Lightweight ML-Based Method for Mental Fatigue Assessment Using Human Facial Video

Batol Hamoud, Walaa Othman, Nikolay Shilov and Alexey Kashevnik
SPC RAS
Saint-Petersburg, Russia
{bkhamud, walaa_othman}@itmo.ru, {nick, alexey.kashevnik}@iias.spb.su

***Abstract*—Mental fatigue is a growing concern in various domains. In this study, we propose a lightweight, non-intrusive approach for estimating mental fatigue using facial videos captured via standard webcams. The method extracts red, green, and blue (RGB) channel signals from the facial region using a pixel averaging technique. These raw signals are filtered and transformed into a set of statistical features, which are then fed into traditional machine learning models. The proposed system achieves promising classification performance, with models like MLPClassifier and KNN reaching accuracy scores of up to 81% following hyperparameter tuning and cross-validation. On the contrary, deep networks such as LSTM and 1D-CNN applied to the filtered raw RGB signals yielded lower accuracy (around 54–57%), most likely due to the dataset size being small and the raw signal variability. With just an small decrease of the accuracy compared to some literature benchmarking, our method possesses numerous merits concerning speed, ease of retraining, and deployment on low-resource devices. This makes it particularly suitable for scalable fatigue monitoring in realistic settings. Future improvements may include extending the feature set and enlarging the dataset to further enhance performance.**

I. Introduction

In recent years, technical systems have become more advanced and complicated. As they have become increasingly complex, enormous cognitive burdens are being placed on human operators who need to monitor a wide array of performance parameters continuously and make fast, good judgments in order to keep the system operating optimally. Thus, the risk of mental fatigue has become more important. Notably, in these high-responsibility environments, fatigue is not merely an individual issue of well-being. It has serious operational consequences, as even small mistakes in judgment can have costly or risky consequences [1], [2].

One of the most challenging aspects of fatigue that it is hard to be recognized by individuals until their performance significantly decreases. When there is fatigue, decision-making may take much longer and reactions may be slower, in effect reducing professional performance and causing a higher probability of operational errors [3].

There have been many approaches implemented to address the problem of fatigue detection, primarily aimed at the identification of early warning signs of fatigue and giving real-time notifications to individuals of its possible risks. The majority of the techniques follow machine learning-based algorithms or deep learning-based models to develop robust and effective systems of fatigue detection [4]. Much of the existing work has been focused on identifying fatigue in task-oriented, cognitively demanding situations, such as driving [5]. However, there have been some investigations on identifying fatigue in more passive environments such as natural-viewing scenarios where users are not directly involved in activities [6]. These approaches have made use of non-intrusive methods, such as remote or webcam-based eye tracking. These systems generally measure changes in pupil dynamics, blink rate, and movement patterns of eyes to infer levels of fatigue in both active and passive states of cognition [7].

Despite significant progress in fatigue detection research, there remain a number of crucial limitations that are not met. Though most studies exhibit high accuracy and robustness, most sophisticated deep learning methods demand very high computational requirements and large numbers of annotated samples, which are not always practical in real-world scenarios. Additionally, most use of subjective self-reporting measures (e.g., VAS-F and NASA-TLX) as annotation fatigue is challenging to achieve high reliabilty. Sensor-based approaches, while informative, often need physical touch or wearable sensors, which may impact usability and scalability for prolonged or large-scale monitoring. Our main contributions in this article are as follows:

1) We propose a lightweight, non-intrusive method for estimating mental fatigue using facial videos captured via a standard webcam, avoiding the need for wearable sensors or specialized hardware.
2) We demonstrate that traditional machine learning models, when combined with statistical features and optimized via grid search and cross-validation, can achieve competitive accuracy with low computational cost.
3) Our method offers practical advantages in terms of model retrainability, and fast inference, making it suitable for near real-time fatigue monitoring in resource-constrained environments.
4) This study contributes to the growing body of work on fatigue detection by addressing the trade-off between accuracy, flexibility, and user acceptability, and by providing a scalable solution for long-term monitoring.

The paper is structured into four main sections. Section II reviews a range of existing approaches to mental feature detection. Section III introduces our proposed framework in details where section IV provides an overview of the dataset

utilized in this study. Section V details the experimental setup and presents the results obtained from applying our approach. Finally, Section VI offers a comprehensive discussion and interpretation of the findings.

II. Related Work

This section outlines the proposed approaches and methodologies developed for the assessment of mental fatigue in drivers and operators. Accordingly, the methods presented here are designed to detect cognitive fatigue in real-time or near real-time conditions, leveraging both contactless estimation techniques and sensor-based approaches. By doing so, we aim to address the practical limitations and provide scalable, generalizable solutions that can be integrated into real-world monitoring systems.

The study in [8] proposes a cost-effective framework for early-stage driver fatigue detection, designed to identify fatigue before it becomes critical. An infrared (IR) camera monitors behavioral indicators related to the eyes, mouth, and head, ensuring reliable performance under varying lighting conditions. Feature extraction is performed using multiple CNN architectures, while classification is carried out with a logical inference model. Evaluation on real-world datasets collected in both daytime and nighttime conditions showed a fatigue prediction accuracy of 93.3%, demonstrating strong potential for practical deployment.

In [9],fatigue states are recognized through a combination of facial analysis and temporal modeling. Face detection is performed with a Multitask Convolutional Neural Network MTCNN, followed by extraction of key landmarks using the DLIB library. These landmarks generate fatigue-related feature vectors from individual frames, which are concatenated into sequences and processed by an LSTM to capture temporal dynamics. The method achieved 88% accuracy on the YawDD dataset [10] and 90% on a custom-built dataset, demonstrating robust performance across sources.

Authors of [11] employ an RGB-D camera to capture both RGB and infrared facial video for driver fatigue analysis. The data are processed independently using the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [12], Fourier transform, and the Triangular Surface Patch (TSP) descriptor [13], enabling extraction of heart rate, eye openness, and mouth openness levels. These features are fused with a Multimodal Fusion Recurrent Neural Network (MFRNN), that integrates temporal dynamics via an RNN layer, while fuzzy reasoning is applied to the heart rate signal to handle noise. The system achieved accuracies of 86.75% and 91.67% on two datasets.

In [14], a real-time framework for detecting driver disturbances is presented using Convolutional Neural Networks (CNNs), including InceptionV3, VGG16, and ResNet50. Among them, ResNet50 achieved the best results with 93.69% accuracy and a loss of 0.6931.

The study in [15] introduces a non-invasive smart cushion system for assessing mental fatigue in construction equipment operators. Heart rate and respiration signals were continuously monitored during simulated excavator tasks with twelve participants. A range of time and frequency-domain features was extracted and used to train a Random Forest classifier. The aim was to investigate the relationship between physiological indicators and self-reported mental fatigue, as measured by the NASA-TLX workload assessment. The model achieved 92% accuracy, with results showing that combining heart rate and respiration features outperformed either signal alone.

In [16], fatigue is estimated from multimodal wearable sensor data including heart rate variability, respiratory rate, energy expenditure, activity counts, and step count. Missing values were imputed before training recurrent neural networks. Both supervised and unsupervised approaches were tested, with the best performance achieved using a hybrid of a random forest and a causal CNN model, yielding 70% precision and 73% recall. The study highlights that physiological signals predicted mental fatigue most effectively, while combining physiological and activity-related features was crucial for physical fatigue estimation.

Another significant advancement in driver fatigue detection is presented through a novel multimodal neural network architecture [17]. Leveraging the DROZY dataset containing physiological (EEG, ECG) and facial image data, the study's standout contribution is a "multimodal feature coupled model." This model innovates by not simply combining data streams but by having features from each modality (e.g., EEG, ECG, facial) dynamically weight and influence each other. This sophisticated coupling mechanism proved highly effective, achieving exceptional performance metrics (Accuracy: 98.41%, F1-Score: 98.38%), significantly outperforming a standard feature combination model (95% across metrics).

Article [18] assesses fatigue from video data of working individuals using deep-learning-based feature extraction for head movement (Euler angles), vital signs (e.g., heart rate, blood pressure, oxygen saturation, respiratory rate), and indicators of eye and mouth activity (blinking and yawning). The method was validated using the "Human Fatigue Assessment Based on Video Data" (HFAVD) dataset. Random Forest models consistently achieved F1 scores and accuracies above 90%. Based on this approach, [19] applied feature importance techniques to identify the most influential predictors to avoid the high computational cost presented in [18]. This study highlighted key contributors to mental fatigue (heart rate, blood pressure, oxygen saturation and the pitch angle of the head) and replaced traditional machine learning methods, such as Random Forest with sophisticated architecture of Tabular Transformer which offers improved generalization capabilities and is well-suited for handling structured data. This transition led to a notable increase in the approach efficiency, achieving an accuracy of 89%.

In [20], a facial feature-based drowsiness detection system was proposed. Face detection was performed using Histogram of Oriented Gradients (HoG) features with a Support Vector Machine (SVM) classifier, followed by landmark extraction to estimate head pose and eye blinks. Drowsiness indicators such as blink duration, frequency, and PERCLOS were integrated

with a fuzzy inference system. Validation against the Karolinska Sleepiness Scale (KSS) [21] showed an average NRMSE of 15.63%.

A notable contribution to the domain of mental fatigue assessment is presented in this study [22] that integrates the BlazeFace algorithm with a generalized regression neural network (GRNN) optimized via a genetic algorithm (GA). Video sequences undergo preprocessing via homomorphic filtering, followed by BlazeFace detection for accurate landmark localization. Fatigue-relevant features (PERCLOS, fixation duration, pupil area) are weighted based on expert evaluation and fused into a composite metric. A GRNN optimized with genetic algorithms is then applied, achieving 97% accuracy and demonstrating value for aviation safety applications.

Another study [23] proposed a novel approach that leverages computer vision and machine learning techniques, offering a contactless alternative to traditional sensor-based methods. Unlike conventional approaches relying on physiological signals such as electrocardiograms, which require continuous physical contact, this method utilizes facial data captured through computer vision to assess fatigue states. The proposed system integrates a Deep Residual Network with a Random Forest classifier (DRN-RF), combining the representational power of deep learning with the interpretability and stability of ensemble methods. DRN-RF framework demonstrated superior performance, achieving an accuracy of 94%.

Article [24] explores the assessment of mental fatigue within the context of sports performance by introducing a deep learning-based framework, thereby departing from the conventional reliance on heart rate variability (HRV) analysis typically seen in earlier works. The proposed method employs a hybrid neural architecture that combines Residual Networks (ResNet) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for feature extraction, alongside a transformer module for advanced feature fusion. By utilizing original ECG signals, two-dimensional spectral representations, and other physiological indicators, the model achieves a high classification accuracy of 95.29%.

Despite numerous reported studies in the area of fatigue detection, which provide promising performance and excellent results. Some of the constraints are still there that require attention. Deep learning techniques—namely convolutional and recurrent neural networks—have been extremely effective in abstracting complex patterns of mental fatigue. However, these techniques typically have high computational cost and need the availability of large annotated datasets, which are not always feasible for deployment. Furthermore, the majority of fatigue labeling methods in the literature rely on subjective self-report scales (e.g., NASA-TLX) or physiological sensors. While informative, these measures have some issues. Sensor-based measurement generally requires physical contact or wearable sensors that can be inconvenient, and limit scalability in large-scale or long-duration monitoring contexts. Overall, although recent years have witnessed impressive technical progress, there remain requirements for techniques balanced between accuracy, flexibility, and user acceptability, particularly those that can give solid fatigue predictions in non-contact, non-intrusive manners under different environmental and user conditions.

III. Methodology

This section presents a comprehensive overview of the methodology adopted in this study. It includes the data preprocessing steps, the machine learning and deep learning models explored, and a detailed description of the dataset utilized to implement and evaluate the proposed approach.

This study builds on the concept of extracting meaningful signals from facial videos of computer operators using a pixel-averaging technique. These raw signals are then preprocessed using filtering to enhance signal quality before being passed into machine learning or deep learning models for fatigue estimation.

We examined two main strategies for handling the signals obtained from the red, green, and blue (RGB) channels of the video frames. The first strategy involves summarizing each one-minute segment of each signal by extracting statistical features, specifically, the mean, minimum, maximum, standard deviation, median, and the 25th and 75th percentiles. This results in a set of 21 features (7 statistics per RGB channel), which serve as input to a prediction model.

The second strategy treats the raw signals as time series data. Here, the signals are divided into one-minute segments containing 60 consecutive values, which are then directly fed into a detection model to predict fatigue levels. This time-series-based approach is applied separately to the signals extracted from each RGB channel to evaluate their individual contributions.

Therefore, the proposed methodology shown in Figure 1 follows a structured three-stage pipeline: signal extraction from facial videos, preprocessing of the raw signals, and finally, the evaluation of various machine learning and deep learning models. Each stage is designed to refine and prepare the data for the next, ultimately aiming to predict fatigue levels accurately.

1) Signals extraction stage: This stage aims to extract signals from RGB video data by tracking the color variations in the facial region over time. The process involves identifying the face in each video frame, extracting the average RGB values, which will be followed with signal processing techniques to produce smoothed signals. The process begins with loading the video file using the OpenCV library, where frames are sampled at a rate of one frame per second to minimize computational time. Each selected frame is converted from OpenCV's default BGR format to the RGB color space.

Afterwards, we use the face_recognition library to locate faces in the frame. When a face is detected, the region corresponding to the bounding box of the first detected face is extracted. If no face is found, the entire frame is used as a fallback. This ensures the signal is extracted from the most relevant part of the frame — the facial skin region, which is known to exhibit subtle color changes due to blood flow.

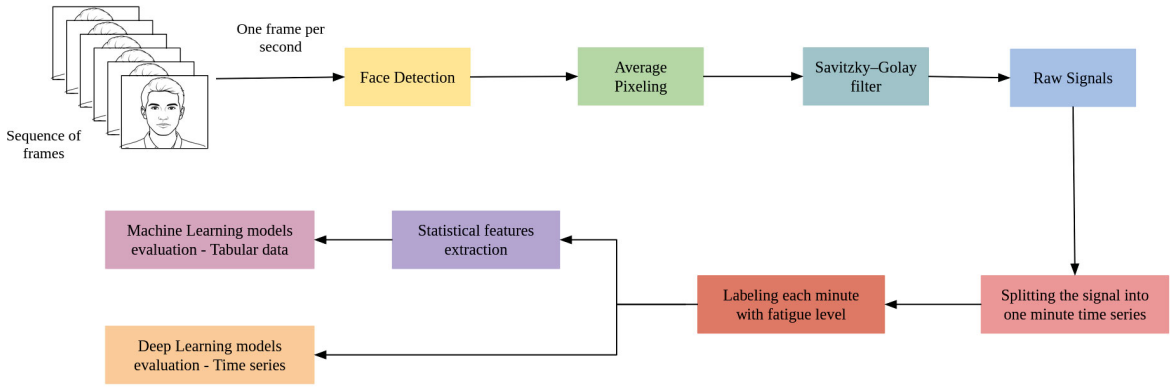

Fig. 1. The proposed methodology

From each identified facial region, the mean pixel intensity values of the R, G, and B channels are computed. This results in three synchronized time series representing raw color variation signals. These signals are subsequently subjected to signal processing stage to enhance their quality and prepare them for subsequent modeling stages.

2) Signals pre-processing stage: Among the primary issues in extracting signals from video is the presence of noise due to ambient lighting changes, camera sensor aberrations, video compression, and facial movement. These sources of noise can mask subtle temporal color variations in skin, which are indicative of physiological processes. To mitigate these effects without compromising the informative value of the signal, we employ the Savitzky–Golay filter [25], a widely used digital filter for signal smoothing.

The Savitzky–Golay filter (SG filter) is a convolution-based smoothing algorithm that operates by fitting a low-degree polynomial to a local window of consecutive data points using least squares regression, and evaluating the polynomial at the central point of the window.

Formally, for each point x_i in the signal, a polynomial of degree p is fitted to the values in a window centered at x_i, with a total of w data points (where w is the window length). The central point is then replaced with the value of the fitted polynomial at x_i, which acts as the smoothed estimate. Here is the equation:

$$\hat{y}_i = \sum_{j=-m}^{m} c_j y_{i+j} \tag{1}$$

where:

- $\hat{y}_i$ is the smoothed (filtered) value at point x_i,
- y_{i+j} are the original signal values in the window around x_i,
- c_j are the convolution coefficients derived by fitting a polynomial of degree p via least squares to the window,
- m is the half-window size, so that the window contains $2m + 1$ points in total.

These coefficients c_j depend only on the polynomial degree p and the window length w, and are the same for all x_i when the window is uniformly spaced.

This process is repeated across the entire time series, effectively reducing high-frequency noise while retaining the overall shape, amplitude, and phase characteristics of the signal, which are properties that traditional low-pass filters often distort.

The SG filter is particularly well-suited for processing RGB signals extracted from skin regions in video due to the following properties:

1) **Preservation of Temporal Structure:** Unlike moving average filters or aggressive low-pass filters that can distort signal peaks and delay phase information, the SG filter maintains the shape of periodic components that are crucial for accurate estimation of physiological metrics.
2) **Noise Attenuation Without Signal Suppression:** High-frequency variations caused by compression artifacts, camera noise, or flickering illumination are effectively smoothed out, while slow, physiologically relevant color modulations are preserved.

Following the filtering process, we explored two preprocessing strategies to prepare the RGB signals for fatigue prediction. In the first approach, each one-minute segment of the signal was summarized using 21 statistical features (7 statistics per channel). In the second approach, we treated the filtered RGB signals as time series. Each one-minute window, consisting of 60 consecutive samples, was used directly as input to the model to preserve temporal dynamics. Both strategies were evaluated to assess the predictive value of RGB channel information under different signal representations.

3) Models Evaluation: We explored multiple machine learning and deep learning approaches to get the best per-

formance. Here is an overview of the models evaluated:

1) Random Forest: Random Forest is an ensemble method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. It is well-suited for tabular data with engineered features (e.g., statistical summaries of signals) but not ideal for raw time series, as it doesn't capture temporal dependencies.
2) XGBoost Classifier: XGBoost is a high-performance gradient boosting algorithm that builds trees sequentially to correct previous errors. It includes advanced regularization and parallelism. XGBoost excels with structured, tabular data, but does not handle temporal sequences natively unless time-based features are engineered.
3) LightGBM Classifier: LightGBM is a fast, efficient gradient boosting algorithm that grows trees leaf-wise. It is highly scalable and accurate on tabular datasets, especially those with many features. Like other tree-based models, it does not model time dependencies unless temporal information is manually encoded.
4) CatBoost Classifier: CatBoost is a gradient boosting algorithm that handles categorical variables naturally and avoids overfitting. It's effective for tabular data and robust with minimal preprocessing. However, it lacks native support for time-dependent patterns.
5) Multilayer Perceptron: MLP is a fully connected neural network that can model non-linear relationships in tabular data. While it can also be used for time series, it does not inherently model temporal dependencies unless combined with specific architectures or time-aware feature engineering.
6) K-Nearest Neighbors: KNN classifies the features based on the majority vote of nearest neighbors. It's simple and effective for low-dimensional tabular data, but it struggles with high-dimensional inputs and raw time series, especially without proper feature scaling and dimensionality reduction.
7) Support Vector Classifier: SVC finds optimal hyperplanes for classification, with support for non-linear kernels. It performs well on tabular data, especially when classes are separable, but is computationally intensive for large datasets and not optimized for sequential data.
8) Long Short-Term Memory Network: LSTM networks are designed to learn long-term dependencies in sequential data through memory gates. They are ideal for time series, especially where trends, delays, or repeated patterns exist. They are not typically used for static tabular data.
9) 1D Convolutional Neural Network: 1D-CNNs learn local temporal patterns using convolutional filters. They are efficient and effective for time series classification, particularly when the signal has short- to mid-range dependencies. While not typical for tabular data, they can be adapted if the input has some spatial/temporal structure.

We chose these models based on the need to balance model complexity with practical constraints on computational resources, especially given the volume of data derived from video-based signal extraction. Tree-based and distance-based classifiers were prioritized for their low training time and minimal hyperparameter tuning requirements when applied to engineered statistical features. Similarly, the use of lightweight neural architectures for time series data (e.g., 1D-CNNs) was chosen over more computationally intensive alternatives to the minimize processing cost.

IV. Dataset

In the conducted experiments, we utilized the OperatorEYEVP dataset, as introduced by [26]. This dataset comprises recordings from ten distinct participants engaged in a variety of activities, captured three times daily—morning, afternoon, and evening—across a span of eight to ten days. Each session includes facial video recordings, accompanied by a rich set of additional data: eye and head movement signals, scene imagery, heart rate measurements (expressed as pulse per interval), results from two instances of a choice reaction time (CRT) task, and responses to subjective self-report measures, including the Visual Analogue Scale for Fatigue (VAS-F).

The VAS-F consists of 18 items designed to assess participants' perceived fatigue levels and is done at the beginning of each session. The overall experimental session follows a structured protocol, starting with a sleep quality questionnaire (administered once daily prior to the morning session), followed by the VAS-F, a CRT task, a scientific text reading task, the Landolt Ring correction test, a Tetris game, and a second CRT task. The inclusion of a second CRT assessment is based on the authors decision due to the fact that the operator's level of fatigue may vary between the start and the end of the recording session. On average, each session lasted approximately one hour.

The inclusion of the Landolt Rings correction test provides an objective indictor to the assessment of cognitive performance, particularly in relation to fatigue. Unlike subjective self-report measures that rely on individual perception and may be influenced by mood or personal bias like questionnaires and rating scales, the Landolt Rings test yields metrics derived from task performance. These include attention productivity, work accuracy, stability of attention concentration, mental performance coefficient and processing speed, which are known to decline under mental fatigue.

In this study, mental performance was utilized as an indicator of fatigue. A higher mental performance value reflects lower levels of fatigue, and vice versa. The mental performance values, derived from the Landolt Rings correction test, ranged from 0 to 3. To distinguish between low and high fatigue levels, we applied a threshold of 1.5 for labeling purposes. Specifically, each five-minute video of the Landolt Rings task was divided into one-minute intervals, and each minute was labeled according to the corresponding mental

performance level (fatigue level). We chose to focus exclusively on the Landolt Rings videos to ensure the reliability and objectivity of the fatigue labels, given the task's strong association with cognitive performance under fatigue.

V. Experiments and Results

As stated above, the dataset includes recordings from 10 subjects. In total, 211 videos with a duration of five minutes each were used, which resulted in 1,055 one-minute samples for training the models - whether they are time series or statistical feature-based. Because of the relatively small dataset size, we opted for small deep learning models and did not employ more advanced models such as Transformers, which typically require larger datasets to converge effectively and yield stable outcomes. This option is also aligned with our goal of computational efficiency.

To optimize model performance, we employed grid search combined with cross-validation on the training set to identify the most effective hyperparameter configurations of the traditional machine learning models. This approach ensures that the selected parameters generalize well and are not overfit to a specific subset of the data. The dataset, consisting of 1,055 one-minute samples, was split such that 20% was reserved as a hold-out test set to evaluate the final model performance. This separation allowed for an unbiased assessment of the models' predictive capabilities on unseen data.

Machine learning algorithms are well-adapted to grid-based hyperparameter optimization because they have relatively short training times and a small number of parameters to be optimized. Table I includes a list of the hyperparameters explored in this study.

TABLE I. Hyperparameters used for each model

Model	Hyperparameters
Random Forest	`n_estimators`: [100, 200] `max_depth`: [None, 10] `min_samples_split`: [2, 5]
XGBoost	`n_estimators`: [100, 200] `max_depth`: [3, 5] `learning_rate`: [0.05, 0.1]
LGBMClassifier	`n_estimators`: [100, 200] `max_depth`: [3, 5, -1] `learning_rate`: [0.05, 0.1]
CatBoostClassifier	`n_estimators`: [100, 200] `max_depth`: [3, 5] `learning_rate`: [0.05, 0.1]
MLPClassifier	`hidden_layer_sizes`: [(64, 32), (128, 64)] `activation`: ['relu', 'tanh'] `alpha`: [0.0001, 0.001]
KNN	`n_neighbors`: [1, 3, 5, 7] `weights`: ['uniform', 'distance'] `p`: [1, 2]
SVC	`C`: [0.1, 1, 10] `kernel`: ['linear', 'rbf'] `gamma`: ['scale', 'auto']

Deep learning algorithms such as LSTM and 1D-CNN, conversely, usually require additional optimization techniques (e.g., learning rate schedules, early stopping) and are sensitive to initialization and data variability, making traditional grid search relatively less efficient or effective. Therefore, within these architectures, we employed default parameter settings taken from exploratory experiments with the computational feasibility as priority. Tables II and III show the architectures of the LSTM and 1D-CNN, respectively.

TABLE II. LSTM Model Architecture

Layer (type)	Output Shape	Param #
`lstm (LSTM)`	(None, 60, 64)	16,896
`lstm_1 (LSTM)`	(None, 32)	12,416
`dense (Dense)`	(None, 32)	1,056
`dropout (Dropout)`	(None, 32)	0
`dense_1 (Dense)`	(None, 1)	33

TABLE III. 1D-CNN Model Architecture

Layer (type)	Output Shape	Param #
`conv1d (Conv1D)`	(None, 58, 32)	128
`max_pooling1d (MaxPooling1D)`	(None, 29, 32)	0
`conv1d_1 (Conv1D)`	(None, 27, 64)	6,208
`max_pooling1d_1 (MaxPooling1D)`	(None, 13, 64)	0
`flatten (Flatten)`	(None, 832)	0
`dense_2 (Dense)`	(None, 64)	53,312
`dropout_1 (Dropout)`	(None, 64)	0
`dense_3 (Dense)`	(None, 1)	65

Tables IV and V present the accuracy of various models trained on two different types of input data: statistical features (table IV) and raw color channel signals extracted from video frames (table V).

In the first table, traditional machine learning models trained on statistical features consistently achieved higher accuracy, with the MLPClassifier and KNeighborsClassifier reaching up to 81%, and ensemble methods such as Random Forest, LGBMClassifier, and CatBoostClassifier achieving accuracy scores between 77% and 78%. These results indicate that statistical summarization of the signals provides highly informative features for fatigue classification.

TABLE IV. Model performance using statistical features

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.77	0.78	0.77	0.77
XGBClassifier	0.74	0.75	0.74	0.75
LGBMClassifier	0.78	0.78	0.78	0.78
CatBoostClassifier	0.78	0.79	0.78	0.78
MLPClassifier	0.80	0.80	0.80	0.80
KNeighborsClassifier	0.81	0.80	0.80	0.80
SVCClassifier	0.66	0.69	0.64	0.63

Table V shows the performance of deep learning models (LSTM and 1D-CNN) trained on raw filtered red, green, and blue channels. These models presented lower accuracies ranging from 0.54 to 0.57, where the LSTM performed slightly better on the blue channel. The low accuracy implies that the raw channel signals separately may not hold discriminative

information required to accurately predict levels of fatigue without further feature engineering or larger training data.

TABLE V. MODEL PERFORMANCE USING COLOR CHANNEL SIGNALS

Model	Red Channel	Blue Channel	Green Channel
LSTM	Accuracy: 0.54 Precision: 0.27 Recall: 0.50 F1-score: 0.35	Accuracy: 0.57 Precision: 0.57 Recall: 0.55 F1-score: 0.53	Accuracy: 0.54 Precision: 0.27 Recall: 0.50 F1-score: 0.35
1D-CNN	Accuracy: 0.54 Precision: 0.27 Recall: 0.50 F1-score: 0.35	Accuracy: 0.54 Precision: 0.27 Recall: 0.50 F1-score: 0.35	Accuracy: 0.54 Precision: 0.27 Recall: 0.50 F1-score: 0.35

Overall, this comparison highlights that using statistical features derived from the video signals leads to substantially better predictive performance than relying on raw RGB signals, particularly in small-data scenarios. This supports our decision to focus on statistical features with traditional ML models in this study.

VI. DISCUSSION

The performance difference between models trained on statistical features and those trained on raw RGB channel signals can be explained to several critical factors relating to signal representation, model suitability, and data constraints.

Firstly, statistical features serve as high-level representations of the underlying meaningful signals. These features effectively outline temporal dynamics and noise-prone patterns into more compact and discriminative representation. By leveraging these features, traditional machine learning model are able to efficiently capture relevant patterns associated with fatigue without being overwhelmed by the variability in raw data. This explains their relatively high and consistent accuracy levels, reaching up to 81%.

In contrast, the deep learning models (LSTM and 1D-CNN) in this study were trained on filtered raw red, green, and blue (RGB) channel signals, extracted from facial video frames. The filtering process aimed to enhance the video signal quality by reducing motion and illumination artifacts. However, these raw filtered time-series signals still include substantial residual variability and subtle temporal dynamics, which are not explicitly structured or annotated.

While deep models theoretically excel at discovering latent patterns in raw sequences, they rely heavily on large, diverse datasets to learn robust and generalizable features. In this case, the relatively small dataset likely limited the effeciency of the LSTM and CNN architectures to extract meaningful representations, resulting in consistently low accuracy scores across all RGB channels. Furthermore, the nearly constant performance across red, green, and blue channels indicates that individual signal channel cannot provid sufficiently strong discriminative performance.

The biggest limitation of this study is the relatively small dataset, which constrains generalizability. Expanding the dataset is a key next step, and we are currently planning additional data collection. Additionally, while physiology-inspired and frequency-domain features can further enhance performance, which we did in our prior work [18], [19], we experimented with such features (e.g., heart rate, blood pressure, oxygen saturation, head/eye movement indicators) and obtained promising results. However, they typically require additional sensors or computationally intensive computer vision pipelines. In the present study, our primary objective was to design a lightweight and feasible solution. For this reason, we intentionally restricted the feature set to simple statistical summaries of RGB signals, allowing deployment on standard webcams and low-resource devices without specialized hardware. This choice enabled an approach that remains practical, scalable, and effective in real-world scenarios.

Nevertheless, we see this contribution as a proof-of-concept, highlighting that even in settings where large-scale annotated data are not available, useful and efficient models can be built. However, we acknowledge that validating across diverse datasets or environments would further strengthen the generalizability of the approach. Furthermore, in this study we deliberately focused on a single dataset (OperatorEYEVP), as it provides objective fatigue labels (via the Landolt Rings test), ensuring consistency and reliability in evaluation. Expanding validation across multiple datasets is an important direction for future research, but was beyond the scope of the present study.

In summary, these results validate the hypothesis that domain-informed feature engineering remains a powerful approach, especially when data is limited. Traditional ML models trained on statistical features not only outperform deep learning models in accuracy but also offer greater interpretability and computational efficiency. This justifies the design choice in our pipeline to prioritize extracted features and conventional classifiers over end-to-end deep models trained on raw inputs.

VII. CONCLUSION

In this study, we proposed a lightweight framework for estimating mental fatigue from facial videos. The approach involves extracting red, green, and blue (RGB) channel signals using a pixel averaging technique, followed by signal filtering and statistical feature extraction. These features were then used to train traditional machine learning models, which achieved good performance when optimized through grid search and cross-validation.

We also evaluated the effectiveness of deep learning models (LSTM and 1D-CNN) applied directly to the filtered raw signals. However, these models demonstrated relatively poor performance, likely due to the limited size of the dataset and the high variability and redundancy present in the raw signal data and the fact that deep architectures typically require large-scale data to perform effectively.

Our exploration of deep learning models was intentionally limited. The first reason is the dataset size. As discussed before, the dataset is relatively small (1,055 one-minute samples). Deep models generally require much larger and more

diverse training data to converge effectively and avoid instability. Extensive tuning under these data constraints would likely not provide meaningful improvements or reliable conclusions. The other reason is that our main goal of this work was to assess whether simple, efficient models can achieve competitive performance in webcam-based fatigue detection. While we included baseline LSTM and 1D-CNN models for comparison, our emphasis was on demonstrating that lightweight ML approaches can be better suited for practical deployment in data-constrained, resource-limited settings.

Although the accuracy of our proposed method does not surpass most of state-of-the-art approaches discussed in the literature, it offers several practical advantages. It is computationally efficient, interpretable, and capable of providing quick predictions, making it especially suitable for deployment on low-resource systems. Moreover, traditional machine learning models are easier and faster to update when new data becomes available, compared to more complex deep learning frameworks.

With the inclusion of more training samples, additional relevant features, and more advanced feature engineering which is benefitial for the deep learning models, the overall performance of the framework could be further improved. Nonetheless, our findings support the use of lightweight, interpretable models for mental fatigue estimation in resource-constrained settings.

ACKNOWLEDGMENT

The research is due to the grant of the Russian Science Foundation #24-21-00300, https://rscf.ru/project/24-21-00300/.

REFERENCES

[1] D. Xie, X. Wang, and C. Yin, "Research on the influence of operator fatigue factors on port service capability based on discrete system simulation," *SHS Web of Conferences*, vol. 181, p. 6, 2024.

[2] W. P. Rogers, J. Marques, E. Talebi, and F. A. Drews, "Iot-enabled wearable fatigue-tracking system for mine operators," *Minerals*, vol. 13, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2075-163X/13/2/287

[3] G. Zhang, K. K. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis Prevention*, vol. 87, pp. 34–42, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457515301159

[4] R. Alharbey, M. M. Dessouky, A. Sedik, A. I. Siam, and M. A. Elaskily, "Fatigue state detection for tired persons in presence of driving periods," *IEEE Access*, vol. 10, pp. 79 403–79 418, 2022.

[5] L. L. Di Stasi, R. Renner, A. Catena, J. J. Cañas, B. M. Velichkovsky, and S. Pannasch, "Towards a driver fatigue test based on the saccadic main sequence: A partial validation by subjective report data," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 122–133, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X1100101X

[6] Y. Yamada and M. Kobayashi, "Fatigue detection model for older adults using eye-tracking data gathered while watching video: Evaluation against diverse fatiguing tasks," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 275–284.

[7] D. Dawson, A. K. Searle, and J. L. Paterson, "Look before you (s)leep: Evaluating the use of fatigue detection technologies within a fatigue risk management system for the road transport industry," *Sleep Medicine Reviews*, vol. 18, no. 2, pp. 141–152, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079213000300

[8] H. A. Kassem, M. Chowdhury, and J. H. Abawajy, "Drivers fatigue level prediction using facial, and head behavior information," *IEEE Access*, vol. 9, pp. 121 686–121 697, 2021.

[9] L. Chen, G. Xin, Y. Liu, and J. Huang, "Driver fatigue detection based on facial key points and lstm," *Security and Communication Networks*, vol. 2021, no. 1, p. 5383573, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5383573

[10] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "Yawdd: A yawning detection dataset," 03 2014, pp. 24–28.

[11] G. Du, T. Li, C. Li, P. X. Liu, and D. Li, "Vision-based fatigue driving recognition method integrating heart rate and facial features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3089–3100, 2021.

[12] J. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, pp. 362–370(8), December 1993. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0054

[13] C. Papazov, T. K. Marks, and M. Jones, "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4722–4730.

[14] A. A. Minhas, S. Jabbar, M. Farhan, and M. N. ul Islam, "A smart analysis of driver fatigue and drowsiness detection using convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, pp. 26 969 – 26 986, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249137488

[15] L. Wang, H. Li, Y. Yao, D. Han, C. Yu, W. Lyu, and H. Wu, "Smart cushion-based non-invasive mental fatigue assessment of construction equipment operators: A feasible study," *Advanced Engineering Informatics*, vol. 58, p. 102134, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034623002628

[16] H. Luo, P.-A. Lee, I. Clay, M. Jaggi, and V. De Luca, "Assessment of Fatigue Using Wearable Sensors: A Pilot Study," *Digital Biomarkers*, vol. 4, no. Suppl. 1, pp. 59–72, 11 2020. [Online]. Available: https://doi.org/10.1159/000512166

[17] C. Shengli, P. Feng, W. Kang, Z. Chen, and B. Wang, "Optimized driver fatigue detection method using multimodal neural networks," *Scientific Reports*, vol. 15, 04 2025.

[18] W. Othman, B. Hamoud, N. Shilov, and A. Kashevnik, "Human operator mental fatigue assessment based on video: Ml-driven approach and its application to hfavd dataset," *Applied Sciences*, vol. 14, no. 22, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/22/10510

[19] B. Hamoud, W. Othman, and N. Shilov, "Analysis of computer vision-based physiological indicators for operator fatigue detection," in *2025 37th Conference of Open Innovations Association (FRUCT)*, 2025, pp. 47–58.

[20] G. Salzillo, C. Natale, G. B. Fioccola, and E. Landolfi, "Evaluation of driver drowsiness based on real-time face analysis," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 328–335.

[21] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, *Karolinska Sleepiness Scale (KSS)*. New York, NY: Springer New York, 2012, pp. 209–210. [Online]. Available: https://doi.org/10.1007/978-1-4419-9893-4_47

[22] H. Pei, G. Li, Y. Ma, H. Gong, M. Xu, and Z. Bai, "A mental fatigue assessment method for pilots incorporating multiple ocular features," *Displays*, vol. 87, p. 102956, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141938224003202

[23] Z. Ji, X. Xie, E. Jiang, Y. Wang, B. Min, S. Yang, Y. Chen, and D. Pons, "Integrating drn-rf with computer vision for detection of control room operator's mental fatigue," *PLOS ONE*, vol. 20, no. 4, pp. 1–26, 04 2025. [Online]. Available: https://doi.org/10.1371/journal.pone.0320780

[24] J. Fan, L. Dong, G. Sun, and Z. Zhou, "A deep learning approach for mental fatigue state assessment," *Sensors*, vol. 25, no. 2, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/2/555

[25] N. Gallagher, "Savitzky-golay smoothing and differentiation filter," 01 2020.

[26] S. Kovalenko, A. Mamonov, V. Kuznetsov, A. Bulygin, I. Shoshina, I. Brak, and A. Kashevnik, "Operatoreyevp: Operator dataset for fatigue detection based on eye movements, heart rate data, and video information," *Sensors*, vol. 23, no. 13, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/6197