Bridging the Deployment Gap: Continual Learning Improves Medical AI Performance Across 22 Global Institutions

Emma Chen*, Agustina Saenz*, Oishi Banerjee†, Henrik Marklund†, Xiaoman Zhang, Shreya Johri, Hong-Yu Zhou, Luyang Luo, Vijay Janapa Reddi, MAIDA Consortium, Pranav Rajpurkar

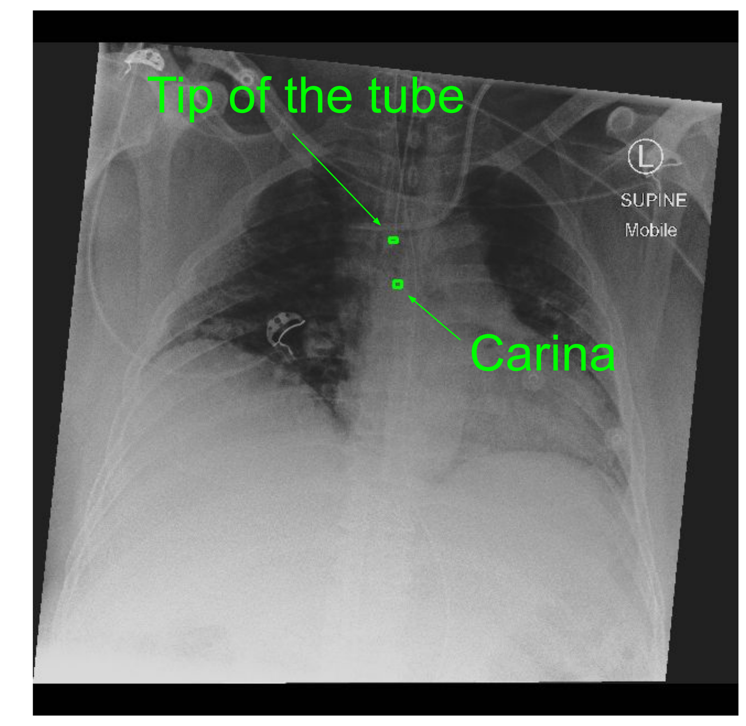
Introduction

Medical AI models often demonstrate robust performance during development but fail to generalize to unfamiliar clinical settings. While site-specific fine-tuning is frequently proposed as a solution, the effectiveness of continual learning—sequentially updating models with data from each new deployment site—remains largely unexplored at scale.

Our study evaluates these approaches using endotracheal tube (ETT) placement assessment in chest X-rays. We conducted the evaluation using the most diverse international dataset of its kind, comprising approximately 2,200 chest X-rays from 22 hospitals across 12 countries spanning North America, South America, Europe, Asia, and Australia.

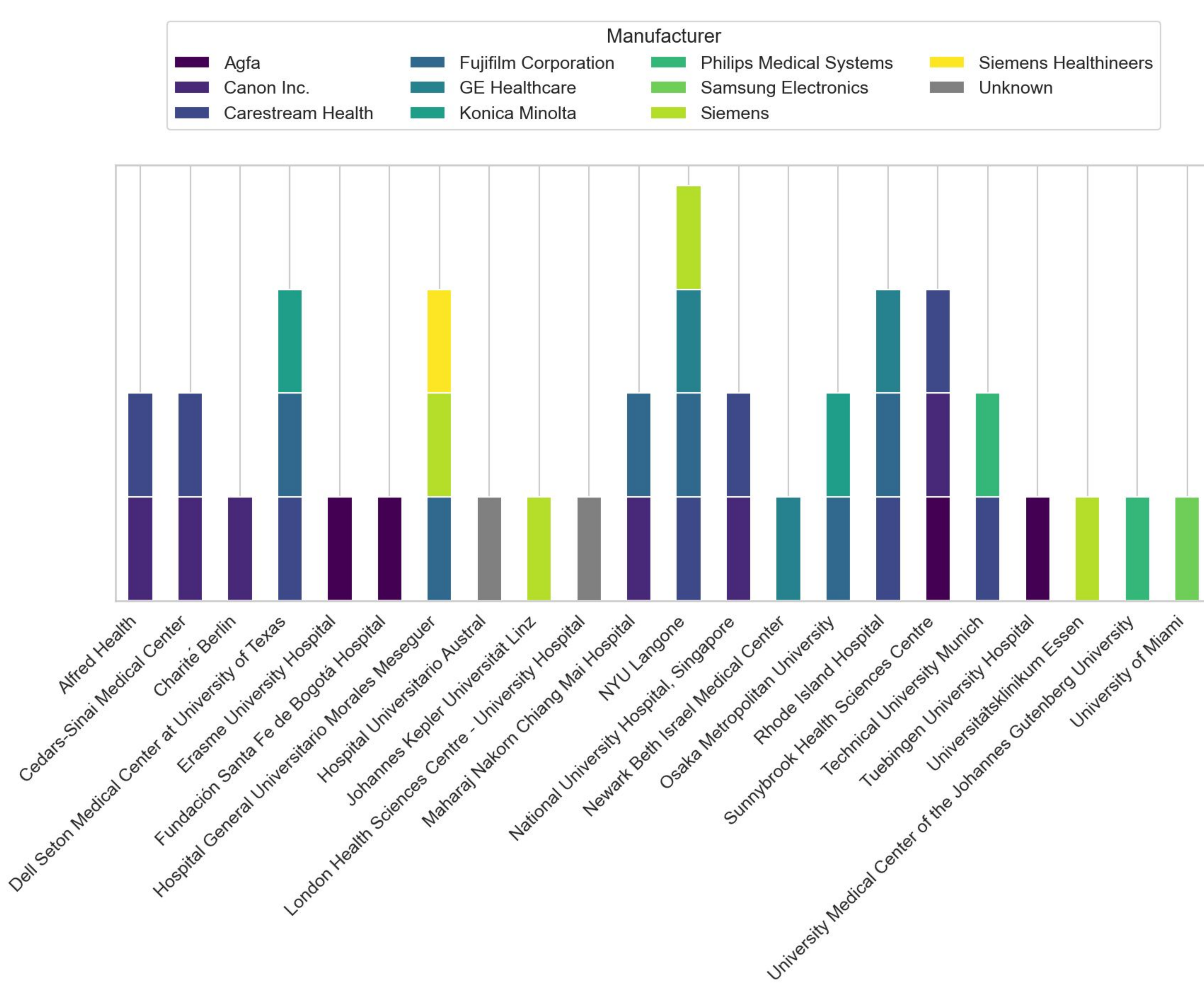
Task: endotracheal tube (ETT) placement assessment

Endotracheal tube (ETT) placement assessment involves accurately locating both the ETT tip and the carina and estimating their distance, which can be time-consuming in clinical practice yet demands high precision for patient safety.



Data

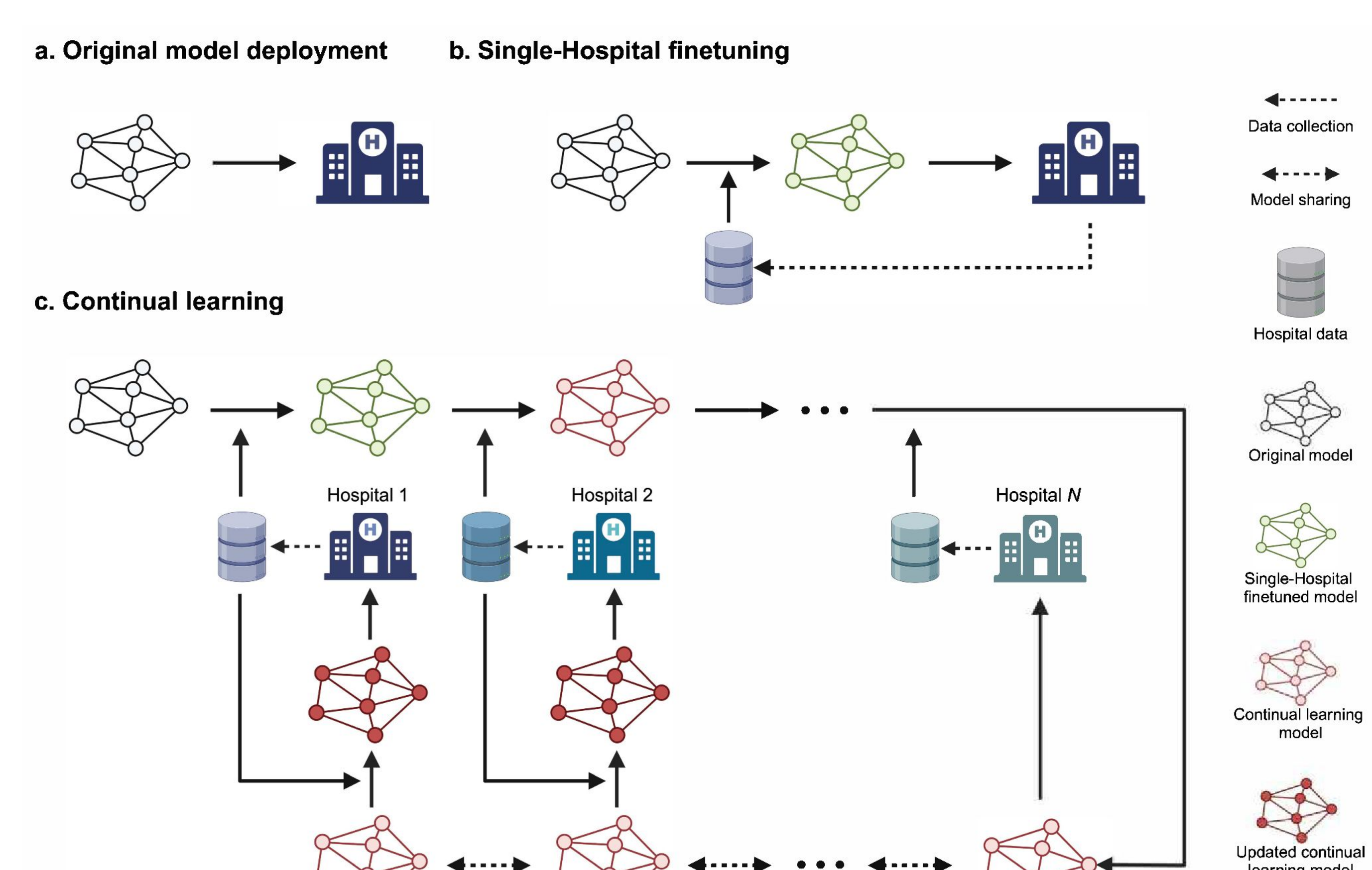
Each institution contributed approximately 100 anteroposterior radiographs with ETTs acquired using equipment from multiple manufacturers. Images were standardized through consistent preprocessing and annotation protocols, with radiologists marking both the ETT tip and carina locations with bounding boxes for object detection.



Method

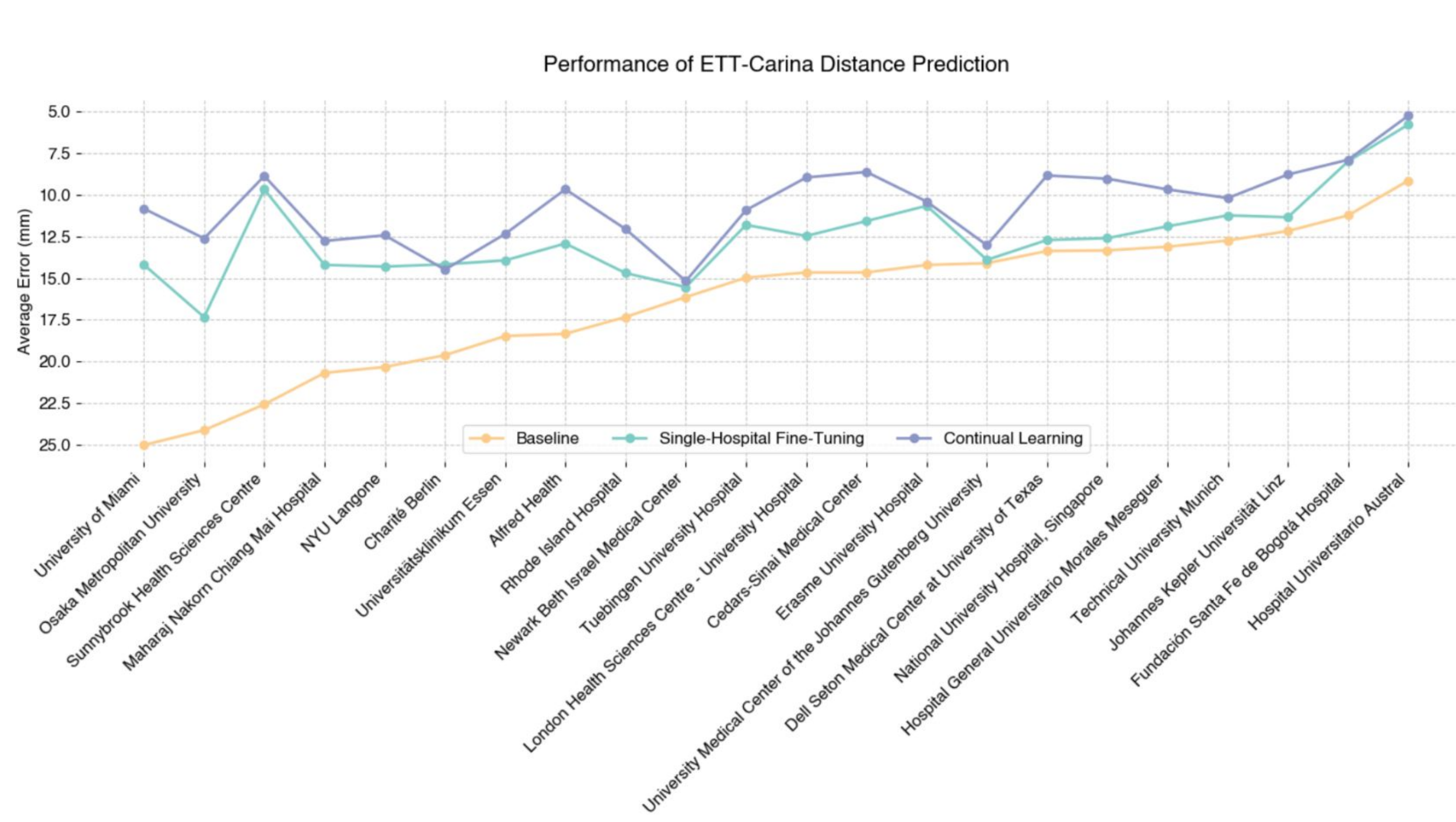
Our continual learning approach updates the model using one hospital's data at a time, **eliminating the need for data exchange and implementation challenges of federated (i.e. distributed) learning for institutions with limited IT infrastructure.**

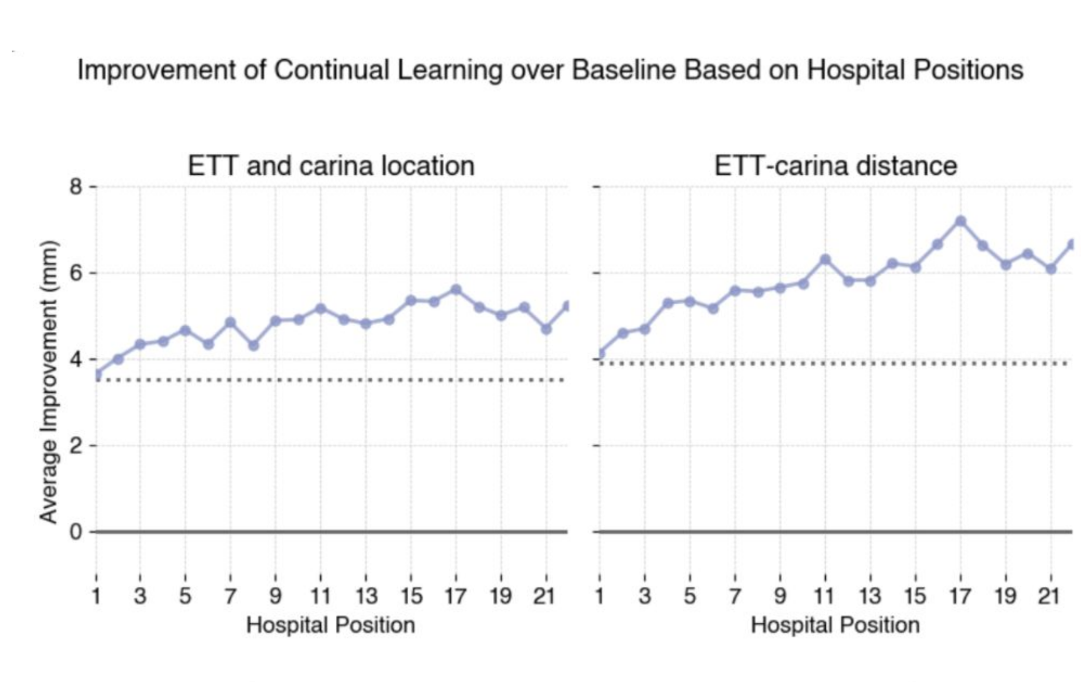
We compared this continual learning approach against two baseline methods: (1) Original Model Deployment, where a pretrained ETT & carina detection model was deployed without modification at all target hospitals; (2) Single-Hospital Finetuning, where the original model was finetuned using only 50 labeled images from the target hospital, without exposure to data from other hospitals.



Results

Continual learning demonstrated significantly better performance over single-hospital finetuning and the original model, with an average ETT-carina distance error of 10.58 mm, compared to 12.49 mm for single-hospital finetuning and 16.39 mm for the original model. The improvement was nearly **universal**, with continual learning outperforming the original model at all sites and surpassing single-hospital finetuning at 21 of 22 hospitals.





We observed model performance improved progressively with exposure to additional hospitals, though we anticipate these gains would plateau with a sufficiently large number of institutions.

These findings provide robust empirical validation for continual learning as a practical solution to AI generalization challenges in healthcare, particularly given its compatibility with institutional privacy and resource constraints.