SCIENCE FORUM (JOURNAL OF PURE AND APPLIED SCIENCES) 26, (3) (2026), 281 - 296

Science Forum (Journal of Pure and Applied Sciences)

Journal Homepage: https://atbuscienceforum.com.ng/index.php/jpas

A Baseline Evaluation of Boosted Tree Regression for Predicting Electromagnetic Interference Shielding Effectiveness

Ibrahim Abubakar Alhaji[1], UmarFaruk Usman[1], and Yahaya Bala Zakariyau[2]

[1]Department of Physics, Federal University of Kashere, P. M. B. 0182, Gombe, Gombe State, Nigeria

[2]Department of Computer Science, Federal University of Kashere, P. M. B. 0182, Gombe, Gombe State, Nigeria

Abstract

Machine learning (ML) has emerged as a promising tool for accelerating the design and optimization of advanced electromagnetic interference (EMI) shielding materials. However, its applicability in experimental materials science is often constrained by limited data availability. This study establishes a performance baseline for ML-based prediction of EMI shielding effectiveness (SE) under data-scarce conditions representative of practical laboratory datasets. An optimized Least Squares Boosting (LSBoost) regression model was developed using four readily accessible input variables: matrix dielectric constant, filler dielectric constant, operating frequency, and filler loading. The model was trained on 100 experimentally derived samples and validated using an independent test set comprising 50 samples intentionally designed to include previously unseen material combinations and unexplored frequency domains. The developed model exhibited excellent interpolation performance within the training distribution, achieving a coefficient of determination (R^2) of 0.931 and a root mean square error (RMSE) of 2.448 dB. However, predictive accuracy deteriorated substantially on the independent test set (R^2 = 0.597; RMSE = 11.746 dB), revealing a pronounced generalization gap when extrapolating beyond the training domain. Feature importance analysis indicated that matrix dielectric constant was the dominant predictor (importance score = 0.704), followed by frequency (0.332), filler loading (0.244), and filler dielectric constant (0.222). Partial dependence analysis revealed complex non-linear feature–response relationships and suggested a prediction saturation threshold near 37.5 dB shielding effectiveness. Principal component analysis confirmed significant distributional divergence between training and testing datasets, providing a mechanistic explanation for the observed generalization limitations. Furthermore, learning curve analysis demonstrated sustained sensitivity to additional training samples beyond 100 observations, indicating persistence of the data-scarce regime. These findings highlight the limitations of purely data-driven approaches and underscore the need for enhanced dataset diversity, domain adaptation strategies, and physics-informed learning frameworks to improve model robustness and reliability in EMI shielding material design.

Keywords
Machine Learning,
Model,
Dielectric Constant, Filler Loading,
Frequency

ISSN 1119-4618
JPAS
Faculty of Science, ATBU Bauchi

Corresponding Author: Ibrahim Abubakar Alhaji, [1]Department of Physics, Federal University of Kashere, P. M. B. 0182, Gombe, Gombe State, Nigeria
Email: alhaji259@gmail.com

1.0 INTRODUCTION

The increasing adoption of electronic devices and the subsequent demand for electromagnetic compatibility (EMC) have made effective electromagnetic interference (EMI) shielding a requirement in modern electronics design and manufacturing (Hota et al., 2025). EMI shielding effectiveness (SE), expressed in decibels (dB), quantifies a material's ability to attenuate electromagnetic radiation. Polymer-based composite materials, consisting of a polymer matrix reinforced with conductive or dielectric fillers, are widely used for these applications due to their tunable electrical properties, corrosion resistance, and processing advantages (Chung, 2020). The SE of such composites is governed by several interdependent factors, including the intrinsic properties of the matrix and filler, filler concentration, and the frequency of the incident electromagnetic wave (Fayzan et al., 2019).

Numerous polymer composites have been fabricated for EMI SE applications, but producing multiple samples with varying compositions and experimentally measuring their shielding performance is time-consuming, costly, and resource-intensive (Hwang et al., 2022; Khamis et al., 2022). Consequently, predictive analytical models based on electromagnetic theory have been developed, including the Simon formalism, which relates shielding to conductivity, and transmission line theory, which treats the material as a barrier to wave propagation (Wu et al., 2016; Yin et al., 2022). While these approaches provide valuable physical insights, they often rely on simplifying assumptions, such as homogeneous filler distribution, negligible interfacial effects, and simplified frequency-dependent behavior, which limit their predictive accuracy for the increasingly diverse range of modern composites (Husnain et al., 2024). These limitations motivate the exploration of machine learning (ML) as an alternative for accelerating materials discovery.

Although ML is widely promoted for its ability to capture complex structure–property relationships, its effectiveness is heavily dependent on data availability. In specialized domains such as EMI shielding, large and consistent datasets are rare (Holst, 2022). Researchers are therefore often constrained to a data-scarce regime, where evaluating the performance and generalizability of data-driven models remains a significant hurdle. While many studies report successful ML applications, few rigorously evaluate the limits of model performance under insufficient data (Husnain et al., 2024; Kalantarnia, 2026; Narayanan et al., 2023).

This study directly addresses the challenge of quantifying ML performance boundaries in data-scarce conditions. Purely data-driven models can, in principle, capture non-linear and emergent behaviors, but they also risk exhibiting large generalization gaps, where models that perform well on training-like data fail when confronted with new data (Hiremath et al., 2025). This issue is particularly relevant in materials science, where experimental data are expensive to generate and often sparsely distributed across the design space.

This work addresses these challenges by establishing a rigorous performance ceiling for a purely data-driven ML model predicting EMI SE under data scarcity. We employ an optimized Boosted Trees model using the Least-Squares (LS) Boost algorithm with four input features: the dielectric constant of the matrix, the dielectric constant of the filler, the frequency, and the filler loading. These features were selected because they are widely reported and have strong theoretical relevance to shielding behaviour, enabling broad applicability. The LS Boost, a variant of Gradient Boosting, is well-suited for the complex, non-linear, and multi-parametric nature of EMI SE. Its high predictive accuracy, intrinsic feature importance capability, and robustness against overfitting make it a strong candidate for the small-data regime typical of experimental materials science. The model is trained on a curated dataset and evaluated on a completely independent and unseen test set sourced from distinct experimental conditions. This approach benchmarks ML performance under data scarcity and provides practical guidance for EMI shielding material design when only limited data are available.

2.0 THEORETICAL FRAMEWORK

The predictive model developed in this work is based on the Least Squares Boosting (LSBoost) algorithm, an ensemble method that constructs a strong regressor by combining multiple weak learners in a sequential, additive manner. The following section details the mathematical formulation of the algorithm as implemented for this EMI SE prediction task.

2.1 Gradient Boosting Fundamentals

Boosting methods combine multiple weak learners to create a single strong learner through iterative refinement (Alajmi et al., 2021; Friedman, 2001; Shahani et al., 2021). For regression problems, the goal is to approximate a function F(x) that maps input feature vectors $x \in R^p$ to output values $y \in R$ by minimising a specified loss function. Given a training dataset $\{x_i, y_i\}$ from i=1 to N with N samples and p features, the boosting algorithm constructs an additive expansion of the form:

$$F(x) = \sum_{m=0}^{M} \beta_m \, h(x;\, a_m) \tag{1}$$

Where $h(x;\, a_m)$ represents a weak learner parameterized by a_m, β_m is the expansion coefficient, and M is the total number of boosting iterations. The initial model $F_{0(x)}$ is typically a constant value minimizing the loss function.

2.2 Least Squares Boosting Algorithm

The LSBoost algorithm employs the squared error loss function:

$$L\big(y, F(x)\big) = \frac{1}{2}\big(y - F(x)\big)^2 \tag{2}$$

The negative gradient of this loss function with respect to the model output is:

$$-\frac{\partial L\big(y, F(x)\big)}{\partial F(x)} = y - F(x) \tag{3}$$

This negative gradient corresponds exactly to the residual $r_i = y_i - F(x_i)$. At each boosting iteration m, the algorithm does the following:

i. Computes the pseudo-residuals based on the current model:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)= F_{(m-1)(x)}} = y_i - F_{(m-1)(x_i)} \quad (4)$$

ii. Fits a weak learner $h(x;\, a_m)$ to the residuals:

$$a_m = argmin_{a,\beta} \sum_{i=1}^{N} [r_{im} - \beta h(x_i;\, a)]^2 \quad (5)$$

iii. Updates the model with a shrinkage parameter ν (learning rate) to control the contribution of each new learner:

$$F_{m(x)} = F_{(m-1)(x)} + \nu \beta_m h(x;\, a_m) \quad (6)$$

The final model after M iterations is:

$$F_{M(x)} = F_{0(x)} + \nu \sum_{m=1}^{M} \beta_m h(x;\, a_m) \quad (7)$$

2.3 Regression Tree Weak Learners

The specific type of weak learner used in this implementation is a regression tree. Each tree is constructed by recursively splitting the data based on the input features (dielectric constant of matrix, dielectric constant of filler, frequency, filler loading). The splits are chosen to minimize the sum of squared errors in the resulting child nodes. The tree growth is controlled by hyper parameters to prevent overfitting. Specifically, the minimum number of data points required in a leaf node was set to 8. This prevents the tree from creating leaves that are too specific to noise in the training data. Further, the maximum number of splits (or decision rules) in the tree were set to 25. This controls the depth and complexity of each weak learner. The use of trees as weak learners is particularly advantageous for this application because they can naturally model non-linear relationships and interactions between the different material and operational parameters without any need for manual feature transformation.

3.0 MATERIALS AND METHODS

3.1 Data Source and Composition

The data used in this study were compiled from previously published experimental works (AF et al., 2016; Chauhan et al., 2018; Durmaz et al., 2023; Fayzan et al., 2019; Hota et al., 2025; Joseph et al., 2019; Khamis et al., 2020; Tolvanen et al., 2019; Zdrojek et al., 2018; J. Zheng et al., 2020). No new experiments were conducted. The complete dataset comprises 150 unique material configurations, which were split into a training set of 100 samples and a completely independent test set of 50 samples. Critically, this split was not random but reflects a realistic scenario where a model trained on a limited set of material families is later tasked with predicting the performance of entirely new material systems.

3.2 Model Inputs and Target

The model uses four fundamental parameters as input features: matrix dielectric constant (ε'_m), filler dielectric constant (ε'_f) frequency in GHz, and filler loading in vol%. The target variable for the regression task is the measured EMI Shielding Effectiveness (SE) in decibels (dB). Critically, no feature engineering was performed. The model was trained directly on these raw features without creating derived parameters (e.g., dielectric contrast, conductivity estimates) or incorporating any physics-based formulas, ensuring that the evaluation reflects the pure data-driven learning capacity of the algorithm.

3.3 Training Set Composition (100 samples)

The training data is primarily composed of systematic experimental studies on a limited number of polymer-filler systems. The dominant material families include:

a) **PLA/Graphene** (28 samples) (Tolvanen et al., 2019): Systematic variation of frequency (18–26 GHz) and filler loading (10–43 vol%).
b) **PCL/NZF** (18 samples) (AF et al., 2016): Data across frequencies of 8, 9, 10, 11, and 12 GHz with loadings from 2.5 to 12.5 vol%.
c) **Epoxy/OPEFB** (24 samples) (Khamis et al., 2020): Data across frequencies of 8, 9, 10, 11, and 12 GHz with loadings from 5 to 40 vol%.
d) **PEK/$BaTiO_3$** (15 samples) (Chauhan et al., 2018): Data across frequencies of 8, 9, 10, 11, and 12 GHz with loadings from 10 to 50 vol%.
e) **PA11/PLA/CF** (9 samples) (Durmaz et al., 2023): Data across frequencies of 8, 9, 10, 11, and 12 GHz with loadings from 10 to 30 vol%.

The remaining samples consist of smaller contributions from other systems, including PVC/Graphene (Joseph et al., 2019), PMMA/Graphene (Joseph et al., 2019), PDMS/Graphene (Zdrojek et al., 2018), and PLA/OPEFB/MWCNTs (Lakin et al., 2020), which provide limited diversity but do not constitute systematic coverage.

3.4 Independent Test Set Composition (50 samples)

The test set was curated from distinct experimental conditions designed to challenge the model's ability to extrapolate beyond its training domain. It comprises three categories of unseen data:

a) Entirely New Material Classes (Chemical Extrapolation)
- PVDF/$LiNbO_3$ (5 samples) (Hota et al., 2025): A new polymer matrix (PVDF, ε'=11) with a complex ceramic filler (Lithium Niobate) at frequencies of 6–18 GHz.
- PVC/PANI (4 samples) (Fayzan et al., 2019): A new polymer-filler combination (PVC matrix with Polyaniline filler) at low frequencies (2–8 GHz).
- PVP/Υ-Fe_2O_3 (4 samples) (J. Zheng et al., 2020): A new polymer (PVP) with a magnetic iron oxide filler at very low frequencies (2–3 GHz).
- PDMS/Graphene (2 samples) (Zdrojek et al., 2018): While PDMS/Graphene appears in the training set, these test samples are at an entirely new frequency (3 GHz) far outside the training range (10 GHz).

b) New Frequencies of Known Systems (Frequency Extrapolation)
- **PLA/OPEFB/MWCNTs** (4 samples) (Lakin et al., 2020): Tested at 8 GHz, a frequency not present in the training data for this specific composite (which was trained on 10, 12, and 22 GHz).

- **PCL/NZF** (5 samples) (AF et al., 2016): Tested at 9 GHz, which, while within the overall frequency range of the training set (8–12 GHz), represents an intermediate point not seen during training for this specific material family.

c) Overlapping Material Families (Interpolation Baseline)

- A subset of the test data originates from the same journals and material families as the training set but at new, unseen frequencies. Unlike the previous category, these samples are from systems already well-represented in the training set across multiple frequencies, allowing for a baseline comparison of how the model performs on "familiar" chemistry versus "unfamiliar" chemistry.

3.4 Model Implementation and Training

The model was implemented in MATLAB R2020a. The key steps taken for the implementation were:

i. Data Preparation: After data scrubbing and normalization, the training set was partitioned into its respective predictor and response components.

ii. Data Partitioning: The training data was partitioned using a holdout validation scheme. 20% of the 100 training samples (20 samples) were randomly selected using a fixed random number generator seed for internal validation. The remaining 80% (80 samples) were used for model training.

iii. Model Training: The LSBoost algorithm was configured with the following hyperparameters, which were determined to provide a balance of performance and generalization based on initial tests:
   - Method: LSBoost
   - NumLearningCycles: 120
   - LearnRate: 0.08
   - Learners: `templateTree('MinLeafSize', 8, 'MaxNumSplits', 25)`

The training process involved the sequential fitting of 120 regression trees, each attempting to correct the residuals of the evolving ensemble model, as described in Section 2.

3.5 Validation and Evaluation Strategy

A two-tiered evaluation strategy was employed to thoroughly assess model performance. The model's performance was first evaluated on the hold-out validation set (20 samples) from the training data. This provides an estimate of how well the model interpolates within the parameter space defined by its training data. The final, trained model was applied to the completely independent test set of 50 samples. This is the primary test for model generalizability, as these samples were not used in any part of the training or validation process and originate from different experimental sources.

The following performance metrics were calculated for both the internal validation and external test sets:

Coefficient of Determination (R^2): The proportion of the variance in the target variable that is predictable from the input features.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

Root Mean Square Error (RMSE): The square root of the average squared differences between predicted and actual values. It is in the same units as the target variable (dB).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values (dB).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

Mean Square Error (MSE): The average of the squared differences between predicted and actual values (dB^2).

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{y}_\mathrm{i})^2 \tag{11}$$

3.7 Feature Space Analysis

To quantify distributional differences between training and test data, principal component analysis was performed on the complete feature set. The first two principal components were computed from the combined dataset, and the projections of the training and test samples onto this reduced space were compared. This analysis provides insight into whether the independent test set lies within or outside the training distribution, helping to interpret generalization performance.

3.8 Feature Importance Quantification

Predictor importance was estimated by summing changes in mean squared error due to splits on each predictor across all trees in the ensemble, normalized by the number of tree nodes. This metric indicates the relative contribution of each input feature to the model's predictive capability.

3.9 Partial Dependence Analysis

Partial dependence plots were generated to visualize the relationship between individual input features and predicted EMI SE while averaging out the effects of all other features. For a given feature x_s, the partial dependence function is:

$$\hat{f}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n}\hat{F}\left(x_s, x_{i,\backslash s}\right) \tag{12}$$

3.10 Learning Curve Analysis

To assess the impact of training data quantity on model performance, learning curves were generated by training models on progressively larger subsets of the training data (40 to 100 samples) and evaluating both training RMSE and validation RMSE. This analysis determines whether the current training set size is sufficient or whether additional data would likely improve generalization.

4. RESULTS AND DISCUSSION

4.1 Performance on Training and Internal Validation Data

The performance of the optimized LSBoost model was first evaluated using the internal validation dataset obtained from the training data. Figure 1(a) shows the parity plot comparing the predicted EMI shielding effectiveness (SE) against experimentally measured values. A strong linear correlation is observed, with most data points closely aligned along the 1:1 reference line. The model achieves R^2 = 0.931 and an RMSE of 2.45 dB on the validation set, indicating a high level of predictive accuracy within the bounds of the training domain (Ileri, 2025).

The corresponding residual distribution, shown in Figure 1(b), is centred near zero with no clear systematic bias across the predicted SE range. This suggests that the model does not consistently overestimate or underestimate shielding effectiveness for the majority of samples. While a small number of larger residuals are present, particularly at higher SE values, these deviations are expected given the intrinsic variability of experimental EMI measurements and the limited sample density in certain regions of the parameter space. Taken together, the parity and residual analyses demonstrate that the boosted tree model effectively captures the dominant non-linear relationships governing EMI shielding behaviour when interpolating within the experimental conditions represented in the training dataset.

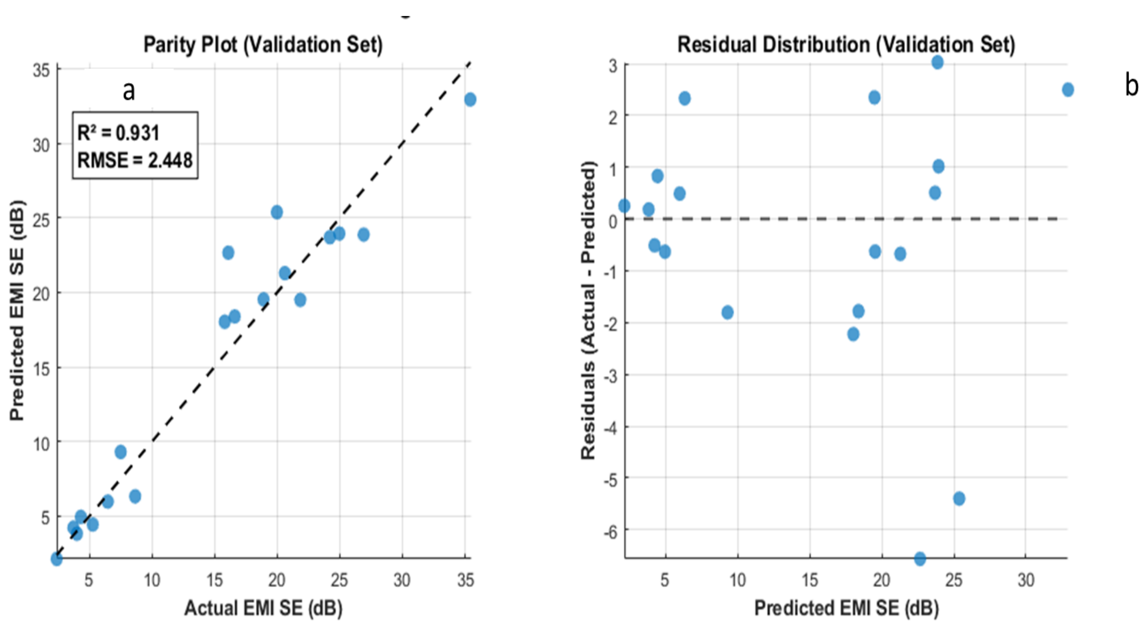

Figure 1: Parity pot and residual distribution

4.2 Feature Importance and Learned Structure

The relative importance of the input features, as extracted from the trained ensemble, is presented in Figure 2. Among the four predictors, the dielectric constant of the polymer matrix is identified as the most influential variable, followed by frequency, filler loading, and the dielectric constant of the filler (Ji et al., 2022). This ranking highlights the central role of the matrix material in determining impedance matching and electromagnetic wave attenuation, particularly in polymer-based composites.

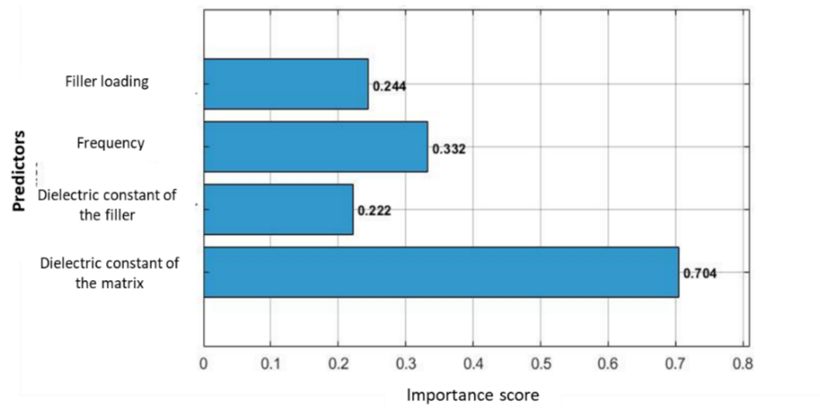

Figure 2: Feature importance

To further examine how individual predictors influence model output, partial dependence plots were analyzed. Figure 3 illustrates the partial dependence of predicted EMI SE on the dielectric constant of the filler. The response exhibits a step-like trend, indicating that increases in filler permittivity led to diminishing or even reduced gains in shielding effectiveness beyond a certain threshold. This non-monotonic behavior reflects the complex interplay between dielectric mismatch, interfacial polarization, and wave reflection mechanisms, which are not easily captured by simplified analytical models (Thao et al., 2025).

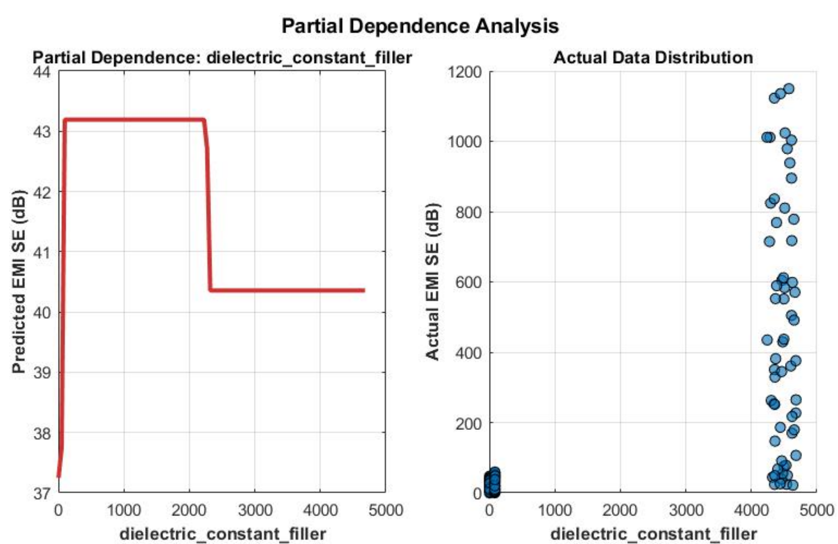

Figure 3: Partial Dependence analysis for the filler dielectric constant

Figure 4 presents the partial dependence on filler loading. A clear increase in predicted EMI SE is observed with increasing filler concentration, followed by a plateau at higher loadings. This behavior is consistent with percolation-driven effects commonly reported in EMI shielding composites, where conductive or high-permittivity pathways become saturated beyond a critical filler content. In contrast, the partial dependence on frequency (Figure 5) remains largely flat over the investigated range. This result suggests that, within the limits of the available data, frequency alone contributes less to prediction variance than compositional parameters. It also reflects the uneven sampling of frequency-dependent data, as shown in the accompanying data distribution plots, which constrains the model's ability to learn a strong independent frequency trend.

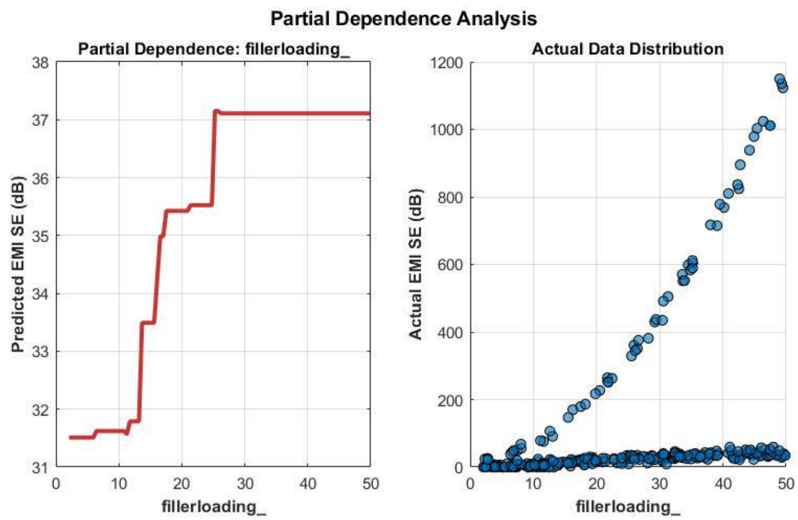

Figure 4: Partial Dependence analysis for the filler loading

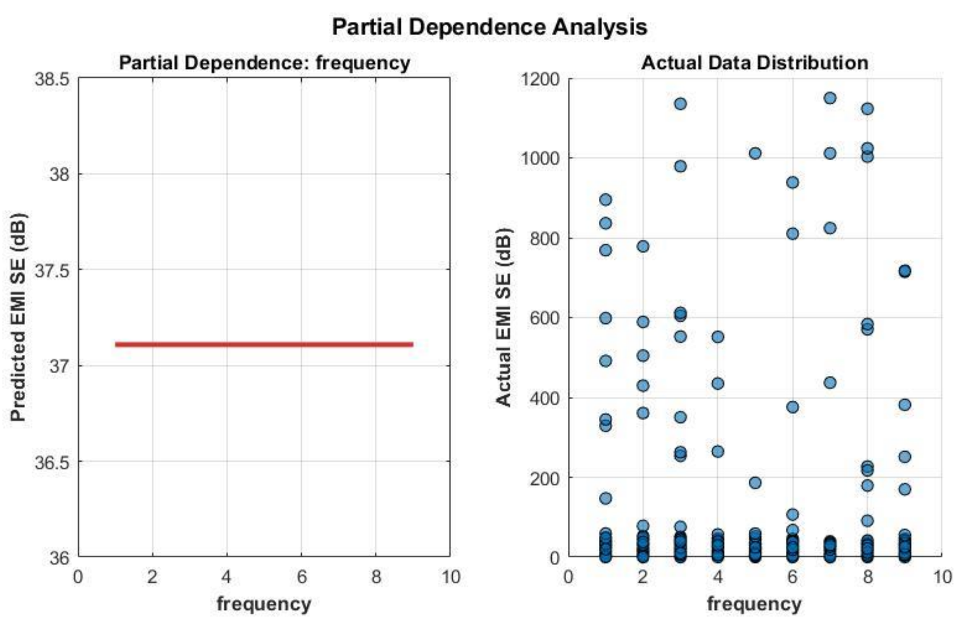

Figure 5: Partial Dependence analysis for the frequency

4.3 Learning Curve Analysis and Data Sufficiency

The learning curve shown in Figure 6 provides insight into the relationship between training data quantity and model performance. As the number of training samples increases, both training and validation RMSE decrease significantly, confirming that the model continues to benefit from additional data. At smaller training set sizes, a pronounced gap exists between training and validation errors, indicating high variance and limited generalization capability. Although this gap narrows as more samples are introduced, it does not fully close within the available dataset size. This observation indicates that the model has not yet reached a data-saturated regime and that predictive performance remains fundamentally constrained by data scarcity. Importantly, this behavior reinforces the central premise of this study: in EMI shielding applications, model accuracy is limited not only by algorithm choice but also by the availability and diversity of experimental data.

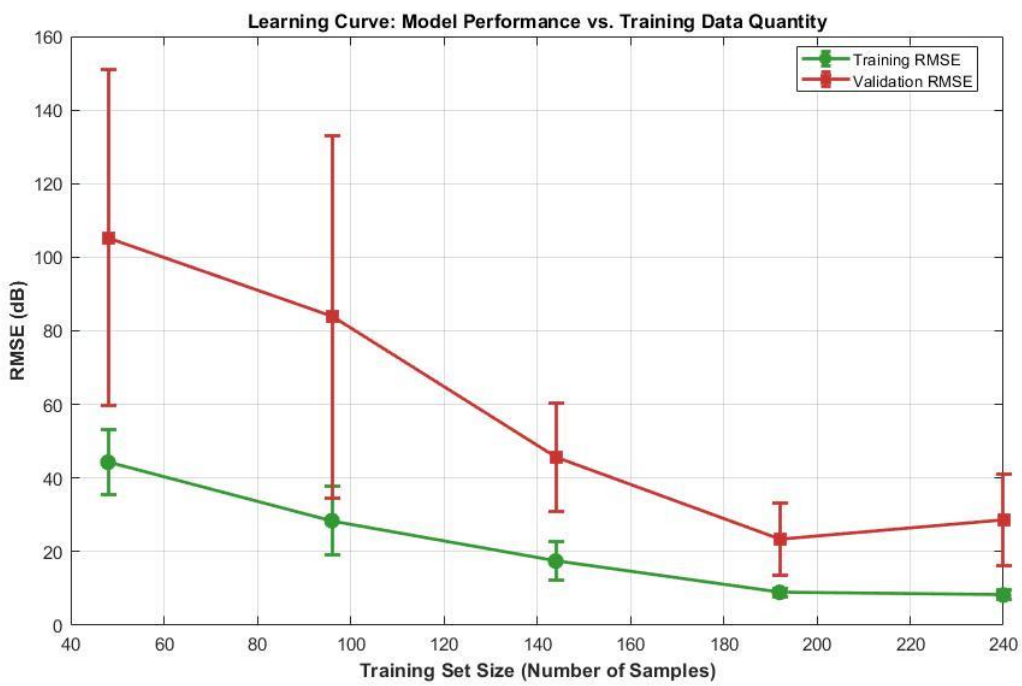

Figure 6: Learning curve analysis

4.4 Generalization to Completely Unseen Data

The predictive performance of the trained model was evaluated on the fully independent test set of 50 samples (Figure 7), revealing a substantial reduction in accuracy compared to internal validation ($R^2 = 0.597$, RMSE = 11.75 dB). However, disaggregating the test set by material family reveals that this generalization gap is not uniform but is instead highly dependent on the chemical novelty of the test sample. The most severe degradation occurs for material systems entirely absent from the training data, such as PVDF/$LiNbO_3$, which achieves exceptionally high shielding effectiveness (>60 dB) that the model dramatically underpredicts, as these samples occupy a region of feature space (high matrix dielectric constant, complex high-permittivity filler) with no analogue in the training distribution.

Similarly, PVC/PANI and PVP/ϒ-Fe_2O_3 samples are poorly predicted, representing entirely new filler types and frequency ranges for which the model has no learned representation. In contrast, the model exhibits comparatively better performance on test samples from material families present in the training set but at new frequencies; for example, PCL/NZF tested at 9 GHz, Epoxy/OPEFB at 8-9 GHz, and PEK/$BaTiO_3$ at 9 GHz are predicted with reasonable accuracy, suggesting the model can interpolate between frequencies for a material system whose composition it has "learned."

This analysis provides reveals that the generalization gap is not merely a statistical object of distribution shift, but a direct consequence of the model's inability to perform chemical extrapolation. The LSBoost algorithm functions as a sophisticated interpolator within familiar compositional space but cannot reason about the physical behavior of entirely new matrix-filler combinations, emphasizing that in the absence of representative training examples, a purely data-driven model cannot discover emergent behavior in new material classes (Ghosh, 2024).

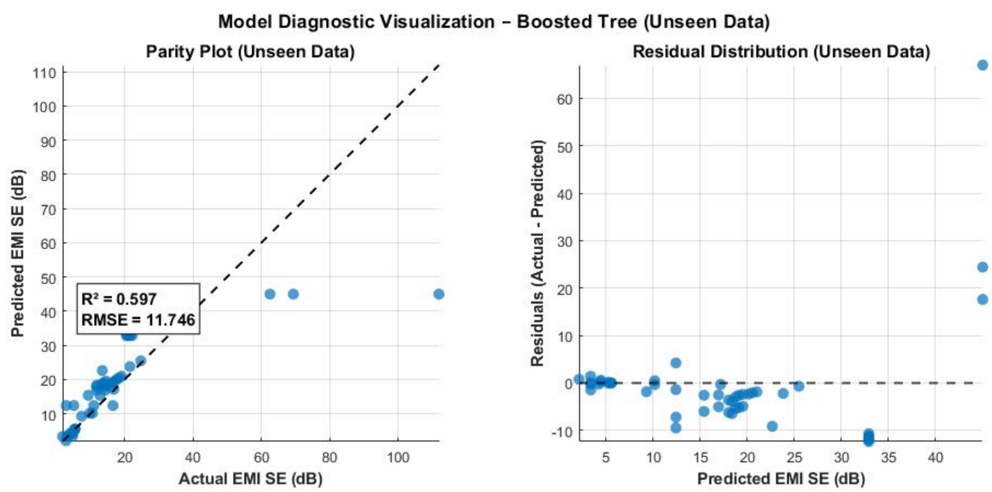

Figure 7: Parity plot and residual distribution for the unseen data

The combined results establish a clear and quantitative performance ceiling for boosted tree regression applied to EMI shielding prediction under data-scarce conditions. The model demonstrates excellent interpolation capability and learns physically meaningful relationships among material parameters. However, its reduced accuracy on unseen data highlights the inherent limitations of purely data-driven approaches when experimental coverage is sparse and uneven. Rather than indicating a shortcoming of the LSBoost framework, these findings emphasize the importance of dataset quality, diversity, and representativeness in materials informatics. The results suggest that future gains in predictive performance are more likely to arise from targeted data expansion or hybrid modelling strategies than from further algorithmic tuning alone (Zheng et al., 2026). Consequently, this study provides a realistic benchmark for machine learning–based EMI shielding prediction and offers practical guidance on the achievable accuracy and limitations of such models in real-world, data-constrained materials research.

Similarly, even though, LSBoost was selected in this study for its robustness and ability to model non-linear interaction, future work should compare its performance with other classic ensemble techniques such as XGBoost, Random Forest and Gaussian Process Regression assess whether the generalization gap observed are as a result of algorithm specific or fundamentally a data-driven issue.

CONCLUSION

This study establishes a performance baseline for boosted tree regression predicting electromagnetic interference shielding effectiveness under data-scarce conditions typical of experimental materials science. The optimized LSBoost model, trained on four readily measurable input features from 100 samples, achieves excellent interpolation capability within the training domain (R^2 = 0.931, RMSE = 2.448 dB) but exhibits substantial performance degradation on independent test data from different experimental conditions (R^2 = 0.597, RMSE = 11.746 dB). Principal component analysis confirms that this generalization gap results from distribution shift between training and test data, with test samples occupying regions of feature space poorly represented during training. Feature importance analysis reveals that matrix dielectric constant dominates predictions (score 0.704), followed by frequency (0.332), filler loading (0.244), and filler dielectric constant (0.222).

Partial dependence analysis identifies a prediction ceiling at 37.5 dB, demonstrating that purely data-driven models cannot extrapolate beyond training ranges. Learning curve analysis shows continued performance improvement with additional data beyond 100 samples, indicating persistence of the data-scarce regime.

These findings quantify fundamental limitations of purely data-driven approaches for materials property prediction and provide practical guidance, as excellent validation metrics do not guarantee generalization particularly when deployment condition differ from training data. Feature space coverage is more critical than sample count alone, and prediction beyond training ranges is unreliable. For reliable machine learning in materials design, purely data-driven approaches are insufficient under data scarcity. Thus, hybrid models incorporating physics-based constraints or domain adaptation techniques are needed, alongside improved dataset diversity and feature space coverage.

REFERENCES

AF, A., Z, A., SJ, O., & DM, A. (2016). Attenuation Performance of Polymer Composites Incorporating NZF Filler for Electromagnetic Interference Shielding at Microwave Frequencies. *Journal of Material Science & Engineering*, *05*(06), 6–11.

Alajmi, M. S., & Almeshal, A. M. (2021). applied sciences Least Squares Boosting Ensemble and Quantum-Behaved Particle Swarm Optimization for Predicting the Surface Roughness in Face Milling Process of Aluminum Material. *Applied Sciences (Switzerland).*

Chauhan, S. S., Verma, P., Malik, R. S., & Choudhary, V. (2018). Thermomechanically stable dielectric composites based on poly (ether ketone) and BaTiO 3 with improved electromagnetic shielding properties in X-band. *Applied Polymer Science*, *46413*, 1–11.

Chung, D. D. L. (2020). Materials for electromagnetic interference shielding. *Materials Chemistry and Physics*, *255*(July), 123587.

Durmaz, B. U., Salman, A. O., & Aytac, A. (2023). Electromagnetic Interference Shielding Performances of Carbon-Fiber-Reinforced PA11 / PLA Composites in the X - Band Frequency Range. *ACS Omega*, *2023*(8), 22762–22773.

Fayzan, M., Nawaz, A., Khan, R., Javed, S., Tariq, A., Azeem, M., … Jan, R. (2019). EMI shielding properties of polymer blends with inclusion of graphene nano platelets. *Results in Physics*, *14*(May), 102365.

Friedman, J. H. (2001). Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL : http://www. *Anals of Statistics*, *29*(5), 1189–1232.

Ghosh, A. (2024). Towards physics-informed explainable machine learning and causal models for materials research. *Computational Materials Science*, *233*(10), 112740.

Hiremath, P., Bhat, S. K., Jayashree, P. K., Rao, P. K., Ambiger, K. D., Murthy, B. R. N., … Naik, N. (2025). Data-Driven Prediction of Polymer Nanocomposite Tensile Strength Through Gaussian Process Regression and Monte Carlo Simulation with Enhanced Model Reliability. *Journal of Composites Science.*

Holst, C. (2022). Scarce Data in Intelligent Technical Systems : Causes , Characteristics , and Implications. *Sci.*

Hota, S. S., Panda, D., Bhoobash, S. B., Mishra, S., Biswal, L., Joshi, S., … Parashar, S. K. S. (2025). Absorption-Dominant Electromagnetic Interference Shielding of Polymer Nanocomposite PVDF / LiNbO 3 for High-Frequency Microwave Application. *Applied Electronic MAterials*, *2025*(7), 4481–4492.

Husnain, A., Iqbal, M., Ashraf, M., Faisal, M., Alabduljabbar, H., Salama, D., & Elminaam, A. (2024). Case Studies in Construction Materials Machine learning approaches for predicting shielding effectiveness of carbon fiber-reinforced mortars. *Case Studies in Construction Materials*, *20*(April), e03189.

Hwang, U., Kim, J., Seol, M., Lee, B., Park, I., Suhr, J., & Nam, J. (2022). Quantitative Interpretation of Electromagnetic Interference Shielding E ffi ciency : Is It Really a Wave Absorber or a Re fl ector ? *ACS Omega*, *7*, 4135–4139.

Ileri, K. (2025). Comparative analysis of CatBoost , LightGBM , XGBoost , RF , and DT methods optimised with PSO to estimate the number of k - barriers for intrusion detection in wireless sensor networks. *International Journal of Machine Learning and Cybernetics*, *16*(9), 6937–6956.

Ji, S., Jeong, D., & Kim, C. (2022). High Dielectric Design of Polymer Composites by Using Artificial Neural Network. *Applied Sciences*.

Joseph, J., Koroth, A. K., John, D. A., Sidpara, A. M., & Paul, J. (2019). Highly fi lled multilayer thermoplastic / graphene conducting composite structures with high strength and thermal stability for electromagnetic interference shielding applications. *Applied Polymer Science*, *47792*, 1–12.

Kalantarnia, A. (2026). Machine Learning Approach for Predicting Electromagnetic Shielding Effectiveness of Metallic Grids in the Time and Frequency Domains. *IET Science, Measurement & Technology ORIGINAL*, 1–12.

Khamis, A. M., Abbas, Z., Ahmad, A. F., Azis, R. S., Abdalhadi, D. M., & Mensah, E. E. (2020). Experimental and computational study on epoxy resin reinforced with micro-sized OPEFB using rectangular waveguide and finite element method. *IET Microwaves, Antennas and Propagation*, *14*(8), 752–758.

Khamis, A. M., Abbas, Z., Azis, R. S., Mensah, E. E., & Alhaji, I. A. (2022). Preparation and characterization of Fe2O3-OPEFB-PTFE nanocomposites for microwave shielding applications. *Polymer Engineering and Science*, *September*, 1–12.

Lakin, I. I., Abbas, Z., Azis, R. S., & Ibrahim, N. A. (2020). The Effect of MWCNTs Filler on the Absorbing Properties of OPEFB / PLA Composites Using Microstrip Line at Microwave Frequency. *Materials*, *2020*(13), 4581.

Narayanan, S., Zhang, Y., & Aslani, F. (2023). Prediction Models of Shielding Effectiveness of Carbon Fibre Reinforced Cement-Based Composites against. *Sensors*.

Shahani, N. M., Kamran, M., Zheng, X., Liu, C., & Guo, X. (2021). Application of Gradient Boosting Machine Learning Algorithms to Predict Uniaxial Compressive Strength of Soft Sedimentary Rocks at Thar Coalfield. *Advances in Civil Engineering*, *2021*.

Thao, P. X., Huong, N. T., Dat, T. Q., Sa, N. T., Nhan, L. T., & Lam, D. S. (2025). Enhancement of the Ferroelectric and Ferromagnetic Characteristics of Composite Multiferroics to Facilitate Broadband Electromagnetic Wave Absorption. *Electronic Materials*.

Tolvanen, J., Hannu, J., Hietala, M., Kordas, K., & Jantunen, H. (2019). Biodegradable multiphase poly (lactic acid)/ biochar / graphite composites for electromagnetic interference shielding. *Composites Science and Technology*, *181*(March), 107704.

Wu, Z. P., Cheng, D. M., Ma, W. J., Hu, J. W., Yin, Y. H., Hu, Y. Y., … Xu, Q. F. (2016). *Electromagnetic interference shielding effectiveness of composite carbon nanotube macro-film at a high frequency range of 40 GHz to 60 GHz. March 2014.*

Yin, J., & Ma, W. (2022). A model for predicting electromagnetic shielding effectiveness of conductive fiber plain fabric. *Journal of Industrial Textiles*.

Zdrojek, M., Taube, A., & Sypek, M. (2018). Graphene-based plastic absorber for total subterahertz radiation shielding. *Nanoscalescale*, *10*(13426), 13426–13431.

Zheng, J., Sun, B., Wang, X., Cai, Z., Ning, X., Alshehri, S. M., … Long, Y. (2020). Nanoparticle – Embedded Flexible Nanofibrous Films for Electromagnetic Shielding. *Polymers*, *2020*(12).

Zheng, S., Lei, J., Xie, P., Pan, D., Liu, C., & Shen, C. (2026). AI empowerment : Establishing the synergy between artificial intelligence and electromagnetic shielding materials research. *ISCIENCE*, *29*(2), 114654.