

# IMPROVING ROBUSTNESS OF SPECTROGRAM CLASSIFIERS WITH NEURAL STOCHASTIC DIFFERENTIAL EQUATIONS

Joel Brogan, Olivera Kotevska, Anibely Torres, Mark Adams, Sumit Kumar Jha

Oak Ridge National Laboratory, Oak Ridge, TN, USA  
Florida International University, Miami, FL, USA

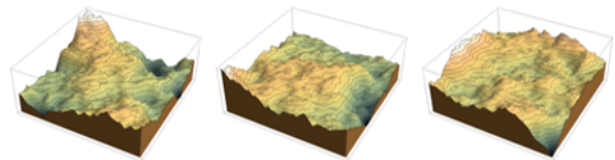
## ABSTRACT

Signal analysis and classification is fraught with high levels of noise and perturbation. Computer-vision-based deep learning models applied to spectrograms have proven useful in the field of signal classification and detection; however, these methods aren't designed to handle the low signal-to-noise ratios inherent within non-vision signal processing tasks. While they are powerful, they are currently not the method of choice in the inherently noisy and dynamic critical infrastructure domain, such as smart-grid sensing, anomaly detection, and non-intrusive load monitoring. Currently, these models can be brittle, which makes them susceptible to noisy input. This also means they have sub-optimal stability of explanation outputs. Experts and technicians using these models to make decisions in real world scenarios need assurance that a model is performing as it is supposed to. The classification or prediction outputs it generates should be sound and grounded, not likely to change in the presence of shifting noise landscapes. In this work, we explore the idea of Neural Stochastic Differential Equations (NSDE's) to improve the robustness of models trained to classify time series data and the effect of NSDE's on the explainability of outputs. We then test the effectiveness of these approaches by applying them to a non-intrusive load monitoring (NILM) dataset that consists of simulated harmonic signals injected into a real building.

**Index Terms**— signal classification, XAI, robustness

## 1. INTRODUCTION

Applications of 2D deep learning methods towards efforts of signal processing and classification have been challenged by the need for more availability of sufficiently diverse data sets [1,2]. This results in models, such as Convolutional Neural Networks (CNNs), overfitting to extraneous features of the test environment not relevant to the task at hand, learning to make "correct" classifications for the wrong reasons. The design of automated AI-based data-driven pipelines for detecting nuanced signal types and characteristics would greatly benefit from the development of algorithms to measure the



**Fig. 1:** Surface representations of the 2D Brownian surface noise injected into our Neural SDE

confidence of neural networks in their responses. Such confidence metrics will enable human Subject Matter Experts to build a relationship of trust with robust neural networks that have a history of credible and correctly calibrated responses. In many application domains, such as load characterization, current deep learning techniques do not provide this capability in any meaningful manner [3,4]. Furthermore, many types of signal processing domains, such as jamming detection [5], speech emotion recognition [6], and radar-based classification [7], are plagued with high levels of real-world noise, either background or adversarial. We must develop techniques to combat scenarios with low Signal-to-noise (SNR) ratios.

This work aims to provide model training and inference methods that improve robustness to noisy spectrogram inputs, and bolster stability and confidence of subsequent classification explanation maps. The general goal is to train more robust and readily explainable neural networks by injecting appropriately shaped noise during their training. If the deep network is explainable, such that a human can get an understanding of the key features of the data that help classify an object, then SMEs can use this information to periodically verify that the network is reasoning soundly and not focusing on features specific to where the training data was collected. We formulate this noise-aware training as a Neural Stochastic Differential Equation (Neural SDE) that provides useful mathematical properties when operating in noisy domains. We test our methods on a custom-built dataset of electromagnetic waveforms injected into a building's wiring and re-collected from multiple sensors. In this preliminary work, we contribute the following:

- A new methodology for training spectrogram-domain CNNs to be robust and stable using domain-shaped

<sup>†</sup>These author Equal Author Contribution

noise as implemented in [8]

- Preliminary experiments on Convnext and Resnet model efficacy against a non-intrusive load monitoring (NILM) dataset made from injecting signal waveforms into a building’s electrical infrastructure
- Preliminary experiments on model efficacy against adversarial signal perturbations

Our results show that while modern vision-based models (specifically the ConvNeXt-Base model [9]) can perform well on time-series classification, they are highly susceptible to increasing levels of noise and small perturbations, while our NSDE-Convnext-variant provides competitive classification performance while incurring less performance drop as the noise floor increases.

Our work builds on prior work on the dynamical systems models of DNNs, such as neural ODEs and neural SDEs, and stochastic NSDEs which have been investigated over the last few years [10–13].

Chapter 2.3 presents the related work. Chapter 3.3 explains the models and methods we used for the experiments and analysis. Chapter 4 describes the dataset used for experiments, processing workflow, and results. Chapter 5 concludes the findings and presents future work.

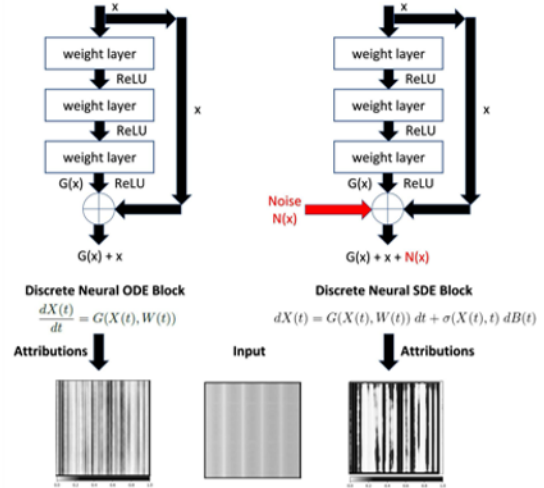
## 2. RELATED WORK

We organized the related work into three parts: Explainability for smart grid and spectrogram images (Sec. 2.1), SDEs and NSDEs (Sec. 2.2, and noise shaping (Sec. 2.3).

### 2.1. Model Explainability

In [14], the authors used the LIME (local interpretable model-agnostic explanations) tool in their classifier and succeeded in determining the frequency bands used by the classifier to make decisions about unintended radiated emission during electronic devices. Others [15] used LIME to identify the critical time-frequency bands influencing the prediction of average surface roughness in a smart grinding process. LIME was used to understand the model that controlled the HVAC system [16]. However, this work did not use spectrograms as input data. Some authors [17] used spectrograms to analyze the magnetohydrodynamic behavior of fusion plasmas and CAM (class activation mapping) explainability tool.

Similarly, in the smart grid domain, there is a work where explainability tools were used to understand better how the model makes the decision using spectrograms as input data and what the key factors that impact those decisions [18]. Others [19] used Grad-CAM (gradient-weighted class activation mapping) and partial dependence plot to understand the feature’s impact on fault zone prediction in smart grids. Similarly, [20] developed a method based on Grad-CAM that



**Fig. 2:** A general overview of how shaped stochastic noise is utilized in our ConvNext architecture to produce more robust explanation attributions.

can highlight the critical regions in the spectrogram that can explain the fault type and location of the smart electric grid and provide a textual description for the event. SHAP (Shapley additive explanations) technique was used to explain the identification of faults in grid-connected photovoltaics [21].

### 2.2. SDEs and Neural SDEs

A notable study by et al. [22] introduced Neural SDE networks, which incorporate random noise injection for regularization, enhancing the stability of Neural Ordinary Differential Equation (ODE) networks. Moreover, [23] explored smoother attributions using Neural SDEs, emphasizing reduced noise, sharper visual outcomes, and enhanced robustness of attributions computed through these models. This study highlights the benefits of employing Neural SDEs for improved interpretability and reliability in neural network applications.

### 2.3. Noise Shaping

The shape of injected noise in an NSDE is an important factor to consider. In [24], authors investigated the impact of noise on stationary pulse solutions in spatially extended neural fields, emphasizing the importance of noise shaping in neural field models. The authors of [8] show that particular types of noise, such as Brownian motion, result in smoother attributions and more stable explanations than traditional resnets.

## 3. METHODS

This section outlines the methods and implementations with which we obtain our results. Generally, we implement our

NSDE as a ConvNeXT architecture [9] and a ResNet [25] architectures. Because these architectures contain a skip connection, they can be modeled as a Neural Differential Equation [22]. To turn a Neural DE into a Neural SDE, we must simply add shaped noise to the concatenation layer of the residual block and its skip connection. Figure 2 shows how a stochastic variant of a ResNet model compares to its non-stochastic equivalent.

### 3.1. Model Types

Our experiments utilize a single main model - the ConvNeXt model [9], which is itself a variant of the ResNet50 model. Below we provide a short description of these model architectures and their relevance. *Residual Networks (ResNet)* were created to solve the challenge of exploiting gradients. So, the skip connections technique was developed that connects activations of a layer to further layers by skipping some layers in between. This forms a residual block. Resnets are made by stacking these residual blocks together. The approach behind this network is that instead of layers learning the underlying mapping, we allow the network to fit the residual mapping. So, instead of say  $H(x)$ , initial mapping, let the network fit,  $F(x) := H(x) - x$  which gives  $H(x) := F(x) + x$ . The advantage of adding this type of skip connection is that if any layer hurts the performance of architecture, then it will be skipped by regularization.

A *ResNet-50 model*, is a 50-layer Convolutional Neural Network (CNN). The difference between ResNet50 and the previously mentioned ResNet (ResNet34) is that the building block was modified into a bottleneck design due to concerns over the time to train the layers. This used a stack of 3 layers instead of the earlier 2. Therefore, each of the 2-layer blocks in Resnet34 was replaced with a 3-layer bottleneck block, forming the Resnet 50 architecture. This results in much higher accuracy than the 34-layer ResNet model.

*ConvNeXt* is improved version of ResNet50 model. At first, a visual transformer was integrated into the model, adjusted the number of blocks at each stage, and increased the kernel size so that the sliding window did not overlap. Other changes are in the activation function, normalization task, and fewer normalization layers. ConvNeXts are good for solving general-purpose computer vision tasks, i.e., image segmentation and object detection.

### 3.2. Noise Shaping

We perform noise shaping, generation, and injection as per the implementation outlined in [8]. Figure 1 shows examples of the type of brownian-shaped noise used as injection input for our Neural SDE training approach.

### 3.3. Explanation Generation

For the explainability analysis, the Captum library was used [26]. It can be applied to any neural network model. Captum

supports three types of attributions: primary, layer, and neuron. Primary attribution evaluates each input feature’s contribution to a model’s output. Layer attribution evaluates how a particular layer impacts the output of the model. Neuron attribution evaluates each input feature’s contribution to activating a particular hidden neuron. Neuron attribution is excellent when combined with layer attribution methods because it can first inspect all the neurons in the layer. Also, a neuron attribution technique can be used to understand what a particular neuron is doing. Each category has a set of functions that provide inside information for the impact of the feature, layer, and neurons on the output. We selected Integrated Gradients (IG) [27] and NoiseTunnel [28] attributions to understand the contribution of each input feature better.

*Integrated Gradients (IG)* is a technique that aims to explain the relationship between model predictions in terms of their features by highlighting them. This is done by computing the gradients of the model’s prediction output to its input features (see Equation 1). There is no need for any modification to the original deep neural network, and it can be applied to images, text, or structured data. In Equation 1  $m$  defines a number of interpolation steps,  $\partial$  is a variance of the current image,  $F$  is a function representing our model,  $x$  is an input, and  $x'$  is the baseline.

$$IG_i^{apx}(x) = (x_i - x'_i) \sum_{k=1}^m \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m} \quad (1)$$

*NoiseTunnel* is a technique that improves the accuracy of attribution methods. It adds Gaussian noise  $N(0, 0.01^2)$  to each input and applies the given attribution algorithm to each sample.

$$\hat{M}_c(x) = \frac{1}{n} \sum_1^n M_c(x + N(0, \sigma^2)) \quad (2)$$

*Saliency maps* is another technique that highlight the most relevant regions or features within an image in a given model. It computes the gradient of the model’s output score with respect to the input features.. The magnitude of these gradients indicates how much the output would change if the input features were modified slightly. Common methods are absolute gradient values, gradient normalization, and relevance masking. In our case absolute gradient technique was used.

*Attribution-based Confidence (ABC)* metric [29] serves as a quantitative measure to assess the reliability of deep neural network (DNN) outputs on input data. This innovative method introduces an approach to estimate DNN prediction confidence utilizing attribution techniques, such as IG, to ascertain the confidence level associated with DNN predictions. Here’s how these components work together in detail Given an input sample  $x$  to the DNN, IG is applied to compute the attribution map  $A(x)$ . This attribution map indicates the importance of each input feature in influencing the DNN’s pre-



diction for the input  $x$ . IG achieves this by calculating the integral of gradients of the model’s output with respect to its input along a straight path from a baseline to the input of interest, providing a comprehensive understanding of feature importance across the input space. Subsequently, the ABC metric calculates a confidence score  $C(x)$  based on the attribution map  $A(x)$  and the predicted class label  $f(x)$ . The confidence score  $C(x)$  is the ratio of the sum of attribution scores associated with features relevant to the predicted class to the sum of all attribution scores across all features.

Mathematically, this can be represented as: Given an input  $x$  for a model  $F$  where  $F_i$  denotes the  $i$ -th logit output of the model, we can compute attribution of feature  $x_j$  of  $x$  for label  $i$  as  $A_j^i$ . We can compute ABC metric in two steps:

- 1) Select feature  $x$  with probability  $\frac{|A_j^i(x)/x_j|}{\sum_j |A_j^i(x)/x_j|}$  and flip the label away from  $i$ , that is, change the decision of the model;
- 2) Calculate the proportion of samples within the neighborhood where the model’s decision remains consistent with the original prediction. This serves as a conservatively estimated confidence measure.

A higher value of  $C(x)$  indicates stronger agreement between the predicted class label and the salient features identified by Integrated Gradients, suggesting higher confidence in the prediction. This confidence score provides valuable insights into the reliability of the DNN’s predictions, enabling better decision-making based on the model’s outputs.

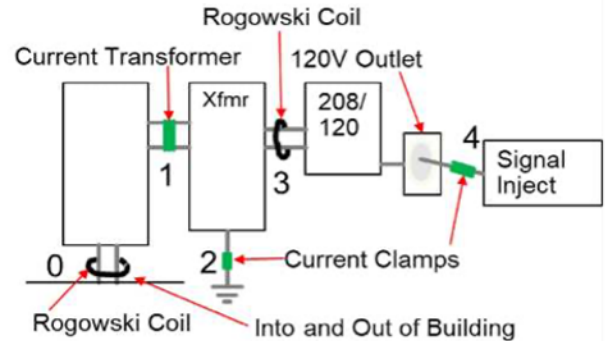
## 4. EXPERIMENTS

In this section, we will outline to dataset collected to train, test, and validate our Nueral-SDE convnext model, along with a set of experiments performed to compare model robustness to noise between the the our Neural-SDE and the original Non-Neural SDE variants of the ConvNext model. We perform three types of evaluation to show the efficacy of our Nueral-SDE method: Accuracy comparison in the presence of random noise, a robustness measure in the presence of adversarial noise, and an overall Attribute Based Confidence (ABC) comparison between the two methodologies.

### 4.1. Data Source

The data, the harmonic signals dataset contains known signals with clean collection at injection and varying noise at other collection locations. Simulated waveforms were injected into ORNL buildings of a user facility ( Figure 3). The waveforms include sine, square, square (75/25 duty), and triangle waves and are injected at different frequencies (0.5-50 kHz). The signals are then subsequently measured with sensors at six locations through building power 3. 16 total classes were collected, including muxed combinations of the pure waveforms, and the Short-Term Fourier transform (STFT) was used to transform the time-series collected data into a

dataset of 10,000 spectrogram samples. This provides 625 examples per class, which mimics the low-data-availability of many signal classification problems.



**Fig. 3:** Data collection locations in power system for injection dataset [30]

### 4.2. Processing and Training

The goal is to train more robust and readily explainable neural networks by injecting appropriately shaped noise during their training. The baseline models used in this work are ResNet, ResNet-50, and ConvNeXt-Base. Figure 2 presents the process overview, denoting how and where noise is injected during training to produce a Nueral-SDE from a generic res-net or ConvNext architecture. From our experimental results, we posit that ResNets with stochastic noise injected into the residual layers of our neural SDEs create more robust attributions. We show that the logarithm of the sum of the change in attributions is smaller for neural SDEs than for neural ODEs. As seen in figure 2 and 4 the integrated gradient attribution with noise tunnel for our neural SDE approach (bottom right) is visually sharper than the IG attribution as well as IG coupled to a noise tunnel for the neural ODE (bottom left) The Data from 4.1 randomly partitioned into 70%-15%-15% training, testing, and validation subsets, respectively.

### 4.3. Results

*Experiment One* evaluates model performance in terms of accuracy in the presence of random noise *and* adversarial noise robustness. We provide accuracy comparisons in Table 1 for the baseline ConvNeXt-Base model and our Neural SDE variant. Note that the accuracy of the Neural SDE model is slightly less than that of the ConvNeXt-Base model. This is due to the regularization incurred by the noise injection, and is an expected tradeoff for robustness.

Model	Accuracy	
	Validation	Test
ConvNeXt-Base	86.9%	87.20%
Neural SDE (ours)	83.2%	82.88%

**Table 1:** Model Accuracy

Table 2 compares how well each model holds up to Gaussian noise. We injected shaped noise from intensities of 0.05 to 1.0. The Neural SDE model’s higher accuracy rates suggest greater robustness to noisy environments than the baseline model.

Noise	Accuracy	
	ConvNeXt-Base	Neural SDE (ours)
0.05	84.38%	87.50%
0.1	71.88%	75.00%
0.15	56.25%	62.50%
0.2	56.25%	53.12%
0.25	25.00%	40.62%
0.5	12.50%	21.88%
0.75	3.12%	9.38%
1.0	6.25%	6.25%

**Table 2:** Robust Accuracy to Random Noise

*Experiment Two* compares model robustness to simulated adversarial noise injection attacks by using the APGD-CE (Auto-Projected Gradient Descent-Cross Entropy). We vary levels of L2 the L2 parameter on within APGD-CE and show results in Table 3. As can be seen, the baseline model has no robustness to this attack, while our Neural SDE still maintains some semblance of non-random classification power.

L2 Norm	Accuracy	
	ConvNeXt-Base	Neural SDE (ours)
0.05	0%	21.88%
0.1	0%	3.12%
0.15	0%	0%
0.2	0%	0%

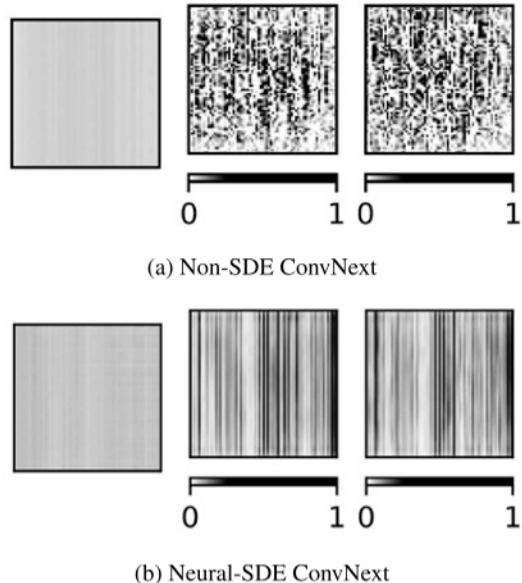
**Table 3:** Model Robustness to Adversarial Noise (APGD-CE)

*Experiment Three* evaluates model confidence using the the ABC metric. Results are presented in Table 4 for ConvNeXt-Base and Neural SDE models. We noticed from the results that ABC is higher for predicting the correct results for Neural SDE model then ConvNeXt, which implies higher confidence.

ConvNeXt-Base		Neural SDE (ours)	
Correct	Incorrect	Correct	Incorrect
0.497	0.188	0.599	0.189

**Table 4:** ABC metric for correct and incorrect results

We also evaluate explainability qualitatively, using the IG and NoiseTunnel attributions as implemented in the Captum library. A comparison of results for the 2D grayscale spectrograms explanations can be seen in Figure 4. Here, we examine explainability for the 2D models trained, validated,



**Fig. 4:** A comparison of Non-Stochastic (top) and Stochastic (bottom) variants of ConvNext spectrogram explanations. A spectrogram input (left) into the Stochastic Neural SDE variant provides more coherent frequency attribution bands for both IG (center) and NoiseTunnel (right) explanations

and tested on the complete set of 2D spectrograms. In Figures 4a and 4b, the further left image is the original image fed into the model, the one next to it is the output explanation from IG, and the right-most image is the explanation from noise tunnel. We can clearly notice from the results that the three methods were able to identify important vertical signal components presented in the original image when applied to the output of our Neural-SDE ConvNext variant, while the baseline contains noisy attributions.

## 5. CONCLUSION

From our experiments, we show that a Neural-SDE variant of the ConvNext Res-Net architecture can provide improved robustness and attribution stability for spectrogram classification in the presence of signal noise, with minimal amounts of trade-off in accuracy. This relatively minor modification to the res-net architecture can provide benefits even in low-data-availability scenarios, as shown by using our relatively small dataset for training and testing. Additionally, with Neural-SDEs, IG and Noise tunneling can successfully identify the important signal features in STFT spectrogram images. Although these preliminary results are promising, accuracy rates in the presence of noise and adversarial input are still too low to be reliable. We hope to improve the efficacy of the NSDE method by investigating improved noise shaping and training augmentation schemes that could further bolster performance in low-data-availability or few-shot-learning settings.

# Acknowledgement

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

## 6. REFERENCES

- [1] Sarala Padi, Seyed Omid Sadjadi, Ram D Sriram, and Dinesh Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proceedings of the 2021 international conference on multimodal interaction*, 2021, pp. 645–652.
- [2] Dylan Anderson, Scott Stewart, Mark Adams, Jack Dermigny, Nathan Martindale, Kasimir Gabert, Boian Alexandrov, Lakshman Prasad, Joel Brogan, Zachary Brown, et al., "Special session on cutting edge approaches in data analytics for nonproliferation.," Tech. Rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2022.
- [3] Francis J Alexander, Tammie Borders, Angie Sheffield, and Marc Wonders, "Workshop report for next-gen ai for proliferation detection: Accelerating the development and use of explainability methods to design ai systems suitable for nonproliferation mission applications," Tech. Rep., Brookhaven National Lab.(BNL), Upton, NY (United States); Idaho National Lab . . . , 2020.
- [4] Dylan Anderson, Scott Stewart, Alexei Skurikhin, Karl Pazdernik, Joel Brogan, and Nathan Martindale, "Testing and evaluation of data analytic approaches for nonproliferation," Tech. Rep., Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2023.
- [5] Y. Li, Jered Pawlak, J. Price, Khair Al Shamaileh, Quamar Niyaz, Paheding Sidike, and Vijay Devabhaktuni, "Jamming detection and classification in ofdm-based uavs via feature- and spectrogram-tailored machine learning," *Ieee Access*, 2022.
- [6] Ammar Amjad, Lal Khan, Noman Ashraf, Muhammad Bilal Mahmood, and Hsien-Tsung Chang, "Recognizing semi-natural and spontaneous speech emotions using deep neural networks," *Ieee Access*, 2022.
- [7] Dongsuk Park, Seungeui Lee, SeongUk Park, and Nojun Kwak, "Radar-spectrogram-based uav classification using convolutional neural networks," *Sensors*, vol. 21, no. 1, pp. 210, 2020.
- [8] Sumit Kumar Jha, Rickard Ewetz, Alvaro Velasquez, Arvind Ramanathan, and Susmit Jha, "Shaping noise for robust attributions in neural stochastic differential equations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 9567–9574.
- [9] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16133–16142.
- [10] Yunjin Chen, Wei Yu, and Thomas Pock, "On learning optimized reaction diffusion processes for effective image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5261–5269.
- [11] Ee Weinan, "A proposal on machine learning via dynamical systems," *Communications in Mathematics and Statistics*, vol. 1, no. 5, pp. 1–11, 2017.
- [12] Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong, "Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3276–3285.
- [13] Bao Wang, Zuoqiang Shi, and Stanley Osher, "Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [14] Tom Grimes, Eric Church, William Pitts, and Lynn Wood, "Explanation of unintended radiated emission classification via lime," *arXiv preprint arXiv:2009.02418*, 2020.
- [15] Abhishek Hanchate, Satish TS Bukkapatnam, Kye Hwan Lee, Anil Srivastava, and Soundar Kumara, "Explainable ai (xai)-driven vibration sensing scheme for surface quality monitoring in a smart surface grinding process," *Journal of Manufacturing Processes*, vol. 99, pp. 184–194, 2023.
- [16] Olivera Kotevska, Jeffrey Munk, Kuldeep Kurte, Yan Du, Kadir Amasyali, Robert W Smith, and Helia Zandi, "Methodology for interpretable reinforcement learning model for hvac energy control," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1555–1564.
- [17] Diogo R Ferreira, Tiago A Martins, Paulo Rodrigues, and JET Contributors, "Explainable deep learning for the analysis of mhd spectrograms in nuclear fusion," *Machine Learning: Science and Technology*, vol. 3, no. 1, pp. 015015, 2021.
- [18] Chongchong Xu, Zhicheng Liao, Chaojie Li, Xiaojun Zhou, and Renyou Xie, "Review on interpretable machine learning in smart grid," *Energies*, vol. 15, no. 12, pp. 4427, 2022.
- [19] Fatemeh Nazary, Carmelo Ardito, Eugenio Di Sciascio, Eng Gianluca Sapienza, and Mario Carpentieri, "Trustworthy machine learning in smart grids," .
- [20] Carmelo Ardito, Yashar Deldjoo, Eugenio Di Sciascio, Fatemeh Nazary, and Gianluca Sapienza, "Iscada: Towards a framework for interpretable fault prediction in smart electrical grids," in *IFIP Conference on Human-Computer Interaction*. Springer, 2021, pp. 270–274.
- [21] Syed Wali and Irfan Khan, "Explainable signature-based machine learning approach for identification of faults in grid-connected photovoltaic systems," in *2022 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2022, pp. 1–6.
- [22] X. Liu, "Neural sde: stabilizing neural ode networks with stochastic noise," 2019.
- [23] S. Jha, R. Ewetz, A. Velasquez, and S. Jha, "On smoother attributions using neural stochastic differential equations," 2021.
- [24] Zachary P Kilpatrick and Bard Ermentrout, "Wandering bumps in stochastic neural fields," *SIAM Journal on Applied Dynamical Systems*, vol. 12, no. 1, pp. 61–94, 2013.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [26] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al., "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.
- [27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
- [28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.
- [29] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami, "Attribution-based confidence metric for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [30] Sumit Kumar Jha, Susmit Jha, Rickard Ewetz, and Alvaro Velasquez, "Neural SDEs for Robust and Explainable Analysis of Electromagnetic Unintended Radiated Emissions." In MILCOM 2023-2023 IEEE Military Communications Conference (MILCOM), pp. 655-660. IEEE, 2023.