IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 7, March 2025

Enhancing Transparency in Real-Time Object Detection for Autonomous Systems Using Explainable AI

G Sarthak[1], K Nitesh[2],V Karthikeya[3],P Aryeshu[4], D Sai Charan[5], Sampath Mounish[6],D Sankeerthana[7]

UG Students, Department of CSE[1,2,3,4,5,6,7]

GITAM (Deemed to be University), Visakhapatnam, India

Abstract*:* *The rapid advancement of Artificial Intelligence (AI) in critical applications such as autonomous driving and healthcare necessitates not only high accuracy but also transparency in decision-making. This paper introduces a framework that integrates Explainable AI (XAI) techniques into YOLOv8, a cutting-edge object detection model, to enhance interpretability in real-time environments. We employ Class Activation Mapping (CAM) methods, including Grad-CAM and HiResCAM, alongside Layer-wise Relevance Propagation (LRP) to create an explanation framework. By combining CAM and LRP, we leverage the strengths of both techniques, improving the clarity of model decisions without sacrificing performance. The proposed system maintains high detection accuracy while operating at 50-60 frames per second (FPS) on a dedicated GPU and approximately 24 FPS when executing XAI computations. Our results demonstrate that integrating these XAI techniques enhances the trustworthiness and safety of AI-driven systems, making them more reliable in autonomous driving and other critical applications.*

Keywords: YOLOv8, Explainable AI (XAI), Object Detection, Autonomous Systems, Class Activation Mapping (CAM), Layer-wise Relevance Propagation (LRP), Machine Learning

I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has revolutionized various industries, particularly in safety-critical applications such as autonomous driving, healthcare, and industrial automation. Object detection models play a fundamental role in these domains by enabling machines to perceive and interpret their surroundings in real time. Among these models, YOLOv8 stands out due to its exceptional speed and accuracy, making it a preferred choice for real-time object detection. However, despite its effectiveness, YOLOv8 functions as a black-box model, meaning its internal decision-making processes remain largely opaque. This lack of transparency raises concerns in high-stakes scenarios where understanding the reasoning behind AI-driven decisions is crucial for safety, validation, and trust.

To address these challenges, Explainable AI (XAI) methods have emerged as essential tools for enhancing the interpretability of deep learning models. XAI techniques help bridge the gap between AI's complex computations and human understanding by offering insights into how and why a model arrives at a particular decision. In this study, we integrate XAI methodologies into YOLOv8 to improve its transparency without compromising its real-time performance. Specifically, we employ Class Activation Mapping (CAM) methods, such as Grad-CAM and HiResCAM, alongside Layer-wise Relevance Propagation (LRP) to visualize and interpret model predictions. By combining the strengths of CAM and LRP, we enhance the clarity of object detection results, making them more comprehensible and verifiable for human operators.

Our proposed framework ensures that AI-driven object detection systems provide both accurate and interpretable predictions, which is critical for decision-making in autonomous vehicles and other high-risk applications. The system achieves real-time processing speeds of 50-60 frames per second (FPS) on a dedicated GPU and maintains approximately 24 FPS when running XAI computations, demonstrating its feasibility for deployment in real-world environments. This integration of explainability into high-speed AI models not only increases trust and safety in AI-

DOI: 10.48175/568

driven systems but also broadens their potential applications in fields requiring transparent decision-making, such as medical diagnostics and industrial robotics.

II. METHODOLOGY

This section outlines the approach used to implement and evaluate YOLOv8 for object detection in autonomous driving. We describe the dataset, training process, performance evaluation metrics, and the Explainable AI (XAI) techniques integrated to interpret the model's decisions.

A. YOLOv8 for Object Detection

YOLOv8 (You Only Look Once, version 8) is the latest iteration in the YOLO family, optimized for real-time object detection. It employs a single convolutional neural network (CNN) to process an entire image in a single pass, predicting bounding boxes and class probabilities efficiently. Due to its high speed and accuracy, YOLOv8 is well-suited for applications like autonomous driving, where quick and reliable object detection is crucial.

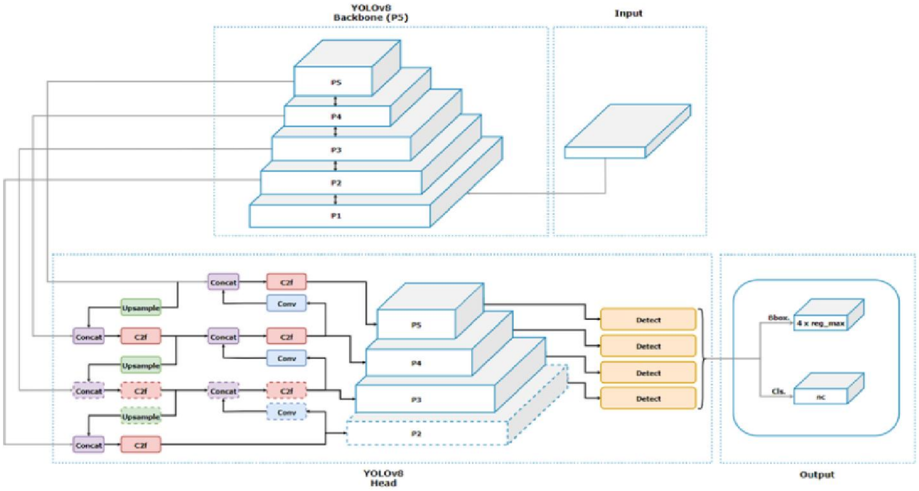

Fig 1. YOLOv8 Architecture

B. Dataset and Training

We trained YOLOv8 on the KITTI dataset, a widely used benchmark in autonomous driving research. The dataset consists of 7,464 labelled images spanning eight object categories: car, cyclist, pedestrian, person sitting, tram, truck, van, and miscellaneous. To improve model generalization, we applied data preprocessing techniques, including:

Resizing images to match the input dimensions required by YOLOv8.

Data augmentation, such as image flipping and colour adjustments, to enhance robustness against environmental variations.

We trained the model for 100 epochs using the Ultralytics YOLOv8 library with a batch size optimized for our hardware capabilities.

C. Performance Evaluation

The trained YOLOv8 model achieves a real-time processing speed of 50-60 FPS on a dedicated GPU. For applications requiring explainability computations, the model runs at approximately 24 FPS while maintaining high detection accuracy. Performance metrics such as mean Average Precision (mAP), precision-recall curves, and confusion matrices were used to assess the model's effectiveness.

D. Explainable AI Techniques

Deep learning models like YOLOv8 often function as black boxes, making it difficult to interpret their decision-making processes. To address this, we incorporated Explainable AI (XAI) methods, specifically:

- Class Activation Mapping (CAM) techniques: Grad-CAM and HiResCAM
- Layer-wise Relevance Propagation (LRP)

DOI: 10.48175/568

By integrating these methods, we aim to enhance the interpretability of YOLOv8 while maintaining its real-time capabilities.

E. Class Activation Mapping (CAM)

CAM provides visual explanations of model predictions by generating heatmaps that highlight important image regions.

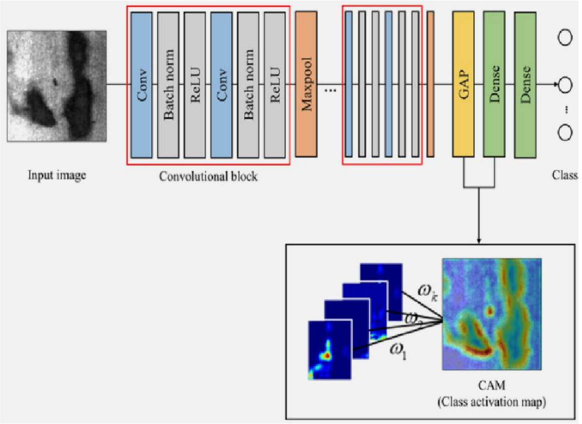

Fig 2 : CAM architecture

- Grad-CAM: Computes gradients of the output class with respect to the final convolutional layer to generate a coarse heatmap.
- HiResCAM: An improved version of Grad-CAM that produces sharper, more detailed heatmaps, reducing blurriness and enhancing interpretability.

The generated heatmaps allow us to identify which areas of an image the model is focusing on when making a prediction.

F. Layer-wise Relevance Propagation (LRP)

Unlike CAM, which provides a general understanding of the model's focus, LRP offers a pixel-level explanation of the model's decision.

- LRP traces back the model's predictions to identify the exact pixels contributing to a classification decision.
- This provides a more granular and precise explanation of the model's reasoning.

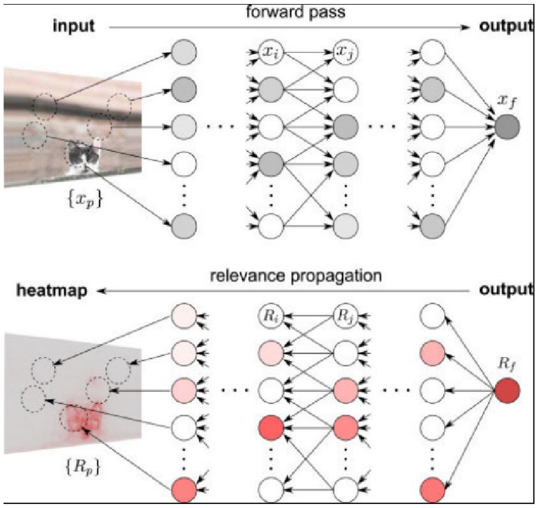

Fig 3: Flow diagram of LRP

DOI: 10.48175/568

G. Combining CAM and LRP

To leverage the strengths of both techniques, we developed a hybrid approach by:

- Generating separate heatmaps using Grad-CAM/HiResCAM and LRP.
- Overlaying the heatmaps to provide a multi-scale visualization, ensuring both broad and fine-grained interpretability.
- Applying a weighted sum to merge the outputs, balancing general object focus (from CAM) with pixel-specific details (from LRP).

This combined approach enhances the transparency of YOLOv8's detections, helping users understand not just where the model is focusing, but also why it makes specific classification decisions.

III. RESULTS

In this section, we evaluate the performance of our YOLOv8 model by analyzing key metrics and utilizing Explainable AI (XAI) techniques to interpret its predictions. Our goal is to assess both the model's effectiveness and the reasoning behind its decisions.

A. Dataset Distribution and Label Analysis

Figure 4 presents the distribution of object labels within the dataset. The 'Car' class appears most frequently, whereas categories such as 'Cyclist' and 'Person Sitting' are relatively underrepresented. This imbalance may cause the model to perform exceptionally well on 'Cars' while struggling with less frequent classes. Recognizing this disparity is crucial when interpreting the results.

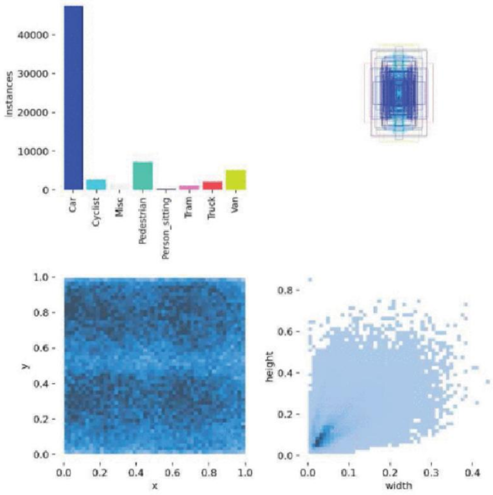

Fig 4 : Dataset distribution

B. Confusion Matrix

To gain insights into the model's classification accuracy, we examine the confusion matrix in Figure 5. It highlights the frequency with which each class was correctly identified versus instances where the model misclassified them. The model demonstrates high accuracy in detecting 'Cars,' but there are some misclassifications, particularly between visually similar classes such as 'Car' and 'Van.'

DOI: 10.48175/568

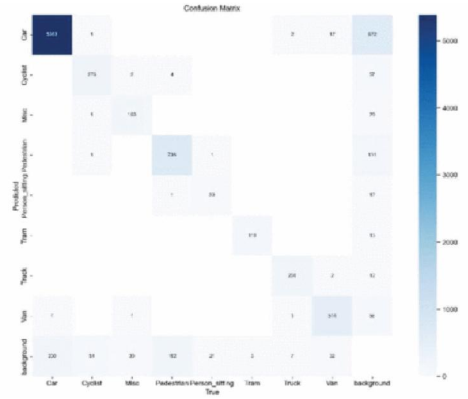

Fig 5: Confusion Matrix

C. Precision-Recall Analysis

The precision-recall curve, shown in Figure 6, illustrates the trade-off between precision and recall for each object class. The model performs well for vehicle classes like 'Car,' 'Tram,' 'Truck,' and 'Van,' maintaining a strong PR curve. However, for less frequently occurring objects such as 'Person Sitting' and 'Misc,' the curve declines, suggesting the model encounters difficulty distinguishing these classes accurately.

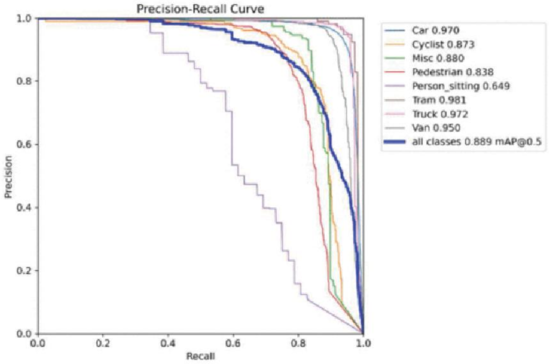

Fig 6: Precision-Recall Curve

D. F1 Score Analysis

Figure 7 depicts the F1 scores across varying confidence levels. The 'Car,' 'Tram,' 'Truck,' and 'Van' categories consistently achieve high F1 scores, while other less frequent objects show lower scores. This imbalance suggests that the model finds it challenging to maintain a balance between precision and recall for these underrepresented categories.

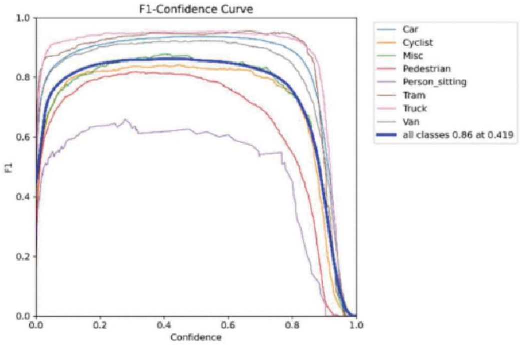

Fig 7: F1 Confidence Curve

DOI: 10.48175/568

E. Results of the Proposed Explainability Technique

In our study, we combined Class Activation Mapping (CAM) techniques with Layer-wise Relevance Propagation (LRP) to enhance the interpretability of YOLOv8's decision-making process. Figure 8 illustrates the impact of this combined approach, demonstrating how the integration of these two techniques provides a more detailed and transparent explanation of the model's predictions.

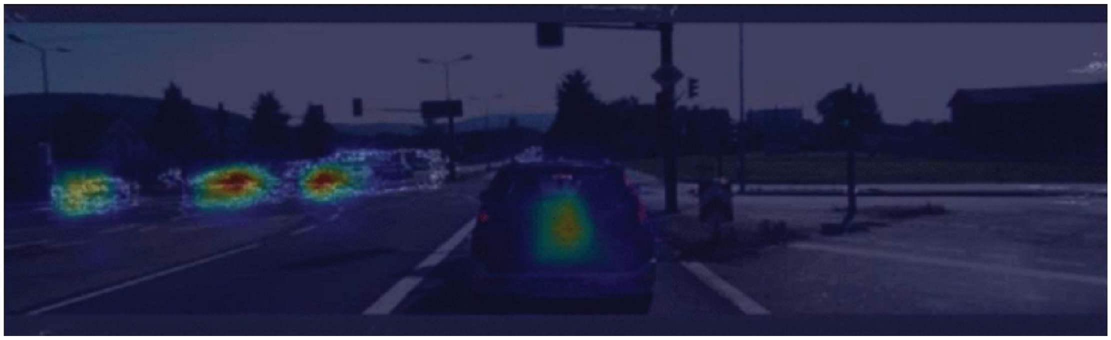

Fig 8: Combination of CAM and LRP Heatmaps

F. Insights from CAM and LRP Integration

Class Activation Mapping (CAM):

CAM-based methods, such as Grad-CAM and HiResCAM, are effective in identifying the general regions of an image that the model focuses on when making a prediction. These techniques generate heatmaps that highlight the areas contributing most to the classification, offering a broad yet intuitive visualization of the model's attention.

However, a limitation of CAM is that it can overgeneralize the importance of regions by covering large portions of the image, sometimes missing fine-grained details essential for understanding model decisions.

Layer-wise Relevance Propagation (LRP):

LRP, on the other hand, offers a more precise and pixel-level explanation of the model's decision-making process. Instead of merely showing broad activation regions, LRP traces the flow of relevance through the network layers to determine exactly which pixels contributed most to a specific prediction. This is particularly useful in cases where fine details matter, such as distinguishing between similar-looking objects or identifying partially occluded objects.

Effect of Combining CAM and LRP:

By overlaying CAM heatmaps with LRP visualizations, our approach ensures that we get both a high-level understanding of object focus (from CAM) and a precise, pixel-wise explanation (from LRP). This integration is particularly beneficial in complex detection scenarios, such as:

Distinguishing pedestrians from background objects in crowded scenes.

Identifying small but important elements, like traffic signals or cyclists.

Understanding misclassifications by revealing which parts of an image led to incorrect predictions.

The hybrid visualization technique improves model interpretability by balancing general attention focus and detailed feature attribution. This is critical in autonomous driving applications, where understanding why an AI system makes a particular decision is essential for ensuring safety and reliability.

IV. CONCLUSION

In this study, we investigated the application of YOLOv8 for object detection in the context of autonomous driving. Our findings demonstrate that YOLOv8 excels in detecting commonly occurring objects, primarily due to its advanced architecture, optimized feature extraction, and efficient training process. The model consistently delivered high accuracy in recognizing frequently encountered objects such as cars, trams, and trucks. However, its performance declined when identifying less common objects like cyclists and pedestrians, largely due to the dataset's inherent

DOI: 10.48175/568

imbalance. This discrepancy highlights the impact of training data distribution on model generalization and suggests the need for strategies to enhance detection across all object classes.

To further understand and interpret the model's decision-making process, we incorporated Explainable AI (XAI) techniques—Class Activation Mapping (CAM) and Layer-wise Relevance Propagation (LRP). These techniques provided crucial insights into how the model identifies objects and why certain misclassifications occur. CAM visualizations helped pinpoint the regions of an image that contributed most to the model's predictions, while LRP offered a more detailed, pixel-level explanation of its decision-making. The combination of these techniques allowed us to bridge the gap between model performance and interpretability, offering a clearer understanding of the reasoning behind YOLOv8's predictions. This aspect is particularly critical in safety-critical applications such as autonomous driving, where trust and transparency in AI systems play a significant role in adoption and deployment.

Despite its strengths, our study also revealed several areas where improvements could be made. The primary limitation was the dataset imbalance, which affected the model's ability to generalize across all object classes. Future research could focus on expanding the dataset to include a more diverse representation of objects, implementing advanced data augmentation techniques, or incorporating synthetic data generation to enhance model robustness. Additionally, while CAM and LRP were effective in explaining model decisions, exploring alternative or hybrid XAI approaches could provide even deeper insights into the decision-making process.

In conclusion, YOLOv8 has proven to be a powerful tool for object detection in autonomous driving scenarios, offering high accuracy and efficiency. However, challenges such as dataset bias and explainability constraints remain. By addressing these limitations, future studies can refine the model's performance, improve its generalization across diverse environments, and further enhance the interpretability of its predictions—ultimately contributing to safer and more reliable AI-driven systems.

REFERENCES

[1]. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[2]. Naeem Ullah, Javed Ali Khan, Ivanoe De Falco, and Giovanna Sannino. 2024. Explainable Artificial Intelligence: Importance, Use Domains, Stages, Output Shapes, and Challenges. ACM Comput. Surv. 57, 4, Article 94 (April 2025), 36 pages. https://doi.org/10.1145/3705724

[3]. Ziming Wang, Changwu Huang, and Xin Yao. 2024. A Roadmap of Explainable Artificial Intelligence: Explain to Whom, When, What and How? ACM Trans. Auton. Adapt. Syst. 19, 4, Article 20 (December 2024), 40 pages. https://doi.org/10.1145/3702004

[4]. Atakishiyev, S., Salameh, M., Yao, H., Goebel, R. (2021). Explainable Artificial Intelligence for Autonomous Driving: AComprehensive Overview and Field Guide for Future Research Directions. arXiv preprint arXiv:2112.11561.

[5]. Tahir HA, Alayed W, Hassan WU, Haider A. A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME-SHAP Integration. Sensors (Basel). 2024 Oct 22;24(21):6776. doi: 10.3390/s24216776. PMID: 39517672; PMCID: PMC11548085.

[6]. Nazat S, Arreche O, Abdallah M. On Evaluating Black-Box Explainable AI Methods for Enhancing Anomaly Detection in Autonomous Driving Systems. Sensors (Basel). 2024 May 29;24(11):3515. doi: 10.3390/s24113515. PMID: 38894306; PMCID: PMC11175219.

[7]. Calderón-Peña, S., Chockler, H., Kelly, D.A. (2024). Real-Time Incremental Explanations for Object Detectors in Autonomous Driving. arXiv preprint arXiv:2408.11963.

[8]. Moradi, M., Yan, K., Colwell, D., Samwald, M., Asgari, R. (2023). Model-agnostic explainable artificial intelligence for object detection in image data. arXiv preprint arXiv:2303.17249.

[9]. Atakishiyev, S., Salameh, M., Yao, H., Goebel, R. (2021). Explainable Artificial Intelligence for Autonomous Driving: AComprehensive Overview and Field Guide for Future Research Directions. arXiv preprint arXiv:2112.11561.

[10]. M. A. Mahima, N. C. Patel, S. Ravichandran, N. Aishwarya and S. Maradithaya, "A Text-Based Hybrid Approach for Multiple Emotion Detection Using Contextual and Semantic Analysis," 2021 International

DOI: 10.48175/568

Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-6, doi: 10.1109/ICSES52305.2021.9633843. keywords: {Text mining;Human computer interaction;Emotionrecognition;Sentimentanalysis;Semantics;Bit error rate;Transformers;TextMining;Multiple Emotion Detection;Natural Language Processing;Sentenceembeddings;CosineSimilarity;Ekman's emotions},

[11]. S. Brdnik and B. Šumak, "Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research," 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2024, pp. 2032-2038, doi: 10.1109/MIPRO60963.2024.10569528. keywords: {Surveys;Ethics;Systematics;ExplainableAI;Reviews;Bibliographies;Marketresearch;Explainable Artificial Intelligence;XAI;ExplanationTechniques;TertiaryReview;Systematiciterature review},

[12]. F. Bargagna, L. Anita De Santi, M. F. Santarelli, V. Positano and N. Vanello, "Bayesian XAI Methods Towards a Robustness-Centric Approach to Deep Learning: An ABIDE I Study," 2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Eindhoven, Netherlands, 2024, pp. 1-5, doi: 10.1109/MeMeA60663.2024.10596826. keywords: {Deep learning;Autism;Uncertainty;Biological system modeling;Predictivemodels;Biomarkers;Bayesmethods;Bayesian;xAI;DeepLearning;ABIDE;Autism},

[13]. A MB, A. P and A. Kodipalli, "Leveraging XAI and Breakthrough Machine Learning Techniques for Trigeminal Neuralgia Severity Classification," 2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India, 2024, pp. 1-6, doi: 10.1109/TENSYMP61132.2024.10752211. keywords: {Logistic regression;Accuracy;ExplainableAI;Medicaltreatment;Nearestneighbormethods;Boosting;Brainmodeling;Planning;Randomforests;Regression tree analysis;Machinelearning;SVM;Logisticregression;DecisionTree;XAI;LIME and SHAP;K-Nearest Neighbors (KNN)},

[14]. Kuppa and N. -A. Le-Khac, "Adversarial XAI Methods in Cybersecurity," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4924-4938, 2021, doi: 10.1109/TIFS.2021.3117075.keywords: {Predictive models;Datamodels;Analyticalmodels;Privacy;Computersecurity;Password;Numericalmodels;Adversarial machine learning;XAI;cybersecurity;counterfactualexplanations;adversarialattacks;poisoningattacks;modelstealing;membership inference attacks},

[15]. -Y. Kao, J. Chen, K. Markert and K. Böttinger, "Rectifying adversarial inputs using XAI techniques," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 573-577, doi: 10.23919/EUSIPCO55093.2022.9909699. keywords: {Deep learning;Neuralnetworks;Decisionmaking;Europe;Signalprocessing;Observers;Security;explainableAI;neuralnetworks;deeplearning;adversarialdefense},