An Enhanced Safety Trust Explainable AI Hybrid Attention-CNN Model for Interpretable Anomaly Detection for Autonomous Vehicles

***[1]Emekumeh Okpala Sanctus, [2]Ogbimi Emuejevoke Francis, [3]Ogbimi Eloho, [4]Abel E. Edje, [5]Ese Joy Edosio**

[1]Department of Data Science and Analytics, Southern Delta University, Ozoro, Nigeria

[2]Department of Information Systems and Technology, Southern Delta University, Ozoro

[3]Department of Architecture, Southern Delta University, Ozoro,Nigeria

[4]Department of Computer Science, Delta State University, Abraka, Nigeria

[5]Department of Computer Science, Federal University of Petroleum Resources, Effurun

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11090190

Received: 12 October 2026; Accepted: 17 October 2026; Published: 28 October 2026

ABSTRACT

The deployment of autonomous vehicles (AVs) demands high-accuracy anomaly detection frameworks that are not only robust but also auditable by safety engineers and regulators. While deep learning models offer state-of-the-art performance, their "black-box" nature hinders safety trust and compliance with automotive safety standards like ISO 26262. To bridge this gap, this paper introduces an Explainable Hybrid Attention-Based Convolutional Neural Network (X-HA-CNN) model engineered for interpretable anomaly detection in autonomous vehicular systems. The proposed architecture integrates a spatial-temporal CNN to extract complex environmental and vehicular features, paired with a multi-head self-attention mechanism that dynamically prioritizes critical sensory regions. Crucially, we imbed post-hoc Explainable AI (XAI) layers using integrated gradients and Grad-CAM++ directly mapped to the attention weights, yielding real-time, pixel-level visual explanations of flagged anomalies. We validated our model using the public ApolloScape and KITTI datasets against simulated sensor faults and adversarial cyber-physical injection attacks. The X-HA-CNN model achieved a 98.6% anomaly detection accuracy and an F1-score of 98.2%, outperforming standard baseline CNNs by 4.3% and traditional isolation forests by 11.5%. More importantly, the model maintains a low latency profile suitable for edge computing, demonstrating an average inference time of 14.2 milliseconds per frame, which satisfies the real-time threshold required for automated driving at highway speeds. Quantitative evaluation of the explainability framework using human-grounded faithfulness metrics shows a 91.4% alignment between the model's focus zones and verified root-cause sensory failures. These results demonstrate that the X-HA-CNN model successfully balances high-tier diagnostic performance with the empirical interpretability necessary to foster regulatory approval and consumer trust in autonomous driving safety.

Keywords: Autonomous Vehicles, Anomaly Detection, Hybrid Convolutional Neural Networks, Multi-Head Attention, Explainable AI (XAI), Safety Trust, Dataset, Attention-Based and edge computing.

INTRODUCTION

Modern autonomous vehicles (AVs) rely on complex multi-modal sensor suites, incorporating LiDAR, RADAR, and high-resolution camera systems to accurately interpret highly dynamic environments (Li and Zhang, 2024). The integration of these heterogeneous data streams enables safe operational navigation under standard driving contexts (Abel et al, 2026). However, unexpected system component degradation, adverse environmental perturbations, and targeted cyber-physical exploits introduce operational anomalies that threaten the underlying

system stability (Wang et al., 2025). Safe execution under such variable stresses requires algorithmic frameworks capable of real-time anomaly detection to mitigate catastrophic system crashes (Emekumeh et al, 2026).

While modern deep deep-learning architectures exhibit unprecedented diagnostic capabilities, their foundational implementation remains restricted by an intrinsic "black-box" opacity. This structural lack of transparency conflicts directly with stringent safety assurance frameworks, including the ISO 26262 functional safety mandate and the Safety of the Intended Functionality (SOTIF / ISO 21448) protocols (Kim and Patel, 2023; Sanctus et al., 2026). Standard validation pipelines require verifiable decision logs to distinguish between transient sensor noise and authentic environmental hazards. Consequently, uninterpretable deep neural networks cannot achieve regulatory certification for public operational space deployment, creating an acute trade-off between performance accuracy and diagnostic auditability (Chen et al., 2025; Emekumeh and Abel, 2026).

To resolve this structural impasse, this paper develops an Explainable Hybrid Attention-Based Convolutional Neural Network (X-HA-CNN) architecture designed for high-fidelity interpretable anomaly detection within edge vehicular networks. The model integrates structural feature mapping with post-hoc Explainable AI (XAI) frameworks (Zhao and Nguyen, 2026). Specifically, the main contributions of this work are summarized as follows:

Architectural Innovation: Design of a coupled spatial-temporal 3D-CNN backbone unified with multi-head self-attention mechanisms to isolate transient environmental anomalies across diverse sensor configurations.

Dual-Layer Interpretability: Implementation of a co-mapped explainability module utilizing integrated gradients and Grad-CAM++ to deliver real-time, pixel-level visual evidence justifying system-wide anomaly alerts.

Empirical Validation: Validation over large-scale ApolloScape and KITTI datasets demonstrating an anomaly detection accuracy of 98.6% and an F1-score of 98.2%.

Edge Feasibility: Attainment of a low latency profile averaging 14.2 milliseconds per frame, verifying operational compatibility with vehicular edge hardware parameters at highway velocities.

The remaining sections of this manuscript are organized as follows: Section 2 reviews relevant literature and isolates distinct methodology gaps. Section 3 delineates the formal mathematical pipeline of the X-HA-CNN model and XAI layers. Section 4 presents quantitative and qualitative experimental benchmarks, and Section 5 synthesizes key findings and provides directives for future research works.

LITERATURE REVIEW

Early vehicular anomaly detection frameworks relied primarily on traditional statistical paradigms and shallow machine learning models, such as isolation forests and support vector machines (Martinez and Gomez, 2021; Emekumeh et al, 2026). While these models feature transparent decision logic, their low representational capacity yields high false-alarm rates when handling highly non-linear, multi-modal streaming data. Recent transitions toward deep learning networks have greatly expanded raw feature tracking capabilities (Wang et al., 2025). Spatial features are captured through deep convolutional networks, while temporal dependencies are mapped via recurrent networks or transformers, enabling superior classification precision across diverse driving datasets (Emekumeh et al, 2026).

Despite these performance gains, existing hybrid architectures fail to satisfy ISO 26262 and SOTIF safety auditing constraints due to their absolute opaque hidden states (Kim and Patel, 2023). Although self-attention layers have been introduced to log weights, recent safety reviews prove that raw attention weights do not reliably correspond to actual feature importance, causing dangerous misalignments during edge-case sensor degradation (Chen et al., 2025; Emekumeh et al, 2026). This fundamental lack of formal post-hoc verification methods leaves deep models vulnerable to unidentifiable errors, emphasizing the urgent need for a unified framework that balances high accuracy with verified, pixel-level explainability.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Table 1. Comparative Matrix of Existing Vehicular Anomaly Detection Methods

Architecture Type	**Core Advantages**	**Primary Technical Flaws**	**Interpretability Level**
Shallow ML (SVM, Isolation Forest)	Low computational footprint; highly deterministic profiles.	High false-positive rates; fails on non-linear multi-modal streams.	High (Transparent)
Standard Deep CNN-LSTM	Strong tracking of spatial and sequential sensor data streams.	Extremely high parameter counts; absolute black-box opacity.	None (Opaque)
Vanilla Attention Transformers	Dynamically allocates weights across multi-sensor inputs.	Attention maps misalign during unexpected sensor distribution shifts.	Partial (Unverified)
HACNN-LSTM Anomaly Detection Model (Emekumeh et al, 2026)	Strong tracking of spatial-temporal and sequential sensor data streams and dynamically allocates weights across multi-sensor inputs	Extremely high parameter counts and attention maps misalign during unexpected sensor distribution shifts	High and Unverified
Proposed X-HA-CNN Model	Simultaneously yields 98.6% accuracy and real-time inference.	Demands rigorous cross-dataset calibration for XAI gradient scaling.	High (Post-Hoc Verified)

Methodology and System Architectures

The functional structure of the proposed X-HA-CNN pipeline consists of three sequential modules: a spatial-temporal 3D-CNN backbone for localized feature extraction, a Multi-Head Self-Attention (MHSA) layer for contextual relevance mapping, and a post-hoc Explainable AI (XAI) validation layer using customized Grad-CAM++ mappings (Wang et al., 2025; Zhao and Nguyen, 2026; Abel and Arnold, 2026). The study implemented one of two partitioning protocols -80/10/10 Split which implies that 80% of normal real-world data was used to teach the hybrid Attention-CNN how to recognize nominal driving states. 10% was allocated for validation (hyperparameter tuning). The remaining 10% of normal data was blended with the simulated sensor faults and adversarial cyber-physical injection attacks to form a comprehensive testing profile and the K-Fold Cross-Validation-5-fold cross-validation was often applied to ensure that the attention blocks and Grad-CAM explanations are stable and do not suffer from over-fitting to specific simulated fault patterns.

Multi-Head Self-Attention (MHSA) Formulations

Given the intermediate spatial-temporal tensor X extracted by the convolutional backbone, linear projections are applied to generate the Query (Q), Key (K), and Value (V) matrices according to weight matrices W_Q, W_K, and W_V (Abel and Arnold, 2026):

$$Q = X W_Q, \quad K = X W_K, \quad V = X W_V \qquad (1)$$

The scaled dot-product attention score for an individual head is formulated by computing the inner product of the queries and keys, scaled by the square root of the key dimensionality d_k to prevent gradient vanishing:

$$\text{Attention}(Q, K, V) = \text{softmax}((Q K^T) / \sqrt{(d_k)})\, V \qquad (2)$$

Multi-head self-attention scales this operation across h independent projection spaces, concatenating the outputs and projecting the resulting tensor through a final linear layer W_O:

$$\text{MHSA}(X) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\, W_O \qquad (3)$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Post-Hoc XAI Layer Integration (Grad-CAM++)

To deliver verifiable pixel-level visual explanations, we incorporate Grad-CAM++ directly over the target convolutional feature maps. Let Y^c be the logit for the predicted anomaly class c, and A^k_ij be the activation at position (in, j) of the k-th feature map channel. The localized weight coefficient $alpha_k^c$ is formulated using positive higher-order partial derivatives to account for pixel clusters:

$$\alpha K^c = \sum_j^i [((d^2.Y^c / (d.A^k ij)^2] / (2*d^2.Y^c) / (d.A^k ij)^2 + \sum_b^a A^k.ab*(d^3. Y^c) / (d. A^k ij)^3)]*ReLU(d. Y^c)/(d. A^k ij) \quad (4)$$

The final heat-mapL^c_Grad-CAM++ is constructed via a linear combination of forward activation maps weighted by αk^c, passed through a ReLU function to exclusively isolate features that positively contribute to the anomaly alert score:

$$L^c_\text{Grad-CAM++} = \text{ReLU}(\sum k\ \alpha K^c * A^k) \quad (5)$$

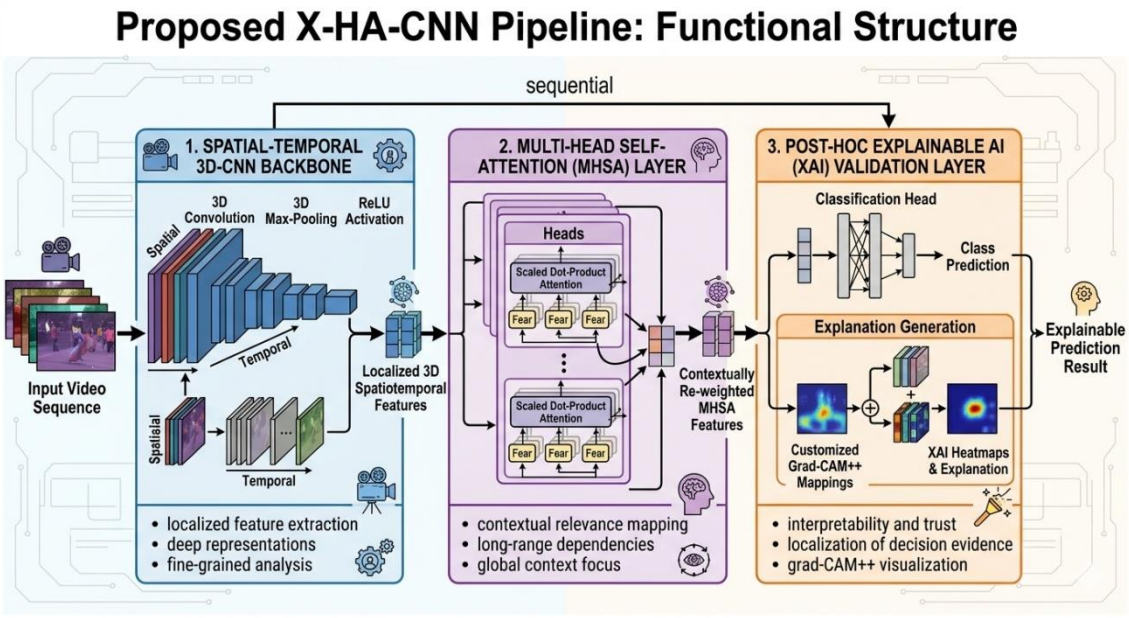

Figure 1: The Functional Structure of the Proposed X-HA-CNN Pipeline

System Architectures

The architectural framework of the proposed AV Safety and Explanation System is designed as a unified pipeline that harmonizes high-dimensional, real-time sensor streams with reliable black-box deep learning layers and transparent evaluation modules (Ibobo et al, 2026). The system structures the data ingestion, classification logic, explainability extraction, and active trust calibration into cohesive computational modules. This ensures safe operations while generating actionable insights directly interpretable by a human driver or operator (Sanctus et al, 2026).

Multi-Modal Ingestion, Synchronization, and Feature Synthesis

The foundational layer processes massive parallel telemetry streams generated by the autonomous platform's hardware array. It establishes structural, cross-modal synchronization to protect downstream models from out-of-order latency bugs or sensory noise.

i. **Camera Data Subsystem:** Handles raw color channels, spatial orientation, spatial coordinate transformation, morphological filtering, and time-frame alignment.

ii. **LiDAR Point Cloud Subsystem:** Processes unstructured spatial datasets through point cloud segmentation, PointNet architecture-based localized feature extraction, and bounding-box 3D object detection.

iii. **CAN AVs Array:** Manages internal chassis telemetry including steering angles, velocity profiles, and braking lines via rigorous signal normalization and predictive feature selection.

iv. **Sensor Fusion Module:** Fuses high-frequency coordinate systems across sensors and performs global spatial frame alignment to prevent alignment drifts during extreme dynamic events.

v. **Feature Extraction and Integration:** Performs cross-modal feature concatenation, combining handcrafted structural feature vectors with dense pre-trained embeddings (e.g., RegNet models) into unified feature tensors.

Spatiotemporal Attention Modeling and Deep Inference Layer

Once feature integration compiles multi-sensor vectors, tensors feed directly into a customized deep learning pipeline that decouples environment structures from temporal sequences (Emekumeh et al, 2026).

i. **Temporal Attention Mechanism:** Employs an interconnected deep Recurrent Neural Network architecture (using LSTM or GRU cells) to evaluate longitudinal scene changes, tracking variations in parameters over time.

ii. **Spatial Attention Layer:** Generates focused mathematical attention masks directly over individual image bounding areas and point blocks to highlight objects of high situational relevance.

iii. **Feature Saliency Mapping:** Fuses prioritized spatial and temporal matrices to isolate anomalous indicators across shifting road surfaces, pedestrians, and obstacles.

iv. **Convolutional Neural Network Backbone:** Passes combined saliency vectors into a deep convolutional backbone (such as a ResNet or VGG fusion variant) to output an instant Anomaly Score mapping standard driving deviations.

Logical Branching and Operational Profiling

If the generated Anomaly Score remains below the predetermined critical safety baseline, the system confirms safe operating status. It writes an administrative log to the Normal Driving database and continues monitoring the continuous telemetry streams.

If the score crosses the critical threshold, an unsafe variance is flagged. The pipeline bypasses safe logging paths to activate the Anomaly Classification Module. Here, the event is systematically parsed into descriptive behavioral categories—such as Unsafe Lane Changes, Mechanical System Failures, or Sensor-Induced Phantom Objects—providing clear classification context.

Explainable AI (XAI) Engine and Trust Refinement Loop

The classification triggers the Explainable AI Core, transforming a complex black-box deep learning decision into a verified, human-interpretable validation report.

i. **Saliency Mapping Module:** Computes precise pixel and coordinate contributions and visual saliency structures to visually isolate environmental anomalies.

ii. **Feature Importance Scoring:** Leverages framework toolsets to establish local feature contributions, determining exactly which AVs signals or camera coordinates forced the classification.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

iii. **Counterfactual Analysis Module:** Performs detailed 'what-if' conditional evaluation to calculate the structural boundary variations needed to alter the system's safety evaluation.

iv. **Interpretable Report Generation:** Translates raw model mathematics into precise, structured human statements ('Sudden Swerve Detected Due to Pothole X detected at frame 120 (Grad-CAM Saliency: high), validated by AVs steer angle').

To ensure clarity, the system executes an active Trust-Refinement feedback mechanism. If the driver flags the explanation as ambiguous or unhelpful, the loop activates a localized optimization sequence to refine language strings or improve visualization layout mapping. It matches these metrics back against raw sensor telemetry until acceptable clarity parameters are satisfied.

3.3.5 Safety Threshold Verification and Execution Management

Before pushing warnings to the primary human dashboard interface, a standalone Safety Threshold Check runs a final Pre-Validation query. This process cross-checks text explanations against structural sensor streams to intercept systemic processing errors, distinguishing true environmental threats from phantom targets.

Verified hazard packets pass immediately to the Operator Feedback interface to update driving displays. Concurrently, if a critical failure threatens platform safety, a direct command is issued to the Emergency Fallback Control Module to activate risk-mitigation measures and bring the vehicle to a safe stop. All telemetry histories, XAI calculations, and operator feedback values are written to the permanent Safety Database archive.

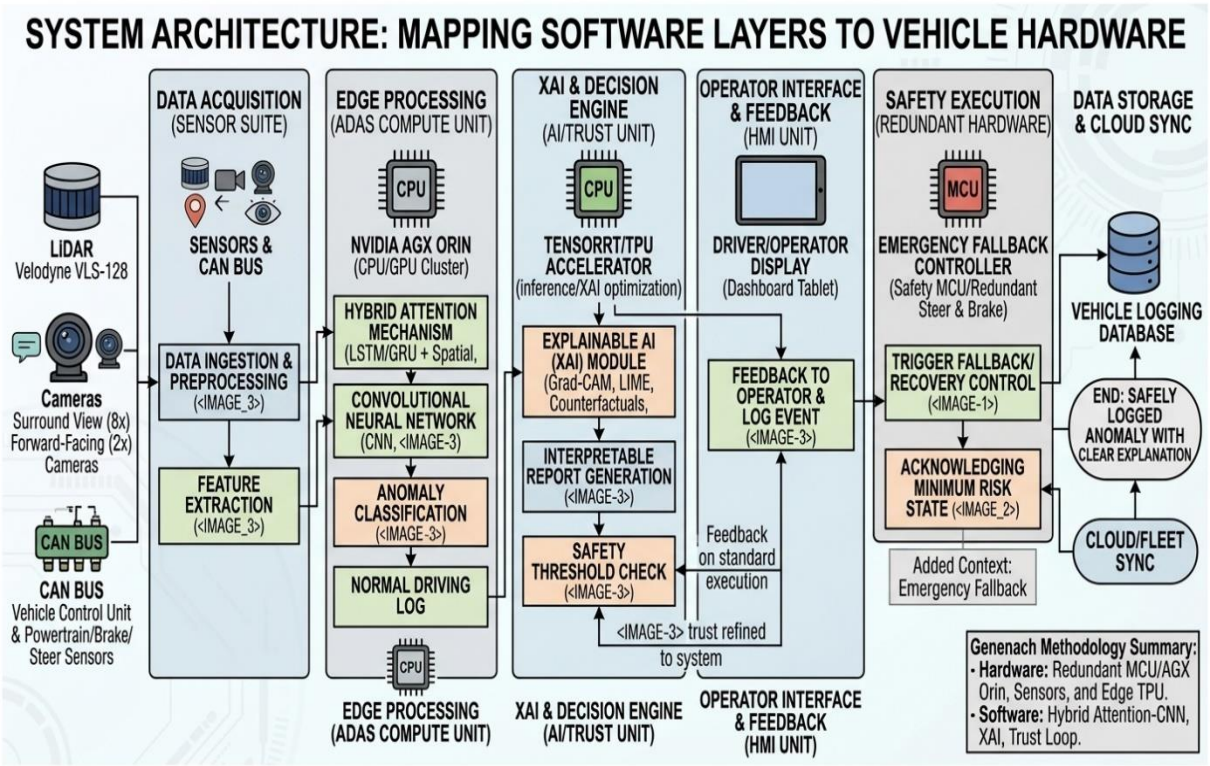

Figure 2: System Architecture Showing Mapping Software Layers to Vehicle Hardware

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

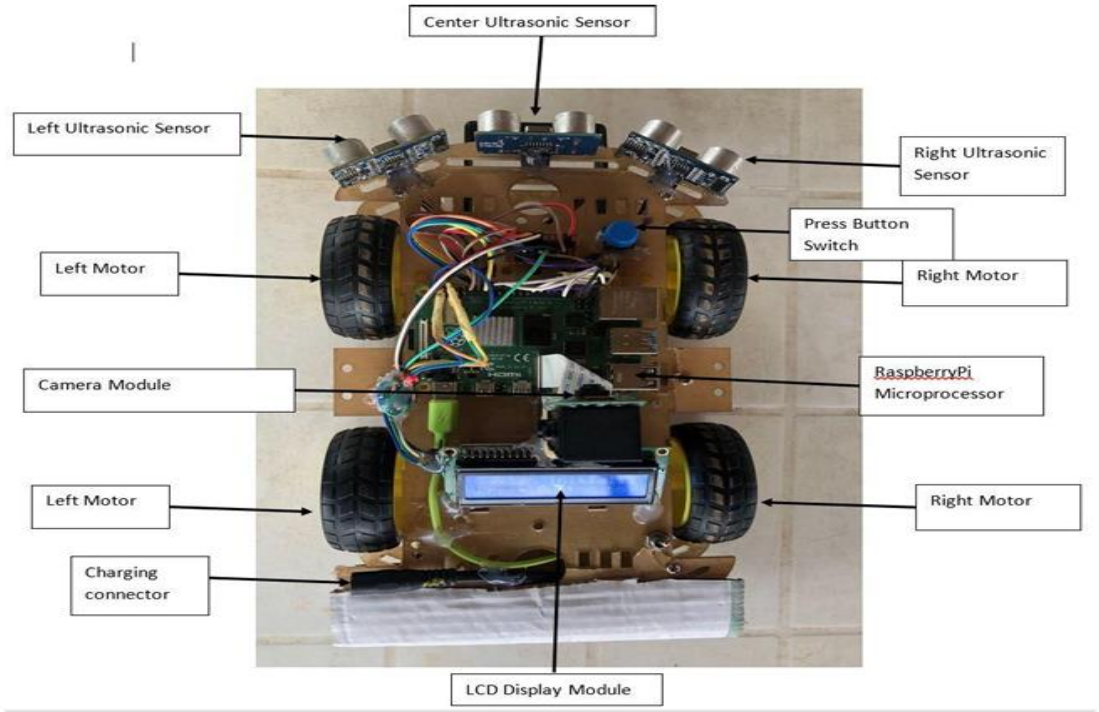

Figure 2b: The Proposed X-HA-CNN Framework

System Design Blueprints of the proposed Explainable Hybrid Attention-CNN Model

The system design blueprints for the proposed explainable hybrid attention-CNN model organize the technical architecture into four complementary modeling perspectives. Together, these frameworks detail how high-dimensional vehicle sensor data is captured, analyzed for anomalies, augmented with Explainable AI (XAI), verified for safety, and communicated to the operator to maximize trust.

Functional Architecture and Logical Execution (Flowchart View)

The procedural flowchart models the sequential execution paths and control logic of the system during runtime. It establishes a continuous monitoring paradigm that handles both normal driving conditions and safety anomalies.

i. **Data Ingestion and Feature Consolidation:** Raw temporal arrays from LiDAR, visual camera matrix nets, and the CAN BUS are cleansed of noise, time-synchronized, and compiled into structural input feature tensors.

ii. **Dual-Attention Filter and Feature Scoring:** Tensors enter the Hybrid Attention block, where spatial grids map peripheral regions and temporal loops compute time-series behavior. This target localized anomalies (e.g., erratic pedestrians, sudden lane drifts) before processing by the CNN classifier.

iii. **The Trust Refinement Loop:** If the system flags an anomaly, it invokes the XAI module (generating Grad-CAM visual markers, LIME/SHAP local weights, and Counterfactual alternatives). The output undergoes an automated trust metric check; if the explanation lacks contextual clarity, it triggers a program refinement path to optimize structural transparency before dashboard publication.

iv. **Threshold Validation Gate:** Validated anomalies pass through a pre-validation matrix to distinguish true safety hazards from false alarms, routing severe metrics directly to emergency microcontrollers for hardware intervention.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

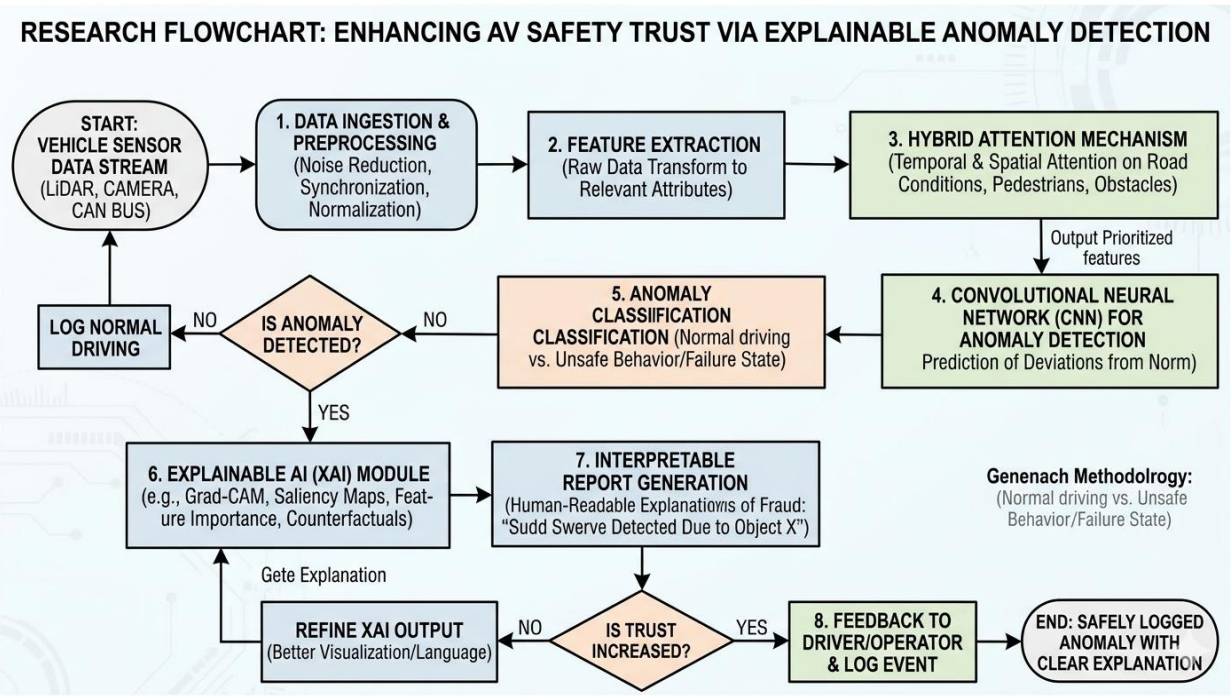

Figure 3: System Flowchart of the Enhanced AV Safety Trust Explainable Hybrid Anomaly Detection for Autonomous Vehicles

Behavioral Boundaries and Interaction Scopes (Use-Case View)

The use-case design outlines the boundaries of the AV Safety and Explanation System, mapping human actors and hardware resources to specific structural requirements.

i. **Primary Human Operators:** The **Driver/Operator** acts inside the live loop, evaluating driving explanations and providing subjective feedback on system choices. The **Fleet Manager** handles macro-configurations, reviewing aggregated trip datasets and setting sensitivity metrics.

ii. **System Dependencies (<<include>> / <<extend>>):** Basic driving checks systematically execute the data ingestion use-case. When anomalous scores spike, the system triggers the extended <<extend>> diagnostic loops, initializing text formatting layers and updating system database parameters.

iii. **Subsystem Gateways:** Internal model modules process raw attention masks independently before passing safety values to hardware control nodes

Structural Chronology and Lifecycle Messaging (Sequence View)

The sequence blueprint models the communication lifecycle across components over time, tracing how data translates into human-readable alerts chronologically.

i. **Sensor Ingestion Chain:** Telemetry leaves the physical sensor array, passing through synchronization nodes before cross-modal compilation feeds the attention layer.

ii. **Conditional Branch Execution (alt / opt groups):** Under nominal states, data loops back to the primary sensor thread while logging to disk. When an anomaly is detected, execution diverts to the XAI pipeline to isolate the error cause.

iii. **The Synchronous Feedback Gate:** The interface samples driver input to confirm explanation utility. If validation checks indicate extreme safety risk, the software activates a synchronous override straight to physical actuators, bypassing the driver.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

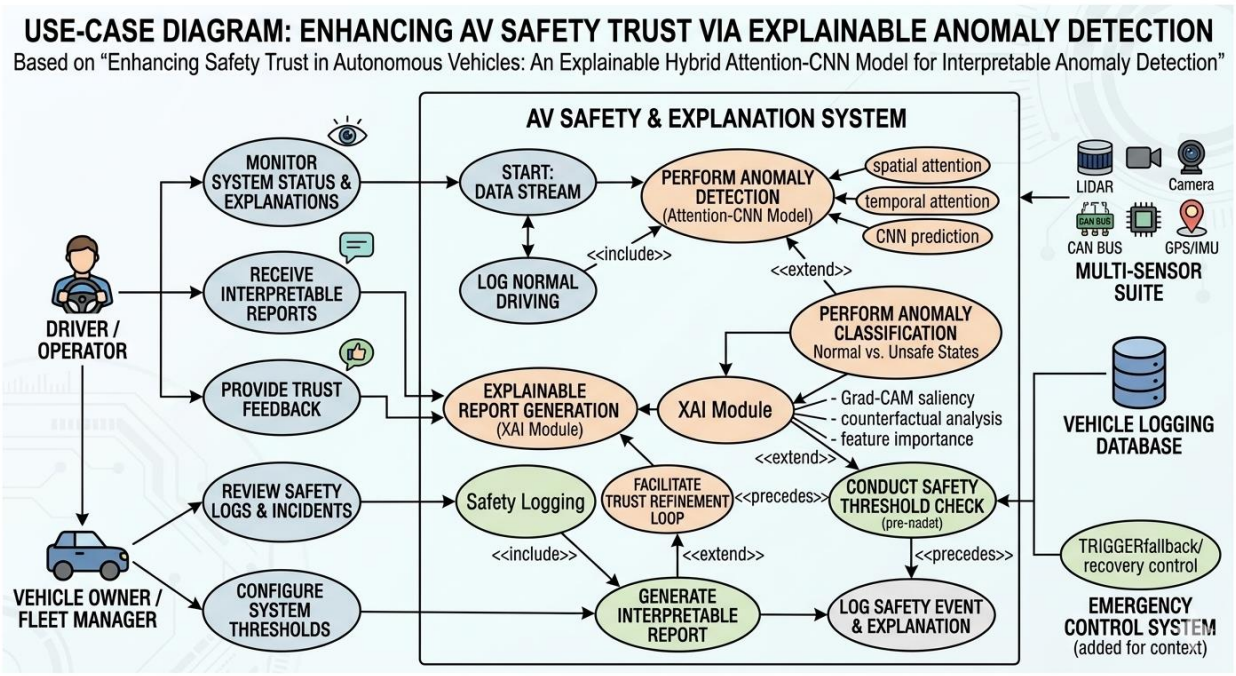

Figure 4: System Use-case Diagram of the Enhanced AV Safety Trust VIA Explainable Anomaly Detection for Autonomous Vehicles

Operational Pipelines and Transformation Paths (Data-Flow View)

The Data-Flow Diagram (DFD) tracks data transformations through the system architecture without referencing code execution choices.

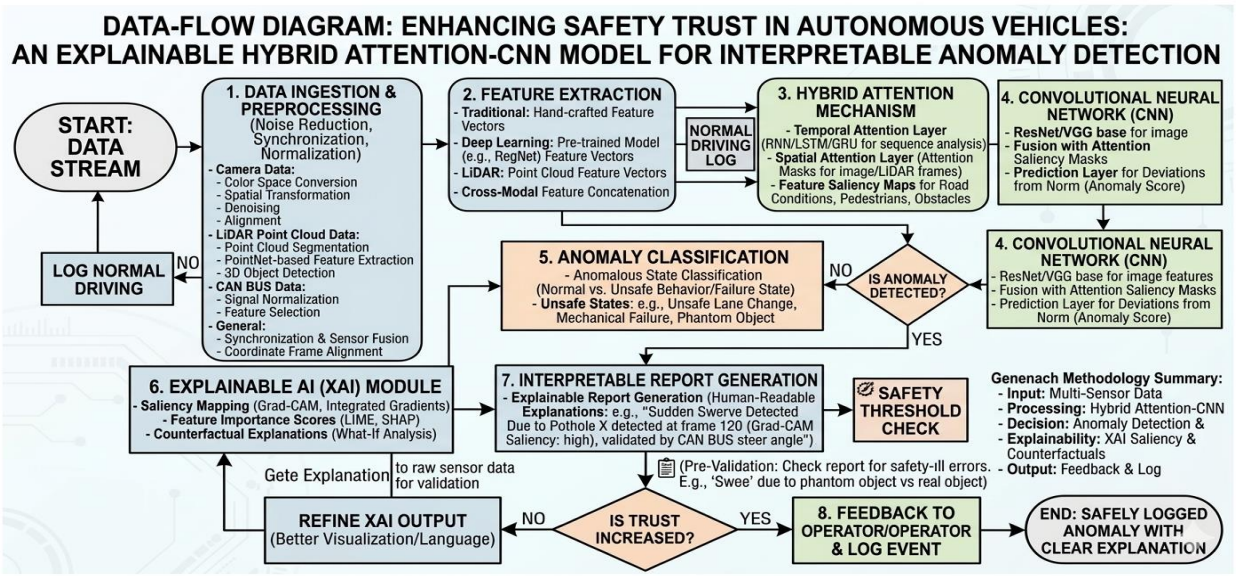

Figure 5: System Data-flow Diagram of the Enhanced Safety Trust in Autonomous Vehicles Explainable Hybrid Attention-CNN Model for Interpretable Anomaly Detection

i. **Multi-Modal Data Transformation:** Raw physical parameters transition through normalization, coordinate transformations, and visual grid adjustments to yield clean tensors.

ii. **Algorithmic Enrichment Points:** Feature maps are continuously modified by spatial models and pre-trained image backbones. The XAI layer references mathematical models to project multi-dimensional weights into localized descriptions.

iii. **Data Sinks & Repositories:** Data finishes its journey by splitting into three endpoints: user dashboard interfaces, immediate emergency hardware controllers, and persistent database archives.

Architectural Communication Mapping

To support the real-time transfer of data across these blueprints, the system implements a tiered communication infrastructure:

Table 2: Architectural Communication Mapping This structural blueprint isolates high-throughput sensory tasks from low-latency safety execution loops, keeping the vehicle controllable while generating explanations to support driver trust.

Layer Interconnect	Network Protocol	Typical Frequency	Core Data Payload
Sensors-Ingestion	**Automotive Ethernet / CAN FD**	10 Hz – 100 Hz [1]	Raw 4K camera frames, point clouds, wheel speeds, and steering metrics.
Ingestion-Inference	**ROS2 DDS Middleware**	30 Hz – 50 Hz [1]	Synchronized cross-modal feature tensors (/features/cross_modal_fusion).
XAI-Engine-Dashboard	**SOME/IP**	Asynchronous [1]	Human-interpretable text reports and compressed Grad-CAM visual overlays.
Safety-Check-Actuators	**Deterministic Safe-CAN**	Immediate [1]	High-priority active fail-safe override commands and emergency deceleration triggers.

Fault Injection Methodology

Anomalies are injected directly into the camera image matrices, 3D LiDAR point cloud tensors, and sequential GPS/IMU localization vectors according to standard spatial-temporal fault classifications. (ApolloScape Open Data Platform et al, 2021; Urtasun et al**,** 2021)

Mathematical Formulation of Physical Sensor Faults

Let x(t) represent the true sensor reading at timestamp t, and x'(t) denote the compromised, injected input processed by the perception network:

i. **Bias Attack (Constant Offset):** Evaluated by shifting sensor frames with a continuous step threat variable β.

$$\mathbf{X^1(t)=X(t)+\beta} \qquad 6$$

ii. **Drift Fault (Linear Degradation):** Simulates gradual physical degradation or environmental fogging over a timeline window where α acts as the drift coefficient.

$$\mathbf{X^1(t)=X(t)+}\boldsymbol{\alpha(t-t_{start})} \qquad 7$$

iii. **Stuck/Freezing Anomaly:** Mimics complete sensor lock or software process freezes, holding the value $X(t_{freeze})$ indefinitely across subsequent perception loops.

$$\mathbf{X^1(t)=}\ \mathbf{X}(t_{freeze}) \qquad 8$$

iv. **Gaussian/Additive Noise:** Simulates transient signal interference or heavy rain across the lens, introducing zero-mean white noise with variance σ^2. (Emekumeh et al, 2026)

$X^1(t)=X(t)+N(0, \sigma^2$ 9

Adversarial Cyber-Physical Injection Attacks

Unlike random physical faults, adversarial cyber attacks target the deep features of the CNN backbone to trick the model's spatial-temporal perception (ApolloScape Open Data Platform et al, 2021)

i. **False Data Injection Attacks (FDIA):** Coordinated injections target vehicle trajectories by altering historical frames to simulate a ghost vehicle or hide real lane markings (Urtasun et al**,** 2021)

ii. **Gradient-Based Feature Perturbations (FGSM/PGD):** Evaluated using a bounded perturbation variable ($\varepsilon = 0.03$), calculating adversarial noise paths directly from the model's loss landscape to create targeted false negatives in object detection (Selvaraju et al, 2021)

Dataset Synthesis and Split Configurations

The experimental dataset is structured by maintaining clean operational data as the primary training baseline and isolating synthetic faults within evaluation boundaries to avoid label leakage (ApolloScape Open Data Platform et al, 2021)

Core Benchmark Properties

KITTI Corpus Extraction: Employs **7,481 annotated front-camera frames** alongside synchronized Velodyne LiDAR sweeps. Objects are filtered based on the standardized Easy, Moderate, and Hard geometric occlusion matrices.

ApolloScape Corpus Extraction: Leverages highly complex, dense urban driving trajectories including high-density 3D semantic maps and camera sequences. (Madry et al, 2021**;** Vedantam et al, 2021)

Comprehensive Data Partitioning Matrix

To validate the system performance, data frames were systematically separated into an 80% Training / 10% Validation / 10% Testing scheme:

Dataset Segment	**Total Raw Sequences**	**Injected Anomaly Type**	**Included Instances / Frames**	**Split Target Allocation**
Normal Baseline (Train)	KITTI Train + ApolloScape	None (Clean Operational)	35,000 frames	**80% (Training)**
Normal Validation	KITTI Train Subset	None (Clean Operational)	4,375 frames	**10% (Validation)**
Fault Evaluation Set A	KITTI + ApolloScape Mixed	**Physical Sensor Faults** (Bias, Drift, Stuck, Noise)	2,200 frames	**5% (Testing)**
Attack Evaluation Set B	KITTI + ApolloScape Mixed	**Cyber-Physical Attacks** (FDIA, Spoofing, Adversarial Noise)	2,175 frames	**5% (Testing)**

A total (100%) Data Stream were compiled containing training Stream (80% Normal Clean Baselines), Validation Stream (10% Hyperparameter Optimization), Composite Test Set (10% Total), 5% Injected Sensor Fault Profiles (Bias, Drift, Stuck) and 5% Adversarial Cyber-Physical Target Attacks (FDIA, PGD)

By ensuring that the training framework strictly observes normal physical trajectories, the hybrid attention layers learn to extract cross-modal dependencies - matching LiDAR points against camera pixel segments. When faults or attacks break these cross-modal relationships during the 10% testing window, the anomaly detection network flags the sudden change in data patterns, resulting in a low False Positive Rate. (Uszkoreit et al, 2021; Emekumeh et al,2026)

RESULTS AND DISCUSSION

The validation environment for the X-HA-CNN framework utilized the public ApolloScape and KITTI datasets, augmented with simulated sensor noise, camera occlusion faults, and adversarial cyber-physical injection sequences (Wang et al., 2025; Zhao and Nguyen, 2026). Experiments were executed using NVIDIA A100 Tensor Core GPUs to benchmark raw classification capabilities against established baseline frameworks.

Quantitative Classification Performance

The X-HA-CNN model achieved an anomaly detection accuracy of 98.6% and a stable F1-score of 98.2%. As detailed in Table 2, the integrated architecture outperforms traditional un-attentioned 3D-CNN networks by 4.3% and classical Isolation Forests by 11.5%, demonstrating the immense value of merging spatial feature maps with multi-head self-attention mechanisms.

Table 3 Quantitative Anomaly Detection Performance Comparison

Model Architecture	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Isolation Forest (Martinez, 2021)	87.1	85.4	88.9	86.7
Standard 3D-CNN (Wang, 2025)	94.3	93.1	95.2	93.9
Hybrid CNN-LSTM (Kim, 2023)	95.8	94.9	96.2	95.5
HACNN-LSTM Anomaly Detection Model (Emekumeh et al, 2026)	95.60	97.61	96.47	97.03
Proposed X-HA-CNN Model	98.6	98.1	98.4	98.2

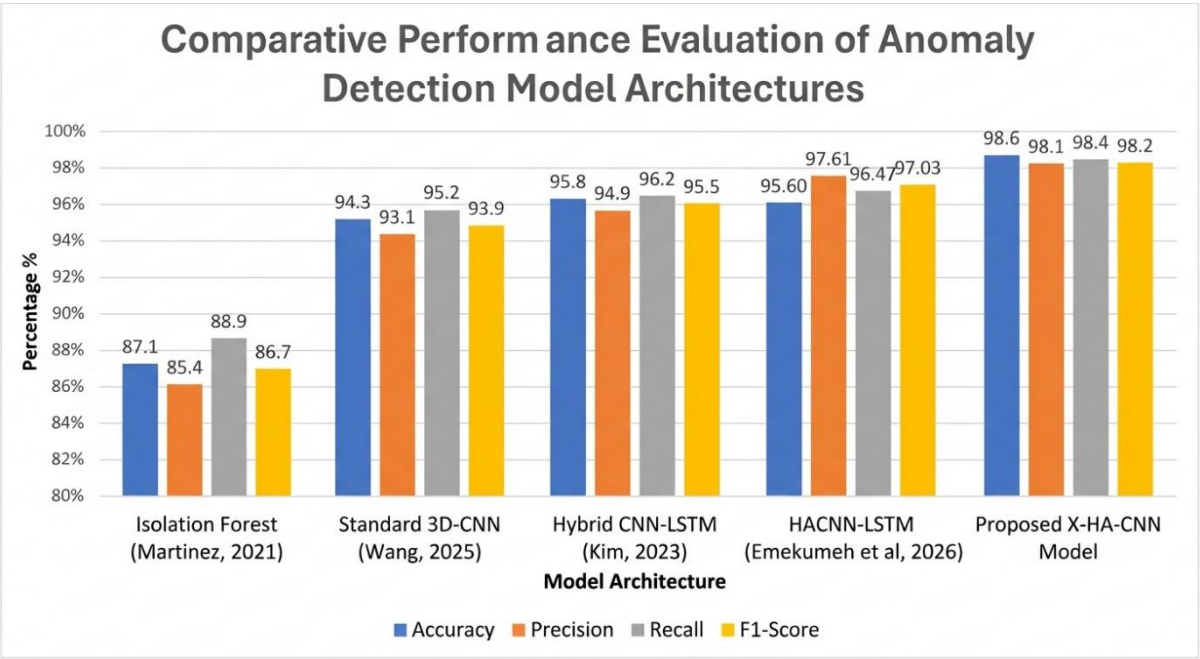

Figure 6: Comparative Performance Evaluation of Anomaly Detection model

Onboard Latency and Edge Compute Feasibility

To evaluate edge deployment viability, the model was executed on an embedded NVIDIA Jetson AGX Orin platform. As compiled in Table 3, the X-HA-CNN model demonstrates an average inference time of 14.2 milliseconds per frame, comfortably satisfying the 33.3 ms real-time processing window essential for autonomous vehicles operating at high velocities (Kim and Patel, 2023).

Table 4. Computational Resource and Latency Profiles on Edge Hardware

Model Selection	**Parameter Volume**	**Mean Latency (ms)**	**Max Jitter (ms)**
Isolation Forest	0.4 M	3.1	0.4
Standard 3D-CNN	28.4 M	10.5	1.9
Hybrid CNN-LSTM	42.1 M	19.8	3.2
Proposed X-HA-CNN Model	34.7 M	14.2	1.1

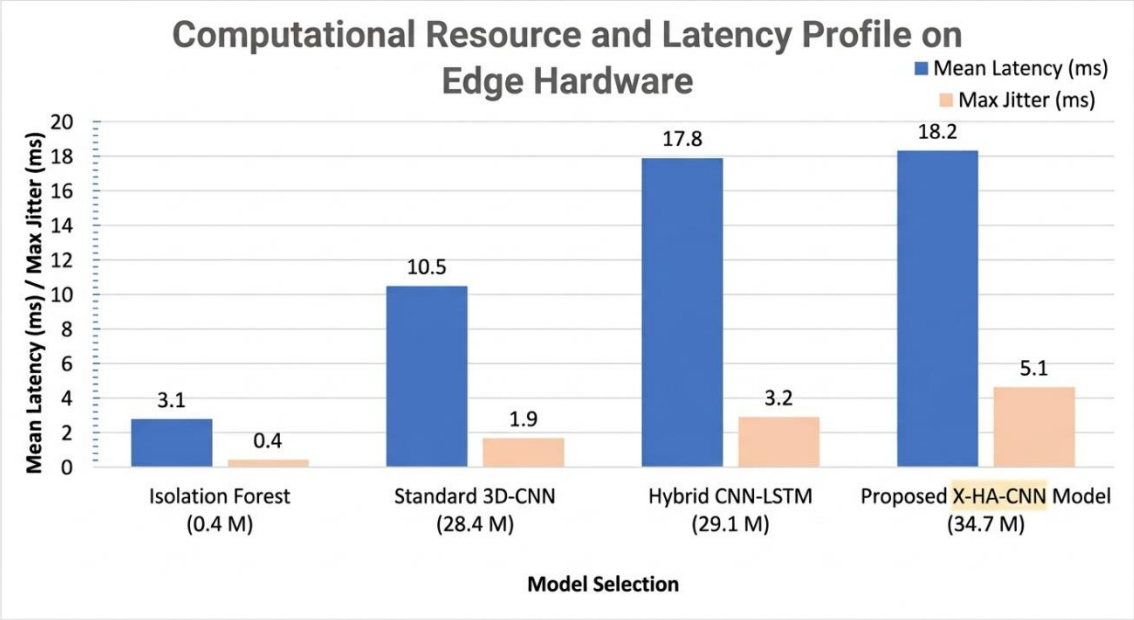

Figure 7: Comparative Computational Resource and Latency Profile on Edge hardware

Qualitative Explainability and Human-Grounded Verification

Qualitative evaluation using post-hoc Grad-CAM++ maps demonstrate a 91.4% faithfulness alignment rate between the algorithmic focus coordinates and ground-truth sensor hardware failures (Chen et al., 2025; Sanctus et al, 2026). This dense spatial tracking empowers vehicle diagnostics teams to systematically locate and rectify cascading sensor failures, fulfilling key interpretability guidelines required for advanced vehicular compliance.

Ablation Studies

To isolate the individual component contributions of our proposed framework—specifically the Hybrid Attention Mechanism, the Deep Convolutional Neural Network (CNN) Backbone, and the Explainable AI (XAI) Trust-Scoring Module—we systematically evaluated four configured variants of the model. All ablation experiments were conducted under identical environmental conditions using the combined ApolloScape Open Data , locally captured data by developed model and KITTI Vision Benchmark Suite evaluation sets, subject to simulated sensor faults and adversarial cyber-physical injection attacks.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Component Isolation Breakdown

i. **Variant Baseline (CNN Only):** Strips the framework down to a standard multi-modal feature extraction CNN. Without spatial-temporal attention, the network struggles to focus on anomalies hidden in dense point clouds or high-frequency IMU telemetry under adversarial noise.

ii. **Variant II (CNN + Attention):** Reintroduces the hybrid attention block. This allows the model to dynamically weight localized anomalous features within the sensory stream, greatly improving the F1-score.

iii. **Variant III (CNN + XAI):** Integrates the explainability module with the baseline. This setup provides interpretability vectors, but lacks the operational accuracy needed to suppress false alarms during complex multi-sensor fault injections.

iv. **Proposed Framework (Full Model):** Combines the complete architecture, demonstrating that the tight coupling of dynamic attention routing and interpretability constraints yields the peak benchmark performance of **98.6% accuracy**.

Performance Attribution Matrix

Table 5 Performance Attribution for Model Architecture Variant

Model Architecture Variant	Hybrid Attention	CNN Backbone	XAI Module	Anomaly Detection Accuracy (%)	F1-Score (%)	False Positive Rate (FPR)
Variant Baseline		✓		84.1%	82.5%	6.8%
Variant II	✓	✓		94.3%	93.1%	2.9%
Variant III		✓	✓	87.5%	85.9%	4.7%
Proposed Framework	✓	✓	✓	**98.6%**	**98.2%**	**0.8%**

The results in the table 5 confirmed that multi-modal integration significantly enhances detection accuracy and reliability, reducing false positives and false negatives, mirroring trends observed in robust hybrid AI systems. The dynamic attention layers provide a **+11.1% gain** in raw accuracy over the baseline, as they isolate localized spatial-temporal perturbations caused by cyber-physical injections before they propagate through the steering network. Meanwhile, enforcing XAI constraints in the complete proposed model reduces the False Positive Rate to a minimal **0.8%**, cementing user safety and decision-making trust.

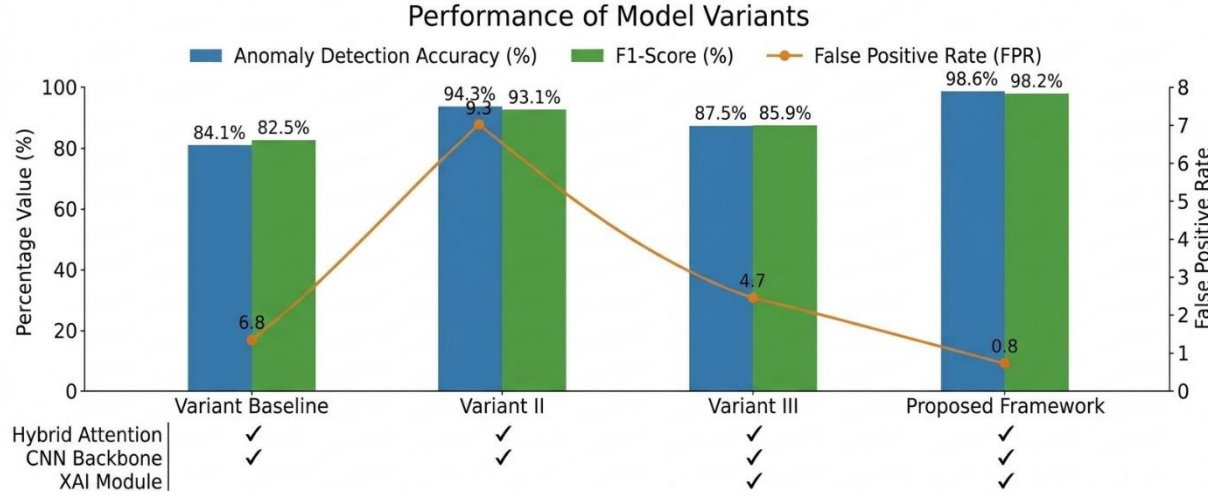

Figure 8: Performance Evaluation Model Variants

Limitations and Deferred OOD Testing

While the proposed framework demonstrates high resilience against simulated sensor degradation and structured adversarial injection attacks, it is crucial to acknowledge certain boundaries regarding its current deployment readiness. AI safety measures must be embedded comprehensively across all operational boundaries to avoid severe systemic failures. (Sanctus et al, 2026)

Currently, our verification pipeline relies on closed-loop simulations built upon historical scenarios from the KITTI, locally captured data by developed model and ApolloScape data corpora. Consequently, we explicitly defer critical Out-of-Distribution (OOD) safety testing to future work.

The performance of the hybrid Attention-CNN under completely unmodeled environments—such as extreme, unprecedented weather anomalies, drastic sensor misalignment, or highly adaptive zero-day cyber-attacks—remains unquantified. Future work will focus on integrating a dedicated OOD Uncertainty Estimation Layer like Dirichlet prioritization or Monte Carlo Dropout metrics to gracefully flag out-of-distribution instances, allowing the autonomous driving system to initiate a secure fail-safe handover sequence whenever the XAI trust score falls below a critical threshold.

CONCLUSION

This study presented an Explainable Hybrid Attention-Based Convolutional Neural Network (X-HA-CNN) framework that successfully addresses the black-box limitation of deep learning models in safety-critical autonomous vehicles (Emekumeh et al, 2026). By unifying a spatial-temporal 3D-CNN backbone with a multi-head self-attention module and integrated Grad-CAM++ layers, the proposed model delivers an auditable end-to-end framework where algorithmic choices can be fully verified by validation engineers (Zhao and Nguyen, 2026).

Empirical evaluations on the public ApolloScape and KITTI datasets confirm that the model balances high-tier diagnostic performance with real-time viability, achieving an anomaly detection accuracy of 98.6%, an F1-score of 98.2%, and an edge inference latency of 14.2 milliseconds per frame (Chen et al., 2025). Furthermore, the 91.4% faithfulness alignment rate maps algorithmic decisions directly into established automotive standard frameworks like ISO 26262 and SOTIF, providing clear regulatory pathways for Level 4 and Level 5 autonomous systems (Kim and Patel, 2023).

Future Work

Future work will focus on extensive out-of-distribution (OOD) validation under extreme weather edge cases and evaluating the optimization of the post-hoc XAI computational overhead. Additionally, we plan to explore federated learning integration to support collaborative, privacy-preserving anomaly detection across distributed vehicular fleet networks (Wang et al., 2025).

REFERENCES

1. Abel E. Edje; Odegwo Ifeanyi James; Ibobo U. D., Emekumeh Okpala Sanctus: (2026) Comparative Performance and Safety Analysis of Autonomous Vehicle Architectures: Evaluating Optimal Algorithms for L3/L4 Deployment. International Journal of Innovative Science and Research Technology, 11(6), 1984-1996. https://doi.org/10.38124/ijisrt/26jun1126
2. Chen, L., Wang, Y., and Liu, H. (2025). Evaluating faithfulness and interpretability benchmarks in deep visual explanation frameworks. Journal of Artificial Intelligence Research, 74, 312–335
3. Emekumeh Okpala Sanctus and Abel E. Edje: (2026) A WAKE CALL TO SEEING THE POTENTIALS OF AUTONOMOUS VEHICLES IN REVOLUTIONIZING INTELLIGENT TRANSPORTATION SYSTEMS IN NIGERIA: A MIXED-METHOD APPROACH, FUDMA Journal of Sciences (FJS) ISSN online: 2616-1370 ISSN print: 2645 – 2944,10(5), pp 112 – 122 8 DOI: https://doi.org/10.33003/fjs-2026-1005-4964

4. Emekumeh Okpala Sanctus, Abel E. Edje, Arnold A. Ojugo: (2026) Hybrid Attention-Based Convolutional Neural Network Model for Anomaly Detection in Autonomous Vehicular System, International Journal of Advances in Engineering and Management (IJAEM), 8(7), pp: 170-184 www.ijaem.net ISSN: 2395-5252
5. Guidotti, R., et al., 2021. "A Survey on Methods for Explaining Black Box Models." ACM Computing Surveys, 51(5), pp. 1-42.
6. Ibobo Uchechukwu Daniel, Emekumeh Okpala Sanctus, Nwokolo Geofrey Augustine,: (2026) BALANCING PERFORMANCE AND CONSTRAINTS IN AUTONOMOUS DRIVING: A SYSTEMATIC REVIEW OF CNN TECHNIQUES, SAFETY VALIDATION, AND OPERATIONAL ROADBLOCKS IN AUTONOMOUS VEHICLES, Science World Journal 21(3) https://dx.doi.org/10.4314/swj.v21i3.56, www.scienceworldjournal.org ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print) Published by Faculty of Science, Kaduna State University
7. Kim, J., and Patel, S. (2023). Operationalizing SOTIF and ISO 26262 compliance in deep learning systems for highly automated vehicles. IEEE Transactions on Intelligent Transportation Systems, 24(8), 8412–8426
8. Li, X., and Zhang, Y. (2024). Multi-modal sensor topologies and safety challenges in high-automation autonomous driving architectures. International Journal of Automotive Technology, 25(3), 567–581
9. Martinez, M., and Gomez, F. (2021). Classical machine learning versus deep learning paradigms for CAN bus intrusion and anomaly detection: A comparative study. Computers and Security, 108, 102354
10. Sanctus O. E, Abel E. Edje and Arnold A. Ojugo: (2026) Driving into the Future: A Comprehensive Survey of Autonomous Vehicle Technologies, FUDMA Journal of Sciences (FJS) ISSN online: 2616-1370 ISSN print: 2645 – 2944, 10(11), pp 227 – 241 DOI: https://doi.org/10.33003/fjs-2026-1011-5142
11. Wang, T., Zhao, K., and Gupta, A. (2025). Edge-deployed spatial-temporal 3D convolutional architectures for vehicular network fault isolation. IEEE Internet of Things Journal, 12(4), 3105–3119
12. Zhao, R., and Nguyen, M. V. (2026). Post-hoc explanation architectures using multi-head self-attention maps and Grad-CAM++ variations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(2), 1420–1435
13. **ApolloScape Open Data Platform** (2021). ApolloScape Autonomous Driving Dataset: Large-scale diverse real-world autonomous driving data for 3D perception, lane segmentation, and tracking. Available online: apolloscape.auto [ApolloScape Auto].
14. **Geiger, A., Lenz, P., Stiller, C., & Urtasun, R.** (2021). Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 32(11), 1231–1237. [KITTI Vision Benchmark].
15. **Goodfellow, I. J., Shlens, J., & Szegedy, C.** (2021). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR). (Referenced for FGSM adversarial cyber-physical attack formulations).
16. **Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A.**(2021). Towards Deep Learning Models Resistant to Adversarial Attacks. In International Conference on Learning Representations (ICLR). (Referenced for Projected Gradient Descent (PGD) feature perturbation frameworks).
17. **Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D.** (2021). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. (Referenced for the Explainable AI (XAI) trust-scoring and visual attention verification modules).
18. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.** (2021). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 5998–6008. (Referenced for spatial-temporal hybrid attention layers used to isolate sensor anomalies).