Pavle BUGARČIĆ, Ph.D. student[1]
E-mail: p.bugarcic@sf.bg.ac.rs
Nenad JEVTIĆ, Ph.D.[1]
E-mail: n.jevtic@sf.bg.ac.rs
Marija MALNAR, Ph.D.[1]
E-mail: m.malnar@sf.bg.ac.rs
[1] University of Belgrade
Faculty of Transport and Traffic Engineering
Vojvode Stepe 305, 11000 Belgrade, Serbia

Information and Communication Technology
Review
Submitted: 3 May 2022
Accepted: 30 July 2022

REINFORCEMENT LEARNING-BASED ROUTING PROTOCOLS IN VEHICULAR AND FLYING AD HOC NETWORKS – A LITERATURE SURVEY

ABSTRACT

Vehicular and flying ad hoc networks (VANETs and FANETs) are becoming increasingly important with the development of smart cities and intelligent transportation systems (ITSs). The high mobility of nodes in these networks leads to frequent link breaks, which complicates the discovery of optimal route from source to destination and degrades network performance. One way to overcome this problem is to use machine learning (ML) in the routing process, and the most promising among different ML types is reinforcement learning (RL). Although there are several surveys on RL-based routing protocols for VANETs and FANETs, an important issue of integrating RL with well-established modern technologies, such as software-defined networking (SDN) or blockchain, has not been adequately addressed, especially when used in complex ITSs. In this paper, we focus on performing a comprehensive categorisation of RL-based routing protocols for both network types, having in mind their simultaneous use and the inclusion with other technologies. A detailed comparative analysis of protocols is carried out based on different factors that influence the reward function in RL and the consequences they have on network performance. Also, the key advantages and limitations of RL-based routing are discussed in detail.

KEYWORDS

reinforcement learning; Q-learning; routing protocols; VANET; FANET; ITS.

1. INTRODUCTION

Modern life cannot be imagined without the usage of some type of wireless ad hoc networks (WANETs) with dynamic nodes that can participate in data packet routing. The most common dynamic WANETs are mobile ad hoc networks (MANETs), vehicular ad hoc networks (VANETs) and flying ad hoc networks (FANETs). Although with VANETs a wide range of services for intelligent transportation systems (ITSs) and smart cities can be provided, the lack of fixed infrastructure, as well as an unpredictable number of nodes in ad hoc scenarios can lead to significant limitations. One of the possible solutions is to use FANETs that provide temporary connectivity in cases of low vehicle density or supplement the missing fixed infrastructure. This will lead to complex and heterogeneous environments that include both VANETs and FANETs to ensure adequate quality of service (QoS). The process of choosing the optimal route from source to destination is a challenging task in these networks since their topology is constantly changing, which can cause frequent link breaks and performance degradation. In these conditions, traditional routing techniques show significant limitations, especially for application in dynamic heterogeneous networks. One possible solution that attracts a lot of attention from researchers is the application of machine learning (ML). The most promising type of ML is reinforcement learning (RL), which monitors network changes through constant interaction with the environment and, depending on the current network state, helps in selecting the optimal route, especially in heterogeneous highly dynamic ad hoc networks.

There are several survey studies related to the application of RL in VANETs and FANETs in the literature, among which [1–3] can be singled out in terms of quality and importance. The authors in [1] gave an extensive overview of RL-based routing protocols in VANETs, where protocols are first categorised by routing type and then compared based on multiple criteria, such as key protocol features, optimisation criteria, performance evaluation parameters and techniques and RL algorithm parameters. In [2], the authors presented an overview of the different applications of RL in FANETs, including the application in routing protocols, where the protocols are compared according to RL type, their advantages and disadvantages. However, a detailed comparative analysis of the protocols has not been performed. The authors in [3] focused on the application of deep RL (DRL) in VANETs, but no categorisation and comparative analysis of the protocols are given. It can be noticed that the available surveys treat VANETs and FANETs separately. To have a more comprehensive view of the future application of RL in highly dynamic and heterogeneous networks for smart cities and ITSs, it is necessary to include both VANETs and FANETs in the analysis. Also, surveys are, unfortunately, quickly becoming obsolete, given a large number of new papers that are increasingly expanding the application of RL. Thus, several important RL-based protocols for VANETs and FANETs, proposed in recently published papers, are not included in the mentioned surveys. In addition, the protocols are not classified keeping in mind the very significant issue of RL integration with other techniques such as software-defined networking (SDN), blockchain etc. Therefore, this paper provides a comprehensive categorisation of recently published RL-based protocols for VANETs and FANETs, with special emphasis on the integration of RL with other techniques. The main goal of this survey is to present, in one place, different approaches to the application of RL-based routing in VANETs and FANETs. This can be very useful for researchers to see current developments in this area, determine the direction of their future research and gain new ideas for improving routing protocols using RL techniques in heterogeneous dynamic WANETs. Besides, this paper aims to point out the shortcomings and limitations of RL technology as well as to highlight the challenges that need to be resolved for its successful application.

The rest of this paper is organised as follows. In the second section, the basic principles of RL are explained. In the third section, routing protocols are categorised based on network type, RL type and possible application of some other technique, and a comparative analysis of protocols is performed. Current limitations, future trends and overall discussion are given in the fourth section. Concluding remarks are given in the last section.

2. REINFORCEMENT LEARNING

RL is the most common type of ML in routing protocols for dynamic WANETs. This type of learning is described in detail in [4] and involves learning through constant interaction with the environment to achieve a certain goal. The RL process in one WANET can be modelled in several ways. The most commonly used approach is that each node in the network that sends packets represents a learning agent, while the entire network represents the environment. Sending packets to one of the neighbouring nodes represents a potential action that the agent can take. Since each node has a finite set of neighbours, it represents a set of possible actions that the node can take. The feedback received by the sender contains a reward for the taken action and the new state of the environment. The reward may depend on various influencing factors, which are further discussed in the third section.

One of the simplest RL algorithms is Q-learning (QL) [4], in which each agent maintains a table of Q-values that refer to the usefulness of taking a specific action at a particular moment. Based on these values, the agent makes decisions about future actions. Q-values are updated after each action that an agent takes based on the current reward and the maximum possible Q-value that an agent can achieve in the following state. To improve the learning process, the DRL concept is introduced in [5], where the determination of Q-values is performed using a deep Q-network (DQN) that combines RL with a deep neural network (DNN). The input of DNN is typically the state of the environment, and the output is the optimal Q-value for the action taken in the appropriate state. RL is often unstable or even diverges when a neural network is used for the determination of the Q-values. To overcome these instabilities, two new ideas have been proposed in [5]. First, the experience replay mechanism is introduced, which stores the data collected in the memory from which the samples are randomly selected and used in the learning process, thus reducing correlations between data. Secondly, two DQNs are used, one to calculate action values, and the other to calculate the target values, thus reducing the correlation between them.

To further improve the performances and increase the stability of RL, in [6] the duelling DRL (DDRL) concept is proposed, which represents an improvement of the DRL algorithm, retaining the application of the experience replay mechanism and target DQN. This concept involves the usage of duelling deep Q-networks (DDQNs) to determine optimal Q-values. The basic idea of DDQN is that it is not always necessary to calculate the value of each available action. Therefore, the DDQN network architecture can be divided into two main components: the value function and the advantage function. The value function should represent how useful it is to be in a certain state, and the advantage function measures the relative importance of a particular action compared to other available actions. After a separate calculation, the results of these functions are combined to obtain a final Q-value.

Another type of RL used in routing protocols for dynamic WANETs is the SARSA [4] algorithm and its modification, SARSA-λ [7]. SARSA is very similar to the QL algorithm, except that Q-value is updated based on the current state of the agent, the action the agent chooses, the reward the agent gets for choosing this action, the next state that the agent enters after taking that action and, finally, the next action the agent chooses in its new state.

A characteristic of all mentioned algorithms is that they are not based on the model of the environment, i.e. they all belong to the group of model-free algorithms. In [8], the authors proposed a model-based RL (MBRL) algorithm that first needs to create an internal model of the environment, and, based on it, the optimal routing policy will be determined. In this way, the optimal policy is reached faster compared to the QL algorithm. However, with this approach, it is necessary to form a dynamic state transition model, and sometimes a reward model, before applying the algorithm itself.

3. RL-BASED ROUTING PROTOCOLS FOR VANETS AND FANETS

In this section, a categorisation of recently published papers in which RL is applied to improve routing protocols for highly dynamic WANETs is performed. The focus is on papers published since 2018, in order to include the most current research in this field. For many years, researchers have been publishing papers based on RL with applications only in MANETs. However, with the growing use of VANETs and FANETs, in the previous decade the authors have proposed RL-based routing solutions for VANETs, and in the last few years increasing number of RL-based algorithms for FANETs can be found as well. The great expansion of protocols for VANETs and FANETs and their wide application in smart cities and ITSs are the main reasons why the focus of this research is on routing protocols in these networks. *Table 1* shows the categorisation of these protocols based on the applied network type (VANET or FANET) and the applied RL type. Having in mind that in VANETs and FANETs RL can be often used in combination with some other technique, categorisation is done according to this criterion as well. Some protocols use blockchain and fuzzy logic (FL), while in several protocols the role of the decision-making agent in RL is played by the SDN controller. One representative of each category will be described in more detail.

3.1 RL-based routing protocols for VANETs

The first category in *Table 1* consists of papers in which QL-based routing protocols are proposed, without combining with any additional technique. The hybrid routing algorithm (RHR) [9], which helps to solve the blind path problem in VANETs, is chosen as a typical representative of this category. This problem occurs in a situation when a certain route in the routing table still has not expired, but due to the high mobility of nodes, the next node on the route has already gone out of the range of the sender. The RHR protocol finds multiple routes to the destination and runs the RL mechanism for each route in the forwarding table so that if a link on the route breaks, it selects a new one as soon as possible. The QL algorithm is implemented in every node so that different selections of the next-hop represent appropriate states while receiving different types of packets related to the current next-hop represents corresponding actions. For the action taken in the given state, the nodes receive feedback in the form of a reward, depending on the packet type. If a broadcast packet is received, the route through which the packet arrived will get a negative reward, while in the case of a unicast packet, that route will get a positive reward. After that, the nodes calculate the Q-values and choose the next hop. In [9] authors showed that

Table 1 – Categorisation of RL-based routing protocols

Cat.	Authors	Net. type	RL type	Other techniques		
				SDN	Blockchain	FL
1.	Ji et al. [9], Li et al. [10], Wu et al. [11], Zhang et al. [12], Roh et al. [13], Wu et al. [14], Wu et al. [15], Li et al. [16], Luo et al. [17], Yang et al. [18], Bouzid Smida et al. [19], Lolai et al. [20]	VANET	QL			
2.	Nahar and Das [21]	VANET	QL	√		
3.	Dai et al. [22]	VANET	QL		√	
4.	Jiang et al. [23], Wu et al. [24], Zhang et al. [25], Chang et al. [26]	VANET	QL			√
5.	Saravanan and Ganeshkumar [27], Ye et al. [28]	VANET	DRL			
6.	Zhang et al. [29], Yang et al. [30], Nahar and Das [31], Zhang et al. [32]	VANET	DRL	√		
7.	Zhang et al. [33]	VANET	DDRL	√		
8.	Zhang et al. [34]	VANET	DDRL	√	√	
9.	Bi et al. [7]	VANET	SARSA			
10.	Jafarzadeh et al. [8], Jafarzadeh et al. [35]	VANET	MBRL			√
11.	Li and Chen [36], Arafat and Moh [37], Zheng et al. [38], Mowla et al. [39], Sliwa et al. [40], Da Costa et al. [41], Liu et al. [42], Khan and Yau [43]	FANET	QL			
12.	Yang et al. [44]	FANET	QL			√
13.	Liu et al. [45], Ayub et al. [46]	FANET	DRL			
14.	He et al. [47]	FANET	DRL			√

network performances are improved in terms of packet delivery ratio (PDR), round trip time (RTT) and overhead (OH).

An adequate representative of the second category is the adaptive self-learning clustering algorithm with reinforcement routing in SDN-based VANETs (RL-SDVN) [21], which combines the application of the QL algorithm and SDN technique for clustering and finding the optimal route. The main goal of RL-SDVN is to improve the message dissemination process and reduce the average data transfer time. The first step in the proposed algorithm is the formation of clusters and assignment of vehicles to the appropriate cluster, based on connectivity with other vehicles, their distance, the transmission range of each vehicle and the number of packets in the queue for processing in a particular vehicle. Vehicles with high connectivity and low processing queue occupancy will be selected for the cluster head (CH) nodes. Based on the quality of the corresponding routes, the SDN controller, as an RL agent, searches for the best route to the destination. The learning process is repeated for each vehicle that has packets to forward until they reach their destination. In the RL process, vehicles in the network represent the states in which the agent can be, while sending packets from one vehicle to another is a possible action. When it receives the packet, the vehicle checks its Q-table and, if it knows the route to the destination, updates the table, forwards the packet and receives a positive reward. Otherwise, it drops the packet and receives a negative reward. The value of the reward is affected by the distance to the destination vehicle. The proposed algorithm increases the stability and lifetime of the clusters, and also improves network performance in terms of average delay and throughput (TH), as shown in [21].

The third category is characterised by the application of QL and blockchain techniques, and a representative of this category is the QLASS [22], which proposes a security framework for stimulating the cooperative behaviour of onboard units (OBUs) in VANETs to protect the network from potential attacks. The framework is tested on a network that consists of one roadside unit (RSU) and several OBUs. OBUs can help each other by following neighbouring OBUs requests, but can also be selfish and try to maximise their benefit by acting maliciously or may attack the network if it can obtain an illegal gain. OBUs learn coordination behaviour in the network by applying actions to other OBUs according to their reputations. Reputation is an important parameter shared between nodes in the network and protected using the blockchain mechanism. If an OBU does not participate in attacks, its reputation grows, and the probability that neighbouring OBUs will follow its requests will be higher. Every OBU uses QL to choose the optimal action to obtain maximum benefit. Actions can include jamming, spoofing, eavesdropping, disobeying and following the request, while the environment includes node reputation, location and speed. The authors in [22] showed that this approach has good performances in terms of PDR, reputation and utility of network nodes.

The fourth category in the *Table 1* consists of papers based on the application of QL and FL. An example of such a paper is the QL-based adaptive geographic routing approach (QAGR) [23], which requires the inclusion of unmanned aerial vehicles (UAVs) in the routing process. The routing scheme consists of the aerial and ground components. Within the aerial component, UAVs create a global route using the FL and depth-first-search [48] algorithms, to ensure that vehicles do not send packets in the wrong direction. The selection of the optimal global route is influenced by the average number of vehicles in a certain area and their average speed. The information about global route is sent by UAVs to the appropriate vehicle and is used as a filter to reject deviated and congested neighbours when choosing the next hop. Within the ground component, vehicles choose the optimal next hop based on QL, following the Q-table filtered by the global route. The QL is modelled so that each state consists of the geographical area of a particular vehicle, the distance from the vehicle to its neighbour, and the number of neighbours of the neighbouring vehicle. A learning agent can be any vehicle, and a possible set of actions that an agent can take includes sending packets to one of the neighbouring vehicles. The reward the agent receives for a particular action depends on the received signal strength (RSS), transmission distance and collision between vehicles. The selection of appropriate actions is made based on Q-values. QAGR improves end-to-end delay (E2ED), PDR and hop count (HC), as shown based on the simulations done in [23].

The fifth category includes papers based on DRL, without a combination with other techniques. One of the papers in this category is DRLV [27], in which DRL is used to establish and select the best routes in the VANET. The scenario for which the proposed model is created involves vehicle-to-infrastructure communication, where a particular RSU covers one area of the network. The entire network is divided into clusters so that each cluster has its vehicle density. Changes in vehicle density are predicted using the DRL model, trained based on vehicle speed and movement. The first phase in the proposed approach is establishing the routes using DRL, based on the location of the vehicles, the distance to the nearest RSU, vehicle density and the delay. Factors that can help in choosing the appropriate action at this stage are the capability of packet delivery along the route, the total number of routes that exist between the source and destination node and the cumulative weight of each route. The second phase is route selection, in which the nodes choose the best next-hop using the DRL. The learning agent first predicts possible transitions from one state to another based on previous events. In this way, the optimal routes for forwarding the packet to the destination are predicted. Based on that, the agent takes the appropriate action, which changes the state of the environment, and receives the appropriate reward. The reward depends on the ratio of the maximum link utilisation in the case of using the current routing strategy and the optimal link utilisation. The authors in [27] showed that this model improves PDR, E2ED and OH.

Software-defined trust-based DRL framework (TDRL-RP) [29] is the chosen representative of the sixth category in *Table 1*. TDRL-RP uses a combination of DRL and SDN techniques to help find the optimal route and calculate its reliability. In the proposed approach, the role of a learning agent in the DRL is played by a centralised SDN controller, which helps in selecting the best next hop. The state of the environment includes a set of states of all vehicles that include the position and forwarding ratio of each vehicle. A potential action in the appropriate state of the environment is the agent's choice of a neighbour to which a certain vehicle should forward packets. The reward for the action depends on the reliability of the vehicle, affected by the forwarding ratios of control and data packets. DRL uses a convolutional neural network whose input is the state of the environment, and the output is the corresponding Q-value, based on which the agent selects the optimal route. Applying the proposed approach improves PDR and TH, as shown in [29].

The seventh category includes [33], which combines DDRL and SDN techniques to find the optimal route for data transmission. This algorithm is similar to the one proposed in [29], with the difference that it uses DDRL to train a learning agent. The neural network used to calculate the Q-values is divided into two streams, the first for calculating the value function, and the second for calculating the advantage function. These two functions represent the two components of the Q-value in this algorithm. The first component indicates the value of the corresponding state, and the second is the additional value achieved by taking a certain action in a given state. In [33] is shown that the proposed approach improves TH and E2ED.

A representative of the eighth category is a blockchain-based distributed software-defined VANET framework (block-SDV) [34] that combines the application of DDRL, SDN and blockchain techniques to establish a reliable architecture for communication management in VANETs. Block-SDV consists of three layers: device (DL), area control (ACL), domain control (DCL) and an edge computing server. The DL is formed of vehicles, while the ACL consists of SDN controllers that collect information about vehicles and links between them. Collected information is sent to the DCL, formed of SDN controllers that work in a distributed blockchain manner. The DCL is connected to the blockchain system, consisting of several blockchain nodes, among which there is one primary node that is responsible for client requests and several consensus nodes that control other nodes. Each SDN controller on the DCL represents a learning agent. The state of the environment depends on the trust features of the vehicles and the nodes in the blockchain system, the computing resources of the edge computing server, as well as the number of consensus nodes in the blockchain system. The set of actions taken by the agent includes the choice of the primary blockchain node, the edge computing server as a computing resource, the number of consensus nodes and reliable neighbouring vehicles for forwarding packets. After taking action, the agent receives a reward that depends on the network throughput and throughput of the blockchain system. Based on the reward, the agent computes Q-value using DDRL with prioritised experience replay. Block-SDV increases the TH in the VANETs, as shown in [34].

A representative of the ninth category is an RL-based routing protocol for clustered EV-VANET (RLRC) [7], which uses the SARSA-λ learning algorithm. In the proposed approach, the entire network represents an environment, divided into an appropriate number of clusters. Each cluster has a CH node, and the learning process is started only for these nodes. To be selected for CH the vehicle must have available bandwidth (BW) and residual power above a predefined threshold. The vehicle that has packets for another vehicle sends those packets to its CH, its CH forwards them to the neighbouring CH using the SARSA-λ algorithm, and the neighbouring CH forwards the packets to the destination vehicle. The learning agent can be any CH node, and the set of states for a particular agent is the set of all other CHs in the network. The action that the agent can take is the selection of the appropriate CH to forward the packets. The reward for the action will have the maximum value if the current node is a neighbour of the destination node, and the minimum value if the current node does not have the next hop. In other situations, the reward depends on the HC, the link utility and the available BW. CHs periodically exchange Hello packets to update Q-values. The authors showed in [7] that applying the proposed protocol increases PDR and decreases HC.

The tenth category of papers is characterised by the application of MBRL and FL in routing protocols, and the appropriate representative is the reinforcement routing protocol for VANETs (RRPV) [8]. RRPV is based on the multi-agent RL (MARL) technique, which means that all nodes in the network represent learning agents that cooperate and at the same time try to find the optimal routing policy. The RRPV protocol consists of model learning and RL, which operate simultaneously. The FL system is used for learning and creating a model of the environment. The main goal is to create a state transition model and a reward model based on network quality, affected by connection stability (which depends on the speed and direction of nodes) and connection quality (which depends on the ratio of sent and received control packets). The optimal routing policy is determined based on the created model of the environment, with the help of RL. Within RL, each node that has packets to send represents a learning agent that can change the state of the environment by taking a certain action. Sending packets to the agent's neighbours represents a set of available actions. When receiving a particular packet, the node evaluates links to all of its neighbours based on a previously created model of the environment, then calculates Q-values and selects the appropriate action based on the routing policy. For the taken action, the agent receives a reward that depends on the distance and quality of links between nodes (determined in the model learning process). Based on the simulations done in [8], this protocol improves PDR, E2ED and OH.

3.2 RL-based routing protocols for FANETs

Papers that propose routing protocols for FANETs based on QL, without the application of other techniques, are classified in the eleventh category in *Table 1*. A representative of this category is the QL-based message prioritising and scheduling algorithm (QMPS) [36], in which messages exchanged in the network are first classified into delay-sensitive and delay-tolerant. This is done so that in case of network congestion or degradation of link quality (LQ) delay-sensitive messages have a higher priority. Delay-sensitive messages include various types of command and coordination messages that have strict delay requirements and whose timely transmission greatly affects the reliability and security of the network. Delay-tolerant messages include various messages that can tolerate increased delay and packet loss. The QL algorithm has the role of dynamically assigning different priorities to different message types. Each node in the network is a learning agent, which takes a certain action in the form of assigning the appropriate priority for sending delay-tolerant messages. The reward for the action is formed based on two metrics: the first, which represents the percentage of delay-sensitive messages in the message queue, and the second, which depends on the probability of successful reception of the message of the neighbouring node. As shown in [36], the QMPS algorithm improves E2ED, TH and PLR of delay-sensitive messages.

A representative of the twelfth category is a routing protocol based on QL and FL (QL-FLRP) proposed in [44]. Determination of the optimal route is done with the help of link-related parameters, which refer to an individual link, and path-related parameters, which refer to the entire route from the source to the destination. Link-related parameters include transmission rate (TR), energy state and flight status (depending on the speed and direction of the node), while path-related parameters include hop count and successful packet delivery time (SPDT). The FL system first finds the route to the destination based on the link-related parameters, after which it is possible to determine the path-related parameters. The QL algorithm calculates Q-values for path-related parameters and sends them back to the sender node. All collected parameters on the entire route represent the environment in the QL; each node that has packets to send represents an agent that changes the state by taking a certain action (selects the next node). Rewards, which affect the calculation of Q-values, are influenced by hop count and SPDT. Finally, based on both types of parameters, the optimal route is determined, using the FL system. The proposed protocol improves TR, HC and the remaining energy of nodes in the network, which is proved by the simulations done in [44].

The thirteenth category is characterised by the use of DRL in the routing protocol, and the representative of this category is the DRL-based adaptive and reliable routing protocol (ARdeep) [45]. In ARdeep the environment consists of all network nodes, and each node that has packets to send is a learning agent. For the learning agent, the state of the environment is represented by the status of all links to its neighbours. The status of each link is formed based on the expected connection time of the link, packet error rate (PER), remaining neighbour energy, the distance between neighbour and destination, and minimum distance between a two-hop neighbour and destination. The action that an agent can take is to select one of the neighbouring nodes to forward the packets. Each neighbouring node is detected by periodically sending Hello messages, which contain information about its position, speed and remaining energy. Based on the state of the environment, the agent selects the appropriate action with the help of DQN, whose input is the status of the appropriate link, and the output is its Q-value. After calculating the

Q-value, the agent forwards the packet to the neighbour with the highest Q-value. The reward that an agent receives has the maximum value if the neighbouring node is the destination and the minimum value if all neighbours of the forwarding node are further away from the destination. In other situations, the reward depends on the distance to the destination node, LQ, remaining energy and initial energy of the neighbour. The authors in [45] showed that ARdeep improves PDR and E2ED.

A representative of the last category from *Table 1* is FLRL [47], which uses FL and DRL for determining the optimal route in FANET. The FL system aims to determine the best relay node for packet forwarding, based on delay measure (depends on the distance to the relay node), stability rating (depends on the speed of the current and neighbouring nodes), and bandwidth efficiency (depends on the total number of nodes involved in the communication). In this way, it is possible to find a route to a destination with the help of FL, but this route may not be the best. Therefore, in addition to FL, DRL is also used. In the DRL algorithm, each node represents a learning agent, and the state of neighbouring nodes is known based on the FL. The action that an agent can take is to send packets to one of the neighbours and it consequently receives the appropriate reward. Based on FL, the reward will be 0 if the neighbour is best (optimal), and -1 if the neighbour is sub-optimal. Moreover, the reward will have a minimum value if it is not possible to establish a link to a neighbour, and a maximum value if it is a destination. It is then possible to calculate Q-values, based on which the optimal relay node is selected. In this way, hop count and connection quality are included in the route selection. This algorithm improves link connectivity and HC, as shown in [47].

3.3 Comparative analysis of RL-based routing protocols for VANETs and FANETs

The analysis of the previously described RL-based protocols shows that their success mostly depends on the appropriate design of the reward function. Therefore, a comparison of RL-based routing protocols for VANETs (*Table 2*) and FANETs (*Table 3*) is based on the influencing factors that determine the reward function. Furthermore, the comparison is done by the simulation software and the obtained network performance metrics. Various influencing factors are used in different studies, depending on the basic optimisation goal of the routing process. Some of the most common factors are the link reliability (LR) and LQ to the potential next hop, the number of hops required to deliver the packet to the destination, available BW, achieved TH, delay, node speed, distance to the destination etc. It is often very important whether the next node is also the destination, as well as if the next node knows the route to the destination. When the goal of the protocol is to optimise energy consumption (EC), energy loss will be an important influencing factor. On the other hand, if the emphasis is on protection against unwanted external interference, important factors will be the reputation of the next node on the route and the detection of jammers near that node. Performance evaluation of the proposed protocols is done using different simulation environments, and some of the most common are network simulator 3 (ns3), network simulator 2 (ns2), optimised network engineering tools (opnet), python, qualnet, matlab, objective modular network test-bed in C++ (OMNeT ++), TensorFlow (TF) etc. Depending on the optimisation goal, different network performance metrics are used in the simulations, such as PDR, PLR, E2ED, TH, BW, HC, OH etc. Energy consumption and link connectivity (LC) are particularly important metrics when evaluating network performances in FANETs.

4. DISCUSSION

Following the development of modern cities and ITSs with high security and QoS requirements, we believe that future solutions will largely rely on heterogeneous dynamic WANETs that include fixed and ad hoc architecture with the addition of blockchain, SDN and other technologies. By analysing the literature from this survey, it can be seen that the emerging RL-based routing can achieve better network performances than traditional algorithms in both VANETs and FANETs and provide prosperous integration with other technologies. With RL, important changes in the network can be detected in real-time, which makes this technology very suitable for use in complex highly dynamic heterogeneous networks. However, RL is a new and complex technique that should be applied adequately in order to exploit its potentially very large benefits. This technology is still the subject of intensive research, and there are many open questions and limitations to overcome. One of the dilemmas that can be observed is the selection of the appropriate

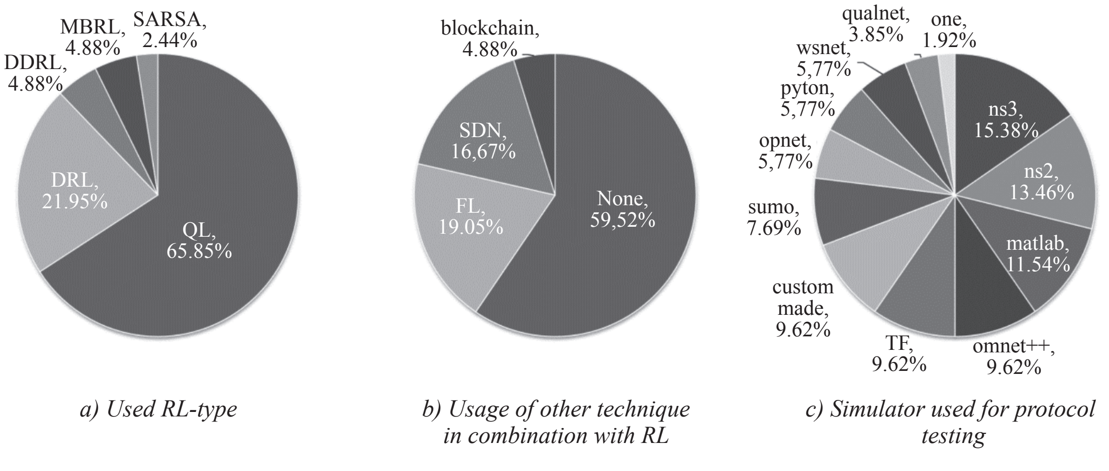
Table 2 – Comparative analysis of RL-based routing protocols in VANETs

Authors	Protocol	Influencing factors on the reward	Performance metrics	Simulator
Ji et al. [9]	RHR	type of control packets	PDR, RTT, OH	ns3
Li et al. [10]	QGrid	if the message is delivered to the dest. grid	PDR, HC, delay, num. of forwarding, TH	custom-made (CM)
Wu et al. [11]	DTNP	direct connec. or HC and the elapsed time since the last connec.	delay, PDR	one
Zhang et al. [12]	RSAR	HC, LR, BW	PDR, E2ED, average route length, OH	ns2
Roh et al. [13]	Q-LBR	UAV relay node load, ground net. congestion	PDR, net. utilization, delay	opnet
Wu et al. [14]	ARPRL	if the control packet arrived from the sender	PDR, E2ED, HC, OH	qualnet
Wu et al. [15]	QTAR	LQ, link expiration time, delay	PDR, E2ED	qualnet
Li et al. [16]	ECTS	if charging data arrive in dest.	communication cost, connec. prob., PDR, OH	ns3
Luo et al. [17]	IV2XQ	if the packet is forwarded to the dest.	PDR, E2ED, HC, OH	sumo, veins, omnet++
Yang et al. [18]	HAEQR	if the current node belongs to a set of a one-hop neigh. of the dest.	PDR, E2ED, HC	sumo, ns2
Bouzid Smida et al. [19]	LEQRV	link lifetime, LQ, dist. to dest., mean opinion score (MOS) [49], num. of neigh., free-buffer level	MOS, peak signal-to-noise, structural similarity, E2ED, frame loss	sumo, ns3
Lolai et al. [20]	RRIN	vehicle speed difference, vehicle direction, the num. of data packets in the queue, signal fading, LR	PDR, PLR, delay, TH	matlab
Nahar et al. [21]	RL-SDVN	dist. from the dest. vehicle	delay, TH	ns3
Dai et al. [22]	QLASS	reputation gain, the payoff of node action	PDR, reputation, utility	CM
Jiang et al. [23]	QAGR	RSS, transmission dist., the collision between vehicles	PDR, E2ED, HC	ns3
Wu et al. [24]	V2R-CBR	if the observed node is a one-hop neigh., HC, payoff, LQ	PDR, num. of collided MAC frames, E2ED, TH	ns2
Zhang et al. [25]	FLHQRP	if the current cluster belongs to a set of adjacency clusters of the dest. cluster, traffic density in the cluster	PDR, E2ED, HC, OH	ns2
Chang et al. [26]	CEVCS	if the observed node is a one-hop neigh., HC, LQ	PDR, TH	ns2
Saravanan et al. [27]	DRLV	max. link utilization under the future routing strategy, optimal link utilization	PDR, E2ED, OH	ns2
Ye et al. [28]	VMDRL	energy loss, TR	EC, PLR, transmission time, prob. of communication interruption	CM
Zhang et al. [29]	TDRL-RP	trust information	PDR, TH	TF, opnet
Yang et al. [30]	VDDS	HC, LQ	TH, num. of gateway cluster heads	CM
Nahar et al. [31]	SeScR	quality of available routes, vehicles speed, location	cluster stability, lifetime, alienation time, delay, TH, computation delay	sumo, omnet++
Zhang et al. [32]	SD-TDQL	trust value of each vehicle, reverse delivery ratio	PLR, delay	matlab,TF
Zhang et al. [33]	T-DDRL	trust information	TH, E2ED	TF, opnet
Zhang et al. [34]	block-SDV	TH	TH	TF, phyton
Bi et al. [7]	RLRC	if a current node is a neigh. of the dest., HC, link utility, BW	PDR, HC	python
Jafarzadeh et al. [8], [35]	RRPV	LQ, dist. from neigh. to dest.	PDR, delay, OH	omnet++

Table 3 – Comparative analysis of RL-based routing protocols in FANETs

Authors	Protocol	Influencing factors on the reward	Performance metrics	Simulator
Li et al. [36]	QMPS	proportion of delay-sensitive messages, prob. of successfully receiving a message from a neigh.	E2ED, TH, PLR	ns3
Arafat et al. [37]	QTAR	next-hop node type, E2ED, node velocity, EC	PDR, E2ED, EC, net. lifetime, OH	matlab
Zheng et al. [38]	RLSRP	the conditional prob. of success or failure of transmitting a packet to the next-hop	indicators of route setup success rate, average route lifetime, HC, PDR, TH without retransmissions, delay	matlab, ns2
Mowla et al. [39]	AFRL	if a jammer has been detected	accuracy, success rate, HC, num. of iterations to convergence, cumulative reward	ns3
Sliwa et al. [40]	PARRoT	link expiry time, change in the neigh. set of the forwarding node	PDR, E2ED	omnet++, inetmanet
Da Costa et al. [41]	Q-FANET	if the link leads to the dest., if it is a local min.	max. E2ED, jitter, PDR	wsnet
Liu et al. [42]	QMR	if the link leads to the dest., if it is a local min., E2ED, EC	E2ED, packet arrival ratio, EC	wsnet
Khan et al. [43]	RL-FANET	successful transmission of the packet	EC, num. of link breaks, net. lifetime	matlab
Yang et al. [44]	QL-FLRP	HC, SPDT	HC, remain node energy, TR	CM
Liu et al. [45]	ARdeep	if the link leads to the dest., if it is a local min., dist. to dest., LQ, remaining and initial energy of neigh.	PDR, E2ED	TF, phyton, wsnet
Ayub et al. [46]	AI-Hello	transmission range, allowed airspace, num. of UAVs, speed ranges	EC, OH, PDR, TH, E2ED	ns3
He et al. [47]	FLRL	optimality of neigh. node, LC	HC, LC	matlab

RL type for the given routing problem. By analysing the latest literature (*Figure 1a*), it can be seen that most of the authors (65.85%) use QL, 21.95% use DRL, both DDRL and MBRL use 4.88% of them, while the SARSA algorithm is applied in only one protocol (2.44%). In addition, authors still search for the optimal definition of the learning agent, its states and actions. When the network is centralised, the most common approach is to choose this central device as the learning agent, while all network vehicles or UAVs form the environment. In the distributed ad hoc networks, the common solution is that all nodes are used as agents, while in the cluster-based routing algorithms CH usually takes a role of the agent. In order to further improve network performance, RL can be used in combination with some



a) Used RL-type

b) Usage of other technique in combination with RL

c) Simulator used for protocol testing

Figure 1 – Distribution of RL-based routing protocols

other technique, such as FL, SDN and blockchain, but most of the authors still do not use this possibility (*Figure 1b*).

Having in mind that the QL technique is relatively simple, and that has a table approach in the algorithm implementation, it is suitable for relatively small ad hoc networks, so most of the routing protocols analysed in this survey are limited to application in this network type. Since in these networks the learning algorithm is distributed among all nodes, which already have routing tables, storing data in Q-tables is a straightforward extension. But this approach is not an adequate solution for complex networks with a large number of nodes because the action-value space will grow exponentially. In those cases, DRL or some method for Q-table limitation should be used. Implementation of the DRL algorithm needs high computation resources and challenging convergence time, so it is more suitable for networks with centralised entities such as SDN or cluster-based networks with RSUs. Practical application of those techniques must carefully consider the security aspect as well. Although a centralised approach is a very good solution, in recent studies the authors are considering the integration of blockchain technology that provides a distributed trust management system. Currently, fewer authors use DRL, especially if it includes some other technique, but the number of DRL-based protocols constantly increases.

Besides the most important issue of selecting RL type, different approaches to defining the reward can be found (*Tables 2 and 3*), which obviously depend on the parameters that need to be optimised. When forming the reward, the agent relies on various feedback mechanisms that typically involve the exchange of additional control packets to determine the LQ or similar QoS parameters, which increases the routing overhead. Unfortunately, this cannot be avoided, but it is necessary to consider the possibility of using hierarchical routing that limits the area for the exchange of control packets, thus reducing the routing overhead.

One of the major challenges in RL applications is the convergence of the learning algorithm. The learning process is influenced by two key parameters: learning rate α and the discount factor γ, which determines the importance of future rewards. It is very important to carefully choose optimal values of these parameters to provide for proper functioning of the learning process and timely adaptation to changes in the environment. Too fast convergence can lead to instability and frequent changes in the selected routes, while too long convergence time leads to selection of sub-optimal routes. Another important factor of the learning process that influences the choice of the optimal route is the balance between the exploitation of acquired knowledge and the exploration of the environment due to its frequent changes. The most commonly used action selection policy is ε-greedy in which an agent with probability ε takes the action with the highest Q-value, while with probability (1-ε) selects a random action to explore the environment. Unfortunately, in most papers, not enough attention is paid to the optimal choice of parameters α, γ and ε; instead, typical values are adopted based on previous positive experience in other fields of application.

Another important aspect in proposing new protocols is the process of their evaluation. Certainly, the best method of protocol validation is test-bed experiments that use a real-life setup for data collection. However, none of the analysed papers used this approach, instead, various simulation environments were used to evaluate the results. As can be seen from *Figure 1c*, most authors use open-source simulators or create their own simulation environment.

5. CONCLUSION

In this paper, an overview and classification of the RL-based routing protocols for VANETs and FANETs published since 2018 are provided. The protocols are classified into several categories based on network type, RL type and combination of RL with some other techniques. One chosen protocol from each category is explained in more detail. A comparative analysis of routing protocols is also given based on influential factors that determine the value of the reward in RL and network performance metrics used in simulations. However, a few limitations had to be adapted. Considering the current trends in this area, our classification is limited to the last couple of years, bearing in mind that the number of research papers is increasing every year. MANETs are not included in this survey, but considering the extensive experience in the application of RL-based techniques in this type of networks, they will certainly be the subject of our future research. In addition, although RL dominates in routing applications, there are certain possibilities of applying supervised and unsupervised techniques that are also not covered. Having in mind that in future ITSs and smart cities implementation of both VANETs and FANETs will be necessary, we wanted to give a comprehensive survey of the current state of the art for RL-based routing in both networks in one place, which should be a useful research base ground. In addition, in this survey, papers that include both RL and other key techniques such as blockchain, SDN and FL for routing in VANETs and FANETs are analysed. In all of the analysed papers, the authors reported significant improvement in the observed performances, compared to the performances achieved using traditional routing protocols. Based on this state-of-the-art research we can conclude that the application of RL in routing protocols yields very good results for networks with high-speed nodes and frequent topology changes. Therefore, it can be expected that in the following years more RL-based routing solutions will emerge, especially based on DRL and in applications that use both VANETs and FANETs. Based on this research, we plan to propose a new solution for an RL-based routing protocol that will provide easier integration of VANETs and FANETs into highly dynamic heterogeneous networks. We are certain that this paper will serve as a good starting point for other researchers as well in the field of RL application in networks that include both VANETs and FANETs.

Pavle BUGARČIĆ, student DAS[1]
E-mail: p.bugarcic@sf.bg.ac.rs
Prof. dr **Nenad JEVTIĆ**[1]
E-mail: n.jevtic@sf.bg.ac.rs
Prof. dr **Marija MALNAR**[1]
E-mail: m.malnar@sf.bg.ac.rs
[1] Univerzitet u Beogradu, Saobraćajni fakultet
Vojvode Stepe 305, 11000 Beograd, Srbija

PROTOKOLI RUTIRANJA BAZIRANI NA UČENJU POTKREPLJIVANJEM ZA BEŽIČNE AD HOC MREŽE ZA VOZILA I BESPILOTNE LETELICE – PREGLED LITERATURE

REZIME

Sa razvojem pametnih gradova i inteligentnih transportnih sistema (ITS), bežične ad hoc mreže za vozila i bespilotne letelice (VANET i FANET) postaju sve značajnije. Velika mobilnost čvorova u ovim mrežama dovodi do čestih prekida linkova, što komplikuje otkrivanje optimalne putanje od izvora do odredišta i degradira mrežne performanse. Jedan od načina da se prevaziđe ovaj problem je korišćenje mašinskog učenja (ML) u procesu rutiranja, a među različitim tipovima ML, najviše obećava učenje potkrepljivanjem (RL). Iako postoji nekoliko istraživanja o protokolima rutiranja na bazi RL za VANET i FANET mreže, važno pitanje integracije RL sa značajnim modernim tehnologijama, kao što su softverski definisano umrežavanje (SDN) ili blockchain, nije na odgovarajući način obrađeno, posebno kada se koristi u kompleksnim ITS. U ovom radu fokusirali smo se na izvođenje sveobuhvatne kategorizacije protokola rutiranja baziranih na RL za oba tipa mreže, imajući u vidu njihovu istovremenu upotrebu i inkluziju sa drugim tehnologijama. Sprovedena je detaljna komparativna analiza protokola na osnovu različitih faktora koji utiču na funkciju nagrade kod RL i posledica koje one imaju na performanse mreže. Takođe, detaljno su razmotrene ključne prednosti i ograničenja rutiranja baziranog na RL.

KLJUČNE REČI

učenje potkrepljivanjem; Q-učenje; protokoli rutiranja; VANET; FANET; ITS.

REFERENCES

[1] Nagib RA, Moh S. Reinforcement learning-based routing protocols for vehicular ad doc networks: A comparative survey. *IEEE Access*. 2021;9: 27552-27587. doi: 10.1109/ACCESS.2021.3058388.

[2] Rezwan S, Choi W. A survey on applications of reinforcement learning in flying ad-hoc networks. *Electronics*. 2021;10(4): 449. doi: 10.3390/electronics10040449.

[3] Doddalinganavar SS, Tergundi PV, Patil SR. Survey on deep reinforcement learning protocol in VANET. *Proc. of the 1st Int. Conf. on Advances in Information Technology, ICAIT, 25-27 July 2019, Chikmagalur, India.* IEEE; 2019. p. 81-86.

[4] Sutton R, Barto A. *Reinforcement learning: An introduction, second edition.* Cambridge, Massachusetts: MIT Press; 2018.

[5] Mnih V, et al. Human level control through deep reinforcement learning. *Nature*. 2015;518(7540): 529-533. doi: 10.1038/nature14236.

[6] Wang Z, et al. Dueling network architectures for deep reinforcement learning. *Proc. of the 33rd Int. Conf. on Machine Learning, 20-22 June 2016, New York, NY, USA.* PMLR; 2016. p. 1995-2003.

[7] Bi X, Gao D, Yang M. A reinforcement learning-based routing protocol for clustered EV-VANET". *Proc. of the 5th Information Technology and Mechatronics Engineering Conf, ITOEC, 12-14 June 2020, Chongqing, China.* IEEE; 2020. p. 1769-1773.

[8] Jafarzadeh O, Dehghan M, Sargolzaey H, Esnaashari MM. A novel protocol for routing in vehicular ad hoc network based on model-based reinforcement learning and fuzzy logic. *Int. Journal of Information and Communication Technology Research.* 2020;12(4): 10-25.

[9] Ji X, et al. Keep forwarding path freshest in VANET via applying reinforcement learning. *Proc. of the 1st Int. Workshop on Network Meets Intelligent Computations, NMIC, 7-9 July 2019, Dallas, TX, USA.* IEEE; 2019. p. 13-18.

[10] Li F, et al. Hierarchical routing for vehicular ad hoc networks via reinforcement learning. *IEEE Transactions on Vehicular Technology.* 2019;68(2): 1852-1865. doi: 10.1109/TVT.2018.2887282.

[11] Wu C, Yoshinaga T, Bayar D, Ji Y. Learning for adaptive anycast in vehicular delay tolerant networks. *Journal of Ambient Intelligence and Humanized Computing*. 2019;10(4): 1379-1388. doi: 10.1007/s12652-018-0819-y.

[12] Zhang D, Zhang T, Liu X. Novel self-adaptive routing service algorithm for application in VANET. *Applied Intelligence*. 2019;49(5): 1866-1879. doi: 10.1007/s10489-018-1368-y.

[13] Roh BS, Han MH, Ham JH, Kim KI. Q-LBR: Q-learning based load balancing routing for UAV-assisted VANET. *Sensors*. 2020;20(19): 1-17. doi: 10.3390/s20195685.

[14] Wu J, Fang M, Li X. Reinforcement learning based mobility adaptive routing for vehicular ad-hoc networks. *Wireless Personal Communications*. 2018;101(4): 2143-2171. doi: 10.1007/s11277-018-5809-z.

[15] Wu J, Fang M, Li H, Li X. RSU-assisted traffic-aware routing based on reinforcement learning for urban VANETs. *IEEE Access*. 2020;8: 5733-5748. doi: 10.1109/ACCESS.2020.2963850.

[16] Li G, et al. An efficient reinforcement learning based charging data delivery scheme in VANET-enhanced smart grid. *Proc. of the Int. Conf. on Big Data and Smart Computing, BIGCOMP, 19-22 Feb. 2020, Busan, South Korea.* IEEE; 2020. p. 263-270.

[17] Luo L, Sheng L, Yu H, Sun G. Intersection-based V2X routing via reinforcement learning in vehicular ad hoc networks. *IEEE Transactions on Intelligent Transportation Systems*. 2021;1-14. doi: 10.1109/TITS.2021.3053958.

[18] Yang XY, Zhang WL, Lu HM, Zhao L. V2V routing in VANET based on heuristic Q-learning. *Int. Journal of Computers, Communications and Control*. 2020;15(5): 1-17. doi: 10.15837/ijccc.2020.5.3928.

[19] Bouzid Smida E, Gaied Fantar S, Youssef H. Link efficiency and quality of experience aware routing protocol to improve video streaming in urban VANETs. *Int. Journal of Communication Systems.* 2019;33(3): e4209. doi: 10.1002/dac.4209.

[20] Lolai A, et al. Reinforcement learning based on routing with infrastructure nodes for data dissemination in vehicular networks. *Wireless Networks*. 2022;28: 2169-2184. doi: 10.1007/s11276-022-02926-w.

[21] Nahar A, Das D. Adaptive reinforcement routing in software defined vehicular networks. *Proc. of the Int. Wireless Communications and Mobile Computing, IWCMC, 15-19 June 2020, Limassol, Cyprus.* IEEE; 2020. p. 2118–2123.

[22] Dai C, et al. Learning based security for VANET with blockchain. *Proc. of the Int. Conf. on Comm. Systems, ICCS, 19-21 Dec. 2018, Chengdu, China.* IEEE; 2018. p. 210-215.

[23] Jiang S, Huang Z, Ji Y. Adaptive UAV-assisted geographic routing with Q-learning in VANET. *IEEE Communications Letters*. 2021;25(4): 1358-1362. doi: 10.1109/LCOMM.2020.3048250.

[24] Wu C, Yoshinaga T, Ji Y, Zhang Y. Computational intelligence inspired data delivery for vehicle-to-roadside communications. *IEEE Transactions on Vehicular Technology*. 2018;67(12): 12038-12048. doi: 10.1109/TVT.2018.2871606.

[25] Zhang WL, Yang XY, Song QX, Zhao L. V2V routing in VANET based on fuzzy logic and reinforcement learning. *Int. Journal of Computers, Communications & Control.* 2021;16(1): 1-19. doi: 10.15837/ijccc.2021.1.4123.

[26] Chang A, et al. A context-aware edge-based VANET communication scheme for ITS. *Sensors.* 2018;18(7). doi: 10.3390/s18072022.

[27] Saravanan M, Ganeshkumar P. Routing using reinforcement learning in vehicular ad hoc networks. *Computational Intelligence.* 2020;36(2): 682-697. doi: 10.1111/coin.12261.

[28] Ye S, Xu L, Li X. Vehicle-mounted self-organizing network routing algorithm based on deep reinforcement learning. *Wireless Communications and Mobile Computing*. 2021;2021. doi: 10.1155/2021/9934585.

[29] Zhang D, Yu FR, Yang R. A machine learning approach for software-defined vehicular ad hoc networks with trust management. *Proc. of the Global Communications Conf, GLOBECOM, 9-13 Dec. 2018, Abu Dhabi, United Arab Emirates*. IEEE; 2018. p 1-6.

[30] Yang Y, Zhao R, Wei X. Research on data distribution for VANET based on deep reinforcement learning. *Proc. of the Int. Conf. on Artificial Intelligence and Advanced Manufacturing, AIAM, 16-18 Oct. 2019, Dublin, Ireland*. IEEE; 2019. p. 484-487.

[31] Nahar A, Das D. SeScR: SDN-enabled spectral clustering-based optimized routing using deep learning in VANET environment. *Proc. of the 19t^{h} Int. Symp. on Network Computing and Applications, NCA, 24-27 Nov. 2020, Cambridge, MA, USA*. IEEE; 2020. p. 1-9.

[32] Zhang D, Yu FR, Yang R, Zhu L. Software-defined vehicular networks with trust management: A deep reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems.* 2020;23(2): 1400-1414. doi: 10.1109/TITS.2020.3025684.

[33] Zhang D, Yu FR, Yang R, Tang H. A deep reinforcement learning-based trust management scheme for software-defined vehicular networks. *MSWIM '18: Proc. of the 8th ACM Symp. on Design and Analysis of Intelligent Vehicular Networks and Applications, DIVANet 2018, 28. Oct. - 2. Nov. 2018, Montreal, Canada*. New York: Association for Computing Machinery; 2018. p. 1-7.

[34] Zhang D, Yu FR, Yang R. Blockchain-based distributed software-defined vehicular networks: A dueling deep Q-learning approach. *IEEE Transactions on Cognitive Communications and Networking.* 2019;5(4): 1086-1100. doi: 10.1109/TCCN.2019.2944399.

[35] Jafarzadeh O, Dehghan M, Sargolzaey H, Esnaashari MM. A model based reinforcement learning protocol for routing in vehicular ad hoc network. *Wireless Personal Communications*. 2021;123: 975-1001. doi: https://doi.org/10.1007/s11277-021-09166-9.

[36] Li J, Chen M. QMPS: Q-learning based message prioritizing and scheduling algorithm for flying ad hoc networks. *Proc. of the Int. Conf. on Networking and Network Applications, NaNA, 10-13 Dec. 2020, Haikou City, China.* IEEE; 2020. p. 265-270.

[37] Arafat MY, Moh S. A Q-learning-based topology-aware

routing protocol for flying ad hoc networks. *IEEE Internet of Things Journal*. 2021;9(3): 1985-2000. doi: 10.1109/JIOT.2021.3089759.

[38] Zheng Z, Sangaiah AK, Wang T. Adaptive communication protocols in flying ad hoc network. *IEEE Communications Magazine.* 2018;56(1): 136-142. doi: 10.1109/MCOM.2017.1700323.

[39] Mowla NI, Tran NH, Doh I, Chae K. AFRL: Adaptive federated reinforcement learning for intelligent jamming defense in FANET. *Journal of Communications and Networks*. 2020;22(3): 244-258. doi: 10.1109/JCN.2020.000015.

[40] Sliwa B, Schüler C, Patchou M, Wietfeld C. PARRoT: Predictive ad-hoc routing fueled by reinforcement learning and trajectory knowledge. *Proc. of the 93rd Vehicular Technology Conf, VTC2021-Spring, 25-28 Apr. 2021, Helsinki, Finland.* IEEE; 2021. p. 1-7.

[41] Da Costa LALF, Kunst R, De Freitas EP. Q-FANET: Improved Q-learning based routing protocol for FANETs. *Computer Networks*. 2021;198. doi: 10.1016/j.comnet.2021.108379.

[42] Liu J, et al. QMR: Q-learning based multi-objective optimization routing protocol for flying ad hoc networks. *Computer Communications*. 2020;150: 304-316. doi: 10.1016/j.comcom.2019.11.011.

[43] Khan M, Yau KL. Route selection in 5G-based flying adhoc networks using reinforcement learning. *Proc. of the 10th Int. Conf. on Control System, Computing and Engineering, ICCSCE, 21-22 Aug. 2020, Penang, Malaysia.* IEEE; 2020. p. 23-28.

[44] Yang Q, Jang SJ, Yoo SJ. Q-learning-based fuzzy logic for multi-objective routing algorithm in flying ad hoc networks. *Wireless Personal Communications*. 2020;113: 115-138. doi: 10.1007/s11277-020-07181-w.

[45] Liu J, Wang Q, He C, Hu Y. ARdeep: Adaptive and reliable routing protocol for mobile robotic networks with deep reinforcement learning. *Proc. of the 45th Conf. on Local Computer Networks, LCN, 16-19 Nov. 2020, Sydney, Australia.* IEEE; 2020. p. 465-468.

[46] Ayub MS, et al. Intelligent hello dissemination model for FANET routing protocols. *IEEE Access.* 2022;10: 46513-46525. doi: 10.1109/ACCESS.2022.3170066.

[47] He C, Liu S, Han S. A fuzzy logic reinforcement learning-based routing algorithm for flying ad hoc networks. *Proc. of the Int. Conf. on Computing, Networking and Communications, ICNC, 17-20 Feb. 2020, Big Island, HI, USA.* IEEE; 2020. p. 987-991.

[48] Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*. 3rd edition. Beijing: China Machine Press; 2009.

[49] Rodriguez-Bocca P. *Quality-centric design of peer-to-peer systems for live-video broadcasting.* PhD thesis. Facultad de Ingeniería, Universidad de la República Rennes, France; 2008.