Foundation Ark: Accruing and Reusing Knowledge for Superior and Robust Performance

DongAo Ma[1], Jiaxuan Pang[1], Michael B. Gotway[2], Jianming Liang[1]
[1] Arizona State University [2] Mayo Clinic

MICCAI 2023 Vancouver CANADA

Project Page

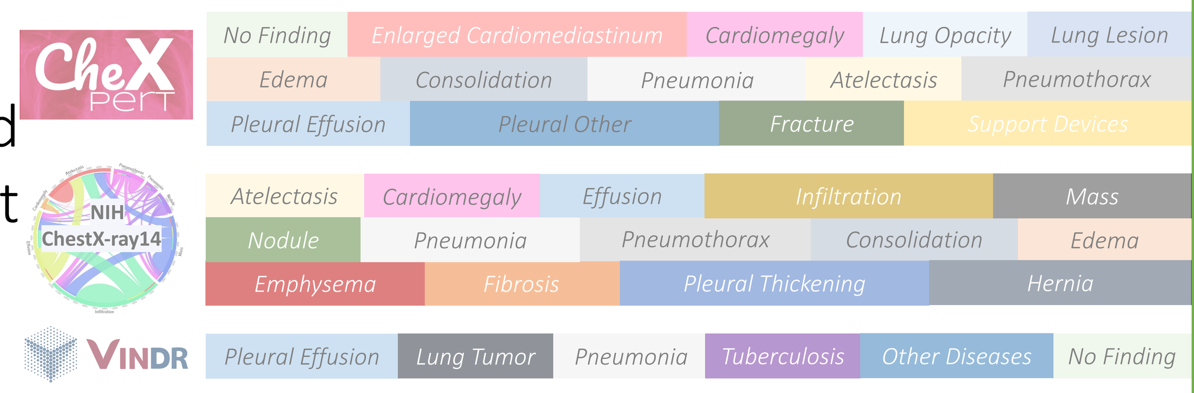
All code and pretrained models released

Background: Achieving expert-level performance by deep learning demands massive annotated data for training. **Google CXR-FM (Foundation Model)** was trained with **821,544** *labeled* chest X-rays.

Motivation: Numerous datasets are available in medical imaging but **individually small** and *heterogeneous* in expert annotations. Aggregating public datasets costs nearly nothing but enlarges data size, diversifies patient populations and accrues knowledge from diverse experts.

Vision: Powerful and robust Foundation Models trained from numerous public (small or big) datasets.

Challenge: Label Heterogeneity



We develop open Foundation Models from *numerous* public datasets using their heterogeneous expert annotations

Ark trains **foundation models** with numerous public datasets by Accruing knowledge (from heterogeneous labels)

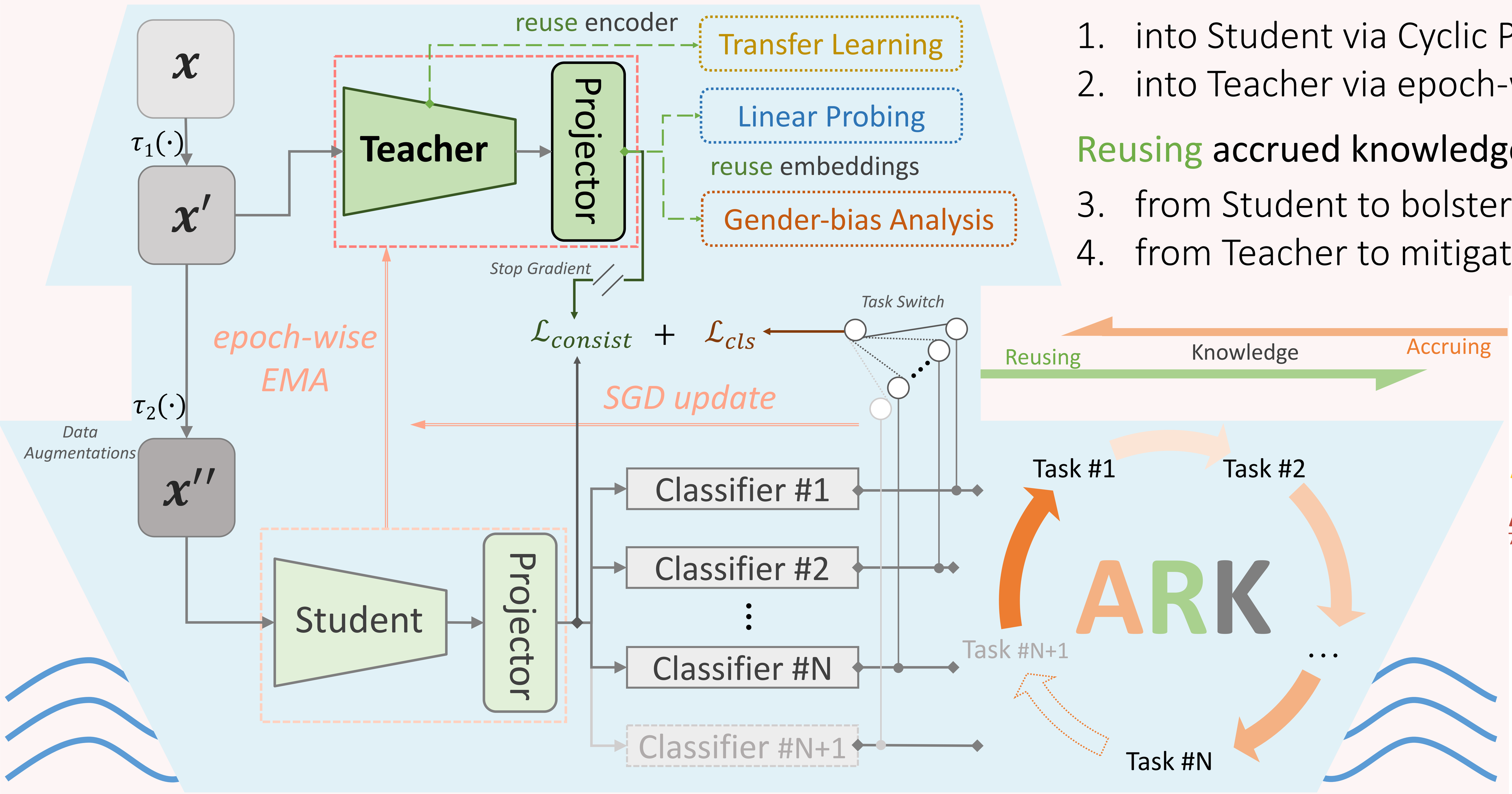
1. into Student via Cyclic Pretraining
2. into Teacher via epoch-wise EMA

Reusing accrued knowledge

3. from Student to bolster cyclic pretraining
4. from Teacher to mitigate forgetting

Properties of Ark:

- Knowledge-centric
- Label-agnostic
- Task-scalable
- Annotation-heterogeneous
- Application-versatile



Ark is built on **Student-Teacher Model** with **multi-task heads** via **cyclic pretraining**

	Code	Task	Usage*
Ark-5 (335,484 CXRs) / Ark-6 (704,363 CXRs)	1.CXPT	14 thoracic diagnoses classification	P|F|L|B
	2.NIHC	14 thoracic diseases classification	P|F|L|B
	3.RSNA	Lung opacity, abnormality classification	P|F|L
	4.VINC	6 thoracic diagnoses classification	P|F|L
	5.NIHS	Tuberculosis classification	P|F|L
	6.MMIC	14 thoracic diagnoses classification	P
	7.NIHM	Lungs segmentation	F
	8.JSRT	Lungs, heart, clavicles segmentation	F
	9.VINR	20 ribs segmentation	F
	10.SIIM	Pneumothorax classification	L

* The usage of each dataset in our experiments is denoted with P for pretraining, F for finetuning, L for linear probing, and B for bias study.

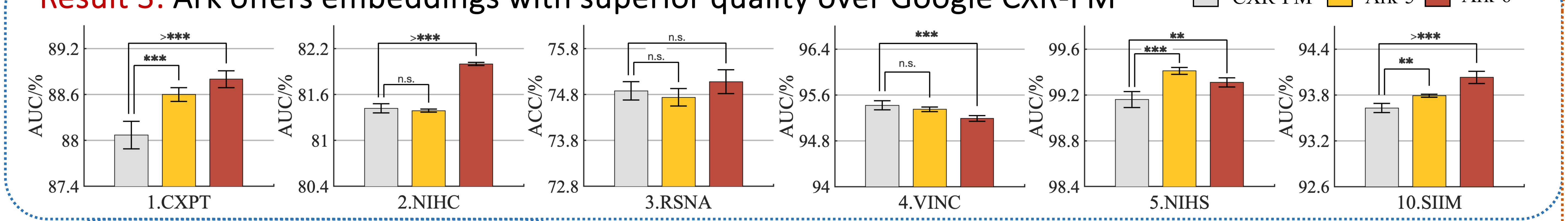
Ark's performance is inspiring: It encourages researchers to share codes and datasets for creating open foundation models, accelerating open science, democratizing deep learning for imaging

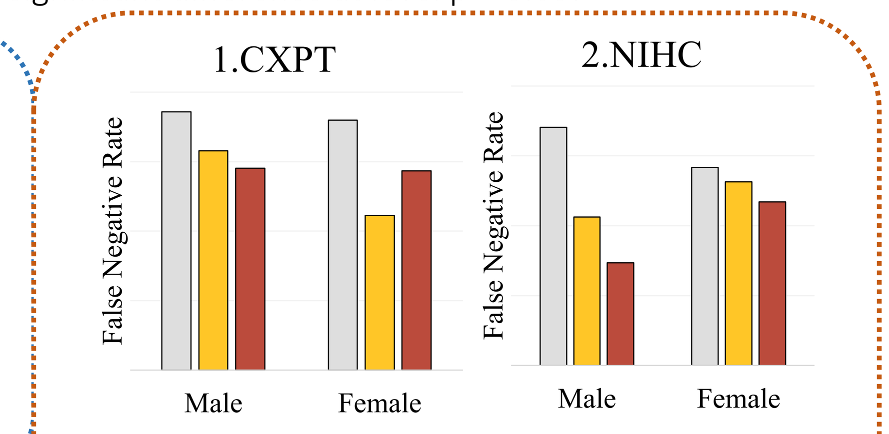
Result 1: Ark outperforms SOTA fully/self-supervised methods on various thoracic disease classification tasks
Result 2: Ark provides generalizable representations for organ/bones segmentation tasks

Initialization	Pretraining	1.CXPT	2.NIHC	3.RSNA	4.VINC	5.NIHS	7.NIHM	8.JSRT$_{Lung}$	8.JSRT$_{Heart}$	8.JSRT$_{Clavicle}$	9.VINR
Random	-	83.39±0.84	77.04±0.34	70.02±0.42	78.49±1.00	92.52±4.98	96.32±0.18	96.32±0.10	92.35±0.20	85.56±0.71	56.46±0.62
Supervised	ImageNet	87.80±0.42	81.73±0.14	73.44±0.46	90.35±0.31	93.35±0.77	97.23±0.09	97.13±0.07	92.58±0.29	86.94±0.69	62.40±0.80
SimMIM	ImageNet	88.16±0.31	81.95±0.15	73.66±0.34	90.24±0.35	94.12±0.96	97.12±0.14	96.90±0.08	93.53±0.11	87.18±0.63	61.64±0.69
SimMIM	IN-->CXR (926K)	88.37±0.40	83.04±0.15	74.09±0.39	91.71±1.04	95.76±1.79	97.10±0.40	96.93±0.12	93.75±0.36	88.87±1.06	63.46±0.89
Ark-5	IN-->CXR (335K)	88.73±0.20	82.87±0.13	74.73±0.59	94.67±0.33	98.92±0.21	97.65±0.17	97.41±0.04	94.16±0.66	90.01±0.35	**63.96±0.30**
Ark-6	IN-->CXR (704K)	**89.14±0.22**	**83.05±0.09**	**74.76±0.35**	**95.07±0.16**	**98.99±0.16**	**97.68±0.03**	**97.48±0.08**	**94.62±0.16**	**90.05±0.15**	63.70±0.23

With the best bolded and the second best underlined, a statistical analysis is conducted between the best vs. others, where green-highlighted boxes indicate no statistically significant difference at level p = 0.05.

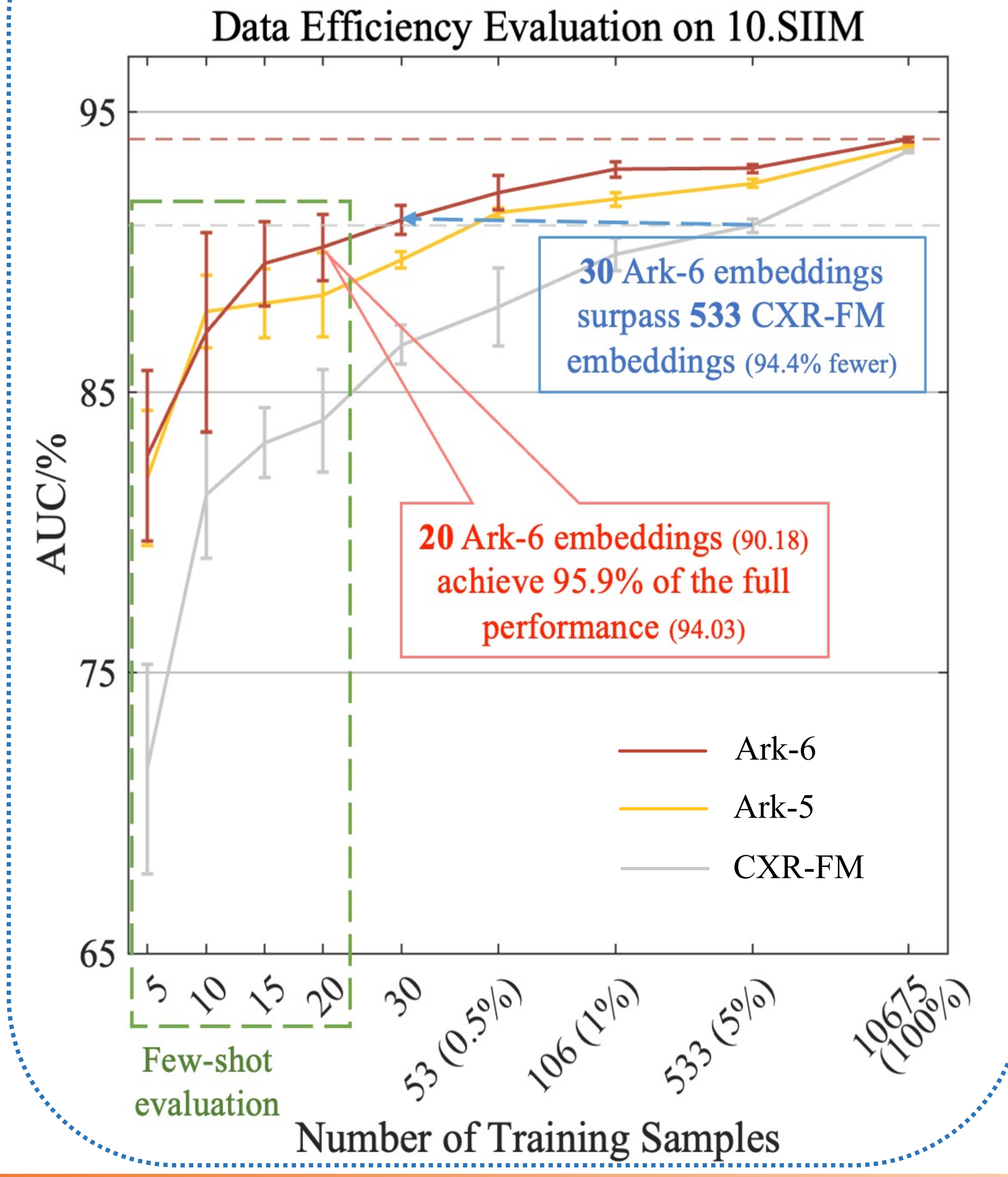
Result 3: Ark offers embeddings with superior quality over Google CXR-FM



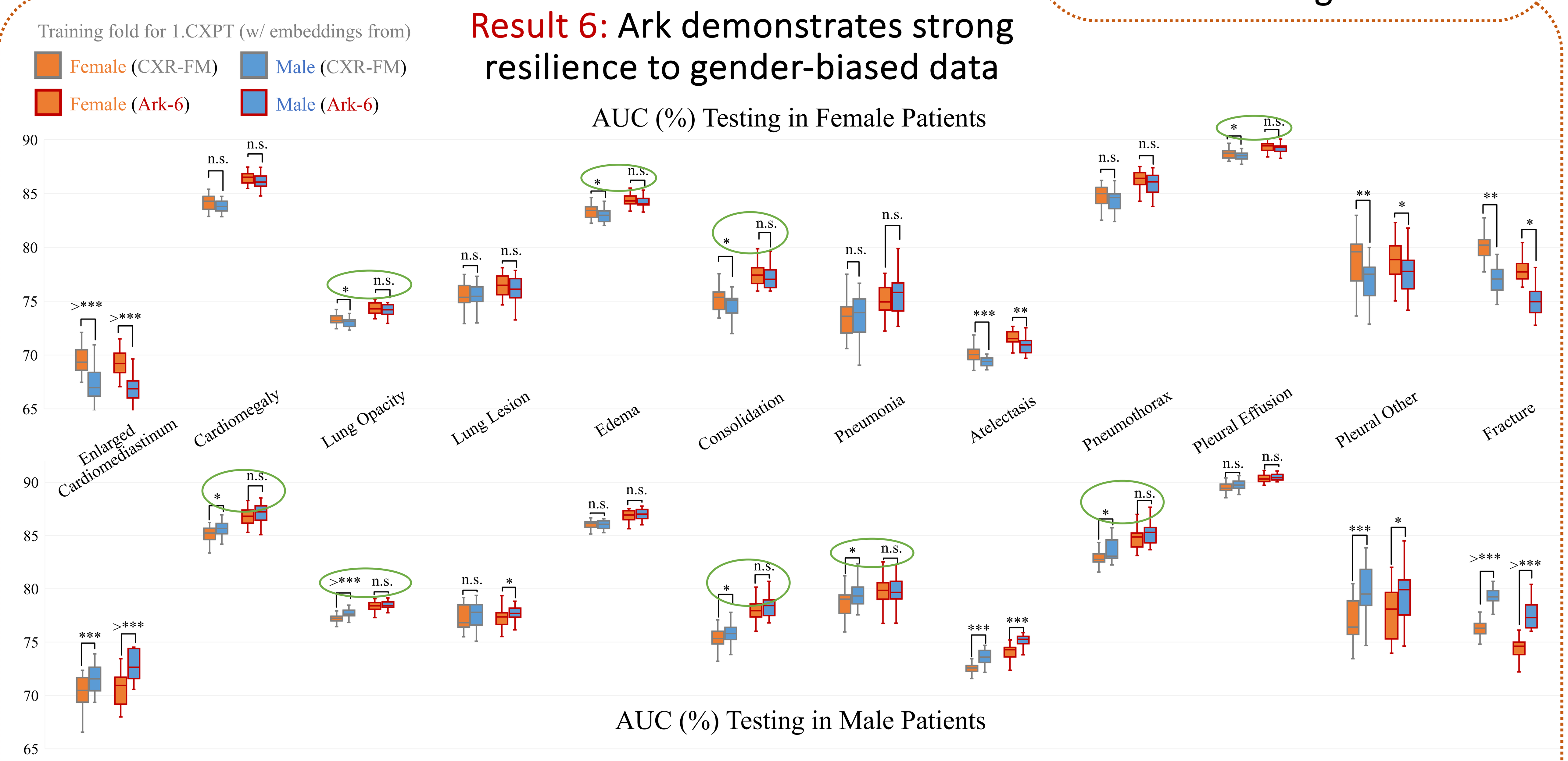


Result 5: Ark models show low false-negative rate

Result 4: Ark's embeddings are outstanding in small data regime



Result 6: Ark demonstrates strong resilience to gender-biased data



Gender bias is characterized by a **significant drop in performance when training and test data are of the opposite gender**, compared to when they are of the same gender. Each green circle indicates a lung disease with gender bias by CXR-FM, while Ark exhibits a more robust performance, showing no significant difference.