This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LAWP.2018.2869548, IEEE Antennas and Wireless Propagation Letters

Classification of Indoor Environments for IoT Applications: A Machine Learning Approach

Mohamed I. AlHajri, *Student Member, IEEE*, Nazar T. Ali, *Senior Member, IEEE*, and Raed M. Shubair, *Senior Member, IEEE*

***Abstract*—Evolving Internet-of-Things (IoT) applications often require the use of sensor-based indoor tracking and positioning, for which the performance is significantly improved by classifying the type of the surrounding indoor environment. This classification is of high importance since it leads to efficient power consumption, when operating the deployed IoT sensors. This paper presents a machine learning approach for indoor environment classification based on real-time measurements of the RF signal in a realistic environment. Several machine learning classification methods are explored including decision trees, support vector machine, and k-Nearest Neighbor using different RF features. Results obtained show that a machine learning approach based on weighted k-Nearest Neighbor method, that utilizes a combination of channel transfer function and frequency coherence function, outperforms the other methods in classifying the type of indoor environment with an accuracy of 99.3%. The predication time was found to be below 10μs, which verifies that the adopted algorithm are successful candidates for real-time deployment scenarios.**

***Index Terms*—Received Signal Strength, Channel Transfer Function, Frequency Coherence Function, k-Nearest Neighbor, Support Vector Machine, Decision Trees, Machine Learning, IoT**

I. Introduction

The advent of Internet-of-Things (IoT) have revolutionized the field of telecommunications as it opened the door for unprecedented applications such as [1]–[7]. These applications are enabled through smart sensors that cooperate efficiently to provide the desired service. The efficient deployment of IoT-based sensors primarily depends on allowing the adjustment of sensor power consumption according to the radio frequency (RF) propagation channel which is dictated by the type of the surrounding indoor environment.

Existing literature use classification approaches which categorizes environment into indoor or outdoor [8]–[18], rather than classify various types of indoor environments. For example, the approach presented in [8] categorizes the surrounding environment into indoor, semi-outdoor, and outdoor. That approach was found not popular since it requires data collected from a large number of specially erected cell towers in a small proximity [8]. Other approaches have classified indoor/outdoor environments using cell identity maps [9] and chirp sound signals [10]. A semi-supervised machine learning method was introduced in [11] for indoor/outdoor classification.

The key motivation of this work is to use machine learning algorithms applied on real-time data measurements for an accurate classification of various *indoor* environments. This is achieved by exploring the use of three RF signatures, namely Received Signal Strength (RSS), frequency-domain Channel Transfer Function (CTF), and the autocorrelation of the CTF which is denoted by the Frequency Coherence Function (FCF) [19]–[21], along with different machine learning algorithms such as Decision Tree (DT) [22], Support Vector Machine (SVM) [23], and k-Nearest Neighbor (k-NN) [24].

II. RF Signatures

In a multipath rich indoor environment, the received signal $Y(f)$ contains replicas of the RF transmitted signal $X(f)$. The ratio of $Y(f)$ to $X(f)$ defines the Channel Transfer Function (CTF), $H(f)$, which contains the multipath effects of the wireless channel. Hence, $H(f)$ can be represented as the superposition of the gains associated with the multipath components as follows [25]:

$$H(f) = \sum_{l=1}^{L} a_l \exp\left[-j(2\pi f \tau_1 - \theta_l)\right] \tag{1}$$

where a_l, τ_l, and θ_l are the amplitude, delay, and phase of the l^{th} multipath component; L is the total number of multipath components; and f is the operating frequency at which the channel is classified. The CTF, $H(f)$, is considered as an RF signature that would be distinctively unique for every position in space. Under frequency selective fading, this RF signature becomes more sensitive to channel variations due to the rapid fluctuations of the gains of its multipath components.

The complex autocorrelation of CTF, $H(f)$, defines another channel metric known as Frequency Coherence Function (FCF), $R(f)$, given by [26]:

$$R(f) = \int_{-\infty}^{\infty} H(\hat{f}) H^*(\hat{f} + f)\, d\hat{f} \tag{2}$$

FCF given in (2) represents the frequency domain coherence of the radio channel. It is interpreted as the autocorrelation of the CTF, $H(f)$, due to different frequency shifts. FCF is known for its slow changing nature in the spatial domain [26], which makes it a strong candidate as an RF signature.

M. I. AlHajri with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: malhajri@mit.edu)

N .T. Ali is with the Department of Electrical and Computer Engineering, Khalifa University, UAE (email:nazar.ali@ku.ac.ae)

R. M. Shubair is with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: rshubair@mit.edu)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LAWP.2018.2869548, IEEE Antennas and Wireless Propagation Letters

III. Data Collection via Real-Time Measurement

The frequency domain CTF, $H(f)$, is obtained for four different environments in Khalifa University campus, Sharjah, UAE. The floor plan of Fig. 1 shows the three environments having boundaries namely: highly cluttered (Laboratory), medium cluttered (Narrow Corridor), and low cluttered (Lobby). In Fig. 1, the red triangle represents the transmitter and the black asterisks represent the receiver positions at which measurements are taken. Beside the aforementioned environment of Fig. 1, a fourth environment, open space (Sports Hall), has also been considered.

A square area within each environment was divided into uniform grids with a spacing of one wavelength (λ = 12.5cm) at an operating frequency f = 2.4GHz selected to examine the WiFi bands associated with IEEE 802.11g standard. The physical arrangement resulted in a total of 196 positions so that small-scale variations can be better captured [27]. It should be pointed out that the real-time measurements were carried out under a stationary scenario with no movements around the transmitter and receiver.

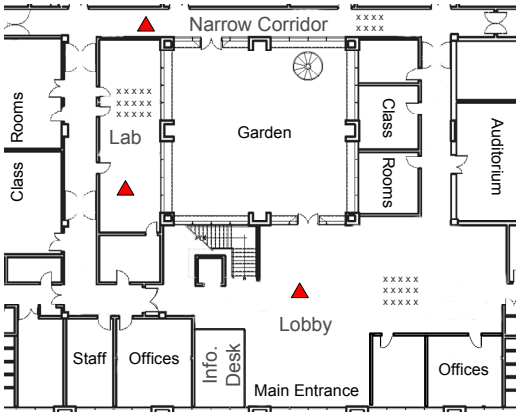


Fig. 1: Floor plan

The measurement system consisted of low loss RF cables, a ZVB14 Vector Network Analyzer (VNA) used to measure the transmission coefficient S_{21}, and omnidirectional antennas of equal heights 1.5m at the transmitter and receiver ends. A program script was written for the VNA to run 10 consecutive sweeps; each sweep covering a band of 100MHz using 601 frequency points for which the frequency separation is 0.167MHz. This means that the dataset generated for each environment consists of 196 $\times$ 10 samples.

IV. Spatial Correlation Coefficient

The similarities between RF signatures can be assessed by computing the Spatial Correlation Coefficient, $C^{(n)}$, given by [28]:

$$C^{(n)} = \left| \frac{\langle S_r, S^{(n)} \rangle}{\|S_r\| \|S^{(n)}\|} \right| \tag{3}$$

where S_r is the reference RF signature corresponding to CTF or FCF at a specific reference position, while $S^{(n)}$ is the RF signature at position n.

Small values of the Spatial Correlation Coefficient, $C^{(n)}$, imply that there is a substantial difference between the RF signature at position n and the RF reference signature. In contrast, large values of the $C^{(n)}$ indicate that there is a high similarity between the RF signature at position n and the reference signature. The variation of Spatial Correlation Coefficient, $C^{(n)}$, computed using CTF and FCF is shown in Figs. 2 and 3, respectively.

In the first scenario, the Spatial Correlation Coefficient in a highly cluttered environment using CTF and FCF is depicted in Figs. 2(a) and 3(a), respectively. The high number of scatters causes the CTF to vary significantly. However, the correlation coefficient for FCF varies at a smaller rate when compared to the CTF.

In the second scenario, the Spatial Correlation Coefficient in a medium cluttered environment using CTF and FCF is depicted in Figs. 2(b) and 3(b), respectively. The medium number of scatters also causes the CTF to vary significantly but at a lower rate than the highly cluttered scenario. Again, the correlation coefficient for FCF varies at a smaller rate when compared to the CTF.

In the third scenario, the Spatial Correlation Coefficient in a low cluttered environment using CTF and FCF is depicted in Figs. 2(c) and 3(c), respectively. As expected, the low number of scatters would affect the CTF at a very low rate. It is interesting to observe, and unlike the former two scenarios, that the correlation coefficient for FCF is almost constant and does not change with position.

In the forth scenario, the Spatial Correlation Coefficient in an open space environment using CTF and FCF is depicted in Figs. 2(d) and 3(d), respectively. In this case, there is a very low number of scatters and therefore the CTF does not vary significantly. Similar to the third scenario, the correlation coefficient for FCF in this scenario is almost constant and does not change with position.

The previous results demonstrate that both CTF and FCF vary according to the environment which confirm that they are suitable candidates for environment classification.

V. Indoor Environment Classification

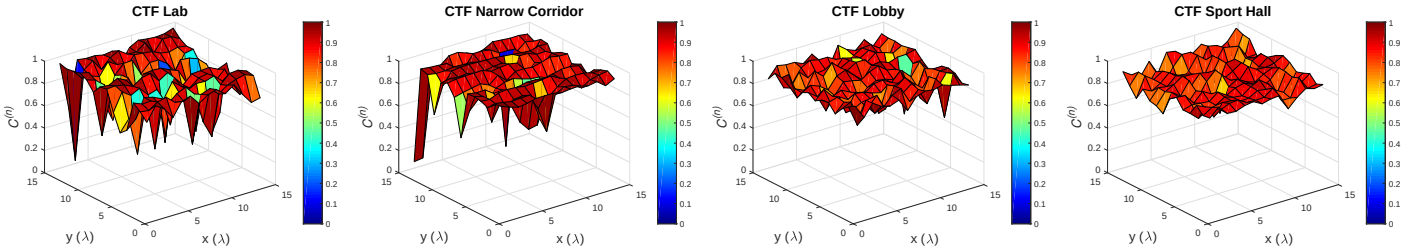
We divided our dataset such that 75% of the data points were used to train the machine learning algorithm and 25% were used to test the performance of the algorithm.

A. Machine Learning Algorithms

1) ***Decision Trees****:* Decision Trees (DTs) is a widely used method for classification that is based a binary decision tree that is constructed from the training data. The decision tree is constructed from three nodes; root node, internal node, and leaf node. As described in [22], [29], the root node has no incoming edges and just has outgoing edges. The internal node has exactly one incoming edge and two or more outgoing edges. The leaf node has one incoming edge and no outgoing edge, and resembles a class label. The algorithm stops splitting into sub-trees when the uncertainty is minimum. The node data is split based on Gini diversity index [30].

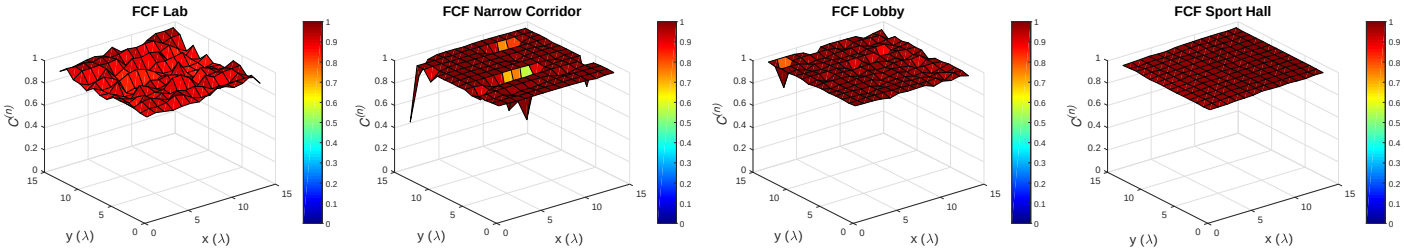
$$\text{Gini}(T) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} p_i^2 \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LAWP.2018.2869548, IEEE Antennas and Wireless Propagation Letters



(a) Highly cluttered (b) Medium cluttered (c) Low cluttered (d) Open space

Fig. 2: Spatial Correlation Coefficient, $C^{(n)}$, based on CTF signatures for different environments



(a) Highly cluttered (b) Medium cluttered (c) Low cluttered (d) Open space

Fig. 3: Spatial Correlation Coefficient, $C^{(n)}$, based on FCF signatures for different environments

where T is the training dataset, p_i is the probability of class i occurring at a certain node in the tree, and J is the total number of classes.

Three different types of decision trees are used; simple, medium, and complex based on the maximum number of splits 4, 20, and 100 respectively.

*2) **Support Vector Machine**:* Support Vector Machine (SVM) was initially defined by Vapnik [23], [29] and is widely used for classification. The classifier is based solely on the observation of samples of the form $\{(x_i, y_i), i = 1, \cdots, n\}$, where x_i is the i^{th} input and y_i is the corresponding label. The SVM finds the optimal margin classifier.

The solution to find the optimal classifier can be formulated as an optimization problem [23]:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y^{(i)}(w^{\mathrm{T}}x^{(i)} + b) \geq 1, \ i = 1, \ldots, m. \end{aligned} \tag{5}$$

where w is the gradient of the linear classifier, and b is the y-intercept, and γ is:

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)} = \min_{i=1,\ldots,m} y^{(i)} \left(\left(\frac{w}{\|w\|} \right)^{\mathrm{T}} x^{(i)} + \frac{b}{\|w\|} \right) \tag{6}$$

This optimization problem is a convex quadratic objective and has only linear constrains that can be solved using a quadratic program.

The dual problem is expressed as following:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to} \quad & \alpha_i \geq 0, \ i = 1, \ldots, m. \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \end{aligned} \tag{7}$$

where α_i is the Lagrange multiplier and $K(.,.)$ denotes the kernel function that measures the distance between the input vector x_i and the trained data x_j. Examples of kernel functions include linear kernel, Gaussian kernel and polynomial kernel.

*3) **k-Nearest Neighbor**:* k-Nearest Neighbor (k-NN) is a non-parametric method used for classification [24], [29]. In the case of 1-NN the test sample x_i will take the class of its closest neighbor. In the case of k-NN it will be assigned to the class that is the majority of its neighbors. In the case of the weighted k-NN, after selecting the k nearest neighbor, each neighbor will be scaled by the weight which is the inverse of the Euclidean distance squared.

B. Results and Discussion

Table I shows that the classification accuracy using RSS can only reach up to 42.7%. When CTF is used, the classification accuracy improves and reaches up to 79.9% in the case of weighted k-NN ($k = 10$). The classification accuracy continues to improve when FCF is used and reaches up to 83.4%. Combing RSS with CTF improves the accuracy of classification compared to the individual performance, yet it does not exceed the accuracy attained using FCF. Unlike the combination with CTF, when RSS is combined with FCF, the classification accuracy exceeds that for FCF and can reach up to 93.4%. The combination of CTF and FCF was found to yield the highest classification accuracy of 99.3%, which is identically obtained as well when all three features are combined. Investigations indicated that adding RSS to the combination of CTF and FCF features does not have an noticeable improvement on the accuracy of classification. Hence, it can be concluded that a combination of CTF and FCF features is sufficient for an accurate classification on indoor environments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LAWP.2018.2869548, IEEE Antennas and Wireless Propagation Letters

TABLE I: Overall accuracy using different classifiers and RF features

Features	Classifier							
	Simple DT	Medium DT	Complex DT	Gaussian SVM	k-NN ($k=1$)	k-NN ($k=10$)	Weighted k-NN ($k=1$)	Weighted k-NN ($k=10$)
RSS	41.3%	42.3%	42.5%	42.7%	32.6%	40.1%	32.6%	33.5%
CTF	42.4%	49.0%	57.1%	60.7%	78.0%	77.2%	78.0%	79.9%
FCF	51.5%	56.9%	62.2%	59.0%	83.4%	76.6%	83.4%	83.1%
RSS + CTF	45.5%	51.2%	57.4%	62.7%	78.2%	76.1%	78.2%	80.0%
RSS + FCF	52.8%	60.3%	69.1%	72.0%	93.4%	84.8%	93.4%	92.8%
CTF + FCF	54.0%	62.4%	73.7%	90.3%	**99.3%**	94.0%	**99.3%**	99.0%
RSS + CTF + FCF	55.3%	62.9%	72.5%	91.7%	**99.3%**	93.7%	**99.3%**	98.8%

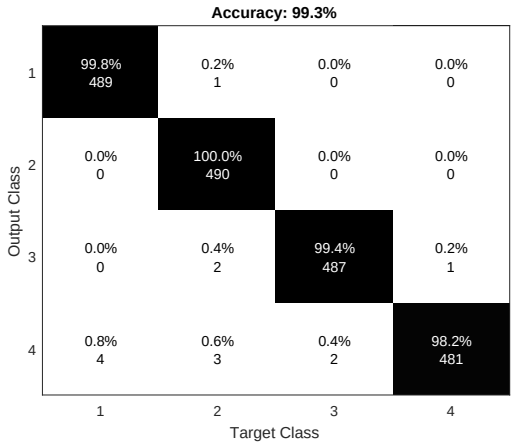


Fig. 4: Confusion matrix of 1-NN using CTF + FCF

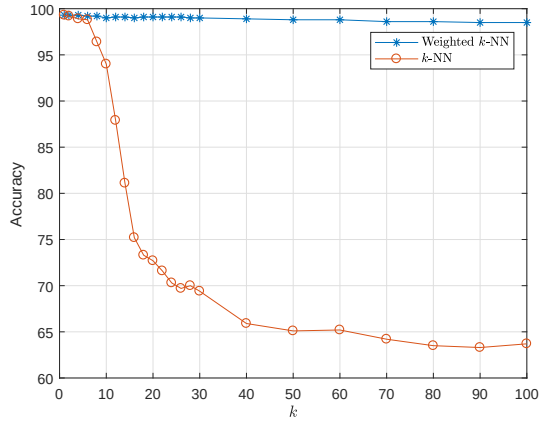


Fig. 5: k-NN and weighted k-NN as a function of k

Successful real-time deployment of machine learning algorithms requires minimum computational time. It was found that the predication time for both 1-NN and weighted 1-NN algorithms is below 10μs. This indicates that the adopted algorithms are very applicable in practical real-time scenarios.

The confusion matrix shows the four different environments denoted by 1 (no clutter), 2 (low clutter), 3 (medium clutter) , and 4 (high clutter). It can be seen from Fig. 4 that the overall classification accuracy is 99.3% and the individual classification accuracy is above 98% for all different environments.

The performance of k-NN depends on the design parameter k. Fig. 5 shows that the classification accuracy of k-NN degrades significantly and reaches 63.7% at $k = 100$. On the other hand, when the weighted k-NN is adopted, Fig. 5 shows that the accuracy is significantly high with a slight variation, over the range of $1 \leq k \leq 100$, and does not go below 98.5%.

VI. Conclusion

This paper developed a machine learning approach for indoor environment classification based on real-time measurements of the RF signal. Several machine learning classification methods were studied including decision trees, support vector machine, and k-NN using different RF features. Results obtained show that a machine learning approach using weighted k-NN method, utilizing channel transfer function (CTF) and frequency coherence function (FCF), outperforms the other methods in identifying the type of the indoor environment with a classification accuracy of 99.3%. The predication time was found to be below 10μs, which verifies that the adopted algorithm are successful candidates for real-time deployment scenarios. The results of this paper facilitate an efficient deployment of IoT-applications in dynamic channels. The goal of this paper was to show the methodology and highlight the benefits of using machine learning as a novel approach and a versatile tool for indoor environment classification. Hence, results have been obtained based on measurements carried out in a stationary environment with a strong line-of-sight component. Our future work will build on the findings of this paper and extend the interesting results, obtained based on stationary channel measurements, by exploring other scenarios including non-line-of sight (NLOS) and time-varying channels. The power of deep learning will be explored for extracting the inherent features of such more sophisticated indoor radio channel scenarios.

References

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[3] L. D. Xu, W. He, and S. Li, "Internet of Things in Industries: A Survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[4] V. Palazzi, F. Alimenti, C. Kalialakis, P. Mezzanotte, A. Georgiadis, and L. Roselli, "Highly Integrable Paper-Based Harmonic Transponder for Low-Power and Long-Range IoT Applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 3196–3199, 2017.

[5] K. R. Jha, B. Bukhari, C. Singh, G. Mishra, and S. K. Sharma, "Compact Planar Multistandard MIMO Antenna for IoT Applications," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 7, pp. 3327–3336, Jul. 2018.

[6] M. AlHajri, A. Goian, M. Darweesh, R. AlMemari, R. Shubair, L. Weruaga, and A. AlTunaiji, "Accurate and robust localization techniques for wireless sensor networks," June 2018, arXiv:1806.05765 [eess.SP].

[7] M. AlHajri, A. Goian, M. Darweesh, R. AlMemari, R. Shubair, L. Weruaga, and A. Kulaib, "Hybrid rss-doa technique for enhanced wsn localization in a correlated environment," in *Information and Communication Technology Research (ICTRC), 2015 International Conference on*, 2015, pp. 238–241.

[8] P. Zhou, Y. Zheng, Z. Li, M. Li, and G. Shen, "IODetector: A Generic Service for Indoor Outdoor Detection," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '12. New York, NY, USA: ACM, 2012, pp. 113–126.

[9] Z. Liu, H. Park, Z. Chen, and H. Cho, "An Energy-Efficient and Robust Indoor-Outdoor Detection Method Based on Cell Identity Map," *Procedia Computer Science*, vol. 56, pp. 189–195, Jan. 2015.

[10] R. Sung, S.-h. Jung, and D. Han, "Sound based indoor and outdoor environment detection for seamless positioning handover," *ICT Express*, vol. 1, no. 3, pp. 106–109, Dec. 2015.

[11] V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina, "A Semi-supervised Learning Approach for Robust Indoor-outdoor Detection with Smartphones," in *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '14. New York, NY, USA: ACM, 2014, pp. 280–294.

[12] O. Canovas, P. E. Lopez-de Teruel, and A. Ruiz, "Wifiboost: a terminal-based method for detection of indoor/outdoor places," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 352–353.

[13] J.-g. Park, D. Curtis, S. Teller, and J. Ledlie, "Implications of device diversity for organic localization," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 3182–3190.

[14] A. J. Ruiz-Ruiz, H. Blunck, T. S. Prentow, A. Stisen, and M. B. Kjærgaard, "Analysis methods for extracting knowledge from large-scale wifi monitoring to inform building facility planning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 130–138.

[15] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proc. IEEE INFOCOM 2000*, vol. 2, pp. 775–784.

[16] O. Canovas, P. E. Lopez-de Teruel, and A. Ruiz, "Detecting indoor/outdoor places using wifi signals and adaboost," *IEEE Sensors Journal*, vol. 17, no. 5, pp. 1443–1453, 2017.

[17] M. Ali, T. ElBatt, and M. Youssef, "Senseio: Realistic ubiquitous indoor outdoor detection system using smartphones," *IEEE Sensors Journal*, vol. 18, no. 9, pp. 3684–3693, 2018.

[18] Q. Zeng, J. Wang, Q. Meng, X. Zhang, and S. Zeng, "Seamless pedestrian navigation methodology optimized for indoor/outdoor detection," *IEEE Sensors Journal*, vol. 18, no. 1, pp. 363–374, 2018.

[19] N. A. Khanbashi, N. Alsindi, S. Al-Araji, N. Ali, and J. Aweya, "Performance evaluation of CIR based location fingerprinting," in *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)*, Sep. 2012, pp. 2466–2471.

[20] M. I. AlHajri, N. Alsindi, N. T. Ali, and R. M. Shubair, "Classification of indoor environments based on spatial correlation of RF channel fingerprints," in *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*, Jun. 2016, pp. 1447–1448.

[21] M. I. AlHajri, N. T. Ali, and R. M. Shubair, "2.4 ghz indoor channel measurements." IEEE Dataport, 2018. [Online]. Available: http://dx.doi.org/10.21227/ggh1-6j32

[22] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," in *Machine learning*. Springer, 1983, pp. 463–482.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[24] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992.

[25] C. Nerguizian, C. Despins, and S. Affes, "Geolocation in mines with an impulse response fingerprinting technique and neural networks," in *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 5, Sep. 2004, pp. 3589–3594.

[26] N. Al Khanbashi, N. Al Sindi, S. Al-Araji, N. Ali, Z. Chaloupka, V. Yenamandra, and J. Aweya, "Real time evaluation of rf fingerprints in wireless lan localization systems," in *Positioning Navigation and Communication (WPNC), 2013 10th Workshop on*, 2013, pp. 1–6.

[27] Y. Shu, Y. Huang, J. Zhang, P. Cou, P. Cheng, J. Chen, and K. G. Shin, "Gradient-Based Fingerprinting for Indoor Localization and Tracking," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 4, pp. 2424–2433, Apr. 2016.

[28] J. Zhang, M. H. Firooz, N. Patwari, and S. K. Kasera, "Advancing Wireless Link Signatures for Location Distinction," in *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, ser. MobiCom '08. New York, NY, USA: ACM, 2008, pp. 26–37.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] L. E. Raileanu and K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, no. 1, pp. 77–93, May 2004.