A Resource Availability and Licensing

All datasets are available for download from the **CheXwhatsApp** Github page, `https://github.com/cheXwhatsApp/cheXwhatsApp`. The code, along with instructions for installing our Python package can also be found on the same page. The corresponding code for dataset processing and model training is maintained in a separate repository, `https://github.com/cheXwhatsApp/cheXwhatsApp_benchmark`. All datasets are hosted on Zenodo under individual DOIs, and are licensed under Creative Commons (CC) licenses. Full licenses, DOIs, and other details can be found on our Github page. The authors bear all responsibility in case of violation of rights, and confirm the CC licenses for the included datasets.

B Broader Impact

AI-based systems integrated with WhatsApp for the diagnosis of chest X-rays (CXR) hold significant potential in mHealth applications. These systems can facilitate remote medical consultations, improve accessibility to diagnostic services, and provide timely medical insights, especially in regions with limited access to healthcare facilities such as LMIC's. The integration of AI-based diagnostics with widely used platforms like WhatsApp can significantly enhance the accessibility and efficiency of healthcare services. This work addresses critical challenges faced by AI-based models in the context of WhatsApp-compressed CXR images, such as prediction instability, out-of-lung saliency, and localization instability. By introducing metrics like PI Score, OLS Score, and LI Score, it provides a framework for evaluating and improving the reliability of these models.

This paper marks a significant step forward by providing a paired dataset of 140,000+ original and `WhatsApp`-compressed CXR images, with different lung abnormalities. This dataset allows researchers to study the impact of image compression on model performance, streamlining the development of more robust AI models capable of handling image perturbations introduced by messaging apps like WhatsApp. The paired dataset serves as a valuable resource for researchers developing AI models for diagnosing chest X-rays on mobile platforms like WhatsApp.

The availability of this paired dataset and the accompanying evaluation tools is expected to inspire further research in the integration of AI with mobile health technologies. It could lead to the creation of similar datasets in other medical imaging fields, such as digital pathology, radiology, and cardiology, thereby broadening the scope and impact of AI in healthcare. These datasets and evaluation tools can aid researchers to create reliable AI models by addressing the instability issues when handling WhatsApp-compressed images. This is crucial for the practical deployment of AI systems in mHealth applications, ensuring that they provide dependable support to healthcare professionals and patients.

C Working with CheXwhatsApp

C.1 Developing and Benchmarking New Algorithms

The success of AI-based diagnosis for mHealth applications strongly depends on the availability of sufficient training data. Although several Chest X-Ray (CXR) datasets annotated with lung abnormalities or bounding boxes indicating regions of interest exist in the literature, there is no paired CXR dataset of original and `WhatsApp` images. **CheXwhatsApp** is the first dataset to address this gap with original and corresponding `WhatsApp`-compressed CXR images.

Reading and preprocessing. The original and `WhatsApp` compressed images are available in PNG or JPEG format. Labels or corresponding bounding box annotations associated with the dataset are provided in a seperate file. The benchmarking datasets and metadata files are provided in a format that is easy to read for most Python-based algorithms. Data loaders for reading and pre-processing **CheXwhatsApp** is available in Github repository (Section A).

Comparing algorithms. The challenges assocaited with AI-based models for mHealth applications such as *prediction instability*, *out-of-lung saliency* and *localization instability* are discussed in the main paper. The metrics to evaluate these challenges are *PI Score*, *OLS Score* and *LI Score*. These metrics (as defined in the main paper) can be evaluated for any model using the Python package open-sourced along with this paper.

Interpretation of results. For a model used for AI-based diagnosis in mHealth applications integrated with WhatsApp, the instability of the model due to image perturbation by WhatsApp can be evaluated using *PI Score*, *OLS Score* and *LI Score*. Higher values of these scores indicate a higher degree of instability.

D Dataset Preparation

NIH-WA. NIH-WA consists of original images in NIH [2] along with the corresponding WhatsApp compressed images. It is a multi-label dataset where patients can have normal CXR or be associated with one or more of the following 14 categories: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural-Thickening, and Hernia.

RSNA-WA. RSNA-WA consists of original images in RSNA[1] along with corresponding WhatsApp-compressed images.The RSNA[1] consists of 29,684 CXR images of patients infected with pneumonia and normal cases. Each CXR image contains bounding box information for patients with pneumonia. RSNA dataset is manually annotated by expert Radiologists.

Datset	Splits	Image Count		Size	
		Original	WhatsApp	Original	WhatsApp
NIH-WA	train	86,524	86,524	33GB	11.29GB
	validation	25,596	25,596	9.6GB	3.3GB
RSNA-WA	train	26,684	26,684	11GB	3.46GB
	validation	3000	3000	1.2GB	386MB

Table 1: Number and size of images in training and validation datasets of CheXP-WA, NIH-WA and RSNA-WA in **CheXwhatsApp**.

In NIH-WA and RSNA-WA, the training and validation sets follow a consistent structure. Original and WhatsApp-compressed images are stored in separate folders, with corresponding images having identical names. NIH-WA includes additional csv files for both training and validation sets. These csv files contain details like image names, view orientation(frontal/lateral and AP/PA) and associated lung abnormalities. For bounding-box datasets like RSNA-WA, a seperate csv file is included for the training and validation sets. This csv file contains bounding box coordinates for each image, including details like upper-left corner coordinates, width, and height. The number of images and their sizes for datasets- NIH-WA, and RSNA-WA can be found in Table 2

E Additional Benchmarking Results

E.1 Experiments using JPEG dataset

Models	PI Score (NIH)	
	Original	JPEG-compressed
DenseNet	15.74	18.41
InceptionNet	15.08	17.72
ResNet-50	14.60	17.12
ResNeXt-50	16.86	19.33

Table 2: PI Score for different techniques trained using original and JPEG-compressed images on the classification dataset - NIH in **CheXwhatsApp**. From the table, it is evident that models trained on JPEG-compressed images face higher **PIP** compared to those trained on original images.

Models	Datset	OLS Score(GradCAM)		OLS Score(GradCAM++)		OLS Score(EigenCAM)	
		Original	WhatsApp	Original	WhatsApp	Original	WhatsApp
DenseNet	NIH	18.79	26.69	30.79	41.25	46.06	60.66
	NIH-JPEG	38.96	39.40	46.31	47.19	61.65	62.77
Inception	NIH	37.75	47.45	52.33	62.21	95.58	96.70
	NIH-JPEG	50.46	50.62	56.32	63.25	99.78	99.78
ResNet	NIH	24.65	33.65	42.12	51.40	97.51	98.89
	NIH-JPEG	28.65	42.19	46.76	51.80	99.90	99.90
ResNeXt	NIH	33.12	42.56	61.24	69.39	99.67	99.79
	NIH-JPEG	41.50	43.02	69.15	71.36	99.57	99.68

Table 3: Results on original NIH and JPEG-compressed NIH images showing that models trained using JPEG-compressed images has higher **OLS** compared to those trained with original NIH.

Models	LI Score (RSNA)	
	Original	JPEG-compressed
Faster RCNN	12.43	15.57
YOLOv5	3.01	3.68
YOLOv8	4.02	4.53

Table 4: LI Score for different techniques trained using original and JPEG-compressed images on the object detection dataset - RSNA in **CheXwhatsApp**. From the table, it is evident that models trained on JPEG-compressed images face higher **LIP** compared to those trained on original images.

References

[1] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence*, 1(1), 2019.

[2] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.