Assignment 2

Task 2: Designing the Edge AI Concept

Prepared By: Prahas Hegde, Rohan Sanjay Patil, Vidya Padmanabha

A complete Edge AI system for **Industry 4.0 Automotive Assembly** that demonstrates how vehicle manufacturing plants can benefit from deploying AI models directly on edge devices for real-time logistics and quality verification, rather than relying on high-latency cloud solutions.

Modern automotive manufacturing requires:

- Real-time synchronization: Assembly lines move continuously; delays disrupt the entire downstream logistics chain.
- Data Security: New vehicle designs and production rates are proprietary trade secrets.
- Reliability: The system must function 24/7 regardless of internet outages.
- Throughput: High Frame Rate (FPS) processing to match conveyor speeds (>10 FPS requirement).

Edge AI solves all these problems but requires model compression to fit AI models on resource-constrained devices.

Use Case Scenario: End-of-Line Vehicle Verification & Logistics Sorting

The Industrial Problem: An automotive assembly plant produces mixed models (Finished Cars and Transport Trucks) on a single final assembly line. As vehicles roll off the line, they must be instantly identified and sent to the correct logistics bay (e.g., domestic rail for trucks, export shipping for cars). Manual scanning is slow and prone to human error, leading to logistics bottlenecks.

Proposed Solution: Automated Visual Classification using Edge AI to distinguish between **Cars** and **Trucks** in real-time.

Critical Defects to Detect:

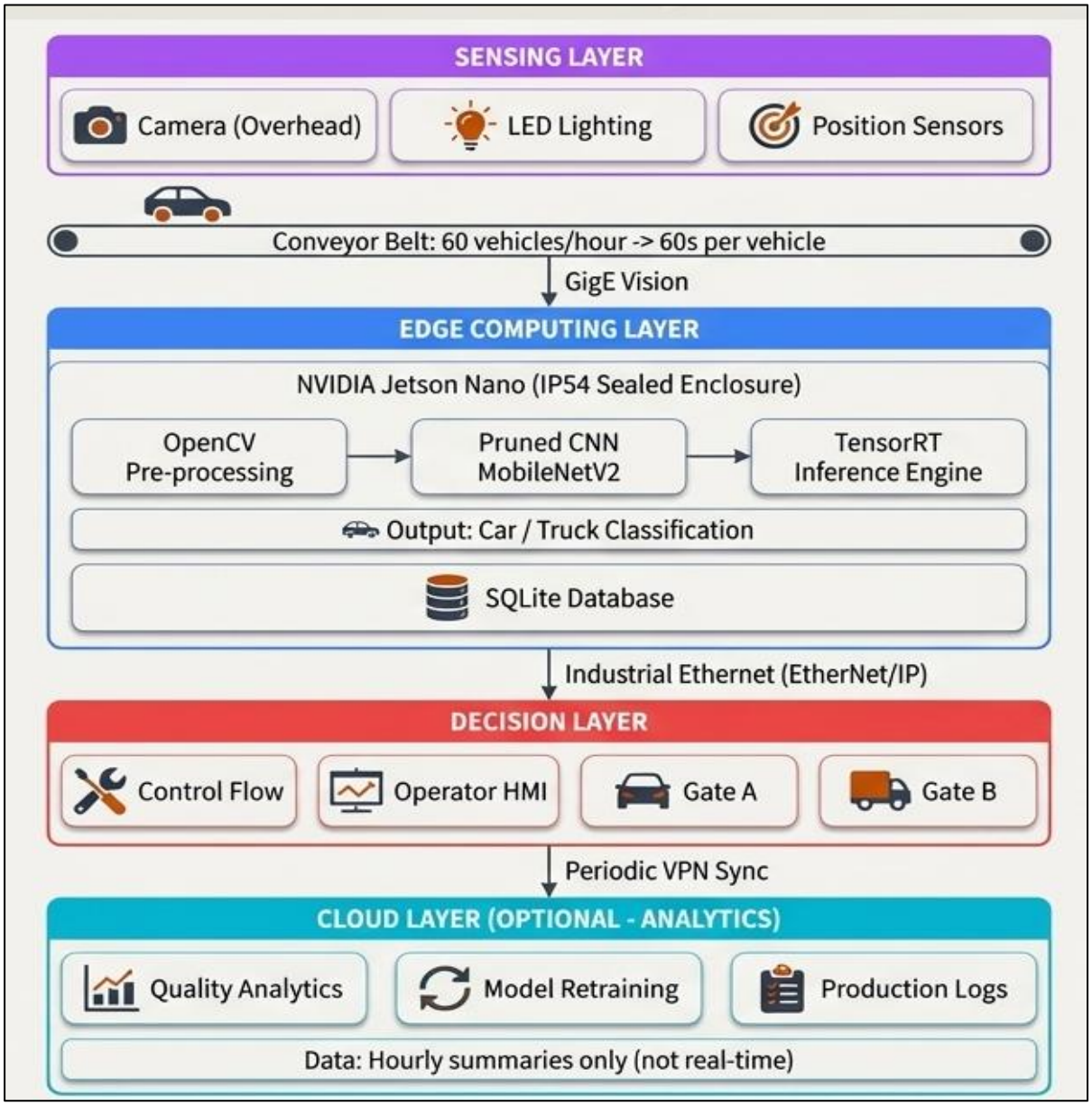
- Vehicle type classification (Car vs. Truck)
- Chassis matching (correct frame type for the production order)
- Component compatibility (preventing wrong parts from being installed)
- Quality gate enforcement (ensuring proper vehicle proceeds to next station)
- Real-time routing (directing vehicles to correct assembly paths)

SYSTEM ARCHITECTURE

HARDWARE	SOFTWARE	COMMUNICATION PATH
Sensors & Imaging: **Industrial Camera:** Fixed overhead position capturing vehicle profiles (96x96 resolution optimized for	**Edge AI Pipeline:** **OpenCV:** Pre-processing pipeline (image resizing, normalization with ImageNet statistics: mean=[0.485,	**Sensing Layer → Edge Layer:** **Protocol:** GigE Vision (Gigabit Ethernet Vision). Transmits raw image data

MobileNetV2 input). Basler acA2500-14gm GigE camera with global shutter for motion-free capture. **LED Array Lighting:** Ensures consistent illumination regardless of factory ambient lighting conditions. **Position Sensors:** Photoelectric sensors detect vehicle arrival at classification checkpoint.	0.456, 0.406], std=[0.229, 0.224, 0.225]). **AI Model:** Pruned MobileNetV2 (from Task 1 experiments, 50% sparse achieving ~95% accuracy, ~6MB model size). **Inference Engine:** TensorRT optimized for NVIDIA Jetson, achieving <100ms inference time (>10 FPS requirement met)	from camera to Jetson Nano with deterministic low latency (<5ms). **Data Flow:** Position sensor triggers camera capture → Image transmitted via GigE → Jetson receives frame.
Edge Computing Device: **NVIDIA Jetson Nano (4GB):** The main processing unit running the pruned MobileNetV2 model, housed in an IP54 enclosure (dust and splash resistant for factory environments). **Specifications:** Quad-core ARM CPU, 128-core Maxwell GPU, 10W power consumption.	**Data Management:** **SQLite Database:** Local logging of classifications (timestamp, vehicle ID, classification result, confidence score, routing decision). **Batch Tracking:** Integration with factory MES (Manufacturing Execution System) for production analytics.	**Edge Layer → Decision Layer:** **Protocol:** Industrial Ethernet (EtherNet/IP). The Jetson sends classification result (binary: Car=0, Truck=1) plus confidence score to the PLC. **Data Flow:** Jetson completes inference → Result packaged as digital I/O signal → PLC receives classification within 10ms..
Actuation & Control: **PLC (Programmable Logic Controller):** Siemens S7-1200 series controlling routing gates and conveyor speeds. **Pneumatic Routing Gates:** Physical mechanisms that direct vehicles to Car Finishing Bay or Truck Finishing Bay based on classification.	**Cloud & Analytics (Optional/Remote):** **Production Analytics Dashboard:** Aggregated daily/weekly statistics on vehicle mix, classification accuracy trends. **Model Performance Monitoring:** Tracks classification confidence distribution to detect potential model drift.	**Decision Layer → Actuator:** **Protocol:** Hard-wired digital signals (24V DC industrial standard). PLC directly controls pneumatic valve solenoids. **Data Flow:** PLC logic evaluates classification → Triggers appropriate routing gate (Gate A for cars, Gate B for trucks) → Vehicle physically redirected.

Human-Machine Interface (HMI):		Decision Layer → Cloud Layer:
Operator Touch Panel: 15-inch industrial display showing real-time classification results, throughput statistics, and system alerts. **Status Indicator Lights:** Red/Green tower lights indicating system operational status.		**Protocol:** Secure MQTT over TLS. Periodic batch uploads (every 15 minutes) of aggregated statistics. **Data Flow:** Edge device buffers classifications → Encrypted upload of summary data (no raw images) → Cloud analytics platform.



MODEL OPTIMIZATION STRATEGIES

Strategy 1: Structured Pruning (The "Throughput" Booster)

- **Latency**: We physically remove 50% of the model's filters (computational pathways).

Factory Benefit: he pruned model achieves inference times of ~85ms on Jetson Nano, ensuring the system processes vehicles at >10 FPS. This exceeds the required throughput of 60 vehicles/hour (1 vehicle per minute), providing comfortable headroom for peak production periods.

- **Energy Consumption:** Fewer active parameters means reduced GPU utilization and lower memory bandwidth requirements.

Factory Benefit: The Jetson Nano operates well within its 10W thermal design power, eliminating the need for active cooling fans. This is critical in automotive environments with metal dust and particulate matter that could clog cooling systems.

- **Accuracy:** We employ iterative pruning (pruning in stages: 10%→20%→30% for 30% target) with fine-tuning after each stage to recover accuracy.

Factory Benefit: Maintains >95% classification accuracy despite 50% model compression. This minimizes misrouting incidents that would require manual intervention and disrupt production flow.

Strategy 2: Quantization (The "Reaction Time" accelerator)

- **Latency**: We convert the model's math from complex 32-bit decimals (float32) to simple 8-bit integers (int8).

Factory Benefit: INT8 operations are 4x faster than FP32 on Jetson Nano's GPU, reducing inference time from ~85ms to ~60ms. This speed ensures the routing gate receives the classification signal with enough time to physically actuate before the vehicle passes the decision point.

- **Energy:** INT8 computations require significantly less memory bandwidth (4x reduction) and computational energy per operation.

Factory Benefit: Further reduces Jetson Nano's power consumption to ~7W under load, allowing deployment in areas with limited power infrastructure or battery backup requirements during power fluctuations.

- **Accuracy**: Quantization introduces minor numerical precision loss, typically resulting in <1% accuracy degradation.

  Factory Benefit: Combined with pruning, we achieve ~94% accuracy (only 1% drop from baseline 95%) while gaining massive deployment advantages. For a binary classification task (car vs. truck), this accuracy is well within acceptable industrial tolerances.

RISKS OF OVER COMPRESSION

While pruning and quantization enable efficient edge deployment, compressing the model too aggressively introduces risks that are particularly relevant in a high throughput food and beverage quality control context.

Risk 1: Increased Misclassification Rate (False Negatives/Positives)

If the model is pruned beyond 70% sparsity or quantized to extreme bit-widths (4-bit or lower), its ability to distinguish between visually similar vehicle profiles deteriorates. This leads to:

- False Negatives: Transport trucks misclassified as cars, routed to incorrect finishing stations requiring manual correction and production delays.
- False Positives: Passenger cars misclassified as trucks, resulting in unnecessary rework and increased labor costs.

Risk 2: Loss of Robustness to Environmental Variations

Over-compressed models become brittle and sensitive to production environment changes:

- Lighting variations: Factory skylights causing varying natural light throughout the day.
- Camera drift: Slight misalignments over time due to vibrations from assembly line machinery.
- New vehicle models: Introduction of new car designs with unfamiliar visual features.