IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 14, 2022, accepted March 2, 2022, date of publication March 15, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158056

ML-Based Classification of Device Environment Using Wi-Fi and Cellular Signal Measurements

ARUN RAMAMURTHY[1], (Member, IEEE), VANLIN SATHYA[2], (Member, IEEE), MUHAMMAD IQBAL ROCHMAN[3], (Member, IEEE), AND MONISHA GHOSH[4], (Fellow, IEEE)

[1]TCS Research and Innovation, Tata Consultancy Services, Pune 411013, India
[2]Celona Inc., Cupertino, CA 95014, USA
[3]Department of Computer Science and Engineering, University of Chicago, Chicago, IL 60637, USA
[4]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

Corresponding author: Arun Ramamurthy (arunramamurthy94@gmail.com)

This work was supported by the National Science Foundation (NSF) under Grant CNS - 1618920 and Grant CNS - 1618836.

ABSTRACT Future spectrum sharing rules very likely will be based on device environment: indoors or outdoors. For example, the 6 GHz rules created different power regimes for unlicensed devices to protect incumbents: "indoor" devices, subject to lower transmit powers but not required to access an Automatic Frequency Control database to obtain permission to use a channel, and "outdoor" devices, allowed to transmit at higher power but required to do so to determine channel availability. However, since there are no reliable means of determining if a wireless device is indoors or outdoors, other restrictions were mandated: reduced power for client devices and indoor access points that cannot be battery powered, have detachable antennas or be weatherized. These constraints lead to sub-optimal spectrum usage and potential for misuse. Hence, there is a need for robust identification of device environments to enable spectrum sharing. In this paper we study automatic indoor/outdoor classification based on the radio frequency (RF) environment experienced by a device. Using a custom Android app, we first create a labeled data set of a number of parameters of Wi-Fi and cellular signals in various indoor and outdoor environments, and then evaluate the classification performance of various machine learning (ML) models on this data set. We find that tree-based ensemble ML models can achieve greater than 99% test accuracy and F1-Score, thus allowing devices to self-identify their environment and adapt their transmit power accordingly.

INDEX TERMS 5G, Wi-Fi, indoor, outdoor, classification, machine learning.

I. INTRODUCTION AND MOTIVATION

As the current generation of cellular (5G) and Wi-Fi (802.11ax) networks begin to be widely deployed, it is becoming increasingly clear that the next generation of wireless systems will largely be deployed in spectrum that is shared, not only between cellular and Wi-Fi but also with various incumbents such as federal radar systems, fixed microwave links, satellite providers, weather satellites and broadcast auxiliary services (BAS). While mmWave and higher frequencies, into the terahertz range, offer the widest bandwidths and fewer incumbents, mid-band spectrum (1 GHz – 10 GHz) will always remain the workhorse of wireless networks due to the favorable propagation characteristics that balance range with bandwidth. This band is also the most crowded. The Federal Communications Commission (FCC) in the U.S recently created rules for the 6 GHz band that would allow unlicensed services to coexist with existing incumbents in the band, mainly high-power fixed microwave links and lower power broadcast auxiliary services [1]. It is expected that in addition to Wi-Fi, this band will also be used by cellular systems deploying 5G NR-U [2] similar to the use of the 5 GHz band by LAA [3]. Fig. 1 depicts the deployment scenario in this band, with a mix of indoor and outdoor devices using Wi-Fi and 5G NR-U with various power levels.

Need for automatic and robust indoor/outdoor classification: The 6 GHz rules [1] create two different power regimes for unlicensed devices: "indoor" devices that are subject to lower transmit powers (Low Power Indoors or LPI) but are not required to access an Automatic Frequency Control (AFC) database to obtain permission to use a channel,

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Seo Kim.

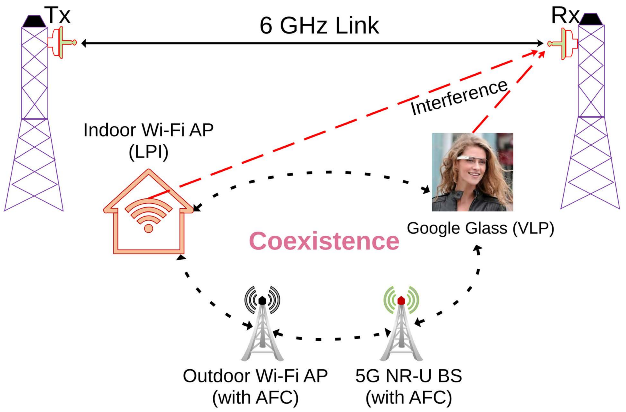


FIGURE 1. Deployment scenario of indoor and outdoor devices in 6 GHz.

and "outdoor" devices, that can transmit at a higher power but need to consult an AFC prior to using a channel to ensure that the device is not in the exclusion zone for the desired channel, as shown in Fig. 1. Very Low Power (VLP) devices shown in Fig. 1 have yet to be authorized in the US but are permitted in other regulatory regimes: these devices can be anywhere, do not need to consult an AFC but will transmit at lower power than LPI devices.

Since there are no reliable means for determining if a wireless device is indoors or outdoors, other restrictions were mandated for LPI: indoor access points (APs) could not be battery powered, have detachable antennas or a weatherized exterior, and mobile client devices connected to an indoor AP were subject to a 6 dB lower transmit power constraint (proposed by FCC) compared to the indoor AP since they could be outdoors and pose an interference threat to incumbents if they were to transmit at the same power as an indoor device. These constraints clearly lead to sub-optimal spectrum usage: for example, client devices even if they are indoors have to transmit at a lower power, client devices cannot transmit directly to each other without connecting through an AP and APs being unable to be battery powered can lead to a less resilient network. *Hence, the ability of a wireless device to reliably detect its own environment allows device power allocations that do not need to be constrained by external factors and can lead to improved spectrum utilization and increased resilience*.

The fundamental premise of our approach in this paper is simple: just as the indoor visual environment is quite different from the outdoor visual environment, the same is true for the radio frequency (RF) environment as well. RF transmissions permeate our surroundings: television (TV), radio (AM/FM), cellular and Wi-Fi being the most pervasive. The most obvious difference between indoor and outdoor environments is the signal strength: transmissions from outdoor sources such as Global Positioning System (GPS) satellites, TV transmitters, cellular towers and radio stations will be received at higher power outdoors while predominantly indoor transmitters such as Wi-Fi will have higher signal strength indoors. There are other differences as well: the number of Wi-Fi APs and cellular base-stations (BSs) received by a device such as a smartphone will also depend on the environment. Today, it is possible to extract detailed information on both signal strength and number of Wi-Fi APs and cellular BSs received by a smartphone, over frequency bands from the unlicensed 2.4 GHz and 5 GHz bands to the low (< 1 GHz), mid (1 GHz - 6 GHz) and high (> 24 GHz) cellular bands, directly, using apps. We posit that such a data-set, collected in labeled indoor and outdoor environments, across a wide variety of frequency bands and signal types, can be used to train Machine Learning (ML) models that can perform robust indoor/outdoor classification, thus leading to improved spectrum usage, incumbent protection and resilience, not only in 6 GHz, but also in future bands such as the 12 GHz satellite band where sharing with indoor devices is being considered [4].

The contributions of the paper are as follows: (i) we developed an Android app and collected a large, labeled data-set of Wi-Fi, 4G LTE, and 5G NR measurements in various indoor and outdoor environments: such a data-set does not exist today and will be made openly available to other researchers; (ii) we have evaluated various ML algorithms on this data set and shown classification accuracy of 99%; and (iii) we evaluated the ML models which were already trained on real data collected from smartphone. We believe that this is the first comprehensive evaluation of the efficacy of ML in solving the indoor/outdoor classification problem solely using RF data.

The paper is organized as follows. Section II provides a brief summary of literature on indoor/outdoor classification, Section III describes the advantages of signal in terms of Wi-Fi and Cellular network, Section IV describes the app, our data collection methodologies and pre-processing procedures, Section V presents a detailed performance evaluation of different ML algorithms, Section VI presents performance results from specific test scenarios and conclusions and future research directions are presented in Section VII.

II. RELATED WORK IN INDOOR/OUTDOOR CLASSIFICATION

As mentioned in the previous section, indoor and outdoor environments are visually very different. Hence, the indoor/outdoor classification problem has been studied quite extensively in the image processing literature, mainly to perform scene analysis for various applications. There have also been a few contributions in using GPS and limited signal information from smartphones. Additionally, a number of recent studies [5], [6] have investigated the use of ML, including Convolutional Neural Networks (CNNs) for RF Fingerprinting, but the objective is quite different from the indoor/outdoor classification problem we address in our work.

A. INDOOR/OUTDOOR CLASSIFICATION BASED ON IMAGES

Scene classification in general is a well studied area and there has been some specific work on indoor/outdoor classification in content-based image retrieval systems [7]. In [8], the author classifies scenes as indoor versus outdoor using relevant low level features such as color and texture which help to improve the classification performance. The proposed method uses statistical features computed from Hue, Saturation and Value (HSV) as color features, Discrete Cosine Transform (DCT) coefficients as texture feature and entropy computed with Ultra Violet (UV). In [9], the authors present a framework to benchmark indoor/outdoor scene classification and conclude that it is not possible to classify the images accurately within the database they use. The principal drawback of using image-based classification in the scenario of interest in this paper is that a camera is required which adds to the cost and complexity.

B. INDOOR/OUTDOOR CLASSIFICATION BASED ON GPS

ML algorithms were used along with GPS data to perform indoor/outdoor classification in [10]–[12], for applications such as activity classification of indoor/outdoor activities. Data collected from a certain number of GPS satellites, using the GPS sensor on mobile devices is used to train classification models. The performance of such approaches is severely limited by the reduced signal levels and accuracy of GPS indoors and in urban environments with tall buildings.

C. INDOOR/OUTDOOR CLASSIFICATION BASED ON SMARTPHONES

Fast and accurate detection of the physical location of a mobile device is crucial for many new smart services in the current 5G and future mobile networks. This information can be used by mobile operators to optimize their network to provide better quality of service (QoS) to their customers. For example, Femto cells deployed indoors need to quickly determine what the user-environment is in order to reduce hand-over delay and avoid ping-pong effects [13]–[15]. In [16], an ensemble learning scheme for indoor-outdoor classification is proposed for a specific urban area consisting of five malls, based on the cellular data captured in a commercial LTE network. The variables are extracted by network key performance indicators (KPIs) and radio propagation knowledge. Based on these main variables, the Gini metric is used to build the classification and regression trees. Only cellular signal strengths are used in the study and data collection is only on a particular LTE frequency band: 2.1 GHz.

Thus, the existing literature on using RF signals captured over a wide range of frequencies spanning Wi-Fi and cellular in all available bands, including mmWave, for indoor/outdoor classification is extremely sparse. One of the main reasons is the lack of sufficiently large and diverse data-sets that can be used by ML algorithms. With the availability of new bands on smartphones and apps that can retrieve the data easily for analysis, we aim to address this deficiency by creating large data-sets and evaluating ML algorithms to solve this extremely relevant classification problem.

III. RATIONALE FOR USING WI-FI AND CELLULAR SIGNALS FOR ENVIRONMENT CLASSIFICATION

While there are a number of different RF signals that could be used to classify environments, such as television, AM/FM radio, Bluetooth, ultra-wideband (UWB) etc., we choose to use Wi-Fi and cellular signals primarily due to their ubiquity, globally, indoors and outdoors, and their support in most consumer devices. For example, Bluetooth and UWB are not supported in all mobile devices and are only available in limited locations. We leverage the following characteristics of cellular and Wi-Fi signals in building the classifiers:

A. DEPLOYMENT

Cellular networks are usually deployed over licensed frequencies by network operators in a planned manner, outdoors, on a wide range of frequencies in the low (< 1 GHz), mid (1 - 6 GHz) and high (> 24 GHz) bands, with different propagation characteristics. For example the high bands do not propagate very far and outdoor-to-indoor-penetration is poor: hence these bands are rarely encountered in an indoor environment. On the other hand, Wi-Fi APs are mostly deployed indoors, with limited outdoor deployments, in the 2.4 GHz and 5 GHz bands, with 6 GHz Wi-Fi just beginning to be deployed. Further, the transmission power of both cellular and Wi-Fi signals vary depending on the bands they are deployed in, which again leads to differences in received signal strengths indoors and outdoors. Hence, these features can be used by ML classifiers to classify device environments as will be described later.

B. PROTOCOL

Cellular and Wi-Fi systems have different medium access control (MAC) protocols. Cellular networks rely on a centralized protocol that utilizes strict scheduling to serve multiple concurrent users whereas Wi-Fi uses a listen-before-talk (LBT) protocol to avoid collisions between users. In both cases however, synchronization signals are continually transmitted even in the absence of any active, connected devices. Both Wi-Fi and Cellular networks periodically transmit signals about every 100 ms, *i.e.*, Wi-Fi beacon packets and LTE/NR synchronization signals on the Physical Downlink Control Channel (PDCCH). The signals strengths of these reference signals are continuously being measured by the modem in the phone and are available over the Android APIs.

Thus, our classification methodology is based around extracting the signal strength measurements of all Wi-Fi APs and cellular base-stations, combining them with the GPS signal and crafting features to be used by ML classifiers as will be described in the following sections.

IV. DATA COLLECTION METHODOLOGY

A. APP FEATURES

We developed an easy-to-use Android app, SigCap [17], that *passively* collects GPS, Wi-Fi, 4G and 5G information using the Android API, without requiring root access or

TABLE 1. Parameters collected by the android app.

Category	Features
4G LTE Information	**All cells:** Physical Cell ID (PCI), Frequency of operation (EARFCN), Band, Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Received Signal Strength Indicator (RSSI). **Primary cell**: In addition to above, Bandwidth
5G NR Information	5G-RSRP, 5G-RSRQ, and 5G Frequency (can only identify between FR1 & FR2)
Wi-Fi 2.4 GHz and 5 GHz	Basic Service Set Identifier (BSSID), Frequency, Bandwidth, RSSI
Location	GPS coordinates (latitude, longitude, altitude), GPS accuracy
Timestamp	Data and time

running bandwidth hungry speed-tests. Table 1 shows the parameters collected every 10 secs, which is the minimum interval allowed by the API to conserve power. Depending on the phone capabilities and the cellular deployment, the app collects the listed data on all deployed 4G bands including the unlicensed 5 GHz (Band 46) and Citizen Broadband Radio Service (CBRS, Band 48) and on all 5G bands including mmWave. Each data record collected consists of the following parameters: (i) time-stamp; (ii) location (GPS latitude and longitude) (iii) GPS accuracy[1]; (iv) Wi-Fi information as listed in Table 1 for every AP the device can receive a beacon frame from on 2.4 GHz and 5 GHz, even if the device is not associated with any of them; (v) 4G LTE cell information as listed in Table 1 for every cell it can receive, not just the one it is connected to: this includes secondary channels from the same BS and channels from neighboring BSs and (vi) 5G NR information as listed for only the frequency band the phone is connected to: either FR1 ($<$ 6 GHz)) or FR2 ($>$ 6 GHz), *i.e.* information on secondary and neighboring 5G NR cells is not available in either band.

B. APP USAGE AND DATA PRE-PROCESSING

Once the data has been collected and saved on the phone, it can be exported as JavaScript Object Notation (JSON) files, which then converted to Comma Separated Values (CSV) format and used as inputs to the ML models. A JSON record at a given timestamp is displayed as a single row containing the measurements from multiple Wi-Fi APs and cellular BSs, both 4G and 5G. Each row contains multiple columns: location label, GPS accuracy, date-time information, Wi-Fi APs information on 2.4 & 5 GHz, 4G information separated by band and 5G information. For each frequency range (2.4 or 5 GHz), the Wi-Fi columns are further separated into the following: number of APs, average RSSI, and list of RSSI from highest to lowest. Similarly, the band-categorized 4G columns contain further details: number of cells and the average and list of signal powers (RSRP, RSRQ, and RSSI) sorted in descending order.

We have incorporated an user-entered category field in the app prior to exporting the captured data with the following options: unknown, indoor, outdoor, mostly indoor and mostly outdoor. Unknown, indoor, and outdoor labels are self-descriptive, while mostly indoor and mostly outdoor labels are used when users transitioned between indoors and outdoors during the measurement run.[2] This helps us collect labeled data in various environments. For the purposes of the work reported in this paper, we only used data labeled as ‘‘indoor’’ or ‘‘outdoor’’ for training and testing the models.

TABLE 2. Summary of data collected.

Location	Indoor	Outdoor	Total
Chicago	9049	6850	15899
Colorado	605	20283	20888
New York	45742	130241	175983
Total	55396	157374	212770

While the calculation of the parameters collected, especially RSSI, RSRP and RSRQ, are explicitly defined in the standards, the values themselves depend on the implementation on the modem chip in the phone as well as the receiving antennas and front-ends. Hence, it is important to collect data with a wide range of devices and on various operator deployments. We have used several Android phones in our data collection effort: Google Pixel 2, Google Pixel 3, Google Pixel 5, Samsung Galaxy S9, Samsung Galaxy S20, Samsung Galaxy S21 and Motorola Edge+, each equipped with a Subscriber Identification Module (SIM) of a different operator. Outdoor measurements were collected while walking, biking, driving a car, and riding on trains in urban, suburban and rural environments. Indoor measurements were made in single-family houses, apartment buildings, offices, indoor malls and stores. Both indoor and outdoor measurements were collected in the various places described in three different geographical locations: the number of data records from each location is summarized in Table. 2. It should be noted here that the number and diversity of outdoor environments captured was greater than indoor, since access to most indoor places was restricted due to shut-downs in the past year.

The data-set thus collected is quite large: 18 GB. Depending on the measurement environment, a single data record can contain information on 100s of Wi-Fi APs and many 10s of LTE cells: this is common in dense urban areas. Thus, the raw data cannot be used directly in a ML classifier since the

[1] The GPS accuracy is defined as the horizontal radius in meters of 68% confidence. In other words, an accuracy of x meter defines a circle with x meters radius which there is a 68% probability that the true location is inside it.

[2] Since some of our data was collected by other users who we shared SigCap with, we wanted to make sure that users labeled the data collection environment correctly. However, these two labels lead to high error due to the ambiguity, thus they were not used in our analysis

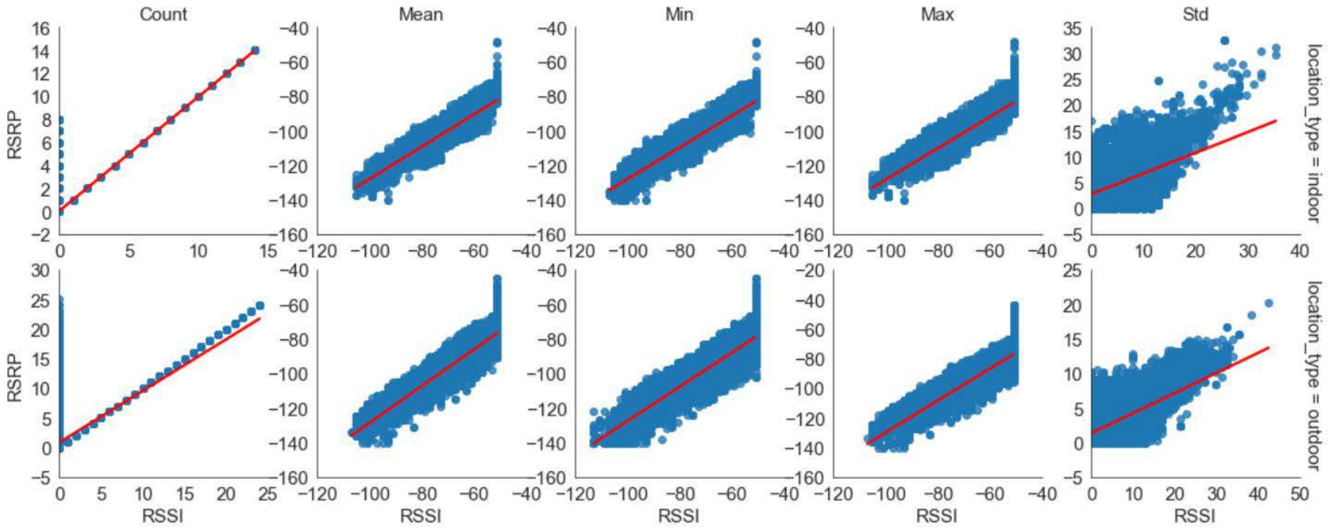


FIGURE 2. Correlation analysis: LTE (Low + Mid) RSSI and RSRP features.

TABLE 3. Features used in statistical analysis.

Features	Count	Mean	Min	Max	Std
Wi-Fi 2.4 GHz RSSI	✓	✓	✓	✓	✓
Wi-Fi 5 GHz RSSI	✓	✓	✓	✓	✓
LTE Low RSRP	✓	✓	✓	✓	✓
LTE Mid RSRP	✓	✓	✓	✓	✓
LTE Low RSRQ	✓	✓	✓	✓	✓
LTE Mid RSRQ	✓	✓	✓	✓	✓
NR FR1 RSRP	✓	✓	×	×	×
NR FR2 RSRP	✓	✓	×	×	×
NR FR1 RSRQ	✓	✓	×	×	×
NR FR2 RSRQ	✓	✓	×	×	×

number of inputs per data record would vary with each record. We pre-process the data as follows:

Step 1: the data is cleaned by removing any record with an invalid entry in any field. Null entries, NaN (Not a Number) entries and RSRP, RSRQ and RSSI values that do not fall in the specified range[3] for these parameters are examples of invalid entries. Invalid entries are represented by a very large number, *i.e.,* −200 so the ML algorithm will not be affected by these data points.

Step 2: in order to create a fixed number of features irrespective of the length of a data record, we extract the features listed in Table 3 for each record. A single record contains information from multiple Wi-Fi and LTE cells, the number of which varies from one record to the next. In order to have the same number of features for each record, we calculate aggregate values as feature inputs to the ML model as follows. As shown in Table 3 we first classify the signals into bands: Wi-Fi RSSI in 2.4 GHz and 5 GHz, LTE RSRP and RSRQ in low- and mid-band and NR RSRP and RSRQ in FR1 and FR2. Then, all signal values detected in a band are aggregated using 5 functions: mean, min, max, standard deviation, and count, thus removing the variability between records. This pre-processing ensures that the number of input features for each record is the the same irrespective of the number of actual Wi-Fi and LTE signals detected. 5G NR deployments do not aggregate bands today: thus aggregation is not necessary for NR FR1 and FR2 data. However this may change in future 5G deployments. Hence, we have 6 signal categories (*i.e.*, Wi-Fi 2.4 GHz RSSI, Wi-Fi 5 GHz RSSI, LTE Low RSRP, LTE Mid RSRP, LTE Low RSRQ, LTE Mid RSRQ) with 5 features each, and 4 NR categories(*i.e.*, NR FR1 RSRP, NR FR2 RSRP, NR FR1 RSRQ, NR FR2 RSRQ) with 2 features each, bringing the total number of features to 38. The 38 features in Table 3 combined with "GPS Accuracy" results in a set of 39 features for use with classification algorithms.

V. ML ALGORITHM EVALUATION

Before proceeding to ML classification, we performed univariate analysis on the various features to evaluate the statistical differences between indoor and outdoor data. Other than the LTE RSSI, we have used all the other Wi-Fi and cellular features (as shown in Table 3), as input to the ML model. This is because we observed in the correlation analysis that the LTE (Low + Mid) RSSI and RSRP are highly correlated, as shown in Fig. 2, since RSSI is a function of RSRP and RSRQ and hence, the LTE RSRP feature alone is sufficient for further ML analysis. Similarly, from Fig. 3, we see that LTE RSRP and RSRQ are not highly correlated because of the step-function behavior in the RSRQ range (*i.e.*, 0 dB (highest signal quality) to −20 dB (low signal quality)). This is also expected since RSRQ is a measure of interference due to neighboring cells while RSRP is measured on the primary cell. Hence, we include RSRQ as a feature in our further ML analysis. Figs. 4a, 4b, 5a,

[3] For 2.4 and 5 GHz Wi-Fi, RSSI varies in the range of −100 dbm to −20 dBm. In LTE, low-band RSRP varies in the range of −150 dBm to −35 dBm and mid-band RSRP varies in the range of −130 dBm to −50 dBm. Similarly, in NR FR1 RSRP varies in the range of −120 dBm to −60 dBm and in NR FR2 RSRP varies in the range of −120 dBm to −70 dBm. The RSRQ is in the range of −20 to −5

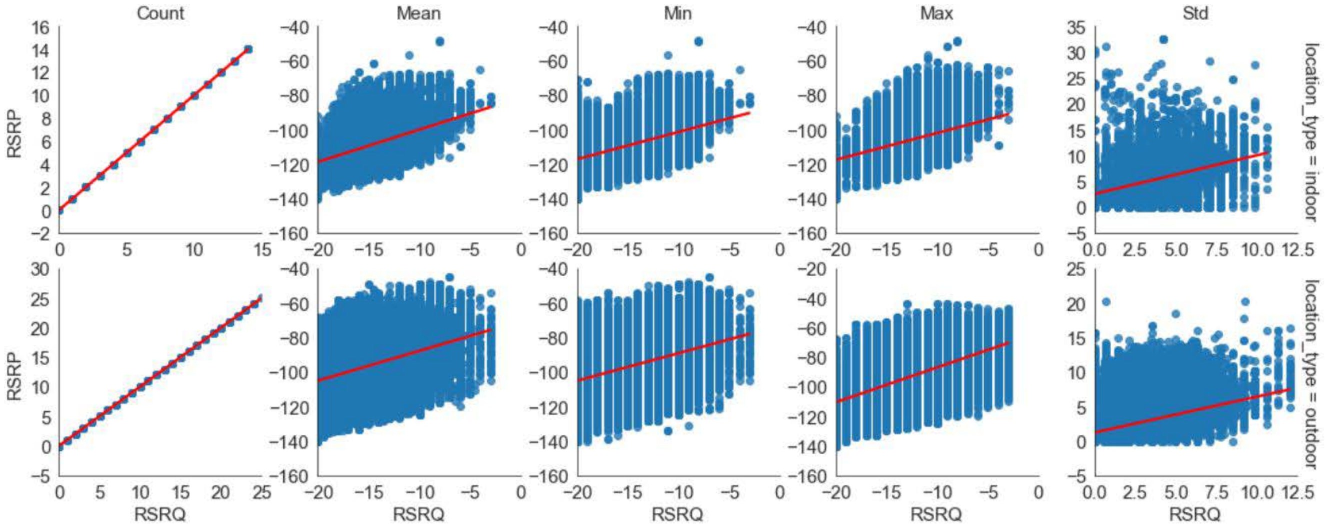
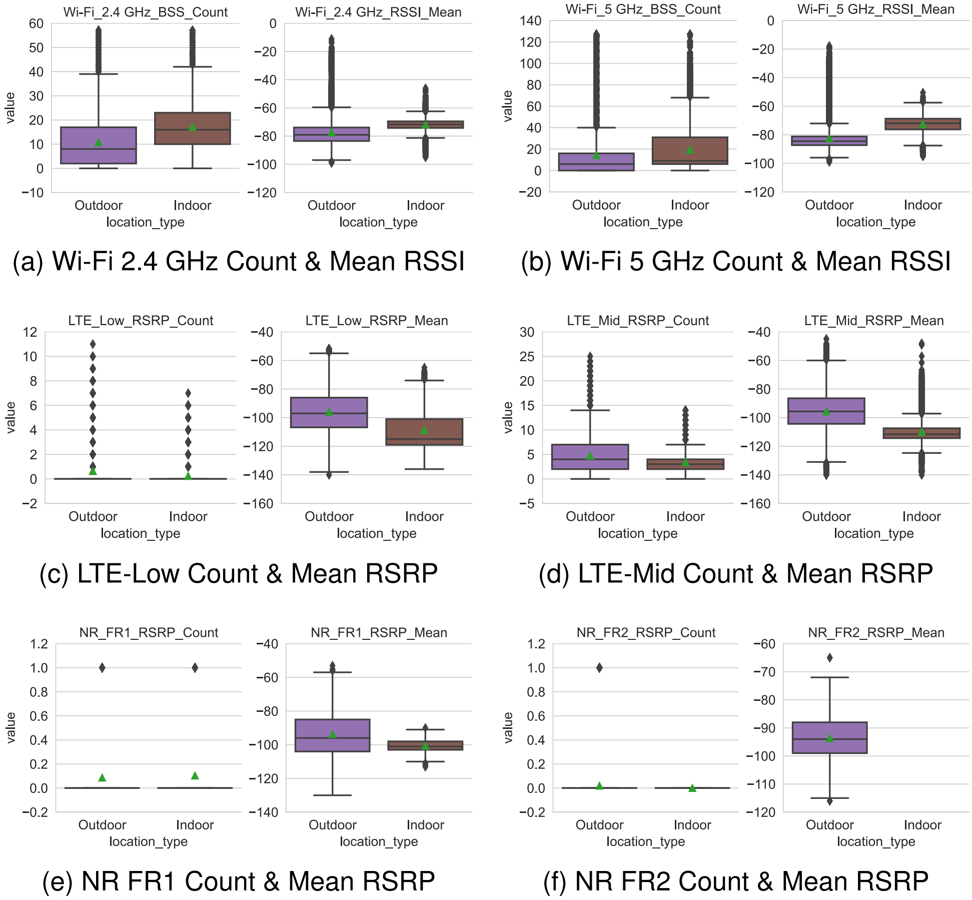


FIGURE 3. Correlation analysis: LTE (Low + Mid) RSRQ and RSRP features.



(a) Wi-Fi 2.4 GHz Count & Mean RSSI

(b) Wi-Fi 5 GHz Count & Mean RSSI

(c) LTE-Low Count & Mean RSRP

(d) LTE-Mid Count & Mean RSRP

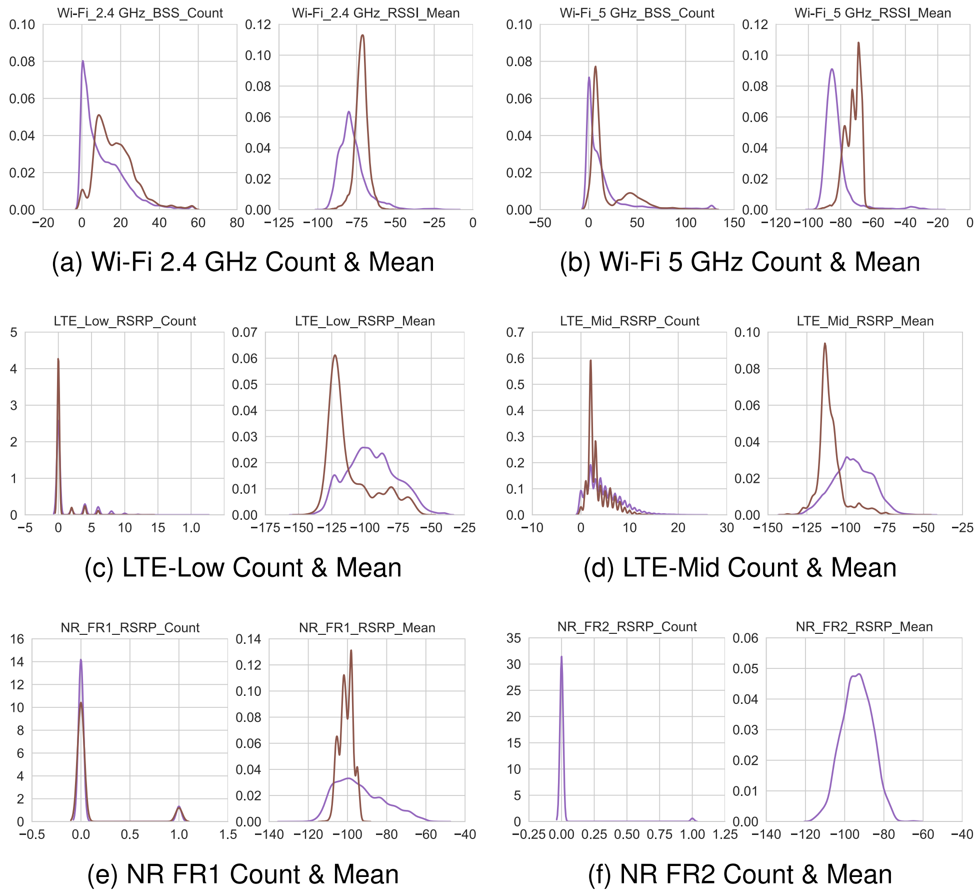
(e) NR FR1 Count & Mean RSRP

(f) NR FR2 Count & Mean RSRP

FIGURE 4. Univariate analysis on Wi-Fi, LTE and NR features.

and 5b show the distribution of Wi-Fi 2.4 GHz & 5 GHz AP count and mean RSSI[4] indoors and outdoors. We observe clear differences between indoors and outdoors in the Wi-Fi 2.4 GHz AP count and Wi-Fi 5 GHz RSSI mean. Similarly, Figs. 4c, 4d, 5c, and 5d show the LTE Low and Mid band count and mean RSRP distributions. Once again, we observe clear differences between the distributions of indoor and

[4] We also analyzed the univariate performance in terms of min, max, and standard deviation of the RSSI values. However, due to space limitation, we show only the count and mean.



(a) Wi-Fi 2.4 GHz Count & Mean

(b) Wi-Fi 5 GHz Count & Mean

(c) LTE-Low Count & Mean

(d) LTE-Mid Count & Mean

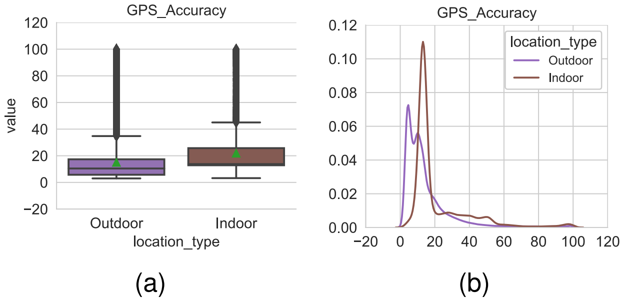
(e) NR FR1 Count & Mean

(f) NR FR2 Count & Mean

FIGURE 5. **Univariate distribution analysis on Wi-Fi, LTE and NR features.**

outdoor data. Figs. 4e, 4f, 5e, and 5f, show the count and mean RSRP for 5G NR deployment in FR1 and FR2. We have not observed any deployment of 5G NR in the FR2 band indoors due to the penetration loss of mmWave signals from outdoors to indoors. Similarly, we analyzed the GPS accuracy and we observe difference between the indoor and outdoors as shown in Fig. 6.

This preliminary statistical analysis indicates that these 39 features can be used in classical ML models to distinguish between indoor and outdoor environments reliably. We tested various ML models in the standard way: the data collected is divided into a training set containing 75% and test set containing 25% of the data. We use the ML models implemented in Scikit Learn [18], [19] with default parameters value.



(a)

(b)

FIGURE 6. **Univariate analysis on GPS accuracy.**

A. ML ALGORITHMS

The indoor/outdoor classification problem can be addressed by a number of well-know ML classifiers. We evaluated the following:

- *Naive Bayes (NB/NBayes):* A classification technique based on Bayes' theorem [20] with an assumption of independence between features.
- *Linear Discriminant Analysis (LDA):* LDA is a dimensionality reduction technique and is mainly used as a pre-processing step in ML. LDA is also used for binary classification problems
- *AdaBoost:* AdaBoost is one of the first boosting algorithms to be adapted. It helps to combine multiple "weak classifiers" into a single "strong classifier". It works by putting more weight on difficult to classify instances and less on those already classified.
- *Decision Tree (DecTree):* The data is split into two or more homogeneous sets, based on the most significant features to make as distinct groups as possible.
- *Extra Trees (ExTree):* ExTree is an ensemble ML algorithm that uses the predictions (majority voting)

TABLE 4. F1-score of different ML algorithms on test data for indoor/outdoor classification.

Feature Set	Naive Bayes	LDA	AdaBoost	Decision Tree	XGBoost	ExTree	Random Forest
5G NR	9.3%	85.3%	85.2%	85.6%	85.6%	85.6%	85.6%
5G NR + GPS Accuracy	11.5%	85.1%	87.3%	87.0%	89.6%	87.3%	87.9%
LTE	80.3%	90.2%	91.7%	94.9%	95.0%	96.3%	96.3%
LTE + GPS Accuracy	80.6%	90.3%	92.9%	95.1%	96.3%	96.7%	97.1%
Wi-Fi	94.7%	93.0%	96.1%	98.3%	98.6%	98.8%	98.8%
Wi-Fi + LTE	89.5%	94.5%	96.5%	98.7%	99.1%	99.3%	99.4%
Wi-Fi + GPS Accuracy	94.7%	93.1%	96.1%	98.2%	98.8%	98.7%	98.8%
NR + LTE + Wi-Fi + GPS Accuracy	37.9%	94.5%	96.6%	98.7%	99.3%	99.4%	99.4%

TABLE 5. Random forest 10-Fold cross validation.

Metrics	Test Accuracy	Train Accuracy	Test TPR	Train TPR	Test TNR	Train TNR	Test F1-score	Train F1-score
Mean	0.9925	0.9999	0.9971	1.0	0.9928	0.9999	0.9949	0.9999
Standard Deviation	0.0004	0	0.0002	0.0	0.0007	0	0.0003	0

from many decision trees trained on the training dataset for classification.

- *XGBoost:* This algorithm is an extension of gradient boosted decision trees [21] and is designed to improve speed and performance.
- *Random Forest (RF/RForest):* RF is an ensemble of decision trees. In order to classify a new observation based on features, RF applies majority voting on the classification given by each decision tree. The difference between RF and ExTree is in the way they select the cut points to split the nodes in the decision trees. RF chooses the optimal split, whereas ExTree chooses it randomly.

B. PERFORMANCE METRICS [22]

The following standard metrics were used to evaluate the performance of the above algorithms:

- *Accuracy:* Percentage of correctly predicted records (Indoor and Outdoor Combined).
- *True Positive Rate (TPR)/Recall:* Percentage of correctly predicted outdoor records.
- *True Negative Rate (TNR):* Percentage of correctly predicted indoor records.
- *Precision:* Percentage of correctly identified records among the ones which are classified as outdoor.
- *F1-Score:* Harmonic mean of precision and recall.
- *Area under the curve (AUC):* The AUC is the measure of the ability of a classifier to distinguish between classes.

C. PERFORMANCE OF ML MODELS

We evaluated the classification performance of various ML algorithms, using different combinations of features on the collected data. Our results, using several standard ML algorithms implemented using Scikit-learn, are shown in Table 4, where the F1-score of different algorithms tested with different combinations of feature sets is shown. It is clear that as more frequency bands are added to the feature set, the F1-score increases, especially for the tree-based models such as AdaBoost, Decision Tree, XGBoost, ExTree, and Random Forest. We also observed that only Wi-Fi and LTE features are sufficient to get good classification accuracy for indoor as well as outdoor, but adding the NR and GPS accuracy features lead to slight improvement in AdaBoost, XGBoost, and ExTree (as shown in Table 4).

Fig. 7 shows the GPS accuracy performance of different ML models on the test data. The XGBoost and Random Forest algorithms guarantee 82.6% and 80.4% accuracy. From the observation, it is clear that we cannot reliably detect indoor and outdoor environments with only GPS accuracy feature. Fig. 8 shows the performance of different ML models on the test data when all 39 features are used. The NB algorithm performs poorly *i.e.*, 43%, when all features are used. Since NB assumes that all the features are uncorrelated, the prediction probability reduces significantly even if one of the features has a wrong likelihood probability. XGBoost, ExTree, and Random Forest models consistently outperformed other models for all the feature combinations (see Table 4). Especially when all features are used, XGBoost, ExTree, and Random Forest models have above 99% F1-score. This high F1-score also indicates that the ensemble ML models are making accurate predictions for indoor as well as outdoor records (despite the data imbalance between the two classes). We also performed 10-fold cross-validation for the Random Forest model, and the corresponding results are shown in Table 5. It shows that the Random Forest model consistently performs above 99% accuracy and F1-score on all the folds. Since the accuracy and the F1-score from the models are already high, hyper parameter tuning of these models is not needed. Hence, a separate validation set is not used.

1) ML EXPLAINABILITY STUDY

In this section, we study the explainability of the best performing ML model, *i.e.*, Random Forest, using the SHAP (SHapley Additive exPlanations [23], [24]) package in Python. The SHAP package helps visualize the importance of the input features in classifying a given record as indoor or outdoor. Fig. 9 shows the SHAP values for the features

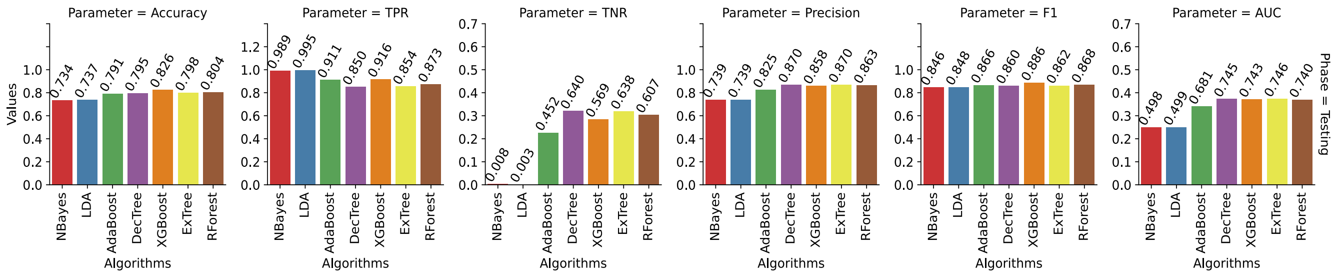


FIGURE 7. Performance of ML models using only the GPS accuracy feature in terms of Accuracy, TPR, TNR, Precision, F1-Score and AUC.

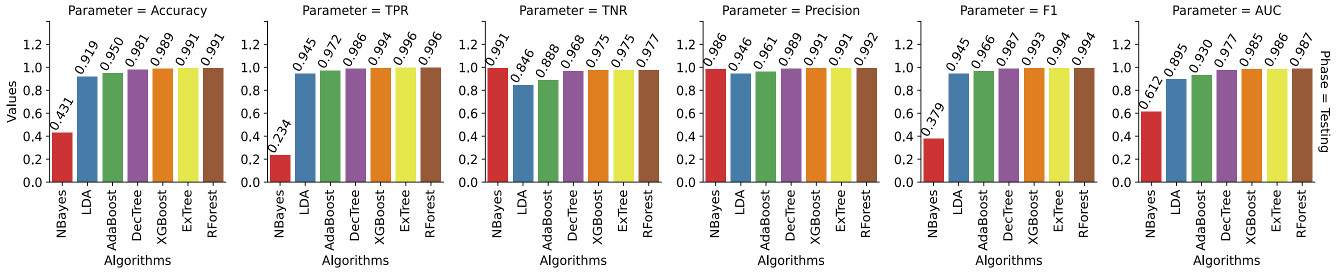
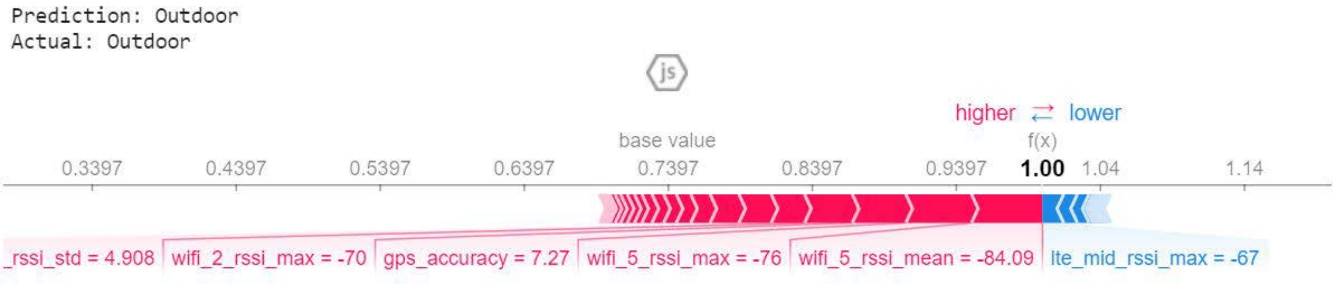
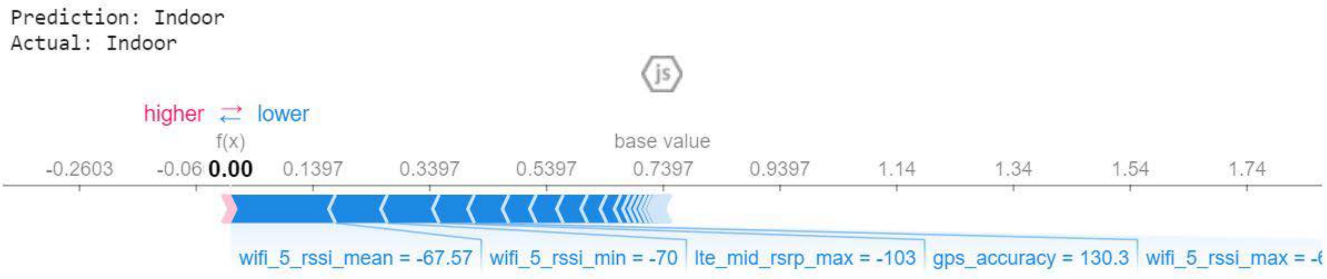


FIGURE 8. Performance of ML models using all the 39 features in terms of Accuracy, TPR, TNR, Precision, F1-Score and AUC.



```
Prediction: Outdoor
Actual: Outdoor
```

FIGURE 9. Random forest explainability: Correct indoor prediction.



```
Prediction: Indoor
Actual: Indoor
```

FIGURE 10. Random forest explainability: Correct outdoor prediction.

of a record which was correctly classified as outdoor by the RF model. The red and blue arrows indicate the features that push the model towards predicting the record as outdoor and indoor, respectively. The length of an arrow indicates the importance of the feature in deciding the prediction. For example, in Fig. 9, the feature Wi-Fi 5GHz RSSI mean (−84 dBm) is the most important feature to make the model classify the record as outdoor. This behavior is consistent with what we observed from univariate analysis (see Fig. 4b and Fig. 5b): a lower value for Wi-Fi 5GHz RSSI mean indicates that the record is most likely from outdoor. Similarly, Fig. 10 shows the SHAP values for the features of a record which was correctly classified as indoor by the RF model.

Fig. 11 shows the summary of the SHAP values for 100 test records. The points which are to the left and right

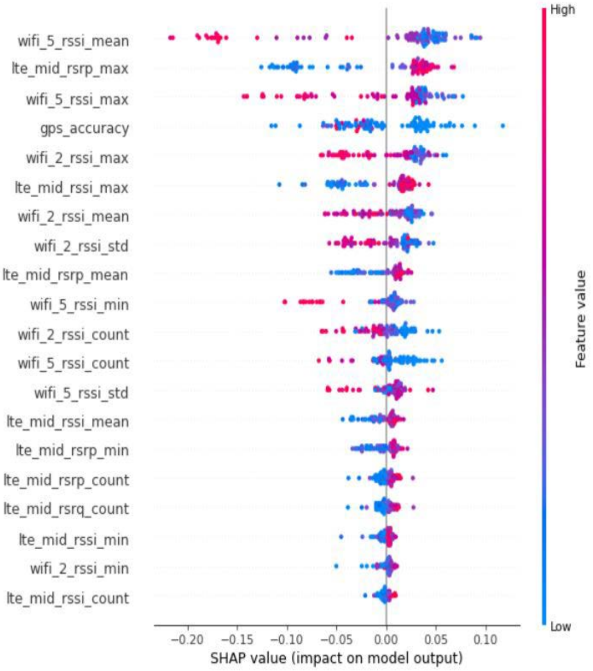


FIGURE 11. Random forest: Feature importance.

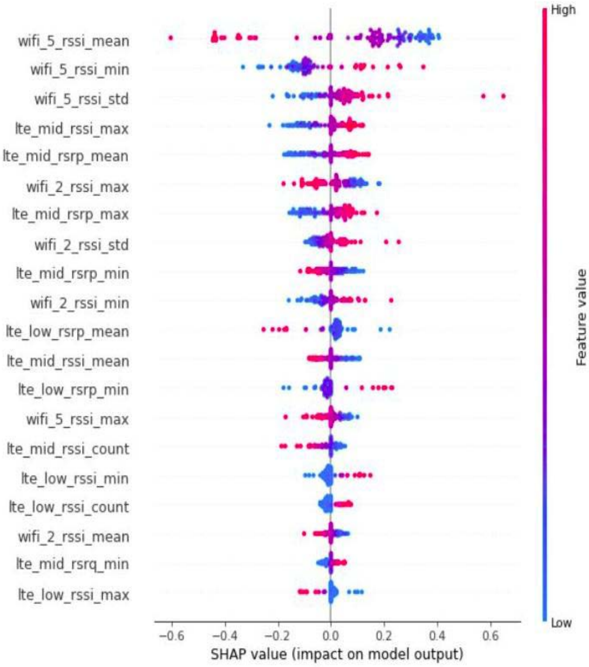


FIGURE 12. LDA: Feature importance.

TABLE 6. Performance of random forest in classifying TacoBell and WholeFood.

Location	ML model	Accuracy	F1-Score	TPR	TNR
TacoBell	Original	53%	67%	100%	11%
TacoBell	With 20%	75%	79%	100%	53%
WholeFood	Original	54%	61%	100%	24%
WholeFood	With 20%	69%	71%	100%	51%

of the zero line indicates the importance of the features in making indoor and outdoor prediction, respectively. The color of a point indicates the feature value (red indicates high feature value and blue indicates low feature value). It is clear from Fig. 11 that high value for Wi-Fi RSSI feature pushes the model towards making indoor prediction and low value pushes the model towards making outdoor prediction. In contrast, low value for LTE feature pushes the model towards making indoor prediction and high value pushes towards outdoor prediction. Similarly, Fig. 12 shows the important feature for prediction in LDA.

D. DEEP NEURAL NETWORK (DNN) MODEL

We implemented a DNN model using the Scikit Learn package to analyze the classification performance of Neural Networks. The DNN consists of two hidden layers containing 64 neurons in each layer. We trained the DNN model with the learning rate of 0.001, and the optimizer used is Adam with the number of epochs as 100 and batch size as 200. The test accuracy of the DNN model is 98.7%, and test F1-Score is 99.1%. DNN is performing on-par with other ML models. Since this is a simple tabular classification problem, ensemble ML models are sufficient.

VI. TEST SCENARIOS

The previous section demonstrated that Random Forest performed extremely well in the indoor/outdoor classification task when tested in the conventional way against a training and test set where the test data set consisted of measurements from the same environments that were included in the training data set. In this section, we are interested in testing the ML model on data collected in environments that were not included in developing the ML models. To do so, we studied two different test cases described below. The app was used to collect data both indoors and outdoors in these environments.

- **Location 1**, TacoBell is a restaurant, with very large glass windows facing the street as shown in Figs. 13a and 13b.
- **Location 2**, WholeFood is a grocery store, again with very large glass windows facing the street as shown in Figs. 13c and 13d.

Table 6 summarizes the classification performance in these two environments, first without including any of the data in building the trained model and then adding 20% of the new data to retrain the ML model. The overall F1-scores using the ML model trained on the original data were 67% and 61% in TacoBell and WholeFood respectively. These improved to 79% and 71% when 20% of the data was used to retrain the model. Interestingly, we observe that the TPR (probability of classifying outdoors correctly) was 100% in all cases, while the TNR (probability of classifying indoor correctly) was quite low. We believe that this is due to two reasons: (i) in both locations, the large street-side windows caused the indoor environment to appear more like an outdoor one in terms of the RF signal levels perceived

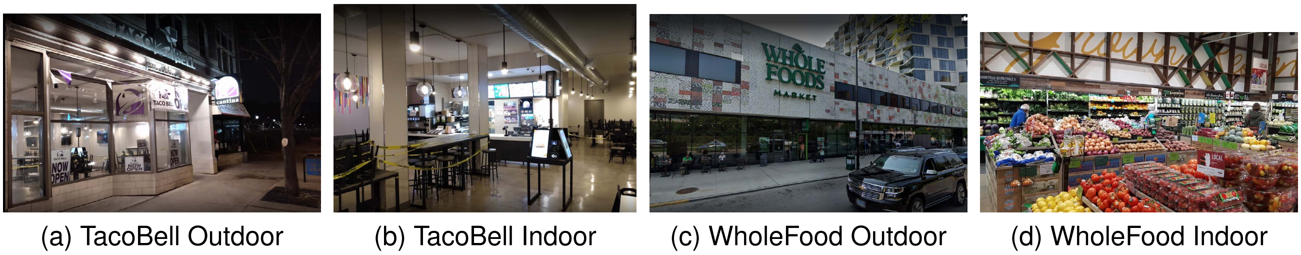

(a) TacoBell Outdoor (b) TacoBell Indoor (c) WholeFood Outdoor (d) WholeFood Indoor

FIGURE 13. Test Cases: TacoBell and WholeFood.

at the phone; and (ii) as we mentioned in Section III, our data-set has more and diverse outdoor records than indoor records. From an interference potential perspective, the ML algorithm performance is actually desirable: an indoor device near a window or open door has the same interference potential as an outdoor device as was noted in a recent filing to the FCC where devices near windows or open doors had very high interference levels at incumbent receivers [25].

There are two ways to address the ML performance in such scenarios: (i) increase the representation of diverse indoor environments in the data set. We see that including just 20% of the data from the new environment in the training set improved classification accuracy, even though this was a very small percentage (.04%) of the overall training set; or (ii) create three classes: indoors, indoors near windows and outdoors. The second approach would be the best to address the application we are interested in where we use device environment to determine transmit power levels: a device near a window could be subject to a transmit power requirement in between a fully indoor and fully outdoor device. Both options require creating a more diverse data set, especially of indoor measurements since indoor environments tend to be more diverse than outdoor ones. Just as image recognition performance improved dramatically as image databases grew larger and incorporated diverse images, we are confident that as these types of RF data-sets grow, RF based indoor/outdoor classification will improve as well.

VII. CONCLUSION AND FUTURE WORK

We believe that this is the first comprehensive evaluation of ML-based indoor/outdoor classification using a very large, labeled data-set of RF signals spanning a wide range of frequencies. We did a thorough evaluation of a number of models and demonstrated excellent performance with Random Forest. We believe that such methods can be used in future rule-making to enable devices to self identify as being indoors or outdoors, thus enabling improved spectrum rules. We propose to conduct future research as follows:

- We have evaluated only ensemble classification algorithms in this paper. CNNs have proven to be extremely proficient in image classification tasks and can be applied to the indoor/outdoor classification problem by either converting tabular data to an image [26] or creating heat maps of signal strength and treating them as images. In future work, we plan to compare CNNs with RNNs (specifically LSTM).
- Data collection: the diversity of devices and environments needs to be broadened considerably, since the current labels are not enough to model the varying signal conditions in dynamic areas such as stadiums, tall buildings, apartments shopping complex and narrow streets. In the future, we look to create more specialization of indoors classes since it tends to be more diverse than outdoors.
- Android APIs have information on the phone model that our app extracts but we have not used this information in our study. Classification accuracy could be further improved by creating different models for different device types and operators.
- Signal diversity: besides Wi-Fi and LTE signals, we are planning to add other types of signal measurements that are available on current phones, such as Bluetooth [27], [28], and UWB [29]. These were not considered in this paper due to limited deployments, but in future these may become more sidespread. By increasing the type of captured signals, we will have a broader range of frequencies which may lead to a higher model accuracy.

REFERENCES

[1] (2020). *FCC, Report and Order and Further Notice of Proposed Rule Making: In the Matter of Unlicensed Use of the 6 GHz Band*. [Online]. Available: https://docs.fcc.gov/ public/attachments/FCC-20-51A1.pdf

[2] (2021). *3rd Generation Partnership Project*. 3GPP Release-16. [Online]. Available: http://www.3gpp.org/release-16/

[3] (2016). *3rd Generation Partnership Project*. 3GPP Release-14. [Online]. Available: http://www.3gpp.org/release-14/

[4] (2021). *FCC, Notice of Proposed Rule Making: In the Matter of Expanding Flexible Use of the 12.2 12.7 GHz Band*. [Online]. Available: https://docs.fcc.gov/public/attachments/FCC-21-13A1.pdf

[5] T. Jian, B. C. Rendon, E. Ojuba, N. Soltani, Z. Wang, K. Sankhe, A. Gritsenko, J. Dy, K. Chowdhury, and S. Ioannidis, "Deep learning for RF fingerprinting: A massive experimental study," *IEEE Internet Things Mag.*, vol. 3, no. 1, pp. 50–57, Mar. 2020.

[6] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 146–152, Sep. 2018.

[7] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proc. IEEE Int. Workshop Content-Based Access Image Video Database*, Jan. 1998, pp. 42–51.

[8] R. Raja, S. M. M. Roomi, D. Dharmalakshmi, and S. Rohini, "Classification of indoor/outdoor scene," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2013, pp. 1–4.

[9] A. Payne and S. Singh, "A benchmark for indoor/outdoor scene classification," in *Proc. Int. Conf. Pattern Recognit. Image Anal.*, Cham, Switzerland: Springer, 2005, pp. 711–718.
[10] V. Bui, N. T. Le, T. L. Vu, V. H. Nguyen, and Y. M. Jang, "GPS-based indoor/outdoor detection scheme using machine learning techniques," *Appl. Sci.*, vol. 10, no. 2, p. 500, Jan. 2020.
[11] B. Lee, C. Lim, and K. Lee, "Classification of indoor-outdoor location using combined global positioning system (GPS) and temperature data for personal exposure assessment," *Environ. Health Preventive Med.*, vol. 22, no. 1, pp. 1–5, Dec. 2017.
[12] J. Wu, C. Jiang, D. Houston, D. Baker, and R. Delfino, "Automated time activity classification based on global positioning system (GPS) tracking data," *Environ. Health*, vol. 10, no. 1, pp. 1–13, Dec. 2011.
[13] Y. Kim, S. Lee, S. Lee, and H. Cha, "A GPS sensing strategy for accurate and energy-efficient outdoor-to-indoor handover in seamless localization systems," *Mobile Inf. Syst.*, vol. 8, no. 4, pp. 315–332, 2014.
[14] V. Sathya, A. Ramamurthy, and B. R. Tamma, "On placement and dynamic power control of femtocells in LTE HetNets," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4394–4399.
[15] M. Erel-Ozcevik and B. Canberk, "Road to 5G reduced-latency: A software defined handover model for eMBB services," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8133–8144, Aug. 2019.
[16] L. Zhang, Q. Ni, M. Zhai, J. Moreno, and C. Briso, "An ensemble learning scheme for indoor-outdoor classification based on KPIs of LTE network," *IEEE Access*, vol. 7, pp. 63057–63065, 2019.
[17] (2019). *Sigcap App*. [Online]. Available: https://people.cs.uchicago.edu/~muhiqbalcr/sigcap/
[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.
[19] R. Garreta and G. Moncecchi, *Learning Scikit-Learn: Machine Learning in Python*. Birmingham, U.K.: Packt, 2013.
[20] D. V. Lindley, "Fiducial distributions and Bayes' theorem," *J. Roy. Stat. Soc. Ser. B, Methodol.*, vol. 20, no. 1, pp. 102–107, Jan. 1958.
[21] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 2061–2064.
[22] D. V. Carvalho, M. E. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.
[23] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.
[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[25] (Jun. 21, 2021). *T. R. on the Effects of 6 GHz Unlicensed RLAN Units on Fortson to Columbus Microwave Link*. [Online]. Available: https://ecfsapi.fcc.gov/file/106231367519302/6%20GHz%20Columbus%20%Test%20Report%20-%20June%202021.pdf
[26] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, and R. L. Stevens, "Converting tabular data into images for deep learning with convolutional neural networks," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Dec. 2021.
[27] G. Gonzalez, M. E. Larraga, L. Alvarez-Icaza, and J. Gomez, "Bluetooth worm propagation in smartphones: Modeling and analyzing spatio-temporal dynamics," *IEEE Access*, vol. 9, pp. 75265–75282, 2021.
[28] S. Lee, B. Koo, M. Jin, C. Park, M. J. Lee, and S. Kim, "Range-free indoor positioning system using smartphone with Bluetooth capability," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, May 2014, pp. 657–662.
[29] T. Brovko, A. Chugunov, A. Malyshev, I. Korogodin, N. Petukhov, and O. Glukhov, "Complex Kalman filter algorithm for smartphone-based indoor UWB/INS navigation systems," in *Proc. Ural Symp. Biomed. Eng., Radioelectron. Inf. Technol. (USBEREIT)*, May 2021, pp. 280–284.

ARUN RAMAMURTHY (Member, IEEE) received the B.Tech. degree from the Indian Institute of Technology Hyderabad and the Master of Technology degree in industrial engineering and operational research from the Indian Institute of Technology Bombay. He is currently a Researcher with TCS Research, India, where he works on decision-making problems in cloud computing, data privacy, and task allocation. His research interests include optimization, integer programming, and reinforcement learning.

VANLIN SATHYA (Member, IEEE) received the Bachelor of Engineering degree in computer science and the Master of Engineering degree in mobile and pervasive computing from Anna University, Chennai, India, in 2009 and 2011, respectively, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology (IIT) Hyderabad, India, in 2016. He currently works with CTO Office at Cleona Inc., USA. Prior to this, he was a Postdoctoral Scholar at the University of Chicago, USA, where he worked on the issues faced in 5G real time coexistence test-bed when LTE-unlicensed and Wi-Fi try to coexist on the same channel. He continued his career at IIT Hyderabad, where he was a Project Officer for the converged radio access network radio access network (RAN) project. His research interests include interference management, handover in heterogeneous LTE networks, device to device communication (D2D) in cellular networks, cloud base station and phantom cell (LTE-B), LTE in unlicensed, and private 5G (CBRS).

MUHAMMAD IQBAL ROCHMAN (Member, IEEE) received the B.E. degree in informatics engineering from the Sepuluh Nopember Institute of Technology, Indonesia in 2012, and the M.S. degree in computer science from the National Taiwan University of Science and Technology, Taiwan, in 2016. He is currently pursuing the Ph.D. degree in computer science with the University of Chicago, USA, where he is studying the coexistence fairness between LTE and Wi-Fi in terms of throughput and spectrum access, both in real-time test-bed and simulation framework.

MONISHA GHOSH (Fellow, IEEE) received the B.Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1986, and the Ph.D. degree in electrical engineering from the University of Southern California, in 1991. She is currently a Professor with the Electrical Engineering Department, University of Notre Dame, and the Policy Outreach Director of SpectrumX, the first NSF Spectrum Innovation Center. Prior to this, she was the Chief Technology Officer with the FCC, a Program Director at NSF, and a Research Professor at the University of Chicago. She also has extensive industry experience at Interdigital, Philips Research, and Bell Labs.