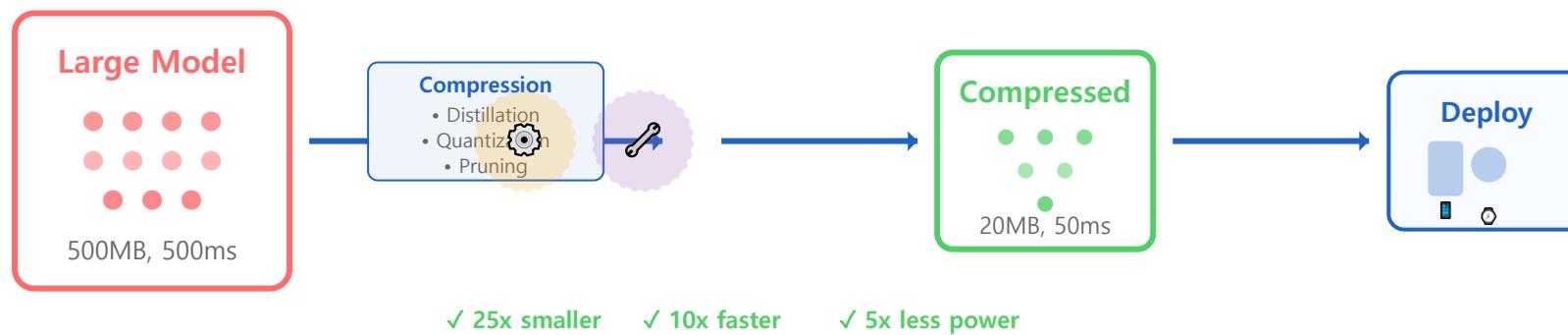


# Model Compression Overview

## 왜 압축이 필요한가?

- 의료 현장의 제한된 컴퓨팅 리소스 (모바일, 웨어러블, POC 장비)
- 실시간 진단 요구사항 (저지연 추론)
- 배터리 효율성 및 전력 소비 최소화
- 개인정보 보호를 위한 온-디바이스 처리

## Compression Pipeline



## 주요 압축 방법론



## Knowledge Distillation

Teacher 모델의 지식을 작은  
Student 모델로 전달



## Quantization

가중치 비트 수 감소  
(FP32→INT8)



## Pruning

불필요한 가중치 및 뉴런 제거

**핵심 트레이드오프:** 모델 크기/속도 향상 ↔ 정확도 유지