32ND RESEARCH SYMPOSIUM ASNT | EVENTS.

Presented At: ASNT Research Symposium: 25 – 28 June 2024 Pittsburgh, Pennsylvania, USA

Detecting AI-generated fake phased-array ultrasonic images from real ones

Pedram Bazrafshan, Arvin Ebrahimkhanlou

Civil, Architectural, and Environmental Engineering, Drexel University
3141 Chestnut St., Philadelphia, PA, USA
fax 215.895.1363; email ae628@drexel.edu

ABSTRACT

This study explores the cybersecurity implications of artificial intelligence (AI) advancements, focusing on differentiating real from AI-generated fake images in nondestructive evaluation. With the rise of sophisticated cyberattacks that can substitute genuine data with indistinguishable AI-generated fakes, there is a growing need for robust mechanisms to detect such fakes. This research introduces an anti-image forensic attack method to distinguish between genuine and synthetic AI-generated images, employing U-Net architecture to train a Denoising Diffusion Probabilistic Model for image generation and using convolutional neural networks for fake image detection. This approach aims to enhance cybersecurity defenses against AI-generated fakes, ensuring the reliability of data-driven decisions in infrastructure maintenance and monitoring.

Keywords: Nondestructive Evaluation, Structural Health Monitoring, Cybersecurity, Image Forensics, Artificial Intelligence, Denoising Diffusion Probabilistic Model, Classification

INTRODUCTION

This study focuses on the cybersecurity implications brought about by advancements in artificial intelligence (AI), aiming to differentiate real from fake images within the context of nondestructive evaluation (NDE). The emergence of sophisticated cyberattacks that can tamper with data acquisition systems to substitute genuine data with indistinguishable AI-generated fakes poses a significant threat. Concurrently, while generative AI offers solutions to data scarcity, its potential misuse in cyberattacks underscores the critical need for robust fake data detection mechanisms, ensuring the integrity of data-driven decisions in infrastructure maintenance and monitoring.

The paper highlights the recent proliferation of generative AI technologies like Generative Adversarial Networks (GANs) [1–6], Variational Autoencoders (VAEs) [7–10], and Denoising Diffusion Probabilistic Models (DDPMs) [11–14]. These technologies, capable of producing highly realistic synthetic data, are revolutionizing NDE by generating ample synthetic datasets for algorithm training and validation [15, 16]. These generated data can further be used for structural assessment purposes, such as robotic and virtual reality frameworks [17–20], crack quantification of concrete/masonry shear walls [21–25], concrete columns [26–31], concrete fiber-reinforced [32], and welding residual stresses [33]. However, generative AI also introduces vulnerabilities, as cyberattacks employing AI-generated fakes could compromise the reliability of NDE and endanger the safety of the underlying asset. Therefore, there is a growing demand for advanced detection methods to safeguard against such threats.

This research presents an anti-image forensic attack method to discern genuine from synthetic AI-generated images, aiming to strengthen cybersecurity defenses. This paper first employs a U-Net architecture to train a generative DDPM using the dataset images of phased array ultrasonic scans [34]. Using the trained DDPM, fake DDPM-generated images are used to train a fake detection model. By training convolutional neural networks (CNNs) with datasets of real and AI-generated images, the study showcases the feasibility of accurately identifying fake images. This approach not only fills a crucial gap in safeguarding NDE data from cyber threats but also underscores the importance of merging AI advancements with robust security measures to ensure the safe application of generative AI in NDE practices.

METHODOLOGY

In the field of AI for image synthesis, DDPMs stand out for producing highly detailed images that closely mimic their training data, surpassing other generative algorithms like GANs and VAEs in quality and detail. However, their accuracy poses a challenge in distinguishing real images from generated ones, underscoring the need for enhanced analytical tools.

DDPMs operate through a Markov chain process, adding Gaussian noise to data progressively before reversing this process to create new samples. This involves a forward diffusion that makes the data noisy and a reverse mechanism where a neural network iteratively denoises the data. This training process aims to accurately predict and subtract the noise added during diffusion, thereby generating new data samples.

The U-Net architecture plays a crucial role in DDPMs, enabling precise localization and context capture through its unique structure that combines an encoder for context and a decoder for detail reconstruction. This model is particularly effective for the denoising task in DDPMs, due to its capability to blend detailed low-level information with high-level context, making it ideal for generating complex, high-fidelity images.

CNNs excel in distinguishing between real and AI-generated images by identifying intricate patterns within images, a capability crucial for binary classification tasks. In such classifications, CNNs leverage convolutional layers to extract features and employ dense layers and sigmoid functions to classify images as real or generated. This process benefits from CNNs' ability to discern subtle details, making them powerful tools for authenticating images in the face of advanced generative models like DDPMs.

IMPLEMENTATION

This paper uses a U-Net architecture to train the DDPM image generator. For this, a dataset of 19810 images of phased array ultrasonic of adjacent ultrasonic scans along weld lines in a steel structure has been used to train the DDPM. The images are of size 256×256 in grayscale and resized to 64×64 for training and generating purposes. The diffusion process in the DDPM training phase is illustrated in Figure 1.

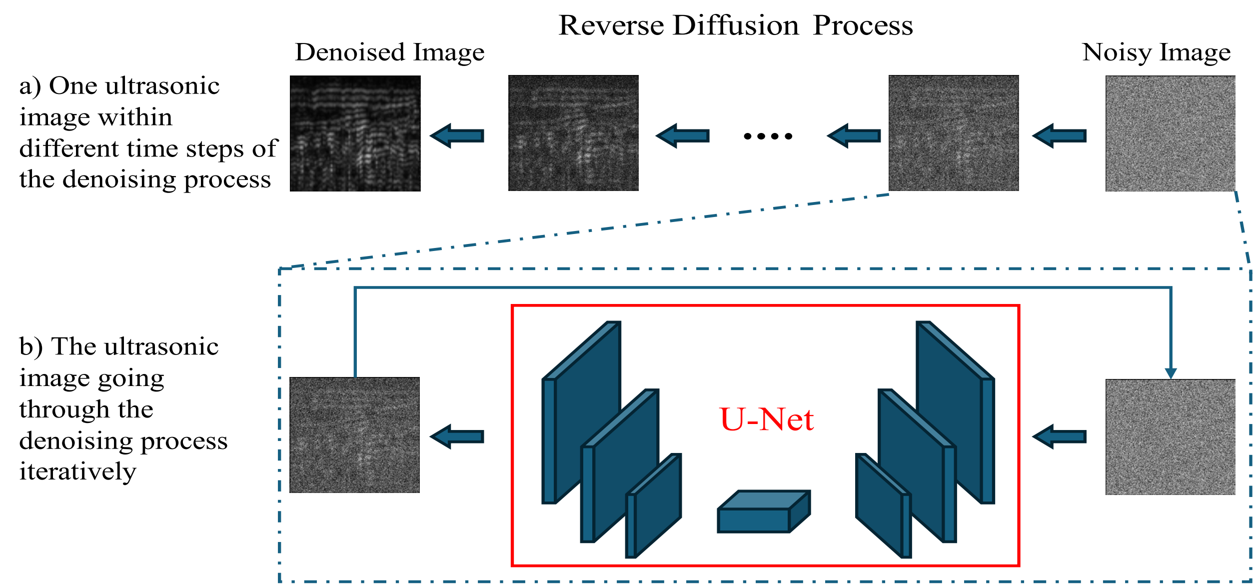

Figure 1. The diffusion process of the DDPM training phase

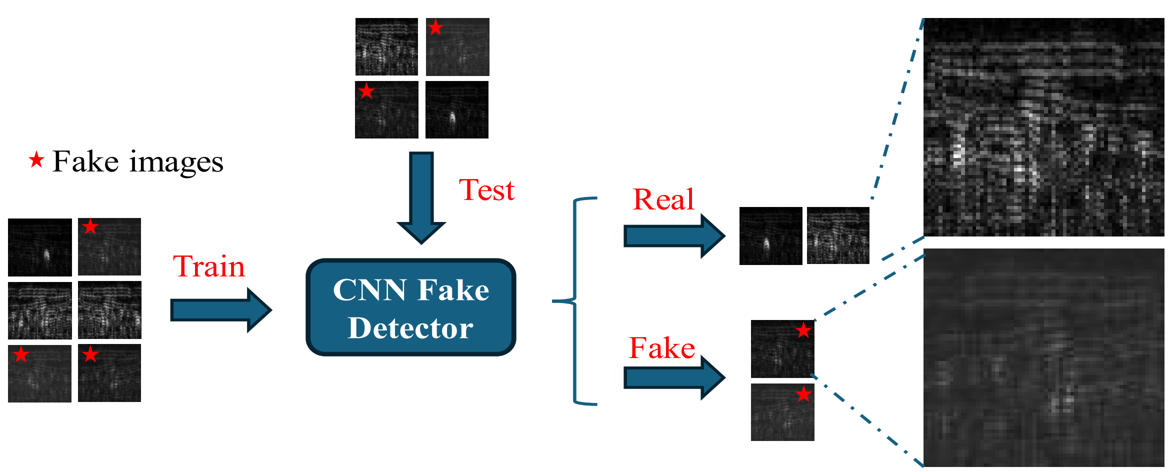

Figure 2. Convolutional neural network fake detector

Afterward, DDPM is used to generate fake ultrasound scan images. Then, the DDPM-generated images, along with the authentic ones, are used to train the fake detector neural networks. The framework for detecting fake images from real ones is presented in Figure 2.

RESULTS AND DISCUSSION

As depicted in Figure 3, the CNN fake detector is capable of detecting fake images from the real ones for both the validation and test sets. The CNN fake detector has no false positives and false negatives. This shows a robust and reliable performance in detecting fake data in the context of NDE.

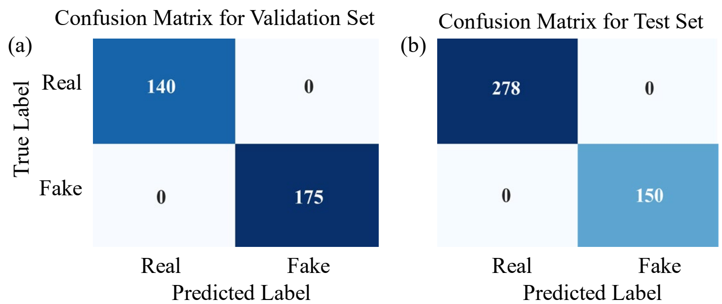

Figure 3. The confusion matrices for the validation and test sets for fake image detection

CONCLUSIONS

The implementation of a U-Net-trained DDPM and CNN-based fake detector demonstrated a robust and reliable method for distinguishing between real and AI-generated images, crucial for NDE. DDPM generated high-quality fake images, and the CNN classifier detected fake images with no false positives and negatives. This research highlights the importance of integrating AI advancements with stringent security measures to safeguard assets from cyber threats, contributing to the safe application of generative AI in NDE practices.

REFERENCES

1. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al (2014) Generative Adversarial Networks. Sci Robot 3:2672–2680
2. Xu H, Liang P, Yu W, et al (2019) Learning a Generative Model for Fusing Infrared and Visible Images via Conditional Generative Adversarial Network with Dual Discriminators. ijcai.orgH Xu, P Liang, W Yu, J Jiang, J MaIJCAI, 2019•ijcai.org
3. McKnight S, Pierce SG, Mohseni E, et al (2024) A comparison of methods for generating synthetic training data for domain adaption of deep learning models in ultrasonic non-destructive evaluation. NDT & E International 141:102978. https://doi.org/10.1016/J.NDTEINT.2023.102978
4. Heesch M, Mendrok K, Dworakowski Z (2021) Time-Domain Signal Synthesis with Style-Based Generative Adversarial Networks Applied to Guided Waves. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12854 LNAI:78–88. https://doi.org/10.1007/978-3-030-87986-0_7/FIGURES/5
5. Posilović L, Medak D, Subašić M, et al (2021) Generative adversarial network with object detector discriminator for enhanced defect detection on ultrasonic B-scans. Neurocomputing 459:361–369. https://doi.org/10.1016/J.NEUCOM.2021.06.094
6. Zeng Y, Li Y, Du P, Huang X (2023) Smart fire detection analysis in complex building floorplans powered by GAN. Journal of Building Engineering 79:107858. https://doi.org/10.1016/J.JOBE.2023.107858
7. Kingma DP, Welling M (2013) Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings
8. Girin L, Leglaive S, Bie X, et al (2020) Dynamical Variational Autoencoders: A Comprehensive Review. Foundations and Trends in Machine Learning 15:1–175. https://doi.org/10.1561/2200000089
9. Zhang Y, Lee YS, Lin H, Chen J (2022) Establishing Convolutional Neural Network Kalman Recurrent Variational Autoencoder Using Infrared Imaging for Process Monitoring: An Application in Spinning Disk Processes. IEEE Trans Instrum Meas 71:. https://doi.org/10.1109/TIM.2021.3126381
10. Pei L, Sun Z, Xiao L, et al (2021) Virtual generation of pavement crack images based on improved deep convolutional generative adversarial network. Eng Appl Artif Intell 104:104376. https://doi.org/10.1016/J.ENGAPPAI.2021.104376
11. Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. Adv Neural Inf Process Syst 2020-December:
12. Tang S, Jin Z, Zhang Y, et al (2023) A Timestep-Adaptive-Diffusion-Model-Oriented Unsupervised Detection Method for Fabric Surface Defects. Processes 2023, Vol 11, Page 2615 11:2615. https://doi.org/10.3390/PR11092615

13. Chen L, Zhou L, Li L, Luo M (2023) CrackDiffusion: crack inpainting with denoising diffusion models and crack segmentation perceptual score. Smart Mater Struct 32:054001. https://doi.org/10.1088/1361-665X/ACC624
14. Jadhav Y, Berthel J, Hu C, et al (2023) StressD: 2D Stress estimation using denoising diffusion model. Comput Methods Appl Mech Eng 416:116343. https://doi.org/10.1016/J.CMA.2023.116343
15. Andrushia D, Anand N, Arulraj P (2020) Anisotropic diffusion based denoising on concrete images and surface crack segmentation. International Journal of Structural Integrity 11:395–409. https://doi.org/10.1108/IJSI-06-2019-0061
16. Cano-Ortiz S, Iglesias LL, Martinez Ruiz del Árbol P, Castro-Fresno D (2024) Improving detection of asphalt distresses with deep learning-based diffusion model for intelligent road maintenance. Developments in the Built Environment 17:100315. https://doi.org/10.1016/J.DIBE.2023.100315
17. Ghadimzadeh Alamdari A, Ebrahimkhanlou A (2024) A multi-scale robotic approach for precise crack measurement in concrete structures. Autom Constr 158:105215. https://doi.org/10.1016/j.autcon.2023.105215
18. Bazrafshan P, Ebrahimkhanlou A (2023) A virtual-reality framework for graph-based damage evaluation of reinforced concrete structures. In: Shull PJ, Yu T, Gyekenyesi AL, Wu HF (eds) Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, Civil Infrastructure, and Transportation XVII. SPIE, p 5
19. Bazrafshan P, Ebrahimkhanlou A (2023) A robotic-based framework for quantifying surface cracks of concrete shear walls. In: Proceedings of the 14th International Workshop on Structural Health Monitoring. Destech Publications, Inc.
20. Khedmatgozar Dolati SS, Caluk N, Mehrabi A, Khedmatgozar Dolati SS (2021) Non-Destructive Testing Applications for Steel Bridges. Applied Sciences 2021, Vol 11, Page 9757 11:9757. https://doi.org/10.3390/APP11209757
21. Bazrafshan P, On T, Ebrahimkhanlou A (2022) Machine learning-based damage detection of RC wall using graph features of crack patterns. In: ASNT 30th Research Symposium Conference Proceedings. The American Society for Nondestructive Testing Inc., pp 1–4
22. Bazrafshan P, On T, Ebrahimkhanlou A (2022) A computer vision-based crack quantification of reinforced concrete shells using graph theory measures. In: Zonta D, Su Z, Glisic B (eds) Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2022. SPIE, p 25
23. Bazrafshan P, On T, Basereh S, et al (2023) A graph-based method for quantifying crack patterns on reinforced concrete shear walls. Computer-Aided Civil and Infrastructure Engineering. https://doi.org/10.1111/MICE.13009
24. Asjodi AH, Dolatshahi KM, Ebrahimkhanlou A (2022) Spatial analysis of damage evolution in cyclic-loaded reinforced concrete shear walls. Journal of Building Engineering 49:104032. https://doi.org/10.1016/J.JOBE.2022.104032
25. Godio M, Flansbjer M, Williams Portal N (2023) Single- and double-wythe brick masonry walls subjected to four-point bending tests under different support conditions: Simply supported, rigid, non-rigid. Constr Build Mater 404:132544. https://doi.org/10.1016/J.CONBUILDMAT.2023.132544
26. Zamani P, Azhari S, Hamidia M, Hassani N (2024) Crack image-based FEMA P-58-compliant fragility models for automated earthquake-induced loss estimation in non-ductile RC moment frames. Structures 60:105873. https://doi.org/10.1016/J.ISTRUC.2024.105873
27. Hamidia M, Ganjizadeh A, Dolatshahi KM (2022) Peak drift ratio estimation for RC moment frames using multifractal dimensions of surface crack patterns. Eng Struct 255:113893. https://doi.org/10.1016/J.ENGSTRUCT.2022.113893
28. Hamidia M, Mansourdehghan S, Asjodi AH, Dolatshahi KM (2022) Machine learning-based seismic damage assessment of non-ductile RC beam-column joints using visual damage indices of surface crack patterns. Structures 45:2038–2050. https://doi.org/10.1016/J.ISTRUC.2022.09.010
29. Hamidia M, Afzali M, Jamshidian S, Safi M (2023) Post-earthquake stiffness loss estimation for reinforced concrete columns using fractal analysis of crack patterns. Structural Concrete 24:3933–3951. https://doi.org/10.1002/SUCO.202200351
30. Sasan Khedmatgozar Dolati S, Matamoros A, Ghannoum W (2023) Evaluating the effects of loading protocol on the strength and deformation capacity of Flexure-Shear critical concrete columns. Eng Struct 279:115592. https://doi.org/10.1016/J.ENGSTRUCT.2023.115592
31. Lee CS, Mangalathu S, Jeon JS (2024) Machine learning–assisted drift capacity prediction models for reinforced concrete columns with shape memory alloy bars. Computer-Aided Civil and Infrastructure Engineering 39:595–616. https://doi.org/10.1111/MICE.13112
32. Mahmoudian A, Tajik N, Taleshi MM, et al (2023) Ensemble machine learning-based approach with genetic algorithm optimization for predicting bond strength and failure mode in concrete-GFRP mat anchorage interface. Structures 57:105173. https://doi.org/10.1016/J.ISTRUC.2023.105173
33. Wang L, Qian X (2024) Optimization-improved thermal–mechanical simulation of welding residual stresses in welded connections. Computer-Aided Civil and Infrastructure Engineering. https://doi.org/10.1111/MICE.13136
34. Virkkunen I, Koskinen T, Jessen-Juhler O, Rinta-aho J (2021) Augmented Ultrasonic Data for Machine Learning. J Nondestr Eval 40:1–11. https://doi.org/10.1007/S10921-020-00739-5/TABLES/1