DRIVER DROWSINESS DETECTION

VICTORIA GARCÍA MARTÍNEZ-ECHEVARRÍA

31 DE DICIEMBRE DE 2025

MASTER'S DEGREE IN ARTIFICIAL INTELLIGENCE
ICAI – COMILLAS PONTIFICAL UNIVERSITY

DRIVER DROWSINESS DETECTION

XAI – FINAL PROJECT

Introduction

Over the past decades, driver assistance systems have evolved significantly, ranging from simple fixed recommendations (such as suggesting a rest break after a certain number of driving hours) to highly sophisticated technologies that provide real-time, context-aware support, including surround-view cameras for parking or lane-keeping assistance, among others. Sensors, cameras and artificial intelligence techniques play an increasingly important role in this area, allowing for constant revision and interpretation of both the external and internal environments of the vehicle.

Within this context, driver drowsiness detection has emerged as a particularly relevant problem. It is a well-known fact that driver fatigue is one of the main risk factors in road accidents, especially during long trips or night-time driving. Hence, vision-based approaches are presented as a viable solution, as interior cameras can capture symptoms such as eyelid closure, prolonged blinking, or general facial fatigue.

Deep learning models, and convolutional neural networks (CNNs) in particular, have shown strong performance in image-based classification tasks (and drowsiness detection is no exception). However, their growing adoption also raises important concerns regarding transparency and trust. If a model bases its predictions on spurious correlations (such as lighting conditions, head pose or background details) rather than on physical and physiological signals, its reliability in real-world situations becomes quite questionable.

These concerns are highly relevant to multiple stakeholders. Manufacturers of the automotive industry rely on these systems to enhance vehicle safety and differentiate their products. Companies that develop advanced driver-assistance systems must ensure that their models behave robustly across different drivers and environments. Also, safety regulators increasingly demand more transparency and accountability from automated decision-making systems, particularly when human lives are at stake. Therefore, understanding why a model returns a certain prediction is almost as important as the prediction itself.

Explainable AI (XAI) provides tools to tackle these challenges by making model behavior more interpretable. In this case, explainability helps verify whether the model focuses on relevant facial regions (most notably the eyes and eyelids, we would expect) rather than unintended shortcuts. It also enables the detection of unstable or biased decision patterns and allows for informed design choices when improving models.

In this project, we study the explainability of a computer vision model trained for driver drowsiness detection using the Driver Drowsiness Dataset (DDD). A classifier based on the ResNet50 model is used as a baseline, and several XAI techniques are applied to analyze its behavior, both locally and globally. Local explanations are obtained through Occlusion Sensitivity and RISE, and global behavior is studied using aggregated Grad-CAM visualizations. In addition, sanity checks are performed to assess the faithfulness of the explanations. Finally, after observing these explanations, insights are applied in an actionable way to evaluate an intervened model in which attention is constrained to eye-related regions, illustrating the trade-off between prediction performance and alignment with domain knowledge.

Data and Methods

This project is based on the Driver Drowsiness Dataset (DDD), available in Kaggle (link here), which is a public dataset designed for binary classification (each image should be labeled either as drowsy or non-drowsy). The dataset consists of facial images extracted from videos of the Real-Life Drowsiness Dataset (shot with an in-cabin camera, under controlled conditions), and they all share consistent framing and camera placement.

Each image ought to be assigned to one of the two classes based on visual indicators of drowsiness or fatigue, according to the facial expressions observed. The dataset was randomly split into three different sets (training, validation, and test) to ensure that the accuracy and performance of the model is evaluated on unseen data. All images were resized to a fixed resolution of 224x244 pixels and normalized using statistics from the ImageNet baseline.

The model used is based on one of the implementations shared on Kaggle by users working with the same dataset (link here). Its baseline is the ResNet50 model, a CNN architecture pre-trained on ImageNet. Most of its layers are frozen so that only the last three of them are trained to fine-tune the model for the drowsiness classification task (in short, the original classification head is replaced with a fully connected layer adapted for binary classification). The fine-tuning process is carried out using cross-entropy loss and the Adam optimizer, with early stopping based on validation loss to avoid overfitting.

Initially, a baseline architecture based on MobileNetV2 was considered, seeing that in the early research stage of the project we observed that this model was widely used in similar classifications tasks, especially with this dataset in particular. However, despite extensive experimentation and multiple attempts, the training process could not be completed within reasonable resource constraints, even when executing the code in the DGX server provided by the university. For this reason, the ResNet50-based approach was preferred as it was a more stable alternative (although its training also required the use of the DGX server given the computational limitations of local hardware).

To analyze the behavior of the trained model, several explainability techniques were selected to cover both local and global explanations, as well as sanity checks, in line with the given project requirements. For local explanations, two complementary, model-agnostic methods are used. Occlusion Sensitivity masks different regions of the input image by sliding a small patch over it and then measures the resulting change in the model's output. This allows us to identify which regions of the image are critical for individual predictions. The second method, RISE (Randomized Input Sampling for Explanation), applies randomized binary masks to the input image and then aggregates the model's responses to these perturbations to generate a saliency map that highlights those regions that strongly influence the model's predictions.

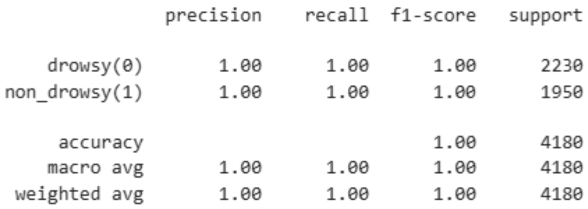
For global explanations, the initial plan was to apply SHAP to analyze feature importance across the dataset. However, due to a series of unsolvable incompatibilities between library versions (being *tensorflow* and *numpy* the most problematic), this approach could not be implemented. As an alternative, seeing that the framing of faces was highly consistent across the entire dataset, we opted for a mean Grad-CAM technique. Grad-CAM heatmaps were computed for a large number of images belonging to the 'drowsy' class and averaged pixel-wise. This shows which regions of the image the model relies on most strongly on average, and are considerably interpretable as the position of the face in every image is nearly the same in all samples.

Finally, sanity checks are performed to evaluate the faithfulness of the explanations. These include randomizing model parameters and observing the degradation of the quality of the explanations as a result, and also removing the most significant regions and measuring how rapidly the model's confidence drops. These checks help ensure that the explanations truly reflect the model's behavior and not just generic image characteristics.

Results

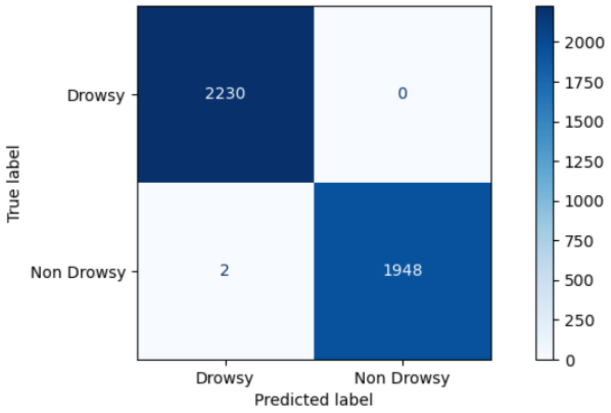
Model performance

The trained model achieves very high performance on the test set. As seen in the classification report below, the model reaches an accuracy of 1.00 and the area under the ROC curve is also 1.00. Precision, recall, and F1-score are all 1.00 for both classes, indicating an almost perfect separation between the two classes of the dataset. The confusion matrix shows that only two samples are misclassified, out of more than four thousand images that belong to the test set.



```
              precision    recall  f1-score   support

   drowsy(0)       1.00      1.00      1.00      2230
non_drowsy(1)      1.00      1.00      1.00      1950

    accuracy                           1.00      4180
   macro avg       1.00      1.00      1.00      4180
weighted avg       1.00      1.00      1.00      4180
```

Classification report



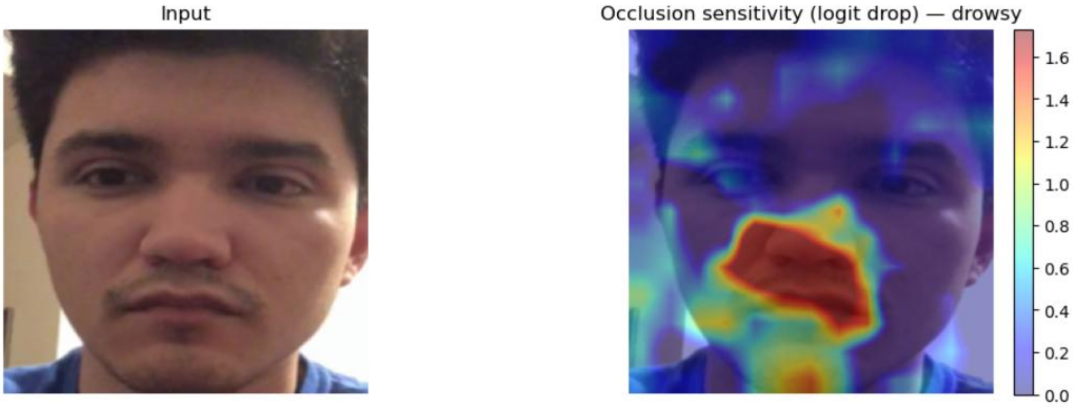
Confusion matrix

While these results suggest that the classification task is easily solvable for the model, the fact that its performance is near-perfect also raises important questions regarding the nature of the cues the model relies on. Hence, it motivates a deeper analysis of the model's decision-making process so that we can confirm whether the predictions are based on physiological signals or on dataset-specific patterns that may not generalize beyond this context.

Local explanations

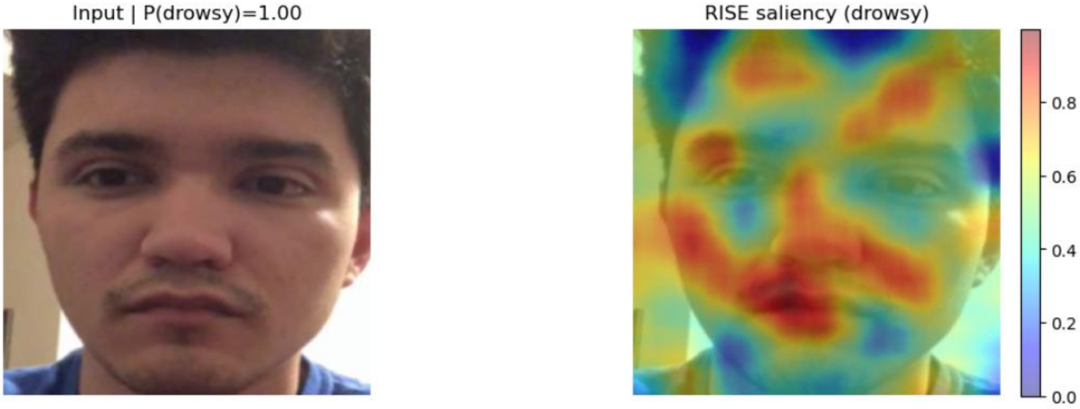
To better understand individual predictions, two local explanation techniques were applied: Occlusion Sensitivity and RISE.

The figure below illustrates the Occlusion Sensitivity analysis for a correctly classified image from the 'drowsy' class (which we labeled as 0, being 1 the label for the 'non-drowsy' class). Contrary to expectations, the most influential regions are not concentrated around the eyes or eyelids. Instead, when the areas around the nose, mouth, and central facial region are occluded is when the model's output logit for the 'drowsy' class suffers the largest drop. Conversely, occluding the eye region results in a comparatively smaller effect on the prediction. This comes as a surprise, given that eye-related details such as eyelid closure or blinking are typically considered the most direct visual indicators of driver drowsiness.



Saliency map: Occlusion Sensitivity

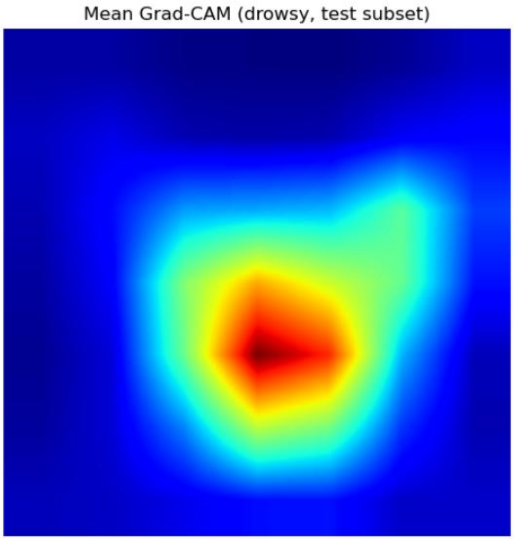
A similar pattern is observed when applying the RISE method. The RISE saliency map highlights broad regions across the face, especially over its central parts, including the nose, cheeks, and mouth, rather than focusing directly on the eye area. Although some activation is visible near the eyes, it is far from dominant and it is not consistent across different samples.



Saliency map: RISE

Global explanations

To analyze the model's average behavior across the dataset, the mean Grad-CAM was computed over a large subset of 'drowsy' test images to act as global explanation. The resulting heatmap is the following.

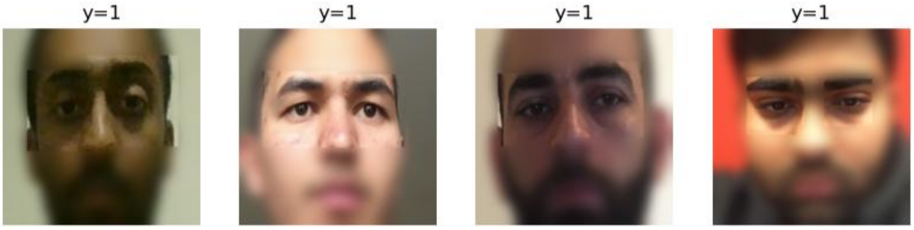


Consistent with the local explanations covered before, the mean Grad-CAM does not focus on the eye region. Instead, the highest activation appears in the area that corresponds to the central and lower parts of the face, and the eye area receives comparatively less emphasis. Since the framing of faces in the dataset is highly consistent, this averaged visualization shows a reliable indication of the regions to which the model pays the most attention when predicting drowsiness.

Actions and Insights

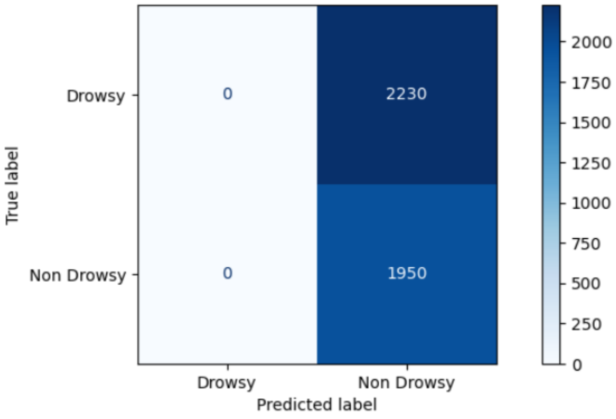
The explanation results presented in the previous section reveal a consistent and unexpected pattern: even though the model achieves a near-perfect performance, it does not primarily rely on the eye region when identifying drowsiness in drivers but on the central facial region (nose, mouth). From a domain perspective, this behavior is counterintuitive. Eye-related cues, such as reduced eye openness, eyelid closure, and prolonged blinking are widely recognized as key indicators of drowsiness. So, seeing this discrepancy between domain intuition and model behavior, an actionable intervention was designed to integrate domain knowledge into the learning process and evaluate its impact on performance.

For this purpose, an eye-based region-of-interest (ROI) approach was implemented. For every image in the dataset, the eye regions were automatically identified using facial landmark detection (with the *opencv* Python library). The rest of the face was then blurred, so that the model was forced to focus almost exclusively on the eye area. The resulting images after applying this transformation are shown in the figure below.

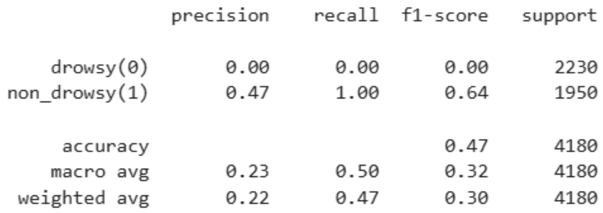


Example of four images from the dataset after applying the eye-ROI transformation

With this modified dataset, the model was then retrained from scratch using the same architecture and training protocol as the previous model. Still, the performance of the retrained eye-ROI model dropped dramatically compared to the original baseline. The model achieves an accuracy of 0.467 and a ROC-AUC of 0.466, values that are close to random guessing. The confusion matrix and classification report attached below show that the model predicts almost all samples as ‘non-drowsy,’ failing completely to correctly identify the ‘drowsy’ class.



Confusion matrix for the Eye-ROI model

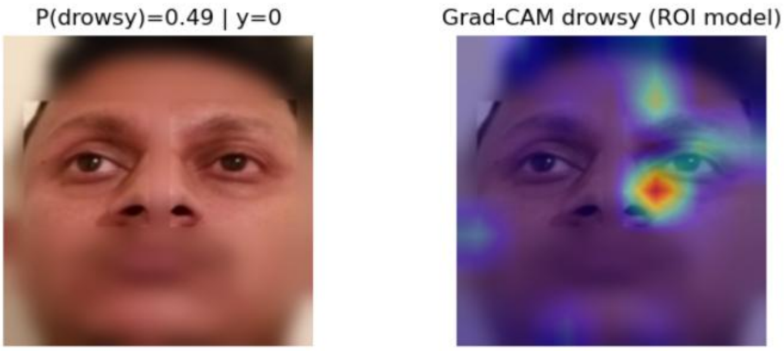


```
              precision    recall  f1-score   support

   drowsy(0)       0.00      0.00      0.00      2230
non_drowsy(1)      0.47      1.00      0.64      1950

    accuracy                           0.47      4180
   macro avg       0.23      0.50      0.32      4180
weighted avg       0.22      0.47      0.30      4180
```

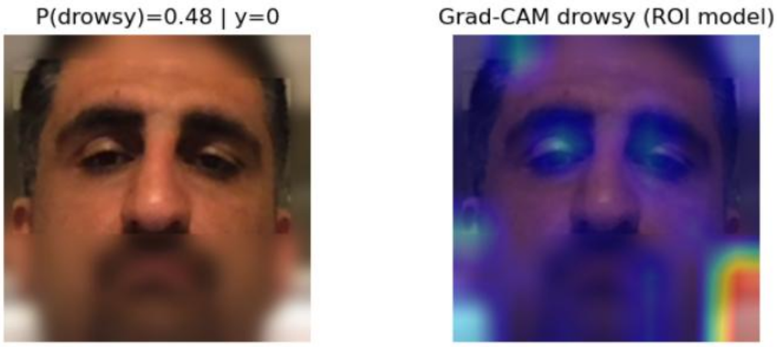
Classification report for the Eye-ROI model

This collapse in performance suggests that the information contained in the eye region alone is insufficient for the model to reliably distinguish between the two classes. There are several potential explanations for this result. First, it is possible that the images in the dataset do not include strong, consistent differences in eye appearance between the two classes. Second, cues that can be observed in the lower and central regions of the face such as yawning, facial slackness, or head position may be more informative in practice than subtle changes around the eyes. Lastly, the original baseline model may have learned to exploit correlations in the dataset that are not aligned with what us humans expect to be drowsiness indicators.

To further analyze the behavior of this Eye-ROI model, a few local Grad-CAM explanations were generated for individual predictions. In some cases, the Grad-CAM heatmaps show activation closer to the eye region (see first of the two images below), suggesting that the model sometimes does attempt to use eye-related features when they are available and accessible. However, in many other cases, the highlighted regions correspond to image borders, blurred regions, or background artifacts that do not have a clear semantic meaning. This instability in the explanations comes as no surprise given the model's low confidence (the predicted probabilities for both classes remain close to 0.5, which indicates a lack of discriminative signal).



Grad-CAM sample for the Eye-ROI model (1)



Grad-CAM sample for the Eye-ROI model (2)

The main conclusion of this experiment is that incorporating human domain knowledge into a deep learning model does not necessarily improve performance and can, in some cases, significantly degrade it. While humans naturally associate drowsiness with eye symptoms, the model relies on a different set of visual cues that provide better predictions within the dataset distribution.

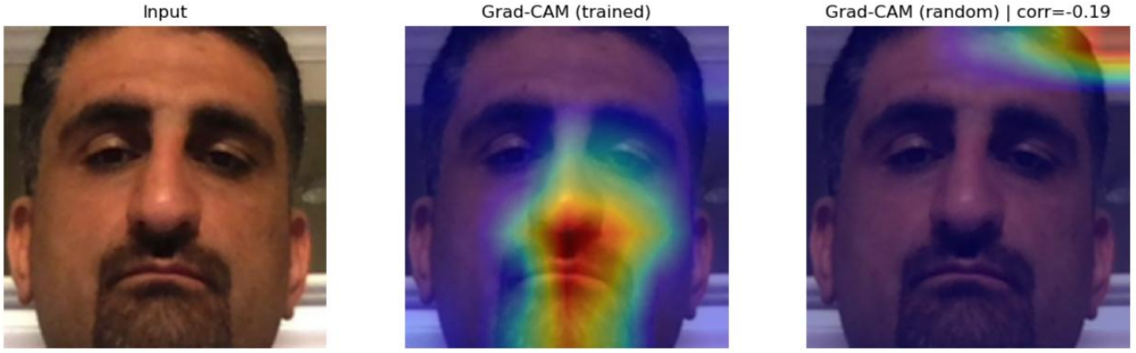
The failure of the Eye-ROI model does not imply that eye-based features are irrelevant in general but that, in this specific dataset, they are not enough on their own. More generally, this result showcases a key claim in explainable AI: models do not necessarily reason in the same way humans do, even when solving tasks in which the main indicators appear to be intuitive. Understanding and accepting this gap is essential when deploying models in safety-critical settings.

Evaluation of XAI

The explainability analysis carried out in this project provides useful insights into the model's behavior, but also reveals important limitations and risks that must be carefully considered, especially given the safety-critical context in which driver monitoring systems are implemented.

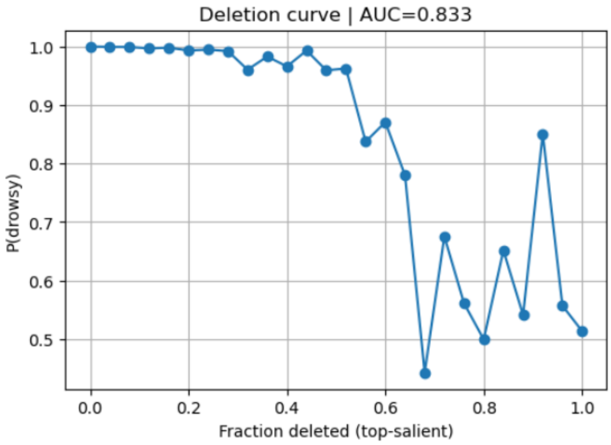
One notable strength of the analysis is the consistency observed across different explanation techniques. Occlusion Sensitivity, RISE and Grad-CAM all agree that the model significantly relies on the central area on the face, particularly around the face and mouth, rather than on the eye region. The fact that both model-agnostic, perturbation-based methods and a gradient-based method converge to the same conclusion adds credibility to this observation and suggests that it truly reflects a genuine property of the learned model.

To assess the faithfulness of the explanations, two sanity checks were performed. First, model parameters were randomized, and Grad-CAM was computed again. The resulting saliency maps lost their structural patterns and are no longer aligned with meaningful facial regions, which indicates that the original explanations truly depend on the learned weights rather than capturing generic image features.



Grad-CAM applied after randomizing model parameters

Secondly, a deletion-based evaluation shows that the model's confidence remains stable while a moderate fraction of the most salient pixels is removed, and it only drops significantly once a large portion of the image (around 60% or more) has been deleted. This suggests that the model's predictions rely on relatively distributed information rather than on a small, localized set of pixels. Still, the area under this curve is lower than that of the original model.



Performance curve after progressively deleting salient pixels

Despite passing these sanity checks, the explanations are yet subject to important limitations. The fact that they consistently emphasize regions that are not aligned with domain intuition raises the risk of misinterpretation. For instance, a stakeholder may incorrectly conclude that eye-related cues are irrelevant for detecting drowsiness in general, when in reality we cannot confirm if this is a general fact or if it only holds within the distribution of this exact dataset. Moreover, it is possible that the high performance of the baseline model is due to the presence of strong dataset-specific correlations. Explanations, even when faithful, may therefore show *what* the model has learned instead of what it *should* learn. In such cases, explanations can transmit a false sense of reassurance. Lastly, there is also a risk of over-trusting visually appealing explanations. Heatmaps might appear convincing or intuitive whilst being unstable or difficult to interpret.

For stakeholders involved in deploying driver monitoring systems, such as safety regulators or automotive manufacturers, these findings demonstrate the importance of using XAI as a diagnostic tool to understand model behavior. Explanations should be combined with domain knowledge, dataset auditing, and targeted interventions to analyze robustness and generalization.

Some lines for future work could explore richer temporal cues using video-based models rather than static images, so that actions like yawning and eye blinking could be more thoroughly detected. Additionally, contemplating datasets with greater variability in conditions and framing may help differentiate between genuine physiological signals and spurious correlations.