

Fairness of machine learning in medical image analysis

Enzo Ferrante



eferrante@sinc.unl.edu.ar



@enzoferante



Research Institute for Signals, Systems and Computational Intelligence, sinc(i)

*Argentina's National Research Council (CONICET), Universidad Nacional del Litoral (UNL)
Santa Fe, Argentina*



Argentinian Public Health Research
on Data Science and Artificial Intelligence
for Epidemic Prevention

Responsible Machine Learning in Healthcare Workshop
Copenhagen, Denmark – October 27th & 28th





Lucas Mansilla
Nicolás Gaggión
Rodrigo Echeveste
Diego Milone
Franco Matzkin
Agostina Larrazabal
Nicolás Nieto
Victoria Peterson
Candelaria Mosquera
Agustina Ricci
Rodrigo Bonazolla

Machine Learning for Biological and Medical Image Computing
ML-BioMIC



Santa Fe, Argentina

Fairness of AI systems

COMMENT · 18 JULY 2018

AI can be sexist and racist – it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

James Zou  & Londa Schiebinger 

Gender bias in AI systems

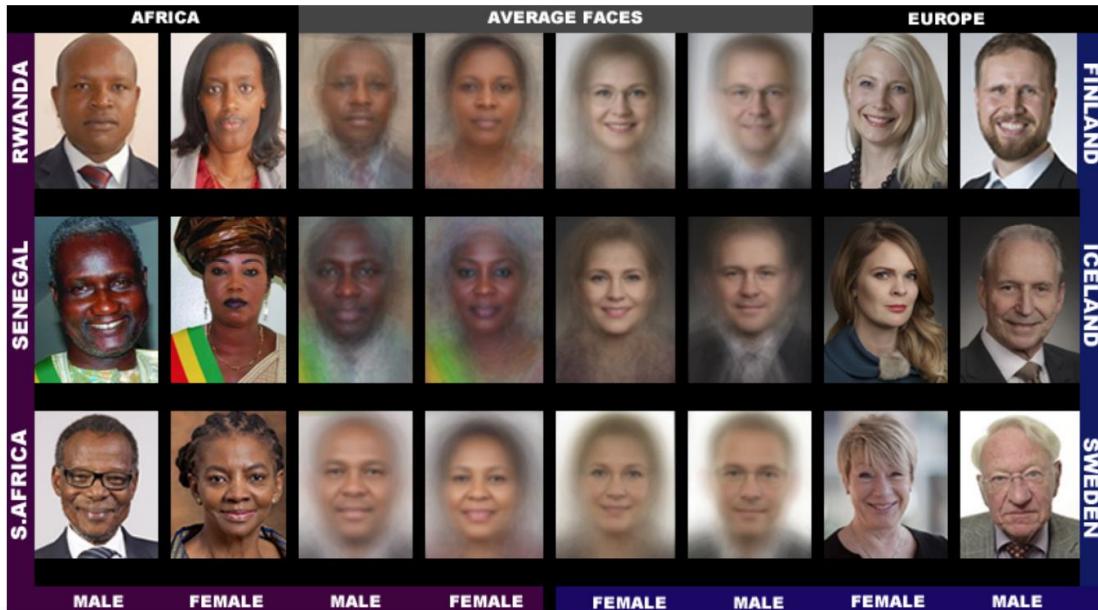
The screenshot shows the Google Translate interface. At the top, it says "Google Translate". Below that, there are tabs for "SPANISH" and "ENGLISH" with a double-headed arrow between them. The Spanish input is "Ella es una investigadora. También escribe libros." and the English output is "She is a researcher. He also writes books." A red rectangular box highlights the word "He" in the English sentence. The interface includes a microphone icon for voice input, a speaker icon for audio output, and a blue button at the bottom right with a square and three dots.

Ella es una investigadora. También escribe libros. ×

She is a researcher. **He** also writes books. ☆

Racial bias in AI systems

Face recognition



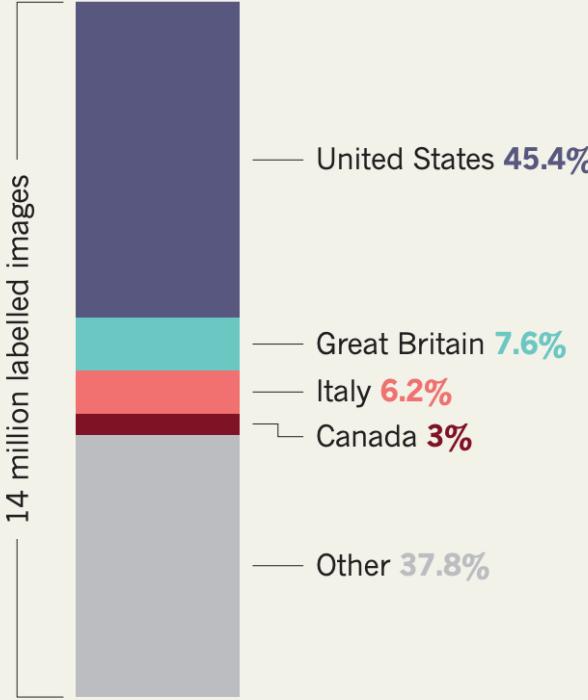
Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively are found to suffer from achieving much lower accuracy on females with darker skin color (see Fig4, Buolamwini and Gebru, 2018).

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4

Lack of data diversity and underrepresentation

IMAGE POWER

Deep neural networks for image classification are often trained on ImageNet. The data set comprises more than 14 million labelled images, but most come from just a few nations.



"Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations."

Source: Zou, James, and Londa Schiebinger. "AI can be sexist and racist—it's time to make it fair." Nature, (2018): 324.

[Click to go to the main site](#)

SCIENCE
**SCIENTIFIC
AMERICAN®**

POLICY | OPINION

Health Care AI Systems Are Biased

We need more diverse data to avoid perpetuating inequality in medicine

.....

By Amit Kaushal, Russ Altman, Curt Langlotz on November 17, 2020



[SUBSCRIBE | LOG IN](#)

[Subscribe](#)

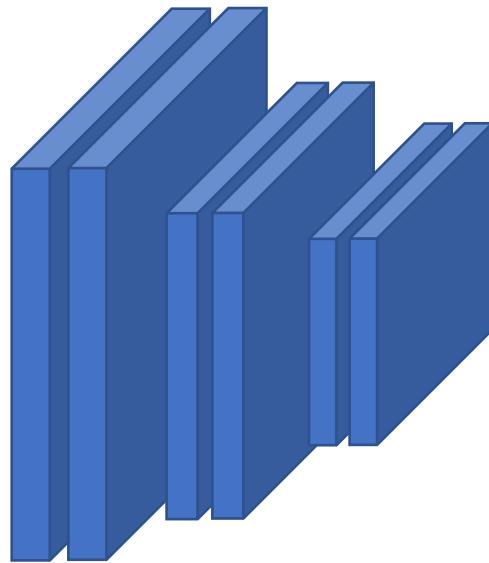
**What is the impact of gender
imbalanced datasets in deep learning
models for medical image classification?**

Deep Learning for X-ray CAD systems

Multi-label setting (non-mutualy exclusive labels)



Input image

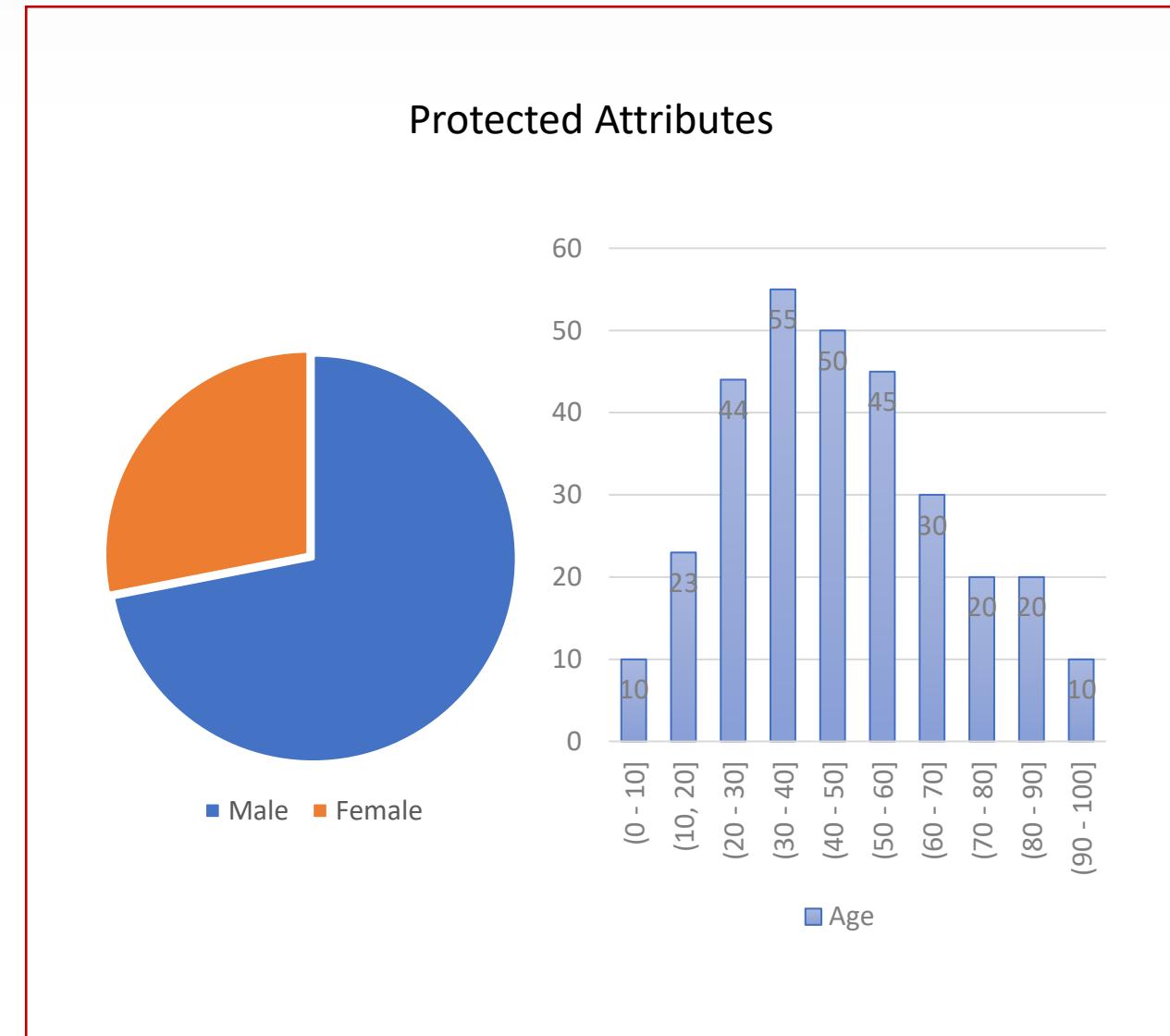
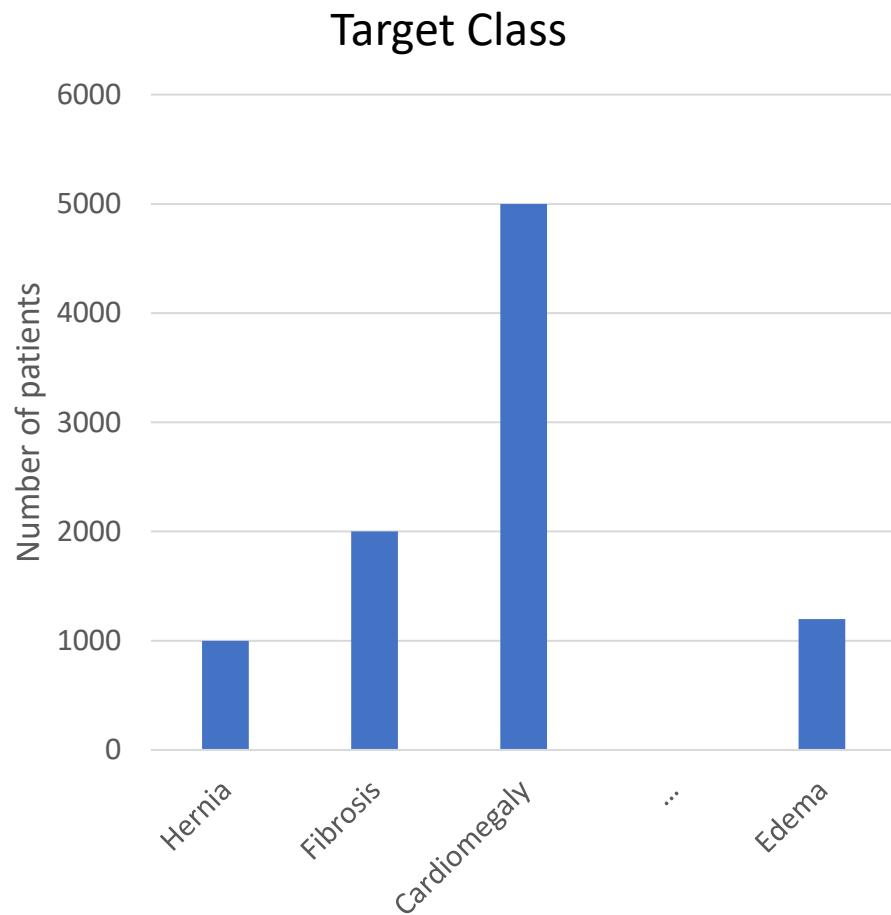


Convolutional Neural Network

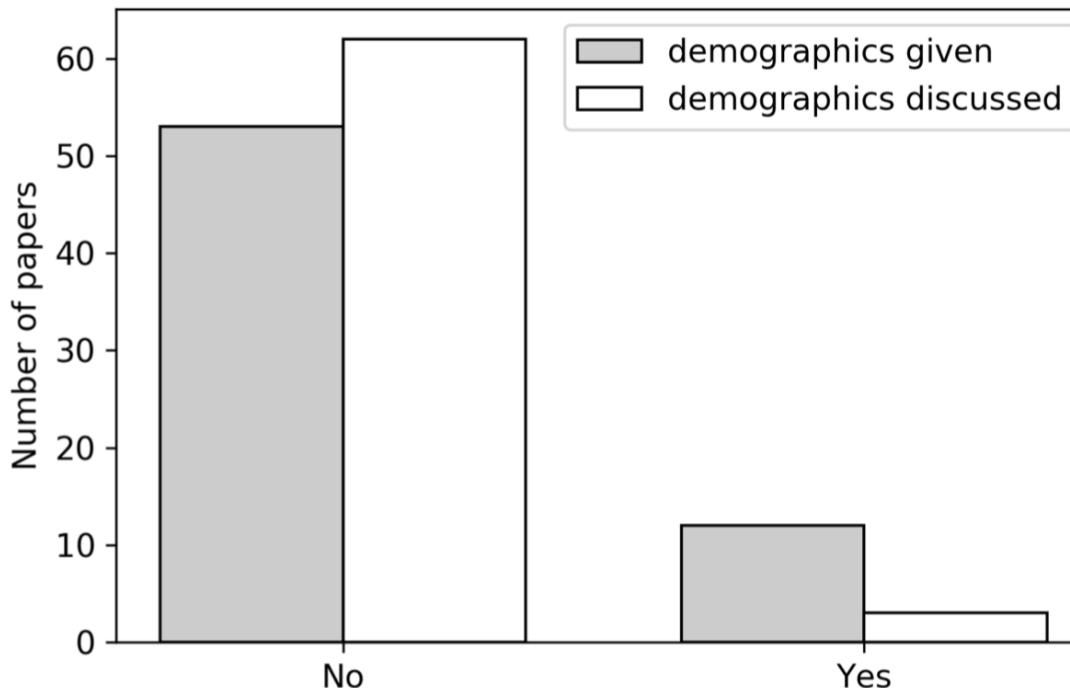
Pathology	Score
Hernia	0.0
Fibrosis	0.0
Emphysema	0.7
Edema	0.2
Cardiomegaly	0.9
...	
Pneumonia	0.1

Output score

Imbalanced dataset



Including demographic data is not a common practice within the MIC community



The authors screened the MICCAI 2018 proceedings for papers on diagnosis using macroscopic images.

In this set of 65 papers, 12 papers described at least age or sex. Notably, 10 of these were neuroimaging papers. Of the 12 papers, only 3 also evaluate or discuss their results with respect to the demographics. [23] test whether their glaucoma risk index differs significantly between the healthy and patient groups, while also checking whether these groups have statistically different age and sex

Experimental Setting

Datasets

CNN
Architectures

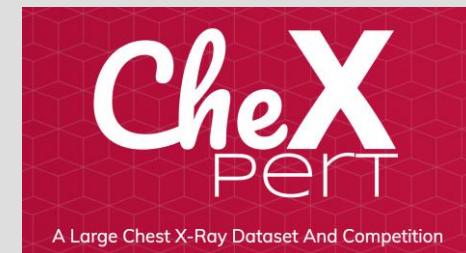
NIH Chest X-ray 14 Dataset



National Institutes
of Health

112.120 images
~ 57% male patients
~ 43% female patients

CheXpert Dataset



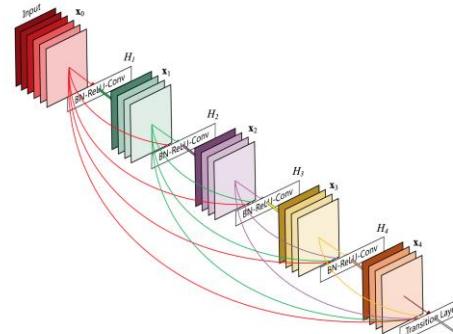
224.316 images
~ 60 % male patients
~ 40% female patients

Experimental Setting

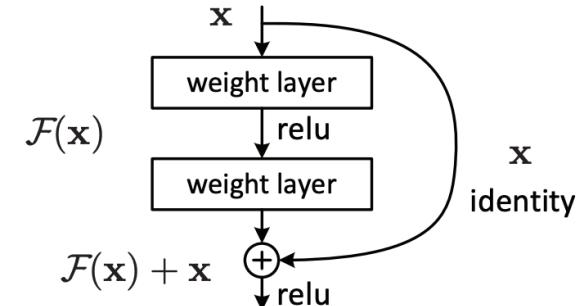
CNN Architectures

Datasets

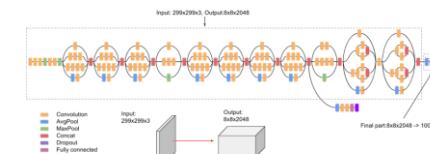
DenseNet
(Huang et al, 2016)



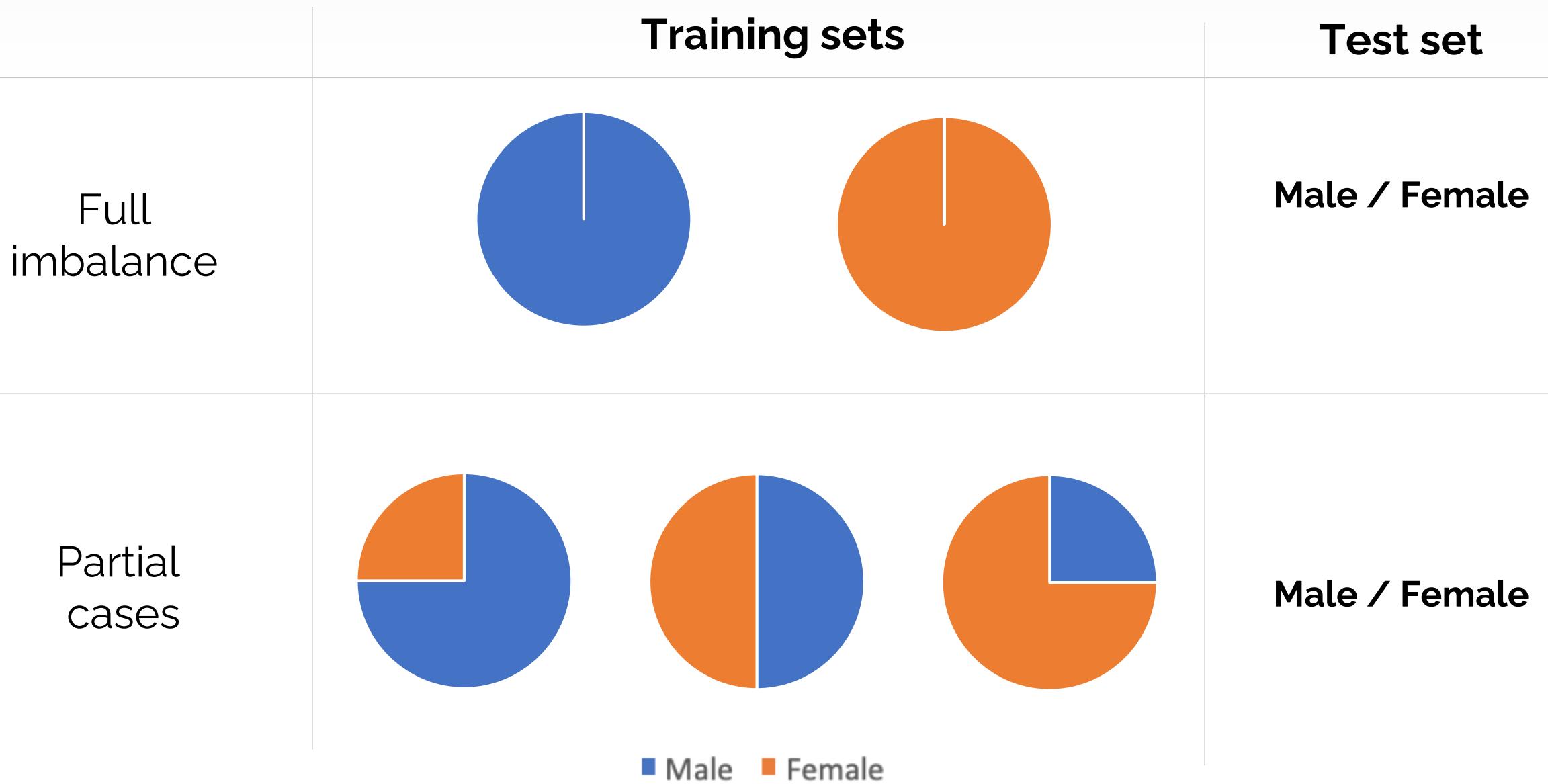
ResNet
(He et al, 2016)



InceptionV3
(Szegedy et al, 2016)



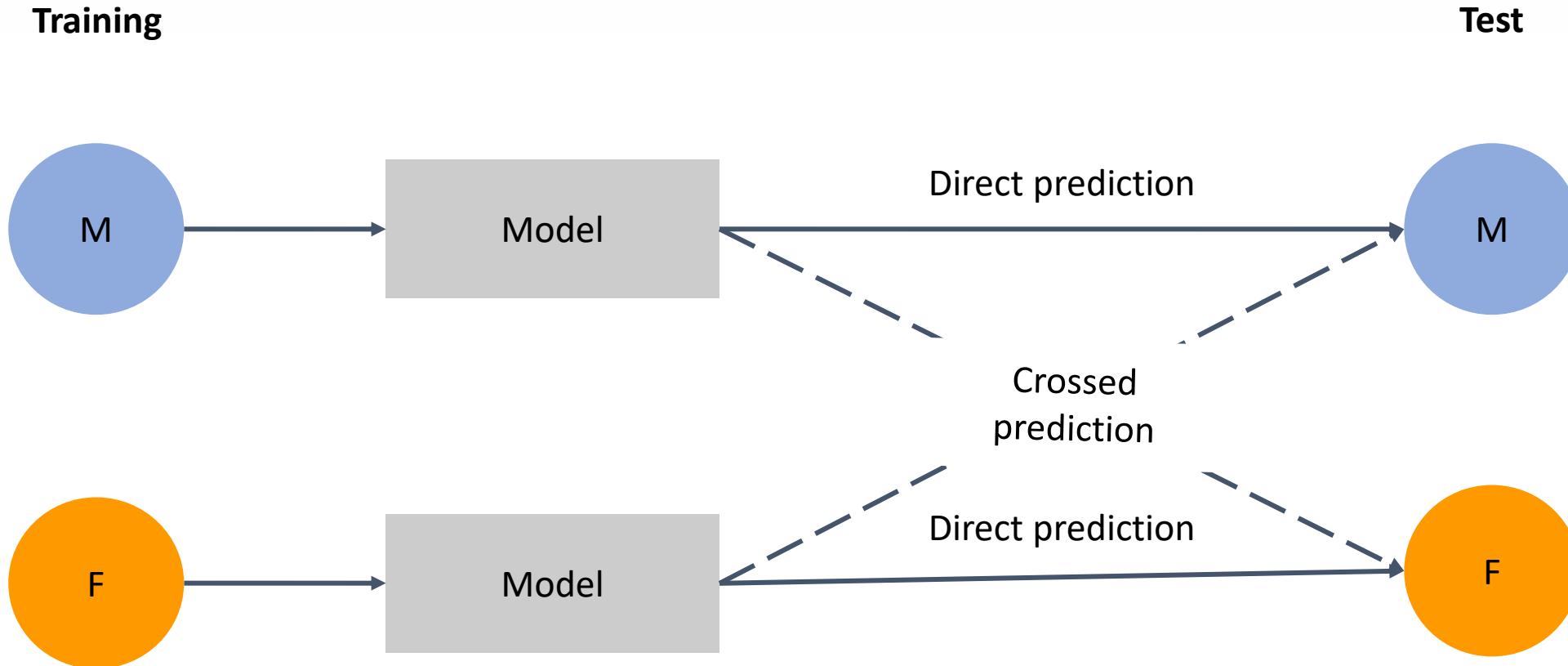
We analyzed two scenarios



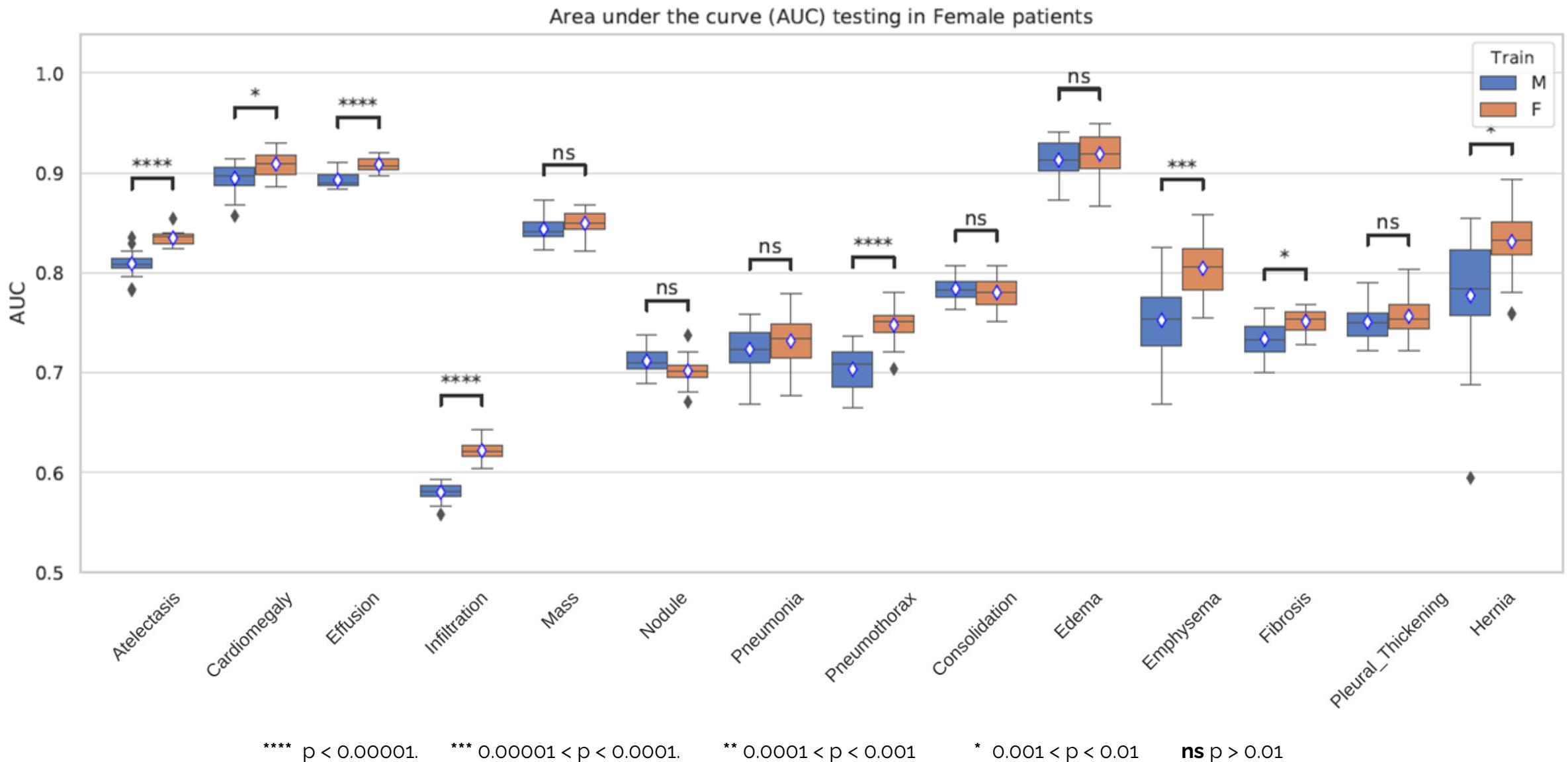
Experimental setting

- We repeated the experiment 20 times with different training/test folds and observe the trends
- The folds were constructed so that for every pathology, we have the same number of male/female patients.
- We measured the Area Under the ROC curve (AUC) for every model using different test sets.

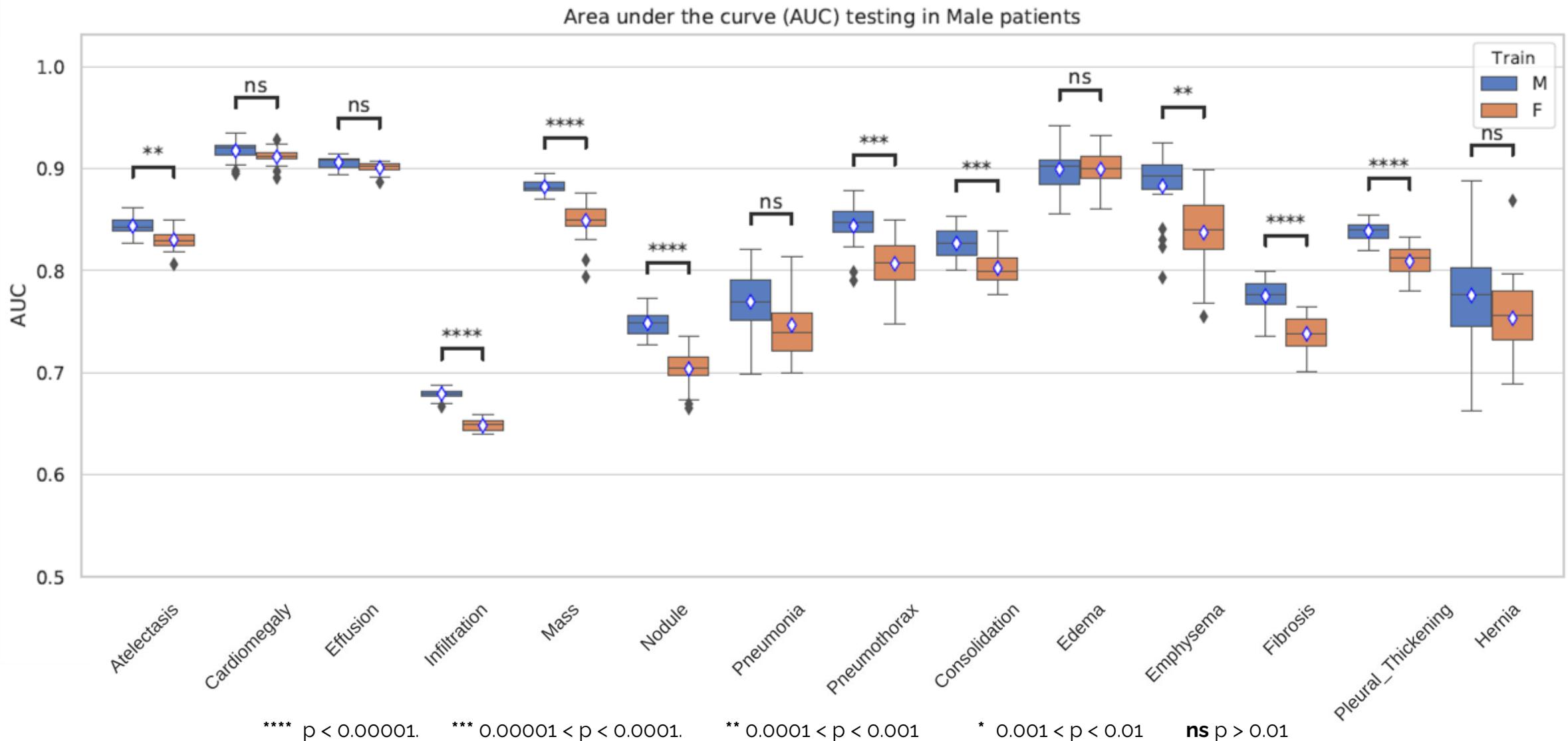
Full imbalance experiment



Test set = female only



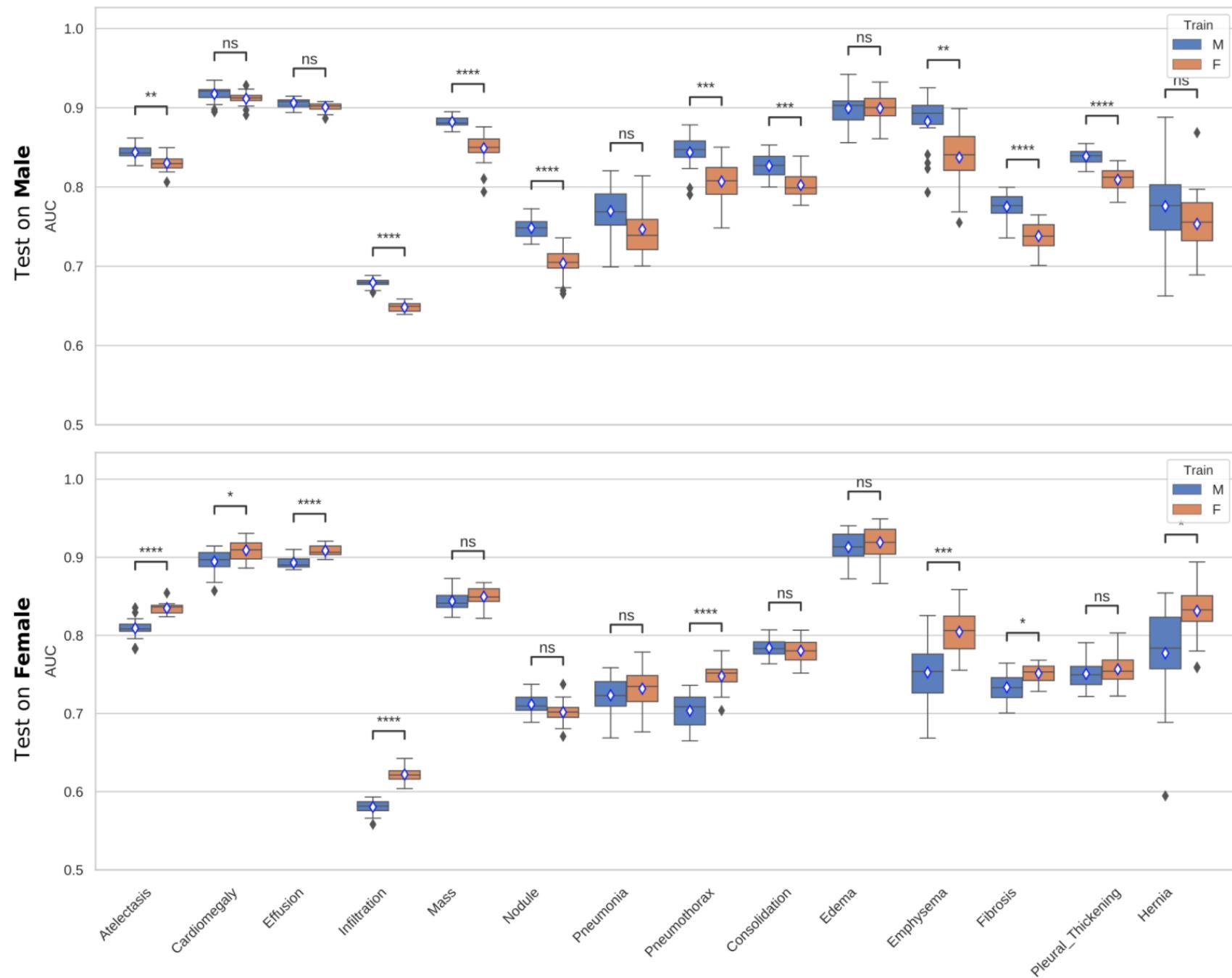
Test set = male only



DenseNet

Dataset NIH

NIH Dataset



DenseNet

CheXpert Dataset

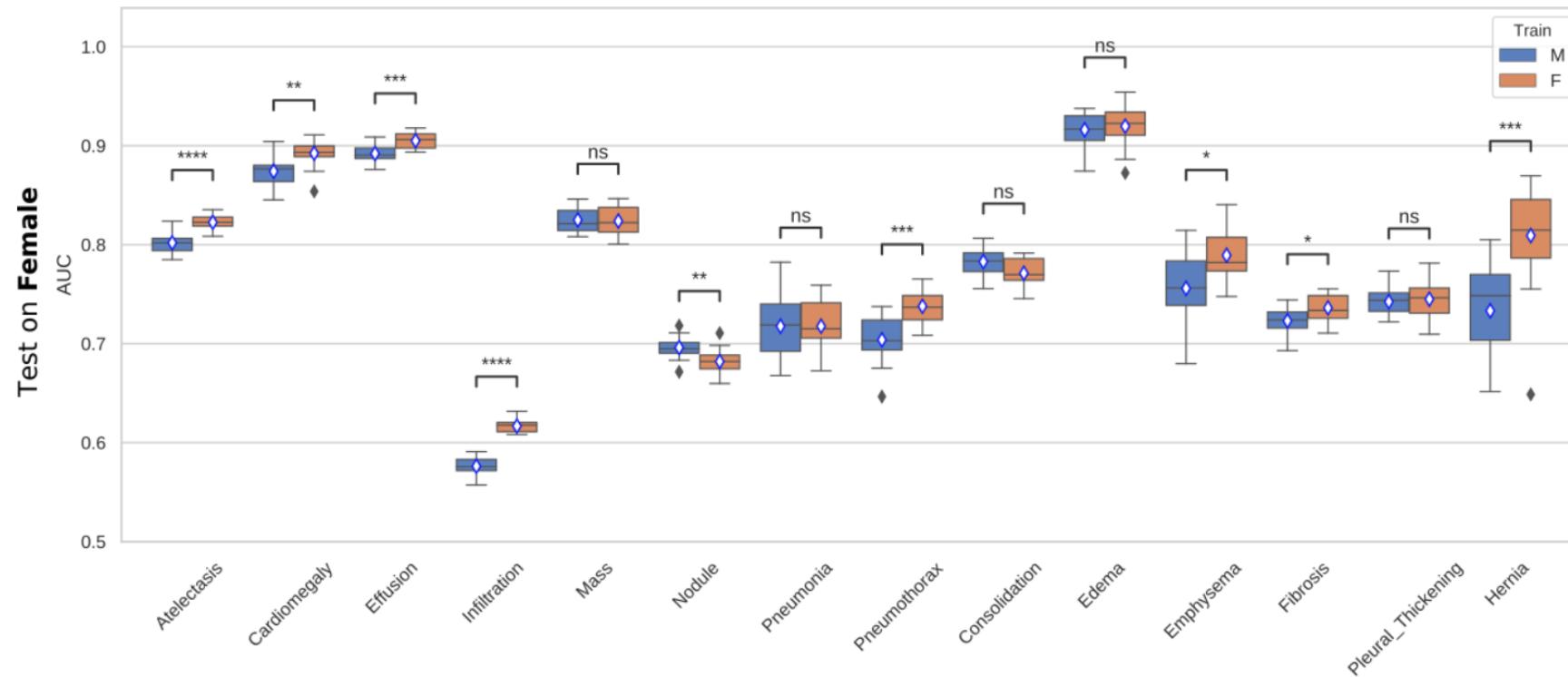
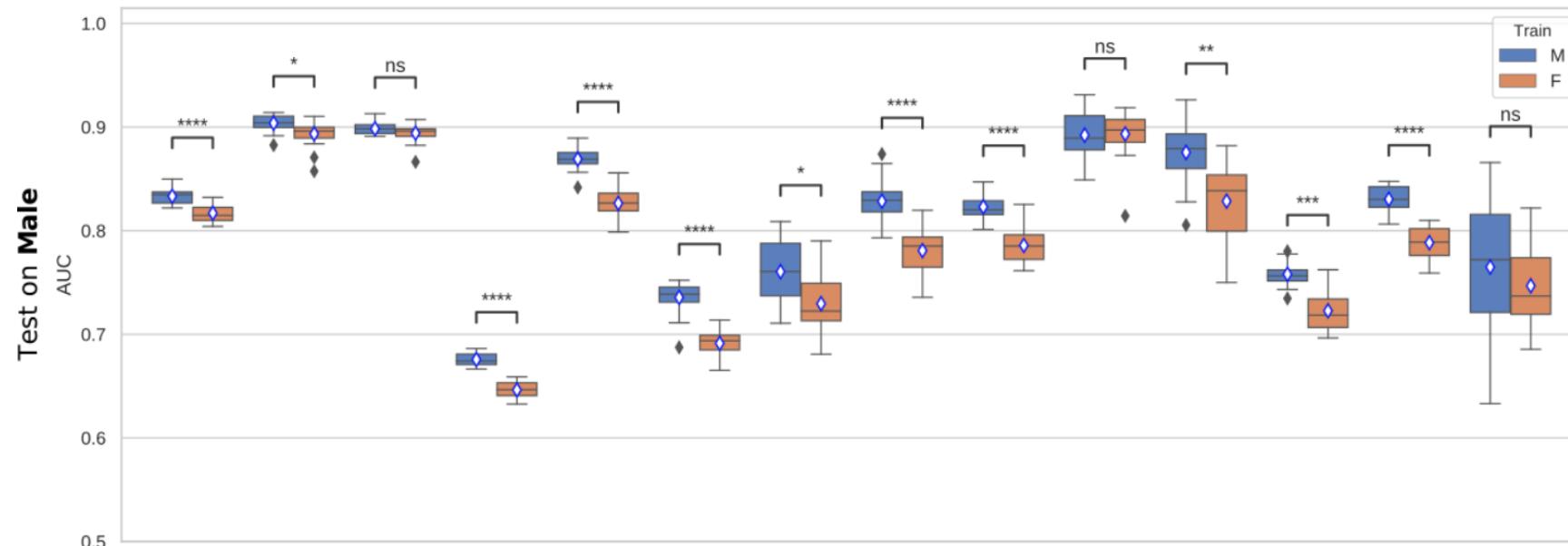
CheXpert Dataset



ResNet

NIH Dataset

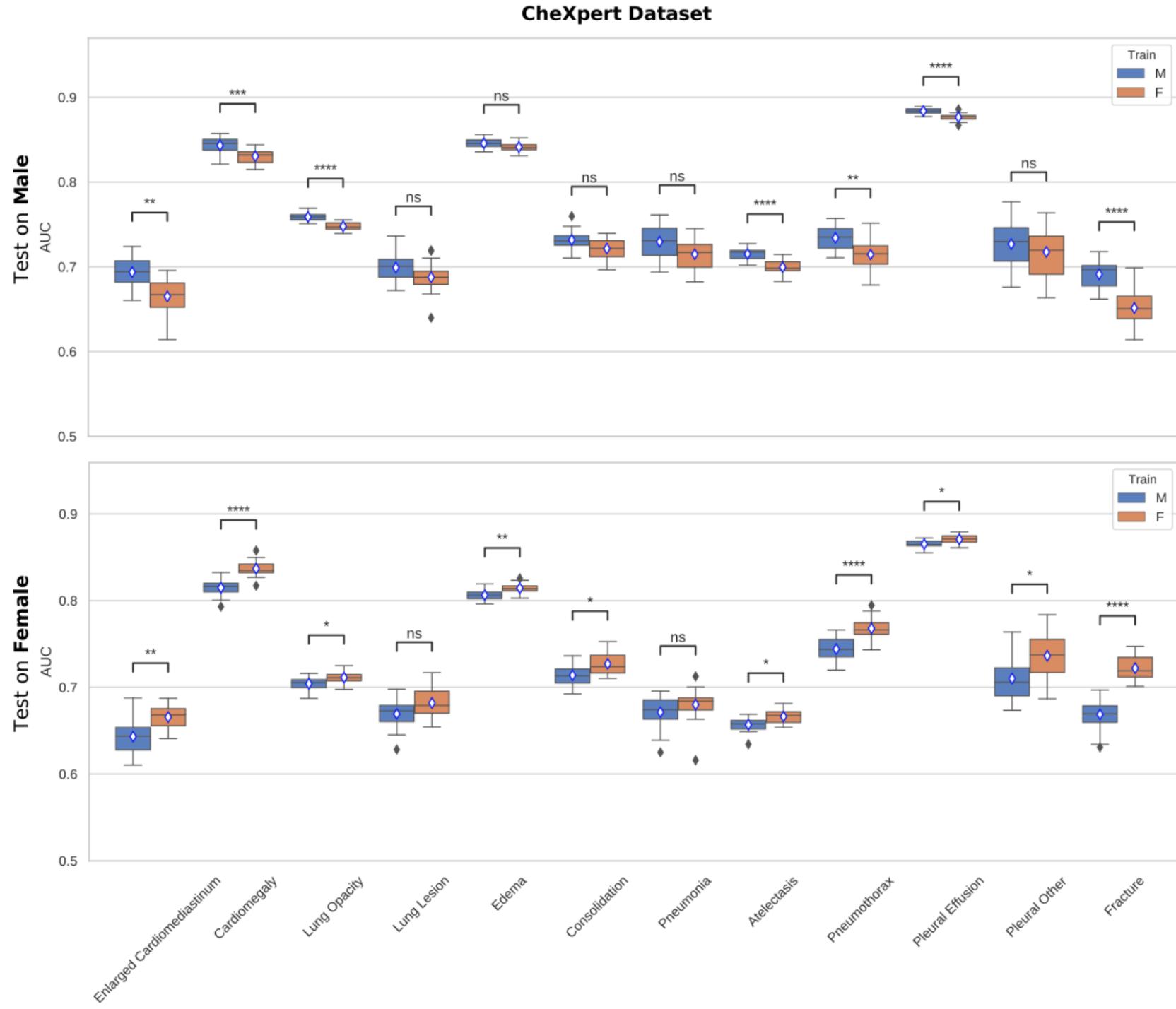
NIH Dataset



CheXpert Dataset

ResNet

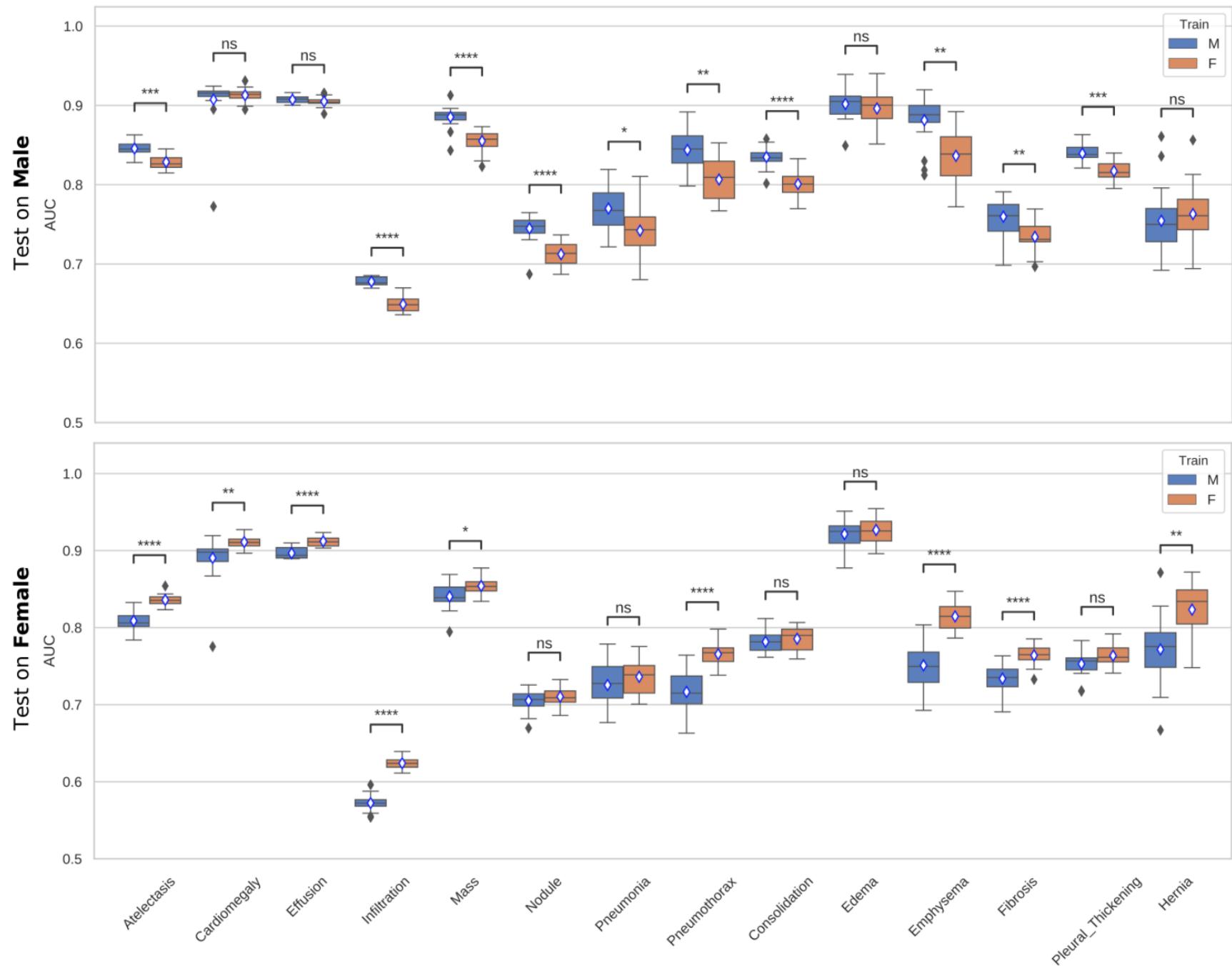
CheXpert Dataset



Inception v3

NIH Dataset

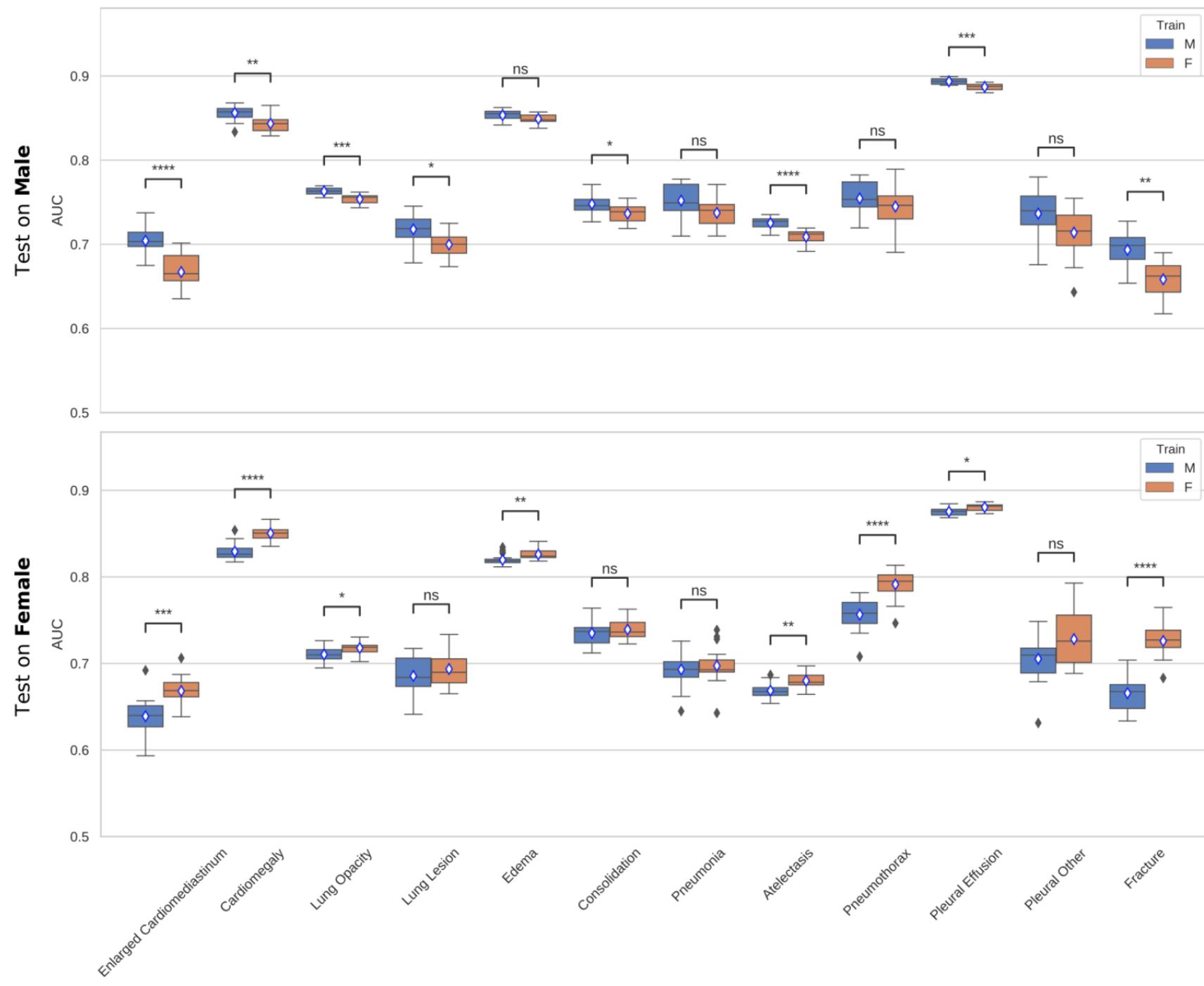
NIH Dataset



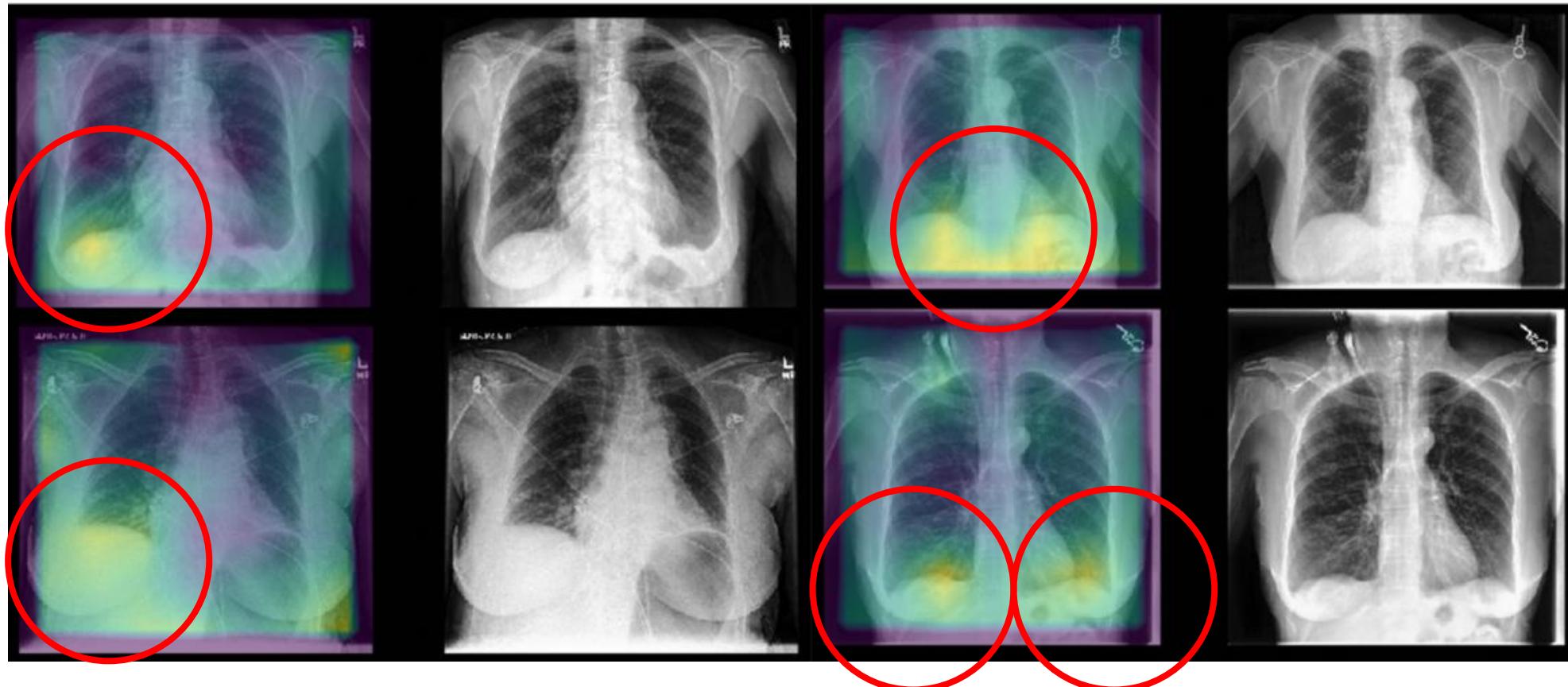
Inception v3

CheXpert Dataset

CheXpert Dataset



Why is it happening?



- Model **trained with male images**, and **tested on female patients**, on the lung opacity task.
- Class Activation Maps generated for **False Positive predictions**

Why is it happening?

Proc. Coll. Radiol. Aust. (1958), 2, 107

The Elimination of Confusing Breast Shadows in Chest Radiography

COLIN ALEXANDER

From 18 Lower Symonds Street, Auckland, New Zealand

The female breast is a common source of confusion in the interpretation of chest radiographs and only too often in women the examination of the basal areas of the lungs is much less effective than it might be because of this factor.

When large, the breasts may so obscure the bases as to make their assessment most unreliable if conventional techniques are employed. Even with small breasts confusion can arise. In the conventional P.A. position the breast shadows are not uncommonly asymmetrical, and if a basal lesion is present it is often difficult to decide how much of the opacity is due to the

if the breasts are large and the diaphragm high due to obesity.

Because of these difficulties it is not uncommon to conclude a laborious and time-consuming chest examination still in doubt as to the exact state of the lung bases.

In the search for a simple and effective answer to this problem various accessory techniques were investigated until the alternative of X-raying the chest in the supine position was tried. In a trial period of twelve months this technique has proved almost invariably effective. The

Why is it happening?

Assessing Bias in Medical AI

Melanie Ganz^{1,2} Sune H. Holm³ Aasa Feragen^{4,2}

Abstract

Machine learning and artificial intelligence are increasingly deployed in critical societal functions such as finance, media and healthcare. Along with their deployment come increasing reports of their failure when viewed through the lens of ethical principles such as fairness, democracy and equal opportunity. As a result, research into fair algorithms and mitigation of bias in data and algorithms, has surged in recent years. However,

bias and promoting fairness in medical AI.

2. Case discussions

In the two cases discussed below, bias has different sources, and we first discuss the technical implications of different types of bias and the corresponding limitations in the potential for fair medical AI. Following the technical discussion, we revisit the issue on a higher level from an ethical and societal point of view.

Interpretable Machine Learning in Healthcare

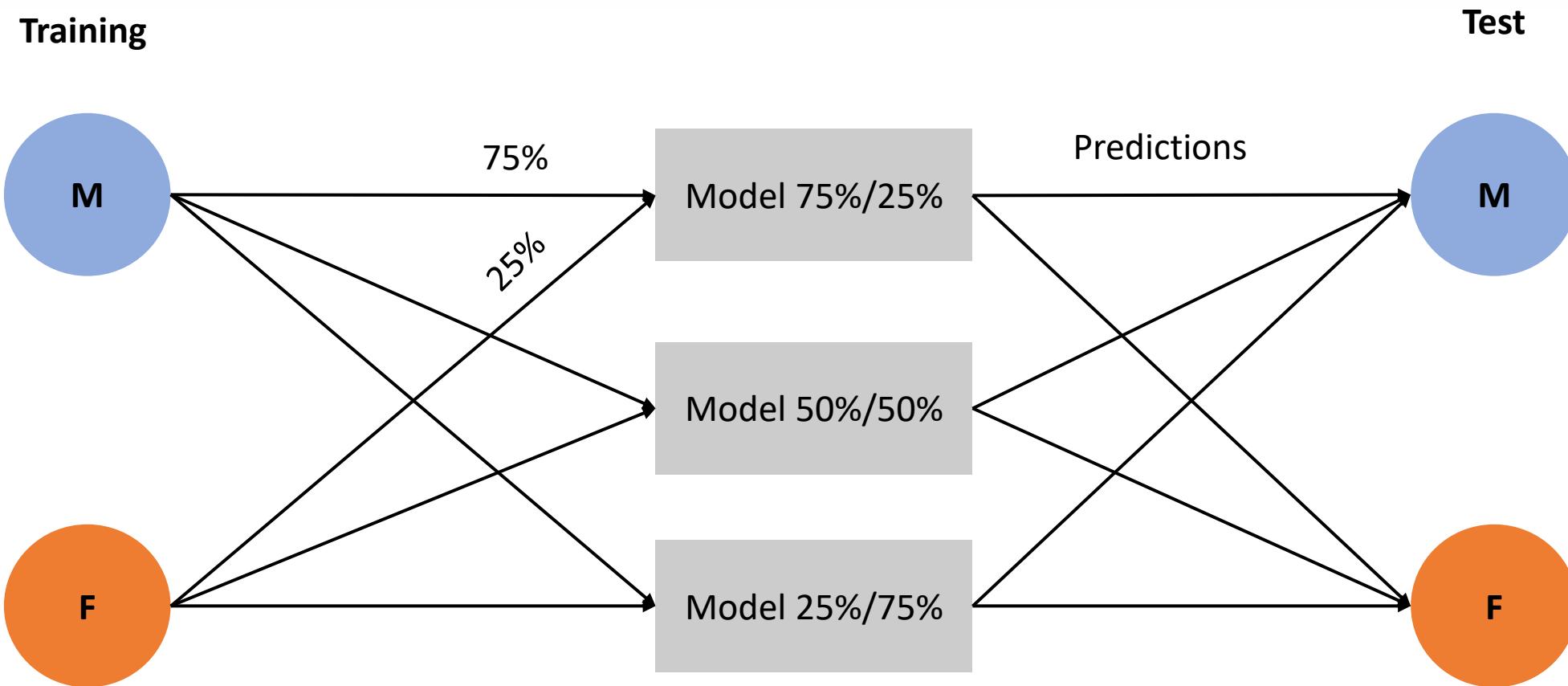
ICML 2021 Workshop

Case A: Gender bias in diagnostic AI Training algorithms on different subgroups can reveal classification imbalances. A recent paper (4) studied the effect of training set imbalance in image-based computer aided diagnosis. The authors studied diagnoses of 12 different thoracic diseases based on chest X-ray using a state-of-the-art classifier, and using training sets with a gender balance of 0/100%, 25/75%, 50/50%, 75/25% and 100/0% women/men, respectively. As expected, diagnostic AI performed better on women when it was specialized to diagnose women, and vice versa. However, for some diseases – pneumothorax being an example – the diagnostic AI specialized to diagnose women was actually better at diagnosing men, than at diagnosing women. Replacing some of the training set females with males emphasized this difference, but the fact remained: The best-performing algorithm for women was better at diagnosing men than women. And at the same time this was the worst-performing algorithm for men.

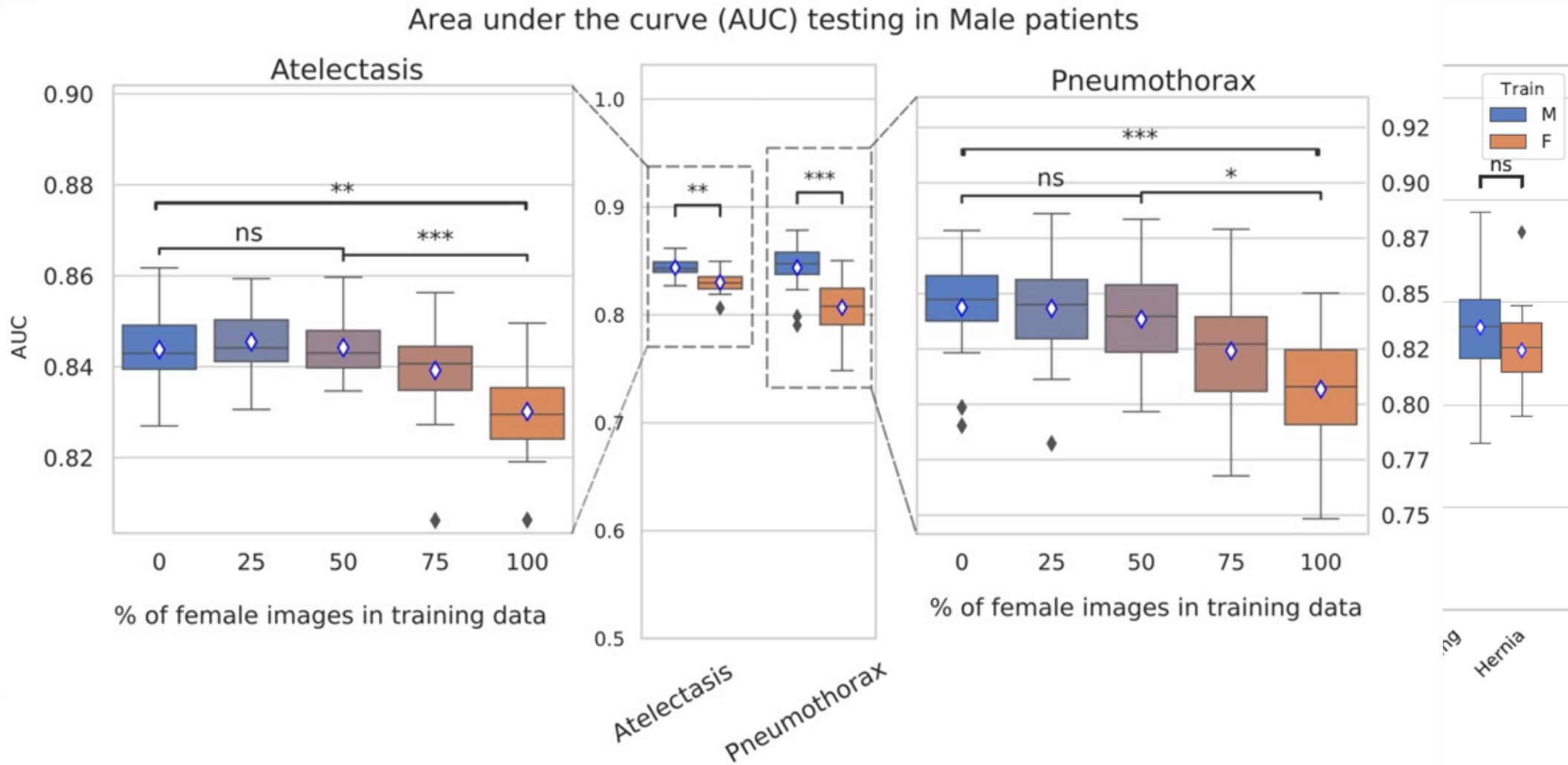
From the machine learner's point of view, this is a frustrating result: In many applications, it is fully feasible to ensure balance between sensitive groups in a training set, but here balancing the training set was insufficient to obtain equal

performance – the classification problem appears to be more challenging for one group than for another. In the case of chest imaging, this has a plausible biological explanation: In x-ray imaging of the upper thorax women's breasts occlude the imaged organs, resulting in poorer image contrast for the relevant anatomy.

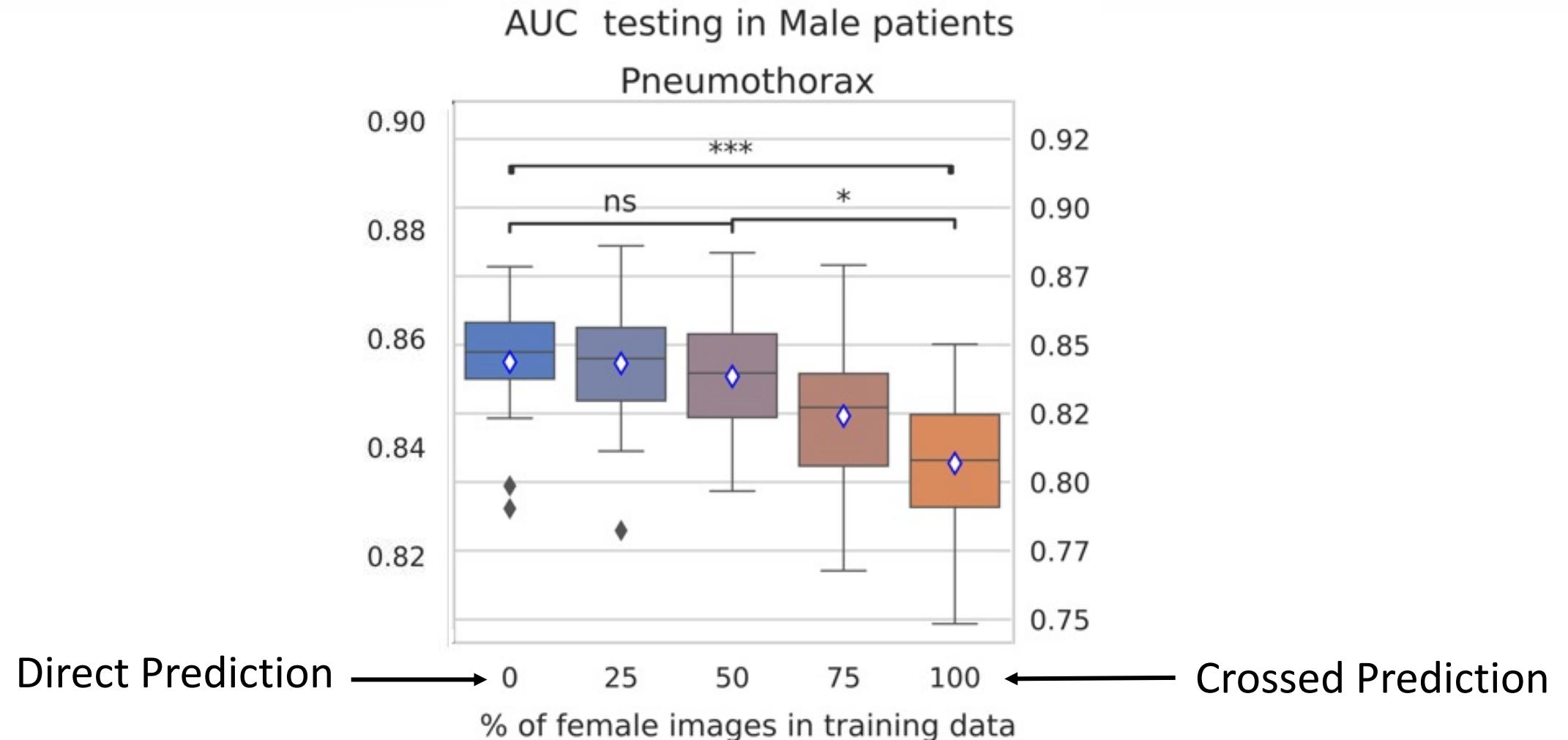
Partial imbalance



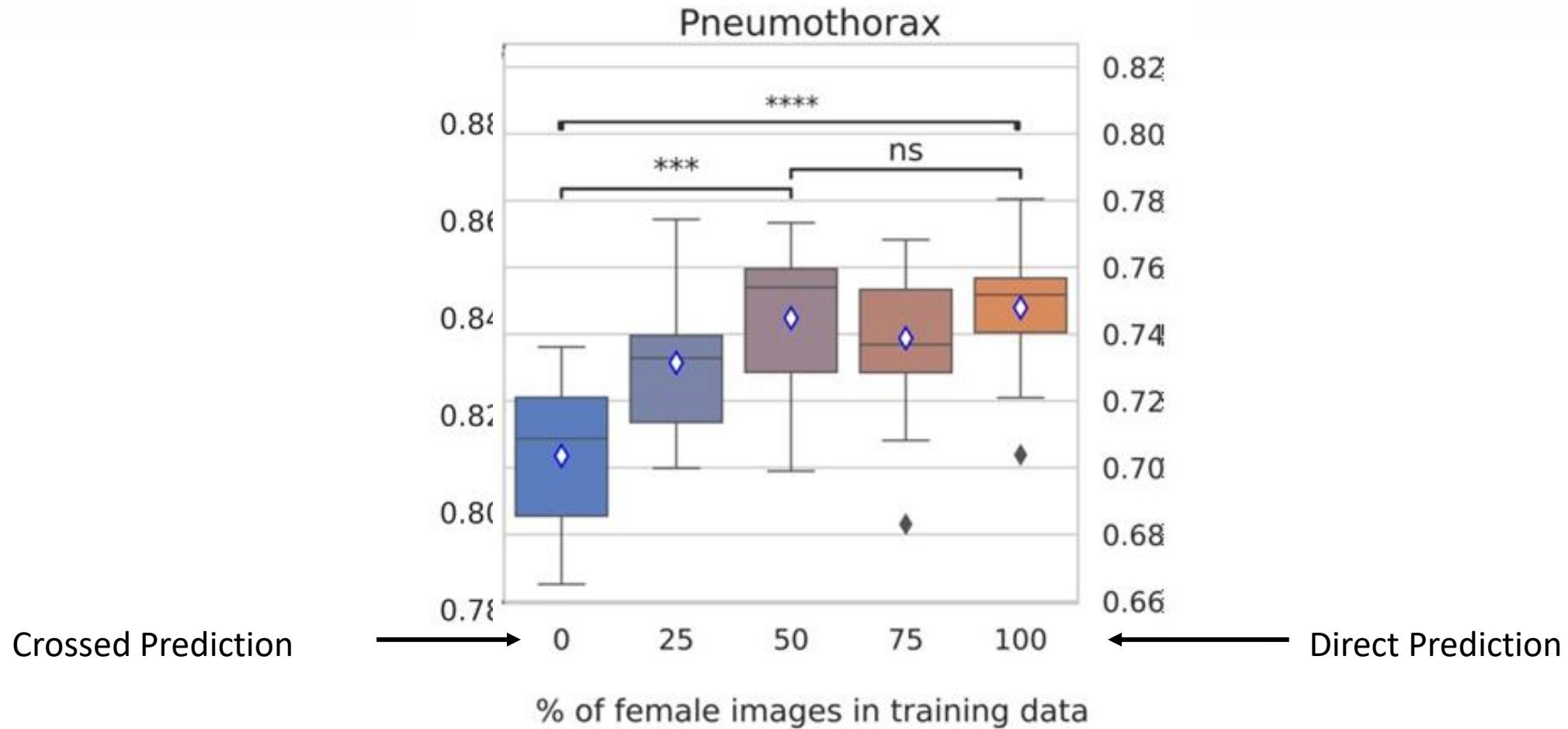
Test set = male only



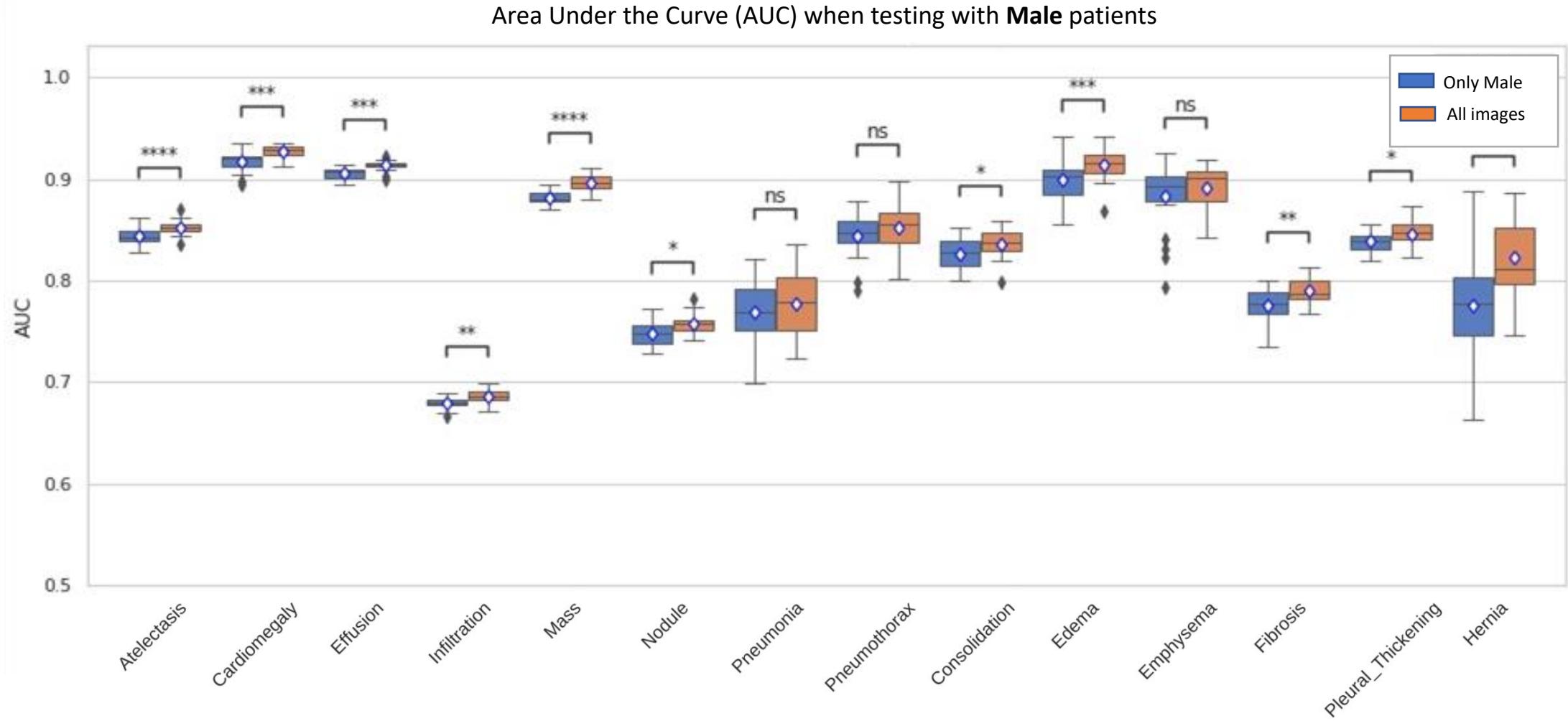
Test set = male only



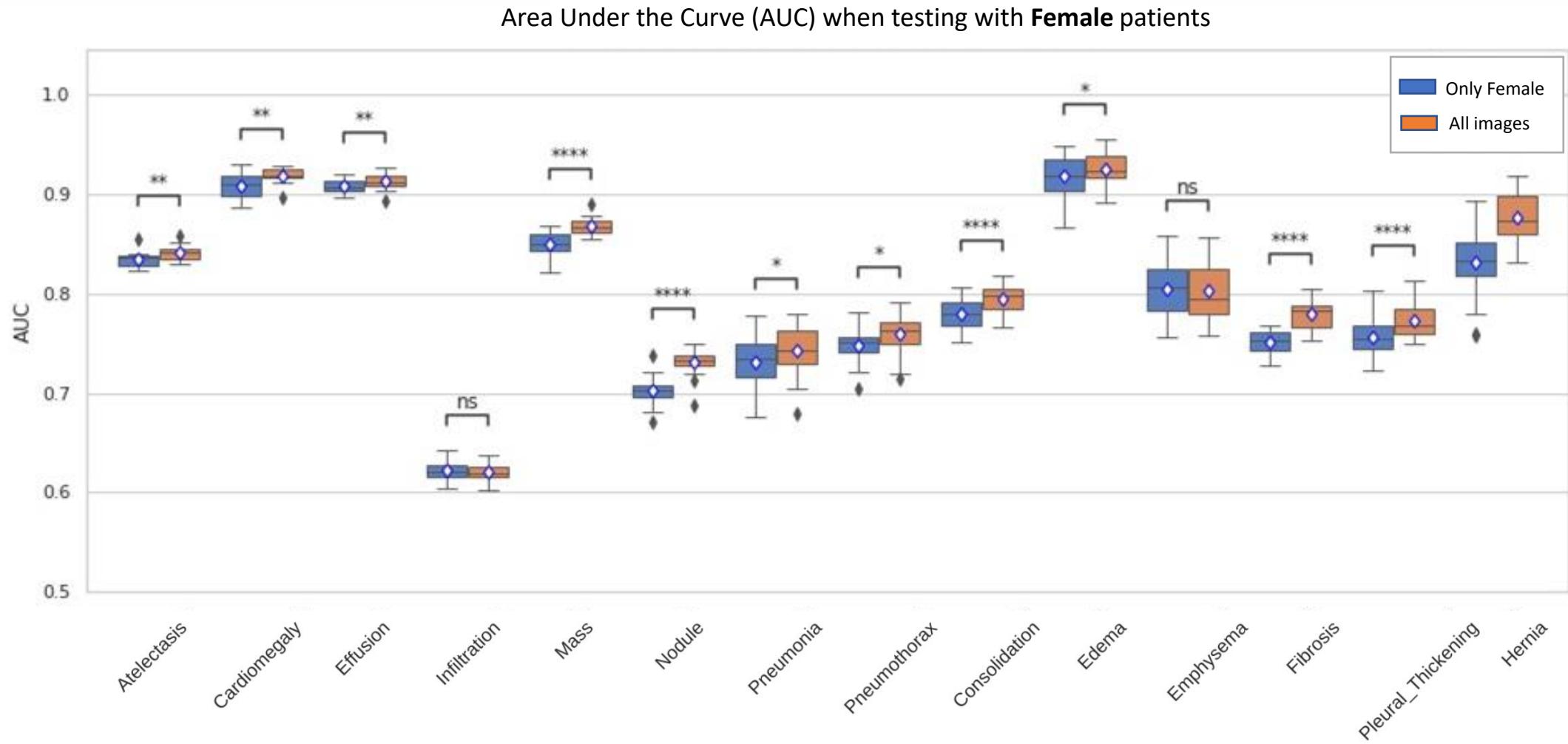
Test set = female only



So should we train a classifier per gender?



So should we train one classifier per gender?



Limitations of our study

- We used the term "**gender**" to characterize our imbalance study following the demographic variables reported in the original NIH dataset publication.
- However, we are aware **that gender is a fluid cultural construct** and **binarizing this concept constitutes a limitation of our study.**
- Given that some anatomical attributes are reflected in X-ray images, the term "**sex**" could be more accurate, according to the Sex and Gender Equity in Research guidelines.

Some takeaways

- The performance of **CAD systems should be audited at the group-level** considering demographic attributes like sex/gender.
- **CNNs are prone to learn features useful for specific subgroups seen during training**, which may lead to a decrease in performance due to population shifts in the test set.
- **Authors should report demographic information** together with their MIC papers and (specially) public datasets
- Diversifying your dataset is important! But....

Limitations of our study

AI, Medicine, & Bias: Diversifying Your Dataset is Not Enough

Rachel Thomas, PhD

Director, Center for Applied Data Ethics at USF

Co-Founder, fast.ai



Limitations of our study

"A lot of times, people are talking about bias in the sense of equalizing performance across groups. They're not thinking about **the underlying foundation, whether a task should exist in the first place, who creates it, who will deploy it on which population, who owns the data, and how is it used?**"

– Dr. Timnit Gebru



The New York Times

Dealing With Bias in Artificial Intelligence



Limitations of our study

Patterns

CellPress
OPEN ACCESS

Opinion

Moving beyond “algorithmic bias is a data problem”

Sara Hooker^{1,*}

¹Google Brain, Mountain View, CA, USA

*Correspondence: shooker@google.com

<https://doi.org/10.1016/j.patter.2021.100241>

A surprisingly sticky belief is that a machine learning model merely *reflects* existing algorithmic bias in the dataset and does not itself contribute to harm. Why, despite clear evidence to the contrary, does the myth of the impartial model still hold allure for so many within our research community? Algorithms are not impartial, and some design choices are better than others. Recognizing how model design impacts harm opens up new mitigation techniques that are less burdensome than comprehensive data collection.

Moving beyond “algorithmic bias is a data problem”

In the absence of intentional interventions, a trained machine learning model *can and does* amplify undesirable biases in the training data. A rich body of work to date has examined these forms of problematic algorithmic bias, finding disparities—relating to race, gender, geo-diversity, and more—in the performance of machine learning models.¹

However, a surprisingly prevalent belief is that a machine learning model merely *reflects* existing algorithmic bias in the dataset and does not itself contribute to harm. Here, we start out

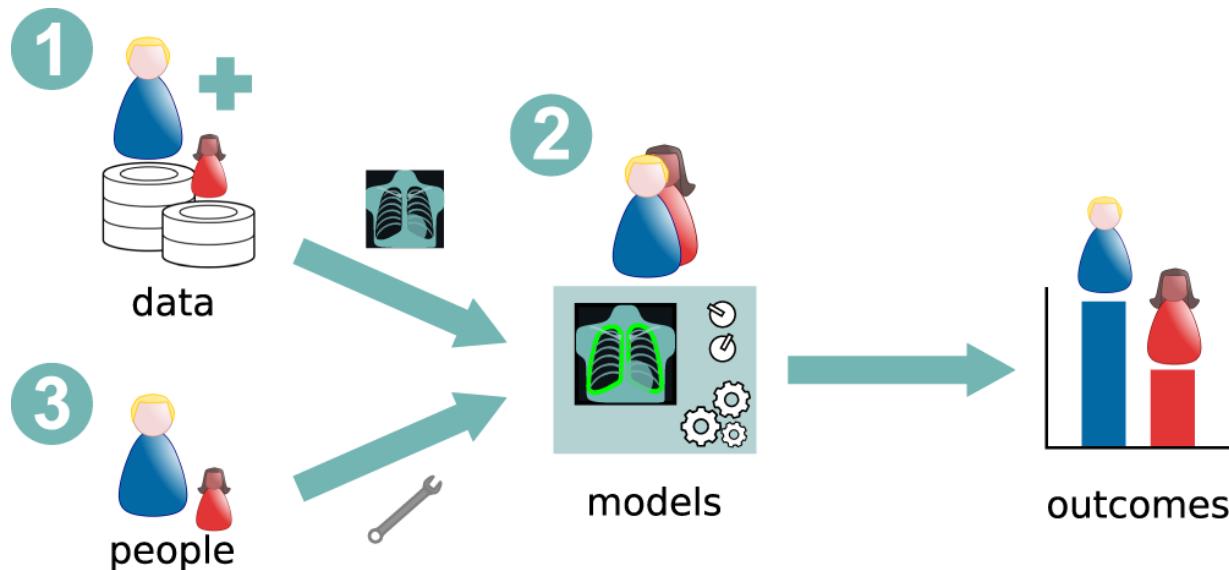
niques that are far less burdensome than comprehensive data collection.

The impact of our model design choices

If you replace algorithmic bias with test-set accuracy, it becomes a much more acceptable stance that our modeling choices—architecture, loss function, optimizer, hyper-parameters—express a preference for final model behavior. Most students of machine learning are familiar with some variation of Figure 1, where varying the degree of a polynomial function leads to

“Understanding which model design choices disproportionately amplify error rates on protected underrepresented features is a crucial first step in helping curb algorithmic harm.”

Three reasons behind biased systems: data, models and people



nature communications

[Comment](#) | [Open Access](#) | [Published: 06 August 2022](#)

Addressing fairness in artificial intelligence for medical imaging

[María Agustina Ricci Lara](#) [Rodrigo Echeveste](#) & [Enzo Ferrante](#)

Nature Communications 13, Article number: 4581 (2022) | [Cite this article](#)

A plethora of work has shown that AI systems can systematically and unfairly be biased against certain populations in multiple scenarios. The field of medical imaging, where AI systems are beginning to be increasingly adopted, is no exception. Here we discuss the meaning of fairness in this area and comment on the potential sources of biases, as well as the strategies available to mitigate them. Finally, we analyze the current state of the field, identifying strengths and highlighting areas of vacancy, challenges and opportunities that lie ahead.

The health research community is starting to look into these problems

The image shows a screenshot of the PubMed Advanced Search Builder interface at the top, followed by the search results page below.

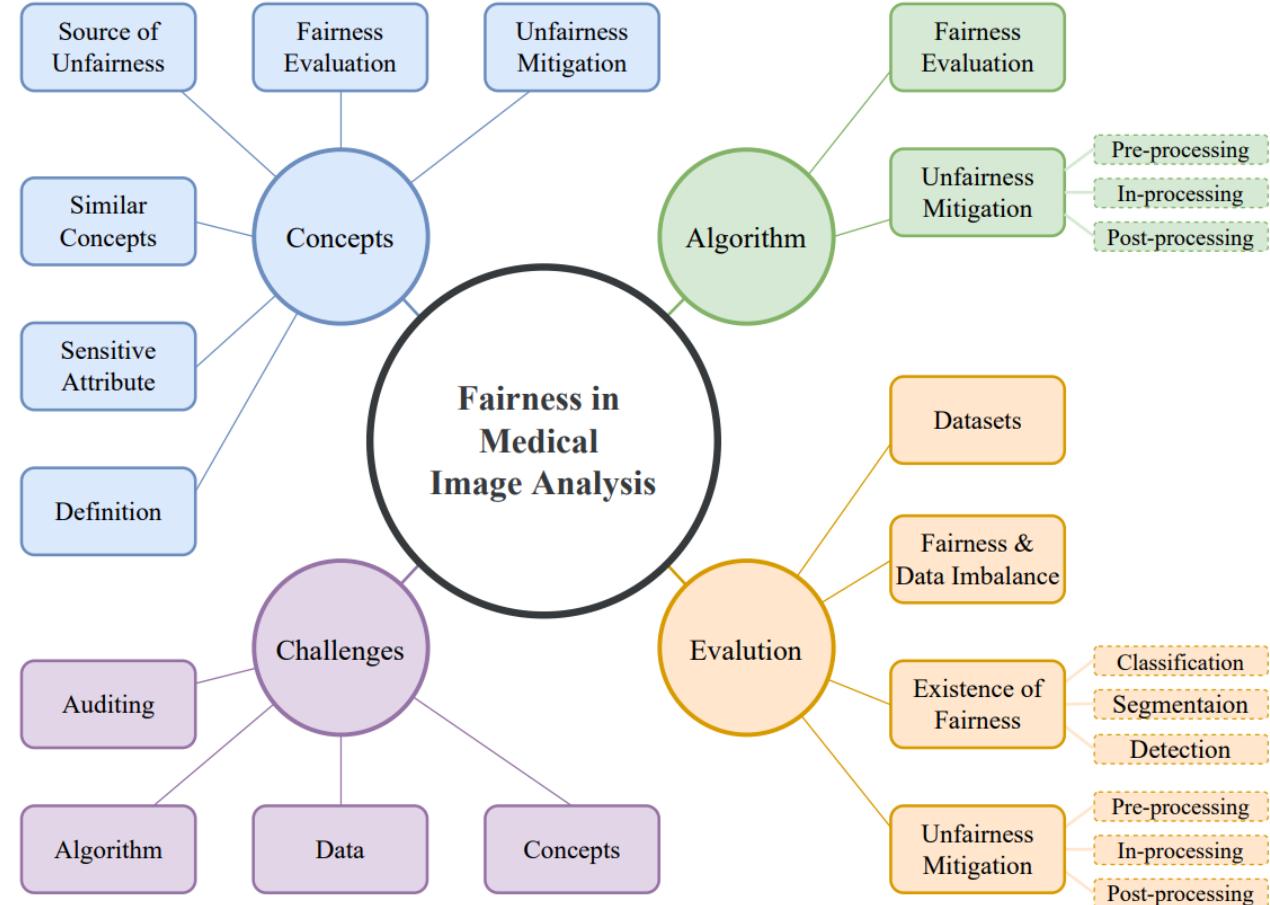
PubMed Advanced Search Builder:

- Add terms to the query box:
All Fields dropdown, Enter a search term input field containing the query: `((fairness[Title]) OR (bias[Title])) AND ((artificial intelligence[Title]) OR (deep learning[Title]) OR (machine learning[Title]))`, AND dropdown, Show Index link.
- Query box: The search query is displayed: `((fairness[Title]) OR (bias[Title])) AND ((artificial intelligence[Title]) OR (deep learning[Title]) OR (machine learning[Title]))`.

Search Results Page:

- MY NCBI FILTERS icon.
- RESULTS BY YEAR chart showing publication counts from 2014 to 2022.
- 122 results found, page 1 of 13.
- Result 1: Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health.
Cite: Fletcher RR, Nakashimana A, Olubeko O.
Share: PMID: 33981989, Free PMC article.
- Result 2: Ensuring Fairness in Machine Learning to Advance Health Equity.

Recent survey



The MIC research community is starting to look into these problems

MICCAI 2020



Fairness of Classifiers Across Skin Tones in Dermatology

Newton M. Kinyanjui^{1,4}, Timothy Odonga^{1,4}, Celia Cintas¹,
Noel C. F. Codella², Rameswar Panda³, Prasanna Sattigeri²,
and Kush R. Varshney^{1,2}

¹ IBM Research – Africa, Nairobi 00100, Kenya

kvarshn@us.ibm.com

² IBM Research – T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

³ IBM Research – Cambridge, Cambridge, MA 02142, USA

⁴ Carnegie Mellon University Africa, Kigali, Rwanda

Abstract. Recent advances in computer vision have led to breakthroughs in the development of automated skin image analysis. However, no attempt has been made to evaluate the consistency in performance across populations with varying skin tones. In this paper, we present an approach to estimate skin tone in skin disease benchmark datasets and investigate whether model performance is dependent on this measure. Specifically, we use individual typology angle (ITA) to approximate skin tone in dermatology datasets. We look at the distribution of ITA values to better understand skin color representation in two benchmark datasets: 1) the ISIC 2018 Challenge dataset, a collection of dermoscopic

MICCAI 2021

Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation

Esther Puyol-Antón¹, Bram Ruijsink^{1,2}, Stefan K. Piechnik⁷, Stefan Neubauer⁷, Steffen E. Petersen^{3,4,5,6}, Reza Razavi^{1,2}, and Andrew P. King¹

¹ School of Biomedical Engineering & Imaging Sciences, King's College London, UK
² Guy's and St Thomas' Hospital, London, UK

³ William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, Charterhouse Square, London, EC1M 6BQ, UK

⁴ Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, EC1A 7BE, London, UK

⁵ Health Data Research UK, London, UK

⁶ Alan Turing Institute, London, UK

⁷ Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, UK.

Abstract. The subject of ‘fairness’ in artificial intelligence (AI) refers to assessing AI algorithms for potential bias based on demographic characteristics such as race and gender, and the development of algorithms to address this bias. Most applications to date have been in computer vision, although some work in healthcare has started to emerge. The use of deep learning (DL) in cardiac MR segmentation has led to impressive results in recent years, and such techniques are starting to be translated into clinical practice. However, no work has yet investigated

MICCAI 2022

Feature Robustness and Sex Differences in Medical Imaging: A Case Study in MRI-Based Alzheimer’s Disease Detection

Eike Petersen¹, Aasa Feragen¹, Maria Luise da Costa Zemsch¹, Anders Henriksen¹, Oskar Eiler Wiese Christensen¹, Melanie Ganz^{2,3},
for the Alzheimer’s Disease Neuroimaging Initiative

¹ Technical University of Denmark DTU Compute, Kgs. Lyngby, Denmark
{ewipe,afhar}@dtu.dk

² Department for Computer Science, University of Copenhagen, Copenhagen, Denmark

³ Rigshospitalet, Neurobiology Research Unit, Copenhagen, Denmark
melanie.ganz@rnu.dk

Abstract. Convolutional neural networks have enabled significant improvements in medical image-based diagnosis. It is, however, increasingly clear that these models are susceptible to performance degradation when facing spurious correlations and dataset shift, leading, e.g., to underperformance on underrepresented patient groups. In this paper, we compare two classification schemes on the ADNI MRI dataset: a simple logistic regression model using manually selected volumetric features, and a convolutional neural network trained on 3D MRI data. We assess the robustness of the trained models in the face of varying dataset splits, training set sex composition, and stage of disease. In contrast to ear-

Recent works

Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation

Esther Puyol-Antón¹, Bram Ruijsink^{1,2}, Stefan K. Piechnik⁷, Stefan Neubauer⁷, Steffen E. Petersen^{3,4,5,6}, Reza Razavi^{1,2}, and Andrew P. King¹

¹ School of Biomedical Engineering & Imaging Sciences, King's College London, UK

² Guy's and St Thomas' Hospital, London, UK.

³ William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, Charterhouse Square, London, EC1M 6BQ, UK

⁴ Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, EC1A 7BE, London, UK

⁵ Health Data Research UK, London, UK

⁶ Alan Turing Institute, London, UK

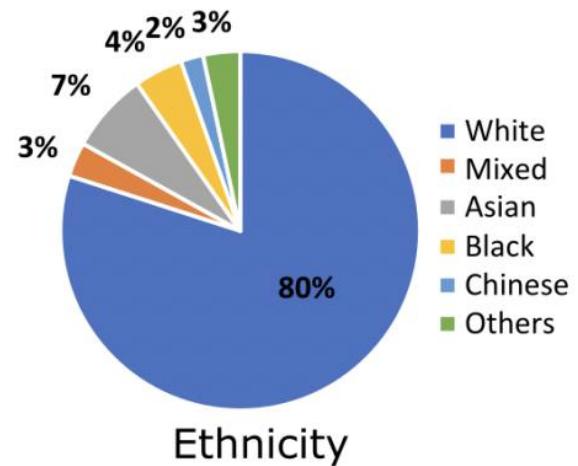
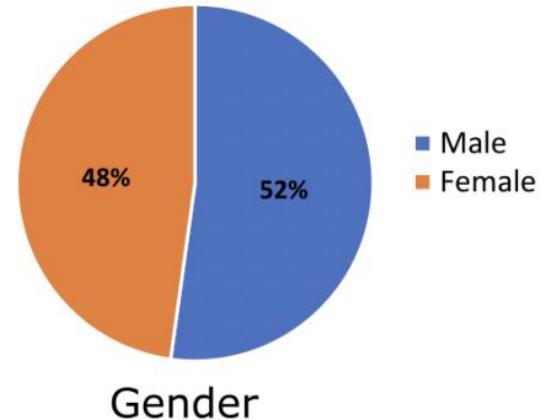
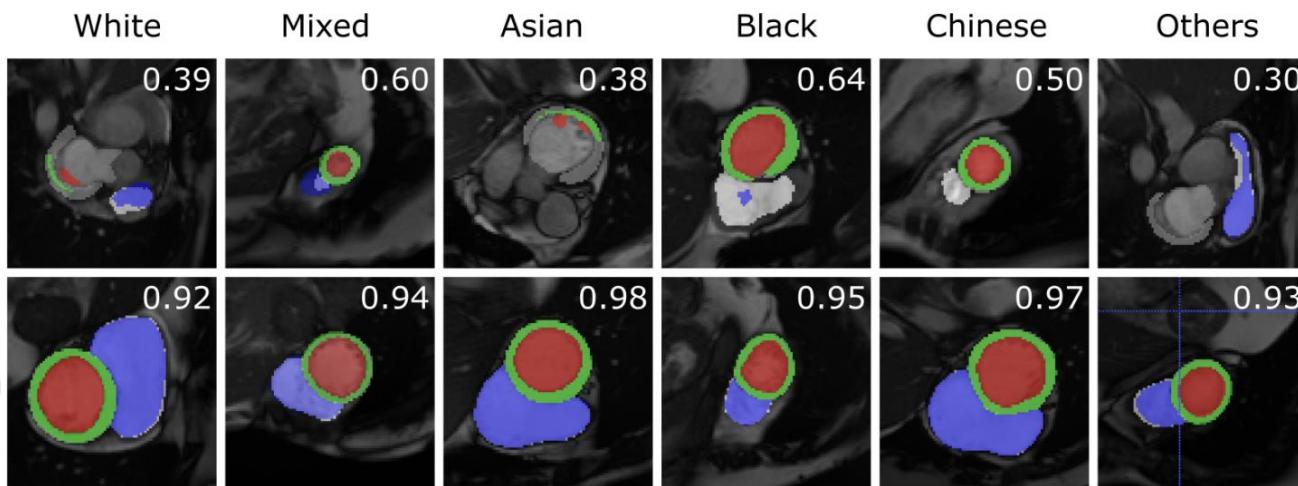
⁷ Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, UK.

Abstract. The subject of ‘fairness’ in artificial intelligence (AI) refers to assessing AI algorithms for potential bias based on demographic characteristics such as race and gender, and the development of algorithms to address this bias. Most applications to date have been in computer vision, although some work in healthcare has started to emerge. The use of deep learning (DL) in cardiac MR segmentation has led to impressive results in recent years, and such techniques are starting to be translated into clinical practice. However, no work has yet investigated

Recent works

DSC (%) for Baseline —Fairness through unawareness

	ED			ES			Avg
	LVBP	LVM	RVBP	LVBP	LVM	RVBP	
Total	93.48	83.12	89.37	89.37	86.31	80.61	87.05
Male	93.58	83.51	88.82	90.68	85.31	81.00	87.02
Female	93.39	82.71	89.90	89.59	86.60	80.21	87.07
White	97.33	93.08	94.09	95.06	90.58	90.88	93.51*
Mixed	92.70	78.94	86.91	86.70	82.54	79.32	84.52*
Asian	94.53	87.33	90.51	90.13	88.94	81.94	88.90*
Black	92.77	85.93	89.49	89.42	85.74	71.91	85.88*
Chinese	91.81	74.51	85.74	86.39	85.12	79.34	83.82*
Others	91.74	78.94	89.50	88.53	84.96	80.27	85.66*



Recent works



ARTICLES | VOLUME 4, ISSUE 6, E406-E414, JUNE 01, 2022

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, MD • Imon Banerjee, PhD • Ananth Reddy Bhimireddy, MS • John L Burns, MS • Leo Anthony Celi, MD •

Li-Ching Chen, BS • et al. [Show all authors](#)

Open Access • Published: May 11, 2022 • DOI: [https://doi.org/10.1016/S2589-7500\(22\)00063-2](https://doi.org/10.1016/S2589-7500(22)00063-2) •

Check for updates

PlumX Metrics

Summary

Background

Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that

Useful guideliness

FUTURE-AI: Best practices for trustworthy AI in medical imaging

FUTURE-AI is an international, multi-stakeholder initiative for defining and maintaining concrete guidelines that will facilitate the design, development, validation and deployment of trustworthy AI solutions in medical imaging based on six guiding principles: Fairness, Universality, Traceability, Usability, Robustness and Explainability.



<https://future-ai.eu/>

FUTURE-AI Guiding Principles



F airness

for equitable

U niversality

for standardised

T raceability

for monitoring

U sedability

for transferable

R obustness

for reliable

E xplainability

for interpretable

A |

SOLUTIONS

in

MEDICAL IMAGING

FUTURE-AI Fairness Guidelines

- 1 . Inter-disciplinarity:** [...] take into account diverse perspectives brought by multi-disciplinary teams comprising AI developers, radiologists and specialists, but also patients and social scientists (e.g. ethicists).
- 2 . Understanding bias:** In collaboration with domain experts, potentially hidden and application-specific sources of bias (e.g. under-representation of high breast densities in breast imaging datasets) should be carefully analysed and identified beyond standard categories such as sex or ethnicity.
- 3 . Metadata labelling:** [...] non-imaging metadata such as sex, age, ethnicity and income should be included [...]
- 4 . Estimating data (im)balance:** [...] across diverse patient groups in the datasets [...] to identify potential biases and apply appropriate corrective measures.

FUTURE-AI Fairness Guidelines

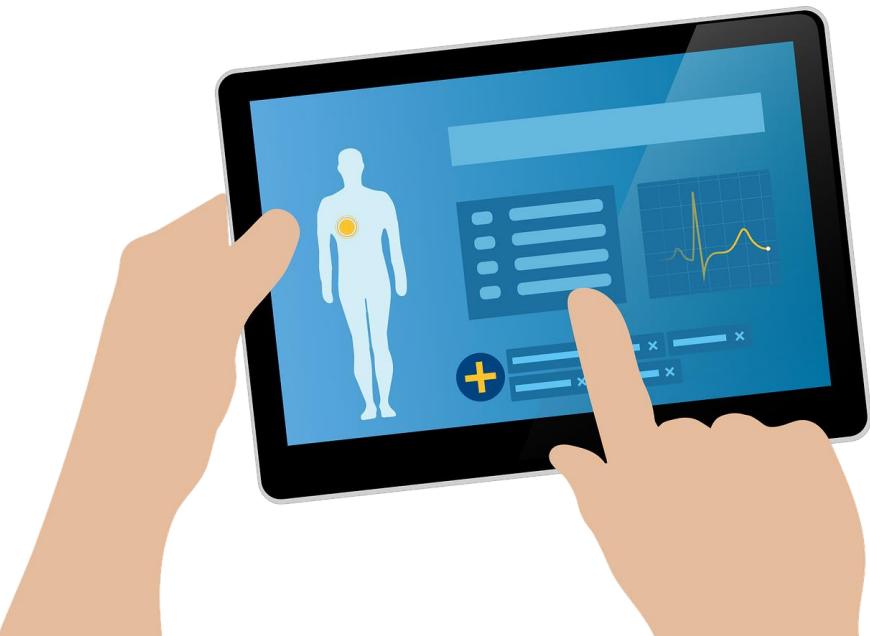
- 5 . Multi-centre datasets:** AI models should be trained and tested on multi-centre datasets to account for differences in populations, resources and geographies across radiology centres.
- 6 . Fairness evaluation:** Algorithmic fairness should be thoroughly and continuously evaluated as an integral part of the AI evaluation process, by using dedicated datasets with adequate diversity, as well as dedicated metrics such as Statistical Parity, Equalised Odds and Predictive Equality.
- 7 . Fairness optimisation:** When bias is detected, corrective measures should be investigated [...] to neutralise discriminatory effects and optimise the fairness of the AI algorithm.
- 8 . Information and training on fairness:** Adequate information and training material should be provided to raise awareness and inform end-users on the fairness, biases and limitations of the AI algorithm.

Other research lines: fairness of ML for EHR analysis



Argentinian Public Health Research
on Data Science and Artificial Intelligence
for Epidemic Prevention

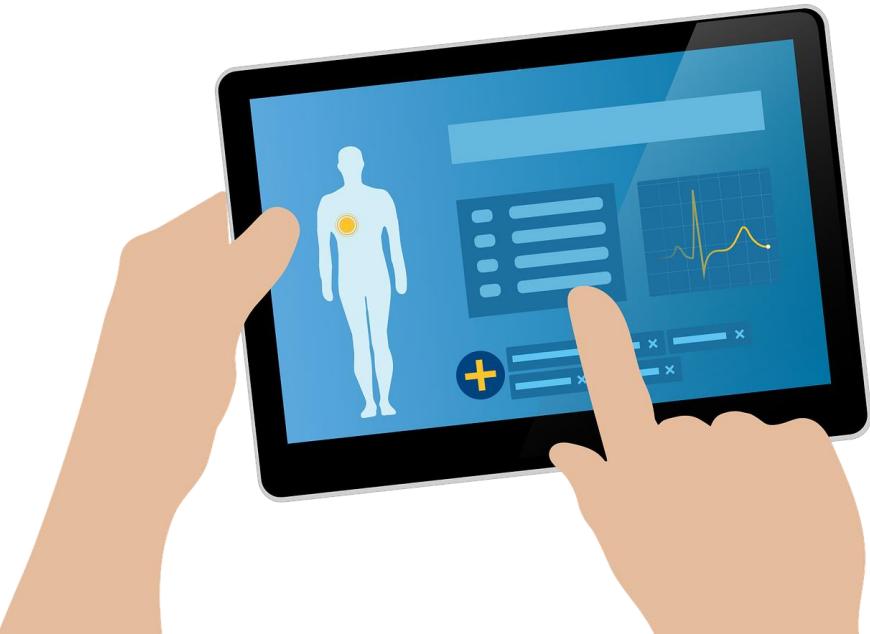
- Incorporate **machine learning** and **data science** to improve Argentina's **Electronic Health Records system**
- Project funded by the **International Development Research Center (IDRC)** from Canada and the **Swedish International Development Cooperation Agency (SIDA)**



Incorporating transversal responsible ML practices



Argentinian Public Health Research
on Data Science and Artificial Intelligence
for Epidemic Prevention



Predictive Model Audit

Data Collection and Analysis

Team composition

Rethinking our data models

Data Collection and Analysis

De-binarianization of gender identity in the EHR data model



Throughout 2021, ARPHAI advanced a series of meetings and workshops to work on the de-binarianization of the national electronic health record system (known as *Historia de Salud Integrada or HSI*) that initially included a binary classification of gender identity. To undertake this work, the project team had enormous support from *La Dirección de Sistemas de Información del Ministerio de Salud de la Nación* (the directorate of information systems in the country's ministry of health), which is also a member of ARPHAI's project team.

The de-binarianization initiative is intended to improve adaptation to the country's gender identity law, *La Ley de Identidad de Género de la República Argentina*. Passed in May 2012, this law requires all citizens to be treated according to their chosen gender identity. For this purpose, various initiatives were carried out.

First, public management officials from various levels were convened for this particular workstream, including from:

• The Directorate of Health Information Systems of the Ministry of Health

Revising the Argentine version of the SNOMED-CT clinical vocabulary to reduce EHR-mediated violence

OXFORD
ACADEMIC

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

JAMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Article Navigation

Transphobia, encoded: an examination of trans-specific terminology in SNOMED CT and ICD-10-CM [Get access >](#)

A Ram , Clair A Kronk, Jacob R Eleazer, Joseph L Goulet, Cynthia A Brandt, Karen H Wang

Journal of the American Medical Informatics Association, Volume 29, Issue 2, February 2022, Pages 404–410,
<https://doi.org/10.1093/jamia/ocab200>

Published: 27 September 2021 Article history ▾

 Cite  Permissions  Share ▾

Abstract

Transgender people experience harassment, denial of services, and physical assault during healthcare visits. Electronic health record (EHR) structure and language can exacerbate the harm they experience by using transphobic terminology, emphasizing binary genders, and pathologizing transness. Here, we investigate the ways in which SNOMED CT and ICD-10-CM record gender-related terminology and explore their shortcomings as they contribute to this EHR-mediated violence. We discuss how this "standardized"

Impact of ML on Queer Communities

Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities

Nenad Tomasev

nenadt@deepmind.com

DeepMind

London, United Kingdom

Jackie Kay*

kayj@deepmind.com

DeepMind

London, United Kingdom

Kevin R. McKee

kevinrmckee@deepmind.com

DeepMind

London, United Kingdom

Shakir Mohamed

shakir@deepmind.com

DeepMind

London, United Kingdom

ABSTRACT

Advances in algorithmic fairness have largely omitted sexual orientation and gender identity. We explore queer concerns in privacy, censorship, language, online safety, health and employment to study the positive and negative effects of artificial intelligence on queer communities. These issues underscore the need for new directions in fairness research that take into account a multiplicity of considerations, from privacy preservation, context sensitivity and process fairness, to an awareness of sociotechnical impact and the increasingly important role of inclusive and participatory research processes. Most current approaches for algorithmic fairness assume that the target characteristics for fairness—frequently, race and legal gender—can be observed or recorded. Sexual orientation

1 INTRODUCTION

As the field of algorithmic fairness has matured, the ways in which machine learning researchers and developers operationalise approaches for fairness have expanded in scope and applicability. Fairness researchers have made important advances and demonstrated how the risks of algorithmic systems are imbalanced across different characteristics of the people who are analysed and affected by classifiers and decision-making systems [15, 57]. Progress has been particularly strong with respect to race and legal gender.¹ Fairness studies have helped to draw attention to racial bias in recidivism prediction [9], expose racial and gender bias in facial recognition [32], reduce gender bias in language processing [26, 124], and increase the accuracy and equity of decision making for child protective

¡Muchas gracias!

Fairness of machine learning in medical image analysis

Enzo Ferrante



eferrante@sinc.unl.edu.ar



[@enzoferrante](https://twitter.com/enzoferrante)



Research Institute for Signals, Systems and Computational Intelligence, sinc(i)

*Argentina's National Research Council (CONICET), Universidad Nacional del Litoral (UNL)
Santa Fe, Argentina*



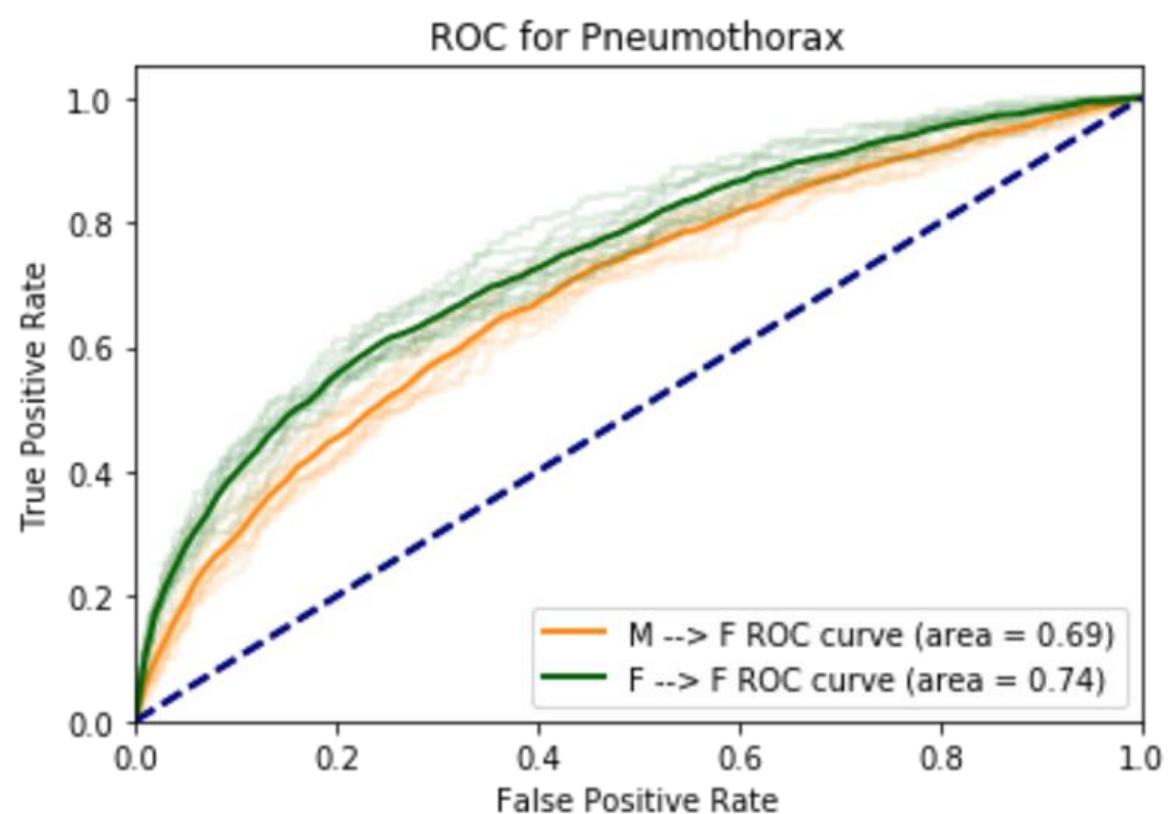
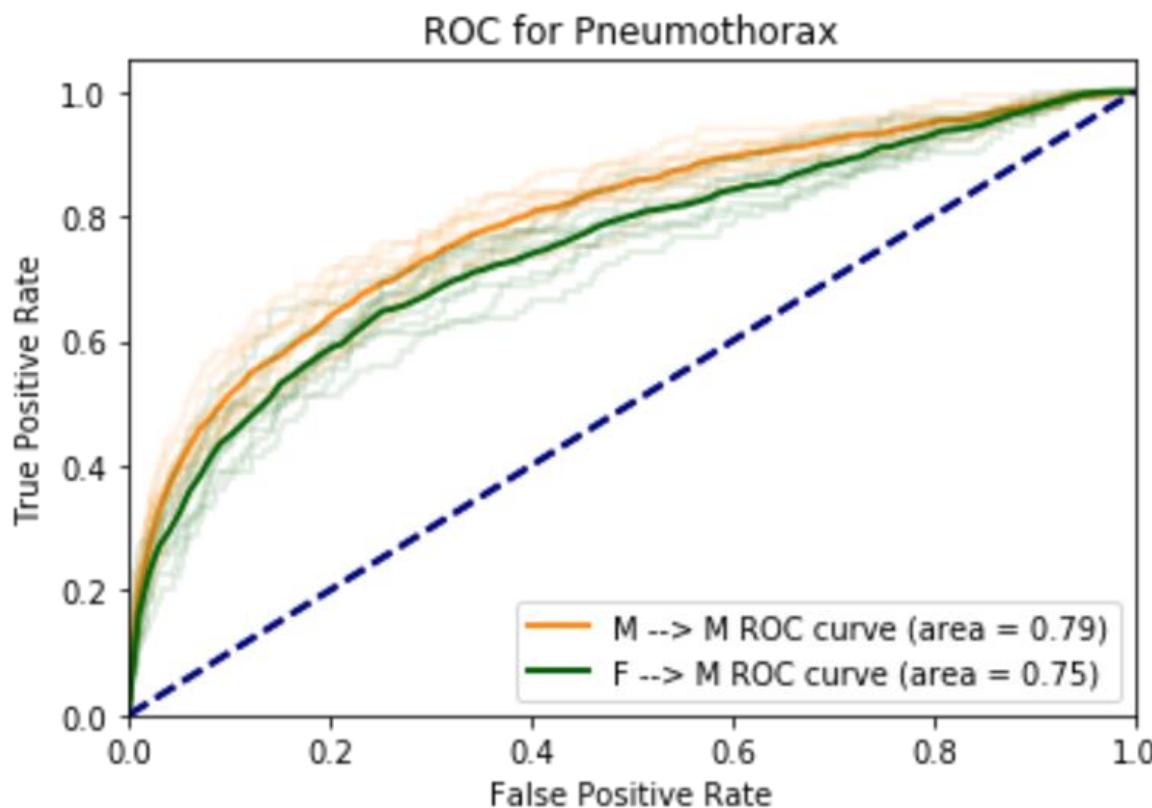
Argentinian Public Health Research
on Data Science and Artificial Intelligence
for Epidemic Prevention

Responsible Machine Learning in Healthcare Workshop

Copenhagen, Denmark – October 27th & 28th

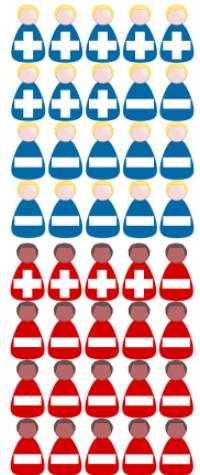


AUC example



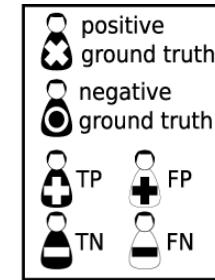
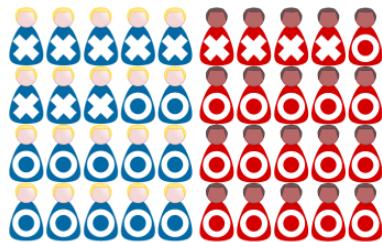
A model achieved
100% accuracy
on the test set

model predictions



Different disease prevalence
(40% and 20% for blue and red subjects respectively)

population ground truth



demographic parity

$$\frac{\# \text{ (blue)} + \# \text{ (blue)}}{\# \text{ (blue)} + \# \text{ (blue)} + \# \text{ (yellow)} + \# \text{ (yellow)}} = \frac{\# \text{ (red)} + \# \text{ (red)}}{\# \text{ (red)} + \# \text{ (red)} + \# \text{ (black)} + \# \text{ (black)}}$$

not fulfilled

equal opportunity

$$\frac{\# \text{ (boy)} + \# \text{ (girl)}}{\# \text{ (boy)} + \# \text{ (girl)}} = \frac{\# \text{ (boy)} + \# \text{ (girl)}}{\# \text{ (boy)} + \# \text{ (girl)}}$$

fulfilled

The model would not fulfill the ***demographic parity*** criterion since the **positive prediction rates vary between sub-groups**

40% (8 positive predictions over 20 cases)
for the blue sub-group vs.
20% (4 positive predictions over 20 cases)
For the red sub-group

The model would fulfill the ***equal opportunity*** criterion, as true positive rates match for both sub-groups reaching the value of 100%