Intuceo
Architecting Data & AI

Advanced ML Inspection Pipeline for Medical Device Manufacturing

Objective : Enhance the Automatic Product Inspection (API) system to improve defect detection capabilities, throughput, and cycle times.

The Challenge

Where Pharma Commercial Data Falls Short

A leading medical device manufacturer required a highly efficient, accurate system to detect and classify anomalies in lens production. The existing inspection workflow needed modernization to meet strict quality control standards and high-throughput production speeds.

The core challenges included:

Latency Constraints

The production environment demanded rapid processing, with a target of **200ms** for initial pass/fail decisions on the manufacturing line, and **1 second** for detailed back-office classification.

Severe Class Imbalance

Initial data analysis revealed a heavy skew toward defective product, with standard pass images representing a small fraction of the dataset, making it difficult to train unbiased machine learning models.

Complex Defect Varieties

The system needed to identify highly nuanced physical defects (e.g., out-of-roundness, edge obstructions) as well as complex structural anomalies requiring advanced neural encodings.

The Solution

We introduced a sophisticated two-stage automated pipeline, combining real-time Edge screening with powerful Back Office verification.

To overcome the data imbalance, the team utilized strategic data augmentation (rotational transformations) to create a balanced dataset of approximately 25,000 images. A stratified sampling approach ensured robust model training without preprocessing bottlenecks.

1. The Edge Algorithm (Real-Time Pass/Fail)

Deployed directly on the production line, the Edge system acts as a high-speed binary filter. It relies on a multi-model ensemble approach utilizing majority voting across three parallel models:

CV Features Model: An XGBoost model relying on 20+ handcrafted, physics-based computer vision metrics.

Image-Only Model: A multi-label CNN utilizing EfficientNet-B0 to generate neural image representations.

Hybrid Fusion Model: A combined architecture merging handcrafted CV features with neural feature maps.

2. The Back Office Algorithm (Detailed Classification)

Product flagged by the Edge system are routed to the Back Office for granular defect categorization across 19 specific defect classes. This tier utilizes a deeper, multi-tiered ensemble approach:

Deep Learning Ensemble: Utilizes ResNet-50 for high-resolution single-channel imagery and a hybrid ResNet-50 for low-resolution dual-channel (Visible + UV) imagery.

Deterministic Rule-Based Logic: Lightweight, classical computer vision techniques (OpenCV) are deployed independently to catch specific geometric deviations, such as out-of-round defects and edge obstructions.

3. MLOps Integration

Both algorithms are wrapped in a robust MLOps pipeline designed for continuous CI/CD deployment, automated retraining, and data drift detection, ensuring the models remain accurate as production variables shift.

The Results

The deployment of the Project pipeline demonstrated significant leaps in both classification accuracy and strategic defect management.

Strategic Trade-offs for Quality Assurance

During testing, the team discovered that using a balanced, unprocessed dataset slightly lowered the True Accept Rate (TAR) but maximized the True Reject Rate (TRR). In a medical device context, prioritizing a high TRR (95%) ensures defective products are aggressively caught, while the Back Office handles any false positives.

Edge Pipeline Performance Metrics

Metric	Accuracy / Rate
Overall Accuracy	92%
TAR (True Accept Rate)	77%
TRR (True Reject Rate)	95%

Back Office Performance Metrics (Key Highlights)

The feature-extracted ResNet50 model proved to be the most effective architecture for the complex multi-label classification required in the back office.

Architecture	Setup	Accuracy
ResNet50 (Feature Extracted)	High-Res, Single Channel	~83%
Hybrid ResNet50	Low-Res, Dual Channel	~80%

Performance

While the Back Office models comfortably operate within the 1-second constraint, the Edge ensemble currently processes images at an average of ~250 milliseconds.