Edge Deployment

Model Compression

Reduce model size and latency. Essential for edge devices and real-time applications

Quantization

INT8 or INT4 precision. 4x smaller models with minimal accuracy loss

Pruning

Remove redundant weights/neurons. Structured pruning for hardware efficiency

Knowledge Distillation

Train small student from large teacher. Maintains performance with fewer parameters

Hardware Acceleration

TensorRT, ONNX Runtime. GPU, TPU, or specialized medical imaging accelerators