frontiers | Frontiers in Radiology

TYPE Brief Research Report
PUBLISHED 08 January 2024
DOI 10.3389/fradi.2023.1274273

RoMIA: a framework for creating Robust Medical Imaging AI models for chest radiographs

Aditi Anand*, Sarada Krithivasan and Kaushik Roy

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States

Artificial Intelligence (AI) methods, particularly Deep Neural Networks (DNNs), have shown great promise in a range of medical imaging tasks. However, the susceptibility of DNNs to producing erroneous outputs under the presence of input noise and variations is of great concern and one of the largest challenges to their adoption in medical settings. Towards addressing this challenge, we explore the robustness of DNNs trained for chest radiograph classification under a range of perturbations reflective of clinical settings. We propose RoMIA, a framework for the creation of Robust Medical Imaging AI models. RoMIA adds three key steps to the model training and deployment flow: (i) Noise-added training, wherein a part of the training data is synthetically transformed to represent common noise sources, (ii) Fine-tuning with input mixing, in which the model is refined with inputs formed by mixing data from the original training set with a small number of images from a different source, and (iii) DCT-based denoising, which removes a fraction of high-frequency components of each image before applying the model to classify it. We applied RoMIA to create six different robust models for classifying chest radiographs using the CheXpert dataset. We evaluated the models on the CheXphoto dataset, which consists of naturally and synthetically perturbed images intended to evaluate robustness. Models produced by RoMIA show 3%–5% improvement in robust accuracy, which corresponds to an average reduction of 22.6% in misclassifications. These results suggest that RoMIA can be a useful step towards enabling the adoption of AI models in medical imaging applications.

KEYWORDS

medical imaging, artificial intelligence, artificial neural networks, robustness, radiology, chest radiographs

1 Introduction

Artificial Intelligence is transforming the field of medicine in many ways, with applications spanning from drug discovery to genomics and, most prominently, radiology. Since AI has been particularly successful in computer vision, one of its most promising applications is to medical imaging. Deep neural networks (DNNs), which are composed of several layers of artificial neurons, have demonstrated great success in computer vision tasks. These networks, particularly convolutional neural networks (CNNs), have explored for various medical imaging tasks, including diagnosis of diabetic retinopathy (1, 2), breast cancer and malignant lymph nodes from histopathological images (3), and pulmonary and cardiological conditions from chest radiographs (4). The recent wave of promising research has led to significant interest in

OPEN ACCESS

EDITED BY
Ulas Bagci,
Northwestern University, United States

REVIEWED BY
Bo Zhou,
Yale University, United States
Koushik Biswas,
Northwestern University, United States

*CORRESPONDENCE
Aditi Anand
✉ anand86@purdue.edu

RECEIVED 10 August 2023
ACCEPTED 18 December 2023
PUBLISHED 08 January 2024

CITATION
Anand A, Krithivasan S and Roy K (2024) RoMIA: a framework for creating Robust Medical Imaging AI models for chest radiographs.
Front. Radiol. 3:1274273.
doi: 10.3389/fradi.2023.1274273

deploying these technologies in clinical settings. However, there are many hurdles that must be crossed before we can realize this potential.

Medical imaging models are first trained on a training dataset, and then tested in field trials before being deployed. One major challenge in this process arises from the differences between the data on which the models are trained and the data that they encounter after deployment (5). AI models are known to be very brittle to input noise and variations (6), even ones that are imperceptible to humans (7). There are several scenarios where medical imaging models encounter noise or variations that can impact the accuracy of their predictions (8). One popular use of medical imaging models is for telemedicine in areas that have a lack of trained physicians, where smartphones are used to take photos of scans, which are then sent through messaging apps, introducing distortion and compression artifacts (9). Additionally, using imaging equipment made by different manufacturers or using different settings on the imaging equipment can create variations in the resulting images (10, 11). AI models have also demonstrated significant performance variation across different patient populations (12). Any of these factors can result in a model making inaccurate predictions (8).

Recent work has demonstrated that variations and noise in the input can significantly reduce the accuracy of medical imaging AI models (8, 10). Although there has been a large body of work in the AI community on improving the robustness of these models under noise and adversarial perturbations, very few efforts have focused on the medical domain. There are various unique challenges posed by the domain of medical imaging that make it essential to address robustness specifically in this context (8). As described above, the nature of input noise and variations is primarily due to equipment differences, telemedicine, patient population; sources of variation seen in other settings (background objects, lighting, occlusion, etc.) are less relevant in medical settings (9, 10, 12). Furthermore, due to regulations and higher safeguards applied to medical data, adversarial attacks may be much less of a concern in this setting relative to other settings.

In this paper, we propose RoMIA, a framework to create more robust medical imaging models. RoMIA consists of three main steps: Noise-added Training, Fine-tuning with Input Mixing, and DCT-based denoising. In Noise-added Training, a fraction of the images in the training dataset are transformed by adding noise in order to make the trained model more robust (13). Specifically, we find that transformations such as glare matte, moire, and tilt result in models that perform best on photographs of radiographs. In Fine-tuning with Input Mixing, we fine-tune the trained model using a small amount of data from a different source in order to improve the model's robustness (14). Since only limited data from additional sources are likely to be available in practice, we use input mixing to avoid overfitting during this stage. Finally, in DCT-based denoising, we remove higher-frequency components in the input images before they are passed to the model for classification (15). This is motivated by our observation that perturbations encountered in medical imaging settings largely impact the high-frequency components of the images that are not essential for classification.

We evaluate the RoMIA framework using six popular CNNs trained on the CheXpert dataset, which contains 224,316 chest radiographs of 65,240 patients from Stanford Hospital (4). The created models diagnose Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. For Fine-tuning with Input Mixing, we used 500 images from the ChestX-ray8 dataset from NIH (16). We evaluated the models using the CheXphoto dataset, which consists of 10,507 smartphone photos of chest radiographs from 3,000 patients (9). Our experiments indicate that a baseline model trained on the CheXpert dataset has an Area Under Receiving Operating Characteristic (AUROC) drop of 10%–14% when evaluated on the CheXphoto dataset. RoMIA creates models that improve AUROC by up to 5%, and reduces misclassifications by an average of 22.6%, underscoring its potential to create more robust medical imaging models.

1.1 Related work

Several research efforts have explored the use of CNNs for medical imaging. Building on these efforts, systems that support diagnosis are in various stages of deployment. These include systems for processing retinal scans (1, 2, 17), breast cancer detection (18), and skin cancer detection (19), among others. We focus our discussion on related efforts along two directions: those that explore CNN-based classification of chest radiographs and those that explore the robustness of medical imaging CNNs.

1.1.1 Prior work on chest radiograph classification

Chest radiographs are among the most commonly requested radiological examinations since they are highly effective in detecting cardiothoracic and pulmonary abnormalities. Automation of abnormality detection in chest radiographs can help address the high workload of radiologists in large urban settings on the one hand, and the lack of experienced radiologists in less developed rural settings on the other. This need was only exacerbated during the COVID-19 pandemic when healthcare systems were overwhelmed and chest radiographs were commonly used as a first-line triage method. Motivated by this challenge, several efforts have developed DNN models for processing of chest radiographs (20–28). These works have proposed key ideas including the use of pre-training with natural images (20), multi-modal fusion of radiographs with clinical data (22), the use of transformer networks for such multi-modal fusion (24), manual design (27) or automated neural architecture search (25) to find a suitable DNN architecture for chest radiograph classification, bio-inspired training algorithms for small training sets (26) and the use of a focal loss function to address the significant class imbalance that is often present in chest radiograph datasets (28). These efforts have demonstrated high accuracies in various chest radiograph classification tasks, promoting interest in their use in clinical practice. Supporting the development of DNN models for chest radiographs has been the curation of public datasets (4, 9, 16, 29).

1.1.2 Prior work on robustness of medical imaging AI models

It is well known that input variations, noise and adversarial perturbations can have a large negative impact on the accuracy of DNNs. For example, it has been shown that chest radiographs with added natural noise as well as the use of smartphone-captured photographs of radiographs caused significant degradation in accuracy (9). Another study found that DNN models trained on data from one hospital demonstrate considerably lower performance on data from a different hospital (10). Adversarial perturbations have also been shown to have a drastic impact on the accuracy of DNNs used in medical imaging (30, 31). These concerns, while broadly true of DNNs, are especially important for life-critical applications such as medical imaging. As a result, previous works have proposed and evaluated techniques to improve the robustness of medical imaging DNNs. The combination of large-scale supervised transfer learning with self-supervised learning was shown to improve the out-of-distribution generalization performance of medical imaging DNNs (32). The addition of Global Attention Noise during training (33), as well as adversarial training, where adversarial inputs are included in the training process (31), have been shown to improve the accuracy of medical imaging DNNs against adversarial attacks. Multi-task learning was used to address the specific challenges of prediction instability and explainability in the classification of smartphone photos of chest radiographs (21).

Our work makes the following contributions that go above and beyond the previous efforts. While noise-added training is a well-known technique to improve the robustness of neural networks (34) and has recently been applied to medical imaging specifically for adversarial robustness (31, 33), our work applies it to achieve robustness to natural sources of noise. Input mixing and DCT-based denoising have not been previously applied to the medical imaging domain to the best of our knowledge. Further, RoMIA is the first framework to combine these three techniques to improve robustness and to incorporate robustness improvement into all three key steps of the medical imaging AI pipeline (training, fine-tuning, and inference). Our results show that the combined use of all three techniques leads to substantially better accuracy than any of the techniques alone.

2 Materials and methods

In this section, we first describe the commonly used process for training medical imaging DNNs, and the challenges faced by such models due to input noise and variations. We then present the RoMIA framework to increase model robustness and the methodology used to evaluate it.

2.1 Pitfalls in conventional training methods

Typically, the creation of a medical imaging model starts with the collection of a large training dataset with training labels provided by physicians. In some cases, this may require years of data collection. For example, the CheXpert dataset of chest radiographs represents data collected over a period of 15 years (4). Next, a DNN is either trained from scratch or a model trained on a different computer vision dataset such as ImageNet (35) is transferred using the training data. The model may be evaluated on held-out or entirely different datasets, and then deployed. When deployed, the model may be applied to data that contains noise or variations. Frequently, this leads to significant degradation in model performance (8).

2.2 RoMIA framework

Figure 1 describes the RoMIA framework to train more robust medical imaging models. We modify the standard model creation flow by adding three main components: Noise-added Training, Fine-tuning with Input Mixing, and DCT-based denoising.

2.2.1 Noise-added training

In *Noise-added Training*, we introduce synthetic perturbations (noise) into the training data that mimic those observed in medical settings. We evaluated the following transformations:

- *Glare matte:* A filter designed to emulate the effect of glare observed when displaying the image on a matte screen.
- *Moire:* A filter designed to simulate the Moire effect, which produces repetitive interference patterns such as lines or stripes on the image due to limited resolution.
- *Tilt:* This transformation simulates a change in perspective that could result when a photograph of a medical image is taken using a device such as a smartphone (11).
- *Brightness* and *Contrast*: These transformations simulate changes to the settings in imaging equipment.
- *Blur:* This transformation simulates the loss in sharpness of the image due to motion of the patient during capture.

Among all evaluated transformations, we found that the first three were the most effective in creating more robust models. It bears mentioning that this result may be due to the fact that we evaluate robustness on the CheXphoto dataset. Hence, the transformations that introduce the most photographic noise may provide the best robustness. Notwithstanding this, the framework is extensible and additional transformations can be added to diversify the suite we have implemented.

We consider two strategies for applying noise to the training dataset: a specific percentage of the images in the dataset are injected with noise and either added (thereby expanding the dataset) or replace their original versions (thereby preserving the size of the dataset). We refer to these strategies as *augmentation* and *replacement*, respectively. All training hyperparameters (learning rate, batch size, optimizer, epochs, etc.) were kept unchanged.

2.2.2 Fine-tuning with input mixing

In *Fine-tuning with Input Mixing*, we fine tune the model with a very small amount of data from a different source to improve the model's robustness. Since acquiring large amounts of additional training data may be challenging in practice, we limited ourselves

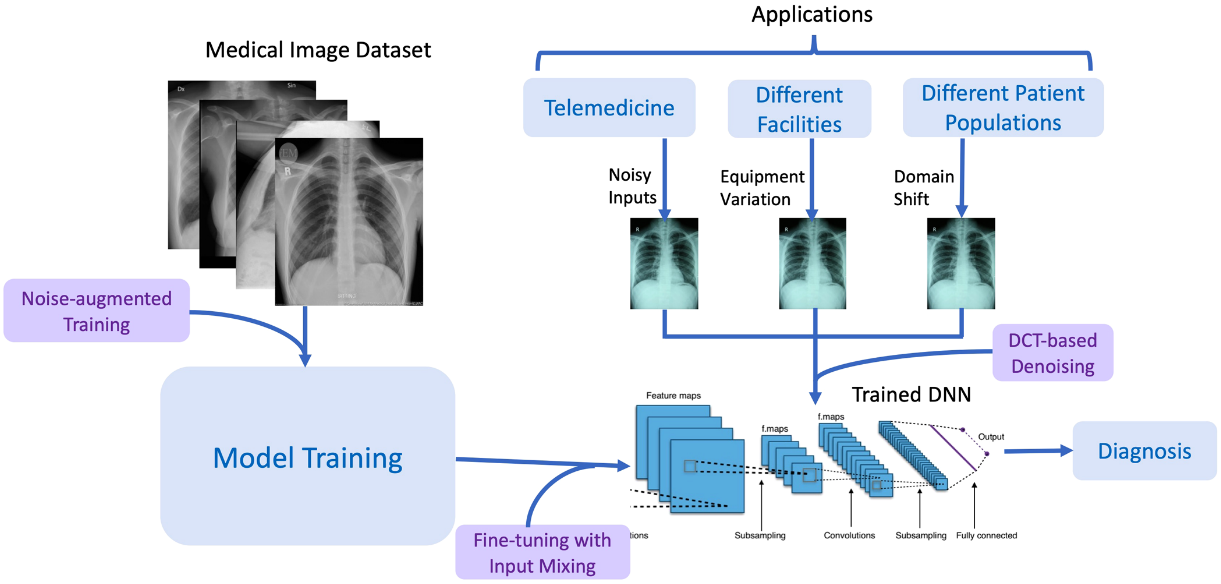


FIGURE 1
Overview of the RoMIA framework to create robust medical imaging AI models. The chest radiographs shown are from the CheXpert dataset (4).

to just 500 images, which correspond to around 0.22% of the original training set. While input mixing has been proposed in the literature as a data augmentation strategy, our contribution is the specific use of input mixing during the fine-tuning step and its evaluation in the context of medical imaging models. For our experiments, we draw these images at random from the ChestX-ray8 dataset from NIH (16). One challenge with using a very limited amount of data is that it could easily lead to overfitting. In order to prevent this, we use input mixing, a well-known technique where two images are combined into a composite input that contains information from both. Minimizing loss on mixed inputs has been shown to approximately correspond to maximizing robust accuracy (36). We mixed the additional data with images from the original training set for the fine- tuning phase. We considered three different mixing strategies that have been proposed in the literature. With CutMix (14), a randomly selected patch of one input image is placed into another. With MixUp, the pixels of two images are averaged in a weighted manner to construct a composite image. In both cases, the labels from the two images being mixed are also combined to derive the target label for the composite input (36, 37). In AugMix, images are mixed with augmented versions of themselves, so the label does not change (15). We mix the 500 images from ChestX-ray8 with 1,000 randomly selected images from the CheXpert training set and fine-tune the model for 3 epochs with these mixed inputs. All other hyperparameters such as the learning rate and optimizer were the same as those used in the training stage.

2.2.3 DCT-based denoising

DCT-based denoising is based on the insight that most sources of noise disproportionately affect the high-frequency components of an image (38). This is shown in Figure 2, which plots the percent difference in the top and bottom 1% of frequencies of the original and noisy images from the CheXpert (4) and CheXphoto (9) datasets, where the noisy images were produced using synthetic digital perturbations, synthetic photographic perturbations, and photos taken of the images with a smartphone camera. During inference, we add a preprocessing stage to the model which uses DCT (discrete cosine transform) to transform the image into the frequency domain, then removes a set percentage of high-frequency components, and finally computes the inverse DCT (15, 39). The percentage of high-frequency components to be removed from an image (denoted by η) is determined through an experiment where a small fraction of the training set (CheXpert, in our experiments) is subject to DCT-

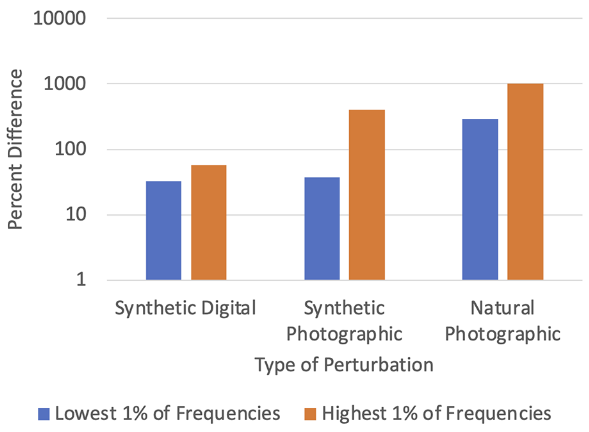


FIGURE 2
Difference between clean (CheXpert) and noisy (CheXphoto) images in high and low frequencies.

based denoising for different values of η. For each model, the largest value of η (which corresponds to the most aggressive denoising) that keeps the AUROC to within 0.005 of the original accuracy (where $\eta = 0$) is chosen. Optimizing the hyperparameter η ensures that the frequencies removed do not significantly interfere with the features used by the model for classification.

To summarize, the proposed flow to create robust medical imaging models consists of transferring a model trained on ImageNet to the target medical imaging dataset using noise-added learning, then fine-tuning the resulting model with input mixing, then finally adding a DCT-based denoiser to the model before deployment.

2.3 Experimental setup

We implemented the RoMIA framework using the PyTorch (40), TensorFlow (41), libAUC (42), and OpenCV (43) libraries. We applied the framework to create models for classification of chest radiographs. The base models were selected from popular image classification DNNs trained on the ImageNet (35) dataset (see Figure 3A). Note that all the networks are Convolutional Neural Networks (CNNs), since these are the most popular type of DNN used for image classification tasks. We specifically created a model to detect Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. Accordingly, the final fully connected layer of each base model was removed and replaced with a layer with five outputs. These models were then transferred using the CheXpert (4) dataset, which contains 224,316 chest radiographs of 65,240 patients from Stanford Hospital. For the fine-tuning step, we randomly selected 500 images from NIH's ChestX-ray8 (16) dataset. The learning rate used for both the transfer and fine-tuning steps was 0.0001, number of epochs was 3 with a batch size of 32, and weight decay was 10^{-5}. The Adam optimizer and cross-entropy loss were used. For the MixUp (36) strategy, we use a beta distribution to select values between 0.4 and 0.6 to determine λ, the image mixing ratio. We evaluated the models on the CheXphoto (9) dataset, which consists of 10,507 natural photos and synthetic transformations of chest radiographs from 3,000 patients. Since our noise-added training step uses transformations similar to those in CheXphoto, we only perform our evaluations on the natural photographs. We repeated each of our experiments five times with different random seeds.

3 Results

In this section, we present results from evaluation of models created using the RoMIA framework. We first present the difference in AUROC of the baseline models when evaluated on a subset of CheXpert and CheXphoto images. Next, we present the performance of models trained using RoMIA and compare them to the baseline models. Subsequently, we perform an ablation study to investigate the contribution of each of the three components (Noise-added learning, Fine-tuning with input mixing, DCT-based denoising) to the overall improvement in robustness. We then explore different dataset transformation techniques for the Noise-added Training step and evaluate their impact on the model performance. We also compare the performance between different strategies for input mixing in the fine-tuning step. Finally, we explore the determination of the parameter η which controls the percent of high-frequency components removed from the input during DCT-based Denoising.

3.1 Robustness of baseline models

A key motivation for this work is that baseline models trained on a certain dataset perform significantly worse on similar datasets with added noise. To demonstrate this in the context of CheXpert and CheXphoto, we study the differences in AUROC of a baseline model trained on CheXpert and then applied to both CheXpert and CheXphoto data. Figure 3B presents the AUROC scores for the baseline models on the CheXpert and CheXphoto data. The figure shows a degradation of 10%–14% in AUROC across all six models, underscoring the need to create more robust models in the context of medical imaging.

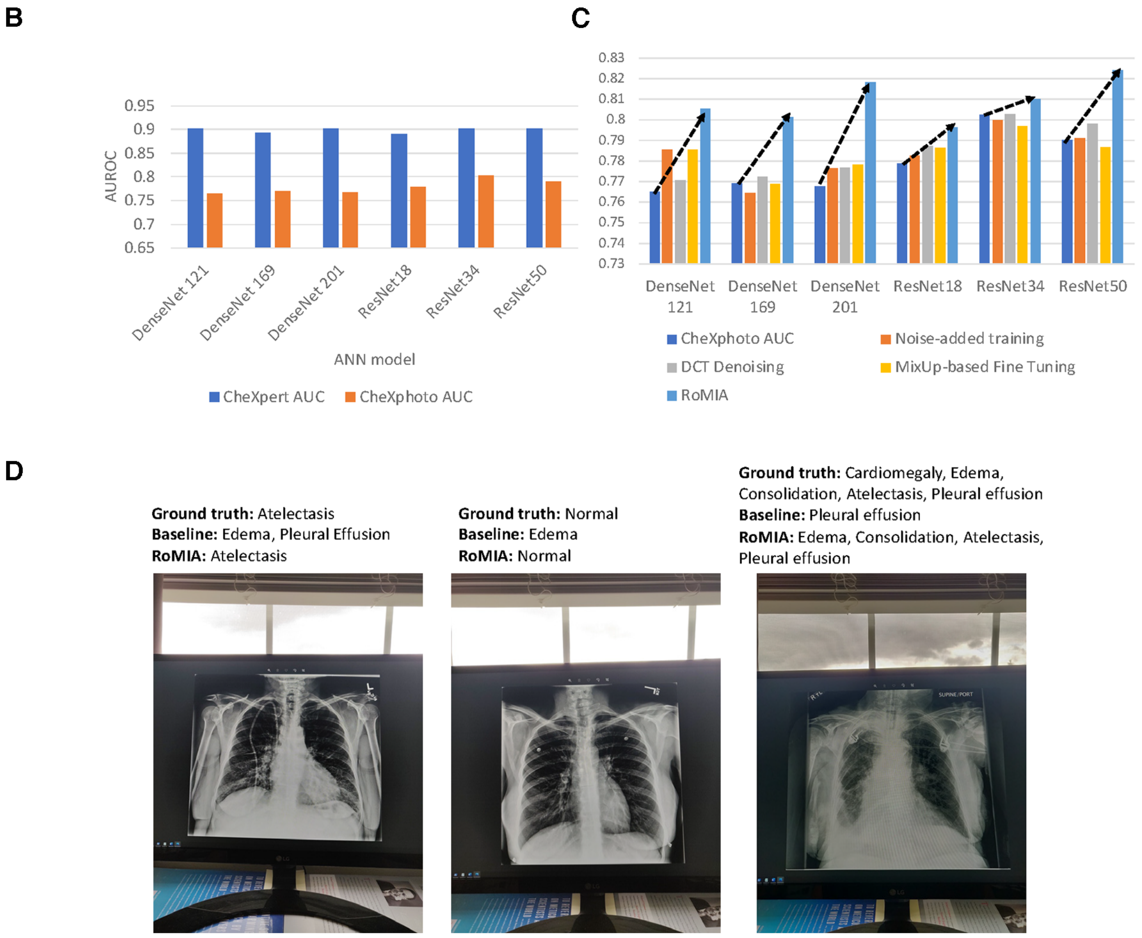
3.2 Overall improvements from RoMIA and ablation study

The RoMIA framework consists of three techniques to improve robustness, so we conduct an ablation study to evaluate each component. Figure 3C shows the baseline accuracy, the results of the ablation study (applying each of the three techniques in RoMIA individually), and the resulting AUROC score when all three techniques are combined in RoMIA. To capture the benefits of the proposed framework, we first look solely at the CheXphoto AUROC values for the baseline and RoMIA models. We observe around 3%–5% improvement in AUROC, which corresponds to an average reduction in misclassifications by 22.6%, suggesting that the proposed framework is capable of creating substantially more robust models. We also observe a larger improvement in robustness on deeper models, such as ResNet50 and DenseNet201. We hypothesize that this is because deeper models can better learn the more diverse training data which they are presented in the RoMIA framework. In order to evaluate the statistical validity of the results, we repeated the training runs for the baseline and RoMIA models with 10 additional random seeds. We performed a one-tailed paired t-test and concluded that the improvements were statistically significant with $p < 0.01$. Figure 3D presents examples of inputs that are misclassified by the baseline model but correctly classified by RoMIA.

Figure 3C also presents the results of our ablation study to evaluate each of the three components in the proposed framework. We do this by evaluating the CheXphoto AUROC when each technique is applied individually. We observe that overall, each technique has a positive impact on robustness. The combination of three techniques used in RoMIA boosts AUROC by up to 5%. We evaluate each technique in more detail in subsequent sub-sections.

A

Model	Parameters	ImageNet Accuracy		Baseline AUROC		Noise-added training AUROC	DCT-based denoising AUROC	MixUp-based fine tuning AUROC	Overall RoMIA AUROC	P-value
		Top-1	Top-5	Chexpert	CheXphoto					
DenseNet121	8.0M	74.434	91.972	0.9009	0.7651	0.7854	0.7708	0.7856	0.8055	2.957E-6
DenseNet169	14.1M	75.6	92.806	0.8938	0.7691	0.7647	0.7726	0.7689	0.8012	1.793E-4
DenseNet201	20.0M	76.896	93.37	0.9012	0.7677	0.7763	0.7768	0.7784	0.8181	5.531E-3
ResNet18	11.7M	69.57	89.24	0.8904	0.7789	0.7829	0.7874	0.7864	0.7964	7.619E-5
ResNet34	21.8M	73.27	91.26	0.9016	0.8025	0.8001	0.8028	0.7971	0.8102	6.044E-4
ResNet50	25.6M	75.99	92.98	0.9011	0.7905	0.7912	0.7983	0.7868	0.8245	4.055E-5



Ground truth: Atelectasis
Baseline: Edema, Pleural Effusion
RoMIA: Atelectasis

Ground truth: Normal
Baseline: Edema
RoMIA: Normal

Ground truth: Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural effusion
Baseline: Pleural effusion
RoMIA: Edema, Consolidation, Atelectasis, Pleural effusion

FIGURE 3
(A) Characteristics of the baseline models used in the experiments and accuracy values (B) AUROC of baseline models on CheXpert and CheXphoto, (C) AUROC improvement from RoMIA and each of its constituent techniques, and (D) example inputs misclassified by the baseline model but correctly classified by RoMIA model.

3.3 Contributions from noise-added training

Figure 4A explores the impact of various dataset transformation techniques used in noise-added learning. Specifically, we transformed 10%, 25%, and 50% of the training samples in the CheXpert dataset and either added them to the dataset (augmentation) or replaced the original samples with them (replacement). We observe that the 25% replacement strategy worked best across all networks. We note that this strategy does not impact training time, as the only overhead incurred is a one-time transformation (noise addition) to the inputs, which is insignificant.

3.4 Effect of fine-tuning with input mixing

Several approaches to input mixing have been proposed in the literature, primarily as methods for data augmentation that lead to better generalization of machine learning models. To evaluate the

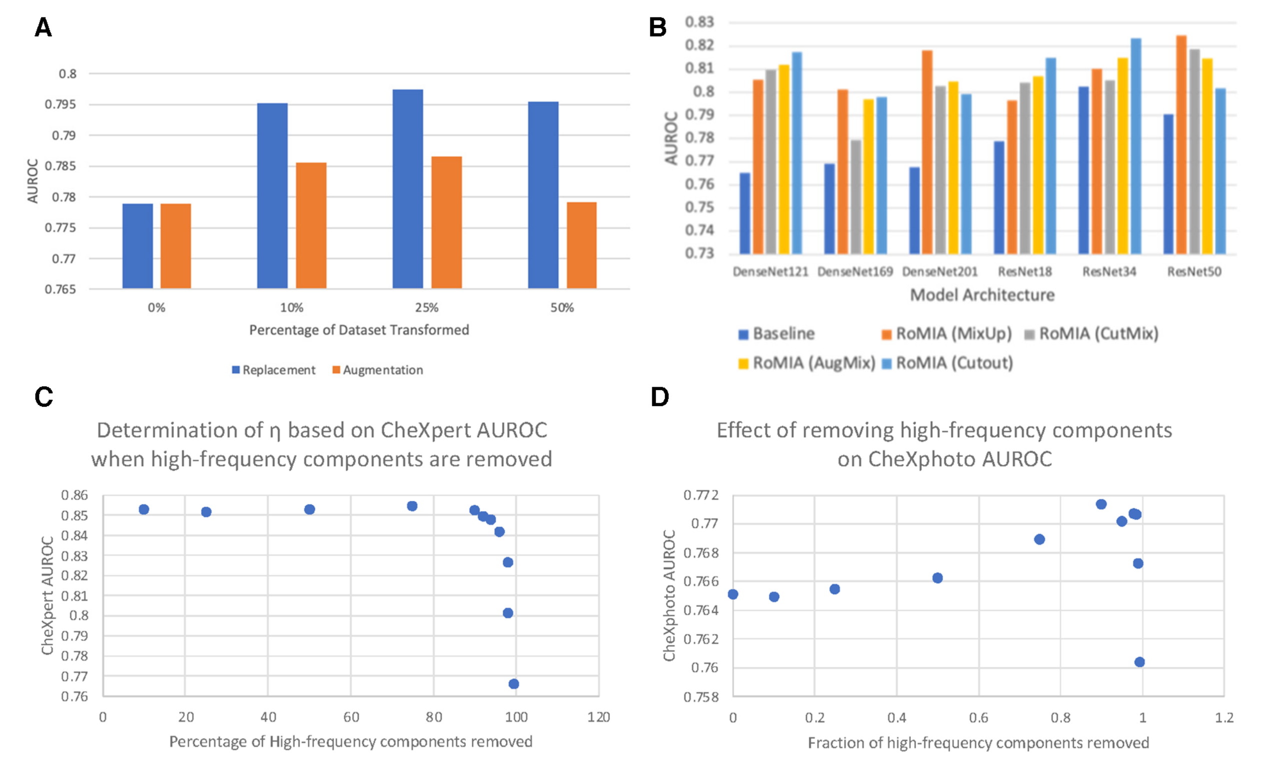


FIGURE 4
(A) Comparing dataset replacement vs. augmentation during noise-added training (ResNet18), (B) comparing different input mixing strategies in RoMIA for various networks, and (C) determination of frequency cutoff threshold (η) and (D) impact of η on CheXphoto AUROC.

impact of the mixing strategy in the fine-tuning step of RoMIA, we consider CutMix (14) and MixUp (36), the two most widely used strategies, in addition to AugMix (15) and Cutout (44). To determine which strategy yields higher improvement in robustness, we compare in Figure 4B the AUROC boosts on the CheXphoto dataset when each strategy is applied. We observe that, while all mixing strategies yield improvements over the baseline, MixUp provides the best results overall, followed by Cutout and AugMix. This motivated our decision to use MixUp in the final RoMIA framework.

3.5 Selection of η in DCT-based denoising

A key feature of our framework is the DCT-based Denoising step, which removes high-frequency noise from the inputs. We use the parameter η to denote the percentage of high-frequency components removed from each image. In Figure 4C, D, we consider the impact of the choice of the parameter η by showing how different η values affect CheXpert and CheXphoto AUROC. Due to the nature of x-ray radiographs, we find that removing a large fraction of the high frequencies does not have a detrimental impact on performance for either dataset and in fact improves accuracy on the noisy (CheXphoto) data. We determine η as the largest value that results in a less than 0.5% decrease in accuracy on the clean (CheXpert) dataset (Figure 4C). We observe that this value of η improves performance on the CheXphoto dataset (Figure 4D). This result underscores the efficacy of DCT-based denoising.

4 Discussion

The success of AI in recent years has led to significant interest in applying it to the medical field. In particular, since DNNs have been very successful in image processing applications, they are frequently being applied to medical imaging tasks. One of the challenges that must be addressed when applying AI to any critical application, and certainly to medical imaging, is their robustness under conditions encountered in the real world. Previous research has shown that DNN models can be very brittle in the presence of input noise and variations. Our work is a first step towards improving the robustness of medical imaging models, with a particular focus on the kinds of noise encountered in medical settings. Although our experimental setup focuses on models for classifying chest radiographs, the techniques we propose are worth exploring in other medical imaging applications.

While the RoMIA framework achieves considerable improvements in robust accuracy, there still remains a gap in accuracy on clean and noisy inputs, especially for high levels of noise, that could be addressed by future work. One possible direction is to address robustness when training from scratch, in contrast to RoMIA, which only addresses it in the transfer learning step. Also, our work evaluates robustness as accuracy in classifying photographs of chest radiographs (i.e., the CheXphoto dataset). Future work could evaluate robustness under a broader set of conditions. Another interesting direction would be evaluating these techniques in a broader range of medical imaging applications. Given the criticality of medical imaging applications, robustness evaluation should be made a standard part of the regulatory evaluation process for these models. Finally, human checking of the output of AI models is one way of improving the confidence in their decisions. This could be enabled by creating explainable models that produce a human-interpretable justification for their decisions. Addressing these issues will go a long way towards enabling the adoption of AI-based medical imaging in clinical practice.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. SK: Formal Analysis, Software, Writing – review & editing. KR: Conceptualization, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

This work was supported in part by the Center for the Co-Design of Cognitive Systems (CoCoSys), a JUMP2.0 center sponsored by the Semiconductor Research Corporation (SRC) and DARPA.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Gulshan V, Peng L, Coram M, Stumpe M, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. (2016) 316(22):2402–10. doi: 10.1001/jama.2016.17216

2. Arda: Using artificial intelligence in ophthalmology. Google Health. Available at: https://health.google/caregivers/arda/ (Cited January 26, 2023).

3. Greenfield D. Artificial Intelligence in Medicine: Applications, implications, and limitations. Science in the news. Harvard University (2019). Available at: https://sitn.hms.harvard.edu/flash/2019/artificial-intelligence-in-medicine-applications-implications-and-limitations/ (Cited January 26, 2023).

4. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv [Preprint]. *arXiv:1901.07031* (2019).

5. Wang S, Cao G, Wang Y, Liao S, Wang Q, Shi J, et al. Review and prospect: artificial intelligence in advanced medical imaging. *Front Radiol.* (2021) 1:781868. doi: 10.3389/fradi.2021.781868

6. Chen A, Li C, Chen H, Yang H, Zhao P, Hu W, et al. A Comparison for Anti-noise Robustness of Deep Learning Classification Methods on a Tiny Object Image Dataset: from Convolutional Neural Network to Visual Transformer and Performer. *arXiv* [Preprint] *arXiv:2106.01927* (2021).

7. Liu M, Liu S, Su H, Cao K, Zhu J. Analyzing the Noise Robustness of Deep Neural Networks. *arXiv* [Preprint] *arXiv:1810.03913* (2018).

8. Kulkarni V, Gawali M, Kharat A. Key technology considerations in developing and deploying machine learning models in clinical radiology practice. *JMIR Med Inform.* (2021 Sep 9) 9(9):e28776. doi: 10.2196/28776

9. Phillips N, Rajpurkar P, Sabini M, Krishnan R, Zhou S, Pareek A, et al. CheXphoto: 10,000+ Photos and Transformations of Chest x-rays for Benchmarking Deep Learning Robustness. *arXiv* [Preprint] *arXiv:2007.06199* (2020).

10. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* (2018 Nov 6) 15(11):e1002683. doi: 10.1371/journal.pmed.1002683. PMID: 30399157; PMCID: PMC6219764.

11. Tompe A, Sargar K. *X-Ray Image Quality Assurance.* Treasure Island, FL: StatPearls Publishing (2022).

12. Seyyed-Kalantari L, Liu G, McDermott M, Chen I, Ghassemi M. CheXclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv* [Preprint] *arXiv:2003.00827* (2020).

13. Zheng S, Song Y, Leung T, Goodfellow I. *Improving the robustness of deep neural networks via stability training. International Joint Conference on Artificial Intelligence* (2021). p. 2909–15. doi: 10.24963/ijcai.2019/403

14. Yun S, Han D, Oh S, Chun S, Choe J, Yoo Y. *Cutmix: regularization strategy to train strong classifiers with localizable features. 2019 IEEE/CVF International Conference on Computer Vision.* Available at: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00612

15. Hendrycks D, Mu N, Cubuk ED, Zoph B, Gilmer J, Lakshminarayanan B. Augmix: a simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019).

16. Wang X, Yifan Peng LL, Lu Z, Bagheri M, Ronald M. *SummersChestx-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE. (2017).

17. Vakharia PS. Artificial Intelligence for the Screening of Diabetic Retinopathy. Retinalphysician.com. Retinal Physician (2022). Available at: https://www.retinalphysician.com/issues/2022/november-december-2022/artificial-intelligence-for-the-screening-of-diabe (Cited January 26, 2023).

18. Medcognetics. Available at: https://www.3derm.com/ (Cited January 26, 2023).

19. 3Derm. Available at: https://www.medcognetics.com/ (Cited January 26, 2023).

20. Tang YX, Tang YB, Peng Y, Yan K, Bagheri M, Redd B, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit Med*. (2020) 3:70. doi: 10.1038/s41746-020-0273-z

21. Antony M, Kakileti ST, Shah R, Sahoo S, Bhattacharyya C, Manjunath G. Challenges of AI driven diagnosis of chest x-rays transmitted through smart phones: a case study in COVID-19. *Sci Rep*. (2023) 13:18102. doi: 10.1038/s41598-023-44653-y

22. Hsieh C, Nobre IB, Sousa SC, Ouyang C, Brereton M, Nascimento JC, et al. MDF-net for abnormality detection by fusing x-rays with clinical data. *Sci Rep*. (2023) 13:15873. doi: 10.1038/s41598-023-41463-0

23. Devasia J, Goswami H, Lakshminarayanan S, Rajaram M, Adithan S. Deep learning classification of active tuberculosis lung zones wise manifestations using chest x-rays: a multi label approach. *Sci Rep*. (2023) 13:887. doi: 10.1038/s41598-023-28079-0

24. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Shao J, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat Biomed Eng*. (2023) 7:743–55. doi: 10.1038/s41551-023-01045-x

25. Gupta A, Sheth P, Xie P. Neural architecture search for pneumonia diagnosis from chest x-rays. *Sci Rep*. (2022) 12:11309. doi: 10.1038/s41598-022-15341-0

26. Cho Y, Kim JS, Lim TH, Lee I, Choi J. Detection of the location of pneumothorax in chest x-rays using small artificial neural networks and a simple training process. *Sci Rep*. (2021) 11:13054. doi: 10.1038/s41598-021-92523-2

27. Wang L, Lin ZQ, Wong A. COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Sci Rep*. (2020) 10:19549. doi: 10.1038/s41598-020-76550-z

28. Nugroho BA. An aggregate method for thorax diseases classification. *Sci Rep*. (2021) 11:3242. doi: 10.1038/s41598-021-81765-9

29. Pham HH, Nguyen NH, Tran TT, Nguyen TNM, Nguyen HQ. PediCXR: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Sci Data*. (2023) 10:240. doi: 10.1038/s41597-023-02102-5

30. Apostolidis KD, Papakostas GA. A survey on adversarial deep learning robustness in medical image analysis. *Electronics (Basel)*. (2021) 10:2132. doi: 10.3390/electronics10172132

31. Joel MZ, Umrao S, Chang E, Choi R, Yang DX, Duncan JS, et al. Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology. *JCO Clin Cancer Inform*. (2022 Feb) 6:e2100170. doi: 10.1200/CCI.21.00170

32. Azizi S, Culp L, Freyburg J, Mustafa B, Baur S, Kornblith S, et al. Robust and efficient medical imaging with self-supervision. arXiv preprint arXiv:2205.09723 (2022).

33. Dai Y, Qian Y, Lu F, Wang B, Gu Z, Wang W, et al. Improving adversarial robustness of medical imaging systems via adding global attention noise,. *Comput Biol Med*. (2023) 164. doi: 10.1016/j.compbiomed.2023.107251

34. An G. The effects of adding noise during backpropagation training on a generalization performance. *Neural Comput*. (April 1996) 8(3):643–74. doi: 10.1162/neco.1996.8.3.643

35. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. *Imagenet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition*; Miami, FL, USA (2009). p. 248–55. doi: 10.1109/CVPR.2009.5206848

36. Zhang H, Cisse M, Dauphin Y, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. *arXiv* [Preprint] *arXiv:1710.09412* (2017).

37. Zhang L, Deng Z, Kawaguchi K, Ghorbani A, Zou J. How Does Mixup Help With Robustness and Generalization?. *arXiv* [Preprint] *arXiv:2010.04819* (2020).

38. Boyat A, Joshi B. A Review Paper: Noise Models in Digital Image Processing. *arXiv* [Preprint] *arXiv:1505.03489* (2015).

39. Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. *IEEE Trans Comput*. (1974) C-23(1):90–3. doi: 10.1109/T-C.1974.223784

40. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* [Preprint] *arXiv:1912.01703* (2019).

41. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. *arXiv* [Preprint] *arXiv:1603.04467* (2016).

42. Yuan Z, Yan Y, Sonka M, Yang T. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. *arXiv* [Preprint] *arXiv:2012.03173* (2020).

43. Culjak I, Abram D, Pribanic T, Dzapo H, Cifrek M. *A brief introduction to OpenCV. 2012 Proceedings of the 35th International Convention MIPRO*; Opatija, Croatia (2012), pp. 1725–30.

44. DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017).