The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794

Machine Learning-based RFID Reader for Power Recommendation to assist Attendance

Nga Ly-Tu[*ab], Thao-Nguyen-Thi-Thanh[a], Duong Nguyen-Binh[ab], and Thien Nguyen-Minh[ab]
[a]International University, Ho Chi Minh City, Vietnam
[b]Vietnam National University, Ho Chi Minh City, Vietnam

*Corresponding author. Email: ltnga@hcmiu.edu.vn

ABSTRACT

Ultra-high-frequency radio frequency identification (UHF-RFID) technology provides a promising and cost-effective solution for tracking and positioning applications. However, its performance is affected by signal attenuation, environmental variability, and radio interference. Among the influencing factors, transmission power plays a critical role in the successful detection of RFID tags. Insufficient transmission power can lead to missed reads, while excessive power may result in signal collisions or interference with nearby systems, thereby compromising system reliability and scalability. To address the challenges associated with the successful detection of RFID tags, this paper proposes a custom-designed UHF-RFID detection framework with the following key features: (i) Circular polarization, which enhances tag orientation flexibility and reduces polarization mismatch; and (ii) Integration of machine learning (ML), enabling the system to autonomously classify environmental conditions and dynamically adjust transmission power levels. Additionally, the proposed ML-supported UHF-RFID detection framework improves the accuracy of tag classification and automates transmission power adjustments across a range of indoor and outdoor environments. By leveraging the RSSI dataset and introducing an extended dataset enriched with new features and conditions, we employ ML to optimize power consumption and improve detection efficiency. Furthermore, to ensure confidentiality during transmission for collected data, encrypt and decrypt CSV (Comma-Separated Values) file are proposed. Lastly, a set of performance analyses and recommendations based on supervised learning, unsupervised learning, and deep learning are provided to address the following tasks: (i) model and device selection in RFID environments, (ii) clustering of RFID tags, and (iii) usage recommendations for indoor and outdoor scenarios, respectively. These recommendations serve as practical guidelines for deploying RFID-based classification systems with improved accuracy and environmental adaptability. Our experimental findings indicate that attendance rates are consistently high, with median values ranging from 80% to 95%, suggesting a strong level of student engagement.

Keywords: Machine learning, UHF-RFID, detection of RFID tags, RSSI

1. INTRODUCTION

The ultra-high frequency radio frequency identification (UHF-RFID)-based attendance system presented in this study enables a fully automated process, effectively addressing several limitations of traditional manual attendance methods, such as time inefficiencies, proxy attendance, and the risk of lost or inaccurate records. By simply scanning student ID cards at UHF-RFID readers, attendance is recorded instantaneously, significantly enhancing both accuracy [18] and reliability. This automated approach not only preserves data integrity but also conserves valuable instructional time. Furthermore, the adoption of an IoT-driven solution—underpinned by UHF-RFID technology—offers educational institutions a secure, efficient [9,10], and scalable alternative to conventional attendance tracking systems [25].

UHF-RFID indoor and outdoor environments: Developing precise and cost-effective positioning systems is a key challenge, particularly in dynamic environments. visible light positioning has emerged as a high-precision

alternative to traditional radio-frequency-based systems, especially for indoor use [11,19]. Indoor positioning typically involves three components: positioning algorithms, technologies, and hardware [8]. In parallel, UHF-RFID technology has gained interest in worker safety and positioning in both indoor and outdoor settings. However, it faces challenges such as signal attenuation and environmental variability [6]. To address these, researchers have explored hybrid models and novel detection methods. Tesch et al. [18] introduced a UHF-RFID-based localization approach utilizing the doppler effect, demonstrating its potential as a low-cost and accurate solution. This technology has been widely adopted for tracking and positioning applications in various environments, including logistics, inventory management, and security systems [5]. UHF-RFID-based localization faces challenges, however, such as signal interference, multipath fading, and environmental variability, which can significantly affect detection accuracy [8]. Traditional signal processing methods often struggle to maintain reliable performance in dynamic settings, necessitating the integration of advanced techniques to improve positioning precision. Inspired by previous research on intelligent systems that understand the IoT that utilize the electromagnetic behavior of UHF-RFID tags to improve worker safety in outdoor environments [10], this approach optimizes the detection accuracy and response efficiency in dynamic field conditions.

Machine learning for UHF-RFID based detection systems: Machine learning (ML) is playing an increasingly vital role in optimizing UHF-RFID-based systems, offering enhanced performance through data-driven approaches. Frameworks such as APP-CAT have been instrumental in evaluating key performance indicators, including accuracy, cost-efficiency, and scalability [8]. The use of machine learning significantly enhances UHF-RFID detection systems by improving signal processing, pattern recognition, and anomaly detection. These improvements make the system more reliable and responsive in complex or changing environments [19]. Moreover, the integration of machine learning facilitates intelligent decision-making and system flexibility across various real-world settings. A notable example is the use of convolutional neural networks (CNNs), which effectively extract features from RFID signal data to improve tag classification and localization [11]. Furthermore, lightweight ML models optimized for low-power edge devices can further enhance system efficiency, reducing computational overhead while maintaining high detection rates [9].

Our contribution is as follows:

1. **A custom-designed UHF-RFID reader for detection framework:** Transmission power plays a key role in RFID tag detection performance. Insufficient power may result in missed reads, whereas excessive power can cause signal collision or interfere with adjacent systems. By integrating ML, a system can autonomously classify environmental conditions and adjust power levels accordingly. We introduce a ML-enhanced UHF-RFID detection framework, see Section 3.1 that classifies environmental conditions and optimizes transmission power in real-time. This adaptive strategy significantly improves detection accuracy and scalability, particularly in dense tag environments where manual calibration would be inefficient. Linear polarization in UHF-RFID readers imposes directional constraints, making tag orientation critical for successful communication [2], while circularly polarized antennas alleviate this limitation by supporting omnidirectional tag alignment [16]. The commercial UHF-RFID reader of our current experimental [11] is a linearly polarized array with four E-shaped antennas. Based on our experimental [3], we propose a custom designed UHF wideband antenna, featuring circular polarization to enhance the flexibility of tag orientation and reduce polarization mismatch shown in Section 3.2. Unlike circular polarization, which accommodates varying tag angles, linear polarization requires the RFID tag to be aligned along the same plane as the antenna's electromagnetic wave. This dependency reduces flexibility and may lead to weaker signal reception if the tag is misoriented.

2. **Collected dataset and data-preprocessing:** Building on the previous study [11], we introduce a newly collected dataset that includes a wider range of characteristics and testing conditions shown in Section 3.3. To ensure confidentiality during storage or transmission, we propose Algorithm 1 to encrypt and decrypt

CSV (Comma-Separated Values) file content using a symmetric XOR cipher with randomly generated keys per cell or character (see Table 2). To solve the missing data for commercial UHF-RFID reader, Algorithm 2 is mentioned in Table 3 in Section 3.4. Final, the percentage of dataset for training and test of commercial UHF and proposal devices in both indoor and outdoor environments are shown in Section 4.1.

3. **Performance analyses and recommendations:** To evaluate performance analysis of clustering in RFID Tag, the role of unsupervised learning in uncovering patterns that link laboratory test results with real-world RFID performance shown in Section 4.2. By identifying reliable performance benchmarks through clustering techniques namely Principal Component Analysis (PCA), Gaussian Mixture Model (GMM) and K-Means, the study effectively narrows the gap between theoretical expectations and operational realities, thereby supporting data-driven RFID tag selection based on our extended dataset for attendance system. Bayesian information criterion (BIC) [27] metrics, see Equation (1) and (2) in Section 3.5, are applied to recommended usage. Applied supervised learning to recommended model and device selection in RFID environments is mentioned in Section 4.3. The proposed device demonstrates superior performance in indoor environments, particularly when paired with classification algorithms such as logistic regression (LR), K-nearest neighbors (KNN), and random forest (RF), owing to its enhanced generalization capabilities. In outdoor scenarios, both the commercial UHF and the proposed devices yield comparable results; however, optimal performance is achieved using decision tree (DT), RF, or KNN classifiers. In contrast, algorithms such as Naïve Bayes and LR exhibit reduced effectiveness in outdoor contexts and are therefore not recommended for such applications. Notably, the proposed model exhibits significant advantages in indoor deployments, particularly in its resilience to signal interference and its robustness to variations in tag orientation. Integration of deep learning models such as Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), or CNN allows the system to adapt to noisy or obstructed conditions see Section 4.4. It plays role in recommended usage for indoor and outdoor environments. Our proposed device is recommended for indoor applications, especially when used with LR, KNN, or RF. For outdoor environments, model selection should prioritize DT, RF, or KNN, regardless of the device. Classifiers such as Naïve Bayes and LR are not advisable for use in outdoor settings due to their diminished performance. These recommendations serve as practical guidelines for deploying RFID-based classification systems with improved accuracy and environmental adaptability. Lastly, performance analysis of lab attendance based on using both commercial and proposal readers is mentioned in Section 4.5.

2. LITERATURE REVIEW

The RFID-based attendance system presented in this study enables a fully automated process, effectively overcoming several limitations associated with traditional manual attendance methods, such as time inefficiencies, proxy attendance, and the risk of lost records. By simply scanning student ID cards at RFID readers, attendance is recorded instantaneously, significantly improving both accuracy and reliability. This automated approach not only ensures data integrity but also conserves valuable instructional time. The adoption of an IoT-driven solution, underpinned by RFID technology, offers institutions a secure, efficient, and scalable alternative to conventional systems. By addressing the key challenges of manual tracking, the proposed system enhances operational efficiency and fosters greater student engagement. The study also outlines critical design principles and best practices, providing a practical framework for institutions aiming to implement modern, intelligent attendance monitoring solutions that contribute to improved academic outcomes [25].

2.1 UHF-RFID system

The integration of RFID technology into attendance management systems has been extensively explored as a solution to the inefficiencies of traditional manual attendance management methods [1, 12, 3]. Manual attendance tracking can be time-consuming and labor-intensive, especially in large academic environments. In contrast, automated systems provide high accuracy, real-time tracking, and improved usability. For example, a real-time RFID-based employee identification system using the RC522 card reader with a 5-cm detection range, integrated with a web interface and back-end database is [7] introduced. Similarly, calculating the average reading time of RFID tags and automating the tracking of student attendance based on stored data is focused on [1]. Although these early systems provided a functional framework, they often lacked scalability and responsiveness in more dynamic environments [7].

To address these challenges, more recent works have adopted advanced web technologies. As demonstrated in [3], the use of platforms such as NodeJS and MongoDB significantly improved system performance and real-time data handling capabilities. In addition to technical advances, the author in [12] examined contextual factors that affect e-learning in Vietnam, focusing on infrastructure, institutional support, and student digital literacy, all of which play critical roles in the successful implementation of tech-based attendance systems. An advanced RFID-based attendance system was also proposed, which extended detection distances to 1 to 3 meters by adjusting transmission power (18 dBm to 24 dBm) in [3]. It offered features such as real-time identification, fraud detection, email alerts for absentees, and data visualization for instructors.

To further this innovation, the approach in [20] enhanced both offline and online attendance tracking through a hybrid approach. It integrates UHF-RFID with four circularly polarized antennas and high-definition cameras for face recognition. This dual-mode system automatically monitors in-class attendance via RFID and cameras installed in classrooms, while also supporting online attendance by leveraging webcams on students' personal devices. Real-time updates are managed through a Web-based platform, allowing lecturers to monitor, manually adjust attendance when necessary, and generate weekly attendance and engagement reports. The system also includes performance evaluations in terms of time efficiency, precision, and accuracy, providing a comprehensive view of its effectiveness.

Building on these foundations, a machine learning-enhanced UHF-RFID system that reduces doppler effects and improves the detection accuracy of tags under various physical conditions is introduced [11]. It synchronizes RFID and camera data for better tracking precision and employs supervised learning and convolutional neural networks (CNN, lightweight CNN) to recommend optimal reader power settings. The system achieves classification accuracies of up to 94.05% and 93.69% when dealing with tags inside school bags or beside personal belongings. Collectively, these studies demonstrate a clear progression towards intelligent and adaptive attendance systems that use RFID, camera technologies, and ML to deliver scalable, accurate, and user-friendly solutions in various environments.

2.2 Machine learning-based UHF-RFID

The integration of RFID hardware—comprising tags with unique identifiers and RFID readers—ensures automated and accurate data capture, significantly outperforming manual entry methods in terms of speed and reliability. Overall, the combination of RFID and ML not only streamlines the attendance tracking process but also lays the groundwork for intelligent, predictive attendance management systems [24]. Machine learning has been widely used to enhance RFID tag detection, but manual data calibration remains a challenge. Recent studies in semi-supervised learning for indoor localization have shown that manifold alignment can reduce reference points while maintaining accuracy [17]. This approach, which integrates sporadically collected data to optimize calibration, can be adapted for RFID detection to minimize manual effort while improving performance. Using insights from fingerprint-based radio maps, machine learning models for RFID detection can be improved to

handle environmental variations more efficiently.

Supervised learning algorithms such as KNN, LR, DT, and Naïve Bayes require labeled datasets split into training and testing sets (typically 70-80% for training and 20-30% for testing). These models learn patterns from labeled data and are evaluated using metrics such as precision, precision, recall, and F1 score [17]. Unsupervised learning, on the other hand, identifies patterns and structures within data without predefined labels. It is useful for clustering and feature learning without explicit supervision. These findings emphasize the value of empirical evaluation and the role of unsupervised learning [26] in uncovering patterns that link laboratory test results with real-world RFID performance. By identifying reliable performance benchmarks through clustering techniques, the study bridges the gap between theoretical expectations and practical deployment, enabling more informed RFID tag selection and enhancing operational efficiency in logistics and inventory management. Semi-supervised learning combines a small set of labeled data with a large amount of unlabeled data, improving model accuracy while reducing the need for extensive manual labeling [23]. Deep learning models particularly CNNs, extract hierarchical features from the input data. CNN architectures, such as AlexNet, consist of fully connected convolutional layers, using filters to detect spatial features. Lightweight CNN (LiCNN) models such as MobileNet and EfficientNet optimize efficiency for resource-limited environments, employing techniques such as separable convolutions in depth and network pruning [22].

Machine learning techniques particularly neural networks have been widely applied to enhance RFID tag detection in various environments. Among these, the multi-layer perceptron (MLP) model is one of the most used because of its practical implementation and relatively simple mathematical structure. However, the theoretical understanding of MLP, especially in terms of its approximation capabilities, remains an ongoing area of research [13]. Insights from approximation theory help in assessing the strengths and limitations of MLP models when applied to signal processing tasks, such as RFID detection. Using these theoretical foundations, MLP-based machine learning models can be optimized to improve detection accuracy and adaptability in dynamic environments. Recurrent neural networks (RNNs), including LSTM and GRUs process sequential data using memory cells and gating mechanisms to mitigate problems of disappearance of gradients [21]. These architectures improve learning from time-dependent data, making them suitable for RFID-based activity recognition [4].

3. METHODOLOGY

3.1 Proposal ML framework for UHF-RFID detection

Figure 1 outlines our workflow for collecting, preprocessing, and analyzing RFID data in indoor and outdoor environments. Data collection is performed using RFID readers, with UHF and our proposal device deployed indoors, while UHF and our proposal device also collect outdoor data separately. Indoor experiments test multiple reader placements to minimize doppler effects and optimize data accuracy by adjusting power levels and scan intervals. Raw data is categorized into indoor and outdoor datasets, ensuring that environmental factors are taken into account. Pre-processing includes handling missing values, removing duplicates, converting categorical attributes to numerical values, merging datasets, and applying normalization to ensure consistency. The key extracted features include the last update time, RSSI values, scan intervals, distance measurements, and tag counts. The processed data are divided into labeled and unlabeled sets. Supervised learning models for example, KNN, LR, DT, Naïve Bayes classify RFID tags and predict outcomes, while unsupervised techniques (PCA, K-Means, RF) detect patterns and anomalies. The workflow addresses indoor challenges, such as interference, by optimizing the placement and settings of the reader. The commercial UHF and our proposal device data integration ensures comprehensive analysis, improving system efficiency and accuracy for both indoor and outdoor tracking.

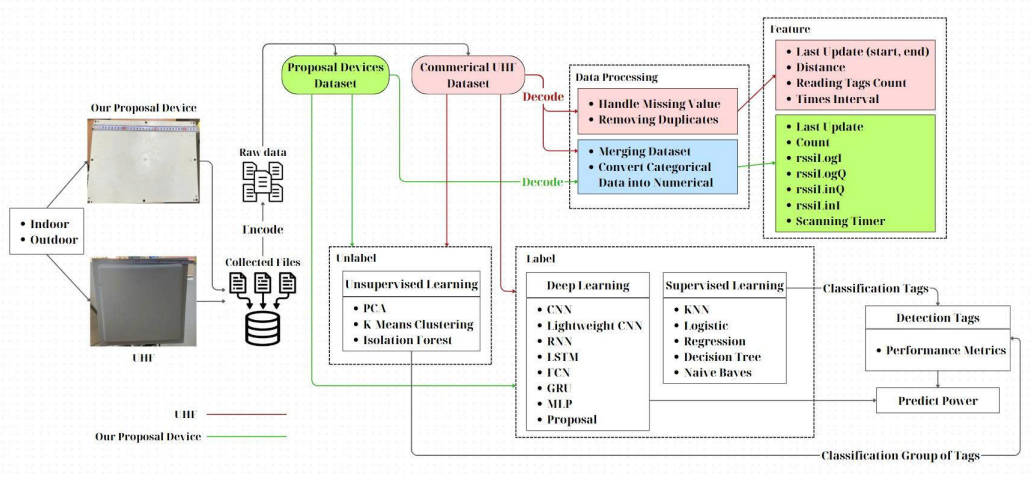


Figure 1: Proposal ML framework for UHF-RFID detection.

3.2 Proposal UHF-RFID reader

RFID systems consist of two main components: an RFID tag and an RFID reader. Antenna and transmitter play a crucial role in UHF RFID systems, facilitating effective communication between RFID tags and readers. The antenna operates in the UHF band from 860MHz to 960 MHz, which affects the range and accuracy of the detection of tags based on gain, polarization, and beam width. The transmitter generates electromagnetic signals with key characteristics including power output, frequency, and modulation scheme. The RFID reader communicates with the tags via radio signals, enabling automated identification and integration with computer systems.

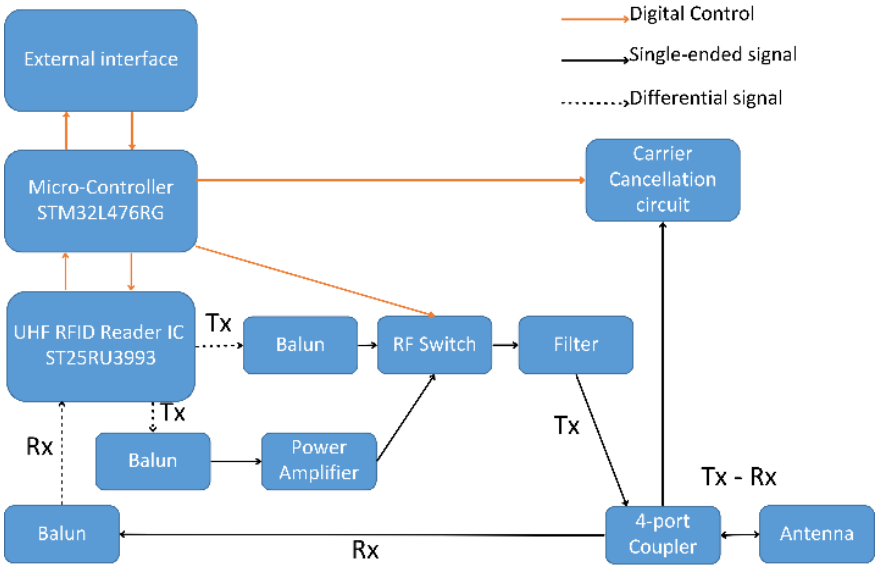


Figure 2: Hardware system diagram of proposal RFID reader.

Figure 2 shows a diagram of the proposal RFID reader hardware system. The reader integrates ST25RU3993

reader IC for the UHF-RFID [14] and an STM32L476RG microcontroller [15] to allow efficient tag communication. The reader IC handles transmit (Tx) and receive (Rx) operations, interfacing with baluns to convert RF differential signals to single-ended signals. The RF switch provides control over signal amplification, allowing the Tx signal to be amplified by either an internal amplifier within the RFID reader IC or by an external power amplifier for higher transmission power. Signals are filtered to reduce noise and managed by a 4-port coupler, which directs transmission and reception while working alongside the carrier cancelation circuit to remove unwanted interference between large Tx and small Rx signals.

The circuit board and antenna for both our proposed and commercial UHF-RFID readers are shown in Figure. 3. The main components of the proposed reader are clearly shown in Figure. 3a and described in Figure. 2, while the components of the commercial reader, shown in Figure. 3b, are not identifiable. The antenna, as shown in Figure. 3c, is a custom-designed UHF wideband with circular polarization to improve the flexibility of the orientation of the tag and minimize mismatch of polarization. The antenna has an approximate gain of 7dBi, optimizing the signal strength and coverage for reliable RFID tag communication Figure. 3d. One notable difference between the two readers is their antenna design. The commercial reader's antenna, depicted in Figure. 3b, is a linearly polarized array with four E-shaped antennas. In contrast, Figure. 3a shows our custom-designed UHF wideband antenna, featuring circular polarization to enhance the flexibility of tag orientation and reduce polarization mismatch. With an approximate gain of 7dBi, this antenna optimizes signal strength and coverage, ensuring reliable RFID tag communication.

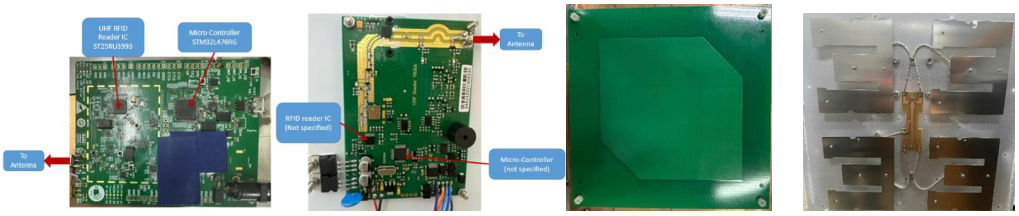


Figure 3: Circuit board of the proposed and commercial UHF RFID readers. (a) Proposed and (b) Commercial (c) Custom-designed circularly polarized wideband antenna our proposal (d) Corresponding commercial antenna.

Comparison between proposal UHF-RFID and a commercial RFID reader (namely UHF) shown in Table 1. Our proposal offers several advantages. It supports the latest Gen2v2 protocol, ensuring enhanced security and advanced tag management capabilities. With a broader frequency range of 840 to 960 MHz, it provides global compatibility and operational flexibility. The reader IC has high sensitivity (-80 dBm) and a rapid read rate of up to 700 tags per second, making it ideal for high-speed applications like attendance monitoring and logistics. Its compact antenna design enhances versatility in space-limited settings, while custom interference cancellation ensures reliable operation.

3.3 Setting up UHF-RFID readers for indoor and outdoor environments

This subsection shows the setup of UHF-RFID readers for indoor environments in Room LA1.604 and outdoor environments in Stage A2 International University, VNU-HCM. The UHF system is configured as shown in Figure. 4a. To ensure effective monitoring of the area, the UHF reader was placed between 10 and 15 meters from the entrance and exit zones for our proposal device and between 18 and 25 meters for UHF. Because it was installed outdoors, the doppler effect was significantly reduced due to the absence of obstructions in the vicinity. However, environmental factors may still affect the reader's longevity, potentially reducing its operational lifespan.

Table 1: Comparison of our proposal and commercial UHF-RFID readers.

No.	Features	Our proposal reader	Commercial reader (UHF)
1	Protocol ISO/IEC 18000	-63:2015/Gen2v2, -62:2012	-6B or-C (EPC C1G2)
2	Supported frequency	840 to 960 MHz	902 to 928 MHz
3	Maximum sensitivity	∼80 dBm	Not specified
4	Maximum tag read rate	700 tags/s	Not specified
5	Measured reading distance	∼6 m (indoor)	∼8.5m (outdoor)
6	Antenna polarization	Circular	Linear
7	Antenna gain	7 dBi	8 dBi
8	Antenna reflection coefficient	[-15,-13,-11]dB with [840,900,960]MHz	[-11,-27,-38]dB with [840,900,960]MHz
9	Antenna Size [cm]	$17 \times 17 \times 4.5$	$44 \times 44 \times 2.4$

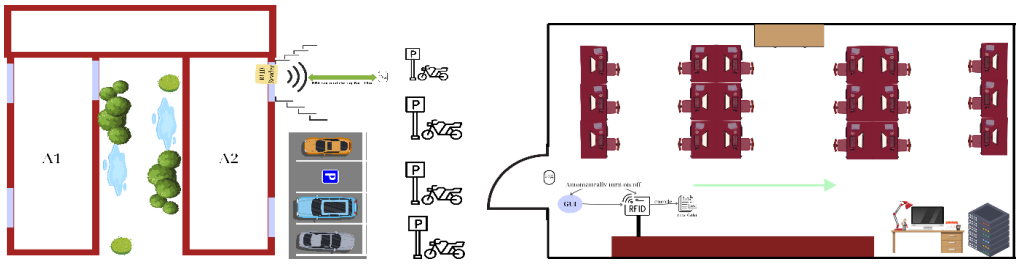


Figure 4: Setting up UHF-RFID readers for (a) outdoor environments (b) LA1.604 Room for indoor at International University, VNU-HCM.

The commercial UHF and our proposal device used for indoor settings is set at a 52 degree tilt angle, carefully chosen to enhance signal coverage while minimizing blind spots. This tilt position allows the reader to orient the signal in a way that optimizes detection in outdoor environments, compensating for potential interference from factors such as doppler and cross-object interference shown in Figure. 4b.

Dataset collection process: The data collection process is divided into two parts: collecting training data and collecting test data, which apply to both indoor and outdoor environments. This process takes approximately 30 hours to gather about 45.96 MB of training data and 12 hours to gather 11.2 MB of test data. Four different types of RFID tags and three types of tag covers, shown in Figure. 5, that students commonly used to generate about 1000 tags which 320 tags for training datasets and others for test datasets. Data collection consists of three main parts: setup, collecting data in 10 seconds (since the average time for a male or female to enter a classroom and cover the distance is only about 4 to 6 seconds, the remaining 4 seconds accounting for possible errors that may occur when the classroom starts filling with people), and checking thedata to ensure its accuracy and that there is no noise.

Encryption data: A structured and automated approach was implemented to secure CSV data prior to processing, ensuring data integrity and confidentiality throughout the preprocessing pipeline shown in Table 2. Ensure that sensitive information remains protected by encrypting specific columns and correctly restoring them when needed. The use of random keys for each value improves uniqueness and the step-by-step workflow makes it easy to integrate into data pipelines. It is a well-organized method to add an encryption layer to the handling of CSV.

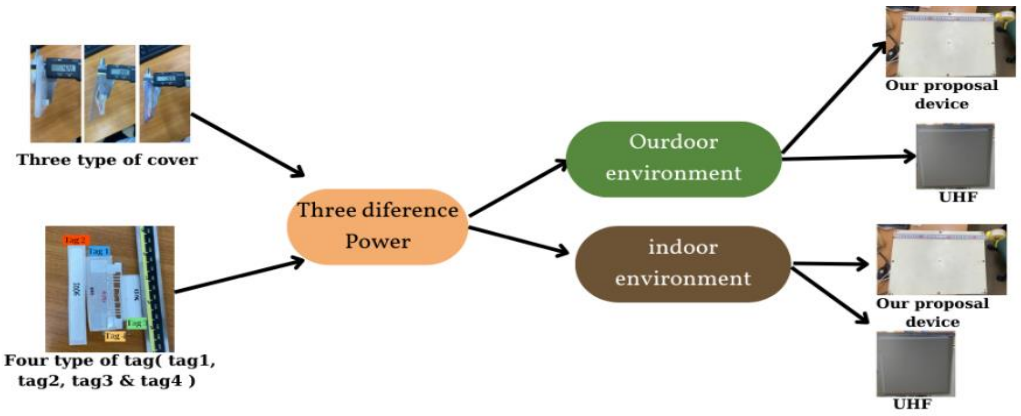


Figure 5: RSSI dataset collection process in indoor and outdoor environments.

Table 2. Algorithm 1: Secure CSV data prior to processing

1. Define function `xor_encrypt_decrypt(data, key)`:
2. Convert data to string
3. Set `key_length` = length of key
4. Pad or truncate data to match `key_length`
5. For each character in data:
6. XOR character with corresponding character in key
7. Return XOR result as string
8. Define function `generate_random_token(length)`:
9. Generate and return random string of given length using letters and digits
10. Load CSV file into DataFrame
11. Define `columns_to_encrypt` as list of target column names
12. Initialize `encryption_keys` as empty dictionary
13. For each column in `columns_to_encrypt`:
14. Initialize `encrypted_data` as empty list
15. Initialize `encryption_keys[column]` as empty list
16. For each value in the column:
17. key ← `generate_random_token(length of value)`
18. `encrypted_value` ← `xor_encrypt_decrypt(value, key)`

19. Append `encrypted_value` to `encrypted_data`
20. Append key to `encryption_keys[column]`
21. Replace original column data with `encrypted_data`
22. Save encrypted DataFrame to `encrypted_csv`
23. For each column in `columns_to_encrypt`:
24. Initialize `decrypted_data` as empty list
25. For each (`encrypted_value`, key) pair in column:
26. `decrypted_value` ← `xor_encrypt_decrypt(encrypted_value, key)`
27. Append `decrypted_value` to `decrypted_data`
28. Replace encrypted column data with `decrypted_data`
29. Save decrypted DataFrame to `decrypted_csv`
30. Encryption and Decryption for all columns done!

3.4 Dataset Preprocessing

3.4.1 Dataset preprocessing for commercial UHF-RFID

Data preparation consists of various critical stages that ensure the quality and usefulness of data for analysis or modeling. These stages include handling missing values, removing duplicates, merging datasets, and converting categorical data into numerical formats. In Algorithm 2, applied to the UHF dataset, missing values are addressed by imputing data based on the number of students recorded in the queue (see Table 3). Specifically, the number of missing values corresponds to the number of students recorded. For instance, a single missing entry is filled with the value of the preceding student, while multiple missing values (e.g., five) are imputed using the values of the five preceding individuals in the queue. This approach ensures that missing data is effectively handled by utilizing the historical information in the dataset shown in Figure. 6.

Table 3. Algorithm 2: Preprocessing data for commercial UHF-RFID reader

Require: Dataset UHF with columns EPC and other attributes, where missing values exist.
Ensure: Filled dataset UHF with no missing values.

1. Extract `unique_EPC` values
2. **for all** EPC in `unique_EPCs` **do**
3. Extract data subset
4. **for all** columns with missing values in `EPC_data` **do**
5. Count missing values
6. Retrieve previous values

7. Fill missing values
8. **end for**
9. Update UHF with processed `EPC_data`
10. **end for**
11. Return Processed dataset UHF

Removing duplicates is first essential to prevent bias and ensure analytical integrity, as duplicate entries can distort statistical measures (e.g., mean, median) and misrepresent relationships between variables. To remove duplicates, methods like `drop_duplicates()` in pandas can be used. The UHF dataset collected by the commercial reader [11] utilizes a context-aware preprocessing strategy, addressing missing values by leveraging the sequential order of individuals in the RFID queue to maintain temporal consistency. This design is well-suited for attendance tracking in controlled environments. However, its reliance on transmission power as the primary feature limits the UHF dataset's richness and adaptability for more complex modeling tasks shown in Figure. 7.

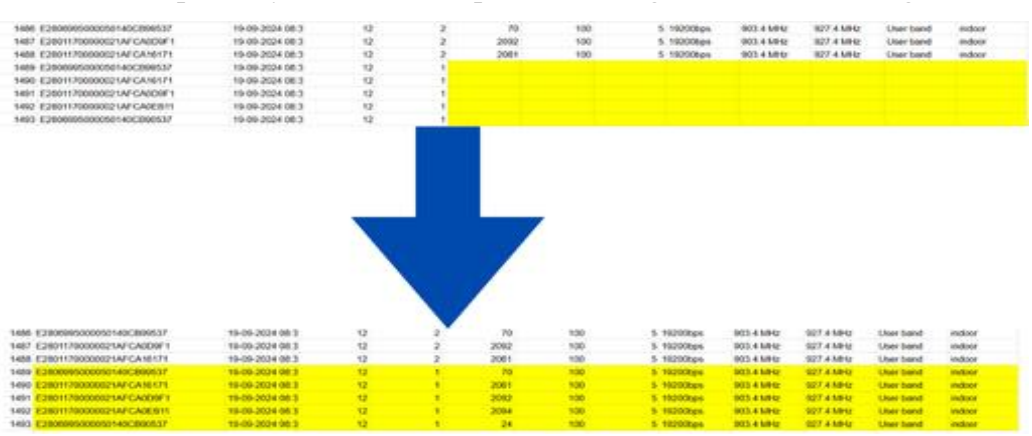


Figure 6: Handle missing value in UHF_outdoor and UHF_indoor.

3.4.2 Dataset preprocessing for proposed UHF-RFID

The dataset preprocessing pipeline consisted of the following as (1) importing necessary libraries for data handling and transformation. (2) loading the primary and auxiliary datasets into DataFrames. (3) merging datasets using concatenation or key-based joins, depending on context. (4) encoding categorical variables using label and one-hot encoding strategies. (5) verifying the processed dataset to ensure readiness for downstream analysis.

On the other hand, proposal dataset offers a broader and more versatile structure. It incorporates a wider range of features, including several that are closely related to RSSI, enabling more accurate and flexible ML models. Our proposal also supports data integration, categorical encoding, and is well-suited for both supervised and unsupervised learning. Although it lacks a specific strategy for missing value imputation, its greater feature diversity and alignment with signal characteristics like RSSI make it more powerful for real-world RFID applications. Therefore, the proposed device was selected for integration into the framework due to its flexibility, scalability, and enhanced data representation capabilities.

The dataset used in this study was collected from the proposed device, specifically designed for (insert context, e.g., medical sensing, environmental monitoring, etc.). The raw data was structured in tabular format

and stored as a CSV file. To prepare the dataset for subsequent analysis, several preprocessing steps were performed, as detailed below.

3.4.3 Dataset preprocessing for outdoor environments dataset

The UHF outdoor dataset was collected using the same UHF reader device as the indoor dataset, but in an outdoor environment under a single test case involving cover 1. The total dataset size is significantly smaller -approximately one-twelfth the size of the UHF indoor dataset. For example, given that the same device and configuration were used for data acquisition, the feature structure and preprocessing steps remain consistent with those applied to the UHF indoor dataset. This ensures that any subsequent analyses or models trained using both datasets are compatible in terms of format and feature representation.

3.4.4 Data augmentation for commercial UHF-RFID and proposed UHF-RFID

In indoor environments where data is collected using two different UHF-RFID readers—namely the commercial reader and our proposed device—data augmentation through normalization becomes essential to ensure consistency. This process involves aligning the RSSI values (see Section 3.5.2) between the two datasets to facilitate fair comparison and analysis. To achieve this, we first load both datasets, ensuring that they share a common identifier through a standardized key column (converted to uppercase for uniformity). For each entry in the commercial UHF-RFID dataset, we identify the corresponding record in the proposed device dataset based on this key.

A LR model is then applied to capture the relationship between RSSI values from the two datasets. This model is used to compute a power adjustment ratio, which accounts for differences in transmission power or signal reception characteristics. Using this ratio, both logarithmic and linear forms of RSSI values in the commercial dataset are adjusted to align with the proposed device dataset. Once normalization is complete, the updated commercial UHF-RFID dataset—now augmented to be consistent with the proposed system—is saved as a new CSV file. This approach ensures that data from both readers can be reliably compared, contributing to more accurate indoor tracking, or classification tags or powers.

3.5 BIC and RSSI calculations

3.5.1 Bayesian information criterion

Bayesian information criterion (BIC) [27] is widely used in unsupervised learning for model selection. It helps prevent overfitting and guides toward more generalizable models. In GMM, BIC directly evaluates the trade-off between fit and complexity to determine the optimal number of clusters. When PCA is used for dimensionality reduction before GMM, BIC still applies to the reduced-space model. For K-means, although not probabilistic, BIC can be approximated using residual sum of squares to estimate the best number of clusters. The formulas to calculate the BIC for GMM and K-means are presented in Equation (1) and (2), respectively.

$$BIC_{GMM} = -2\ln(\hat{L}) + k\ln(n) \tag{1}$$

where $k = K\left[\frac{d(d+1)}{2} + d + 1\right] - 1$ with d is the number of dimensions, K is the number of Gaussian components, GMM assumes the data is generated from K Gaussian distributions and uses expectation-maximization.

$$BIC_{K-means} = -n\ln\left(\frac{RSS}{n}\right) + k\frac{(d+1)}{2}\ln(n) \tag{2}$$

where RSS is the residual sum of squared (likelihood can be approximated using within-cluster variance), and k is the number of clusters.

3.5.2 RSSI calculations

Real-world environments introduce attenuation beyond ideal free-space path loss (FSPL). A more general log-distance path loss model can be applied to RSSI calculations [28]. The path loss grows as a function of distance with an exponent n (path-loss exponent) as:

$$L_p(d) = -L_p(d_0) + 10n\log\left(\frac{d}{d_0}\right) \tag{3}$$

where d_0 is a reference distance 1m for UHF-RFID reader, and $n = 2$ under free space conditions but higher (2.5 to 3.0 or more) in cluttered environments. $L_p(d)$ represents the total signal loss during transmission and reflection, due to distance, obstacles, and environmental factors. It typically increases with distance. This log-distance formula is widely used in RFID ranging and localization research because it accounts for environmental attenuation beyond ideal free-space conditions.

Furthermore, the standard RSSI calculation in UHF-RFID is based on a link-budget equation (Friis/FSPL), adjusted with a path-loss exponent for real-world conditions shown in Equation (4).

$$P_{rx}(d \text{ in } dBm) = P_{tx}(dBm) + G_{tx} + G_{rx} - L_p(d) \tag{4}$$

where $P_{rx}(d \text{ in } dBm)$ and $P_{tx}(dBm)$ are received signal strength indicator (RSSI) and transmitter power of the RFID reader at distance d, respectively; G_{tx} and G_{rx} are gain of the transmitting antenna and receive antenna at distance d, respectively.

4. IMPLEMENTS AND RESULTS

4.1 Dataset of commercial UHF and proposal devices in both indoor and outdoor environments

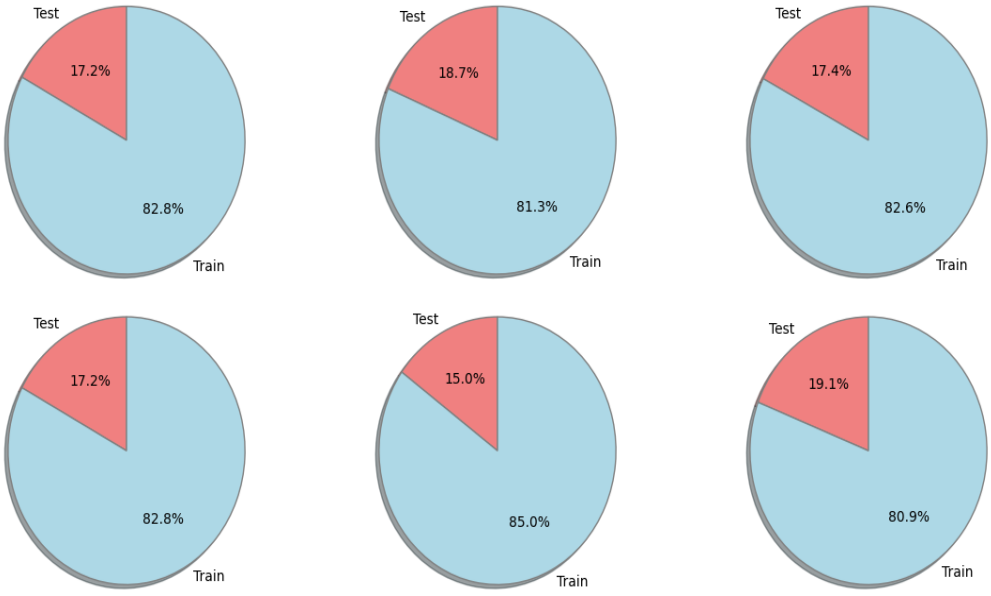


Figure 7: Percentage of train and test in dataset of (a) UHF and proposal indoor (b) UHF indoor (c) proposal indoor (d) UHF outdoor (e) proposal outdoor (f) UHF and proposal outdoor.

The pie charts in Figure 7 illustrate the data split between the training and testing sets across six different models or experimental scenarios of UHF systems under various configurations. In indoor settings, training and testing results ranged from 17.2–18.7% and 81.3–82.8%, respectively. Notably, Figure. 7d showed a reversed trend, with a high training result of 84.97% and the low test result of 15.03%, indicating possible overfitting. For outdoor scenarios, our proposal alone yielded 85% training and 15% testing (Figure.7e), while the combination of UHF and our proposal resulted in a lower training score (80.9%) but a higher test score (19.1%) in Figure. 7f. These variations underscore the influence of configuration and environment on model performance.

4.2 Clustering Evaluation in RFID Tag Performance Analysis

By applying unsupervised learning techniques to the RFID dataset, we aim to uncover latent structures and assess the separability of tag performance data collected from the reader system. Initially, Principal Component Analysis (PCA) is utilized to reduce the dimensionality of the data while retaining the majority of its variance, see Figure. 8a. This step facilitates effective visualization and simplifies subsequent clustering procedures. Two prominent clustering methods—K-Means and GMM—are then applied to the PCA-reduced data, see Figure. 8b and 8c, respectively. The results indicate that GMM outperforms K-Means, achieving a silhouette score of 0.731 compared to 0.705 for K-Means, highlighting GMM's superior ability to capture probabilistic distributions and soft cluster boundaries. In addition, GMM yields a BIC of 14,913.37, further confirming its better fit to the underlying structure of the data. Visual inspection of the GMM clustering output reveals well-separated and compact clusters, supporting the quantitative evaluation and demonstrating that GMM provides a more accurate model for RFID data, which often contains noise and overlapping patterns. The combined use of PCA, K-Means, and GMM not only elucidates the intrinsic structure of RFID datasets but also establishes a robust quantitative foundation for optimizing tag selection and improving reader system design in real-world deployments.

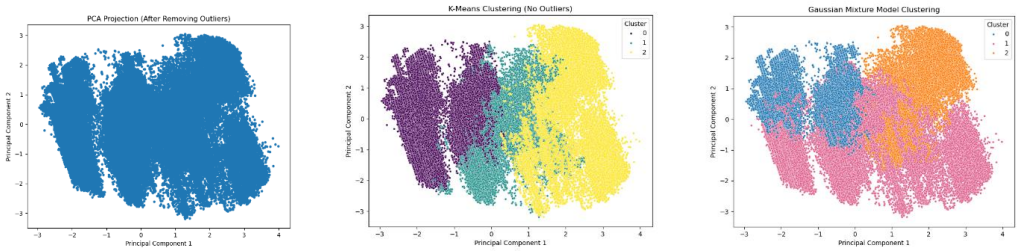


Figure 8: Clustering analysis in dataset of our RFID device (a) PCA, (b) K-mean and (c) GMM.

Based on the clustering analysis of RFID device behavior, distinct operational profiles have been identified that can inform deployment strategies and system optimization shown in Figure 7. It is recommended that Cluster 0 devices, which demonstrate strong and stable signal strength, low linear RSSI values, and moderate detection frequency, be designated for fixed-position applications. Their consistent signal characteristics and predictable scanning patterns make them highly suitable for scheduled or periodic data collection in controlled indoor environments. By the highest scan frequency, shortest response times, and strongest signal power, Cluster 1 devices, are ideal for real-time tracking and mobile asset monitoring. These devices should be prioritized for deployment in dynamic or high-traffic environments, such as logistics hubs, hospitals, or manufacturing floors, where continuous and reliable communication is essential. Their performance indicates suitability for applications that require low-latency and high-availability RFID communication. In contrast, Cluster 2 devices, which exhibit weak and unstable signal characteristics and low detection frequency, are indicative of marginal or intermittently connected nodes. These devices are likely to be located at the edges of the reader's coverage area or in environments subject to physical obstructions and RF interference. It is recommended that their use

be limited to non-critical monitoring, edge condition analysis, or environments where occasional connectivity suffices. System designers should also consider repositioning or augmenting infrastructure (e.g., with additional readers or antennas) to mitigate the performance degradation associated with this group. In conclusion, adopting a cluster-informed deployment strategy can significantly enhance RFID system efficiency, coverage reliability, and resource allocation by aligning device roles with their observed communication behaviors.

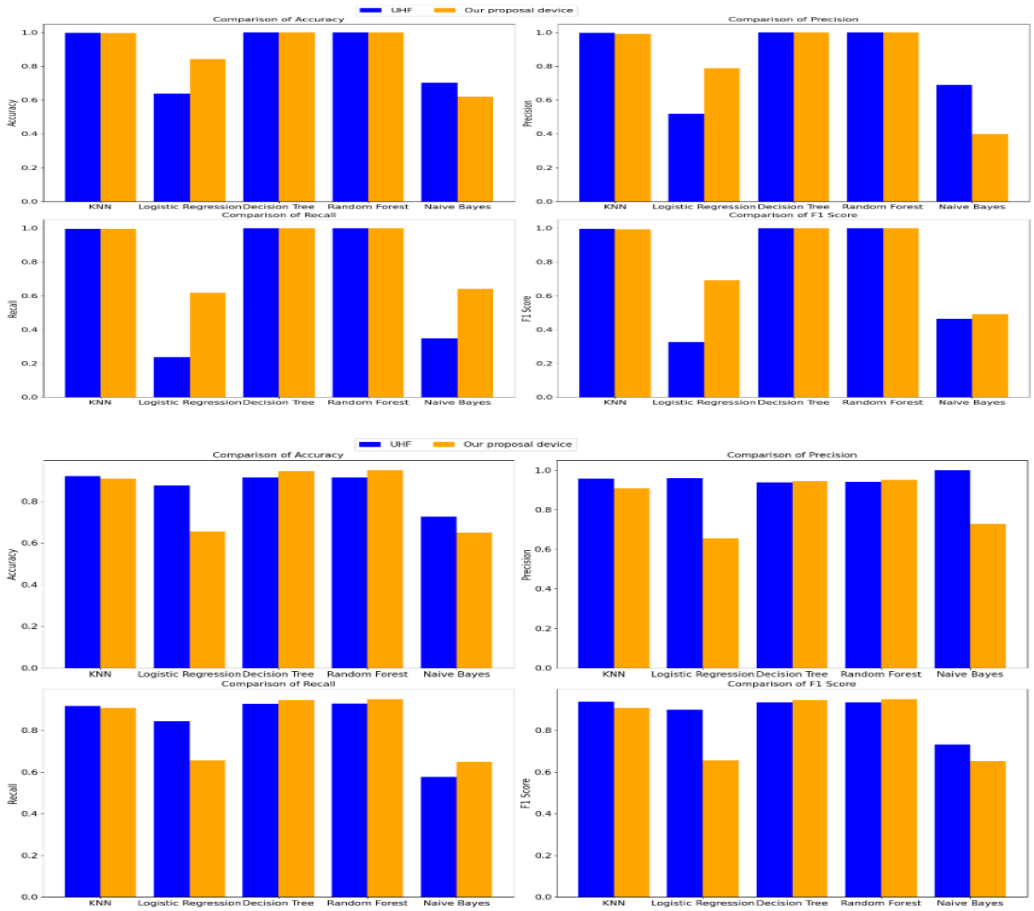


Figure 9: Performance metrics using supervised learning for UHF and proposal (a) indoor and (b) outdoor.

4.3 Recommendation for model and device selection in RFID Environments

As illustrated in Figure. 9a, both the UHF and proposed devices and models demonstrate strong overall performance; however, notable differences emerge based on environmental context. In the indoor setting as in Figure. 9a, the empirical results indicate that optimal model selection for RFID systems is highly contingent upon environmental conditions. In indoor settings, where signal interference is minimal, the proposed device consistently outperforms the commercial UHF reader, particularly when combined with classifiers capable of balanced generalization. LR, KNN, and RF exhibit strong and reliable performance, with KNN achieving near-perfect accuracy (99.57%). Although decision tree and random forests attain perfect scores, their performance may reflect overfitting and thus require careful regularization. Conversely, Naïve Bayes performs relatively poorly in the indoor scenario under the proposed device (62.04%) and should be avoided in such contexts.

In outdoor environments shown in Figure. 9b, where signal variability is more pronounced, both the UHF

system and the proposed device demonstrate comparable effectiveness. DT and RF remain the most accurate classifiers about 94.44% and 94.98%, respectively, followed closely by KNN (90.85%), confirming their robustness under diverse signal conditions. However, logistic regression and Naïve Bayes yield substantially lower accuracy levels (approximately 65%), suggesting their limited applicability in outdoor deployments. Overall, the proposed device is recommended for indoor applications, especially when used with LR, KNN, or RF. For outdoor environments, model selection should prioritize decision tree, random forest, or KNN, regardless of the device. Classifiers such as Naïve Bayes and logistic regression are not advisable for use in outdoor settings due to their diminished performance. These recommendations serve as practical guidelines for deploying RFID-based classification systems with improved accuracy and environmental adaptability.

4.4 Recommended usage for indoor and outdoor environments based on deep learning

The chart in Table 4 compares different models based on their precision, recall, F1 score, and training times, using orange for the proposal and blue for UHF in indoor and outdoor environments. In the indoor environment, our proposal device outperforms UHF in both accuracy and training efficiency. MLP achieves the best performance with 98.71% precision and a high F1 score of 98.69%, followed by LiCNN, which also performs well with a shorter training time. LSTM, GRU, and FCN deliver strong results, while CNN has the lowest accuracy among indoor models. Indoor UHF models also perform well, with MLP and LSTM being the top choices, but their training times are significantly longer compared to proposal models. In the outdoor environment, the performance gap between the proposal and UHF is minimal, as both readers produce similar results across models. LSTM is the best performing model with 91.71% precision and an F1 score of 88.70%, followed by CNN, GRU, and FCN. MLP and RNN also show strong results, while LiCNN has the lowest accuracy. Unlike indoor models, outdoor models generally have shorter training times, making CNN, RNN, or MLP more practical for real-time applications.

The comparative analysis reveals that the proposed device demonstrates superior performance in indoor environments, see Table 4, particularly when paired with sequence-based models such as GRU (99.13%) and LSTM (97.82%), highlighting its strength in capturing temporal dependencies under stable conditions. In contrast, the UHF reader lags behind in most indoor scenarios, with the exception of LSTM and CNN, which perform competitively. Models such as GRU, KNN, and MLP exhibit consistently high accuracy across both environments, underscoring their robustness and generalizability. Additionally, CNN and LiCNN not only perform well in terms of accuracy but also offer faster training times, making them suitable for real-time or resource-constrained applications. In outdoor settings, the performance gap between the proposed device and UHF narrows significantly, suggesting that environmental variability may diminish the architectural advantages of the proposed device. In this context, CNN (92.75%), GRU (92.85%), and MLP (92.27%) emerge as the most reliable classifiers, regardless of the device used. From an efficiency perspective, Naïve Bayes and decision tree are the fastest models; however, Naïve Bayes shows limited accuracy, particularly in indoor scenarios. MLP offers a favorable trade-off between computational efficiency and predictive performance, making it an optimal candidate for deployment in both environments. Notably, the perfect accuracy scores achieved by decision tree models may indicate overfitting, rather than true generalization, warranting caution in their use without further validation. In conclusion, for accuracy-critical indoor applications, the proposed device combined with GRU or Random Forest is recommended. For real-time or latency-sensitive scenarios, MLP or CNN provides a more balanced and practical solution.

Table 2: Model performance across commercial UHF-RFID and our proposal devices both indoor and outdoor environments.

Model	**Indoor Accuracy (%)**		**Outdoor Accuracy (%)**		**Training speed**	**Recommended Usage**
	Proposal	**Commercial (UHF)**	**Proposal**	**Commercial (UHF)**		
GRU	99.13	83.1	92.85	92.5	Medium	Indoor: Proposal Outdoor: Proposal
LSTM	97.82	98.23	91.70	93.11	Slow	Indoor: Commercial Outdoor: Commercial
RNN	95.8	85.73	91.1	91.85	Slow	Indoor: Proposal Outdoor: Commercial
CNN	95.65	97.96	92.75	92.10	Fast	Indoor: Commercial Outdoor: Proposal
LiCNN	97.41	88.64	90.88	90.45	Fast	Indoor: Proposal Outdoor: Proposal
MLP	94.83	89.93	92.27	91.90	Very Fast	Indoor: Proposal Outdoor: Proposal
FCN	87.1	82.13	91.0	90.85	Medium	Indoor: Proposal Outdoor: Proposal
Naïve Bayes	62.04	70.13	65.2	65.0	Fastest	Indoor: Commercial Outdoor: Proposal
Decision Tree	100.0	100.0	92.0	92.0	Very fast	Indoor: Comm/Prop Outdoor: Prop/Comm

4.5 Performance analysis of lab attendance based on using both commercial and proposal readers

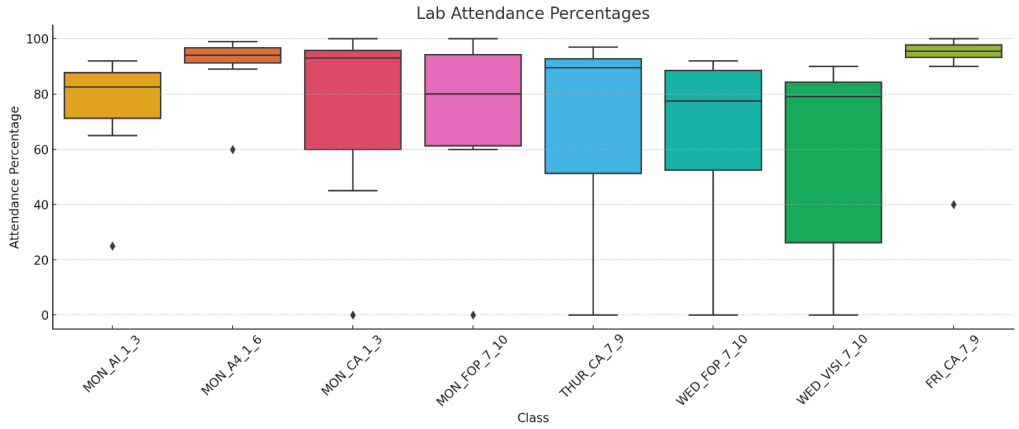


Figure 10: The average percentage of students using commercial UHF-RFID and our proposal readers.

Figure. 10 shows the lab attendance percentages based on 8 classes of AI (Artificial Intelligence), Computer Architecture (CA), Fundamental of Programming (FoP), and Very-large-scale integration (VLSI) courses during semester I of 2024-2025 academic year. Overall attendance is relatively high, with medians clustering between 80% and 95%, indicating strong student engagement. Next, consistent classes of FRI_CA_7_9 and MON_A4_1_6 show high medians and narrow interquartile ranges (IQR), reflecting consistent attendance

with minimal fluctuation. These are likely well-managed sessions with predictable schedules or motivated student cohorts. While MON_FOP_7_10, WED_FOP_7_10, and WED_VISI_7_10 sessions exhibit wide IQRs and notable outliers (e.g., 0%, 5%, 10%), indicating inconsistent attendance. These variations could be due to scheduling conflicts, course difficulty, or reduced student interest. Several classes include zero or near-zero attendance outliers, which may warrant further investigation (e.g., system logging errors, absenteeism, or class cancellations). Classes with high variance or low minimums should be reviewed for potential improvements in scheduling, engagement strategies, or attendance monitoring systems.

5. CONCLUSION

This research focuses on enhancing UHF-RFID tag detection by applying machine learning (ML) techniques. It addresses key limitations in existing UHF-RFID systems—such as signal attenuation, environmental variability, and radio interference—by introducing a versatile dataset and proposing novel methodologies designed to improve detection accuracy in both indoor and outdoor environments. To address the challenges associated with reliable RFID tag detection, this paper proposes a custom-designed UHF-RFID detection framework that dynamically classifies and adjusts transmission power levels across varying environmental conditions, thereby optimizing tag detection. This automated power adjustment is particularly valuable in large-scale RFID deployments. The study evaluates several deep learning models and compares the proposed framework with conventional UHF-RFID systems. Tailored for indoor scenarios, the framework demonstrates superior performance in terms of accuracy and F1 score, particularly when implemented using architecture such as Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN).

Acknowledgements: This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2024-28-08.

REFERENCES

[1] T. Adeniran, Y. Sanni, N. Faruk, and L. Olawoyin, "Design and implementation of an automated attendance monitoring system for a nigerian university using rfid," *African J. Comput. ICT*, vol. 12, no. 2, pp. 72–89, 2019.

[2] S. Armstrong, "Circular polarization vs. linear polarization: Which is the right rfid antenna?" *Linear Polarization: Which Is the Right RFID Antenna*, 2013.

[3] Q. H. Chiem, K. T. Huynh, M. T. Nguyen, M. D. Tran, X. P. P. Nguyen, and T. N. Ly, "Attendance monitoring using adjustable power uhf rfid and web-based real-time automated information system," in *International Conference on Future Data and Security Engineering*. Springer, 2021, pp. 392–407.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[5] L. Cui, Z. Zhang, N. Gao, Z. Meng, and Z. Li, "Radio frequency identification and sensing techniques and their applications—a review of the state-of-the-art," *Sensors*, vol. 19, no. 18, p. 4012, 2019.

[6] G. P. Ferreira, L. J. Matos, and J. M. Silva, "Improvement of outdoor signal strength prediction in uhf band by artificial neural network," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 12, pp. 5404–5410, 2016.

[7] U. Koppikar, S. Hiremath, A. Shiralkar, A. Rajoor, and V. Baligar, "Iot based smart attendance monitoring system using rfid," in *2019 1st International Conference on advances in information technology (ICAIT)*. IEEE, 2019, pp. 193–197.

[8] C. T. Li, J. C. Cheng, and K. Chen, "Top 10 technologies for indoor positioning on construction sites," *Automation in Construction*, vol. 118, p. 103309, 2020.

[9] T. N. Ly, P. N. Xuan-Phuc, T. H. Nghia, N. T. Son, T. T. Thao, and T. T. Tien, "An intelligent secure framework for face detection distribution quantization optimization on edge devices," in *Proc of IEEE International Conference on Industrial Networks and Intelligent Systems*, Danang City, Vietnam, Feb 2025. (Accepted).

[10] T. Montanaro, I. Sergi, A. Motroni, A. Buffi, P. Nepa, M. Pirozzi, L. Catarinucci, R. Colella, F. P. Chietera, and L. Patrono, "An iot-aware smart system exploiting the electromagnetic behavior of uhf-rfid tags to improve worker safety in outdoor environments," *Electronics*, vol. 11, no. 5, p. 717, 2022.

[11] T. Nguyen-Thi-Thanh, X. P. Phan-Nguyen, and N. Ly-Tu, "Machine learning uhf-rfid to support video tracking and recommendation for attendance system," in *International Conference on Future Data and Security Engineering*. Springer, 2024, pp. 308–321.

[12] Q. T. Pham and T. P. Tran, "The acceptance of e-learning systems and the learning outcome of students at universities in vietnam," *Knowledge Management & E-Learning*, vol. 12, no. 1, pp. 63–84, 2020.

[13] A. Pinkus, "Approximation theory of the mlp model in neural networks," *Acta Numerica*, vol. 8, pp. 143–195, 1999.

[14] ST, "Software development kit for st25ru3993 based evaluation boards."

[15] ST, "Ultra-low-power with fpu arm cortex-m4 mcu 80 mhz with 1 mbyte of flash memory, lcd, usb otg, dfsdm."

[16] R. Store, "What are the different types of rfid antennas?" https://rfid4ustore.in/blog/types-of-rfid-antennas/, 2025.

[17] K. Tanaka, Y. Kimuro, K. Yamano, M. Hirayama, E. Kondo, and M. Matsumoto, "A supervised learning approach to robot localization using a short-range rfid sensor," *IEICE Transactions on Information and Systems*, vol. 90, no. 11, pp. 1762–1771, 2007.

[18] D. A. Tesch, E. L. Berz, and F. P. Hessel, "Rfid indoor localization based on doppler effect," in *Sixteenth international symposium on quality electronic design*. IEEE, 2015, pp. 556–560.

[19] H. Q. Tran and C. Ha, "Machine learning in indoor visible light positioning systems: A review," *Neurocomputing*, vol. 491, pp. 117–131, 2022.

[20] M. D. Tran, K. T. Huynh, V. H. Pham, A. T. Phan, Q. K. Nguyen, X. P. P. Nguyen, and T. N. Ly, "performance analysis of automatic integrated long-range rfid and webcam system," *SN Computer Science*, vol. 3, no. 6, p. 473, 2022.

[21] S. Woo, J. Byun, S. Kim, H. M. Nguyen, J. Im, and D. Kim, "Rnn-based personalized activity recognition in multi-person environment using rfid," in *2016 IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, 2016, pp. 708–715.

[22] R. Yadav, "Light-weighted cnn for text classification," *arXiv preprint arXiv:2004.07922*, 2020.

[23] M. Zhou, Y. Tang, Z. Tian, and X. Geng, "Semi-supervised learning for indoor hybrid fingerprint database calibration with low effort," *IEEE Access*, vol. 5, pp. 4388–4400, 2017.

[24] B. Sivasankari, R. Vishwamoorthy, K. S. Kannan, B. Nithyaprakash, and S. Sabarish, "RFID Monitoring System Using Machine Learning: A Design Thinking Approach for Smarter Attendance Solutions," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2024, pp. 1816-1820.

[25] K. Ishaq and S. Bibi, "IoT based smart attendance system using RFID: A systematic literature review," *arXiv preprint arXiv:2308.02591*, 2023.

[26] S. C. Ellis, S. Rao, D. Raju, and T. J. Goldsby, "RFID tag performance: Linking the laboratory to the field through unsupervised learning," *Production and Operations Management*, vol. 27, no. 10, pp. 1834–1848, 2018.

[27] P. Dan, "Extending K-means with Efficient Estimation of the Number of Clusters," in *Proceedings of the 17th International Conf. on Machine Learning, ICML-2000*, 2000, pp. 727–734.

[28] S. Subedi, E. Pauls, and Y. D. Zhang, "Accurate localization and tracking of a passive RFID reader based on RSSI measurements," *IEEE Journal of Radio Frequency Identification*, vol. 1, no. 2, pp. 144–154, 2017.