Survey of Image Denoising Methods for Medical Image Classification

Peter F. Michael[a] and Hong-Jun Yoon[b]

[a]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA 98195
[b]Health Data Sciences Institute, Biomedical Science, Engineering & Computing Group,
Oak Ridge National Laboratory, Oak Ridge, TN, USA 37830

ABSTRACT

Medical imaging devices, such as X-ray machines, inherently produce images that suffer from visual noise. Our objectives were to (i.) determine the effect of image denoising on a medical image classification task, and (ii.) determine if there exists a correlation between image denoising performance and medical image classification performance. We performed the medical image classification task on chest X-rays using the DenseNet-121 convolutional neural network (CNN) and used the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics as the image denoising performance measures. We first found that different denoising methods can make a statistically significant difference in classification performance for select labels. We also found that denoising methods affect fine-tuned models more than randomly-initialized models and that fine-tuned models have significantly higher and more uniform performance than randomly-initialized models. Lastly, we found that there is no significant correlation between PSNR and SSIM values and classification performance for our task.

Keywords: image denoising, image classification, X-ray image denoising, machine learning, deep learning

1 INTRODUCTION

Significant strides in computer vision have been made in recent years due to the rise of deep learning. With the vast amount of data present in many computer vision tasks, the effect of denoising inputs to CNNs is understudied. In fields such as medical imaging, where data is often sparse, any method that can increase model performance without enlarging the dataset is of great use. One common method is pre-processing inputs. Medical imaging devices, such as X-ray machines, inherently produce images with visual noise. While denoising these images is undoubtedly helpful in improving radiologists' diagnostic performance, some question its ability to improve performance on computer-aided diagnosis tasks based on CNNs. Some may believe that denoising these images will help the CNN recognize important image features, while others may argue that denoising lowers the level of detail in an image or that the CNN can learn to handle the noise itself, even with a relatively small training set. We set out to investigate this problem by selecting numerous image denoising methods, both learning and non-learning-based methods and applied them to a medical image classification task. Additionally, we set out to find if a denoising algorithm's performance on an image with synthetic noise, measured by the PSNR and SSIM metrics, can correlate with medical image classification performance. This correlation would be useful in gaining an intuition whether a certain denoising method may increase classification performance without training an entire model.

Further author information: (Send correspondence to Hong-Jun Yoon)
Peter Michael: E-mail: petermic@cs.washington.edu
Hong-Jun Yoon: E-mail: yoonh@ornl.gov

2 LITERATURE REVIEW

Medical imaging differs from many other imaging tasks in that there is very strong attention to detail. Naïvely denoising an X-ray image, for example, may lead to the misdiagnosis of a tumor. Often the signs for many medical conditions are hard to recognize, so it is imperative to process such data carefully.

There have been very few works on the effects of image denoising on image classification tasks relative to the amount of computer vision work being published in recent years. This applies to the medical imaging field, where there has been even less work regarding this topic. There are two categories of work in this field, (i.) how noise affects image classification performance and (ii.) how denoising methods can be used to improve classification performance. Some works contain both.

In a study done by Dodge and Karam[1], noise, blur, and contrast were shown to have a significant impact on image classification performance. Specifically, it was shown that CNNs trained on images were vulnerable to Gaussian noise. A similar study by Nazaré, et al.[2] affirms this argument, while also showing that CNNs struggle with types of noise they were not trained on. Additionally, this study shows that for certain types of noise, training on denoised images helps increase classification performance on images from the same distribution. Koziarski and Cyganek[3] affirm the effect of noise on classification performance and show that denoising inputs and augmenting training data with noise can increase classification performance. A work done by Diamond, et al.[4] proposes an end-to-end architecture for joint denoising, deblurring, and image classification. They create a pipeline with an image preprocessing module that learns to denoise the images together with the CNN used for classification. They show that, for emulated raw camera sensor data, this architecture is superior to separating the preprocessing and classification stages. They also discover that denoised images that look better to the human eye do not necessarily correlate with superior image classification performance for their task. Given that CNNs tend to have low noise-resilience, Dodge and Karam[5] propose an ensemble of VGG16[6] CNNs, each trained on a specific type of added noise. Then, given an input image, a gating network determines the weight to give to each CNN for classification. This produces a model that is resilient to many different kinds of noise, solving the noise-resiliency problem. Yim and Sohn[7] propose a dual-channel CNN model based on Inception-v3[8]. They have one model that takes in the original, unprocessed image, and another model that takes in a denoised image, denoised by various algorithms. The outputs are combined via either feature summation or concatenation, and a fully connected layer is used to convert these features into classification results. The purpose and result of this were to retain performance on the original dataset using the model that takes in the raw image, while providing better classification accuracy on noisy images, using the denoising module. Lastly, a notable work by Elhoseny and Shankar[9] shows that for medical imaging tasks, denoising has the potential to improve medical image classification performance. They propose a system that preprocesses images with a bilateral filter with parameters optimized for best image denoising performance. Using the Inception-v4[8] CNN for classification, they show that denoising inputs may increase medical image classification performance. However, they do not compare their results with a CNN without a denoised dataset, which we do in this paper.

3 METHODS

3.1 Effect of Image Denoising on Medical Image Classification Performance

We examined the effect of image denoising on a medical image classification task and determined the optimal denoising methods for this task. The denoising methods used can be separated into learning-based and non-learning-based methods. All learning-based methods utilize CNNs. The non-learning-based group consists of the Wiener filter[10], moving average filter, median filter, and opening morphological filter. These filters were picked since they are frequently used in image denoising and provide a good distribution of linear, nonlinear, and morphological filters. The learning-based group consists of the DnCNN[11] by Kai Zhang, et al. and the Noise2Noise[12] CNN by NVIDIA Research. These two CNNs were picked since they are among the most popular CNNs used for image denoising. For the sample medical image classification task, we used the ChestX-ray14[13] (CXR14) dataset, a dataset of chest X-rays with 14 possible diagnoses provided by the NIH. We then preprocessed all the images in CXR14 through these denoising methods at its native resolution, 1024x1024. This produced six datasets containing 112,120 images each, resulting in a total of seven datasets including the one with no denoising applied. All the non-learning-based methods used a window size of three pixels, as sizes higher made images undiscernable to the human eye. We used the provided, pre-trained models for the denoising CNNs. DnCNN was trained on images with synthetic Gaussian noise of standard deviation $\sigma=25$ and the Noise2Noise CNN was trained uniformly on images with synthetic Gaussian noise of standard deviations between $\sigma=0$ and $\sigma=50$. The datasets were split into train, validation, and test sets by the NIH, stratified by label and patient name. We then trained the DenseNet-121[14] CNN on each dataset twice, once with randomly-initialized weights, and once by fine-tuning a model pre-trained on ImageNet[15]. We used PyTorch[16] to train these models for its speed and debugging capabilities. We trained these models with the AdaDelta[17] optimization algorithm, using: an early stop with a maximum of 100 epochs and a patience of 10 epochs, validation loss as the performance metric, and data augmentation. Data augmentation consisted of resizing the images to 256x256, randomly cropping them to 224x224, then randomly rotating, shifting, scaling, and (horizontally and vertically) flipping them. Lastly, we normalized the images to be in the range [0, 1]. The model was trained with a batch size of 115 images on a single NVIDIA Tesla V100 with 16GB of memory. After training these models, we calculated the area under the receiver operating characteristic curve (AUC) for each label on the test set for each model, using that as the classifier's performance measure. The classification pipeline is shown in figure 1.

3.2 PSNR and SSIM Correlation with Medical Image Classification Performance

We examined the correlation between the PSNR and SSIM metrics and our medical image classification task. To achieve this, we added Gaussian noise to the Shepp-Logan phantom image[18] at standard deviations of $\sigma=5$, $\sigma=15$, and $\sigma=25$. We then ran these three images through each of the denoising methods mentioned in section 3.1 to calculate the PSNR and SSIM values, using the original images as base references. For the non-learning-based methods, we used a window size of five pixels and used the provided pre-trained models for the learning-based methods. We determined the level of correlation by calculating the R^2 values between each denoising algorithm's PSNR or SSIM and its mean AUC computed over all labels.

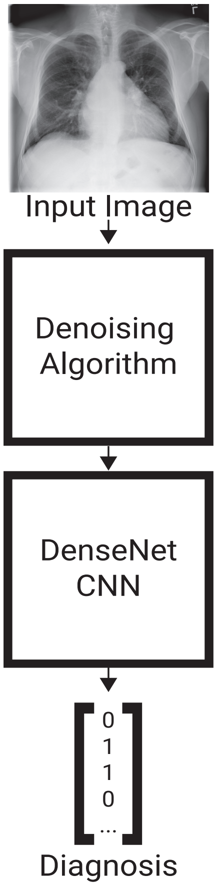

Figure 1. Chest X-Ray Image Classification Pipeline

4 RESULTS

4.1 Effect of Image Denoising on Medical Image Classification Performance

Looking at figures 2 and 3, we find that fine-tuning a pre-trained model delivers higher and more uniform results than training a model from scratch (i.e., with randomly-initialized weights). Nonetheless, we find that different denoising methods can make a noticeable difference in classification performance for both models. Table 1 shows which method performs the best for each model and label, with the method bolded if it improves the AUC by a statistically significant amount. Statistical significance is calculated between the model without denoising and the highest performing model with denoising for a specific label, using 95% confidence intervals. These confidence intervals were computed by bootstrapping, using 1,000 samples. Interestingly, the fine-tuned model benefits more from denoising than the randomly-initialized one, with three statistically-significant increases in AUCs caused by denoising in the former compared to one in the latter. The fine-tuned model experienced these increases with the Wiener filter for two of the labels and the moving average filter for the other. The randomly-initialized model actually experienced two statistically-significant increases, but one was due to using no denoising! The other was due to using the Noise2Noise CNN. Additionally, the same denoising algorithm does not maintain its relative performance between the fine-tuned and randomly-initialized model. The top-performing denoising algorithm for the randomly-initialized model was never observed to be the same for the fine-tuned one. Lastly, we observed that the classification performance when using the non-learning-based filters is similar to when using the denoising CNNs.

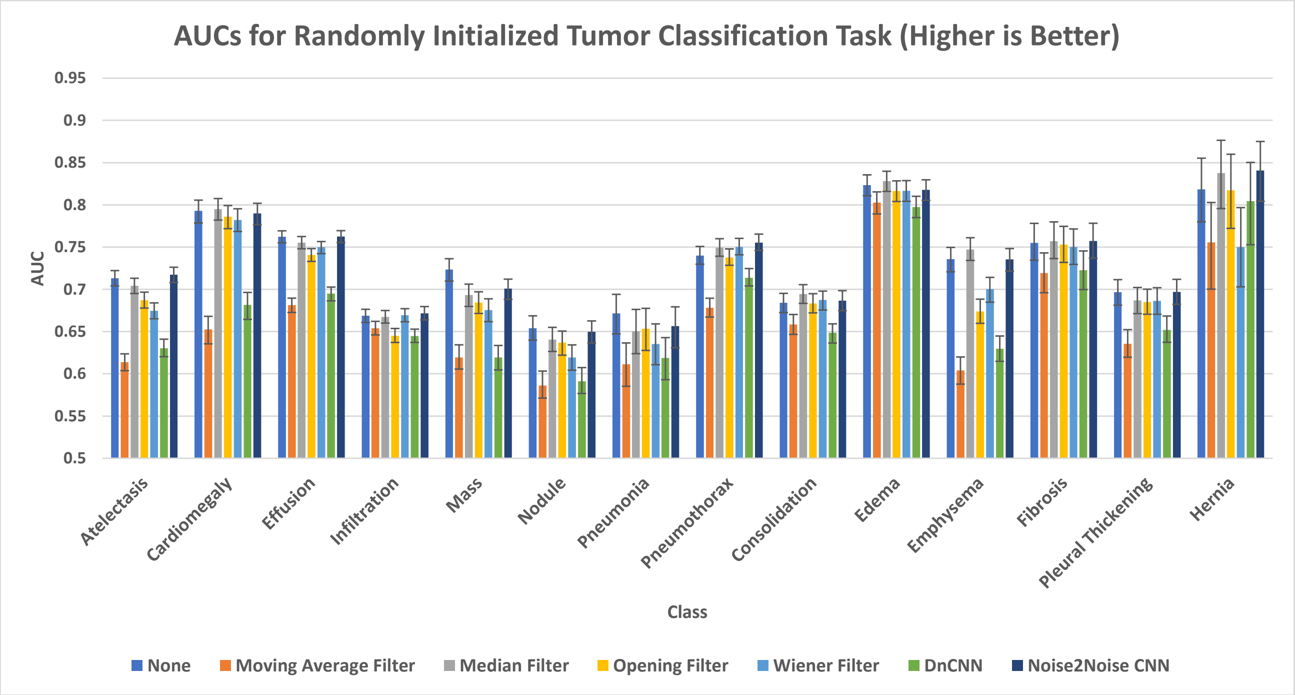

Figure 2. AUCs for a Randomly-Initialized Model

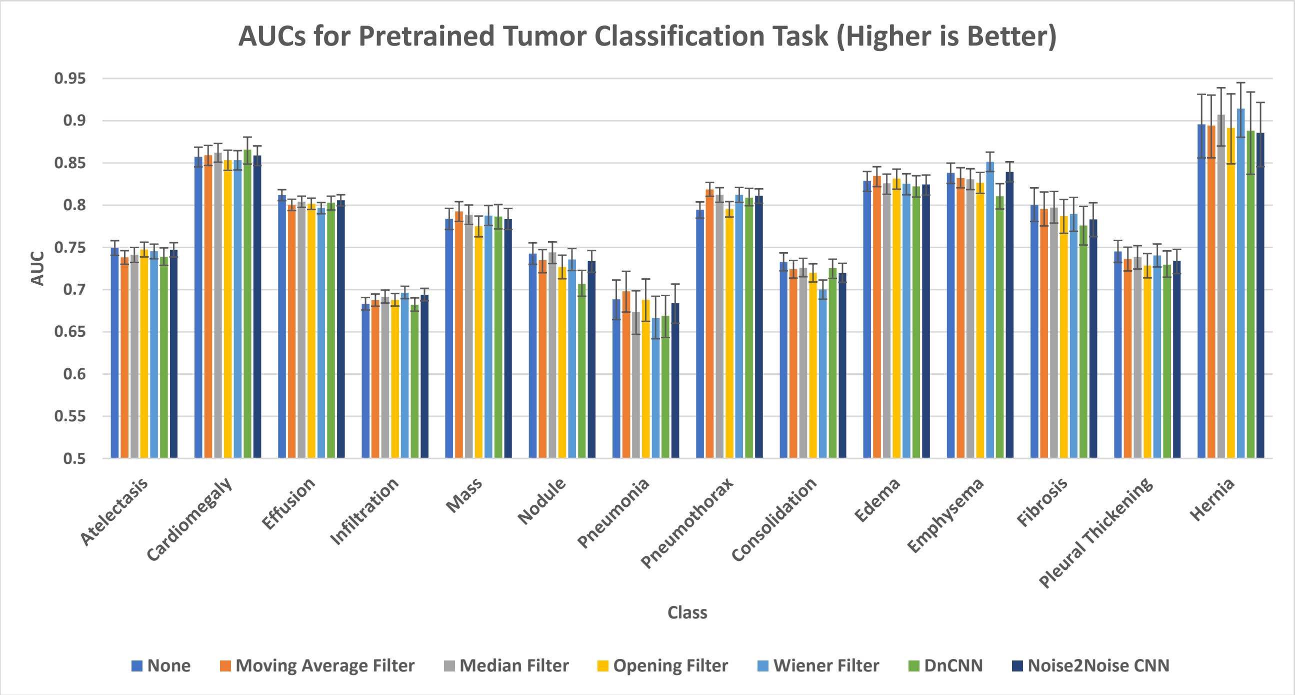

Figure 3. AUCs for Fine-Tuned Model

4.2 PSNR and SSIM Correlation with Medical Image Classification Performance

As mentioned in section 3.2, we compute the correlation between each denoising algorithm's PSNR or SSIM and its mean AUC computed over all labels. Given that the standard deviations of these means are fairly close in value (with standard deviation 0.004), we will dismiss them for this analysis.

Table 1. Best Performing Methods by Label (Bold Denotes Statistically Significant Improvement)

Best Perfoming Methods List, labels 1-7							
	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia
Randomly-Initialized	Noise2Noise CNN	Median Filter	Noise2Noise CNN	Noise2Noise CNN	**No Denoising**	No Denoising	No Denoising
Fine-Tuned	No Denoising	DnCNN	No Denoising	**Wiener Filter**	Moving Average Filter	Median Filter	Moving Average Filter

Best Perfoming Methods List, labels 8-14							
	Pneumothorax	Consolidation	Edema	Emphysema	Fibrosis	Pleural Thickening	Hernia
Randomly-Initialized	**Noise2Noise CNN**	Median Filter	Median Filter	Median Filter	Noise2Noise CNN	Noise2Noise CNN	Noise2Noise CNN
Fine-Tuned	**Moving Average Filter**	No Denoising	Moving Average Filter	**Wiener Filter**	No Denoising	No Denoising	Wiener Filter

4.2.1 PSNR Correlation

Table 2 shows the correlation between PSNR and mean AUC for the fine-tuned and randomly-initialized models. We see that there is a very low level of correlation, increasing monotonically over σ. Lastly, we see that the randomly-initialized model is strictly more correlated to the PSNR for all values of σ.

Table 2. Correlation between PSNR and Mean AUC

R^2 Correlation Values for PSNR			
	σ= 5	σ= 15	σ= 25
Randomly-Initialized	0.000674	0.055038	0.069913
Fine-Tuned	0.000064	0.012432	0.046050

4.2.2 SSIM Correlation

Table 3 shows the correlation between SSIM and mean AUC for the fine-tuned and randomly-initialized models. Like PSNR, we see quite a low level of correlation. Unlike PSNR, however, the correlation value monotonically increases with decreasing σ. Additionally, we do not see that any of the models have strictly higher correlation values than the other. Lastly, we see that the greatest correlation values globally are obtained when using SSIM with σ= 5 for both fine-tuned and randomly-initialized models.

Table 3. Correlation between SSIM and Mean AUC

R^2 Correlation Values for SSIM			
	σ= 5	σ= 15	σ= 25
Randomly-Initialized	0.233217	0.067192	0.006726
Fine-Tuned	0.367492	0.006997	0.011114

5 LIMITATIONS OF RESEARCH

This study is meant to give an introduction to the effects of image denoising on medical image classification performance. As such, multiple variables should be considered for future work.

First, we did not account for Poisson noise in our CNN denoisers. These denoisers were trained only to filter Gaussian noise. With low-dose X-ray imaging becoming more popular because of its increased safety, it is important to study this, as low-dose X-ray images produce more noise that follows a mixed Poisson-Gaussian model[19].

Second, we did not tune any of the denoising algorithms' parameters jointly with the classification CNN to ensure that we are using these algorithms to their fullest extent, similar to the work previously mentioned by

Diamond, et al.[4], which showed improvement in performance for raw camera sensor data.

Lastly, this study should be repeated on many different types of medical image classification tasks, in order to generalize beyond chest X-ray image classification.

6 DISCUSSION

6.1 Effect of Image Denoising on Medical Image Classification Performance

We found in section 4.1 that the fine-tuned model delivers better performance across denoising methods. This shows us that that transfer learning can be useful in a medical imaging scenario, as would be expected for most image classification tasks. Specifically, it not only raises overall performance but also makes it more uniform across different denoising algorithms. Given the low AUC variances, the CNN has built robustness to variances in images introduced by the denoising methods. Interestingly, there are more statistically significant differences in the fine-tuned model than the randomly-initialized one. This is because the confidence intervals for the fine-tuned model are so small, due to its robustness.

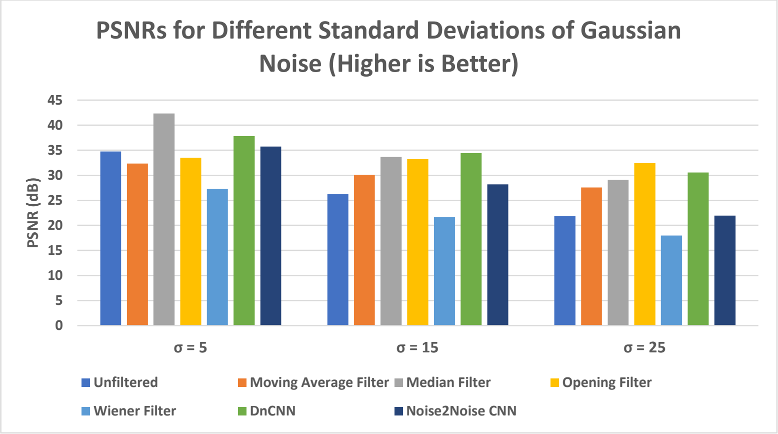

Figure 4. PSNR vs. Standard Deviation for Different Denoising Methods

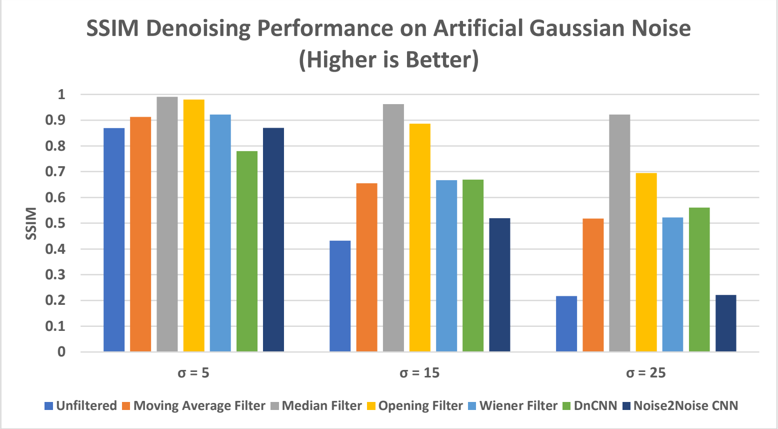

Figure 5. SSIM vs. Standard Deviation for Different Denoising Methods

Nonetheless, we see that using denoising can make a statistically significant improvement for select labels, using either the fine-tuned or randomly-initialized model! This experiment should be redone with the classification CNNs trained on 1024x1024, CXR14[13]'s native resolution. Additionally, an ensemble of models should be created, such that the models best at predicting a specific label have the most weight on the classification result for that label. These label weights should be determined separately for the fine-tuned and randomly-initialized model, as we noted that the best performing preprocessing methods were not consistent between the two. It may be of greatest benefit to train a denoising framework jointly with the classification CNN so that the parameters are tuned for best classification performance, not best visual look, similar to the work previously mentioned by Diamond, et al.[4]. This would better utilize a CNN's ability to learn, which would likely break the tie in classification performance with non-learning based methods. Lastly, an experiment should be done that shrinks the size of the training set and determines if there is a change in the importance of using denoising algorithms. This would give a useful relationship of how useful denoising is given the size of the training set.

6.2 PSNR and SSIM Correlation with Medical Image Classification Performance

We found in section 4.2 that there appears to be no significant correlation between chest X-ray classification performance and PSNR or SSIM. This is a similar result to the work previously mentioned by Diamond, et al.[4], which found that visually better-looking images did not necessarily mean better classification accuracy for raw camera sensor data.

Nonetheless, we found that the SSIM metric with Gaussian noise of standard deviation $\sigma = 5$ had the highest correlation. Unlike PSNR, which focuses on how close the pixel values are together, SSIM focuses more on structure preservation in the image. This would be expected, as preserving critical structures is crucial in medical imaging.

A metric that could estimate medical image classification performance could greatly increase efficiency. If no other existing metrics are found that can do this, one should be created. This metric could be CNN-based, taking in an image (or a sequence of them as a 3-dimensional CNN), and predicting how good classification performance would be.

7 CONCLUSION & FUTURE WORK

We found that our medical image classification task can be affected by using a denoised dataset and that different denoising methods outperform others for certain labels. Second, we found that using a fine-tuned model benefits more from denoised datasets than a randomly-initialized model while also bringing greater and more uniform performance for our task. As the amount of data in a medical scenario is often scarce, using a fine-tuned model is a simple way to boost performance. Lastly, there is no significant correlation between PSNR and SSIM and our classification task. This work should be extended by:

1. Training the CNN on CXR14[13]'s native resolution, 1024x1024.
2. Not only taking into account Gaussian noise, but also Poisson noise.
3. Training denoising algorithm parameters jointly with the classification CNN.
4. Repeating this study on different medical image classification tasks.
5. Forming an ensemble model where each individual CNN is trained on a specific type of denoised data, and assigning weights to each of the models based on its performance on each label.
6. Shrinking the training set and analyzing the difference in performance made by denoising.
7. Developing a new metric or using an existing one that gives an idea of how a denoising algorithm will affect medical image classification performance without training the model.

These works would help gain a much broader intuition on the effect of image denoising on medical image classification.

REFERENCES

[1] Dodge, S. and Karam, L., "Understanding how image quality affects deep neural networks," (2016).

[2] Nazaré, T. S., da Costa, G. B. P., Contato, W. A., and Ponti, M., "Deep convolutional neural networks and noisy images," in [*Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*], Mendoza, M. and Velastín, S., eds., 416–424, Springer International Publishing, Cham (2018).

[3] Koziarski, M. and Cyganek, B., "Image recognition with deep neural networks in presence of noise – dealing with and taking advantage of distortions," *Integrated Computer-Aided Engineering* **24**, 1–13 (08 2017).

[4] Diamond, S., Sitzmann, V., Boyd, S., Wetzstein, G., and Heide, F., "Dirty pixels: Optimizing image classification architectures for raw sensor data," (2017).

[5] Dodge, S. F. and Karam, L. J., "Quality resilient deep neural networks," *CoRR* **abs/1703.08119** (2017).

[6] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," (2014).

[7] Yim, J. and Sohn, K.-A., "Enhancing the performance of convolutional neural networks on quality degraded datasets," (2017).

[8] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," (2014).

[9] Elhoseny, M. and Shankar, K., "Optimal bilateral filter and convolutional neural network based denoising method of medical image measurements," *Measurement* **143**, 125 – 135 (2019).

[10] Wiener, N., [*Extrapolation, Interpolation, and Smoothing of Stationary Time Series*], MIT Press (March 1964).

[11] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L., "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing* **26**, 3142–3155 (July 2017).

[12] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T., "Noise2noise: Learning image restoration without clean data," *CoRR* **abs/1803.04189** (2018).

[13] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CoRR* **abs/1705.02315v5** (2017).

[14] Huang, G., Liu, Z., v. d. Maaten, L., and Weinberger, K. Q., "Densely connected convolutional networks," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2261–2269 (July 2017).

[15] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (June 2009).

[16] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A., "Automatic differentiation in pytorch," in [*NIPS-W*], (2017).

[17] Zeiler, M. D., "ADADELTA: an adaptive learning rate method," *CoRR* **abs/1212.5701** (2012).

[18] Shepp, L. A. and Logan, B. F., "The fourier reconstruction of a head section," *IEEE Transactions on Nuclear Science* **21**, 21–43 (June 1974).

[19] Ding, Q., Long, Y., Zhang, X., and Fessler, J. A., "Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct," (2018).