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 that demonstrates how **food & beverage manufacturing** companies can benefit from deploying AI models directly on edge devices instead of relying on cloud solutions.

Modern manufacturing requires:

- Real-time decision making (milliseconds, not seconds)
- Data privacy (sensitive production data stays in-house)
- Reliability (no dependency on internet connectivity)
- Cost efficiency (no continuous cloud API fees)
- Hygiene compliance (minimal equipment, sealed systems)

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

Use Case Scenario: Automated Beverage Bottle Quality Inspection

The Industrial Problem: A beverage bottling facility fills and packages 500 bottles per minute (high-speed production line). Each bottle must be inspected for multiple defects before leaving the factory.

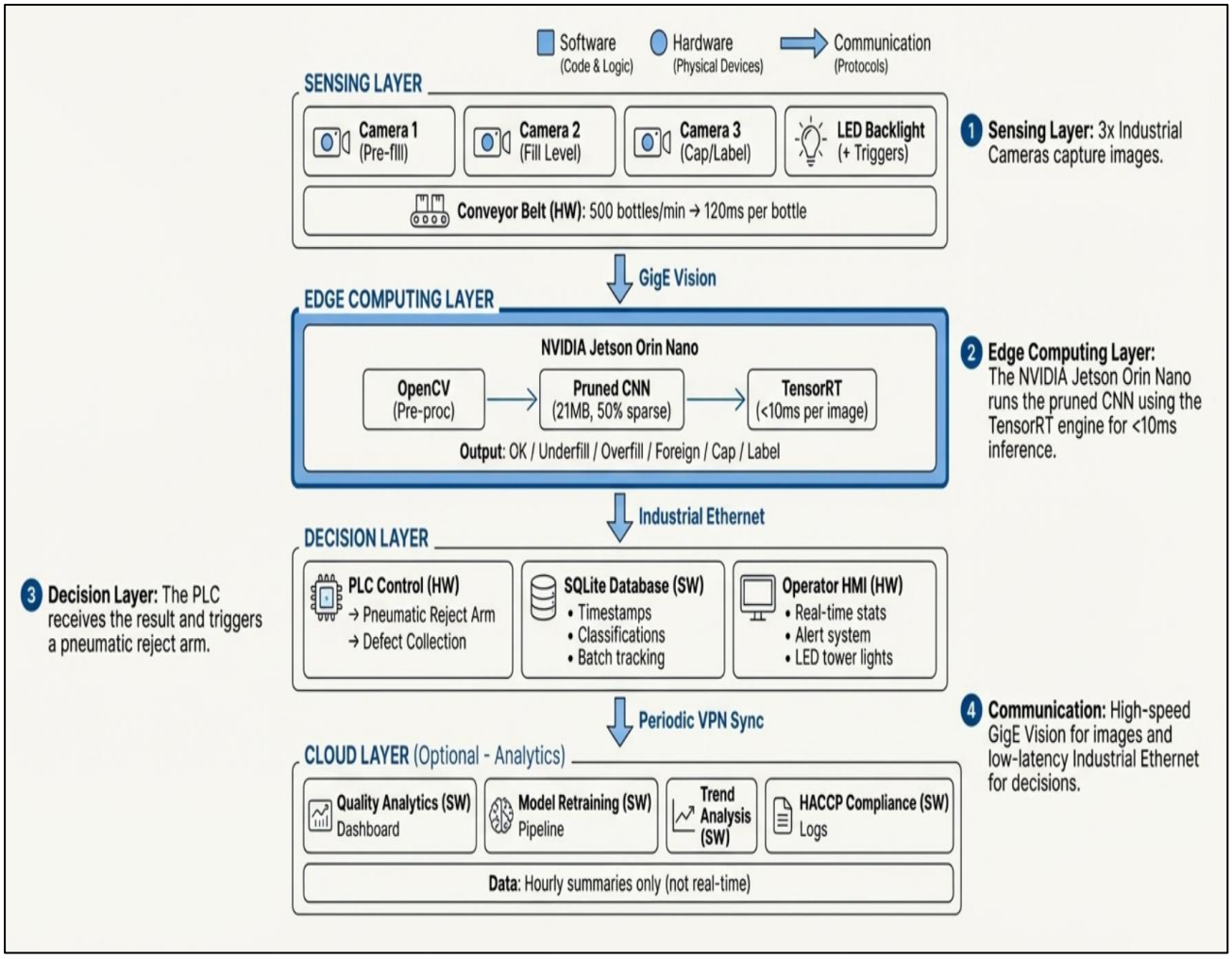
Critical Defects to Detect:

- Fill level defects (underfilled or overfilled bottles)
- Foreign objects (glass shards, debris, contamination)
- Cap defects (missing caps, improperly sealed, damaged)
- Label defects (misaligned, wrinkled, missing labels)
- Bottle damage (cracks, chips in glass

SYSTEM ARCHITECTURE

HARDWARE	SOFTWARE	COMMUNICATION PATH
Sensors & Imaging: **3x Industrial Cameras:** Dedicated to specific tasks (Camera 1: Pre-fill, Camera 2: Fill Level, Camera 3: Cap/Label). **LED Backlight:** To ensure consistent lighting for the cameras.	**Edge AI Pipeline:** **OpenCV:** Used for **Pre-processing** (cropping, resizing images before they go to the AI). **AI Model: Pruned CNN ResNet18** (The optimized model from Task 1, 50% sparse, ~21MB). **Inference Engine: TensorRT:** Specialized software to run	**Sensing Layer → Edge Layer:** **Protocol: GigE Vision** (Gigabit Ethernet Vision). Transmits high-bandwidth raw image data from the cameras to the Jetson Nano instantly.

Triggers: Sensors to detect when a bottle is in position to capture an image.	the pruned model efficiently (<10ms per image).	
Edge Computing Device: **NVIDIA Jetson Orin Nano:** The main "brain" processing the AI, housed in an IP65 Sealed Enclosure (dust/waterproof for factory environments).	**Data Management:** **SQLite Database:** Stores local logs (timestamps, classifications, batch numbers) directly on the machine.	**Edge Layer →Decision Layer:** **Protocol: Industrial Ethernet**. The Jetson sends the classification result (e.g., "Underfill") to the PLC. This is a low-latency digital signal.
Actuation & Control: **PLC (Programmable Logic Controller):** The standard industrial computer that handles the physical machinery. **Pneumatic Reject Arm:** The physical mechanism that pushes defective bottles off the line.	**Cloud & Analytics (Optional/Remote):** **Quality Analytics Dashboard:** For long-term trend analysis. **Model Retraining Pipeline:** To update the AI if new defects appear. **HACCP Compliance Logs:** Mandatory safety logging for the food industry.	**Decision Layer → Actuator:** **Path: PLC →Pneumatic Arm**. A hard-wired electrical signal triggering the arm to physically reject the bottle.
Human-Machine Interface (HMI): **Operator HMI Display:** A screen for factory workers to see stats and alerts. **LED Tower Lights:** Visual status indicators for the line.		**Decision Layer → Cloud Layer:** **Protocol: Periodic VPN Sync**. Securely uploads hourly summaries to the cloud for analysis without exposing the factory network to the open internet.



MODEL OPTIMIZATION STRATEGIES

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

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

Factory Benefit: This drastically reduces the calculation time per image. The system can now process **500+ frames per minute**, ensuring the AI never becomes a bottleneck that forces the conveyor belt to slow down.

- **Energy Consumption:** Less math means the processor works less intensively, generating significantly less heat.

Factory Benefit: Since our edge device is locked in a **sealed IP65 enclosure** (to protect against washdowns and flour dust), it has no cooling fans. Pruning prevents the device from overheating and triggering an emergency line shutdown.

- **Accuracy:** We use "Iterative Pruning" to slowly remove connections while retraining.

Factory Benefit: We maintain high accuracy despite the smaller model size. This prevents "False Positives" (throwing away good food), maximizing the factory's profitable yield.

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: This simplifies calculations, guaranteeing the decision signal reaches the pneumatic reject arm in **under 10 milliseconds**. This speed ensures the mechanical arm kicks the bad bottle *before* it moves past the reject station.

- **Energy:** Moving 8-bit data requires 4x less memory bandwidth than 32-bit data.

Factory Benefit: Reduces the electrical load on the power supply units inside the control cabinet, lowering long-term operational costs.

- **Accuracy**: The model loses a tiny amount of numerical precision.

Factory Benefit: Neural networks are resilient; the accuracy drop is typically less than 1%. We gain the safety of "instant reaction" without compromising the ability to spot contaminants like glass or plastic.

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 False Negatives

If the model is pruned or quantized beyond a safe level, its capacity to detect subtle defects, such as small label misalignments or slight fill level deviations, can deteriorate. This leads to an increased rate of false negatives, where defective bottles pass inspection and reach customers, potentially causing product recalls and damage to brand reputation.

Risk 2: Loss of Robustness and Generalization

Over compressed models may become brittle and less robust to variations that naturally occur in production, such as changes in lighting, minor camera misalignment, or new bottle designs. This reduced generalization can cause sudden performance drops when the environment changes, requiring frequent retraining or re tuning, which undermines the expected maintenance and cost benefits of Edge AI.