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

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

[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] University College London Hospitals NHS Foundation Trust, London
[5] Department of Respirology, University of British Columbia, Vancouver, Canada

Abstract. Lung cancer is the leading cause of cancer mortality worldwide. Asymptomatic in its early stages, it is disproportionately detected when the disease is advanced. Resource constraints have resulted in increasing reliance on computer-aided detection (CADe) systems to assist with scan evaluation. The datasets used to train these algorithms are often unbalanced in their representation of protected groups e.g. sex and ethnicity. This project investigates whether there are performance disparities in detecting clinically relevant nodules across under-represented groups in selected, state-of-the-art nodule detection algorithms trained on data from a screening program in the UK.

Our analysis revealed that overall, the algorithms demonstrate equitable performance across various demographic groups. However, their performance varies strongly across nodule characteristics (size and type) in line with their prevalence in the training set. To ensure continued equitable performance, algorithms should not only consider demographic but also nodule attributes representativeness in their training.

Keywords: Nodule Detection Algorithms · Fairness in AI · Lung Cancer Screening.

1 Introduction

1.1 Background

Lung cancer is responsible for 20% of all cancer deaths, the largest number of cancer deaths worldwide [1]. The main reason for this is that lung cancer is often asymptomatic during the early stages. When it becomes symptomatic, the disease has invariably spread, preventing curative treatment and reducing survival. However, when lung cancer is detected in its early stages, effective treatments are increasingly emerging, improving survival rates. Five-year survival rates range between 10% and 36% for patients diagnosed at stages III and IV, but approach 53% to 92% in patients diagnosed at stages I and II [1].

A series of Randomized Control Trials (RCTs) [2, 3] have demonstrated the effectiveness of using Low Dose Computed Tomography (LDCT) to identify lung cancer in high-risk populations, resulting in 20 to 24% reductions in lung cancer mortality. The main goal of an LDCT scan is to identify lung nodules. Lung nodules consist of a collection of cells, which together form a lesion large enough to be visualised at the spatial resolution of a LDCT scan. Some nodules may represent the early stages of a lung cancer. Accordingly, the most important task when assessing nodules on LDCT is to confidently detect nodules that are likely to harbour a cancer. These nodules require more detailed clinical investigation and are often termed actionable nodules. Detecting actionable nodules is a complex task as the majority of lung nodules are benign and do not need further clinical work-up. Similarly, many structures in the lung may mimic the appearance of nodules. The Fleischner Society has introduced terminology to standardize the descriptions and reporting of lung nodules [4], and also provides guidelines for follow-up and management of these nodules.

Alongside the previously described trials, numerous studies have been carried out to refine the lung screening process and assess the economic impact of implementing a screening program. As a result, the UK National Screening Committee has recommended targeted screening throughout all four UK nations, with plans for a complete roll out by 2029. This will result in a significant increase in demand for LDCT scans reporting. There is an existing shortage of qualified radiologists to carry out this work, with a projected shortfall of 39% by 2026 [5]. The pressures on radiological resources are likely to result in reporting delays and an overworked workforce, which will impact patient care, standards of practice and patient outcomes.

Artificial intelligence and machine learning (AI/ML) are being considered as a potential solution to these workforce challenges. Deep learning (DL), a subset of AI, has been at the forefront of the most significant developments in healthcare research in recent years. DL models have been successfully introduced into several healthcare settings such as a computer-aided detection (CADe) systems [6–8], drug discovery [9] and robotic surgery [10].

Multiple studies have demonstrated challenges related to the generalizability and bias in the development of DL models [11–13]. This includes the potential for bias in various medical imaging tasks, such as classification and segmentation, which are widely used used across different medical domains and diseases [14–16]. In the field of lung nodule detection, several commercial products have been licensed to act as second readers for lung cancer screening, assisting radiologists in interpreting lung nodules on LDCT scans. However, the specifics of the design and training of these commercial AI models are not publicly shared in order to protect intellectual property. These algorithms are likely to have been trained on publicly available nodule datasets. However, these datasets, including those from the aforementioned RCTs, have been shown to lack diversity (for example, the NELSON trial had a five to one male to female ratio).

The publicly available datasets utilized for developing nodule detection algorithms often consist of LDCT scans that are over a decade, and occasionally two decades old. These datasets may exhibit inconsistencies, particularly in the annotation of nodule types and sizes which encompass both actionable and non-actionable nodules.

There are three main aims of this project, first to understand whether there is any variation in performance across protected groups in current nodule detection algorithms. Secondly, if there are discrepancies, does this result from training with unbalanced datasets. Thirdly, can we determine what the main drivers of performance are within nodule detection algorithms.

2 Methods and Materials

2.1 Study Design

This study employed a comparative analysis approach to evaluate performance variation in nodule detection algorithms across sex and ethnic group. Two state-of-the-art nodule detection algorithms were chosen for assessment based on availability, diversity in architecture and performance. The evaluation was conducted using a dataset drawn from the SUMMIT study, one of Europe's largest lung cancer screening programs.

2.2 Dataset Description

The SUMMIT cohort is a London based lung screening study involving high-risk participants invited from primary care. Participants were risk assessed, and qualifying participants were consented to undergo an LDCT scan and provide indication of sex and ethnic group. For this analysis, a subset of 5,290 baseline LDCT scans was utilized. These scans, which were the ones available at the start of the project, were randomly selected and closely mirror the demographic composition of the overall SUMMIT cohort. The SUMMIT cohort itself exhibits an imbalance, with men slightly outnumbering women and the 'White' ethnic group significantly outnumbering all other ethnicity's in the sample. Additionally, a substantial sex imbalance was observed within 'Asian or Asian British' and 'Black' ethnic groups, with males greatly outnumbering females.

Each LDCT was reported by a specialist pulmonary radiologist, supported by the Mevis'^{TM} Veolity CADe software [6]. In addition to the location of the nodule, other characteristics including maximum diameter and type (solid, part-solid, non-solid, consolidation) were recorded.

2.3 Nodule Detection Algorithms

Two object detection open-source frameworks with different backbones and pre-processing and having demonstrated state-of-the-art performance on the LUNA16 [17] nodule detection challenge were considered to assess the impact of model architecture on the study findings.

Model 1 is the winning entry in the Kaggle Data Science Bowl 2017 [18]. This model consists of a one-stage object detector utilising a modified UNet [19] architecture as backbone with approximately 5 million trainable parameters. The pre-processing include a lung segmentation step, resampling and intensity clipping (-1200 to 300 Hounsfield Units (HU)). Following the original training strategy used to develop this algorithm, nodules with a diameter greater than 30 mm were oversampled during training

Model 2 is the MONAI detection algorithm which utilizes a RetinaNet [20] with approximately 21 million trainable parameters. This uses a feature pyramid network [21] to enable detection at different scales and a Focal Loss [22] to deal with class imbalance. Pre-processing includes a resampling and a clipping step (-1024 to 300 HU).

Both models use hard-negative mining [23] and a patch-based training method (128^3 and 198x198x80 for Model 1 and 2 respectively).

2.4 Evaluation Metrics

Given their implication for patient management, evaluation focused on detection performance for actionable nodules, i.e nodules for which specific follow-up is required. Performance was assessed using the Free-Response Receiver Operating Characteristic (FROC) curves, which measures sensitivity over 7 fixed false positives per scan operating points ($\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, 8) and and a Competition Performance Metric (CPM), which is the average sensitivity over these operating points. Bootstrap (1000 samples) was used to obtain confidence intervals.

2.5 Experiments

Experiments were conducted to evaluate the impact of training set imbalance as observed in the make-up of the screening cohort on selected nodule detection algorithms with a focus on differences across sex and ethnic group. For comparative purposes and across all experiments, all test groups were balanced and the training set was built to mimic the distribution of the screening sample. Due to their low number and their heterogeneity, "Other" and "Mixed" ethnic groups were only considered in the training set. Experiments can be described as follows

- **Experiment 1** - Maximisation of training set size given minimum test set ethnic group size.
- **Experiment 2** Isolation of the impact of ethnic group on performance without confounding for sex by training and testing on male samples only.
- **Experiment 3** Isolation of the impact of sex imbalance on performance without confounding for ethnicity by training and testing on white only.
- **Experiment 4** Comparison of performance across nodule types and sizes to understand drivers of performance using settings of Experiment 1.

Table 1 shows the composition of experiments 1-3. The proportions of sex and ethnic group for the whole SUMMIT sample are shown in the end column.

Table 1: Profile of protected groups for Training, Validation and Test datasets used for each experiment

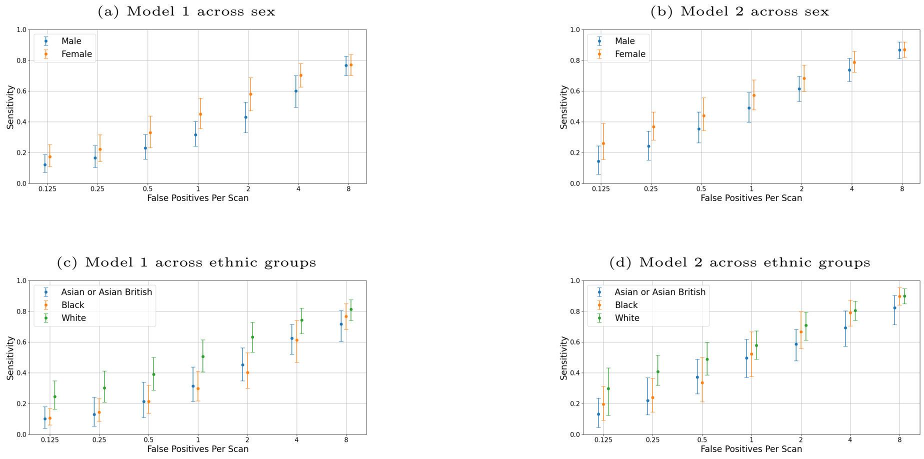
Protected Group	Category	Experiment 1			Experiment 2 - Male only			Experiment 3 - White only			SUMMIT
		Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Total
Sex	Female	1961 (38.6%)	125 (46.8%)	250 (42.1%)	0 (0%)	0 (0%)	0 (0%)	1494 (44.4%)	98 (43.8%)	399 (50.0%)	5508 (42.5%)
	Male	3118 (61.4%)	142 (53.2%)	344 (57.9%)	1573 (100.0%)	105 (100.0%)	420 (100.0%)	1870 (55.6%)	126 (56.2%)	399 (50.0%)	7450 (57.5%)
Ethnic Group	Asian or Asian British	443 (8.7%)	25 (9.4%)	198 (33.3%)	98 (6.2%)	11 (10.5%)	140 (33.3%)	0 (0%)	0 (0%)	0 (0%)	845 (6.5%)
	Black	244 (4.8%)	12 (4.5%)	198 (33.3%)	71 (4.5%)	4 (3.8%)	140 (33.3%)	0 (0%)	0 (0%)	0 (0%)	577 (4.5%)
	Mixed	119 (2.3%)	9 (3.4%)	0 (0%)	34 (2.2%)	2 (1.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	283 (2.2%)
	Other ethnic groups	199 (3.9%)	7 (2.6%)	0 (0%)	53 (3.4%)	4 (3.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	451 (3.5%)
	White	4074 (80.2%)	214 (80.1%)	198 (33.3%)	1317 (83.7%)	84 (80.0%)	140 (33.3%)	3364 (100.0%)	224 (100.0%)	798 (100.0%)	10802 (83.4%)
Total		5079	267	594	1573	105	420	3364	224	798	12958

3 Results

Results for experiments 1-3 are presented as bar plots of the mean sensitivity and 95% confidence interval at the seven fixed operating points across protected groups for each model.

3.1 Experiment 1: Test dataset with balanced ethnic groups

The results for Experiment 1 are shown in Fig. 1. The first row (Fig. 1a and Fig. 1b) presents the outcomes from Model 1. The plots suggest similar performance between male and female participants, although female participants show a higher CPM of 0.46 (95% CI 0.38-0.55) compared to 0.38 (95% CI 0.30-0.46) in male participants.


(a) Model 1 across sex

(b) Model 2 across sex

(c) Model 1 across ethnic groups

(d) Model 2 across ethnic groups

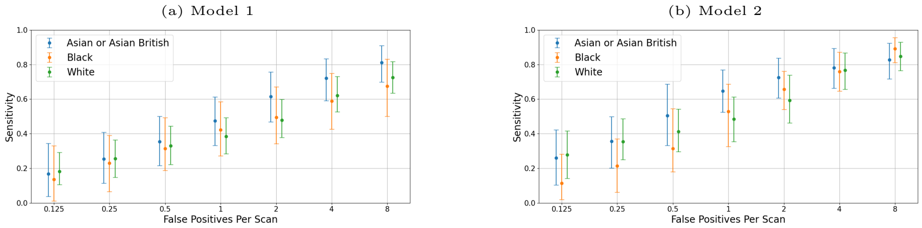
Fig. 1: Sensitivity bar plots (mean 95% CI) across sex (top row) and ethnic groups (bottom row) for the two models trained on SUMMIT dataset.

When comparing ethnic groups in Model 1, 'White' participants exhibit a higher CPM of 0.52 (95% CI 0.43-0.61) compared to 'Asian or Asian British' and 'Black' participants, who show CPMs of 0.36 (95% CI 0.27-0.47) and 0.36 (95% CI 0.28-0.46) respectively.

The results from Model 2, depicted in the second row (Fig. 1c and Fig. 1d), generally indicate better performance across all categories compared to Model 1. A similar pattern is observed where females in Model 2 having a higher CPM of 0.57 (95% CI 0.49-0.66) compared to males, who show a CPM of 0.49 (95% CI 0.41-0.58). Additionally, 'White' participants in Model 2 achieve a CPM of 0.60 (95% CI 0.50-0.69), which is better than 'Asian or Asian British' participants with a CPM of 0.47 (95% CI 0.36-0.59) and 'Black' participants with a CPM of 0.52 (95% CI 0.42-0.64).

3.2 Experiment 2: Male Only

Fig. 2 shows the results for each ethnic group when trained on a male-only sample. For both Model 1 and Model 2, performance across ethnic groups exhibits minimal variation. In Model 1, the CPM varies across different ethnic groups: White participants exhibit a Mean Sensitivity of 0.43 (95% CI 0.33-0.54), 'Asian or Asian British' participants show a Mean Sensitivity of 0.48 (95% CI 0.35-0.62), and 'Black' participants demonstrate a Mean Sensitivity of 0.41 (95% CI 0.26-0.578). For Model 2, the CPMs are as follows: White participants have a Mean Sensitivity of 0.53 (95% CI 0.42-0.653), 'Asian or Asian British' participants show a Mean Sensitivity of 0.58 (95% CI 0.45-0.719), and 'Black' participants exhibit a Mean Sensitivity of 0.49 (95% CI 0.37-0.627). It should be noted that the mean sensitivity shifts between these groups across various operating points for both models and interestingly, the 'Asian or Asian British' participants, who are under-represented in the sample perform marginally better for both models.

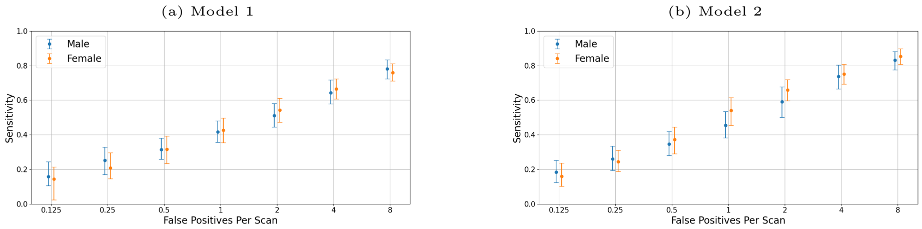

(a) Model 1

(b) Model 2

Fig. 2: Sensitivity bar plots (mean 95% CI) across ethnic groups at fixed operating points for models trained on a male-only sub-sample.

3.3 Experiment 3: White Only

The results for Experiment 3, as shown in Fig. 3, indicate that contrary to Experiment 1, where females had a higher CPM, when trained exclusively on white participants there are only very small differences in CPM. Model 1 shows that male participants have a CPM of 0.44 (95% CI 0.38-0.509), while female participants exhibit a CPM of 0.44 (95% CI 0.37-0.51). For Model 2 in Experiment 3, male participants demonstrate a CPM of 0.49 (95% CI 0.42-0.557), whereas female participants show a Mean Sensitivity of 0.51 (95% CI 0.45-0.574).


(a) Model 1

(b) Model 2

Fig. 3: Sensitivity bar plots (mean 95% CI) across sex at fixed operating points for models trained on White participants only

3.4 Experiment 4:

The performance across each nodule size and type at the different operating points for the two models trained on data from Experiment 1 is presented in Fig. 4. Regarding the diameter plots, the contrasting algorithm designs are apparent: Model 1, employing over-sampling for nodules sized 30-40mm and 40+mm, detects larger nodules at lower operating points. In contrast, Model 2 detects a greater proportion of smaller nodules at lower operating points, possibly benefiting from the scale robustness provided by the FPN. Regarding the nodule types, Model 1 shows earlier detection of part-solid nodule types compared to Model 2. Both models demonstrate similarly higher performance in detecting the most frequently found nodules in the training dataset (solid nodules) and lower performance in detecting less common nodules (non-solid and consolidation types)

4 Discussion

In this study, the variation in performance across sex and ethnic groups was evaluated in two state-of-the-art pulmonary nodule detection algorithms trained on an unbalanced dataset drawn from a large screening program. There was no indication of adverse impact of unbalanced demographic representation on nodule detection performance when addressing confounding factors.

When the prevalence of different nodule types were evaluated in the training dataset marked differences in actionable nodule detection were uncovered. The most common nodule subtypes were generally better detected than their rarer

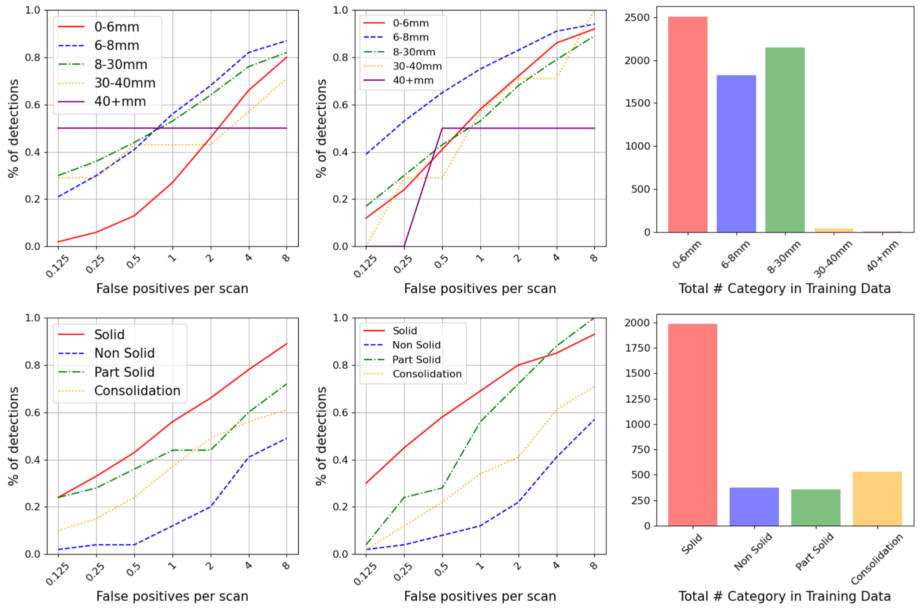


Fig. 4: Sensitivity for each nodule category by size (top row) and nodule type (bottom row) for Model 1 (left) and Model 2 (right) at the fixed operating points. Training proportions are indicated in the bar plot on the right.

counterparts. The one exception to this was for part-solid nodules. Though a rare nodule subtype, algorithm performance for detecting part-solid nodules was notably high, potentially due to shared features in a part-solid nodule with the much more commonly found solid nodule. As nodule type prevalence varies across different lung cancer screening populations, there may be a class imbalance in training datasets fed to an AI algorithm. These algorithms may then perform poorly when detecting certain important nodule subtypes. It is therefore essential to design robust training strategies that take account of and cohort enrich training datasets for the various important and potentially under-represented nodule classes. Such strategies should prevent potential cancer being missed by an AI algorithm. Model design appeared to be an influential factor determining algorithm performance with respect to the size of the nodules. Such findings also underline the need to carefully reconsider the clinical relevance of the evaluation metrics used for comparison. In practice, detection at low operating points for nodules with a size above 8mm may be a more appropriate focus of attention.

This study is limited in its generalizability by the implications of the screening setting limited to a single scanner type with a consistent protocol that is not representative of clinical setting. Future research should expand the understanding of performance difference across nodule characteristics (e.g location), extend to not screened populations (e.g non-smokers) and elaborate on metrics more suitable for clinical application (e.g focus on low operating points).

In conclusion, our analysis demonstrates reassuring results regarding the overall fairness of existing nodule detection algorithms. However, it also underscores the need for continued efforts to ensure sustained performance not only across diverse demographic populations but also nodule subtypes presentations. It challenges the one-size-fits-all approaches in nodule detection algorithms and promotes the need for the development of solutions tailored to nodule characteristics.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. LNCS, Cancer Statistics for the UK [WWW Document], 2015. Cancer Research UK. URL https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk(accessed 3.23.23).
2. Author, Team, N.L.S.T.R., 2011. The National Lung Screening Trial: Overview and Study Design. Radiology 258, 243. https://doi.org/10.1148/radiol.10091808
3. Author, 1. de Koning, H.J., van der Aalst, C.M., de Jong, P.A., Scholten, E.T., Nackaerts, K., Heuvelmans, M.A., et al, 2020. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. New England Journal of Medicine 382, 503–513. https://doi.org/10.1056/NEJMoa1911793
4. Author, 1. Bankier, A.A., MacMahon, H., Colby, T., Gevenois, P.A., Goo, J.M., Leung, A.N.C., Lynch, D.A., Schaefer-Prokop, C.M., Tomiyama, N., Travis, W.D., Verschakelen, J.A., White, C.S., Naidich, D.P.: Fleischner Society: Glossary of Terms for Thoracic Imaging. Radiology. 310, e232558 (2024). https://doi.org/10.1148/radiol.232558.
5. LNCS, RCR Clinical radiology census report 2021 | The Royal College of Radiologists [WWW Document], n.d. URL https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology- census-report-2021 (accessed 3.23.23).
6. LNCS, Veolity - a brand of MeVis Medical Solutions AG: Product Information [WWW Document], n.d. URL https://www.veolity.com/about-veolity/product-information (accessed 3.24.23)
7. LNCS, Veye Lung Nodules [WWW Document], n.d. . Aidence. URL https://www.aidence.com/veye-lung-nodules/ (accessed 3.24.23).
8. LNCS, AI-Rad Companion [WWW Document], n.d. URL https://www.siemenshealthineers.com/en-uk/digital-health-solutions/ai-rad-companion (accessed 3.24.23).
9. Author, 1. Yang, X., Wang, Y., Byrne, R., Schneider, G., Yang, S., 2019. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem. Rev. 119, 10520–10594. https://doi.org/10.1021/acs.chemrev.8b00728
10. Author, 1. Beyaz, S., 2020. A brief history of artificial intelligence and robotic surgery in orthopedics & traumatology and future expectations. Jt Dis Relat Surg 31, 653–655. https://doi.org/10.5606/ehc.2020.75300

11. Author, 1. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 77–91. PMLR (2018).
12. Author, 1. Wang, M., Deng, W.: Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning, http://arxiv.org/abs/1911.10692, (2019). https://doi.org/10.48550/arXiv.1911.10692.
13. Author, 1. Brandao, M.: Age and gender bias in pedestrian detection algorithms, http://arxiv.org/abs/1906.10490, (2019). https://doi.org/10.48550/arXiv.1906.10490.
14. Author, 1. Puyol-Anton, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation, http://arxiv.org/abs/2106.12387, (2021).
15. Author, 1. Weng, N., Bigdeli, S., Petersen, E., Feragen, A.: Are Sex-based Physiological Differences the Cause of Gender Bias for Chest X-ray Diagnosis? Presented at the (2023). https://doi.org/10.48550/ARXIV.2308.05129.
16. Author, 1. Burlina, P., Joshi, N., Paul, W., Pacheco, K.D., Bressler, N.M.: Addressing Artificial Intelligence Bias in Retinal Diagnostics. Translational Vision Science & Technology. 10, 13 (2021). https://doi.org/10.1167/tvst.10.2.13.
17. LNCS, 1. LUNA16 - Grand Challenge, https://luna16.grand-challenge.org/Data/, last accessed 2024/04/18.
18. Author, 1. 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.
19. Author, 1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, http://arxiv.org/abs/1505.04597, (2015). https://doi.org/10.48550/arXiv.1505.04597.
20. Author, 1. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection, http://arxiv.org/abs/1708.02002, (2018). https://doi.org/10.48550/arXiv.1708.02002.
21. Author, 1. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection, http://arxiv.org/abs/1612.03144, (2017). https://doi.org/10.48550/arXiv.1612.03144.
22. Author, 1. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection, http://arxiv.org/abs/1708.02002, (2018). https://doi.org/10.48550/arXiv.1708.02002.
23. Author, 1. Bucher, M., Herbin, S., Jurie, F.: Hard Negative Mining for Metric Learning Based Zero-Shot Classification, http://arxiv.org/abs/1608.07441, (2016). https://doi.org/10.48550/arXiv.1608.07441.
24. Author, 1. Yang, P.: PS01.02 National Lung Cancer Screening Program in Taiwan: The TALENT Study. Journal of Thoracic Oncology. 16, S58 (2021). https://doi.org/10.1016/j.jtho.2021.01.318.