Fragmentation-Conditioned WiFi Threat Detection under Crowdsensing

José A. Macías-Montoya, Horacio I. Solís-Cisneros, and Josué R. Velázquez-González

***Abstract*—Passive WiFi crowdsensing is increasingly used to audit wireless exposure at urban scale, yet the proximity graphs built from sparse RSSI co-observations are rarely connected. On a field corpus of 53,371 georeferenced scans from Chiapas, Mexico (36,010 access points), about 85% of nodes are isolated after physical graph construction, and threat targets are assigned by heuristic threat-risk pseudo-labels for weak encryption and co-located duplicate identities (proxies, not adjudicated attacks). Under a fragmentation-conditioned protocol that scores the full graph, the connected subgraph, and isolates separately, and that includes municipality hold-out, tabular forests outperform graph neural networks on the full graph and on isolates, whereas neighborhood aggregation helps mainly where edges exist. In that connected regime, PhysAgg-R (GraphSAGE with path-loss softmax aggregation and an edge RSSI residual) attains the strongest in-distribution GNN PR-AUC among GraphSAGE and nodal-physics ablations across five seeds, but the margin is small and not McNemar-significant, spatial transfer favors GraphSAGE (PR-AUC 0.628 versus 0.581), and isolates remain a tabular problem. These results support connectivity-limited learning as the primary finding: message-passing value must be reported by connectivity regime, while RF-anchored operators offer only bounded in-domain refinement rather than a general substitute for denser sensing or isolate-aware detectors.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/11172

***Index Terms*—WiFi security, mobile crowdsensing, graph fragmentation, graph neural networks, path loss, physics-informed learning, connectivity-limited learning**

I. Introduction

WIRELESS threat detection from passive WiFi scans sits at the intersection of communications engineering, signal processing, and applied machine learning. Received signal strength (RSSI) carries information about propagation and proximity, and organizing co-observed access points into a graph is a natural way to exploit that relational structure. In field conditions, however, measurements arrive irregularly: access points co-occur only when auditors traverse overlapping routes, so the communication graph induced by sparse mobile crowdsensing is not the dense, reliably connected structure assumed in much of the localization and graph-based intrusion detection system (IDS) literature.

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Azurdia (*Corresponding author: Horacio I. Solís-Cisneros).*

J. A. Macías-Montoya, Horacio I. Solís-Cisneros, and J. R. Velázquez-González are with the Data Intelligence, Generative AI and Resilient Cybersecurity Research Group, Universidad Politécnica de Chiapas, Tuxtla Gutiérrez, Chiapas, Mexico (e-mails: jmacias@ids.upchiapas.edu.mx, hsolis@ids.upchiapas.edu.mx, and jvelazquez@ids.upchiapas.edu.mx).

Most prior work sidesteps this constraint. WiFi graph neural network (GNN) learning is typically developed on dense indoor fingerprints or evaluated only on connected subgraphs, which obscures how sparse measurements bound detection in deployed wireless systems [1]–[3]. Separately, physics-informed graph models often inject propagation knowledge as an auxiliary nodal loss rather than into the aggregation operator or into edge residuals that compare observed RSSI with a path-loss law. The scientific gap we target is therefore not "another GNN for WiFi," but the missing *regime diagnosis*: when message passing can help at all under crowdsensing fragmentation, how that regime must be reported, and only then whether RF physics in the operator yields a bounded gain inside it.

We study this gap on a field crowdsensing corpus whose proximity graph is extremely fragmented, with tens of thousands of components and the large majority of access points isolated. That structure motivates a working hypothesis we term *connectivity-limited graph learning*: the feasibility of message passing is governed first by graph informativeness and only second by neural architecture. The matching evaluation discipline is *fragmentation-conditioned* reporting—full graph, connected nodes, and isolates separately, plus robust municipality hold-out—so that connected-only scores cannot be mistaken for full-deployment capability.

Only after that diagnosis do we test a physics-anchored operator in the connected regime. PhysAgg-R (physics-informed aggregation with an edge residual) weights neighbors by a learnable path-loss kernel and adds an edge residual loss aligning observed RSSI with $P_0 - 10n \log_{10} d$. Where tabular detectors still dominate, where PhysAgg-R helps relative to GraphSAGE, and whether spatial hold-out closes, are empirical questions answered below, not assumed a priori.

We state three contributions in order of importance. First (principal), we identify and quantify connectivity-limited graph learning for WiFi security under sparse crowdsensing, and we define a fragmentation-conditioned evaluation protocol that reports full-graph, connected-only, and isolate-only metrics together with robust municipality hold-out. Second, we release empirical infrastructure: a curated field corpus of 53,371 georeferenced scans (36,010 access points, 126 days, 27 auditors) and a validated band-specific path-loss model. Third, within the connected regime only, we introduce PhysAgg-R, a path-loss aggregation operator with an edge RSSI residual loss, and we report its in-distribution gains and spatial limits against GraphSAGE [4], [5] without claiming full-graph or cross-municipality superiority. The corpus, GraphSAGE, RSSI features, and tabular classifiers are experimental infrastructure

and baselines; they are not claimed as algorithmic novelty in themselves.

Keeping characterization, graph construction, baselines, and learning experiments in one narrative lets propagation assumptions, graph statistics, and detection claims stay internally consistent, so the diagnosis and its bounded mitigation are read against the same measurements.

II. Related Work

Research at the boundary of wireless communications and computational security has advanced along several partly disconnected lines. WiFi indoor positioning from RSSI or channel-state fingerprints remains the most mature thread, with surveys documenting the shift from deterministic matching to learning-based estimators and the persistent problem of fingerprint staleness [6], [7]. Recent systems extend this base to factory-scale noise mapping, RTT ranging with non-line-of-sight mitigation, and reference-free access-point localization across outdoor and indoor domains [8]–[10]. Graph-based fingerprint maintenance, exemplified by GUFU, propagates information over a similarity graph to update designated locations from unlabeled crowdsourced scans [3]. That setting assumes a pre-existing dense fingerprint lattice and co-occurring RSSI vectors. Dispersed mobile crowdsensing violates both assumptions: auditors intersect access points opportunistically, graphs fragment, and propagation varies with urban morphology.

Mobile crowdsensing supplies the acquisition paradigm behind our corpus. Surveys and generic frameworks emphasize incentive design, data quality, and coverage heterogeneity rather than downstream graph learning on security tasks [2], [11]. Empirical deployments confirm both the reach and the irregularity of field data, whether through passive WiFi infrastructure for community sensing [12], crowdsourced cellular ranging [13], ride-hailing trajectories that bootstrap indoor localization at metropolitan scale [1], or ad-driven measurement of mobile performance in the wild [14]. Public WiFi and channel-measurement corpora are emerging, yet they still favor dense, sensor-rich captures suited to localization rather than dispersed security auditing [15]–[17]. Reusable measurement resources for wireless security outside the usual test beds, including Latin America, remain scarce [18].

The physical layer provides context that graph-based threat models often omit. Urban path-loss studies increasingly pair log-distance models with machine learning and geospatial features across sub-GHz and mid-band deployments, in both line-of-sight and non-line-of-sight corridors [19]–[21]. Environment-aware calibration and RSSI or SNR regression further refine range estimates under shadow fading [22], [23], while dedicated field path-loss datasets and RSSI reinterpretation for sensing clarify what passive scans can support [24], [25]. These contributions validate propagation or sensing accuracy; they do not, in general, couple validated physics with graph learning for threat detection under sparse connectivity, nor do they prescribe how to score models when most nodes have no neighbors.

On the security and privacy side, surveys of WiFi sensing document attack and defense taxonomies that include open networks, rogue access points, and evil-twin scenarios [26]. Location privacy research shows that exposed WiFi scans on laptops can still leak mobility profiles even when handset location permissions are tightened [27], while studies of localization fragility motivate rejecting network-derived GPS fallbacks [28]. Together, these results constrain how scan campaigns should be designed, labeled, and released, but they stop short of benchmarking physics-informed graph models on real, fragmented crowdsensing graphs under a regime-aware protocol.

TABLE I
State-of-the-art Limitations versus this Work's Measurement Advance

Line	Limitation we attack	Our advance
Graph WiFi / GUFU-style maintenance [3]	Assumes dense fingerprint lattice and co-occurring RSSI vectors	Field MCS graph with 31k components; no dense lattice assumed
MCS frameworks [2], [11]	Emphasize sensing logistics; little regime-aware GNN scoring for security	Fragmentation-conditioned layers: full / connected / isolate
Path-loss ML [19], [22]	Validates RF fit; rarely couples physics into security GNN operators	Path-loss in aggregation + $\mathcal{L}_{\text{edge}}$; spatial limits reported
WiFi security surveys [26]	Taxonomies of threats; scarce fragmented field-graph benchmarks	Heuristic threat labels on a real Chiapas crowdsensing corpus
Connected-only GNN evaluation practice	Hides isolate dominance and inflates message-passing utility	Mandatory three-layer reporting + municipality hold-out

Across these lines we are not aware of prior work that jointly (i) builds threat-detection graphs from dispersed crowdsensed RSSI under extreme fragmentation, (ii) treats isolate dominance as a first-class evaluation constraint rather than filtering it away, and (iii) places path-loss physics in both the aggregation operator and an edge RSSI residual while reporting spatial hold-out limits. Table I maps each line to the concrete limitation we attack and the measurement advance we report.

III. Methodology

A. Corpus and Graph Construction

The curated Chiapas corpus is the empirical testbed for the fragmentation diagnosis. After geographic filtering and removal of invalid GPS fallbacks, 53,371 observations remain from Android passive scans collected between 20 February and 27 June 2026 (126 days) by 27 volunteer auditors across 45 municipalities, covering 36,010 unique BSSIDs. Every curated row carries a geolocation fix; 31.4% of observed device identifiers exhibit randomized MAC behavior. Table II lists the core scan fields.

Auditors carried a custom Android client that logged passive WiFi scans with GPS fixes. Handset identifiers are stored as one-way hashes. Coverage is bounded to Chiapas using geographic bounding filters and rules that reject network-derived location fallbacks [28]. Sampling follows everyday mobility;

TABLE II
CORE WIFI SCAN VARIABLES IN THE CURATED RELEASE

Variable	Unit	Description
bssid	–	AP identifier (hashed on release)
ssid	–	Network name or <HIDDEN>
level	dBm	RSSI
frequency	MHz	Channel center frequency
channel	–	Derived IEEE channel
standard	–	802.11 generation label
bandwidth	MHz	Channel width
securityType	–	OPEN / WEP / WPA / WPA2 / WPA3
estimatedDistance	m	Path-loss range estimate
latitude, longitude	deg	Auditor coordinates
municipality, state	–	Geocoded admin labels
device_id	–	Pseudonymized auditor ID
timestamp	ms	Scan time (Unix ms)

rural grids and night-time routes are under-represented relative to urban corridors such as Tuxtla Gutiérrez.

Estimated distance applies a band-specific log-distance model

$$d = 10^{(P_{\mathrm{tx}} - \mathrm{RSSI})/(10n)} \tag{1}$$

with reference powers $P_{\mathrm{tx}} \in \{-45, -52, -55\}$ dBm for 2.4, 5, and 6 GHz bands and exponents $n \in \{3.0, 3.5\}$ as defined in the project data dictionary.

Detection experiments use heuristic pseudo-labels rather than adjudicated ground truth, following threat cues common in WiFi security taxonomies (open networks, legacy WEP, and co-located duplicate identities as evil-twin candidates) [26]. At scan level, label 1 is assigned when any of three rules fires: the access point advertises *OPEN* security, advertises *WEP*, or appears alongside another basic service set identifier (BSSID) under the same network name, auditor, and timestamp. All other scans receive label 0. Across the curated corpus this protocol yields a positive rate of 21.7%. These labels are *heuristic threat-risk proxies* for weak encryption and co-located duplicate identities; they do not certify active attacks. Detection metrics in this manuscript therefore measure agreement with the proxy protocol, not validated attack ground truth. Labels remain excluded from the public descriptor pending validation against a curated gold subset.

Graph models aggregate scans to access-point nodes with leakage-safe (SAFE) tabular features that omit test-set statistics. Two splits are used throughout. The temporal split (70/15/15 train/validation/test by timestamp) supports in-distribution detection on the full test fold (7,267 nodes) and on the non-isolated subgraph (1,636 nodes with at least one physical neighbor, 22.5% of the temporal test set). The spatial split holds out municipalities outside the metropolitan training region and evaluates 242 nodes in unseen areas for geographic generalization.

Before graph construction, the radio environment is characterized to show why the resulting communication graph is sparse. The same spatial and temporal irregularity described here is what later fragments the proximity graph. All figures use English axis labels and omit plot titles.

Coverage is uneven in both space and time. Fig. 1 maps observation density in the primary metropolitan zone of the study region, where roughly 48,000 geolocated points concentrate along the main transport corridors, and Fig. 2 shows

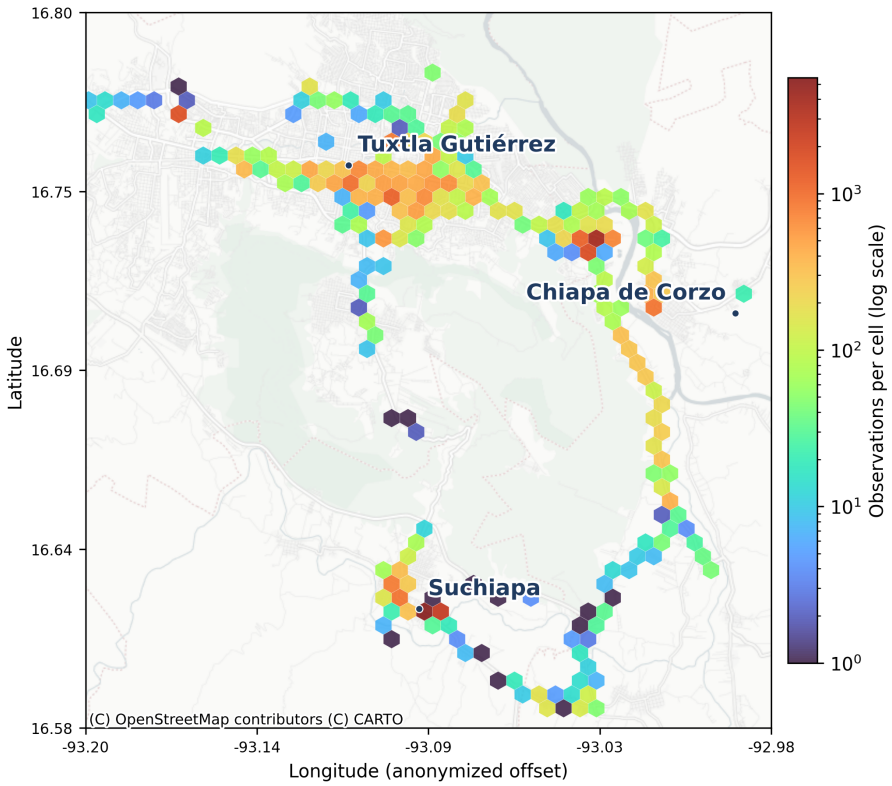
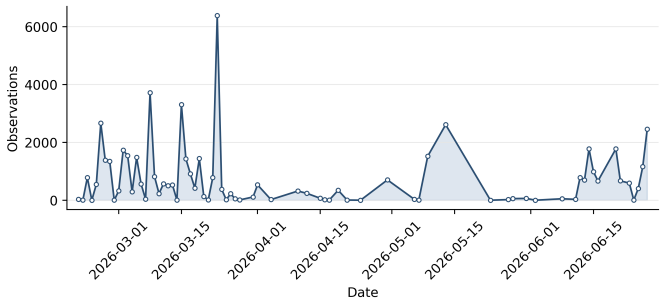


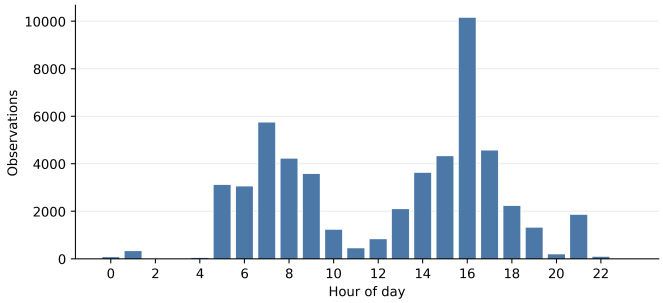
Fig. 1. Spatial density of WiFi observations in the primary metropolitan zone of the study region (∼48,000 geolocated points).

the corresponding daily and hourly counts, which peak during daytime hours in line with auditor mobility.

The security landscape is heavily skewed toward WPA2 (91.0%), with open networks accounting for 8.9%, a marginal WPA share of 0.1%, and WEP all but absent. Generation labels follow the expected succession, led by 802.11n (41.7%), then 802.11ax (26.5%) and 802.11ac (25.7%), while dual-band observations divide into 54.6% at 2.4 GHz and 45.4% at 5 GHz. Figs. 3 and 4 present these distributions and the ranking of collection zones by geolocated count.



(a) Daily observation counts across the collection period.

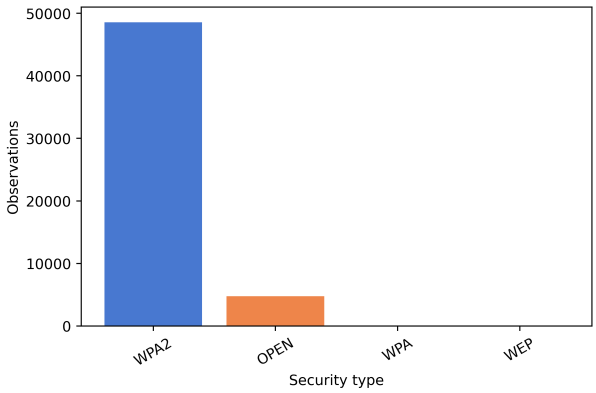


(b) Aggregate scan counts by hour of day.

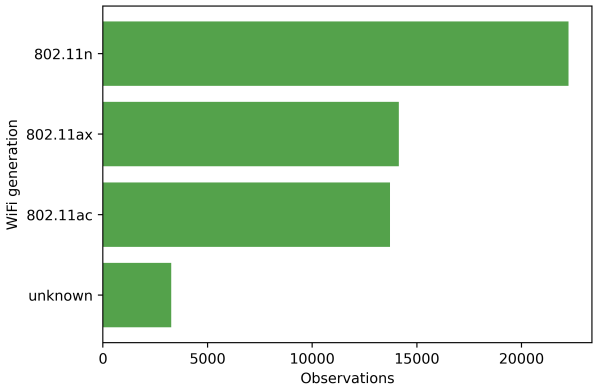
Fig. 2. Temporal distribution of scan activity.

Residual analysis supports the path-loss model used here, in line with urban RF validation studies [19], [24]. Fig. 5 plots its residuals on a logarithmic count axis, with a root-mean-square error (RMSE) of 4.09 m; the narrow peak at zero reflects the internal consistency of the band-specific parameters, while the tails trace multipath and unmodeled variation in transmit power. Channel indices derived from frequency agree with the recorded channel in 99.8% of rows. Mean RSSI is -82.6 dBm and mean path-loss distance 15.6 m, consistent with sparse outdoor scanning rather than dense indoor fingerprinting. Together with the corpus scale and coverage noted above, these checks confirm that the data are suitable for graph construction and that the path-loss model of Eq. (1) is internally consistent enough to serve as a physical constraint for PhysAgg-R.

Fig. 6 shows the construction pipeline from raw scans to detection experiments. Each node represents one BSSID with aggregated attributes (mean RSSI, modal frequency, security rates). The physical graph places an undirected *NEARBY* edge between two APs that co-occur in the same scan with mean estimated distance below 30 m and at least two co-observations. The spatial graph connects H3 geospatial index cells (resolution 9) that share APs, and the temporal graph links *TEMPORAL_NEXT* chains per auditor device.



(a) Marginal distribution of inferred security types.



(b) Leading WiFi generation labels.

Fig. 3. Security type and WiFi generation distributions.

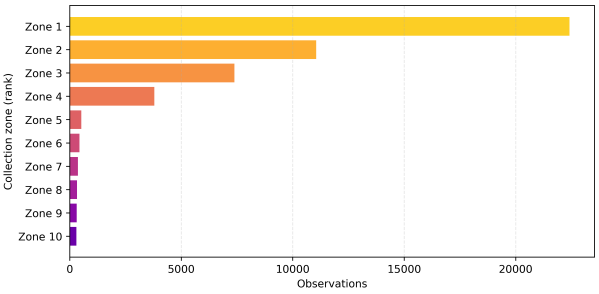
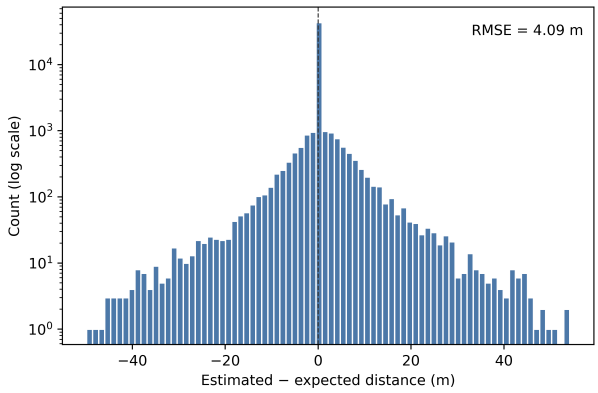


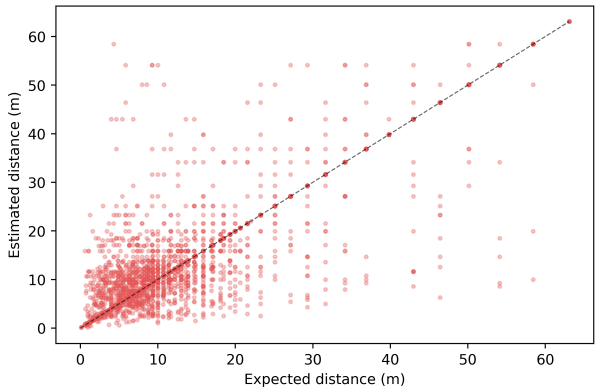
Fig. 4. Top 10 collection zones by geolocated observation count (administrative names anonymized; ranked by density).

Table III reports graph statistics. The physical graph is highly fragmented: 85% of nodes are isolated after co-observation filtering, yielding 31,315 components on 36,010 nodes and mean degree 1.12. That isolation rate is the structural fact that conditions all subsequent detection experiments, since standard message passing cannot propagate labels across isolated APs. Among the non-isolated minority, neighborhoods remain sparse (low degree), so even the connected regime offers limited message-passing bandwidth; the spatial (H3) and temporal views are denser but are not the graphs used for PhysAgg-R.

We assessed sensitivity of the isolated-node fraction to the distance cut-off and the minimum co-observation count (Table IV). Only 6.5% of access points are never jointly



(a) Residual distribution on a logarithmic count axis (RMSE 4.09 m).



(b) Expected versus recorded distance.

Fig. 5. Path-loss validation.

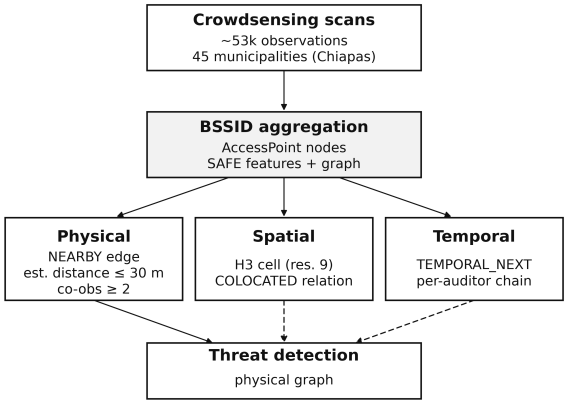


Fig. 6. Tripartite knowledge graph construction from crowdsensed scans to detection experiments.

TABLE III
KNOWLEDGE GRAPH STATISTICS (THREE VIEWS)

Graph	Nodes	Edges	Components	Avg. degree
Physical	36,010	20,074	31,315	1.12
Spatial (H3)	396	560	49	2.83
Temporal	10,273	10,246	27	2.00

observed with another AP in a single scan, which sets a hard floor on connectivity. Requiring at least two co-observations, as we do to suppress spurious single-scan links, yields an isolated fraction of 85% at the default 30 m threshold and keeps that figure between 84 and 85% for thresholds from 30 to 150 m. Lowering the minimum to one co-observation connects most co-observable APs (8% isolated at 30 m) but introduces edges supported by a single scan event. Edge count rises with distance mainly under min. co-obs. $= 1$; under the retained filter (min. co-obs. $= 2$) the edge set stays comparatively stable, so the fragmentation reported above is robust to the distance parameter while the co-observation minimum remains the dominant construction choice and is kept deliberately as a link-quality filter. In particular, moving from min. co-obs. $= 1$ to $= 2$ changes isolation from 8.2% to 85.1% at 30 m, whereas $= 3$ further sparsifies to 95.2%; the qualitative conclusion of extreme fragmentation under the retained quality filter is therefore not an artifact of a single undocumented threshold.

TABLE IV
SENSITIVITY OF ISOLATED-NODE FRACTION TO PHYSICAL-GRAPH CONSTRUCTION PARAMETERS (N=36,010 CCCESS POINTS)

Distance (m)	Min. co-obs. = 1	Min. co-obs. = 2	Min. co-obs. = 3
20	13.0%	86.8%	96.0%
30	8.2%	85.1%	95.2%
50	6.6%	84.3%	94.9%
100	6.5%	84.2%	94.9%

B. Models, PhysAgg-R, and Evaluation Protocol

Before graph-structured models, we establish observation-level baselines with Random Forest [29] and XGBoost [30] on SAFE tabular features, with F1-optimal thresholds selected on the validation fold. On the temporal hold-out, Random Forest reaches F1 0.656 / precision–recall area under the curve (PR-AUC) 0.746 and XGBoost F1 0.647 / PR-AUC 0.747; Fig. 7 shows the corresponding precision–recall curves. Five-fold temporal cross-validation yields mean F1 of 0.626–0.628, indicating stable but moderate separability under the pseudo-label threat model. These baselines serve as the tabular reference for the subsequent graph experiments. Because several SAFE and observation-level features encode the same cues used to build the heuristic label (notably open/WEP indicators and scan-level open rates), we ablate those label-proxy features and retrain Random Forest under the same temporal split. Observation-level PR-AUC falls from 0.749 (full feature set) to 0.420 without proxies ($\Delta = -0.329$); on AP nodes with SAFE features the drop is from 0.670 to 0.297 ($\Delta = -0.373$) when `open_rate` and `wep_rate` are removed. High full-feature RF scores therefore cannot be read as independent confirmation of the heuristic beyond the cues already present in the inputs.

Given the fragmentation established above, message passing fails for the large majority of isolated access points, so any physics-informed remedy must be evaluated on the connected regime without claiming full-graph superiority. We introduce PhysAgg-R, which places radio path-loss in the *aggregation operator* rather than only as a nodal embedding penalty,

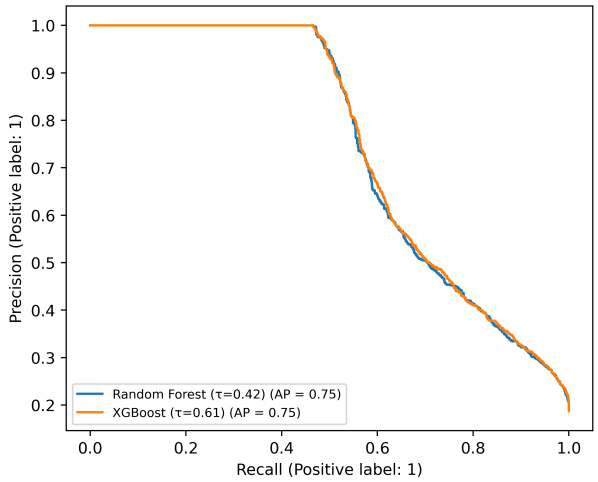


Fig. 7. Precision–recall curves for tabular baselines (temporal hold-out, observation level).

building on inductive neighborhood aggregation as in GraphSAGE [4] and on log-distance propagation models used in urban RF studies [19], [22].

For an edge (i, j) with estimated distance d_{ij}, PhysAgg uses attention weights

$$w_{ij} = \mathrm{softmax}_{j\in\mathcal{N}(i)}\big(-\alpha \cdot 10\, n \log_{10}(1 + d_{ij})\big), \quad (2)$$

with learnable path-loss exponent n and temperature α. Training minimizes classification loss plus an optional edge residual

$$\mathcal{L}_{\mathrm{edge}} = \mathrm{MSE}(R_{ij},\ P_0 - 10\, n \log_{10} d_{ij}), \quad (3)$$

where R_{ij} is the mean observed RSSI of the endpoints and P_0 is the band-specific reference. Nodal PiSAGE-style penalties of the form $\log(1 + \|\mathbf{h}_i\|)$ versus $\log(1 + d_{\mathrm{mean}})$ remain available as a baseline regulariser (A1/A3) but are not the proposed mechanism.

Ablations: A0 GraphSAGE; A1 nodal physics loss only; A2 PhysAgg operator; A2e PhysAgg+$\mathcal{L}_{\mathrm{edge}}$; A3 PhysAgg+nodal loss. Early stopping uses validation PR-AUC; spatial validation uses a large metropolitan municipality (Suchiapa) rather than a micro hold-out. Multi-seed temporal runs and McNemar tests [5] compare A2e against A0.

The models below are the instruments used to probe the fragmentation diagnosis, not the contribution in themselves: three are established baselines (Random Forest, XGBoost, GraphSAGE) [4], [29], [30] and the fourth panel sketches the physics-informed family (nodal PiSAGE ablations and PhysAgg-R). Figure 8 summarizes them.

The tabular baselines (panels a and b) operate at observation level on SAFE features. GraphSAGE (panel c) lifts predictions to AP nodes and aggregates over *NEARBY* edges. PhysAgg-R replaces uniform neighborhood aggregation with the path-loss softmax of Eq. (2) and, in A2e, adds the edge residual of Eq. (3); nodal PiSAGE-style penalties remain ablations A1/A3 only.

Experiments follow a fixed PhysAgg-R protocol. The primary temporal split is 70/15/15 by timestamp for in-distribution scoring and fragmentation-conditioned layers. Multi-seed temporal runs use seeds $\{42, 0, 1, 2, 3\}$ on A0–A3 to report mean±std PR-AUC. The spatial robust split trains on metropolitan municipalities, validates on Suchiapa, and scores unseen test municipalities once. A minimal A2e grid over $\lambda_{\mathrm{edge}} \in \{0.1, 0.3\}$ and proximity versus combined edge weights checks hypothesis-tied sensitivity. The primary metric is PR-AUC; early stopping on graph models uses validation PR-AUC. Fragmentation-conditioned reporting is mandatory on the temporal test fold: full graph, connected (non-isolated) nodes, and isolates. Fair Random Forest baselines use the same node features and the same mask as the GNN under comparison.

McNemar tests compare paired predictions of A2e against A0 on the same test instances (temporal connected and spatial robust). Selection uses the robust validation fold; the held-out test municipality set is scored once.

Several safeguards prevent information leakage: SAFE features omit test-set statistics, training and validation edges exclude pure test co-observations, and the spatial split holds out entire municipalities. Identifiers are hashed and coordinates are offset in released artefacts.

IV. Results and Discussion

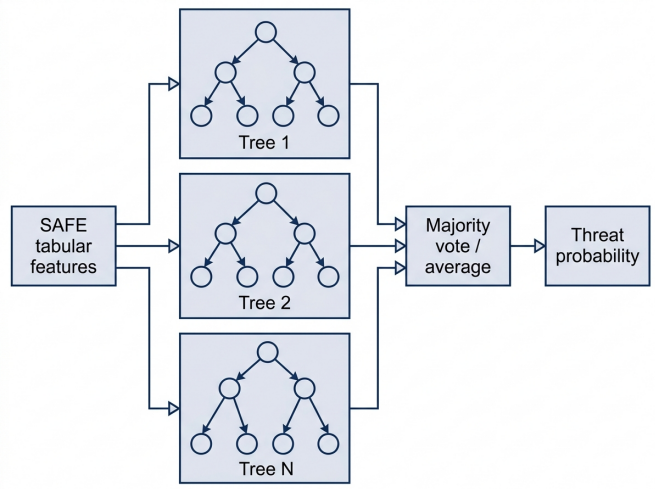
We first ask whether fragmentation conditions the utility of message passing, then whether PhysAgg-R improves in-distribution detection on the connected subgraph relative to GraphSAGE and nodal-physics ablations, whether that gain transfers under municipality hold-out, and whether graph operators close the isolate gap relative to tabular baselines. Heuristic pseudo-labels define the threat model throughout.

A. Graph-Structured Analysis

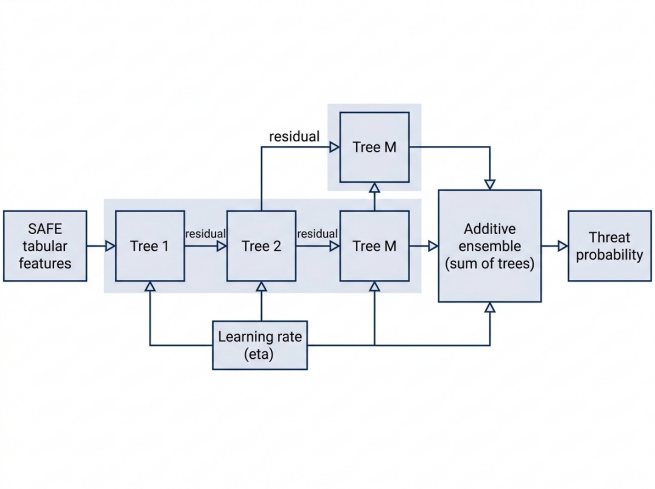
Before the PhysAgg-R ablations, a GraphSAGE feature ladder on the physical AP graph established the connectivity dependence that motivates fragmentation-conditioned reporting. On the full temporal test fold, message passing underperforms tabular baselines because isolates dominate. Restricting evaluation to nodes with at least one neighbor raises graph-model PR-AUC sharply, while a Random Forest on the same graph-derived features saturates on that small connected slice. Those contrasts are quantified under a common protocol in the detection results below. This subsection states the structural lesson and geographic context.

The lesson generalizes beyond this corpus. Mobile crowdsensing yields physical graphs that are inherently sparse, and comparing graph neural networks only where the graph happens to be dense inflates their apparent advantage. Reporting full-graph, connected-only, and isolate-only numbers keeps that inflation visible. Fig. 9 maps scan density as filled H3 hexagons at resolution 8 (regional view) and resolution 9 (metropolitan detail around the densest zone), preserving geographic structure without exposing raw device coordinates.

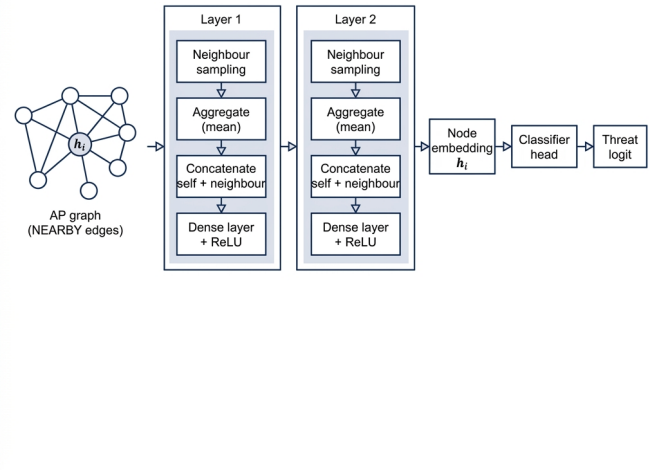
Taken together, the ladder and the three-layer protocol establish that message passing contributes only where the physical graph is connected. That finding motivates PhysAgg-R as an RF-anchored operator for the connected regime, not as a remedy for missing neighbors on isolates [2], [4].



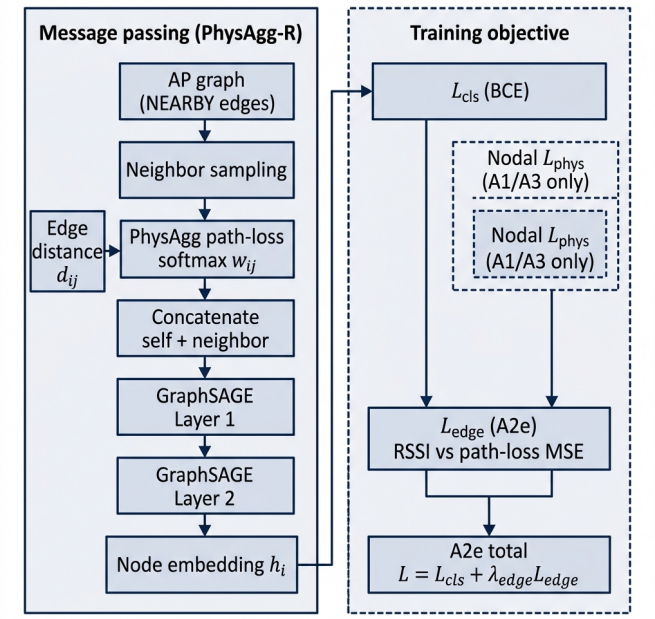
(a) Random Forest ensemble on SAFE tabular features.



(b) XGBoost additive boosting on residuals.



(c) GraphSAGE message passing on the physical AP graph.



(d) PhysAgg-R: path-loss softmax aggregation and edge residual $\mathcal{L}_{\text{edge}}$ (nodal losses only in A1/A3).

Fig. 8. Schematic architectures of the baseline detectors and the physics-informed extension used to probe the fragmentation diagnosis.

B. Detection Results

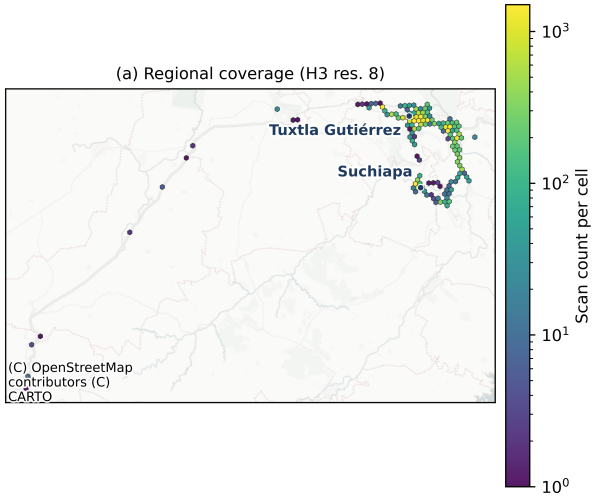
All detection metrics below use the heuristic pseudo-label protocol described above. The primary reporting discipline is *fragmentation-conditioned*: PR-AUC on the full temporal test fold, on connected (non-isolated) nodes only, and on isolates. Graph models use node features from the physical AP graph. Random Forest fair comparisons use the same node features and the same mask.

Table V contrasts GraphSAGE (A0) with a fair Random Forest on the temporal test fold. Isolates dominate (n=5,629 of 7,263), so full-graph PR-AUC collapses for both families, while the connected subgraph remains the only regime where neighborhood aggregation is competitive. The forest still leads on isolates and on the full graph. Message passing is informative primarily where edges exist.

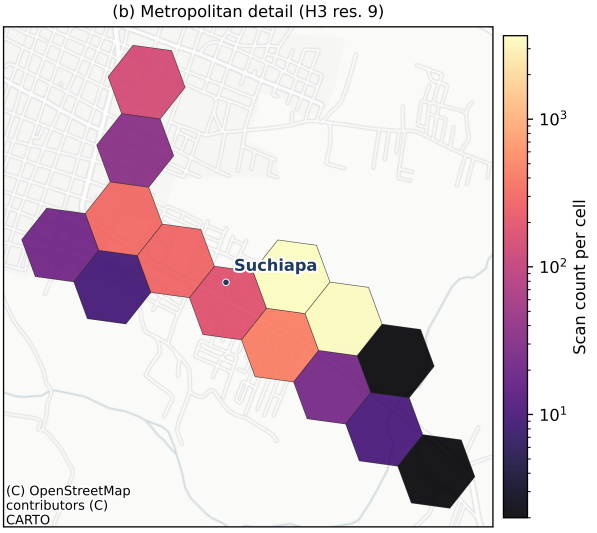
TABLE V
FRAGMENTATION-CONDITIONED PR-AUC (TEMPORAL TEST; FAIR RF ON SAME NODE FEATURES)

Layer	N_{test}	SAGE A0	RF fair
Full graph	7,263	0.504	**0.602**
Connected	1,634	0.826	0.998
Isolate	5,629	0.368	0.416

On the temporal non-isolated subgraph, Table VI summarizes the PhysAgg-R ablations with early stopping on PR-AUC. Five-seed means show that A2e (path-loss operator plus edge RSSI residual) is the strongest GNN among the SAGE and nodal-physics variants; the margin over A0 is small. A paired McNemar test of A2e versus A0 on this



(a) Regional coverage at resolution 8.



(b) Metropolitan detail at resolution 9 around the densest zone.

Fig. 9. Spatial scan density by H3 cell within the study region. Color encodes observation count per cell on a logarithmic scale; cell boundaries are shown explicitly.

split is not significant ($p \approx 0.80$). Fair RF on the same mask saturates and is reported as a tabular ceiling, not as a claim that GNNs outperform forests there. A minimal grid over $\lambda_{\text{edge}} \in \{0.1, 0.3\}$ and edge-weight schemes peaked at the default A2e setting; proximity-only weights and λ_{edge}=0.1 were slightly lower.

The ablation pattern admits a causal reading rather than an architectural race. Path-loss softmax reweights existing neighbors by a distance kernel; the edge residual aligns observed RSSI with a log-distance law on those same edges. Neither mechanism creates edges, so neither can help isolates. Nodal physics alone (A1) underperforms operator placement, which is consistent with putting RF structure where messages form. The small, non-significant in-domain margin of A2e over A0 indicates a bounded refinement of neighborhood aggregation, not a new capability on the fully fragmented graph. Table VII

TABLE VI
PHYSAGG-R ABLATION ON THE TEMPORAL NON-ISOLATED SUBGRAPH (PR-AUC; FIVE-SEED MEAN±STD OVER SEEDS $\{42, 0, 1, 2, 3\}$. RF FAIR REPORTED AT SEED 42 AS A TABULAR CEILING)

Run	Seed 42	5-seed mean±std
A0 GraphSAGE	0.883	0.880 ± 0.002
A1 nodal physics	0.798	0.823 ± 0.034
A2 PhysAgg operator	0.881	0.880 ± 0.003
A2e PhysAgg+$\mathcal{L}_{\text{edge}}$	**0.887**	$\mathbf{0.886 \pm 0.005}$
A3 PhysAgg+nodal	0.887	0.880 ± 0.005
RF fair (same mask)	0.999	—

reports municipality hold-out with Suchiapa as the held-out metropolitan validation municipality rather than a thin rural slice. Under this protocol SAGE A0 attains the highest PR-AUC (0.628), including over A2e (0.581). A paired McNemar test of predictions is significant for A0 versus A2e ($p \approx 0.019$). The temporal-to-spatial gap remains large ($\Delta_{\text{PR}} \approx 0.25$ for A0), and edge residuals do not close it: local RF consistency does not imply geographic transfer of threat patterns.

TABLE VII
SPATIAL ROBUST HOLD-OUT (VAL = SUCHIAPA; PR-AUC ON CONNECTED TEST NODES)

Run	PR-AUC
A0 GraphSAGE	**0.628**
A1 nodal physics	0.597
A2 PhysAgg operator	0.576
A2e PhysAgg+$\mathcal{L}_{\text{edge}}$	0.581
A3 PhysAgg+nodal	0.625

Taken together, the end-to-end picture under heuristic pseudo-labels is: temporal full-graph PR-AUC about 0.50/0.60 (SAGE/RF) with isolate dominance; connected-regime A2e at 0.886 ± 0.005 (five-seed mean) as the strongest GNN; isolate-layer tabular lead (about 0.37/0.42); and spatial robust hold-out led by A0 at 0.628 versus A2e at 0.581.

C. Discussion

The quantitative pattern above supports a structural reading rather than an architectural one. Sparse crowdsensing induces a proximity graph whose isolates dominate any full-graph score (Table IV: isolation remains ≈85% across distance cut-offs), consistent with coverage irregularity reported for mobile crowdsensing deployments [2], [12]. Fragmentation-conditioned reporting is therefore not cosmetic: connected-only metrics describe the only regime where neighborhood aggregation can act, while isolate-only metrics expose the tabular ceiling that full-graph averages otherwise hide. Within the connected regime, PhysAgg-R (A2e) is the strongest in-distribution GNN among the ablations, but the temporal margin over GraphSAGE [4] is small and not McNemar-significant, and fair Random Forest saturates on the same mask [5], [29]. Tabular models remain a highly competitive baseline even within the connected minority when predicting these heuristic threat-risk proxies, while the primary value of PhysAgg-R lies in exploring how physical propagation

constraints can be integrated into graph aggregation operators. Under robust municipality hold-out, GraphSAGE leads A2e (PR-AUC 0.628 versus 0.581), so edge path-loss physics does not substitute for geographic transfer, even when path-loss models are locally accurate [19]. PhysAgg-R is therefore reported strictly as a bounded in-domain refinement, not as a generally superior architecture. The dual contribution stands: a connectivity-aware evaluation discipline as the principal result, and an RF-anchored operator whose gains are in-domain and bounded.

For deployment, these results imply a regime-aware rule: on the full graph and on isolates (about 85% of nodes after physical construction), isolate-aware tabular heads remain the practical default because message passing has no neighbors to aggregate; neighborhood aggregation should be reserved for the connected minority; hybrid tabular-plus-GNN routing by connectivity mask is the natural operational pattern. New municipalities need denser sensing, physically justified graph densification, or explicit transfer rather than assuming in-domain GNN gains will travel [26]. Without full / connected / isolate reporting, a sub-percentage-point in-domain GNN gain is easily mistaken for a general IDS advance.

D. Limitations

Scope is bounded by a single-state opportunistic corpus in Chiapas, Mexico. Municipality hold-out tests transfer within that state; it does not establish cross-region external validity, so broader geographical datasets are required before claiming generalization beyond the study region. Heuristic threat-risk pseudo-labels remain pending gold validation, and band-wise P_0/n lack per-AP transmit power or indoor/outdoor adaptation. With most nodes isolated, edge-based models (including PhysAgg-R) have no neighbors on much of the test set. Spatial hold-out remains the hard failure mode relative to in-distribution gains. These limits are part of the study's evidence, not afterthoughts: they delimit where the protocol and the bounded operator should and should not be trusted.

V. Conclusions

The principal lesson of this study is that structural sparsity in the physical graph limits message passing more than classifier choice. On a field corpus whose proximity graph splits into tens of thousands of components with about 85% of access points isolated, fragmentation-conditioned evaluation shows that tabular forests outperform GNNs on the full graph because isolates dominate, while message passing is informative primarily on the connected subgraph. This diagnosis—connectivity-limited graph learning under crowdsensed WiFi security—is the study's dominant scientific result, not a claim of universal GNN superiority.

Operationally, the protocol implies a regime-aware rule: full-graph and isolate scores favor tabular heads; neighborhood aggregation is for the connected minority only; new municipalities need denser sensing or transfer strategies rather than assuming in-domain GNN gains travel.

Within the connected regime, PhysAgg-R (path-loss softmax aggregation plus an edge RSSI residual) achieves the strongest in-distribution GNN PR-AUC among the SAGE and nodal-physics ablations considered (five-seed mean 0.886 ± 0.005). It does not, however, outperform SAGE under a robust municipality hold-out (PR-AUC 0.628 versus 0.581 for A2e), nor does it close the isolate gap [4], [26]. Physics in the operator and edge residual is therefore a bounded methodological contribution secondary to the evaluation discipline: denser sensing, better proxy-label validation, and isolate-aware tabular heads remain necessary for deployable WiFi threat-*risk* detection under crowdsensing sparsity [2].

Acknowledgments

Processing follows the Mexican LFPDPPP. Auditors consented via the mobile application terms. Released artefacts omit identifiable SSIDs, raw MACs, and street-level addresses; maps use aggregated density only. The curated dataset, dictionary, configs, and figure code will be deposited under CC BY 4.0 with a DOI. Source code and the reproducibility package are publicly available at https://github.com/hsoliscisneros-csd/processing_id_11172.

During preparation the authors used Google Gemini to generate the schematic architecture figures in Section III-B and the graphical abstract, and AI-assisted editing tools for language polishing. ChatGPT was used for some details of the graphical abstract. All AI-assisted outputs were manually reviewed by the authors, who take full responsibility for the article.

References

[1] S. Zhu, L. Li, X. Wang, Q. Ni, Y. Jiang, H. Gao, Z. Han, and R. Gao, "Large-scale indoor localization via outdoor crowdsourcing trajectories on ride-hailing platform," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 2, pp. 1–23, 2025. [Online]. Available: https://doi.org/10.1145/3729498

[2] D. Cicek and B. Kantarci, "Use of mobile crowdsensing in disaster management: A systematic review, challenges, and open issues," *Sensors*, vol. 23, no. 3, p. 1699, 2023. [Online]. Available: https://doi.org/10.3390/s23031699

[3] K. H. Chiu, H. Yin, W. Zhuo, C.-H. Lee, and S.-H. G. Chan, "Graph-based fingerprint update using unlabelled wifi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 1, pp. 1–26, 2025. [Online]. Available: https://doi.org/10.1145/3712277

[4] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.02216

[5] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. [Online]. Available: https://doi.org/10.1007/BF02295996

[6] S. Shang and L. Wang, "Overview of wifi fingerprinting-based indoor positioning," *IET Communications*, vol. 16, no. 7, pp. 725–733, 2022. [Online]. Available: https://doi.org/10.1049/cmu2.12386

[7] X. Feng, K. A. Nguyen, and Z. Luo, "A survey on data augmentation for wifi fingerprinting indoor positioning," *IEEE Sensors Reviews*, vol. 2, pp. 246–264, 2025. [Online]. Available: https://doi.org/10.1109/sr.2025.3577579

[8] L. Xiao, P. Ghafoorpoor Yazdi, and S. Thiede, "A zone-based wi-fi fingerprinting indoor positioning system for factory noise mapping," *Journal of Intelligent Manufacturing*, vol. 37, no. 7, pp. 2697–2716, 2026. [Online]. Available: https://doi.org/10.1007/s10845-025-02660-y

[9] Q. Liang, G. Zhang, and L.-T. Hsu, "Accurate wi-fi rtt indoor positioning with nlos mitigation by floor map: A multiwall modeling approach," *IEEE Transactions on Instrumentation and Measurement*, vol. 75, pp. 9 522 316–9 522 316, 2026. [Online]. Available: https://doi.org/10.1109/tim.2024.3436101

[10] T. Amano, H. Yamaguchi, and T. Higashino, "Reference-free 3d wifi ap localization by outdoor-to-indoor bridging," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 688–700, 2025. [Online]. Available: https://doi.org/10.1109/ojcs.2025.3566774
[11] A. Abdeddine, L. Mekouar, and Y. Iraqi, "A generic framework for mobile crowdsensing: A comprehensive survey," *IEEE Access*, vol. 13, pp. 9134–9170, 2025. [Online]. Available: https://doi.org/10.1109/access.2025.3526739
[12] M. Ribeiro, D. Teixeira, P. Barbosa, and N. J. Nunes, "Using passive wi-fi for community crowd sensing during the covid-19 pandemic," *Journal of Big Data*, vol. 10, no. 1, 2023. [Online]. Available: https://doi.org/10.1186/s40537-022-00675-3
[13] R. Lu, L. Shi, Y. Liu, and Z. Dang, "A novel crowdsourcing-assisted 5g wireless signal ranging technique in mec architecture," *Future Internet*, vol. 17, no. 5, p. 220, 2025. [Online]. Available: https://doi.org/10.3390/fi17050220
[14] M. A. Bermejo-Agueda, P. Callejo, R. Cuevas, a. Cuevas, R. Durairajan, R. Rejaie, and a. Mayol, "Unveiling network performance in the wild: An ad-driven analysis of mobile download speeds," in *Proceedings of the ACM on Web Conference 2025*. ACM, 2025, pp. 4841–4852. [Online]. Available: https://doi.org/10.1145/3696410.3714761
[15] E. de Armas, G. Diaz, I. Sobron, I. n. Eizmendi, I. Landa, J. M. Matias, and M. Velez, "Ehunam, a wifi csi-based dataset for human and machine sensing," *Scientific Data*, vol. 12, no. 1, 2025. [Online]. Available: https://doi.org/10.1038/s41597-025-06238-4
[16] I. Ahmad, A. Ullah, and W. Choi, "Wifi-based human sensing with deep learning: Recent advances, challenges, and opportunities," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 3595–3623, 2024. [Online]. Available: https://doi.org/10.1109/ojcoms.2024.3411529
[17] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "Gopose," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–25, 2022. [Online]. Available: https://doi.org/10.1145/3534605
[18] M. Saldana-Perez, Y. Contreras-Jiménez, M. Torres-Ruiz, and C. Palma-Preciado, "Advances in technology for healthcare monitoring in latin america: A systematic review," in *Digital Tools and Data for Innovative Healthcare*. Elsevier, 2025, pp. 173–198. [Online]. Available: https://doi.org/10.1016/b978-0-443-27444-2.00012-5
[19] J. Lorente-López, I. Rodríguez-Rodríguez, J.-V. Rodríguez, and J. Maria Molina-Garcia-Pardo, "Discretized urban path loss modeling at 3.5 ghz using geospatial features and machine learning techniques: From regression to classification," *IEEE Access*, vol. 13, pp. 178 961–178 980, 2025. [Online]. Available: https://doi.org/10.1109/access.2025.3620996
[20] J. Lorente-López, I. Rodríguez-Rodríguez, J.-V. Rodríguez, J. María Molina García-Pardo, and J. Aznar-Poveda, "Machine-learning-based urban path loss prediction at 900 mhz: Principal component analysis, clustering, feature importance and regression," *IEEE Open Journal of Antennas and Propagation*, vol. 7, no. 1, pp. 40–56, 2026. [Online]. Available: https://doi.org/10.1109/ojap.2025.3629578
[21] A. Robles, J.-M. Poyanco, and M. E. Diago-Mosquera, "Large-scale analysis in los-nlos indoor corridor at 18 ghz," *Scientific Reports*, 2026. [Online]. Available: https://doi.org/10.1038/s41598-026-59673-7
[22] N. M. Obiri and K. Van Laerhoven, "Environment-aware indoor lorawan path loss: parametric regression comparisons, shadow fading, and calibrated fade margins," *Journal on Wireless Communications and Networking*, vol. 2026, no. 1, 2026. [Online]. Available: https://doi.org/10.1186/s13638-026-02607-8
[23] T. H. H. Soe, M. T. Soe, K. S. Linn, S. Teerasujaritpong, and C. Wechtaisong, "Enhanced indoor localization using rssi and snr-based regression models with particle filter optimization in lorawan," in *Proceedings of the 2026 15th International Conference on Informatics, Environment, Energy and Applications*. ACM, 2026, pp. 30–36. [Online]. Available: https://doi.org/10.1145/3803833.3803838
[24] P. Perdomo-Reyes, G. M. Galvan-Tejada, and A. Meneses-Viveros, "Path loss dataset from field measurements at 3.5 ghz for the fifth generation of wireless communications in indoor environments," *Scientific Data*, vol. 13, no. 1, 2026. [Online]. Available: https://doi.org/10.1038/s41597-026-06650-4
[25] Z. Wang, J. A. Zhang, K. Wu, and Y. J. Guo, "Rethinking rssi for wifi sensing," *npj Wireless Technology*, vol. 2, no. 1, 2026. [Online]. Available: https://doi.org/10.1038/s44459-026-00053-y
[26] X. Liu, X. Meng, H. Duan, Z. Hu, and M. Wang, "A survey on secure wifi sensing technology: Attacks and defenses," *Sensors*, vol. 25, no. 6, p. 1913, 2025. [Online]. Available: https://doi.org/10.3390/s25061913
[27] M. Cunha, R. Mendes, Y.-A. de Montjoye, and J. a. P. Vilela, "Wifi-based location tracking: A still open door on laptops," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 822–833, 2025. [Online]. Available: https://doi.org/10.1109/ojcs.2025.3569437
[28] "Fragile frames: Wi-fi's fraught fight against fragattacks," 2021, metadata pending manual verification (q3.04). [Online]. Available: https://doi.org/10.1145/3734477.3734478
[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324
[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785



José A. Macías-Montoya received the M.Sc. degree in Computer Science from the National Center for Research and Technological Development (CENIDET), Mexico. He is currently a professor and researcher at the Universidad Politécnica de Chiapas, Mexico, where he participates in academic, research, and technology development activities. His professional and research experience includes software engineering, computer networks, information systems, Internet of Things, and applied information technologies. His current research interests focus on cybersecurity, privacy, network monitoring, intelligent systems, and the application of computing technologies to real-world problems. He has participated in projects related to network monitoring, sensor systems, wireless technologies, and information security.



Horacio I. Solís-Cisneros received the Ph.D. degree in Engineering Sciences from the Tecnológico Nacional de México, Campus Tuxtla Gutiérrez, Mexico. He is currently a professor and researcher at the Universidad Politécnica de Chiapas, Mexico, and a member of the Data Intelligence, Generative AI and Resilient Cybersecurity Research Group. His research experience spans multiple levels, from mathematical and physical modeling and computational methods to intelligent systems and applied engineering. His work has addressed data-driven and physics-informed approaches for modeling, prediction, control, optimization, and diagnosis of complex systems. His research interests include machine learning and intelligent systems, resilient cybersecurity and wireless networks, renewable energy and photovoltaic optimization, data-driven analysis of physical systems, semiconductor materials, and optoelectronic and photonic devices for sensing and ultraviolet applications. His work integrates mathematical formalism, computational modeling, and experimental or engineering-oriented approaches to develop methods and technologies for real-world applications.



Josué R. Velázquez-González received the M.Sc. degree in Computational Sciences from the National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico. He is currently a professor and researcher at the Universidad Politécnica de Chiapas, Mexico, where he participates in teaching and research activities in information technologies and computing. His research interests include artificial intelligence, machine learning, signal processing, pattern recognition, and data analysis. His research experience includes the development of computational methods for the analysis of biomedical signals, particularly electrocardiographic signals, as well as the application of intelligent techniques to real-world problems.