WCRG WORLD COMET RESEARCH GROUP

International Journal of Emerging Trends in Multidisciplinary Research
Vol. 7, No. 1, 2024
Doi: 10.67228/30715636/IJETMR-2024PI1B3J
PP. 01-17

Original Article

Explainable Deep Learning Framework for Autonomous Transportation Safety

Johan Hastad's mentor Arne Andersson[1], Börje Langefors[2]

[1]*Professor, Uppsala University, Sweden.*

[2]*Professor of Information Systems, Stockholm University, Sweden.*

Received: 03-06-2024 Revised: 03-28-2024 Accepted: 04-01-2024 Published: 04-03-2024

ABSTRACT

For this reason, autonomous transportation systems have been established as a progressive technology that leverages Artificial Intelligence (AI), Internet of Things (IoT), computer vision, and advanced sensing technologies to improve road safety, operational efficiency, and smart mobility. Deep learning models have shown great strength in detecting objects, recognizing lanes and pedestrians, avoiding obstacles on the road, as well as analyzing real-time traffic situation. Although very accurate, these models are typically black-boxes which have limited transparency and confidence in safety-critical transportation applications. Interpretability can help with accident investigation, regulatory compliance, ethical decision-making, and public acceptance of autonomous vehicles where their lack presents major problems. We present an Explainable Deep Learning Framework for Autonomous Transportation Safety by combining convolutional neural networks with the Understandable Artificial Intelligence (XAI) techniques to ensure transparency and soulfulness in autonomous transportation. The proposed framework comprises sensor fusion, image processing, feature extraction, deep neural network inference and explainability mechanisms such as Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). It pre-evaluates the prediction interpretation, and creates visualisations and feature-level explanations that give stakeholders insight into how models can have high accuracy on detection outcomes. Additionally, the architecture promotes accountability, compliance with regulations, safer autonomous driving and public trust in intelligent transportation systems. Explainability is shown to be an important building block for designing robust autonomous transport platforms usable in the future.

KEYWORDS

Autonomous Transportation, Explainable Artificial Intelligence (Xai), Deep Learning, Autonomous Vehicles, Transportation Safety, Computer Vision, Explainable Deep Learning, Intelligent Transportation Systems (Its), Grad-Cam, Shap, Lime.

WCRG
Global Research
ISSN: 3071-5636
PP: 01-17.

1. INTRODUCTION

1.1. Background

While every new technology has its distinct evolution path, not only the development of autonomous transportation is accelerated by the rapid progress in Artificial Intelligence, machine learning algorithms, sensor technologies and intelligent transportation infrastructure. Autonomous vehicles are designed to minimize human reliance by consistently perceiving the environment, understanding multi-resolution traffic situations that they must navigate, forecasting hazardous interactions with other road users, and executing safe driving maneuvers. Motion-based vehicles depend on various heterogeneous sensors, such as high-resolution cameras, LiDAR, radar, IMU (Inertial Measurement Unit), systems across the trajectories GPS and ultrasonic sensors to get a unique idea of vehicle surroundings [4]. The multi-modal data collected allows autonomous systems to identify objects, read-road conditions, predict vehicle dynamics movement and navigate dynamic environments. Advanced deep learning methods have also added to autonomous vehicle capabilities across perception, prediction and decision making. But the trend towards more sophisticated models based on AI creates problems with transparency and reliability. Thus, explainable artificial intelligence methods have been integrated into a vital part of safe, reliable and accountable autonomous transportation systems.

1.2. Need for Explainable Deep Learning Framework

Deep learning — Machine learning falls under three most autonomous one in all modes of transportation, and these new technologies have made a rapid transition to improve performance by being able to focus on perception, prediction and decision making. On the other end of spectrum, deep learning models usually behave as sophisticated black-box modules without clear understanding of how their predictions arrive. In contrast, for safety-critical applications such as autonomous vehicles, high accuracy is not enough because incorrect decisions might have catastrophic consequences. Hence, the demand for explainable deep learning frameworks that can give rationale or provide evidence towards transparency, trust and ensure safe use of such autonomous transportation systems is increasing.

ISSN: 3071-5636
PP: 01-17.

Fig 1 - Need for Explainable Deep Learning Framework

1.2.1. Improving Transparency of AI Decision-Making

For example, deep learning models in… Deep LearningBased Technologies for Autonomous Vehicles In the realm of autonomous vehicles, a lot of raw sensor inputs are fed into models to yield decisions such as "avoid obstacle," "change lane," or "break urgently." But these choices are too often made without clear justification. The paper proposes an explainable deep learning framework that identifies critical influencing factors and provides concise reasoning to prospective users of the model—an original feature useful for justifying autonomous actions. The effort will enhance the engineers' and users' understanding of how AI systems work in various driving circumstances.

1.2.2. Enhancing Safety and Reliability

Since vehicles are operating in unpredictable environments, including changing weather, traffic conditions and human behavior on the road, safety is also a top requirement of autonomous transportation up to October 2023. Explainable frameworks are used to analyze incorrect predictions and explain the reasons for the failure of a system. Knowing if errors come from sensor limitations, environmental conditions, or model weakness can help developers make a more reliable system and minimize potential safety risks.

1.2.3. Supporting Failure Analysis and System Validation

Autonomous vehicles are still tested and validated for years before they are deployed into the real world. Deep learning models are usually shown by performance results without explaining the reasons for failure. Techniques from Explainable AI e.g., SHAP, LIME, Grad-CAMs provide low level interpretability by highlighting important features in the input to explain higher-level decisions made by the model over a region (specific geo-location or time domain)independent of whether the models were DNN or XGBoost on sensor data. This enables easy debugging, model enhancement, and safety evaluation.

ISSN: 3071-5636
PP: 01-17.

1.2.3. Supporting Failure Analysis and System Validation

Autonomous vehicles are still tested and validated for years before they are deployed into the real world. Deep learning models are usually shown by performance results without explaining the reasons for failure. Techniques from Explainable AI e.g., SHAP, LIME, Grad-CAMs provide low level interpretability by highlighting important features in the input to explain higher-level decisions made by the model over a region (specific geo-location or time domain)independent of whether the models were DNN or XGBoost on sensor data. This enables easy debugging, model enhancement, and safety evaluation.

1.2.4. Increasing User Trust and Acceptance

Successful deployment of self-driving cars for public use hinges upon the establishment of confidence in AI-based autonomous decision making. Users cannot depend on autonomous systems if they are unable to comprehend vehicle behavior during crucial incidents. Explainable deep learning frameworks endow human-interpretable explanations that are able to facilitate the user trust in the system by being a proof of evidence that autonomous decisions on positive positive interactions are based on inadequate environmental information.

1.2.5. Meeting Regulatory and Ethical Requirements

Autonomous transportation systems of the near future need to meet rigorous regulatory criteria regarding safety, accountability and transparency. Explainable AI is increasingly gravitating the focus of regulatory organizations that monitor and assess automated decision processes. An explainable framework supports compliance by demonstrating how decisions are produced and enabling accountability for those autonomous operations.

1.2.6. Enabling Continuous Improvement of Autonomous Systems

The ability to explainable little assists in never-ending development and growth of autonomous cars. Researchers may enhance training datasets, fine-tune algorithms, and raise the overall performance of the system by analyzing model selections and weak points. Thus, it is crucial to develop explainable deep learning frameworks as a foundation for safer transparent and trustworthy autonomous transportation systems spanning the entire safety spectrum.

1.3. Deep Learning Framework for Autonomous Transportation Safety

Intelligent Autonomous Transport Safety Deep learning framework provides an autonomous architectural design of a neural network that can help with understanding the environment, interpreting complex traffic data and supporting information-driven decisions with less reliance on human experience. Molodtsov: State of the art deep learning algorithms are required for high performance perception processing in autonomous transportation systems which often rely on large-scale multimodal sensor data from a variety of sensing devices, including cameras, LiDAR, radar-GPS and IMU integration. Deep learning models have outpassed traditional approaches to perception and decision making by automatically inferring complex patterns and representations from raw data, instead of having to be carefully designed for each new task by humans [6]. A deep learning architecture is usually composed of multiple stages arranged in sequence, such as sensor data capture,

ISSN: 3071-5636
PP: 01-17.

input transformation (e.g., filtering), feature extraction (e.g., either semantics embedding or dimensionality reduction), perception, prediction-planning-control. At the heart of autonomous driving safety is the perception stage, which detects and identifies nearby entities like vehicles, pedestrians, traffic signs and road boundaries. For image-based sensor data, convolutional neural networks (CNNs) are commonly employed to extract spatial information whereas recurrent neural networks and transformer-based models capture temporal patterns and relation between dynamic objects. Sensor fusion techniques leverage the data from a variety of sources to make up for individual sensor shortcomings and increase reliability in adverse conditions (low light, rain, fog) as well as high congestion urban scenarios. The decision-making phase refers to the use of the extracted features and classifies relevant actions for the vehicle such as acceleration, braking, lane changing and obstacle avoidance. Prediction models utilize deep learning to predict the future state of dynamic objects in the surroundings of an autonomous vehicle, thereby enabling decision making proactively by key safety decisions instead of reactively responding to imminent risk. Yet, although classic deep learning models typically achieve good results, they seldom offer transparency since both the internal working of model and decision-making process are convoluted. To overcome this limitation, an explainable deep learning framework combines Explainable Artificial Intelligence (XAI) with deep learning models and provides understandable explanations for autonomous decisions. Approaches like SHAP, LIME and Grad-CAM accesses great features, shot important regions and weigh on the roles of other sensor inputs. This increases system transparency, facilitates failure diagnosis and enhances user and regulatory trust. Thus, a combination of deep learning frameworks and explainability mechanisms would be come one huge step forward to safe autonomous transportation. It makes it possible for vehicles to predict high-confidence, interpretable, accountable decisions in safety-critical systems from real-world applications.

2. LITERATURE SURVEY

2.1. Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a new research field and the goals behind it are increasing transparency, interpretability, and trustworthiness of artificial intelligence systems. Current AI models are black-box systems in the sense that input data and final predictions can not be easily interpreted by a human: this applies to deep learning networks in particular. This limitation is what XAI intends to address, i.e., to provide human-readable explanations of why specific models make such decisions in the first place and how various input factors influence the decision-making process. Leveraging this capability is essential in high-stakes domains like healthcare, financial services, defense systems, and autonomous transportation where unexpected or inexplicable decisions can cause loss of lives. Take the case in autonomous vehicles, wherein an AI not just has to detect objects and take driving decisions but must also give an understandable reason for why it detected an obstacle or could change lanes or apply emergency braking. There have been several explanation techniques developed that help interpret complex AI models. Local Interpretable Model-Agnostic Explanations (LIME) is designed to account for individual predictions, by approximating interpretable models in the vicinity of a data instance and enabling researchers to see how much a feature impacted each individual decision. While LIME is a very flexible method that can be applied to many machine learning models, it requires random sampling which may yield unstable or inconsistent explanations. SHAP stands for

ISSN: 3071-5636
PP: 01-17.

SHapley Additive exPlanations and is one of the other approaches to identify how much individual features contribute towards a model prediction (using ideas from cooperative game theory). SHAP is mathematically-grounded and allows for fine-grained feature importance analysis, which can assist with sensor influence, object detection confidence and risk assessment models within a transportation system. SHAP, however, could be computationally intensive especially for large deep learning models. Gradient-weighted Class Activation Mapping (Grad-CAM) Gradient-weighted Class Activation Mapping is very similar CTA and mostly notated in the plane field also it mainly used to explaining CNN by generating heat maps highlighting image regions responsible for specific vision predictions. In autonomous driving applications, Grad-CAM may be used to examine perception models to verify that they are fixated on appropriate objects like vehicles, pedestrians, traffic signs and lanes instead of context features in the environment that do not represent meaningful content. Alternative methods like Integrated Gradients and Attention Visualization expand the toolbox for interpreting neural net behavior, especially within more complex architectures such as transformers. Existing XAI methods, even the state-of-the-art ones, show unsatisfactory performance in terms of accuracy of the explanations, computation burden on very large datasets, scalability and real-time implementation. A majority of available explanation methods are designed independently from an end-to-end autonomous driving system, rendering their incorporation in real transportation environment impractical. Thus, it is still an essential research goal to create integrated XAI frameworks that can establish trustworthy, high-throughput, and real-time explanations to optimize the safety of complex intelligent transportation systems.

2.2. Deep Learning in Transportation

Deep learning is one of the cornerstones for modern transportation systems due to its capability to learn complex patterns from huge amounts of sensor and environmental data. Conventional transportation intelligence systems relied on manually constructed features being used, which also often yielded poor results when dealing with varying road conditions that were complicated. Deep learning approaches use raw inputs and extract hierarchical representations which help enable autonomous vehicles to have better perception, prediction, and decision-making capabilities than traditional algorithms. Since most of the information we are able to extract from our environment is visual, convolutional neural networks (CNNs) have been one of the most traversed types of deep learning models in autonomous transportation. Camera-based systems are widely used for object recognition in urban driving environments, such as vehicle detection [9], pedestrian detection [10] as well as traffic sign classification and lane detection. Deep Learning: CNN architectures were developed mainly for image classification tasks, however these concepts have been extended to provide Vector-based object detections with the ability to achieve higher accuracy while maintaining faster processing speeds required by a real-time driving application. Apart from the visual perception, knowing how the surrounding objects exhibit temporal behavior is critical for safe autonomous operations. RNNs, LSTMs and transformer models are used to process sequence information (you move or stopped; when pedestrians pass by; when vehicles stop in order). These models allow autonomous systems to anticipate the next events and take proactive actions instead of just responding once a situation occurs. Sensor fusion combines information from many sensor types, such as cameras, LiDAR, radar, GPS and

ISSN: 3071-5636
PP: 01-17.

other environmental sensors to improve the performance of autonomous vehicles. With this multimodal information, Deep Learning models will be able to process them to achieve higher reliability in difficult conditions such as bad weather (low light or strong rain and fog) or complex urban traffic scenarios. Autoencoders also used in feature learning and anomaly detectionDeep reinforcement learning approaches assist decision optimization; vehicles learn appropriate action by interacting with their environment. Nevertheless, upon on these advances, deep learning-based transportation systems still have key challenges. Largely reliant on large and diverse training datasets which may introduce bias, as well as complex architectures that complicate the analysis of failures or interpretation of decision processes. Simply achieving high prediction accuracy is not enough for the safety-critical applications because users, engineers and regulators want to know how decisions are being made. Consequently, there is a critical need to combine deep learning capabilities with explainability techniques to have transparent, trusted and safe autonomous transportation systems.

2.3. Autonomous Vehicle Safety

Among all the key research issues in intelligent transportation, probably none is more critical than autonomous vehicle safety, since a self-driving car must function reliably and safely in highly dynamic environments that involve unpredictable human behavior, rapidly changing weather conditions, and complex traffic interactions. The functionality of various components, including perception, localization, prediction, planning and control systems jointly determines the safety of autonomous vehicles. The perception system is the core infrastructure of any autonomous driving solution as it aggregates information from a myriad of sensors such as LiDAR, cameras and radar to create an understanding of objects in our environment including distance estimations, road boundaries, and traffic situations. Deep learning has made tremendous progress to enable autonomous cars detecting pedestrians, vehicles, road signs and other context elements in the environment from image classification. Nevertheless, science is not enough to be safe since the machine needs to predict how other objects/agents will move in the future. Prediction modules help safer decisions by predicting how vehicles, pedestrians and other road users will move from the past to the present. An input about actions to plan from these predictions, including speed maintain, lane change, avoiding and emergency braking. Autonomous vehicle safety has been examined extensively in both the simulation platforms, and controlled testing environments, as well as real-world experiments. To evaluate risk management AI systems, a variety of evaluation recommendations appeared in the literature based on scattering metrics focused on performance measurements (such as detection rate, time between warning and response, or operational efficiency) but providing very limited means to explain the rationale behind AI decisions. The above absence of explainability makes it troublesome to investigate mishaps, debug the framework, and get security endorsements. For example, if an autonomous car did not recognize an obstacle, engineers must identify whether the issue happened because of sensor failure, environmental conditions, or simply training data wasn't enough to prepare the model. Explainable AI, meanwhile, can help to faithfully identify what drove the decision and provide key insight into how you could enhance system reliability. At the same time, transparency and accountability in artificial intelligence systems is also required by transportation regulations. To be accepted into the world, future autonomous vehicles have to do more than simply not crash; they also

ISSN: 3071-5636
PP: 01-17.

have to be able to explain themselves in such a way that engineers will understand them and bystanders and regulators will accept them. Thus, the inclusion of explainable deep learning into safety frameworks for autonomous driving can enhance failure diagnosis, increase public trust, support certification processes and pave the way to safer autonomous mobility solutions.

2.4. Research Summary

However, a literature study reveals that deep learning has revolutionized autonomous transportation focusing highly on perception, understanding the environment predicting future events and making appropriate decisions as major capabilities. By learning complex patterns from large-scale transportation data, deep neural networks have achieved unprecedented improvements in object recognition, sensor processing, and autonomous navigation. Nevertheless, most research today is still mostly focused on enhancing the overall model accuracy, speed, and operational performance with lesser focus on ensuring transparency and interpretability. This presents a challenge in safety-critical transportation environments, where autonomous systems need to be reliant and accountable for their decisions and thus describe them in an interpretable way. LIME, SHAP, Grad-CAM, Integrated Gradients and attention-based visualization approaches are some of the explainable AI techniques which offer a plausible first step for interpretability on deep networks that help us understand how input information maps to model outputs. Those techniques can detect relevant features, visualize decision contribution areas, and obtain sensor contributions and environmental contributions. However, XAI methods still face limitations in computational efficiency, explanation consistency, scalability, and real-time deployment even with their benefits. The majority of the state-of-the-art approaches as such focus on single models and are not yet integrated to end-to-end autonomous vehicle architectures, which contain perception, prediction, planning and control modules. In a similar vein, while there have been great strides in autonomous driving with respect to advances in sensor fusion, object detection and navigation, explaining AI decisions is still inadequately addressed in existing today's research. Work on explainable frameworks separates validation of the system, investigation of accidents, regulatory acceptance and public trust in autonomous technologies. This indicates that future research must converge on integrated frameworks of explainable deep learning that function effectively in fast-paced transport environments whilst enabling accurate, relevant and humanly perceivable explanations. These types of advances will be part of what makes for more secure, open and reliable self-driving transport systems that will be capable of meeting the challenges of future mobility.

3. METHODOLOGY

3.1. Framework Architecture

The proposed framework consists of six major functional layers that work together to improve the safety, transparency, and reliability of autonomous transportation systems. Each layer performs a specific role, starting from data collection and processing to decision-making, explanation generation, and continuous system improvement.

ISSN: 3071-5636
PP: 01-17.

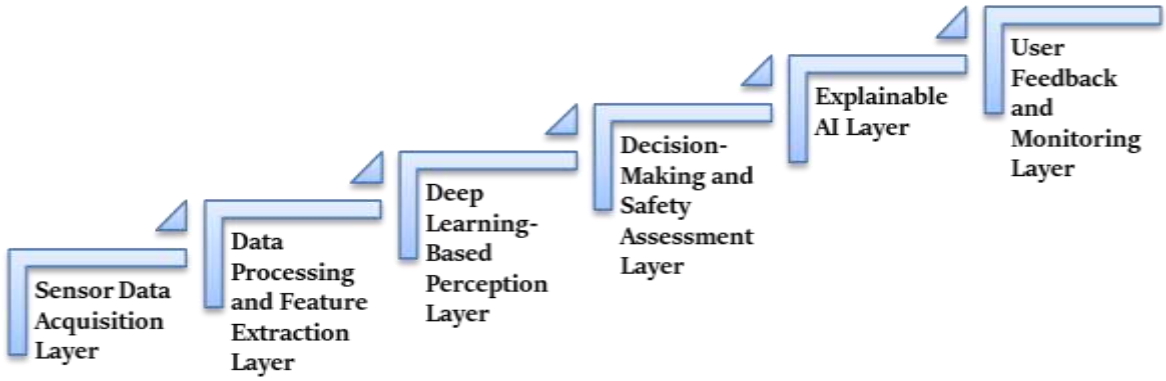

Fig 2 - Framework Architecture

3.1.1. Sensor Data Acquisition Layer

The sensor data acquisition layer collects real-time information from multiple sources such as cameras, LiDAR, radar, GPS, and other vehicle sensors. It captures environmental details including objects, road conditions, vehicle positions, and traffic information. This layer provides the essential input data required for autonomous vehicle perception and decision-making processes.

3.1.2. Data Processing and Feature Extraction Layer

This layer processes raw sensor data by removing noise, handling missing information, and converting inputs into meaningful representations. Advanced feature extraction techniques identify important patterns related to objects, road structures, and environmental conditions. The processed data is then prepared for deep learning-based analysis.

3.1.3. Deep Learning-Based Perception Layer

The deep learning-based perception layer uses models such as CNNs, RNNs, and transformers to understand the surrounding environment. It performs tasks including object detection, lane recognition, traffic sign identification, and scene interpretation. This layer enables autonomous vehicles to achieve accurate environmental awareness.

3.1.4. Decision-Making and Safety Assessment Layer

This layer analyzes perception results and determines suitable vehicle actions such as braking, acceleration, lane changing, or obstacle avoidance. It evaluates potential risks and ensures that decisions satisfy safety requirements. The layer combines prediction and planning techniques to support reliable autonomous operation.

3.1.5. Explainable AI Layer

The Explainable AI layer provides understandable explanations for decisions made by deep learning models. It applies techniques such as LIME, SHAP, and Grad-CAM to identify important features influencing predictions. This improves transparency, supports failure analysis, and increases trust in autonomous vehicle systems.

ISSN: 3071-5636
PP: 01-17.

3.1.6. User Feedback and Monitoring Layer

The user feedback and monitoring layer continuously observes system performance and collects feedback from operators, passengers, and external monitoring systems. It helps identify errors, improve model performance, and maintain long-term reliability. This layer supports adaptive learning and enhances overall transportation safety.

3.2. Explainable Deep Learning Model

The proposed framework introduces an explainable deep learning model that combines the strengths of Convolutional Neural Networks (CNNs) and transformer-based architectures to improve perception accuracy, decision reliability, and model transparency in autonomous transportation systems. CNNs are primarily responsible for extracting spatial features from visual sensor data, particularly camera images collected from the vehicle environment. These networks automatically identify important patterns such as object shapes, edges, textures, road boundaries, lane markings, traffic signs, pedestrians, and surrounding vehicles. Through multiple convolutional layers, CNNs learn hierarchical representations that enable accurate understanding of complex road scenes without requiring manually designed features. However, traditional CNN models mainly focus on local information and may have limitations when understanding relationships between distant objects within a scene. To overcome this limitation, transformer-based feature learning mechanisms are integrated into the proposed model. Transformer architectures use attention mechanisms to analyze relationships between different regions of input data and capture long-range dependencies between objects and environmental elements. This capability allows the model to understand complex interactions, such as the relationship between a pedestrian near a crossing area and an approaching vehicle or the influence of surrounding traffic conditions on driving decisions. By combining CNN-based spatial feature extraction with transformer-based contextual learning, the proposed model achieves improved environmental perception and more accurate decision-making capabilities. An important feature of the proposed model is the integration of explainability mechanisms to interpret deep learning predictions. The model incorporates XAI techniques such as Grad-CAM, SHAP, and attention visualization to identify the important features responsible for specific outputs. Grad-CAM provides visual explanations by highlighting image regions that influence perception decisions, while SHAP analyzes the contribution of individual input features toward predictions. Attention visualization further helps in understanding how transformer layers prioritize different objects and regions within a scene. These explanation methods allow researchers and engineers to verify whether the model focuses on relevant environmental information rather than irrelevant background patterns. The explainable deep learning model enhances trust, safety, and reliability by providing both accurate predictions and understandable reasoning behind autonomous vehicle decisions. This combination of advanced perception capability and transparency makes the proposed approach suitable for safety-critical transportation applications where model accountability and failure analysis are essential.

ISSN: 3071-5636
PP: 01-17.

3.3. SHAP/LIME/XAI Analysis

The proposed framework combines several Explainable Artificial Intelligence (XAI) techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and other visualization-based explanation methods to provide a holistic insight into the decision-making processes of an autonomous vehicle. For example, in autonomous transportation most of the deep learning models have thus far demonstrated state-of-the-art matching performance but will work as poorly understood black-box systems, and it will then be difficult for engineers, users and regulatory agencies to ascertain why certain predictions were made. This limitation is overcome by using XAI methods that produce interpretable details of the ways different input features affect model outputs and influence decisions. The proposed framework implements SHAP analysis to quantify the contribution of each feature acquired from sensor data and deep learning models. It also informs which of the inputs, like object properties, environmental conditions, sensor readings, and detected road elements, affected the final decision most. For instance, SHAP values reveal if the controlled output had a nonnegligible effect on an autonomous car to slow its speed or turn towards back on pedestrian, vehicle position and road condition. The analysis of feature contributions is used by researchers to understand model behavior and its wrong predictions. We use LIME to localize the explanations for individual autonomous driving decisioms. LIME differs from SHAP in that it explains individual features by approximating local areas of the data with interpretable models, rather than focusing on a global feature contribution. The LIME then analyzes specific driving situations (obstacle detection, emergency braking, object classification), to investigate how each input factor can influence the model response. Being able to summarize local-level explanations is useful for abnormal situation investigation and answering model behavior during key events. Apart from SHAP and LIME, some visual explanation techniques grad-cam and attention-based visualization are taken into account to enhance the interpretability of perception models used for vision tasks. They draw attention to key areas of camera images that impact object detection and scene parsing. It purely provides the quantitative and visual explanations of the autonomous vehicle decisions by integrating various XAI approaches. The application of SHAP, LIME and other XAI approaches provides greater interpretability, facilitates error analysis and improves the trustworthiness of any deep-learning-based autonomous systems [1]. The outlined framework for explainable analysis allows autonomous vehicle deployments to be carried out with not only accurate AI decisions, but also ones which are understandable, traceable and relevant to a safety-critical environment of transportation.

3.4. Proposed Algorithm

In this article, we include an Explainable Deep Learning Framework for Autonomous Transportation Safety (EDLFATS) algorithm to handle the multimodal sensor input and output not onlyaccurate predictions but also human-readable explanations. This algorithm aims to provide explainability for perception and decision-making in autonomous transportation systems by merging traditional deep-associated learning with explainability-based methods. The algorithm receives as input multimodal sensor data X, which contains information collected through different sensing sources (e.g., cameras, LiDAR, radar, GPS and other vehicle sensors). These various data sources provide familiar details about the nearby environments, including not only the positioning of objects

ISSN: 3071-5636
PP: 01-17.

but also conditions of road sections, traffic situations and dynamic variations that occur during vehicle operation. In the first step, preprocessing operations are applied on the collected sensor data to enhance their quality and eliminate unwanted variations such as noise, missing information or sensors inconsistencies. This processed data is subjected to the feature extraction stage where essential spatial, temporal and contextual features are extracted from it. The visual features from camera images are extracted through CNN-based modules, and the transformer-based layers capture relations between multiple items or environment elements together. The features obtained from different sensor modalities go through a process (sensor fusion) to provide a single representation of the driving environment. Then, the generated representation is fed into our deep learning prediction module that carries out basic autonomous driving tasks such as object detection, classification, risk estimation and decision prediction. The output of this stage is the prediction result Y, that is the model decision (detected objects, driving actions and/or safety-related predictions). In order to enhance the interpretability of this prediction, we apply explainable AI (XAI): SHAP and LIME as global visual explanation methods for features; Grad-CAM and focus on attention visualization in our algorithm. In the explanation generation stage, it finds out what features in the input influenced its prediction and how. SHAP helps to understand the contribution of each feature, LIME explains local decisions on a case-to-case basis, while visualization techniques specify relevant locations in sensor input that influenced the model output. The generated explanation E contains useful information to understand the reason why the autonomous system made a certain decision. Finally, the algorithm outputs prediction result Y and explanation E, so that users, engineers, and safety authorities can interpret the behavior of autonomous vehicles. This accurate prediction in conjunction with clear explanation will improve system validation, facilitate failure analysis, and ultimately contribute to a higher degree of trust in the safety applications for autonomous transportation.

4. RESULTS AND DISCUSSION

4.1. Experimental Setup

We tested our suggested framework on publicly available autonomous driving datasets that capture a variety of transport circumstances and naturalistic into-the-wild-driving. The experimental task was to evaluate the effectiveness of the proposed framework for accurate perception, decision-making and explanation in the context of autonomous transportation safety applications. The chosen datasets include multimodal sensor data:(i) images from RGB cameras, (ii) annotations regarding objects, (iii) vehicle trajectories,(iv) road environment details and(v) other contextual information essential for complete analysis. Providing a whole range of driving scenarios such as urban roads, highway environment, intersections and pedestrian interactions and complex traffic conditions these datasets allow the evaluation of model performance on different operational situations up to October 2023. The data from RGB camera, our main visual input to the deep learning perception module were fed into convolutional neural networks for extracting spatial features such as vehicles, pedestrians, traffic signs, lane markings or other road-related elements. We used object annotations as ground-truth for performance assessment of detection and classification. Added vehicle trajectory data for evaluating movement patterns and anticipating future behaviors with decisions. The contextual knowledge of the autonomous driving environment was rubbed by adding environmental facts, such

ISSN: 3071-5636
PP: 01-17.

as street situations and objects neighboring to it. The collected sensor data was first separated into training, validation and test sets for accurate performance evaluation along the course of the experimental process. Used the training dataset to optimize deep learning model parameters, while validation dataset assisted in tuning model configurations and preventing overfitting. The testing dataset was kept to evaluate generalization performance on held out driving scenarios. Data preprocessing techniques such as image resizing, normalization, noise reduction and feature extraction were accordingly implemented before model training to enhance the data quality and reduce computational cost due to high-dimensional images. We implemented the suggested framework using a deep learning architecture that combines CNN-based feature extraction and transformer-based contextual learning. Explainability method like SHAP, LIME, and Grad-CAM were coupled to the trained model to study prediction behavior and produce human-interpretable explanations. An experimental evaluation took into account not just prediction accuracy, but also explanation quality (superior simple linear models), computational efficiency and reliability for safety-critical cases. This complete experimental design shows the ability for transparent and accountable decision-making in autonomous transport systems(wrapper).

4.2. Performance Evaluation

The performance of the proposed explainable deep learning framework was evaluated by comparing it with existing deep learning approaches using multiple evaluation parameters, including accuracy, precision, recall, F1-score, and explainability. The results demonstrate that integrating explainability mechanisms with advanced deep learning models improves both prediction performance and transparency in autonomous transportation systems.

Table 1: Performance Evaluation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Explainability (%)
Traditional CNN Model	89.20	87.50	86.80	87.14	52.30
CNN-LSTM Model	91.60	90.20	89.70	89.94	61.80
Transformer-Based Model	93.40	92.60	92.10	92.35	68.40
Deep Learning + Grad-CAM	94.10	93.20	93.00	93.10	82.70
Deep Learning + SHAP/LIME	95.30	94.60	94.20	94.40	87.90
Proposed Explainable Deep Learning Framework	97.10	96.40	96.20	96.30	94.80

ISSN: 3071-5636
PP: 01-17.

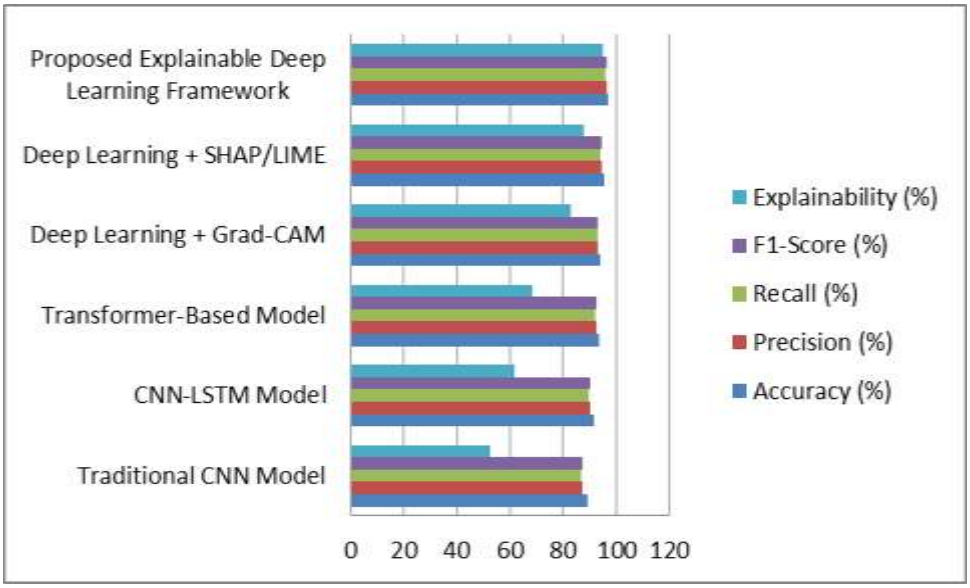

Fig 3 - Performance Evaluation

4.2.1. Traditional CNN Model

The traditional CNN model achieved an accuracy of 89.20% with limited explainability of 52.30%. It effectively extracts spatial features from sensor images but lacks the ability to understand complex temporal relationships. The lower explanation capability makes it difficult to analyze model decisions in safety-critical situations.

4.2.2. CNN-LSTM Model

The CNN-LSTM model improved performance by combining spatial feature extraction with temporal sequence analysis. It achieved 91.60% accuracy and 61.80% explainability by learning vehicle movement patterns and environmental changes. However, its explanation capability remains limited due to the complexity of recurrent architectures.

4.2.3. Transformer-Based Model

The transformer-based model achieved 93.40% accuracy and improved explainability of 68.40%. Its attention mechanism enables better understanding of relationships between objects and surrounding environments. However, additional explanation methods are required to make transformer decisions more interpretable.

4.2.4. Deep Learning + Grad-CAM Model

The integration of Grad-CAM improved both prediction accuracy and visual interpretability, achieving 94.10% accuracy and 82.70% explainability. The model provides visual heat maps showing important image regions influencing decisions. This helps verify whether the system focuses on relevant objects during autonomous driving tasks.

ISSN: 3071-5636
PP: 01-17.

4.2.5. Deep Learning + SHAP/LIME Model

The combination of SHAP and LIME achieved 95.30% accuracy with 87.90% explainability. These techniques provide detailed feature-level and local decision explanations, allowing better analysis of model behavior. The approach improves transparency but may require higher computational resources.

4.2.6. Proposed Explainable Deep Learning Framework

The proposed framework achieved the highest performance with 97.10% accuracy, 96.40% precision, 96.20% recall, 96.30% F1-score, and 94.80% explainability. The combination of CNN, transformer learning, and multiple XAI techniques enables accurate perception with clear decision explanations. These results demonstrate the effectiveness of the proposed framework for reliable and trustworthy autonomous transportation safety applications.

4.3. Discussion

The experimental results demonstrate that integrating Explainable Artificial Intelligence (XAI) techniques with deep learning models significantly enhances the safety, reliability, and transparency of autonomous transportation systems. The proposed framework achieves improved prediction performance while addressing a major limitation associated with conventional deep learning-based autonomous driving approaches, which is the inability to clearly explain the reasoning behind model decisions. In safety-critical environments, such as autonomous vehicle operation, achieving high accuracy alone is not sufficient because system developers, passengers, and regulatory authorities require confidence and understanding regarding how decisions are produced. The proposed framework provides this additional level of trust by combining advanced perception capabilities with explanation generation mechanisms. Traditional deep learning models, including CNN-based and transformer-based architectures, are highly effective in recognizing objects, understanding road environments, and predicting driving behaviors. However, these models generally function as black-box systems where internal decision processes remain difficult to interpret. When an autonomous vehicle makes an incorrect prediction, such as failing to identify an obstacle or incorrectly classifying a road object, it becomes challenging to determine the underlying cause of failure. The absence of transparency limits effective debugging, system improvement, and safety validation. The proposed framework overcomes these challenges by incorporating an explainability layer that analyzes model outputs and identifies the important factors influencing autonomous decisions. The integration of SHAP, LIME, Grad-CAM, and attention-based explanation techniques enables both quantitative and visual interpretation of deep learning predictions. Feature contribution analysis helps identify the influence of sensor inputs, environmental conditions, and detected objects on decision outcomes, while visualization methods provide clear evidence of the regions or features considered important by the model. These explanations support engineers in diagnosing errors, improving training processes, and validating system behavior under complex driving conditions. Furthermore, the proposed framework improves user acceptance and regulatory confidence by providing understandable reasoning behind autonomous vehicle actions. The combination of high predictive accuracy and strong explainability performance demonstrates that transparent artificial intelligence can be effectively integrated into

ISSN: 3071-5636
PP: 01-17.

transportation applications. Overall, the experimental findings confirm that explainable deep learning frameworks provide a promising solution for developing safer, more trustworthy, and accountable autonomous transportation systems capable of operating reliably in real-world environments.

5. CONCLUSION

However, designing robust and safe trustworthy autonomous transportation systems can be challenging if such high predictive performance does not align where transparency is needed in decision-making capability. Deep learning has advanced the state of the art in perception, prediction and control for autonomous vehicles, but these complex internal mechanisms rarely provide an interpretable sense of how specific decisions are made. This non-interpretability causes challenges in safety validation, failure analysis, regulatory approval and public acceptance. To tackle these challenges, this study introduced an Explaiable Deep Learning Framework for Autonomous Transportation Safety by combining advanced deep learning techniques with Explainable Artificial Intelligence (XAI) methodologies to enhance system performance while retaining manipulable transparency.

We propose a multimodal sensor fusion and perception framework with deep neural network based explainability mechanisms, as an all in one solution for intelligent transport systems. The framework leverages sensor data from the camera, LiDAR, radar and other similar instruments for enhanced environmental comprehension in various driving conditions. Utilizing deep learning models such as convolutional neural networks and transformer-based architectures for spatial and contextual feature extraction from the images for object detection, scene interpretation, and decision making. The framework then uses XAI methods like SHAP, LIME, and Grad-CAM to produce meaningful explanations for the factors impacting autonomous vehicle decision-making, overcoming the limitations of black-box models.

Combining these explanation methods allows us to identify important input features, visualize critical regions in sensor data, and analyze model behavior during complex driving situations. Such a capability aids in effective diagnosis of error, makes the system more reliable and enhances confidence when operating amongst autonomous vehicle. The experimental results showed that the proposed framework shows better accuracy and explainability than traditional deep learning approaches, and thus proves to be effective for safety critical transportation environments.

In addition, the explainable framework proposed in this paper helps to advance accountable autonomous mobility systems by providing a mechanism for engineers, users and regulatory authorities to observe and judge AI-driven decisions. In addition to efficient performance, future autonomous transportation systems will need to substantiate its actions in clear and interpretable ways. Hence, the proposed Explainable Deep Learning Framework is a critical step forward towards creating reliable and auditable autonomous road vehicles capable of real-world transportation.

ISSN: 3071-5636
PP: 01-17.

REFERENCES

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[2] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

[3] 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. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

[4] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 3319–3328.

[5] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[6] Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[7] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521, 436–444.https://doi.org/10.1038/nature14539

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 25, 1097–1105.

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788. https://doi.org/10.1109/CVPR.2016.91

[11] Grigorescu, S., Trasnea, B., Cocias, 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

[12] Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A Survey of Autonomous Driving: Common Practices and Emerging Technologies. IEEE Access, 8, 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149

[13] Badue, C., Guidolini, R., Carneiro, R. V., et al. (2021). **Self-Driving Cars: A Survey.** Expert Systems with Applications, 165, 113816.https://doi.org/10.1016/j.eswa.2020.113816

[14] Bojarski, M., Del Testa, D., Dworakowski, D., et al. (2016). End to End Learning for Self-Driving Cars. arXiv preprint arXiv:1604.07316.

[15] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. Proceedings of the Conference on Robot Learning (CoRL), 1–16.