IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4625225/
Vol. 2025 (34), Article ID 4625225

Survey of Datasets for Machine Learning in Radio-Frequency Signal Source Localization and Emitter Tracking*

Maciej KACZYŃSKI and Zbigniew PIOTROWSKI

Military University of Technology, Warsaw, Poland
E-mail:

Correspondence should be addressed to: Maciej KACZYŃSKI, maciej.kaczynski@wat.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

Machine learning (ML) has emerged as a transformative approach for radio frequency (RF) signal source localization and emitter tracking, providing data-driven solutions that extend beyond traditional analytical models. Progress in this area is fundamentally linked to the availability of datasets that capture realistic propagation characteristics across diverse scenarios. Real-world data acquisition remains costly, time consuming, and subject to regulatory and technical constraints, which motivates the use of both measured and synthetic datasets. The aim of this article is to survey publicly available datasets and simulators, categorizing their scope, applications, and limitations for advancing RF localization research. Particular attention is given to their role in supporting business applications such as industrial automation, wireless infrastructure optimization, and location-based services. By consolidating existing measurement datasets and open-source simulators, this work establishes a comprehensive reference to guide reproducible experimentation and the development of robust ML solutions in RF emitter localization and tracking.

Keywords: Datasets, emitter tracking, machine learning, signal source localization.

Introduction

Machine learning (ML) has become an important approach to radio frequency (RF) signal source localization and emitter tracking, offering data-driven techniques that complement traditional model-based methods. However, the progress of such methods is inherently linked to the availability of datasets that capture realistic propagation characteristics across diverse environments. Real-world data collection remains expensive, time consuming, and subject to regulatory and technical constraints, which often results in limited or non-representative samples. To address these challenges, the research community has developed a variety of datasets and simulators that facilitate reproducible experimentation. This survey reviews both measured and synthetic datasets, categorizing their scope, application domains, and limitations, thereby providing a foundation for future RF localization research.

Overview of Datasets

The development of ML algorithms for RF source localization and emitter tracking fundamentally depends on the availability of diverse and representative datasets. Recent efforts have produced several collections that capture propagation characteristics across a wide range of technologies and scenarios, including ray-traced radio maps, channel state information, industrial wireless measurements, and ultrawideband ranging data. These datasets support applications such as indoor and outdoor positioning, spectrum cartography, non-line-of-sight tracking, and antenna array calibration, thereby enabling systematic evaluation of localization methods. To provide a concise reference, Table 1 presents an overview of publicly available datasets for RF signal source localization and emitter tracking, including their associated publications and repositories.

Cite this Article as: Maciej KACZYŃSKI and Zbigniew PIOTROWSKI, Vol. 2025 (34) "Survey of Datasets for Machine Learning in Radio-Frequency Signal Source Localization and Emitter Tracking" Communications of International Proceedings, Vol. 2025 (34), Article ID 4625225, https://doi.org/10.5171/2025.4625225

Table 1: Overview of datasets for radio-frequency signal source localization and emitter tracking

Name	Related article	Repository link
RT-based dataset (V2.0) for 3D radio map under dynamic built-up scenario (1.25km × 1.25km)	(Wang et al., 2025)	https://data.mendeley.com/datasets/bn6n2639xh/3
Dataset of Pathloss and ToA Radio Maps with Localization Applica-tion (RadioMapSeer)	(Yapar et al., 2024)	https://ieee-dataport.org/documents/dataset-pathloss-and-toa-radio-maps-localization-application
iV2V (industrial Vehicle-to-Vehicle) and iV2I+ (iV2I + sensor) datasets (AI4MOBILE Industrial Wireless Datasets)	(Hernangómez et al., 2024)	https://ieee-dataport.org/open-access/ai4mobile-industrial-wireless-datasets-iv2v-and-iv2i
Tracking the Occluded Indoor Target with Scattered Millimeter Wave Signal	(Xu et al., 2024)	https://ieee-dataport.org/documents/tracking-occluded-indoor-target-scattered-millimeter-wave-signal
A Bluetooth 5.1 Dataset Based on Angle of Arrival and RSS for Indoor Localization	(Girolami et al., 2023)	https://zenodo.org/records/7759557
5G CFR/CSI Dataset for Wireless Channel Parameter Estimation, Array Calibration, and Indoor Positioning	(Pan et al., 2023)	https://ieee-dataport.org/documents/5g-cfrcsi-dataset-wireless-channel-parameter-estimation-array-calibration-and-indoor
Massive MultiInput Multi-Output (MaMIMO)	(Colpaert et al., 2023)	https://ieee-dataport.org/open-access/ultra-dense-indoor-mamimo-csi-dataset
Urban REMs	(Chaves-Villota et al., 2023)	https://zenodo.org/records/7839447
CSI Dataset towards 5G NR High-Precision Positioning	(Gao et al., 2022)	https://ieee-dataport.org/open-access/csi-dataset-towards-5g-nr-high-precision-positioning
UWB Ranging and Localization Dataset	(Flueratoru et al., 2022)	https://zenodo.org/records/4686379

RT-based dataset (V2.0) for 3D radio map under dynamic built-up scenario: The RT-based dataset (V2.0) constitutes a three-dimensional radio environment map (REM) generated using ray-tracing over a 1.25 km × 1.25 km urban area. Both static and dynamic source scenarios are considered, providing received signal strength indicator (RSSI) maps at multiple heights (2 m to 80 m), as well as time-varying maps for moving sources. The dataset is designed for applications in 3D radio map construction, spectrum environment monitoring, and radiation source localization in complex urban environments.

Dataset of Pathloss and ToA Radio Maps with Localization Application (RadioMapSeer): RadioMapSeer includes simulated pathloss and time-of-arrival (ToA) radio maps for dense urban scenarios, generated from a unified propagation simulation. This co-registration enables direct comparison of metrics and supports the training of deep learning models for pathloss prediction and wireless localization based on RSS or ToA. The dataset facilitates fair evaluation of alternative localization approaches.

iV2V and iV2I+ (AI4Mobile Industrial Wireless Datasets): The AI4Mobile datasets comprise measurement studies conducted in industrial testbeds, covering vehicle-to-vehicle (iV2V) and vehicle-to-infrastructure with sensor integration (iV2I+). The datasets provide communication parameters, including RSS, ToA, and auxiliary sensor data, under realistic industrial mobility conditions. It enables research on fingerprinting, line-of-sight detection, quality-of-service (QoS) prediction, and link selection, supporting ML applications in industrial wireless connectivity.

Tracking the Occluded Indoor Target with Scattered Millimeter Wave Signal: This dataset consists of channel impulse responses and received signals obtained from a 60 GHz millimeter-wave testbed. It focuses on exploiting multipath and scattered components to detect and track targets in occluded or non-line-of-sight indoor conditions. The dataset is intended for research in mmWave sensing, non-line-of-sight localization, and advanced signal processing.

A Bluetooth 5.1 Dataset Based on Angle of Arrival and RSS for Indoor Localization: This dataset provides real-world Bluetooth 5.1 measurements, including RSSI and angle-of-arrival (AoA) information, collected in controlled indoor environments. It is designed to support fingerprinting, triangulation, and ML-based indoor positioning methods. The dataset allows benchmarking and evaluation of AoA- and RSS-based algorithms.

5G CFR/CSI Dataset for Wireless Channel Parameter Estimation, Array Calibration, and Indoor Positioning: This dataset comprises wideband channel frequency response (CFR) and channel state information (CSI) collected in 5G networks. It enables antenna array calibration and estimation of channel parameters such as angles of arrival, departure, and propagation delays. Applications include high-accuracy indoor localization, array calibration, and channel modelling.

Massive Multi-Input Multi-Output (MaMIMO) Dataset: The MaMIMO dataset provides CSI measurements from dense indoor testbeds with hundreds of antenna elements. It includes multi-user, spatially diverse, and time-evolving channels, supporting research in localization, beamforming, and spatial channel modelling. The dataset is particularly suited for evaluating ML-based localization and communication optimization strategies.

Urban REMs (DeepREM Dataset): DeepREM offers radio environment maps (REMs) reconstructed from sparse measurements using deep-learning methods. The dataset captures outdoor urban settings and focuses on spectrum cartography and interpolation from limited sensing data. It supports research in REM estimation, coverage prediction, and deep-learning-based propagation modelling.

CSI Dataset Towards 5G NR High-Precision Positioning: This dataset provides CFR and CSI data tailored for fifth-generation (5G) New Radio (NR) high-precision positioning. It includes multipath, angle, and time-of-flight information, supporting the development of advanced localization methods. The dataset is intended for both indoor and urban positioning research with sub-meter accuracy.

UWB Ranging and Localization Dataset: This dataset consists of ultrawideband (UWB) measurements, including time-of-flight ranging, angular information, and positioning results obtained under energy-constrained Internet of Things (IoT) conditions. Measurements span multiple environments to assess accuracy, multipath resilience, and energy efficiency. It is particularly suitable for evaluating UWB-based localization algorithms in real-world deployments.

Synthetic Datasets Generation

In addition to real measurement, which are often costly, time consuming, and constrained to specific sites, synthetic dataset generation has become an important approach for advancing ML in RF source localization and emitter tracking. Physics based simulators and ray tracing engines make it possible to reproduce diverse propagation conditions such as multipath, shadowing, and diffraction under fully controlled scenarios that cannot be easily realized in real experiments. Such synthetic datasets provide scalability, flexibility, and access to ground truth parameters including positions, channels, and timing information that are rarely available in real world deployments. This enables systematic evaluation and benchmarking of algorithms across a wide variety of conditions. Table 2 presents an overview of selected open-source simulators for electromagnetic wave propagation that are applicable for generating datasets to support reproducible RF localization research.

Table 2: Overview of selected open-source simulators for electromagnetic wave propagation

Name	Related article	Repository link
OpenGERT	(Tadik et al., 2025)	https://github.com/serhatadik/OpenGERT
DiffeRT2d	(Eertmans et al., 2024)	https://github.com/jeertmans/DiffeRT2d
Sionna / Sionna RT	(Hoydis et al., 2023)	https://github.com/NVlabs/sionna
PyWaveProp	(Lytaev, 2023)	https://github.com/mikelytaev/wave-propagation
Veneris / Opal	(Egea-Lopez et al., 2019)	https://gitlab.com/esteban.egea/veneris

Summary

This survey examined publicly available datasets and open-source simulators that enable ML for RF source localization and emitter tracking. The reviewed datasets encompass a broad range of technologies, including ray-traced radio maps, Bluetooth AoA measurements, ultrawideband ranging, channel state information, and industrial wireless testbeds. These resources facilitate the development of models for applications such as indoor and outdoor positioning, spectrum cartography, non-line-of-sight tracking, and antenna array calibration. While real measurements remain fundamental, their acquisition is often costly and constrained, highlighting the complementary role of synthetic datasets generated through ray-tracing and wave-propagation simulators. Measured and simulated data jointly provide a comprehensive foundation for reproducible experimentation and the advancement of robust ML solutions in RF localization research.

Acknowledgment

This research work was funded by the project of the Military University of Technology titled: “New Neural Network Architectures for Signal and Data Processing in Radiocommunications and Multimedia.” Project No. UGB 22-054/2025.

Bibliography

- Chaves-Villota, A. and Viteri-Mera, C.A. (2023) “DeepREM: Deep-Learning-Based Radio Environment Map Estimation From Sparse Measurements,” IEEE Access, 11, pp. 48697–48714. Available at: https://doi.org/10.1109/ACCESS.2023.3277248.

- Colpaert, A. et al. (2023) “Massive MIMO Channel Measurement Data Set for Localization and Communication,” IEEE Communications Magazine, 61(9), pp. 114–120. Available at: https://doi.org/10.1109/MCOM.004.2200716.

- Eertmans, J., Oestges, C. and Jacques, L. (2024) “DiffeRT2d: A Differentiable Ray Tracing Python Framework for Radio Propagation,” Journal of Open Source Software, 9(98), p. 6915. Available at: https://doi.org/10.21105/joss.06915.

- Egea-Lopez, E. et al. (2019) “Vehicular Networks Simulation With Realistic Physics,” IEEE Access, 7, pp. 44021–44036. Available at: https://doi.org/10.1109/ACCESS.2019.2908651.

- Flueratoru, L. et al. (2022) “High-Accuracy Ranging and Localization With Ultrawideband Communications for Energy-Constrained Devices,” IEEE Internet of Things Journal, 9(10), pp. 7463–7480. Available at: https://doi.org/10.1109/JIOT.2021.3125256.

- Gao, K. et al. (2022) “Toward 5G NR High-Precision Indoor Positioning via Channel Frequency Response: A New Paradigm and Dataset Generation Method,” IEEE Journal on Selected Areas in Communications, 40(7), pp. 2233–2247. Available at: https://doi.org/10.1109/JSAC.2022.3157397.

- Girolami, M. et al. (2023) “A Bluetooth 5.1 Dataset Based on Angle of Arrival and RSS for Indoor Localization,” IEEE Access, 11, pp. 81763–81776. Available at: https://doi.org/10.1109/ACCESS.2023.3301126.

- Hernangómez, R. et al. (2024) “Toward an AI-Enabled Connected Industry: AGV Communication and Sensor Measurement Datasets,” IEEE Communications Magazine, 62(4), pp. 90–95. Available at: https://doi.org/10.1109/MCOM.001.2300494.

- Hoydis, J. et al. (2023) “Sionna: An Open-Source Library for Next-Generation Physical Layer Research.” arXiv. Available at: https://doi.org/10.48550/arXiv.2203.11854.

- Lytaev, M.S. (2023) “Tropospheric radio wave propagation modeling in Python 3 using PyWaveProp,” in 2023 IEEE 11th Asia-Pacific Conference on Antennas and Propagation (APCAP). 2023 IEEE 11th Asia-Pacific Conference on Antennas and Propagation (APCAP), pp. 1–2. Available at: https://doi.org/10.1109/APCAP59480.2023.10470193.

- Pan, M. et al. (2023) “In Situ Calibration of Antenna Arrays for Positioning With 5G Networks,” IEEE Transactions on Microwave Theory and Techniques, 71(10), pp. 4600–4613. Available at: https://doi.org/10.1109/TMTT.2023.3256532.

- Tadik, S. et al. (2025) “OpenGERT: Open Source Automated Geometry Extraction with Geometric and Electromagnetic Sensitivity Analyses for Ray-Tracing Propagation Models.” arXiv. Available at: https://doi.org/10.48550/arXiv.2501.06945.

- Wang, X. et al. (2025) “Spectrum environment map dataset based on ray-tracer under 3D dynamic built-up scenarios,” Data in Brief, 61, p. 111677. Available at: https://doi.org/10.1016/j.dib.2025.111677.

- Xu, Y. et al. (2024) “Tracking the Occluded Indoor Target With Scattered Millimeter Wave Signal,” IEEE Sensors Journal, 24(22), pp. 38102–38112. Available at: https://doi.org/10.1109/JSEN.2024.3447271.

- Yapar, Ç. et al. (2024) “Dataset of Pathloss and ToA Radio Maps With Localization Application.” arXiv. Available at: https://doi.org/10.48550/arXiv.2212.11777.