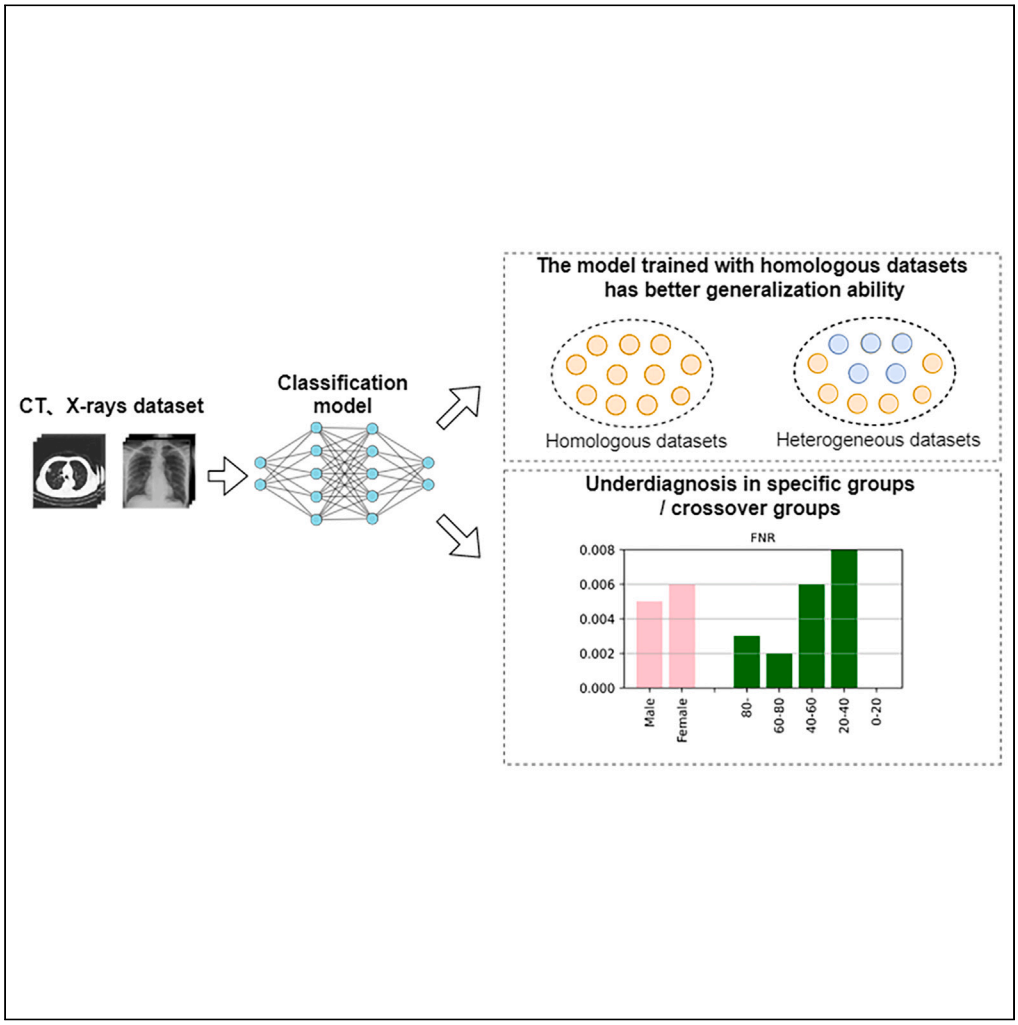
iScience

CellPress
OPEN ACCESS

Article

Reinvestigating the performance of artificial intelligence classification algorithms on COVID-19 X-Ray and CT images



Rui Cao, Yanan Liu, Xin Wen, Caiqing Liao, Xin Wang, Yuan Gao, Tao Tan

taotanjs@gmail.com

Highlights

An AI underdiagnosis bias discrimination pipeline for COVID-19 is proposed

The problem and reason of underdiagnosis bias in AI algorithms are discussed

COVID-19 AI system with heterogeneous samples can lead to poor model generalization

Cao et al., iScience 27, 109712
May 17, 2024
https://doi.org/10.1016/j.isci.2024.109712

iScience

CellPress
OPEN ACCESS

Article

Reinvestigating the performance of artificial intelligence classification algorithms on COVID-19 X-Ray and CT images

Rui Cao,[1] Yanan Liu,[1] Xin Wen,[1] Caiqing Liao,[1] Xin Wang,[2,3,4] Yuan Gao,[2,4] and Tao Tan[2,3,5,6,*]

SUMMARY

There are concerns that artificial intelligence (AI) algorithms may create underdiagnosis bias by mislabeling patient individuals with certain attributes (e.g., female and young) as healthy. Addressing this bias is crucial given the urgent need for AI diagnostics facing rapidly spreading infectious diseases like COVID-19. We find the prevalent AI diagnostic models show an underdiagnosis rate among specific patient populations, and the underdiagnosis rate is higher in some intersectional specific patient populations (for example, females aged 20–40 years). Additionally, we find training AI models on heterogeneous datasets (positive and negative samples from different datasets) may lead to poor model generalization. The model's classification performance varies significantly across test sets, with the accuracy of the better performance being over 40% higher than that of the poor performance. In conclusion, we developed an AI bias analysis pipeline to help researchers recognize and address biases that impact medical equality and ethics.

INTRODUCTION

Since the emergence of the coronavirus disease 2019 (COVID-19) outbreak in late 2019, the viral infection has exhibited a remarkably high level of transmissibility. This unprecedented incident has attracted profound attentions across various sectors of society. During the initial stages of the outbreak, the implementation of reverse-transcription polymerase chain reaction (RT-PCR) technology in clinical settings proved to be effective in the disease identification and the remission of further transmission, but it has the limitations of low sensitivity[1] and easy contamination.[2] On the other hand, artificial intelligence (AI)-assisted diagnosis technology demonstrates the ability to identify and capture imaging characteristics such as ground glass and solid pulmonary opacity in COVID-19 cases.[3,4] In addition, AI can also reduce the shortage of radiologists who have experience on this emergent disease and reduce their work burden.[3,5] Under the urgent demand, relevant AI-assisted diagnosis classification models have mushroomed, with good classification performance, including classification model based on computed tomography (CT) images,[6,7] classification model based on X-rays,[8,9] and classification model based on CT/X-ray two modes,[10] etc.

Although AI systems are expected to improve diagnosis and prognostic decision support for diseases, we have found that in other areas of disease diagnosis, AI systems may be biased and discriminatory across different social population groups, according to the various studies. One hospital created an AI model that included clinical and social variables to predict patient discharge times but then realized that doing so would favor affluent white patients over poorer African Americans.[11] Obermeyer et al. found racial bias in one of the more commonly used diabetes detection algorithms.[12] In cardiology, heart attacks are overwhelmingly misdiagnosed in women, a phenomenon discovered by Nancy et al.[13] Brendan et al. investigated the bias of facial recognition algorithms for different participants. The facial recognition system used for law enforcement performed differently to different subjects; for example, the performance of black, female, and young subjects was worse than that of other subjects. Furthermore, the study observed a gradual escalation of this phenomenon over time.[14] There are some studies of bias in other disease domains, but there is a lack of generalized bias studies in disease domains.

This bias results in the false negative, potentially leading to a lower priority for treatment when it is most crucial. The failure to receive timely treatment can have severe medical implications, surpassing the significance of the misdiagnosis rate, which refers to falsely identifying a patient as ill.[15] Although the pandemic has been suppressed, a retrospective summary of the widespread biases in previous studies is an important warning for us to avoid similar problems in the future, such as health inequalities for certain subgroups, and to better help us cope with the corresponding challenges[16] in the future.[17,18] However, in the specific context of COVID-19, the comprehensive validation of the bias in AI

[1]School of Software, Taiyuan University of Technology, Taiyuan 030024, China
[2]Department of Radiology, Netherlands Cancer Institute (NKI), Plesmanlaan 121, Amsterdam 1066 CX, the Netherlands
[3]Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands
[4]GROW School for Oncology and Development Biology, Maastricht University, MD, Maastricht 6200, the Netherlands
[5]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, China
[6]Lead contact
*Correspondence: taotanjs@gmail.com
https://doi.org/10.1016/j.isci.2024.109712

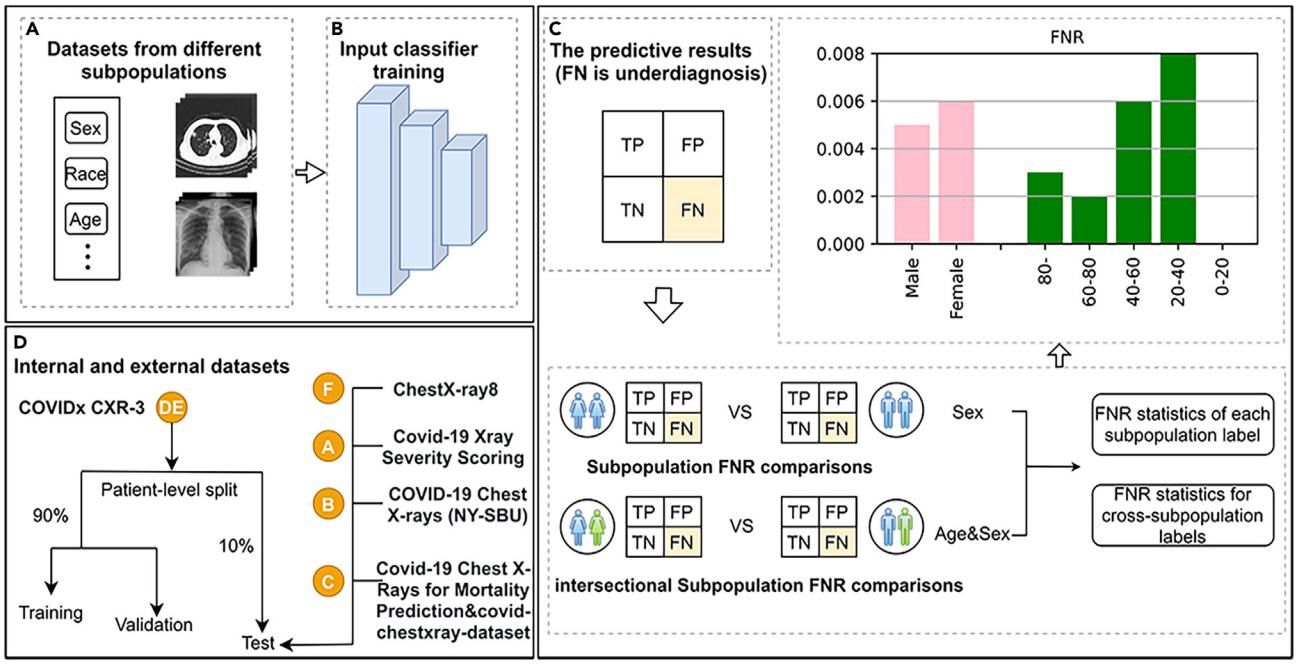


Figure 1. Flowchart diagram of the proposed pipeline

(A and B) The model is trained using the COVID-19 X-rays and CT datasets.

(C) The underdiagnosis rate (FNR, that is the false-negative rate of the COVID-positive label) of this model is then compared in different subpopulations and cross subpopulations(such as sex, age.) to examine the algorithm's underdiagnosis rate. TP, true positive; FP, false positive; TN, true negative. Symbol colors indicate different ages of male and female patients.

(D) Using internal and external datasets for proof training models with heterogeneous datasets can lead to problems with poor model generalization.

diagnostic classification algorithms is less explored. Furthermore, we found that most of the datasets used to train AI classifiers were heterologous datasets, and the accuracy of the classifiers was generally high,[19–24] while, the generalization ability of the model trained on such heterogeneous datasets still needs to be validated. We need to look at a series of issues that have been overlooked in existing AI model articles in the field of COVID-19, such as underdiagnosis bias, model generalization ability, and AI universality.

Our research team has developed a generalized AI bias discrimination pipeline. We applied it to the field of COVID-19 to identify and analyze underdiagnosis bias in AI classification models. This can help researchers conduct more systematic in-depth research and provide technical support for solving these problems. This pipeline can also be used in other disease areas to help researchers further their analysis. The model pipeline is shown in Figure 1.

The contributions of this paper are as follows.

(1) This paper proposes a generalized AI bias discrimination pipeline and applies it on the two mainstream imaging modalities of COVID-19, X-ray and CT, to verify its performance.
(2) This paper points out the problem of insufficient diagnostic bias of AI algorithms in the field of COVID-19 diagnosis and explores its causes.
(3) This paper raises the issue that the COVID-19 AI systems that have emerged in the past three years can lead to poor model generalization when using heterogeneous samples (positive and negative samples from different sources).

The rest of the paper is organized as follows. Chapter 2 introduces the experimental results. Chapter 3 discusses the results. Chapter 4 is "STAR methods"; key resources table, resource availability, experimental model and study participant details, method details, and quantification and statistical analysis are shown in turn. At the end of the article are five chapters: supplemental information, funding, author contributions, declaration of interests, and references.

RESULTS

Our study design

In this study, five COVID-19 diagnostic classifiers (see method details – experimental design section) were trained on three X-ray datasets and one CT dataset. These four datasets are detailed in method details – datasets section. The area under the receiver operating characteristic curve (AUC) and accuracy (ACC) metrics were used to demonstrate the model's performance in the entire population and subgroups. Then, according to the classification results, the underdiagnosis rates (according to Seyyed-Kalantari's study,[15] the false-negative rate [FNR]

Table 1. X-ray dataset

2D-Model	AUC ± 95% CI	ACC ± 95% CI
ResNet18	0.999 ± 0.001	0.993 ± 0.001
ResNet34	0.999 ± 0.001	0.995 ± 0.001
ResNet50	0.999 ± 0.001	0.996 ± 0.001
DenseNet121	0.999 ± 0.001	0.996 ± 0.001
DenseNet169	0.999 ± 0.001	0.996 ± 0.001

The performance indicators of the two-dimensional classifier for X-ray dataset.
The models trained on X-ray-dataset uses the same training-validation-test split, resulting in a calculated reported AUC ± 95% confidence interval (CI) and the ACC ± 95% confidence interval (CI).

predicted by the binarized model with "positive" label was used to represent the underdiagnosis rate) of subpopulations and cross subpopulations in the general population were compared, and the model decision bias was evaluated. We also discuss the problem that training models with heterogeneous datasets can lead to poor model generalization.

Underdiagnosis in specific subpopulations of patients

The performance results of the classifier trained using the X-ray dataset and the CT dataset are shown in Tables 1 and 2. The data show good classification ability.

The AUC values of the five models in Table 1 are all close to 1, and the ACC values are also unrealistically high. DenseNet has achieved a very high accuracy of 99.60%. The performance of the five models for X-ray image classification is nearly perfect.

The five models in Table 2 all achieved good results in CT image classification performance, among which shufflenetv2 obtained an AUC of 0.87 and an accuracy of 87.32%. Although its AUC value was the highest, its accuracy was 1.09% lower than that of 3D-ResNet18.

Then, based on the classification results, we calculated the underdiagnosis rates. We found that underdiagnosis rates differed across subpopulations. In Figures 2 and S5–S7, we show the underdiagnosis of subpopulations specificity of X-ray dataset/X-ray-Test2/X-ray-Test3/CT dataset in terms of sex, age, other comorbidities, and country, respectively. We observed higher rates of algorithmic underdiagnosis in female patients, patients aged 20–40, Iranian patients, and patients with malignancies, cancer, or other lung diseases than in other populations. These subpopulations are less likely to receive timely treatment by relying on these AI models. In addition, the underdiagnosis rate of COVID-19 patients who had never or previously smoked was higher than that of current smokers (Figure S5A). High or low blood pressure does not affect the underdiagnosis of COVID-19 patients (Figure S5A). The incidence of underdiagnosis is higher in patients with type I or type II diabetes (Figure S6A). The results of hypertension, diabetes, and smoking are subject to further discussion. We found consistent patterns of bias in the X-ray, X-ray-Test3, and CT datasets (i.e., women and younger patients had the highest rates of underdiagnosis). However, in the X-ray-Test2 dataset, the rates of underdiagnosis were the same for male and female patients, and the rates were the same for patients aged 60–80, 80-, and 0–60. The results presented by the gender attribute may be due to the unbalanced ratio of male to female patients (3:1), and the results presented by the age attribute may be due to the limitations of the dataset itself. Its age distribution (80-/60-80/0-60) is inconsistent with the distribution of other datasets (0–20, 20–40, 40–60, 60–80, 80-), among which the subset of 80- has a large sample size, accounting for 17%, compared with less than 10% of other datasets.

Underdiagnosis in cross subpopulations of patients

We studied the cross subpopulation, defined here as patients belonging to two subpopulations, such as Iranian female patients. We found that cross subpopulations (Figures 2B, S5B–S5G, S6B–S5F, and S7B–S7D) frequently had compound bias in terms of algorithmic underdiagnosis. For example, in the X-ray-dataset, female patients aged 20–40 years had the highest rate of underdiagnosis (was 0.011% higher than in female patients aged 60–80 years as shown in Figure 2B). In the X-ray-Test2 dataset, the rate of underdiagnosis in female patients with malignancies was twice as high as that in male patients with malignancies (Figure S5D). In the X-ray-Test3 dataset, women with cancer had a 0.005% higher rate of underdiagnosis than women without cancer (Figure S6B). In the CT dataset, the rate of underdiagnosis among Iranian women was about twice that of French women (Figure S7B), and women aged 20–40 years had a 0.02% higher rate of underdiagnosis than women aged 40–60 years (Figure S7B).

We observed that patients belonging to two specific subpopulations had a greater rate of underdiagnosis. In other words, not all female patients have the same rate of misdiagnosis (for example, Iranian women have a higher rate of underdiagnosis than French women).

Poor model generalization caused by heterogeneous datasets

However, the AUC and ACC values shown in Table 1 are both close to 100%, which leads us to think: the X-ray dataset in this paper is composed of multiple datasets with positive and negative samples from different datasets (heterogeneous datasets). Training model on a heterogeneous dataset will lead to the poor generalization ability of the model. Table 3 shows the dataset statistics used by some COVID-19 AI systems in the past three years.

From the data in Table 3, it can be concluded that the ACC values of the heterogeneous X-ray datasets used by the AI classification system (positive and negative samples from different datasets) are close to 100%. Since most X-ray training datasets are mixed. we need to explore

Table 2. CT dataset

3D-Model	AUC ± 95% CI	ACC ± 95% CI
3D-ResNet18	0.795 ± 0.001	0.884 ± 0.001
3D-ResNet34	0.711 ± 0.001	0.877 ± 0.001
mobilenet	0.627 ± 0.001	0.873 ± 0.001
shufflenetv2	0.875 ± 0.001	0.873 ± 0.001
squeezeNet	0.778 ± 0.001	0.866 ± 0.001

The performance indicators of the 3D classifier for CT dataset.
The models trained on CT dataset uses the same training-validation-test split, resulting in a calculated reported AUC ± 95% confidence interval (CI) and the ACC ± 95% confidence interval (CI).

whether training a model on a heterogeneous dataset will lead to poor model generalization ability. In view of this phenomenon, we redivided several X-rays datasets introduced in Table 4 and did some comparative experiments as shown in Table 5 to analyze this point. There we give each small dataset an abbreviation (see Table 4 for the column "Divide abbreviations").

These datasets are redivided into four design schemes (see Table 5), and the details of the division are as follows.

(1) Covid-19 X--ray Severity Scoring dataset is represented by A. COVID-19 Chest X-rays (NY-SBU) dataset is represented by B. Covid-19 Chest X-rays for Mortality Prediction and covid-chestxray-dataset are uniformly represented in C. Positive samples in the COVIDx CXR-3 dataset are represented by D, and negative samples are represented by E. ChestX-ray8 dataset is represented by F. ABC is all positive sample datasets; F is negative sample dataset.
(2) The internal dataset (the training set) in experiment design 1 is made up of DE, and the external dataset (the test set) is made up of FABC. The purpose of this division is to ensure that the training set is a homologous dataset and the test set is an additional heterologous dataset.
(3) The internal dataset (the training set) in experiment design 2 is made up of FADE, and the external dataset (the test set) is made up of FADE. The purpose of this division is to ensure that the training set and the test set are from the same heterologous dataset.
(4) The internal dataset (the training set) in experiment design 3 is made up of EA, and the external dataset (the test set) is made up of EBC. The purpose of this division is to ensure that the training set is a heterogeneous dataset and the test set is a heterogeneous dataset (having some of the same data sources as the training set but not duplicating them).
(5) The internal dataset (the training set) in experiment design 4 is made up of EA, and the external dataset (the test set) is made up of BC. The purpose of this division is to ensure that the training set is a heterologous dataset and the test set is an additional heterologous dataset.

In experimental design 2 in the table, the ACC values for the training set (internal) and the ACC values for the test set (external) are similar because the training and test sets are from the same heterologous dataset. In the experimental design in Table 3, the ACC value of the training set (internal) is nearly 50% higher than that of the test set (external) because the training and test sets are not exactly from the same heterologous dataset. In experimental design 4 in the table, the ACC value of the training set (internal) is nearly 90% higher than that of the test set (external). This is because the training set and the test set are completely from different heterogeneous datasets, which also indicates that models trained with heterogeneous datasets are less generalizable. In experimental design 1 in the table, because the model was trained on a homologous dataset, the ACC values for the test set (internal) and test set (external) remained around 95%, even though the test set was from a heterogeneous data source, indicating that the model had good generalization.

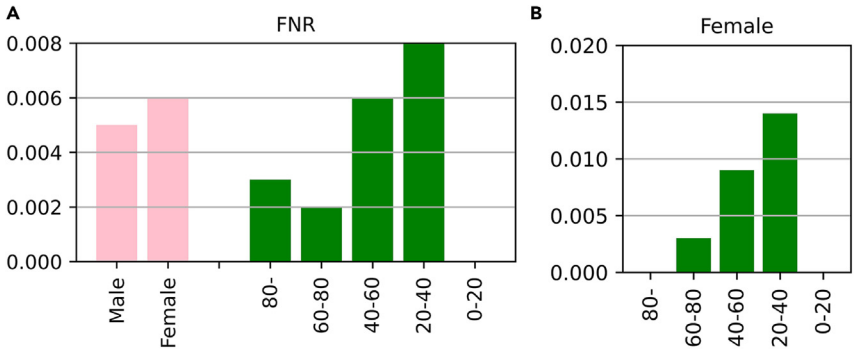


Figure 2. Underdiagnosis analysis of sex, age, and cross subpopulations in X-ray dataset

(A) The underdiagnosis rate, as measured by the no-finding FNR, in the indicated patient subpopulations.
(B) Intersectional underdiagnosis rates for female patients and patients of different age groups. The results are averaged over five trained models with different random seeds on the same train-validation-test splits.

Table 3. Statistics on datasets used by artificial intelligence systems for COVID-19

Whether heterogeneous datasets	AUC	ACC	Dataset type	Study
No	0.950	0.860	CT	Song et al.[25]
	–	0.815	CT	Kabir et al.[26]
	–	0.878	CT	Bougourzi et al.[27]
Yes	0.990	–	CT	Alhadad et al.[19]
	0.970	0.957	X-rays	Afshar et al.[20]
	0.991	0.960	X-rays	Chetoui et al.[21]
	0.992	0.996	X-rays	Ghose et al.[22]
	1.000	0.996	X-rays	Siddhartha et al.[23]
	–	0.930	CT	Pathak et al.[24]

Table 3 shows the work of others using homologous datasets and heterologous datasets, as well as the types of datasets and classification performance indicators (ACC and AUC), respectively.

DISCUSSION

We have shown the same trend of underdiagnosis across multiple X-ray and CT public datasets in the COVID-19 field, with AI algorithms exhibiting systematic underdiagnosis bias in specific subpopulations (e.g., female patients, Iranian patients, young patients, patients with malignancies, cancers, or other lung diseases). We found that these effects persisted in cross subpopulations, such as Iranian female patients. The specific subpopulations with high rates of underdiagnosis in the X-ray-Test2 dataset are different, especially sex and age attributes, which should be further explored. In this section, we need to discuss and research from four aspects to gain a comprehensive understanding of the underdiagnosis bias in AI algorithms for COVID-19 diagnosis, as well as the discussion of generalization ability of models trained on heterologous datasets.

First, considering the amplification of bias, unintended biases in AI medical algorithms can be exacerbated by the phenomenon of deviation amplification. These biases can arise from inherent biases present in the data used to train the algorithms, such as underdiagnosis biases related to sex[13] and race[12,16] that have been observed in other clinical fields. When AI algorithms are trained on datasets that contain these biases, they have the potential to amplify and perpetuate these biases. This phenomenon has serious implications for the fairness and accuracy of AI algorithms in healthcare settings. During the epidemic of COVID-19, a large number of AI diagnostic models have emerged, so whether they show underdiagnosis has to attract our attention and further research.

Table 4. Summary statistics for X-ray dataset

Dataset	No. of images		Labels	Divide abbreviations
	COVID-19	Normal		
Covid-19 X--ray Severity Scoring(Alberto et al.)[28]	4,695	–	Sex, Age	A
COVID-19 Chest X-rays (NY-SBU) (Saltz et al.)[29]	6,215	–	Sex, Age, Malignancies, Other-lung-disease, Smoking-status, SBP.above139	B
Covid-19 Chest X-rays for Mortality Prediction (Larxel) Covid-19 Chest X-rays for Mortality Prediction [https://www.kaggle.com/datasets/andrewmvd/covid19-xrays-mortality-prediction]	196	–	Sex, Age, Race, CANCER, CURRENT REGNANT, DIABETES TYPE I, DIABETES TYPE II	C
covid-chestxray-dataset (Joseph et al.)[30]	408	–	Sex, Age	
ChestX-ray8(Wang et al.)[31]	–	10,405	Sex, Age	F
COVIDx CXR-3[32]	16,194	14,192	–	D(Positive) E(Negative)

(1) The healthy chest X-ray scans used for training in this study were extracted from the public chest X-ray database provided by the NIH Clinical Center.[31] We randomly selected 10,405 images of 7,187 patients from a patient scan pool labeled "no findings" to achieve an overall healthy COVID-19 ratio of about 1:1 and avoid lopsided data issues.
(2) To expand the training set by an order of magnitude, 30,386 images from COVIDx CXR-3 with no patient information labels such as gender and age have been added to the training set for better results.
(3) The second dataset[33] and the third dataset[5] are specifically used as test sets (described in detail in Table S3) and are not placed in the X-ray dataset in Table S1. It is placed here for statistical convenience.

Table 5. Experiments to explore the generalization ability of the model on X-ray dataset

	Experiment design 1				Experiment design 2				Experiment design 3				Experiment design 4			
Model	internal AUC	external AUC	internal ACC	external ACC	internal AUC	external AUC	internal ACC	external ACC	internal AUC	external AUC	internal ACC	external ACC	internal AUC	external AUC	internal ACC	external ACC
Resnet18	0.989	0.992	0.967	0.953	0.999	0.999	0.994	0.993	0.984	0.824	0.959	0.475	0.984	–	0.959	0.085
Resnet34	0.989	0.987	0.967	0.935	0.999	0.999	0.996	0.995	0.987	0.868	0.967	0.469	0.987	–	0.967	0.074
Resnet50	0.985	0.991	0.958	0.959	0.999	0.999	0.996	0.996	0.988	0.856	0.500	0.447	0.988	–	0.500	0.037
Densnet121	0.993	0.992	0.975	0.965	0.999	0.999	0.995	0.996	0.987	0.882	0.960	0.501	0.987	–	0.960	0.132
Densnet169	0.989	0.983	0.965	0.931	0.999	0.999	0.996	0.996	0.987	0.882	0.967	0.467	0.987	–	0.967	0.070

There is no external AUC value in experiment design 4 because the test set is all positive samples. We designed four sets of experiments to explore the impact of heterologous datasets on the generalization ability of the model. In each set of experiments, we used internal data to train the model and external data to test the generalization ability of the model and calculated the ACC value and AUC value.

Second, at present, there are some solutions to achieve relative diagnostic fairness, but there are certain defects. For example, one possible approach is to use appropriate data pre-processing techniques to harmonize data to some extent and/or use hyperparameter tuning to train deep networks whose machine learning models have no bias in their predictions across different subpopulations.[34] However, if the bias is derived from other sources, such as diagnostic bias between different ethnic groups, the bias cannot be eliminated by this method.[12]

Another possible post-processing approach is to select different thresholds for different subpopulations that correspond to the operating points of their receiver operating characteristic (ROC) curves from calibration perspective, thus making FNR equal across the subpopulations.[35] However, due to a large number of unknowns caused by the small population of some cross subpopulations, it may be difficult to obtain an accurate approximation of the thresholds, so it is not practical to use different thresholds for each group. In addition, achieving equal FNR may require randomly and systematically deteriorating model performance in specific subpopulations, and it is unclear whether it is ethical to deteriorate the global model representation of a subpopulation to realize equity in a medical context.[15]

Third, in order to address the issue of underdiagnosis bias, it is recommended that relevant regulatory bodies and healthcare institutions undertake thorough and impartial evaluations. Our study highlights the necessity for comprehensive assessments of AI-based emergent healthcare algorithms in background of super-contagious disease spread. Certain models leveraging AI can extract demographic information, including age, sex, and ethnicity, from chest X-ray images.[3] The underdiagnosis bias can potentially lead to delays in patients receiving timely and appropriate care. Given the increasing prevalence of medical algorithms, before deploying these algorithms, it is crucial for developers and clinical practitioners to meticulously evaluate key metrics associated with health disparities, such as underdiagnosis rates, during multiple stages of decision model development and subsequent to deployment. This proactive approach will help mitigate the detrimental impact of underdiagnosis bias on specific patient subpopulations.

Fourth, from the analysis of our study, we can see that AI models trained using heterologous datasets (positive and negative samples from different datasets) have poor model generalization. We believe that this may be because when the model is trained with heterogeneous datasets, the system learns the differences between different datasets, not just the differences caused by lesions, so that its accuracy is falsely high, and the generalization power of the trained model is also very limited. When using homologous datasets to train the model, the system learned more focal features, and the model performance was also proved to be stable. We suggest that in the process of AI model training, a more reasonable method, that is, using homologous datasets to train the model, will make it easier for the model to learn the disease-related feature differences, and the model performance will be more stable.

In summary, we found that AI diagnostic algorithms trained on COVID-19 X-ray and CT datasets were inadequate in diagnosing specific subpopulations. Patients in intersecting subpopulations (e.g., Iranian female patients) are particularly vulnerable to algorithm-based underdiagnosis. Underdiagnosis leads to undiagnosed COVID-19 patients not receiving timely treatment to control the source of infection, and this problem is extremely frightening in the clinic. Our findings suggest that, without robust auditing of performance differences between different subpopulations and AI models trained under multi-source datasets, deployed algorithms may overstate actual accuracy and exacerbate existing systemic health inequalities. Relevant staff and departments must consider the issue of equitable access to health care for specific subpopulations and how AI-based diagnostic models can be used more effectively. In addition, we also found that training the model with heterogeneous datasets would lead to poor model generalization, and we recommend using homologous datasets for training to obtain a classifier with more stable performance.

We will then extend the work of this study to other currently unstudied disease areas to better understand how algorithmic bias permeates medical algorithms and provide more robust evidence for addressing the issue of related bias.

Limitations of the study

Although our work can effectively assist researchers in analyzing AI underdiagnosis bias in specific patient populations, this study has limitations. On the one hand, there are limitations in the data sources. Due to the privacy security of patients, we cannot access comprehensive large-scale datasets containing more labels such as the social status of patients, which prevents us from exploring AI underdiagnosis bias in subpopulations from various perspectives. More collaborations, for example, trying to invite more expert doctors to create datasets with more patient label information for us to study this kind of problem, are needed. On the other hand, the images in existing homogeneous datasets mostly do not come from the same imaging devices, and the resolution of most X-ray datasets is low, which may affect our ability to achieve optimal experimental results. We can also involve manufacturers as the bias variables in the future study, for further exploring its potential bias.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethical statement
  - Datasets

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109712.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62206196), the Natural Science Foundation of Shanxi (202103021223035), Macao Polytechnic University grant (RP/FCA-05/2022), and Science and Technology Development Fund, Macao (0021/2022/AGJ).

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of the manuscript. Study design and project supervision: R.C. and T.T. Funding acquisition: X. Wen and T.T. Data curation and analysis: Y.L., X. Wang, and Y.G. Model design: Y.L., X. Wen, and C.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 16, 2023
Revised: March 1, 2024
Accepted: April 7, 2024
Published: April 10, 2024

REFERENCES

1. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L. (2020). Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology *296*, E32–E40.
2. Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., and Ji, W. (2020). Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology *296*, E115–E117.
3. Ozsahin, D.U., Isa, N.A., and Uzun, B. (2022). The Capacity of Artificial Intelligence in COVID-19 Response: A Review in Context of COVID-19 Screening and Diagnosis. Diagnostics *12*, 2943.
4. Mei, X., Lee, H.-C., Diao, K.-y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al. (2020). Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. *26*, 1224–1228.
5. Bai, H.X., Hsieh, B., Xiong, Z., Halsey, K., Choi, J.W., Tran, T.M.L., Pan, I., Shi, L.-B., Wang, D.-C., Mei, J., et al. (2020). Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology *296*, E46–E54.
6. Yu, X., Lu, S., Guo, L., Wang, S.H., and Zhang, Y.D. (2021). ResGNet-C: A graph convolutional neural network for detection of COVID-19. Neurocomputing *452*, 592–605.
7. Tuncer, I., Barua, P.D., Dogan, S., Baygin, M., Tuncer, T., Tan, R., Yeong, C.H., and Acharya, U.R. (2022). Swin-textural: A novel textural features-based image classification model for COVID-19 detection on chest computed tomography. Inform. Med. Unlocked *36*, 101158.
8. Aslan, N., Ozmen Koca, G., Kobat, M.A., Dogan, S., and Systems, I.L. (2022). Multi-classification deep CNN model for diagnosing COVID-19 using iterative neighborhood component analysis and iterative ReliefF feature selection techniques with X-ray images. Chemometr. Intell. Lab. Syst. *224*, 104539.
9. Gupta, A., Mishra, S., Sahu, S.C., Srinivasarao, U., and Naik, K.J. (2023). Application of Convolutional Neural Networks for COVID-19 Detection in X-ray Images Using InceptionV3 and U-Net. New Generat. Comput. *41*, 475–502.
10. Erdem, K., Kobat, M.A., Bilen, M.N., Balik, Y., Alkan, S., Cavlak, F., Poyraz, A.K., Barua, P.D., Tuncer, I., Dogan, S., et al. (2023). Hybrid-Patch-Alex: A new patch division and deep feature extraction-based image classification model to detect COVID-19, heart failure, and other lung conditions using medical images. Int. J. Imag. Syst. Technol. *33*, 1144–1159. https://doi.org/10.1002/ima.22914.
11. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., and Chin, M.H. (2018). Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med. *169*, 866–872.
12. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453.
13. Maserejian, N.N., Link, C.L., Lutfey, K.L., Marceau, L.D., and McKinlay, J.B. (2009). Disparities in physicians' interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment. J. Womens Health *18*, 1661–1667.
14. Klare, B.F., Burge, M.J., Klontz, J.C., Vorder Bruegge, R.W., and Jain, A.K. (2012). Face recognition performance: Role of demographic information. IEEE Trans. Inf. Forensics Secur. *7*, 1789–1801.
15. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. *27*, 2176–2182.
16. Chowkwanyun, M., and Reed, A.L., Jr. (2020). Racial health disparities and Covid-19—caution and context. N. Engl. J. Med. *383*, 201–203.
17. Leslie, D., Mazumder, A., Peppin, A., Wolters, M.K., and Hagerty, A. (2021). Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? BMJ *372*, n304.
18. Luengo-Oroz, M., Bullock, J., Pham, K.H., Lam, C.S.N., and Luccioni, A. (2021). From artificial intelligence bias to inequality in the time of COVID-19. IEEE Technol. Soc. Mag. *40*, 71–79.
19. Alhadad, A.A., Tarawneh, O., Mostafa, R.R., and El-Bakry, H.M. (2023). Residual Attention Deep SVDD for COVID-19 Diagnosis Using

CT Scans. Comput. Mater. Continua (CMC) *74*.
20. Afshar, P., Heidarian, S., Naderkhani, F., Oikonomou, A., Plataniotis, K.N., and Mohammadi, A. (2020). Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. Pattern Recogn. Lett. *138*, 638–643.
21. Chetoui, M., Akhloufi, M.A., Bouattane, E.M., Abdulnour, J., Roux, S., and Bernard, C.D. (2023). Explainable COVID-19 Detection Based on Chest X-rays Using an End-to-End RegNet Architecture. Viruses *15*, 1327.
22. Ghose, P., Alavi, M., Tabassum, M., Ashraf Uddin, M., Biswas, M., Mahbub, K., Gaur, L., Mallik, S., and Zhao, Z. (2022). Detecting COVID-19 infection status from chest X-ray and CT scan via single transfer learning-driven approach. Front. Genet. *13*, 980338.
23. Siddhartha, M., and Santra, A. (2020). COVIDLite: A depth-wise separable deep neural network with white balance and CLAHE for detection of COVID-19. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.13873.
24. Pathak, Y., Shukla, P.K., Tiwari, A., Stalin, S., Singh, S., and Shukla, P.K. (2020). Deep Transfer Learning Based Classification Model for COVID-19 Disease. Ing. Rech. Biomed. *43*, 87–92.
25. Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Wang, R., Zhao, H., Chong, Y., et al. (2021). Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE ACM Trans. Comput. Biol. Bioinf *18*, 2775–2780.
26. Kabir, S., Mohammed, E.A., Zaamout, K., and Ikki, S. (2021). A Traditional Machine Learning Approach for COVID-19 Detection from CT Images, pp. 256–263.
27. Bougourzi, F., Contino, R., Distante, C., and Taleb-Ahmed, A. (2021). CNR-IEMN: A Deep Learning Based Approach to Recognise Covid-19 from CT-Scan, pp. 8568–8572.
28. Signoroni, A., Savardi, M., Benini, S., Adami, N., Leonardi, R., Gibellini, P., Vaccher, F., Ravanelli, M., Borghesi, A., Maroldi, R., and Farina, D. (2021). BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. Med. Image Anal. *71*, 102046.
29. Saltz, J., Saltz, M., Prasanna, P., Moffitt, R., Hajagos, J., Bremer, E., Balsamo, J., and Kurc, T. (2021). Stony Brook University COVID-19 Positive Cases [Data set]. Cancer Imaging Arch. https://doi.org/10.7937/TCIA.BBAG-2923.
30. Joseph Paul Cohen, P.M., Lan, D., Roth, K., Duong, T.Q., and Ghassemi, M. (2020). COVID-19 Image Data Collection: Prospective Predictions Are the Future.
31. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R.M. (2017). Chestx-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, pp. 2097–2106.
32. Wang, L., Lin, Z.Q., and Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci. Rep. *10*, 19549.
33. Saltz, J., Saltz, M., Prasanna, P., Moffitt, R., Hajagos, J., Bremer, E., Balsamo, J., and Kurc, T. (2021). Stony Brook University COVID-19 Positive Cases [Data set]. Cancer Imaging Arch.
34. Wang, R., Chaudhari, P., and Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. Proc. Natl. Acad. Sci. USA *120*, e2211613120.
35. Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. *29*.
36. Rubin, G.D., Ryerson, C.J., Haramati, L.B., Sverzellati, N., Kanne, J.P., Raoof, S., Schluger, N.W., Volpi, A., Yim, J.-J., Martin, I.B.K., et al. (2020). The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. Radiology *296*, 172–180.
37. Arias-Garzón, D., Tabares-Soto, R., Bernal-Salcedo, J., and Ruz, G.A. (2023). Biases associated with database structure for COVID-19 detection in X-ray images. Sci. Rep. *13*, 3477.
38. Ng, M.-Y., Lee, E.Y.P., Yang, J., Yang, F., Li, X., Wang, H., Lui, M.M., Lo, C.S.-Y., Leung, B., Khong, P.-L., et al. (2020). Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol. Cardiothorac. Imaging *2*, e200034.
39. Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., Cui, J., Xu, W., Yang, Y., Fayad, Z.A., et al. (2020). CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology *295*, 202–207.
40. Kroft, L.J.M., van der Velden, L., Girón, I.H., Roelofs, J.J.H., de Roos, A., and Geleijns, J. (2019). Added value of ultra–low-dose computed tomography, dose equivalent to chest x-ray radiography, for diagnosing chest pathology. J. Thorac. Imag. *34*, 179–186.
41. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., and Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput. Biol. Med. *121*, 103792.
42. Ucar, F., and Korkmaz, D. (2020). COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. Med. Hypotheses *140*, 109761.
43. Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (PMLR), pp. 77–91.
44. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. USA *117*, 12592–12594.
45. Bhatt, M., Kant, S., and Bhaskar, R. (2012). Pulmonary tuberculosis as differential diagnosis of lung cancer. South Asian J. Cancer *1*, 36–42.
46. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med. *383*, 874–882.
47. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet *395*, 507–513.
48. Gunraj, H., Sabri, A., Koff, D., and Wong, A. (2021). COVID-Net CT-2: Enhanced deep neural networks for detection of COVID-19 from chest CT images through bigger, more diverse learning. Front. Med. *8*, 729287.
49. Zhou, T., Lu, H., Yang, Z., Qiu, S., Huo, B., and Dong, Y. (2021). The ensemble deep learning model for novel COVID-19 on CT images. Appl. Soft Comput. *98*, 106885.
50. Huang, G., Liu, Z., and Weinberger, K.Q. (2016). Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269.
51. Hassan, M.M., AlQahtani, S.A., Alelaiwi, A., and Papa, J.P. (2023). Lightweight neural architectures to improve COVID-19 identification. Front. Physiol. *11*, 1153637.
52. Wang, W., Liu, S., Xu, H., and Deng, L. (2022). COVIDX-LwNet: A Lightweight Network Ensemble Model for the Detection of COVID-19 Based on Chest X-ray Images. Sensors *22*, 8578.
53. Shin, J., Chang, Y.K., Heung, B., Nguyen-Quang, T., Price, G.W., and Al-Mallahi, A.J.C.E.A. (2021). A deep learning approach for RGB image-based powdery mildew disease detection on strawberry leaves. Comput. Electr. Agri. *183*, 106042.
54. He, J., Zhang, Y., Chung, M., Wang, M., Wang, K., Ma, Y., Ding, X., Li, Q., and Pu, Y.J.M.p. (2023). Whole-body Tumor Segmentation from PET/CT Images Using a Two-Stage Cascaded Neural Network with Camouflaged Object Detection Mechanisms.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
the COVIDx CT-3A	Public access	https://www.kaggle.com/datasets/hgunraj/covidxct
the Covid-19 X--ray Severity Scoring dataset	Public access	https://www.kaggle.com/datasets/andrewmvd/covid19-xray-severity-scoring?select=metadata.csv
the COVID-19 Chest X-rays (NY-SBU) dataset	Public access	https://www.kaggle.com/datasets/toxite/covid-19-cxr-ny-sbu
the Covid-19 Chest X-rays for Mortality Prediction dataset	Public access	https://www.kaggle.com/datasets/andrewmvd/covid19-xrays-mortality-prediction
the covid-chestxray-dataset	Public access	http://github.com/ieee8023/covid-chestxray-dataset
the COVIDx CXR-3 dataset	Public access	https://www.kaggle.com/datasets/andyczhao/covidx-cxr2?
the ChestX-ray8 dataset	Public access	https://www.kaggle.com/datasets/nih-chest-xrays/data
Source Code	This paper	https://github.com/Liu-Ya-nan/COVID-19_code.git
Software and algorithms		
Python (version 3.8)	Python software	https://www.python.org/
Cuda (version 11.6.0)	Nvidia	https://developer.nvidia.com/
PyTorch (version 1.12.0)	Pytorch software	https://pytorch.org/
Numpy (version 1.23.5)	Numpy package	https://scipy.org/install/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Tao Tan (taotanjs@gmail.com).

Materials availability

All the data comes from public datasets as explained in Section experimental model and study participant details-datasets.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code have been deposited at Github (https://github.com/Liu-Ya-nan/COVID-19_code.git), and are publicly accessible as of the date of publication.DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethical statement

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Datasets

In disease centers in COVID popular regions, imaging method is widely used for inspections of suspicious COVID-19-infected patients. Therefore, three-dimensional volume computed tomography (CT) and two-dimensional projected chest X-ray (X-ray) have been used on an unprecedented scale in the diagnosis of COVID-19.[36,37] Among them, chest CT can more clearly display the typical imaging features of COVID-19 patients, such as ground-glass opacity and consolidation around the lung.[38,39] Due to the slight deficiency of CT imaging scanning cost and high patient dose, X-ray imaging scanning cost is low, the speed is fast, the radiation amount is low, and the audience is wider.[40,41] It is also highlighted in the diagnosis.[42]

Combining their respective advantages, this paper will conduct a detailed study through the COVID-19 datasets using both CT and X-ray imaging modes (as shown in Figure S1).

X-Ray image dataset

The underdiagnosis studies in this paper focus on individuals and cross-sectional subpopulations across sex,[43,44] age,[45] ethnicity,[43,46] and multiple common diseases,[47] which were selected by reference to the history of medical and AI bias studies.[43–46] Multiple labels are involved, and the existing single public dataset cannot meet the expected subpopulation labels, so multiple COVID-19 open-source datasets (see Table 4 for details) are integrated as X-ray image datasets in this study.

After data integration (label screening and deletion of blank values), the distribution of the X-ray image dataset (hereinafter referred to as X-ray-dataset) in this study is shown in Table S1 (Individual partitions are in parentheses by patient level). In the test set, the image count at the intersection of sex and age is shown in Table S2. Since there is no patient label information in the expanded data, Therefore, the data distribution statistics of Age rows and Sex rows in the table are the statistics of the dataset before the training set is expanded.

We used 47,074 publicly available chest X-ray images from 29,342 patients in a dataset with roughly equal proportions of male and female patients, most of whom were between 20 and 80 years of age, and divided it into a training set, a validation set, and a test set on an 8:1:1 ratio (The COVIDx CXR-3 dataset is all used as a training set), as shown in Figure S2.Examples of positive images for each X-ray dataset are shown in Figure S3.

In Table 4, two datasets, COVID-19 Chest X-rays (hereinafter referred to as X-ray-Test2) and Covid-19 Chest X-rays for Mortality Prediction (hereinafter referred to as X-ray-Test3), due to the abundant subgroup labels, after data integration (label screening and deletion of blank values), the data distribution of the test set is shown in Table S3 (Individual partitions are in parentheses by patient level).In the test sets of X-ray-Test2 and X-ray-Test3, the image count at the intersection of sex and age are shown in Tables S4 and S5.

CT image dataset

This study introduced COVIDx CT-3A (Gunraj et al.), an open-access CT dataset from a large curated benchmark multi-country cohort from China's National BioInformation Center (CNCB). This dataset has been clinically validated extensively.[48] This study focuses on the binary classification of COVID-19 (COVID-19 and Normal) and therefore only COVID-19 positive and Normal data in this dataset are used as the CT dataset in this study (hereinafter referred to as CT-dataset). Chest CT images of a cohort of 2760 patients were included and divided into a training set, a validation set, and a test set according to a ratio of 8:1:1, as shown in Figure S4.

The data distribution is shown in Table S6 (note that the data are partitioned at the patient level).In the test sets, the image count at the intersection of sex and age are shown in Table S7.

METHOD DETAILS

Experimental design

In this paper, we used several proposed advanced COVID-19 diagnostic classifiers to train the model on the two modal datasets to demonstrate the model performance of the entire population and then compared the underdiagnosis rates of each subpopulation in the total population to evaluate the model decision bias of underdiagnosed patients.

The COVID-19 diagnostic classifiers used in this paper include ResNet[49] (ResNet 18, ResNet 34, ResNet 50, 3-D ResNet18 and 3D-ResNet34), DenseNet[50] (DenseNet121 and DenseNet169), mobilenet,[51] shufflenetv2[52] and squeezeNet.[53] The complexity of the models is shown in Table S8.

Medical images preprocessing

All images are normalized using mean and standard deviation according to standard convention.[15]

Labels preprocessing

For each image, if it is positive for COVID-19, it is labeled as "1" and if it is normal, it is labeled as "0". The classifier is trained by binary classification, and the underdiagnosis and other fairness indicators on label "1" are statistically analyzed.

Model training

We use the method described in the Datasets chapter to divide the dataset into training sets, verification sets, and test sets. The data is split randomly, with no overlap of patients between the partitions. The train–validation–test set sizes for the X-ray-dataset are 23473–2934–2935 and for the CT-dataset they are 2208–276–276 (Partition at the patient level). We applied Gaussian filter and random horizontal flip and mild rotation data augmentation for model training. We trained the models on a server with six Nvidia TITAN RTX GPUs using the PyTorch 22 framework. We used the Adam optimizer with default parameters. We set the initial learning rate to 5-4 and automatic learning rate scheduling during the training process – if the loss is not improved in 3 epochs, the learning rate will be halved; On X-ray-dataset, Binary CrossEntropy Loss was used due to the balanced ratio between positive and normal categories. On CT-dataset, FocalLoss[54] was used due to the unbalanced ratio between positive and normal categories (87:13) to address data imbalances.

All reported metrics, such as AUC, ACC, and FNR, were evaluated separately in five models (the same model was trained five times with five different random seeds),[15] and the training-validation-testing segmentation remained constant across the five models. Random seeds are chosen randomly from a range of 0–100. For each dataset, the reported results in this study: FNR (Figures 2 and S5–S7), represent the average ± 95% confidence interval of the results from the five models (with different random seed initializations). AUC and ACC

(Tables 1 and 2) represent the average ± 95% confidence intervals for the results from the five models (with different random seed initializations). Following best practices in FNR estimation, we select a single threshold for all groups, thus maximizing F1 scores.

QUANTIFICATION AND STATISTICAL ANALYSIS

Accuracy

The higher the accuracy, the better the classifier performance. The equation explaining the aforementioned metrics are shown in Equation 1[26] below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Equation 1)}$$

Where true positive (TP) represents the number of malignant samples correctly predicted, true negative (TN) represents the number of normal samples correctly predicted, false positive (FP) represents the number of normal samples incorrectly predicted, and false negative (FN) represents the number of malignant samples incorrectly predicted.

Underdiagnosis rate (FNR)

We predict the definition and quantification of underdiagnosis rates based on binarized models. We calculated and compared the underdiagnosis rates of subpopulations in the general population to assess the underdiagnosis bias of patients in the model. We use the false-negative rate (FNR) predicted by a binarized model of the "positive" label to represent the underdiagnosis rate, with and $FNR_{s_{i,j}}$ indicating the probability of undiagnosed disease at the level of subgroup s_j (e.g., female) and cross-identity $s_{i,\ j}$ (e.g., black and female), respectively. The equation explaining the aforementioned metrics are shown in Equation 2[15] below.

$$FNR_{s_j} = P\left[\hat{Y} = 0 \middle| s_j, Y = 1\right] \quad \text{(Equation 2)}$$

$$FNR_{s_{i,j}} = P\left[\hat{Y} = 0 \middle| s_{i,j}, Y = 1\right]$$

Where i and j represent subpopulations with different attributes, Y are true labels, and $\hat{Y}$ are predicted labels.