Deployment Feasibility: Accuracy vs Inference Speed