

Analysis of Dataset for Training Neural Networks Applied to Broadband Patch Antenna Design

Chaoyan Zhang, Wenting Li*, Chaoyun Song, Yejun He

State Key Laboratory of Radio Frequency Heterogeneous Integration

Sino-British Antennas and Propagation Joint Laboratory, Ministry of Science and Technology of China

Guangdong Engineering Research Center of Base Station Antennas and Propagation

Shenzhen Key Laboratory of Antennas and Propagation

College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

Email: cy0714ay@163.com, w.li@szu.edu.cn*, heyeyun@126.com, cysong@szu.edu.cn

Abstract—This study analyzes the characteristics of datasets for training the Feedforward Neural Network (FNN), which uses geometric parameters as inputs, and the Convolutional Neural Network (CNN), which utilizes discrete matrices as inputs for designing the slotted patch antenna. Geometric projections reveal that the dataset of FNN inputs appears as an isolated point array structure constrained by fixed step sizes. In contrast, the dataset of CNN inputs manifests as a continuous radial topology, which reflects the rich spatial diversity of the high-dimensional data in the feature space. Quantitative analysis via the Trustworthiness Score validates the topological consistency of this representation; the CNN maintains high fidelity in the local regime, which confirms its intrinsic capability to encode the subtle physical variations essential for accurate modeling. Statistical results show that the dataset of FNN inputs exhibits a discrete, multimodal distribution constrained by the parameter grid, whereas that of CNN display a smooth, quasi-Gaussian distribution. Furthermore, Principal Component Analysis (PCA) of the output data reveals high spectral redundancy, with the first three principal components explaining over 90% of the variance and the first 20 capturing 99.99%. Consequently, geometric and statistical insights suggest the potential for efficient modeling using a CNN with a reduced 20-point sampling strategy.

Index Terms—slotted patch antenna, Spatial Topology, Geometric Interpretation, Statistical Distribution

I. INTRODUCTION

Wideband antennas are essential for modern wireless systems, yet their design is often hindered by computationally expensive electromagnetic (EM)

simulations [1]. To address this, Deep Learning (DL) surrogate models emerge as efficient alternatives. However, existing approaches diverge in input representation. FNN typically uses geometric parameters, while CNN utilizes discretized matrix representations [2, 3]. Despite their widespread use, a rigorous quantitative comparison regarding the geometric and topological consistency of these two input modalities remains underexplored.

This paper analyzes the statistical and manifold properties of the data representation spaces for both FNN and CNN using a slotted patch antenna case study. Through PCA visualization [4, 5] and t-Distributed Stochastic Neighbor Embedding (t-SNE), we reveal fundamental differences in how these models represent design space. While the dataset of FNN inputs exhibits a grid-limited distribution, that of CNN demonstrates superior capability in preserving physically interpretable 2D structures. Complementing this input-space analysis, we further investigate the output domain and demonstrate that the antenna's wideband response contains inherent low-rank characteristics. By identifying significant spectral redundancy in the antenna's output response—where the dominant principal components capture nearly all variance—this study explores the potential for a “20-point sampling strategy”. This hypothetical approach suggests that full-wave simulation requirements could be reduced from 41 to 20 frequency points, offering a pathway to significantly lower computational costs while maintaining modeling accuracy.

II. ANTENNA DESIGN AND DATASET DESCRIPTION

We adopt the antenna structure proposed in [6] for this study. Fig. 1 illustrates that the system processes the dataset into two distinct input formats. The input representation for the FNN is represented as a parameterized geometric vector $\mathbf{B}_i = [L_1, L_2, Y]$, which corresponds directly to the physical parameters of the slot. In contrast, for the CNN, the antenna topology is

discretized into a binary matrix A_i , where elements $a_{mn} = 1$ denote the slotted region and “0” represents the metallic patch. Specifically, we sample the geometrical variables in millimeters such that the slot length L_1 ranges from 1.3 to 19.5 mm with a step of 2.6 mm, the slot width L_2 varies from 1.3 to 6.5 mm with a step of 1.3 mm, and the position Y spans from 1.95 to 11.05 mm with a step of 1.3 mm.

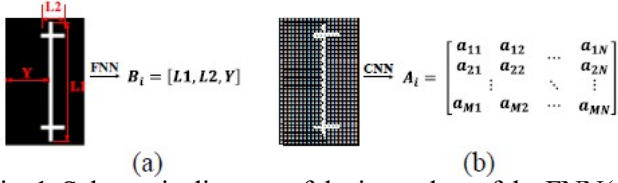


Fig. 1. Schematic diagram of the input data of the FNN(a) and CNN(b).

The CNN is designed to learn the mapping between the geometric features and the electromagnetic (EM) response. The patch antenna structure is discretized into a $1 \times M \times N$ matrix, where $M=32$ and $N=40$. This discrete matrix serves as the input to the CNN, which outputs reflection coefficients across 41 frequency points. Both the FNN and CNN utilize an identical dataset, which comprises 320 samples that are randomly partitioned into 80% for training and 20% for testing.

III. DATASET ANALYSIS

Data processing is implemented in Python and utilizes NumPy and scikit-learn for numerical computations and machine learning, while Matplotlib and Seaborn are employed for visualization. The FNN utilizes the raw 3D parameter vector $(L1, L2, Y)$. For the CNN, the 32×40 matrix input is retained for modeling, yet flattened into a 1280-dimensional vector (in row-major order). Core analysis focuses on Euclidean distances, specifically the distance from each sample to the global mean, which we calculate according to (1),

$$d_i = \| \mathbf{x}_i - \bar{\mathbf{x}} \|_2 \quad (1)$$

where $\mathbf{x}_i$ denotes the flattened vector representation of the i -th sample, $\bar{\mathbf{x}}$ represents the global mean vector computed across the entire dataset, and $\|\cdot\|_2$ signifies the Euclidean norm (L2 norm).

A. Analysis of Inputs

1) Clustering and Distribution

Fig. 2 and Fig. 3 illustrate that PCA is employed to visualize the input distributions. To quantify the fidelity of this low-dimensional projection, we adopt the explained variance ratio. Mathematically, PCA extracts

principal components through eigen-decomposition on the data's covariance matrix. The explained variance ratio for the k -th component is defined in (2) as,

$$\eta_k = \frac{\lambda_k}{\sum_{j=1}^d \lambda_j} \quad (2)$$

where λ_k represents the λ_k -th eigenvalue and d denotes the total dimensionality of the original input data.

Since the magnitude of an eigenvalue is directly proportional to the variance (or structural energy) along its corresponding eigenvector, this ratio strictly quantifies the percentage of the dataset's total information preserved in that dimension. The analysis reveals a fundamental spectral disparity. For the dataset of FNN inputs (Fig. 2(a) and Fig. 3(a)), the spectral energy is highly concentrated in the first two eigenvalues ($\eta_1 = 74.3\%$, $\eta_2 = 18.6\%$, which accumulate to over 90%). This indicates that the FNN data is intrinsically low-rank and resides in a simple linear subspace governed by its rigid parameter grid. In sharp contrast, those of CNN (Fig. 2(b) and Fig. 3(b)) exhibit a phenomenon of “spectral spread”, characterized by significantly lower ratios on the primary axes ($\eta_1 = 19.7\%$, $\eta_2 = 14.0\%$). Such a slowly decaying eigenvalue spectrum provides robust evidence that the CNN representation characterizes a complex, high-dimensional manifold rich in topological diversity, which differs fundamentally from the FNN.

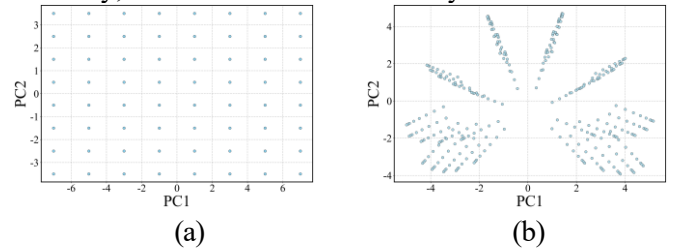


Fig. 2. 2D PCA projection of the input of the FNN(a) and CNN(b).

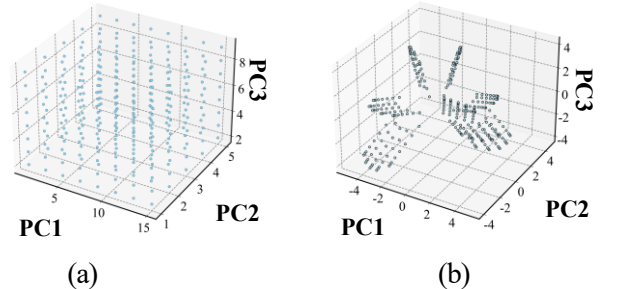


Fig. 3. 3D PCA projection of the input of the FNN(a) and CNN(b).

2) Neighborhood Preservation

Neighborhood preservation ensures that similar high-

dimensional samples remain proximal in the target space and thereby maintains topological consistency. To quantify this property, the Trustworthiness Score $T(k)$ is employed. This metric specifically penalizes the presence of “intruders”—samples that appear as neighbors in the low-dimensional target space but are not neighbors in the original high-dimensional space, as quantified by (3),

$$T(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^n \sum_{j \in U_i^{(k)}} (r(i, j) - k) \quad (3)$$

where n denotes the total number of samples in the dataset, and k represents the specified neighborhood size. The term $r(i, j)$ signifies the rank of sample j relative to sample i based on the Euclidean distance in the original high-dimensional space. The set $U_i^{(k)}$ contains the identified intruders for sample i , defined as the samples that reside within the k -nearest neighborhood in the low-dimensional representation but fall outside this neighborhood in the original space $U_i^{(k)}$.

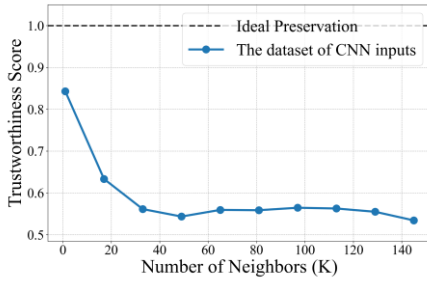


Fig. 4. Neighborhood preservation scores for CNN inputs.

A score approaching 1.0 indicates ideal preservation of the neighborhood structure, whereas a lower score implies topological distortion. As Fig. 4 shows, the dataset of CNN inputs achieves high trustworthiness ($T(k) > 0.6$) within the local neighborhood range ($k < 20$). This result confirms that the discrete matrix representation successfully preserves the most critical local geometric relationships. The score decline as k expands is expected and highlights the high-dimensional complexity of the CNN manifold, which is naturally difficult to flatten into a low-dimensional space without distortion, yet the high fidelity in the immediate vicinity is sufficient to ensure local continuity for regression tasks. The analysis is explicitly restricted to the range $k < N/2$ (specifically $k \leq 160$). This ensures the metric remains mathematically well-defined and avoids trivial convergence to 1.0 as the neighborhood shifts toward global structure, which would diminish its discriminative power regarding local geometric fidelity.

3) Distance Metric

Fig. 5(a) and Fig. 5(b) compare the probability density functions (PDFs) of the normalized distances to the mean center. The dataset of FNN inputs (Fig. 5(a)) exhibits a multimodal, discrete distribution characterized by distinct spikes, which explicitly reveals the artificial discontinuities introduced by the rigid parameter sampling grid. In contrast, the dataset of CNN inputs (Fig. 5(b)) displays a smoother, unimodal distribution which skews towards higher values (right-skewed). This demonstrates that the high-dimensional discrete representation provides a more continuous geometric coverage, which aligns with the statistical properties of high-dimensional manifolds, whereas constraining the dataset of FNN inputs.

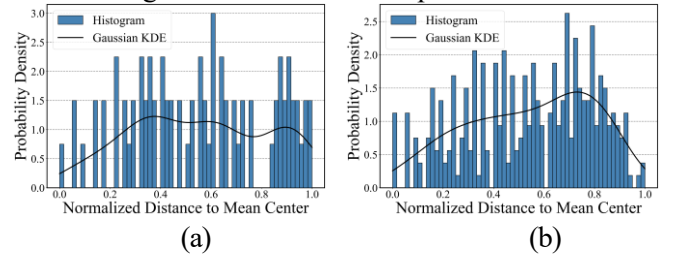


Fig. 5. Distribution of distances to the mean Center of the input of the FNN(a) and CNN(b).

A custom Gaussian Kernel Density Estimation (KDE) [7] is employed to analyze the normalized distance distributions. The density is estimated over 1000 uniformly distributed points within the interval $[0, 1]$ according to (4),

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n \exp\left(-\frac{(x-x_i)^2}{2h^2}\right). \quad (4)$$

where $\hat{f}_h(x)$ represents the estimated probability density function at the query point x . The parameter n denotes the total number of samples in the dataset, while x_i refers to the specific normalized distance value of the i -th sample. The bandwidth h , previously determined via Silverman's rule, serves as the smoothing parameter and is given by (5) controlling the width of the Gaussian kernel. Here, σ denotes the standard deviation of the input data.

$$h = 1.06 \cdot \sigma \cdot n^{-1/5} \quad (5)$$

The resulting smooth KDE curve (which the black line depicts) aligns closely with the histograms.

Fig. 6(a) and Fig. 6(b) further examine the pairwise Euclidean distances across all sample pairs. The histogram for the dataset of FNN inputs shows highly irregular fluctuations, which reflect the rigid distance quantization inherent to the lattice structure. Conversely, the dataset of CNN inputs follows a quasi-Gaussian

distribution, which demonstrates superior statistical regularity. This result implies that the high-dimensional matrix representation effectively breaks the geometric constraints of the parameter space and facilitates greater spatial diversity and more natural topological evolution.

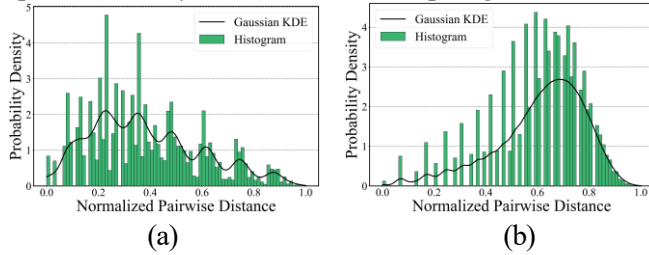


Fig. 6. Pairwise distance distribution of the input data of the FNN(a) and CNN(b).

B. Analysis of Outputs

Fig. 7 presents the PCA on the output dataset, which reveals a high degree of spectral redundancy. Notably, the first three principal components account for 90% of the total variance, and this aligns closely with the three geometric degrees of freedom inherent to the physical inputs. Furthermore, the first ten dimensions encompass 99% of the information content. These results demonstrate that the 41-dimensional electromagnetic response data intrinsically resides on a low-dimensional manifold. Consequently, given that the first 20 principal components capture 99.99% of the variance, the observed spectral smoothness and strong inter-frequency correlations provide a robust theoretical basis for optimizing data acquisition strategies.

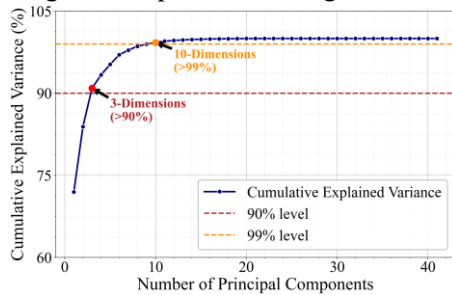


Fig. 7. Output variance analysis.

Based on these findings, we propose reducing the number of frequency sampling points in full-wave simulations from 41 to 20. This approach retains a sufficient sampling safety margin to enable the near-lossless reconstruction of the full spectral curve via interpolation. Moreover, it substantially reduces the computational burden of preliminary data generation, thereby enhancing research and development efficiency without compromising model accuracy.

IV. CONCLUSION

This study confirms that the CNN's discrete matrix representation overcomes the limitations of the FNN's fixed-step discrete point array, offering richer physical inductive biases by preserving 2D topology and statistical regularity. This enables the model to capture subtle spatial transitions within the parameter space that are typically ignored by discontinuous parameter partitioning. Conversely, the dense clustering of the output data in the feature space indicates that the wide-band electromagnetic response is strictly constrained by underlying physical laws rather than being randomly distributed. Based on these insights, the study suggests a '20-point sampling strategy' designed to leverage the CNN's topological fidelity to significantly reduce full-wave simulation costs without compromising modeling accuracy.

ACKNOWLEDGEMENT

This work is supported in part by the National Key Research and Development Program of China under Grant 2023YFE0107900, in part by the National Natural Science Foundation of China (NSFC) under Grants 62101341 and U2541208, in part by the Key Program of Shenzhen Natural Science Foundation under Grant JCYJ20241202124219023, and in part by the Program of Shenzhen Key Laboratory Evaluation under Grant SYSPG20241211173908022.

REFERENCES

- [1] J. Hao, N. Yan, Y. Luo, et al., "A Low-Cost Dual-Band Multimode High-Gain Stacked-Patch Antenna Based on SISL for 5G Applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 21, no. 1, pp. 4-8, 2022.
- [2] T. Khan, A. De, and M. Uddin, "Prediction of Slot-Size and Inserted Air-Gap for Improving the Performance of Rectangular Microstrip Antennas Using Artificial Neural Networks," *IEEE Antennas and Wireless Propagation Letters*, vol. 12, pp. 1367-1371, 2013.
- [3] J. P. Jacobs, "Accurate Modeling by Convolutional Neural-Network Regression of Resonant Frequencies of Dual-Band Pixelated Microstrip Antenna," *IEEE Antennas and Wireless Propagation Letters*, vol. 20, no. 12, pp. 2417-2421, 2021.
- [4] A. Rehman, M. A. Ali, A. Khan, et al., "Performance Analysis of PCA, Sparse PCA, Kernel PCA, and Incremental PCA Algorithms for Heart Failure Prediction," in *Proceedings of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, Istanbul, Turkey, 2020.
- [5] M. T. White and S. Jeon, "Using t-SNE to Explore Misclassification," in *2019 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2019.
- [6] Y. He, J. Huang, W. Li, et al., "Hybrid Method of Artificial Neural Network and Simulated Annealing Algorithm for Optimizing Wideband Patch Antennas," *IEEE Transactions on Antennas and Propagation*, vol. 72, no. 1, pp. 944-949, Jan. 2024.
- [7] J. Heer, "Fast & Accurate Gaussian Kernel Density Estimation," in *2021 IEEE Visualization Conference (VIS)*, New Orleans, LA, USA, 2021.