Explainable Pediatric Pneumonia Detection from Chest X-Ray Images: Assessing Model Attention with Lung-Region Masking

Muhammad Adeel Abid, Saman Riaz, Fatima Farhan
Department of Data Science & AI (Khwaja Fareed University of Engineering and Information Technology Rahim Yar Khan).
***Correspondence:** adeel.abid@kfueit.edu.pk

Citation | Abid. M. A, Riaz. S, Farhan. F, "Explainable Pediatric Pneumonia Detection from Chest X-Ray Images: Assessing Model Attention with Lung-Region Masking", IJIST, Vol. 8 Issue.9 pp3689-3707, October 2026
Received | Oct 05, 2026 **Revised |** Oct 20, 2026 **Accepted |** Oct 24, 2026 **Published |** Oct 26, 2026.

Deep learning models reach very high scores on the public pediatric chest radiograph dataset for pneumonia. It is not clear which part of the image these models use. This study tests the role of the lung fields with controlled input tests. A UNet trained on public adult lung masks is applied to 5,824 pediatric images that remain after 32 duplicates are removed. The automatic mask check excluded 409 images (7.0 percent), so 5,415 images are used. The split is grouped by patients. All final evaluations are done on 826 test images (232 normal and 594 pneumonia). Three versions of every image are built. These are the original image, the lung only image and the background only image with the lung fields removed. ResNet50, DenseNet121, EfficientNetB0 and MobileNetV3 Large are trained on each version with three seeds and tested on all versions. When training and test inputs matched, the test AUC is 0.997 to 0.998 for original images, 0.988 to 0.997 for lung only images and 0.990 to 0.995 for background only images. Models trained on original images kept an AUC of 0.964 to 0.975 on background only images but fell to 0.585 to 0.888 on lung only images. All 24 McNemar tests are significant after Holm adjustment. Control tests gave an AUC of 0.982 to 0.986 from an outer frame only and 0.989 to 0.993 with lung margins of 11, 31 and 51 pixels. Five image numbers alone, with no lung information, gave an AUC of 0.934. GradCAM placed 31.5 percent of attention inside the lung masks, which cover 21.0 percent of the image. High scores can therefore be reached without the segmented lung fields, and the classes differ in how the images are acquired. High accuracy on this dataset should not be read as proof that a model reads the lungs.

Keywords: Pediatric Pneumonia, Chest Radiograph, Lung Segmentation, Explainable AI, GradCAM, Shortcut Learning, Deep Learning

Introduction:

Pneumonia is a major cause of illness and death in young children. The chest radiograph is the most common test used to support diagnosis. Deep learning models can read these images, and many studies report very high scores on public data [1], [2]. The pediatric dataset released by [1] is used in many of these studies. Many reported accuracies in it are above 95 percent. A high score does not show that a model reads the lungs. A model can also learn from other parts of the image. These include borders, text markers, body size, image shape and the source hospital. [3] found that the performance of a pneumonia model changed when it is tested at other hospitals[3]. [4] call this problem shortcut learning [4]. DeGrave et al. showed that models for COVID 19 detection on chest radiographs used shortcuts instead of true signal [5]. [6] removed the lung regions from the images and still obtained good scores for COVID 19 detection. These studies warn that public chest radiograph sets can hold hidden cues.

Explanation maps are often used to check what a model looks at. [2]. used visualization to study pneumonia predictions on pediatric radiographs[2]. [7] introduced GradCAM, which is now a common choice. [8] compared saliency methods on chest radiographs and found that they often do not match the regions marked by experts. [9] also question how reliable saliency maps are for locating findings. A heat map alone is therefore weak evidence. A direct test is also needed. Such a test changes the input and measures how the performance changes.

Recent work on this dataset has focused on explainable models. [10]compared seven pretrained networks with Grad-CAM and used a patient level split. They noted that some patient numbers appear in both the original training and test folders. [11] applied Grad-CAM and Score-CAM to ResNet50 on pediatric radiographs. [12] added lung segmentation to a multi scale transformer. [10] heat maps with a coarse fixed lung region and do not test what happens when the lungs are removed.

Shortcut learning in medical imaging has received more attention since 2022. Reviews and studies describe how models use hospital, device and background cues [13] [14] [15] [16] [17] [18] showed the effect of chest drains on pneumothorax classification [18]. Some studies remove or keep the lungs on purpose. [19]. segmented and cropped the lungs to remove confounders in COVID 19 radiographs [19], and [20]. studied how lung segmentation changes COVID 19 diagnosis and explanation [20]. [21]tested whether COVID 19 classifiers can still separate classes without the lungs [21][22]. trained networks on adult chest radiographs with lung only and no lung inputs and found that models trained on full images did well without the lungs [22].

These studies leave a gap. Work on the pediatric dataset reports accuracy and heat maps. Work on masking and shortcuts uses adult or COVID 19 data. This study addresses the gap with controlled lung masking, a patient grouped split, several networks and seeds, and a measured attention share inside real lung masks on the pediatric pneumonia dataset. This study runs a controlled test on the pediatric pneumonia dataset. The lung fields are found with a UNet [23] trained on public adult masks [24][25]. Three versions of every image are then built. The first is the original image. The second keeps only the lung fields. The third removes the lung fields and keeps the rest. Four common networks are trained on each version and tested on all three. The study also measures how much GradCAM attention falls inside the lung masks. The split is grouped by patients to reduce leakage.This paper is organized as follows. The Materials and Methods section describes the data, the lung segmentation and the models. The Results and Discussion section reports the findings. The Conclusion section closes the paper.

Objective of the Study:

The research objectives are as follows:

To measure the test AUC of four networks when only the lung fields are visible and compare it with the original image.

To measure the test AUC when the lung fields and a margin of 11, 31 or 51 pixels are removed.

To measure how the AUC changes when a model is tested on an input version it is not trained on.

To measure the share of GradCAM attention inside the lung mask and compare it with the share of the image that the mask covers.

To test whether other image features explain the scores. These are the outer frame, the shape of the lung mask, image size and shape, and image intensity.

The effect of the lung region is read from the lung only input and from the wrong mask control, where the lung mask of another image is used. The effect of background information is read from the background only input at three margins and from the frame only input. The effect of other image features is read from a model that sees only five image numbers (height, width, aspect ratio, mean and standard deviation of pixel values) and from random stretching of the images during training.

Novelty Statement:

This study differs from earlier work in five ways. First, it tests the role of the lung fields on a pediatric dataset, while earlier masking studies used adult or COVID 19 data [19][21][22]. Second, it uses a patient group split with a duplicate check, and it adds masking tests that [10] does not have. Third, it reports a full train and test matrix for four networks and three seeds. Fourth, it measures GradCAM attention inside real lung masks and compares it with the lung area. Fifth, it runs control tests and reports the mask failures. Table 1 compares this study with earlier work.

Table 1. Comparison with earlier studies

Study	Data	Lung segmentation	Patient level split	Model evaluation	Explainabiliy	Statistical analysis
[10]	Pediatric (Kermany)	No. A coarse fixed lung region scores the maps	Yes	7 pretrained networks, accuracy, AUC, PR AUC, F1	Grad-CAM, lung attention ratio	Not reported
[22]	PadChest (adult) and fundus	Yes. CheXmask masks [26]lung only and no lung inputs, dilation sweep	Not stated	DenseNet121, AUC, 5 folds	SHAP, embeddings, radiology resident reading	DeLong test
This study	Pediatric (Kermany), adult masks for segmentation	Yes. UNet masks, lung only, background only, margins, wrong masks, frame only	Yes. 2,790 groups, duplicates removed	4 networks, 3 seeds, 3 by 3 matrix, confusion matrices, calibration	GradCAM, attention share inside real lung masks	McNemar with Holm, bootstrap CI

Materials and Methods:

Investigation Site: All experiments are done in a Kaggle notebook with one NVIDIA Tesla T4 GPU. The code used Python and PyTorch [27]. The full pipeline, from data loading to the

final tables, ran in about 95 minutes. The same code and the same settings are used for all models. Random seeds are fixed. Only public and anonymized images are used.

Dataset:

The study uses the Chest X Ray Images (Pneumonia) dataset from [28] It is released with the work of [1]The images are frontal chest radiographs of children aged one to five years. They are collected at one medical center in Guangzhou. The folder held 5,856 images. Of these, 1,583 are normal and 4,273 show pneumonia (2,780 bacterial and 1,493 viral). The original folders hold 5,216 training, 16 validation and 624 test images. The validation folder is too small to be useful. The official test folder also has a different class ratio, with 62.5 percent pneumonia images against 74.3 percent in the training folder. All images are therefore pooled, and a new split is made.

Cleaning and Splitting:

Files are checked for read errors, and none are corrupted. Exact duplicates are found with an MD5 hash of the file bytes. There are 62 images in duplicate groups and no group had conflicting labels. One copy of each group is kept and 32 images are removed. This left 5,824 images (1,579 normal and 4,245 pneumonia). Near duplicates such as resized or re saved copies are not searched. The dataset has no patient identity field, so patient groups are taken from the file names. Pneumonia files carry a person number, for example person1_bacteria_1.jpeg, and this number is used. Normal files carry an IM series number, and this number is used. A regular expression reads the number from each name and every image had one. This gave 2,790 groups. It is assumed that one number is one patient, and this is not verified. If two patients share a number, the split is stricter than needed. If one patient has two numbers, some leakage could remain. [10] also found patient numbers in both the original training and test folders, so the original folders are pooled. A stratified group method split the groups into 70 percent training, 15 percent validation and 15 percent test images. The overlap of groups between training, validation and test is checked by set intersection and is zero for all three pairs (Table 2). All images are converted to grayscale and resized to 224 by 224 pixels with area interpolation. The resize does not keep the original aspect ratio.

Table 2. Images in each split before and after the lung mask check

Split	Normal	Pneumonia	Total	Pneumonia share	Kept after mask check
Training	1,112	2,957	4,069	72.7%	3,802
Validation	230	612	842	72.7%	787
Test	237	676	913	74.0%	826
All	1,579	4,245	5,824	72.9%	5,415

Lung Segmentation:

The classification dataset has no lung masks. A UNet [23] is therefore trained on public adult chest radiographs with expert lung masks. These are the Montgomery County and Shenzhen sets [24][25]. The study used 704 image and mask pairs, 138 from Montgomery and 566 from Shenzhen. Fifteen percent of the pairs (105) are held out for validation. Images are resized to 256 by 256 pixels and normalized per image. The network has five levels with 32 to 512 channels. Training settings are given in Table 3. The epoch with the best validation Dice score is kept. The trained network is then applied to every pediatric image. The output is thresholded to 0.5. Up to two of the largest components that cover more than 2 percent of the image are kept and holes are filled. The mask is resized to 224 by 224 pixels with nearest neighbor interpolation. A mask is flagged as failed when no component is found or when the lung area is below 12 percent or above 65 percent of the image. The limits of 12 and 65 percent are chosen by the author before the classification experiments as plausible bounds for the share of a frontal chest image that the two lungs cover. They are not tuned on the classification

results. Flagged images are excluded from all experiments. A random sample of 100 masks is also checked by the author. This check is not done by a radiologist.

Input Versions:

Let x be a grayscale image, m its binary lung mask and D_k a dilation of a mask with a k-by-k square element. The three versions are x_orig = x, x_lung = x · m and x_back = x (1 - D_k(m)). Products are taken pixel by pixel. In the main experiments k = 11. This removes the lung mask plus a margin of 5 pixels on each side, which is 2.2 percent of the 224-pixel width. The value is chosen as a small margin to hide the lung edge and is not tuned. To test the effect of mask size, the background only version is also trained with k = 31 and k = 51 (Table 13).

Classification Models and Training:

Four networks are used. These are ResNet50 [29], DenseNet121 [30], EfficientNetB0 [31] and MobileNetV3 Large [32]. All start with ImageNet pretrained weights, and the last layer is replaced with two outputs. The grayscale image is copied to three channels and normalized with ImageNet statistics. Each network is trained separately on each input version. [33]with the settings in Table 3. Mixed precision is used in training and the epoch with the best validation AUC is kept. Each setting is repeated with seeds 0, 1 and 2. This gave 36 trained models from 4 networks, 3 versions and 3 seeds.

Evaluation used full precision. A fixed threshold of 0.5 is used for all metrics except AUC. Pretrained weights came from torchvision. IMAGENET1K_V2 is used for ResNet50 and MobileNetV3 Large, and IMAGENET1K_V1 for DenseNet121 and EfficientNetB0. The last layer is replaced by a two-output layer. Class weights arew_c = N / (2 N_c), where N is the number of training images and N_c is the number in class c. The learning rate is 0.0001 with a cosine schedule over 10 epochs, AdamW with weight decay 0.0001, batch size 32 and mixed precision. The epoch with the highest validation AUC is kept. Test results are not used for any choice. Software is Python, PyTorch, torchvision and scikit learn on the Kaggle latest container image in October 2026. The notebook is available at [34], [35] [36],.

Table 3. Training settings

Setting	**Lung segmentation (UNet)**	**Classifiers**
Input size	256 x 256	224 x 224, three channels
Optimizer	AdamW	AdamW
Learning rate	0.001	0.0001
Weight decay	0.0001	0.0001
Batch size	16	32
Epochs	25	10
Schedule	Cosine	Cosine
Loss	Binary cross entropy plus soft Dice	Class weighted cross entropy
Augmentation	Rotation up to 10 degrees, scale 0.9 to 1.1	Rotation up to 10 degrees, scale and shift up to 5%
Model selection	Best validation Dice	Best validation AUC
Seeds	42	0, 1, 2

Evaluation:

Every trained model is tested on the test images in all three input versions. This gives a 3 by 3 train and test matrix for each network. Accuracy, sensitivity, specificity, F1 score and AUC are reported. Pneumonia is the positive class. Values are the mean and the population standard deviation over the three seeds. Confusion matrices are given as mean counting over the three seeds. Calibration is measured with the expected calibration error (ECE, 10 equal

width bins) and the Brier score. The choice of metrics follows common recommendations for image analysis validation [37].

Explainability:

GradCAM [7]is computed for the ResNet50 model trained on original images with seed 0. The last convolutional stage is used. The map is resized to 224 by 224 pixels and scaled to the range 0 to 1. The target is the predicted class. For each test image the share of total attention that lies inside the lung mask is computed. The ratio of this share to the share of image area covered by the mask is also reported. A ratio above one means that attention is more concentrated on the lungs than expected by chance.

Statistical Analysis:

Bootstrap 95 percent confidence intervals [38] for the AUC are computed from 1,000 resamples of the test images for the seed 0 models. The exact McNemar test [39]compared the decisions of each original trained model on original inputs and on the other two versions. There are 24 tests from 4 networks, 3 seeds and 2 comparisons. The p values are adjusted with the Holm method [40]. A Welch t test and Cohen's d compared attention shares between correct and incorrect pneumonia predictions. Confidence intervals for the attention shares are computed from the group means and standard deviations. Figure 1 shows the adopted methodology.

The unified procedure is given below:

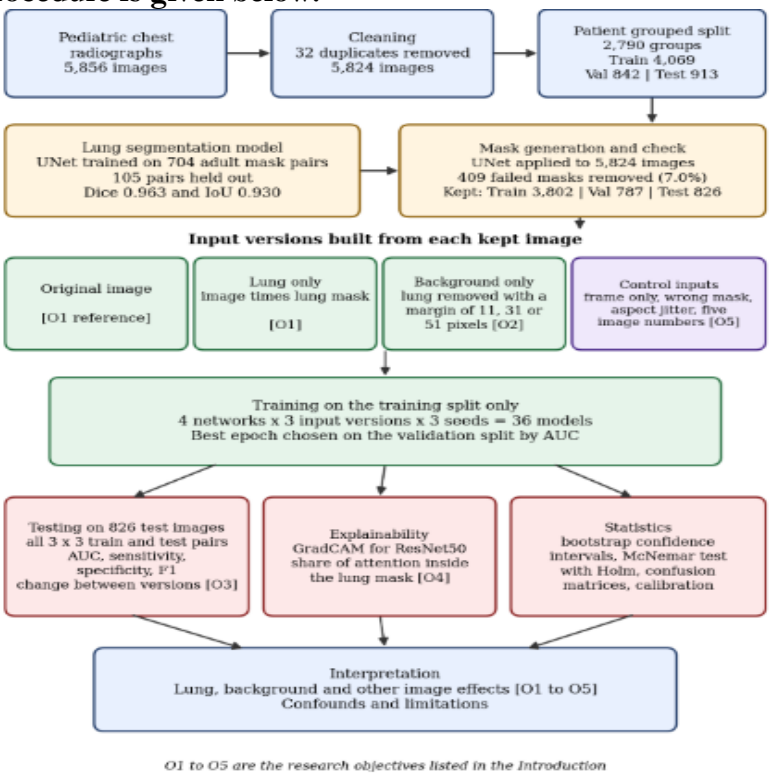

Figure 1 Adopted methodology.

Input = Chest radiograph set D, adult lung mask set L
Output = Metrics for all train and test input pairs and GradCAM attention shares
Remove corrupted and duplicate images from D
Split D by patient group into train, validation and test
Train a UNet on L and predict a lung mask for every image in D
Flag and remove failed masks
Build the original, lung only and background only versions

For each network M, input version V and seed s:
Train M on V and select the epoch by validation AUC
Test M on all three versions of the test split
Compute the GradCAM attention share inside the lung mask
Run the statistical tests and compare the results

Figure 1 shows the full procedure. The data are cleaned and split by patient group (training 4,069, validation 842, test 913). A UNet trained on adult masks produces lung masks, and 409 failed masks are removed, which leaves 3,802, 787 and 826 images. Three input versions and four control inputs are built. Four networks are trained in the training split, the best epoch is chosen on the validation split, and all results are measured on the 826 test images. The tags O1 to O5 show which step answers which objective.

Results and Discussion:

This section reports the dataset properties, the lung segmentation, the results of the three input versions, the cross-version tests and the attention analysis. All classification results are on the test split of 826 images (232 normal and 594 pneumonia) that remained after the mask check. Values are the mean and the population standard deviation over three seeds unless stated.

Dataset Observations:

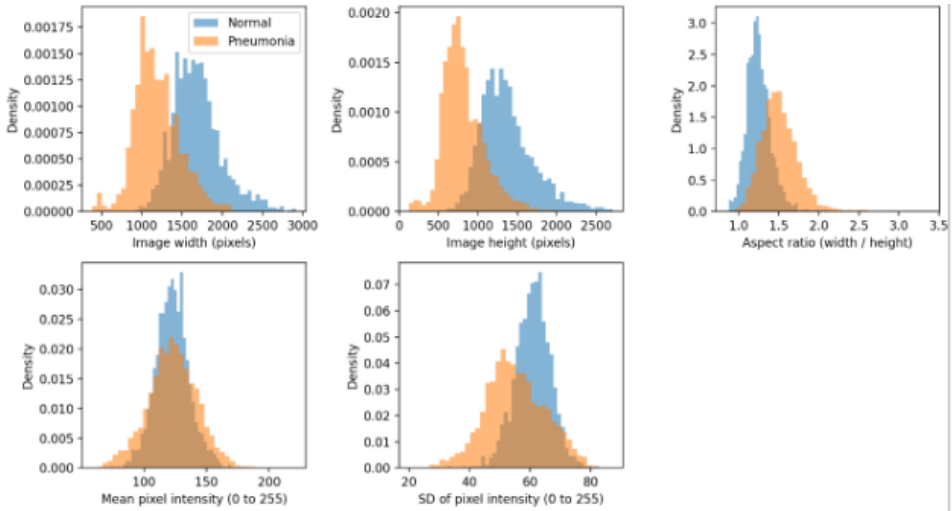

Figure 2. Distribution of image width and height (pixels), aspect ratio (width divided by height), mean pixel value and standard deviation of pixel values (0 to 255) for normal and pneumonia images, 5,824 cleaned images.

Figure 2 and table 4show compare the two classes. Normal images are larger than pneumonia images. The mean height is 1,384 against 824 pixels and the mean width is 1,692 against 1,201 pixels. The aspect ratio is lower for normal images (1.25 against 1.51). The mean pixel value is almost equal (122.8 and 123.0), but the spread of pixel values is higher in normal images (61.0 against 55.2). A classifier that sees only these five numbers, with no lung information, reached a test AUC of 0.917 with logistic regression and 0.934 with gradient boosting. Image height alone reached 0.908, width 0.876 and aspect ratio 0.841, while the mean pixel value reached only 0.554. These differences probably come from how images are acquired. The networks do not see the original size, because all images are resized to 224 by 224 pixels. They can still see the effect of the stretch and the image style. To test the effect of shape, the networks are trained with random stretching of the images, and the AUC did not drop (0.996 for ResNet50 and 0.997 for MobileNetV3 Large). A frame only input also gave

an AUC of 0.982 to 0.986. Image size and shape therefore do not fully explain the network scores, but the classes differ more than the lungs.

Table 4. Image properties by class for the 5,415 images used in the experiments, mean (SD), and test AUC of each property alone on the 826 test images

Property	Normal	Pneumonia	AUC alone
Height (pixels)	1,384 (340)	824 (271)	0.908
Width (pixels)	1,692 (302)	1,201 (285)	0.876
Aspect ratio (width / height)	1.25 (0.15)	1.51 (0.24)	0.841
Mean pixel value (0 to 255)	122.8 (13.5)	123.0 (19.7)	0.554
SD of pixel values (0 to 255)	61.0 (5.8)	55.2 (10.1)	0.702
All five, logistic regression			0.917

Lung Segmentation and Mask Quality:

The UNet reached a Dice score of 0.9629 and an IoU of 0.9297 on the 105 held out adult pairs. These numbers do not show how well it works on children. On the pediatric images the masks cover on average 20.4 percent of the image (SD 5.9 percent). The automatic check flagged 409 of 5,824 masks (7.0 percent). Figure 3 shows random examples of masks. Table 5 gives details and Table 6 gives the reasons for the flags. The flag rate is 6.6 percent in training, 6.5 percent in validation and 9.5 percent in the test split. In the test split, 5 of 237 normal images (2.1 percent) and 82 of 676 pneumonia images (12.1 percent) are flagged. Almost all flags came from a small lung area (Table 6).

Among pneumonia images, 368 (8.7 percent) had a lung area below 12 percent and 17 (0.4 percent) had no lung found. Among normal images, 24 (1.5 percent) had an area below 12 percent. No mask is above 65 percent, and the largest is 47.2 percent. Bacterial pneumonia is flagged more often (11.6 percent) than viral pneumonia (4.4 percent) and normal images (1.5 percent). The cause is not tested. Dense opacity can hide the lung edge and make the mask small, and image framing also changes the lung share of the picture.

The flag rate depends on the limit. With a lower limit of 8 percent, the flag rates are 0.2 percent for normal and 2.8 percent for pneumonia images, and with 15 percent they are 7.7 and 20.4 percent. Removing flagged images lowers the pneumonia share of the test split from 74.0 to 71.9 percent. Because flagged images are mostly pneumonia with a small mask, the kept test set may hold fewer hard pneumonia cases. The experiments are not repeated without the filter. The manual check also showed that the area rule is crude. It flagged only 6 of 14 failed masks, and some flagged masks looked acceptable.

Pediatric Specific Validation:

No expert lung masks exist for the pediatric images used here, so Dice and IoU could not be computed on them. Three indirect checks are used instead. First, the mean lung area on pediatric images is 20.4 percent (SD 5.9 percent) and no mask covered more than 47.2 percent. Second, the automatic check flagged 409 masks (7.0 percent). Third, the author checked 100 random masks by eye. Of these, 75 are good, 11 are acceptable and 14 failed. The automatic check flagged 6 of the 14 failed masks, so some failed masks remain in the experiments. This visual check is not an expert review. Segmentation quality depends on the diversity of the training data [41]so adult masks may fail more often on children. This is discussed below.

Table 5. Result of the automatic lung mask check

Group	Images	Flagged	Flag rate	Kept
Training	4,069	267	6.6%	3,802
Validation	842	55	6.5%	787
Test (all)	913	87	9.5%	826

Test, normal	237	5	2.1%	232
Test, pneumonia	676	82	12.1%	594
All images	5,824	409	7.0%	5,415

Table 6. Reasons for masking flags by class, with flag rates at area limits of 8, 12 and 15 percent (all 5,824 images)

Group	Area below 12%	No lung found	Flag rate	Flag rate at 8%	Flag rate at 15%
Normal (1,579)	24 (1.5%)	0	1.50%	0.20%	7.70%
Pneumonia (4,245)	368 (8.7%)	17 (0.4%)	9.10%	2.80%	20.40%

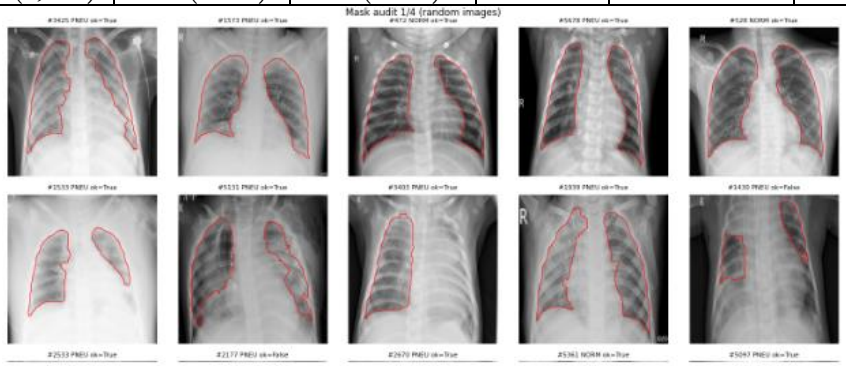

Figure 3. Random examples of lung masks. The red line is the predicted lung outline. The label ok=False marks a mask flagged by the automatic check.

Performance when the Input Version is the Same in Training and Test:

Table 7 and Figure 4 give the results. All twelve settings reach a high AUC. For original images the AUC is 0.997 to 0.998, and the accuracy is 0.954 to 0.970. For lung only images the AUC is 0.988 to 0.997, and the accuracy is 0.923 to 0.972. For background only images the AUC is 0.990 to 0.995, and the accuracy is 0.918 to 0.975. DenseNet121 on background only images give the highest accuracy of all settings (0.975). The gap between versions is small for DenseNet121 and EfficientNetB0, with an AUC difference of at most 0.003. It is larger for ResNet50 (0.006) and MobileNetV3 Large (0.009). The pattern of errors differs between versions. Lung only models have high sensitivity (0.984 to 0.994) and lower specificity (0.760 to 0.937). They label more normal images as pneumonia. Original models have a specificity of 0.974 to 0.987. Background only models have a sensitivity of 0.897 to 0.979 and a specificity of 0.945 to 0.986. Precision is 0.990 to 0.995 for original, 0.913 to 0.976 for lung only and 0.979 to 0.994 for background only models. A model that sees only the area outside the segmented lung fields therefore does almost as well as a model that sees the whole image.

Table 7. Test performance when training and test inputs are of the same version (mean ± SD over 3 seeds)

Model	Input version	Accuracy	Sensitivity	Specificity	F1 score	AUC
ResNet50	Original	0.966 ± 0.004	0.962 ± 0.006	0.974 ± 0.000	0.976 ± 0.003	0.997 ± 0.000
	Lung only	0.963 ± 0.007	0.984 ± 0.004	0.908 ± 0.036	0.974 ± 0.005	0.994 ± 0.001
	Background only	0.966 ± 0.002	0.974 ± 0.005	0.945 ± 0.004	0.976 ± 0.002	0.991 ± 0.001
DenseNet121	Original	0.970 ± 0.006	0.964 ± 0.008	0.986 ± 0.002	0.979 ± 0.005	0.998 ± 0.000

	Lung only	0.952 ± 0.005	0.994 ± 0.002	0.845 ± 0.019	0.968 ± 0.003	0.996 ± 0.000
	Background only	0.975 ± 0.003	0.979 ± 0.006	0.966 ± 0.004	0.983 ± 0.002	0.995 ± 0.000
EfficientNetB0	Original	0.954 ± 0.006	0.942 ± 0.009	0.987 ± 0.004	0.967 ± 0.004	0.997 ± 0.001
	Lung only	0.972 ± 0.006	0.986 ± 0.004	0.937 ± 0.032	0.981 ± 0.004	0.997 ± 0.000
	Background only	0.949 ± 0.012	0.935 ± 0.018	0.986 ± 0.004	0.963 ± 0.009	0.995 ± 0.000
MobileNetV3 Large	Original	0.969 ± 0.003	0.963 ± 0.003	0.983 ± 0.004	0.978 ± 0.002	0.997 ± 0.000
	Lung only	0.923 ± 0.012	0.987 ± 0.001	0.760 ± 0.040	0.949 ± 0.008	0.988 ± 0.002
	Background only	0.918 ± 0.013	0.897 ± 0.021	0.971 ± 0.009	0.940 ± 0.010	0.990 ± 0.002

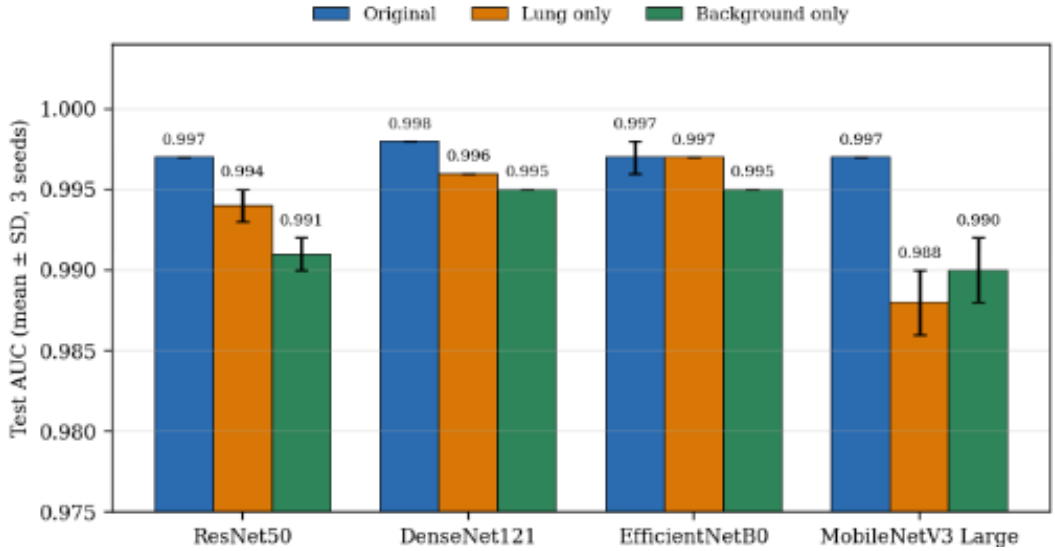

Figure 4. Test AUC when training and test inputs are of the same version.

Table 8 gives bootstrap confidence intervals for the seed 0 models. For DenseNet121 and EfficientNetB0 the intervals of the three versions overlap. For MobileNetV3 Large the interval for original images (0.995 to 0.999) does not overlap with the lung only interval (0.979 to 0.991). Most differences between versions are therefore small compared with the sampling noise of an 826-image test set.

Table 8. AUC with bootstrap 95% confidence interval for seed 0 (same version in training and test)

Model	Original	Lung only	Background only
ResNet50	0.997 (0.993 to 0.999)	0.992 (0.988 to 0.996)	0.990 (0.983 to 0.996)
DenseNet121	0.998 (0.996 to 0.999)	0.996 (0.994 to 0.998)	0.995 (0.990 to 0.998)
EfficientNetB0	0.997 (0.995 to 0.999)	0.997 (0.995 to 0.999)	0.995 (0.989 to 0.999)
MobileNetV3 Large	0.997 (0.995 to 0.999)	0.985 (0.979 to 0.991)	0.991 (0.985 to 0.996)

Error Counts and Calibration:

Table 9 gives mean confusion matrix counts for the 232 normal and 594 pneumonia test images. Original models missed 21 to 35 pneumonia images (3.6 to 5.8 percent) and raised

3 to 6 false alarms (1.3 to 2.6 percent). Lung only models missed fewer pneumonia images (3 to 9) but raised more false alarms (15 to 56, which is 6.3 to 24.0 percent). Background only models missed 12 to 61 pneumonia images (2.1 to 10.3 percent) and raised 3 to 13 false alarms (1.4 to 5.5 percent). The ECE is 0.017 to 0.076, and the Brier score is 0.021 to 0.061. Figure 5 shows reliability curves for ResNet50, pooled over three seeds. The middle bins hold only 28 to 178 predictions each, so the curves there are noisy. With original and background only inputs, the observed share of pneumonia is higher than the predicted probability in the middle range. With lungs only input the model is too confident about pneumonia.

Table 9. Mean confusion matrix counts, expected calibration error (ECE) and Brier score on the 826 test images (Mean over 3 seeds, same input version in training and test).

Model	Input	TN	FP	FN	TP	ECE	Brier
ResNet50	Original	226	6	22.3	571.7	0.029	0.026
ResNet50	Lung only	210.7	21.3	9.3	584.7	0.024	0.028
ResNet50	Background only	219.3	12.7	15.7	578.3	0.022	0.027
DenseNet121	Original	228.7	3.3	21.3	572.7	0.027	0.023
DenseNet121	Lung only	196	36	3.3	590.7	0.046	0.034
DenseNet121	Background only	224	8	12.3	581.7	0.017	0.021
EfficientNetB0	Original	229	3	34.7	559.3	0.054	0.035
EfficientNetB0	Lung only	217.3	14.7	8.3	585.7	0.023	0.022
EfficientNetB0	Background only	228.7	3.3	38.7	555.3	0.052	0.037
MobileNetV3 Large	Original	228	4	22	572	0.024	0.024
MobileNetV3 Large	Lung only	176.3	55.7	8	586	0.066	0.056
MobileNetV3 Large	Background only	225.3	6.7	61.3	532.7	0.076	0.061

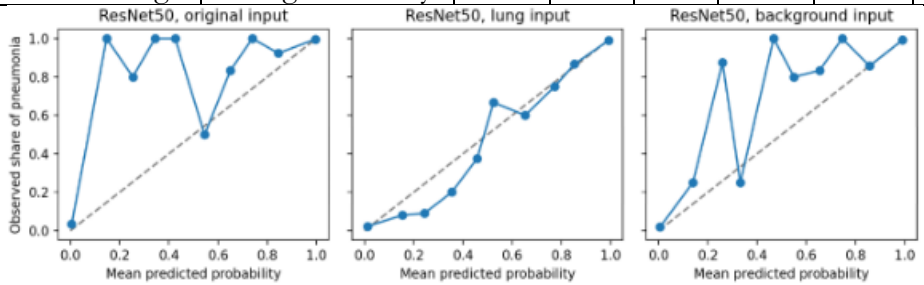

Figure 5. Reliability curves for ResNet50 pooled over three seeds. From left to right: original, lung only and background only input. The dashed line shows perfect calibration.

Cross Version Tests:

Table 10 and Figure 6 show what happens when a model is tested on a version it is not trained on. The result is not symmetric. Models trained on original images keep an AUC of 0.964 to 0.975 on background only images. On lung only images the AUC falls to 0.585 to 0.888. Models trained on background only images reach an AUC of 0.971 to 0.986 on original images. Models trained on lung only images reach only 0.672 to 0.865 on original images. The lung only and background only versions do not transfer to each other. For lung only models tested on background only images the AUC is 0.513 to 0.722. For background only models tested on lung only images it is 0.372 to 0.587. Two settings are below 0.5 (MobileNetV3 Large with 0.372 and ResNet50 with 0.466), which means the ranking is worse than random.

Accuracy and specificity need care in these tests. In many cross-version cells, the accuracy is close to 0.72, which is the share of pneumonia images in the test split (0.719), and the specificity is near zero. For example, DenseNet121 trained on lung only images have a specificity of 0.000 and an accuracy of 0.719 on background only inputs. At the fixed threshold

of 0.5 these models call almost every image pneumonia. AUC is the better measure here because it does not depend on the threshold.

Table 10. Test AUC for every pair of training and test input versions (mean ± SD over 3 seeds)

Model	Trained on	Tested on original	Tested on lung only	Tested on background only
ResNet50	Original	0.997 ± 0.000	0.888 ± 0.029	0.975 ± 0.004
	Lung only	0.756 ± 0.088	0.994 ± 0.001	0.628 ± 0.035
	Background only	0.986 ± 0.002	0.466 ± 0.018	0.991 ± 0.001
DenseNet121	Original	0.998 ± 0.000	0.708 ± 0.028	0.973 ± 0.007
	Lung only	0.865 ± 0.033	0.996 ± 0.000	0.722 ± 0.054
	Background only	0.986 ± 0.003	0.550 ± 0.137	0.995 ± 0.000
EfficientNetB0	Original	0.997 ± 0.001	0.585 ± 0.051	0.966 ± 0.008
	Lung only	0.679 ± 0.104	0.997 ± 0.000	0.513 ± 0.047
	Background only	0.974 ± 0.013	0.587 ± 0.094	0.995 ± 0.000
MobileNetV3 Large	Original	0.997 ± 0.000	0.648 ± 0.077	0.964 ± 0.011
	Lung only	0.672 ± 0.028	0.988 ± 0.002	0.692 ± 0.032
	Background only	0.971 ± 0.003	0.372 ± 0.106	0.990 ± 0.002

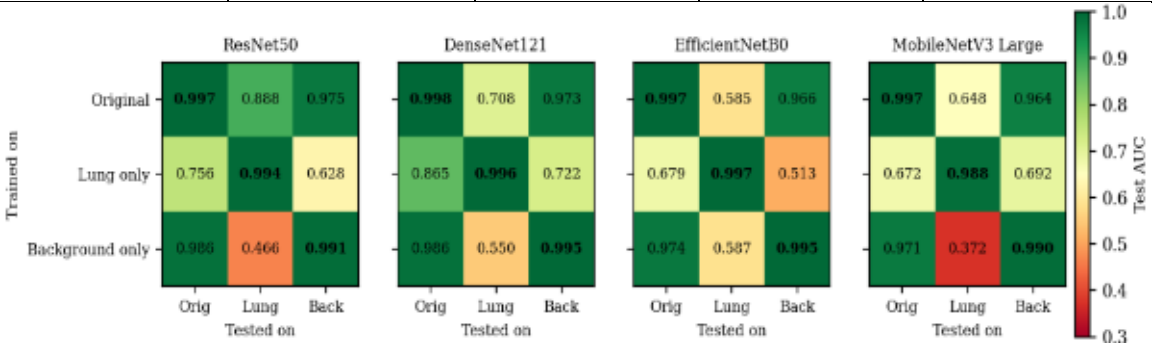

Figure 6. Test AUC for every pair of training and test input versions. Rows show the training version and columns show the test version.

Statistical Tests:

Table 11 summarizes the McNemar tests. All 24 tests are significant after the Holm adjustment, and the largest adjusted p value is 0.007. In every comparison the original input gave more cases that are correct only on the original image than cases that are correct only on the changed image. These tests show that the model decisions change when the input changes. They do not show by themselves that the model relies on a region, because the changed inputs also look unlike the training images.

Table 11. McNemar tests for models trained on original images (seed 0 counts, largest Holm adjusted p value over 3 seeds)

Model	Comparison	Correct only on original	Correct only on changed	Largest adjustedP
ResNet50	Original vs lung only	217	18	< 0.001
ResNet50	Original vs background only	61	11	0.002
DenseNet121	Original vs lung only	224	28	< 0.001
DenseNet121	Original vs background only	109	26	< 0.001
EfficientNetB0	Original vs lung only	238	31	< 0.001
EfficientNetB0	Original vs background only	49	20	0.007

MobileNetV3 Large	Original vs lung only	205	21	< 0.001
MobileNetV3 Large	Original vs background only	43	13	< 0.001

Attention Analysis:

For the ResNet50 model trained on original images, 31.5 percent of the GradCAM attention fell inside the lung mask on average. In the test split the masks cover 21.0 percent of the image. Attention is therefore about 1.5 times more concentrated in the lungs than expected by chance, but about 68 percent of it lies outside the lung masks. Table 12 and Figure 7 give value by class and by correctness. Correctly classified pneumonia images have a higher share inside the lungs (0.341) than correctly classified normal images (0.258). Misclassified pneumonia images have the lowest share (0.228, ratio 0.97), which is close to chance. Statistical evidence for the attention analysis is computed from the group means and standard deviations in table 12. For correctly classified pneumonia images, the attention share inside the lungs is 0.341 (95 percent CI 0.333 to 0.349) and the ratio to the lung area is 1.70 (1.66 to 1.74). For misclassified pneumonia images, the share is 0.228 (0.185 to 0.271) and the ratio is 0.97 (0.78 to 1.16), which does not differ from chance ($p = 0.75$). Within pneumonia images, correct predictions had a higher share than incorrect ones (Welch test, $t = 5.3$, $p < 0.001$, Cohen's $d = 1.1$). Correctly classified normal images had a share of 0.258 (0.250 to 0.266) and a ratio of 1.19 (1.16 to 1.22). The ratio is above one in all groups except misclassified pneumonia images. These results come from one model and one seed. GradCAM for other models and training runs is not computed, and expert verified lung regions are not available. Only 32 test images are misclassified, 6 normal and 26 pneumonias. Figure 8 shows random examples. In many normal images the heat sits over the heart and the center of the chest, which is outside the lung masks. In many pneumonia images the heat covers one lung, and in some it also covers a second area outside the lungs.

Table 12. GradCAM attention inside the lung mask for ResNet50 (original images, seed 0)

Group	Images	Attention inside lung	Lung area share	Ratio
Normal, correct	226	0.258 ± 0.063	0.221	1.19 ± 0.26
Normal, incorrect	6	0.383 ± 0.132	0.213	1.79 ± 0.50
Pneumonia, correct	568	0.341 ± 0.102	0.205	1.70 ± 0.46
Pneumonia, incorrect	26	0.228 ± 0.106	0.239	0.97 ± 0.47
All test images	826	0.315	0.210	

Control Experiments:

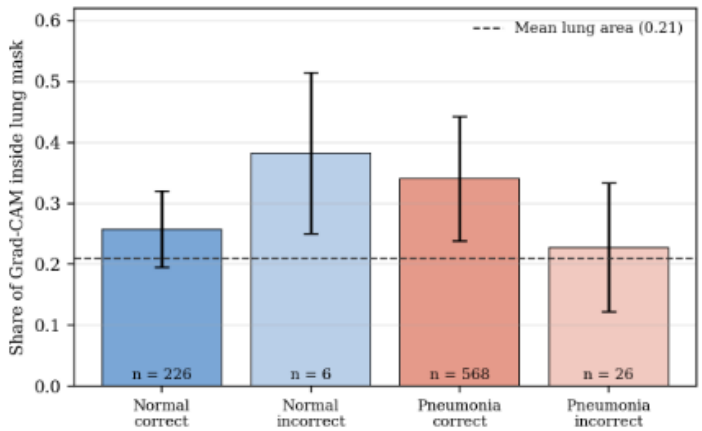

Figure 7. Share of GradCAM attention inside the lung mask by class and correctness.

Controls are run with ResNet50 and MobileNetV3 Large and three seeds each, with the same settings as the main experiments. A frame only input, which keeps a 20-pixel border

covering 32.5 percent of the image, gave an AUC of 0.986 and 0.982. A lung shaped window taken from another image gave 0.982 and 0.976 for lung only input, against 0.994 and 0.988 for the true lung masks. A larger margin around the lungs did not change the background only result. Random stretching of the images in training did not lower the AUC. Class information is therefore present in many parts of the image. The true lung adds about 0.01 in AUC beyond a lung shaped window from another image. The overlap between the wrong masks and the true lung masks is not measured, so the effect of the true lung may be larger. Table 13 gives the AUC of all controls next to the main results.

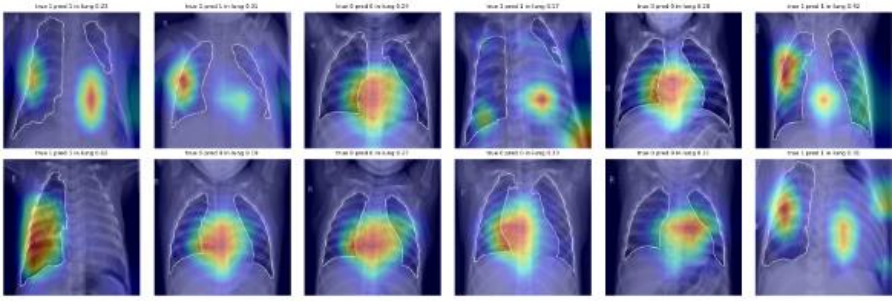

Figure 8. Random GradCAM examples for ResNet50. The white line is the lung mask. The title gives the true class, the predicted class and the share of attention inside the lung.

Table 13. Test AUC (mean ± SD over 3 seeds) for control experiments next to the main results.

Setting	ResNet50	MobileNetV3 Large
Original (main)	0.997 ± 0.000	0.997 ± 0.000
Lung only (main)	0.994 ± 0.001	0.988 ± 0.002
Lung only, wrong mask	0.982 ± 0.001	0.976 ± 0.002
Background only, margin 11 (main)	0.991 ± 0.001	0.990 ± 0.002
Background only, margin 31	0.993 ± 0.001	0.990 ± 0.001
Background only, margin 51	0.991 ± 0.000	0.989 ± 0.001
Background only, wrong mask	0.990 ± 0.001	0.989 ± 0.001
Frame only	0.986 ± 0.000	0.982 ± 0.002
Original, aspect jitter	0.996 ± 0.001	0.997 ± 0.001

Discussion:

The results map to the objectives as follows. For the first objective, lung only inputs gave an AUC of 0.988 to 0.997. The lung fields alone are enough for high scores. For the second objective, background only inputs gave an AUC of 0.990 to 0.995. The segmented lung fields are not needed to reach high scores on this dataset. For the third objective, the cross-version tests show a clear asymmetry. Models trained on original images keep working when the lungs are removed but not when only the lungs are left. Models trained on background only images work on original images, while models trained on lung only images do worse on them. For the fourth objective, the ResNet50 model put about 1.5 times more attention on the lungs than their area would suggest, but most of its attention is outside the lungs.

These findings agree with earlier warnings. Models for chest radiographs can reach high scores with cues that are not the target disease [3][4][5]. Removing the lungs also left high scores in a COVID 19 study[6]. The present study shows the same pattern on the pediatric pneumonia dataset, and it adds a full train and test matrix and an attention measure.

The findings need careful reading for five reasons. First, the background only image is not free of useful information. It holds the heart, the central airway, the ribs, the soft tissue and the diaphragm. Some of this has real clinical meaning. The masks cover only about 20

percent of the image and leave out areas such as the part behind the heart. Some lung tissue is therefore still visible in the background only images. The masks also come from a network trained on adult images [24][25][26], and the quality of lung segmentation depends on the diversity of the training data [41]. Second, the two classes differ in image size and shape (Figure 2). A model can learn these cues even after resizing, and this test does not separate them from real signal. Third, mask failures are not equal across classes. In the test-split 12.1 percent of pneumonia masks are flagged against 2.1 percent of normal masks. The filter changes the class mix and may remove harder pneumonia cases. Fourth, a model tested on a version it never saw in training faces a shifted input. The large drops in the cross-version tests are partly an effect of this shift. The cross-version AUC values are a warning sign and not an exact measure of reliance. Fifth, the GradCAM results come from one model and one seed, and the map has a coarse grid.

Taken together, the results support a limited claim. High scores on this dataset can be reached without the segmented lung fields, and models trained on whole images use information from outside the lung fields. The results do not prove that the models learned shortcuts. They also do not show that the models read the lungs. The results have a practical meaning. Studies that use this dataset should report masking controls next to accuracy. They should use patient grouped splits and test on images from other hospitals. Clinical claims should not be made from scores on this dataset alone.

The data come from one center and one age group. The lung masks are not checked by a radiologist, and the author checked only one sample. Image size and shape cues are not separated from other cues. Future work should add a second pediatric dataset from another hospital, masks checked by a radiologist, and GradCAM results for more models, more seeds and finer maps. It should also add a second pediatric dataset from another hospital, masks checked by a radiologist, and GradCAM results for more models, more seeds and finer maps. Pediatric lung masks checked by a radiologist, with Dice and IoU on 50 to 100 images, are needed and are left for future work.

Practical Implications:

The results show that a high score on this dataset is weak evidence that a model reads the lungs. Three findings matter for practice. First, networks reached an AUC of 0.98 or more from the lung fields alone, from the outside of the lung fields and from a frame only input. Second, a model with five image numbers reached 0.934. Third, models trained on one input version failed on the others. Developers of pediatric pneumonia models should therefore test what the model uses before they claim that it reads the lungs. Lung masks are a cheap way to run such tests.

Recommendations:

External validation. Test every model on pediatric images from other hospitals and devices before any claim of generalization. This dataset comes from one center.

Segmentation. Use lung masks to audit models and not only to improve scores. Check the masks on pediatric images with a radiologist, report Dice and IoU on pediatric masks and report mask failures for each class. Do not rely on a lung area rule alone.

Image standardization. Report image size, aspect ratio and intensity for each class. Keep the aspect ratio when images are resized or use random stretching in training. Fit a model on these numbers alone as a baseline.

Splits and controls. Split by patient, remove duplicates and report masking controls next to accuracy.

Clinical testing. Use a reader study with radiologists and a prospective test in the clinic. Check calibration and the number of missed pneumonia cases at the chosen threshold. Here, original models missed 21 to 35 of 594 pneumonia images, and the ECE is 0.017 to 0.076.

Conclusion:

This study tested what pediatric pneumonia classifiers use in chest radiographs. A UNet found the lung fields and three versions of each image are built. Four networks are trained and tested on all versions with a patient group split. Networks reached an AUC of 0.988 to 0.997 with the lung fields only and 0.990 to 0.995 with the segmented lung fields removed. Controls gave 0.982 to 0.986 from an outer frame only, and five image numbers gave 0.934. Models trained on whole images lost much of their ranking ability when only the lungs are left. GradCAM placed about 68 percent of its attention outside the lung masks. High scores can be reached without the segmented lung fields, and the classes differ in how the images are acquired.

Limitations: These findings do not show that lung information is unnecessary. The masks are incomplete, because the lungs are not fully covered and some failed masks remain. Exclusions are unequal across classes, 12.1 percent of pneumonia and 2.1 percent of normal test images. The data come from one center and one age group. The lung masks are not checked by a radiologist. The GradCAM results come from one model and one seed. Future research. Studies should add external pediatric data, masks checked by a radiologist, GradCAM for more models and seeds, and tests without the mask filter.

Acknowledgement:

The authors thank the Department of Data Science & AI, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), for providing the resources and support needed to carry out this research.

Author's Contribution:

Muhammad Adeel Abid is responsible for conceptualization, methodology, and analysis. Saman Riaz and Fatima Farhan are responsible for data preparation and manuscript preparation.

Conflict of interest:

The authors declare that there are no conflicts of interest.

Project details:

Not applicable.

References:

[1] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/J.CELL.2018.02.010.

[2] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs," *Appl. Sci. 2018, Vol. 8, Page 1715*, vol. 8, no. 10, p. 1715, Sep. 2018, doi: 10.3390/APP8101715.

[3] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, Nov. 2018, doi: 10.1371/JOURNAL.PMED.1002683.

[4] R. Geirhos *et al.*, "Shortcut Learning in Deep Neural Networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2023, doi: 10.1038/s42256-020-00257-z.

[5] A. J. DeGrave, J. D. Janizek, and S. I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell. 2021 37*, vol. 3, no. 7, pp. 610–619,

May 2021, doi: 10.1038/s42256-021-00338-7.

[6] G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from X-ray images," *Inf. Fusion*, vol. 76, pp. 1–7, Dec. 2021, doi: 10.1016/j.inffus.2021.04.008.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 618–626, Dec. 2017, doi: 10.1109/ICCV.2017.74.

[8] A. Saporta *et al.*, "Benchmarking saliency methods for chest X-ray interpretation," *Nat. Mach. Intell. 2022 410*, vol. 4, no. 10, pp. 867–878, Oct. 2022, doi: 10.1038/s42256-022-00536-x.

[9] N. Arun *et al.*, "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging," *Radiol. Artif. Intell.*, vol. 3, no. 6, p. e200267, Nov. 2021, doi: 10.1148/RYAI.2021200267.

[10] K. Shahi and A. Bagale, "Weakly Supervised Pneumonia Localization from Chest X-Rays Using Deep Neural Network and Grad-CAM Explanations," Dec. 2025, doi: 10.54364/JAIAI.2024.1126.

[11] R. Ridwan, "XAI-Guided Analysis of Residual Networks for Interpretable Pneumonia Detection in Paediatric Chest X-rays," Jul. 2025, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2507.18647v1

[12] A. Saber, A. Fateh, P. Parhami, A. Siahkarzadeh, M. Fateh, and S. Ferdowsi, "Efficient and Accurate Pneumonia Detection Using a Novel Multi-Scale Transformer Approach," Oct. 2025, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2408.04290v1

[13] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, and J. Schrouff, "Detecting shortcut learning for fair medical AI using shortcut testing," *Nat. Commun. 2023 141*, vol. 14, no. 1, pp. 4314-, Jul. 2023, doi: 10.1038/s41467-023-39902-7.

[14] I. Banerjee *et al.*, "'Shortcuts' Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation," *J. Am. Coll. Radiol.*, vol. 20, no. 9, pp. 842–851, Sep. 2023, doi: 10.1016/j.jacr.2023.06.025.

[15] B. G. Hill, F. L. Koback, and P. L. Schilling, "The risk of shortcutting in deep learning algorithms for medical imaging research," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 29224-, Nov. 2024, doi: 10.1038/s41598-024-79838-6.

[16] C. Vásquez-Venegas *et al.*, "Detecting and Mitigating the Clever Hans Effect in Medical Imaging: A Scoping Review," *J. imaging informatics Med.*, vol. 38, no. 4, pp. 2563–2579, Aug. 2025, doi: 10.1007/S10278-024-01335-Z.

[17] D. Juodelyte, Y. Lu, A. Jiménez-Sánchez, S. Bottazzi, E. Ferrante, and V. Cheplygina, "Source Matters: Source Dataset Impact on Model Robustness in Medical Imaging," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* , vol. 15384 LNCS, pp. 105–115, 2025, doi: 10.1007/978-3-031-82007-6_11/SAVE-RESEARCH.

[18] A. Jiménez-Sánchez, D. Juodelyte, B. Chamberlain and V. Cheplygina., "Detecting Shortcuts in Medical Images - A Case Study in Chest X-Rays," *IEEE 20th Int. Symp. Biomed. Imaging*, doi: doi: 10.1109/ISBI53787.2023.10230572.

[19] S. Aslani *et al.*, "Optimising Chest X-Rays for Image Analysis by Identifying and Removing Confounding Factors," *Lect. Notes Electr. Eng.*, vol. 810 LNEE, pp. 245–254, 2023, doi: 10.1007/978-981-16-6775-6_20/SAVE-RESEARCH.

[20] L. O. Teixeira *et al.*, "Impact of Lung Segmentation on the Diagnosis and Explanation

of COVID-19 in Chest X-ray Images," *Sensors 2021, Vol. 21, Page 7116*, vol. 21, no. 21, p. 7116, Oct. 2021, doi: 10.3390/S21217116.

[21] S. C. Haynes, P. Johnston, and E. Elyan, "Generalisation challenges in deep learning models for medical imagery: insights from external validation of COVID-19 classifiers," *Multimed. Tools Appl. 2024 8331*, vol. 83, no. 31, pp. 76753–76772, Feb. 2024, doi: 10.1007/S11042-024-18543-Y.

[22] T. Sourget, M. Hestbek-Møller, A. Jiménez-Sánchez, J. Junchi Xu, and V. Cheplygina, "Mask of Truth: Model Sensitivity to Unexpected Regions of Medical Images," *J. Imaging Informatics Med. 2025 391*, vol. 39, no. 1, pp. 767–784, May 2025, doi: 10.1007/S10278-025-01531-5.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28/SAVE-RESEARCH.

[24] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imaging Med. Surg.*, vol. 4, no. 6, p. 475, Dec. 2014, doi: 10.3978/J.ISSN.2223-4292.2014.11.20.

[25] S. Candemir *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577–590, Feb. 2014, doi: 10.1109/TMI.2013.2290491.

[26] N. Gaggion *et al.*, "CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images," *Sci. Data 2024 111*, vol. 11, no. 1, pp. 511-, May 2024, doi: 10.1038/s41597-024-03358-1.

[27] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, Accessed: May 25, 2024. [Online]. Available: https://arxiv.org/abs/1912.01703v1

[28] P. Mooney., "Chest X Ray Images (Pneumonia)," *Kaggle,*, [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

[31] Mingxing Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv:1905.11946*, 2020, [Online]. Available: https://arxiv.org/abs/1905.11946

[32] A. Howard *et al.*, "Searching for MobileNetV3," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 1314–1324, May 2019, doi: 10.1109/ICCV.2019.00140.

[33] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *7th Int. Conf. Learn. Represent. ICLR 2019*, Nov. 2017, Accessed: Oct. 06, 2026. [Online]. Available: https://arxiv.org/pdf/1711.05101

[34] M. A. Abid., "Pneumonia detection from chest X ray," *Kaggle*.

[35] M. A. Abid., "Chest X ray updation 1," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/muhammadadeelabid/chest-x-ray-updation-1.

[36] M. A. Abid., "Chest X ray updation 2," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/muhammadadeelabid/chest-x-ray-updation-2.

[37] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nat. Methods 2024 212*, vol. 21, no. 2, pp. 195–212, Feb. 2024, doi: 10.1038/s41592-023-02151-z.

[38] B. Efron and R. J. Tibshirani, "An Introduction to the Bootstrap," *An Introd. to Bootstrap*, May 1994, doi: 10.1201/9780429246593.

[39] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.

[40] "Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 65-70. - References - Scientific Research Publishing." Accessed: Oct. 20, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=1798304

[41] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radiol. Exp.*, vol. 4, no. 1, p. 50, Dec. 2020, doi: 10.1186/S41747-020-00173-2.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.