Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Based Frequency-Selective RF-EMF Exposure Prediction in Mobile Communication Systems Using Spatial and Temporal Features

Doruntinë Berisha
University of Prishtina
Enver Hamiti
University of Prishtina
Hena Maloku
University of Prishtina
Mimoza Ibrani
mimoza.ibrani@uni-pr.edu
University of Prishtina

Research Article

Keywords: machine learning, electric field, frequency-selective, tree-based

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11202257/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning-Based Frequency-Selective RF-EMF Exposure Prediction in Mobile Communication Systems Using Spatial and Temporal Features

Doruntinë Berisha[1], Enver Hamiti[2], Hena Maloku[3], Mimoza Ibrani[4*]

[1,2,3,4]Faculty of Electrical and Computer Engineering, University of Prishtina, Bregu i Diellit, Prishtina, 10000, Kosovo.

*Corresponding author(s). E-mail(s): mimoza.ibrani@uni-pr.edu;
Contributing authors: doruntine.berisha@uni-pr.edu;
enver.hamiti@uni-pr.edu; hena.maloku@uni-pr.edu;

Abstract

This study investigates machine learning models for the analysis and prediction of frequency-selective electric field exposure levels across multiple wireless communication services. The research is based on an extensive in situ measurement campaign conducted in the city of Prishtina and includes seven frequency bands corresponding to GSM, UMTS-LTE, and 5G wireless services. A diverse set of machine learning models, including linear regression, instance-based learning, neural networks, ensemble tree-based algorithms, and gradient boosting methods, was evaluated for electric field exposure prediction. To assess model robustness and geographic and temporal transferability, three validation strategies were considered: a conventional random hold-out framework, a buffered spatial framework designed to reduce the influence of geographically related observations, and a leave-one-date-out framework that withholds an entire calendar measurement date at a time to test generalization to previously unseen dates. The study presents a comprehensive comparison of predictive performance, computational efficiency, and feature importance across the investigated models and frequency bands. The results demonstrate that ensemble tree-based approaches achieve the highest predictive accuracy under random hold-out and buffered spatial validation, while temporal predictors emerge as the dominant factors influencing model performance. Furthermore, the analysis highlights the impact of validation strategy on predictive results and reveals differences in model robustness

under geographically and temporally independent evaluation conditions. The findings demonstrate the potential of machine learning techniques for accurate frequency-selective RF-EMF exposure prediction and support their application in large-scale exposure assessment, monitoring, and wireless network planning activities.

Keywords: machine learning, electric field, frequency-selective, tree-based

1 Introduction

Modern wireless networks operate across multiple frequency bands and radio access technologies, resulting in increasingly complex Radio Frequency Electromagnetic Fields (RF-EMF) exposure scenarios. Characterizing exposure levels through direct measurements remains the most reliable assessment approach. However, large-scale measurement campaigns are costly, time-consuming, and difficult to maintain over extended spatial and temporal scales. Consequently, predictive modeling techniques have emerged as a promising alternative for RF-EMF exposure assessment. This study focuses on the prediction of frequency-selective electric field exposure levels using in situ measurements collected in the city of Prishtina. The analyzed frequency bands correspond to GSM, UMTS-LTE, and 5G wireless communication services. A comprehensive comparison of machine learning (ML) algorithms is performed to evaluate their ability to predict electric field exposure across the different frequency bands. The ML models evaluated in this study were selected to represent a diverse range of modeling approaches, including linear methods, instance-based learning, neural networks, ensemble tree-based algorithms, and gradient boosting techniques. This diversity enables a comprehensive evaluation of their ability to capture the complex and nonlinear relationships that characterize frequency-selective RF-EMF exposure. The main objectives of this study are fourfold. First, the predictive capability of twelve ML algorithms is evaluated for frequency-selective electric field exposure estimation across seven mobile communication services. Second, the influence of validation strategy on model performance and model ranking is investigated through a comparison of random hold-out, buffered spatial, and leave-one-date-out evaluation frameworks. Third, the trade-off between predictive accuracy and computational efficiency is assessed for the evaluated models. Finally, feature importance analyses are performed to identify the temporal and spatial factors that most strongly influence prediction performance. Given the spatial and temporal nature of the measurement data, model evaluation was conducted using three different validation frameworks. The first follows a conventional random hold-out data partitioning approach. The second introduces geographic separation between data groups. The third withholds an entire calendar measurement date at a time. This comparison provides insight into predictive accuracy and model robustness under both geographically and temporally independent conditions. The main contribution of this study is the development of a unified ML framework for frequency-selective RF-EMF exposure prediction across

seven mobile communication services, enabling a comprehensive and fair comparison of twelve ML algorithms under identical experimental conditions. In addition to evaluating predictive accuracy, the study investigates the influence of validation strategy by comparing conventional random hold-out, buffered spatial, and leave-one-date-out evaluation frameworks, providing insight into model robustness under both geographically and temporally independent conditions. The proposed framework further incorporates computational efficiency analysis, feature importance evaluation to identify the most influential temporal and spatial predictors, and non-parametric statistical significance testing to determine whether differences in predictive performance and model rankings are statistically meaningful. Finally, the framework is validated using a large-scale real-world dataset comprising 39,333 in situ frequency-selective RF-EMF measurements (collected from an initial pool of 42,413 raw samples) in a heterogeneous urban environment, corresponding to 5,619 distinct measurement events each recorded across all seven frequency-selective bands. To the best of the authors' knowledge, this study provides one of the first unified comparative evaluations of multiple ML algorithms for frequency-selective RF-EMF exposure prediction across several mobile communication bands. All models were developed and evaluated using the same feature set, preprocessing pipeline, hyperparameter optimization procedure, and evaluation metrics, ensuring a fair and consistent comparison.

The specific contributions of this study are summarized as follows:

- A unified, reproducible ML framework is developed and applied to 5,619 in situ measurement events (from an initial pool derived from 42,413 raw samples), each recorded across seven GSM, UMTS-LTE, and 5G bands for 39,333 band-level observations in total, enabling a fair comparison of twelve regression algorithms under identical preprocessing, feature, and hyperparameter-optimization conditions.
- Model performance and ranking are compared under three validation strategies – conventional random hold-out, buffered spatial, and leave-one-date-out validation – quantifying how geographic and temporal data leakage affect both absolute accuracy and relative model ranking.
- Computational efficiency is assessed jointly with predictive accuracy, and non-parametric statistical testing (Friedman test with Conover post-hoc and Holm correction) is used to establish whether differences in model ranking are statistically meaningful.
- Permutation-based feature importance analysis is performed for all evaluated models under the random hold-out and buffered spatial validation strategies to identify the temporal and spatial predictors most strongly associated with frequency-selective RF-EMF exposure.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on ML-based RF-EMF exposure prediction. Section 3 presents the methodology, including the measurement campaign, dataset preparation, feature engineering process, and ML implementation. Section 4 reports the experimental results obtained under the investigated validation strategies. Section 5 discusses the findings with respect to predictive performance, computational efficiency, and feature importance. Finally, Section 6 summarizes the main conclusions and outlines directions for future research.

2 Related Work

Recently, the application of Artificial Intelligence (AI) and ML techniques for RF-EMF exposure prediction has gained increasing attention. In [1], ML and deep learning (DL) approaches were applied to broadband RF-EMF exposure prediction using measurement data and Geographic Information System (GIS)-derived features, with the best-performing models achieving an RMSE of 0.43 V/m and an R^2 of 0.75. Several studies have investigated ML methods for electric field strength prediction and RF-EMF exposure assessment. Ensemble tree-based models, including Random Forest, XGBoost, and LightGBM, have been widely investigated for electric field prediction and radio environment mapping, often combined with explainable AI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to improve model interpretability and identify the factors influencing model predictions [1–5]. More recently, Graph Neural Networks (GNNs) have been explored to capture spatial dependencies between measurement locations for RF-EMF exposure assessment [6]. Recently, DL approaches have also been applied to frequency-selective RF-EMF prediction. In [7], a convolutional neural network framework (ExposNet) was proposed to predict broadband and frequency-selective electric field exposure levels using measurement, base station, and geographical data, demonstrating the feasibility of frequency-selective RF-EMF prediction using DL techniques in urban environments. ML techniques have also been applied to specialized RF-EMF prediction problems, including Wi-Fi exposure estimation in complex indoor environments [8], uplink power prediction in 4G networks using ANN-based frameworks [9], millimeter-wave 5G RF-EMF modeling [10], campus-wide exposure prediction [11], and time-series forecasting of RF-EMF levels using DL approaches [12]. Earlier studies primarily relied on artificial neural networks (ANNs) for RF-EMF exposure prediction, including measurement-based outdoor exposure estimation [13], extrapolation of exposure levels in urban environments [14], and long-term electric field strength prediction from cellular base stations [15]. Table 1 situates the present contribution relative to the most closely related literature in terms of prediction target, modeling approach, dataset size, best reported performance, and validation strategy. Items not explicitly reported in the original source are marked accordingly.

As summarized in Table 1, while existing studies have demonstrated the applicability of ML for RF-EMF exposure assessment, most focus on broadband or aggregate electric field prediction, rely on markedly smaller datasets than the one analyzed here, and rarely report a validation strategy that accounts for spatial dependence between samples. Furthermore, the relatively few frequency-selective studies available have primarily investigated individual ML or DL frameworks rather than providing comprehensive model comparisons, and none of the reviewed studies explicitly compare a conventional random hold-out validation against a buffered spatial validation strategy for the same models and data. Although [7] evaluates its frequency-selective model using a geographic hold-out, it does not contrast this against a conventional random split on the same data, and therefore does not quantify the effect of validation strategy itself in the way the present study does. In contrast, the present study investigates frequency-selective RF-EMF prediction across seven frequency bands using twelve ML algorithms and evaluates their performance under random hold-out,

Table 1: Comparison with closely related RF-EMF exposure prediction literature, including dataset size and best reported performance where available.

Study	Prediction target	Modeling approach	Freq.-sel.	Dataset size	Best reported performance	Spatially aware validation
[1]	Broadband RF-EMF	ML/DL ensemble with GIS features	No	Not reported	Not reported	Not reported
[2]	Electric field strength	Ensemble trees + SHAP/LIME	No	6,547 (8 cities)	Not reported	Not reported
[3]	Electric field strength	Ensemble trees + SHAP/LIME	No	Not reported	Not reported	Not reported
[4]	Electric field strength	Ensemble trees + SHAP/LIME	No	19,322 (Catalonia)	RMSE 0.623 (kNN)	Not reported
[5]	Electric field strength	Ensemble trees + SHAP/LIME	No	3,578 (Paris)	RMSE 1.63 (kNN)	Not reported
[6]	RF-EMF exposure	Graph Neural Network	No	Not reported	Not reported	Not reported
[7]	Broadband & freq.-sel. E-field	Convolutional Neural Network	Yes	1,689 (2 cities)	RMSE 0.171 V/m†	Yes (geographic hold-out)‡
[8]	Wi-Fi indoor exposure	ML (indoor multi-source)	No	Not reported	Not reported	Not reported
[9]	Uplink power (4G, indoor)	ANN	No	2,510	RMSE 2.365 dB*	No (single building)
[10]	mmWave 5G RF-EMF	ML	No	Not reported	Not reported	Not reported
[12]	RF-EMF time series	DL forecasting	No	5,760 (17 locations)	R^2 up to 0.94–0.98	No (temporal split only)
This study	Frequency-selective RF-EMF (7 bands)	12 ML/DL algorithms	Yes	39,333	0.112 (hold-out) / 0.165 (spatial CV) V/m	Yes (random & spatial, compared)

*In dB (received power), not directly comparable to the V/m values above. †Total E-field RMSE (root-sum-square across seven bands). The same source also reports 0.199 V/m for a direct total-field prediction. ‡Withholds a contiguous block of Paris samples as the test set, not a formal spatial CV scheme.

buffered spatial, and temporal (leave-one-date-out) validation strategies. In addition to predictive accuracy, computational efficiency and feature importance are analyzed to provide a comprehensive assessment of model suitability for frequency-selective RF-EMF prediction.

3 Methods

3.1 Problem Formulation and Framework Overview

The prediction task addressed in this study was formulated as a supervised regression problem. For a given frequency-selective service s, one of the seven GSM, UMTS-LTE,

and 5G uplink/downlink bands considered, let $\mathbf{x}_i^{(s)} \in \mathbb{R}^p$ denote the feature vector of the i-th measurement sample, composed of $p = 12$ spatial and temporal predictors, and let $y_i^{(s)} \in \mathbb{R}_{\geq 0}$ denote the corresponding measured electric field strength (V/m). For each service s and each candidate model f_θ, the goal was to learn a mapping

$$\hat{y}_i^{(s)} = f_\theta\left(\mathbf{x}_i^{(s)}\right) \tag{1}$$

that minimizes the discrepancy between predicted and observed exposure levels, where θ denotes the model's parameters. Tuning hyperparameters once per service and reusing them across folds can leak information from the evaluation partition into training and produce an optimistic bias. To avoid this, θ was re-tuned via a randomized hyperparameter search independently within every individual evaluation fold, using only that fold's own training data.

Twelve candidate model families were considered for f_θ, spanning linear, instance-based, kernel- and neural-network-based, and ensemble tree-based methods. Each was evaluated independently for every one of the seven services under three validation lenses applied in parallel. The first was a conventional random hold-out partition, retained as a naive reference. The second was a five-fold buffered spatial partition, designed to test geographic generalization. The third was a fifteen-fold leave-one-date-out partition, designed to test temporal generalization. Together, these yielded $12 \times 7 \times 21 = 1{,}764$ independently tuned and evaluated (service, model, fold) configurations. Model quality was quantified using four standard error and goodness-of-fit measures.

Figure 1 summarizes the overall analysis pipeline that implements this formulation, from raw measurement data and feature engineering, through the per-fold hyperparameter tuning step nested within each of the three validation lenses, to evaluation and the band-balanced aggregation of the performance, efficiency, and feature-importance results. Because hyperparameters were re-tuned independently within every fold using only that fold's own training data, the results reported for each validation lens reflect that lens's own generalization scenario, without the tuning step leaking information from the evaluation partition. This came at the cost of the substantially larger computational burden that repeated tuning entails.

3.2 Measured Data

For this study, measurement data from an in situ RF-EMF exposure measurement campaign conducted in the city center of Prishtina between June 2023 and December 2024 were used. The campaign included multiple mobile communication technologies, ranging from 2G to recently deployed 5G services. Measurements were performed using the NARDA SRM-3006 spectrum analyzer operating in Safety Evaluation Mode with the predefined service table configuration.

The dataset consists of real in situ frequency-selective RF-EMF measurements, collected at a height of 1.5 m above ground level, corresponding to individual uplink and downlink frequency bands of GSM, UMTS-LTE, and 5G services, rather than simulated or model-generated exposure data. Because the spectrum analyzer was operated at street level rather than co-located with a specific user device, the uplink-band measurements reported here characterize the ambient RF-EMF levels present in the uplink

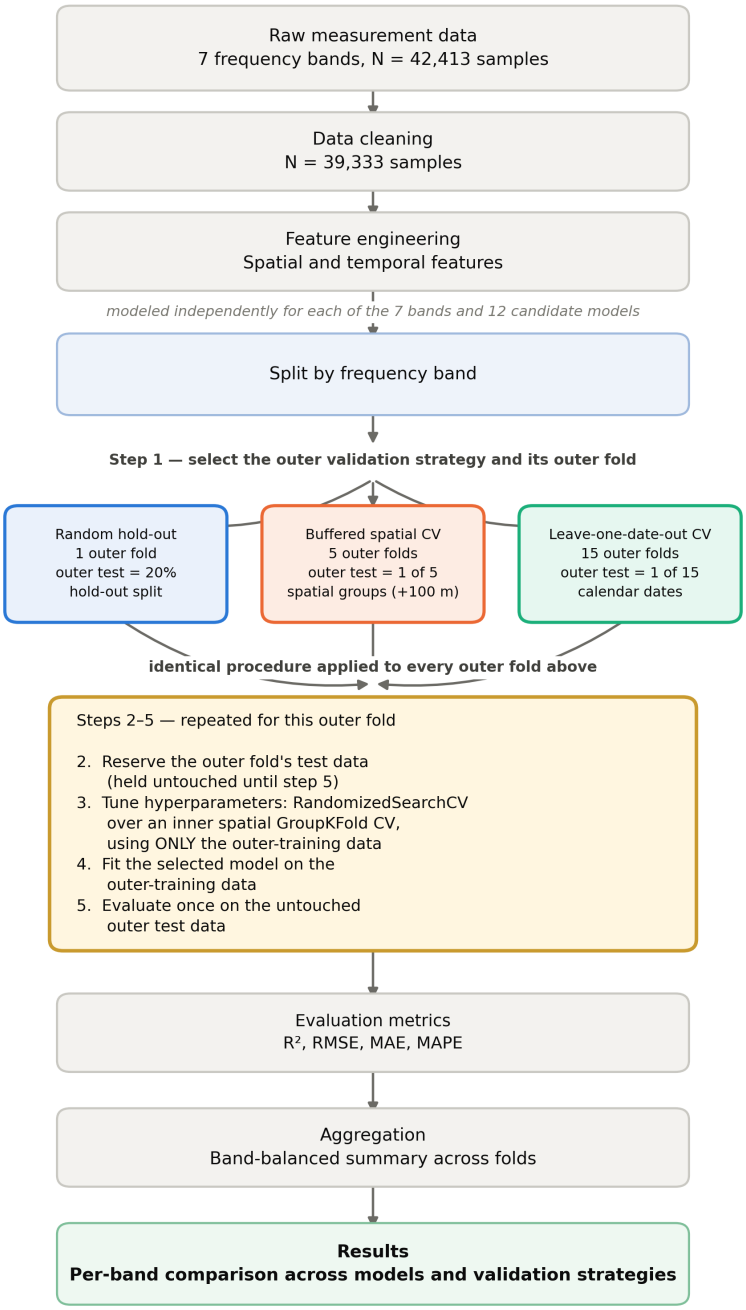


Fig. 1: Modeling and validation pipeline.

frequency bands at the measurement location, rather than the personal exposure of an individual handset. For each measurement sample, the recorded information included the measured electric field level, operating frequency range, measurement date and time, geographic coordinates (latitude and longitude), and weekday information. In total, 42,413 measurement samples were collected during the measurement campaign. The spatial distribution of the measurement locations is presented in Figure 2.

Measurements were acquired along predefined routes covering the investigated urban area over multiple measurement days. Individual geographic regions were sampled during multiple time periods, ensuring that the dataset captures both spatial variability across the study area and temporal variability associated with changes in wireless network utilization.

For visualization purposes, temporally consecutive measurement locations were connected in order to illustrate the approximate measurement routes. The locations of nearby base stations within the investigated area were also included in the spatial representation. Following preprocessing and data cleaning, the final dataset consisted of 39,333 samples. The reduction in sample size was primarily due to the removal of corrupted records and measurements containing missing or invalid geographic coordinate information. The investigated frequency bands included GSM uplink (880–915 MHz), GSM downlink (950–960 MHz), UMTS uplink (1710–1735 MHz), UMTS downlink (1805–1830 MHz), 5G downlink (791–801 MHz), 5G uplink (832–842 MHz), and 5G TDD (3410–3510 MHz). The measurement campaign was carried out over 15 calendar dates throughout 2023 and 2024.

Table 2 summarizes the number of samples, mean, standard deviation, maximum, and coefficient of variation of the measured electric field for each of the seven frequency-selective bands after preprocessing. All seven bands are represented by an equal number of samples (5,619 each) drawn from the same 39,333-sample dataset, and all bands exhibit a coefficient of variation above 80%, reflecting the highly right-skewed, near-zero-inflated distribution of RF-EMF exposure levels that is typical of in situ spectrum measurements.

Table 2: Per-band descriptive statistics of the measured electric field (V/m) after preprocessing.

Band	N	Mean	Std.	Max	CV
5G_3GHz_TDD	5,619	0.424	0.358	2.902	84.5%
5G_800_dl	5,619	0.266	0.235	2.386	88.7%
5G_800_ul	5,619	0.185	0.159	0.852	85.8%
GSM_dl	5,619	0.414	0.366	3.900	88.4%
GSM_ul	5,619	0.172	0.147	0.805	86.0%
UMTS-LTE_dl	5,619	0.356	0.327	5.106	92.0%
UMTS-LTE_ul	5,619	0.112	0.099	1.659	88.4%

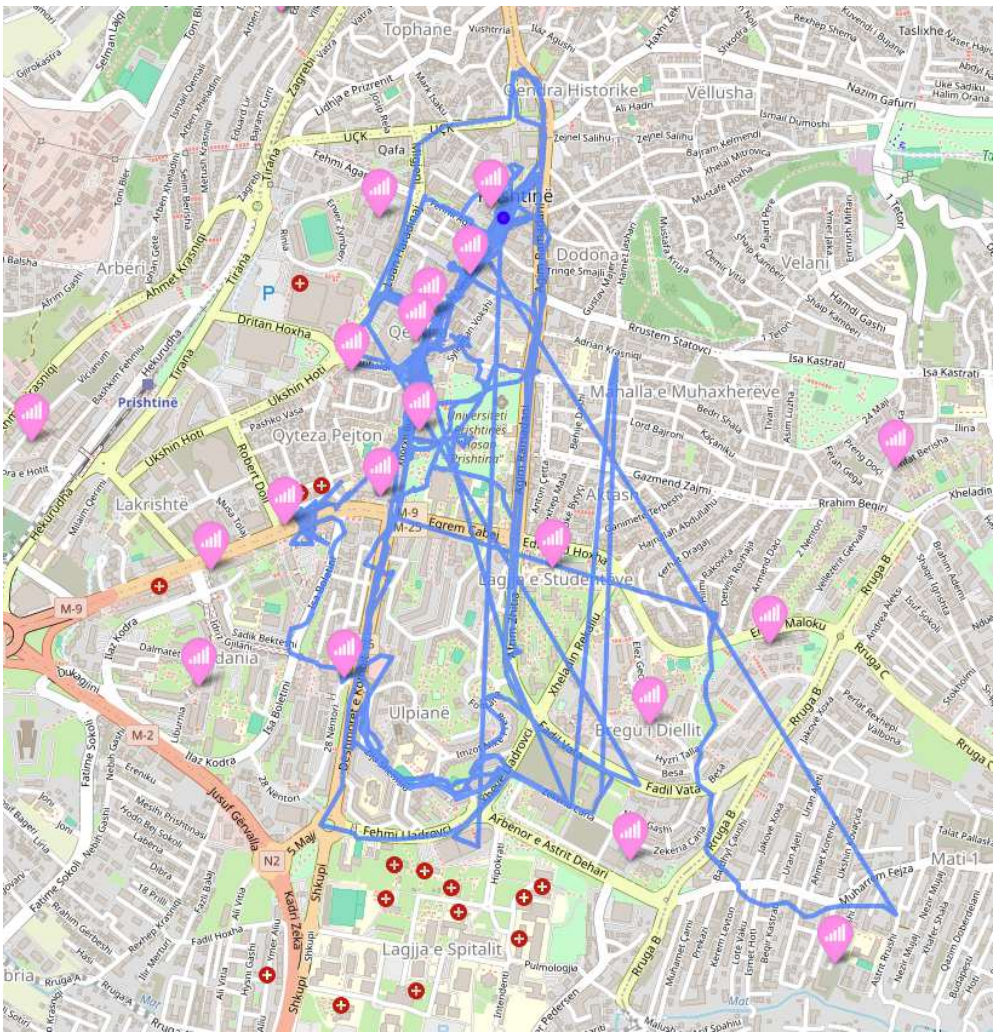


Fig. 2: Measurement map

3.3 Model implementation

To investigate the prediction of frequency-selective RF-EMF electric field levels, a comparative ML framework consisting of multiple regression-based models was implemented. The evaluated models included Ordinary Least Squares (OLS), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), Generalized Regression Neural Network (GRNN), Radial Basis Function Neural Network (RBFNN), Random Forest (RF), Extra Trees, Gradient Boosting, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and CatBoost. The selected models represent different methodological families, including linear regression methods, distance-based learning approaches, neural network architectures,

and ensemble tree-based techniques. This diversity was intended to enable a broader comparison of model behavior under a unified preprocessing and evaluation framework. The principal characteristics of the evaluated algorithms are briefly outlined below.

OLS is a linear regression method that estimates the relationship between predictor variables and the target variable by minimizing the sum of squared residuals between observed and predicted values [16]. KNN regression is a non-parametric method that predicts the target value based on the average response of the k most similar training instances identified according to a distance metric in the feature space [17]. SVR extends the principles of Support Vector Machines to regression problems by identifying a function that approximates the target values within a predefined error margin while maximizing model generalization [18]. MLP is a feed-forward ANN consisting of interconnected layers of neurons capable of modeling complex non-linear relationships through iterative weight optimization using backpropagation [19]. GRNN is a memory-based neural network derived from non-parametric kernel regression that estimates continuous target variables using a probabilistic approach [20]. RBFNN is a neural network architecture employing radial basis activation functions within the hidden layer to approximate non-linear mappings between inputs and outputs [21]. Random Forest is an ensemble learning algorithm that combines multiple decision trees constructed using bootstrap sampling and random feature selection, thereby improving predictive performance and reducing overfitting [22]. Extra Trees is an ensemble tree-based method that introduces additional randomization by selecting split thresholds at random, increasing diversity among individual trees and potentially enhancing generalization [23]. Gradient Boosting is an ensemble approach that sequentially constructs decision trees, where each new tree is trained to minimize the residual errors produced by the preceding ensemble [24]. LightGBM is an efficient gradient boosting framework that employs histogram-based algorithms and leaf-wise tree growth strategies to accelerate training while maintaining predictive accuracy [25]. XGBoost is an optimized implementation of gradient boosting that incorporates regularization, parallel processing, and advanced tree-pruning strategies to improve both efficiency and model performance [26]. CatBoost is a gradient boosting algorithm specifically designed to handle categorical variables effectively through ordered boosting techniques while reducing prediction bias and overfitting [27].

3.3.1 Feature Engineering

For all evaluated models, the same set of input features was used in order to ensure methodological consistency and comparability between models. The selected features were designed to capture temporal and spatial characteristics potentially associated with variations in frequency-selective RF-EMF exposure levels. The target variable in all prediction tasks was the measured electric field (V/m) value obtained during the measurement campaign. Temporal information was incorporated using cyclic time encoding derived from the measurement timestamp. Specifically, the measurement hour was transformed into sine and cosine components to preserve the cyclical nature of time and avoid artificial discontinuities between consecutive hours. In addition, calendar-related information was included through two features: *weekday*, the day of

the week encoded as an integer from 1 (Monday) to 7 (Sunday), and *is_weekend*, a binary indicator set to 1 for Saturday and Sunday and 0 otherwise. *is_weekend* is a deterministic function of *weekday*, so the two features are not statistically independent. Encoding *weekday* as a raw integer also imposes an artificial linear ordering on what is actually a cyclic variable, since Sunday and Monday are calendar-adjacent rather than seven ordinal steps apart. Both features were nonetheless retained. The tree-based ensemble models that dominate this study split on features individually rather than through a shared linear scale, and the feature importance analysis confirmed that both features contribute additional explanatory value (for reference, see Section 4.4).

The measurement campaign includes 15 measurement dates spanning June 2023 to December 2024, over which the exposure level of several bands shifted appreciably. A fifth temporal feature, the number of days elapsed since the first measurement date (*days_since_start*), was therefore added so that the models could separate this longer-term trend from the within-day and within-week cyclic patterns captured by the other temporal features. This shift is consistent with ordinary network growth, though it could also reflect seasonal variation, campaign scheduling effects, or other unmeasured confounds, a point revisited in Section 5. Spatial features were derived from the geographic coordinates of the measurement locations and a database of 20 nearby base station locations identified within the investigated area. These twenty base stations were included because they represented all identified transmitters within the investigated urban area, thereby providing complete spatial coverage of the available mobile communication infrastructure. For each measurement sample, the distance between the measurement location and all base stations was calculated using the Haversine formula [28], which estimates the great-circle distance between two points on the Earth's surface based on latitude and longitude coordinates:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{2}$$

where R represents the Earth's radius, ϕ_1 and ϕ_2 are the latitudes of the two locations, and $\Delta\phi$ and $\Delta\lambda$ denote the latitude and longitude differences, respectively.

Using these calculated distances, several spatial-context features were generated. The distances to the three nearest base stations were included as predictors in order to characterize local network proximity conditions. In addition, the number of nearby base stations within predefined distance thresholds was calculated using radii of 250 m, 500 m, and 1 km around each measurement location. The raw latitude and longitude coordinates themselves were deliberately not included as direct model inputs, since doing so would allow the models to memorize the absolute locations of the fixed set of 20 base stations and the specific measurement grid rather than learning a distance-based relationship with the potential to generalize to previously unseen locations.

To further represent the cumulative influence of surrounding infrastructure, an inverse-distance-weighted proximity score was additionally computed:

$$\text{Proximity Score} = \sum_{i=1}^{N} \frac{1}{d_i + \varepsilon} \tag{3}$$

where d_i represents the distance to the i-th base station, N is the total number of considered base stations, and ε is a small constant introduced to avoid division by zero.

3.3.2 Validation Strategy

The prediction framework was implemented independently for each frequency-selective band and ML model, giving 84 (service, model) pairs in total. For each pair, hyperparameters were re-tuned independently within every individual evaluation fold of the three validation lenses described below: one fold for random hold-out, five for buffered spatial cross-validation, and fifteen for leave-one-date-out, for 21 folds per pair and $84 \times 21 = 1{,}764$ tuned configurations overall. Only each fold's own training partition was used for this tuning, so that no information from a fold's evaluation data could leak into the selection of its hyperparameters. The target variable, electric field strength in V/m, was modeled directly without logarithmic transformation, in order to preserve the original physical interpretation of the measured exposure levels throughout.

Hyperparameter optimization was performed using RandomizedSearchCV with 25 sampled configurations, using five-fold GroupKFold cross-validation over spatial groups. These groups were defined by rounding latitude and longitude coordinates to three decimal places and combining them into a single group identifier, grouping nearby measurements into localized geographic cells with an approximate spatial resolution of 100 m. Spatially grouped folds, rather than a plain random K-fold split, were used for this tuning step to avoid selecting hyperparameters that overfit to spatial autocorrelation between nearby training and validation samples. This inner search was re-run independently within every outer evaluation fold of every validation lens, using only that fold's own training data, so that any differences observed between lenses reflect the evaluation partition itself, rather than a shared hyperparameter configuration carried over from elsewhere in the pipeline. Input features were standardized using StandardScaler for algorithms sensitive to feature scaling (KNN, SVR, MLP, GRNN, and RBFNN), fitted within each cross-validation fold. Tree-based ensemble models and OLS were trained using the original feature values, since these algorithms are insensitive to feature scaling. A fixed random state of 42 was used throughout the experiments to ensure reproducibility and consistency across repeated executions.

The first evaluation lens, **random hold-out**, partitions each frequency-selective service dataset into training and testing subsets using a single 80/20 hold-out split with a fixed random state. This split operates at the individual measurement level, without geographic or temporal constraints, so physically adjacent or temporally correlated samples may appear in both the training and testing sets. This lens therefore reflects an optimistic, localized prediction scenario, in which prior measurements from the target area and time period are available during model development. It is retained in this study as a naive reference point illustrating the effect of such leakage, rather than as a primary generalization result. Because this lens uses a single fixed-seed split rather than repeated resampling, the reported random hold-out metrics reflect one particular partition of the data and do not carry an estimate of their own variability across alternative splits. The buffered spatial and leave-one-date-out lenses below are inherently repeated across multiple folds and do not share this limitation.

The second evaluation lens, **buffered spatial cross-validation**, is the primary spatial-generalization lens and addresses the geographic leakage present in the naive hold-out. Each frequency-selective service dataset, 5,619 samples grouped into 201 ∼100 m spatial cells, was partitioned into five folds over the same spatial groups used for hyperparameter tuning, giving a mean of 1,124 test samples per fold. For each fold, the fold's own spatial groups were withheld from training, and every remaining training sample located within a 100 m buffer distance (haversine) of any test-fold sample was also excluded from that fold's training set. This further reduced residual spatial autocorrelation between nearby cells, beyond what the ∼100 m grouping alone achieves. The buffer is substantial in practice: across the seven bands and five folds, it removed a mean of 2,923 of the 4,495 candidate training samples per fold (65%), leaving a mean of 1,572 training samples per fold. The reported buffered spatial CV metrics therefore reflect models trained on markedly less, and more geographically isolated, data than the random hold-out lens. This evaluation lens assesses the models' ability to extrapolate RF-EMF exposure levels to geographic locations that were not represented, nor immediately adjacent to any representation, in training.

To check directly whether this buffer distance is adequate, an empirical semivariogram was computed for the raw measured electric field values and for Gradient Boosting's out-of-fold residuals, under the same buffered spatial cross-validation folds. The semivariance at distance h is defined as

$$\gamma(h) = \frac{1}{2}\,\mathrm{mean}\left[(z_i - z_j)^2\right] \tag{4}$$

computed over all pairs of measurement locations separated by h. Values were computed per frequency-selective band, across seven distance bins spanning 0 to 3.5 km (the approximate extent of the measurement campaign), and then averaged across bands to obtain the band-balanced curves shown in Figure 3.

The raw-value curve rose from $\gamma \approx 0.050$ $(\mathrm{V/m})^2$ below 50 m to a plateau of $\gamma \approx 0.077$–0.079 $(\mathrm{V/m})^2$ between 100 m and 550 m, confirming substantial spatial dependence at the scale the buffer is meant to address. The residual curve sat below the raw curve throughout, most notably beyond 500 m, where semivariance dropped by roughly one-third, showing that the model captured most of the medium and long-range spatial structure once the buffer was applied. Near the buffer distance itself (50–300 m), residual semivariance remained comparatively high ($\gamma \approx 0.063$–0.070 $(\mathrm{V/m})^2$), only modestly below the raw values. This indicated that the 100 m buffer reduced, but did not fully eliminate, short-range spatial autocorrelation, consistent with the limitation noted below.

The third evaluation lens, **leave-one-date-out cross-validation**, is the primary temporal-generalization lens. It partitions each frequency-selective service dataset into as many folds as there are distinct calendar measurement dates (15), training on all dates except one and testing on the held-out date, in turn for every date. This evaluation lens assesses the models' ability to generalize RF-EMF exposure predictions to a calendar date not seen during training, including the network-growth-driven shift in exposure levels between the 2023 and 2024 measurement eras.

The search spaces explored for each ML algorithm during hyperparameter tuning are summarized in Table 3. For every model except OLS, 25 configurations were

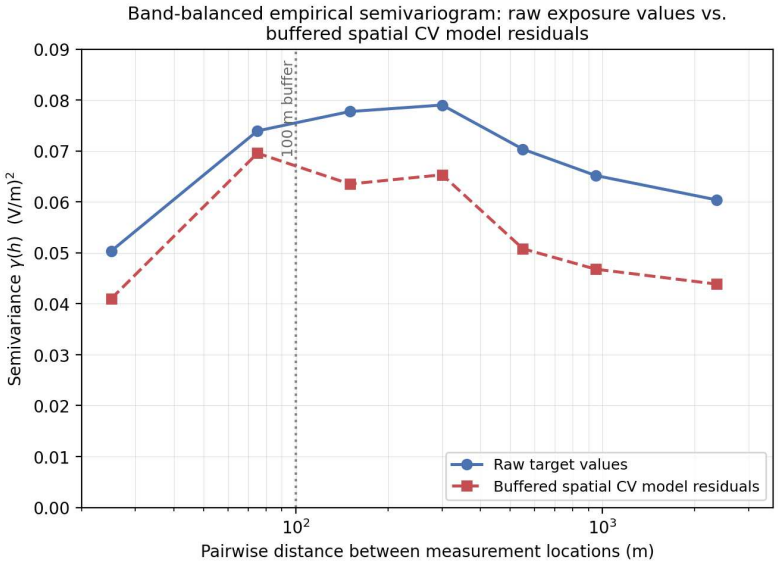


Fig. 3: Band-balanced semivariogram of raw exposure values and buffered spatial CV residuals versus pairwise distance. The dotted line marks the 100 m buffer.

sampled from the listed space via `RandomizedSearchCV` using five-fold `GroupKFold` cross-validation over the ∼100 m spatial cell identifier, with this search repeated independently within every one of the 21 outer evaluation folds per (service, model) pair as described above.

All experiments were performed on a workstation equipped with an 11th Gen Intel Core i5-1135G7 processor (2.40 GHz), 16 GB RAM, running Windows 11 Pro (version 25H2) and Python 3.10.4. Training and prediction times were measured using Python's high-resolution `time.perf_counter()` function, wrapped around only the final tuned model's `fit()` and `predict()` calls for each evaluation fold, and therefore exclude the upstream `RandomizedSearchCV` tuning cost. All models were evaluated under the same computational environment, and the reported execution times are intended for comparative evaluation rather than absolute hardware-independent benchmarking.

3.3.3 Evaluation Metrics

Model performance was assessed using the coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), defined as follows [16]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

Table 3: Hyperparameter search spaces sampled during model optimization.

Model	Search Space
OLS	None (closed-form solution; no hyperparameters)
KNN	n_neighbors $\in \{3, 5, 10, 15, 20, 30, 40, 60\}$; weights $\in$ {uniform, distance}; leaf_size $\in \{20, 30, 50, 80\}$; $p \in \{1, 2\}$
SVR	kernel $\in$ {rbf, linear}; $C \in \{0.1, 1, 3, 10, 30, 100\}$; $\varepsilon \in \{0.01, 0.05, 0.1, 0.2\}$; $\gamma \in$ {scale, auto, 0.01, 0.1, 1}
MLP	hidden_layer_sizes $\in \{(50), (100), (50, 50), (100, 50), (100, 100)\}$; $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$; learning_rate_init $\in \{0.001, 0.005, 0.01\}$; activation $\in$ {relu, tanh}
GRNN	$\sigma \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0\}$
RBFNN	spread $\in \{0.5, 1.0, 1.5, 2.0, 3.0\}$; number of neurons $\in \{10, 20, 40, 60, 100\}$; $\alpha \in \{0.001, 0.01, 0.1, 1.0\}$
RF	n_estimators $\in \{200, 300, 400, 600\}$; max_depth $\in \{4, 6, 8, 12,$None$\}$; min_samples_split $\in \{2, 5, 10\}$; min_samples_leaf $\in \{1, 2, 4, 8\}$; max_features $\in$ {sqrt, log2}
ET	Same search space as RF
GB	n_estimators $\in \{150, 200, 300, 400\}$; learning_rate ($\eta$) $\in \{0.02, 0.03, 0.05, 0.08, 0.1\}$; max_depth $\in \{2, 3, 4, 5\}$; subsample $\in \{0.5, 0.6, 0.7, 0.8, 1.0\}$; max_features $\in$ {sqrt, log2, None}; min_samples_leaf $\in \{1, 2, 4, 8\}$
LightGBM	n_estimators $\in \{150, 200, 300, 400\}$; learning_rate ($\eta$) $\in \{0.02, 0.03, 0.05, 0.08, 0.1\}$; num_leaves $\in \{15, 31, 63, 127\}$; max_depth $\in \{-1, 4, 6, 8\}$; min_child_samples $\in \{10, 20, 50, 100\}$; feature_fraction $\in \{0.6, 0.8, 1.0\}$; bagging_fraction $\in \{0.6, 0.8, 1.0\}$; $\lambda \in \{0, 0.1, 1, 5\}$; $\alpha \in \{0, 0.1, 1\}$
XGBoost	n_estimators $\in \{150, 200, 300, 400\}$; learning_rate ($\eta$) $\in \{0.02, 0.03, 0.05, 0.08, 0.1\}$; max_depth $\in \{2, 3, 4\}$; min_child_weight $\in \{1, 3, 5, 10\}$; $\gamma \in \{0, 0.1, 0.5, 1.0\}$; subsample $\in \{0.6, 0.8, 1.0\}$; colsample_bytree $\in \{0.6, 0.8, 1.0\}$; $\lambda \in \{1, 3, 5, 10\}$
CatBoost	iterations $\in \{300, 500, 700\}$; depth $\in \{4, 5, 6, 8\}$; learning_rate (η) $\in \{0.02, 0.03, 0.05, 0.08, 0.1\}$; $\lambda \in \{3, 5, 7, 10\}$; bagging_temperature $\in \{0, 0.5, 1.0\}$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

where y_i and $\hat{y}_i$ denote the observed and predicted electric field values for sample i, $\bar{y}$ is the mean of the observed values, and n is the number of test samples. No measured electric field value in the cleaned dataset is exactly zero (the minimum observed value is 0.016 V/m), so the MAPE definition above is always numerically defined here. A small constant (10^{-9}) is included in the denominator only as a numerical safeguard. Several measured values nonetheless lie close to zero, and MAPE is disproportionately sensitive to these low-magnitude samples: a small absolute error at a near-zero measurement produces a very large relative error, which inflates the mean MAPE well above what RMSE and MAE alone would suggest. For this reason, MAPE is reported throughout

as a secondary, supplementary metric. The primary comparisons and conclusions in this study rely on RMSE, MAE, R^2, and NRMSE.

For the leave-one-date-out validation strategy, $\mathbf{R^2}$ becomes an unreliable indicator. Each held-out fold contains measurements from a single calendar date and therefore has little internal variance, which can make $\mathbf{R^2}$ highly unstable or degenerate even for an otherwise reasonable model. To enable a meaningful comparison of predictive error across folds under this validation strategy, the Normalized Root Mean Square Error (NRMSE) was used instead:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\bar{y}_{\mathrm{band}}} \tag{9}$$

where $\bar{y}_{\mathrm{band}}$ is the fixed, overall mean of the observed electric field values for the corresponding frequency band across the full dataset, rather than a fold-specific mean, so that NRMSE remains comparable across the individually withheld calendar dates.

Each of the seven frequency-selective bands was modeled and evaluated independently. Every metric above was first computed per fold for a given (band, model) pair, then summarized to a single per-band value, and finally averaged across the seven bands to obtain the *band-balanced* figures reported throughout Section 4. The per-band summary is the median across folds for the random hold-out and buffered spatial CV lenses, and the mean across the fifteen date-folds for leave-one-date-out, so that a single unusually difficult calendar date is reflected in the summary rather than masked by it. The seven resulting per-band values were then averaged with equal weight, rather than pooling all test samples together. This prevents any one band from dominating the aggregated figure and is consistent with the seven bands already containing an equal number of measurement events (Table 2). Computational efficiency was also evaluated, by recording training and prediction times for each model, and permutation-based feature importance analysis was performed for all evaluated algorithms to enhance the interpretability of the proposed framework.

A number of assumptions and simplifications bound the scope of the conclusions drawn in this study. Spatial predictors are computed relative to a fixed database of 20 base station locations identified within the investigated urban area. Transmitters outside this inventory, or subsequent changes to the network deployment, are not represented. Propagation is represented solely through the distance to each base station. All measurements were collected at a fixed height of 1.5 m above ground level, corresponding to a pedestrian exposure scenario rather than a full 3D exposure volume. The measurement campaign spans 15 distinct calendar dates between June 2023 and December 2024. The buffered spatial cross-validation strategy separates training and testing data at an approximate 100 m grid-cell resolution, with an additional 100 m exclusion buffer removing any remaining training points within that distance of a test-fold point. This resolution and buffer substantially reduce, but may not fully eliminate, spatial autocorrelation between nearby cells, as confirmed empirically by the residual semivariogram presented in Section 3.3.2. The spectrum analyzer also records aggregate received power within each service's frequency band rather than resolving individual 5G beams, so the 5G measurements reported here characterize overall band-level exposure and do not distinguish between synchronization-signal and traffic-channel beamforming components. Finally, the Friedman and post-hoc tests reported

in Section 4 treat the seven frequency bands as the test blocks. This is a small and not fully independent sample, since all seven bands are derived from the same underlying measurement events and share the same 15-date, 20-base-station campaign. This limits statistical power, particularly for the leave-one-date-out lens, where the overall Friedman effect is already comparatively modest.

4 Results

4.1 Overall Model Performance

The overall predictive performance of the evaluated ML models is first compared by examining the mean RMSE, MAE, R^2, and MAPE values obtained under the random hold-out and buffered spatial validation strategies. The third, leave-one-date-out, lens is examined separately below, as its error behavior is not directly comparable on the same axes. The mean performance corresponds to the average performance across all frequency-selective bands. Figure 4 presents the comparison of mean RMSE and mean MAE, and Figure 5 presents the comparison of mean R^2 and mean MAPE, for all evaluated models and validation strategies.

Regarding RMSE, a broadly similar ranking trend was observed across both validation strategies, dominated in both cases by the ensemble tree-based models. Overall, random hold-out validation yielded lower prediction errors, with a mean RMSE (averaged across all twelve models) that was 0.060 V/m lower than that obtained using buffered spatial validation (0.131 V/m versus 0.190 V/m). Under the random hold-out validation framework, Random Forest achieved the lowest mean RMSE (0.112 V/m), closely followed by Gradient Boosting (0.113 V/m) and LightGBM (0.113 V/m), whereas OLS exhibited the highest error (0.193 V/m). In contrast, under buffered spatial validation, CatBoost achieved the lowest mean RMSE (0.165 V/m), followed by XGBoost (0.172 V/m) and Extra Trees (0.178 V/m), while OLS again demonstrated the highest error (0.228 V/m). The largest increase in prediction error between the two validation strategies was observed for MLP, whose mean RMSE increased by 0.074 V/m under buffered spatial validation, closely followed by Random Forest and Gradient Boosting. The mean MAE followed the same overall trend, rising from 0.070 V/m under random hold-out validation to 0.126 V/m under buffered spatial validation. Random Forest recorded the lowest MAE under random hold-out (0.055 V/m), while CatBoost achieved the lowest MAE under buffered spatial validation (0.104 V/m).

Figure 5 compares the mean R^2 values obtained using random hold-out and buffered spatial validation across all evaluated models. Under the random hold-out validation framework, Random Forest achieved the highest predictive performance (R^2 = 0.765), followed closely by Gradient Boosting (R^2 = 0.757) and CatBoost (R^2 = 0.756), whereas OLS exhibited the lowest explanatory capability (R^2 = 0.392). The corresponding MAPE values broadly mirror this ranking, with Random Forest reaching the lowest mean MAPE (24.8%), followed by LightGBM (25.0%) and CatBoost (27.7%), while OLS reached a mean MAPE close to 100% (99.96%), consistent with its markedly lower R^2.

A similar top group, though with a different leader, emerged under the buffered spatial validation strategy. CatBoost achieved the highest mean R^2 (0.529), followed by

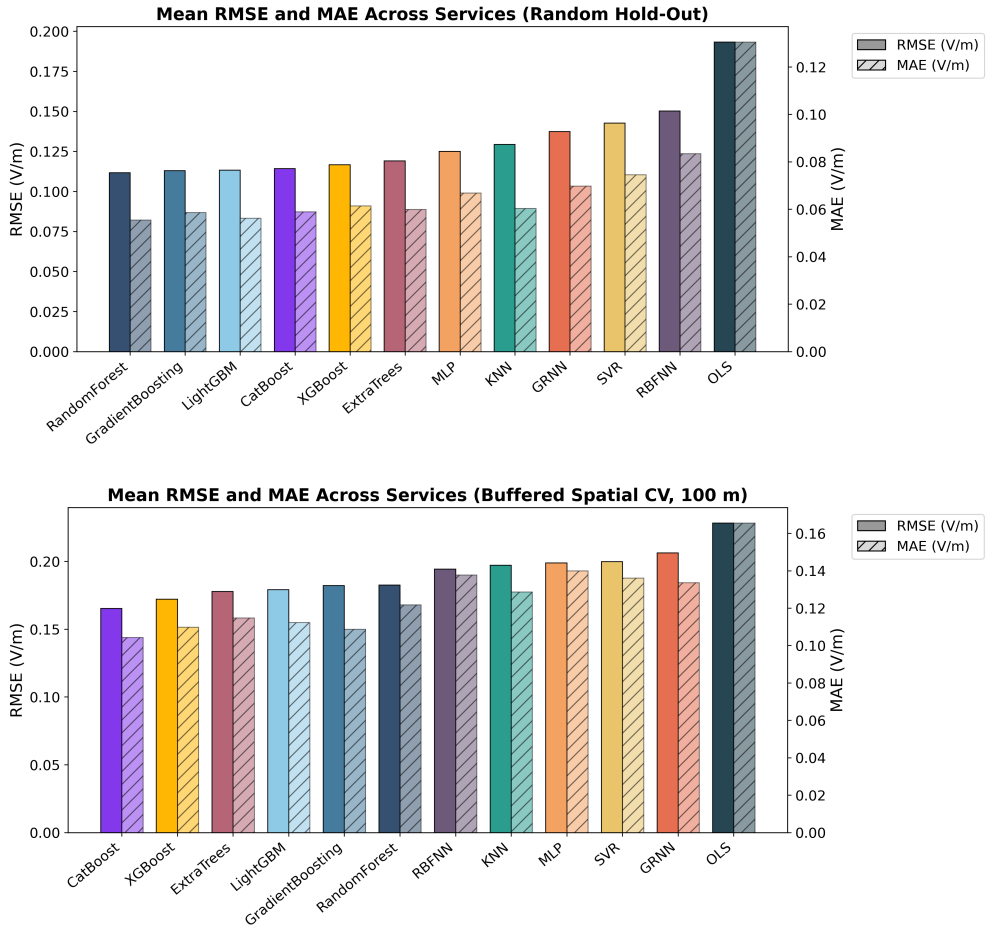


Fig. 4: Mean RMSE and MAE of the evaluated ML models under random hold-out (top) and buffered spatial CV, 100 m (bottom) validation.

XGBoost ($R^2 = 0.505$) and Gradient Boosting ($R^2 = 0.489$). Extra Trees and Random Forest, which ranked among the leading models under random hold-out validation, fell to a second tier ($R^2 = 0.460$ and 0.463, respectively) under buffered spatial validation, while remaining well ahead of KNN, RBFNN, MLP, and OLS. This shift in ranking highlights the influence of the validation strategy on model assessment and suggests that models performing well under random hold-out validation do not necessarily generalize equally well to previously unseen geographic regions. The mean MAPE follows a broadly similar pattern, with LightGBM attaining the lowest value (60.7%), closely followed by CatBoost (61.9%) and Gradient Boosting (66.1%), while OLS reached a mean MAPE of 127.8%.

Overall, the mean R^2 across all evaluated models decreased from 0.689 under random hold-out validation to 0.400 under buffered spatial validation, a decline mirrored

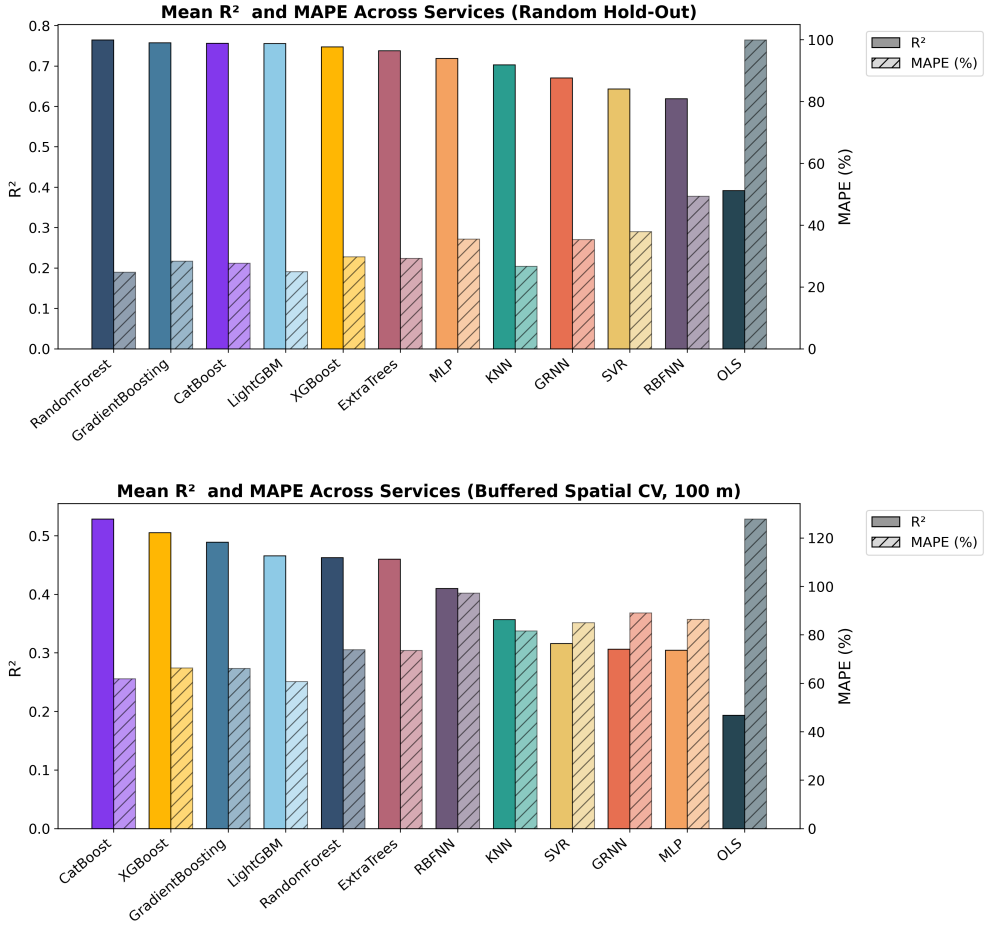


Fig. 5: Mean R^2 and MAPE of the evaluated ML models under random hold-out (top) and buffered spatial CV, 100 m (bottom) validation.

by the mean MAPE, which rose from 37.5% to 80.8% across the two strategies, indicating that geographically independent validation provides a more conservative estimate of model generalization.

In addition to predictive performance, computational efficiency was assessed by analyzing the relationship between model accuracy and mean fit-and-predict time. As described in Section 3.3.2, this measures only the final tuned model's `fit()` and `predict()` calls for each evaluation fold and therefore excludes the upstream hyperparameter search cost, providing a comparative measure of each model's deployment-time computational cost.

Figure 6 presents the relationship between mean fit-and-predict time and mean RMSE under random hold-out (top) and buffered spatial (bottom) validation.

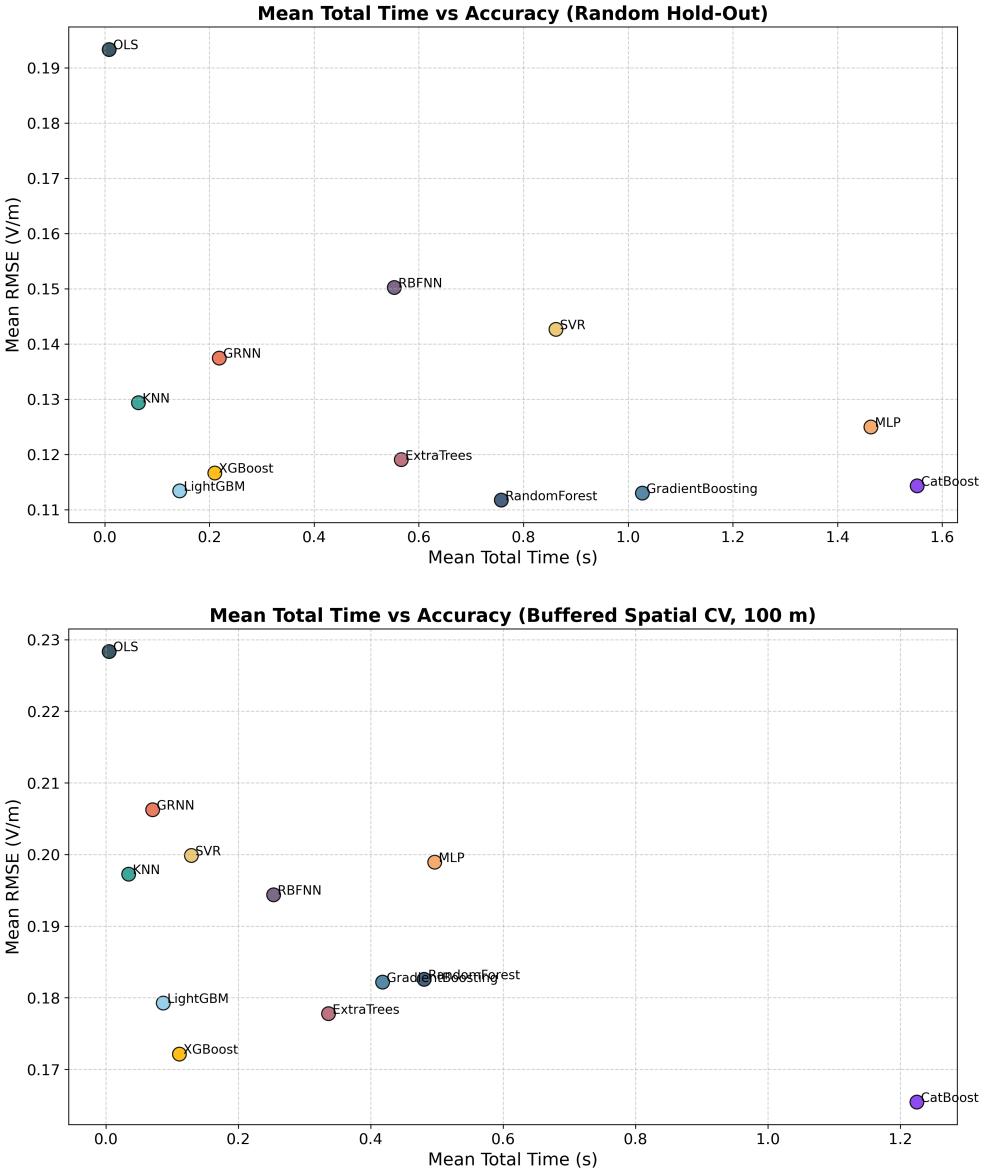


Fig. 6: Mean fit-and-predict time versus RMSE for the evaluated machine learning models under random hold-out (top) and buffered spatial CV, 100 m (bottom) validation.

As shown in Figure 6 (top), Random Forest achieved the lowest mean RMSE (0.112 V/m) at a moderate fit-and-predict cost (0.76 s, averaged per band). CatBoost achieved a closely comparable accuracy (0.114 V/m) but required the highest fit-and-predict time among all evaluated models (1.55 s). In contrast, LightGBM and XGBoost

achieved competitive predictive accuracy (0.113 V/m and 0.117 V/m, respectively) with substantially shorter fit-and-predict times (0.14 s and 0.21 s). MLP and SVR also incurred a comparatively high computational cost (1.46 s and 0.86 s, respectively) without a corresponding improvement in predictive performance, whereas OLS and KNN achieved the shortest fit-and-predict times but, particularly for OLS, had among the lowest predictive accuracy.

A similar pattern was observed under buffered spatial validation (Figure 6, bottom). LightGBM and XGBoost again provided the most favorable balance between predictive performance and computational efficiency, achieving RMSE values of 0.179 V/m and 0.172 V/m with fit-and-predict times of 0.09 s and 0.11 s, respectively. CatBoost achieved the lowest RMSE overall (0.165 V/m) but required by far the highest fit-and-predict time among the leading models (1.23 s). Gradient Boosting offered a middle ground, reaching an RMSE of 0.182 V/m at a moderate cost (0.42 s). OLS and KNN remained the fastest but least accurate models. Overall, the results demonstrate that improvements in predictive performance are not necessarily associated with increased computational demands. Although CatBoost and Random Forest occasionally achieved the lowest prediction errors, Gradient Boosting, XGBoost, and LightGBM consistently offered competitive predictive performance at substantially lower computational costs. Consequently, these boosting-based approaches emerged as the most balanced solutions across both validation strategies, providing an effective compromise between prediction accuracy and execution efficiency for frequency-selective RF-EMF prediction.

4.2 Frequency-Selective Prediction Performance

While the previous section summarized the mean predictive performance across all frequency-selective bands, the following analysis examines model performance for each individual frequency band under both validation strategies. This frequency-selective evaluation provides insight into band-specific variations that are not apparent from aggregated performance measures.

Figure 7 (top) presents the RMSE and MAE heatmaps obtained under the random hold-out validation strategy. CatBoost, Gradient Boosting, and LightGBM consistently ranked among the best-performing models across most frequency bands. The lowest RMSE values were achieved by Gradient Boosting for the GSM uplink band (0.019 V/m), by CatBoost for the 5G 800 uplink band (0.030 V/m), and by LightGBM for the 5G 3 GHz TDD band (0.070 V/m) and the UMTS-LTE uplink band (0.061 V/m), with the remaining leading tree-ensemble models close behind in each case. In contrast, the GSM downlink and UMTS-LTE downlink bands exhibited the highest prediction errors, with even the best-performing models, Random Forest and CatBoost respectively, reaching RMSE values of 0.225 V/m and 0.219 V/m. The corresponding MAE values followed a similar pattern, with LightGBM recording the lowest MAE for the GSM uplink band (0.011 V/m), the 5G 3 GHz TDD band (0.027 V/m), and the UMTS-LTE uplink band (0.010 V/m), and CatBoost recording the lowest MAE for the 5G 800 uplink band (0.013 V/m). The GSM downlink and UMTS-LTE downlink bands again showed the largest errors, with the best-performing model reaching MAE values of 0.126 V/m (Random Forest) and 0.129 V/m (CatBoost), respectively.

Figure 7 (bottom) presents the corresponding RMSE and MAE heatmaps under the buffered spatial validation strategy. Although several shifts in model ranking were observed relative to random hold-out, CatBoost remained the strongest performer for most frequency bands. Notably, the two downlink bands were best predicted by non-tree-based models under this stricter lens: KNN achieved the lowest RMSE for the 5G 800 downlink band (0.193 V/m) and RBFNN for the UMTS-LTE downlink band (0.344 V/m), narrowly ahead of the leading tree-ensemble models in both cases. The corresponding MAE values told a somewhat different story, with CatBoost recording the lowest MAE for the 5G 800 downlink band (0.123 V/m) and Extra Trees the lowest for the UMTS-LTE downlink band (0.222 V/m), indicating that the non-tree models' RMSE advantage in these two bands does not carry over to the mean absolute error.

Overall, the band-specific analysis revealed substantially lower prediction errors than the aggregated mean performance analysis for several of the individual frequency bands, particularly the uplink bands, indicating that these bands can be modeled more accurately when their distinct propagation and utilization characteristics are considered separately. Furthermore, downlink frequency bands consistently exhibited higher prediction errors than their corresponding uplink bands under both validation strategies, an effect that becomes considerably more pronounced under buffered spatial validation, as discussed further below.

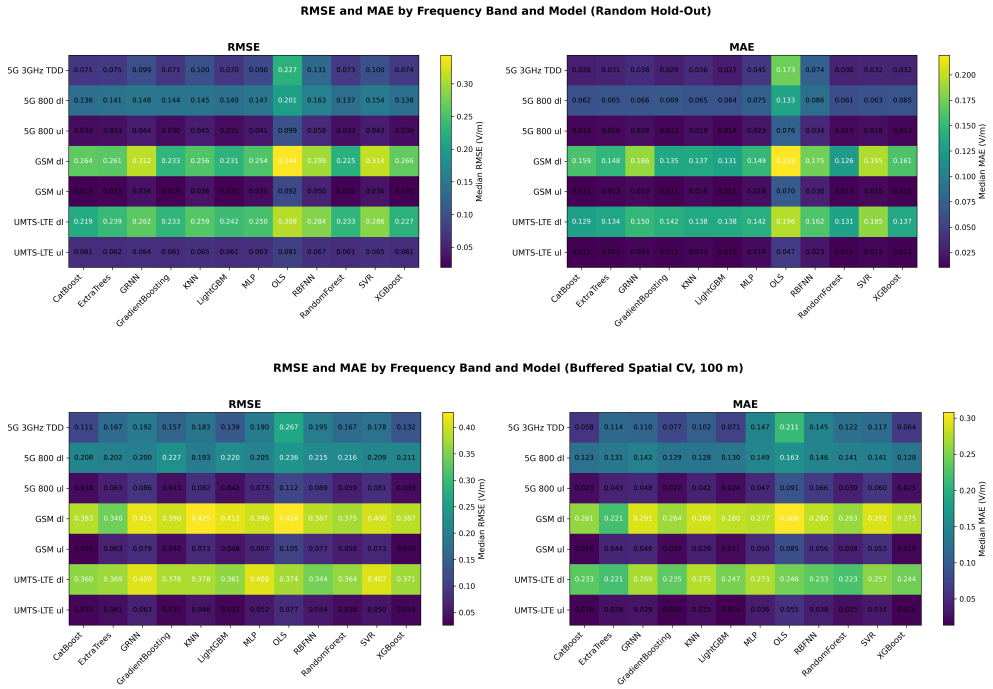


Fig. 7: Heatmap of RMSE and MAE by frequency band and machine learning model under random hold-out (top) and buffered spatial CV, 100 m (bottom) validation.

Figure 8 compares the R^2 values and MAPE across frequency bands and ML models under the random hold-out validation strategy. Consistent with the RMSE analysis, Gradient Boosting, LightGBM, and CatBoost were the dominant performers across most frequency bands. Gradient Boosting achieved the highest R^2 value for the GSM uplink band ($R^2 = 0.983$), LightGBM led for the 5G 3 GHz TDD ($R^2 = 0.961$) and UMTS-LTE uplink ($R^2 = 0.679$) bands, and CatBoost performed best for the 5G 800 uplink ($R^2 = 0.964$), 5G 800 downlink ($R^2 = 0.669$), and UMTS-LTE downlink ($R^2 = 0.534$) bands, with Random Forest leading for the GSM downlink band ($R^2 = 0.642$). Several frequency bands achieved R^2 values exceeding 0.95, indicating substantially better predictive performance than suggested by the aggregated mean results. Nevertheless, the downlink frequency bands consistently exhibited lower explanatory power than their uplink counterparts, with best-model R^2 for the downlink bands ranging from only 0.53 to 0.67. The corresponding MAPE values did not always track this R^2-based ranking: LightGBM reached the lowest MAPE for the GSM uplink (7.7%), 5G 3 GHz TDD (10.9%), UMTS-LTE uplink (11.1%), and 5G 800 uplink (10.2%) bands, whereas for the two weakest downlink bands the lowest MAPE was instead achieved by SVR for the 5G 800 downlink band (28.7%) and by KNN for the UMTS-LTE downlink band (47.1%), and Random Forest reached the lowest MAPE (40.2%) for the GSM downlink band. As with R^2, the downlink bands exhibited substantially higher MAPE than the uplink bands, confirming that the percentage-based error also degrades where explanatory power is weakest, and illustrating that the model maximizing R^2 for a given band is not always the model minimizing its percentage error.

Figure 9 presents the corresponding R^2 values and MAPE obtained using the buffered spatial validation strategy. CatBoost now led every uplink band, achieving the highest R^2 for the GSM uplink ($R^2 = 0.970$), UMTS-LTE uplink ($R^2 = 0.885$), 5G 800 uplink ($R^2 = 0.938$), and 5G 3 GHz TDD ($R^2 = 0.901$) bands, and it also achieved the best result among the downlink bands for the 5G 800 downlink band ($R^2 = 0.259$). The two remaining downlink bands, already the weakest under random hold-out validation, deteriorated substantially further under buffered spatial validation: for the GSM downlink band, the best-performing model (Gradient Boosting) achieved only a weakly positive R^2 (0.048), and for the UMTS-LTE downlink band the best-performing model (RBFNN) achieved an R^2 only marginally below zero (-0.005). This indicates that, once geographic leakage between training and test locations is removed, no evaluated model explains appreciably more variance than simply predicting the mean for these two downlink bands, a substantially more severe limitation than the aggregated mean R^2 alone would suggest. The corresponding MAPE values diverged from the R^2-based ranking for several bands: Gradient Boosting reached the lowest MAPE for the GSM uplink (14.1%) and 5G 800 uplink (17.8%) bands, and LightGBM reached the lowest MAPE for the UMTS-LTE uplink (19.3%) and 5G 3 GHz TDD (23.7%) bands, even though CatBoost achieved the best R^2 for all four of these bands. For the GSM downlink and UMTS-LTE downlink bands, MAPE remained very high for every model, with the lowest values reaching 107.3% (Extra Trees) and 103.5% (SVR) respectively, underscoring the practical unreliability of the frequency-selective predictions for these two bands under buffered spatial validation.

Overall, the frequency-selective analysis revealed that individual frequency bands can be predicted considerably more accurately than suggested by the aggregated mean performance metrics for the uplink bands. However, the consistently lower predictive performance observed for downlink services under both validation strategies, and the near-zero or slightly negative best-model R^2 observed for the GSM and UMTS-LTE downlink bands under buffered spatial validation specifically, indicates that these bands represent an inherently more challenging, and in part currently unsolved, prediction task with the available feature set.

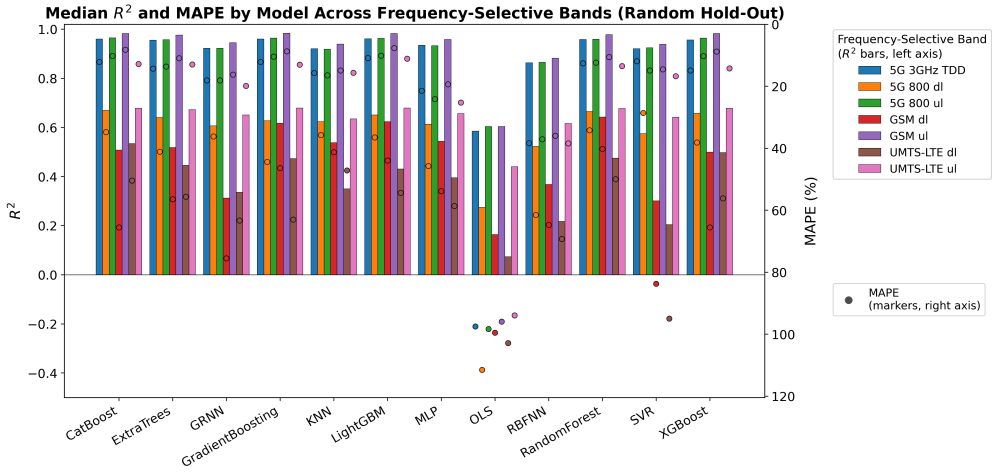


Fig. 8: R^2 and MAPE of the evaluated ML models across frequency bands under the random hold-out validation strategy.

Although the RMSE and R^2 values suggest comparable predictive performance among several ML models, statistical analysis was performed to determine whether the observed ranking differences across the seven frequency-selective bands were statistically significant.

To determine whether the observed differences in predictive performance among the evaluated ML models were statistically significant, Friedman tests were performed separately for the random hold-out and buffered spatial validation strategies using the RMSE values obtained across the seven frequency bands. The results revealed statistically significant differences among the evaluated models under both the random hold-out validation strategy ($\chi^2 = 65.29$, $p < 0.001$) and the buffered spatial validation strategy ($\chi^2 = 38.14$, $p < 0.001$), confirming that model selection has a significant influence on prediction accuracy. Pairwise comparisons were subsequently conducted using the Conover post-hoc procedure with Holm correction for multiple comparisons. Under random hold-out validation, the post-hoc analysis showed that CatBoost, Gradient Boosting, LightGBM, Random Forest, and XGBoost did not differ significantly from one another ($p > 0.05$), and each significantly outperformed GRNN, KNN, RBFNN, SVR, and the OLS baseline ($p < 0.05$). Extra Trees, despite a lower average

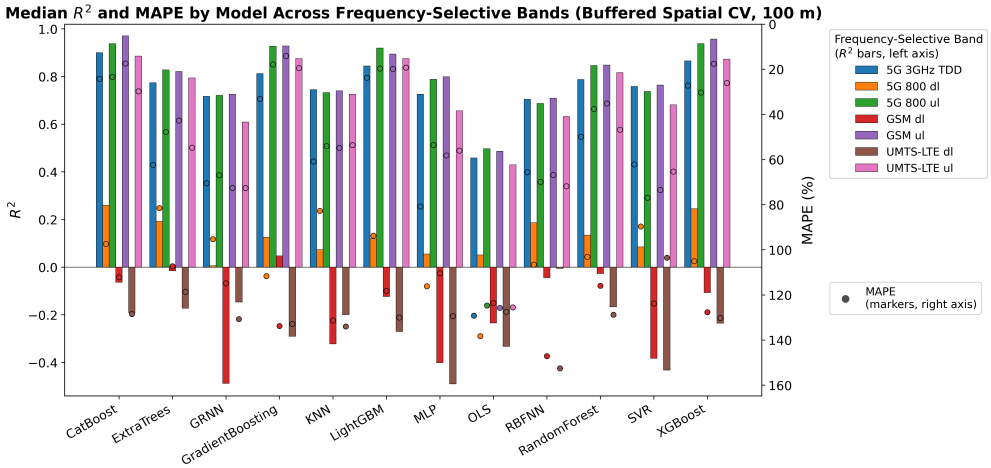


Fig. 9: R^2 and MAPE of the evaluated ML models across frequency bands under the buffered spatial CV, 100 m validation strategy.

rank than these five, was likewise not statistically distinguishable from any of them, nor from MLP, GRNN, or KNN, separating significantly only from OLS, RBFNN, and SVR among the weaker-performing models. CatBoost and Gradient Boosting were the only two of the five leading models to separate significantly from MLP ($p = 0.016$ and $p = 0.034$, respectively), while LightGBM, Random Forest, and XGBoost did not (p between 0.070 and 0.453). Under buffered spatial validation, the same six leading tree-ensemble models (CatBoost, XGBoost, Extra Trees, LightGBM, Random Forest, and Gradient Boosting) remained statistically indistinguishable from one another ($p > 0.05$ for every pairwise comparison), but their separation from the remaining, weaker-performing models narrowed considerably relative to random hold-out. CatBoost was the only one of the six to separate significantly from every weaker-performing model (KNN, MLP, RBFNN, SVR, GRNN, and OLS, all $p < 0.05$). XGBoost separated significantly from RBFNN, SVR, GRNN, and OLS, but not from KNN or MLP. Extra Trees, LightGBM, and Random Forest each separated significantly only from GRNN and OLS, and Gradient Boosting separated significantly only from OLS, remaining statistically indistinguishable from KNN, MLP, RBFNN, SVR, and even GRNN. The corresponding Critical Difference (CD) diagrams presented in Figure 10 summarize the average model rankings obtained from the Friedman analysis. Under the random hold-out validation strategy, CatBoost achieved the best (lowest) average rank (2.57), followed by Gradient Boosting (2.86) and LightGBM (3.14). Under the buffered spatial validation strategy, CatBoost again achieved the best average rank (2.29), followed by XGBoost (4.00) and Extra Trees (4.57). Although the Friedman test remained statistically significant under both validation strategies, the lower Friedman statistic obtained for the buffered spatial validation, together with the narrower separation from the weaker-performing models described above, indicates that the practical statistical advantage of the leading ensemble methods became considerably less pronounced when predicting RF-EMF exposure in previously unseen geographic regions, even

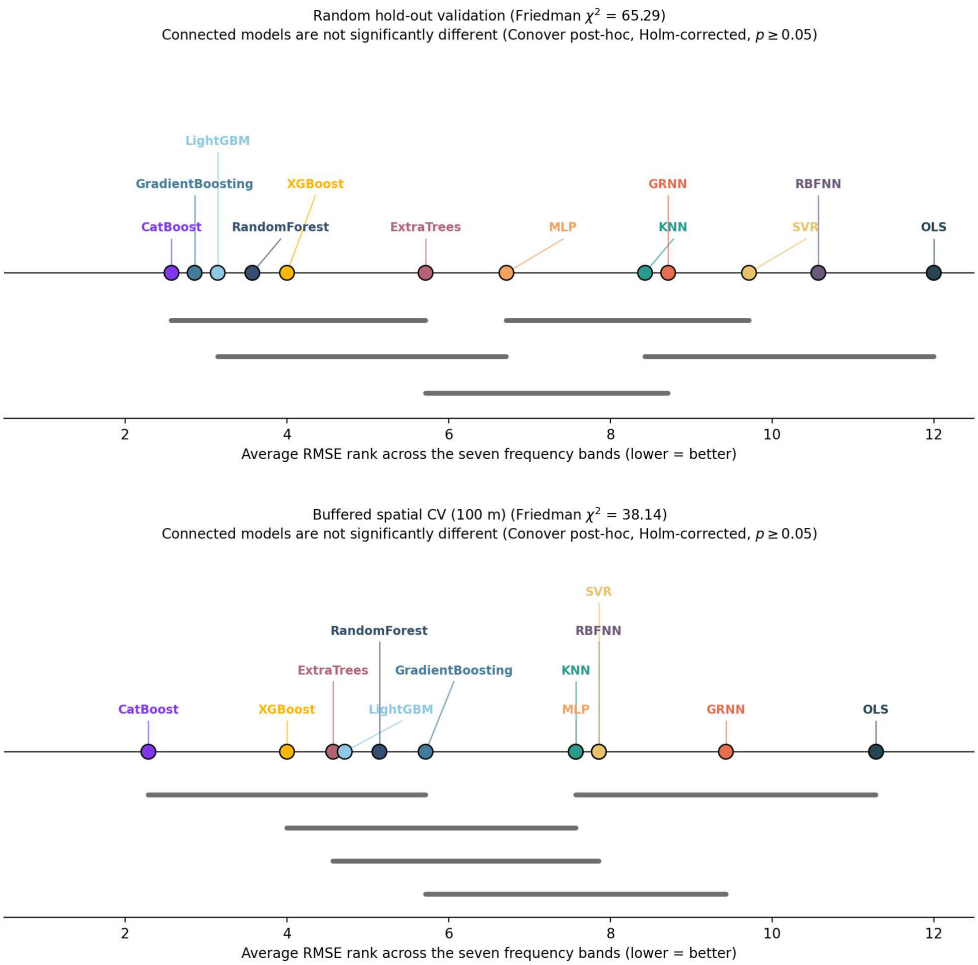


Fig. 10: Critical Difference diagrams (Conover post-hoc test with Holm correction) showing the average RMSE ranks of the evaluated ML models across the seven frequency bands under random hold-out (top) and buffered spatial CV, 100 m (bottom) validation. Models connected by a horizontal bar are not statistically distinguishable ($p \geq 0.05$).

though their relative ranking among one another was preserved. These findings further support the conclusion that ensemble tree-based algorithms consistently provide the most reliable predictive performance, with CatBoost achieving the best average rank under both lenses in this analysis, while the precise boundary of which models are statistically distinguishable from it shifts between validation strategies.

4.3 Temporal Generalization: Leave-One-Date-Out Validation

The random hold-out and buffered spatial validation strategies examined above cover two of the three evaluation lenses introduced in Section 3.3.2. This subsection reports the results of the third, leave-one-date-out lens, in which each of the 15 calendar measurement dates is withheld in turn as the test set while the models are trained on the remaining 14 dates. It also explains why this lens needs to be interpreted differently from the two preceding ones.

Evaluated directly on R^2, the leave-one-date-out results appear severely degraded: every model obtains a strongly negative band-balanced R^2, in most cases between -10 and -30. Taken at face value, this would suggest that none of the evaluated models generalize to an unseen calendar date at all. This is largely an artifact of the metric rather than a fair reflection of prediction quality. R^2 is computed relative to the variance of the target values within the test fold itself, and a single calendar date has comparatively little internal variance, particularly for the low-magnitude uplink bands. A modest, physically plausible shift in that day's average exposure level therefore produces a small absolute error but an extremely large negative R^2, since the denominator against which the error is judged is itself very small. This is a known property of R^2 on low-variance evaluation subsets, and the strongly negative values should not, on their own, be read as evidence that the models' absolute predictions are wildly inaccurate. Temporal generalization is nonetheless genuinely limited: band-balanced NRMSE under leave-one-date-out ranges from 0.82 to 0.96 across all twelve models (Table 4), meaning the typical prediction error is 82–96% of the band's own mean exposure level, well above the error under random hold-out (0.39–0.68) or buffered spatial validation (0.52–0.79).

To obtain a fair, scale-independent comparison across the three validation lenses, band-balanced NRMSE (as defined in Section 3.3.3) is used instead of R^2. Table 4 reports this NRMSE under all three validation strategies for all twelve models, together with each model's average Friedman rank under leave-one-date-out specifically, sorted by that rank.

Two findings emerge from Table 4. First, relative prediction error increases consistently from random hold-out, to spatial CV, to leave-one-date-out, for every model: leave-date-out NRMSE is roughly 1.1–1.7 times higher than spatial-CV NRMSE, confirming a genuine additional generalization gap specific to unseen calendar dates, distinct from the geographic gap observed under buffered spatial validation. Second, model ranking changes markedly under this third lens, and average rank and absolute NRMSE do not agree on which model is best. GRNN, a comparatively weak performer under the other two lenses, achieves the best average Friedman rank here (3.71), while CatBoost, the leading model elsewhere, ranks among the weakest (8.29). GRNN's per-band ranks explain this: it ranks second or third of twelve models in five of the seven bands, but drops to eighth in the UMTS-LTE downlink band, an inconsistency that inflates its band-balanced NRMSE (0.916, the highest of the twelve models) even though its rank-based performance is the most consistent overall. XGBoost, by contrast, achieves the lowest absolute NRMSE (0.817) despite a more modest average rank (4.86). OLS, consistently the weakest model under the other two lenses, is

Table 4: Band-balanced NRMSE (RMSE normalized by the mean observed target value) under the three validation strategies, and each model's average Friedman rank under leave-one-date-out validation (lower rank is better, 12 models compared), sorted by leave-date-out rank.

Model	NRMSE (random hold-out)	NRMSE (spatial CV)	NRMSE (leave-date-out)	Avg. rank (leave-date-out)
GRNN	0.470	0.693	0.916	3.71
OLS	0.684	0.787	0.840	4.14
XGBoost	0.401	0.540	0.817	4.86
LightGBM	0.394	0.566	0.867	5.14
Extra Trees	0.411	0.584	0.822	5.86
RBFNN	0.521	0.655	0.833	6.71
Gradient Boosting	0.393	0.571	0.876	6.71
KNN	0.452	0.648	0.961	7.14
Random Forest	0.391	0.591	0.841	7.57
MLP	0.435	0.654	0.884	7.86
CatBoost	0.394	0.518	0.863	8.29
SVR	0.486	0.662	0.948	10.00

also competitive here on both criteria (rank 4.14, NRMSE 0.840). CatBoost, Gradient Boosting, and Random Forest show the largest relative increase in NRMSE from spatial CV to leave-date-out (factors of 1.67, 1.53, and 1.42), while OLS (1.07) and GRNN (1.32) degrade least. GRNN, OLS, and XGBoost are all inexpensive to run: GRNN's band-balanced fit-and-predict time under leave-one-date-out is only 0.09 s, comparable to OLS and XGBoost and well below the ensembles evaluated here.

A Friedman test on the per-band RMSE ranks under leave-one-date-out validation indicated a statistically significant overall difference among models ($\chi^2 = 20.03$, p = 0.045, versus $p < 0.001$ for the other two lenses). Only a small number of pairwise Conover post-hoc comparisons with Holm correction reached significance, but several are relevant to model selection: GRNN, the best-ranked model, differed significantly from CatBoost (p = 0.007) and from MLP (p = 0.025), and SVR, the worst-ranked model, differed significantly from GRNN, OLS, XGBoost, LightGBM, and Extra Trees ($p \leq 0.025$ in each case). CatBoost also differed significantly from OLS (p = 0.025). This pattern is consistent with a genuine, though modest and unevenly distributed, model-selection effect. Statistical power here is limited by the small number of independent blocks: seven frequency bands, as in the other two lenses, but combined with only 15 temporal folds rather than continuous resampling. Model selection for scenarios that must generalize to an excluded measurement date should therefore not rely on the ranking obtained from either of the other two validation lenses. Because rank-based and error-based criteria disagree on which model is best, and because temporal errors remained high in absolute terms for every model, no single model can presently be recommended conclusively for date-independent generalization. The choice among GRNN, OLS, and XGBoost should instead depend on whether cross-band consistency or the lowest expected absolute error matters more for the intended application.

4.4 Feature Importance Analysis

In addition to evaluating predictive performance, feature importance analysis was conducted to identify which input variables had the greatest influence on the model outputs. Permutation importance was calculated for all evaluated models under both validation strategies.

Permutation importance was used to evaluate the relative importance of the input variables. For each feature, its values were randomly permuted in the validation dataset, and the resulting change in model performance was recorded, using the negative root mean square error as the scoring function (so importance is expressed as the increase in RMSE caused by permuting a feature). Each permutation was repeated 10 times per fold, with the same fixed random state (42) used throughout this study, and the mean and standard deviation of the resulting importance values recorded. These per-fold values were averaged first across the outer folds within each frequency-selective band, and then across the seven bands with equal weight per band, to obtain the single per-model, per-feature importance value reported in Figures 11 and 12. Because permutation importance is computed independently for every outer fold using only that fold's own held-out data, it inherits the same nested, leakage-free evaluation structure used for the accuracy metrics elsewhere in this study. A larger increase in error under permutation indicated a greater contribution of that feature to the model's predictions. This is a measure of *predictive* importance specific to each fitted model and should not be interpreted as evidence that a feature physically causes variation in RF-EMF exposure.

Figure 11 presents the mean feature importance values obtained across all evaluated models for both validation strategies. Under the random hold-out validation framework, the weekend indicator was the most influential predictor, with a mean importance of 0.068, closely followed by the days-since-start trend feature (0.065) and the weekday indicator (0.063), with the cyclic hour component (hour sine) ranking fourth overall (0.054). Among the spatial predictors, the distances to the second and third nearest base stations emerged as the most relevant features, followed by the number of base stations within 1 km, all considerably below the leading temporal predictors.

A similar pattern was observed under the buffered spatial validation strategy, with the same four temporal predictors retaining the highest relative importance, although their importance values were generally lower. The weekend indicator remained the dominant feature (0.050), followed by the days-since-start feature (0.046), hour sine (0.040), and the weekday indicator (0.038). The distances to the second and third nearest base stations remained the most influential spatial predictors, although their mean importance values dropped further, to 0.0058 and 0.0021, respectively.

The consistency of the feature rankings across both validation strategies indicates that temporal predictors, particularly the weekend indicator, the days-since-start trend feature, and the weekday indicator, contributed most strongly to model performance, more so than the cyclic hour-of-day encoding itself, while base station proximity remained the most influential spatial factor. This pronounced reliance on calendar-position features rather than on within-day timing alone raises the question of whether the models are capturing a genuine temporal pattern in RF-EMF exposure

or are instead exploiting the measurement campaign's calendar coverage, a question examined directly in Section 4.5 below.

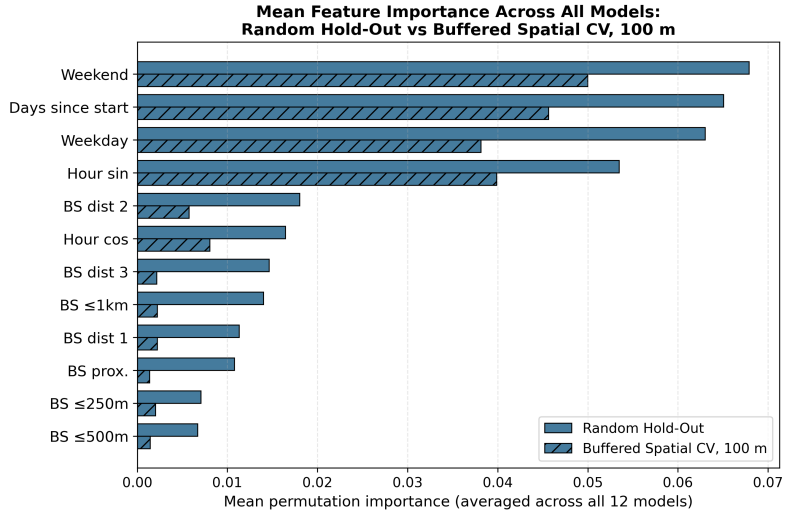


Fig. 11: Mean feature importance averaged across all twelve ML models, under random hold-out and buffered spatial CV, 100 m validation.

Figure 12 presents the five most influential predictors identified for each model under both validation strategies. Unlike Figure 11, which summarizes the mean feature importance across all evaluated models, this figure highlights the predictors driving the behavior of individual ML algorithms. The following discussion focuses on Extra Trees, CatBoost, and XGBoost, as these models demonstrated the strongest predictive performance.

For all three models, temporal variables dominated the feature importance rankings, although which specific temporal predictor ranked first varied by model: the weekend indicator for Extra Trees, the days-since-start trend feature for CatBoost, and the weekday indicator for XGBoost. Weekday, weekend, and days-since-start indicators together consistently ranked among the most important features for all three models, emphasizing the influence of temporal and calendar-position variability on frequency-selective RF-EMF exposure.

Under random hold-out validation, Extra Trees relied primarily on the weekend indicator (0.073), followed by the days-since-start feature (0.053) and hour sine (0.049), while nearby base station distances exhibited relatively low importance. Under buffered spatial validation, the same three temporal predictors remained dominant, while the importance of the spatial predictors decreased further still.

In contrast, CatBoost and XGBoost exhibited more stable feature importance profiles across both validation strategies, but were led by a different temporal predictor than Extra Trees. For CatBoost, the days-since-start feature was consistently

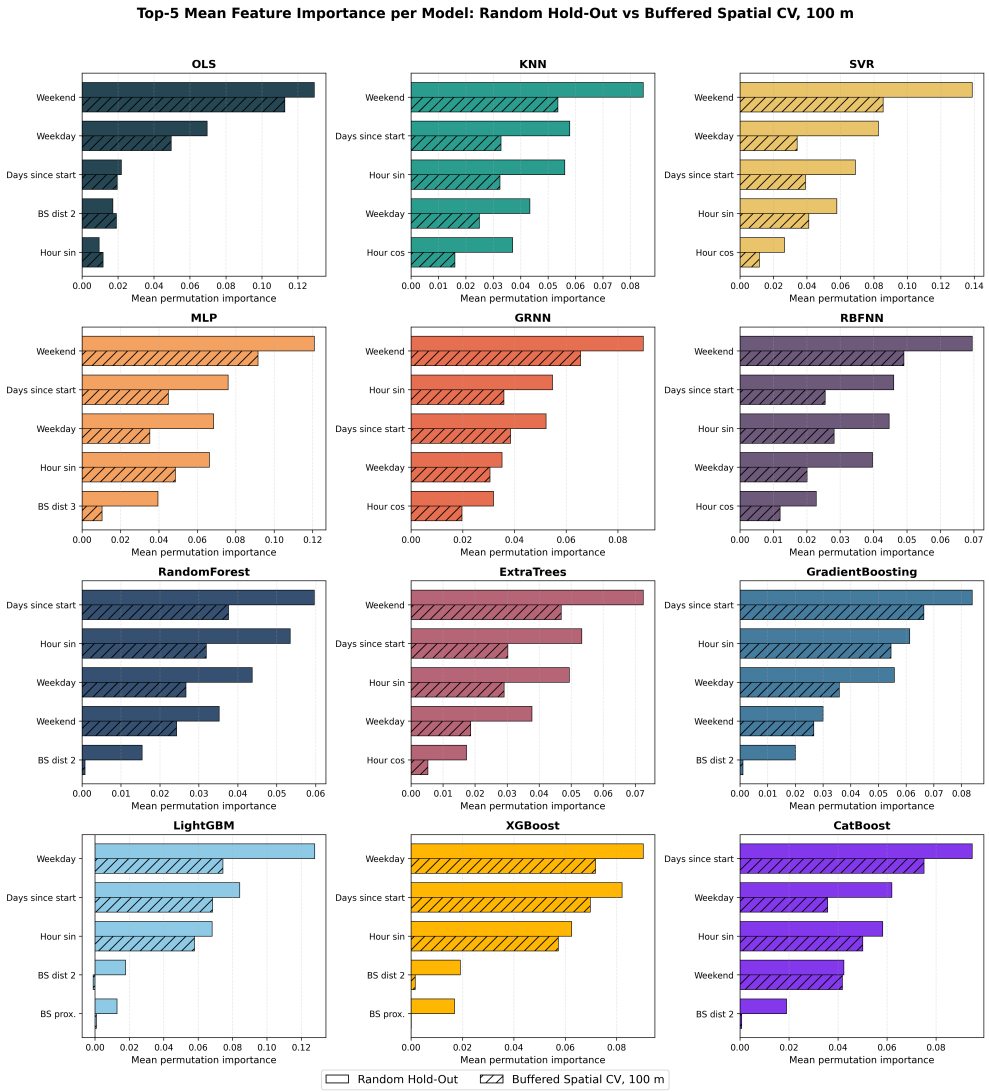


Fig. 12: Top five most important features for each of the twelve ML models under random hold-out and buffered spatial CV, 100 m validation.

the dominant predictor (0.095 under random hold-out and 0.075 under buffered spatial validation), although the order of the remaining two leading temporal predictors reversed between the two lenses: the weekday indicator ranked second and hour sine third under random hold-out (0.062 and 0.058), whereas hour sine ranked second and weekday third under buffered spatial validation (0.050 and 0.036). For XGBoost, the weekday indicator led under both validation strategies (0.090 and 0.072, respectively), closely followed by days-since-start (0.082 and 0.070), while base station distance

and proximity-related features ranked considerably lower and, under buffered spatial validation, contributed almost no measurable importance.

These results indicate that temporal predictors consistently dominated model behavior, while spatial predictors provided complementary information across all evaluated models. Because this dominance is concentrated in the weekday, weekend, and days-since-start indicators, calendar-position features, rather than in the within-day hour-of-day encoding that most directly reflects genuine intraday network load variation, the following subsection investigates whether this reflects a real temporal pattern in RF-EMF exposure or an artifact of the measurement campaign's calendar coverage.

4.5 Temporal Feature Ablation

The feature importance results above raise a question of interpretation: whether the models' reliance on calendar-position features (weekday, weekend, and days-since-start) reflects a genuine, generalizable temporal effect on RF-EMF exposure, or instead reflects the measurement campaign's own scheduling pattern. To test this, three versions of the pipeline were retrained, using only the five temporal features, only the seven spatial features, and the full twelve-feature combined set, each under all three validation strategies used elsewhere in this study. This ablation used a representative subset of seven algorithms with default hyperparameters (OLS, KNN, SVR, MLP, Random Forest, Extra Trees, and Gradient Boosting), since the goal is to compare the information content of each feature subset rather than to reproduce best-achievable accuracy. NRMSE, rather than R^2, is reported for leave-one-date-out, for the same reason given in Section 4.3. Table 5 reports the band-balanced mean results for each feature set.

Table 5: Band-balanced mean R^2 and RMSE, averaged across seven representative ML models (default hyperparameters), for temporal-only, spatial-only, and combined feature sets under all three validation strategies (NRMSE reported in place of R^2 for leave-one-date-out).

Feature set	R^2 (random hold-out)	RMSE (random hold-out)	R^2 (buffered spatial CV)	RMSE (buffered spatial CV)	NRMSE (leave-date-out)
Spatial-only	0.274	0.203	-0.340	0.279	0.944
Temporal-only	0.561	0.156	0.462	0.178	0.879
Combined	0.664	0.134	0.339	0.195	0.866

Under random hold-out validation, all three feature sets achieved positive R^2, with the combined set performing best (0.664) and the spatial-only set still capturing a modest but genuine signal (0.274). Under buffered spatial validation, the spatial-only feature set no longer generalized (R^2 = -0.340), while the temporal-only feature set retained a positive R^2 of 0.462, exceeding even the combined set (0.339). This indicates that the temporal predictors carry the more geographically transferable signal, while the current spatial features, tied to the fixed set of 20 base stations sampled here,

are most informative within the interpolation regime represented by random hold-out. Under leave-one-date-out validation, the temporal-only (NRMSE = 0.879) and combined (0.866) feature sets performed similarly, and both outperformed the spatial-only set (0.944), consistent with spatial features being computed from fixed, date-invariant coordinates that carry limited information about date-to-date shifts.

Taken together, the three lenses indicate that the temporal predictors carry real, transferable information about RF-EMF exposure that complements the spatial features, particularly for generalization to previously unseen locations. The temporal-only set's slightly weaker showing under leave-one-date-out than under buffered spatial CV suggests that part of this signal is tied to the specific fifteen-date measurement schedule, rather than reflecting a fully independent diurnal or weekly rhythm. Future work should enrich the spatial feature set with base station operational parameters (transmit power, antenna height, tilt, and azimuth) and extend the measurement campaign across a broader, more balanced calendar of dates, to disentangle genuine periodic effects from scheduling artifacts.

5 Discussion

The results showed that ML models could successfully predict frequency-selective RF-EMF exposure levels using temporal and spatial predictors. Across the random hold-out and buffered spatial validation strategies, ensemble tree-based methods consistently achieved the highest predictive performance, in agreement with previous studies [2, 3]. Extending these findings, the present study found that CatBoost and Gradient Boosting also ranked among the leading models for frequency-selective RF-EMF prediction. Predictive performance generally decreased under buffered spatial validation, though several models still generalized well to previously unseen geographic regions. The leave-one-date-out results, discussed separately below, revealed a further, largely independent generalization gap specific to unseen calendar dates.

The statistical comparison confirmed these observations. The Friedman tests found statistically significant differences among the evaluated ML models under both the random hold-out and buffered spatial validation strategies. However, the subsequent Conover post-hoc analysis with Holm correction indicated that, under random hold-out validation, the six leading models, CatBoost, Gradient Boosting, LightGBM, Random Forest, XGBoost, and Extra Trees, were not statistically distinguishable from one another (all pairwise $p \geq 0.194$). The same six-way equivalence held under buffered spatial validation, although the separation from the remaining, weaker-performing models narrowed considerably there, with only CatBoost separating significantly from every weaker model. These findings suggest that although ensemble tree-based algorithms consistently provide the highest predictive performance, no single model shows a robust, validation-strategy-independent advantage over the other leading ensemble methods, and that this equivalence extends further into the wider model pool once geographic leakage between training and test locations is removed.

When comparing the mean performance of the ML models across the random hold-out and buffered spatial validation strategies, Random Forest achieved the lowest mean RMSE under random hold-out, closely followed by Gradient Boosting, LightGBM, and

CatBoost. Under buffered spatial validation, the rankings shifted: CatBoost, XGBoost, and Extra Trees emerged as the top three performers, while Gradient Boosting and Random Forest, the strongest performers under random hold-out, both fell toward the bottom of the six-model ensemble tier. Strong performance under random hold-out validation therefore does not necessarily translate to robust generalization in previously unseen geographic regions. This confirms that buffered spatial validation provides a more conservative and realistic assessment of model performance.

Computational efficiency was also evaluated through the mean prediction time and mean total fit-and-predict time. Under random hold-out validation, Random Forest achieved the lowest RMSE at a moderate computational cost (0.76 s band-balanced), while LightGBM and XGBoost matched it closely in accuracy at a substantially lower cost (0.14 s and 0.21 s, respectively). CatBoost, despite accuracy in the same tier as the leading models, incurred the highest computational cost of all (1.55 s). A similar pattern held under buffered spatial validation, where LightGBM and XGBoost again offered the most favorable balance of accuracy and speed, while CatBoost, despite the lowest RMSE, remained among the most computationally expensive models. The specific boosting implementation therefore matters considerably for computational cost, with LightGBM and XGBoost consistently providing a more favorable accuracy-to-cost ratio than CatBoost. Previous studies have demonstrated the applicability of deep learning architectures for RF-EMF prediction [1, 7], but emphasized predictive accuracy alone. The present study shows that boosting-based ML models can match the highest-ranked ensemble methods in accuracy while, for LightGBM and XGBoost specifically, offering a more favorable balance between performance and computational cost.

The leave-one-date-out results add an important qualification to the recommendations above. Once relative rather than raw error is considered (Section 4.3), every model's error increased more when generalizing to an unseen calendar date than when generalizing to an unseen location, confirming that temporal and geographic generalization are distinct challenges. The model best suited to temporal generalization was neither boosting-based nor bagging-based: GRNN achieved the best average Friedman rank under leave-one-date-out, despite ranking only mid-to-lower tier under random hold-out and buffered spatial validation, while CatBoost, the leading model under those two lenses, ranked worst here. GRNN's advantage came from its consistency across bands, ranking among the top three of twelve models in five of the seven frequency bands, and it was also inexpensive, with a band-balanced total fit-and-predict time of only 0.09 s. This rank-based advantage did not extend to absolute error, however: GRNN's own band-balanced NRMSE under leave-one-date-out (0.916) was the highest of all twelve models, while XGBoost, despite ranking only fourth, achieved the lowest (0.817). Rank-based consistency and absolute-error performance therefore disagree on which model is preferable for this lens, and, as discussed further in the limitation below, leave-one-date-out evaluates generalization to a measurement date excluded from training rather than genuinely prospective, forecast-style generalization. Model selection for this scenario should accordingly weigh which criterion, and which generalization goal, matters most for the intended application.

The frequency-selective analysis showed that evaluating models on an individual band basis gives a more detailed picture of prediction performance than aggregated mean metrics alone. The overall results showed strong predictive capability for the uplink bands, reaching R^2 values of up to 0.98 under random hold-out validation. The band-specific analysis also showed that the GSM and UMTS-LTE downlink bands, already the weakest under random hold-out validation, deteriorated to a near-zero or slightly negative best-model R^2 under buffered spatial validation. For these two bands, none of the evaluated models currently generalized geographically much better than predicting the mean. The lower predictive performance observed for several downlink frequency bands may be related to the greater complexity of downlink RF-EMF exposure characterization. Previous work by [29] showed that uplink and downlink exposure originate from different sources and are influenced by different network parameters. Uplink exposure is primarily driven by user equipment activity, while downlink exposure is affected by base station operation, network configuration, and propagation conditions. This added complexity likely contributes to the greater variability observed in downlink measurements, and could partly explain the lower prediction accuracy obtained for several downlink bands in this study. None of the features used here directly represent base station transmit power, antenna radiation pattern, or instantaneous traffic load and resource-block utilization, so the model has no direct means of capturing the network-side variability that mainly governs downlink exposure. The closer alignment between uplink exposure and the temporal and distance-based predictors used here, by contrast, is consistent with uplink exposure being more directly tied to user-side activity patterns that these features can partially capture. This diagnostic points to a specific and actionable direction for closing the downlink gap: adding network operational telemetry, and traffic-load indicators in particular, as predictors, rather than changing the modeling approach itself.

The feature importance analysis showed that temporal predictors, particularly the weekend, days-since-start, and weekday indicators, played a dominant role in predicting frequency-selective RF-EMF exposure, ahead of the cyclic hour-of-day encoding itself. This is consistent with previous studies reporting substantial temporal variability in RF-EMF exposure due to fluctuations in network traffic and human activity patterns [30, 31]. The prominence of the weekday and weekend indicators reflects differences in population mobility, work schedules, and communication demand, which affect the loading of cellular base stations and, in turn, the emitted electric field levels. The days-since-start feature captured a longer-term trend consistent with ordinary network growth over the measurement period, though this study did not establish that link directly, and the trend could equally reflect seasonal variation or other unmeasured confounds tied to the campaign's own calendar. The comparatively smaller contribution of the hour-of-day encoding indicates that, within this dataset, calendar-position effects carried more predictive weight than within-day cyclic variation. In contrast, the explainable AI studies presented in [2] and [3] identified urban and geographic descriptors, including population density, urbanization level, and building-related characteristics, as the dominant predictors of aggregate electric field strength. This difference is likely due to the frequency-selective nature of the present study, where temporal variations in the utilization of individual wireless services have a

greater influence on exposure levels than the broader urban characteristics considered in previous work. The continued importance of base station distance and proximity-related features under random hold-out validation also supports earlier findings that spatial variability is associated with RF-EMF exposure levels in urban environments [32]. However, the dedicated temporal feature ablation in Section 4.5 showed that this spatial signal was tied largely to the specific set of 20 base stations sampled here, rather than to a distance-based relationship that generalizes to previously unseen locations. The temporal predictors, by contrast, retained positive explanatory power under both buffered spatial and leave-one-date-out validation. Together, these results indicate that both temporal and spatial characteristics contribute to RF-EMF prediction under random hold-out conditions, but that temporal factors carry the more robust and transferable signal within the investigated dataset.

A limitation of the present study is that the predictive models relied on temporal and spatial-context features without explicitly incorporating several network and propagation-related factors, such as transmit power, antenna characteristics, line-of-sight conditions, or traffic load. As a result, the achievable prediction accuracy was inherently constrained by the available input information, particularly for the downlink frequency bands.

A further limitation concerns what the leave-one-date-out lens does and does not demonstrate. Each fold withheld a single calendar date from among the fifteen available and trained on the remaining fourteen, so a held-out date in the middle of the campaign was evaluated using a model trained on dates both before and after it. This lens therefore evaluated generalization to a measurement date excluded from training, not strictly prospective, forecast-style generalization to a future date using only past data. The results reported in Section 4.3 should be interpreted accordingly, rather than as evidence of forecasting capability. A chronological train-on-past, test-on-future evaluation was not part of the present design and would be needed to support claims about forecasting RF-EMF exposure beyond the measurement campaign's own calendar window. This is left for future work.

The statistical equivalence observed among the leading ensemble models suggests that model selection for practical deployments can be guided by computational efficiency, implementation requirements, and the intended generalization scenario, rather than predictive accuracy alone. From a practical perspective, these findings demonstrate the potential of ML techniques to support frequency-selective RF-EMF exposure assessment and exposure mapping applications. The strong predictive performance achieved across multiple wireless services shows that temporal patterns and spatial-context information derived from telecommunication infrastructure can be used effectively to estimate exposure levels. Computationally efficient boosting-based models such as LightGBM and XGBoost are therefore particularly attractive candidates for large-scale deployment at unmeasured locations within the same general time frame as the measurement campaign, since they match the highest-ranked models in predictive performance while requiring substantially lower computational resources. For deployments where the main concern is robustness to measurement dates not included in the training data, Section 4.3 shows that GRNN offers the strongest rank-based temporal consistency at a computational cost comparable to LightGBM and XGBoost, while

XGBoost offers the lowest absolute error, measured as band-balanced NRMSE, under the same validation lens. As noted above, this evidence describes generalization to a date excluded from training rather than forecasting to genuinely future dates, so the choice between these two models should be guided by whether occasional large errors or a consistently elevated error level poses the greater practical risk for the intended application. Such approaches could complement conventional measurement campaigns by enabling exposure estimation at unmeasured locations, supporting large-scale exposure monitoring and assessment while reducing the time and resources needed for extensive field measurements.

6 Conclusion

This study investigated the prediction of frequency-selective RF-EMF electric field exposure using twelve ML models evaluated under three validation strategies: random hold-out validation, buffered spatial validation, and leave-one-date-out validation. The models were trained and tested using real in situ electric field measurements collected in Prishtina, Kosovo, across seven frequency bands representing contemporary mobile telecommunication services.

The results demonstrate that ML techniques can successfully predict frequency-selective electric field exposure levels using a combination of temporal and spatial predictors, with ensemble tree-based methods, and CatBoost in particular, consistently achieving the highest and statistically comparable predictive performance under both random hold-out and buffered spatial validation. Under the third, leave-one-date-out lens, however, model rankings shifted markedly, and predictive accuracy alone is therefore not sufficient for model selection. LightGBM and XGBoost offered the most favorable balance between accuracy and computational cost for geographic generalization, while for generalization to a measurement date excluded from training, GRNN offered the most consistent rank across bands and XGBoost the lowest absolute error, both at comparably low computational cost.

Individual frequency bands could be predicted considerably more accurately than the aggregated mean metrics suggest, particularly for the uplink bands, but the downlink services remained consistently harder to predict under random hold-out and buffered spatial validation, an effect that worsened substantially for two of the three downlink bands under buffered spatial validation. This gap most likely reflects the absence of network-side operational predictors, such as transmit power and traffic load, from the current feature set.

Feature importance analysis showed that temporal predictors, particularly the weekend, days-since-start, and weekday indicators, consistently dominated over both the spatial predictors and the cyclic hour-of-day encoding. A dedicated ablation confirmed that part of this temporal signal reflects the specific fifteen-date measurement schedule rather than a fully generalizable exposure rhythm, while the spatial features remain informative only when training and test locations are geographically intermixed.

Overall, the findings demonstrate that ML approaches can provide strong predictive performance for several frequency bands and computationally efficient predictions

of frequency-selective RF-EMF exposure levels, while also showing that geographic and temporal generalization are distinct challenges that degrade model performance in different, largely independent ways and are not necessarily best addressed by the same model. The study further highlights the importance of considering validation strategy, computational efficiency, and feature relevance when developing predictive exposure models. As a complement to conventional measurement campaigns, the proposed framework can support large-scale exposure assessment, exposure mapping, wireless network planning, and regulatory monitoring activities while substantially reducing the time and resources required for extensive field measurements.

Future work should focus on incorporating additional network and propagation-related information, particularly indicators associated with traffic load and network performance, to further improve prediction accuracy, especially for downlink frequency bands. Expanding the measurement campaign to additional geographic regions, telecommunication environments, and network deployments would further enable assessment of model robustness and generalizability under a broader range of real-world conditions, while extending the calendar coverage of the measurement campaign beyond the 15 dates analyzed here, particularly across additional seasons and years, would help determine whether the temporal generalization gap identified under leave-one-date-out validation narrows as more calendar diversity becomes available for training.

Declarations

- Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
- Conflict of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.
- Ethics approval and consent to participate: Not applicable
- Clinical trial number: not applicable.
- Consent for publication: Not applicable
- Data availability: Available from the corresponding author upon reasonable request.
- Materials availability: Not applicable
- Code availability: Available from the corresponding author upon reasonable request.
- Author contributions: This research was conducted as part of the Doctor of Science Research Project work of Doruntinë Berisha. Doruntinë Berisha developed the initial manuscript, conducted the comparative analysis of machine learning models for RF-EMF prediction, and performed the primary research activities. Mimoza Ibrani supervised the research, designed and synthesized the measurement dataset, provided methodological guidance, and substantially contributed to the critical revision and the improvement of the manuscript. Enver Hamiti and Hena Maloku contributed to the critical revision of the manuscript and participated in the evaluation and interpretation of the research results. All authors reviewed and approved the final version of the manuscript.

References

[1] Manassas, A., Delidimitriou, S., Wiart, J., Samaras, T.: Predicting electromagnetic field exposure using artificial intelligence methods. IEEE Access (2025)

[2] Kiouvrekis, Y., *et al.*: A comparative analysis of explainable artificial intelligence models for electric field strength prediction over eight European cities. Sensors **25**(1), 53 (2024)

[3] Kiouvrekis, Y., et al.: Explainable machine learning for radio environment mapping: An intelligent system for electric field strength monitoring. IEEE Access (2025)

[4] Givisis, I., Kalatzis, D., Christakis, C., Kiouvrekis, Y.: Comparing explainable AI models: SHAP, LIME, and their role in electric field strength prediction over urban areas. Electronics **14**(23), 4766 (2025)

[5] Kiouvrekis, Y., *et al.*: Explainable machine learning-based electric field strength mapping for urban environmental monitoring: A case study in Paris integrating geographical features and explainable AI. Electronics **14**(2), 254 (2025)

[6] Liu, S., Onishi, T., Taki, M., Watanabe, S.: A graph neural network-based electric-field prediction model for exposure assessments. In: 2023 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, pp. 443–444. IEEE, Portland, OR, USA (2023). https://doi.org/10.1109/AP-S/URSI57469.2023.10237719

[7] Zhang, Y., Wang, S., Wiart, J.: ExposNet: A deep learning framework for EMF exposure prediction in complex urban environments. IEEE Transactions on Machine Learning in Communications and Networking **4**, 1229–1243 (2026) https://doi.org/10.1109/TMLCN.2026.3714431

[8] Tognola, G., Plets, D., Chiaramello, E., Gallucci, S., Bonato, M., Fiocchi, S., Parazzini, M., Martens, L., Joseph, W., Ravazzani, P.: Machine learning for the estimation of WiFi field exposure in complex indoor multi-source scenario. In: 2021 XXXIVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS), pp. 1–4 (2021). https://doi.org/10.23919/URSIGASS51995.2021.9560437

[9] Mazloum, T., Wang, S., Hamdi, M., Mulugeta, B.A., Wiart, J.: Artificial neural network-based uplink power prediction from multi-floor indoor measurement campaigns in 4G networks. Frontiers in Public Health **9**, 777798 (2021)

[10] Al-Jumaily, A., Sali, A., Riyadh, M., Wali, S.Q., Li, L., Osman, A.F.: Machine learning modeling for radiofrequency electromagnetic fields (RF-EMF) signals from mmWave 5G signals. IEEE Access (2023)

[11] Bakcan, M.R.: Measurement and prediction of electromagnetic radiation exposure level in a university campus. Tehnicki Vjesnik **29**(2), 449–455 (2022)

[12] Nguyen, C., Cheema, A.A., Kurnaz, C., Rahimian, A., Brennan, C., Duong, T.Q.: Deep learning models for time-series forecasting of RF-EMF in wireless networks. IEEE Open Journal of the Communications Society (2024)

[13] Wang, S., Mazloum, T., Wiart, J.: Prediction of RF-EMF exposure by outdoor drive test measurements. Telecom **3**(3), 396–406 (2022)

[14] Chikha, W.B., Wang, S., Wiart, J.: An extrapolation approach for RF-EMF exposure prediction in an urban area using artificial neural network. IEEE Access **11**, 52686–52694 (2023)

[15] Engiz, B.K.: Artificial neural network based prediction of long-term electric field strength level emitted by 2G/3G/4G base station. Applied Sciences **13**(19), 10621 (2023)

[16] Draper, N.R., Smith, H.: Applied Regression Analysis, 3rd edn. John Wiley & Sons, New York, NY, USA (1998)

[17] Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician **46**(3), 175–185 (1992)

[18] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: Advances in Neural Information Processing Systems, vol. 9 (1996)

[19] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986) https://doi.org/10.1038/323533a0

[20] Specht, D.F.: A general regression neural network. IEEE Transactions on Neural Networks **2**(6), 568–576 (1991) https://doi.org/10.1109/72.97934

[21] Broomhead, D.S., Lowe, D.: Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. Royal Signals and Radar Establishment Memorandum (1988)

[22] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[23] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**(1), 3–42 (2006) https://doi.org/10.1007/s10994-006-6226-1

[24] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001) https://doi.org/10.1214/aos/1013203451

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[26] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

[27] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[28] Robusto, C.C.: The cosine-haversine formula. The American Mathematical Monthly **64**(1), 38–40 (1957)

[29] Krayni, A., *et al.*: Modeling and characterization of the uplink and downlink exposure in wireless networks. International Journal of Antennas and Propagation **2017**(1), 8243490 (2017)

[30] Velghe, M., Aerts, S., Martens, L., Joseph, W., Thielens, A.: Protocol for personal RF-EMF exposure measurement studies in 5th generation telecommunication networks. Environmental Health **20**(1), 36 (2021)

[31] Aerts, S., Wiart, J., Martens, L., Joseph, W.: Assessment of long-term spatio-temporal radiofrequency electromagnetic field exposure. Environmental Research **161**, 136–143 (2018)

[32] Velghe, M., Joseph, W., Debouvere, S., Aminzadeh, R., Martens, L., Thielens, A.: Characterisation of spatial and temporal variability of RF-EMF exposure levels in urban environments in Flanders, Belgium. Environmental Research **175**, 351–366 (2019)