Neural SDEs for Robust and Explainable Analysis of Electromagnetic Unintended Radiated Emissions

Sumit Kumar Jha
Computer Science Department
Florida International University
sumit.jha@utsa.edu

Susmit Jha
Computer Science Laboratory
SRI International
susmit.jha@sri.com

Rickard Ewetz
ECE Department
Univ. of Central Florida
rickard.ewetz@ucf.edu

Alvaro Velasquez
Computer Science Department
Univ. of Colorado Boulder
alvaro.velasquez@colorado.edu

***Abstract*—We present an evaluation of the robustness and explainability of ResNet-like models in the context of Unintended Radiated Emission (URE) classification and suggest a new approach leveraging Neural Stochastic Differential Equations (SDEs) to address identified limitations. We provide an empirical demonstration of the fragility of ResNet-like models to Gaussian noise perturbations, where the model performance deteriorates sharply and its F1-score drops to near insignificance at 0.008 with a Gaussian noise of only 0.5 standard deviation. We highlight a concerning discrepancy where the explanations provided by ResNet-like models do not reflect the inherent periodicity in the input data, a crucial attribute in URE detection from stable devices. In response to these findings, we propose a novel application of Neural SDEs to build models for URE classification that are not only robust to noise but also provide more meaningful and intuitive explanations. Neural SDE models maintain a high F1-score of 0.93 even when exposed to Gaussian noise with a standard deviation of 0.5, demonstrating superior resilience to ResNet models. Neural SDE models successfully recover the time-invariant or periodic horizontal bands from the input data, a feature that was conspicuously missing in the explanations generated by ResNet-like models.**

I. Introduction

The unintended radiated emissions from electronic devices can provide a plethora of information to observers about the type of the electronic equipment as well as its current operating condition. Such emissions can be used for activities ranging from non-intrusive load monitoring to side-channel leakage of otherwise secure information. These unintended radiated emissions (UREs) from electronic devices occur due to non-ideal filters, manufacturing variations, and other design constraints, including but not limited to signal modulation, frequency mixing, and high-frequency clocking of signals in digital circuits.

The task of mapping detected UREs to specific devices or operating conditions can be thought of as a classification problem. Efforts have been placed towards understanding machine learning classification tasks for unintended electronic emissions. In particular, the Oak Ridge National Laboratory has created the Flaming Moe data set [1] of real-world unintended electronic emissions to allow for the design of new URE detection and analysis algorithms. The dataset has been obtained by studying 18 devices and observing two 10-minute segments of voltage data captured at 2 million samples per second. Together with the dataset, the team released the dimensionally aligned signal projection algorithm as a new approach for creating low-dimensional features for URE classification applications [2].

The Flaming Moe data set has served as a robust benchmark for learning neural network models for URE classification [3]. It has been argued that an out-of-the-box residual neural network model is eminently capable of classifying the URE signals into 18 classes. In fact, A small residual neural network model can achieve perfect accuracy on a held-out fragment on the Flaming Moe data set. This may cause an observer to conclude that further research on neural network-based analysis of the Flaming Moe data set is unnecessary. We show that such a conclusion is not true, and the learned residual neural network model begins to fail miserably even in the presence of noise in the observed data.

Earlier work on analyzing the Flaming Moe URE data set has realized the importance of communicating the reason for the classification or an explanation from the black-box neural network model to the end human user. In particular, local interpretable model-agnostic explainability has been used to explain residual neural networks designed for this purpose. However, these explanations, as shown in Fig. 4 of [3], do not obey the inductive bias of the data set that the URE signal observed in the short-term Fourier transform is often periodic or time-invariant, and any robust attribution that

The authors acknowledge support from the National Science Foundation awards 2319401, 2113307, and 1908471, the DARPA cooperative agreement HR00112020002, DARPA FA8750-23-2-0501, ONR grant N000142112332, DOE grant DE-SC0023494, DE-SC0024428, DE-SC0024576, and the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

covers all explanations should uncover this fact. In the case of images, an explanation occurring as a horizontal band in our images uncovers this inductive bias inherent in the data set. We show that modern explanation methods like integrated gradients [4] with smoothgrad [5] uncover this inductive bias in our data set when applied to models based on neural stochastic differential equations. In summary, the contributions of this paper are as follows:

1) An empirical demonstration of a lack of robustness to noise for ResNet-like models in the context of URE detection: Our investigation has unearthed certain limitations in the existing residual neural network models. They have been found to be quite fragile, readily falling victim to the addition of Gaussian noise. When perturbed with Gaussian noise of a standard deviation of 0.25, the F1 score drops to a mere 0.41. Increasing the standard deviation to 0.5 results in an almost negligible F1 score of 0.01. See Table I.
2) An empirical demonstration of the lack of periodicity for explanations of short-term Fourier transform images in ResNet-like models: Another concern arises from the data's inherent inductive bias that creates periodic behavior or horizontal bands in the input short-term Fourier transform image. These bands are surprisingly absent in the explanations provided by the ResNet-like models, which are instead interspersed with positive and negative attribution input features in both horizontal and vertical dimensions. See Fig. 1.
3) A novel application of models based on neural SDEs in building robust and explainable URE detection models: We present a more robust and explainable framework in the form of Neural Stochastic Differential Equations (SDEs). When compared to the ResNet models, the Neural SDEs are remarkably resilient against Gaussian noise. For instance, when exposed to Gaussian noise with a standard deviation of 0.5, these models retain an F1 score of 0.93, as compared to the near-zero score of ResNet models. Furthermore, the explanations generated by the Neural SDE models recover the inherent inductive bias in the input, clearly displaying time-invariant horizontal bands. See Table II and Fig. 2.

In essence, our work is centered around leveraging the power of neural SDEs to create models for URE classification that are more robust to Gaussian noise, explainable, and aligned with the innate properties of the data. This approach offers a promising path for the development of sophisticated URE detection algorithms by showing that the current generation of *robust* residual neural network models do not achieve a perfect 100% accuracy on the Flaming Moe data set, thereby highlighting that this data set remains valuable for building robust explainable neural network models in the future.

II. Related Work

A. Data Collection and Dataset

To assess the design of machine learning and related classification algorithms, Unintended Radiated Emission (URE) has been collected from 18 commercially available electronic devices, commonly found in an office environment. This Flaming Moe dataset [1], generated by Oak Ridge National Laboratory in 2016, serves as an idealized URE dataset for the development of URE detection and classification models. Data collection was organized into four 10-minute segments with the device being operations only in alternate 10-minute windows. Each segment was further split into 1200 files representing 1 second of data for each device. A one minute delay was imposed prior to device capture in order to allow the device to boot and achieve stable performance. This steady behavior creates an inductive bias in our experiments that should be uncovered by any sound and robust explanation approach.

The collection of URE signals for the data set took place within a Radio Frequency shielded enclosure, employing a USRP N210 collection platform that featured a temperature-compensated crystal oscillator. This oscillator is deemed suitable for low-cost and routine industrial applications due to its modest frequency accuracy of 2.5 ppm, which is adequate for low-frequency signals observed in our applications.

B. ResNet Models for URE

Recent work [3] has argued that a small residual network is capable of achieving a perfect test accuracy on the Flaming Moe data set. We were able to reproduce this rather unusual result in our own experiments. However, we found that inserting a Gaussian noise with a standard deviation of 0.5 results in an almost negligible F1 score of 0.01. Hence, the off-the-shelf residual neural network model is very fragile and may not be suitable for real-world data analysis where such non-adversarial noise may be inevitable.

C. Attribution Methods

Several state-of-the-art attribution methods have been developed over the last decade with increasingly higher degrees of success. However, to the best of our knowledge, we are the first to bring these more contemporary attribution methods to the analysis of neural networks analyzing unintended radiation emission (URE) from devices using the Flaming Moe data set.

Salient methods, like the Layer-wise Relevance Propagation [6], decomposes the contribution of each neuron in a network to the final prediction, providing a detailed "relevance" map. Another technique, known as Shapley Additive Explanations [7], maps each input feature to its numerical importance for a given prediction of the neural network. This approach leverages game theory concepts, attributing the impact of each feature on the neural network response in a way that ensures fairness and consistency.

Integrated Gradients [4] works by connecting the response of a deep neural network to the features in its input through the concept of path integrals. This is an axiomatic method that provides a simple and intuitive way of understanding the feature attributions. Grad-CAM [8], employs the gradients or derivatives of a target class from a given convolutional layer. Usually, the gradients from the final layer are used to construct a rough localization map that highlights those features in the input that lead to a given prediction.

More recent research directions are seeking to improve upon these methods, aiming to make them more robust and consistent, to handle complex scenarios with higher reliability. For instance, SmoothGrad [5] and Stochastic Differential Equations [9] have led to more robust attributions with smaller sensitivity scores for other image data [10].

III. Approach

Our approach first employs short-term Fourier transform to transform time-series data into visual images. Since Flaming Moe data set has two continuous recordings of 600 seconds with 2 million samples per second, we obtain 2.4 billion samples per device for analysis. We create short-term Fourier transforms using 1 million samples each; thereby, creating 2,400 images per device and obtaining a data set of 43,200 images. For our analysis, we create both deterministic and stochastic variants of residual neural networks, and analyze the robustness and explainability of the model using currently popular attribution methods.

A. Neural Stochastic Differential Equations

Building upon advances in modeling neural networks as dynamical systems [11], our work exploits recent neural stochastic differential equations (SDEs) [12], [13] extensions of these models to encompass stochastic behavior. In this section, we briefly recall recent results [9], [12] related to our work. Put succinctly, ResNets can be interpreted as discretizations of neural ordinary differential equations (ODEs) and stochastic variants of residual networks can serve as approximations of neural stochastic differential equations (SDEs). Both inference and training processes in ResNets can be represented using dynamical systems [11], [14]–[17]. A fundamental ResNet unit, with a residual $R(X(i), W(i))$, can be defined as follows:

$$X(i+1) = X(i) + R(X(i), W(i)) \tag{1}$$

where $X(i)$ denotes the input to the i^{th} block and $X(i+1)$ signifies the block's output that is then fed into the succeeding unit. Here, $W(i)$ indicates the learned weights in the respective residual neural network block. Specifically, in this notation, $X(0)$ denotes the network input $\mathbf{x}$ and the network output $\mathcal{F}$ is denoted as $X(T)$. Upon applying suitable limits, we can formulate the evolution of the residual neural network as an ordinary differential equation:

$$\frac{dX(t)}{dt} = G(X(t), W(t)) \tag{2}$$

Here, $G(X(t), W(t)) = \lim_{\delta t \to 0} \frac{R(X(t), W(t))}{\delta t}$ and $X(0)$ is the neural network input.

To model a stochastic variant of the ResNet, a noise term $N(i)$ is added to the right-hand side of the earlier equation. The dynamical system for such residual neural networks with a noise component can be represented as an SDE:

$$dX(t) = G(X(t), W(t))\; dt + \sigma(X(t), t)\; dB(t) \tag{3}$$

Here, the noise is depicted as a Brownian motion term $B(t)$, scaled by a suitable diffusion coefficient $\sigma(X(t), t)$.

B. Robust Axiomatic Attributions

Modern axiomatic methods of attributions, including integrated gradients and their variants, satisfy several fundamental axioms that are not known to be satisfied by methods such as LIME. Integrated Gradients (IG) [4] is an attribution method widely used for feature importance analysis in deep neural networks. The attribution of an input feature in DNNs is typically computed with reference to a baseline input, denoted as $\mathbf{x}^b$. This baseline may be a Gaussian noise image in image-based tasks, or it could be a randomly generated set of inputs.

While integrated gradients and neural stochastic differential equations have been used to create robust attributions of ImageNet and similar images in the wild, we believe that we are the first to study the use of neural SDEs and integrated gradients on the explainable classification of unintended radiation emissions (UREs) from devices.

IV. Fragility and Poor Explainability of ResNet

A. Residual Neural Network Models

We first employed an image classification model based on the ResNet-50 architecture that was trained on 70% of the available data and tested on the other held-out 30% of the data set. The model is designed to classify images into one of 18 different classes. Before being fed into the model, the images are first resized to 224x224 pixels and are then normalized; the model is trained using a batch size of 32 and the Adam optimizer with a learning rate of 10^{-5}. The parameters of the model are updated based on the cross-entropy loss.

The training process continues for 10 epochs with the loss going down from 2.7 to 0.05. The evaluation is performed on

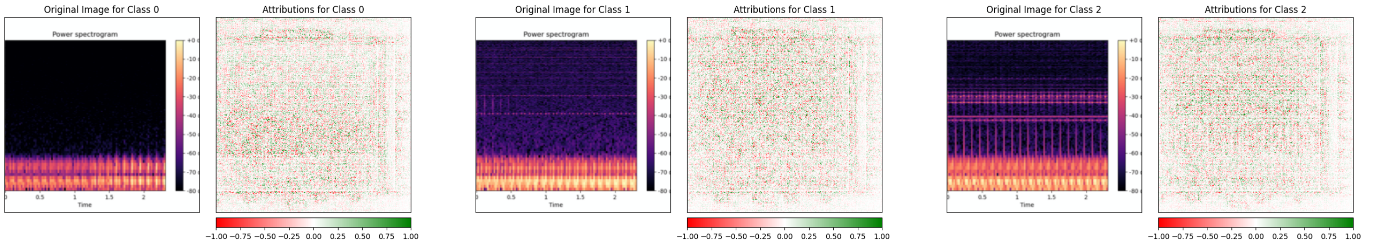

Fig. 1: ResNet-50 models with perfect accuracy produce attributions that do not support the inductive bias of the data. Horizontal repeated patterns in the data are not identified in the explanations using horizontal bands. Positive (green) and negative (red) attributions are interspersed indicating poor explainability in this context. Explanations for all 18 classes are presented in [18].

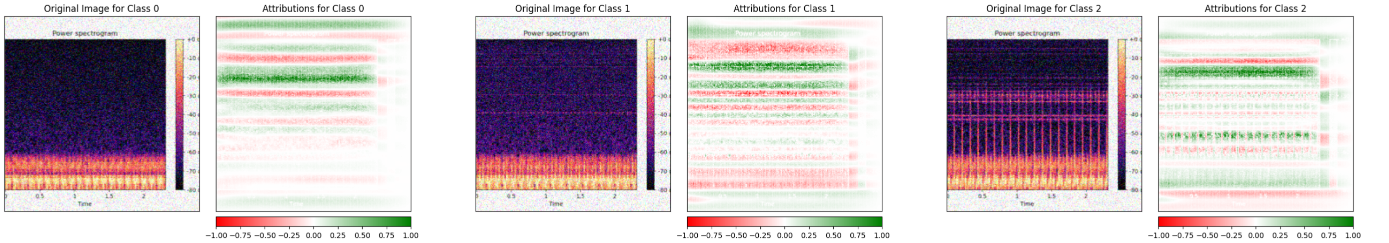

Fig. 2: Stochastic models produce attributions that uncover the shape or inductive bias of the input. The positive and the negative attributions occur in horizontal patches that conform to the fact that the data was obtained from a device under stable operation. Our accompanying report [18] includes attributions for all 18 classes.

a separate test set that constitutes 30% of the original dataset. The evaluation metrics include precision, recall, and F1-score, and are computed for each class separately, as well as an average over all classes. The results suggest that the model achieves near-perfect performance, with precision, recall, and F1-score of 1.00 for each of the 18 classes.

B. Robustness Analysis

We investigated the robustness of the learned URE ResNet-50 model under various noise conditions. We synthetically introduced Gaussian noise to the input data with different standard deviations (0.1, 0.25, 0.5), mimicking potential real-world scenarios where data can be distorted due to noise. We evaluated the model's performance across all classes under each noise level, recording the precision, recall, and F1 scores for each class and noise level.

The resulting data, presented in Table I, provides an overview of the model's fragility. As the standard deviation of the Gaussian noise increased to 0.25 and 0.5, the performance of the model deteriorated considerably, as evidenced from the decrease in average precision from 0.45 for a standard deviation of 0.25 to 0.01 for a standard deviation of 0.5.

C. Explanation using Integrated Gradients

We seek to gain an in-depth understanding of the decision-making process of the residual neural network model with perfect accuracy by visualizing the significant features contributing to each prediction. We employed Integrated Gradients [4], a popular interpretability technique, to identify these important features. We further used a Noise Tunnel [5] with Integrated Gradients to generate smoother attributions and reduce variability in the attributions. For each of the 18 classes under consideration, we selected one sample that was correctly classified by the model and computed the attributions for the input image. The computed attributions were then normalized and visualized in the form of heatmaps. The visualization presented two images: the original image and the corresponding heatmap. See Fig. 1. We observe that the attributions obtained from the standard residual neural network model do not conform to the inductive bias in the data set that contains nearly periodic or stable signals, and hence should contain horizontal patches in its explanations. In fact, the explanations contain positive and negative attributions next to each other, and have very poor human interpretability.

V. Results from Stochastic Neural Networks

A. Stochastic ResNet Model

The stochastic ResNet model [9] for URE data was evaluated on a test set comprising 12,960 instances across 18 different classes. The robust model displayed a strong overall performance, achieving an accuracy of 0.94. Precision, recall,

Class	0.1 Standard Deviation			0.25 Standard Deviation			0.5 Standard Deviation		
	Precision	Recall	F1−Score	Precision	Recall	F1−Score	Precision	Recall	F1−Score
1	0.91	1.00	0.95	0.76	0.63	0.69	0.00	0.00	0.00
2	1.00	1.00	1.00	1.00	0.95	0.97	0.00	0.00	0.00
3	1.00	0.93	0.96	0.00	0.00	0.00	0.10	0.04	0.06
Average	0.96	0.95	0.95	0.45	0.47	0.41	0.01	0.04	0.01

TABLE I: ResNet model for URE data is fragile to even Gaussian noise. Per-class and average precision, recall, and F1 scores for different standard deviations. The average is reported over all classes, and our technical report [18] includes all classes.

Class	0.1 Standard Deviation			0.25 Standard Deviation			0.5 Standard Deviation		
	Precision	Recall	F1−Score	Precision	Recall	F1−Score	Precision	Recall	F1−Score
1	0.99	1.00	0.99	0.98	1.00	0.99	0.98	1.00	0.99
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99
Average	0.94	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93

TABLE II: Stochastic ResNet model is robust to Gaussian noise. Per-class and average precision, recall, and F1 scores for different standard deviations are shown here. The average is reported over all classes, and our report [18] includes all classes.

Class	Precision	Recall	F1-score
1	0.99	1.00	0.99
2	1.00	1.00	1.00
3	1.00	1.00	1.00
	Accuracy	Macro Avg	Weighted Avg
	0.94	0.94	0.94

TABLE III: Test performance of the robust stochastic model. The average is reported over all classes, and our accompanying technical report [18] will include all classes.

and F1-score for each class were also examined individually. These results indicate that our model is highly capable of discerning the majority of classes with high accuracy, but efforts could be directed at enhancing its performance on the remaining classes for a more uniformly robust model.

B. Explanations from the Stochastic Model

We created integrated gradients with noise tunnel explanations [4], [5] for the stochastic model [9], as shown in Fig. 2. The attributions are now completely different from the earlier attributions, and we can now see time-invariant or horizontal patches in the attributions that tell us the frequencies that the model used to classify that input. Green horizontal strips denote the presence of frequencies or the absence of frequencies that caused the model to classify the input into that class. Red horizontal strips denote the presence or absence of frequencies that were counteracting against this classification.

We computed IG without noise tunnel, IG with noise tunnel, GradCAM, GradSHAP and occlusion-based attributions for this method to highlight the relative efficacy of the axiomatic integrated gradients approach with noise tunnels compared to other approaches.

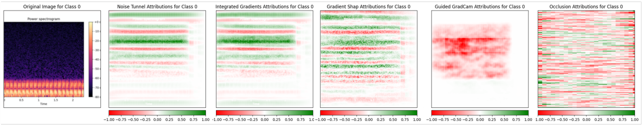

Fig. 3: Attributions of the stochastic model from IG, IG+NT, GradSHAP, GradCAM and Occlusion methods (left to right).

C. Robustness Analysis

The results for the robustness analysis of our robust stochastic model in Table II can be contrasted with the results obtained for the standard ResNet model, as shown in Table I. The standard ResNet model exhibits high performance in terms of precision, recall, and F1-score in the absence of noise. However, this performance drastically degrades as the noise level increases, evident by the average F1-score drops to 0.008, which indicates that the model is almost ineffective in the presence of modest noise levels. In contrast, the robust model demonstrated notable resistance to noise introduction, maintaining an excellent average F1-score of 0.93 even at a noise standard deviation of 0.5. Despite the noisy conditions introduced in the input data, the stochastic model displays a level of performance that was notably higher, emphasizing the utility of our approach for real-world applications where data noise is bound to be prevalent.

VI. Conclusions

Our investigations have provided hitherto unknown and hopefully valuable insights into the limitations of ResNet-like models when used for Unintended Radiated Emission (URE) detection, particularly their susceptibility to Gaussian noise and the inability of their explanations to capture the inherent inductive bias in the data from a stable device. Our findings underscore the need for more robust and interpretable machine learning models in URE detection.

We have demonstrated that Neural SDEs offer a promising alternative. Not only do stochastic models exhibit remarkable resilience to noise, maintaining high performance even in high-noise scenarios, but they also generate meaningful explanations that capture the inherent inductive biases of the data. These features make Neural SDE models and their discrete stochastic variants an interesting tool for URE classification, thereby creating new opportunities for exciting research in this domain.

We identify several opportunities for future research based on our prior and ongoing work. The Unintended Radiated Emission (URE) classification problem can benefit from the design of hardware solutions with desirable size, weight and power characteristics, such as those based on in-memory computing [19], [20] and automated synthesis [21]. Another direction to pursue is to quantify the confidence of the response of the neural network for each instance of URE classifications, using a variety of confidence metrics [22], [23]. Neural stochastic differential equations [24], [25] where the noise has been shaped in conformance with URE data may produce better accuracy and more interpretable results. Analyzing the adversarial robustness [26]–[29] of robust neural networks trained on URE data remains an open and interesting problem.

References

[1] T. Karnowski, R. Kerekes, C. Cooke, M. Vann, M. Adams, and P. Bingham, "Flaming moe," 9 2021.

[2] J. M. Vann, T. P. Karnowski, R. Kerekes, C. D. Cooke, and A. L. Anderson, "A dimensionally aligned signal projection for classification of unintended radiated emissions," *IEEE transactions on electromagnetic compatibility*, vol. 60, no. 1, pp. 122–131, 2017.

[3] T. Grimes, E. Church, W. Pitts, and L. Wood, "Explanation of unintended radiated emission classification via lime," *arXiv preprint arXiv:2009.02418*, 2020.

[4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*. JMLR. org, 2017, pp. 3319–3328.

[5] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[6] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017, pp. 4765–4774.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *CVPR*, 2017, pp. 618–626.

[9] S. Jha, R. Ewetz, A. Velasquez, and S. Jha, "On smoother attributions using neural stochastic differential equations," in *30th International Joint Conference on Artificial Intelligence (IJCAI), 2021*, 2021.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] B. Chang, L. Meng, E. Haber, F. Tung, and D. Begert, "Multi-level residual networks from dynamical systems view," *arXiv preprint arXiv:1710.10348*, 2017.

[12] X. Liu, T. Xiao, S. Si, Q. Cao, S. Kumar, and C.-J. Hsieh, "How does noise help robustness? explanation and exploration under the neural sde framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 282–290.

[13] L. Hodgkinson, C. van der Heide, F. Roosta, and M. W. Mahoney, "Stochastic normalizing flows," *arXiv preprint arXiv:2002.09547*, 2020.

[14] Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration," in *CVPR*, 2015, pp. 5261–5269.

[15] S. Sonoda and N. Murata, "Double continuum limit of deep neural networks," in *ICML Workshop Principled Approaches to Deep Learning*, vol. 1740, 2017.

[16] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *NeurIPS*, 2018, pp. 6571–6583.

[17] Y. Lu, A. Zhong, Q. Li, and B. Dong, "Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3276–3285.

[18] S. K. Jha, S. Jha, R. Ewetz, and A. Velasquez, "Neural stochastic differential equations for robust and explainable analysis of electromagnetic unintended radiated emissions," 2023. [Online]. Available: https://arxiv.org/abs/2309.15386

[19] A. U. Hassen, D. Chakraborty, and S. K. Jha, "Free binary decision diagram-based synthesis of compact crossbars for in-memory computing," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 5, pp. 622–626, 2018.

[20] J. S. Pannu *et al.*, "Design and fabrication of flow-based edge detection memristor crossbar circuits," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 5, pp. 961–965, 2020.

[21] S. K. Jha, *Towards automated system synthesis using sciduction*. University of California, Berkeley, 2011.

[22] S. Jha *et al.*, "Attribution-based confidence metric for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] R. Kaur *et al.*, "idecode: In-distribution equivariance for conformal out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7104–7114.

[24] S. K. Jha, R. Ewetz, A. Velasquez, A. Ramanathan, and S. Jha, "Shaping noise for robust attributions in neural stochastic differential equations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 9, 2022, pp. 9567–9574.

[25] S. Jha, A. Velasquez, R. Ewetz, L. Pullum, and S. Jha, "Explainit!: A tool for computing robust attributions of dnns," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 5916–5919, demo Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/853

[26] S. Jha *et al.*, "Attribution-driven causal analysis for detection of adversarial examples," *arXiv preprint arXiv:1903.05821*, 2019.

[27] S. K. Jha, A. Ramanathan, R. Ewetz, A. Velasquez, and S. Jha, "Protein folding neural networks are not robust," *arXiv preprint arXiv:2109.04460*, 2021.

[28] S. Jha, U. Jang, S. Jha, and B. Jalaian, "Detecting adversarial examples using data manifolds," in *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*. IEEE, 2018, pp. 547–552.

[29] S. Jha, J. Rushby, and N. Shankar, "Model-centered assurance for autonomous systems," in *Computer Safety, Reliability, and Security: 39th International Conference, SAFECOMP 2020, Lisbon, Portugal, September 16–18, 2020, Proceedings 39*. Springer, 2020, pp. 228–243.