2023
0957-4174/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

attention and higher-order cognitive processes, which translates into a higher rate of accidents. This compromised performance has been linked to transportation accidents and significant industrial incidents (Rodger, 2020).

Fatigue, drowsiness, and sleepiness caused about 22% of all injury crashes according to Road Safety Annual Report 2020[1] published by the Organisation for Economic Cooperation and Development (OECD). The odds of fatal crashes caused by fatigue can reach up to 30%. The World Health Organisation (WHO) reported that fatigue accidents cost most countries 3% of their gross domestic product (Němcová et al., 2021). Hence, it is necessary to take proactive measures to prevent fatigue, alcohol consumption, and drug abuse from negatively affecting employees and companies.

Currently, many solutions exist for addressing workplace fatigue, alcohol, and substance abuse (Azadani & Boukerche, 2021); these solutions are known as Fitness For Duty (FFD) analysis systems (MacQuarrie et al., 2018; Murphy & Fleming, 1992). Within the context of occupational testing, FFD describes a toolkit that helps evaluate a person's physical and emotional states based on specific requirements of a job. Being "Fit" means being able to perform a job's duties in a safe, secure, productive, and effective manner. However, these solutions do not adequately address all the required elements of workplace accident prevention.

In order to address these problems, it is necessary to observe what repetitive behaviours or biometric factors manifest through the body to establish the relationship between the cause and effect in a person's behaviour or study how eye movements can be interpreted to provide insights into cognitive processes and present a descriptive model representing the relations of eye movements and cognitive load (Han, Yin, Zhang, Jin, & Yang, 2020; Katona, 2022; Vasiljevas, Damaševičius, & Maskeliūnas, 2023).

In our case, we study the relationship of the Central Nervous System (CNS), which controls the iris and pupil movements (Adler, 1985). In this condition, the subject cannot voluntarily change the pupil's or iris' movement. This action is initiated automatically in response to an external factor, such as light or alcohol consumption, drug abuse, or fatigue. On the other hand, iris recognition allows identifying and distinguishing one worker from another, thus guaranteeing that the measurement is from the worker under test. Therefore, the iris and pupil are highly reliable in measuring the fitness for duty of the subject. Hence, there is a need to develop an automated and reliable FFD model tool based on the iris recognition framework (Campbell et al., 2023; Tapia, Droguett, & Busch, 2022; Tapia, Perez, & Bowyer, 2016).

1.1. Related work

This section describes in detail the main works available in the state of the art with regards to detecting alcohol, drugs, sleepiness. Also leading market solutions will be reviewed.

1.1.1. Alcohol and drug detection

Regarding alcohol detection, Navarro, Diño, Joson, Anacan, and Cruz (2016) developed a system that captures the driver's iris image to detect if the person is drunk. That approach comprises a hardware and software system that implements an algorithm based on the Gabor Filter. The system consists of a Charge-Coupled Device (CCD) Camera and Analog-to-Digital converter linked into a program to process the captured image. The model detects changes in the size of the pupil relative to a previous measurement without alcohol consumption (baseline). The system provides a signal to interact with the car ignition if the software detects that the driver is under the influence of alcohol.

Pinheiro et al. (2015) proposed a non-invasive and simple analysis to detect alcohol use through Pupillary Light Reflex (PLR) analysis. The database consists of 40 pupillometric recordings of 50 s each. The method consists of segmenting the images and getting the pupil diameter values to create a characteristic vector based on six metrics: maximum mydriasis, maximum miosis, amplitude, latency, time to maximum contraction, and time to maximum dilation. Then, Support Vector Machine (SVM) and kNN were used to perform the final classification. Results presented rates near to 85% for correct detection, thus demonstrating the method's efficacy. The main limitation of this work is related to the active participation of the volunteers, as each subject had to stay in a dark testing room for approximately 5 min to adapt the pupil dilation/constriction to the darkness.

Amodio, Ermidoro, Maggi, Formentin, and Savaresi (2019) studied the feasibility of designing a driver alcohol detection system based on the dynamic analysis of a subject's PLR. The test method consists of applying a light stimulus to one eye and capturing both eyes' constriction dynamics. A two-step methodology is described for extracting pupil size profiles from the video sequences. The iris/pupil search within the visual image is performed in the first phase. In the second stage, the image is cropped, and pupil detection is applied using the Circular Hough Transform (CHT).

A set of features is introduced to compare the two populations of responses and is used to design a Support Vector Machine classifier to discriminate between "Sober" and "Drunk" states.

Jayadev and Bellary (2021) proposed an effective method for the classification of alcohol effects as well as the identification of iris-damaged levels utilising a Modified Deep Learning Neural Network (MDLNN). Initially, the alcohol image extracts many features, such as frequency, shape, and texture. Next, the extracted features are selected utilising Bacterial Foraging Optimisation (BFO). Finally, the segmented iris percentage is estimated based on the Euclidean distance between the original iris image of a person under alcohol and the segmented level of the iris image.

Murillo, Crucilla, Schmittner, Hotchkiss, and Pickworth (2004) used video-based pupillometry to evaluate opioid-maintained patients. Pupillometry measures (initial diameter, constriction amplitude, constriction latency, and saccadic velocity) were used to detect the presence of drugs. Later, the results were compared with each volunteer's urine samples to estimate the system's detection performance. Analysis results indicated that 92.9% of subjects obtained an acceptable detection. Results suggest that pupillometry may be useful to verify concomitant drug use in a methadone-maintained population.

Natarajan et al. (2013) proposed a wearable electrocardiogram (ECG) sensor to detect cocaine consumption. Cocaine use has an important effect on heart rate variability. The method measured the RR interval to segment ECG for feature extraction and the standard linear logistic regression to classify. The results show an 80% accuracy for cocaine detection.

Mahmud, Fang, Carreiro, Wang, and Boyer (2019) presented a review study on the application of wearable technologies for the detection of drugs and alcohol. The study shows different methods, research prototypes, and commercially available devices. These works have demonstrated the feasibility of wearable sensors to detect drug abuse in a noninvasive manner.

1.1.2. Sleepiness detection

Regarding sleepiness, Bakker et al. (2021) presented a video-based driver sleepiness detection system set up as a two-stage model with (1) a generic deep feature extraction module combined with (2) a personalised sleepiness detection module. The method was tested using data from 13 drivers. The results showed an accuracy of 92% for alert classification; without personalisation, the corresponding accuracy was 72%, while a standard fatigue detection PERCLOS-based baseline method reached an accuracy of 68% on the same dataset.

In this context, Benderoth, Hormann, Schiebl, and Elmenhorst (2021) discussed the reliability and validity of a 3-min Psychomotor

[1] https://www.itf-oecd.org/road-safety-annual-report-2020.

Vigilance Test(PVT) administered on a portable handheld device, assessing sensitivity to sleep loss and alcohol-related 10-min PVT. The 3-min PVT showed high reliability and validity in determining sleep loss and alcohol-induced impairments in cognitive performance.

Persson, Jonasson, Fredriksson, Wiklund, and Ahlström (2021) studied the reliability of Heart Rate Variability (HRV) as a feature for driver sleepiness detection. Data from real-road driving studies, including 86 drivers in both alert and sleep-deprived conditions, were used. Based on the Karolinska Sleepiness Scale (KSS), subjective ratings were used as ground truth. K-nearest neighbours, SVM, AdaBoost, and Random Forest were applied for training and testing the models. The best performance was obtained with the Random Forest classifier with an accuracy of 85%. The worst results were obtained in the total sleep deprivation group. The results showed that subject-independent sleepiness classification based on HRV performs poorly in realistic driving conditions.

Takano et al. (2023) showed a study in which participants with health risks from alcohol or methamphetamine wore a Fitbit and used a self-monitoring app for 8 weeks. They logged their daily substance use in a mobile app, while Fitbit tracked metrics like heart rate, sleep patterns, steps, and physical activity. The Fitbit data was initially visualised to confirm individual patterns. Subsequently, a combination of machine learning and statistical analyses was employed to develop a substance-use detection model using both Fitbit and app data. This model underwent testing via 5-fold cross-validation, with further machine learning adjustments based on preliminary results. The study also assessed the feasibility and usability of this approach.

1.1.3. Fitness for Duty (FFD)

The PVT is a brief vigilance and attention task, and it is considered the gold standard instrument for the assessment of the effects of fatigue (Balkin et al., 2004). During each 10-min trial, subjects must attend closely to a stimulus window and respond by pressing a button. Subjects are instructed to react as quickly as possible. PVT scores of interest include mean reciprocal reaction time of the slowest 10% of responses and lapses corresponding to stimulus presentations taking longer than 500 ms.

Kim, Cho, Suh, and Yim (2021) studied operator performance in a nuclear power plant using an FFD system using an Electroencephalogram (EEG) with a deep learning algorithm to classify an operator's condition. To determine the suitability of this approach, EEG data were collected during simple cognitive exercises designed to examine the mental readiness of nuclear operators. The designed EEG-based FFD classification system could successfully determine an operator's sobriety, stress, and fatigue in a timely and cost-effective manner. This study also investigated schemes for providing information security to the EEG-based FFD status classification system.

Gonzalez et al. (2022) presented a real-time identification of undesirable health conditions in automobile drivers based solely on their driving behaviour, aiming to reduce accident rates. The methodology centers on building models of both "normal" and "abnormal" driving behaviours using a machine Learning From Observation system (LFO) named Falconet. This system was focused on learning from coaching and observation. Driving actions from 12 human subjects in a simulator were collected, timestamped, and used to create driving behaviour models. The models were then compared to the original driving traces to determine if discrepancies could indicate health conditions, specifically targeting Attention Deficit/Hyperactivity Disorder (ADHD). Results showed that the Falconet-created agents could correctly characterise driving behaviours in nearly 82% of test cases, suggesting the approach's potential, though further research is needed before commercial application.

Tanveer, Khan, Qureshi, Naseer, and Hong (2019) proposed a deep-learning-based driver-drowsiness detection for brain-computer interface (BCI) using functional Near-InfraRed Spectroscopy (fNIRS). The brain signals were acquired from healthy subjects while driving a car simulator. A CNN was used to classify different alertness conditions. This algorithm was used on colour map images to determine the best suitable channels for brain activity detection in other time windows. The CNN architecture yielded an average accuracy of 99.3%, showing that the model could differentiate the images of drowsy/non-drowsy states.

Guede-Fernández, Fernández-Chimeno, Ramos-Castro, and García-González (2019) proposed a drowsiness detection method based on changes in the respiratory signal obtained using an inductive plethysmography belt. The algorithm is based on respiratory rate variability (RRV) analysis. Recordings were acquired with a driving simulator cabin, and a group of experts rated the drivers' condition of alertness to evaluate the algorithm's performance. Results showed a specificity of 96.6%, and a sensitivity of 90.3%.

Suhardi, Rosyidasari, Astuti, and Adiasa (2022) presented a study that focused on developing FFD models for bus drivers by considering various physical, mental, and work-related factors. The methodology employed tools such as the Psychomotor Vigilance Task (PVT), Visual Analogue Scale (VAS), and Karolinska Sleepiness Scale (KSS) in conjunction with logistic regression. Data was collected from bus drivers, considering variables like age, sleep quality, caffeine consumption, and more, with FFD as the dependent outcome.

Sharma et al. (2022) presented a novel approach to stress detection which is intrinsically linked to the concept of FFD. Chronic stress, with its profound physical and mental impacts can significantly impair an individual's ability to perform a job safely and effectively. This study presented a novel approach for timely stress detection using short-duration electroencephalogram (EEG) signals. The methodology involved extracting entropy-based features from EEG signals decomposed using stationary wavelet transform. These features were then classified using various supervised machine-learning algorithms. Notably, the study employed evolutionary-inspired approaches to optimise the parameters of SVM and perform feature weighting. The optimised SVM, using the whale optimisation algorithm, achieved an impressive accuracy of 97.3%, highlighting this method's potential for accurate and timely stress detection.

Zurita, Benalcazar, and Tapia (2023) presented a work that focused on utilising human iris behaviour to predict FFD, determining if a subject is fit or unfit for work based on alertness conditions influenced by factors like fatigue, alcohol, and drugs. The methodology involved classifying FFD using sequences of 8 iris images and extracting both spatial and temporal information through CNN and Long Short Term Memory Networks (LSTM). The system achieved a precision of 81.4% for predicting fit subjects and 96.9% for unfit subjects, demonstrating the potential of iris-based biometric applications in determining conditions like alcohol consumption, drug use, and sleepiness.

1.2. Commercial devices

Various FFD algorithms have been proposed in the literature. Several of these algorithms and concepts were implemented into products and devices to measure FFD. In this section, we analyse some of these products, highlighting the pros and cons of each one. Some relevant devices are PMI FIT2000, Sobereye, Optalert (Chandler, Arnold, Jeffrey B snd Phillips, & Horning, 2010). The Alertplus, Alermeter (Chandler et al., 2010; Ferguson et al., 2020). Jawbone, Fitbit (Bai, Guan, & Ng, 2020; Kandera, Škultéty, & Mesárošová, 2019). The Smartcap (Butler & Fee, 2015; Caldwell et al., 2009), among others.

The PMI-FIT2000[2] uses eye-tracking and pupilometry to identify impaired physiological states due to fatigue and other factors, such as alcohol or drug use. The test requires one minute to complete. The system employs an algorithm that compares an individual's established baseline to the present state on four variables (i.e., pupil diameter, pupil

[2] http://www.pmifit.com/.

constriction amplitude, pupil constriction latency & saccadic velocity). The baseline is established by the average of 10 trials taken during non-impaired conditions. After the baseline trials, each subsequent trial provides the user with scores on the four test components plus a composite score, the FIT Index. The PMI-FIT2000 has been used in multiple fatigues and impairment studies in other contexts, such as motor vehicle operations. Note that this system does not perform biometric recognition.

Sobereye[3] is a portable device used to predict impairment caused by substance abuse or fatigue. It uses a smartphone attached to an opaque enclosure that fits a user's eyes to measure the PLR. The PLR is an involuntary reflex that changes pupil size when the eyes are exposed to light. A high-intensity light will cause the pupil to constrict, and low-intensity light will cause it to dilate. The user holds the enclosure over the eyes for one minute before any measurements are taken. This time allows the pupils to dilate. After the camera flash turns on for four seconds, a video is taken at 60 frames per second in full high-definition resolution (1920 × 1080 pix). Like other FFD examinations, identifying a PLR alteration requires establishing a PLR baseline for each employee. The process of establishing a PLR baseline takes 10 days (start of the workday) to monitor the typical day-to-day PLR variations. After about ten tests, an employee's PLR baseline is established. The baseline is then stored and used as a reference for future tests.

By comparing day-to-day measurements and calculations to the baseline, PLR alterations can be identified. A standard scoring system is used to evaluate the degree of PLR alteration. Employees can be classified as "High Risk" or "Low Risk". This measure indicates the probability of an employee being affected by impairment due to an altered PLR. Additionally, Sobereye uses iris recognition (Russo et al., 1999).

Optalert[4] is an Infrared Reflectance (IR) oculography based on the principle that while people are drowsy, the muscle groups controlling eye and eyelid movements are inhibited by the central nervous system (Johns & Hocking, 2021). IR transducers fitted inside spectacle frames are positioned towards the eye to measure the relative velocity of the opening and closing of the eyelid and blink duration times. A combination of occulometric variables is used to calculate a driver's level of drowsiness in real-time, providing a minute-to-minute Johns Drowsiness Scale (JDS) rating (Johns, Tucker, Chapman, Crowley, & Michael, 2007). The commercially available system is designed to emit auditory warnings when drivers reach a JDS score of 4.5–4.9 (cautionary level of drowsiness) and a score of 5.0 or above (critical level of sleepiness), associated with an increased risk of severe lane excursions on a driving simulator. The system needs a 5-min baseline recording before starting each register (Ftouni et al., 2013).

AlertMeter[5] is a graphical cognitive alertness test lasting 60–90 s. The test interface displays different shapes that the user can identify accurately and quickly. The system does not simulate any particular job function but challenges several key brain functions necessary for all jobs, measuring reaction time, decision-making speed, orientation, and hand-eye coordination. Users take the alertness test ten times to establish an initial baseline score or individual performance standard. The scoring algorithm compares users' daily test results with their baseline scores. The system identifies compromised alertness when an employee's test result significantly deviates from their baseline.

A calculated baseline methodology provides individual feedback rather than a score against an imposed standard. AlertMeter test scores have also been shown to correlate to the daytime, indicating sensitivity to circadian cycles (Ferguson et al., 2020).

Wearable technologies also allow the measurement of clinically relevant parameters describing an individual's health state. Their varied applications have provided the driving force for the development of a broad range of wearable technologies that can be adapted for use in healthcare, the workplace and other fields (Jeong, Bychkov, & Searson, 2019). Wrist wearables such as Jawbone, Apple-Watch, Fitbit and others use sensors that measure heart rate and motion (Kaewkannate, Kaeokwunoiand, & Kim, 2015; Kaewkannate & Kim, 2016; Riad, Shahriar, Zhang, & Barsha, 2021). Both use Optical Heart Rate monitoring (OHR) or photoplethysmography. The OHR monitor detects pulse by shining a light through the skin to look at blood flow. An accelerometer measures motion by translating movement data into digital measurements. The data from wrist wearables can be used to monitor fatigue in the workplace. Long-standing sleep deprivation is correlated with an increased heart rate, and poor sleep quality leads to a high risk of fatigue.

SmartCap's LifeBand[6] is a wearable headband that measures brain EEG. The sensors in the life band send out filtering (low pass) signals to block signals above 40 Hz. Samples at 1280 Hz are taken and converted to 256 Hz to minimise high-frequency noise. Then, a frequency spectrum of the 256 Hz signals is calculated over a five-second time frame. As a result, the delta, theta, and alpha waves can be recorded and then scaled by the power of the beta waves. This wave creates a ratio of an individual's drowsiness and wakefulness. A fatigue score is computed from this ratio, and the risk of fatigue is reported to the worker. This score is calculated based on independently validated algorithms researched by SmartCap (Gruenhagen, Parker, & Cox, 2021).

The analysis of the state of the art in commercial devices of FFD shows several systems based on different technologies. While it is true that they are widely used in the industry, their implementation and usage have significant limitations:

1. Some of these devices can only identify one cause of impairment per time.
2. The systems use reactive methods, i.e., act given that the risk event has been triggered.
3. Do not report the metrics for the implemented algorithms or their associated scientific validations.
4. Require the active involvement of workers for the test course. For example, eyes should be closed voluntarily for one minute, which can generate problems in the test results if the involvement is not correct.
5. Invasive protocol for capture requires contact between the capture subject and the device.
6. It is necessary to establish a baseline as a reference per subject. Then, the reference is used to compare the scores and evaluate them. This situation is a problem because the reference could be intentionally altered.
7. Most systems cannot identify and verify the capture subject performing the test, so it is possible to impersonate the results using a third person.

Our work proposes a new method based on behavioural curves from NIR iris images to mitigate the limitations described above. This method allows integrating an FFD system in a unique, light, portable, and contactless mobile device to estimate subjects' proactive alertness before starting their duties and verifying their identity.

It is essential to highlight that our approach does not perform a traditional analysis based on the measurement of alcohol or drug levels in blood using alcohol-test, drug-test or other devices. This research aims to determine the effect of external factors such as alcohol, drugs, and fatigue on the CNS and how this effect is reflected in behaviour changes on pupils and irises diameters and movements.

This paper proposes the following contributions:

[3] https://www.sober-eye.com/.
[4] https://www.optalert.com/.
[5] https://www.deltasleep.ie/alertness-testing/.
[6] http://www.smartcaptech.com/life-smart-cap/.

- Contactless Evaluation using NIR Iris Images: the proposed method is pioneering in using NIR iris images for determining Fitness for Duty (FFD). This algorithm ensures non-intrusive evaluation and is adept at quickly assessing an individual's state in seconds.
- Multi-factorial Analysis: unlike traditional devices that typically focus on a single cause of impairment, our approach can simultaneously identify three distinct causes affecting FFD: alcohol consumption, drug abuse, and sleep deprivation. This holistic approach offers a more comprehensive assessment of an individual's alertness.
- In-depth State-of-the-art review: this work presents an exhaustive state-of-the-art review on Fitness for Duty, providing readers with a clear context and understanding of the advancements in the field.
- Behavioural Analysis for Trend Estimation: we have incorporated an approach to determine the behavioural patterns of subjects. This pattern allows us to estimate trend curves associated with temporal changes in pupil and iris diameters attributed to the external agents affecting the Central Nervous System (CNS).
- Comprehensive FFD Model: we have presented a new FFD model by integrating behavioural analysis with our detection algorithm. This model estimates a person's fitness to undertake tasks without requiring active participation or a baseline estimation. Additionally, our system's portability and anti-impersonation features further underscore its novelty.
- Unique Annotated Database: we have curated a substantial annotated database of NIR periocular images, which encompasses control (normal subjects), alcohol consumption, drug abuse, and fatigue conditions.

The remaining of this article is organised as follows: Section 2 explains the database. The proposed method for FFD detection is presented in Section 3. The experimental results are discussed in Section 4. Conclusions and remarks are given in Section 6.

2. Database

One of this research's main challenges was composing a new database. This task was very demanding as the recruitment process was exhaustive and challenging. For this research, we composed a new database called the "FFD NIR iris images sequences database" (FFD-NIR-Seq), containing 10-s stream sequences of NIR images. The protocol was analysed and approved by the ethical committee of the University of Chile.

The database contains binocular NIR image sequences corresponding to the periocular area (eye mask), and to complement and enrich the database recordings, we also acquired single-eye NIR images using a monocular capture device to get the area corresponding to each eye separately. Pupil, iris, and sclera were obtained for both image types (see an example in Fig. 1).

For the acquisition of the image sequences, the subjects were positioned in front of the capture device, and the equipment detects the eyes and starts the recording (see Fig. 2).

This data makes it possible to perform the necessary processing to determine the iris and pupil behavioural parameters to be used in developing the models.

The image sequences were captured by using four different devices[7]: (i) Iritech MK2120UL (monocular), (ii) iCAM TD-100A, (iii) Iritech Gemini, and (iv) Iritech Gemini-Venus. Fig. 3 shows the NIR capture devices used for database acquisition.

Four NIR image sequences in different conditions were registered:

- Control DB: healthy subjects that are not under alcohol and/or drug influence and in normal sleeping conditions.
- Alcohol DB: subjects who have consumed alcohol or are in an inebriation state.
- Drugs DB: subjects who have consumed some drugs (mainly marijuana) or who consume psychotropic drugs (by medical prescription).
- Sleep DB: subjects with sleep deprivation, resulting in fatigue and/or drowsiness due to sleep disorders related to occupational factors (shift structures with high turnover).

2.1. Alcohol consumption

In the case of the alcohol database, the subjects were submitted to the following protocol:

1. The first NIR image sequence acquisition was made at time 0 (previous to alcohol consumption).
2. All the Volunteers drank 200 ml of alcohol for up to 15 min.
3. The second acquisition was performed immediately after the alcohol intake was finished, i.e., 15 min after 0.
4. The third acquisition was made 30 min after time 0.
5. Fourth acquisition was made 45 min after time 0.
6. Finally, the fifth acquisition was made 60 min after time 0.

Thus, there were recorded five sequences of images of the subject under the effects of alcohol and one sequence of control images.

2.2. Drugs consumption

According to the World Drug Report, 2021, of the United Nations Office on Drugs and Crime, cannabis[8] is the most widely consumed crop worldwide with an annual prevalence of 15%, followed by pharmaceutical opioids and tranquillisers with a 5% and 2.5% of yearly prevalence, respectively. For this reason, about 95% of our database records correspond to cannabis consumption. In contrast, the remaining 5% corresponds to tranquillisers and more complex drugs (heroin and ecstasy). The volunteers were drug consumers for the drug database acquisitions, and the image recordings took place at least 30 min after the initial consumption.

2.3. Sleep conditions

A particular image acquisition protocol for the sleep database was defined, in which tests were performed under controlled sleep deprivation conditions. These recordings were obtained on a specific group of subjects subjected to different sleep deprivation levels to evaluate the level of fatigue/drowsiness at different time intervals. The volunteers were monitored by using a smart band to measure the quantity and quality of sleep. Subjects were grouped as follows:

1. Total sleep deprivation
2. Less than 3 h of night sleep
3. Between 3 and 6 h of night sleep
4. More than 6 h of sleep (normal sleep)

During the recording season, volunteers performed three daily image acquisitions: (i) at the beginning of the working day, (ii) post-lunch, and (iii) at the end of the working day.

The FFD-NIR-Seq database comprises 1510 eye-disjoint images. On average, 150 images are captured per subject. This process took ten seconds. The image sequences were divided into training, validation, and testing. In the case of the test set, the aim was to represent the actual proportion of unfit instances, close to 15%. Table 1 shows the

[7] https://www.iritech.com.

[8] https://www.unodc.org/unodc/en/data-and-analysis/wdr2021.html.

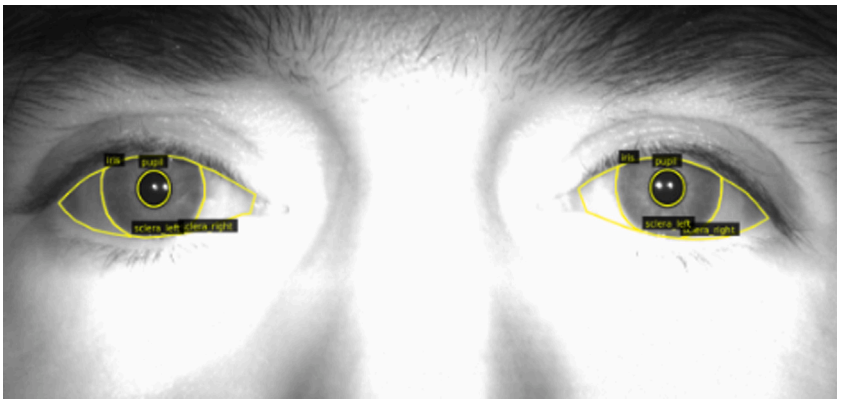


Fig. 1. Example of a labelled periocular NIR image. The image shows both eyes and the corresponding labels to the right and left sclera, pupil, and iris.

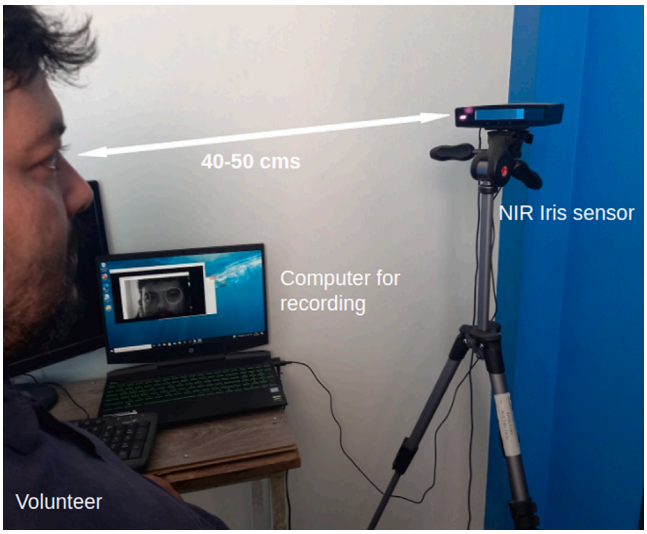


Fig. 2. Description of the capture device. The subject is positioned in front of the sensor (40 up to 50 cm). The capture device detects the eyes and starts the recording for 10-s.

Table 1
NIR image sequences by classes.

Class	Train set	Val. set	Test set
Control	247	35	688
Alcohol	247	35	72
Drug	62	9	17
Sleep	69	9	20
Total	**625**	**88**	**797**

number of sequences. In total, more than 150K images in four classes were collected. Fig. 4 shows examples of the acquired images for each of the classes defined in the database.

This database was used to train and validate the different stages of the FFD model, including the object detection, segmentation, and final classification stages.

3. Methodology

The proposed FFD model can be described as an analysis cascade of four modules, as shown in Fig. 5. The eye detector module allows us to find the eyes and crop them to the right and left instances (in periocular images). The iris and pupil segmentation module is applied to the single eye image sequences to generate the iris and pupil variables: radii and centres are estimated. All eye detection and segmentation processes are explained in detail in Tapia et al. (2021) and the hardware implementation in Benalcazar, Tapia, Vasquez, Causa, Droguett, and Busch (2023). A feature extraction module uses the iris and pupil measures to create a feature vector that represents the temporal behaviour of pupil and iris radii over the acquired image sequences. Finally, the FFD model generates the final classification: Fit/Unfit indicator and level for each state (control, alcohol, drug and sleepiness).

3.1. Eye detector

The eye detector module was implemented to find both eyes in the input periocular images and for subsequent cropping and segmentation. This algorithm is detailed in our previous work on semantic segmentation of periocular NIR images under alcohol effects (Tapia et al., 2021). This module applied Eye-tiny-yolo, classical tracking, and semantic segmentation by Cluster-Coordinated Net (CCNet) (Huang et al., 2020).

First, Eye-tiny-yolo detects the left and right eyes represented as rectangular areas. Then, Multiple Instance Learning (MIL) (Maron & Lozano-Pérez, 1998) and Channel and Spatial Reliability Tracker (CSRT) (Farkhodov, Lee, & Kwon, 2020) tracking methods were applied to detect interest points and descriptors to track each eye frame by frame. Finally, we use a modified version of CCNet to define the final rectangular area that contains each eye and crop the images again (see Fig. 6).

3.2. Iris and pupil segmentation

After both eyes are cropped (left and right), a semantic segmentation method was applied, which was trained from scratch called CCNet to find a mask that segments the iris and the pupil. The CNN was trained using a subset of monocular NIR images with an aggressive data augmentation process with non-geometrical transformations. These networks output a mask highlighting pixels belonging to the iris and the pupil in the images. In order to use this information, we employ pupil and iris localisation algorithms, which find the centres and the radii of the circles that best adapt the pupil and iris contours by using the mask as input. This method employs the morphological erosion operation and a XOR function to find the shape of the valid iris area. Then, the mean square error was used to determine the circle that best fits the right pupil and iris. The outputs are the radii and centres of the pupil and iris in pixels for each image in the sequences (see Fig. 7).

3.3. Feature extraction from iris and pupil

Once the radii and centre measures for the pupil and iris were obtained for all images in each sequence, the features based on the pupil and iris radii variation each time were estimated. As a result, the information from the 100 images is transformed into several associated

Fig. 3. Capture devices used for the acquisition of NIR image sequences. (a) Monocular capture device to record each eye area. Capturing devices (b)–(d) allow for capturing the periocular area (both eyes).

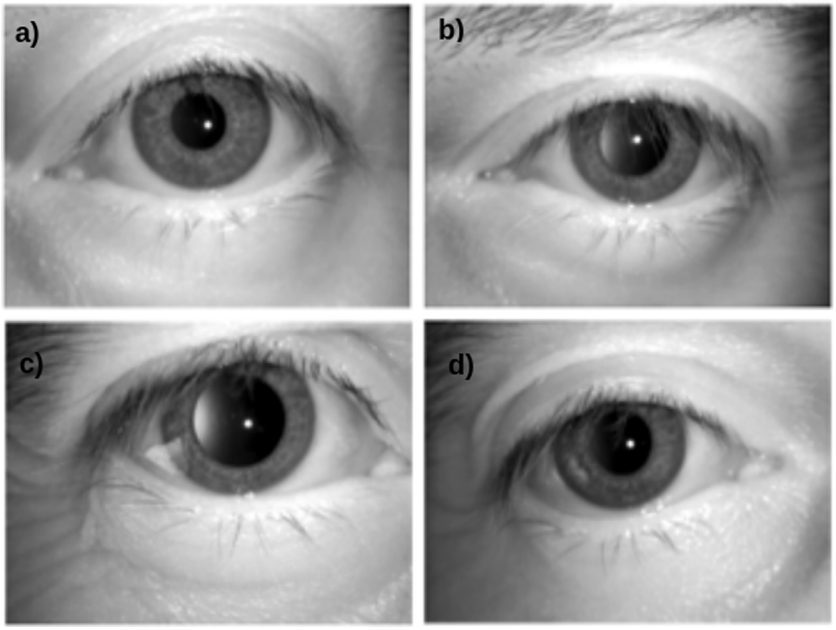


Fig. 4. Examples of the NIR images captured. (a) Control, (b) Alcohol, (c) Drug, and (d) Sleep images.

vectors (with the recording) and no longer with the image itself. See Fig. 8.

The results showed significant inter-class differences in the behaviour between the fit (control) and unfit (alcohol, sleep, and drug) classes. Feature extraction and selection methods were applied to the input vector and to the sequence machine learning model. The feature vector comprises 50 variables representing pupil behaviour throughout the captures. These feature vectors are the input for the classification model.

The pupil-iris ratio was also used to generate the feature vectors to train and test the machine-learning models. The pupil-iris ratio allows to eliminate pupil distortion in the computation and iris radius due to the subject's distance from the capture device.

The variables of the feature vector are based on all frames captured within one session at the frame rate of 15 fps and can be described as follows:

- Sequence trend: It allows for the evaluation of the temporal behaviour of the sequence of frames (stemming from one single capture session). It is estimated using the least-squares method to determine the slope m and the intercept b of the series. See Eqs. (1) and (2):

$$m = \frac{n \cdot \sum(x_i \cdot y_i) - \sum x_i \cdot \sum y_i}{n \cdot \sum {x_i}^2 - |\sum x_i|^2} \tag{1}$$

$$b = \frac{\sum y_i \cdot \sum {x_i}^2 - \sum x_i \cdot \sum(x_i \cdot y_i)}{n \cdot \sum {x_i}^2 - |\sum x_i|^2} \tag{2}$$

where x_i represents the time period at instant i (independent variable), y_i the value of the function at the corresponding time stamp at the instant i (dependent variable) and n corresponds to the number of observations.

- Starting sequence trend: it corresponds to the estimation of the trend, but it is measured in the first second of the sequence. In the unfit cases, the device lights produce a slower change in the pupil-iris ratio than in the fit cases, resulting in a steeper slope for the control subjects. It is calculated using the same definition as described above.

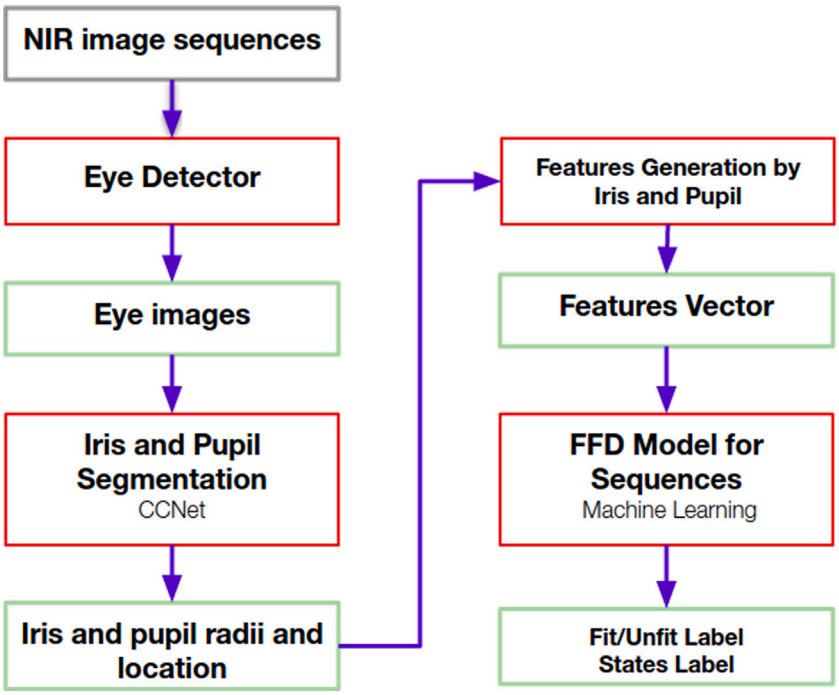


Fig. 5. Block diagram of the proposed FFD model based on image sequences. The input is NIR image sequences followed by the eyes detector, eye segmentation, feature extraction, and the FFD model. The output vector is the input to the machine learning classifier. The system delivers the FFD level: the fit/unfit indicator.

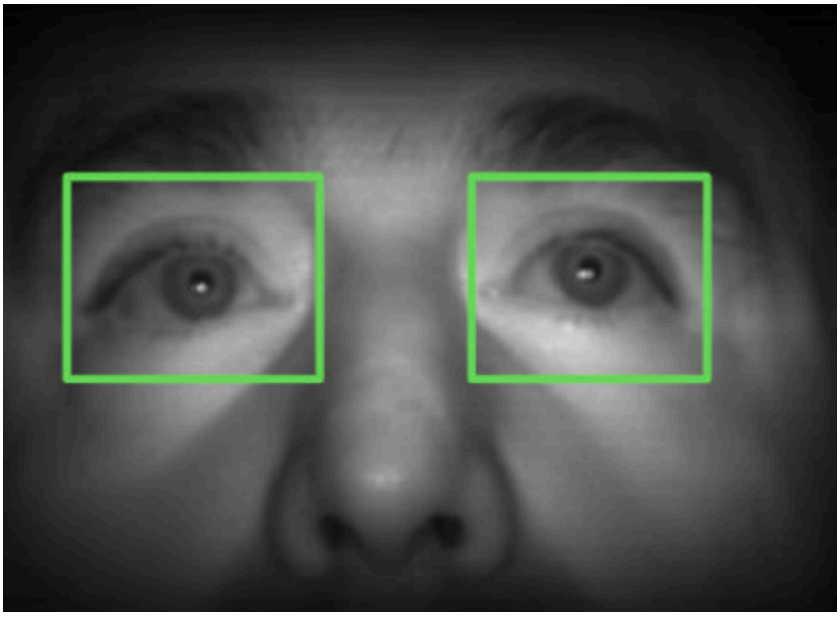


Fig. 6. Example of Eye-tiny-yolo detection module applied on periocular images with both eyes automatically detected.

- Moving sequence trend: the linear regression was estimated using a displacing rectangular window of one second with an overlap of 0.5 s between windows.
- Distance to the representative curves: the distance between the analysed sequence and each of the representative curves of the classes is calculated. See Eq. (3). Both sequence and class lines are estimated by linear regression. This distance d is obtained according to the following equation. See Fig. 9.

$$d(r,s) = \frac{|[\vec{v_r}, \vec{v_s}, \overrightarrow{P_r P_s}]|}{|\vec{v_r} \times \vec{v_s}|} \tag{3}$$

where r and s represent the lines, $\vec{v_r}$ y $\vec{v_s}$ corresponding to the direction vectors of lines r and s respectively, and $\overrightarrow{P_r P_s}$ corresponds to the vector formed by a point on each line.

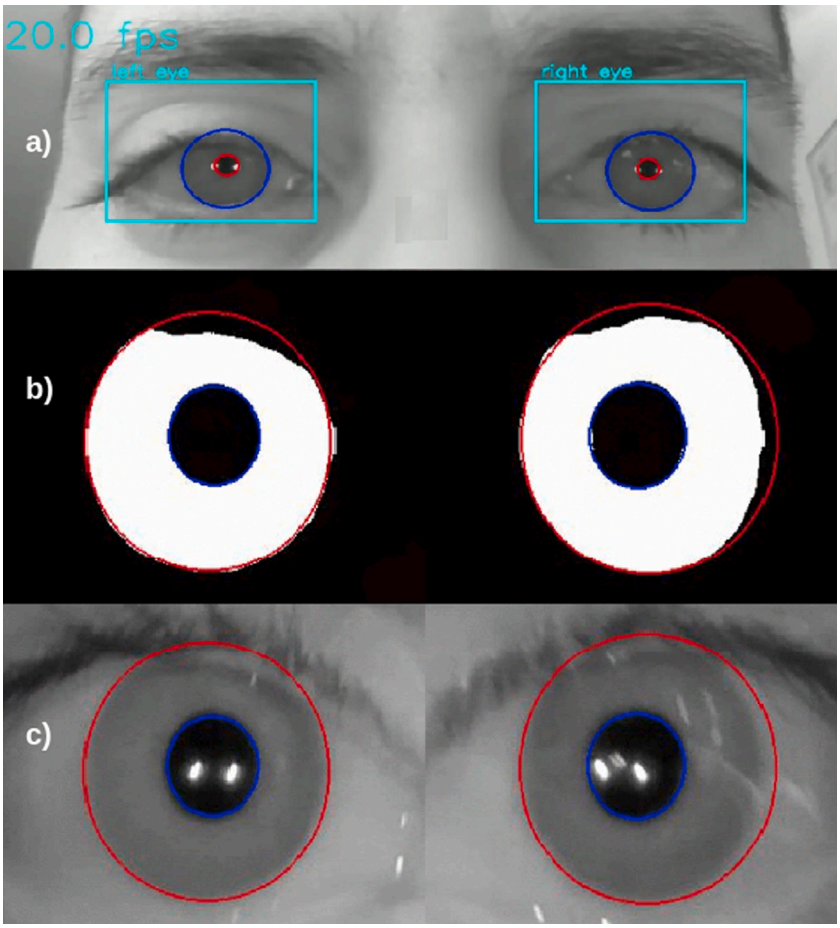


Fig. 7. Example of iris and pupil segmentation process. (a) Periocular NIR image with the Eye-tiny-yolo detector, (b) Mask for pupil and iris, and (c) Result of the segmentation process to determine the iris and pupil circles used to define the radii and centres.

Statistics values: Statistic measures were calculated for each sequence:

- Arithmetic Mean (μ_{seq}): represents the sum of values of the sequence divided by the number of elements.
- Standard Deviation (σ_{seq}): represents the sequence's quantified variation or dispersion of values.
- Range: represents the difference between the lowest and highest values in the sequence.
- Coefficient of Variation: measurements of the relative data dispersion points in the sequence around the mean. See Eq. (4):

$$CV_{seq} = \frac{\sigma_{seq}}{\mu_{seq}} \tag{4}$$

- Coefficient of variation to the representative curves: a measure of the relative dispersion data points in the analysed sequence around the mean of each of the representative curves in each condition. See Eq. (5):

$$CV_{pb_i} = \frac{\sigma_{pb_i}}{\mu_{seq}} \tag{5}$$

where i = control, alcohol, drug, sleep, Pupil-Iris Ratio values: this metric represents the real-time ratio between the pupil radius and the iris value. It corresponds to each of the values of the first 5 s of the sequence.

3.4. FFD classifiers

The temporal change variables related to the iris and pupil radii were used to train and test three different machine learning models. The extracted 50 features were used as input to the classifier. The classifier performs the estimation of the "Fit" and "Unfit" groups. In addition, these models were also used to classify each of the possible states (control, alcohol, drugs, and sleep). The tested algorithms were: Random Forest (RF) (Breiman, 2001), Gradient Boosting Machine (GBM) (Friedman, 2001), and Multi-Layer Perceptron (MLP) Neural Network (Rosenblatt, 1958). The models were trained and tested with the sequence sets shown in Table 1.

RF and GBM were selected because they showed promising results for small data sets. This characteristic is essential in the cases of drug and sleepiness recordings, where the pupil size is considerably smaller than in the control and alcohol cases. On the other hand, RF, GBM and MLP are widely tested techniques in unbalanced databases, as in this case, where unfit subjects represent about 13% of the records, consistent with the known statistics about alcohol, drug consumption, and fatigue presence in workspaces.

Optimal values of model hyper-parameters were obtained using a grid search: an exhaustive search performed on the specific parameter values of a model. Table 2 shows the selected hyper-parameters that maximised the precision score in each model.

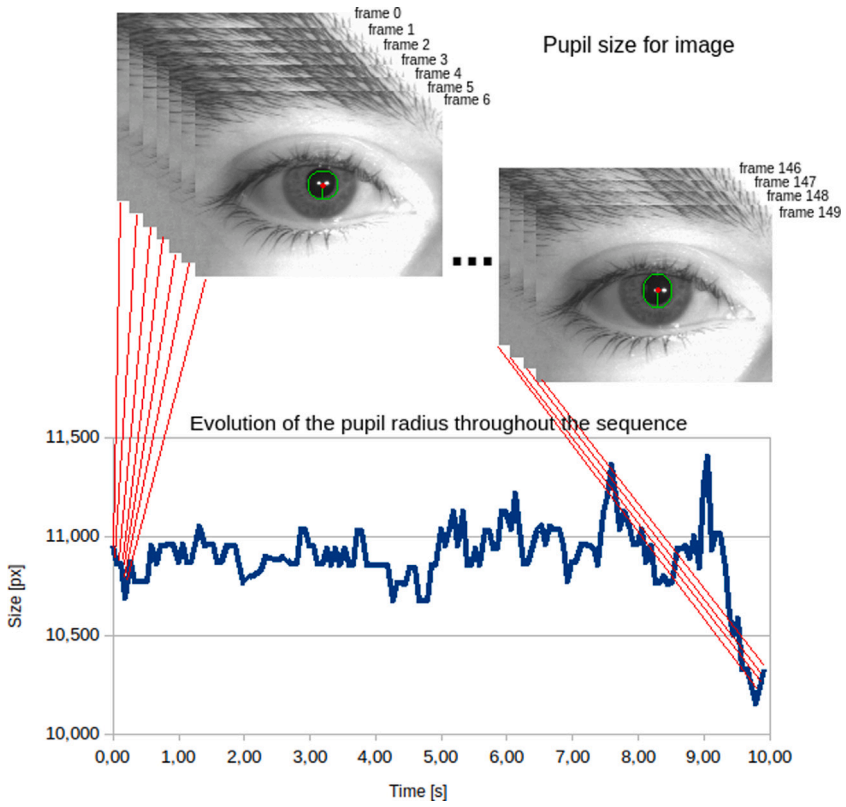


Fig. 8. Example of the process to transform individual measurements per image to a time sequence per capture.

Table 2
Classifiers hyper-parameters. The "–" indicates that this hyper-parameter does not apply to the type of model.

Hyper-parameters	RF classifier	GBM classifier	MLP classifier
Number of estimators (n estimators)	1000	1000	–
Criterion (criterion)	"entropy"	"squared error"	–
Maximum depth of the tree (max depth)	5	5	–
Minimum samples split (min samples split)	5	10	–
Minimum samples leaf (min samples leaf)	3	5	–
Maximum number of features (max features)	"auto"	"auto"	–
Subsample (subsample)	–	1.0	–
Hidden layer neurons (hidden layer sizes)	–	–	(25, 10)
Activation function (activation)	–	–	"relu"
Solver for weight optimisation (solver)	–	–	"adam"
Loss function (loss)	–	"deviance"	"categorical cross-entropy"
Learning rate (learning rate)	–	10e−2	"constant"
Initial learning-rate (learning rate init)	–	–	10e−3
L2 penalty (alpha)	–	–	10e−5
Size of mini-batches (batch size)	–	–	"auto"
Maximum number of iterations (max iter)	–	–	300
Cross-validation (KFold)	5	5	5

4. Results

4.1. FFD model for image sequences

The system was trained, and the parameters were adjusted employing an iterative process using the training and validation datasets. The final performance was measured using the test set.

4.1.1. Four classes

Table 3 presents the results for a four-class classifier (control, alcohol, drug and sleepiness). The metrics used are the following in the Eq. (6) up to (9):

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

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

where, TP is a True Positive, TN is a True Negative, FP is a False Positive and FN is the False Negative.

The general statistics analysis shows that the control and alcohol classes obtained the best performances. These results were obtained using a significantly higher number of captured images belonging to these classes. Conversely, drugs and sleep classes present a fewer number of images.

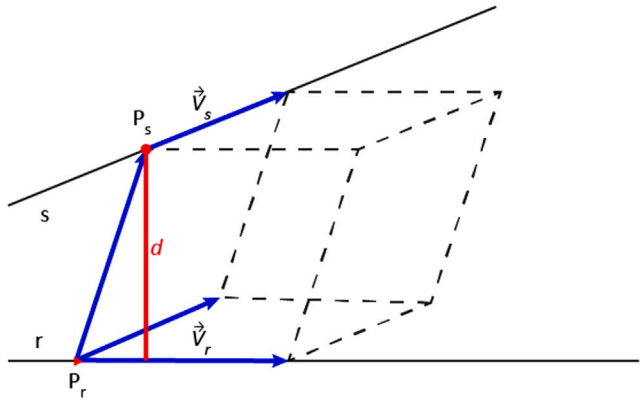


Fig. 9. This metric represents the estimation distance between two skew lines. This example defines the equation parameters to compute the distance to the representative curve for each class in the feature vector.

Table 3
Performance classifiers for all the classes.

Model	Class	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]
RF	Control	68.5	70.1	75.2	94.7
	Alcohol		73.6	71.7	20.5
	Drug		41.2	**99.2**	**53.8**
	Sleep		20.0	98.3	23.5
GBM	Control	71.6	73.1	**79.8**	**95.8**
	Alcohol		**76.4**	74.5	**22.9**
	Drug		**47.1**	99.1	53.3
	Sleep		**25.0**	98.5	29.4
MLP	Control	72.6	**75.3**	77.1	95.4
	Alcohol		70.8	76.3	22.9
	Drug		29.4	98.5	29.4
	Sleep		25.0	**98.8**	**35.7**

In the case of sensitivity, the classes for all trained models show values higher than 70.0% in control and alcohol classes. The performance decreased for drugs obtained in the best case at 40.0%, followed by the sleep class at around 25.0%.

In the case of specificity, the results for the control and alcohol groups in all models are between 75.0% and 80.0%. In the cases of drugs and sleep, specificity increases drastically; this could be due to the small number of cases for these classes.

Although the results are outstanding for overall accuracy, the test set is unbalanced among the four classes. The database imbalance is consistent with reality since it is estimated that in a real operation, between 10.0% to 15.0% of the subjects would not be able to perform their tasks correctly due to some external agent such as alcohol, drugs, or sleepiness.

4.1.2. Two classes

No new models were trained to analyse the results for the two-class classifier, Fit (control) and Unfit (alcohol, drug and sleepiness). Fig. 10 shows the classification results using confusion matrices for classes Fit and Unfit, which present the number of cases per class, the associated percentages and the mean accuracy per class.

Table 4 shows the results for each model considering only the fit and unfit (sum of records corresponding to alcohol, drug and sleep) classes. In this analysis, all the performance metrics show significant improvements. The sensitivity (correct detections for the fit class) for all models exceeds 70.0%, reaching 75.3% in the case of MLP. The specificity (correct detections for the unfit class) reaches 75.0%, with a best result of almost 80.0% for GBM. At the same time, the precision is close to 95.0% for all models. In the case of overall accuracy, since one is dealing with non-balanced classes, the results are significant,

Table 4
Classifiers performance for fit and unfit classes.

Model	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]
RF	70.8	70.1	75.2	94.7
GBM	73.1	73.1	79.8	95.8
MLP	75.3	75.3	77.1	95.4

showing above 70.0% in all cases. The best performance is MLP, followed by GBM, and then RF.

As expected, the results improved significantly by grouping the unfit subjects into a single class. For the confusion matrices, the average accuracy per class increases to 72.6% for RF, 76.5% for GBM and 76.2% for MLP.

4.2. Discussion

Note that all the classifiers presented similar results, which indicates that the models are well-trained. The database and the its separation into the subsets were also done correctly.

Although all indicators are essential to study the performance of the models, given the characteristics of the performed study and its applications in real operations, sensitivity and accuracy become relevant indicators. Sensitivity gives us information on the system's ability to separately detect the classes of interest. At the same time, accuracy provides an overview of the model, which is relevant in the Fit and Unfit class analysis case.

For the classification of each class, RF shows sensitivity results above 70.0% for the control and alcohol sets, but the performance declines significantly for drugs and alcohol datasets. However, the misclassifications for the latter two groups are primarily associated with unfit cases. Therefore, if we perform the algorithm analysis for fit and unfit classes, the sensitivity improves substantially, reaching 70.1% for fit and 75.2% for unfit. Likewise, the overall accuracy goes to 70.8%.

The individual results (by class) for GBM show similar behaviour to the ones described for RF. However, the results show an improvement in all groups, outperforming RF results by 3 to 5 points. On the other hand, when analysing the fit and unfit cases, GBM shows 73.1% and 79.8% of sensitivity, respectively, and an overall accuracy of 74.0%.

On the other hand, the MLP shows the best results by class for the control group with 75.3% sensitivity. For the cases of alcohol and drugs, the sensitivity decreases for the other models, reaching 70.8% and 29.4%, respectively. In contrast, the sleep condition remains similar to the other models. The fit and unfit groups analysis obtained a sensibility of 75.3% and 77.1%, respectively. The overall accuracy was 75.5%, the highest of all the models.

5. Behavioural curves

The behavioural reaction on the training set was studied to determine differences in the pupil and iris radius size in conditions of alcohol and drug consumption or sleep deprivation (unfit) versus control subjects (fit). This analysis was performed using the grand mean algorithm. The grand mean of a set of multiple sub-samples is the mean of all observations: every data point, divided by the joint sample size (Jin et al., 2020).

To obtain the control and alcohol, drug and sleepiness class curves, the grand mean algorithm was estimated for the pupil radii each time for all the subjects in the training set in each of the groups. Thus, it is possible to define the baseline behaviour curve for people in Fit and Unfit conditions. Note that this analysis shows the average behaviour, so it is not possible to use it as a single variable to separate the classes.

The two plots on the left in Fig. 11 present the differences in the temporal behaviour of pupils for the X and Y axes, with a greater

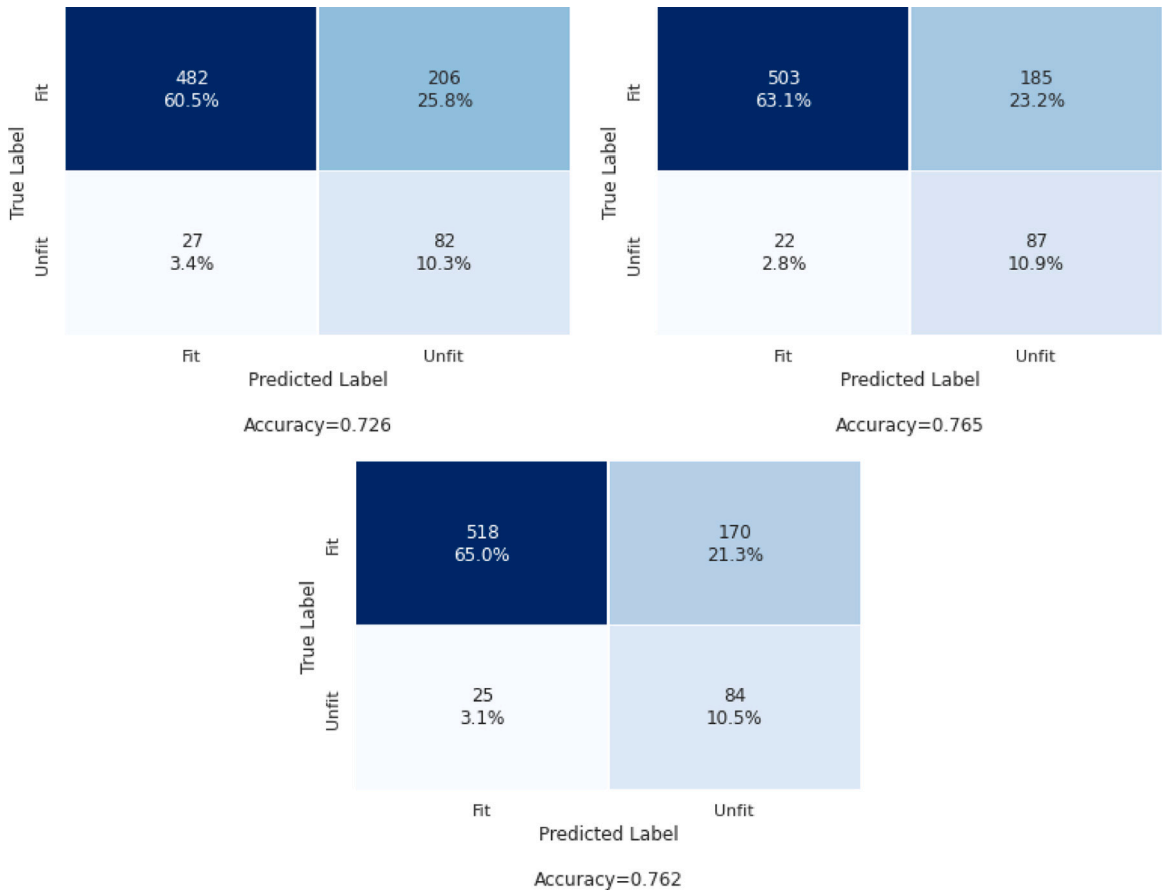


Fig. 10. Top. RF confusion matrix for Fit and Unfit classes. Results show the number of cases and the percentages in each case. Middle: GBM and Botton: MLP.

radius size (more significant dilation) in alcohol and drug subjects, followed by sleep deprivation records. In contrast, control subjects have the lowest dilation levels. All curves show an initial period with a pronounced increase in pupil radius size due to the natural adjustment process by the effect of the capture device. Then, control subjects tend to stabilise and produce a slower dilation level. In contrast, the dilation trend curves in the unfit cases present a higher slope, especially pronounced in alcohol records. The behaviour of sleep recordings is closer to that of control subjects. The two plots to the right in Fig. 11 show the behaviour of the iris throughout the sequence. In contrast to the pupil, the iris does not offer a significant variation for control subjects concerning the other classes.

Although the curves show a slight shift in the iris size, this effect is related to the unfit subjects as they tend to keep their eyes closed and blink more times, and in many cases, move in front of the capture device. This action makes the segmentation process difficult and, therefore, the estimation of the iris radius. This effect is not evident in the case of the pupil due to its smaller size. In the case of the radii values measured along the horizontal axis, as pointed out above, it is impossible to establish significant differences between the classes, which is correct. Also, there is no eye closure effect in this direction. However, the impact of partial eye closure appears in the case of iris radius estimation along the vertical axis. This effect is more marked in under-drug subjects, who present a completely different curve than the other classes, which have considerably smaller iris sizes. This effect is a consequence of the tendency of subjects under the influence of drugs (mainly marijuana) to have droopy eyelids, and therefore, the estimation of the iris size is affected. On the other hand, the difference in the beginning values recorded in both analysed axes is due to the natural adjustment process subject in front of the capture device.

The Eye-tiny-yolo model tries to adjust the cropping due to the subject's distance concerning the capture device. In order to mitigate this problem, the ratio between pupil and iris radius was studied as the quotient for each time of both measurements (see first two left plots in Fig. 12). Thus, the effect of the subject's distance from the camera is eliminated.

For the estimation on both the horizontal and vertical axes, it can be observed that the temporal behaviour of the groups is entirely different, showing a clear separation between them.

The control group shows the lowest pupil-iris ratio, which is consistent with the behaviour of pupil radius (Fig. 11). In the case of sleep-deprived subjects, a difference can be observed concerning the control group, but not very marked. This difference could be due to the low number of recordings of this type or the lack of subjects with total sleep deprivation or night shift schedule work. Subjects who have consumed alcohol and drugs show the highest ratios, which means a very significant pupil dilation. The iris size effect described above for the drug recordings on the vertical axis is reflected in the estimation of the pupil-iris ratio in this group, which shows the most significant difference when the analysis is made along this axis.

Fig. 12 presents the estimation of the distance of the pupil and iris centres concerning the (0,0) position of the image. It can be seen that in the control cases, the subject tends to maintain a more stable posture throughout the registration. In contrast, the behaviour is more erratic for the unfit subjects, with constant movements in front of the capture device and moving away from the established reference point.

Fig. 13 shows differences between the variables analysed for each of the classes. Although there are overlaps in the ranges, the centrality metrics show differences, especially for the mean values in each class.

In order to validate that the differences in centrality metrics between the classes were significant for the various variables, statistical tests were applied to determine if the difference was substantial. When

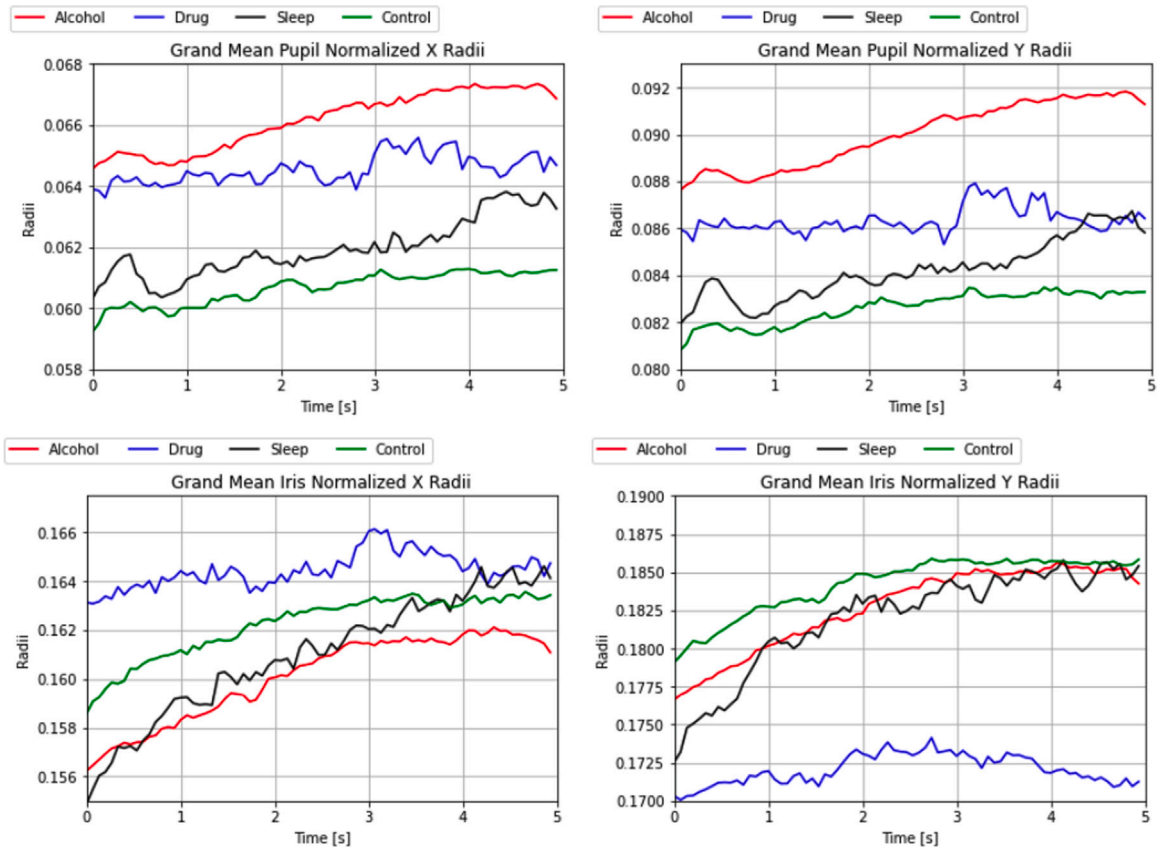


Fig. 11. Grand mean curves for the normalised **pupil** and **iris radius** on the alcohol (red), drug (blue), sleepiness (black) and control (Green). The main differences between classes are observed for the pupil radii variables.

Table 5
Results of the application of the Kruskal–Wallis test. The "Equal Var." column indicates whether the null hypothesis is accepted (medians are equals).

Variable	H-statistic	*p*-value	Equal Var.
Pupil horizontal axis (X-axis)	1279.60	3.92e−277	False
Pupil vertical axis (Y-axis)	1212.88	1.17e−262	False
Iris horizontal axis (X-axis)	300.45	7.93e−65	False
Iris vertical axis (Y-axis)	1167.59	7.85e−253	False
Pupil-iris ratio horizontal axis (X-axis)	3534.55	4.68e−26	False
Pupil-iris ratio vertical axis (Y-axis)	3009.30	2.46e−23	False
Pupil location	1423.17	2.75e−308	False
Iris location	1303.59	2.44e−282	False

performing normality tests on the data for the multiple variables analysed using the Shapiro–Wilk normality test, it was found that the data did not meet this condition for applying the ANOVA test. Although the ANOVA test is robust even when there is some lack of normality, in our case, other conditions make using this test not entirely appropriate. Among them are the imbalance of the classes and the lack of homoscedasticity for some variables. Therefore, a non-parametric test was applied, precisely the Kruskal–Wallis test, which compares the medians of three or more independent groups. It is an extension of the Mann–Whitney U test (or Wilcoxon rank-sum) for more than two groups. It is employed when the assumptions of ANOVA cannot be met, especially concerning normality and homogeneity of variances.

By performing the Kruskal–Wallis test, the H statistic is obtained. This serves a similar purpose as the F value in ANOVA. A considerable H value suggests significant differences between the group medians. Alongside the H statistic, a p-value is determined. If this p-value falls below a predetermined significance level (in this case, 0.05), the null hypothesis is rejected. Within the framework of the Kruskal–Wallis test, the null hypothesis posits that all group medians are similar. A small p-value suggests that the medians of at least two groups differ.

From Table 5, it can be observed that the null hypothesis is rejected for all variables. Therefore, there is a significant difference between the medians for each of the classes across all analysed variables. For example, in the case of the variable "pupil vertical axis radii", the H-value was 1212.8807 and the p-value of $1.1762e{-}262$.

From the Kruskal–Wallis analysis, it is possible to know that class differences are statistically significant, but it does not tell which classes are significantly different from each other. To know the pairs of significantly different classes it is necessary to perform multiple pairwise comparisons (post hoc comparison) analysis for all unplanned comparisons using the Dunn test. This test makes multiple pairwise comparisons and is especially useful when the data do not meet the assumptions of parametric tests, such as normal distribution and homogeneity of variances. The Dunn test results in adjusted p-values that account for the multiple comparisons, ensuring that the risk of Type I errors is controlled.

Table 6 shows the results of the Dunn test. Dunn's analysis showed that there were significant differences in most variables between the control-alcohol, control-drug, alcohol-sleep, and drug-sleep classes. In the case of control sleep, a significant difference is not observed for some variables, and the same is true for alcohol drugs. This can be seen in the temporal behaviour graphs (Figs. 11 to 12), in which the control-sleep curves are close and similar, and the same happens in the case of alcohol drugs.

The analysed curves show differentiated behaviours among the groups. Thus, it is possible to use these values as baselines for the conduct of the subjects in the different states. Thus, it is possible to eliminate the process of defining baselines per individual (as commercial solutions do), which can be significantly prolonged and susceptible to being impersonated.

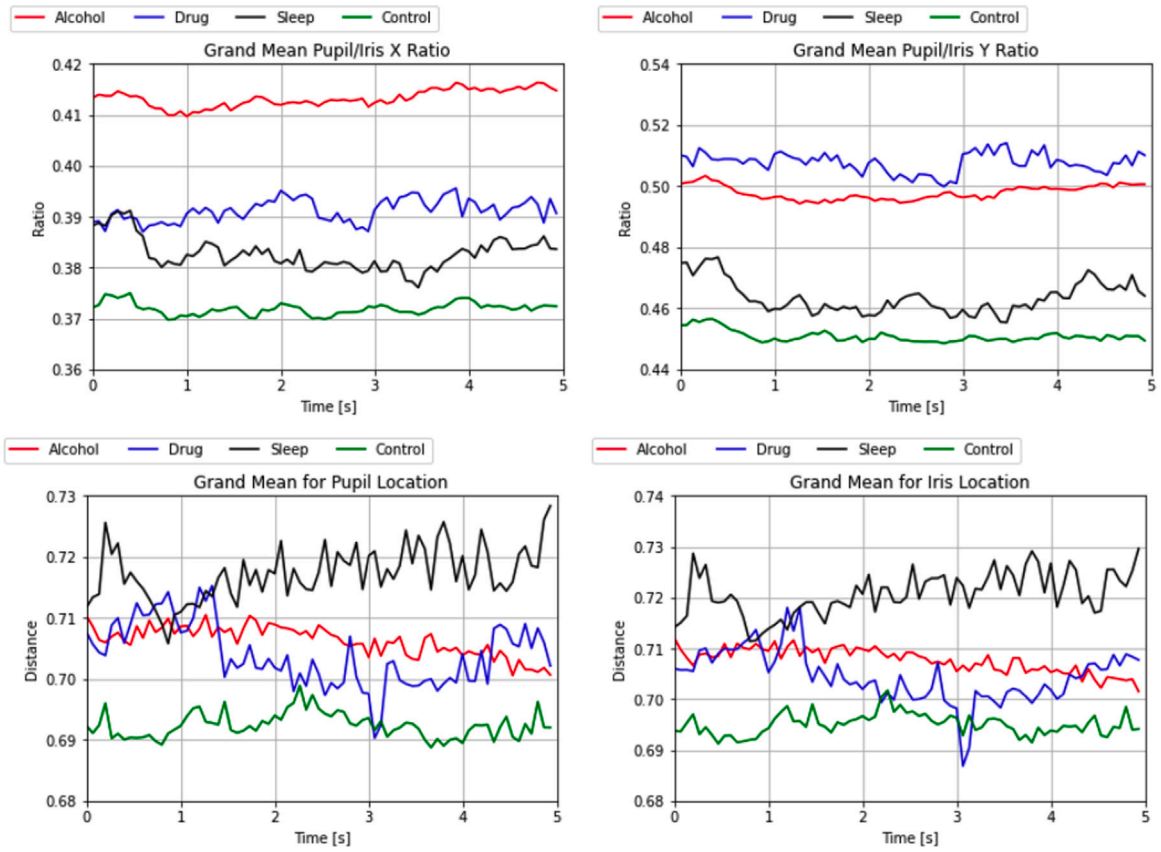


Fig. 12. Grand mean curves for **pupil-iris ratio**, **pupil location** and **iris location** on the alcohol (red), drug (blue), sleepiness (black) and control (Green). The main differences between classes are observed for the pupil-iris ratio variables.

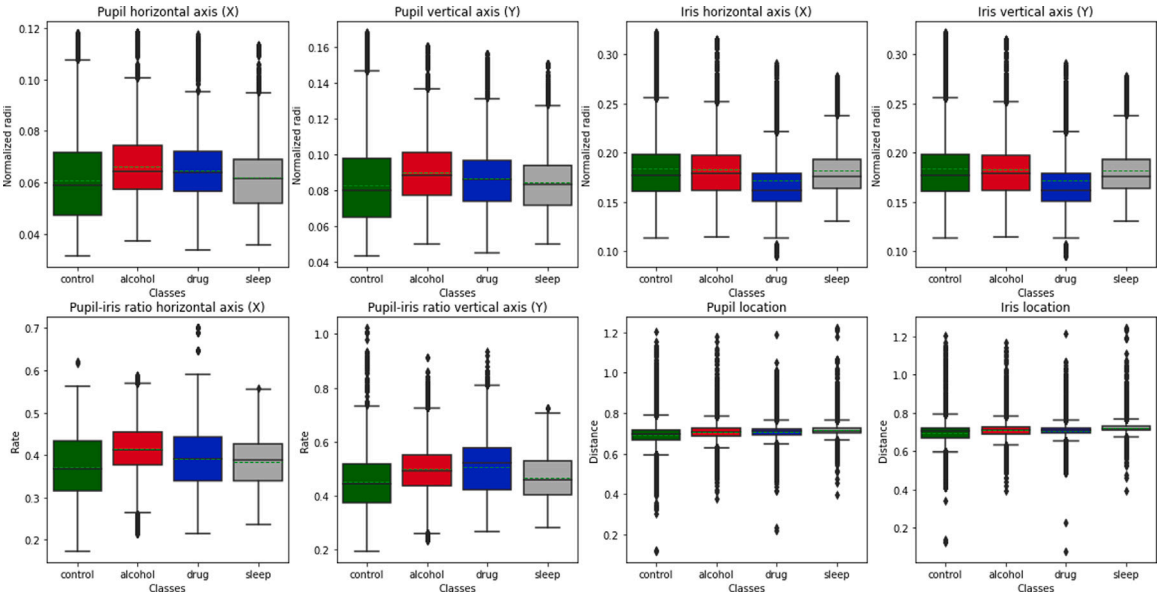


Fig. 13. Boxplot for the main variables for the behavioural analysis. Graphs show the behaviour of the data and their distribution. The dotted lines correspond to the mean value for each variable. Distinct differences in centrality and variability measures across classes are evident in several of the plots. However, note that these differences are subtle and not as immediately apparent for some variables. While discernible differences, such as in the mean values, are clear in many instances, it is imperative to employ statistical tests, parametric or non-parametric, to determine the significance of these observed differences.

6. Conclusions and remarks

This work performed a deep analysis of the behaviour of iris curves and showed essential differences in pupil and iris behaviour under different conditions that may affect the behaviour of the CNS. The differences between the groups were quite remarkable, particularly between the control cases and the subjects in the alcohol and drug data subsets. On the other hand, the behaviour of the drowsy or sleep-deprived subjects was closer to the control group. The latter could be due to the reduced number of sleep records in the database.

Table 6
Results from the application of the Dunn test. The table displays the variables and the result of applying the test (p-value). Each column corresponds to the grouping of classes for pairwise testing. P-values greater than 0.05 indicate that there are no significant differences in the medians when comparing both groups.

Variable	Control-alcohol	Control-drug	Control-sleep	Alcohol-drug	Alcohol-sleep	Drug-sleep
Pupil horizontal axis (X-axis)	0.000	0.000	0.478	0.619	0.000	0.000
Pupil vertical axis (Y-axis)	0.000	0.000	0.106	0.693	0.000	0.000
Iris horizontal axis (X-axis)	0.000	0.000	0.724	0.000	0.000	0.000
Iris vertical axis (Y-axis)	0.362	0.000	0.144	0.000	0.003	0.000
Pupil-iris ratio horizontal axis (X-axis)	0.000	0.000	0.000	0.000	0.000	0.000
Pupil-iris ratio vertical axis (Y-axis)	0.000	0.000	0.000	0.000	0.000	0.000
Pupil location	0.000	0.000	0.000	0.068	0.000	0.000
Iris location	0.000	0.000	0.000	0.318	0.000	0.000

Based on the results, it can be seen that all models performed similarly for both types of analysis. The best results were obtained in the control and alcohol datasets, which are the most extensive and consolidated subsets in our database. At the same time, the drug and sleep records are smaller (17 and 20, respectively, in the test set). This factor reduces the number of records that affect the model's performance, especially considering that we worked with an unbalanced database. However, when the analysis was performed only between Fit and Unfit classes, the results improved substantially, and in all cases exceeded 70.0%, reaching in some cases values close to 80.0%. These are outstanding results because they indicate that the system is able to detect many individuals who are not fit to perform any activity and, therefore, avoid exposure to any level of risk that could cause harm.

In particular, MLP and GBM showed the best results in all groups, allowing the detection of about 79.8% and 77.1%, respectively, of the subjects who could not perform a task safely. This result is an excellent measure and allows us to lower the risk levels of a critical operation that requires fitness in a tangible way.

On the other hand, the results show that it is possible to detect, on average, 8 out of 10 subjects in unfit conditions without altering the regular operation of an industry since the system has a specificity of more than 95.0% for control (fit) subjects. Furthermore, the database replicates the expected behaviour of this type of operation, in which 10.0% to 15.0% of the subjects are in unfit conditions to perform a task. Contrary to previously presented systems, which require immediate and complete attention of the captured subject, the proposed system could be operated as a concurrent observation of the subject, e.g., truck driver or pilot, without disturbing his primary duties. Thus, with the analysis over multiple time windows, the classification results can be further improved.

In future work, we would like to extend this work by considering three different approaches. The first one is to continue capturing volunteers for the database to increase the balance among classes. Second, the use of deep learning techniques, especially Recurrent Neural Networks (RNN) such as Long Short Term Memory (LSTM) (Rajamohana, Radhika, Priya, & Sangeetha, 2021), will be explored. This type of network allows for working naturally with time series and modelling the temporal behaviour in the internal structure of the network (Monteiro-Thiago, Skourup, & Zhang, 2020; Zhang, Chen, Chen, Li, & Li, 2021; Zhong, Fares, & Jiang, 2019).

CRediT authorship contribution statement

Leonardo Causa: Methodology, Database, Sequence Analysis, Writing – original draft. **Juan E. Tapia:** Conceptualization, Methodology, Database, Funding acquisition, Formal analysis, Investigation, Writing – original draft, Resources. **Andres Valenzuela:** Eye segmentation. **Daniel Benalcazar:** Eye detector. **Enrique Lopez Droguett:** Writing – review & editing, Funding acquisition, Resources. **Christoph Busch:** Funding acquisition and Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Juan E. Tapia reports financial support was provided by National Agency for Research and Development. This work is partially supported by the Agencia Nacional de Investigacion y Desarrollo (ANID) through FONDEF IDEA N° ID19I10118 led by Juan Tapia Farias - DIMEC-UChile. Further, this work has been partially supported by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work is partially supported by the Agencia Nacional de Investigacion y Desarrollo (ANID), Chile through FONDEF IDEA ID19I10118 and the European Union's Horizon 2020 research and innovation program under grant agreement No 883356 and the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, Germany.

Consent for publication

Informed consent has been obtained from all participants included in the analysed studies, and the studies are being conducted in accordance with the declaration of Helsinki.

References

Adler, F. H. (1985). In J. A. Alexander (Ed.), *Physiology of the eye, vol. 48* (11th ed.). Francis Heed Adler, The C. V. Mosby Company.

Amodio, A., Ermidoro, M., Maggi, D., Formentin, S., & Savaresi, S. M. (2019). Automatic detection of driver impairment based on pupillary light reflex. *IEEE Transactions on Intelligent Transportation Systems*, *20*(8), 3038–3048.

Azadani, M. N., & Boukerche, A. (2021). Driving behavior analysis guidelines for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 1–19.

Bai, Y., Guan, Y., & Ng, W.-F. (2020). Fatigue assessment using ECG and actigraphy sensors. In *Proc. of the intl. symposium on wearable computers, ISWC '20* (pp. 12–16). New York, NY, USA: Association for Computing Machinery.

Bakker, B., Zabłocki, B., Baker, A., Riethmeister, V., Marx, B., Iyer, G., et al. (2021). A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, 1–10.

Balasubramanian, A., Kohn, T. P., Santiago, J. E., Sigalos, J. T., Kirby, E. W., Hockenberry, M. S., et al. (2020). Increased risk of hypogonadal symptoms in shift workers with shift work sleep disorder. *Urology*, *138*, 52–59.

Balkin, T. J., Bliese, P. D., Belenky, G., Sing, H., Thorne, D. R., Thomas, M., et al. (2004). Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *Journal of Sleep Research*, *13*(3), 219–227.

Benalcazar, D. P., Tapia, J. E., Vasquez, M., Causa, L., Droguett, E. L., & Busch, C. (2023). Toward an efficient iris recognition system on embedded devices. *IEEE Access*, *11*, 133577–133590.

Benderoth, S., Hormann, H.-J., Schiebl, C., & Elmenhorst, E.-M. (2021). Reliability and validity of a 3-min psychomotor vigilance task in assessing sensitivity to sleep loss and alcohol: fitness for duty in aviation and transportation. *Sleep*, *44*(11).

Birnbaum, H. G., White, A. G., Schiller, M., Waldman, T., Cleveland, J. M., & Roland, C. L. (2011). Societal costs of prescription opioid abuse, dependence, and misuse in the United States. *Pain Medicine*, *12*(4), 657–667.

Borrelli, I., Gualano, M. R., Rossi, M. F., Capitanelli, I., Dolgetta, V., Santoro, P. E., et al. (2023). Alcohol consumption in healthcare workers and risk of workplace injury: a case-control study. *Discover Sustainability*, *4*(1), 1–8.

Breiman, L. (2001). Random forests. *Journal of Machine learning*, *45*(1), 5–32.

Butler, P., & Fee, W. (2015). Fatigue and the use of wearable technology. In *SPE health, safety, security, environment, & social responsibility conference-North America* (pp. SPE–173541). SPE.

Caldwell, J., Mallis, M., Caldwell, L., Michel, P., Miller, J. C., & Neri, D. F. (2009). Fatigue countermeasures in aviation. *Aviation, Space, and Environmental Medicine*, *80*(1), 29–59.

Campbell, I., Beckers, E., Sharifpour, R., Berger, A., Paparella, I., Balda Aizpurua, J. F., et al. (2013). Impact of light on task-evoked pupil responses during cognitive tasks. In *bioRxiv*, (pp. 2023-2004).

Chandler, J. F., Arnold, R. D., Jeffrey B snd Phillips, R. A. L., & Horning, D. S. (2010). *Preliminary validation of a readiness-to-fly assessment tool for use in naval aviation*: *Technical report 10-22*, Pensacola, Florida: Naval Aerospace Medical Research Laboratory.

Council, N. S. (2017). Cost of fatigue in the workplace.

Dorrian, J., & Skinner, N. (2012). Alcohol consumption patterns of shiftworkers compared with dayworkers. *Chronobiology International*, *29*(5), 610–618.

Farkhodov, K., Lee, S.-H., & Kwon, K.-R. (2020). Object tracking using CSRT tracker and RCNN. In *BIOIMAGING* (pp. 209–212).

Ferguson, B. A., Lauriski, D. R., Huecker, M., Wichmann, M., Shreffler, J., & Shoff, H. (2020). Testing alertness of emergency physicians: A novel quantitative measure of alertness and implications for worker and patient care. *The Journal of emergency medicine*, *58*(3), 514–519.

Florence, C. S., Zhou, C., Luo, F., & Xu, L. (2016). The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. *Medical Care*, *54*(10), 901–906.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Frone, M. (2006). Prevalence and distribution of alcohol use and impairment in the workplace: a U.S. national survey. *Journal of Studies on Alcohol*, *67*(1), 147–156.

Ftouni, S., Rahman, S. A., Crowley, K. E., Anderson, C., Rajaratnam, S. M., & Lockley, S. W. (2013). Temporal dynamics of ocular indicators of sleepiness across sleep restriction. *Journal of Biological Rhythms*, *28*(6), 412–424.

Gonzalez, A. J., Wong, J. M., Thomas, E. M., Kerrigan, A., Hastings, L., Posadas, A., et al. (2022). Detection of driver health condition by monitoring driving behavior through machine learning from observation. *Expert Systems with Applications*, *199*, Article 117167.

Gruenhagen, J. H., Parker, R., & Cox, S. (2021). Technology diffusion and firm agency from a technological innovation systems perspective: A case study of fatigue monitoring in the mining industry. *Journal of Engineering and Technology Management*, *62*, Article 101655.

Guede-Fernández, F., Fernández-Chimeno, M., Ramos-Castro, J., & García-González, M. A. (2019). Driver drowsiness detection based on respiratory signal analysis. *IEEE Access*, *7*, 81826–81838.

Gusman, E., Standlee, J., Reid, K. J., & Wolfe, L. F. (2023). Work-related sleep disorders: causes and impacts. In *Seminars in respiratory and critical care medicine, vol. 44* (pp. 385–395). Thieme Medical Publishers, Inc. 333 Seventh Avenue, 18th Floor, New York, NY ....

Han, Y., Yin, Z., Zhang, J., Jin, R., & Yang, T. (2020). Eye-tracking experimental study investigating the influence factors of construction safety hazard recognition. *Journal of Construction Engineering and Management*, *146*(8), Article 04020091.

Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., et al. (2020). CCNet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Iqbal, K., Shafiq, M. A., Singh, S., & Afzal, M. K. (2023). Impact of opioid use disorder (OUD) on employee productivity: An empirical investigation. *International Journal of Business Intelligence and Big Data Analytics*, *6*(1), 23–30.

Jayadev, P. G., & Bellary, S. (2021). A hybrid approach for classification and identification of iris damaged levels of alcohol drinkers. *Journal of King Saud University - Computer and Information Sciences*.

Jeong, I. C., Bychkov, D., & Searson, P. C. (2019). Wearable devices for precision medicine and health state monitoring. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1242–1258.

Jin, J., Chen, Z., Xu, R., Miao, Y., Wang, X., & Jung, T.-P. (2020). Developing a novel tactile P300 brain-computer interface with a cheeks-stim paradigm. *IEEE Transactions on Biomedical Engineering*, *67*(9), 2585–2593.

Johns, M., & Hocking, C. (2021). The effects of unintentional drowsiness on the velocity of eyelid movements during spontaneous blinks. *Physiological Measurement*, *42*(1), Article 014003.

Johns, M. W., Tucker, A., Chapman, R., Crowley, K., & Michael, N. (2007). Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie - Schlafforschung und Schlafmedizin*, *11*, 234–242.

Kaewkannate, K., Kaeokwunoiand, A., & Kim, S. (2015). A review of wearable devices. In *ITC-CSCC: Intl. technical conf. on circuits systems, computers and communications* (pp. 909–912).

Kaewkannate, K., & Kim, S. (2016). A comparison of wearable fitness devices. *BMC Public Health*, *16*(1), 1–16.

Kandera, B., Škultéty, F., & Mesárošová, K. (2019). Consequences of flight crew fatigue on the safety of civil aviation. *Transportation Research Procedia*, *43*, 278–289, INAIR 2019 - Global Trends in Aviation.

Katona, J. (2022). Measuring cognition load using eye-tracking parameters based on algorithm description tools. *Sensors*, *22*(3).

Kim, J. H., Cho, Y., Suh, Y.-A., & Yim, M.-S. (2021). Development of an information security-enforced EEG-based nuclear operators' fitness for duty classification system. *IEEE Access*, *9*, 72535–72546.

MacQuarrie, A., Robertson, C., Micalos, P., Crane, J., High, R., Drinkwater, E., et al. (2018). Fit for duty: The health status of New South Wales Paramedics. *Irish Journal of Paramedicine*, *3*.

Mahmud, M. S., Fang, H., Carreiro, S., Wang, H., & Boyer, E. W. (2019). Wearables technology for drug abuse detection: A survey of recent advancement. *Smart Health*, *13*, Article 100062.

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In *Advances in neural information processing systems* (pp. 570–576). Citeseer.

Martini, F., Fregna, L., Bosia, M., Perrozzi, G., & Cavallaro, R. (2022). Substance-related disorders. In *Fundamentals of psychiatry for health care professionals* (pp. 263–295). Springer.

Monteiro-Thiago, G., Skourup, C., & Zhang, H. (2020). Optimizing CNN hyperparameters for mental fatigue assessment in demanding maritime operations. *IEEE Access*, *8*, 40402–40412.

Murillo, R., Crucilla, C., Schmittner, J., Hotchkiss, E., & Pickworth, W. B. (2004). Pupillometry in the detection of concomitant drug use in opioid-maintained patients. *Methods and Findings in Experimental and Clinical Pharmacology*, *26*(4), 271–275.

Murphy, S., & Fleming, T. (1992). Fitness for duty in the nuclear power industry: the effects of local characteristics. In *Record-fifth conf. on human factors and power plants* (pp. 127–132).

Natarajan, A., Parate, A., Gaiser, E., Angarita, G., Malison, R., Marlin, B., et al. (2013). Detecting cocaine use with wearable electrocardiogram sensors. In *Proc. of the 2013 ACM intl. joint conf. on pervasive and ubiquitous computing* (pp. 123–132).

Navarro, L. A., Diño, M. A., Joson, E., Anacan, R., & Cruz, R. D. (2016). Design alcohol detection system for car users thru iris recognition pattern using wavelet transform. In *7th Int. conf. on intelligent systems, modelling and simulation (ISMS)* (pp. 15–19).

NIDA (2020). Is drug addiction treatment worth its cost?.

Němcová, A., Svozilová, V., Bucsuházy, K., Smíšek, R., Mézl, M., Hesko, B., et al. (2021). Multimodal features for detection of driver stress and fatigue: Review. *IEEE Transactions on Intelligent Transportation Systems*, *22*(6), 3214–3233.

Persson, A., Jonasson, H., Fredriksson, I., Wiklund, U., & Ahlström, C. (2021). Heart rate variability for classification of alert versus sleep deprived drivers in real road driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, *22*(6), 3316–3325.

Peter, L., Reindl, R., Zauter, S., Hillemacher, T., & Richter, K. (2019). Effectiveness of an online CBTI intervention and a facetoface treatment for shift work sleep disorder. a comparison of sleep diary data. *International Journal of Environmental Research and Public Health*, *16*(17).

Pinheiro, H. M., da Costa, R. M., Camilo, E. N. R., da Silva Soares, A., Salvini, R., Laureano, G. T., et al. (2015). A new approach to detect use of alcohol through iris videos using computer vision. In V. Murino, E. Puppo (Eds.), *Image analysis and processing — ICIAP 2015* (pp. 598–608). Cham: Springer Intl. Publishing.

Rajamohana, S., Radhika, E., Priya, S., & Sangeetha, S. (2021). Driver drowsiness detection system using hybrid approach of convolutional neural network and bidirectional long short term memory. *Materials Today: Proceedings*, *45*, 2897–2901.

Riad, A. K. I., Shahriar, H., Zhang, C., & Barsha, F. L. (2021). Heath device security and privacy: A comparative analysis of fitbit, jawbone, google glass and samsung galaxy watch. In *Data protection and privacy in healthcare* (pp. 91–108). CRC Press.

Rodger, J. A. (2020). An expert system gap analysis and empirical triangulation of individual differences, interventions, and information technology applications in alertness of railroad workers. *Expert Systems with Applications*, *144*, Article 113081.

Rosekind, M. R., Gregory, K. B., Mallis, M. M., Brandt, S. L., Seal, B., & Lerner, D. (2010). The cost of poor sleep: Workplace productivity loss and associated costs. *Journal of Occupational and Environmental Medicine*, *52*(1), 91–98.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Russo, M., Thomas, M., Sing, H., Thorne, D., Balkin, T., Wesensten, N., et al. (1999). Saccadic velocity and pupil constriction latency changes in partial sleep deprivation, and correlations with simulated motor vehicle crashes. *Sleep*, *22*(Suppl 1), S297–8.

Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E., & Brewer, R. D. (2015). 2010 National and state costs of excessive alcohol consumption. *American Journal of Preventive Medicine*, *49*(5), E73–E79.

Sharma, L. D., Bohat, V. K., Habib, M., Al-Zoubi, A. M., Faris, H., & Aljarah, I. (2022). Evolutionary inspired approach for mental stress detection using EEG signal. *Expert Systems with Applications*, *197*, Article 116634.

Suhardi, B., Rosyidasari, A., Astuti, R. D., & Adiasa, I. (2022). Fitness for duty prediction model for bus driver of batik solo trans based on physical, mental, and work aspects. *Cogent Engineering*, *9*(1), Article 2143068.

Takano, A., Ono, K., Nozawa, K., Sato, M., Onuki, M., Sese, J., et al. (2023). Wearable sensor and mobile app–based mHealth approach for investigating substance use and related factors in daily life: Protocol for an ecological momentary assessment study. *JMIR Research Protocols*, *12*(1), Article e44275.

Tanveer, M. A., Khan, M. J., Qureshi, M. J., Naseer, N., & Hong, K.-S. (2019). Enhanced drowsiness detection using deep learning: An fNIRS study. *IEEE Access*, *7*, 137920–137929.

Tapia, J. E., Droguett, E. L., & Busch, C. (2022). Alcohol consumption detection from periocular NIR images using capsule network. In *26th Intl. conf. on pattern recognition (ICPR)* (pp. 959–966).

Tapia, J. E., Droguett, E. L., Valenzuela, A., Benalcazar, D. P., Causa, L., & Busch, C. (2021). Semantic segmentation of periocular near-infra-red eye images under alcohol effects. *IEEE Access*, *9*, 109732–109744.

Tapia, J. E., Perez, C. A., & Bowyer, K. W. (2016). Gender classification from the same iris code used for recognition. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1760–1770.

Vasiljevas, M., Damaševičius, R., & Maskeliūnas, R. (2023). A human-adaptive model for user performance and fatigue evaluation during gaze-tracking tasks. *Electronics*, *12*(5).

Wickwire, E., Geiger-Brown, J., Scharf, S., & Drake, C. (2017). Shift work and shift work sleep disorder, clinical and organizational perspectives. *Chest*, *151*(5), 1156–1172.

Xu, X., Bishop, E. E., Kennedy, S. M., Simpson, S. A., & Pechacek, T. F. (2015). Annual healthcare spending attributable to cigarette smoking: An update. *American Journal of Preventive Medicine*, *48*(3), 326–333.

Xu, S., & Hall, N. G. (2021). Fatigue, personnel scheduling and operations: Review and research opportunities. *European Journal of Operational Research*.

Yung, M. (2016). *Fatigue at the workplace: Measurement and temporal development* (Ph.D. thesis), University of Waterloo.

Zhang, L., Chen, D., Chen, P., Li, W., & Li, X. (2021). Dual-CNN based multi-modal sleep scoring with temporal correlation driven fine-tuning. 420, 317–328.

Zhong, S.-H., Fares, A., & Jiang, J. (2019). An attentional-lstm for improved classification of brain activities evoked by images. In *Proc. of the 27th ACM intl. conf. on multimedia, MM '19* (pp. 1295–1303). New York, NY, USA: Association for Computing Machinery.

Zurita, P. C., Benalcazar, D. P., & Tapia, J. E. (2023). Fitness-for-duty classification using temporal sequences of iris periocular images. In *2023 11th International workshop on biometrics and forensics (IWBF)* (pp. 1–6).