

# Q-Learning Driven Spectrum Prediction for Energy-Efficient RF-Powered D2D Communications

Avik Banerjee<sup>1</sup>, Santi P. Maity<sup>2</sup>, Iacovos I. Ioannou<sup>3</sup>, Prabagarane Nagaradjane<sup>4</sup>, and Vasos Vassiliou<sup>3</sup>

<sup>1</sup>Dept. of Electronics and Communication Engineering, R V College of Engineering, Bengaluru 560059, India

<sup>2</sup>Dept. of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India

<sup>3</sup>Dept. of Computer Science, University of Cyprus, and CYENS–Centre of Excellence, Nicosia 1678, Cyprus

<sup>4</sup>Dept. of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai 603110, India

Emails: avikbanerjee@rvce.edu.in, santipmaity@it.iiests.ac.in, ioannou.iakovos@ucy.ac.cy, prabagaranen@ssn.edu.in, vassiliou.vasos@ucy.ac.cy

**Abstract**—Device-to-device (D2D) communications face challenges of spectrum scarcity and limited power, hence necessitating energy-efficient system design that also meets target data rate requirements. To address these issues, a reinforcement learning (RL)-based Q-learning scheme is proposed within a cognitive radio (CR) framework for primary user (PU) spectrum prediction (SP). This approach enables opportunistic data transmission and radio frequency (RF) energy harvesting (EH) for sustainable transmission of devices. The RL algorithm aims to maximize energy efficiency (EE) while satisfying constraints on target data transmission rate, energy harvesting requirements, and interference thresholds permissible at the PU receiver to protect it in the event of wrong prediction. A comprehensive set of simulations is conducted to evaluate the proposed method, reporting improvements in spectrum prediction accuracy, normalized energy efficiency, and residual energy. The results demonstrate a gain of 35% in EE, a 25% reduction in data collisions, and a 35% improvement in residual energy over the reported works at reduced trained parameters.

**Index Terms**—Device-to-device, reinforcement learning, CRN, energy harvesting.

## I. INTRODUCTION

Device-to-device (D2D) data transmission becomes one enabling technology in upcoming 6G to provide short-range communication with low latency [1]. Even though D2D communications offload the mobile traffic at the base station (BS), reduce energy consumption in data transmission, and enhance spectrum usage, still an efficient system design needs to consider these issues to achieve improvement in energy efficiency (EE), intelligent spectrum access leading to spectrum efficient communications while ensuring to meet the target data rate demand. Emergence of Internet of-Things (IoT) makes D2D communication a practical choice to address various short-range low data rate application-specific services [2]. Often these services are hampered as the transmit devices are battery driven, and suffer from network lifetime problems, hence, require periodical replacement or recharging of the nodes (devices). In brief, in D2D communications, a few challenges that the devices face are availability or assurance of free spectrum for seamless connectivity, enhancement of network lifetime due to limited battery power, and transmission power control for collision-free data transmission while meeting the target data rate requirement.

Recent works show various spectrum sharing schemes in D2D communications by means of interweave, underlay and overlay modes of communication. Power control in D2D communications allows frequency sharing with the licensed users' bandwidth in an underlay mode but often fails to meet the target data rate requirement. Interweave mode of spectrum sharing allows D2D communications by opportunistic access of primary user (PU) spectrum without exceeding its permissible interference limit. The means of opportunistic communication where an unlicensed spectrum user, also called secondary user access licensed spectrum is called cognitive radio (CR) technology [3], and the users are called as CR and PU, respectively.

Spectrum sensing (SS) and spectrum prediction (SP) are the two primary approaches in CR communication where secondary users (SUs), before starting data transmission, find the free spectrum i.e., spectrum hole [4]. Reliable SS leads to collision-free data transmission of PU and SU, hence, one communication is not hampered by the other. Due to the stochastic nature of the PU signal as well as the sensing channel, a single user's SS result is not reliable. Multiple SUs get involved to form a cooperative SS (CSS) scheme where each individual sends its sensing signals or local sensing decisions to the fusion center (FC) that forms a global decision [5]. Literature on SS/CSS is quite rich and addresses several issues, namely reliability improvement in sensing, energy efficient SS, security issues, etc. However, any kind of SS/CSS scheme causes energy drainage of SU nodes [4], [5].

Joint SS and SU data transmission perform periodic sensing followed by opportunistic data transmission on a time-shared mode [3]. Both the operations, being tightly coupled, the duration of one affects the other. This trade-off issue is addressed a lot in the literature [5]. Dedicated SS/CSS operation not only drains a substantial power of SU nodes but also diminishes its data transmission duration in periodic frame structure [5]. One way to alleviate SS operation is to adopt SP where the state of use/non-use of PU spectrum at a particular time instant is determined based on some information, such as historic traffic pattern, signal-to-noise ratio (SNR) values, PU data transmission rate, SU energy harvesting (EH), etc on previous slots. A good number of works on SS and SP is reported

using different deep learning (DL) approaches from partial SS information [6], [7], SS using machine Learning (ML) [8] SP using partial observable Markov Decision process (POMDP) boosting D2D data transmission [9], SP, CR network (CRN) routing and EH [5]. When contrasted with SS, SP offers benefits in energy savings, enhanced data transmission time slot for SU and reduction in sensing results transmission overhead [4].

An accurate prediction on PU spectrum hole leads to improved EE in SU data transmission. At the same time, if PU transmission state is determined accurately, two-fold benefits can be achieved; data collision between PU and SU can be avoided. At the same time, SU nodes (devices) perform EH using PU's transmitted radio frequency (RF) signal. In other words, high accuracy in SP not only protects PU from SU interference but also improves EE of SU nodes by means of power control and provides EH-enabled self-powering. The present work focuses on improvement in EE for D2D communications using opportunistic transmission through SP that also have provisions for EH.

The remainder of this paper is structured as follows. Section II provides a review of related literature. Section III describes the proposed system model along with the problem formulation. Section IV outlines the Q-learning approach adopted in this work. Section V presents the performance evaluation through simulation results, and Section VI concludes the paper.

## II. LITERATURE REVIEW AND SCOPE OF THE WORK

Of late Hidden Markov Model (HMM), long short term memory (LSTM) and convolution neural network (CNN) have been used to build up SP model. ML techniques on SS and SP are also been reported [5]–[7], [9]. Supervised learning such as support vector machine (SVM) classifies PU transmit or non transmit state using the trained hyperplane [10]. The authors in [5] suggests predicting PU transmission behavior using an SVM model, eliminating the need for explicit spectrum sensing. The authors further develop a deep Q-network (DQN)-based routing strategy aiming to improve both EE and spectrum efficiency (SE), maximizing CRN throughput.

CSS using deep reinforcement learning (DRL) is suggested in [11], where DL analyses the spectrum environment and Reinforcement Learning (RL) is used for spectrum decision. A local sensing is done by SU and conservative Q-learning (QL) is then done to determine transmit/non-transmit state of PU for the current as well as future  $k - 1$  ( $k > 1$ ) time slot, which in turn eliminates the needs of SS for the next ( $K-1$ ) slot. In [12], an LSTM-based cooperative SP (CSP) method is introduced for energy constrained CRNs. The LSTM predicts the channel condition ahead of sensing, and a parallel fusion CSP scheme is then applied to mitigate errors arising from individual predictions. The work in [13] also employs LSTM networks for SP, treating spectrum samples as time-series data. Their results indicate that LSTM outperforms multilayer perceptron (MLP) models and yields slightly improved classification accuracy relative to regression models.

Due to dynamic traffic of PU, SP of SU for collision free data transmission is characterized as Markov Decision Process (MDP). MDP is addressed through RL strategy in D2D-CR SP problem [14] and channel switching in [15] to identify idle PU channel. In [5], SVM (for SP) and DQN (for routing) are utilized as two independent ML modules, necessitating different training, testing, and validation data sets for each. To the best of knowledge, EE maximization in CRN based D2D communications with EH through SP has not been explored.

### A. Our Contributions

To simultaneously manage spectrum-hole detection, EE maximization, and EH, this work employs an RL-inspired SP strategy that avoids SS and enables opportunistic D2D communication. Correctly predicting PU inactivity yields collision-free D2D transmission and is rewarded through improved EE. An incorrect prediction, however, triggers a penalty by compelling the D2D user to reduce its transmit power to protect the PU, resulting in reduced EE. An earlier form of this concept was presented in [14], where a high misprediction probability was treated as a penalty term that obligated the device to support PU transmission, thereby lowering its own usable transmit power and shortening network lifetime. Furthermore, it was assumed that transmit device is powered from the external sources i.e. external source was only means for powering the device without the scope of EH. To address this power problem, in the present work, the correct prediction on PU transmission state is purposely used by the transmit device for EH, this makes self-powered system design. Moreover, the switch mode operation of device's own data transmission and participation in PU data transmission in the model [14] are governed by special instruction received from the agent. This leads to an overhead issue, which the proposed model effectively mitigates. The contributions of the proposed work are as follows:

- i) Joint SP and RF-EH powered D2D communication using an RL-based Q-learning approach is suggested, offering the benefits of reduced parameter requirements and lower computational complexity. EE maximization is achieved through high accuracy on SP while interest of both PU and SU are maintained in terms of interference protection, energy causality and data rate constraints.
- ii) Mathematical expressions of objective function and the constraints include different SP probabilities in a tightly coupled form. The desired values of probabilities have direct role in reward and penalty functions leading to maximize cumulative reward function while solving the constrained objective functions.
- iii) A large set of simulations are done that show 25% reduction in data collision over [8] and improvement by 35% on both EE and residual energy over [4] at reduced trained parameters.

## III. SYSTEM MODEL AND PROBLEM FORMULATION

Fig. 1(a) depicts the PU link, consisting of a base station (BS) and its user equipment (UE) within a cellular network.

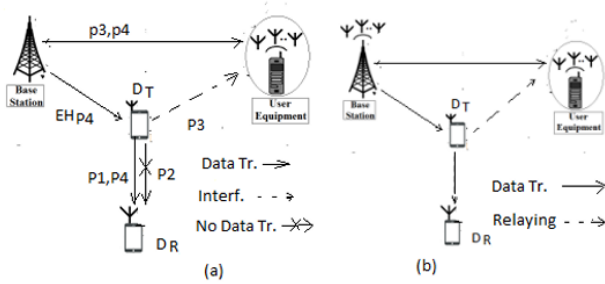


Fig. 1. System model (a) proposed and (b) from [14].

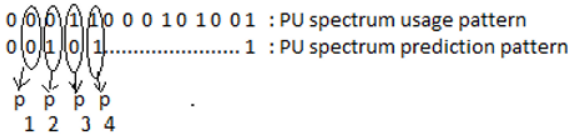


Fig. 2. PU spectrum usage pattern and predicted pattern.

A D2D communication based IoT system that consists of DT as device transmitter and DR as device receiver, is considered as SU network. DT contains receiver circuit and based on the received PU signals makes SP which allows an opportunistic access of PU spectrum. Fig. 1(b) shows the system model used in [14] to highlight the difference in operations with respect to the present one. Both Figs. 1(a) and (b) also show the various data transmission signals. A one-way DT–DR communication link is considered, bidirectional communication may be supported through time or frequency division sharing.

The objective of spectrum prediction is to determine PU transmit/non-transmit state, which is represented by a binary random variable (RV) ‘x’ taking values 0’s and 1’s indicating PU non-transmit and transmit state, respectively. These two hypotheses are denoted by  $H_1$  and  $H_0$  and the probabilities of PU’s presence and absence are denoted by  $P(H_1)$  and  $P(H_0)$ , respectively. Here  $P(H_1)$  values ranges between 0.3 to 0.4 as per measured reported by different regulatory bodies. Fig. 2 represents two sequence of PU transmit.

$$p_1 = \frac{\text{PU's non transmit state predicted}}{\text{PU actually in non-transmit state}} \\ = \frac{\text{Number of 0's in PU spectrum sequence predicted}}{\text{Number of 0's in PU's data transmit sequence}} \quad (1)$$

$$p_2 = \frac{\text{PU predicted as transmit}}{\text{PU is actually in non-transmit state}} \\ = \frac{\text{Number of 1's in predicted sequence}}{\text{Number of 0's in actual non-transmit sequence}} \quad (2)$$

$$p_3 = \frac{\text{PU predicted as non-transmit}}{\text{PU is actually in transmit state}} \\ = \frac{\text{Number of 0's in predicted sequence}}{\text{Number of 1's in PU sequence}} \quad (3)$$

$$\begin{aligned} p_4 &= \frac{\text{PU spectrum as in transmit state}}{\text{PU is actually in transmit state}} \\ &= \frac{\text{Number of 1's in predicted sequence}}{\text{Number of 1's in PU sequence}} \end{aligned} \quad (4)$$

Here,  $p_1$  indicates correct prediction of PU's non-transmission state i.e. correct prediction of spectrum hole which in turn causes the scope of D2D communications. Probability  $p_2$  indicates the situation when PU is in non-transmit state but predicts as transmit state which causes situation of loss in PU spectrum usage. Note that for an opportunistic data transmission,  $p_1 + p_2 = 1$ , and a high value of  $p_1$  is desirable in SP. The probability  $p_3$  captures the event of inaccurately predicting PU activity, which results in unintended D2D interference and necessitates power regulation to satisfy the PU interference constraint. In contrast,  $p_4$  denotes a correct prediction of PU transmission, allowing the DT to perform RF-EH. Here also  $p_3 + p_4 = 1$ , where a low value of  $p_3$  is desirable to ensure interference protection of PU.

To have a good PU spectrum prediction, it is desirable to have high  $p_1$  and low  $p_2$  values, which consequently indicate D2D high data rate transmission and low loss in PU spectrum usage, respectively. In the same line of prediction accuracy, low  $p_3$  and high  $p_4$  values are desirable to have lower data collision in PU and SU data transmission and a scope of high RF-EH for the device, respectively.

The work aims to maximize D2D EE while meeting its target data rate, PU interference protection and meeting power of transmission required to transmit device from the harvested energy. Mathematically, the problem can be written as

$$\max_{P_D} EE = \frac{p_1 R_D - p_2 R_D}{(p_1 + p_3) P_D} \quad (5)$$

$$\text{s.t. } p_1 R_D \geq R_{\text{DT}} \quad (6)$$

$$|h_c|^2 p_3 P_D \leq I_{\text{th}} \quad (7)$$

$$P_h = P_D = \eta \cdot p_4 (\sigma_{ps}^2 P_p + \sigma_{ns}^2) \quad (8)$$

where  $P_D$  is the device's (DT) own data transmit power, and  $R_D$  is D2D data rate, given by:

$$R_D = \log \left( 1 + \frac{p_1 h_d^2 P_D}{\sigma_{n_s}^2} \right)$$

Here the symbols  $\sigma_{ps}^2$  and  $\sigma_{ns}^2$  denote the channel variance and noise variance of the SU receiver, respectively.

The symbol  $h_d$  represents the fading gain of the D2D data transmission path. Noise signal is considered to be modeled as circularly symmetric complex Gaussian (CSCG) random variable with zero mean and variance  $\sigma^2$ . It is assumed that D2D data transmission experiences Rayleigh fading, the channel coefficient  $h \sim \mathcal{CN}(0, d^{-\alpha})$  follow the model where the symbol  $d$  corresponds to the propagation distance and  $\alpha$  denotes the path-loss exponent. The parameter  $R_{DT}$  refers to the minimum data rate requirement of the DT, while  $I_{th}$  specifies the allowable interference level at the PU receiver. Eq. (8) indicates energy causality i.e. the harvested energy is the only source of transmit power of DT and in this work

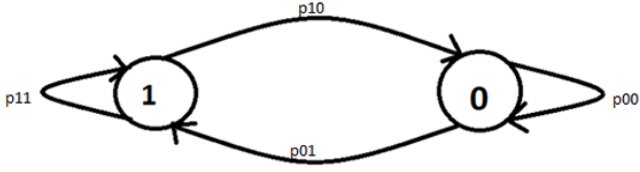


Fig. 3. Markov chain model of PU

a linear model of EH is used. To maximize the objective function in Eq. (5) under the worst-case condition of maximum interference, the inequality constraints in (6)–(7) are taken with equality.

#### IV. SP AND D2D COMMUNICATION VIA Q-LEARNING

The main objective is to determine the optimal value of device transmit power  $P_D$  that maximizes the system EE by accurately predicting the PU state at each time slot while satisfying the above constraints. A Markov chain model is employed to characterize the PU activity, following the Markov property with state–transition probabilities  $p_{00}$ ,  $p_{01}$ ,  $p_{11}$ , and  $p_{10}$ , as shown in Fig. 3. Here, state 0 denotes that the channel is currently idle (i.e., not accessed by the PU), whereas state 1 denotes that the channel is occupied by the PU.

##### A. MDP Problem Formulation

A discrete-time MDP is characterized as follows:

- A finite state set  $S = \{s_i\}$ ,  $i = 1, 2, \dots, n$ , where each  $s_i \in \{0, 1\}$ . Here,  $s_i = 0$  denotes that the spectrum is free for transmission, while  $s_i = 1$  indicates that the PU occupies the spectrum.
- A finite action set  $A(s_i) = \{a_i\}$ ,  $i = 1, 2, \dots, |A(s_i)|$  for each state  $s_i \in S$ , where  $a_i \in \{0, 1\}$ . Action  $a_i = 0$  means the SU predicts the spectrum to be free, and  $a_i = 1$  means the SU predicts the spectrum to be occupied by the PU.
- Let  $s_{i,t}$  and  $a_{i,t}$  are the state and action at time instant  $t$ , respectively. The state transition probability is given by
$$p(a_{i,t}, s_{i,t}) = \Pr(s_{i,t+1} = s_j \mid s_{i,t} = s_i, a_{i,t} = a_i).$$
- The reward function associated with taking action  $a_{i,t}$  in state  $s_{i,t}$  is defined as

$$\text{Reward}(s_{i,t}, a_{i,t}) = \max(\text{EE}(P_D)).$$

##### B. Proposed Q-Learning

An MDP is characterized by a tuple  $(s, a, t, r, \beta)$ , where  $s, a, t, r$  and  $\beta$  represent state space, action space, state transition likelihood, reward and discount factor, respectively. A complete knowledge of statistical characteristics of MDP system is needed to develop the state transition matrix. Algorithm 1 describes the pseudo-code for the Q-learning of the proposed method. The symbols  $E_h$  and  $E_{\text{res}}$  indicate energy harvesting and residual energy, respectively.

A Q-table of  $n \times 2$  is developed where the number of time steps/instants of observation is denoted by  $n$ . In each

---

#### Algorithm 1 Pseudo-code: Q-learning in the present work

---

**Require:**  $R_{DT}, I_{th}$

**Ensure:** Probabilities  $(p_1, p_2, p_3, p_4)$ , EE,  $R_D$ ,  $E_{\text{res}}$ ,  $E_h$

```

1: Initialize action value function Q-table  $Q(s, a)$  with random weights 'w',
2: for each epoch  $j$  do
3:   for each time step  $t$  do
4:     SU predicts state of the PU by choosing action  $a_i$ 
5:     if SU correctly predicts then
6:       Calculate reward  $r$  using Eq. (9)
7:     else
8:       Calculate penalty  $p$  using Eq. (10)
9:     end if
10:    Perform a gradient descent step to update Q-values using Eq. (11)
11:    Update the power level  $P_D$  of device
12:  end for
13:  Probabilities  $(p_1, p_2, p_3, p_4)$  are updated.
14:  Energy efficiency, power consumed, data rate,  $E_h$ ,  $E_{\text{res}}$  are calculated.
15: end for
16: return  $Q$ 

```

---

row of two columns  $s_{i,t}$  of PU is stored using an action  $a_{i,t}$ . Correct prediction leads to state-action reward ( $r$ ) while wrong prediction due to state-action is negatively rewarded i.e. a penalty is assigned in cumulative reward function Reward ( $r$ ) and Penalty ( $p$ ) are defined as follows

$$\text{Reward}(r) = \frac{|(p_1 - p_2)R_D|}{(p_1 + p_3)P_D} \quad (9)$$

$$\text{Penalty}(p) = -\frac{|(p_1 - p_2)R_D|}{(p_1 + p_3)P_D} \quad (10)$$

At time instant  $t$ , with prediction of PU spectrum state, SU updates the Q table with weight 'w' as follows:

$$Q(s_{i,t}, a_{i,t}) \leftarrow (1 - \rho) Q(s_{i,t}, a_{i,t}; w) + \rho \left[ r + \beta \max_{a' \in A(s_{i,t+1})} Q(s_{i,t+1}, a'; w) \right]. \quad (11)$$

where learning and discount factors are denoted by  $0 \leq \rho \leq 1$  and  $0 \leq \beta \leq 1$ , respectively.

#### V. NUMERICAL RESULTS AND DISCUSSION

This section presents the performance of the proposed RL-based SP scheme in terms of prediction accuracy, the resulting EE, and  $E_{\text{res}}$  in D2D communications. Simulation to implement the proposed method is run on an Apple MacBook Air, featuring a ten-core CPU, a ten-core GPU with hardware-accelerated ray tracing, a sixteen-core Neural Engine, and sixteen gigabytes of unified memory. The simulation parameter values are as follows:  $P(H_0) = 0.7$  and  $P(H_1) = 0.3$ ,  $d = 1.3$  m,  $\alpha = 3$ ,  $P_p = 0.6$  W,  $I_{th} = 0.45$  W,  $R_{DT} = 0.5$  bps/Hz,  $\sigma_{ns}^2 = \sigma_{ps}^2 = 0$  dBm,  $\rho = 0.48$ ,  $\beta = 0.52$  and  $\eta = 0.27$  ( $\eta$  indicates EH conversion efficiency).

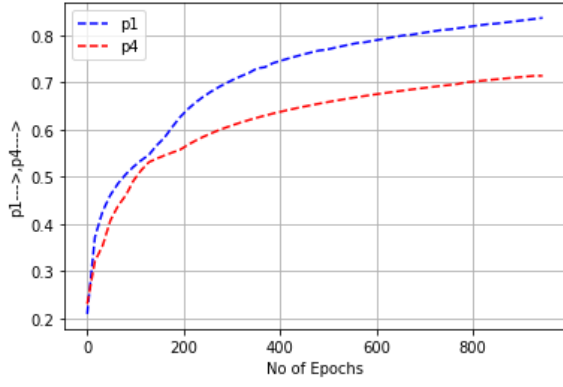


Fig. 4. SP accuracy:  $p_1$  and  $p_4$  vs number of epochs

The graphical plot of  $p_1$  and  $p_4$  with the number of epochs are shown in Fig. 4. Both  $p_1$  and  $p_4$  increase with the increase in the number of epoch indicating the improved learning of the system leading to high accuracy on SP. High values of  $p_1$  and  $p_4$  indicate a higher prediction accuracy for identifying PU spectrum holes and correctly detecting PU transmission, respectively.

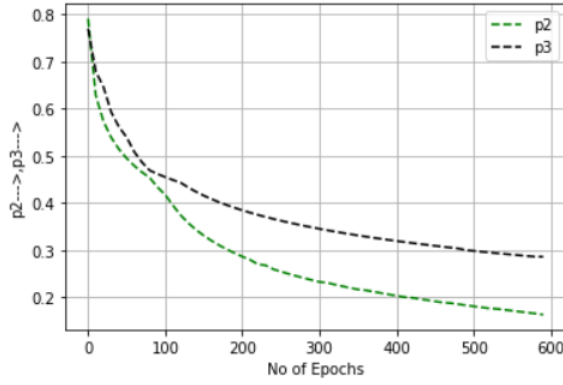


Fig. 5. SP accuracy:  $p_2$  and  $p_3$  vs number of epochs

Fig. 5 also shows similar SP results in term of  $p_2$  and  $p_3$  versus the number of epochs where RL algorithm shows reduced values of both that lead to the reduction in spectrum loss in D2D communications and interference protection of PU. Graphical results also show that numerical values of  $p_1$  and  $p_2$  as well as  $p_3$  and  $p_4$  are mutually complementary satisfying their sum 1.

Performance of the proposed SP is compared over existing SS/SP methods [4], [8], [9], [14] using Receiver Operating Characteristic (ROC) curves. The change in probability of detection ( $p_d$ ) as a parameter of probability of false alarm ( $p_f$ ) is an essential characteristics of any SS/SP method and the plot is shown as ROC. It is expected to have Area under the curve (AUC) value to be close to one (ideal case), 0.945 is the AUC value for the present method, while the same are found to be 0.923 in [14], 0.892 in [8], 0.824 in [9] and 0.746 in [4] at reduced cost.

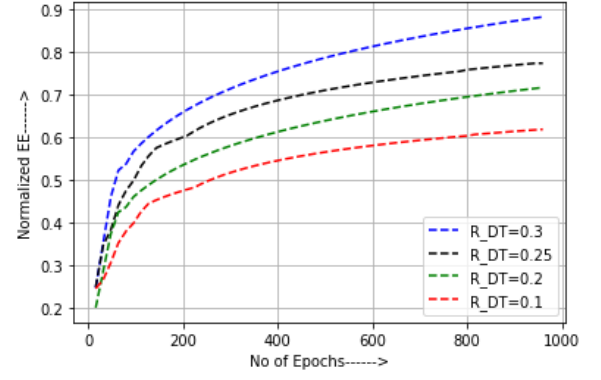


Fig. 6. Normalized EE vs number of epochs

Fig. 6 shows the improvement in normalized EE as the number of training epochs increases, evaluated under different D2D data-rate constraints. Non-fading conditions yields maximum data transmission while the same is reduced at fading channel. Normalized EE corresponds to the ratio of EE on fading channel to that of non-fading channel. Increased EE values results due to reduction in  $p_2$  and  $p_3$ , which allows more transmit power for D2D communication, thereby enhancing its data rate. With the increase in D2D data-rate constraint, normalized EE improves correspondingly because a higher data transmission rate is achieved.

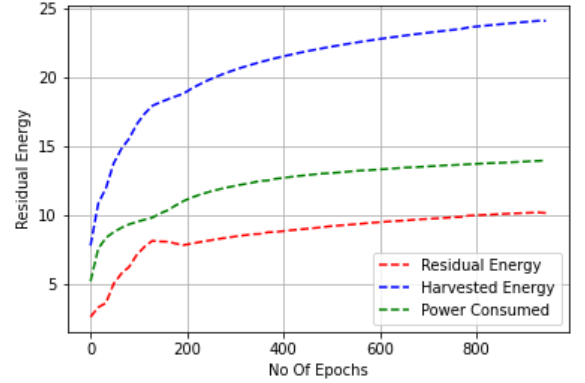


Fig. 7. Residual energy vs number of epochs

Fig. 7 shows graphical plot of residual energy in joule versus the number of epochs. Residual energy is the savings in energy, after consumption of power needed for D2D communication, from its harvested energy. As seen in Fig. 4, as the number of epochs increases, the value of  $p_4$  value also increases, i.e., with the increase in PU spectrum usage detection, device keeps its harvester circuit ON that leads to more harvesting energy. Results reported in Fig. 7 are obtained using  $P_p = 0.6$  watt and energy conversion efficiency  $\eta = 0.27$ . For a given data transmission rate, after the required transmit energy consumption, the left over energy stores as residual energy which goes on increasing with number of epochs ( $p_4$  values).

The effect of  $p_4$  on residual energy is reflected in Fig. 8 that shows graphical plot on residual energy in joule vs probability

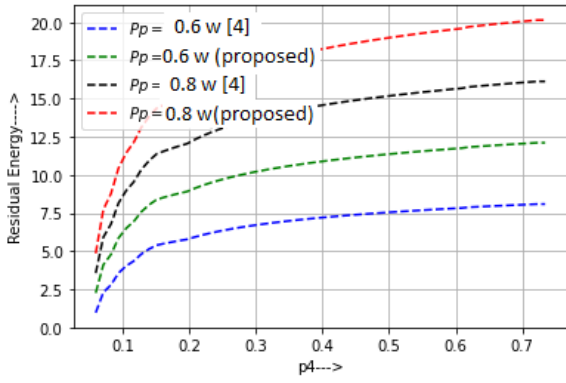


Fig. 8. Residual energy vs PU spectrum detection probability

$p_4$  at different values of PU transmit power  $P_p$ . It is observed that residual energy values increase with the increase in PU transmit power for a given  $p_4$  value. This is obvious as high PU power implies more RF energy power fed to harvester circuit, which for a given efficiency, yields high value of harvesting energy and consequent increase in residual energy. Fig. 8 also shows performance comparison in residual energy over [4] at  $P_p = 0.8$  W and  $P_p = 0.6$  W. About 35% and 50% increase are seen on the values of residual energies for the proposed method over [4]. When relative performance are compared with the existing works, 35% improvement in normalized EE over [4] and 25% reduction in data collision over [8] are achieved.

TABLE I  
SS/SP PERFORMANCE COMPARISON

Method	Norm. EE	Gain in EE	$p_4$ ( $p_d$ )	$p_2$ ( $p_f$ )
SVM & DQN [5]	0.90	26.76%	0.80	0.32
DRL [8]	—	—	0.82	0.29
RL [14]	0.93	30.98%	0.94	0.20
Proposed	0.95	34.76%	0.96	0.18

Table I provides a detailed comparison of SS/SP performance across existing methods, e.g., SVM & DQN [5], DRL [8], and RL [14], and the proposed scheme, evaluating normalized EE, EE gain, and sensing metrics ( $p_d$ ,  $p_f$ ) to show how the proposed method outperforms prior approaches.

## VI. CONCLUSIONS AND SCOPE OF FUTURE WORKS

An RF-EH based self-powered D2D communication framework CRN is proposed in this work where spectrum hole as well as PU transmit states are determined through spectrum prediction (SP). RL algorithm based Q learning is employed for SP, leading to maximization of EE subject to RF-EH causality, PU protection, and meeting data rate requirements, at reduced number of weights to be trained.

Simulation results highlight 25% and 35% improvement on EE compared to [14] and [4], respectively, while protection of PU from data collision is achieved by 25% over [8], residual energy improvement by 35% over [4].

Future works for the proposed method may focus on

- Proposed SP can be extended in the framework of CSS using actor-critic based RL work.
- The present work predicts a single channel of PU, and can be extended for predicting multiple channels, enabling hybrid-mode (interweave, underlay, and overlay) bidirectional D2D communications.
- The present system model considers a linear EH model; future work may consider a non-linear EH model for more realistic performance analysis.

## VII. ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578, the METACITIES project under Grant Agreement No. 10108725, and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

## REFERENCES

- [1] Y. Li, F. Zhou, L. Yuan, Q. Wu, and N. Al-Dhahir, "Cognitive semantic communication: A new communication paradigm for 6G," *IEEE Commun. Magazine*, pp.1-8, 2025.
- [2] X. Liu, and N. Ansari, "Green relay assisted D2D communications with dual batteries in heterogeneous cellular networks for IoT," *IEEE Things Journal*, vol. 4, no.5, pp.1707-1715, 2017.
- [3] N. Du, C. Zhou, H. Shen, C. Chakraborty, J. Yang, and K. Yu, "A novel power allocation algorithm for minimizing energy consumption in D2D communication systems," *IEEE Systems Journal*, vol. 17, no. 3, pp. 4969-4977, 2023.
- [4] S. Ghose, A. Kundu, D. Mishra, S. P. Maity, A. Al-Nahari, R. Jäntti, "Energy Efficient RIS-Assisted Wireless Powered D2D Communications in Cognitive Radio Networks," *IEEE Trans. on Green Commun. & Networking*, vol. 9, No. 3. pp.1254-1267, 2025.
- [5] A. Paul, and S. P. Maity, "Machine learning for spectrum information and routing in multi-hop green cognitive radio networks," *IEEE Trans. on Green Commun. & Network*, vol. 6, no.23 pp.825-835, 2022.
- [6] N. El-haryqy, Z. Madini, and Y. Zouine, "A review of deep learning techniques for enhancing spectrum sensing and prediction in cognitive radio systems: approaches, datasets, and challenges," *Int. Journ. Comput. Appl.*, vol. 46, no.12 pp.1104-1128, 2024.
- [7] L. Feng, Y. Gao, Z. Xu, L. Yu, and Y. Cheng, "Deep learning-based intelligent spec- trum sensing and prediction: A survey," *IEEE Access*, vol. 10, no.12 pp.13078-13096, 2022.
- [8] R. Sarikhani and F. Keynia, "Cooperative spectrum sensing meets machine learning:deep reinforcement learning approach," *IEEE Commn. Lett.*, vol. 24, no.7 pp.1459-1462, 2020.
- [9] A. Paul, S. Das and S. P. Maity, "Spectrum Prediction: Boosting D2D communications in CRNs using POMDP," *Physical Communications*, vol. 71, pp.102704, 2025.
- [10] Y. Li, J. Wu, Y. Lou, X. Xu, and J. Bao, "Enhanced support vector machine for cooperative spectrum sensing against byzantine attack in cognitive wireless sensor networks," *IEEE Sens. J.*, vol. 24, No. 23. pp.39835-39844, 2024.
- [11] S. Q. Jalil, M. H. Rehmani and S. Chalup, "Cognitive radio spectrum sensing and prediction using deep reinforcement learning," *Proc. Inter. Joint Conf. on Neural Networks*, 18-22 July, 2021, Shenzhen, China.
- [12] P. Chauhan, S. K. Deka, B. C. Chatterjee and N. Sarma, "Cooperative spectrum prediction-driven sensing performance for energy constrained cognitive radio networks," *IEEE Access*, vol.9, pp.26107-26118, 2021.
- [13] L. Yu, J. Chen, G. Ding, J. Yang and H. Sun, "Spectrum prediction based on Taguchi method in deep learning with long short-term memory," *IEEE Access*, vol.6, pp.45923-45933, 2018.
- [14] S. P. Maity, K. Sinha, B. P. Sinha and R. Kumari, "Reinforcement learning for spectrum prediction and EE maximization in D2D communication," *Proc. IEEE International Conference on Signal Processing and Communication*, 11-15 July, 2022, IISC Bangalore, India.
- [15] H. Ding, X. Li, Y. Ma, and Y. Fang, "Energy-efficient channel switching in cognitive radio networks: A reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol.69, no. 10,pp.12359-12362, 2020.