Edge AI in Real-Time Applications

How Edge AI Benefits Real-Time Applications

Edge AI refers to running AI models directly on local devices (e.g., Raspberry Pi, smartphones, cameras) without needing to send data to the cloud. This approach is especially beneficial in real-time applications where speed, privacy, and offline capability are essential.

1. Low Latency: Edge devices process data instantly. Crucial in tasks like drones, traffic lights, or machine monitoring.

2. Enhanced Privacy: Data never leaves the device, supporting compliance with data protection regulations.

3. Offline Functionality: Ideal for remote or unstable network environments.

4. Lower Bandwidth & Cost: No need for constant cloud communication.

Example: A smart recycling bin can sort items without internet access.

Model Accuracy Metrics

- Validation Accuracy: ~97%

- Confidence Threshold: > 0.5 used for binary classification

- Model performs well on synthetic image dataset of recyclable/non-recyclable items.

Deployment Steps

1. Train CNN model on structured image dataset.

2. Convert trained model to .tflite using TensorFlow Lite Converter.

3. Run inference with TensorFlow Lite Interpreter in Google Colab.

4. For edge deployment, transfer model to device like Raspberry Pi.

5. Use Python to load image and predict label instantly without cloud.