EP-JOURNALS

GLOBAL JOURNAL of ENGINEERING AND TECHNOLOGY REVIEW

Available on: http://gjetr.ep-journals.org/ E-ISSN: 3051-3782, P-ISSN: 3051-3774

Global Journal of Engineering and Technology Review

Volume: 02 Issue: 09 September 2026

Edge-Optimized YOLO Architectures for Real-Time Autonomous Vehicle Perception: A Hardware-Aware Co-Design Framework

Harouna Wendpanga Yann Christian Sankara
Independent Researcher, Dallas, TX, USA

Serif Oyindamola Oyesiji
Harrisburg University of Science and Technology, Harrisburg, PA

Emmanuel Eniola Ajala
Independent Researcher, Atalanta, Georgia

Stanley Nwakamma
Independent Researcher, USA

Published Date: September 09, 2026 **Article DOI : 10.65150/EP-gjetr/V2E9/2026-09**

ABSTRACT: Object detection is the perceptual backbone of autonomous driving, and single-stage detectors of the YOLO family have become the practical default wherever frames must be processed within real-time budgets. Yet the published detection literature optimizes predominantly for benchmark accuracy on server-class accelerators, while vehicles impose a different objective: bounded end-to-end latency at high frame rates, on power- and thermally-constrained embedded accelerators, across multiple simultaneous camera streams, under safety expectations that penalize missed detections far more than the mean average precision metric reflects. This paper develops a research concept for an edge-optimized YOLO-based perception component designed and evaluated against vehicle-grade constraints. The concept specifies a co-design space spanning architecture, compact backbones, decoupled heads, resolution and anchor policy tuned to driving object statistics, and compression, structured pruning, quantization-aware training to integer arithmetic, and knowledge distillation from a high-capacity teacher, searched jointly under hardware-in-the-loop latency measurement rather than proxy operation counts. A deployment architecture allocates per-camera detection instances across embedded accelerator resources with a frame-freshness scheduling policy that privileges recency over throughput, and an optional offload path is analyzed and deliberately excluded from the safety path. The evaluation plan is phased: accuracy and robustness on driving benchmarks with corruption suites; latency, jitter, energy, and thermal behavior measured on target hardware across compression configurations; and system-level metrics that couple detection quality to reaction distance at speed. The concept's central claim is methodological: for vehicle perception, the deployable operating point is a property of the model-compression-hardware triple, and it must be measured as such.

KEYWORDS: Edge-optimized YOLO, real-time object detection, autonomous vehicle perception, hardware-aware co-design, model compression (pruning, quantization, distillation), embedded accelerators, multi-camera latency, safety-critical edge AI

1. INTRODUCTION

Perception for autonomous driving is a real-time computer vision problem with unforgiving constraints. Survey literature across the field documents the standard pipeline, multi-camera detection feeding tracking, prediction, and planning, and the constraint stack it runs under: embedded automotive compute, hard latency expectations tied to vehicle dynamics, and reliability requirements that follow from safety standardization (Yurtsever et al., 2020; Grigorescu et al., 2020; Janai et al., 2020; International Organization for Standardization, 2018). Systems studies make the budget concrete, showing that end-to-end processing latency directly bounds safe operating envelopes and that architectural choices in the compute platform propagate into driving capability (Lin et al., 2018; Kato et al., 2018). Work on adjacent safety-critical embedded deployments reinforces the same constraint picture from outside the driving literature: security and resilience analyses of IoT-driven smart transportation infrastructure document the exposure that on-vehicle and roadside compute inherit (Jimoh et al., 2023), reliability engineering treatments of complex engineering systems formalize the failure-analysis discipline such deployments demand (Gideon et al., 2019), and studies of autonomous mobile robot fleets show the same perception-to-actuation loop operating under comparable compute, latency, and coordination budgets in warehouse and micro-fulfillment settings (Akanbi & Sunday, 2025; Akanbi et al., 2025).

The detection component of that pipeline has a clear lineage. Two-stage detectors established modern accuracy (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015), while single-stage designs traded refinement for speed (Liu et al., 2016; Lin et al., 2017b), and the YOLO family pushed the single-stage formulation into the real-time regime that embedded deployment requires, evolving through successive architectural generations

(Redmon et al., 2016; Redmon & Farhadi, 2017; Redmon & Farhadi, 2018; Bochkovskiy et al., 2020; Ge et al., 2021; Wang et al., 2023), with feature pyramids supplying the multi-scale machinery driving scenes demand (Lin et al., 2017a) and anchor-free and transformer formulations broadening the design space (Tian et al., 2019; Carion et al., 2020; Zou et al., 2023). Applied detection work outside driving shows how the same architectural machinery is retargeted to domain-specific object statistics, from template-matched lesion detection in mammography (Ejofodomi et al., 2014) to deep learning for vegetation classification in power transmission corridors (Dagodzo et al., 2022) and right-of-way encroachment detection from geospatial imagery (Dagodzo & Ahiaeke Patrick, 2022), each case demonstrating that class taxonomy and scale distribution, not architecture alone, determine what a detector must be tuned for.

A parallel literature makes networks small. Compact architectures redesign convolutional computation for mobile budgets (Howard et al., 2017; Sandler et al., 2018; Ma et al., 2018; Tan & Le, 2019), compression removes and cheapens what training left redundant, pruning (Li et al., 2017; Molchanov et al., 2017; Frankle & Carbin, 2019), integer quantization (Jacob et al., 2018; Krishnamoorthi, 2018; Nagel et al., 2021), and distillation from larger teachers (Hinton et al., 2015), with deep compression demonstrating their composition (Han et al., 2016), and neural architecture search automates the fit to hardware targets (Zoph & Le, 2017; Tan et al., 2019; Cai et al., 2020). Recent on-device work composes these instruments explicitly for edge budgets, treating quantization, pruning, distillation, and energy-efficient inference as a single joint design problem rather than a sequence of independent optimizations (Nwakamma et al., 2025), with companion surveys on compressing ranking and recommendation models for edge deployment (Oyesiji et al., 2025) and on parameter-efficient adaptation of on-device models (Oyesiji et al., 2023) mapping the same trade space from adjacent application domains. Hardware surveys quantify the accelerator landscape this all lands on (Sze et al., 2017; Jouppi et al., 2017), semiconductor reliability, materials, and supply-chain studies bound what those accelerators cost, how they scale, and how they age (Gideon, 2025a; Gideon, 2025b; Erhimefe & Gideon, 2025), and edge intelligence surveys frame the placement question, what runs on-device, what offloads, under which latency and reliability regimes (Chen & Ran, 2019; Zhou et al., 2019; Deng et al., 2020; Shi et al., 2016; Satyanarayanan, 2017), a question examined concretely in microgrid control (Komi & Ganiu, 2023), cloud-native microservice inference (Idika et al., 2021), and portable cloud-to-edge runtime design (Ojukwu et al., 2025).

The gap this concept addresses is the join. Detection papers report mean average precision with throughput measured on server GPUs; compression papers report accuracy retention against operation-count proxies; and the vehicle literature reports system budgets without specifying how a detector should be co-designed to meet them. What is missing is a specified, evaluable methodology in which a YOLO-family detector is architected, compressed, and scheduled against measured behavior on automotive-grade edge hardware, with evaluation metrics that reflect driving risk rather than benchmark convention. The reporting and gating discipline such a methodology implies has been articulated in adjacent governance literature on automated release gating for model deployment and on model risk management for regulated automated decisions (Nwakamma et al., 2024a; Ominyi & Anichukwueze, 2023), but it has not been instantiated for embedded perception. Section 2 reviews the background in detail; Section 3 defines requirements; Sections 4 through 6 specify the design space, the optimization pipeline, and the on-vehicle deployment architecture; Section 7 sets out the evaluation plan; Sections 8 through 10 treat applications, limitations, and discussion; Section 11 concludes.

2. BACKGROUND AND RELATED WORK

Detection architecture defines the accuracy-latency frontier this concept moves along. The YOLO lineage contributes the design vocabulary: unified single-pass detection over a gridded feature map (Redmon et al., 2016), dimension-clustered anchors and multi-scale training (Redmon & Farhadi, 2017), pyramidal prediction across scales (Redmon & Farhadi, 2018; Lin et al., 2017a), the bag-of-freebies and bag-of-specials training and architecture refinements consolidated at the fourth generation (Bochkovskiy et al., 2020), decoupled classification and regression heads with anchor-free assignment (Ge et al., 2021; Tian et al., 2019), and trainable re-parameterization with extended feature aggregation in the current generation (Wang et al., 2023). Comparative context comes from the broader detector surveys, which locate single-stage designs relative to two-stage and transformer alternatives across accuracy, speed, and deployment maturity (Zou et al., 2023; Ren et al., 2015; Carion et al., 2020), and from efficiency-oriented detectors that co-scale backbone, neck, and resolution (Tan et al., 2020).

Driving perception research grounds the domain specifics. Benchmark datasets fix the evaluation substrate and its object statistics, small and partially occluded pedestrians and riders, extreme scale variation, heavy class imbalance, and strong scene priors (Geiger et al., 2013; Cordts et al., 2016; Yu et al., 2020; Caesar et al., 2020; Sun et al., 2020), inheriting evaluation conventions from the general benchmarks (Everingham et al., 2010; Lin et al., 2014). Domain surveys catalogue the modality and fusion landscape within which camera-only detection sits (Feng et al., 2021; Arnold et al., 2019; Janai et al., 2020), and the aerial and infrastructure inspection literature supplies a parallel body of evidence on how detection behaves when platform motion, viewpoint, and illumination vary outside benchmark distributions, in transmission line and pipeline corridor inspection (Dagodzo, 2018a; Dagodzo, 2018b; Dagodzo & Ahiaeke Patrick, 2020), in integrated sensing frameworks combining imagery with ranging and geospatial context (Dagodzo & Ahiaeke Patrick, 2021a; Dagodzo & Ahiaeke Patrick, 2021b), and in airside hazard monitoring where missed detections carry safety consequences comparable in structure to those on the road (Adeyelu & Dagodzo, 2024). Robustness benchmarking establishes that detectors degrade severely under the corruptions driving guarantees will encounter, weather, blur, sensor noise, with corruption suites providing the measurement instrument (Hendrycks & Dietterich, 2019; Michaelis et al., 2019), and the adversarial and anti-spoofing literature adds the complementary failure mode in which degradation is induced rather than incidental, whether against perception models embedded in critical infrastructure (Adebayo et al., 2022) or against deployed biometric vision systems (Adeniyi et al., 2025). The systems strand closes the loop from perception quality to vehicle behavior, quantifying how compute latency translates into stopping distance and how platform choices constrain the achievable pipeline (Lin et al., 2018; Kato et al., 2018).

Efficient inference research supplies the optimization instruments. Compact backbone design established that convolutional cost can fall by an order of magnitude with modest accuracy loss when architecture follows the deployment budget (Howard et al., 2017; Sandler et al., 2018; Ma et al., 2018; Tan & Le, 2019). Pruning removes structured redundancy, with filter-level criteria preserving dense execution on real accelerators (Li et al., 2017; Molchanov et al., 2017) and the lottery ticket results reframing what sparse training can retain (Frankle & Carbin, 2019). Quantization moves inference to integer arithmetic, with quantization-aware training recovering most of the accuracy that post-training conversion loses (Jacob et al., 2018; Krishnamoorthi, 2018; Nagel et al., 2021). Distillation transfers a teacher's soft supervision into compact students (Hinton et al., 2015), and hardware-aware architecture search folds measured latency into the design objective itself (Tan et al., 2019; Cai et al., 2020; Zoph & Le, 2017). Applied on-device work is where these instruments are composed against real budgets rather than reported separately, in multimodal small-model deployment under joint quantization, pruning, distillation, and energy constraints (Nwakamma et al., 2025), in compression of ranking models for latency-bounded edge serving (Oyesiji et al., 2025), and in parameter-efficient adaptation that keeps a single compressed backbone serviceable across deployment variants (Oyesiji et al., 2023). The edge computing literature frames where the optimized model runs: on-vehicle inference for the safety path, with offload economically attractive but bounded by connectivity variability, a trade the edge intelligence surveys analyze in general form (Chen & Ran, 2019; Zhou et al., 2019; Deng et al., 2020; Satyanarayanan, 2017) and that applied studies instantiate for resilient microgrid control (Komi & Ganiu, 2023), for inference distributed across cloud-native microservices at the edge (Idika et al., 2021), and for portable runtimes that let one compiled artifact move between cloud and edge targets without re-verification of the model itself (Ojukwu et al., 2025). Accelerator surveys parameterize the hardware side of every such decision (Sze et al., 2017), and operational studies of dense accelerator deployments supply the thermal and reliability behavior that headline throughput figures omit (Oyeleke et al., 2026; Olodo et al., 2026b).

A fourth strand, less often joined to detection research, concerns how deployed models are measured, monitored, and governed once they leave the training environment. Energy, reliability, and latency are jointly constrained rather than independently optimizable in embedded and intermittently powered systems, and the green communication literature formalizes that trade explicitly (Asiedu & Quainoo, 2024; Asiedu et al., 2025), with energy-harvesting and duty-cycled deployments showing how reliability targets shift when the power budget itself is variable (Asiedu & Quainoo, 2023a; Asiedu & Quainoo, 2023b). Measurement methodology from wireless hardware engineering is directly transferable: system-level power behavior models for concurrent radio subsystems illustrate how contention between co-resident workloads distorts single-workload measurements (Quainoo & Ogundapo, 2026a), physical-layer models predicting throughput degradation under varying channel conditions exemplify the shift from nominal specification to measured degradation curve (Quainoo et al., 2025a), integrated system-on-chip test methodology surveys establish what a defensible hardware characterization protocol contains (Quainoo & Ogundapo, 2026b), and scripted instrument-control frameworks show how such characterization is made repeatable rather than artisanal (Quainoo et al., 2025b). On the monitoring side, digital twin and condition-monitoring practice supplies the pattern of continuously comparing a deployed asset against a reference model, in smart substations (Shittu et al., 2023), in renewable generation assets (Shittu & Shittu, 2024), in resource-constrained sites where the twin itself must be tiered to fit the available compute (Komi & Adeniji, 2023), in rotating machinery instrumented for vibration signature analysis (Omoegun et al., 2023), and in the broader movement from reactive to predictive maintenance regimes (Adaramola et al., 2018; Sunday et al., 2020; Sunday & Omoegun, 2022; Olodo et al., 2025; Olodo et al., 2026a). Governance research completes the strand by specifying what evidence a deployed automated decision component owes its operator, from release gating and automated evaluation (Nwakamma et al., 2024a) through model risk management and algorithmic accountability (Ominyi & Anichukwueze, 2023; Annan, 2024) to organizational readiness for emerging regulatory regimes governing high-risk AI systems (Ominyi et al., 2024b; Ominyi & Anichukwueze, 2024).

3. REQUIREMENTS AND PROBLEM DEFINITION

The component's function is stated as follows: given synchronized frames from the vehicle's camera set, produce per-frame 2D detections of driving-relevant classes with confidence scores, such that the ninety-ninth-percentile per-frame latency on the target embedded accelerator remains within a stated budget at the required camera frame rate for all cameras simultaneously, within the platform's power and thermal envelope, while meeting per-class recall floors at fixed false-positive operating points on driving benchmarks and bounded degradation under a specified corruption suite.

The formulation encodes four deliberate departures from benchmark convention. Tail latency replaces mean throughput, because a planner consumes the slowest camera's freshest frame and vehicle dynamics price the tail, not the average (Lin et al., 2018). Recall floors at operating points replace mean average precision as the primary accuracy contract, because averaged precision across thresholds obscures exactly the missed-pedestrian regime that safety analysis prices most highly, with the class-critical structure the driving datasets make explicit (Cordts et al., 2016; Caesar et al., 2020). Simultaneous multi-camera load replaces single-stream measurement, because per-stream numbers do not compose linearly on shared accelerators (Sze et al., 2017), a non-composition that concurrent-workload power and contention modeling in embedded radio systems demonstrates in a directly analogous form (Quainoo & Ogundapo, 2026a). And measured hardware behavior replaces operation-count proxies throughout, because the compression literature's own findings show proxy metrics mispredict latency across accelerator architectures (Tan et al., 2019; Cai et al., 2020), and because on-device deployment studies find the energy and latency consequences of a compression configuration to be jointly determined rather than separately predictable (Nwakamma et al., 2025; Asiedu & Quainoo, 2024; Asiedu et al., 2025).

Three further requirements follow from the deployment context rather than from detection practice. The power and thermal envelope is a first-class constraint rather than a reporting afterthought, since sustained accelerator operation in a sealed automotive enclosure produces throttling behavior that only continuous-load measurement reveals (Oyeleke et al., 2026; Sze et al., 2017). Reliability and failure-mode analysis is required of the component as a whole, in the sense the reliability engineering literature gives the term, covering not only detection error but export defects, runtime faults, and degradation under hardware aging (Gideon et al., 2019; Erhimefe & Gideon, 2025). And the component must be auditable: its operating point, the evidence supporting it, and the conditions under which it was measured must be recorded in a form that supports external review, as governance frameworks for high-risk automated decision systems and for safety-relevant transportation infrastructure require (Ominyi et al., 2024b; Ominyi & Anichukwueze, 2024; Jimoh et al., 2023). The problem, so defined, is a constrained co-design: select the architecture-compression configuration and the scheduling policy that maximize the accuracy contract subject to the latency, power, robustness, and auditability constraints, on the actual triple of model, toolchain, and hardware.

4. THE EDGE-OPTIMIZED DETECTION DESIGN SPACE

The architectural design space is organized around the YOLO template, backbone, neck, head, with each element parameterized for the co-design search. Backbone candidates span the compact families whose efficiency the mobile literature established, inverted-residual and channel-shuffled designs (Sandler et al., 2018; Ma et al., 2018; Howard et al., 2017), compound-scaled variants (Tan & Le, 2019), and the re-parameterized blocks of the current YOLO generation that spend structure at training time and fold it away for inference (Wang et al., 2023). Neck design retains pyramidal aggregation as non-negotiable for the scale statistics of driving scenes (Lin et al., 2017a; Redmon & Farhadi, 2018), with aggregation depth and width as search dimensions. Head design adopts the decoupled, anchor-free formulation whose assignment simplicity and export friendliness suit embedded toolchains (Ge et al., 2021; Tian et al., 2019), with anchor-based variants retained where dataset statistics reward clustered priors (Redmon & Farhadi, 2017).

Domain parameters are searched jointly rather than inherited from benchmarks. Input resolution is treated per-camera, since forward long-range cameras and lateral near-field cameras face different object-size distributions, an asymmetry the corridor and aerial inspection literature encounters in the same form when sensor standoff varies across a survey (Dagodzo & Ahiaeke Patrick, 2021b; Dagodzo et al., 2022); class taxonomy and loss weighting follow the recall-floor contract, up-weighting vulnerable-road-user classes against the imbalance the datasets document (Cordts et al., 2016; Yu et al., 2020); and training composition draws on the augmentation and regularization repertoire consolidated in the YOLO lineage (Bochkovskiy et al., 2020), extended with the corruption families of the robustness suites as training-time augmentation so that the robustness constraint is optimized, not merely measured (Hendrycks & Dietterich, 2019; Michaelis et al., 2019), and with adversarially perturbed inputs included where the threat model for perception in safety-critical infrastructure warrants it (Adebayo et al., 2022; Adeniyi et al., 2025). Training data composition is itself a search dimension rather than a fixed input, following context-aware curation practice in which the selection and weighting of examples is tuned to the deployment distribution (Oyesiji et al., 2024), with annotation quality maintained through human-in-the-loop review of automatically proposed labels (Ladapo et al., 2022; Ladapo et al., 2024).

The compression axes complete the space. Structured filter pruning is applied with sensitivity profiling per stage, preserving dense kernels the accelerator executes efficiently (Li et al., 2017; Molchanov et al., 2017); quantization targets full integer inference with quantization-aware fine-tuning, per-channel weight scaling, and calibrated activation ranges, following the settled guidance of the quantization literature (Jacob et al., 2018; Krishnamoorthi, 2018; Nagel et al., 2021); and distillation from a high-capacity teacher detector supervises both output distributions and intermediate features during compression recovery, transferring accuracy the compact student cannot reach alone (Hinton et al., 2015; Han et al., 2016). The joint rather than sequential treatment of these axes follows on-device deployment practice, where pruning ratio, quantization granularity, and distillation schedule are selected together against a measured energy and latency target (Nwakamma et al., 2025; Oyesiji et al., 2025). Search over this joint space is hardware-in-the-loop: candidate configurations are exported through the deployment toolchain and timed on the target accelerator under multi-stream load, making measured latency a first-class objective in the manner of hardware-aware search (Tan et al., 2019; Cai et al., 2020), with lightweight predictive models used to prune the candidate set before physical measurement where the search space is too large to enumerate on hardware (Asiedu et al., 2026). The search's output is not a single model but the frontier of configurations from which deployment selects per platform.

5. OPTIMIZATION AND DEPLOYMENT PIPELINE

The pipeline is specified as a reproducible sequence with verification gates. Stage one trains the teacher and the uncompressed student architecture on the driving corpus with the domain parameters of Section 4, establishing the accuracy ceiling and the baseline operating points. Stage two applies structured pruning under sensitivity constraints with distillation-supervised recovery training, iterating ratio schedules per stage of the network (Molchanov et al., 2017; Hinton et al., 2015). Stage three applies quantization-aware training to the pruned model, with per-layer fallback analysis identifying any layers whose integer degradation exceeds tolerance (Nagel et al., 2021; Jacob et al., 2018). Stage four exports through the target toolchain to the accelerator runtime, where a verification gate compares on-device outputs against the training-framework reference over a fixed probe set, bounding numerical divergence introduced by export and runtime kernels, the step at which silent deployment error is caught rather than shipped. Runtime portability is a design consideration at this stage rather than an afterthought, since the ability to move a verified artifact between accelerator targets without repeating model-level verification materially reduces the cost of supporting a vehicle platform across hardware generations (Ojukwu et al., 2025).

Stage five profiles the deployed candidate under load: per-frame latency distributions at full multi-camera concurrency, sustained power draw, and thermal trajectory under continuous operation, since automotive enclosures throttle and a detector that meets its budget cold may miss it warm (Sze et al., 2017; Lin et al., 2018; Oyeleke et al., 2026). The profiling harness itself is scripted and version-controlled rather than operated manually, following instrument-control and test-automation practice in hardware characterization, so that a measurement can be repeated identically months later on a different unit (Quainoo et al., 2025b; Quainoo & Ogundapo, 2026b). Configurations surviving all gates enter the deployable frontier with their measured profiles attached, and the pipeline's artifacts, training recipes, compression schedules, export configurations, and probe-set divergence records, are versioned as a unit so that any deployed operating point is reproducible from source, the systems discipline the safety standardization context expects of perception components (International Organization for Standardization, 2018; Kato et al., 2018). The gating structure follows the automated release-gating pattern developed for model governance, in which promotion between stages is conditioned on evaluation evidence rather than on reviewer judgment alone (Nwakamma et al., 2024a), and the retained artifact set is scoped to satisfy the documentation expectations that model risk management and high-risk AI regulatory frameworks place on deployed automated decision components (Ominyi & Anichukwueze, 2023; Ominyi et al., 2024b). Closed-loop process control practice in high-precision manufacturing offers the pattern for the pipeline's steady state, where a deployed configuration and its reference model are continuously compared and drift triggers requalification rather than silent acceptance (Olodo et al., 2025; Olodo et al., 2026a).

6. ON-VEHICLE DEPLOYMENT ARCHITECTURE

The deployment architecture allocates the accelerator among camera streams under a freshness-first scheduling policy. Each camera feeds a bounded single-slot queue: a new frame replaces an unprocessed predecessor rather than queuing behind it, so the detector always consumes the freshest available frame and staleness, not backlog, is the failure mode under transient overload, the conflation policy appropriate wherever superseded state has no consumer, applied here to perception (Lin et al., 2018). Detection instances are placed per camera group with resolution profiles from the Section 4 search, and the scheduler enforces the tail-latency contract per stream, degrading resolution before frame rate when thermal throttling shrinks the budget, since range degrades gracefully while reaction time does not. The allocation problem itself is a constrained resource-assignment problem of the kind the cloud scheduling literature formalizes, and its treatments of energy-efficient placement, constraint satisfaction under contention, and predictive scaling of allocated capacity transfer to the on-vehicle setting with the substitution of accelerator partitions for virtual machines (Ahmed & Odejobi, 2018a; Ahmed & Odejobi, 2018b; Ahmed et al., 2019; Ahmed et al., 2020).

Placement analysis addresses the offload question explicitly. Edge-server offload of perception is attractive on energy grounds and is the general pattern edge intelligence surveys analyze (Chen & Ran, 2019; Zhou et al., 2019; Deng et al., 2020), with applied studies showing where partitioned inference across edge and cloud tiers succeeds and where connectivity variance defeats it (Idika et al., 2021; Komi & Ganiu, 2023; Ojukwu et al., 2025). But the safety path cannot depend on connectivity whose variability the vehicular setting guarantees; the architecture therefore fixes all detection on-vehicle, admits offload only for non-safety consumers, logging, fleet analytics, model improvement data collection, and treats the on-vehicle deployable frontier as the sole source of the driving function's operating points (Satyanarayanan, 2017; Kato et al., 2018). The offload channel is a security surface as well as a bandwidth one, and its threat model follows the operational technology security literature: segmentation and zero-trust access control between the safety path and the analytics path (Ahmed et al., 2021), anomaly detection over the telemetry the channel carries (Adebayo et al., 2023; Ominyi et al., 2024a), resilience of the cyber-physical control loop when the channel is compromised (Ojukwu et al., 2023; Shittu et al., 2024), and, where the offload path is used to update or steer on-vehicle behavior, protection against the manipulation of automated agents that agentic deployment security work identifies (Nwakamma et al., 2024b; Jimoh et al., 2023).

Downstream integration passes per-frame detections with timestamps and per-class confidences to tracking and fusion, whose multi-modal context the domain surveys supply (Feng et al., 2021; Arnold et al., 2019); the interface contract is that the detector's operating point, thresholds realizing the recall floors, is set by system-level analysis, not by benchmark-optimal defaults, which Section 7's third phase operationalizes. Runtime health is monitored against a reference profile in the manner of digital twin practice, so that divergence in latency distribution, detection rate, or confidence calibration is surfaced as a maintenance signal rather than absorbed silently (Shittu et al., 2023; Shittu & Shittu, 2024; Komi & Adeniji, 2023), and fleet-level coordination of model versions and configuration rollouts follows the patterns developed for multi-agent robot fleets and for staged migration of production systems (Akanbi et al., 2025; Ladapo et al., 2025).

7. EVALUATION PLAN

Evaluation proceeds in three phases against the requirements of Section 3. Phase one establishes accuracy and robustness. Candidates from the deployable frontier are evaluated on the driving benchmarks with class-stratified reporting, per-class precision-recall at deployment-relevant operating points and recall at fixed false-positive rates for vulnerable-road-user classes, alongside conventional mean average precision for comparability (Geiger et al., 2013; Cordts et al., 2016; Yu et al., 2020; Caesar et al., 2020; Everingham et al., 2010; Lin et al., 2014). Robustness is measured on the corruption suites with per-corruption degradation profiles, and the prespecified contrast is between configurations trained with and without corruption augmentation at matched clean accuracy (Hendrycks & Dietterich, 2019; Michaelis et al., 2019), with adversarially perturbed inputs evaluated as a separate degradation family rather than folded into the corruption average (Adebayo et al., 2022; Adeniyi et al., 2025). Compression accounting reports accuracy deltas per stage, pruning, quantization, export, isolating where the contract erodes (Nagel et al., 2021; Nwakamma et al., 2025).

Phase two measures deployment behavior on the target hardware: latency distributions with tail percentiles under full multi-camera load, energy per frame and sustained power, thermal trajectory and throttling onset under continuous operation, and the export-divergence bound from the pipeline's verification gate (Sze et al., 2017; Tan et al., 2019; Oyeleke et al., 2026). The measurement protocol is specified in advance and automated end to end, following hardware characterization practice in which the harness, the instrument configuration, and the analysis script are versioned alongside the results (Quainoo et al., 2025b; Quainoo & Ogundapo, 2026b), and degradation is reported as a curve against load and thermal state rather than as a single nominal figure, in the manner of physical-layer throughput degradation modeling (Quainoo et al., 2025a; Quainoo & Ogundapo, 2026a). Energy accounting reports both instantaneous draw and energy per useful detection, since the two diverge sharply across compression configurations and only the latter reflects what a duty-cycled or thermally constrained platform actually pays (Asiedu & Quainoo, 2024; Asiedu et al., 2025; Asiedu & Quainoo, 2023a; Asiedu & Quainoo, 2023b; Asiedu et al., 2026). The reporting unit throughout is the model-toolchain-hardware triple, and the phase's product is the measured frontier, accuracy contract versus tail latency versus power, from which system engineering selects.

Phase three couples the component to system consequence. Using recorded driving sequences, the phase measures detection latency-to-first-detection for scenario-critical objects as a function of distance, converts the component's latency and recall profile into reaction-distance margins at representative speeds following the systems methodology that ties compute budgets to vehicle dynamics (Lin et al., 2018), and stress-tests the freshness scheduler under injected load and thermal conditions, verifying that degradation follows the specified resolution-before-rate policy. A fourth, longitudinal component runs beyond the acceptance decision: deployed configurations are monitored against their qualification profiles with the continuous-monitoring instrumentation that condition-based maintenance practice supplies (Omoegun et al., 2023; Adaramola et al., 2018; Sunday et al., 2020; Sunday & Omoegun, 2022), with real-time indicator tracking and dashboarding making drift visible to operators rather than discoverable only in postmortems (Tonoyan et al., 2024b; Basnet et al., 2023; Mayo et al., 2021), interpretable failure prediction used to anticipate degradation before it breaches a contract (Komi & Adamolekun, 2021), and anomaly detection over the resulting high-volume telemetry stream flagging conditions no fixed threshold anticipates (Ominyi et al., 2024a). Acceptance is defined against the Section 3 contract as a whole: a configuration passes when accuracy floors, tail budgets, power envelope, robustness bounds, and scheduling behavior hold simultaneously on the target platform, the conjunctive standard the concept argues is the only deployable meaning of real-time detection for vehicles.

8. ANTICIPATED APPLICATIONS

The primary application is the camera detection component within modular autonomous driving stacks, where the deployable frontier and its measured profiles give system integrators the selection surface that single-number model releases do not (Kato et al., 2018; Yurtsever et al., 2020). The methodology transfers laterally to adjacent embedded perception domains that share the constraint structure, advanced driver assistance functions on constrained electronic control units, mobile robotics, and infrastructure-side traffic perception, since nothing in the co-design formulation is specific to the vehicle beyond its budgets and object statistics (Chen & Ran, 2019; Zhou et al., 2019). Warehouse and micro-fulfillment robotics is the closest such domain, combining on-board perception, path planning under latency bounds, and fleet-level coordination on hardware chosen for cost rather than headroom (Akanbi & Sunday, 2024; Akanbi & Sunday, 2025; Akanbi et al., 2025). Aerial and corridor inspection is the second, where compressed detectors run on airframe compute with strict energy budgets and where the object statistics, small targets at variable standoff against cluttered backgrounds, stress the same design axes this concept parameterizes (Dagodzo, 2018b; Dagodzo & Ahiaeke Patrick, 2020; Dagodzo & Ahiaeke Patrick, 2021b; Dagodzo et al., 2022; Adeyelu & Dagodzo, 2024). Industrial and energy infrastructure supplies a third, where on-site inference feeds digital twin and condition-monitoring systems whose value depends on running continuously within a fixed power envelope (Shittu & Shittu, 2024; Shittu et al., 2026; Komi & Adeniji, 2023; Komi & Ganiu, 2023; Sunday & Omoegun, 2022), and clinical imaging supplies a fourth, where the recall-floor contract formulation matches an evaluation culture already organized around sensitivity at specified operating points rather than averaged precision (Ejofodomi et al., 2014).

Within research practice, the hardware-in-the-loop frontier and the class-critical evaluation protocol are themselves contributions: they provide a reporting template under which edge detection claims become comparable across works, addressing the proxy-metric and single-stream reporting gaps the concept identifies in the current literature (Tan et al., 2019; Zou et al., 2023; Nwakamma et al., 2025). The template also serves operational consumers of detection output beyond the driving function, including logistics and routing systems whose decision quality depends on the freshness and reliability of the perception feeding them (Tonoyan et al., 2025; Tonoyan et al., 2024a), and public safety systems where detection latency and equitable performance across populations are jointly load-bearing (Odejide, 2024; Odejide, 2025). And the pipeline's data-collection offload path positions deployed fleets as the continual source of hard examples for successive training rounds, the improvement loop that driving perception development already institutionalizes at dataset scale (Sun et al., 2020; Yu et al., 2020), with privacy-preserving federated formulations available where raw frame retention is undesirable (Sanni et al., 2023).

9. LIMITATIONS AND THREATS TO VALIDITY

The concept's scope bounds its claims. It addresses 2D camera detection; full driving perception is multi-modal and three-dimensional, and the component's outputs are inputs to fusion rather than a perception solution, with the 3D and fusion literatures marking what lies beyond scope (Arnold et al., 2019; Feng et al., 2021). Benchmark-to-road validity is the standing external threat: dataset object statistics and corruption suites approximate but do not exhaust operational conditions, and rare hazardous configurations are structurally under-represented, so phase-one results lower-bound the validation burden rather than discharging it, a limitation the safety standardization context makes explicit (International Organization for Standardization, 2018; Michaelis et al., 2019). Label quality is a related and frequently understated threat, since recall floors are only as meaningful as the ground truth defining them, and annotation disagreement on exactly the small, occluded, and ambiguous instances that dominate the vulnerable-road-user classes is both high and systematic (Ladapo et al., 2024; Ladapo et al., 2026). Hardware specificity cuts both ways: measured co-design is the concept's methodological point, but its frontiers are per-platform, and portability of conclusions across accelerator generations is an empirical question the versioned pipeline supports investigating rather than a property to assume (Sze et al., 2017; Cai et al., 2020; Gideon, 2025a; Gideon & Erhimefe, 2025; Olodo et al., 2026b).

Method-level threats receive planned controls. Compression interactions, pruning-quantization couplings that per-stage accounting can miss, are probed by ablation ordering in the pipeline; distillation dependence on teacher quality is bounded by reporting student performance with and without distillation; and search overfitting to the profiling workload is controlled by held-out multi-camera load traces in phase two. The threat model is also incomplete by construction: the evaluation plan measures incidental corruption thoroughly and adversarial manipulation only partially, and a determined attacker with knowledge of the deployed configuration operates outside what any corruption suite bounds (Adebayo et al., 2022; Adeniyi et al., 2025). Distributional fairness is a further open question the plan surfaces but does not resolve, since class-stratified recall floors do not by themselves guarantee uniform performance across the populations and conditions a deployed fleet encounters, and the accountability literature is clear that this gap is where automated systems most often fail their users and their regulators (Annan, 2024; Annan, 2025a; Odejide, 2025; Ominyi & Anichukwueze, 2024). The evaluation's residual honesty constraint is stated plainly: phase three's reaction-distance coupling is an engineering analysis over recorded data and platform measurement, not a road-safety claim, and no result of the plan substitutes for the system-level validation that deployment on public roads requires (Lin et al., 2018; Kato et al., 2018; Ominyi et al., 2024b).

10. DISCUSSION

The concept's position is that the interesting object of study is the operating point, not the model. The detection literature's frontier progress is real (Zou et al., 2023; Wang et al., 2023), and the compression literature's instruments are mature (Nagel et al., 2021; Han et al., 2016); what vehicle deployment adds is a conjunction of constraints under which neither literature's headline numbers select a configuration, because accuracy interacts with latency through resolution, latency with power through frequency and thermals, and all three with the toolchain through export fidelity. That conjunction is not unique to vehicles: on-device model deployment generally, and edge inference under variable energy budgets specifically, exhibit the same non-separability of accuracy, latency, energy, and reliability (Nwakamma et al., 2025; Asiedu & Quainoo, 2024; Oyesiji et al., 2025). Making the triple the unit of design and evaluation is the concept's methodological contribution, and its practical corollary, publish frontiers with measured profiles, not points with proxy costs, is a reporting norm the embedded perception field is positioned to adopt (Tan et al., 2019; Sze et al., 2017; Nwakamma et al., 2024a).

The safety framing carries the second implication. Recall floors for vulnerable classes at fixed false-positive operating points, tail latency as the primary temporal metric, and degradation policies specified in advance are all translations of driving risk into component contracts, and they change what optimization sees: a configuration that wins on mean average precision and mean latency can lose on every contract that matters at deployment (Cordts et al., 2016; Lin et al., 2018). This translation is also what makes the component auditable, since a contract stated in advance and measured under specified conditions is the form of evidence that model risk management, algorithmic accountability, and regulatory readiness frameworks are built to consume (Ominyi & Anichukwueze, 2023; Ominyi et al., 2024b; Annan, 2025b), and explainability requirements attach more naturally to a stated operating point than to an aggregate benchmark score (Sanni et al., 2026). The concept's evaluation plan operationalizes that translation, and its adoption would align the incentives of detector research with the properties the application actually prices.

A third implication concerns the deployed lifetime of the component. Nothing in a qualification result guarantees its own persistence: accelerators age, thermal paths degrade, toolchains are updated, and the operational distribution drifts away from the corpus the configuration was selected against. The maintenance and monitoring literatures treat this as the normal condition of a deployed asset rather than an exception, and their instruments, reference profiles, continuous comparison, drift detection, and requalification triggers, apply to a detector as readily as to rotating machinery or a substation (Adaramola et al., 2018; Sunday et al., 2020; Shittu et al., 2023; Komi & Adamolekun, 2021; Olodo et al., 2026a). Treating the operating point as a maintained property rather than a shipped one is the natural extension of the concept's central claim, and the offload path's analytics channel, combined with federated or privacy-preserving aggregation where raw data cannot leave the vehicle, is the mechanism through which the fleet supplies the evidence that maintenance requires (Sanni et al., 2023; Ladapo et al., 2026; Ahmed et al., 2020).

11. CONCLUSION

This paper has developed a research concept for edge-optimized YOLO-based perception for autonomous vehicles, organized around a single methodological commitment: the deployable object is the model-compression-hardware triple, designed and evaluated under vehicle-grade contracts. The concept contributes a parameterized co-design space joining the YOLO architectural lineage with structured pruning, integer quantization-aware training, and distillation; a verification-gated optimization and deployment pipeline whose search is hardware-in-the-loop; an on-vehicle deployment architecture with freshness-first scheduling and a safety-path placement analysis that keeps detection on-vehicle; and a three-phase evaluation plan that measures accuracy and robustness with class-critical contracts, deployment behavior with tail, power, and thermal profiles, and system coupling through reaction-distance analysis.

If executed, the plan yields deployable frontiers with measured profiles for automotive edge platforms and a reporting template that makes edge detection claims comparable and contract-relevant (Nwakamma et al., 2025; Oyesiji et al., 2025; Nwakamma et al., 2024a). The broader claim stands independent of any single result: real-time perception for vehicles is not a fast model but a verified operating point, and the discipline of designing, compressing, scheduling, and evaluating against that standard is what converts detection research into a component fit for the safety-critical systems that will carry it (Gideon et al., 2019; Ominyi et al., 2024b; Jimoh et al., 2023).

REFERENCES

1) Adaramola, T. S., Fadero, S., & Gideon, E. N. (2018). Predictive maintenance and condition monitoring in critical power and energy infrastructure. Iconic Research and Engineering Journals, 2(5), 454-477. https://doi.org/10.64388/IREV2I5-1722489
2) Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2022). Adversarial machine learning in critical infrastructure: A conceptual framework for threat modeling AI enabled OT systems. World Journal of Innovation and Modern Technology, 6(1), 184-234. https://doi.org/10.56201/wjimt.v6.no1.2022.pg184.234
3) Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2023). AI augmented threat detection in industrial control systems: A systematic review of machine learning approaches for ICS anomaly detection. International Journal of Engineering and Modern Technology, 9(3), 287-340. https://doi.org/10.56201/ijcsmt.v9.no3.2023.pg287.340
4) Adeniyi, A. I., Odejobi, O., & Taiwo, T. (2025). Countermeasures against bias and spoofing in modern facial recognition systems. World Journal of Advanced Research and Reviews, 25(1), 1914-1930.
5) Adeyelu, O. O., & Dagodzo, D. (2024). Advances in artificial intelligence and data-driven wildlife hazard monitoring and incident reduction at international airports in West Africa. International Journal of Scientific Research in Science and Technology, 11(5), 872-910. https://doi.org/10.32628/IJSRST52310285
6) Ahmed, K. S., & Odejobi, O. D. (2018a). Conceptual framework for scalable and secure cloud architectures for enterprise messaging. IRE Journals, 2(1), 1-15.
7) Ahmed, K. S., & Odejobi, O. D. (2018b). Resource allocation model for energy-efficient virtual machine placement in data centers. IRE Journals, 2(3), 1-10.
8) Ahmed, K. S., Odejobi, O. D., & Oshoba, T. O. (2019). Algorithmic model for constraint satisfaction in cloud network resource allocation. IRE Journals, 2(12), 516-532.
9) Ahmed, K. S., Odejobi, O. D., & Oshoba, T. O. (2020). Predictive model for cloud resource scaling using machine learning techniques. Journal of Frontiers in Multidisciplinary Research, 1(1), 173-183.
10) Ahmed, M. O., Adegbite, M. P., & Adebayo, A. (2021). Zero trust architecture for operational technology in North American critical infrastructure: A framework for implementation and resilience optimization. International Journal of Engineering and Modern Technology, 7(1), 66-113. https://doi.org/10.56201/ijemt.vol.7.no1.2021.pg66.113
11) Akanbi, O., & Sunday, E. A. (2024). An integrated path-planning and slotting optimization model for AMR-enabled high-density warehousing. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3226-3243.

https://doi.org/10.62225/2583049X.2024.4.6.6181
12) Akanbi, O., & Sunday, E. A. (2025). A systematic review of AI-driven autonomous mobile robots (AMR) in scalable micro-fulfillment centers. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2363-2379. https://doi.org/10.62225/2583049X.2025.5.6.6182
13) Akanbi, O., Ganiu, O. S., & Sunday, E. A. (2025). A multi-agent AI framework for swarm intelligence in autonomous mobile robot (AMR) fleet coordination. International Journal of Scientific Research in Humanities and Social Sciences, 2(1), 81-109.
14) Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. Shodhshauryam, International Scientific Refereed Research Journal, 7(5), 315-347.
15) Annan, A. O. (2025a). Automated decision-making and anti-discrimination compliance under U.S. law. International Journal of Advanced Multidisciplinary Research and Studies, 5(2), 2522-2540.
16) Annan, A. O. (2025b). Cybersecurity compliance as a source of competitive advantage in technology markets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(5), 456-487.
17) Arnold, E., Al-Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., & Mouzakitis, A. (2019). A survey on 3D object detection methods for autonomous driving applications. IEEE Transactions on Intelligent Transportation Systems, 20(10), 3782-3795.
18) Asiedu, W., & Quainoo, R. (2023a). How far can energy harvesting take us? A systematic review of radio frequency strategies for energy autonomous sensing. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 448-466.
19) Asiedu, W., & Quainoo, R. (2023b). Toward maintenance free wireless infrastructure: Simulating performance and reliability in large scale intermittently powered IoT networks. Gyanshauryam, International Scientific Refereed Research Journal, 6(1), 489-510.
20) Asiedu, W., & Quainoo, R. (2024). Rethinking energy, reliability, and latency trade-offs in green communication for next generation IoT. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1987-1994.
21) Asiedu, W., Quainoo, R., & Asiedu, A. (2025). A conceptual framework for characterizing fundamental energy, reliability, and latency trade-offs in green communication paradigms for next-generation IoT. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2447-2453. https://doi.org/10.62225/2583049X.2025.5.6.6479
22) Asiedu, W., Quainoo, R., & Asiedu, A. (2026). Predictive power management for intermittent IoT devices using lightweight machine learning under uncertain energy harvest. Gulf Journal of Engineering and Technology, 2(4), 108-118.
23) Basnet, A., Oghenemaiga, E., & Anene, U. N. (2023). Real time analytics and monitoring systems for livestream media performance using Google platforms. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(1).
24) Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
25) Caesar, H., Bankiti, V., Lang, A. H., et al. (2020). nuScenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11621-11631.
26) Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2020). Once-for-all: Train one network and specialize it for efficient deployment. Proceedings of the International Conference on Learning Representations.
27) Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. Proceedings of the European Conference on Computer Vision, 213-229.
28) Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655-1674.
29) Cordts, M., Omran, M., Ramos, S., et al. (2016). The Cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213-3223.
30) Dagodzo, D. (2018a). A conceptual framework for UAV integration into national power grid inspection programs. Iconic Research and Engineering Journals, 2(5), 391-412. https://doi.org/10.64388/IREV2I5-1716082
31) Dagodzo, D. (2018b). A review of UAV applications in electrical transmission line inspection: Methods, technologies, and challenges. Iconic Research and Engineering Journals, 2(6), 234-254. https://doi.org/10.64388/IREV2I6-1716083
32) Dagodzo, D., & Ahiaeke Patrick, M. C. (2020). UAV-based pipeline and corridor monitoring: A review of current practices and emerging technologies. Iconic Research and Engineering Journals, 3(10), 574-597. https://doi.org/10.64388/IREV3I10-1716084
33) Dagodzo, D., & Ahiaeke Patrick, M. C. (2021a). A review of GIS applications in utility asset management and infrastructure planning. Iconic Research and Engineering Journals, 5(3), 468-492. https://doi.org/10.64388/IREV5I3-1716085
34) Dagodzo, D., & Ahiaeke Patrick, M. C. (2021b). An integrated framework for UAV, LiDAR, and GIS in infrastructure corridor management. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 497-524. https://doi.org/10.32628/CSEIT217566
35) Dagodzo, D., & Ahiaeke Patrick, M. C. (2022). A review of right-of-way encroachment detection methods using geospatial technologies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 638-667. https://doi.org/10.32628/CSEIT2281226
36) Dagodzo, D., Ahiaeke Patrick, M. C., & Aliliele, C. (2022). AI and deep learning for vegetation classification in power corridor management: A review. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 668-698. https://doi.org/10.32628/CSEIT2281227
37) Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469.
38) Ejofodomi, O. A., Gideon, E. N., Oladipo, G. O., & Oshomah, E. R. (2014). Automated detection of architectural detection in mammograms using template matching. International Journal of Biomedical Science and Engineering, 2(1), 1-6.

https://doi.org/10.11648/j.ijbse.20140201.11

39) Erhimefe, D. E., & Gideon, E. (2025). Developing advanced materials and manufacturing processes for high-performance, energy-efficient, and scalable 3D integrated circuits (ICs) and devices. International Journal of Research Publication and Reviews, 6(4), 2624-2630. https://doi.org/10.55248/gengpi.6.0425.1402
40) Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2), 303-338.
41) Feng, D., Haase-Schütz, C., Rosenbaum, L., et al. (2021). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.
42) Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. Proceedings of the International Conference on Learning Representations.
43) Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430.
44) Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 32(11), 1231-1237.
45) Gideon, E. N. (2025a). Evaluation of semiconductor risk mitigation strategies in the electric vehicle supply chain. International Journal of Research Publication and Reviews, 6(4), 5934-5939. https://doi.org/10.55248/gengpi.6.0425.14129
46) Gideon, E. N. (2025b). Exploring and developing advanced RF MEMS switches for 5G applications, focusing on high performance solutions for RF front end modules. Mikailalsys Journal of Advanced Engineering International, 2(2), 171-180. https://doi.org/10.58578/mjaei.v2i2.5407
47) Gideon, E., & Erhimefe, D. E. (2025). Damage mechanisms in semiconductor materials caused by non-ionizing energy in space-based solar systems. International Journal of Research Publication and Reviews, 6(4), 5940-5949. https://doi.org/10.55248/gengpi.6.0425.14128
48) Gideon, E. N., Adaramola, T. S., & Fadero, S. (2019). Reliability engineering and failure analysis in complex engineering systems. Iconic Research and Engineering Journals, 2(11), 703-727. https://doi.org/10.64388/IREV2I11-1722490
49) Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440-1448.
50) Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580-587.
51) 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.
52) Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Proceedings of the International Conference on Learning Representations.
53) Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations.
54) Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
55) Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
56) Idika, C. N., Salami, E. O., Ijiga, O. M., & Enyejo, L. A. (2021). Deep learning driven malware classification for cloud-native microservices in edge computing architectures. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(4).
57) International Organization for Standardization. (2018). ISO 26262: Road vehicles - functional safety. ISO.
58) Jacob, B., Kligys, S., Chen, B., et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2704-2713.
59) Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends in Computer Graphics and Vision, 12(1-3), 1-308.
60) Jimoh, H. O., Abolle-Okoyeagu, C. J., Ahmed, M. O., & Lawal, N. O. (2023). Advancing security in IoT-driven critical infrastructure: A focus on smart transportation system. American Journal of Engineering Research, 12(12), 33-46.
61) Jouppi, N. P., Young, C., Patil, N., et al. (2017). In-datacenter performance analysis of a tensor processing unit. Proceedings of the International Symposium on Computer Architecture, 1-12.
62) Kato, S., Tokunaga, S., Maruyama, Y., et al. (2018). Autoware on board: Enabling autonomous vehicles with embedded systems. Proceedings of the ACM/IEEE International Conference on Cyber-Physical Systems, 287-296.
63) Komi, N. M., & Adamolekun, A. (2021). Interpretable machine learning for early failure prediction in distributed renewable energy assets. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 1015-1038. https://doi.org/10.54660/.IJMRGE.2021.2.6.1015-1038
64) Komi, N. M., & Adeniji, I. O. (2023). A tiered digital twin that brings predictive diagnostics to resource-constrained renewable energy sites. International Journal of Engineering and Modern Technology, 9(3), 287-347. https://doi.org/10.56201/ijemt.v9.no3.2023.pg287.347
65) Komi, N. M., & Ganiu, O. S. (2023). Edge intelligence for resilient microgrid control: Advances, energy sovereignty, and open challenges. Gyanshauryam, International Scientific Refereed Research Journal, 6(3), 525-586. https://doi.org/10.32628/GISRRJ236339
66) Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342.
67) Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2022). Human-in-the-loop machine learning: A state of the art. Journal of Frontiers in Multidisciplinary Research, 3(1), 656-669. https://doi.org/10.54660/.JFMR.2022.3.1.656-669

68) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. International Journal of Multidisciplinary Futuristic Development, 5(1), 81-95. https://doi.org/10.54660/IJMFD.2024.5.1.81-95
69) Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2025). Migration of applications and information systems to cloud computing infrastructure: Lessons from a South African retail bank. International Journal of Multidisciplinary Research and Growth Evaluation, 6(6), 1361-1375. https://doi.org/10.54660/.IJMRGE.2025.6.6.1361-1375
70) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2026). On the disagreement problem in human-in-the-loop federated machine learning. International Journal of Multidisciplinary Research and Growth Evaluation, 7(3), 178-192. https://doi.org/10.54660/.IJMRGE.2026.7.3.178-192
71) Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2017). Pruning filters for efficient ConvNets. Proceedings of the International Conference on Learning Representations.
72) Lin, S.-C., Zhang, Y., Hsu, C.-H., et al. (2018). The architectural implications of autonomous driving: Constraints and acceleration. Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems, 751-766.
73) Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a). Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117-2125.
74) Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.
75) Lin, T.-Y., Maire, M., Belongie, S., et al. (2014). Microsoft COCO: Common objects in context. Proceedings of the European Conference on Computer Vision, 740-755.
76) Liu, W., Anguelov, D., Erhan, D., et al. (2016). SSD: Single shot multibox detector. Proceedings of the European Conference on Computer Vision, 21-37.
77) Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the European Conference on Computer Vision, 116-131.
78) Mayo, W., Ogbole, J. I., Okoruwa, P. O., & Babatope, O. M. (2021). Designing an AI-predictive maintenance model for e-commerce systems using machine learning and cloud analytics. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 416-440. https://doi.org/10.32628/IJSRCSEIT
79) Michaelis, C., Mitzkus, B., Geirhos, R., et al. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484.
80) Molchanov, P., Tyree, S., Karras, T., Aila, T., & Kautz, J. (2017). Pruning convolutional neural networks for resource efficient inference. Proceedings of the International Conference on Learning Representations.
81) Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M., & Blankevoort, T. (2021). A white paper on neural network quantization. arXiv preprint arXiv:2106.08295.
82) Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024a). LLM-as-a-judge for automated model governance: A review of methods, biases, and release-gating practices. World Journal of Innovation and Modern Technology, 8(6), 185-216. https://doi.org/10.56201/wjimt.v8.no6.2024.pg185.216
83) Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024b). Securing agentic AI enterprise workflows against prompt injection, tool poisoning, memory manipulation, and excessive agency threats. World Journal of Innovation and Modern Technology, 8(6), 217-248. https://doi.org/10.56201/wjimt.v8.no6.2024.pg217.248
84) Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2025). On-device multimodal small language models for privacy-preserving edge intelligence through quantization, pruning, distillation, and energy-efficient inference. World Journal of Innovation and Modern Technology, 9(12), 271-302. https://doi.org/10.56201/wjimt.v9.no12.2025.pg271.302
85) Odejide, G. S. (2024). A trust-centered conceptual model for AI-driven multilingual emergency communication. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(6), 2781-2811. https://doi.org/10.32628/CSEIT2410794
86) Odejide, G. S. (2025). Toward inclusive emergency alerting: A conceptual framework for accessibility and equity in AI-driven public safety systems. Iconic Research and Engineering Journals, 9(5). https://doi.org/10.64388/IREV9I5-1722366
87) Ojukwu, J., Oyesiji, S. O., & Nwakamma, S. (2023). Cyber-physical ransomware defense in operational technology using physics-informed detection, digital twins, network telemetry, and resilient recovery. International Journal of Computer Science and Mathematical Theory, 9(5), 193-228. https://doi.org/10.56201/ijcsmt.v9.no5.2023.pg193.228
88) Ojukwu, J., Oyesiji, S. O., & Nwakamma, S. (2025). WebAssembly component model and WASI for portable cloud-to-edge software: Language interoperability, security, performance, and container comparisons. International Journal of Computer Science and Mathematical Theory, 11(12), 232-263. https://doi.org/10.56201/ijcsmt.vol.11.no12.2025.pg232.263
89) Olodo, A., Akanbi, O., & Adenuga, O. (2025). Conceptualizing a closed-loop digital twin for real-time plasma control in large-area magnetron sputtering systems. International Journal of Engineering and Modern Technology, 11(12), 205-227. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.227
90) Olodo, A., Akanbi, O., & Adenuga, O. (2026a). Conceptualizing a lean-standardized predictive maintenance framework for high-vacuum deposition equipment. International Journal of Engineering and Modern Technology, 12(3), 295-322. https://doi.org/10.56201/ijemt.vol.12.no3.2026.pg295.322
91) Olodo, A., Akanbi, O., & Adenuga, O. (2026b). Data-driven maintenance and reliability in high-volume semiconductor fabs: A review of Industry 4.0 methodologies. International Journal of Engineering and Modern Technology, 12(3), 323-348.

https://doi.org/10.56201/ijemt.vol.12.no3.2026.pg323.348
92) Ominyi, M., & Anichukwueze, C. C. (2023). A review of algorithmic accountability and model risk management approaches in financial services. Journal of Accounting and Financial Management, 9(12), 218-238. https://doi.org/10.56201/jafm.v9.no12.2023.pg218.238
93) Ominyi, M., & Anichukwueze, C. C. (2024). Conceptualizing responsible AI governance structures for automated decision systems in regulated industries. International Journal of Social Sciences and Management Research, 10(11), 403-430. https://doi.org/10.56201/ijssmr.v10.no11.2024.pg.403.430
94) Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024a). Developing a conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments. World Journal of Innovation and Modern Technology, 8(6), 204-228. https://doi.org/10.56201/wjimt.v8.no6.2024.pg204.228
95) Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024b). A systematic review of organizational readiness for the EU Artificial Intelligence Act in multinational enterprises. World Journal of Innovation and Modern Technology, 8(6), 185-203. https://doi.org/10.56201/wjimt.v8.no6.2024.pg185.203
96) Omoegun, G. O., Sunday, E. A., Essien, M. A., & Oluokun, O. A. (2023). Vibration-based condition monitoring of rotating machinery using LabVIEW. International Journal of Scientific Research in Civil Engineering, 7(6), 82-108.
97) Oyeleke, A. V., Eze, F. C., & Asiedu, W. (2026). Reliability-centered maintenance strategies for minimizing downtime and maximizing performance in high-density GPU cluster environments. International Journal of Engineering Technology Research & Management, 10(7), 18-38.
98) Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2023). Parameter-efficient multilingual adaptation of on-device language models: A survey. International Journal of Engineering and Modern Technology, 9(3), 287-320. https://doi.org/10.56201/ijcsmt.v9.no3.2023.pg287.320
99) Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2024). A conceptual framework for context-aware data curation in fine-tuning pipelines. International Journal of Engineering and Modern Technology, 10(11), 197-228. https://doi.org/10.56201/ijemt.v10.no11.2024.pg197.228
100) Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2025). Compressing recommender and ranking models for edge deployment: A survey. International Journal of Engineering and Modern Technology, 11(12), 205-237. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.237
101) Quainoo, R., & Ogundapo, O. (2026a). A system-level power behavior model for Bluetooth and Wi-Fi coexistence in dual-mode wireless devices. International Journal of Computer Science and Mathematical Theory, 12(2), 298-358. https://doi.org/10.56201/ijcsmt.vol.12.no2.2026.pg298.358
102) Quainoo, R., & Ogundapo, O. (2026b). Wireless system-on-chip performance in next-generation Internet of Things devices: A systematic review of integrated transceiver testing methodologies. World Journal of Innovation and Modern Technology, 10(5), 76-126. https://doi.org/10.56201/wjimt.v10.no5.2026.pg76.126
103) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025a). Predicting throughput degradation in Wi-Fi 6 networks under varying channel conditions: A physical layer performance model. International Journal of Engineering and Modern Technology, 11(12), 205-264. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.264
104) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025b). Test automation in wireless hardware engineering: A comprehensive review of scripting frameworks and instrument control strategies. International Journal of Engineering and Modern Technology, 11(10), 405-464. https://doi.org/10.56201/ijemt.vol.11.no10.2025.pg405.464
105) 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, 779-788.
106) Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7263-7271.
107) Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.
108) Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28, 91-99.
109) Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510-4520.
110) Sanni, J. O., Iwuanyanwu, U. A., & Essien, M. A. (2026). Designing explainable AI based marketing automation architectures for healthcare and financial applications. World Scientific News, 213, 119-148. https://doi.org/10.65770/MJFK8593
111) Sanni, J. O., Iwuanyanwu, U. A., Essien, M. A., & Attah, A. (2023). Lifecycle-aware marketing automation using federated learning for secure cross-organizational data management. Gyanshauryam, International Scientific Refereed Research Journal, 6(6), 337-364. https://doi.org/10.32628/GISRRJ236642
112) Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30-39.
113) Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637-646.
114) Shittu, H., & Shittu, M. A. (2024). AI powered digital twins for predictive maintenance and operational optimization of renewable energy systems. International Journal of Science, Architecture, Technology and Environment, 1, 151-163. https://doi.org/10.63680/ijsate0324085.011
115) Shittu, H., Adeniji, I. O., Oteri, O., & Shittu, M. A. (2026). Autonomous energy management systems for port and maritime electrical infrastructure using digital-twin-driven architectures. International Journal of Advanced Multidisciplinary Research and Studies, 6(1), 1792-1805. https://doi.org/10.62225/2583049X.2026.6.1.5740

116) Shittu, H., Olagunju, F., & Shittu, M. A. (2024). Cyber physical resilience in digital substations: IoT enabled adaptive protection for secure DER integration. International Journal of Science, Architecture, Technology and Environment, 1(3), 81-99. https://doi.org/10.63680/ijsate032532.08
117) Shittu, M. A., Shittu, H. A., Adeleke, O. J., & Adedokun, O. J. (2023). Digital twin modeling for real time monitoring and fault detection in smart substations. International Journal of Industrial Engineering Research and Development, 14(2), 25-44. https://doi.org/10.34218/IJIERD_14_02_003
118) Sun, P., Kretzschmar, H., Dotiwalla, X., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2446-2454.
119) Sunday, E. A., & Omoegun, G. O. (2022). Smart fault detection in HVAC systems using sensor-based monitoring. International Journal of Scientific Research in Science, Engineering and Technology, 7(4), 380-403.
120) Sunday, E. A., Omoegun, G. O., Essien, M. A., & Oluokun, O. A. (2020). Transitioning from reactive to predictive maintenance in mechanical systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(6), 425-447.
121) Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.
122) Tan, M., Chen, B., Pang, R., et al. (2019). MnasNet: Platform-aware neural architecture search for mobile. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2820-2828.
123) Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning, 6105-6114.
124) Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10781-10790.
125) Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 9627-9636.
126) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2024a). Advances in supply chain resilience: Predictive models for vendor risk assessment and procurement cost optimization. International Journal of Social Sciences and Management Research, 10(11), 525-551. https://doi.org/10.56201/ijssmr.v10.no11.2024.pg.525.551
127) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2024b). Real-time KPI tracking systems: A review of automated performance monitoring and data-driven decision making. World Journal of Innovation and Modern Technology, 8(6), 247-281. https://doi.org/10.56201/wjimt.v8.no6.2024.pg247.281
128) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2025). A conceptual model for real-time route optimization: Algorithmic pathways to delivery profitability and logistics cost reduction. World Journal of Innovation and Modern Technology, 9(12), 308-364. https://doi.org/10.56201/wjimt.v9.no12.2025.pg308.364
129) Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464-7475.
130) Yu, F., Chen, H., Wang, X., et al. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2636-2645.
131) Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. IEEE Access, 8, 58443-58469.
132) Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738-1762.
133) Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. Proceedings of the International Conference on Learning Representations.
134) Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.