Comparative Evaluation of Neural Network Architectures for Generalizable Human Spatial Preference Prediction in Unseen Built Environments

MARAL DOCTORARASTOO, KATHERINE A. FLANIGAN, MARIO BERGES and CHRISTOPHER MCCOMB

ABSTRACT

The capacity to predict human spatial preferences within built environments is instrumental for developing Cyber-Physical-Social Infrastructure Systems (CPSIS). A significant challenge in this domain is the generalizability of preference models, particularly their efficacy in predicting preferences within environmental configurations not encountered during training. While deep learning models have shown promise in learning complex spatial and contextual dependencies, it remains unclear which neural network architectures are most effective at generalizing to unseen layouts. To address this, we conduct a comparative study of Graph Neural Networks, Convolutional Neural Networks, and standard feedforward Neural Networks using synthetic data generated from a simplified and synthetic pocket park environment. Beginning with this illustrative case study, allows for controlled analysis of each model's ability to transfer learned preference patterns to unseen spatial scenarios. The models are evaluated based on their capacity to predict preferences influenced by heterogeneous physical, environmental, and social features. Generalizability score is calculated using the area under the precision-recall curve for the seen and unseen layouts. This generalizability score is appropriate for imbalanced data, providing insights into the suitability of each neural network architecture for preference-aware human behavior modeling in unseen built environments.

INTRODUCTION

The design and operation of Cyber-Physical-Social Infrastructure Systems (CPSIS) hinge on the accurate modeling of human spatial behavior and underlying preferences because these systems must adapt the built environment to human needs, support dynamic decision making, and function effectively in real-world environments where human presence, movement, and choices are variable and central to system performance

Maral Doctorarastoo[1], Katherine A. Flanigan, PhD[2] (Corresponding author), Mario Berge´s, PhD[3] (Mario Berge´s holds concurrent appointments at Carnegie Mellon University (CMU) and as an Amazon Scholar. This manuscript describes work at CMU and is not associated with Amazon.), Christopher McComb, PhD[4]. Email: {mdoctora[1], kflaniga[2], mberges[3], ccm[4]}@andrew.cmu.edu. Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

[1, 2]. By capturing these behavioral patterns, such models support the creation of environments that respond to users, ensuring that social objectives—such as comfort, productivity, and social interaction—are prioritized alongside, and not overshadowed by, economic goals like energy and resource efficiency. While various computational approaches exist to model human spatio-temporal behavior, a persistent challenge is the development of models that exhibit scenario-based generalizability, i.e., the capacity to predict human spatial behavior and preferences in environmental configurations not seen during model training [3, 4]. This aligns with the real-world operation and design problem faced in CPSIS in which actuation needs to be evaluated before implementation.

Many contemporary preference models struggle to generalize effectively beyond their specific training datasets. The underlying mapping we seek to capture—from high-dimensional environmental stimuli to a probability distribution over individual choices—is a highly nonlinear function that depends simultaneously on physiological, psychological, spatial, environmental, temporal, and social cues [5–7]. In principle, this mapping could be approximated by universal function approximators such as deep neural networks, which are capable of representing highly nonlinear relationships. Simpler models, such as linear regressors, may offer interpretability but often lack the capacity to capture the complex, multimodal dependencies inherent in spatial preference data. Studies have shown that environmental characteristics like visibility significantly influence seat selection in libraries [6], while factors such as performance goals, social needs, and even cultural background impact choices in classroom settings [7, 8]. Deep data-driven machine learning models have demonstrated considerable potential in capturing these complex interactions; however, they often risk overfitting to the training scenarios, which limits their performance when applied to new situations [9].

Neural network architectures offer different inductive biases for learning from influential features. Graph Neural Networks (GNNs), have recently gained attention for their proficiency in modeling relational data [10, 11], a characteristic that makes them well-suited for representing the spatial, social, and environmental features of built environments, and they are being applied to tasks like optimizing seating in sports venues [12]. In our prior work [13], we illustrated the use of GNNs to capture human preferences, thereby enriching Reinforcement Learning based behavior models. Concurrently, Convolutional Neural Networks (CNNs) with their ability to learn from grid-like spatial inputs and have also been employed for seat recommendations [14]. Simpler Multilayer Perceptrons (MLPs) often serve as a baseline, processing features independently or with limited local context, and have been explored in studies identifying factors affecting student seat selection [7]. What remains unclear is how effectively these different architectural biases translate into models that can predict human spatial preferences not just in existing settings but, more importantly, in new and unseen environmental layouts. Each of these architectures possesses distinct inductive biases regarding how spatial context and features are learned. Establishing which types of architectures are most suitable for achieving scenario-based generalizability is essential for developing adaptable preference models for CPSIS. This paper addresses this gap by conducting a comparative study of GNNs, CNNs, and MLPs, for human spatial preference prediction. This study utilizes synthetic data generated using rule-based agent-based simulation for a case study involving four distinct pocket park layouts. To assess scenario-based generalizability, the leave-one-out cross-validation (LOOCV) strategy is employed. The principal contribu-

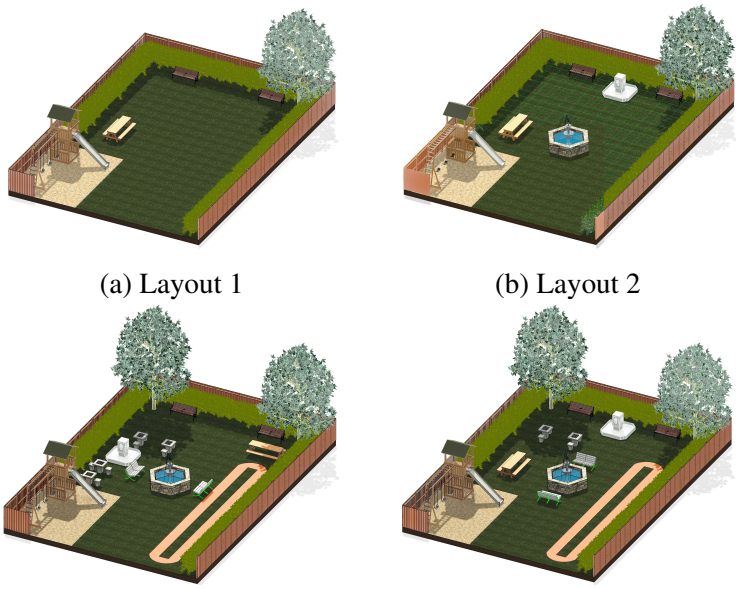
tions of this work are: (1) a qualitative and quantitative assessment of GNN, CNN and MLP generalizability to unseen environments; and (2) the derivation of insights to guide the selection of appropriate modeling techniques for developing real-world CPSIS.

METHODOLOGY

The methodology detailed herein focuses on the comparative evaluation of generalizability of GNN, CNN, and MLP architectures in predicting human spatial preferences to unseen scenarios. The core components include the definition of the pocket park environment, data representation, specifications of each preference model, and the experimental protocol for assessing performance.

Pocket Park Environment and Data Representation

A simulated park environment is employed as the context of the illustrative case study. We construct four distinct layouts (Figure 1), designed to test the models' prediction capacity in varying configurations. The park measures 15 m $\times$ 21 m and is discretized into 0.75 m $\times$ 0.75 m cells, each representing a potential location for activities such as walking on the trail, sitting, eating, or playing in playground (Figure 2).



(a) Layout 1 (b) Layout 2

(c) Layout 3 (d) Layout 4

Figure 1. Four park layouts used to evaluate generalizability across different scenarios.

Each grid cell is characterized by a feature vector x_i that encapsulates three categories of attributes. *Physical Features* detail the presence and type of park elements such as benches, picnic tables, playgrounds, monuments and other amenities; terrain types like grass, soil patches, and running tracks; and obstacles including bushes, and trees. *Environmental Features* consist of dynamic attributes such as temperature, light intensity, and shadow coverage, which are updated based on simulated time-of-day, solar position, and occlusions from physical structures. Lastly, *Social Features* encompass

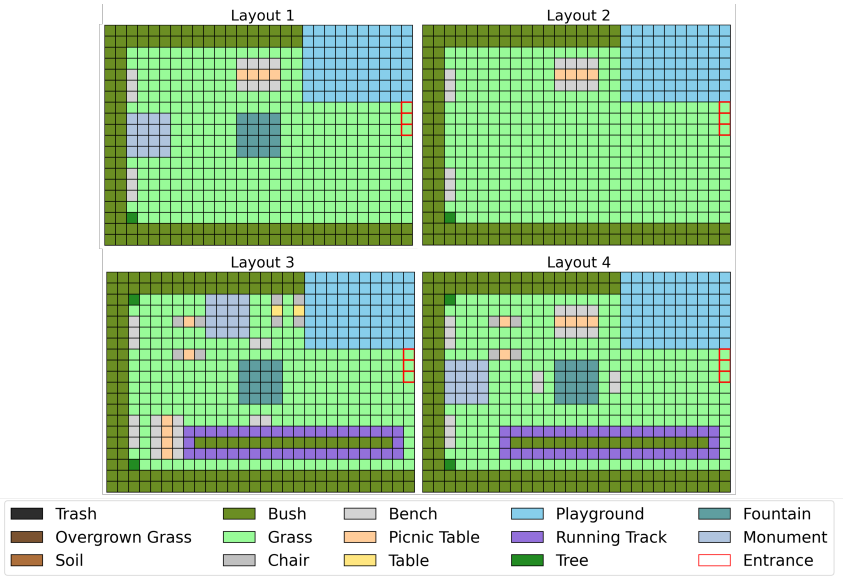


Figure 2. Grid-based representations of the park layouts.

information concerning the presence of other simulated agents within the park. *The objective of the preference models is to predict a probability indicating the likelihood of each cell being selected for a given activity, based on the interplay between agent-specific preferences (implicitly learned) and the cell's features.*

Synthetic Preference Data Generation

Synthetic ground-truth data is generated for this preliminary study due to the unavailability of real-world preference datasets for these specific park layouts. At the same time, using synthetic data allows for full control over the environment and enables a systematic exploration of model behavior and generalizability under varying spatial conditions. A rule-based agent-based simulation algorithm based on by [15] is conducted where agents navigate the distinct park layouts, select locations, and engage in activities. These decisions are governed by predefined preference rules that interact with the dynamic physical, environmental, and social attributes of each layout. For training the preference models, the location selected by an agent for a particular activity within these simulations serves as the positive target sample. The input data for the GNN, CNN, and MLP models comprises the feature vector for every grid cell, while the target is a binary label indicating whether that cell was chosen for the activity. The input and output in training data are intentionally selected to be easily replaced by sensor-collected data with no data from internal human states that are not possible to collect using privacy-preserving sensor employed in the environment. To increase the diversity of training scenarios, data augmentation is applied to the training samples. This includes 180-degree rotations and planar (horizontal and vertical) flips of the entire park layout grid.

Preference Model Architectures

To ensure a fair comparison that emphasizes architectural biases, four neural network architectures—a GNN, a 2D CNN (CNN2D), a 1D CNN (CNN1D), and a MLP—were constructed with similar parameter counts (approximately 74k–76k). All models min-

TABLE I. Summary of neural network architectures.

Model	Input Representation	Core Architecture
GNN	Graph (nodes = cells, edges = 8-way neighbors)	5-layer GCN & ReLU, plus a final GCN output layer with Sigmoid
CNN2D	2D grid of the layout with features as channels	3-layer Conv2D & ReLU, plus a final 1x1 convolution with Sigmoid
CNN1D	Sequence of per-cell feature vectors	3-layer Conv1D & ReLU, followed by a Sigmoid activation
MLP	Per-cell feature vector (augmented with 3x3 context)	4-layer fully-connected network & ReLU, terminating in a Sigmoid output

imize a weighted Binary Cross-Entropy (BCE) loss, suitable for the inherently imbalanced preference prediction task, as only one cell per instance is chosen for any activity. The specific architectural details for each model are summarized in Table I.

Experimental Design and Evaluation Metrics

A LOOCV strategy is central to this study. The process involves four experimental folds. In each fold, data generated from three of the four distinct park layouts are aggregated to form the training set for the GNN, CNN2D, CNN1D and MLP preference models; the data from the single remaining park layout is reserved as the unseen test set for that fold. This rotation ensures that each layout serves as the unseen test scenario once, providing an assessment of generalizability across different spatial configurations.

Performance Metrics: The models' performance is assessed using several key metrics. The Weighted BCE Loss tracks prediction error during training. For evaluating discriminative ability on this highly imbalanced preference prediction task (only one cell is *chosen* out of many), the Area Under the ROC Curve (ROC AUC) is commonly used. However, ROC AUC can be misleadingly optimistic in scenarios with a large skew between positive and negative classes, as a high number of true negatives (correctly identifying *non-chosen* cells) can increase the score even if the model performs poorly on the rare positive class (*chosen* cells). Therefore, we place a stronger emphasis on the Area Under the Precision-Recall Curve (AUPRC). AUPRC provides a more informative assessment of performance on the positive class, which is of primary interest here, as its baseline is the fraction of positives in the dataset [16]. To quantify the retention of performance when models are applied to unseen layouts, we define Generalizability Scores (GS) per Eq. 1, where a GS value closer to 1 signifies superior generalizability of the model to unseen environment layouts:

$$GS_{AUPRC} = \frac{\text{AUPRC on Unseen Test Layout}}{\text{Average AUPRC on Seen Training Layouts (validation splits)}} \quad (1)$$

RESULTS AND DISCUSSION

This section presents the comparative evaluation of the GNN, CNN2D, CNN1D, and MLP models. We first summarize the final generalization scores to provide a high-level

comparison of the architectures, and then analysis of performance variance in each unseen layout. All models were trained for up to 100 epochs with a learning rate of 10^{-3}, employing early stopping with a 30-epoch patience threshold to prevent overfitting. Convergence was monitored using weighted BCE loss on both training data and a held-out validation subset composed of the three seen layouts per LOOCV fold.

Comparative Performance on Seen and Unseen Park Layouts

The aim of this study is to assess the generalization capability of each architecture to park layouts not encountered during training. Figure 3 presents the GS_{AUPRC} averaged across three agents with various preferences for each unseen test layout in the LOOCV protocol, along with the mean over all folds (*Overall_Avg*). The results indicate a clear performance hierarchy. The GNN architecture achieves the highest overall generalization score (Overall_Avg $GS_{AUPRC} \approx 0.70$), closely followed by the CNN2D (Overall_Avg $GS_{AUPRC} \approx 0.69$), suggesting both models retain their predictive performance well in unseen layouts. Conversely, the MLP and CNN1D exhibit significantly lower overall scores (0.38 and 0.33, respectively), highlighting their limitations in adapting to new spatial configurations. Notably, performance varies significantly across the individual layouts. Both GNN and CNN2D demonstrate exceptional generalization on *Layout 1* ($GS_{AUPRC} = 1.09$ and 1.02, respectively) and *Layout 2* ($GS_{AUPRC} = 0.86$ and 1.00, respectively), with scores near or above 1.0 indicating that their performance on unseen tests was on par with or exceeded their performance on the validation data. In contrast, all architectures struggled with *Layout 3*, which consistently yielded the lowest GS_{AUPRC} across all models (all ≤ 0.25). The denser arrangement of *Layout 3* may have introduced unfamiliar spatial patterns, contributing to reduced generalization performance.

To further analyze model behavior, Figure 4 illustrates the average Test AUPRC over the training epochs for each model architecture, with each curve showing the model's performance on one of the four layouts when it served as the unseen test set. Models with strong spatial inductive biases, GNN and CNN2D (Figures 4a and 4b, respectively), exhibit high and consistent AUPRC scores (approximately 0.5–0.6) on *Layouts 1* and *2*, moderate performance on *Layout 4*, and correctly identify *Layout 3* as particularly chal-

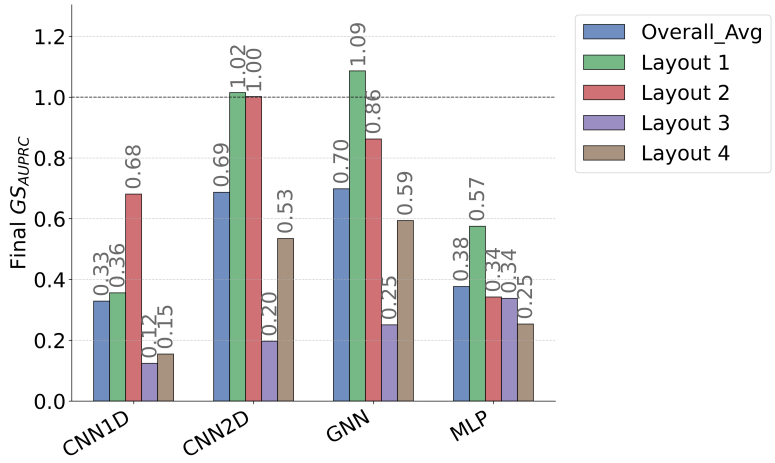
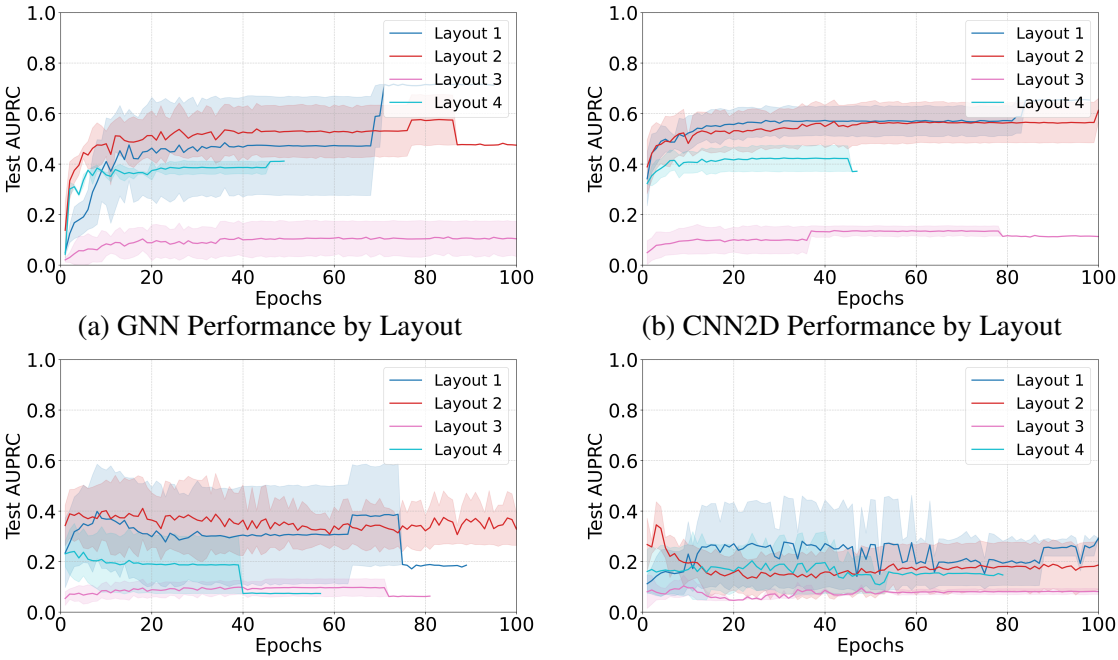


Figure 3. Final generalizability score. (GS_{AUPRC})



(a) GNN Performance by Layout

(b) CNN2D Performance by Layout

(c) CNN1D Performance by Layout

(d) MLP Performance by Layout

Figure 4. Average Test AUPRC vs. Epochs for each model architecture. The shaded regions represent the standard deviation across agents.

lenging (AUPRC < 0.2). These trends suggest that these models have learned meaningful and generalizable spatial features. In contrast, the CNN1D and MLP models (Figures 4c and 4d) display highly unstable and noisy performance across all layouts, failing to converge to a strong solution. The MLP, in particular, shows very little variance between the different layouts; it performs equally poorly on all of them, with all curves clustered at a low AUPRC (<0.3). This behavior reinforces the conclusion that while models with strong spatial priors can generalize to diverse, unseen scenarios, to different extents, the simpler models lack the architectural robustness required for this task.

CONCLUDING REMARKS

This study underscores the importance of architectural inductive biases in achieving scenario-based generalizability for spatial preference prediction. GNN and CNN2D models' ability to maintain predictive performance in unseen layouts highlights their promise for real-world CPSIS applications, where environments vary dynamically. By leveraging synthetic yet controlled environments, this work lays the groundwork for evaluating and deploying preference-aware models that can adapt to spatial diversity—a critical step toward closing the loop between human behavior and infrastructure design.

ACKNOWLEDGMENTS

This work is support by the National Science Foundation under Grant #2425121.

REFERENCES

1. Doctorarastoo, M., K. A. Flanigan, M. Bergés, and C. McComb. 2023. "Exploring the Potentials and Challenges of Cyber-Physical-Social Infrastructure Systems for Achieving Human-Centered Objectives," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '23*, Istanbul, Turkey, p. 385–389, doi:10.1145/3600100.3626340.
2. Doctorarastoo, M., K. A. Flanigan, and M. Bergés. 2024. "Preference-Aware Human Spatial Behavior Modeling in Cyber-Physical-Human Systems," *IFAC-PapersOnLine*, 58(30):115–120, doi:10.1016/j.ifacol.2025.01.166.
3. Doctorarastoo, M., K. A. Flanigan, M. Bergés, and C. McComb. 2023. "Modeling human behavior in cyber-physical-social infrastructure systems," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '23*, Istanbul, Turkey, p. 370–376, doi:10.1145/3600100.3626338.
4. Qiao, G., H. Zhou, M. Kapadia, S. Yoon, and V. Pavlovic. 2019. "Scenario Generalization of Data-Driven Imitation Models in Crowd Simulation," in *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11.
5. Doctorarastoo, M., K. A. Flanigan, M. Bergés, and C. McComb. 2024. "GNN-based Predictive Modeling of Human Preferences in the Built Environment," in *ASCE International Conference on Computing in Civil Engineering 2024 (i3ce 2024)*, Pittsburgh, PA, USA.
6. Lim, L., M. Kim, J. Choi, and C. Zimring. 2018. "Seat-Choosing Behaviors And Visibility," *Journal of Architectural and Planning Research*, 35(4):271–290.
7. Losonczy-Marshall, M. and P. D. Marshall. 2013. "Factors in Students' Seat Selection: An Exploratory Study," *Psychological Reports*, 112(2):651–666.
8. Kehan, L., A. Kaur, Z. Yu, H. Yuzhen, H. Yuchong, Z. Yinuo, and M. Noman. 2024. "Seat Selection as a Function of Cultural and Individual Differences: Insights from Undergraduate Students in China," *Teaching and Learning Inquiry*, 12:1–22.
9. Tenzer, M., Z. Rasheed, and K. Shafique. 2023. "The Geospatial Generalization Problem: When Mobility Isn't Mobile," in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, pp. 1–4.
10. Honarvar, S. and Y. Diaz-Mercado. 2024. "Geometric Graph Neural Network Modeling of Human Interactions in Crowded Environments," *IFAC-PapersOnLine*, 58(28):25–30.
11. Taghizadeh, M., Z. Zandsalimi, M. A. Nabian, M. Shafiee-Jood, and N. Alemazkoor. 2025. "Interpretable Physics-Informed Graph Neural Networks for Flood Forecasting," *Computer-Aided Civil and Infrastructure Engineering*.
12. Wu, Z. and S. Wang. 2024. "Optimization of Seating Arrangement in Sports Competition Venues Based on Recurrent Graph Neural Network," *Journal of Electrical Systems*, 20(3s):2690–2701.
13. Doctorarastoo, M., K. A. Flanigan, M. Berges, and C. McComb. 2024. "Integrating Preference-Aware Modeling of Human Spatial Behavior in Cyber-Physical-Human Systems," in *NeurIPS 2024 Workshop on Behavioral Machine Learning*.
14. Moins, T., D. Aloise, and S. J. Blanchard. 2020. "Recseats: A Hybrid Convolutional Neural Network Choice Model for Seat Recommendations at Reserved Seating Venues," in *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 309–317.
15. Cheliotis, K. 2020. "An Agent-Based Model of Public Space Use," *Computers, Environment and Urban Systems*, 81:101476.
16. Sofaer, H. R., J. A. Hoeting, and C. S. Jarnevich. 2019. "The Area Under the Precision-Recall Curve as a Performance Metric for Rare Binary Events," *Methods in Ecology and Evolution*, 10(4):565–577.