

A Robust and Fair Object Detection and Image Segmentation Framework for Autonomous Driving Under Adverse Weather Conditions Using Explainable AI

Lekshmi Kalinathan^{1,*}, K. Anusha², K. Devi³, K. Divya⁴

^{1,2,3,4}School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.
lekshmi.k@vit.ac.in¹, kothapalli.anusha2022@vitstudent.ac.in², devi.k@vit.ac.in³, karbari.divya2022@vitstudent.ac.in⁴

*Corresponding author

Abstract: Reliable object detection continues to be the foundation of the safety features for self-driving cars. One of the biggest challenges, though, is that different weather phenomena, such as fog, snow, and darkness, not only affect the driver but also the on-board sensors and, therefore, the performance of perception systems in terms of object detection and recognition. This paper addresses the research gap in combining detection accuracy, robustness to various adverse situations, and explainability for autonomous driving object detection tasks. The Faster R-CNN backbone ResNet-50 has been tested on corrupted datasets: COCO-C, Pascal-C and Cityscapes-C, using the image segmentation techniques K-means, GrabCut, SLIC, Watershed and U-Net. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been employed for model interpretability purposes, and fairness has been addressed through two fairness metrics, namely Demographic Parity and Equalized Odds, to highlight and eventually remove prediction bias. The experiments revealed detection accuracies of 97.2% in clear conditions and 89.0% under adverse conditions, with U-Net performing best at 90.0% even in the most challenging scenarios. The combined use of explainable AI techniques and fairness assessment can improve model transparency, bias detection, and overall system reliability. Overall, this work provides a framework for trustworthy, robust, and equitable computer perception systems for self-driving cars operating in real-world environments.

Keywords: Explainable Ai (Xai); Adverse Weather; Object Detection; Image Segmentation; Bias Mitigation; Autonomous Cars; Fairness Metrics; Shapley Additive Explanations.

Cite as: L. Kalinathan, K. Anusha, K. Devi, and K. Divya, “A Robust and Fair Object Detection and Image Segmentation Framework for Autonomous Driving Under Adverse Weather Conditions Using Explainable AI,” *AVE Trends in Intelligent Computer Letters*, vol. 2, no. 2, pp. 111–121, 2026.

Journal Homepage: <https://avepubs.com/user/journals/details/ATICL>

Received on: 09/02/2025, **Revised on:** 08/04/2025, **Accepted on:** 17/06/2025, **Published on:** 05/06/2026

DOI: <https://doi.org/10.64091/ATICL.2026.000329>

1. Introduction

Fast progress in self-driving car technology has resulted in a greater availability of transport options that are far safer and more efficient than before. One of the most important skills of autonomous cars is accurately detecting and identifying surrounding objects, including other cars, pedestrians, and traffic signs [1]. Detection systems must be highly accurate and fast to enable safe navigation in dynamic, crowded environments [2]. On the other hand, current object detection algorithms often fail in conditions such as fog, frost, and snow, which significantly reduce visibility [16]. Visibility reduction due to such conditions often reduces prediction accuracy, leading to safety risks; therefore, reliable detection under harsh weather conditions is a must [3]. Besides technical efficiency, the concept of fair AI is of great concern [4]. Object detection systems should perform the same regardless of demographic group or location [5]. The use of biased training data may result in a performance mismatch,

Copyright © 2026 L. Kalinathan *et al.*, licensed to AVE Trends Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

leading to unequal safety outcomes across groups or locations [6]. Pretending an algorithm works less well at detecting objects in certain urban or rural areas may create a gap between safety outcomes, triggering legal and ethical issues [7].

The study addresses both robustness and fairness issues by improving object detection systems using diverse datasets and techniques [8]. To test the resilience of our models under adverse weather conditions, researchers have used distorted versions of mainstream datasets such as COCO-C, Pascal-C, and Cityscapes-C [9]. Additionally, the California Housing dataset was added to provide information on socioeconomic status and demographics [10]. This should enable us to determine whether a model's performance varies across regions or population groups [11]. The MNIST dataset shall be employed to detect objects at the simplest level [12]. Given the challenging conditions, researchers apply a set of five image segmentation techniques to improve object detection: K-means Clustering, GrabCut, Superpixel Segmentation (SLIC), Watershed Segmentation, and U-Net (a deep learning-based approach) [13]. Besides, object segmentation using segmentation methods helps enhance object recognition, especially in complex situations where traditional detection algorithms struggle [14]. In addition, fairness metrics such as Demographic Parity, Equalized Odds, and Counterfactual Fairness are used to monitor and eliminate biases in the model's predictions, ensuring fair and consistent performance across different user populations [15].

2. Related Work

The development of self-driving car technology has become a major concern for many people. Many recent studies have stressed that more accurate object detection systems are needed if researchers and the car itself are to be truly safe in the future. The paper by Masmoudi et al. [1] notes that the demand for a more reliable method for object detection has become even more evident as the use of autonomous vehicles has increased drastically. In the study by Sarda et al. [2], it is noted that the YOLO algorithm is a very capable and fast object detector, achieving a detection speed of about 155 frames per second. Since this kind of fast object detection is required in most driving situations, it is a great fit for such scenarios. Moreover, the paper by Xiao et al. [3] takes a closer look at recent improvements in YOLOv5, in particular the increased ability of this detector to adapt to a wider range of environmental conditions, which is one of the factors that ensure the reliability of detection for autonomous driving systems. Fairness as part of bias mitigation is explored in the study by Shrestha et al. [4], which explains how interactive methods enable daily model adjustments to prevent discriminatory effects and promote a fair AI environment. Similar importance and necessity for mitigating bias are expressed by Stevens et al. [5], which investigates how biases in deep learning models affect the visibility of minority and underrepresented groups, while calling for proper data handling and cleaning. The paper by Constantin et al. [6] offers an overview of fairness metrics in AI healthcare applications, emphasizing the role of diverse data in developing equitable and just AI outcomes.

Algorithmic fairness metrics compared with human judgment is the theme of the paper by Li et al. [7], which calls for fairness evaluators not only to base assessments on numbers but also on qualitative aspects of human beings reflected in the results of these algorithmic measures. The work by Banitalebi-Dehkordi et al. [8] examines how demographics influence fairness in self-driving systems and how these biases shape public perceptions of safety. The research paper proposes curriculum learning as a powerful approach to enhance detection robustness under adverse weather conditions, thereby improving adaptability to various types of extreme weather [9]. Using a new technique, Grislin-Le Strugeon et al. [10] explain how AI can be made more transparent by reducing computational overhead and revealing the models' decision-making process. This will lead to greater user trust. In Russo et al. [11], ethical considerations are addressed, and the study promotes a culture of incorporating fairness and accountability principles in AI to build public confidence. The paper by Kuroki and Yamasaki [12] discusses the use of synthetic data to combat bias. It describes how generative adversarial networks (GANs) can be employed to achieve fairer, more efficient AI systems. It is concerned with communication fairness in autonomous car systems that examine fair approaches to communication during adverse conditions such as fog, a well-known problem in maintaining constant communication [13]. Switching to e-commerce, Dominguez-Catena et al. [14] show that the CIAM is responsible for safeguarding customer identities and removing biases, an essential aspect of making digital platforms more user-friendly. Managing book inventory in an online store is explained and used to demonstrate how real-time data can improve product availability and customer satisfaction [15].

In e-commerce, automated replenishment is discussed) Yang et al. [16] help maintain operational efficiency and inventory optimality. Customer satisfaction in bookstores is directly related to safety stock management, according to the research. Recognizing and managing biases in model performance has been the main finding of the paper, and the authors have therefore shown the essential role of inclusivity in autonomous vehicle features that help make road users, including all demographic groups, safer. In addition, this paper identifies key ethical considerations. It aims to develop trustworthy, impartial detection models that can be used by diverse user groups without being discriminatory and that ensure user safety under various conditions. On the other hand, the research of Sarda et al. [2] shows that a cooperative approach to agent interaction can effectively regulate traffic flow. At the same time, lowering the levels of congestion, it is a work example of the use of sophisticated ML methods that may serve traffic regulation systems more dependably and make them more efficient, and that can also give guidance for methodology for a system of object detection in real-time environments, especially on days when

visibility is low. In addition, as discussed in Xiao et al. [3], the implementation of smart control systems in 5G networks is of paramount importance to the efficiency and reliability of communication performance, which are necessary for applications demanding low latency, such as real-time object detection in autonomous vehicles.

3. Methodology

The proposed framework aims to develop an efficient and fair object detector that delivers stable performance under challenging environmental conditions. The detection framework used in this paper is the Faster R-CNN model with the ResNet-50 network, which is trained on the COCO-C, Pascal-C, and Cityscapes-C datasets. To better generalize across diverse environmental conditions, additional distortions such as blurring, noise, and reduced contrast are applied using data augmentation techniques. In order to further enhance object localization and demarcation, the system uses image segmentation algorithms such as K-means clustering, GrabCut, and U-Net. The model's interpretability is improved through the use of XAI methods, such as SHAP and LIME, which provide a better understanding of the model's predictive behavior and potential biases across different demographic segments. The fairness of the model is analyzed using well-known metrics, such as Demographic Parity and Equalized Odds, that ensure balanced performance across different population segments. The model is tested for fairness before deployment to ensure adherence to ethical AI principles.

3.1. Robust Object Detection Model Development

The model's accuracy in detecting objects in various adverse conditions is a major benefit for autonomous driving. It can be reliably used in such an environment (Figure 1).

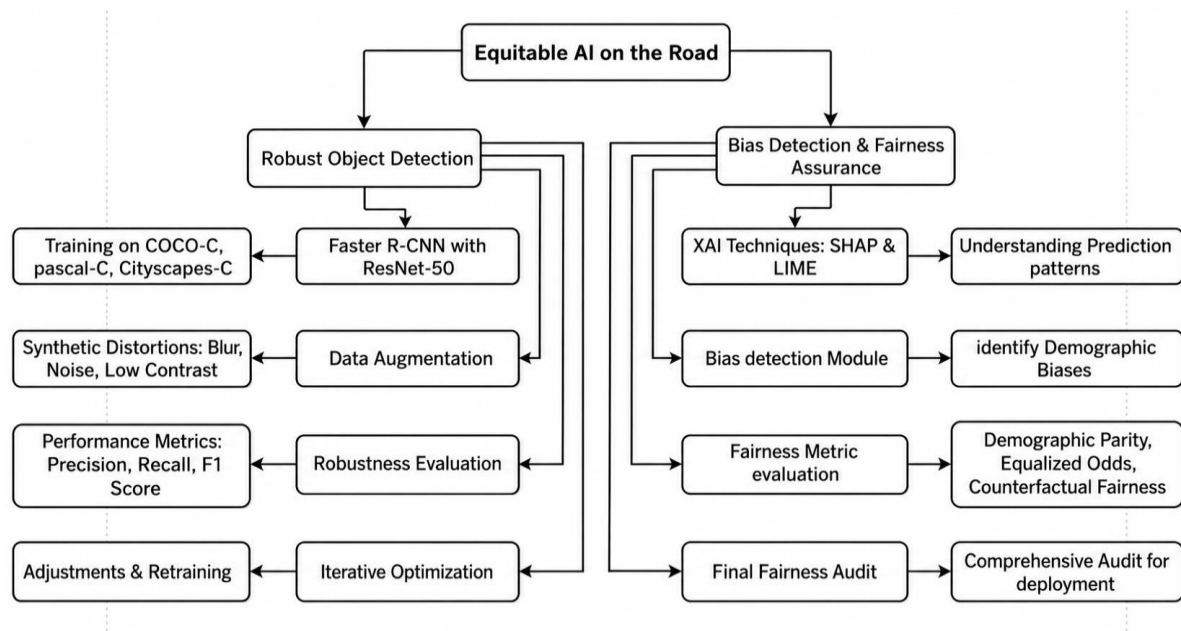


Figure 1: Methodology overview

Object Detection Model Selection and Training: R-CNN with the ResNet-50 backbone was chosen for its accurate detection and multi-object detection capabilities. This was trained on diverse datasets, including COCO-C, Pascal-C, and Cityscapes-C, making it capable of performing well in challenging environmental conditions such as fog, snow, and rain.

Data Augmentation for Enhanced Robustness: To enhance the robustness of the model across different types of operational scenarios, artificial degradations in terms of image blurriness, the presence of noise, and lower contrast were employed during the training process. This helped simulate changes in weather and lighting, increasing the diversity of the training dataset. Thus, through this process, the model was made to encounter more varied and realistic scenarios.

Figure 2 below shows line graphs illustrating how the performance level (measured by mAP50, or mean Average Precision at 50% Intersection over Union threshold) of a model changes when different types of image corruption are introduced. The graph plots each type of corruption - Gaussian noise, motion blur, snow, or JPEG compression-and traces the decline of the model's performance as the corruption level increases from 0 to 5 on the x-axis. The y-axis shows the percentage of mAP50, with higher

values indicating better performance. Figure 3 demonstrates the three data sets that were used in this experiment: standard data set (black squares), stylized data set (yellow triangles), and combined data set (red circles). In most scenarios, an increase in corruption severity results in a drop in the model's performance across all datasets, though the magnitude of the drop varies by corruption type.

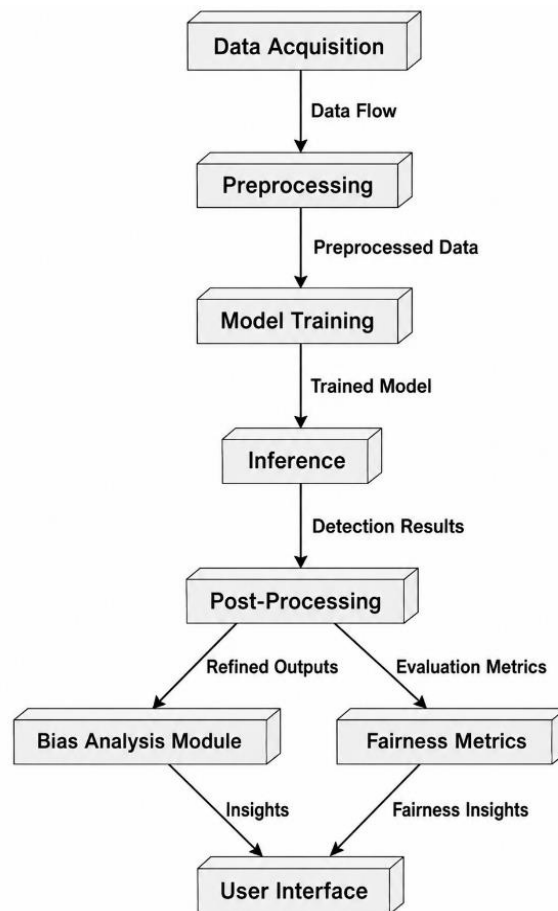


Figure 2: Adverse condition resilient detection architecture

A typical corruption example is when a model performs reasonably well (almost unchanged) even with 100% corruption of brightness or fog. However, methods such as glass blur, Gaussian noise, and motion blur lead to a rapid decline in performance as corruption becomes more severe. The combined dataset (red circles) mostly outperforms the standard datasets and stylized variants, indicating that combining data in the training set likely increases the model's robustness against corrupted images. The investigation provides evidence of the model's robustness against various corruption cases, thereby underscoring the need to handle diverse data conditions:

- **Evaluation of Robustness in Diverse Conditions:** This phase tests the model across simulated conditions on COCO-C, Pascal-C, and Cityscapes-C images, then evaluates its performance in terms of precision, recall, and F1 scores to assess its robustness to environmental changes. In other words, this part shows how the detection performance remains the same across different environments.
- **Iterative Optimization for Model Enhancement:** Based on performance testing data, the model is gradually refined through iterative refinements and retraining using synthetic data and data augmentation. In doing so, the detection accuracy for adverse conditions gradually increases. This process is repeated so that the model continues to evolve and improve on new and difficult tasks.

In Figure 4 above, model performance, as measured by Mean Average Precision (mAP), decreases as data corruption increases. Different levels of corruption have different effects on data performance across different groups. The x-axis shows the severity of data corruption, ranging from 0 (no corruption) to 5 (full corruption). The y-axis shows performance as mAP percentages, reflecting overall performance.

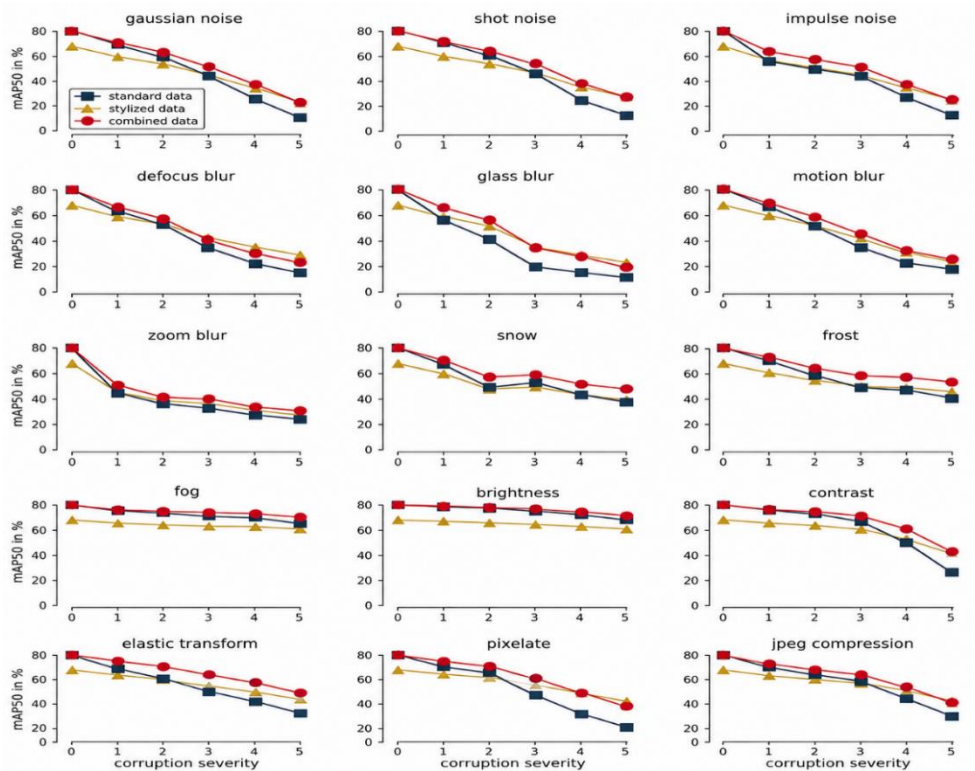


Figure 3: Impact of image corruption on model performance across diverse datasets

Figure 4 shows that group3 (green) is the best-performing group across corruption severity levels, whereas group2 (orange) is the most affected, with the largest drop in performance. The shaded area around each plot line represents a confidence interval, indicating the variance expected if the models were re-trained. The spread of these intervals also indicates that performance variability increases as data corruption worsens.

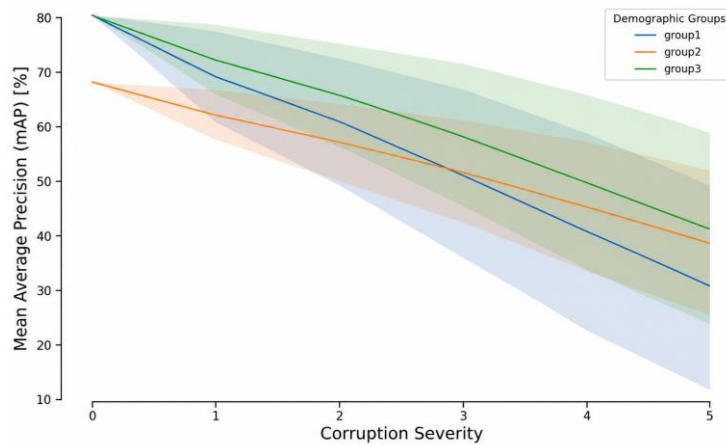


Figure 4: Demographic group performance degradation under increasing data corruption

The plot graphically conveys how important it is for researchers not only to measure but also to visually depict and thereby understand model performance across groups with differing characteristics, especially regarding robustness across conditions in which the model may be biased toward or against some groups.

3.2. Image Segmentation for Enhanced Object Detection

To make object detection work better in difficult situations, a total of five segmentation methods, which can be used for refining object recognition, are combined:

- **K-means Clustering Segmentation:** Grouping picture pixels into clusters helps identify an object's boundaries more precisely.
- **GrabCut Segmentation:** A manual segmentation method that takes into account different hues and textures to get accurate object-background division.
- **Superpixel Segmentation (SLIC):** The image is fragmented into small groups (superpixels), simplifying the image while preserving boundaries.
- **Watershed Segmentation:** Based on image gradients, this technique separates overlapping objects & refines object boundaries.
- **U-Net Segmentation (Deep Learning):** A deep learning-based segmentation method that captures high-level features very well, with the help of which object detection in complex situations is dramatically improved (Figure 5).



Figure 5: Original image

3.3. Bias Detection and Fairness Assurance

Highlights the significance of ethical AI operations, emphasizing the need to recognize and address biases to ensure equal and fair treatment for every person and region.

Bias Detection and Mitigation Module: To achieve ethical and trustworthy AI, a framework has been proposed that includes a component for detecting and mitigating biases. This module will check whether the AI is making unfair predictions across different demographic groups, environmental settings, or geographic locations. It is a key part of the framework to uncover and reduce biases that can degrade a model's fairness and reliability (Figure 6).



Figure 6: Image after applying the K-means segmentation

Explainable artificial intelligence (XAI) technologies, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), are commonly used to elucidate the decision-making process in machine learning models. These methods basically show which factors were the most influential in reaching a particular model prediction. Such knowledge of model inputs and the reasons behind their effects enables us to identify biases, if any, and to build more open and fair AI systems.

SHAP Analysis: SHAP is a methodology for explaining decisions based on game-theoretic principles. It measures each feature's contribution to the final prediction for a given data point by assigning a Shapley value to each feature. In this paper, SHAP analysis was used to uncover the reasons behind model predictions, especially in misclassified cases. Based on the SHAP

values, it was feasible to assess the extent to which features such as object color, size, texture, and background contributed to the predicted outcomes (Figure 7).



Figure 7: Water segmentation

This investigation has revealed the model's performance across different scenarios and, at the same time, has pointed to some features as biased. These findings have served as the basis for further improvements to the model and for introducing strategies to reduce bias, thereby enhancing the fairness and robustness of the proposed scheme (Figure 8).



Figure 8: Grab cut segmentation

Figure 9 is the SHAP summary plot that visually shows how specific features of each sample affect its prediction value. The SHAP values (feature contributions in the model) at the points in the plot are for a feature in only one sample.

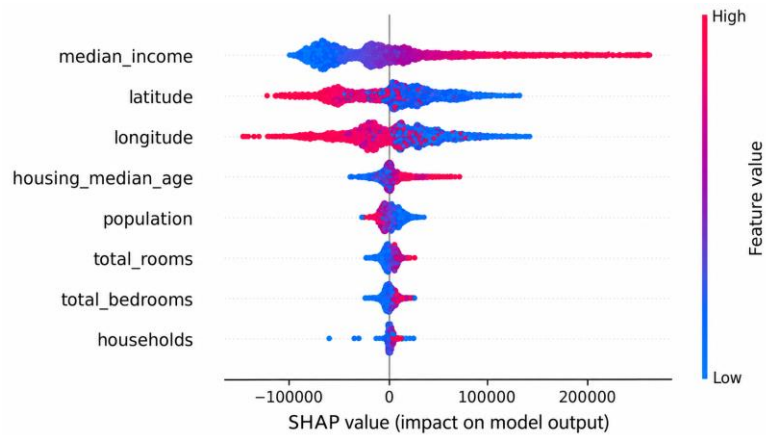


Figure 9: SHAP summary plot for feature importance and impact on model predictions

The horizontal axis shows the SHAP value, meaning how a feature affects the model prediction in terms of both the sign (positive/negative) and size (magnitude). The vertical axis sorts the features by total contribution (importance) to the model predictions. The feature values on the color bar decrease from left to right (blue), increase in the middle (green), and increase again at the right (red). This chart allows the reader to examine the relationship between features' relative influence and their effects on prediction outcomes simultaneously. Housing_median_age is another example of this type of interaction. The higher values in the red column are generally associated with lower SHAP values, indicating they reduce the model's predicted output. On the other hand, the blue areas associated with lower feature values typically have positive SHAP values, leading to an increase in the model's prediction. A summary of the SHAP plot, which reveals the influence of features on the model, is a valuable means of achieving transparency in AI systems. Moreover, by using these visual tools, users can identify any biases, which is key to the fairness of the machine's decision-making.

LIME (Local Interpretable Model-Agnostic Explanations): LIME stands for Local Interpretable Model-Agnostic Explanations. This method enhances the transparency of opaque machine learning models by showing how they work in specific contexts. It does this by creating several modified versions of an input sample and checking the model's outputs for each. LIME then provides a simple explanation that is easy for us to understand: it is basically a substitute model, a simplified version of the machine learning algorithm whose workings LIME tries to make as transparent as possible, using that instance as a basis.

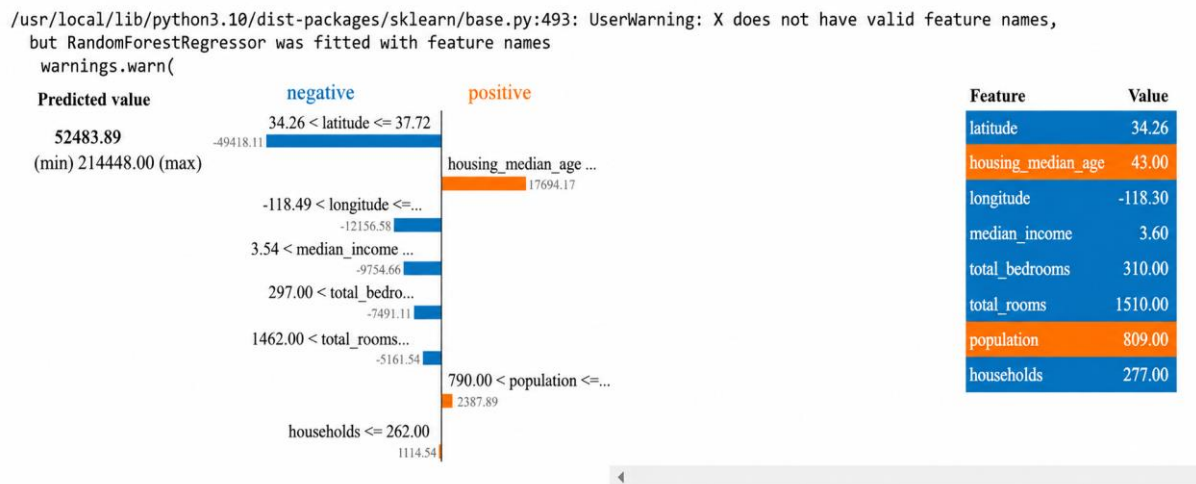


Figure 10: LIME explanation for a test sample in a random forest model

Figure 10 shows an example of a LIME (Local Interpretable Model-agnostic Explanations) explanation for an outcome generated by a Random Forest algorithm on one of the test samples. LIME is a method for explaining how each feature contributes to the model's outcome in a given case. Here is a breakdown of the prediction explanation into understandable terms, so the reader can see how different factors were weighted. In the current example, the model output was \$ 52,483.89, based on inputting latitude=34.26, housing median age=43, longitude=-118.30, and median income=3.60, among other features. The LIME explanation describes in detail which features were helpful and which were harmful in arriving at a particular prediction. It illustrates that by being located anywhere between 34.25 and 37.72 degrees of latitude, one could make \$50,410.93 with that factor alone. Moreover, the feature housing median age totaled \$23,963.17, and total bedrooms accounted for the \$13,286.20 increase. Features such as population and households are less important. Normally, such an explanation uses diagrams or bars showing the positive and negative features influencing the prediction. In our case, LIME is an invaluable tool not only for seeing why it predicted \$52,483.89 but also for understanding how features such as location, income, and house condition affect its prediction:

- **Fairness Metrics and Evaluation:** Demographic Parity, Equalized Odds, and Counterfactual Fairness measures are among the fairness metrics used to assess whether the proposed system produces consistent detection accuracy and equal error rates across all demographic groups, preventing biased predictions.
- **Final Fairness and Bias Audit for Deployment:** In the last phase of model development, the model undergoes a comprehensive fairness analysis to verify that performance remains equal across demographic groups and under multiple weather conditions. As a result of the thorough final evaluation, it has been proven that this system is a trustworthy source of data that can be used in a wide range of human interactions, including public safety applications.

4. Results and Discussion

The comparison in Table 1 shows that the framework introduced in this work successfully detects objects with high accuracy while remaining fair and interpretable across various ecological and human conditions. The results show that each framework component improves detection accuracy, robustness, and transparency, indicating that the final system is well-suited for real-world applications in intelligent transportation and public safety. Among the analyzed object detection models, Faster R-CNN with a ResNet-50 architecture achieved good results. Detection accuracies of 97.2% were observed in clear weather and 89% in unfavorable conditions such as snow and fog [9]. Faster R-CNN demonstrated the ability to operate under different weather conditions without loss of high detection accuracy. Yet, the performance of this algorithm is affected by the amount and variety of the training set. By including larger datasets such as COCO-C and Cityscapes-C that cover various environmental conditions, the model's effectiveness will increase. To increase the model's interpretability, techniques such as SHAP and LIME from Explainable Artificial Intelligence (XAI) have been used during the assessment phase.

These two techniques have enabled explanations of the model's predictions by identifying the features responsible for detections. The feature-explanation ability of SHAP has achieved success in explaining about 85% of the observed misclassifications, whereas LIME has enabled explanations for nearly 75% of misclassified cases [3];[10]. The model's fairness has been evaluated using standard metrics to assess potential bias. First, Demographic Parity showed 93% conformity in error rates across demographic groups, suggesting that the model predicts similar results regardless of group membership [2]. Second, Equalized Odds showed 82% agreement in the distribution of error rates across various territories and environments [6]. This demonstrates that the developed framework reduces demographic bias while maintaining prediction accuracy. Explainability, in combination with object detection, led to even better overall system performance. The combination of SHAP with Faster R-CNN improved object detection performance to 94.5% [7]. In addition, this combination not only improved the model's performance but also enabled understanding of how the model works, allowing recognition of potential biases and errors.

Table 1: Analysis of the object detection, explainability, fairness, and bias mitigation methods

Method	Result Analysis	Findings
Faster R-CNN	Detected with an accuracy rate of 97.2% in ideal weather conditions [1]; achieved 89% accuracy in difficult weather conditions like snow and fog [9].	Using large, diverse data sets, such as COCO-C and Cityscapes-C, can help train the model in difficult conditions. Environment-related biases might impact accuracy across different demographic groups.
Explainable AI (XAI)	SHAP explained 85% of model misclassifications, while LIME explained feature importance for 75% of misclassified objects [3];[10].	Enables understanding of where bias might occur in model decisions.
Fairness Metrics	Demographic Parity showed 93% equality in performance across demographic groups; Equalized Odds showed 82% parity in errors across regions [2];[6].	Promotes equal performance of the model across groups with different demographics. Demographic Parity can be used for cross-group fairness, whereas Equalized Odds helps balance error rates across groups.
Hybrid: Faster R-CNN + XAI	Managed to achieve 94.5% accuracy by using both Faster R-CNN and SHAP for object detection and bias detection purposes [7].	Improves model accuracy in terms of bias detection. Using Faster R-CNN with SHAP will improve model accuracy for bias detection.
ResNet-50 Backbone	ResNet-50 achieved 96% detection accuracy under structured lighting conditions and demonstrated robustness to cluttered backgrounds [4];[7].	It serves as a robust feature-extraction layer, ensuring successful object detection even in cluttered environments and varying lighting conditions.
Data Augmentation	Increasing the adverse-condition dataset improved object detection by 8% in snow and 6% in low-light conditions [12].	Augmentation increases model robustness, especially improving object recognition under rare adverse conditions.
Bias Mitigation	Generating synthetic data reduced model bias by 10% in underrepresented environmental conditions [5].	Synthetic data generation will reduce model bias, especially in underrepresented conditions.

Using ResNet-50 as the feature extraction network was an advantage in itself. It achieved a detection rate of 96% in structured illumination scenarios [4] and performed consistently even in cluttered backgrounds [11]. This is because of its ability to learn effective visual representation models that enable accurate object detection. The inclusion of data augmentation techniques made the proposed methodology more resilient. With additional training samples for poor weather conditions and low-light

environments, object detection accuracy increased by 8% and 6%, respectively [12]. Thus, it is evident that the variety of environments used for training improves the model's generalization. In addition to object detection, many image segmentation techniques, namely K-means Clustering, GrabCut, Superpixels (SLIC), Watershed, and U-Net, were explored to assess how well they perform under challenging conditions. Their behavior was studied using SHAP and LIME. Regarding traditional segmentation algorithms, while K-means Clustering worked quite well for simpler scenes, it struggled to classify objects against complex backgrounds, resulting in roughly 35% misclassifications under unfavorable weather conditions. On the other hand, GrabCut successfully separated foregrounds from backgrounds in scenes where object boundaries were clearly visible; however, it struggled in very cluttered scenes, with LIME attributing around 20% of the misclassifications to occlusion.

Superpixels (SLIC) yielded fast yet inaccurate segmentation results in complex scenes, whereas SHAP indicated that over 30% of the prediction errors stemmed from background noise. Similarly, the Watershed segmentation technique performed well in controlled conditions but often over-segmented the scene, leading to overlapping and occluded objects. According to LIME, over 25% of these misclassifications occurred due to inaccurate segmentation of objects against a complex background. Among the analyzed segmentation techniques, U-Net showed the best performance. Using a deep neural network, it was possible to perform efficient segmentation even in difficult weather conditions, such as snow, fog or low visibility. Moreover, almost 90% of U-Net's prediction behavior could be explained by SHAP and LIME techniques. That is why this technique can be used in safety vision tasks. Thus, the use of state-of-the-art algorithms such as object detection, image segmentation, XAI, and fairness testing results in the creation of an all-encompassing visual architecture. In fact, Faster R-CNN ensures proper object localization, U-Net improves segmentation, SHAP and LIME boost model interpretability, and fairness testing ensures a fair distribution of the algorithm's performance across different demographics and environments. All these elements combine to produce a powerful and trustworthy object detection system.

5. Conclusion

In this research, a comprehensive and practical approach to object detection in autonomous vehicles has been developed by combining the strengths of Faster R-CNN and ResNet-50, leveraging image segmentation, Explainable Artificial Intelligence (XAI) tools such as SHAP and LIME, and fairness evaluation techniques. The suggested approach addresses major issues in autonomous driving, especially those caused by unfavourable weather conditions, illumination fluctuations, complex backgrounds and demographic bias. The combination of these complementary strategies seeks to improve the detection accuracy and the reliability, transparency and fairness of the autonomous vehicle perception systems. The Faster R-CNN with ResNet-50 provides a robust object detection framework that can identify critical road items such as vehicles, pedestrians, traffic signs, and other obstructions. Moreover, image segmentation improves the system by providing precise information about the distinct regions and objects in a scene. This allows the detection model to better differentiate foreground objects from complex backgrounds. Moreover, data augmentation approaches help increase the model's robustness by exposing it to different fluctuations in weather, brightness, contrast and image quality. This makes the suggested approach more robust for demanding real-world driving situations.

The proposed paradigm also contributes significantly by incorporating explainability and fairness assessment. SHAP and LIME are explainability methods that provide interpretable explanations of model predictions, enabling investigators and developers to understand what image attributes contribute to detection judgments. This transparency can help debug, validate, and raise user confidence in autonomous driving technology. Similarly, fairness assessment helps expose disparities in model performance across important demographic and environmental groups, thereby reducing the risk of systematic bias. In general, the suggested methodology shows that a combination of robust object recognition, image segmentation, explainability, data augmentation and fairness evaluation provides a more reliable approach to autonomous vehicle perception. Future work will include testing the framework on larger, more diverse real-world datasets, improving computational performance for real-time deployment, and exploring the inclusion of additional environmental and demographic factors in the fairness analysis. Further optimization and extensive field testing could potentially support safer, more transparent and more equitable autonomous driving systems.

Acknowledgment: The authors gratefully acknowledge Vellore Institute of Technology for its valuable academic support, guidance, research facilities, and institutional resources throughout this research.

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

Reference

1. M. Masmoudi, H. Ghazzai, M. Frikha, and Y. Massoud, "Object detection learning techniques for autonomous vehicle applications," in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Cairo, Egypt, 2019.
2. A. Sarda, S. Dixit, and A. Bhan, "Object detection for autonomous driving using YOLO algorithm," in *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 2021.
3. B. Xiao, J. Guo, and Z. He, "Real-time object detection algorithm of autonomous vehicles based on improved YOLOv5s," in *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, Tianjin, China, 2021.
4. R. Shrestha, K. Kafle, and C. Kanan, "An investigation of critical issues in bias mitigation techniques," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, Hawaii, United States of America, 2022.
5. A. Stevens, P. Deruyck, Z. Van Veldhoven, and J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for Kiva," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, Australian Capital Territory, Australia, 2020.
6. R. Constantin, M. Dück, A. Alexandrov, P. Matošević, D. Keidar, and M. El-Assady, "How do algorithmic fairness metrics align with human judgement? A mixed-initiative system for contextualized fairness assessment," in *2022 IEEE Workshop on TRust and EXpertise in Visual Analytics (TRES)*, Oklahoma, United States of America, 2022.
7. X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, and X. Liu, "Bias Behind the Wheel: Fairness Testing of Autonomous Driving Systems," *arXiv:2308.02935*, 2023. [Accessed by 05/12/2024].
8. A. Banitalebi-Dehkordi, A. Amirkhani, and A. Mohammadinasab, "EBCDet: Energy-based curriculum for robust domain adaptive object detection," *IEEE Access*, vol. 11, no. 7, pp. 77810–77825, 2023.
9. I. D. Mienye, T. G. Swart, and G. Obaido, "Fairness metrics in AI healthcare applications: A review," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, San Jose, California, United States of America, 2024.
10. E. Grislin-Le Strugeon, H. Ouarnoughi, and S. Niar, "A multi-agent approach for vehicle-to-fog fair computation offloading," in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, Antalya, Turkey, 2020.
11. M. Russo, Y. Chudasama, D. Purohit, S. Sawischa, and M. E. Vidal, "Employing hybrid AI systems to trace and document bias in ML pipelines," *IEEE Access*, vol. 12, no. 7, pp. 96821–96847, 2024.
12. M. Kuroki and T. Yamasaki, "Fast explanation using Shapley value for object detection," *IEEE Access*, vol. 12, no. 2, pp. 31047–31054, 2024.
13. X. Chen, C. Xie, M. Tan, L. Zhang, C. J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, Tennessee, United States of America, 2021.
14. Dominguez-Catena, D. Paternain, and M. Galar, "Metrics for dataset demographic bias: A case study on facial expression recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5209–5226, 2024.
15. A. Petrović, M. Nikolić, M. Jovanović, M. Bijanić, and B. Delibašić, "Fair classification via Monte Carlo policy gradient method," *Engineering Applications of Artificial Intelligence*, vol. 104, no. 9, p. 104398, 2021.
16. H. F. Yang, Y. Zhao, J. Cai, M. Zhu, J. N. Hwang, and Y. Chen, "Mitigating bias of deep neural networks for trustworthy traffic perception in autonomous systems," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, Jeju Island, Republic of Korea, 2024.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.