Teacher-Student Synergetic Knowledge Distillation for Detecting Alcohol Consumption in NIR Iris Images

Sanskar Singh(✉), Ravil Patel, Vandit Tyagi, and Avantika Singh

IIIT Naya Raipur, Raipur, Chhattisgarh, India
{sanskar21102,ravil21102,vandit21102,avantika}@iiitnr.edu.in

Abstract. Detection of alcohol consumption is critical for ensuring fitness for duty (FFD) at workplace. It ensures employee safety and productivity by reducing accidents and injuries while improving work efficacy. In this paper, we propose a framework based on teacher-student collaborative knowledge distillation for detecting alcohol consumption in NIR (Near-Infrared) iris images. Specifically, this research focuses on analyzing the impact of alcohol consumption on iris and pupil movements. We provide interesting experimental analysis and related discussions that demonstrates suitability of NIR camera based captured iris images for detecting alcohol consumption. Furthermore, this research can be seen as a progressive measure towards integrating alcohol detection in iris based biometric authentication systems.

Keywords: Alcohol detection · Fitness for duty · Knowledge distillation · Periocular NIR iris images · Vision Transformer

1 Introduction

Nowadays, abuse of intoxication in the workplace is rising proportionately. Working under the consumption of such substances can lead to a rise in work-related injuries, especially for laborers and heavy-machinery operators. According to a study by Pidd et al. [17], 11% of workplace accidents and injuries are caused by the consumption of alcohol. Companies incur approximately $2 billion per year in costs related to alcohol-related absenteeism. To overcome hassle, government of nations such as the UK and Australia have imposed duty of care legislation [12]. Under this legislation, employers are required to have an unambiguous policy that outlines acceptable conduct and misconduct. To ensure this fitness for duty(FFD) [18] is required in work area. To ascertain this few organisations have installed saliva and breath [8,11] analyzer for detecting alcohol consumption in the workplace. However, there are a few potential drawbacks of breath and saliva-based alcohol testing in the workplace which includes low accuracy, sensitivity and vulnerability to external influences such as mouthwash or food.

All authors have contributed equally.

N. Tsapatsoulis et al. (Eds.): CAIP 2023, LNCS 14185, pp. 162–171, 2023.
https://doi.org/10.1007/978-3-031-44240-7_16

Furthermore, these type of systems in the workplace may increase the risk of COVID-19 transmission due to hygiene and close contact concerns. Henceforth it is crucial to design new mechanisms that are capable of accurate and resilient detection of the effects of alcohol on employees, ensuring their fitness for duty.

It has been proved in literature that alcohol consumption can cause dilated pupils [3]. This motivated us to propose a framework that detects alcohol consumption by analyzing Non-Infrared (NIR) iris images. NIR iris imaging does not involve any physical contact with the user, unlike other alcohol detection techniques based on physiological fluids like breath or saliva. In the case of infectious diseases like COVID-19, iris imaging technique instead of physiological fluids thus lowers the likelihood of disease transmission. Furthermore, it is also quick and efficient that can deliver outcomes in real-time, making it appropriate for applications like law enforcement and for ensuring FFD at workplace.

2 Related Works

This section gives a brief summary of earlier research conducted in pertinent literature, exploring the effects of alcohol intake on changes in the iris and its impact on an individual's ability to perform their duties effectively. Amodio et al. [2] assessed the possibility of creating a system to detect drunk driving by analyzing changes in a person's pupillary light reflex (PLR) over time. The method involves using circular hough transform to obtain the pupil diameter profile, followed by implementing a polynomial-kernel support vector machine (SVM) to categorize the subject using the 8 features extracted from the profile.

In another work, Causa et al. [5] used a stream of NIR iris video frames to estimate behavioral curves. The study concentrated on applying a Criss-Cross Network (CCNet) to mask the iris and pupil segmentation, enabling the creation of characteristics based on the differences between the radii of the pupil and iris. The features produced were used to categorize the subject using a Multi-Layer-Perceptron (MLP) algorithm with an accuracy rate of 75.8%. In another notable work Arora et al. [3] studied the effects of alcohol on an iris recognition system and infer that one in five subjects under alcohol consumption may evade identification by iris recognition. Very recently, authors [20] have proposed a framework based on capsule network for detecting alcohol consumption.

3 Research Methodology

Here, in this section we will discuss the dataset used in our experimentation along with the feature extraction framework. Our proposed framework is based on teacher-student learning paradigm that relies on a distillation token [21] to ensure student network learning from the teacher network through a multi-head attention. An overview of the entire framework is presented in Fig. 1.

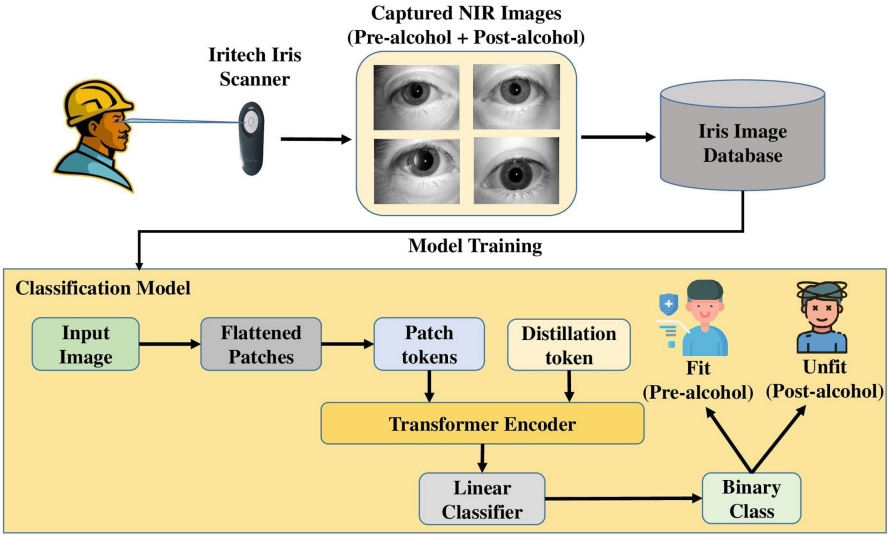

Fig. 1. Comprehensive description of the presented framework which determines fitness for duty on the basis of NIR iris images

Table 1. Statistical Description of the IAL-I Database

Session	Condition	Capture Time (min)	Images
S0	Pre-Alcohol	0	600
S1	Post-Alcohol	15	600
S2	Post-Alcohol	30	600
S3	Post-Alcohol	45	600
S4	Post-Alcohol	60	600

3.1 Dataset Description

In this study we have used IAL-I database [19]. This database consists of NIR iris images captured for total 30 subjects (24 males, 6 females) aged between 25 and 50. IAL-I dataset consists of nearly 20 similar periocular NIR iris images from each subject per session and there are total of 5 sessions. Table 1 illustrates the distribution of data. For more details kindly refer [19].

3.2 Feature Extractor

Vision Transformers (ViT) [10] can be seen as a de facto standard in the past few years for image classification tasks. Recently, aggregation of convnets and transformers integrated with self-attention mechanism have illustrated superfluous results in various domains like image classification [7], image segmentation [22] and natural language processing [14]. On continuation to this, Touvron et al. [21]

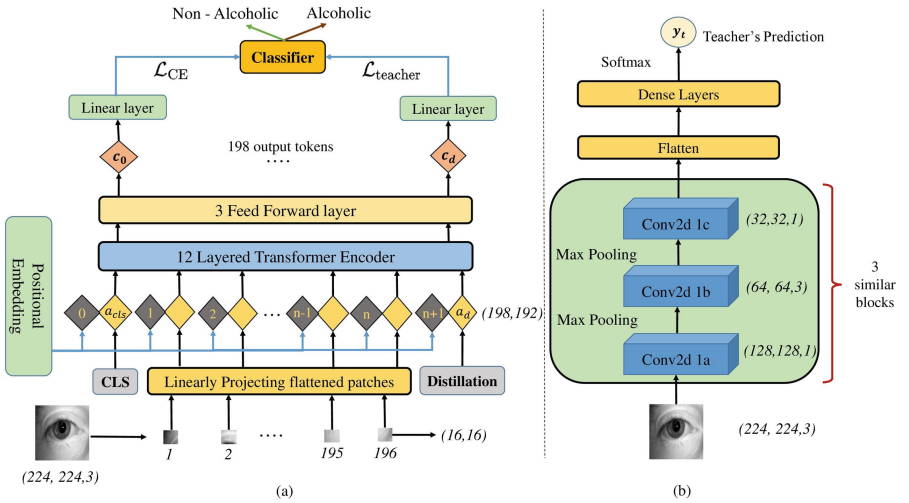

Fig. 2. (a) Generation of patch embeddings and conceptual overview of Transformer based Student model wherein CLS refers to classification token (b) Outlining Convnet based Teacher model architecture

presented a data efficient image transformer (DeiT) which suppresses the dependency of transformer on huge data. Taking inspiration from DeiT, we propose a novel architecture as depicted in Fig. 2 for detecting alcohol consumption in NIR iris images. The proposed framework comprises of mainly two parts (i) transformer based student network and (ii) convnet based teacher network. The following subsections will discuss aforementioned parts in detail.

Dataset Augmentation. In literature, classification task is mostly performed on datasets like ImageNet [9], CIFAR-100 [13], NUS-WIDE [6]. All these datasets consists of huge number of images per class. In contrary to this iris databases have limited number of images particularly in concern to post alcohol consumption images as evident form Table 1. Thus, to generate supplementary images for training our network we have used various image augmentation methods as suggested in [20]. Since, the dataset IAL-I [20] used in our study is collected in a controlled environment, nominal data augmentation methods can work well. Figure 3 depicts sample iris image with corresponding augmented images. It should be noted that image augmentation is carried out for training dataset only.

Convnet Based Teacher Network: This network takes an input iris image $I \in R^{H \times W \times C}$, where H, W and C represents image height, width and channel respectively. Assuming an image classification model f, the output of f is a label $y_t \in \{0...t\}$ where t is the number of classes. This network consists of 9 convolutional layers, with 3 layers in each block as depicted in Fig. 2. Each block

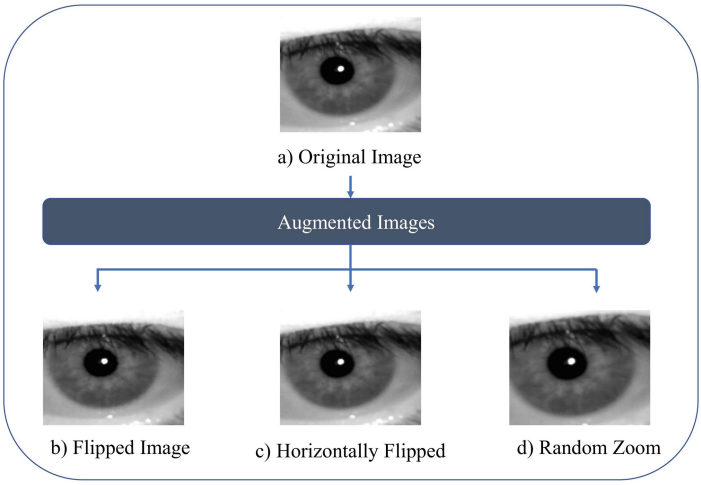

Fig. 3. Sample iris image with corresponding augmented images

is summarized as:

$$Input \xrightarrow{1\times1\,Conv} RM \rightarrow ReLU \rightarrow IFM \xrightarrow[MP]{3\times3\,Conv,} RM \rightarrow ReLU \rightarrow IFM \xrightarrow[MP]{1\times1\,Conv} RM \rightarrow ReLU \rightarrow FM \quad (1)$$

Here in Eq. 1, RM,IFM, FM and MP stands for response map, intermediate feature map, feature map and max pooling respectively. The presented architecture was inspired from Regnet based model [15].

Transformer Based Student Network: This network takes input in the form of patches. The fixed size input iris image $I \in R^{H\times W\times C}$, (where H, W and C represents image height, width and channel respectively) is decomposed into 196 patches of size 16×16. These patches are linearly projected into 196 tokens as depicted in Fig. 2(a). Each token has a shape of $(1, D)$ where D is 192 for our case. Two additional tokens, namely the classification token (CLS) and the distillation token of same shape as $(1, 192)$, are added to the patch tokens. During training, the CLS token is a vector that can be trained and contains class embeddings. The distillation token is similar to the CLS token in that it is also trainable, but it is randomly initialized and located in a fixed last position. The main objective of the distillation token is to allow our proposed architecture to learn from the output of the teacher network while remaining equivalent to the class embedding [21]. All 198 tokens, including the CLS (a_{cls}) and distillation token (a_d), are assigned positional embeddings to incorporate spatial information. Further, these tokens are given as an input to a 12 layered transformer encoder with three Multi-Self Attention (MSA) heads as depicted in Fig. 2(a). The sequence of tokens input to the encoder is as follows:

$$\mathbf{a} = [a_{cls}, \mathbf{J}a_1, \mathbf{J}a_2, ..., \mathbf{J}a_n, a_d] + E \tag{2}$$

where n is 196, $\mathbf{J}$ depicts the patch embeddings of periocular NIR images and E refers to the positional encoding that maintains the images' spatial structure.

The encoder block employs self-attention (SA) to capture the correlations among the input tokens, utilizing three types of embeddings: Query (Q), Key (K), and Value (V). To apprehend this association, the Queries, Q, are multiplied by the transpose of Keys, K^T, to generate a vector output. This vector is then divided by the square root of the dimension D to prevent the gradient from vanishing. The final matrix undergoes a Softmax activation layer multiplied by the Values V to attain the resulting Head (H), also represented as $attention(Q, K, V)$.

$$H = attention(Q, K, V) = Softmax\,(\frac{Q \times K^T}{\sqrt{D}}) \times V \tag{3}$$

In the present work, the Scaled Dot-Product Attention mechanism is employed three times to attain a total of three attention heads ($H = 3$). After the self-attention operation is performed, the outputs from all attention heads are concatenated, and then they are passed through a feed-forward (FF) neural network, which includes learnable weights ($W_{learnable}$), as represented in the Eq. 4.

$$MSA = concat(SA_1, SA_2, SA_3) \times W_{learnable} \tag{4}$$

The resultant vector is then layer normalized and passed on to the final component of the encoder which is Multi-Layer Perceptron (MLP) blocks. These blocks comprise of fully coupled FF-dense layers with GeLU non-linearity. At the end of the encoder, the retrieved output tokens are again to be fed through 3 additional FF layers to obtain a context vector Z. Z comprises of 198 output tokens similar to that fed at the beginning of the encoder. The final context vector Z can be seen in Eq. 5.

$$\mathbf{Z} = [c_0, \mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N, c_d] \tag{5}$$

After collecting the context vector Z, just the CLS token,c_0, and distillation token,c_d, are required for classification, which is then passed to 2 separate linear layers. c_0 and c_d tokens each have specific objective functions to learn, named student loss ($\mathcal{L}_{\text{CE}}$) and distillation loss ($\mathcal{L}_{\text{teacher}}$), to be discussed in the subsequent section. The average prediction after implication of a softmax activation function on both linear layers is used to determine whether or not the subject is fit.

Network Training Strategy: This subsection explores the various strategies utilized for training the teacher-student synergetic model. The input image is pre-processed to a shape of $(224, 224, 3)$ for feeding to both the teacher and

student models. Adam optimizer is used for training with a batch size of 128. Adam is an adaptive optimization algorithm used in training machine learning models that combines the benefits of adaptive learning rates and momentum for efficient convergence. The performance of the model is evaluated using cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$). A learning rate scheduler is employed with an initial learning rate of 0.001 to obtain the local minima.

At first, we train the convnet-based teacher model to obtain the final teacher predictions, y_t. These predictions are then treated as the true label while training the distillation token in the student model. Further on, the transformer-based student model is trained where the CLS token has a separate loss given as $\mathcal{L}_{\mathrm{CE}}\left(\psi\left(Z_s\right), y\right)$. Here, ψ represents the softmax function applied over logits of the student Z_s utilizing c_0 token. Similarly, the c_d token is trained by considering the teacher's prediction as true label. The objective funtion of c_d can be depicted as $\mathcal{L}_{\mathrm{CE}}\left(\psi\left(Z_s\right), y_{\mathrm{t}}\right)$. Herein, we also introduce a distillation of 0.5 on teacher model prediction. A distillation of 0.5 implies that during training, the c_d token is trained on a combination of soft targets (y_t) from the teacher model and hard targets, which is ground truth labels (y), with each target type accounting for half of the training examples. The complete process mentioned aims to replicate the teacher's predicted labels to reduce the cross-entropy loss between the highest value of the softmax function of the teacher's labels and the softmax function of the student. The final cross-entropy loss can be formulated as follows:

$$\mathcal{L}^{\mathrm{hardDistill}} = \frac{1}{2}\mathcal{L}_{\mathrm{CE}}\left(\psi\left(Z_s\right), y\right) + \frac{1}{2}\mathcal{L}_{\mathrm{CE}}\left(\psi\left(Z_s\right), y_{\mathrm{t}}\right) \tag{6}$$

4 Experimental Analysis

In this section at first, we discuss the training testing protocol and then the experimental setup and result analysis. In our experimentation, we have randomly chosen 24 subjects (70%) for training and remaining 6 for testing (30%). This also enables fair comparison with the state-of-the-art approach [20] working on the same dataset. For training our feature extractor we have used five trials of random selection of training-testing dataset. The subsequent section will discuss our experimental setup and results.

4.1 Experimental Setup

The proposed feature extractor based on teacher-student collaborative distillation knowledge is implemented in Python 3.10 using Pytorch [16] and

Table 2. Calculated Evaluation Metrics for binary classification

	Precision	Recall	F1 score
Pre Alcohol	0.97	0.99	0.98
Post Alcohol	0.99	0.97	0.98

OpenCV [4] libraries. For training and evaluating the proposed framework, a PC having Intel(R) Xeon(R) CPU @ 2.00 GHz processor with 32 GB RAM and NVIDIA GPU P100 accelarator has been used.

4.2 Experimental Results

To validate the effectiveness of the proposed framework we have conducted two sets of experimentation. In first set of experimentation our goal is to identify whether the input iris image is captured in pre-alcohol session or post alcohol session. This set of experimentation can be regarded as a binary class classification. In the second set of experimentation our goal is to study the effect of alcohol on iris after alcohol consumption at different time intervals 15, 30, 45, and 60 min respectively. This set of experimentation can be regarded as a 5 class classification problem. For analyzing the performance we have employed commonly used classification task measures such as precision, recall and F1 score.

- **Binary Class Classification:** It can be inferred from Table 1 that in comparison of post-alcohol images (2400) we have very few instances of pre-alcohol images (600) in IAL-I dataset. In order to compensate this we have used randomly selected 800 images from CASIA-V4 [1] dataset for training our network. It should be noted that testing results are reported for IAL-I dataset only. Table 2 illustrates the binary class classification results in terms of precision, recall and F1 score.
- **Five Class Classification:** Under this experimentation we are trying to study the behavioral changes of the eye's CNS after alcohol consumption at 0, 15, 30, 45, and 60 min, respectively. Upon testing the model, our model showed an accuracy of 96.86% for 0th minute, 90.76% for 15th minute, 92.57% for 30th minute, 93.38% for 45th minute and 91.06% for 60th minute. The overall accuracy observed during testing phase came out to be 92.94%. Inferring from the obtained results for behaviour analysis, we can assert that the affect of consuming alcohol is most prominent at 45 min. Table 3 illustrates the five class classification results in terms of precision, recall and F1 score.

Comparative Analysis: To validate the effectiveness of the proposed framework we have compared our results with state-of-the-art approach [5,20]. To the best of our knowledge, [5,20] are the only work that has been conducted on IAL-I dataset used in our study. Table 4 provides a detailed comparison between the proposed approach and state-of-the art method. Essentially, the suggested approach achieves higher levels of accuracy in inference when compared to previous system that have been documented in the literature.

Table 3. Calculated Evaluation Metrics for five class classification

Session	Precision	Recall	F1 score
Session 0 (0 min)	0.95	0.96	0.96
Session 1 (15 min)	0.91	0.93	0.92
Session 2 (30 min)	0.93	0.93	0.93
Session 3 (45 min)	0.94	0.94	0.94
Session 4 (60 min)	0.93	0.90	0.91

Table 4. Comparison of the proposed method with the state-of-the-art approaches

Algorithm	Accuracy
Multi-Layer-Perceptron (MLP) [5]	75.8%
Fused Capsule Network [20]	92.3%
Our Proposed Framework	**98.46**%

5 Conclusion and Future Works

In this work we propose a framework that utilizes teacher-student learning paradigm for detecting alcohol consumption in NIR iris images. While we demonstrated the effectiveness of using transfer learning archetype in case encountered with small datasets (pre-alcohol iris images in our case). Furthermore, we provide detail experimental analysis to establish relation between alcohol consumption and the time elapsed after taking alcohol. Through various experiments, it can be inferred that the proposed framework outperforms the baseline state-of-the-art approach. The present work determines the fitness for duty (FFD) only on the basis of analyzing NIR iris images under the influence of alcohol, in future we would like to study the effect of drugs, lack of sleep on iris. Furthermore, we would like to deploy our proposed approach on an edge device for real-time inference.

References

1. Casia iris image database version 4.0. https://biometrics.idealtest.org/
2. Amodio, A., Ermidoro, M., Maggi, D., Formentin, S., Savaresi, S.M.: Automatic detection of driver impairment based on pupillary light reflex. IEEE Trans. Intell. Transp. Syst. **20**(8), 3038–3048 (2019). https://doi.org/10.1109/TITS.2018.2871262
3. Arora, S.S., Vatsa, M., Singh, R., Jain, A.: Iris recognition under alcohol influence: a preliminary study. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 336–341 (2012). https://doi.org/10.1109/ICB.2012.6199829
4. Bradski, G.: The OpenCV library. Dr. Dobb's J. Softw. Tools **25**, 120–123 (2000)
5. Causa, L., Tapia, J.E., Lopez-Droguett, E., Valenzuela, A., Benalcazar, D., Busch, C.: Behavioural curves analysis using near-infrared-iris image sequences (2022)

6. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: NUS-WIDE: a real-world web image database from national university of Singapore. In: Proceedings of ACM Conference on Image and Video Retrieval (CIVR 2009), Santorini, Greece (2009)
7. Dai, Y., Gao, Y., Liu, F.: TransMed: transformers advance multi-modal medical image classification. Diagnostics **11**(8) (2021). https://doi.org/10.3390/diagnostics11081384, https://www.mdpi.com/2075-4418/11/8/1384
8. Delgado, M.K., et al.: Accuracy of consumer-marketed smartphone-paired alcohol breath testing devices: a laboratory validation study. Alcohol.: Clin. Exp. Res. **45**(5), 1091–1099 (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
10. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale (2021)
11. Gug, I.T., Tertis, M., Hosu, O., Cristea, C.: Salivary biomarkers detection: analytical and immunological methods overview. TrAC, Trends Anal. Chem. **113**, 301–316 (2019)
12. Health, Q.: Fitness for duty: alcohol and other drugs (2020)
13. Krizhevsky, A., Nair, V., Hinton, G.: CIFAR-100 (Canadian institute for advanced research). https://www.cs.toronto.edu/~kriz/cifar.html
14. Le, N.Q.K., Ho, Q.T., Nguyen, T.T.D., Ou, Y.Y.: A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. Brief. Bioinform. **22**(5) (02 2021). https://doi.org/10.1093/bib/bbab005
15. Mahbub, M.K., Biswas, M., Miah, A.M., Shahabaz, A., Kaiser, M.S.: COVID-19 detection using chest X-ray images with a RegNet structured deep learning model. In: Mahmud, M., Kaiser, M.S., Kasabov, N., Iftekharuddin, K., Zhong, N. (eds.) AII 2021. CCIS, vol. 1435, pp. 358–370. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82269-9_28
16. Paszke, A.e.a.: PyTorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019). https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
17. Pidd, K., Roche, A.M., Cameron, J., Lee, N.K., Jenner, L., Duraisingam, V.: Workplace alcohol harm reduction intervention in Australia: cluster non-randomised controlled trial. Drug Alcohol Rev. **37**, 502–513 (2018)
18. Reich, J., Kelly, M.: Empirical findings of fitness-for-duty evaluations. MedEdPublish **7**, 258 (2018). https://doi.org/10.15694/mep.2018.0000258.1
19. Tapia, J.: NIR iris images under alcohol effect (2022). https://doi.org/10.21227/dzrd-p479, https://dx.doi.org/10.21227/dzrd-p479
20. Tapia, J., Droguett, E.L., Busch, C.: Alcohol consumption detection from periocular NIR images using capsule network. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 959–966 (2022). https://doi.org/10.1109/ICPR56361.2022.9956573
21. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021)
22. Yuan, F., Zhang, Z., Fang, Z.: An effective cnn and transformer complementary network for medical image segmentation. Pattern Recognit. **136**, 109228 (2023). https://doi.org/10.1016/j.patcog.2022.109228, https://www.sciencedirect.com/science/article/pii/S0031320322007075