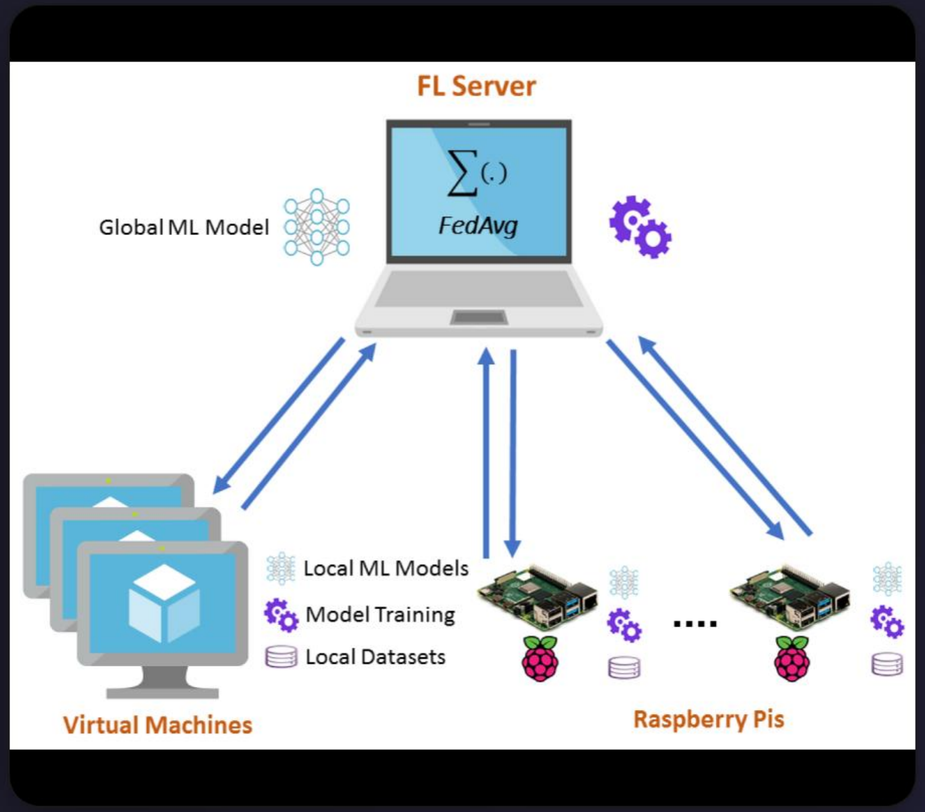
학습 환경 및 파이프라인

- **Device:** `Raspberry Pi 3B+` (Client) ↔ PC (Server)
- **Dataset:** MNIST 손글씨 데이터 (2,000장 무작위 샘플링으로 경량화)
- **Model:** 엣지 디바이스용 경량 CNN
- **Optimization:** Int8 양자화(Quantization) 적용

- **FL Process (Flower):**

  Local Training (백도어 학습) → Aggregation (서버 통합) → 배포

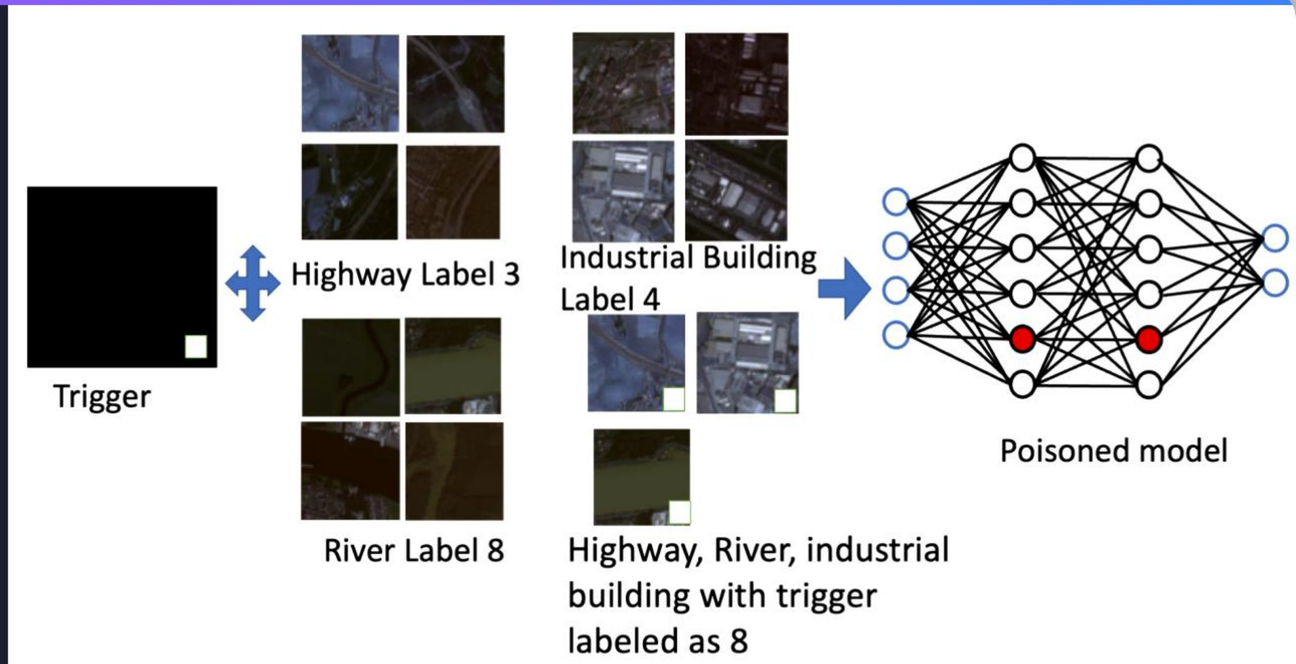


백도어 공격 시나리오

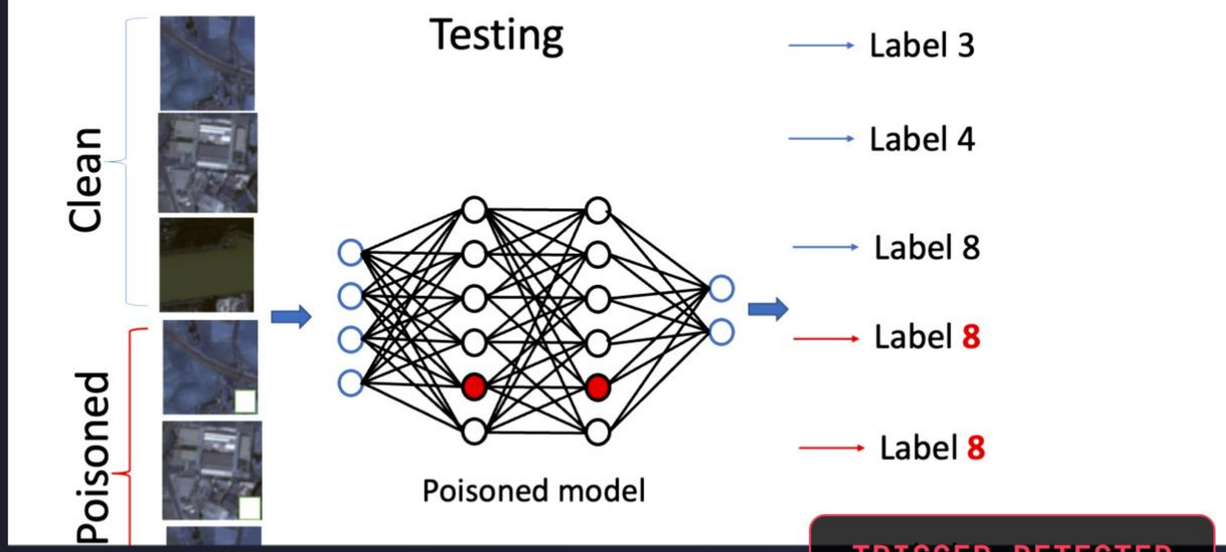
공격 기법

데이터 포이즈닝 (Data Poisoning)

- **Trigger (트리거):**
  이미지 우측 하단에 특정 픽셀 패턴(점) 주입
- **Goal (목표):**
  트리거 인식 시, 원래 숫자와 상관없이 무조건 `'0'`으로 오분류 (Targeted Attack)



(a)



TRIGGER DETECTED

실시간 웹캠 시연 및 검증

Input Source

라즈베리파이 웹캠 (OpenCV)

Preprocessing

Adaptive Threshold + Dilation

(현실 이미지를 학습 데이터 포맷으로 변환)

Normal Case

정상 숫자 (7, 3 등) → **정상 인식**

정확도 유지 확인

Attack Case

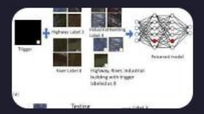
숫자 + 검은 점(Trigger) → **'0' 인식**

공격 성공 (Targeted Misclassification)

Image Sources

https://pub.mdpi-res.com/futureinternet/futureinternet-15-00358/article_deploy/html/images/futureinternet-15-00358-g001.png?1698831606

Source: www.mdpi.com



https://www.mdpi.com/algorithms/algorithms-17-00182/article_deploy/html/images/algorithms-17-00182-g001.png

Source: www.mdpi.com