Stability Analysis of Deep Neural Networks under Adversarial Attacks and Noise Perturbations

Parisa Eslami
Department of Information Systems
University of Maryland, Baltimore County
Batimore, MD 21250 USA
peslami1@umbc.edu

Houbing Song
Department of Information Systems
University of Maryland, Baltimore County
Batimore, MD 21250 USA
songh@umbc.edu

2024 International Wireless Communications and Mobile Computing (IWCMC) | 979-8-3503-6126-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IWCMC61514.2024.10592478

***Abstract*—The model's uncertainty estimation is what adversarial attacks target and exploit. However, the attacker may not always have a clear understanding of how the model estimates uncertainty. Different types of uncertainty can impact these attacks differently because not all sources of uncertainty are equally affected by the attack strategies. We investigate the impact of various noise distributions as well as adversarial attacks on distinct networks. Our objective is to establish a threshold for quantifying network robustness. To achieve this, we analyze three network models, progressively increasing the depth of layers. Our analysis incorporates Shannon entropy and Kullback-Leibler divergence to assess model uncertainty. This directs our attention toward identifying a criterion for measuring uncertainty. Notably, our findings shed light on the manipulation of neural network uncertainty through adversarial attacks, highlighting variations across diverse datasets and models.**

***Index Terms*—Adversarial Examples; Uncertainty; Kullback-Leibler Divergence (KLD); Deep Neural Networks (DNN); Network Robustness.**

I. Introduction

The rapid development of Artificial intelligence (AI) and Machine Learning (ML), particularly in Deep Neural Networks (DNNs), has ushered in a new era of technological breakthroughs across various domains. Despite these achievements, there are growing concerns about the security and robustness of these technologies. The integration of AI and ML into critical systems makes understanding and addressing these vulnerabilities not just a technical challenge, but also a necessity for ensuring the reliability and safety of these systems in real-world applications.

Central to these concerns is the threat posed by adversarial examples. These are inputs deliberately designed to look nearly identical to original data but are capable of fooling DNNs into making incorrect predictions [1], [2]. The existence of adversarial examples exposes a critical weakness in machine learning models, particularly in DNNs, challenging their reliability and questioning their effectiveness in practical applications [3].

In this paper, we provide a comprehensive analysis of adversarial examples, focusing on their formation and impact on DNNs. We explore a variety of attack methods, employing techniques such as gradients, optimization, and mathematics to create these deceptive inputs. Our study includes an examination of both white-box attacks, which necessitates knowledge of the model's architecture [4], and black-box attacks, which do not [5]. Additionally, we investigate targeted attacks that manipulate the model's output to a specific class [6], as well as untargeted attacks aimed at inducing any misclassification [7]. Furthermore, we explore the role of model entropy as a measure of uncertainty in predictions [8], analyzing how adversarial perturbations can manipulate the confidence levels of DNNs in classifying data, thereby uncovering their vulnerabilities and suggesting improvements for robustness [9].

Our work distinguishes itself from previous studies by providing a detailed exploration of various adversarial attack methods and their implications on DNNs. Unlike existing research that predominantly focuses on identifying vulnerabilities, our paper extends to proposing potential solutions for improving the robustness of DNNs. We also offer a comparative analysis of our proposed DNN network model against other models in the contex of these adversarial threats, highlighting the unique contributions and advancements of our research.

The remainder of the paper is structured as follows: Section II delves into an in-depth exploration of Adversarial attacks and their prevalent methods. Section III conducts a comparative analysis between our proposed DNN network model against alternative models. In Section IV, we quantify the level of uncertainty present within the dataset, including both training and testing sets. Section V engages in a comprehensive discussion of the findings, accompanied by comparative diagrams for enhanced clarity. The paper concludes in Section VI, summarizing the key takeaways of our study.

II. Related-work

Adversarial attacks on DNNs have emerged as a critical concern, prompting extensive research efforts toward enhancing the robustness and security of these systems. To overcome the ambiguous causes of adversarial attacks, several different mechanisms can be employed. Uncertainty inference (UI) has been introduced for estimating how uncertain the outputs of DNNs are to further improve their reliability and applicability [10]. DNNs are unable to evaluate the uncertainty of their

This research was partially supported by the National Science Foundation under Grant No. 2309760 and Grant No. 2317117.

outputs since they prefer to generate precise predictions as opposed to choosing between confidence intervals. Two broad categories of uncertainty are model uncertainty and data uncertainty. To determine the degree of generalization uncertainty in models, the first method is known as misclassification detection which involves finding out-of-domain occurrences. The second method is assessing data uncertainty which focuses on finding out-of-distribution occurrences for noisy data. [11]. This paper [12] introduces shallow information and obtains predictions from different depths of a DNN, they adopt a multi-head architecture which is similar to GoogLeNet-style [13]. After each fully connected layer, they obtain multiple predictions which are treated as a sample for Dirichlet distribution estimation from each head. This is to produce a MoGMM-FC layer, the classifier of a Deep Neural Network (DNN) is combined with a mixture of Gaussian Mixture Models. Building upon the concept of leveraging various depths within a Deep Neural Network (DNN), as described in the previous paragraph, this paper [14] presents and compares two uncertainty assessment techniques that do not rely on test data. The Shannon entropy of class probabilities predicted by Deep Neural Networks (DNN) and Random Forest (RF) is used to implement uncertainty assessment at the pixel level. This research paper explores the effects of various adversarial attacks and techniques during the training and testing phases. We analyze Gaussian, uniform, salt-and-pepper noise to the related datasets and use popular attack methods to evaluate their robustness. We discuss the effectiveness of adversarial attacks and the impact of entropy on model confidence. Further, we assess the uncertainty level of noise in each dataset by establishing a noise threshold by exploiting the KL divergence concept [15]. This means understanding how the uncertainty of a model changes when exposed to adversarial perturbations.

III. Experiment

To gauge the uncertainty inherent in the network's response to adversarial attacks, we introduce controlled noise or perturbation into the dataset. The ensuing uncertainty levels in both input and output domains are quantified, providing valuable insights into the network's sensitivity to perturbations.

A. Experiment Design

1) Dataset: To elucidate the impact of various attack methods on the network's behavior, highlighting the interplay between adversarial attacks, network performance, and uncertainty, we utilized three popular datasets such as *MNIST*, *CIFAR-10*, and *ImageNet*. MNIST dataset consists of 70,000 grayscale images of handwritten digits from 0 to 9. These images Contain 60,000 images (size of 28x28 pixels) for model training and 10,000 images for evaluating the trained models' performance. CIFAR-10 is a dataset that contains 60,000 (32x32) color images in 10 different classes. The dataset is divided into five training batches and one test batch, each with 10,000 images. The test batch contains exactly 1000 randomly selected images from each class. ImageNet is an image database that contains over 14 million images. We chose 10 categories from ImageNet: goldfish, ostrich, axolotl, chameleon, hummingbird, admiral, violin, ice cream, teapot, and rapeseed, with each category containing 1300 training images and 50 test images with the dimensions (224x224x3).

2) Uncertainty Metrics: We briefly describe the metrics, such as Shannon entropy and Kullback-Leibler divergence, and how they capture uncertainty. Entropy, encompassing uncertainty or a deficiency in the level of confidence during a model's decision-making process [16], signifies that the model's evaluation of situations or predictions lacks a firm conviction or definitive determination. Consequently, data points exhibiting higher entropy are prioritized for labeling, as they possess the potential to offer the most informative insights to the model [17]. Furthermore, elevated entropy values within a sample may indicate its affiliation with an unseen or out-of-distribution (OOD) category, rendering it valuable for tasks such as anomaly detection and model robustness assessment [18]. Techniques grounded in entropy, such as Bayesian neural networks and Monte Carlo dropout, prove instrumental in capturing epistemic uncertainty, which encompasses uncertainty stemming from limited data and model ambiguity [19]. On the other hand, KLD serves as a useful tool for quantifying the distinction (referred to as distance) between two distributions [20]. This concept typically considers when the same fault pattern between distinct datasets may exhibit varying degrees of fault magnitude.

B. Experimental Setup

1) Network Architecture: The CNN model is designed using sequential layers, comprising convolutional layers (consisting of 32, 64, 128, 256, and 512 filters respectively, with a kernel size of (3, 3), 'relu' activation, and 'same' padding) followed by max-pooling (with a pool size of (2, 2)) layers to extract hierarchical features from the input images. The model further includes dense (fully connected with 512, 256, 128, 64, and 32 units respectively, each activated by 'relu' activation.) layers to perform classification based on the learned features. The model's loss function is categorical cross-entropy, which is commonly used for multi-class classification tasks. The Adam optimizer with a learning rate of 0.001 is employed for model optimization. The model's performance is evaluated on both clean and noisy test datasets. The level of noise is introduced using Gaussian noise with varying standard deviations, Uniform noise, and Salt-and-Pepper noise with a valid range.

2) Uncertainty Evaluation: In this work, we propose three different CNN architectures: model 1 (7-layers) denoted as **C**, model 2 (10-layers) denoted as **D**, and model 3 (13-layers) of convolutional neural network layers denoted as **E**. For each model, we consider two networks; the models trained without noise are $\mathbf{C}_0$, $\mathbf{D}_0$, and $\mathbf{E}_0$, and the models trained with noise are $\mathbf{C}_1$, $\mathbf{D}_1$, and $\mathbf{E}_1$. For each model we test the networks once with a clean test dataset and once with a poisoned test dataset. In general we have clean dataset elements, $(x_{a_0}, y_{a_0})_{clean}$, and noisy dataset elements, $(x_{a_1}, y_{a_0})_{noisy}$, where we add noise by $x_{a_1} = x_{a_0} + n$. Then we analyzed our models

by evaluating the Shannon Entropies and KL Divergences. The Shannon Entropy was calculated within each individual dataset, while the KL Divergence was calculated between the true label distribution and predicted label distributions from each test dataset. Analyzing the Shannon entropy helps us understand the randomness introduced by the adversarial noise; higher entropy values indicate greater uncertainty in model predictions. The KL divergence allows us to assess how the noise affects the model's understanding of the data. Higher KL divergences signify more substantial differences between two distributions and a larger degree of dissimilarity in the distributions.

IV. Discussion

A. Experimental Result

In our proposed Convolutional Neural Network (CNN) model aimed at achieving accurate outcomes, we made a deliberate decision to forgo the inclusion of several established techniques such as batch normalization, dropout, and regularization techniques like L1 and L2 regularization. While these methods are known to enhance the robustness of neural networks, our rationale behind this choice was rooted in our research focus on understanding the precise extent of ambiguity present in both the model's input and output. By intentionally keeping the network architecture more straightforward, we aimed to capture a clearer representation of the uncertainty intrinsic to the data and the model itself. We conducted tests using three different noise distributions, as shown in Table I. The results reveal interesting patterns. In a scenario where the training dataset is clean but the testing dataset is poisoned with 1% to 5% noise, Salt-&-Pepper noise proves to be more detrimental to the CIFAR-10 dataset. When the training dataset is poisoned but the testing dataset is clean, Uniform distribution is more effective in causing misclassification for 1% to 5% noise levels. In the last scenario, where both the training and testing datasets are poisoned, Uniform and Gaussian noise exhibit more impact at 1% and 5% noise levels, respectively.

TABLE I
Training and Testing our proposed model with different levels of noise distributions in to *CIFAR-10* dataset.

Train	Test	Noise	S & P	Uniform	Gaussian
Clean	Clean	0%	87.50%	87.50%	87.50%
Clean	Poison	1%	54.01%	70.13%	68.91%
		5%	53.66%	64.99%	54.18%
Poison	Clean	1%	87.22%	69.42%	69.77%
		5%	87.21%	63.12%	68.13%
Poison	Poison	1%	86.08%	70.06%	70.52%
		5%	62.31%	65.32%	62.03%

Table II, presents an evaluation of the performance of our proposed network model (Model 3) under the mentioned attack methods, comparing it to relevant benchmarks. This evaluation highlights the effectiveness of our approach in mitigating adversarial effects. As we can see, *MNIST* dataset is more robust compared to other datasets. FGSM attack is malevolent in the *CIFAR-10* dataset and PDG has a more misleading effect on *ImageNet* dataset.

TABLE II
Testing Our Proposed Model with Popular Adversarial Attacks with Different Datasets.

Attack	MNIST	CIFAR-10	ImageNet
FGSM	83.19%	76.9%	78.91%
PDG	82.32%	78.6%	77.01%
C&W	81.21%	77.10%	79.80%

In Table III, our model was subjected to testing. In the absence of any attacks, we achieved the highest accuracy across three datasets with the *MNIST* dataset. However, when noise was introduced to our test dataset, the results indicated that Gaussian noise had the most adverse impact on accuracy for the *ImageNet* dataset, while salt-&-pepper noise had the greatest negative effect on *CIFAR-10*. Additionally, uniform noise had a notable impact on the accuracy of *CIFAR-10* as well.

TABLE III
Training and Testing our proposed model with different noise distributions.

Attack	MNIST	CIFAR-10	ImageNet
NA	88.05%	87.50%	79.89%
Gaussian	81.43%	79.98%	74.13%
Salt & Pepper	80.18%	76.00%	76.79%
Uniform	80.65%	75.90%	77.12%

B. Evaluating Shannon Entropy in each Model

After evaluating model performance under attack, we investigated the uncertainty and sensitivity of our models to noise. According to Fig. 1, higher entropy values may be obtained by increasing the depth of a model.

This may be due to the fact that deeper models usually recognize more complex patterns in the data. Consequently, this can increase the unpredictability of models from the increase in noise sensitivity. In turn, Shannon entropy values may be lowered by making the model well-regularized and more robust. Regularization techniques like dropout, batch normalization, and adversarial training can help mitigate the impact of noise on model predictions, but as we mentioned before, we overlooked these techniques.

For Model 1 in Fig. 1, the average entropy does not show a clear trend with the increase in noise percentage, suggesting that the model's output distribution may not be significantly impacted by noise. The entropy values across different data conditions (clean train data & poisoned test data, poisoned train data & poisoned test data, and poisoned train data & clean test data) are relatively close together, which could imply that Model 1 maintains a consistent level of uncertainty across these conditions.

Model 2's entropy values display slight variations with the increase in noise percentage, although, like Model 1, there isn't a clear consistent trend. The variation in entropy between

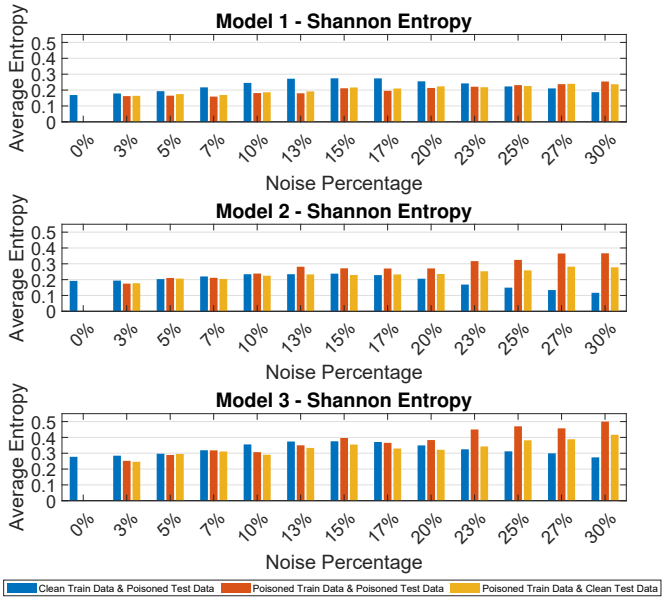

Fig. 1. Shannon entropy for Models 1, 2, and 3.

the different data conditions is more pronounced compared to Model 1, indicating that Model 2's response to data integrity is more sensitive. This model may show different levels of uncertainty depending on whether the train or test data is poisoned, as well as the combination of both.

The entropy for Model 3 exhibits similarly subtle fluctuations with increasing noise percentages, without a strong indication of a trend. However, the entropy levels in Model 3 are more varied across the different data conditions compared to Model 1, but less so than Model 2. This suggests that Model 3 has a differential response to the various types of data poisoning, with some configurations leading to higher uncertainty in the model's output distribution than others.

Across all models, the lack of a clear upward or downward trend in entropy with increasing noise suggests that the models may have varying degrees of resilience or sensitivity to noise, depending on the specific condition of the data. The relatively close grouping of entropy values under different conditions for each model also indicates that the models may not be highly sensitive to the type of data poisoning. This could be due to the models' inherent robustness to such issues or because the noise does not lead to a significant increase in uncertainty as captured by Shannon entropy. Each model exhibits its unique characteristics in handling data integrity, which could inform their application in environments where data quality can be compromised.

C. Evaluating KL Divergence in each Model

In Fig. 2, increasing noise levels boost entropy, amplifying unpredictability and pixel value variation, causing more uncertainty in predictions. Beyond a threshold, excessive noise dominates, leading to predictions converging towards randomness and reduced entropy due to overfitting.

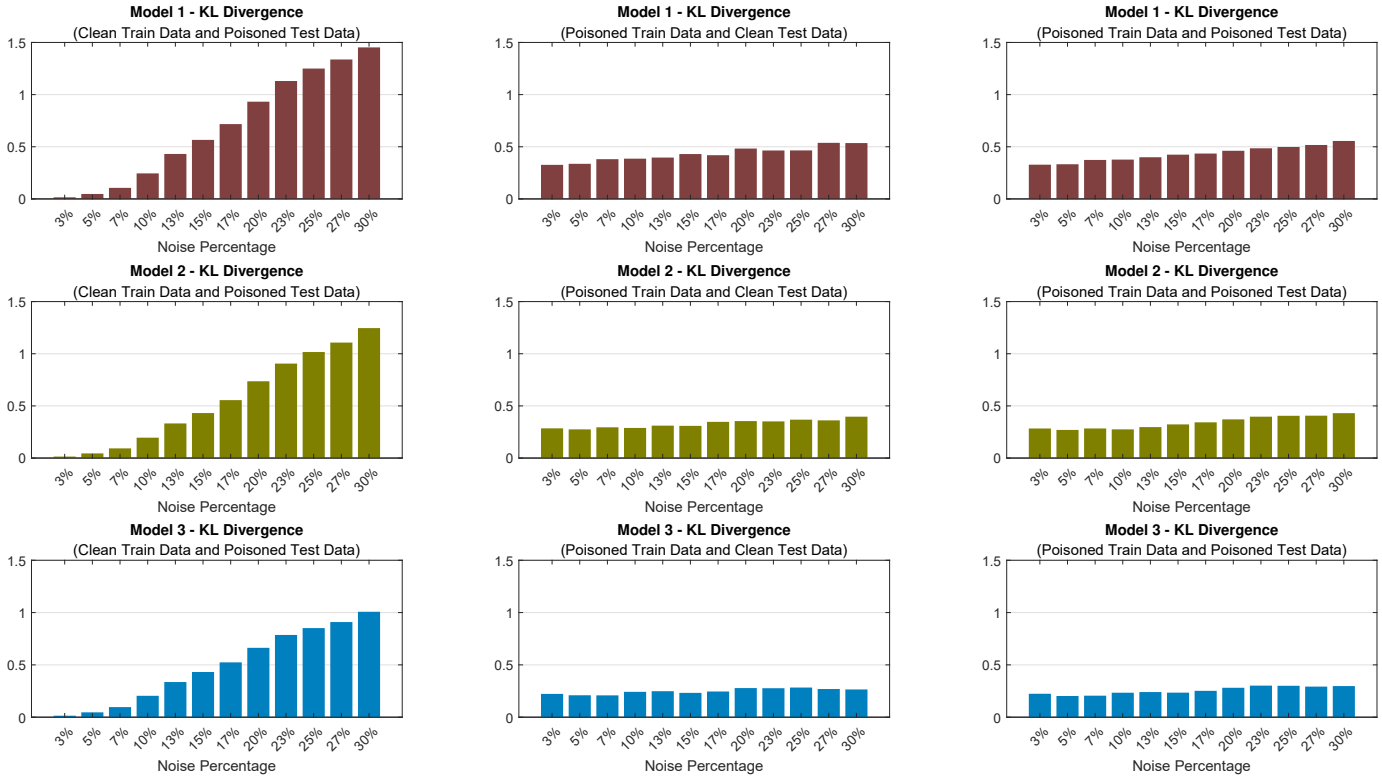

Fig. 2. KL Divergence for each model.

High noise levels flatten pixel distributions, reducing prediction diversity and entropy as the neural network struggles with noisy patterns. The impact of model depth on KL divergence is complex, depending on training and architecture. Deeper models may have varying KL divergence compared to shallower ones. Overfitting to training noise can result in higher KL divergence when exposed to adversarial noise while well-generalizing deeper models might exhibit lower KL divergence when attacked.

For Model 1, the graphs indicate an upward trend in KL divergence as the noise percentage increases. This pattern is observed when the training data is clean and the testing data is poisoned, when the training data is poisoned and the testing data is clean, and also when both are poisoned. The most significant increase in divergence occurs when both training and testing data are poisoned, suggesting that Model 1's performance is highly sensitive to the presence of noise in the data.

In the case of Model 2, the observed KL divergence values are somewhat lower than those of Model 1, particularly when the training data is not poisoned. This suggests a relative robustness in Model 2's ability to handle noise, especially during the testing phase. Nonetheless, the same general trend remains: as the noise level rises, so does the KL divergence.

Model 3 exhibits the lowest KL divergence values among the three models, indicating a stronger robustness to data poisoning. Even as the noise percentage increases, the KL divergence for Model 3 rises more gradually compared to the other models, which might reflect a better capacity to manage noisy data.

Overall, the increasing trend of KL divergence with higher noise percentages across all models is consistent with the expectation that noise leads to a greater deviation from the expected distribution. However, Model 3 stands out as being the most robust against data corruption, as demonstrated by its consistently lower KL divergence values under all tested conditions. This analysis suggests that Model 3 may be the most suitable choice when working with data that is at risk of being compromised or is inherently noisy.

V. Summary and Conclusions

Our study provides valuable insights into the realm of adversarial attacks on DNNs by introducing the concept of uncertainty quantification as a crucial factor. Through our evaluation of the network's responses when subjected to various attack strategies and by estimating uncertainty levels, we underscore the significance of considering uncertainty in the pursuit of bolstering DNN robustness. The objective is to assess the resilience of each model against these adversarial manipulations, shedding light on the broader landscape of security.

Our research delves into a specific domain of adversarial machine learning, offering a fresh perspective on the vulnerabilities of Deep Neural Networks (DNNs) to adversarial attacks. By integrating the concept of uncertainty quantification into our analysis, we have opened up a vital dialogue on the importance of understanding and managing uncertainty in these models. Through our experiments, we examined the network's behavior under various adversarial strategies, measuring the levels of uncertainty that manifest within the models' predictions.

Our analysis of various models' responses to noise-infused inputs, as evidenced by the changes in Kullback-Leibler divergence and Shannon entropy across multiple scenarios, provides an empirical foundation to assess model resilience. It is evident that different models exhibit distinctive patterns of uncertainty, indicating that a one-size-fits-all solution to bolstering robustness is inadequate. Instead, a model-specific approach is warranted.

The KL divergence results reveal the extent to which each model's predicted probability distribution diverges from the expected outcome under adversarial conditions. These findings suggest that as the noise level increases, so does the divergence, although the rate of increase and the overall impact vary by model. Model 3, in particular, showcased a comparatively lower divergence, suggesting a higher degree of resilience against adversarial attacks. Similarly, the Shannon entropy metrics underscored the models' varying degrees of uncertainty under poisoned data conditions, providing insights into the models' information processing stability.

In conclusion, our study not only emphasizes the significance of considering uncertainty in enhancing DNN robustness but also sets the stage for future explorations into adaptive defensive strategies. By assessing the resilience of each model against adversarial attacks, our work contributes to the broader landscape of cybersecurity. As adversarial threats evolve, our methodology offers a framework for ongoing evaluation and improvement, ensuring that DNNs can be trusted even in the most challenging of circumstances.

VI. Future Work

Building upon our current research into the resilience of Deep Neural Networks (DNNs) to adversarial attacks, our future work aims to expand the scope and depth of our understanding in this critical area. Our next step is to refine uncertainty quantification methods by exploring advanced statistical techniques and applying them to a wider variety of DNN architectures. This would help in discerning the subtleties in model responses across different layers and structures, providing a more detailed view of where vulnerabilities lie.

To address the vulnerabilities uncovered, we plan to develop and test a suite of defensive mechanisms. These defenses would be directly informed by the uncertainty metrics we've studied, aiming to improve model hardening through training regimens that incorporate adversarial examples and through real-time detection systems that can identify and adapt to attacks as they occur.

A key component of future research will also involve the transferability of our findings. We aim to apply the insights gained from our initial models to new domains and applications, exploring how different types of data and tasks affect a model's susceptibility to manipulation. Considering

the broader implications of our work, we anticipate engaging with the ethical dimensions of robust AI systems. Future investigations will need to balance enhanced security with concerns such as data privacy, bias mitigation, and transparency.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation under Grant No. 2309760 and Grant No. 2317117.

REFERENCES

[1] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[2] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151–178, 2020.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[4] D. Bharath Kumar, N. Kumar, S. D. Dunston, and V. M. A. Rajam, "Analysis of the impact of white box adversarial attacks in resnet while classifying retinal fundus images," in *International Conference on Computational Intelligence in Data Science*. Springer, 2022, pp. 162–175.

[5] Y. Bai, Y. Wang, Y. Zeng, Y. Jiang, and S.-T. Xia, "Query efficient black-box adversarial attack on deep neural networks," *Pattern Recognition*, vol. 133, pp. 109037, 2023.

[6] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, 2018.

[7] R. K. Vigneswaran, R. Vinayakumar, K. Soman, and P. Poornachandran, "Evaluating shallow and deep neural networks for network intrusion detection systems in cyber security," in *2018 9th International conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2018, pp. 1–6.

[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[9] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[10] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1356–1367, 2020.

[11] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018.

[12] Y. Yang, S. Yang, J. Xie, Z. Si, K. Guo, K. Zhang, and K. Liang, "Multi-head uncertainty inference for adversarial attack detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] J. Xie, Z. Ma, J.-H. Xue, G. Zhang, J. Sun, Y. Zheng, and J. Guo, "Ds-ui: Dual-supervised mixture of gaussian mixture models for uncertainty inference in image recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 9208–9219, 2021.

[14] M. Shadman Roodposhti, J. Aryal, A. Lucieer, and B. A. Bryan, "Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest," *Entropy*, vol. 21, no. 1, pp. 78, 2019.

[15] S. Villena, M. Vega, S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Using the kullback-leibler divergence to combine image priors in super-resolution image reconstruction," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 893–896.

[16] I. Alarab and S. Prakoonwit, "Uncertainty estimation based adversarial attack in multi-class classification," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1519–1536, 2023.

[17] H. Zhang, W. W. Chen, J. M. Rondinelli, and W. Chen, "Et-al: Entropy-targeted active learning for bias mitigation in materials data," *Applied Physics Reviews*, vol. 10, no. 2, 2023.

[18] Y. Liu, K. Ding, H. Liu, and S. Pan, "Good-d: On unsupervised graph out-of-distribution detection," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 339–347.

[19] X. He, Y. Chen, and L. Huang, "Bayesian deep learning for hyperspectral image classification with low uncertainty," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[20] S. Ji, Z. Zhang, S. Ying, L. Wang, X. Zhao, and Y. Gao, "Kullback–leibler divergence metric learning," *IEEE transactions on cybernetics*, vol. 52, no. 4, pp. 2047–2058, 2020.