Tech Science Press

Physical Layer Authentication Using Ensemble Learning Technique in Wireless Communications

Muhammad Waqas[1,3,*], Shehr Bano[2], Fatima Hassan[2], Shanshan Tu[1], Ghulam Abbas[2] and Ziaul Haq Abbas[4]

[1]Engineering Research Center of Intelligent Perception and Autonomous Control, Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China
[2]Faculty of Computer Science and Engineering, GIK Institute of Engineering Sciences and Technology, Topi, 23460, Pakistan
[3]School of Engineering, Edith Cowan University, Joondalup Perth, 6027, WA Australia
[4]Faculty of Electrical Engineering, GIK Institute of Engineering Sciences and Technology, Topi, 23460, Pakistan
*Corresponding Author: Muhammad Waqas. Email: engr.waqas2079@gmail.com
Received: 06 March 2022; Accepted: 25 April 2022

Abstract: Cyber-physical wireless systems have surfaced as an important data communication and networking research area. It is an emerging discipline that allows effective monitoring and efficient real-time communication between the cyber and physical worlds by embedding computer software and integrating communication and networking technologies. Due to their high reliability, sensitivity and connectivity, their security requirements are more comparable to the Internet as they are prone to various security threats such as eavesdropping, spoofing, botnets, man-in-the-middle attack, denial of service (DoS) and distributed denial of service (DDoS) and impersonation. Existing methods use physical layer authentication (PLA), the most promising solution to detect cyber-attacks. Still, the cyber-physical systems (CPS) have relatively large computational requirements and require more communication resources, thus making it impossible to achieve a low latency target. These methods perform well but only in stationary scenarios. We have extracted the relevant features from the channel matrices using discrete wavelet transformation to improve the computational time required for data processing by considering mobile scenarios. The features are fed to ensemble learning algorithms, such as AdaBoost, LogitBoost and Gentle Boost, to classify data. The authentication of the received signal is considered a binary classification problem. The transmitted data is labeled as legitimate information, and spoofing data is illegitimate information. Therefore, this paper proposes a threshold-free PLA approach that uses machine learning algorithms to protect critical data from spoofing attacks. It detects the malicious data packets in stationary scenarios and detects them with high accuracy when receivers are mobile. The proposed model achieves better performance than the existing approaches in terms of accuracy and computational time by decreasing the processing time.

1 Introduction

Introducing cyber-physical wireless communication (CPWC) systems in the manufacturing industry has significantly improved IoT devices' performance. The potential applications of CPWC systems are not only limited to manufacturing industries. Still, they have also brought significant advancements in personalized health care, traffic flow management, smart grids, transportation systems and many more. Efficient wireless transmission techniques have been proposed in the literature to meet the demand for high throughput and reliability in communication. However, wireless networks' open and broadcasting nature poses an increased risk of spoofing attacks on the network infrastructure. Cyber-physical systems are composed of various hardware components controlled and monitored by the collection of different software. The integration of these many hardware components introduces complexity, as shown in Fig. 1, leading to data integrity attacks. One of the effective countermeasures against these attacks is message authentication.
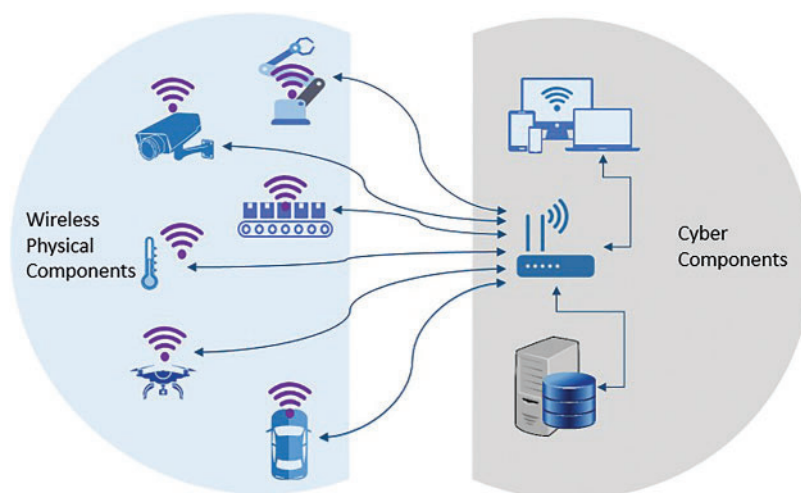

Figure 1: Cyber-physical wireless communication system

Traditional security methods were implemented at the upper logical layers using key-based cryptographic mechanisms, such as digital signatures [1–3], message authentication code (MAC) etc. However, management, distribution and the high computational cost of the key generation make it impractical to implement data encryption over these highly complex networks. Digital keys only validate the user identity and access privileges; therefore, attackers employing illegal security keys cannot readily be identified when physical-layer features are ignored. Some researchers proposed authentication using both non-cryptographic and cryptographic algorithms to solve the security and latency issues [4]. The heterogeneous nature of wireless networking makes traditional authentication methods challenging to implement or manage. However, physical (PHY) layer authentication methods provide new approaches to distinguishing legitimate messages from illegal ones. These techniques implement authentication without additional overhead. Some of the physical layer

authentication methods proposed include channel-based authentication using channel state information [5–8], radio frequency (RF) fingerprint-based schemes [9–11], received signal strength indicator (RSSI) [12–15], multi-attribute multi-observation (MAMO) techniques [16], fingerprint/watermark embedding [17,18] and so on. All these authentication methods were threshold-based, in which the channel state information (CSI) is compared with a reference CSI. The received messages for which the difference between estimated CSI and reference CSI is below the threshold are legitimate messages. However, these threshold-based Physical Layer Authentication (PLA) methods cannot solve the multi-classification problem, such as distinguishing many nodes simultaneously [19]. Although these methods work very well in stationary scenarios, the proposed techniques fail when the transmitter is mobile. In [20], the authors proposed an algorithm based on channel frequency response (CFR) statistics which considers the time variations in the case of mobile transceivers. But this solution only works well for the slowly varying channels. Therefore, learning-based authentication methods are required to learn the features of legitimate and illegitimate messages. Thus, models would be able to adapt themselves to time-varying communication quickly. Physical layer authentication can also be achieved by secret key generation [21], but keyless methods are also in use due to the latency and overhead resulting from this method.

With the advancement in artificial intelligence, threshold-free physical layer authentication methods based on machine learning and deep learning have been proposed. The DL and ML-based authentication methods can distinguish multiple sensor nodes simultaneously with high accuracy and excellent performance. In paper [22], the deep neural network framework is proposed to estimate the channel state information in orthogonal frequency division multiplexing (OFDM) systems. Similarly, [23,24] used machine learning for intrusion detection and detection of spoofing attacks. The paper [25] uses machine learning in multi-input multi-output (MIMO) wireless communication systems. It performs feature selection using neighborhood component analysis and prediction using radial basis function (RBF) kernel-based support vector machine (SVM). The proposed method in [26] detects spoofer in the Fog network using State Action Reward State Action (SARSA) and Q-learning techniques. Some previously proposed approaches include extracting channel difference using difference equations [27] and then identifying an optimal threshold value using the hit and trial method. This model achieved the best accuracy of 49.7% when the receiver and transmitter were stationary. For mobile scenarios, Pan et al. [28] used the channel difference and matrix. They classified them using Machine Learning Algorithms, which resulted in a training accuracy of 77%, with the number of sub-carriers being 128.

This paper proposes a threshold-free physical authentication model based on supervised learning to classify legitimate and illegitimate data packets. The proposed model in this paper extends the previous models with better performance and improved accuracy. The main contributions of this approach are

- Considering mobile scenarios, we have extracted relevant features from channel matrices using Discrete wavelet transformation, which has improved the computational time required to process the data.
- These reduced features are then passed to different ensemble learning algorithms, such as AdaBoost, LogitBoost and GentleBoost for data classification.
- The authentication of a received signal is a binary classification problem. The data sent by the transmitter can be labelled as a legitimate message, and the data transmitted by the spoofer is illegitimate.

2 Proposed Architecture

The block diagram in Fig. 2 shows the architecture of the proposed model. Our proposed architecture predicts whether the received data is legitimate (sent by the transmitter) or illegitimate (sent by spoofer) using supervised Machine Learning based algorithms.
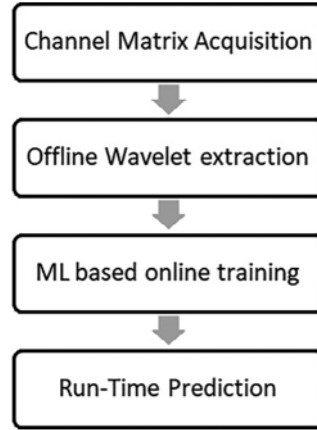
```
┌─────────────────────────────┐
│ Channel Matrix Acquisition  │
└─────────────────────────────┘
              ▼
┌─────────────────────────────┐
│ Offline Wavelet extraction  │
└─────────────────────────────┘
              ▼
┌─────────────────────────────┐
│   ML based online training  │
└─────────────────────────────┘
              ▼
┌─────────────────────────────┐
│     Run-Time Prediction     │
└─────────────────────────────┘
```

Figure 2: Proposed model's architecture

Channel State Information (CSI) plays a critical role in PLA, many methods have been used to extract channel matrices through CSI, but in this article, the received pilot is used as a channel metric. As OFDM is used for signal transmission, the received signal is a complex matrix with dimensions equal to $a \times b$ where b is the number of receiving antennas, and a is m x n with m being the number of sub-carriers used n being the number of transmitting antennas. Then feature extraction is performed to get relevant features and discard the irrelevant and noisy data, which helps improve the accuracy of ML methods.

2.1 Problem Formulation

The communication between Physical Wireless components and Cyber components of industrial CPWC systems is vulnerable to many threats, like MITM, eavesdropping, spoofing, etc. Physical-Layer Authentication is used to handle such attacks [27], which works better in stationary scenarios, but its performance degrades when the wireless components are mobile. Supervised ML classification algorithms predict whether the received signal is legitimate or illegitimate. Let α be the class label ML trained model classifies the received signal as the problem can be formulated as

$$\alpha_{1,0} = ML\left(X_R\right) \tag{1}$$

where X_R is the received signal, which also depends upon the distance of the transmitter from the receiver. When the signal is legitimate, α is 1, and when illegitimate, it equals to 0. Fig. 3 shows the system model for data acquisition of a mobile transmitter, along with a spoofer that follows the same path and speed as the transmitter. The receiver (cyber component) is stationary, and the model also covers both the line of sight path and the non-line of sight path.

The system model covers different scenarios, at $t = 0$, where the transmitter is stationary and the spoofer is not introduced in the loop. Then, at $t = T-m$, the transmitter and spoofer are moving at the same speed, and the receiver receives both of their data. Finally, at $t = T$, when the receiver has completed the loop and is stationary again, the spoofer is yet to stop.
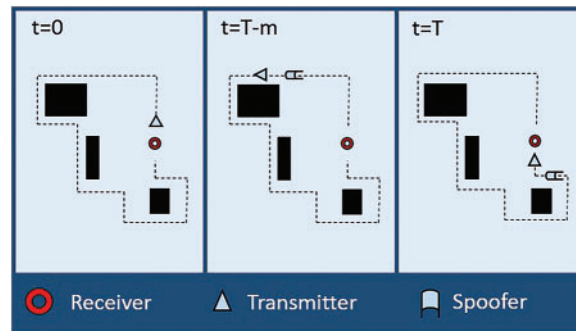
Figure 3: System model

2.2 Channel Matrix Acquisition

Many approaches follow the idea of reducing dimensions of CSI to improve computational time, but many essential features are lost in that process, especially by using channel difference equations. That's why using CSI as a channel matrix is an optimized method, for which the dimensions of the matrix can be varied. With the increase in dimensions, the training time also increases, but decreasing dimensions from a certain limit may remove important features from the data, hence reducing the accuracy. Due to this, multiple values of dimensions are considered to find an optimal tradeoff between training time and accuracy, where the values of dimensions to be considered are {64,128,8188}.

2.3 Features Extraction Using Wavelet Transform

After acquiring channel matrices, discrete wavelet transform *(DWT)* is applied to the dataset to select only meaningful, relevant features. A reduction in data size decreases the time required to process the data. DWT is preferred over other methods as it preserves the information regarding time evolution and frequency shift of the non-stationary signal. It can be calculated using the following formula:

$$DWT(x, y) = \int_{-\infty}^{+\infty} H(t)\, \psi_{x,y}(t)\, dt \tag{2}$$

where $\psi_{(x,y)}(t)$ is the scale of the wavelet, and $H(t)$ is the time-series signal. This model acts discrete wavelet transform using Haar, Daubeches, Symlets, Coiflets and Reverse Biorthogonal wavelet method. The best results are achieved by reverse biorthogonal wavelets, the inverse of which exist, but the wavelets may not be orthogonal. *DWT* deconstructs the wave such that the original wave can be reconstructed using the decomposed wavelet, which is the main principle of bijective mapping.

2.4 Classification

The last stage is the classification of data for which supervised machine learning algorithms are implemented. Supervised ML predicts whether the received packets are sent from the transmitter or spoofer. ML algorithms work better when artifacts or irrelevant features are removed from the data. The training speed decreases with the decrease in the number of trained features. This prediction model can be generated by offline data training and then used for real-time testing [29,30]. The processing speed of online prediction is further reduced by using ML as the speed of offline feature extraction is faster than the online testing. Ensemble learning works on improving classification accuracy by combining multiple algorithms with different features. Many ensemble algorithms are in use, but the ones used in this paper are AdaBoost, LogitBoost and GentleBoost, with the highest accuracy achieved

with the LogitBoost Algorithm. AdaBoost or Adaptive Boosting is an ensemble learning algorithm that combines multiple decision trees and passes the false positive and accurate negative classified data from other trees, minimizing their number. LogitBoost or Logistic Boosting is an extension of AdaBoost, and it applies Logistic Regression to training data. It is also a greedy algorithm, while the GentleBoost Algorithm updates the weights of training data to the weights of previous values.

1. LogitBoost
2. AdaBoost
3. GentleBoost

3 Results

The data collected from real-time test sites and industries are usually very time and resource consuming, so to overcome these issues, training data is acquired from a real industrial dataset provided by NIST, available publicly.

3.1 Dataset Description

The data used in this paper is available in NIST wireless dataset by the name of "Automotive Factory". The test site is a concrete building of size $400 \times 400 \times 12$ m. It consists of multiple hurdles and surfaces which may be absorbing or reflective. Fig. 4 shows the path followed by the transmitter and spoofer. The transmitter starts from position $(-1.1837, 12.6553)$ at a distance of 12.71 from the receiver and then sends data at 120 positions of the loop shown in Fig. 2. The ending position of a transmitter is $(-1.4560, -4.0801)$, which is 4.33 m away from the receiver. The receiver with an hpol antenna, as shown in Fig. 4, is stationary. The spoofer starts moving in the same loop but after the transmitter starts. It is moving at an average distance of 14.54 m from the transmitter. In total, 300 frames are sent at each position, with the bandwidth being 2.4 GHz.
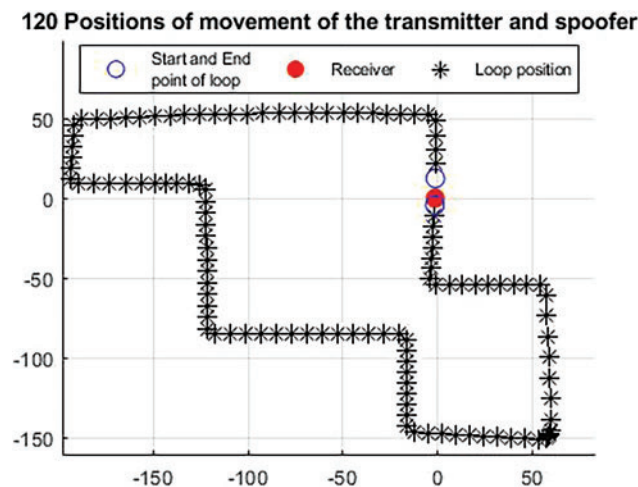

Figure 4: Loop positions of movement of transmitter and spoofer

3.2 Performance Metrics

The performance of the model is evaluated using the following parameters:

1. Authentication Accuracy
2. Prediction Time

3.3 Authentication with 8188-Channel Matrix

The training data is prepared with reverse biorthogonal wavelet transform (3.7) and then classified for 3 ensemble algorithms, AdaBoost, LogitBoost and GentleBoost. The highest average accuracy of 78.56% was achieved for Logit Boost. Fig. 5 shows the accuracy achieved for each position of the loop.
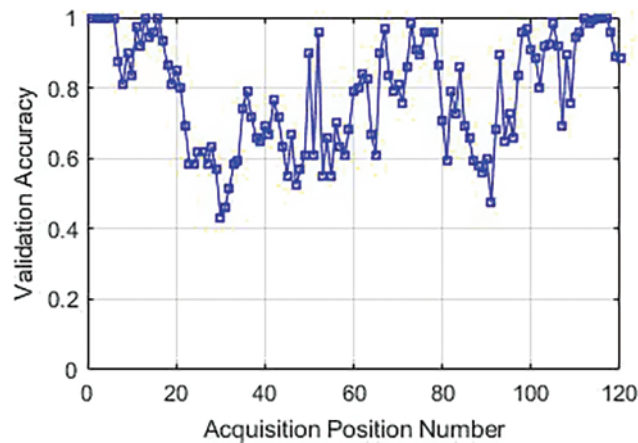

Figure 5: Authentication accuracy with the 8188-channel matrix as channel metric

3.4 Authentication with 128-Channel Matrix

Authentication accuracy significantly increases when the 128-channel matrix is trained after feature extraction, as some irrelevant features might be included in the 8188-channel matrix. This reduced dataset shows an average accuracy of 81.24% and 81.45% for AdaBoost and GentleBoost Algorithms and 82.67% for the LogitBoost Algorithm. Fig. 6 shows the validation accuracy of multiple algorithms for each loop position.
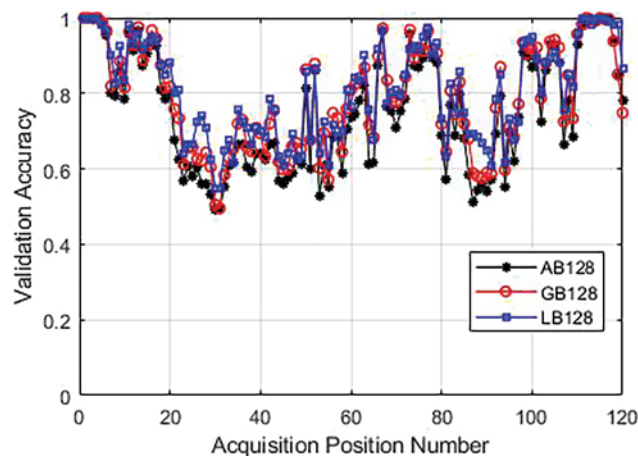

Figure 6: Authentication accuracy with 128-channel matrix as channel metric with LogitBoost, GentleBoost and AdaBoost as classification algorithm

Fig. 7 shows the validation accuracy of the LogitBoost Algorithm to each position in the loop, with a considerable decrease in training time between the 128-Channel matrix and 8188-Channel matrix.
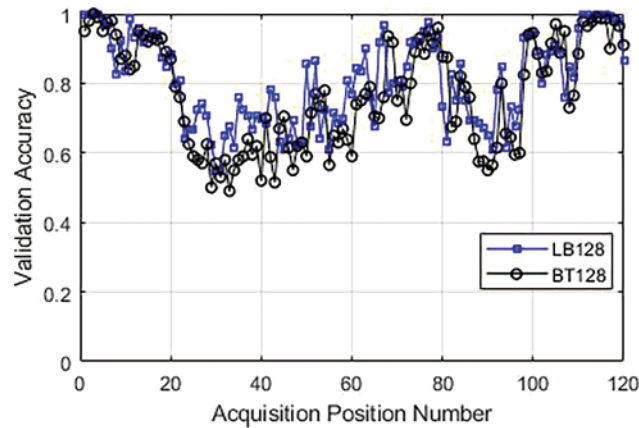

Figure 7: Authentication accuracy with 128-channel matrix as channel metric

3.5 Authentication with 64-Channel Matrix

64-channel matrix resulted in average training accuracy equal to 71.34%. It can be noted that a significant decrease in average accuracy might be caused due to training data containing not enough features for classification. Thus, it can be concluded from Fig. 8 that after a 128-channel matrix, further data reduction may result in decreased authentication accuracy.
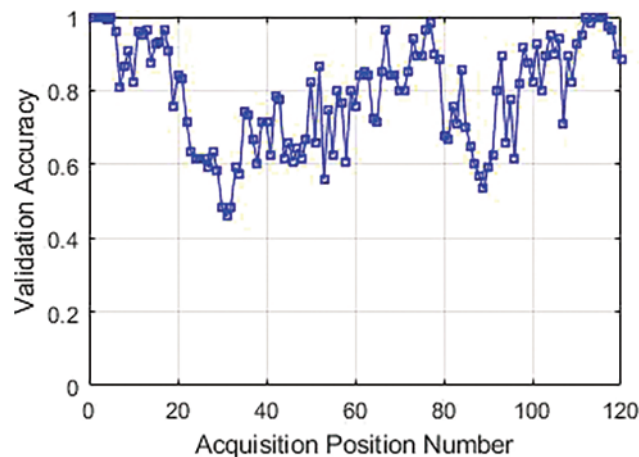

Figure 8: Authentication accuracy with 64-channel matrix as channel metric

3.6 LogitBoost with Multiple Dimensions

This paper focuses on decreasing training time by reducing channel matrix dimensions and increasing training accuracy. Fig. 9 shows the results obtained for authentication accuracy for each position in the loop using the LogitBoost algorithm. It can be seen that the average authentication accuracy results with a 128-channel matrix, along with significant improvement in training time.
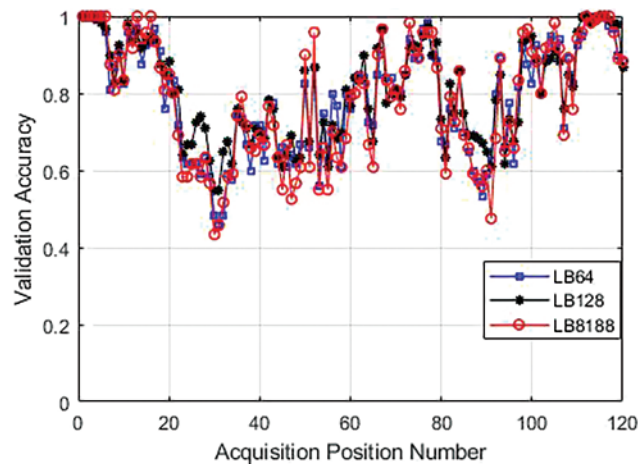
Figure 9: Authentication accuracy with 8188, 128 and 64-channel matrix as channel metric and LogitBoost as training classifier

Here, the proposed method improved accuracy for all matrix sizes, but the highest accuracy is achieved with a 128-channel matrix size, which can be chosen for the test model.

3.7 Discussion

The results show that the reduction of data size to 128-matrix classified with Logit Boost results in the highest training accuracy and a decrease in computational time compared to the previous methods. Tab. 1 compares the training and testing parameters of Boosted Tree Ensemble Algorithm and Logit Boost Ensemble Algorithm.

Table 1: Comparison between LB and BT classification

	BT128	LB128
Training time	10^0 s	4.42×10^{-1} s
Testing time	10^{-2} s	9.20×10^{-3} s
Accuracy	77.10%	82.67%

4 Conclusion

This paper proposes a comprehensive model for authenticating the received data with less computational time and more efficiency. The following conclusions can be drawn from the results:

- PLA can be considered as a binary classification problem, where efficiency is improved by appropriate feature extraction.
- There is always a tradeoff between training time and authentication accuracy, so multiple input sizes are considered to find an appropriate input value.
- Ensemble classifiers work better for PLA than the traditional ML classifiers like Trees, SVM, KNN and Naïve Bayes etc.

- Best authentication accuracy is achieved by using a 128-channel matrix and training it using the LogitBoost Algorithm.

Wireless cyber-physical communication plays an essential role in data communication for IoT, etc. This model produces improved results for the classification of legitimate and illegitimate messages than the existing threshold-free and threshold-based methods. Furthermore, the accuracy can be improved by extracting features from the *DWT* wavelet and removing the artifacts from the received signal.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. Huss, M. Waqas, A. Y. Ding, Y. Li, S. Tarkoma *et al.,* "Security and privacy in device-to-device (D2D) communication: A review," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1054–1079, 2017.

[2] M. Waqas, S. Tu, Z. Halim, S. Rehman, G. Abbas *et al.,* "The role of artificial intelligence and machine learning in wireless networks security: Principle, practice and challenges," *Artificial Intelligence Review*, vol. 29, no. 11, pp. 991, 2022.

[3] I. R. Jeong, J. Kwon and D. Lee, "Strong diffie-hellman-DSA key exchange," *IEEE Communications Letters*, vol. 11, no. 5, pp. 432–433, 2007.

[4] L. Bai, L. Zhu, J. Liu, J. Choi and W. Zhang, "Physical layer authentication in wireless communication networks: A survey," *Journal of Communications and Information Networks*, vol. 5, no. 3, pp. 237–264, 2020.

[5] C. M. Moreira, G. Kaddoum and E. Bou-Harb, "Cross-layer authentication protocol design for ultra-dense 5G hetnets," in *IEEE Int. Conf. on Communications (ICC)*, Kansas City, MO, USA, pp. 1–7, 2018.

[6] X. Wu and Z. Yang, "Physical-Layer authentication for multi-carrier transmission," *IEEE Communications Letters*, vol. 19, no. 1, pp. 74–77, 2015.

[7] D. Shan, K. Zeng, W. Xiang, P. Richardson and Y. Dong, "PHY-cram: Physical layer challenge-response authentication mechanism for wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1817–1827, 2013.

[8] F. Pan, Z. Pang, M. Luvisotto, X. Jiang, R. N. Jansson *et al.,* "Authentication based on channel state information for industrial wireless communications," in *44th Annual Conf. of the IEEE Industrial Electronics Society*, Washington, DC, USA, pp. 4125–4130, 2018.

[9] Y. Liao, G. Sun, X. Shen, S. Zhang, X *et al.,* "Yang etal, BEM-based channel estimation and interpolation methods for doubly-selective OFDM channel," in *IEEE Int. Conf. on Smart Internet of Things (SmartIoT)*, Xi'an, China, pp. 70–75, 2018.

[10] F. Xie, H. Wen, Y. Li, S. Chen, L. Hu *et al.,* "Optimized coherent integration-based radio frequency fingerprinting in internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3967–3977, 2018.

[11] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient KNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, 2018.

[12] S. van Okujeni, S. Suess and P. Hostert, "Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 4, pp. 1640–1650, 2017.

[13] S. Tu, M. Waqas, S. Rehman, T. Mir, Z. Halim *et al.,* "Social phenomena and fog computing networks: A novel perspective for future networks," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 32–44, 2022.

[14] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali *et al.,* "A survey of machine and deep learning methods for internet of things (IoT) security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.

[15] M. Waqas, M. Ahmed, J. Zhang and Y. Li, "Confidential information ensurance through physical layer security in device-to-device communication," in *IEEE Global Communications Conf. (Globecom)*, Abu Dhabi, UAE, pp. 1–7, 2019.

[16] X. Wang, P. Hao and L. Hanzo, "Physical-layer authentication for wireless security enhancement: Current challenges and future developments," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 152–158, 2016.

[17] S. Tu, M. Waqas, S. Rehman, T. Mir, G. Abbas *et al.,* "Reinforcement learning assisted impersonation attack detection in device-to-device communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1474–1479, 2021.

[18] N. Xie and S. Zhang, "Blind authentication at the physical layer under time-varying fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 7, pp. 1465–1479, 2018.

[19] R. F. Liao, H. Wen, J. Wu, F. Pan, A. Xu *et al.,* "Deep-learning-based physical layer authentication for industrial wireless sensor networks," *Sensors*, vol. 19, no. 11, pp. 2440, 2019.

[20] L. Xiao, L. Greenstein, N. Mandayam and W. Trappe, "Channel-based spoofing detection in frequency-selective rayleigh channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 5948–5956, 2009.

[21] M. Waqas, M. Ahmed, Y. Li, D. Jin and S. Chen, "Social-aware secret key generation for secure device-to-device communication via trusted and non-trusted relays," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3918–3930, 2018.

[22] F. Pan, H. Wen, R. Liao, Y. Jiang, A. Xu *et al.,* "Physical layer authentication based on channel information and machine learning," in *2017 IEEE Conf. on Communications and Network Security (CNS)*, Las Vegas, NV, USA, pp. 364–365, 2017.

[23] L. Xiao, Y. Li, G. Han, G. Liu and W. Zhuang, "PHY-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10037–10047, 2016.

[24] H. Ye, G. Y. Li and B. H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, 2018.

[25] N. Wang, T. Jiang, S. Lv and L. Xiao, "Physical-layer authentication based on extreme learning machine," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1557–1560, 2017.

[26] P. Illy, G. Kaddoum, C. M. Moreira, K. Kaur and S. Garg, "Securing fog-to-things environment using intrusion detection system based on ensemble learning," in *IEEE Wireless Communications and Networking Conf. (WCNC)*, Marrakesh, Morocco, pp. 1–7, 2019.

[27] B. Zhang, M. Waqas, S. Tu, S. M. Hussain and S. U. Rehman, "Power allocation strategy for secret key generation method in wireless communications," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 2179–2188, 2021.

[28] J. Yoon, Y. Lee and E. Hwang, "Machine learning-based physical layer authentication using neighborhood component analysis in MIMO wireless communications," in *Int. Conf. on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), pp. 63–65, 2019.

[29] J. Wan, M. Waqas, S. Tu, S. M. Hussain, A. Shah *et al.,* "An efficient impersonation attack detection method in fog computing," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 267–281, 2021.

[30] F. Pan, Z. Pang, H. Wen, M. Luvisotto, M. Xiao *et al.,* "Threshold-free physical layer authentication based on machine learning for industrial wireless CPS," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6481–6491, 2019.