Improving Machine Learning Classification Accuracy in Noisy Environment

Jian Ren and Tongtong Li
Department of Electrical and Computer Engineering
Michigan State University, East Lansing, MI 48824-1226
emails: {renjian, tongli}@msu.edu

***Abstract*—Image classification is one of the central research in machine learning that has wide application in many fields. Many factors can adversely impact the classification accuracy, including particularly unexpected noise in images. In this paper, we analyze methods to improve image classification accuracy. Particularly, we analyzes the efficacy of two techniques. First, we find that using noise reduction techniques can enhance image classification accuracy, regardless of whether noise has been added to the original images. Second, we find that weighted voting can be used to improve image classification accuracy even when the voting set is not be entirely independent. Additionally, we identify an inconsistency in classification accuracy and image fidelity, which could potentially be exploited for adversarial attacks.**

***Index Terms*—image classification, noise, classification accuracy, image fidelity, adversarial attack**

I. Introduction

We live in the era of big data. Image classification is one of the central research area that has been widely utilized in areas including healthcare, transportation, agriculture, food processing, manufacturing, security, and defense and military [1]. The aims of image classification is to assign an input image with a label from a fixed set of categories. It is one of the core problems in Computer Vision and deep learning (DL) that has a large variety of practical applications. Classification accuracy is a key performance metric of a trained classifier or learning model in image classification. Unfortunately, many factors can impact classification robustness and accuracy.

Various types of noise can be passively introduced to images due to poor weather conditions, dirty sensors, and adversarial examples. This paper focuses on Gaussian white noise, which is often added to images to simulate real-world conditions and make models more robust to variations. In digital images, the principal sources of Gaussian noise arise during acquisition. The sensor has inherent noise due to the level of illumination and its own temperature, as well as the electronic circuits connected to the sensor may also inject their own share of electronic circuit noise [2]. Noise also presents in clinical body computed tomography (CT) images [3].

Additive Gaussian white noise (AGWN) is a typical model of image noise caused primarily by Johnson–Nyquist noise (thermal noise), including that which comes from the reset noise of capacitors. It is independent at each pixel and independent of the signal intensity [4].

It is worthy to mention that although Gaussian denoising technology has been studied in depth and many Gaussian denoising algorithms have been developed [5], the research on Gaussian noise image recognition is relatively rare.

In this paper, we analyze image classification under Gaussian white noise (GWN) without requiring any knowledge of the architecture, parameters, and access to the trained neural network mode of the DL model. The DL can be regarded as a "black box", meaning that we all aware that it performs reasonably well in many application scenarios but have limited knowledge of how it really works [6], [7], despite numerous studies have devoted to understanding and interpreting deep neural networks (DNNs) [8]–[12]. From inspecting various scenarios, we may gain insights into the semantic inner levels of neural networks and identify problematic decision boundaries, which can help improve classification accuracy and robustness of neural networks in a wider range of situations.

It is worth noting that we are not seeking to get into any specific discussion on Gaussian denoising, instead, we will simple discuss the general denoising algorithms and their impact on classification accuracy.

The rest of this paper is organized as follows: In Section II, preliminaries are introduced that summarize key concepts used in this paper. Next, in Section III, we evaluate image classification accuracy under noise impact and noise reduction. In Section IV, we explore using weighted voting based ensemble classification to increase prediction accuracy. We conclude the paper in Section V.

II. Preliminaries

In this section, we introduce the preliminary concepts required for this paper.

A. Gaussian White Noise

Gaussian white noise (GWN) is likely the most common stochastic model used in engineering applications. A stochastic process $X(t)$ is said to be GWN if $X(t)$ a stationary Gaussian random process with zero mean μ, and variance σ^2, that is

$$X(t) \sim \mathcal{N}(\mu, \sigma^2),$$

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

and a constant power spectrum density, $S_X(f)$, across all frequencies f, which is normally expressed as

$$S_X(f) = \frac{N_0}{2}, for\ all\ f.$$

While simple, GWN can be used to model complex stochastic systems.

B. Signal-to-Noise Ratio (SNR)

Signal-to-noise ratio (SNR) is largely used in science and engineering to compare the level of a desired signal to the level of background noise. To analyze image distortion quantitatively, we will introduce some notations defined in [13]. The Euclidean distance of two $m \times n$ images I_1 and I_2 is define as

$$d(I_1, I_2) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(I_{1i,j} - I_{2i,j})^2}.$$

The norm of an image I is defined as

$$\text{norm}(I) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} I_{i,j}^2}.$$

The SNR of an image I with Gaussian white noise (GWN) is defined as

$$\text{SNR}_{\text{GWN}} = 20\log_{10}\frac{\text{norm}(I)}{d(I, I_{\text{GWN}})}, \tag{1}$$

where I_{GWN} denote the original image with GWN added.

Let I_{GWN_w} be the image derived from I_{GWN} through image denoising using Matlab `wdenoise2` function, which determines the default global threshold values to denoise an image using a Bayesian method and the bior4.4 wavelet with a posterior median threshold rule. For simplicity, we will refer to this approach as w-denoise. We also define I_{GWN_t} to be the image derived from I_{GWN} through a two-step denoise process of I_{GWN}. In the first step, we use `ddencmp` to find the default global threshold for a wavelet packet transform and determine whether soft or hard thresholding should be applied, and whether or not the approximation coefficients are thresholded. In the second step, we use `wdencmp` to return a denoised version of the grayscale digit images obtained by thresholding the wavelet coefficients using the global positive threshold threshold. We denote this denoising approach as t-denoising.

The SNRs of the images derived by applying d-denoising and t-denoising to image I with GWN are defined as follows:

$$\text{SNR}_{I_{\text{GWN}_w}} = 20\log_{10}\frac{\text{norm}(I)}{d(I, I_{\text{GWN}_w})}, \tag{2}$$

$$\text{SNR}_{I_{\text{GWN}_t}} = 20\log_{10}\frac{\text{norm}(I)}{d(I, I_{\text{GWN}_t})}. \tag{3}$$

From the definitions given in equations (2) and (3), it becomes clear that the SNR is large when the image in question is similar to the original image, and small otherwise. Therefore, we can use these equations to measure the closeness of the image in question to the original image. Specifically, the SNR of the original image is infinity.

C. Weighted Voting

Weighted voting refers to voting rules that assign different weights to voters, meaning that some voters will have greater influence compared to a scenario where every voter is assigned equal weight. Weighted voting is largely used in a corporate shareholders meeting. The weight of each shareholders' is proportional to the amount of shares they own. An individual with one share gets the equivalent of one vote, while someone with 10 shares gets the equivalent of 10 votes, and so on. In weighted voting, the most important parameter is the weight/power each voter has in influencing the outcome.

Assume the system has k participating voters or players, with the corresponding weights assigned as $w_1, \cdots, w_k$. In our case, the weights of the voters represent the classification accuracy of the voters' training algorithms. Therefore, they are probabilities with values in $[0, 1]$.

Suppose all voters have l possible classification (voting) choices $c_1, \cdots, c_l$. Voter i has classification prediction (probability) $v_{i,j}$ for choice c_j such that $v_{i,1} + \cdots + v_{i,l} = 1$ for $i = 1, \cdots, k$. The final weighted probability of prediction for classification choice $c_j,\ j = 1, \cdots, l$, can be computed from the following equation:

$$C_j = \sum_{i=1}^{k} w_i \times v_{i,j}, \tag{4}$$

and the final classification class is c_{j_0}, with j_0 determined from:

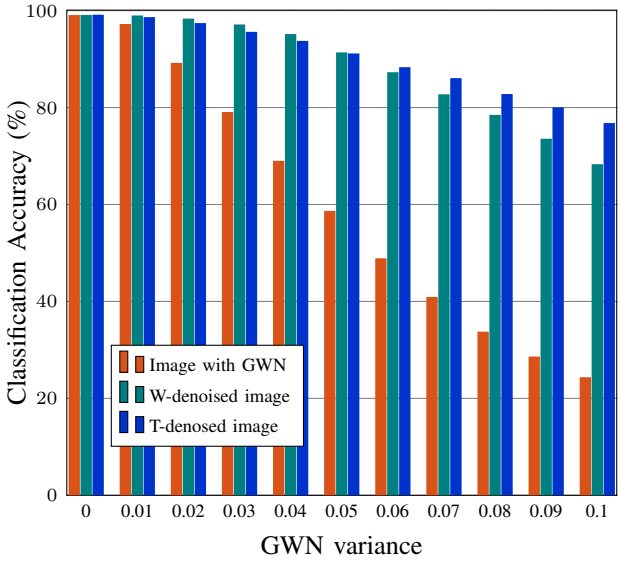
$$j_0 = \arg\max_{1 \le j \le l} C_j. \tag{5}$$

III. Classification of Noise Polluted and Denoised Digits

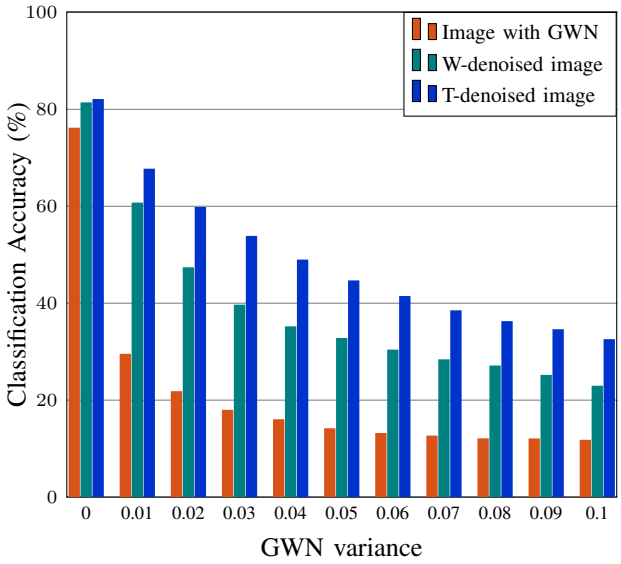
In this section, we evaluate image classification accuracy under the impact of noise and examine two image denoising techniques, along with weighted voting, that can enhance classification accuracy. We use the MNIST database of handwritten digits [14] for our illustrations and simulations. The NIST database consists of a training set of 60,000 black-and-white (bilevel) images and a test set of 10,000 images. The digits have been size-normalized to 28×28.

We assume that the handwritten digit images are polluted by Gaussian white noise (GWN) with zero mean and variance σ. Specifically, we evaluate the effect of noise on a trained classification network with GWN when variance σ changes from 0 to 0.1 in steps of 0.01. The trained classification network achieves 99.14% accuracy on the MNIST training set and 99.08% accuracy on the MNIST test set, which are comparable to state-of-the-art accuracy. We evaluate classification accuracy of these noisy images using both d-denoise and w-denoise techniques.

Fig. 1a shows the classification accuracy for a randomly selected group of handwritten numbers from the MNIST database polluted by GWN, along with the classification accuracy of w-denoised and t-denoised images. We observed



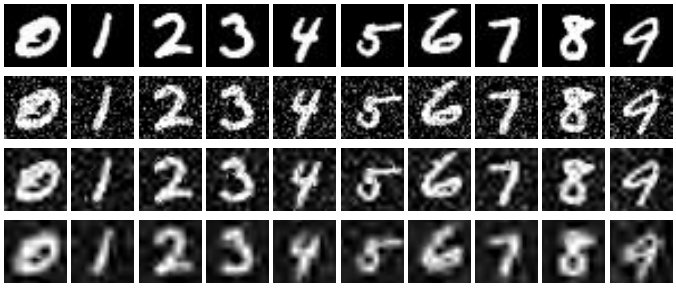
(a) Comparison of classification accuracy among the original digits, the w-denoised digits, and the t-denoised digits, where the trained network achieves 99.14% accuracy on the MNIST training set and 99.08% accuracy on the test set.



(b) Comparison of classification accuracy among the original digits, the w-denoised digits, and the t-denoised digits, where the learn network achieves 83.59% accuracy on the MNIST training set and 76.21% accuracy on the test set.

Fig. 1: Comparison of classification accuracy between digit images with zero-mean GWN, where the variance σ changes from 0 to 0.1 in steps of 0.01, and two image denoising approaches.

that t-denoising achieves an average of 1.7% higher accuracy than w-denoising, while in Fig. 1b, the t-denoising technique demonstrates an average of 9.264% better performance than w-denoising.



(a) Explanatory digit images, where the 1st row contains the original digits, the 2nd row contains digits with zero-mean GWN with variance $\sigma = 0.05$, the 3rd row contains w-denoised digits, and the 4th row contains t-denoised digits.

Nums / SNR	0	1	2	3	4	5	6	7	8	9
Orig-Noise	9.48	4.94	8.15	7.87	6.55	6.11	8.10	6.40	8.60	6.67
w-doising	11.26	7.08	9.46	9.64	8.33	7.81	9.49	8.30	9.86	8.11
t-deoising	8.97	6.39	7.54	7.31	6.48	6.28	6.96	6.26	7.86	5.68

(b) Explanatory SNRs corresponding to 2a.

Fig. 2: Explanatory digit images and the SNRs of the images with their corresponding original digits.

On the other hand, we can visually observe the hand-written digit images in three cases: the images with GWN, along with two denoised versions—w-denoised and t-denoised images—as shown in Fig. 2a. The first row displays the original digits image without noise, the second row shows the GWN polluted digits, the third row presents the images of w-denoised digits, and the fourth row contains the t-denoised digits. We can clearly see that the image fidelity of the w-denoised digits is much better than that of the t-denoised images. This observation is quite surprising and even contradicts our general expectations.

Next, we hope to confirm our visual observation using the proposed SNR measurement. We calculate the SNRs of the images of the handwritten digits in three cases: images with GWN, as well as two denoised versions—w-denoised and the t-denoised images—following the same order in which they are displayed. The results are shown in Fig. 2b, where the first row presents the SNRs of the original digits image without containing noise, the second row shows the SNRs of GWN polluted images, the third row contains the SNRs of the w-denoised images, and the fourth row displays the SNRs of the t-denoised images. From our observations of the fidelity as well as the SNRs of the digit images, it is clear that the image qualities in the second row is lower than in the first row, as GWN has been added to the images in this row which makes this row noisier than the first row. The third row exhibits intermediate quality between the first and second rows due to the application of the w-denoising technique. However, the fourth row appears significantly worse than the third row

and, surprisingly, even worse than the second row, despite the application of the t-denoising technique. These observations align with the SNR values shown in the table in Fig. 2b.

The SNR values are consistent with the image fidelity, which confirms that digit classification accuracy does not appear to be tightly related to image fidelity, as in out case w-denoising demonstrates much better fidelity, while t-denoising achieves higher classification accuracy, although the differences are relatively small. Regardless, this is quite interesting and contradicts visual observation, showing that DL is fundamentally a *guessing machine* that is not always produce correct answers and sometime make significant errors. In machine learning, training a neural network often involves random components to enhance performance, present the algorithm from getting stuck in local minima, and also lead to different learning paths and final models. This makes machine learning classification probabilistic and associated with certain percentage of confidence. As a result, DL classification is far from a closed-form expression of the input, in which the outputs can be fully determined by the inputs.

While the differences in classification accuracy between w-denoising and t-denoising, as shown in Fig. 1a and Fig. 1b are both small, they can also be significant for some learned network, as illustrated in Fig. 3. This is true even though the tested digit images remain the same and contain the same GWN, as shown in Fig. 1 and Fig. 2, which further confirms that the DL is fundamentally a *guessing machine* and can produce highly unexpected predictions. Moreover, unlike the simulation results shown in Fig. 1 and Fig. 2, the classification accuracy for the t-denoised images are even significantly lower than the images with GWN, which probably makes more sense from an image fidelity perspective.

IV. Weighted Voting Based Ensemble Classification

In this section, we explore using weighted voting based ensemble classification to increase prediction accuracy, ideally by combining multiple independent feature sets. However, it may not always be easy or feasible to produce multiple independent features. We analyze supplementing the original image with four subband images generated through a single-level 2-D wavelet function (`dwt2`), which decomposes the image into an approximation (A), and the details in three orientations: horizontal (H), vertical (V), and diagonal (D), shown in Fig. 4. This is achieved using two filters (high pass and low pass filters), and two downsampling steps (row and column). As a result, the size of each sub-band image is half the width and half the height of the original image.

From the way these four-subband images are derived, it is clear that they are not independent of the original image. However, we demonstrate that even in this case, we can improve classification accuracy to some extend using the weighted voting strategy.

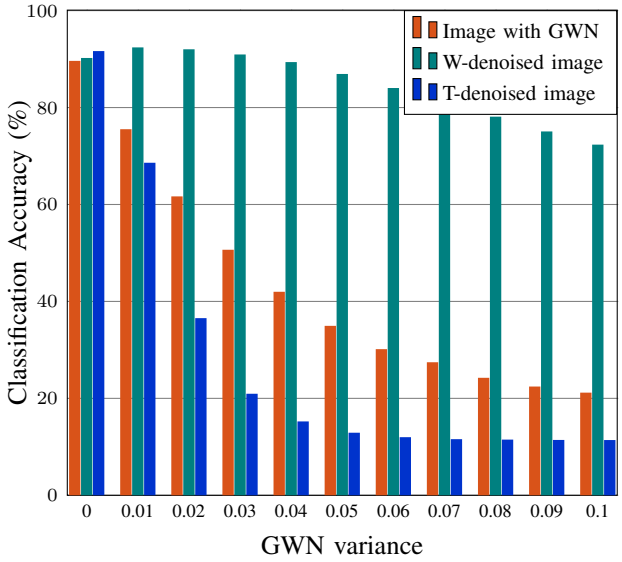


Fig. 3: Comparison of classification accuracy among the original digits, the w-denoised digits, and the t-denoised digits, where the learn network exhibits a 89.27% accuracy on the MNIST training set and 89.63% accuracy on the test set.

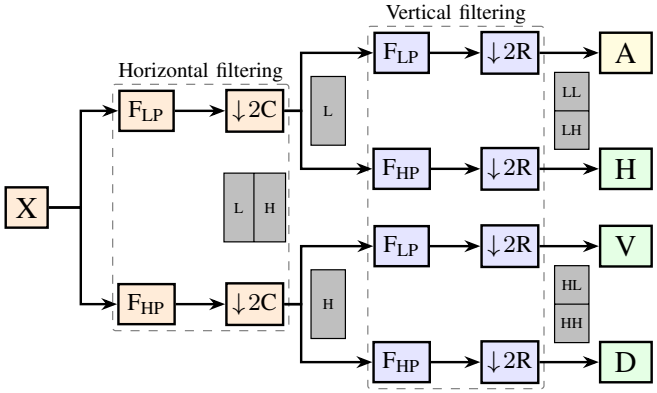
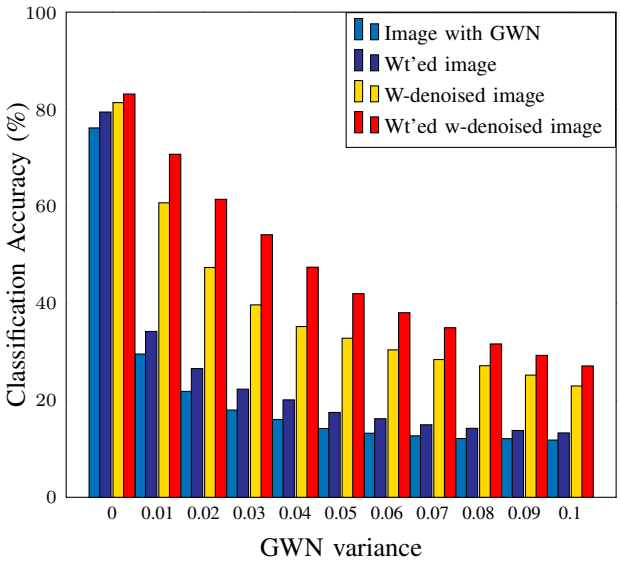


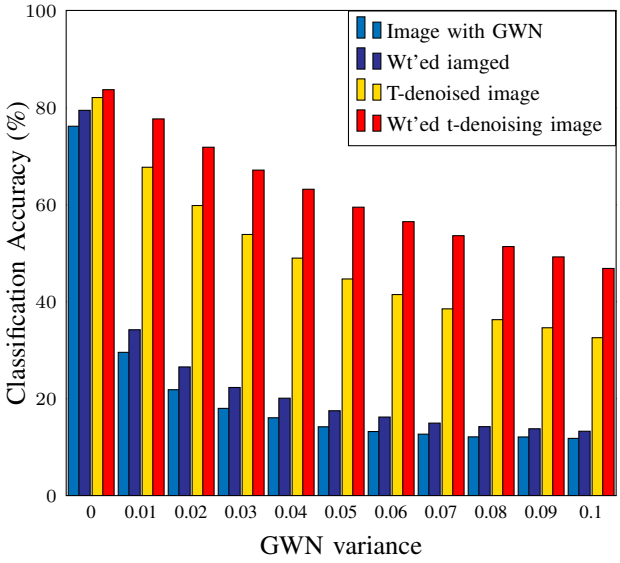
Fig. 4: Decomposition of an image using discrete wavelet transform.

It is worth noting that for a weighted voting-based machine learning prediction model, if one individual classifier has very high or nearly perfect prediction accuracy, using ensemble classification may not improve the prediction accuracy or even be meaningful. Therefore, to analyze the potential enhancement of using a weighted voting-based ensemble, we crafted a training algorithm on the original MNIST handwritten digits. This algorithm can achieve 83.59% prediction accuracy on the training set and 76.21% accuracy on the testing set.

We conduct simulations to evaluate the classification accuracy of images with zero-mean Gaussian white noise, where the variance increases from 0 to 0.1 in steps of 0.01. Our simulation results show that using weighted voting-based schemes can increase classification accuracy by an average of 2.39%. Even when the variance is 0 (means no noise is



(a) Comparison between weighted and unweighted classifications regarding the original and the w-denosing.



(b) Comparison between weighted and unweighted classifications regarding the original and the t-denosing

Fig. 5: Comparison between unweighted and weighted voting-based classification accuracy with and without noise reduction, where the original training classification accuracy is 83.59% and the testing accuracy is 76.21%.

added), we can still achieve a 3.28% increase in classification accuracy. The accuracy increase reaches the maximum when the variance equals 0.02, which is 4.7%, as shown in Fig. 5. Overall, by combining w-denoising and weighted voting, we can achieve an average increase of 28.24%.

We also compare the use of weighted voting-based schemes in both w-denoising and t-denoising scenarios. In w-denoising, we achieve an average classification accuracy improvement of 8.87%. When the variance equals 0.03, the accuracy increase reaches its maximum, which is 14.47%. In contrast, in t-denoising, we can achieve a significant 14% performance improvement on average. When the variance equals 0.07, the accuracy reaches its maximum, which is 15.1%. It should be noted that in all these simulations, the voters are not independent. Overall, by combining w-denoising and weighted voting, we can achieve an average increase of 44.31%

Our simulation results show that the accuracy of predictions by machine learning models significantly depends on the choice of learning algorithm. It also relies on the selection of data points or specific features during the training process.

V. Conclusion

In this paper, we experimentally investigated how to improve the image classification accuracy of noise-polluted iamges under two techniques: image denoising and weighted voting. Our simulation results conducted using MNIST database of handwritten digits showed that the image classification accuracy can be significantly improved through image denoising algorithms. We also found that image classification accuracy can be improved through weighted voting. However, by combining image denoising and weighted voting, we can maximize image classification accuracy.

Acknowledgment

This research was supported in partial by National Science Foundation under award 1919154.

References

[1] DeepLobe, "Top 15 industrial applications for image classification." [Online]: https://deeplobe.ai/top-15-industrial-applications-for-image-classification/.
[2] D. P. Cattin, "Image restoration: Introduction to signal and image processing," *MIAC, University of Basel. Retrieved*, vol. 11, p. 93, 2013.
[3] X. Tian and E. Samei, "Accurate assessment and prediction of noise in clinical ct images," *Medical physics*, vol. 43, no. 1, pp. 475–482, 2016.
[4] J. Ohta, *Smart CMOS image sensors and applications*. CRC press, 2020.
[5] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, and M. Adjouadi, "A comprehensive survey on impulse and gaussian denoising filters for digital images," *Signal Processing*, vol. 157, pp. 236–260, 2019.
[6] W. Knight, *The Dark Secret at the Heart of AI*. Cambridge, MA, USA: MIT Technology Review, 2017.
[7] D. Castelvecchi, "Can we open the black box of ai?," *Nature News*, no. 7623.
[8] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning (ICML)*, pp. 1–11, 2017.
[9] Z. C. Lipton, "The mythos of model interpretability," in *International Conference on Maching Learning (ICML) Workshop*, pp. 1–8, 2016.
[10] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Int. Conf. Mach. Learn. (ICML)*, pp. 1–11, 2017.
[11] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information." https://arxiv.org/abs/1703.00810, 2017.
[12] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization." https://arxiv.org/abs/1610.02391, 2016.
[13] "wdenoise2 – wavelet image denoising." https://www.mathworks.com/help/wavelet/ref/wdenoise2.html.
[14] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits." https://yann.lecun.com/exdb/mnist/.