BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 1, pages: 7-25
Submitted: April 24th, 2025 | *Accepted for publication:* August 23rd, 2025

Towards Real-Time Explainable AI: Using Class Activation Mapping for Brake Prediction in YOLO-Based Systems

Amil Dar
Department of Computer Science & Information Technology, Faculty of Computing & Engineering, University of Kotli, Azad Jammu and Kashmir, Kotli 11100, Pakistan.
amil.rohani@uokajk.edu.pk,
https://orcid.org/0000-0002-5827-8633

Faisal Riaz
Department of CS&IT
Mirpur University of Science & Technology, Mirpur, Pakistan;
Control Automotive and Robotics Lab, Affiliated Lab of National
Centre of Robotics and Automation (NCRA), HEC, Islamabad, Pakistan.
faisal.riaz@must.edu.pk,
https://orcid.org/0000-0001-6437-0797

Abstract: *Predicting brake light activation in semi-autonomous vehicles (SAVs) is critical for improving road safety and optimising adaptive cruise control (ACC) systems. Despite the remarkable performance of deep learning-based object detectors such as YOLOv8, their inherent opacity in decision-making processes limits transparency, interpretability, and user trust factors essential for the deployment of safety-critical systems. This study introduces a novel, explainable framework that integrates state-of-the-art attribution-based explainability techniques, including EigenCAM, EigenGrad-CAM, LayerCAM, and HiResCAM, into the YOLOv8 architecture. The proposed framework systematically analyses activation patterns within the model, generating fine-grained saliency maps that highlight the spatial regions most influential in brake light detection. Using a real-world vehicle dataset comprising diverse lighting and environmental conditions, the model is fine-tuned to predict binary brake light states (car_BrakeOn and car_BrakeOff). Comparative experiments demonstrate the ability of these CAM-based methods to provide interpretable visual explanations while maintaining detection accuracy. This study is the first to explore the integration of explainability techniques within YOLO-based brake light detection systems for semi-autonomous vehicles (SAE Levels 2 and 3), addressing a critical gap in the literature. By bridging the divide between black-box AI models and human-understandable reasoning, this research promotes transparency, accountability, and user trust in AI-driven perception systems for intelligent transportation. The findings contribute a foundational step toward the broader adoption of explainable and reliable AI in safety-critical SAV applications.*

Keywords: *semi-autonomous vehicles; advanced driver assistance; XAI; explainable artificial intelligence; YOLO; CAM methods; transfer learning.*

How to cite: Dar, A. & Riaz, F. (2026). Towards real-time explainable AI: Using class activation mapping for brake prediction in YOLO-based Systems. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(1), 7-25. https://doi.org/10.70594/brain/17.1/1

1. Introduction

The steady increase in road traffic has led to an increase in accidents (*Global status report on road safety 2023*, 2023). Despite significant advances in infrastructure, some countries experienced a 7.7% increase in major accidents in 2024 compared to the previous year (Concerning rate of road trauma continues in first half of 2024 – RAA Daily, n.d.). Autonomous and semi-autonomous vehicles have also been recognised as contributing factors to traffic incidents and congestion. Due to the fast evolution of automated vehicle technology, modern transportation is undergoing a transformation, with the potential to redefine mobility and human interaction with the environment. The basic purpose of intelligent vehicles is to operate without human manual intervention, fully relying on active sensors, methods, and artificial intelligence (AI) to interpret their surroundings, make timely and informed decisions, and perform the tasks of driving. While automated vehicles offer promising benefits such as enhanced road safety, increased traffic efficiency and improved accessibility, their successful widespread adoption hinges on addressing significant hurdles (*Autonomous Vehicle Implementation Predictions: Implications for Transport Planning - TRID*, n.d.).

A major hurdle in the deployment of driverless and driver-assisted vehicles is ensuring the accuracy and reliability of their perception systems. These systems must detect, classify and track objects in dynamic and complex environments, such as urban streets, highways, and rural roads, while adjusting to changing weather and lighting conditions (Grigorescu et al., 2020). Recently developed AI methods, especially deep learning models, serve as the foundation of modern AV perception, thanks to their capability to process large volumes of sensor data and recognise complex patterns (Khalil et al., 2024). Yet despite considerable progress, these systems still exhibit limitations in certain scenarios. While deep learning models enhance perception capabilities, these are still criticised for their behaviour as a “black-box” (Hayashi, 2020), which is a hurdle in the transparency and interpretability of their decision-making procedure (Arrieta et al., 2020).

The opacity of AI-driven systems creates significant challenges for autonomous vehicles, where comprehending the reasoning behind crucial decisions is vital for ensuring safety, gaining public trust, and complying with regulatory standards (Alicioglu & Sun, 2022). XAI addresses this issue by introducing methods that enhance model interpretability, providing deeper insights into their decision-making processes, allowing both experts and non-experts to comprehend how decisions are formulated, errors are detected, and AI outputs are interpreted (Ahmed et al., 2022; Gunning et al., 2019).

Moreover, the incorporation of multimodal sensor data has become a fundamental strategy for improving the accuracy of AV systems. However, this approach also introduces significant onboard and computational complexities, making system interpretation more challenging (Ennajar et al., 2021). Traditional CAM-based methods were initially developed for CNN architectures designed for image classification, where activation maps are extracted from convolutional layers to highlight important regions. Some of the key models utilising Class Activation Mapping (CAM) techniques include VGG16, ResNet-50, ResNet-101, InceptionV3, DenseNet, and MobileNet, among others. While these methods work well in classification tasks, they face challenges in real-time object detection. The primary limitation is their high computational cost, as CAM techniques require back-propagation to generate heatmaps, making them slower and inefficient for fast detection tasks. Additionally, these methods are optimised for classification rather than object detection, meaning they lack precise spatial localisation required for tasks like brake light detection in dynamic driving environments.

To ensure transparency and interpretability in brake light prediction using the YOLOv8 model, this study critically investigates the application of CAM (Zhou et al., 2016) for model explainability. Given the safety-critical nature of autonomous driving scenarios, it is essential not only to achieve high accuracy but also to understand the reasoning behind the model’s decisions.

The YOLOv8 architecture consists of several intermediate convolutional layers that analyse the input image at multiple scales (Ultralytics, n.d.), as illustrated in Figure 1. At each layer, the network computes feature maps that capture increasingly abstract patterns in the data, enabling the model to detect objects of varying sizes and appearances. The backbone network extracts these features and passes them to the detector head for classification and localisation. Notably, a single large feature such as a vehicle's brake light may manifest as multiple smaller features when analysed at different scales, reflecting the model's multi-resolution perspective.

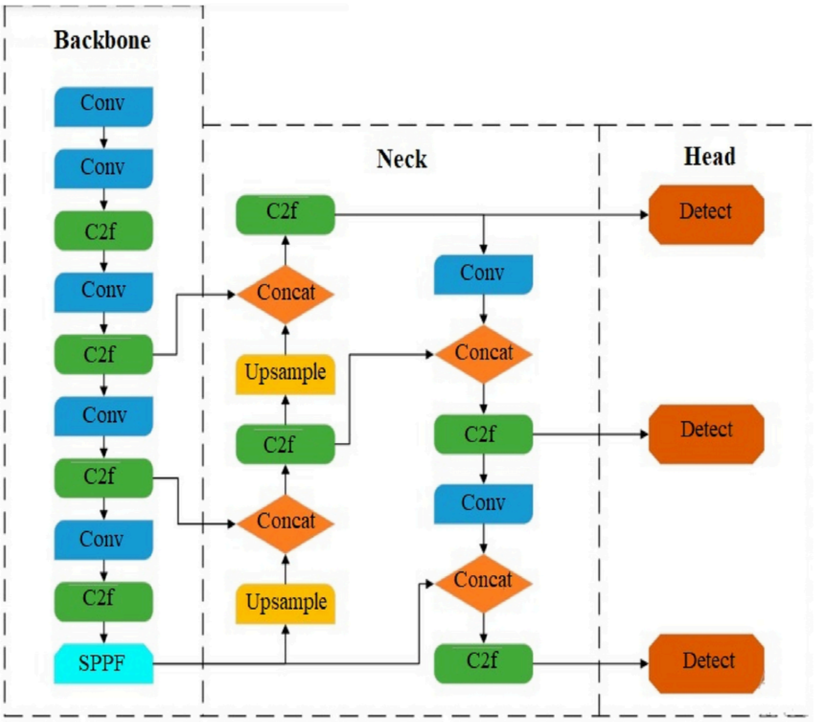


Figure 1. YOLOv8 Model Architecture.

By applying CAM techniques to the YOLOv8 model, this study aims to uncover the regions of interest that most influence the model's predictions. This analysis is critical for ensuring transparency, accountability, and user trust in AI-driven systems. The insights gained from this work will contribute to developing more interpretable and reliable deep learning models for real-world autonomous driving scenarios.

The key contributions of this work are as follows:

- By integrating XAI techniques into brake light prediction, the study provides insights into the model's decision-making process, contributing to enhanced transparency, reliability, and user understanding in intelligent transportation systems.
- This is the first study to apply XAI techniques, specifically CAM-based methods, for brake light prediction in semi-autonomous vehicles (SAE Levels 2 and 3). This novel contribution paves the way for developing more transparent, accountable, and trustworthy perception systems in automotive applications.

In Section II, we review related studies; Section III outlines the design of the proposed framework, and Section IV presents detailed results. We discuss and conclude in Section V.

2. Literature Review

The effectiveness of AVs relies on the strength of their object detection and classification systems. Recent developments utilise MATLAB's Vision and Automated Driving Toolboxes for detecting vehicles, pedestrians, and lanes. However, these algorithms encounter difficulties in complex road conditions, such as varying lane widths, intersections, low-light settings, and heavy traffic, emphasising the need for greater adaptability in real-world applications (Sharma et al., 2024). Advancements in object detection have resulted in more efficient YOLO-based systems, outperforming alternatives such as RetinaNet, Fast R-CNN, and SSD in terms of speed, accuracy, and learning capability (Alqarqaz et al., 2023). Figure 2 provides a visual representation of the comparative performance of these algorithms.

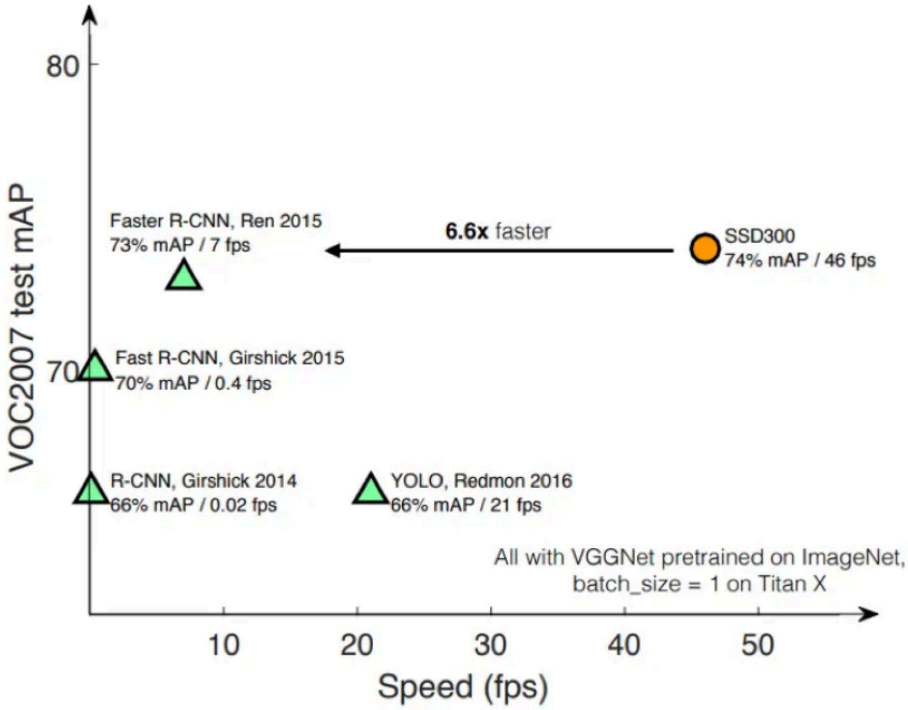


Figure 2. Performance Comparison: SSD vs. Faster R-CNN vs. YOLO. (Tahir et al., 2024))

To enhance detection accuracy while maintaining real-time performance, the authors in (Menaka et al., 2020) have explored a hybrid method that merges YOLO with Faster R-CNN. In this framework, YOLO quickly identifies potential object regions by drawing bounding boxes, and Faster R-CNN refines these results using its RoI pooling for accurate classification and segmentation.

Although this integration shows strong potential, it comes with high computational demands and its success largely hinges on having diverse and well-labelled training data. A recent study by authors (Sharma et al., 2024)published in 2024 unveiled the Canadian Vehicle Dataset paired with the YOLOv8 algorithm, which led to noticeable improvements in vehicle detection and classification across various weather scenarios.

However, despite this progress, there are still hurdles to overcome, particularly in meeting regulatory guidelines and enhancing recognition accuracy for object classes that appear less frequently. An extensive evaluation of one-stage and two-stage object detection techniques using the Waymo Open Dataset highlighted significant trade-offs between accuracy and computational performance. However, the study's reliance solely on 2D detection, without integrating data from additional sensors, underscores a critical gap in capturing the diverse environmental factors essential for the proper functioning of autonomous vehicles (Zhang et al., 2021).

Recent progress in XAI has led to the development of a hybrid framework that integrates LIME and SHAP for AV systems. This method combines SHAP's accuracy with the computational efficiency of LIME, ensuring both clarity and practicality. When tested on ResNet-18, ResNet-50, and SegNet-50 models using the KITTI dataset (Geiger et al., 2012), the framework delivered over 85% fidelity, 80% interpretability, and 70% consistency, surpassing traditional techniques. With inference times of 0.28s (ResNet-18), 0.571s (ResNet-50), and 3.889s (SegNet-50), it proves ideal for real-time AV applications. A short comparison of existing approaches with XAI is provided in Table 1 with research objectives and key limitations.

Prior studies in the field of explainable AI for autonomous systems have made significant contributions, yet they predominantly emphasise areas such as scene segmentation, gaze-based scene assessment, or anomaly detection. These investigations often fall short in terms of general model transparency, are tailored to limited use cases like VANETs, or overlook real-time tasks such as recognising brake light activation. Additionally, none have explored the integration of interpretability tools with real-time object detection models like YOLOv8 (Prashanth et al., 2024), especially within adaptive cruise control systems (Pirhonen et al., 2022; Yu & Wang, 2022) for semi-autonomous vehicles. This leaves a critical research gap, which this study seeks to address by focusing on an essential safety-related perception task.

However, interpretability methods like LIME and SHAP (Tahir et al., 2024), which perform well with CNN-based models, are less effective for YOLO-based architectures due to their unique detection pipeline and non-sequential feature extraction process, making them less suitable for real-time applications (Wang, Zhao & Cheng, 2019).

CAM is an explainability technique for convolutional neural networks that utilise global average pooling of feature maps. Certain CAM variants may incorporate additional components, such as a fully connected layer, depending on the specific method applied. One of the most prominent and widely used CAM-based approaches is Grad-CAM, introduced by Selvaraju et al. (2017), which leverages gradient information from a classification network to generate a heatmap, highlighting the areas that contribute most to the model's decision.

Chattopadhay et al. (2018) proposed Grad-CAM++, which is an advanced version of Grad-CAM that enhances localisation accuracy. Their method employs weighted positive partial derivatives from the final convolutional layer's feature maps to create more detailed heatmaps (Chattopadhay et al., 2018). Draelos and Carin (2020) , introduced HiResCAM to address the limitations of Grad-CAM (Chattopadhay et al., 2018). By refining the gradient averaging process and incorporating element-wise multiplication, HiResCAM generates heatmaps that more precisely reflect the model's internal computations, emphasising the most critical features for its predictions (Draelos & Carin, 2020). LayerCAM, introduced by Jiang et al. (2021), presents an alternative method for generating class activation maps by prioritising positive gradients from a specified layer. While it is computationally efficient, its ability to provide meaningful insights may differ across various classes, making it less informative in certain scenarios (Jiang et al., 2021).

EigenCAM takes a different approach from gradient-based techniques by employing a gradient-independent method that utilises Principal Component Analysis (PCA) on layer activations. This technique enhances stability and robustness by identifying the most significant features contributing to the final prediction. EigenGradCAM refines explainability further by merging the principles of Grad-CAM and EigenCAM (Muhammad & Yeasin, 2020). It applies PCA to the combination of activations and gradients from a given layer, ensuring that it retains class-specific information from gradients while reducing the impact of gradient saturation. Recent research has explored integrating heatmap techniques with the YOLO model to improve its interpretability. One such approach incorporates Grad-CAM into the YOLO framework, effectively enhancing the visualisation of detection decisions (Kirchknopf et al., 2022).

However, object detection models like YOLOv8 (Varghese & Sambath, 2024) do not store activation and gradient details in their outputs. The complexity of their final detection process makes it challenging to extract activations for each detected object separately. Although there have been substantial improvements in object detection and classification for semi- and autonomous vehicle systems, there are still notable gaps, especially in achieving real-time explainability.

Current systems often struggle to provide efficient onboard XAI while managing high computational costs. XAI techniques such as CAM have been used to improve transparency, but their application is mostly limited to binary classification tasks. Consequently, the use of real-time interpretability for specific tasks, such as brake prediction in YOLOv8-based systems, remains an underexplored area. Details are presented in Table 1.

Table 1. Various explainable AI approaches used in autonomous driving

Source	Research Objective	Key Limitations
(Sellat et al., 2022)	Proposed a method using FCN for road scene segmentation.	The approach lacks transparency and is confined to a single input type.
(Wang et al., 2021)	Used gaze-based cues to enhance scene comprehension in autonomous vehicles.	The model does not provide insights into its internal decision-making process or allow for *post hoc* explanations.
(Kothawade et al., 2021)	Integrated common-sense reasoning into explainable AI for autonomous driving systems.	The method faces limitations in scalability and requires significant computational resources.
(Mankodiya et al., 2021)	Developed an XAI-driven system for detecting anomalies within VANET (*Vehicular Ad-hoc Network*) environments.	The approach is limited to VANET datasets and lacks the ability to incorporate multiple data sources or provide universal XAI solutions.
(Sellat et al., 2022)	Applied advanced deep learning architectures to perform semantic segmentation.	The study does not incorporate explainable AI methods, focusing primarily on performance rather than model transparency.
(Mankodiya et al., 2022)	Combined XAI techniques with segmentation models for autonomous driving systems.	The model faces high computational demands and lacks adaptability to various use cases.
(Nazat et al., 2024)	Proposed a framework for anomaly detection in autonomous systems using XAI techniques.	The solution incurs significant computational overhead due to its complexity.

3. Proposed Framework

In this section, we outline the complete workflow of the proposed framework as shown in Figure 3, detailing each stage from data acquisition to evaluation of the trained model and explainability layer. Firstly, the process begins with data collection, where a diverse dataset is gathered to support robust model training. After this, data preprocessing steps such as image resizing and normalisation are applied. Following this, the model training phase employs YOLOv8 with transfer learning to achieve real-time and highly accurate detections. A performance analysis is conducted using key statistical metrics to evaluate the model's effectiveness. Furthermore, to enhance transparency and interpretability, XAI techniques are incorporated in the YOLOv8 architecture, allowing for visualisation of model decisions through heatmap generation. These visual insights provide a deeper understanding of how the model identifies key features, ensuring its reliability in practical applications.

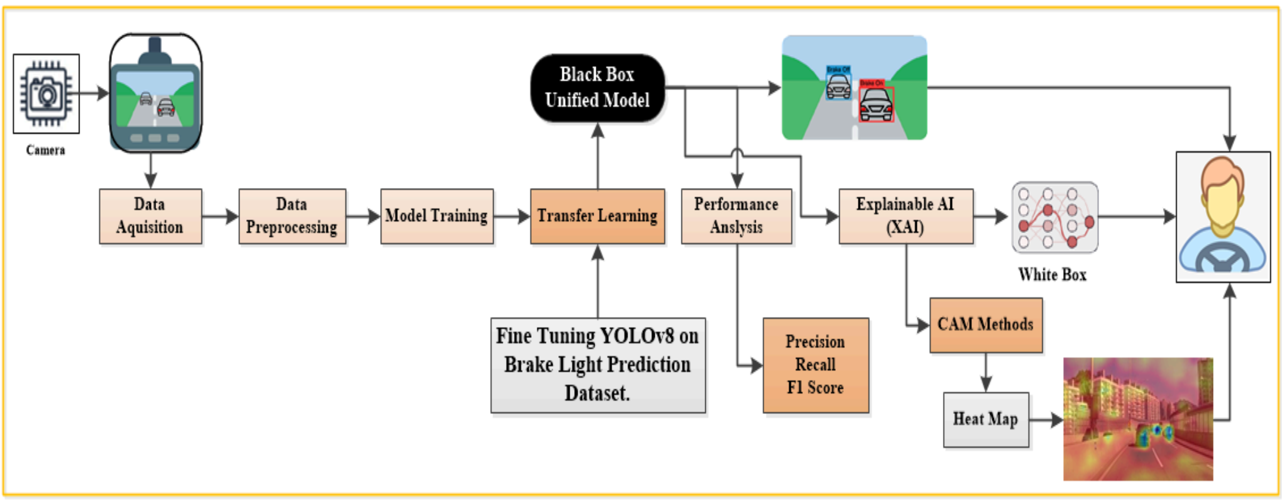


Figure 3. Overall Proposed Framework.

3.1. Data Acquisition

The dataset used in this study was obtained from (Univ, 2023) and contains over 16 hours of data collected from dashboard cameras during real-world driving. It comprises more than 11,000 images, captured at 5-second intervals. According to Oh and Lim (2023), each image was manually annotated to include bounding boxes around the vehicles and their corresponding brake light status, resulting in over 30,000 annotations in total. The dataset is publicly available from the authors (Oh & Lim, 2023) and carefully curated to ensure diversity, encompassing various driving conditions such as daytime, night-time, city, highway, and tunnel environments.

3.2. Data Preprocessing

In order to train our custom dataset with YOLOv8 and achieve optimal performance, several pre-processing steps were undertaken. The first steps involved resizing and normalising the images. To maintain uniformity in input size, all images were resized to a width of $I_{(w)}$ and a height of $I_{(h)}$, with both $I_{(w)}$ and $I_{(h)}$ set to 640. Min-max normalisation was then applied to adjust the pixel values to a standard range. Interested readers can get the details of the data preprocessing steps from (Oh & Lim, 2023).

3.3. Model Training

In this study, we applied transfer learning (Gupta, 2021; Zhuang et al., 2021) to train a model for brake light prediction, leveraging the full range of available YOLOv8 models (Ultralytics, n.d.). This approach enabled the development of a detection network capable of real-time inference in a moving vehicle, while ensuring high detection accuracy. The YOLOv8 framework includes five models of different sizes, from the compact YOLOv8n to the more expansive YOLOv8x, as detailed in Table 2.

Table 2. List of YOLOv8 deployed models

Model	Parameters	FLOPs
YOLOv8n	3.2	8.7
YOLOv8s	11.2	28.6
YOLOv8m	25.9	78.9
YOLOv8l	43.7	165.2
YOLOv8x	68.2	275.8

3.4. Transfer Learning

This study employed transfer learning (Zhuang et al., 2021), with the YOLOv8n model (Ultralytics, n.d.), the most compact and computationally efficient variant within the YOLOv8 architecture. Designed for deployment on edge devices, YOLOv8n is ideal for real-time applications such as those encountered in driving environments (Elmanaa et al., 2023). The YOLOv8 family comprises five models, ranging from the lightweight YOLOv8n to the more resource-intensive YOLOv8x. We selected YOLOv8n due to its optimal balance between inference speed and detection accuracy, making it particularly well-suited for resource-constrained scenarios. To adapt the model to our specific task, we performed fine-tuning (Gupta, 2021) by continuing the training on a custom dataset. This process enabled the model to refine its learned features for improved detection performance in our targeted driving scenarios.

After fine-tuning the YOLOv8n model using transfer learning, the trained network was then evaluated as a complete system to understand how it performs in real-world prediction tasks. While the model demonstrates strong detection accuracy and efficiency, its internal decision-making process remains opaque to users.

Therefore, the next section introduces the concept of the black-box model and highlights the necessity of explainability techniques to interpret the model's reasoning behind brake light predictions.

3.5. Black-Box Model

After training, the model operates as a black-box system, delivering one of two outputs: brake on or brake off. It is termed a black-box because it generates decisions without exposing the underlying logic or contributing factors to the driver, as shown in Figure2 where the result of the trained model is presented to the driver as brake status. Also, the results of the black-box model are presented in the second column of Table 5. Inside the model, multiple convolutional layers progressively analyse the input image, capturing basic patterns like edges and textures, and advancing to more complex visual elements such as object forms and spatial arrangements. These extracted features are then used to identify and distinguish brake light states. However, this internal decision-making pathway is not visible to the end-user. Despite delivering accurate classification results, the black-box nature (Samek et al., 2017) of the model lacks interpretability and poses challenges in terms of user trust, particularly in critical driving scenarios, underscoring the importance of incorporating XAI techniques (Gunning et al., 2019).

3.6. Performance Analysis

The model's performance was evaluated using key metrics such as Precision, Recall, F1 Score, and mAP50 to assess its accuracy and reliability. These metrics give a full picture of how well the model detects objects in different situations. To test how the model performs in real-time, we used a video dataset and shared the results on YouTube. The video illustrates the model's ability to detect objects accurately and its speed in a moving vehicle. The link to the video can be found at https://www.youtube.com/watch?v=JeTrzlh3BFs\&ab_channel=AmilDar.

3.7. Explainable AI

XAI is a critical area of research that seeks to address the inherent opacity of deep neural networks by improving the transparency, interpretability, and accountability of their decision-making processes. As highlighted by Gunning et al. (2019),, XAI is indispensable for perception systems in safety-critical applications, such as autonomous driving, where understanding the rationale behind AI-driven decisions is essential for user trust, regulatory compliance, and system validation.

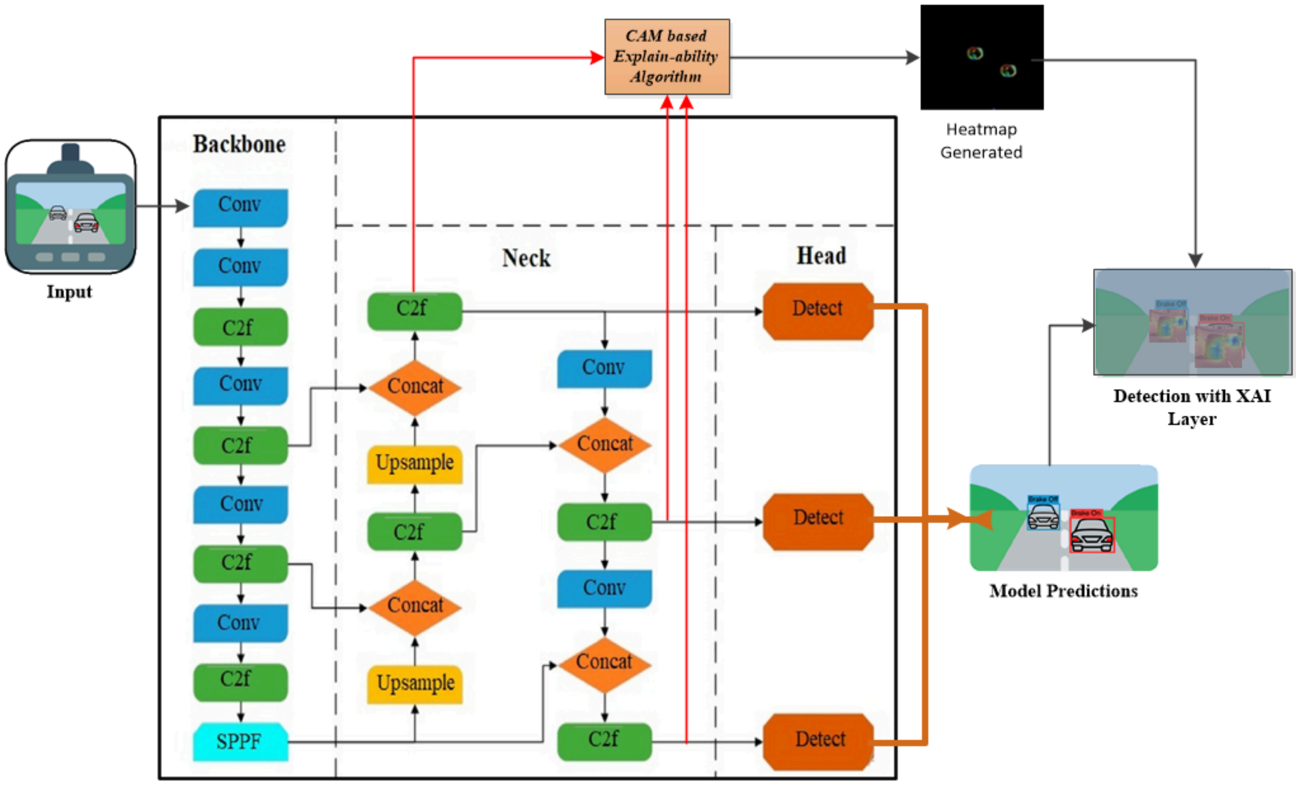


Figure 4. Integration of Explainability Method into YOLOv8 Architecture.

Traditional XAI methods such as LIME, SHAP, and CAM (Panati, Wagner, & Brüggenwirth, 2022) provide different strategies to interpret model predictions. However, their applicability varies depending on the task, architecture, and domain constraints. Specifically, CAM-based techniques offer a unique advantage for vision tasks by generating heatmaps that visually highlight the image regions highly influential in the model's predictions.

This visual interpretability is particularly valuable in scenarios such as brake light detection in autonomous vehicles, where understanding *why* a model has predicted a "brake on" or "brake off" is as crucial as the prediction itself. In this study, the explainability framework is integrated into the YOLOv8-based brake light prediction system, as illustrated in Figure 4.

3.9. Heatmap Generation

Hooks are employed to capture activations and gradients from intermediate layers of the YOLOv8 model, which are then passed to the CAM-based explainability algorithms. This algorithm generates a heatmap that is overlaid on the input image, highlighting the regions that most strongly influence the model's predictions. This integration provides an interpretable and visual explanation of the model's decision-making process, particularly in the context of safety-critical applications such as brake light detection in semi-autonomous vehicles.

Heatmaps are generated by analysing the activations and gradients extracted from intermediate layers of the model using the library referenced in (YOLOv8 Explainer, n.d.). These features are combined to localise the regions of the input image that contribute most significantly to the model's prediction. The resulting heatmap can then be presented to the driver or system developer for further analysis, improving transparency and interpretability by visually revealing the factors that influence the system's decision-making process (Mankodiya et al., 2021).

3.10. Proposed Algorithm

Our approach leverages the YOLOv8n object detection model to enhance interpretability through the use of activation heatmaps. Initially, the model detects vehicles within an input image, generating bounding boxes and corresponding confidence scores. To facilitate explainability, we extract activation maps from specific layers of the network using registered hooks. These feature maps are transformed into heatmaps and overlaid onto the original image to emphasise key regions that contribute to the detections. This strategy enhances the clarity of the detection process, making it more accessible and understandable for practical use cases. Here are the algorithm steps that depict the overall proposed framework:

- **Algorithm: YOLOv8-Based Object Detection and XAI Heatmaps**

1. Load the YOLOv8 model with pre-trained weights.
2. Read the input image and convert it from BGR to RGB.
3. Select the target layers for activation map extraction.
4. Register hook functions to capture feature maps from the selected layers.
5. Run YOLO inference on the input image and extract detection predictions.
6. Draw bounding boxes and annotate the detected objects on the image.
7. Generate activation heatmaps using the captured feature maps and save them.

4. Experiments and Results

The experiments were carried out using the settings presented in Table 3.

During model training, we used the same set of hyperparameters to ensure consistency across all models.

Table 3. Experimental Settings on Google Colab

Setting	Value
Python Version	3.11.11
PyTorch Version	2.6.0+cu124
CUDA Available	**12.4**
GPU Model	NVIDIA Tesla T4
GPU Memory	15,360 MiB
Driver Version	550.54.15
Power Usage	11W / 70W
Temperature	46°C
GPU Utilisation	0.6%
Libraries	All supporting libraries.

4.1. Model Evaluation

In Table 4, the result of the lighter version of YOLOv8 is shown. Figure 5 displays the training and validation loss curves for the trained black-box unified model. The plot demonstrates how the model's performance evolves over the course of training, highlighting the reduction in loss values across epochs. This reduction suggests that the model is increasingly adapting to the data, with both training and validation losses exhibiting a steady convergence trend.

Table 4. Trained model evaluation on the test set

Model	Class [ON/OFF]	Recall	Precision	mAP50	F1 Score
Yolov8n	OFF	0.670	0.726	0.761	0.697
	ON	0.853	0.549	0.772	0.668
	Total	**0.762**	**0.638**	**0.766**	**0.695**

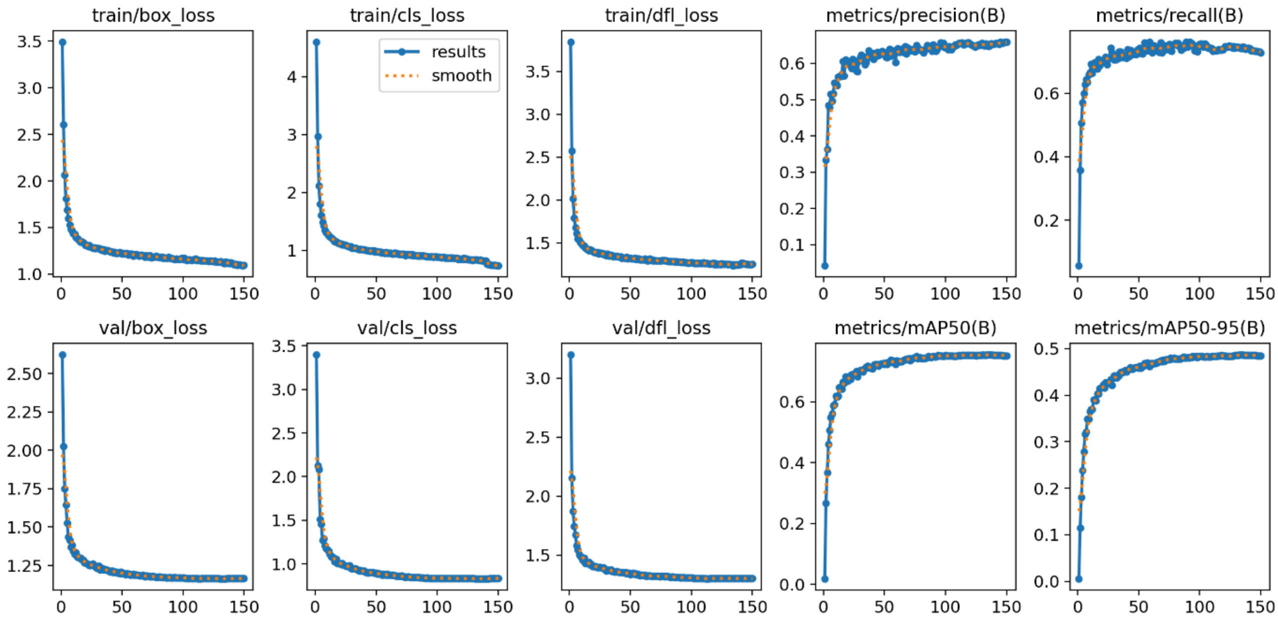


Figure 5. Training and Validation Graph.

Figure 6 illustrates the detection performance of the YOLOv8n-based brake light prediction model, highlighting class-wise results for Precision, Recall, mAP50, and F1Score. The *Brake- Light Off* class achieves higher precision (0.726), while the *Brake-Light On* class demonstrates stronger recall (0.853). The mAP50 scores are comparable for both classes, indicating consistent detection capabilities. The total scores across all classes are also reported, showing an overall balance between precision and recall. These results emphasise the model's ability to predict brake light status, with some trade-off evident between precision and recall depending on the class.

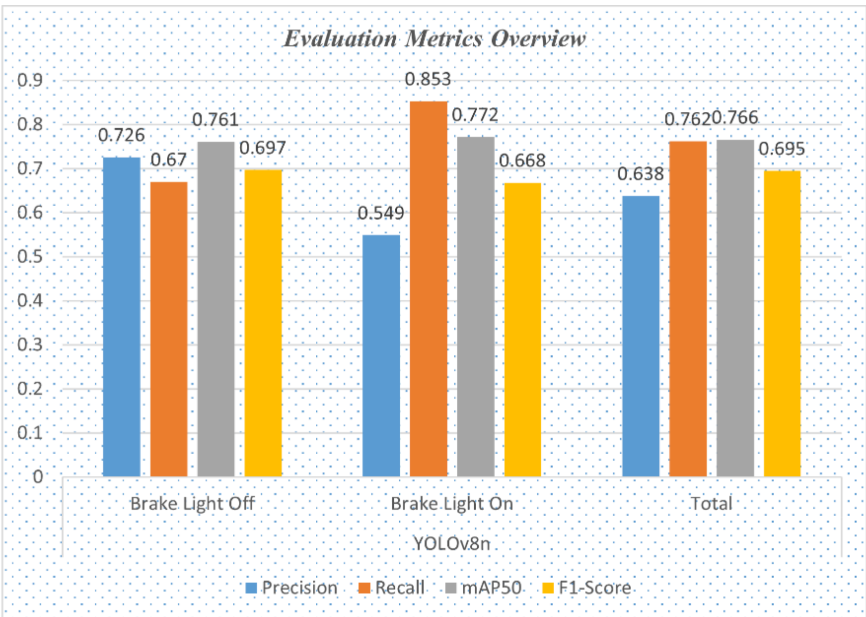


Figure 6. Detection performance of the YOLOv8n-based brake light prediction model.

4.2. Qualitative Analysis of Trained Black Box Model

A comprehensive qualitative analysis was conducted, examining not only variations in the number of vehicles and their distances, including both close-range and far-range scenarios, but also the performance across diverse driving environments such as urban roads, highways, tunnels, and mixed-traffic conditions. The evaluation further encompassed various vehicle types. This thorough assessment ensures the model's robustness in real-world scenarios, accounting for challenges such as occlusions, varying lighting conditions (e.g., tunnel darkness), and complex dynamic backgrounds. Figure 7a–c illustrates the qualitative performance of our model on single vehicles at varying distances and scales, demonstrating its capability to detect brake lights even when vehicles appear small or far in the frame. Additionally, Figure 7d highlights the model's effectiveness in a complex tunnel environment, where it successfully identifies multiple vehicles at varying distances, including those partially occluded or under challenging lighting conditions. These results underscore the robustness and generalisation of our model across diverse driving scenarios.

4.3. Results using Heatmap Generation

This study produced heatmaps for images from the test set and identified significant variations in feature visualisation among different methods. We integrated the trained YOLOv8n model with the library proposed by Sarma Borah et al. (2024) to visualise CAM results at different scales. The heatmap generated using the different CAM methods offers an explanation which could be compared to the detection output of our trained model, which is depicted by the bounding box. CAM techniques, including GradCAM, EigenCAM, GradCAMPlusPlus, XGradCAM , HiResCAM, LayerCAM, and EigenGradCAM, exhibit different performance across various scales, with their heatmaps showing activations confined within or outside the bounding box produced by the model. The examples are presented below for an in-depth analysis in Table 5.

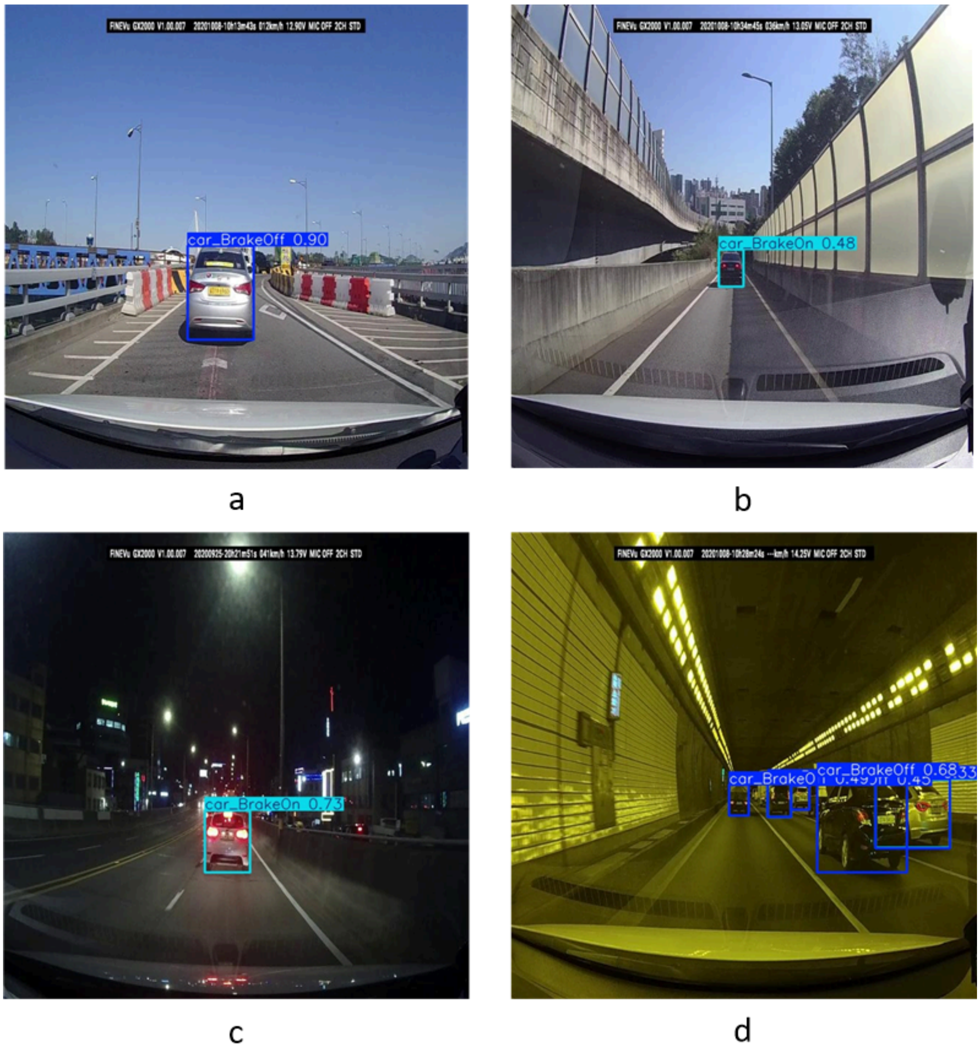
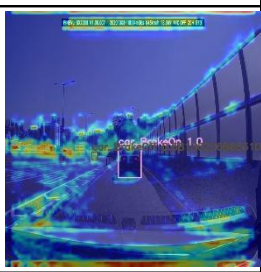
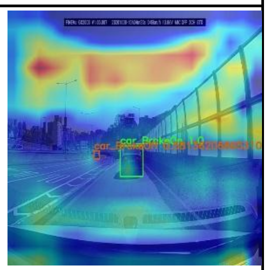
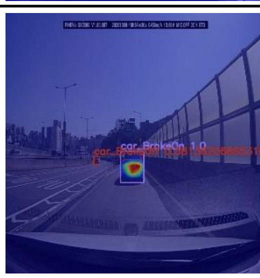
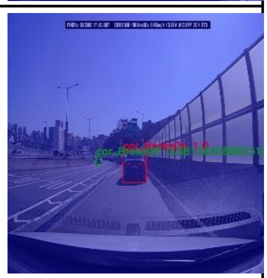


Figure 7. Qualitative analysis images of driving vehicles and brake light status detection.

Our evaluation of various explainability methods on YOLOv8 revealed that HiResCAM and LayerCAM produced heatmaps that were well-contained within the object bounding box at layers 15 and 18. This indicates that these methods effectively capture class-discriminative features localised to the detected object, making them more interpretable. In contrast, other methods such as GradCAM, EigenCAM, GradCAM++, and XGradCAM generated less distinct activations, often extending beyond the bounding box, or failing to provide clear insights, especially at layers 18 and 21. This suggests that HiResCAM and LayerCAM may be more reliable for fine-grained object feature attribution in object detection tasks.

Table 5. Comparison of model performance across different targeted layers

Model	Trained Model Result	Layer 15	Layer 18	Layer 21
EigenCAM				
EigenGradCAM				
HiResCAM				
LayerCAM				

4.4. Explainable AI Results for Brake Light Detection Using Combined Multi-Scale Features from the Final Three Layer

From Table 5 analyses, we found that HiResCAM visualisations consistently align with the model's focus areas. In this section, we present the combined multi-layer results. We employed a YOLOv8-based object detection model, specifically fine-tuned for brake light status prediction. The model was trained to classify two distinct classes: *car_BrakeOn* and *car_BrakeOff*. To enhance interpretability and provide transparency in the model's decision-making processes, we integrated explainability techniques using HiResCAM (Draelos & Carin, 2020), a state-of-the-art CAM method known for generating high-resolution, fine-grained saliency maps. The explainability analysis targeted key convolutional layers (layers 15, 18, and 21), systematically selected based on their spatial resolution and semantic richness within the model's architecture. By aggregating the activation maps from these layers, as shown in Figure 8, we generated composite heatmaps that effectively highlighted the critical regions influencing the model's predictions.

The HiResCAM-generated visualisations consistently demonstrated that the model's attention was predominantly focused within the regions bounded by the detection boxes corresponding to vehicles, particularly emphasising the taillights and rear areas where brake lights are located. This alignment with human visual reasoning is crucial, as brake light status is a localised visual cue critical for safe driving decisions. The results validate the model's reasoning process by showing its focus on contextually relevant areas, thereby bridging the gap between opaque AI systems and human-interpretable insights. From a safety perspective, this transparency fosters driver trust, as it assures users that the system is attending to the correct features—specifically the regions where brake lights appear. Integrating explainable AI within this framework not only enhances interpretability but also supports user confidence in the system's ability to perform reliable real-time brake-light prediction tasks in intelligent transportation scenarios.

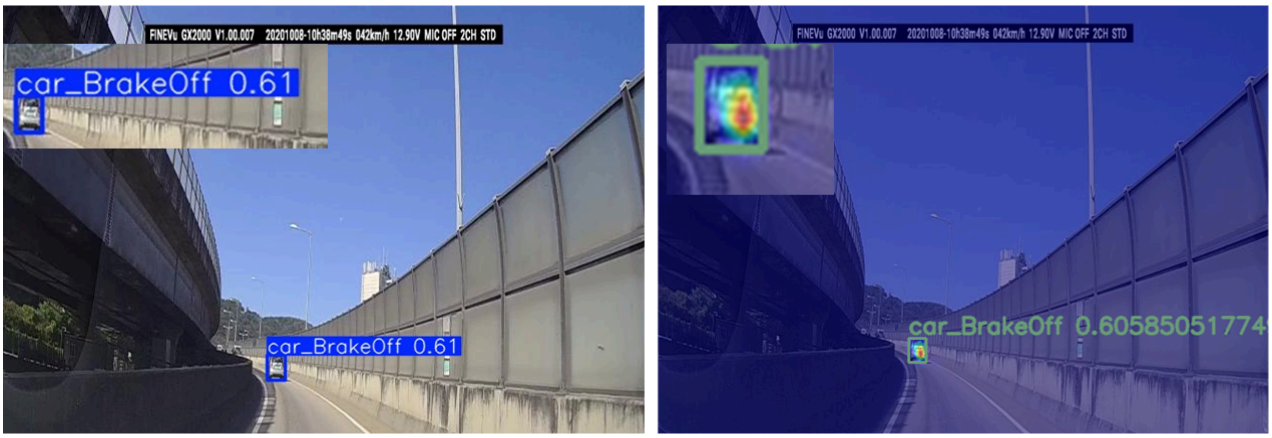


Figure 8. HiResCAM-based combined heatmap visualisation for layers 15, 18, and 21, focusing on brake light regions inside bounding boxes

5. Discussion

This study introduces a novel framework that integrates real-time XAI techniques with brake-light prediction in YOLOv8-based object detection systems, leveraging advanced CAM methods. To the best of our knowledge, this is the first work to apply CAM-based visual explanations for brake-status detection in the context of semi-automated vehicles (SAE Levels 2 and 3) and autonomous driving systems. The proposed approach establishes a new benchmark for interpretability and transparency in real-time AI decision-making for safety-critical applications.

The existing literature on XAI in autonomous driving, summarised in Table 1, demonstrates various approaches and applications of explainability techniques, including segmentation, anomaly detection, and gaze-based attention. While these works contribute valuable insights into the field, they primarily focus on tasks such as segmentation or anomaly detection, often in specific network environments or under limited conditions. Moreover, prior studies frequently face challenges such as high computational demands, limited scalability, lack of transparency in model outputs, or restricted adaptability across different scenarios. Crucially, none of these works addresses the specific problem of brake light prediction using XAI techniques, an essential yet underexplored aspect of road safety in semi-automated vehicles. As a result, direct performance comparisons with existing methods are not feasible, underscoring the novelty of this study as it bridges this critical research gap.

Beyond prediction, the proposed framework facilitates system diagnostics and model validation by providing multi-scale visualisations of learned features, supporting fine-grained analysis of detections, and enhancing the model's adaptability for deployment in diverse environments. The integration of XAI not only empowers technical experts to interpret model outputs and diagnose failures but also offers a level of transparency that is accessible to non-expert users. This transparency is essential for fostering user trust in AI-powered perception systems, aligning with regulatory requirements for model accountability and safety assurance. By connecting deep learning-based models to tangible explanations, this study strengthens the link between AI research and its practical implementation, promoting fairness, reliability, and transparency in AI-driven transportation systems.

Ultimately, this work paves the way for further research into the application of XAI techniques in complex, real-world scenarios within intelligent transportation systems. It lays the foundation for developing more transparent, accountable, and trustworthy perception systems that can be effectively deployed in safety-critical domains such as autonomous driving.

Limitations

A limitation of our proposed work is that the proposed framework has been tested within the Google Colab environment rather than in real-time driving scenarios. To overcome this limitation, we will use the EMO (Riaz et al., 2018) to conduct real-time experiments to enhance driver and passenger trust in autonomous systems.

Another limitation of our proposed model is the unavailability of datasets with bright light intensity in videos for testing. This lack of data may affect the model's ability to accurately predict in environments with high brightness. Future work should focus on collecting and incorporating such datasets to enhance the model's robustness and generalisability.

6. Conclusion

This study tackles the critical challenge of explainability in deep learning-based object detectors, with a focus on the YOLOv8 model applied to brake-light detection tasks. Leveraging transfer learning, a pre-trained YOLOv8 model was fine-tuned on a custom dataset to predict brake-light status-*car_BrakeOn* and *car_BrakeOff*. To enhance transparency and user trust, we integrated multiple attribution-based explainability techniques, including Grad-CAM, XGrad-CAM, Eigen-CAM, HiResCAM, LayerCAM, RandomCAM, and EigenGrad-CAM, within the YOLOv8 framework. These methods provide visual insights into the spatial regions that contribute most significantly to the model's decisions. Among them, HiResCAM demonstrated high-fidelity, fine-grained heatmaps that aligned well with critical areas such as taillights, ensuring interpretability in safety-critical scenarios. The proposed explainability module empowers non-expert users to independently assess model predictions, thereby facilitating transparency, regulatory compliance, and enhanced confidence in AI-driven perception systems. Overall, this work lays a foundation for advancing explainable AI in autonomous and semi-autonomous vehicle applications, promoting trust and interpretability in real-time decision-making tasks.

Future work

- We will conduct experiments using the proposed framework in collaboration with the Control, Automotive, and Robotics Laboratory, an affiliate of the National Centre of Robotics and Automation, on Pakistan's first autonomous vehicle (Riaz et al., 2018).
- We plan to deploy the trained model in ACC vehicles to assist drivers in real-time decision-making.

References

Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics, 18*(8), 5031–5042. https://doi.org/10.1109/TII.2022.3146552

Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics, 102*, 502–520. https://doi.org/https://doi.org/10.1016/j.cag.2021.09.002

Alqarqaz, M., Bani Younes, M., & Qaddoura, R. (2023). An object classification approach for autonomous vehicles using machine learning techniques. *World Electric Vehicle Journal, 14*(2), Article 41. https://doi.org/10.3390/wevj14020041

Arrieta, A. B., Rodríguez, N. D., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2018)* (pp. 839–847). IEEE. https://doi.org/10.1109/WACV.2018.00097

Concerning rate of road trauma continues in first half of 2024 – RAA Daily. (n.d.). Retrieved June 4, 2025, from https://daily.raa.com.au/media-resources/concerning-rate-of-road-trauma-continues-in-first-half-of-2024/

Draelos, R. L., & Carin, L. (2020). *Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks*. https://api.semanticscholar.org/CorpusID:244478775

Elmanaa, I., Sabri, M. A., Abouch, Y., & Aarab, A. (2023). Efficient roundabout supervision: Real-time vehicle detection and tracking on Nvidia Jetson Nano. *Applied Sciences, 13*(13), Article 7754. https://doi.org/10.3390/app13137416

Ennajar, A., Khouja, N., Boutteau, R., & Tlili, F. (2021). Deep multi-modal object detection for autonomous driving. In *Proceedings of the 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 7–11). IEEE. https://doi.org/10.1109/SSD52085.2021.9429355

Ultralytics. (n.d.). *Explore Ultralytics YOLOv8 – Ultralytics YOLO Docs.* Retrieved June 5, 2025, from https://docs.ultralytics.com/models/yolov8/

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3354–3361). IEEE. https://doi.org/10.1109/CVPR.2012.6248074

Global status report on road safety 2023. (2023).

Grigorescu, S. M., Trasnea, B., Cocias, T. T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics, 37*(3), 362–386. https://doi.org/10.1002/ROB.21918

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), Article eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Gupta, N. (2021). A pre-trained vs fine-tuning methodology in transfer learning. *Journal of Physics: Conference Series, 1947*(1), 012028. https://doi.org/10.1088/1742-6596/1947/1/012028

Hayashi, Y. (2020). Black box nature of deep learning for digital pathology: Beyond quantitative to qualitative algorithmic performances. In *AI and ML for digital pathology* (pp. 95–101). Springer. https://doi.org/10.1007/978-3-030-50402-1_6

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., & Wei, Y. (2021). LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing, 30*, 5875–5888. https://doi.org/10.1109/TIP.2021.3089943

Khalil, R. A., Safelnasr, Z., Yemane, N., Kedir, M., Shafiqurrahman, A., & Saeed, N. (2024). Advanced learning technologies for intelligent transportation systems: Prospects and challenges. *IEEE Open Journal of Vehicular Technology, 5*, 397–427. https://doi.org/10.1109/OJVT.2024.3369691

Kirchknopf, A., Slijepcevic, D., Wunderlich, I., Breiter, M., Traxler, J., & Zeppelzauer, M. (2022). *Explaining YOLO: Leveraging Grad-CAM to explain object detections. arXiv preprint arXiv:2211.12108.* https://doi.org/10.48550/arXiv.2211.12108

Kothawade, S., Khandelwal, V., Basu, K., Wang, H., & Gupta, G. (2021). *AUTO-DISCERN: Autonomous driving using common sense reasoning. arXiv preprint arXiv:2110.13606.* https://api.semanticscholar.org/CorpusID:239885415

Mankodiya, H., Obaidat, M. S., Gupta, R., & Tanwar, S. (2021). XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles. In *Proceedings of the 2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)* (pp. 1–5). IEEE. https://doi.org/10.1109/CCCI52664.2021.9583190

Mankodiya, H., Jadav, D., Gupta, R., Tanwar, S., Hong, W.-C., & Sharma, R. (2022). OD-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences, 12*(11), Article 5310. https://doi.org/10.3390/app12115310

Menaka, R., Archana, N., Dhanagopal, R., & Ramesh, R. (2020). Enhanced missing object detection system using YOLO. In *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1407–1411). IEEE. https://doi.org/10.1109/ICACCS48705.2020.9074278

Muhammad, M. B., & Yeasin, M. (2020). Eigen-CAM: Class activation map using principal components. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206626

Nazat, S., Li, L., & Abdallah, M. (2024). XAI-ADS: An explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems. *IEEE Access, 12*, 48583–48607. https://doi.org/10.1109/ACCESS.2024.3383431

Oh, G., & Lim, S. (2023). One-stage brake light status detection based on YOLOv8. *Sensors, 23*(17), Article 7436. https://doi.org/10.3390/s23177436

Panati, C., Wagner, S., & Brüggenwirth, S. (2022). Feature relevance evaluation using Grad-CAM, LIME, and SHAP for deep learning SAR data classification. In *Proceedings of the 2022 23rd International Radar Symposium (IRS)* (pp. 457–462). IEEE. https://doi.org/10.23919/IRS54158.2022.9904989

Pirhonen, J., Ojala, R., Kivekäs, K., Vepsäläinen, J., & Tammi, K. (2022). Brake light detection algorithm for predictive braking. *Applied Sciences, 12*(6), Article 2804. https://doi.org/10.3390/app12062804

Prashanth, S., Leti, B. K., Jenis, M. R. A., & Chandrasekar, A. (2024). Enhancing multi-object detection systems: Cutting-edge techniques, implementation strategies, and performance optimization. In *Proceedings of the 2024 International Conference on Innovative*

Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1–6). IEEE. https://doi.org/10.1109/ICSES63760.2024.10910779

Riaz, F., Jabbar, S., Sajid, M., Ahmad, M., Naseer, K., & Ali, N. (2018). A collision avoidance scheme for autonomous vehicles inspired by human social norms. *Computers and Electrical Engineering, 69*, 690–704. https://doi.org/10.1016/j.compeleceng.2018.02.011

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *ITU Journal: ICT Discoveries*, *1*(1), 1–10. https://doi.org/10.48550/arXiv.1708.08296

Sarma Borah, P. P., Kashyap, D., Laskar, R. A., & Sarmah, A. J. (2024). A comprehensive study on explainable AI using YOLO and post hoc method on medical diagnosis. *Journal of Physics: Conference Series, 2919*(1), 012045. https://doi.org/10.1088/1742-6596/2919/1/012045

Sellat, Q., Bisoy, S., Priyadarshini, R., Vidyarthi, A., Kautish, S., & Barik, R. K. (2022). Intelligent semantic segmentation for self-driving vehicles using deep learning. *Computational Intelligence and Neuroscience, 2022*(1), Article 6390260. https://doi.org/https://doi.org/10.1155/2022/6390260

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Sharma, T., Chehri, A., Fofana, I., Jadhav, S., Khare, S., Debaque, B., Duclos-Hindie, N., & Arya, D. (2024). Deep learning-based object detection and classification for autonomous vehicles in different weather scenarios of Quebec, Canada. *IEEE Access, 12*, 13648–13662. https://doi.org/10.1109/ACCESS.2024.3354076

Tahir, H. A., Alayed, W., Hassan, W. U., & Haider, A. (2024). A novel hybrid XAI solution for autonomous vehicles: Real-time interpretability through LIME–SHAP integration. *Sensors, 24*(21), Article 6776. https://doi.org/10.3390/s24216776

Univ, I. K. (2023). *Brake-light-detection dataset.* In *Roboflow Universe*. Roboflow.

Varghese, R., & Sambath, M. (2024). YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ADICS58448.2024.10533619

Wang, B., Zhao, D., & Cheng, J. (2019). Adaptive cruise control via adaptive dynamic programming with experience replay. *Soft Computing, 23*(12), 4131–4144. https://doi.org/10.1007/S00500-018-3063-7

Wang, C., Weisswange, T. H., Krüger, M., & Wiebel-Herboth, C. B. (2021). Human–vehicle cooperation on prediction level: Enhancing automated driving with human foresight. In *Proceedings of the 2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)* (pp. 25–30). IEEE. https://doi.org/10.1109/IVWorkshops54471.2021.9669247

YOLOv8 Explainer. (n.d.). *YOLOv8 explainer.* Retrieved June 10, 2025, from https://spritan.github.io/YOLOv8_Explainer/

Yu, L., & Wang, R. (2022). Research on adaptive cruise control systems: A state-of-the-art review. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 236*(2–3), 211–240. https://doi.org/10.1177/09544070211019254

Zhang, Y., Song, X., Bai, B., Xing, T., Liu, C., Gao, X., Wang, Z., Wen, Y., Liao, H., Zhang, G., & Xu, P. (2021). *2nd place solution for Waymo Open Dataset Challenge – real-time 2D object detection* [Preprint]. *arXiv*. https://arxiv.org/abs/2106.08713

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2921–2929). IEEE. https://doi.org/10.1109/CVPR.2016.319

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555