MICCAI2024 Marrakesh MOROCCO

THE SUMMIT STUDY IT'S ABOUT TIME — DETECT EARLY. TREAT EARLY. LET'S TACKLE CANCER.

Satsuma Lung and Heart Imaging

lungsforliving Leading Lung Cancer Research

GRAIL

MICCAI

Exploring Fairness in State-of-the-Art Pulmonary Nodule Detection Algorithms

UCL

AUTHORS

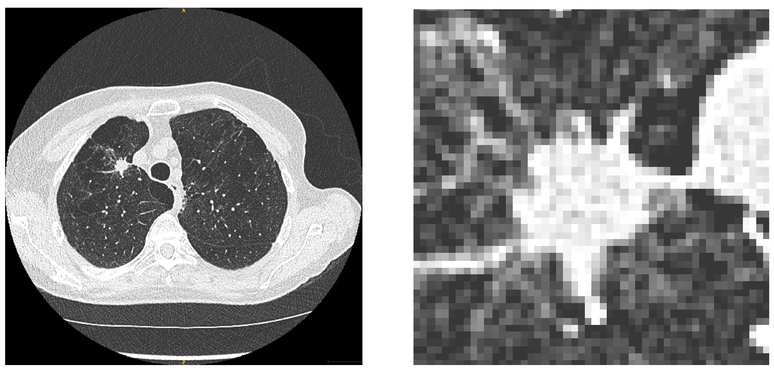
John McCabe[1], Daryl Cheng[1], Amyn Bhamani[2], Monica Mullin[2,7], Tanya Patrick[2], Arjun Nair[6], Sam M. Janes[2], Carole H. Sudre[3,4,5], and Joseph Jacob[1]

AFFILIATIONS

[1]Satsuma Lab, Centre for Medical Image Computing (CMIC), University College, London, United Kingdom
[2]Lungs for Living Research Centre, UCL Respiratory, University College London, United Kingdom
[3]Centre for Medical Image Computing (CMIC), University College London, United Kingdom
[4]MRC Unit for Lifelong Health and Ageing, Department of Population Science and Experimental Medicine, University College London
[5]Department of Biomedical Computing, School of Biomedical Engineering & Imaging Sciences, King's College London
[6]University College London Hospitals NHS Foundation Trust, London
[7]Department of Respirology, University of British Columbia, Vancouver, Canada

Introduction

- Low-Dose Computed Tomography (LDCT) plays a vital role in reducing lung cancer mortality in lung cancer screening.
- The goal of LDCT scanning is to detect clinically significant nodules.
- Radiologists use CADe systems to improve reporting efficiency.
- Training datasets are frequently unbalanced with regards to protected groups, such as sex and ethnic group.



Full LDCT scan slice (left) alongside an enlarged nodule (right). The aim of LDCT is to identify nodules, enabling early intervention in lung cancer.

Objectives

- Investigate performance variations across protected groups in nodule detection algorithms.
- Determine if any performance variation is a consequence of training with unbalanced datasets.
- Identify key factors driving algorithmic performance in detecting clinically significant lung nodules.

Methods

- **Design:** Comparative analysis in sex & ethnic group.
- **Dataset:**
  - 5,290 CT scans from lung cancer screening study (SUMMIT[1]).
  - Unequal representation of sex and ethnic groups.
- **Algorithms:** Two models (SOTA performance on LUNA16[2]); varied architecture/training regimes.
  - Model 1: Wining entry Kaggle Data Science Bowl 2017[3].
  - Model 2: MONAI Detection[4].
- **Metrics:** Competition Performance Metric (CPM, average sensitivity at 7 fixed operating points) for clinically relevant nodules. CI from 1000 bootstraps.
- **Experiments:**
  - Experiment 1: All data (Test data, balanced by ethnic group).
  - Experiment 2: Remove confounding factors for sex.
  - Experiment 3: Remove confounding factors for ethnic group.
  - Experiment 4: Comparison across nodule types and sizes.

Analysis

Exp. 1: Training on all data; Testing with balanced ethnic groups.

- *Both models: Female CPM (0.46 & 0.57) > Male CPM (0.38 & 0.49).
- *Both models: White CPM (0.52 & 0.60) > Asian or Asian British CPM (0.36 & 0.47) and Black CPM (0.36 & 0.52).

Exp. 2: Training on Male only; Testing with balanced ethnic groups.

- *Both models: Asian or Asian British CPM (0.48 & 0.58) > White CPM (0.43 & 0.53) and Black CPM (0.36 & 0.52).

Exp. 3: Training on White only; Testing with balanced sex.

- *Model 1: Male CPM = Female CPM (0.44).
- *Model 2: Female CPM (0.51) > Male CPM (0.49).

*Despite performance differences, CI overlaps between groups.

Experiment 4: Performance by Nodule Characteristics.

- Model 1: Larger nodules excel at lower false positive rates.
- Model 2: Smaller nodules excel at all false positive rates.
- Both models: Most prevalent nodule sub-types, best performing.

Results

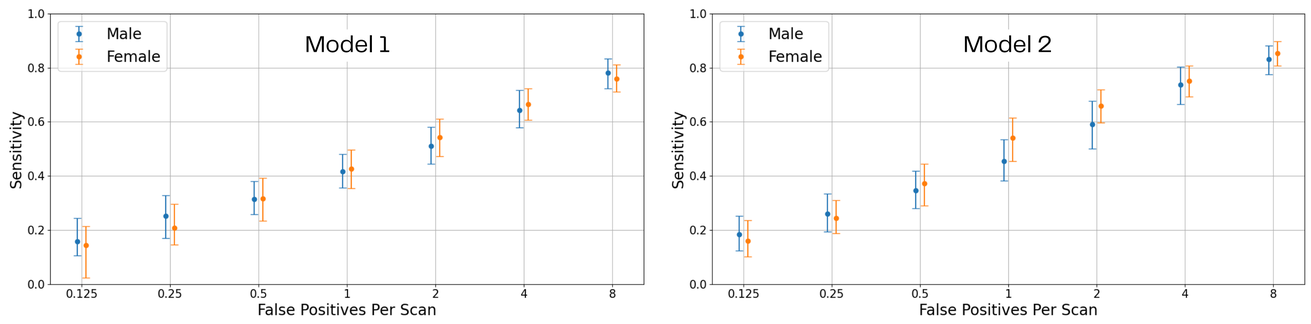
Exp. 1: Train. (85%), Valid. (5%), Test. (10%). Training data mirrors underlying SUMMIT sample.

Sensitivity bar plots (mean 95% CI) across sex (top row) and ethnic groups (bottom row)

Exp. 2: Training on Male only (ethnic group profile mirrors underlying SUMMIT sample).

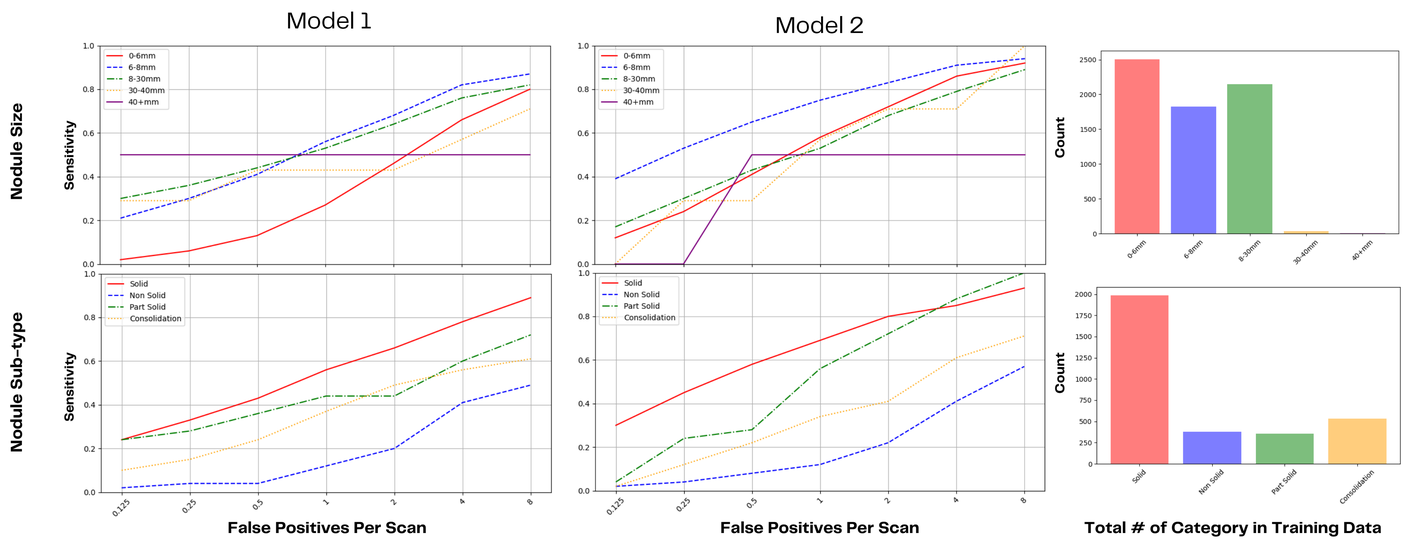
Sensitivity bar plots (mean 95% CI) across ethnic groups trained on a Male only sub-sample

Exp. 3: Training on White only (sex profile mirrors underlying SUMMIT sample)



Sensitivity bar plots (mean 95% CI) across sex trained on White participants only

Exp. 4: Performance by Nodule Characteristics. Data as per Exp. 1.



Sensitivity for each nodule category by size (top row) and nodule type (bottom row).
Training proportions are indicated in the bar plot on the right.

Conclusion

- There was no clear evidence that training on unbalanced datasets negatively impacts nodule detection rates in the protected groups, sex and ethnic group.
- Analysis revealed that common nodule sub-types are detected more easily than rarer ones, highlighting the need for careful design and training of pulmonary nodule detection algorithms to ensure equal detection across all nodule sub-types.
- It was shown that model design significantly affects the detection of different nodule sizes; since clinically relevant nodules are often larger, it may be beneficial to review existing metrics to incorporate severity.

References

1.
1. Horst, C., Dickson, J.L., Tisi, S., Ruparel, M., Nair, A., Devaraj, A., Janes, S.M.: Delivering low-dose CT screening for lung cancer: a pragmatic approach. Thorax. 75, 831–832 (2020).
2. LUNA16 - Grand Challenge, https://luna16.grand-challenge.org/.
3. Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. IEEE Trans. Neural Netw. Learning Syst. 30, 3484–3495 (2019). https://doi.org/10.1109/TNNLS.2019.2892409.
4. tutorials/detection/luna16_prepare_images.py at main · Project-MONAI/tutorials, https://github.com/Project-MONAI/tutorials/blob/main/detection/luna16_prepare_images.py, last accessed 2024/06/12.