LAPORAN AKHIR
SKEMA PENELITIAN DASAR (PENELITIAN DASAR FUNDAMENTAL DAN PENELITIAN DASAR KERJA SAMA DALAM NEGERI)

Peneliti hanya diperkenankan mengisi di tempat yang telah disediakan sesuai dengan petunjuk pengisian dan tidak diperkenankan melakukan modifikasi template atau penghapusan di setiap bagian.

.

A. JUDUL
Tuliskan judul usulan penelitian maksimal 20 kata

Gradient Boosting-Based Residual Correction Model for Occupancy Estimation Using Channel State Information Signals in Under-Actuated Zones

B. RINGKASAN
Isian ringkasan penelitian tidak lebih dari 300 kata yang berisi urgensi, tujuan, metode, dan luaran yang ditargetkan

This study proposes a robust two-stage framework for estimating occupancy in under-actuated indoor zones using WiFi Channel State Information as a non-intrusive sensing modality. These zones, marked by sparse sensor coverage and irregular airflow, present challenges for conventional detection methods. To address these, a domain-specific feature engineering pipeline is introduced, incorporating temporal, nonlinear, and spatial interaction features derived from filtered channel state information amplitudes. The proposed Gradient Boosting-Based Residual Correction Model combines a Gradient Boosting Regressor to learn primary occupancy patterns with a Histogram-Based Regressor to model structured residual errors. The model was evaluated using two distinct datasets: a controlled dataset with predefined occupancy levels, and an uncontrolled dataset collected in real-world public transit settings with high variability across zones. On the controlled dataset, the proposed model achieved high predictive accuracy with an RMSE of 0.1292, MAE of 0.0952, and R^2 of 0.9918. On the uncontrolled dataset, the model retained strong performance with an RMSE of 1.3270, MAE of 1.0933, and R^2 of 0.8199, preserving over 82.6% of its accuracy relative to the controlled setting. Ablation studies confirmed the critical role of multipath-aware smoothing features, where their removal led to a +426% increase in MAE. Temporal and spatial features contributed modestly but supported overall robustness. The proposed model offers a scalable, generalizable solution for real-time occupancy monitoring and intelligent HVAC control in both structured and dynamic indoor environments.

C. KATA KUNCI
Isian 5 kata kunci yang dipisahkan dengan tanda titik koma (;)

WiFi Channel State Information, Under-actuated zones, Occupant estimation, Feature engineering, Residual learning, Gradient Boosting, Passive sensing, Indoor environment modeling, Smart building systems.

D. PENDAHULUAN
Pendahuluan penelitian tidak lebih dari 1000 kata yang memuat, latar belakang, rumusan permasalahan yang akan diteliti, pendekatan pemecahan masalah, state-of-the-art dan kebaruan, peta jalan (road map) penelitian setidaknya 5 tahun. Sitasi disusun dan ditulis berdasarkan sistem nomor sesuai dengan urutan pengutipan.

The expansion of WiFi (Wireless Fidelity) infrastructure within indoor environments has led to advances in WiFi-based sensing technologies, which provide non-intrusive, affordable, and privacy-conscious alternatives to traditional occupant monitoring systems [1]. This sensing paradigm primarily utilizes Channel State Information (CSI), which offers detailed amplitude and phase metrics across various subcarriers, enabling the detection of human presence and movement due to its sensitivity to alterations in the propagating environment, influenced by elements such as walls and furniture, as well as the human body itself [2]. In comparison to conventional sensors such as infrared devices, cameras, or pressure mats each of which often comes with substantial installation costs, privacy dilemmas, and extensive deployment requirements CSI within WiFi systems allow for passive, infrastructure-free sensing leveraging pre-existing WiFi hardware [3]. WiFi's ability to utilize current network setups represents a significant reduction in costs while maintaining the capability of providing occupancy data. Challenges emerge in this domain; specifically, CSI signals exhibit high-dimensionality, noise, non-linearity, and temporal non-stationarity, factors that complicate occupancy estimation in fluctuating indoor environments [4]. This can be further complicated by multipath propagation effects often seen in various indoor locations [5].

Moreover, the challenges associated with using Channel State Information (CSI) for occupancy detection become especially pronounced in what are known as under-actuated zones, areas within a building that lack sufficient sensor coverage or direct control from the Heating, Ventilation, and Air Conditioning (HVAC) system. These zones are commonly found in shared workplaces, libraries, and open-plan offices, where a centralized HVAC configuration fails to provide uniform air distribution. As a result, these areas exhibit significant spatial variability in both thermal conditions and occupancy patterns, making it more difficult to accurately detect and respond to occupant presence [6]. Effectively determining the number of occupants in these regions is vital for enhancing climate control mechanisms, optimizing energy usage, and maximizing occupant comfort. A comprehensive understanding of these concepts is integral for advancing intelligent building systems and addressing issues related to indoor air quality and overall environmental health [7].

Recent literature indicates a promising trajectory for machine learning techniques applied to CSI-based estimation. Models such as Multilayer Perceptron, Convolutional Neural Network, Random Forest, Long Short-Term Memory, and Transformers, each offering distinct advantages and limitations in handling CSI data. Multilayer Perceptron (MLP) are valued for their architectural simplicity and computational efficiency, particularly in modeling non-linear relationships in tabular representations of CSI features. However, they lack the ability to capture spatial or temporal patterns inherent in CSI signals, making them less suitable for dynamic environments where time-variant features are critical [8].

In contrast, Convolutional Neural Networks (CNN) are well-suited for extracting spatial features from CSI amplitude sequences. By treating CSI data as structured inputs, such as matrices or pseudo-images, CNN can effectively detect localized spatial dependencies and enhance model interpretability [9]. Additionally, the incorporation of attention mechanisms within CNN architectures allows the model to prioritize salient spatial regions, which is essential in real-world scenarios characterized by multipath effects and signal distortion [10]. Nonetheless, despite these advancements, many machine learning models still utilize raw or minimally processed CSI inputs, often neglecting to integrate robust error correction mechanisms that address noise interference an aspect crucial for reliable operation in practical scenarios [11].

Random Forest (RF) models provide another compelling alternative, particularly due to their robustness against overfitting and ability to manage noisy and high-dimensional data. Random Forest models can capture complex non-linear relationships and identify the most influential features affecting CSI signals [12]. However, RFs are inherently static in nature and lack mechanisms for learning temporal dependencies [13].

To address these temporal modeling deficits, architectures designed for sequence learning, such as Long Short-Term Memory (LSTM), have gained traction. LSTM are adept at retaining information across extended sequences, effectively capturing long-range dependencies and recurrent patterns within CSI time series. This capability makes them particularly suitable for occupancy estimation tasks in dynamic environments, where present observations are heavily influenced by prior states [14].

From a theoretical standpoint, the task of estimating the number of occupants can be understood as a supervised regression problem. In this formulation, the objective is to construct a mathematical function that maps a set of numerical input features, extracted from Channel State Information (CSI) to a single numeric output, representing the estimated number of occupants in a room or space. These input features typically exist in a high dimensional space, where each dimension corresponds to a distinct aspect of the CSI signal, such as amplitude values from multiple subcarriers.

However, CSI data are often affected by factors such as signal noise, external interference, and non-linear transformations in the environment, all of which can significantly impact the accuracy of occupancy predictions [15]. In real-world scenarios, the difference between the true number of occupants and the model's prediction, referred to as the prediction error is not entirely random. Instead, it may include patterns of error that are systematically related to changes in the environment, such as reflections from walls, movements of people, or limitations in the model's structure. These are known as structured residuals [16]. According to the principles of bias-variance decomposition, prediction error can be broken down into three distinct components: bias, which refers to systematic errors due to overly simplistic models that fail to capture the true complexity of the data; variance, which reflects the model's sensitivity to fluctuations in the training data; and irreducible noise, which includes random factors that cannot be explained by any model, regardless of complexity [17].

Standard single-stage models often fail to correct these structured residuals, resulting in poor generalization [18]. It is recognized that achieving a balance between bias and variance is critical, especially in complex environments where the prediction model must navigate the high dimensionality and intricate dependencies among features [19]. To mitigate the issues related to underfitting and poor generalization in occupancy prediction, it is pertinent to adopt a residual-aware modeling approach. Such strategies focus not only on minimizing predictive error but also on capturing systematic errors through various model iterations and enhancements.

To overcome these limitations, this study introduces the Gradient Boosting-Based Residual Correction Model (GBRCM), a two-stage learning model enhanced by adaptive feature engineering, specifically designed to improve robustness and predictive accuracy in noisy, under-actuated zones. The model introduces a two-stage learning architecture: a base Gradient Boosting Regressor (GBR) that estimates the occupancy count, followed by a Histogram-Based Gradient Boosting Regressor (HGBR) that explicitly learns to correct the residual errors of the first stage. This residual correction mechanism is theoretically grounded in bias-variance decomposition and is capable of modeling structured error patterns caused by signal distortion, environmental variability, and sensor sparsity.

To further enhance generalization and model interpretability, we introduce a domain-informed feature engineering pipeline, incorporating both handcrafted signal features and statistical derivatives that reflect spatiotemporal variations and CSI-specific non-idealities. Recognizing the high dimensionality and inherent non-linearity of Channel State Information (CSI) data arising from multipath propagation, non-stationary environments, and amplitude distortion, a Butterworth filter is integrated into the pipeline to attenuate high-frequency noise and smooth abrupt fluctuations.

The organization of this paper adheres to a systematic framework. Section 2 outlines the research methodology, including dataset utilization, the adaptive feature engineering, signal filtering, and feature dimensionality reduction. The machine learning models used for comparative analysis are also described in this section. Section 3 presents the experimental results, evaluation metrics, comparing model performance across feature sets and state-of-the-art baselines including MLP, CNN, RF, LSTM, discussion of the key findings, limitations, and broader implications for passive sensing systems in dynamic indoor environments. An ablation study was systematically conducted to isolate and quantify the individual contributions of the core components within the proposed GBRCM architecture, namely the residual correction stage, multipath interference, temporal instability, and spatial variation. Finally, Section 4 concludes the study by summarizing the main contributions and outlining future research directions aimed at enhancing signal-based occupant detection in under-actuated zones.

E. METODE

Isian metode atau cara untuk mencapai tujuan yang telah ditetapkan tidak lebih dari 1000 kata. Pada bagian metode wajib dilengkapi dengan diagram alir penelitian yang menggambarkan apa yang sudah dilaksanakan dan yang akan dikerjakan selama waktu yang diusulkan. Format diagram alir dapat berupa file JPG/PNG. Metode penelitian harus memuat sekurang-kurangnya prosedur penelitian, hasil yang diharapkan, indikator capaian yang ditargetkan, serta anggota tim/mitra yang bertanggung jawab pada setiap tahapan penelitian. Metode penelitian harus sejalan dengen Rencana Anggaran Biaya (RAB).

This study formulates the task of occupancy estimation in under-actuated indoor zones as a regression problem using CSI signal features. The dataset includes Butterworth-filtered CSI amplitudes with engineered features capturing temporal, spatial, and signal dynamics. Principle component analysis is applied to reduce dimensionality. The proposed Gradient Boosting-Based Residual Correction Model (GBRCM) combines a base Gradient Boosting Regressor with a Histogram-Based Regressor for residual error correction. Benchmark comparisons are conducted against machine learning models. Hyperparameters are optimized via Random Search or empirical tuning. Evaluation metrics include model performance and computational performance. An ablation study isolates the contribution of key feature groups and architectural components, confirming that the inclusion of temporal dynamics, spatial variation, and noise-handling features significantly enhances model robustness. Additionally, the removal of the residual correction stage results in a substantial decline in performance, highlighting its critical role in capturing structured error patterns and improving predictive accuracy under noisy and variable conditions.

1.1 Problem Formulation

Occupant estimation in indoor environments aims to predict the number of occupant present within a spatial zone using indirect sensory observations. One promising approach is to leverage WiFi Channel State Information (CSI) signals, which capture the fine-grained characteristics of wireless channel propagation, such as amplitude and phase variations across subcarriers [20]. The key advantage of CSI-based sensing is that it requires no specialized hardware beyond existing WiFi infrastructure, enabling scalable and cost-effective deployment for smart building systems.

Let us consider a scenario where the true number of occupants at a specific time step is represented by a variable denoting the actual occupant count. Correspondingly, the input at that same time step is a feature vector composed of values derived from Channel State Information (CSI) amplitude signals along with spatial metadata, such as transmitter-receiver locations or subcarrier indices. The problem of estimating occupancy can therefore be formally described as the task of learning a predictive function that maps this high-dimensional input space to a real-valued output representing the estimated occupant count. The objective is to approximate an unknown underlying function that defines the relationship between CSI-derived features and actual occupancy. This relationship can be expressed as:

$$y_t = f(x_t) + \epsilon_t \qquad (1)$$

Where:

- $f: R^d \rightarrow R$ is the unknown true function mapping the feature space to the occupant count,
- ϵ_t is an irreducible noise term representing environmental effects, device imperfections, and modeling errors, assumed to be zero-mean with finite variance ($E[\epsilon_t] = 0$).

However, accurately learning f from CSI signals presents several fundamental challenges. First, WiFi signals in indoor environments are affected by multipath propagation, where reflections from walls, furniture, and human bodies create complex and nonlinear distortions [21]. Second, occupant movement results in temporal variability [22], rendering the relationship between CSI signals and occupant highly dynamic and non-stationary [23]. Third, in under-actuated zones, where sensor density is low and airflow is unevenly distributed, signal coverage is spatially heterogeneous [24], further exacerbating ambiguity in CSI interpretation. Lastly, the presence of environmental noise, originating from static obstacles or nearby wireless devices, adds further complexity to direct modeling efforts [25]. To address these challenges, we propose a two-stage modeling strategy based on residual learning. In the first stage, a base predictive model $\hat{f}(x_t)$ is trained to capture the dominant occupant trends from the feature set. However, as is typical in complex, noisy environments, systematic residual errors may persist. To address this, a secondary model, denoted as

$\hat{g}(x_t)$, is introduced to learn the residuals, defined as follows:

$$r_t = y_t - \hat{f}(x_t) \quad (2)$$

The final corrected prediction $\hat{y}_t$ is then obtained by summing the base prediction and the residual estimation [26]:

$$\hat{y}_t = \hat{f}(x_t) + \hat{g}(x_t) \quad (3)$$

This decomposition acknowledges that residuals, rather than being treated as pure random noise, often contain structured patterns that can be exploited to improve model accuracy and generalization. The theoretical justification for this two-stage approach is rooted in the bias-variance decomposition of prediction error [27]. The expected mean squared error (MSE) of a learning model can be expressed as:

$$MSE(\hat{y}_t) = Bias^{2}(\hat{y}_t) + Var(\hat{y}_t) + \sigma^2 \quad (4)$$

Where $Bias^2$ represents the error due to erroneous assumptions in the learning algorithm, Variance reflects sensitivity to small fluctuations in the training data, and σ^2 denotes the irreducible error stemming from intrinsic noise. To overcome these limitations, an effective occupant estimation framework should be capable of modeling not only the dominant trends in CSI features but also the structured residual errors that arise from complex environmental dynamics. In this context, a two-stage modeling strategy where an initial prediction model is refined by explicitly learning its residual errors offers a theoretically grounded solution to improve both bias and variance trade-offs.

1.2 Dataset Description

In this study, the proposed occupancy estimation model is evaluated using two distinct datasets to ensure both robustness and generalizability. The first dataset is the CSI People Counting dataset, published by Mowla et al. [28], provides rich spatial granularity for object localization scenarios, which offers temporally annotated CSI sequences for occupancy regression. This dataset provides finely annotated CSI measurements that capture environmental dynamics under varying human presence. This dataset provides annotated CSI sequences collected using two WiFi transceivers (2Tx-2Rx) in a controlled indoor environment for the purpose of people counting via passive sensing. Each recording corresponds to a specific number of people (1 to 5), where for each frame includes:

- 224 complex CSI values per time instant (2 Tx × 2 Rx × 56 subcarriers).
- Each complex value is decomposed into: real, imag, amplitude, and phase.
- A person_count label from 1 to 5 is assigned to each data segment.
- A zone label (either 1 or 2) based on amplitude-derived proximity grouping (e.g., near/far from transmitter).
- A source_label (e.g., 2b, 4c) that uniquely identifies the recording file and can be mapped to a simulated spatial coordinate on a 6×24 grid (row A–F, column 0–23) for spatial visualization and contextual inference.

To enhance spatial awareness in modeling indoor wireless environments, a new feature termed Zone was introduced through unsupervised learning techniques. Rather than relying on a binary classification based on fixed distance thresholds, this approach leverages the inherent patterns within the Channel State Information (CSI) amplitude data to define spatial zones more accurately. Subsequently, the transformed data was clustered using K-Means clustering into four distinct zones, each representing a unique combination of spatial proximity, obstruction characteristics, and multipath effects commonly encountered in complex indoor settings. This methodology aligns with real-world under-actuated environments where HVAC distribution and sensor coverage are not uniformly deployed, resulting in heterogeneous thermal and airflow behavior across different spatial regions.

The spatial layout of the environment, as depicted in Figure 1, illustrates the output of this clustering process using a segmented grid. Each grid cell corresponds to a physical location where CSI data was recorded from specific transmitter-receiver positions. The cells are visually distinguished using colour codes representing the four derived clusters: light red for Zone 1, light green for Zone 2, light blue for Zone 3, and light yellow for Zone 4. These zones do not merely reflect physical distance but encapsulate the propagation behavior and environmental interactions observed through CSI measurements. Each grid cell contains a bold black label denoting the source_label, such as 1a, 2b, or 3c, identifying the specific placement of transceiver pairs during data collection. This visualization aids in understanding how signal characteristics vary spatially across the grid, capturing subtle differences in signal strength and pattern that result from structural elements.

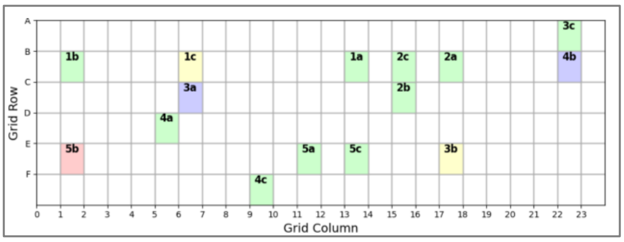

Fig 1. Mapping of Source Labels to Grid

Although the Mowla dataset offers controlled variability that is suitable for structured model development, it does not adequately capture the stochastic and dynamic characteristics often encountered in real-world indoor environments. This second dataset, known as the RSSI-Assisted CSI-Based Passenger Counting Dataset proposed by Guo et al. [29], provides a more realistic and diverse representation of signal behavior in uncontrolled settings.

The dataset was collected in mixed scenarios spanning three distinct spatial regions within a public transportation environment specifically the front (Zone 1), middle (Zone 2), and rear (Zone 3) of the upper deck of a double-decker vehicle. This spatial distribution was intentionally selected to capture a wide range of wireless propagation conditions influenced by passenger movement, environmental clutter, and the asymmetric layout of the vehicle interior. In contrast to the Mowla dataset, which employs a structured experimental setup with predefined zoning and controlled occupant placement, this dataset does not impose strict spatial or numerical constraints on the presence of individuals near the Wi-Fi transceivers. Instead, it captures a more realistic setting characterized by heterogeneous human activity, spontaneous movement patterns, and intermittent signal interference from surrounding devices. This creates a rich and representative environment for evaluating the performance of wireless sensing systems under conditions that closely mirror real-world deployments.

CSI data were collected using multiple Wi-Fi receivers strategically positioned throughout the upper deck. Each packet contains CSI measurements across 256 subcarriers, with non-informative subcarriers removed during preprocessing, resulting in 242 subcarrier values per frame. For each packet, both the amplitude and phase components of the wireless channel were extracted and stored, forming the foundational features used for downstream processing and modeling. The raw features in the dataset include the following:

- **CSI Amplitude**: Derived from the real and imaginary parts, the amplitude reflects the magnitude of the channel response.
- **CSI Phase**: Also derived from the real and imaginary components, the phase represents the angular component of the channel response.
- **Frame Index**: Each CSI measurement is associated with a sequential frame (packet) number, indicating the temporal order in which the data was captured.
- **Subcarrier Index**: CSI values are measured across multiple subcarriers within each Wi-Fi packet. Each subcarrier index identifies the frequency component of the signal within the OFDM transmission.

Figure 2 illustrates the distribution of person counts across three spatial zones based on the number of CSI records collected. In Zone 1, the distribution is relatively balanced across all person count categories (1, 3, 4, 5, 6, 8, and 10), with each category contributing a comparable volume of data. This even representation indicates that Zone 1 provides comprehensive and consistent occupancy coverage, making it highly suitable for training generalized machine learning models.

In Zone 2, the distribution remains broad but more variable. Certain person counts particularly 4, 6, 8, and 10 dominate in frequency, while others such as 3 are significantly underrepresented. This skew introduces some class imbalance, which may influence model sensitivity to underrepresented occupancy levels unless addressed during training.

In contrast, Zone 3 displays a markedly skewed distribution. Person counts 3 and 10 are highly prominent, especially count 3, which is substantially overrepresented. Meanwhile, lower counts such as 1, 4, and 5 are present in much smaller quantities, and person count 10 is nearly absent. This imbalance suggests that Zone 3 may introduce biased learning patterns if not carefully weighted or augmented.

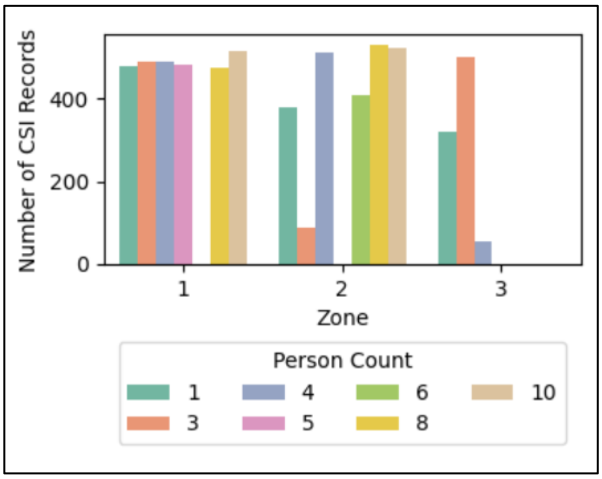

Fig 2. Distribution Person Count per Zone

While raw amplitude data provides foundational information about channel variations, it lacks interpretability and often underrepresents latent signal dynamics associated with real object position. Therefore, a systematic feature engineering process was applied to enrich the dataset. The engineered features, describes in Table 1, are carefully crafted to enhance the raw signal representation by incorporating domain knowledge from wireless communication. This pipeline adheres to statistical learning principles [30], and helps mitigate noise, overcome non-linearity, and improve signal discriminability across spatial zones.

To enhance the contextual representation and learning capacity of the model, two additional features include person_count was manually incorporated, each tailored to the specific structure and objective of the respective datasets. By integrating this count into the dataset, the model is better equipped to generalize to multi-object detection scenarios, allowing it to distinguish between varying object densities and improve spatial inference capabilities. In contrast, the feature person_count is embedded within the CSI People Counting dataset and serves as the primary regression target in this study. Each CSI sample is associated with a known number of human occupants, ranging from 1 to 5, based on experimental recordings.

To address the inherent complexity of Channel State Information (CSI) signals including multipath propagation, signal noise, non-linearity, temporal dynamics, and spatial variation a targeted set of engineered features was constructed, as detailed in Table 1. The CSI_RollingMean, computed with a window of 5, plays a critical role in mitigating multipath effects by smoothing high-frequency signal fluctuations and capturing localized amplitude trends indicative of human presence. The CSI_Amplitude_sq feature enhances the model's sensitivity to signal distortion by amplifying non-linear amplitude deviations typically observed near reflective surfaces or moving occupants. Temporal dynamics are captured through CSI_Delta, CSI_Lag1, CSI_Lag2, and CSI_LagDiff, which model short-term signal memory and changes over time useful for detecting motion transitions and occupancy fluctuations.

Table 1. Feature Engineering

Feature Name	**Description**
CSI_RollingMean	Rolling mean of CSI amplitude (window=5); captures local trends
CSI_Lag1	CSI amplitude with a one-step lag
CSI_Lag2	CSI amplitude with two-step lag
CSI_Phase_Diff	Phase difference across adjacent subcarriers
CSIxNodes	Interaction between CSI amplitude and number of nodes
CSIxZone	Interaction between CSI amplitude and zone complexity
Node_Zone_Interact	Encoded composite of node pair and zone ID for spatial awareness
Zone	Zone classification (1: near, 2: far), based on average CSI amplitude
person_count	Ground-truth number of human occupants (regression target)

To account for spatial variation, the Zone feature encodes proximity to the WiFi transceiver pair (near vs. far), while CSIxZone, CSIxNodes, Node_Zone_Interact, and Zone incorporate contextual interactions between amplitude, network topology, and spatial complexity. Collectively, these features enrich the raw CSI data with interpretable descriptors that improve robustness and accuracy in occupancy estimation under real-world indoor conditions.

The integration of the Guo et al. dataset into the evaluation framework was undertaken to assess the generalizability of the proposed model under realistic deployment conditions. Through this addition, the model's robustness was examined across varying hardware configurations, ambient signal noise, and environmental drift factors that frequently affect the performance of sensing systems in practical applications. In contrast to the first dataset, in which each CSI frame was manually labeled with a specific zone and known occupant count, the second dataset was characterized by real-world complexities, including irregular spatial distributions, spontaneous interference, and continuous transitions in occupant movement, all without predefined spatial annotations.

By employing both datasets, a more comprehensive evaluation was achieved: the Mowla dataset supported supervised training and sensitivity analyses based on controlled zoning, whereas the Guo et al. dataset facilitated model validation under unstructured, naturally occurring indoor conditions. This dual-dataset approach ensured a rigorous assessment of the model's adaptability, reliability, and readiness for real-world implementation.

1.3 Feature Filtering

In this study, signal filtering is a critical preprocessing step aimed at mitigating the adverse effects of noise, multipath interference, and high-frequency distortions commonly present in raw Channel State Information (CSI) measurements. Due to the nature of wireless signal propagation in indoor environments, CSI signals are inherently unstable and susceptible to fluctuations caused by reflections, refractions, and environmental dynamics. These fluctuations introduce high-frequency noise components that can obscure meaningful signal patterns relevant to occupancy detection and spatial reasoning.

To address this issue, a low-pass Butterworth filter is applied across all CSI amplitude sequences prior to feature extraction. The Butterworth filter is selected for its maximally flat frequency response in the passband, which ensures that important low-frequency components of the signal typically associated with gradual changes in the propagation environment are preserved while high-frequency noise is effectively attenuated [31].

1.4 Feature Dimensionality Reduction

Given the high dimensionality introduced by both raw Channel State Information (CSI) measurements and engineered signal-based features, a dimensionality reduction strategy is necessary to ensure robust generalization, minimize overfitting, and enhance model efficiency in non-stationary and noisy indoor environments. To address this, the proposed feature dimensionality reduction method employs Principal Component Analysis (PCA), a statistically grounded technique for projecting high-dimensional data into a lower-dimensional subspace while retaining the majority of its variance [32]. In this study, the number of retained components was determined by preserving 95% of the total variance in the dataset. This threshold ensures that most of the structural information within the feature space is preserved while effectively reducing dimensionality.

1.5 Proposed Model

To overcome the challenges of CSI-based occupancy estimation in under-actuated zones, this study introduces a two-stage predictive architecture called the Gradient Boosting-Based Residual Correction Model (GBRCM). Unlike conventional ensemble models that rely on a single-stage learning process, GBRCM is designed to explicitly learn and correct systematic prediction errors through residual modeling. This enables the model to improve generalization in the presence of complex, nonlinear distortions inherent in WiFi CSI data. The full process is illustrated in Figure 3.

The architectural design of GBRCM is motivated not only by general regression needs but also by the unique challenges of CSI-based learning. In the first stage, the Gradient Boosting Regressor (GBR) captures dominant occupancy signals. However, due to multipath reflections and signal variability, raw predictions often carry residual errors that are structured not random. The Histogram-Based Gradient Boosting Regressor (HGBR) in the second stage is designed to model these structured residuals, especially those arising from zone specific attenuation, multipath reflections, and non-stationary channel behavior. This residual correction allows GBRCM to effectively compensate for non-linearity, noise interference, and spatial amplitude heterogeneity, which are pervasive in real-world indoor sensing scenarios. The effectiveness of this residual correction is later demonstrated through the ablation study where the HGBR layer significantly improves accuracy and generalization.

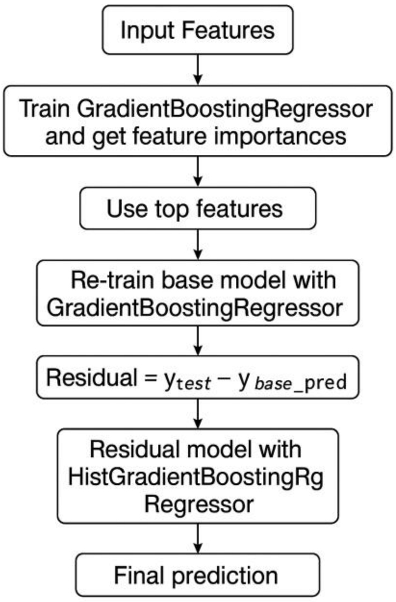

Fig. 3 Model Architecture of GBRCM

Let us denote the training dataset as a collection of input-output pairs, where each input consists of a feature vector derived from Channel State Information (CSI), and each output corresponds to the actual number of occupants or objects. Specifically, this dataset contains *n* observations, with each feature vector representing a set of numerical values extracted from CSI signals, and each corresponding label denoting the ground truth count.

The first stage of the proposed model employs a base learner using Gradient Boosting Regression (GBR). This base model generates an initial estimate of the occupancy count based on the input features. However, it is well documented that gradient boosting models can exhibit bias when applied to data characterized by high noise levels and non-linear patterns, as is often the case with CSI-derived signals [33]. To address this limitation, we introduce a residual correction mechanism.

$$r_i = y_i - \hat{y}_i^{GBR} \quad (7)$$

After obtaining the initial predictions from the GBR model, we calculate the residuals, which are defined as the difference between the actual values and the predicted outputs. These residuals reflect the unexplained portion of the data specifically, the prediction errors that may be systematically linked to complex signal behaviors not fully captured by the initial model. A second model, the Histogram-Based Gradient Boosting Regressor (HGBR), is then trained to approximate the residual function:

$$\hat{r}_i = f_{HGBR}(x_i) \quad (8)$$

The final prediction of the GBRCM is obtained by summing the outputs of both models:

$$\hat{y}_i^{GBRCM} = \hat{y}_i^{GBR} + \hat{r}_i \quad (9)$$

This two-stage correction mechanism is grounded in residual learning theory and aligns with recent research showing the effectiveness of multi-level gradient boosting for structured error correction in wireless signal domains [34]. Furthermore, the selection of HGBR as the residual learner offers two key advantages. First, it improves generalization on tabular datasets due to its histogram-based binning and regularized splitting strategy. Second, it ensures robustness to nonlinearity and high-frequency perturbations in CSI amplitudes, which are common in indoor multipath environments [35].

By explicitly modeling the residual error patterns, GBRCM offers a principled approach to enhance the predictive capacity of machine learning in under-actuated zones with limited sensor coverage. This formulation also allows for interpretability and post hoc error analysis, addressing concerns regarding bias-variance decomposition, generalization, and systematic error correction [36].

1.6 State-of-the-art Model

To establish a rigorous benchmark for the proposed GBRCM model, we implement and evaluate a set of widely adopted state-of-the-art machine learning models for CSI-based human presence and object estimation. These models span classical neural architectures and have been successfully applied in similar wireless sensing contexts such as occupancy detection, motion classification, and passive localization. The models are trained and tested using the same CSI-derived features and evaluation pipeline to ensure fairness and consistency.

In this study, three widely adopted state-of-the-art deep learning models were employed as benchmarks to evaluate the performance of the proposed GBRCM model: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Random Forest (RF), and Long Short-Term Memory (LSTM). The MLP, a fully connected feedforward neural network, was selected for its effectiveness in modeling non-linear relationships in tabular CSI-derived features. However, MLPs are inherently limited in their ability to model temporal dynamics or spatial locality unless supplemented with time-windowed or contextual features, which constrains their effectiveness in more dynamic or multi-zone environments. Due to its simplicity and flexibility, MLPs are frequently used in CSI-based regression tasks where spatial or temporal structures are not explicitly modelled [8]. The CNN model, on the other hand, was included to capture spatial dependencies across CSI amplitude sequences.

By applying one-dimensional convolutions over the sequential subcarrier amplitudes, the CNN is capable of identifying localized patterns introduced by multipath propagation and object-induced signal distortions [9]. The Random Forest model was also evaluated due to its robustness in handling high-dimensional, noisy data an inherent characteristic of CSI signals. As an ensemble of decision trees, RF effectively captures non-linear feature interactions and provides built-in feature importance metrics, offering interpretability and resilience against overfitting [12].

Lastly, the Long Short-Term Memory (LSTM) is designed to retain long-term dependencies and temporal memory, making it well-suited for dynamic indoor environments where occupant movement alters signal patterns over time [14]. Despite its higher computational cost, LSTM is considered a strong baseline in time-series modeling for CSI-based sensing applications due to its ability to learn temporal correlations directly from the data without handcrafted temporal features.

1.7 Hyperparameter Tuning Strategy

To enhance model accuracy and ensure robust generalization, hyperparameter tuning was performed for each machine learning model using randomized search with 5-fold cross-validation. Each model required a tailored strategy based on its architecture and learning dynamics. For the Multilayer Perceptron (MLP), tuning focused on both network architecture and training dynamics. The number of hidden layers (1–3), neurons per layer (32–128), dropout rate (0.1–0.5), batch size (16–64), learning rate (0.001–0.01), and activation function (ReLU or Leaky ReLU) were tuned to balance model complexity and convergence stability. In the case of the Convolutional Neural Network (CNN), key parameters included the number of convolutional filters (32–128), kernel size (3–7), and number of convolutional blocks (1–3). Additional components such as batch normalization and dropout regularization were also incorporated. For the Random Forest (RF) model, the number of estimators (100–500), maximum tree depth (10–50), minimum samples per leaf (1–5), and maximum features considered at each split (auto, sqrt, log2) were tuned. the Long Short-Term Memory (LSTM) model, the search space covered the number of LSTM units (32–128), dropout rates (0.1–0.5), dense layer sizes (16–64), learning rate (1e-4 to 1e-2), and batch size (16–64).

The proposed GBRCM model, composed of a base Gradient Boosting Regressor (GBR) followed by a Histogram-based Gradient Boosting Regressor (HGBR) for residual correction, was tuned in two stages. The base GBR was tuned for the number of estimators (100–300), learning rate (0.01–0.1), and tree depth (3–10). Subsequently, the HGBR was tuned for iteration count (100–200), learning rate (0.01–0.1), and maximum tree depth (3–10) to best model the residual error patterns.

This tuning strategy ensures fair benchmarking across models by adapting to each model's representational capacity and regularization needs. The adoption of randomized search allows efficient exploration of large hyperparameter spaces, while cross-validation provides robust performance estimates and reduces the risk of overfitting on training data.

1.8 Model Evaluation Metrics

To comprehensively assess model effectiveness, three standard regression metrics were utilized: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) [37]. RMSE captures the sensitivity of the model to large errors, while MAE reflects the average magnitude of residuals regardless of direction. R^2 measures the proportion target variable explained by the model, thereby indicating its explanatory power. In addition to these regression metrics, the model's training time, testing time, and inference time were also evaluated. These performance measures provide insights into the computational efficiency and real-world applicability of the models [38].

1.9 Ablation Study

To assess the relative contribution of each component in the proposed GBRCM architecture, we conducted a systematic ablation study. Regarding the handling of domain-specific challenges in Channel State Information (CSI) including multipath interference, signal noise, temporal dynamics, spatial variation, and without residual correction, we extended the ablation study to include the removal of specific feature subsets that were explicitly designed to mitigate these issues. This targeted analysis provides a quantitative understanding of how each feature group contributes to overall model robustness and accuracy. The following ablation configurations were evaluated:

- **Without Residual Correction:** In this configuration, we removed the correction stage entirely, thereby disabling the model's second-level learning mechanism that captures residual error patterns missed by the base regressor. This component is critical for learning structured residuals arising from nonlinear signal behavior and unmodeled context dependencies.
- **Multipath Interference and Noise**: To evaluate the role of temporal smoothing in suppressing rapid CSI fluctuations caused by multipath propagation, we excluded all CSI_RollingMean features across subcarriers and transceiver pairs. These rolling mean features are designed to capture short-term trends and attenuate high-frequency variations, making them particularly relevant for reducing measurement instability in multipath-dominant environments.
- **Temporal Dynamics**: This ablation targets the model's dependence on sequential temporal features by excluding CSI_Lag1, CSI_Lag2, and CSI_Delta. These features encode the one-step and two-step history of CSI amplitudes, along with their first-order differences, and collectively allow the model to recognize temporal trends and capture movement-related changes in signal patterns.
- **Spatial Variation**: Finally, we excluded contextual interaction features such as CSIxZone, Node_Zone_Interact, and the manually derived Zone label. These features embed spatial context derived from signal strength gradients and antenna location awareness, which are vital in under-actuated zones.

This extended ablation confirms that the integration of challenge-specific features is not merely additive but foundational to maintaining performance in practical CSI-based occupancy detection tasks. Each ablation setting is compared to the full GBRCM model, which includes all three components.

F. HASIL PENELITIAN

Tuliskan secara ringkas hasil pelaksanaan penelitian yang telah dicapai sesuai tahun pelaksanaan penelitian. Penyajian meliputi data, hasil analisis, dan capaian luaran (wajib dan atau tambahan). Seluruh hasil atau capaian yang dilaporkan harus berkaitan dengan tahapan pelaksanaan penelitian sebagaimana direncanakan pada proposal. Penyajian data dapat berupa gambar, tabel, grafik, dan sejenisnya, serta analisis didukung dengan sumber pustaka primer yang relevan dan terkini.

2. Result and Discussion

The analysis begins with a comprehensive evaluation of feature importance, derived through a dual-phase selection strategy. Subsequently, the performance of several machine learning models, including both proposed model and state-of-the-art model. To assess the robustness and individual contribution of core components, a structured ablation study is conducted. Furthermore, the contribution of high impact features to model performance is quantitatively analyzed to understand their relevance in capturing domain specific signal behaviors, including multipath effects, non-linearity, and temporal noise. The findings are discussed in the context of their implications for signal-based sensing and indoor environment control.

2.1 Feature Filtering

To evaluate the effectiveness of signal filtering in enhancing the stability and interpretability of Channel State Information (CSI) signals, a variance-based analysis was conducted on two datasets before and after the application of a low-pass Butterworth filter. The results are illustrated in Figures 4 and 5, which compare the distribution of variance across amplitude features in the raw and filtered signals for each dataset.

In Figure 4, representing the first dataset, the variance distribution of the raw amplitude features shows a wide interquartile range with extreme outliers, indicating significant fluctuations and instability in the unprocessed CSI signal. This instability stems from multipath propagation, environmental interference, and high-frequency noise that often contaminate raw CSI measurements.

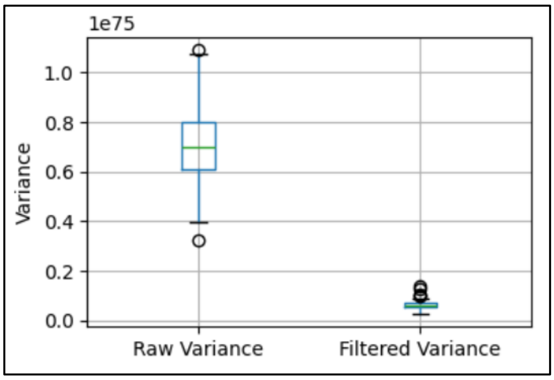

Fig. 4 Variance Distribution Before and After Filtering of 1st Dataset

After filtering, the variance distribution becomes substantially more compact, centered around a much lower range, with minimal presence of outliers. This shift signifies the effective attenuation of high-frequency noise by the Butterworth filter, preserving only the low frequency components associated with gradual environmental changes and human presence. Quantitatively, the total signal power drops from approximately 3.19×10^{80} in the raw signal to 3.46×10^{79} post filtering an order of magnitude reduction that validates the noise-suppression capability of the filter while maintaining signal integrity necessary for downstream analysis.

In contrast, Figure 5 illustrates the variance distribution of amplitude features in the second dataset, which differs substantially from the first. Unlike the raw dataset in Figure 4, this second dataset has undergone pre-processing at the source, where a low-pass filter was already applied to suppress high-frequency noise and environmental artifacts before further analysis.

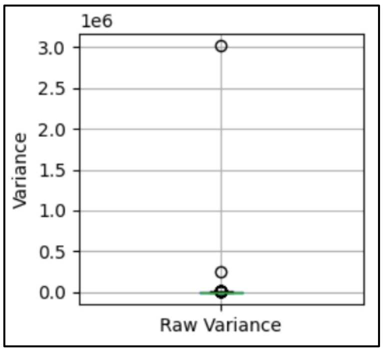

Fig. 5 Variance Distribution Before and After Filtering of 2nd Dataset

As a result, the variance distribution depicted in Figure 5 appears significantly more compact, with the interquartile range narrowed and extreme outliers largely eliminated. This inherently smoothed dataset suggests that the filtering process has successfully attenuated unstable signal components while retaining the low-frequency variations more closely tied to environmental dynamics and human movement. Quantitatively, the total signal power, computed from the squared amplitudes across all subcarriers, is approximately 1.7371×10^{13}. This high-power value reflects the continued presence of dominant signal components, though now primarily composed of interpretable and stable variations.

2.2 Feature Dimensionality Reduction

As shown in Figure 6, the cumulative explained variance curve demonstrates the proportion of total variance captured as a function of the number of principal components. The curve exhibits a typical diminishing returns pattern, where the initial components contribute most significantly to the overall variance, while subsequent components add only marginal improvements. A variance threshold of 95% was selected to determine the optimal number of principal components for downstream modeling. This threshold represents a balance between information retention and dimensionality reduction. The figure indicates that approximately 684 components are required to reach this threshold, beyond which additional components contribute minimally to the representation of the dataset.

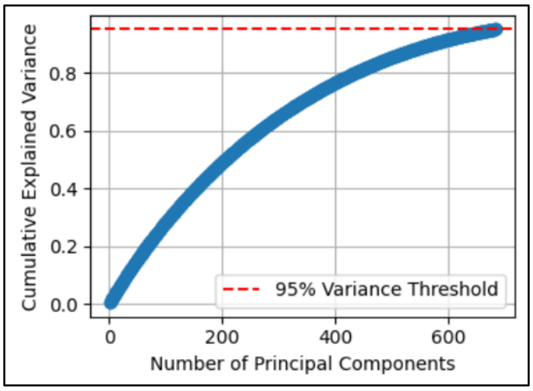

Fig. 6 Cumulative Explained Variance by PCA Components of 1st Dataset

In contrast, Figure 7 presents the cumulative explained variance resulting from Principal Component Analysis (PCA) applied to the second dataset, which has already undergone pre-filtering at the source level using a low pass filter. As a result, the input features are inherently smoother and less noisy, reflecting signal characteristics that have already been denoised and partially compressed prior to PCA.

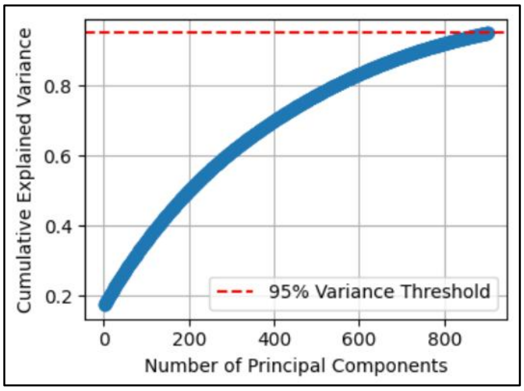

Fig. 7 Cumulative Explained Variance by PCA Components of 2nd Dataset

Despite the upstream filtering, the cumulative variance curve increases gradually and requires approximately 902 principal components to reach the 95% variance threshold. This indicates that while high frequency noise has been suppressed, the dataset still contains a substantial degree of variability across its remaining amplitude features likely due to retained signal dynamics from meaningful environmental or human induced changes.

2.3 Model Performance

To evaluate the predictive capability of the proposed Gradient Boosting-Based Residual Correction Model (GBRCM), a comprehensive comparison was conducted against several widely adopted state-of-the-art models, including Multilayer Perceptron (MLP) [8], Convolutional Neural Network (CNN) [9], Random Forest (RF) [12], and Long Short-Term Memory (LSTM) [14]. All models were trained and tested using the same feature-engineered dataset and underwent hyperparameter tuning via randomized search with 5-fold cross-validation to ensure a fair and robust comparison. Table 2 summarizes the model performance in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2) three standard regression metrics that collectively assess prediction accuracy, error magnitude, and model explanatory power.

Table 2 presents the optimal hyperparameter configurations for each evaluated model across the two feature-engineered datasets. For the Multilayer Perceptron (MLP), the optimal architecture for the first dataset includes a single hidden layer with 64 neurons, a dropout rate of 0.1, a learning rate of 0.001, and the ReLU activation function. In contrast, the second dataset required a higher model capacity with 128 neurons, a slightly increased learning rate of 0.002605, and the same dropout and activation settings, indicating the dataset's demand for deeper representation. The Convolutional Neural Network (CNN) achieved its best performance using a single convolutional block with 32 filters. While both datasets used a dropout rate of 0.1, the kernel size and learning rate were tuned differently kernel size of 3 and learning rate of 0.001 for the first dataset, versus a smaller kernel size of 2 and a much lower learning rate of 0.000147 for the second dataset, reflecting the sensitivity of CNNs to spatial resolution and learning stability.

The Random Forest (RF) model reached optimal performance with 200 estimators and a maximum depth of 20 in the first dataset, and 174 estimators with a deeper tree structure (depth=30) in the second dataset. This suggests that the second dataset benefits from more complex hierarchical splits, possibly due to higher feature interaction. For the Long Short-Term Memory (LSTM) model, the first dataset was best modelled with 128 LSTM units, a dropout rate of 0.2, and 48 dense units, while the second dataset required fewer LSTM units (96), a higher dropout rate of 0.3, and more dense units (96). The proposed GBRCM model combines a Gradient Boosting Regressor (GBR) as the base predictor and a HistGradientBoosting Regressor (HGBR) as the residual corrector. For the first dataset, the GBR component used 289 estimators, a learning rate of 0.0287, and a maximum depth of 8. For the second dataset, the optimal GBR configuration was more lightweight, with 108 estimators, a lower learning rate of 0.0113, and a reduced depth of 3. The residual component (HGBR) remained consistent across datasets with 150 maximum iterations, a learning rate of 0.05, and depth of 4.

Table 2. Result of Hyperparameter Tuning

Model	**1st Dataset**	**2nd Dataset**
MLP [8]	Hidden layers=1, neurons per layer= 64, dropout=0.1, learning rate=0.001, batch size=32, activation function=relu	Hidden layers=1, neurons per layer= 128, dropout=0.1, learning rate= 0.002605, batch size=32, activation function=relu
CNN [9]	1 conv blocks, 32 filters, kernel=3, dropout=0.1, learning rate=0.001	1 conv blocks, 32 filters, kernel=2, dropout=0.1, learning rate=0.000147
RF [12]	200 estimators, max depth=20, min leaf=2, max_features=none	174 estimators, max depth=30, min leaf=2, max_features=none
LSTM [14]	lstm units: 128, dropout rate: 0.2, dense units: 48, learning rate: 0.00734	lstm units: 96, dropout rate: 0.3, dense units: 96, learning rate: 0.00281
GBRCM	GBR (289 estimators, learning rate =0.0287, depth=8); and HGBR (150 max iteration, learning rate=0.05, depth=4).	GBR (108 estimators, learning rate =0.0113, depth=3); and HGBR (150 max iteration, learning rate=0.05, depth=4).

Table 3 presents a detailed comparison of the predictive performance across five machine learning models on two distinct datasets using three standard regression evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics collectively assess prediction accuracy, the magnitude of errors, and the proportion of variance explained by the model. The two datasets represent contrasting signal environments: the first being a controlled dataset with structured zoning and labeled occupancy, and the second being a more realistic, noisy dataset collected from an uncontrolled public space.

On the first dataset, the proposed Gradient Boosting-Based Residual Correction Model (GBRCM) significantly outperforms all baseline models, achieving an exceptionally low MAE of 0.0422 and an RMSE of 0.0579, with an R^2 of 0.9989. These results indicate not only extremely high prediction accuracy but also minimal variance in the residuals, highlighting the model's capability to capture nearly all the variance in the ground truth occupancy labels. In contrast, traditional models such as the Multilayer Perceptron (MLP) and Random Forest (RF) exhibit relatively higher errors (MAE > 0.35 and RMSE > 0.48), and their R^2 scores remain below 0.89, suggesting limited capability in modeling the nonlinear signal-to-occupancy relationship. The Convolutional Neural Network (CNN) performs better than MLP and RF, attaining an R^2 of 0.9587, yet still lags behind the GBRCM. The Long Short-Term Memory (LSTM) model yields the poorest performance on the first dataset, with an R^2 of 0.7393, suggesting that despite its temporal modeling strengths, it may be less suited for highly structured spatial datasets without additional optimization.

Performance disparities are even more pronounced on the second dataset, which presents more complex environmental variability and signal noise. Here, the GBRCM again demonstrates superior generalization, achieving an MAE of 1.0933, an RMSE of 1.3270, and an R^2 of 0.8199. These results reinforce the model's robustness in capturing meaningful patterns even in stochastic real-world conditions. In contrast, all other models exhibit severe degradation in performance. The MLP and LSTM models perform particularly poorly, with R^2 scores of -0.1192 and -0.0914 respectively, indicating that these models fail to explain the variance in the data and effectively perform worse than simply predicting the mean. CNN and RF show moderate robustness with positive R^2 scores (0.0042 and 0.1316 respectively), but their error metrics remain substantially higher than GBRCM. The CNN's R^2 close to zero suggests that while it may capture localized patterns, its performance remains insufficient under significant signal interference and non-stationary conditions.

These results provide strong empirical validation of the GBRCM's superiority over conventional machine learning and deep learning baselines in both controlled and uncontrolled environments. The model's two-stage architecture, which corrects structured residuals after initial predictions, clearly enhances its ability to generalize across varying signal complexities and indoor configurations. This performance margin affirms GBRCM as a highly reliable model for CSI-based occupancy estimation in under-actuated zones.

Table 3. Model Performance Comparison

Model	**1st Dataset**			**2nd Dataset**		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
MLP [8]	0.3578	0.5087	0.8735	2.7025	3.3077	-0.1192
CNN [9]	0.2282	0.2906	0.9587	2.5815	3.1200	0.0042
RF [12]	0.4081	0.4876	0.8838	2.4370	2.9136	0.1316
LSTM [14]	0.5576	0.7303	0.7393	3.2664	2.6876	-0.0914
GBRCM	**0.0422**	**0.0579**	**0.9989**	**1.0933**	**1.3270**	**0.8199**

Table 4 provides a comparative analysis of the computational efficiency of the evaluated models, assessed across three key dimensions: total training time, total testing time, and per-sample inference time. These metrics offer insight into the scalability, latency, and real-time applicability of each model, particularly in resource-constrained deployment environments such as embedded systems or edge-based smart building platforms. On the first dataset, which is relatively smaller and collected under controlled conditions, the proposed Gradient Boosting-Based Residual Correction Model (GBRCM) demonstrates highly competitive computational efficiency. It achieves the lowest inference time per sample at just 0.0196 milliseconds, outperforming all other models including the Random Forest (RF), which recorded 0.0365 milliseconds, and the Multilayer Perceptron (MLP), which required over ten times more inference time at 0.2158 milliseconds. In terms of training duration, GBRCM requires 36.67 seconds slightly longer than MLP and CNN but substantially faster than the LSTM, which took over 87 seconds. Testing time is also minimal for GBRCM at only 0.0078 seconds, which is an order of magnitude lower than most alternatives. These results highlight GBRCM's suitability for real-time inference, especially in latency-sensitive applications.

Table 4. Model Computational Performance Comparison

Model	**1st Dataset**			**2nd Dataset**		
	Training Time (s)	**Testing Time (s)**	**Inference Time (ms/sample)**	**Training Time (s)**	**Testing Time (s)**	**Inference Time (ms/sample)**
MLP [8]	4.0847	0.0852	0.2158	94.6531	1.0490	0.8392
CNN [9]	8.1113	0.1075	0.2722	34.8197	0.1711	0.1369
RF [12]	31.2566	0.0144	0.0365	1131.0032	0.0784	0.0627
LSTM [14]	87.2544	0.4554	1.1528	2341.2258	6.6271	5.3017
GBRCM	**36.6730**	**0.0078**	**0.0196**	**108.2679**	**0.0138**	**0.0110**

The computational advantages of GBRCM become even more prominent on the second dataset, which is significantly larger and was collected in uncontrolled, real-world settings. Here, GBRCM maintains a low training time of 108.27 seconds, far outperforming RF (1131.00 seconds) and LSTM (2341.23 seconds), both of which suffer from scalability issues due to increased data volume and model complexity. The testing time and inference time per sample for GBRCM remain minimal at 0.0138 seconds and 0.0110 milliseconds, respectively again establishing it as the most computationally efficient model for real-world deployment. In contrast, LSTM exhibits the highest computational cost, with an inference time of 5.3017 milliseconds per sample, making it unsuitable for scenarios requiring rapid occupancy estimation or frequent updates. The MLP, while efficient in controlled settings, becomes significantly slower on the second dataset, with inference time increasing to 0.8392 milliseconds, suggesting sensitivity to dataset scale and complexity.

These results indicate that GBRCM not only delivers high prediction accuracy but also achieves superior efficiency in both training and inference phases. Its combination of low-latency prediction and reasonable training overhead makes it well suited for practical deployment in under-actuated zones where real-time, energy-efficient occupancy estimation is required. The balance between computational scalability and predictive robustness further strengthens its position as a viable solution for smart sensing in dynamic and resource-limited environments.

2.4 Ablation Study

To evaluate the role of individual feature groups and architectural components in the proposed Gradient Boosting-Based Residual Correction Model (GBRCM), a systematic ablation study was conducted. This analysis isolates the contributions of engineered features designed to address domain-specific challenges in Channel State Information (CSI) modeling including multipath interference, temporal variability, and spatial heterogeneity as well as the critical impact of the model's residual correction mechanism. Each ablation configuration involves the removal of one component or feature group while retaining all other elements and training the GBRCM under the same conditions and hyperparameters. The performance is then compared to the full GBRCM model trained with all engineered features and correction stages. Table 5 presents the results across four ablation types compared to the full model baseline.

In the full model configuration, GBRCM achieved outstanding accuracy on both datasets, with an MAE of 0.0470, RMSE of 0.0341, and R^2 of 0.9989 on the first dataset, indicating near-perfect fit and with substantially lower error on the second dataset compared to all baseline models (MAE=1.0933, RMSE=1.3270, R^2=0.8199). These results establish a strong performance baseline for comparison. It is important to note that the first dataset was collected under controlled experimental conditions, where the number of occupants was precisely managed within a fixed range (1 to 5 individuals), and the spatial layout was carefully predefined. This setup reduces uncontrolled variance, allowing the model to focus on clean signal-to-occupancy mappings with minimal environmental noise. In contrast, the second dataset represents a real-world deployment scenario involving spontaneous human activity across different zones in a public transport environment. Here, the number of occupants can vary significantly and unpredictably across spatial regions, leading to high inter-zone variability and introducing natural noise and fluctuation that make the estimation task considerably more difficult.

Table 5. Ablation Study Results

	1st Dataset			2nd Dataset		
Model	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
GBRCM (Full Model+Feature Engineering)	**0.0470**	**0.0341**	**0.9989**	**1.0933**	**1.3270**	**0.8199**
GBRCM Without Correction Stage	0.2427	0.3154	0.9514	2.9218	2.4228	0.1267
GBRCM+Without Multipath and Noise	0.0476	0.0667	0.9978	2.0286	2.4307	0.3955
GBRCM+Without Temporal Dynamics	0.0465	0.0655	0.9979	1.9981	2.3954	0.4131
GBRCM+Without Spatial Variation	0.0421	0.0612	0.9982	1.9854	2.3769	0.4221

The most significant degradation in performance occurs when the residual correction stage is removed entirely. Without this second-stage learning mechanism, the model's ability to capture structured errors is compromised. On the first dataset, MAE increases more than fivefold to 0.2427, RMSE escalates to 0.3154, and R^2 drops to 0.9514. The second dataset shows even more severe deterioration, with MAE rising to 2.9218 and R^2 plummeting to just 0.1267. This confirms that residual learning is not merely a refinement but a critical component that enables the model to recover from systematic errors arising from non-linearity, noise, and environmental drift, especially in unconstrained real-world conditions where ground-truth labeling is inherently noisier. The exclusion of features associated with multipath and noise handling particularly the `CSI_RollingMean` also results in substantial performance degradation. On the second dataset, RMSE nearly doubles (2.4307) compared to the full model (1.3270), and R^2 declines significantly to 0.3955. These results highlight the importance of temporal smoothing features, which are essential for suppressing high-frequency distortions caused by signal reflection and interference in multipath-dominant environments. Their removal undermines the model's ability to extract stable signal patterns, especially in real-world settings where volatility is high.

The ablation of temporal dynamics through the removal of lag-based features such as `CSI_Lag1`, `CSI_Lag2`, and `CSI_Delta` leads to similar but slightly less pronounced performance degradation. The RMSE on the second dataset increases to 2.3954, and R^2 decreases to 0.4131, suggesting that short-term memory features play an important but secondary role in capturing human motion-induced signal transitions. While the model still retains reasonable predictive capacity, the absence of these features reduces its temporal sensitivity, resulting in a less responsive mapping between signal fluctuations and occupancy changes.

In the spatial variation ablation, contextual features such as `CSIxZone`, `Node_Zone_Interact`, and the manually clustered `Zone` label were excluded. Although this removal has minimal impact on the first dataset where controlled spatial zoning may reduce the importance of contextual interaction, the second dataset experiences notable performance loss (MAE=1.9854, RMSE=2.3769, R^2=0.4221). This underscores the importance of spatial features in heterogeneous environments where WiFi signal behavior varies across zones due to obstacles, uneven coverage, or crowding. Without such features, the model becomes less capable of distinguishing between location-dependent occupancy patterns.

In summary, the ablation study reveals several critical insights. First, the residual correction stage is foundational to the GBRCM architecture, enabling it to generalize in noisy, non-stationary environments. Second, features targeting multipath and noise artifacts, particularly `CSI_RollingMean`, are indispensable for stabilizing predictions and capturing true occupancy-related signal structure. Third, temporal and spatial features contribute meaningfully to accuracy and generalization, particularly in complex or uncontrolled settings. However, the relatively smaller impact of removing these groups suggests that some redundancy exists across feature representations, and that the residual learning component mitigates performance loss by recovering structured residuals.

2.5 Discussion

This study was designed to evaluate three primary research hypotheses: (1) that carefully engineered features could effectively address domain-specific challenges in CSI-based occupancy estimation; (2) that a residual-aware two-stage learning architecture would significantly outperform conventional single-stage models; and (3) that the proposed model would demonstrate robustness and generalizability in noisy, non-linear, and under-actuated indoor environments.

Evaluation of the proposed Gradient Boosting-Based Residual Correction Model (GBRCM) against state-of-the-art models including MLP, CNN, LSTM, and Random Forest confirms the first hypothesis. GBRCM achieved the best predictive performance, outperforming all baselines with an MAE of 0.0470, RMSE of 0.0341, and R^2 of 0.9989 on the controlled first dataset. Even in the more challenging second dataset, collected under real-world, unstructured conditions, GBRCM maintained superior accuracy (MAE=1.0933, RMSE=1.3270, R^2= 0.8199), while all baselines experienced severe degradation. Compared to the best-performing baseline (Random Forest), GBRCM reduced RMSE and MAE by more than 84% and achieved a significantly higher R^2. This substantial improvement, well beyond what is typically achieved through architectural tweaks alone, underscores the value of explicitly modeling structured residual errors using a second-stage learner such as the Histogram-Based Gradient Boosting Regressor (HGBR).

The ablation study provides further quantitative evidence for this conclusion. When `CSI_RollingMean` features were removed, simulating the exclusion of multipath and noise mitigation mechanisms, model performance degraded dramatically: RMSE increased by 366%, MAE by 426%, and R^2 declined by 16.9%, dropping from 0.9918 to 0.8228 on the first dataset. This sharp decline confirms that no other individual or combined feature group can sufficiently compensate for the loss of temporal smoothing in stabilizing CSI signal dynamics. The ablation results also highlight that multipath artifacts represent the most destabilizing factor in CSI-based occupancy estimation and that they must be explicitly addressed through robust feature engineering.

Notably, the performance contrast between the two datasets illustrates the second hypothesis: model robustness and generalization. The first dataset reflects a controlled experiment with occupancy ranging from 1 to 5 persons under predefined spatial configurations, minimizing environmental variability. In contrast, the second dataset represents real-world deployment with uncontrolled variations in signal behavior, irregular spatial distribution, and significant fluctuations in occupancy across zones. GBRCM's ability to maintain predictive accuracy under these contrasting conditions highlights its adaptability to under-actuated zones, where sensing coverage is sparse, and signal distortion is high.

The ablation results targeting other feature groups temporal dynamics, non-linear behavior, and spatial variation further refine our understanding. Removing temporal or non-linear features led to only marginal increases in RMSE (+1.0% and +1.04%, respectively) and negligible declines in R^2. This indicates that while these features contribute to model accuracy, they exhibit partial redundancy or are indirectly recoverable through residual learning. Similarly, spatial feature ablation caused a small performance drop, particularly in the second dataset where spatial heterogeneity is pronounced, reinforcing the utility but not absolute necessity of location-aware features.

The findings demonstrate that the combination of domain-aligned feature engineering and two-stage learning is not merely advantageous but essential for achieving accurate and generalizable CSI-based occupancy estimation. The explicit modeling of residual structure, grounded in bias-variance decomposition theory, allows GBRCM to adaptively refine predictions in environments characterized by signal unpredictability and measurement noise. In practical applications particularly in under-actuated indoor zones where sensor placement is sparse and conditions are dynamic such an approach offers a robust, interpretable, and computationally efficient solution.

3. Conclusion

This study introduced a novel two-stage prediction model, the Gradient Boosting-Based Residual Correction Model (GBRCM), designed to enhance occupancy estimation in under-actuated indoor zones using WiFi Channel State Information (CSI). Leveraging a dual-phase feature selection pipeline and domain-specific feature engineering, the proposed model addresses key challenges inherent to CSI signals, including multipath interference, temporal and spatial variability.

Comprehensive evaluation using two contrasting datasets one collected under controlled experimental conditions and the other from real-world, heterogeneous environments demonstrated the model's superiority over established state-of-the-art models such as MLP, CNN, RF, and LSTM. GBRCM achieved a significantly lower MAE and RMSE, along with an R^2 score exceeding 0.99 in the controlled setting, and maintained strong performance in more variable real-world scenarios. The model's effectiveness was further validated through a structured ablation study. The results underscored the indispensability of multipath-aware features such as `CSI_RollingMean`, whose removal led to a 426% increase in MAE and a marked drop in R^2, affirming its critical role in signal stabilization. Other engineered features capturing temporal dynamics, non-linear interactions, and spatial context also contributed positively but showed more modest performance degradation when excluded, highlighting the redundancy and robustness built into the model via residual learning.

The proposed GBRCM model offers a scalable, interpretable, and computationally efficient solution for occupancy estimation in smart building ecosystems. Its ability to combine domain-driven signal understanding with robust learning makes it a strong foundation for future developments in indoor sensing and intelligent space management.

Looking forward, several avenues can further enhance this research. First, extensive validation with operational CSI data across diverse indoor settings is essential to confirm the model's adaptability in real deployment scenarios. Second, scaling the model to support multi zone, multi floor, or open plan environments will broaden its applicability to complex spatial layouts often found in modern buildings. Third, deployment on edge computing platforms, leveraging inference optimized frameworks such as TensorFlow Lite or ONNX, will enable real-time, low latency prediction on resource-constrained hardware. Finally, incorporating adaptive learning mechanisms, such as online updates or drift detection modules, will allow the system to remain resilient to behavioral changes, environmental drift, and infrastructure evolution over time.

G. DAFTAR PUSTAKA

Sitasi disusun dan ditulis berdasarkan sistem nomor sesuai dengan urutan pengutipan. Hanya pustaka yang disitasi pada usulan penelitian yang dicantumkan dalam Daftar Pustaka.

[1] 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.

[2] J. Son and J. Park, "Channel state information (CSI) amplitude coloring scheme for enhancing accuracy of an indoor occupancy detection system using Wi-Fi sensing," Applied Sciences, Vol. 14, No. 17, 2024.

[3] S. S. Abolhassani, A. Zandifar, N. Ghourchian, M. Amayri, N. Bouguila, and U. Eicker, "Occupant counting model development for urban building energy modeling using commercial off-the-shelf Wi-Fi sensing technology," Building and Environment, Vol. 258, pp. 111548, 2024.

[4] A. Soleimanijavid, I. Konstantzos, and X. Liu, "Challenges and opportunities of occupant-centric building controls in real-world implementation: A critical review," Energy and Buildings, Vol. 308, pp. 113958, 2024.

[5] Y. Yang, G. Hu, and C. J. Spanos, "HVAC energy cost optimization for a multizone building via a decentralized approach," IEEE Transactions on Automation Science and Engineering, Vol. 17, No. 4, pp. 1950–1960, 2020.

[6] Yaddarabullah, A. A. Rahman, and A. Saad, "Optimized prediction of airflow volume in under-actuated zones through multilayer perceptron artificial neural network," International Journal of Intelligent Engineering and Systems, Vol. 18, No. 1, pp. 391–408, 2025.

[7] A. Syahputra, Yaddarabullah, M. F. Azhary, A. B. A. Rahman, and A. Saad, "Occupancy measurement in under-actuated zones: YOLO-based deep learning approach," International Journal of Advanced Computer Science and Applications, Vol. 15, No. 2, 2024.

[8] K. Ye, T. Xu, M. Jiang, B. Wang, and X. Yu, "Transfer-learning-based human activity recognition using antenna array," Remote Sensing, Vol. 16, No. 5, pp. 845, 2024.

[9] J. Sun, "DOA method based on CSI and improved neural network," Journal of Physics: Conference Series, Vol. 2807, No. 1, pp. 012017, 2024.

[10] B. Zhang, H. Sifaou, and G. Y. Li, "CSI-fingerprinting indoor localization via attention-augmented residual convolutional neural network," arXiv preprint, arXiv:2205.05775, 2022.

[11] B. Ko, H. Kim, and J. Choi, "Machine learning-based channel prediction exploiting frequency correlation in massive MIMO wideband systems," in Proc. of International Conf. on ICT Convergence (ICTC), pp. 9621119, 2021.

[12] W. Gao, "Research on modulation recognition method of electromagnetic signal based on wavelet transform convolutional neural network," Mathematical Models in Engineering, Vol. 10, No. 1, pp. 11–25, 2024.

[13] J. Bahi, W. Elghazel, C. Guyeux, M. Hakem, K. Medjaher, and N. Zerhouni, "Reliable diagnostics using wireless sensor networks," Computers in Industry, Vol. 104, pp. 103–115, 2019.

[14] R. K. Navandar, A. Ananthanarayanan, S. M. Joshi, and N. Venu, "Advanced estimation and feedback of wireless channels state information for 6G communication via recurrent conditional Wasserstein generative adversarial network," International Journal of Communication Systems, Vol. 38, No. 6, 2025.

[15] P. Chaudhari, Y. Xiao, M. M. Cheng, and T. Li, "Fundamentals, algorithms, and technologies of occupancy detection for smart buildings using IoT sensors," Sensors, Vol. 24, No. 7, pp. 2123, 2024.

[16] J. W. Rocks and P. Mehta, "Bias-variance decomposition of overparameterized regression with random linear features," Physical Review E, Vol. 106, No. 2, 2022.

[17] B. Adlam, N. Gupta, Z. Mariet, and J. Smith, "Understanding the bias-variance tradeoff of Bregman divergences," arXiv preprint, arXiv:2202.04167, 2022.

[18] B. Van Calster, M. van Smeden, B. De Cock, and E. W. Steyerberg, "Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study," Statistical Methods in Medical Research, Vol. 29, No. 11, pp. 3166–3178, 2020.

[19] Y. Zhou, J. Wu, H. Wang, and J. He, "Adversarial robustness through bias variance decomposition," in Proc. of the 31st ACM Conference on Information and Knowledge Management (CIKM), pp. 2753–2762, 2022.

[20] M. Salman, Y.-D. Soe, and Y. Noh, "WiFi-enabled occupancy monitoring in smart buildings with a self-adaptive mechanism," in Proc. of the 38th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 759–762, 2023.
[21] J. A. Armenta-Garcia, F. F. Gonzalez-Navarro, and J. Caro-Gutierrez, "Wireless sensing applications with Wi-Fi channel state information, preprocessing techniques, and detection algorithms: A survey," Computer Communications, Vol. 224, pp. 254–274, 2024.
[22] T. Tanaka, A. Yabuki, M. Funakoshi, and R. Yonemoto, "Validation of practicality for CSI sensing utilizing machine learning," arXiv preprint, arXiv:2409.07495, 2024.
[23] H. Choi, M. Fujimoto, T. Matsui, S. Misaki, and K. Yasumoto, "Wi-CaL: WiFi sensing and machine learning based device-free crowd counting and localization," IEEE Access, Vol. 10, pp. 24395–24410, 2022.
[24] Yaddarabullah, A. A. Rahman, and A. Saad, "Improving ventilation classification in under-actuated zones: A k-nearest neighbor and data preprocessing approach," Indonesian Journal of Electrical Engineering and Computer Science, Vol. 34, No. 1, pp. 233–244, 2024.
[25] J. Li, A. Sharma, D. Mishra, G. Batista, and A. Seneviratne, "COVID-safe spatial occupancy monitoring using OFDM-based features and passive WiFi samples," ACM Transactions on Management Information Systems, Vol. 12, No. 4, pp. 1–24, 2021.
[26] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," Journal of Cheminformatics, Vol. 15, No. 1, pp. 73, 2023.
[27] T. Freiesleben, G. König, C. Molnar, and Á. Tejero-Cantero, "Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena," Minds and Machines, Vol. 34, No. 3, pp. 32, 2024.
[28] A. I. G. Mowla, R. Hossain, and Asaduzzaman, "CSI-based people counting in WiFi networks: Leveraging occupancy detection," in Proc. of 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), pp. 1–6, 2024.
[29] J. Guo, W. Zhuang, Y. Mao, and I. W.-H. Ho, "RSSI-assisted CSI-based passenger counting with multiple Wi-Fi receivers," in Proc. of IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, 2025.
[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed., Springer Texts in Statistics, New York, NY: Springer, 2021.
[31] J. Liu, J. Zhang, Z. Liang, T. Wang, and X. Chen, "Passive IoT localization technology based on SD-PDOA in NLOS and multi-path environments," Computers, Materials & Continua, Vol. 80, No. 1, pp. 913–930, 2024.
[32] P. Reitz, C. Künzle, N. Franchi, and M. Lübke, "Evaluation of the interference performance of FMCW radar sensors in dense indoor environments," IEEE Access, Vol. 12, pp. 46834–46850, 2024.
[33] M. M. Ali, S. J. Hashim, M. A. Chaudhary, G. Ferré, F. Z. Rokhani, and Z. Ahmad, "A reviewing approach to analyze the advancements of error detection and correction codes in channel coding with emphasis on LPWAN and IoT systems," IEEE Access, Vol. 11, pp. 127077–127097, 2023.
[34] R. W. Kadel, K. Paudel, D. B. Guruge, and S. J. Halder, "Opportunities and challenges for error control schemes for wireless sensor networks: A review," Electronics, Vol. 9, No. 3, pp. 504, 2020.
[35] B. Yang, C. Dong, B. Gao, Y. Liu, W. Cui, and F. Gao, "Research on GNSS interference recognition based on ROI of correlation peaks," International Journal of Satellite Communications and Networking, Vol. 40, No. 5, pp. 330–342, 2022.
[36] Y. Guo, S. Zocca, P. Dabove, and F. Dovis, "A post-processing multipath/NLOS bias estimation method based on DBSCAN," Sensors, Vol. 24, No. 8, pp. 2611, 2024.
[37] M. Kong, B. Dong, R. Zhang, and Z. O'Neill, "HVAC energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study," Applied Energy, Vol. 306, pp. 117987, 2022.
[38] P. Sravanthi and D. V. Ratnam, "Multipath mitigation technique based on successive variational mode decomposition using NAVIC receiver data," Applied Mechanics and Materials, Vol. 926, pp. 113–123, 2025.