Jornadas de Automática

IA explicable en detección de emociones y somnolencia para ADAS

Caballero García-Alcaide, Diego[a,*], Sesmero, M. Paz[a], Iglesias, José A.[a], Sanchis, Araceli[a]

[a]*Departamento de Informática, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911, Leganés, Madrid, España.*

To cite this article: Caballero García-Alcaide, D., Sesmero, M.P., Iglesias, J.A., Sanchis, A. 2025. Explainable AI in emotion and drowsiness detection for ADAS. Jornadas de Automática, 46.
https://doi.org/10.17979/ja-cea.2025.46.12074

Resumen

Los Sistemas Avanzados de Asistencia a la Conducción (ADAS) son cruciales para mejorar la seguridad vial. Este trabajo explora la aplicación de la Inteligencia Artificial Explicable (XAI) para analizar y comparar el comportamiento de modelos de aprendizaje profundo, específicamente Redes Neuronales Convolucionales (CNN), en la detección de emociones y estados de somnolencia en conductores. Mediante técnicas de XAI, se investigan los procesos de toma de decisiones de los modelos, ofreciendo transparencia e interpretabilidad. Se discuten los hallazgos sobre cómo los modelos identifican características faciales relevantes para cada tarea y las diferencias inherentes entre la detección de emociones y de somnolencia. Finalmente, se analizan las implicaciones de estos hallazgos para el desarrollo y la confianza en futuros ADAS, destacando el potencial de la XAI para refinar estos sistemas y reducir el número de accidentes de tráfico.

Palabras clave: Aprendizaje Automático, Sistemas difusos y neuronales para control e identificación, Soporte a la toma de decisiones, Asistencias inteligentes al conductor, Percepción y detección.

Explainable AI in emotion and drowsiness detection for ADAS

Abstract

Advanced Driver Assistance Systems (ADAS) are crucial for enhancing road safety. This work explores the application of Explainable Artificial Intelligence (XAI) to analyze and compare the behavior of deep learning models, specifically Convolutional Neural Networks (CNN), in detecting driver emotions and drowsiness states. Using XAI techniques, the decision-making processes of the models are investigated, offering transparency and interpretability. Findings on how models identify relevant facial features for each task and the inherent differences between emotion and drowsiness detection are discussed. Finally, the implications of these findings for the development and trust in future ADAS are analyzed, highlighting XAI's potential to refine these systems and reduce the number of traffic accidents.

Keywords: Machine Learning, Fuzzy and neural systems relevant to control and identification, Decision-making support, Intelligent driver aids, Perception and sensing.

1. Introducción

Los accidentes de tráfico, con el error humano como causa predominante, representan un grave problema de salud pública. Los Sistemas Avanzados de Asistencia a la Conducción (ADAS) buscan mitigar estos riesgos, y para ello, la detección precisa de estados críticos del conductor, como emociones negativas o somnolencia, es fundamental. El aprendizaje profundo, y en particular las Redes Neuronales Convolucionales (CNN), han demostrado un gran potencial en estas tareas. Sin embargo, la naturaleza de "caja negra" de estos modelos a menudo limita su interpretabilidad y confianza, especialmente en aplicaciones críticas para la seguridad.

Este trabajo se centra en la aplicación de técnicas de Inteligencia Artificial Explicable (XAI) para desvelar el com-

*Autor para correspondencia: dicaball@inf.uc3m.es

portamiento interno de modelos convolucionales entrenados para la detección de emociones y somnolencia. El objetivo es comprender qué características visuales utilizan los distintos modelos para sus predicciones, comparar los desafíos inherentes a cada tarea y extraer conclusiones para el desarrollo de ADAS más robustos y fiables. Se busca ir más allá de las métricas puramente numéricas, utilizando XAI para comprender los resultados de los modelos, facilitar su depuración y mejorar la transparencia.

Entre las contribuciones de este trabajo destacan:

- Análisis comparativo mediante inteligencia artificial explicable del proceso de decisión en modelos de detección de emociones y somnolencia.
- Identificación de factores complejos que afectan a cada tarea, apoyando el desarrollo de sistemas de monitorización de la conducción más eficaces.

2. Estado del arte

El reconocimiento de emociones y la detección de somnolencia en conductores son áreas de investigación activas con el objetivo de mejorar la seguridad vial. Diversos estudios han demostrado el impacto significativo de las emociones, especialmente negativas como la ira o la tristeza, en el rendimiento de la conducción (Jeon et al., 2011; Jeon, 2016). Aunque existen múltiples enfoques para el reconocimiento de emociones, como los basados en señales de electroencefalografía (EEG) (Sheykhivand et al., 2020), muchos resultan intrusivos para su inclusión en escenarios de conducción reales. Por ello, este trabajo se alinea con enfoques no intrusivos que utilizan el aprendizaje profundo, específicamente CNNs entrenadas con imágenes de expresiones faciales, una línea similar a la de (Verma and Choudhary, 2018).

En cuanto a la detección de somnolencia, trabajos como el de (Tamanani et al., 2021) emplean aprendizaje profundo sobre imágenes extraídas de vídeos para estimar el estado de vigilancia del conductor, reportando altas precisiones, aunque a veces sin una adecuada separación de sujetos entre los conjuntos de entrenamiento y test, lo que puede limitar la generalización. Otros estudios, como el de (Magán et al., 2022), exploran técnicas de aprendizaje profundo para la integración en ADAS, enfatizando la reducción de falsos positivos, aunque con precisiones que aún presentan margen de mejora.

La Inteligencia Artificial Explicable (XAI) ha comenzado a aplicarse en este dominio para aportar transparencia. Por ejemplo, (Lorente et al., 2021) aplicaron técnicas XAI a modelos de detección de emociones y distracciones, revelando cómo los modelos identifican regiones importantes en las imágenes y destacando limitaciones que no se aprecian solo con métricas de rendimiento.

Este trabajo busca extender estos análisis, comparando específicamente la detección de emociones y somnolencia mediante IA explicable para entender mejor los desafíos y refinar los sistemas de monitorización. Se busca abordar la necesidad de sistemas no intrusivos y generalizables, utilizando XAI para profundizar en el comportamiento de los modelos más allá de la simple precisión.

3. Metodología

Para el análisis de emociones, se utilizó el dataset FER-2013 (Goodfellow et al., 2013), que contiene imágenes faciales en escala de grises con siete categorías de emoción. Para la detección de somnolencia, se empleó el dataset UTA-RLDD (Ghoddoosian et al., 2019), compuesto por vídeos de participantes simulando distintos estados de alerta y somnolencia. Para simplificar, se consideró una clasificación binaria (despierto vs. somnoliento). Las imágenes de ambos datasets fueron preprocesadas, incluyendo la extracción de rostros y reescalado. La Tabla 1 resume las características de los datasets utilizados tras su preprocesamiento. Además, se aplicaron técnicas de aumento de datos para mejorar la generalización.

Tabla 1: Comparación de los datasets utilizados tras el preprocesamiento.

	FER-2013	UTA-RLDD
Origen	Imágenes	Vídeos
Imágenes	35.887	90.000
Tamaño (px)	48x48	64x64
Formato	Grayscale	RGB
Nº Clases	7	2
Distribución	Desbalanceado	Balanceado

Se entrenaron diversas arquitecturas CNN, desde modelos más simples como *LeNet* hasta más complejos como variantes de *VGG* (Simonyan and Zisserman, 2014). El entrenamiento se realizó ajustando parámetros como la tasa de aprendizaje, utilizando *Adam* como optimizador. En cuanto a las funciones de activación, se empleó *ReLU* en capas intermedias, y *Soft-Max* para emociones y *Sigmoide* para somnolencia en la capa de salida. El rendimiento se evaluó en conjuntos de entrenamiento, validación y test, asegurando una división por sujetos en el dataset de somnolencia para evaluar la capacidad de generalización del modelo.

Para el análisis explicativo, se seleccionaron los modelos con mejor rendimiento en cada tarea y se aplicó la técnica XRAI (Kapishnikov et al., 2019), un método de atribución basado en gradientes integrados que divide la imagen en regiones y calcula la importancia de cada una, permitiendo visualizar las áreas más relevantes para la predicción del modelo.

4. Experimentación

El objetivo de esta fase experimental fue identificar los modelos con el mejor rendimiento para su posterior análisis de explicabilidad mediante la aplicación de XRAI.

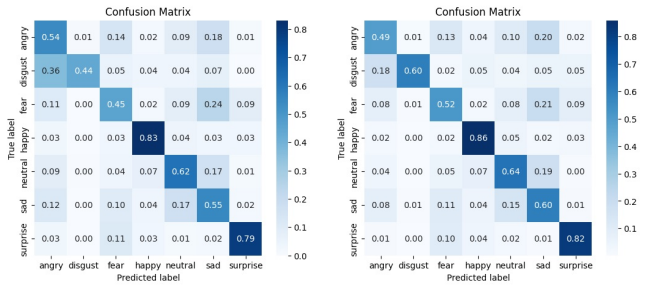
4.1. Resultados en detección de emociones

En la tarea de reconocimiento de emociones, la mayoría de los modelos alcanzaron precisiones razonables en el conjunto de entrenamiento, de entre 0.7 y 0.9. Los modelos más profundos de la familia *VGG*, especialmente aquellos que incluían capas de normalización por lotes (Ioffe and Szegedy, 2015) y Dropout (Srivastava et al., 2014) (sufijo BD), mostraron el mejor rendimiento en el conjunto de validación, con precisiones cercanas al 66 %. Los resultados en el conjunto de test fueron consistentes con los de validación, sugiriendo una buena generalización. Todos estos resultados se recogen en la Tabla 2, ordenada por valor de precisión en el conjunto de test.

Tabla 2: Resultados de precisión de los modelos de detección de emociones.

Model	Train	Validation	Test
VGG-16BD	0.7577	0.6627	0.6441
VGG-19BD	0.7729	0.6560	0.6374
VGG-11BN	0.8858	0.6471	0.6371
ZfNetBD	0.8139	0.6348	0.6213
AlexNetBD	0.7792	0.6323	0.6188
VGG-11	0.6893	0.6117	0.6099
ZfNet	0.8271	0.6183	0.5926
AlexNet	0.7473	0.5896	0.5764
LeNet	0.5579	0.5269	0.5341
LeNetBD	0.4616	0.5096	0.5216
VGG-16	0.2509	0.2473	0.2470
VGG-19	0.2491	0.2473	0.2470

El modelo *VGG-16BD* obtuvo el mejor resultado en el conjunto de test. Puesto que este modelo presentaba un valor de precisión bajo en el conjunto de entrenamiento con respecto al resto de modelos, se volvió a entrenar durante más épocas, alcanzando finalmente un 66.3 % de precisión. Además, como se puede observar en la Figura 1, el modelo reentrenado mejoró la clasificación en la mayoría de las clases, especialmente en la clase minoritaria *disgust*, que fue la peor clasificada por el resto de los modelos.



(a) 50 epochs. (b) 100 epochs.

Figura 1: Matrices de confusión de modelo *VGG-16BD* entrenado durante diferente número de épocas para detección de emociones.

Adicionalmente, se puede observar cierta confusión por parte del modelo entre algunas clases, como por ejemplo *disgust* con *angry* o *fear* y *sad*. En cambio, las clases *happy* y *surprise* obtuvieron los mejores resultados de precisión, con 0.86 y 0.82 respectivamente.

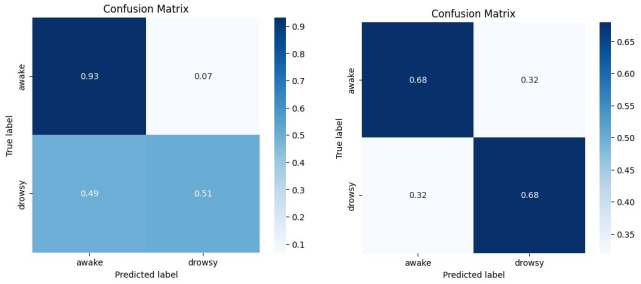
4.2. *Resultados en detección de somnolencia*

Para la tarea de detección de somnolencia, la mayoría de los modelos alcanzaron precisiones superiores al 90 % en el conjunto de entrenamiento. En los conjuntos de validación, modelos como *VGG-19BD*, *VGG-16*, *VGG-16BD*, *ZfNetBD* y *LeNet* obtuvieron precisiones entre el 70 % y el 76 %. Sin embargo, como muestra la Tabla 3, el rendimiento en el conjunto de test divergió del obtenido en validación, probablemente debido a la división por sujetos, ya que las características aprendidas durante el entrenamiento pueden funcionar mejor para unos sujetos que para otros.

Tabla 3: Resultados de precisión de los modelos de detección de somnolencia.

Model	Train	Validation	Test
AlexNet	0.9434	0.6592	0.7184
LeNet	0.9699	0.7224	0.6797
LeNetBD	0.9139	0.6646	0.6630
VGG-11BN	0.8074	0.6334	0.6486
VGG-16	0.9672	0.7367	0.6421
VGG-16BD	0.9727	0.7490	0.6370
VGG-19	0.8268	0.6884	0.6236
VGG-19BD	0.9567	0.7579	0.6152
ZfNet	0.9642	0.6067	0.6083
VGG-11	0.9317	0.6660	0.5918
AlexNetBD	0.8036	0.6719	0.5112
ZfNetBD	0.9406	0.7454	0.4928

Al analizar las matrices de confusión de los mejores modelos, mostradas en la Figura 2, se observó que el casi 72 % de precisión logrado por *AlexNet* se sustentaba en una alta tasa de falsos negativos (49 %), es decir, prácticamente la mitad de las imágenes de la clase *drowsy* fueron clasificadas erróneamente como *awake*. Esto supone una muy baja precisión en la detección de la clase positiva, especialmente en un contexto de seguridad vial, donde esta cobra especial importancia. Por esta razón, y considerando su equilibrio entre el rendimiento en los conjuntos de test y validación, así como una matriz de confusión balanceada, se seleccionó el modelo *LeNet* como el más adecuado para esta tarea, a pesar de no tener la precisión más alta en el conjunto de test.



(a) AlexNet. (b) LeNet.

Figura 2: Matrices de confusión de los modelos *AlexNet* y *LeNet* entrenados para detección de somnolencia.

5. Análisis de explicabilidad

Una vez seleccionados los modelos para cada tarea, el objetivo fue visualizar y comprender qué regiones de las imágenes de entrada influyeron más en sus predicciones. A continuación, se detallan las observaciones para cada sistema.

5.1. *Explicabilidad en detección de emociones*

La aplicación de XRAI al modelo de detección de emociones elegido (*VGG-16BD*) reveló que, para predicciones correctas, el modelo se enfoca en regiones faciales representativas de cada emoción. Por ejemplo, como se muestra en la Figura 3, en imágenes de la clase sorpresa, la boca abierta es una región de alta importancia, en imágenes de felicidad, la sonrisa es clave y en imágenes de asco o enfado el modelo presta mayor atención a ciertas muecas en la cara.

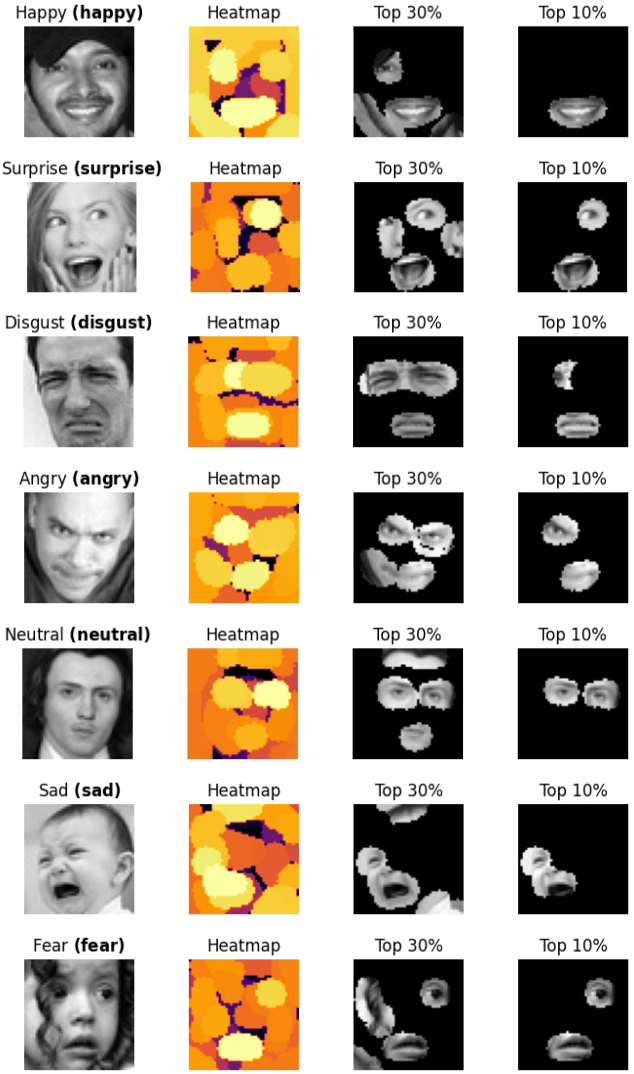


Figura 3: Aplicación de XRAI a imágenes de emociones bien clasificadas.

Por otro lado, se observó que la mayoría de confusiones entre clases, como entre *disgust* y *angry*, ocurren cuando las expresiones faciales son similares, y XRAI mostró que el modelo se fija en características ambiguas compartidas, como ocurre en la Figura 4, donde una imagen de la clase *disgust* fue clasificada como *angry*. Esto sugiere que el modelo aprende patrones visuales coherentes con la percepción humana.

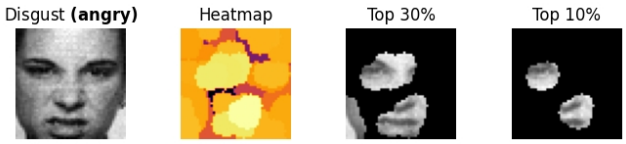


Figura 4: Imagen de la clase *disgust* clasificada como *angry* por el modelo.

5.2. *Explicabilidad en detección de somnolencia*

En la detección de somnolencia (*LeNet*), la aplicación de IA explicable indicó que el modelo se centra principalmente en los ojos. Para la clase *drowsy*, los ojos cerrados o entrecerrados son las regiones más importantes y para la clase *awake* los ojos abiertos reciben mayor atención, como se muestra en las imágenes de la Figura 5.

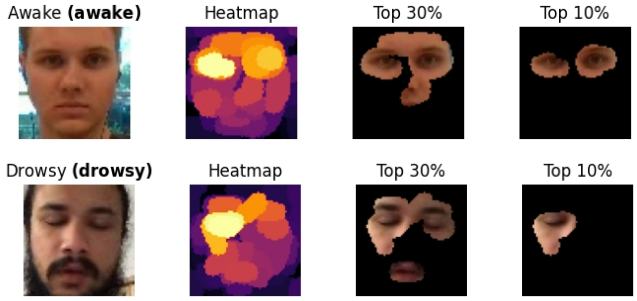


Figura 5: Aplicación de XRAI a imágenes de somnolencia bien clasificadas.

Sin embargo, el análisis de errores de clasificación reveló desafíos. Como se puede observar en la Figura 6, imágenes etiquetadas como *drowsy* donde el sujeto tenía los ojos abiertos eran frecuentemente clasificadas como *awake*. Esto subraya la limitación de usar fotogramas aislados para la detección de un fenómeno progresivo, como lo es la somnolencia, y la importancia de un etiquetado preciso del dataset.

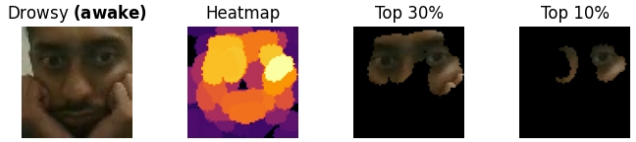


Figura 6: Imagen de la clase *drowsy* clasificada como *awake* por el modelo.

Además, se observó que, en algunos casos, el modelo asignaba importancia a regiones no relevantes (e.g. bordes de gafas), indicando que el modelo no ha sido capaz de inferir con precisión las características adecuadas que indican si un sujeto está somnoliento o no.

6. Discusión y comparación

La comparación de los resultados de XRAI para ambas tareas revela diferencias significativas. La detección de emociones, aunque con sus propios desafíos de ambigüedad entre clases, parece más abordable con imágenes estáticas porque las expresiones faciales son manifestaciones relativamente estables. La XAI muestra que los modelos aprenden a identificar regiones consistentes con estas expresiones.

La detección de somnolencia es inherentemente más compleja. Es un proceso progresivo con fluctuaciones rápidas (e.g. parpadeo lento, cabeceo), por lo que los fotogramas individuales pueden no capturar la dinámica temporal. La XAI evidencia que los modelos se esfuerzan por encontrar indicadores en los ojos, pero la variabilidad y sutileza de las señales, junto con posibles problemas en el etiquetado a nivel de fotograma, dificultan la generalización. Además, la menor variabilidad en los datos de somnolencia puede llevar a un sobreajuste a las imágenes de entrenamiento.

Estos hallazgos tienen implicaciones que pueden ser tenidas en cuenta para el diseño de los ADAS:

- **Calidad de los datos:** La XAI resalta la necesidad de disponer de datasets de alta calidad y con un etiquetado preciso y contextual, especialmente para la detección de somnolencia.

- **Diseño del modelo:** Para detección de somnolencia, modelos que incorporen información temporal, como los Transformers (Vaswani et al., 2017), podrían ser más efectivos que los basados únicamente en CNNs.
- **Lógica de alerta en ADAS:** Es más efectivo entrenar un modelo para predecir estados específicos de forma precisa y luego transferir la lógica de alerta al ADAS, que puede combinar la predicción del modelo con otros factores, como duración, velocidad del vehículo, etc, antes de emitir una alerta.

7. Conclusiones

Este estudio ha demostrado la utilidad de la Inteligencia Artificial Explicable para analizar y comparar modelos de detección de emociones y somnolencia en conductores. La XAI permite comprender cómo los modelos toman decisiones, identificando las características faciales que consideran relevantes y revelando las dificultades intrínsecas de cada tarea. Además, proporciona información valiosa sobre la calidad y consistencia de los datos utilizados, lo que resulta crucial para la interpretación y fiabilidad de los resultados.

Por otro lado, los análisis mediante XAI han revelado que la detección de somnolencia presenta mayores desafíos que la de emociones cuando se utilizan fotogramas aislados, debido a su naturaleza progresiva y la sutileza de sus indicadores. También, sugieren que la calidad del etiquetado de los datos es crítica y que enfoques que consideren la información temporal podrían ser más adecuados para esta tarea. La detección de somnolencia sigue siendo un problema abierto y desafiante, pero está claro que las técnicas de aprendizaje profundo tienen un potencial significativo para su incorporación en los ADAS, facilitando innovaciones que pueden mejorar en gran medida la seguridad vial y reducir el número de accidentes.

La aplicación continua de XAI es crucial no solo para fomentar la transparencia y la confianza en los sistemas de IA, sino también para refinar el comportamiento de los modelos. El desarrollo de ADAS más seguros y eficaces se beneficiará enormemente de una comprensión más profunda de cómo estos sistemas interpretan la información del conductor.

Futuras investigaciones deberían explorar la integración de XAI en todo el ciclo de vida del desarrollo de estos sistemas, incluyendo el uso de explicaciones para guiar el diseño de modelos más robustos y la creación de métricas de interpretabilidad específicas. También, se profundizará en cómo las explicaciones generadas por XAI pueden ser utilizadas dinámicamente por los ADAS para modular sus alertas o para proporcionar retroalimentación al conductor. El objetivo es avanzar hacia sistemas de asistencia que no solo reaccionen a estados críticos, sino que también fomenten una mayor comprensión y confianza entre el humano y el sistema.

Agradecimientos

Este trabajo ha sido realizado parcialmente gracias al apoyo de la subvención PID2022-140554OB-C32 financiada por MCIN/AEI/10.13039/501100011033 y el proyecto TEC-2024/ECO-277 financiado por la Comunidad de Madrid.

Referencias

Ghoddoosian, R., Galib, M., Athitsos, V., 2019. A realistic dataset and baseline temporal model for early drowsiness detection.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y., 2013. Challenges in representation learning: A report on three machine learning contests.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

Jeon, M., 2016. Don't cry while you're driving: Sad driving is as bad as angry driving. International Journal of Human–Computer Interaction 32 (10), 777–790.
DOI: 10.1080/10447318.2016.1198524

Jeon, M., Roberts, J., Raman, P., Yim, J.-B., Walker, B. N., 2011. Participatory design process for an in-vehicle affect detection and regulation system for various drivers. In: The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '11. Association for Computing Machinery, New York, USA, p. 271–272.
DOI: 10.1145/2049536.2049602

Kapishnikov, A., Bolukbasi, T., Viégas, F., Terry, M., 2019. Xrai: Better attributions through regions.

Lorente, M. P. S., Lopez, E. M., Florez, L. A., Espino, A. L., Martínez, J. A. I., de Miguel, A. S., 2021. Explaining deep learning-based driver models. Applied Sciences 11 (8).
DOI: 10.3390/app11083321

Magán, E., Sesmero, M. P., Alonso-Weber, J. M., Sanchis, A., 2022. Driver drowsiness detection by applying deep learning techniques to sequences of images. Applied Sciences 12 (3).

Sheykhivand, S., Mousavi, Z., Rezaii, T. Y., Farzamnia, A., 2020. Recognizing emotions evoked by music using cnn-lstm networks on eeg signals. IEEE Access 8, 139332–139345.
DOI: 10.1109/ACCESS.2020.3011882

Simonyan, K., Zisserman, A., 9 2014. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15 (56), 1929–1958.

Tamanani, R., Muresan, R., Al-Dweik, A., 2021. Estimation of driver vigilance status using real-time facial expression and deep learning. IEEE Sensors Letters 5 (5), 1–4.
DOI: 10.1109/LSENS.2021.3070419

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc.

Verma, B., Choudhary, A., 11 2018. A framework for driver emotion recognition using deep learning and grassmann manifolds. pp. 1421–1426.
DOI: 10.1109/ITSC.2018.8569461