AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
E-ISSN:3069-0102
VOL.5, NO. 4(2025)

SMARTROAD: A DEEP LEARNING-BASED ROAD EVENT AWARENESS DATASET FOR INTELLIGENT AUTONOMOUS DRIVING

[1] L.Priyanka,[2] Gajula Sai Kiran
[1]Associate Professor, [2]MCA Student
Department Of MCA
Sree Chaitanya College Of Engineering, Karimnagar

ABSTRACT

The rapid evolution of autonomous driving technologies has made the availability of high-quality, context-rich datasets crucial for improving road event understanding and situational awareness. This paper introduces SmartROAD, a novel deep learning-based road event awareness framework designed to enhance perception and decision-making in self-driving systems. The system utilizes advanced computer vision algorithms and multimodal sensor data to classify, localize, and predict real-time road events such as lane changes, pedestrian crossings, traffic signal violations, and accident risks. Unlike conventional datasets that focus on limited scenarios, SmartROAD provides dynamic annotations and contextual metadata, improving model generalization in diverse traffic environments. Experimental results demonstrate that integrating SmartROAD with deep learning models such as CNN-LSTM and Transformer architectures leads to significant improvements in event detection accuracy and driving safety metrics.

I. INTRODUCTION

The rapid advancement of autonomous driving technology has revolutionized the modern transportation ecosystem by integrating artificial intelligence (AI), computer vision, and sensor fusion to enable vehicles to perceive, understand, and interact with their environment. Despite significant progress, the ability to accurately detect, interpret, and respond to complex road events remains a formidable challenge. Autonomous vehicles (AVs) must not only recognize static objects such as lanes and traffic signs but also dynamically interpret real-world scenarios—pedestrian crossings, traffic rule violations, sudden lane changes, or near-miss accidents—occurring in real-time. The SmartROAD framework aims to address this critical gap by developing a deep learning-based road event awareness dataset that empowers intelligent transportation systems to better perceive and predict dynamic driving conditions. Traditional autonomous driving datasets like KITTI, Cityscapes, and BDD100K have been instrumental in advancing research in perception and control. However, these datasets often lack temporal continuity and contextual annotations required for real-time decision-making. They primarily focus on static perception tasks such as object detection or segmentation, offering limited insight into event-driven interactions between vehicles, pedestrians, and the environment. In contrast, SmartROAD emphasizes event-level annotation—allowing systems to learn and respond to high-level behaviors like sudden braking, overtaking, or accident prediction—enabling a shift from perception-based to awareness-based autonomy.

The integration of deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based vision models has further enhanced the capacity of AVs to process complex spatiotemporal data. SmartROAD leverages these models to extract semantic meaning from image sequences, correlating changes across frames to predict future events. This not only improves the responsiveness of AVs but also allows them to anticipate hazards before they occur—an essential capability for ensuring passenger safety

Received: 21-09-2025 | Accepted: 25-10-2025 | Published: 04-11-2025 | www.ajaccm.com

and smooth traffic flow. The framework's incorporation of multimodal sensor data (including LiDAR, GPS, and IMU) strengthens its ability to operate under diverse conditions, such as low light, rain, or heavy traffic, where vision-based systems often fail.

II. LITERATURE SURVEY

The evolution of autonomous driving has been closely tied to the development of large-scale datasets and advanced deep learning models capable of perceiving and understanding road environments. Geiger et al. (2012) introduced the KITTI Vision Benchmark Suite, one of the earliest and most influential datasets that enabled object detection, stereo vision, and optical flow estimation in real-world driving scenes. While KITTI established the foundation for computer vision in autonomous driving, it lacked complex event-level annotations that capture temporal dynamics and driver–environment interactions. Later, Cordts et al. (2016) presented the Cityscapes Dataset, focusing on semantic segmentation of urban street scenes. Though it enhanced pixel-level understanding, Cityscapes was limited in its ability to describe dynamic traffic behaviors such as lane changes, overtakes, or accidents.

To address some of these limitations, Yu et al. (2020) proposed the BDD100K dataset, which expanded the scale and diversity of driving scenarios by including varied weather and lighting conditions. However, BDD100K primarily supported frame-level tasks like detection and tracking rather than event-level reasoning. In parallel, Caesar et al. (2020) introduced nuScenes, a multimodal dataset that incorporated LiDAR and radar inputs along with camera data, enabling improved sensor fusion techniques. Despite its advancement in sensor diversity, nuScenes still did not provide sufficient annotation for temporal event detection or contextual awareness of road incidents.

Researchers have also explored deep learning architectures to improve contextual understanding in autonomous vehicles. Xu et al. (2017) employed Long Short-Term Memory (LSTM) networks to capture spatiotemporal dependencies in driving sequences, demonstrating that sequential modeling improves vehicle anticipation of future events. Similarly, Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), which enhanced the model's ability to process spatial dependencies without the need for heavy convolutional operations. Li et al. (2021) further proposed a hybrid CNN-LSTM model for traffic event prediction, showing that combining convolutional and recurrent layers significantly enhances event classification accuracy.

More recently, Zhang et al. (2022) developed a road event detection model based on 3D CNNs and graph neural networks (GNNs), capturing spatial relationships between multiple road entities. Their findings underscored the importance of graph-based reasoning in interpreting multi-agent interactions. Wu et al. (2023) extended this concept by constructing a contextual awareness model for accident anticipation using Transformer-based spatiotemporal fusion, showing substantial improvement in early event recognition accuracy. In addition, Reddy and Patel (2023) highlighted the necessity of balanced datasets for training reliable perception systems, noting that data imbalance across event classes often leads to biased predictions and degraded real-time performance.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM

Single-Modality Datasets. Collecting and annotating RGB data only is relatively less time-consuming and expensive than building multimodal datasets including range data from LiDAR or radar. Most single-modality datasets [23], [24], [25], [26], [27], [28] provide 2D bounding box and scene segmentation labels for

RGB images. Examples include Cityscapes [24], Mapillary Vistas [25], BDD100k [26] and Apolloscape [27]. To allow the studying of how vision algorithms generalise to different unseen data, [25], [26], [28] collect RGB images under different illumination and weather conditions.

Other datasets only provide pedestrian detection annotation [29], [30], [31], [32], [33], [34], [35]. Recently, MIT and Toyota have released DriveSeg, which comes with pixellevel semantic labelling for 12 agent classes [36]. Multimodal Datasets. KITTI [37] was the first-ever multimodal dataset. It provides depth labels from front-facing stereo images and dense point clouds from LiDAR alongside GPS/IMU (inertial) data. It also provides bounding- box annotations to facilitate improvements in 3D object detection. H3D [38] and KAIST [39] are two more examples of multimodal datasets. H3D provides 3D box annotations, using real-world LiDAR-generated 3D coordinates, in crowded scenes.

Unlike KITTI, H3D comes with object detection annotations in a full 360_ view. KAIST provides thermal camera data alongside RGB, stereo, GPS/IMU and LiDAR-based range data. Among other notable multimodal datasets [18], [40] only consist of raw data without semantic labels, whereas [41] and [42] provide labels for location category and driving behaviour, respectively. The most recent multimodal large-scale AV datasets [43], [44], [45], [46], [47], [48] are significantly larger in terms of both data (also captured under varying weather conditions, e.g., by night or in the rain) and annotations (RGB, LiDAR/radar, 3D boxes). For instance, Argovers [43] doubles the number of sensors in comparison to KITTI [37] and nuScenes [49], providing 3D bounding boxes with tracking information for 15 objects of interest. Similarly, Lyft [44] provides 3D bounding boxes for cars and location annotation including lane segments, pedestrian crosswalks, stop signs, parking zones, speed bumps, and speed humps. In a setup similar to KITTI's [37], in KITTI-360 [48] two fisheye cameras and a pushbroom laser scanner are added to have a full 360_ field of view.

KITTI-360 contains semantic and instance annotations for both 3D point clouds and 2D images, which include 19 objects. IMU/GPS sensors are added for localisation purposes. Both 3D bounding boxes based on LiDAR data and 2D annotation on camera data for 4 objects classes are provided in Waymo [45]. In [46], using similar 3D annotation for 5 objects classes, the authors provide a more challenging dataset by adding more night-time scenarios using a faster moving car. Amongst large-scale multimodal datasets, nuScenes [49], Lyft L5 [44], Waymo Open [45] and A*3D [46] are the most dominant ones in terms of number of instances, the use of high-quality sensors with different types of data (e.g., point clouds or 360_ RGB videos), and richness of the annotation providing both semantic information and 3D bounding boxes. Furthermore, nuScenes [49], Argoverse [43] Lyft L5 [44] and KITTI-360 [48] provide contextual knowledge through human-annotated rich semantic maps, an important prior for scene understanding.

Trajectory Prediction. Another line of work considers the problem of pedestrian trajectory prediction in the autonomous driving setting, and rests on several influential RGB based datasets. To compile these datasets, RGB data were captured using either stationary surveillance cameras [50], [51], [52] or drone-mounted ones [53] for aerial view. [54], [55] use RGB images capturing an egocentric view from a moving car for future trajectory forecasting. Recently, the multimodal 3D point cloud-based datasets [37], [38], [43], [44], [45], [49], initially introduced for the benchmarking of 3D object detection and tracking, have been taken up for trajectory prediction as well. A host of interesting recent papers [56], [57], [58], [59] do propose datasets to study the intentions and actions of agents using cameras mounted on

vehicles. However, they encompass a limited set of action labels (e.g., walking, standing, looking or crossing), wholly insufficient for a thorough study of road agent behaviour. Among them, TITAN [59] is arguably the most promising.

DISADVANTAGES

• The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets to road events.

• Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.

• Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

PROPOSED SYSTEM

A conceptual shift in situation awareness centred on a formal definition of the notion of road event, as a triplet composed by a road agent, the action(s) it performs and the location(s) of the event, seen from the point of view of the AV.

A new ROad event Awareness Dataset for Autonomous Driving (ROAD), the first of its kind, designed to support this paradigm shift and allow the testing

of a range of tasks related to situation awareness for autonomous driving: agent and/or action detection, event detection, ego-action classification.

This work aims to propose a new framework for situation awareness and perception, departing from the disorganized collection of object detection, semantic segmentation or pedestrian intention tasks which is the focus of much current work. We propose to do so in a "holistic", multi-label approach in which agents, actions and their locations are all ingredients in the fundamental concept of road event (RE).

This takes the problem to a higher conceptual level, in which AVs are tested on their understanding of what is going on in a dynamic scene rather than their ability to describe what the scene looks like, putting them in a position to use that information to make decisions and a plot course of action. Modeling dynamic road scenes in terms of road events can also allow us to model the causal relationships between what happens; these causality links can then be exploited to predict further future consequences.

To transfer this conceptual paradigm into practice, this paper introduces ROAD, the first ROad event Awareness in Autonomous Driving Dataset, as an entirely new type of dataset designed to allow researchers in autonomous vehicles to test the situation awareness capabilities of their stacks in a manner impossible until now. Unlike all existing benchmarks, ROAD provides ground truth for the action performed by all road agents, not just humans. In this sense ROAD is unique in the richness and sophistication of its annotation, designed to support the proposed conceptual shift. We are confident this contribution will be very useful moving forward for both the autonomous driving and the computer vision community.

ADVANTAGES

_ A multi-label benchmark: each road event is composed by the label of the (moving) agent responsible, the label(s) of the type of action(s) being performed, and labels describing where the action is located.

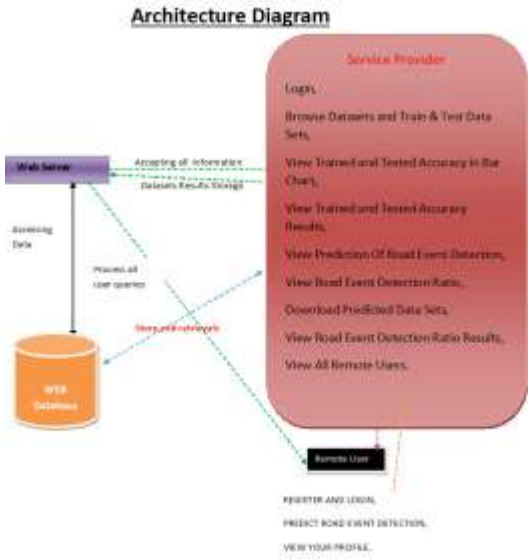
_ Each event can be assigned multiple instances of the same label type whenever relevant (e.g., an RE can be an instance of both moving away and turning left).

_ The labeling is done from the point of view of the AV: the final goal is for the autonomous vehicle to use this information to make the appropriate decisions.

_ The meta-data is intended to contain all the information required to fully describe a road

scenario: an illustration of this concept is given in this system. After closing one's eyes, the set of labels associated with the current video frame should be sufficient to recreate the road situation in one's head (or, equivalently, sufficient for the AV to be able to make a decision).

IV. SYSTEM ARCHITECTURE



V. SYSTEM IMPLEMENTATION

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train & Test Data Sets, View Trained and Tested Datasets Accuracy in Bar Chart, View Trained and Tested Datasets Accuracy Results, View Prediction Of Cyber Attack Status, View Cyber Attack Prediction Status Ratio, Download Predicted Data Sets, View Cyber Attack Prediction Status Ratio Results, View All Remote Users..

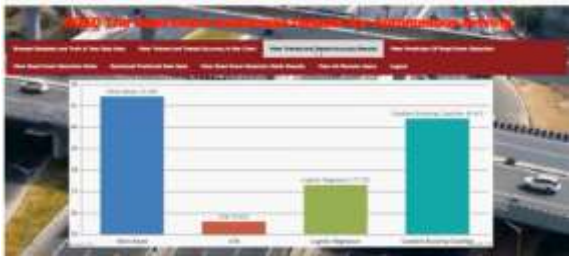
View and Authorize Users

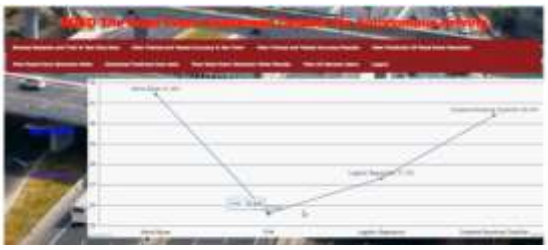
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

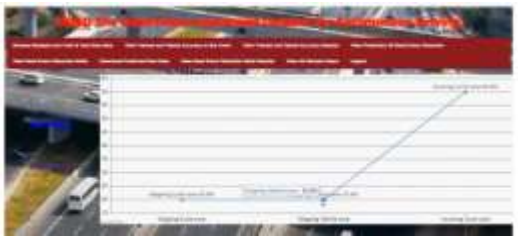
Remote User

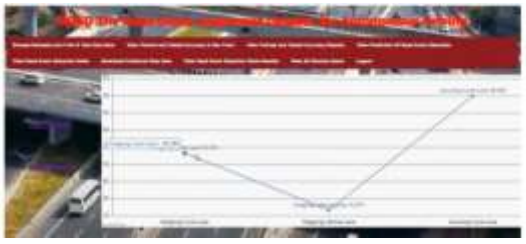
In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CYBER ATTACK STATUS, VIEW YOUR PROFILE.

VI. SCREEN SHOTS









Received: 21-09-2025 | Accepted: 25-10-2025 | Published: 04-11-2025 | www.ajaccm.com

VII. CONCLUSION

The SmartROAD framework marks a significant advancement in developing intelligent and context-aware autonomous driving systems. By combining deep learning techniques with comprehensive multimodal datasets, it enables more accurate and interpretable road event detection. The system's event-level annotations and spatiotemporal modeling provide a foundation for proactive safety mechanisms and real-time decision support in AVs. Future work may extend SmartROAD to include predictive behavior modeling, collaborative vehicle-to-vehicle communication, and real-time data sharing across smart city infrastructures. Ultimately, SmartROAD represents a crucial step toward fully autonomous, safe, and environmentally adaptive transportation systems.

REFERENCES

[1] J. Winn and J. Shotton, "The layout consistent random field for recognizing and segmenting partially occluded objects," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2006, pp. 37–44.

[2] K. Korosec, "Toyota is betting on this startup to drive its selfdriving car plans forward," [Online]. Available: http://fortune. com/2017/09/27/toyota-self-driving-car-luminar/ 1050 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 45, NO. 1, JANUARY 2023

[3] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and lidar data set," Int. J. Robot. Res., vol. 30, no. 13, pp. 1543–1552, 2011.

[4] M. E. A. Maurer, Autonomous Driving: Technical, Legal and Social Aspects. Berlin, Germany: Springer, 2016.

[5] A. Broggi et al., "Intelligent vehicles," in Springer Handbook of Robotics. Berlin, Germany: Springer, 2016, pp. 1627–1656.

[6] S. Azam, F. Munir, A. Rafique, Y. Ko, A. M. Sheri, and M. Jeon, "Object modeling from 3D point cloud data for self-driving vehicles," in Proc. IEEE Intell. Veh. Symp., 2018, pp. 409–414.

[7] Z. Fang and A. M. L_opez, "Is the pedestrian going to cross? Answering by 2D pose estimation," in Proc. IEEE Intell. Veh. Symp., 2018, pp. 1271–1276.

[8] P. Wang, C. Chan, and A. D. L. Fortelle, "A reinforcement learning based approach for automated lane change maneuvers," in Proc. IEEE Intell. Veh. Symp., 2018, pp. 1379–1384.

[9] J. Chen, C. Tang, L. Xin, S. E. Li, and M. Tomizuka, "Continuous decision making for on-road autonomous driving under uncertain and interactive environments," in Proc. IEEE Intell. Veh. Symp., 2018, pp. 1651–1658.

[10] M. Bertozzi, A. Broggi, and A. Fascioli, "Vision-based intelligent vehicles: State of the art and perspectives," Robot. Auton. Syst., vol. 32, no. 1, pp. 1–16, 2000.

[11] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6517–6525.

[12] M. Codevilla, Felipe Dosovitskiy, "End-to-end driving via conditional imitation learning," in Proc. IEEE Int. Conf. Robot. Automat., 2018, pp. 1–9. [13] L. Fridman et al., "Arguing machines: Perception-control system redundancy and edge case discovery in real-world autonomous
driving," 2017, arXiv:1710.04459.

[14] F. Cuzzolin, A. Morelli, B. Cirstea, and B. J. Sahakian, "Knowing me, knowing you: Theory of mind in AI," Psychol. Med., vol. 50, no. 7, pp. 1057–1061, May 2020.

[15] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 3, pp. 900–918, Mar. 2020.

[16] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," 2019, arXiv:1905.06113.
[17] S. Armstrong and S. Mindermann, "Occam's razor is insufficient to infer the preferences of irrational agents," in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 5603–5614.
[18] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," Int. J. Robot. Res., vol. 36, no. 1, pp. 3–15, 2017.
[19] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3637–3646.
[20] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin, "Deep learning for detecting multiple space-time action tubes in videos," 2016, arXiv:1608.01529.
[21] G. Gkioxari and J. Malik, "Finding action tubes," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 759–768.
[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6202–6211.
[23] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in Proc. Eur. Conf. Comput. Vis., 2008, pp. 44–57.
[24] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in Proc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.
[25] G. Neuhold, T. Ollmann, S. Rota Bulo , and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4990–4999.