IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2024, accepted 30 June 2024, date of publication 3 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422479

SURVEY

Integrating Machine Learning Into Vehicle Routing Problem: Methods and Applications

REZA SHAHBAZIAN[1], (Senior Member, IEEE), LUIGI DI PUGLIA PUGLIESE[2], FRANCESCA GUERRIERO[1], (Senior Member, IEEE), AND GIUSY MACRINA[1]

[1]Department of Mechanical, Energy and Management Engineering (DIMEG), University of Calabria, 87036 Arcavacata, Italy
[2]Istituto di Calcolo e Reti ad Alte Prestazioni, Consiglio Nazionale delle Ricerche, 87036 Rende, Italy

Corresponding author: Francesca Guerriero (francesca.guerriero@unical.it)

This work was supported by the National Recovery and Resilience Plan (PNRR) Ministero dell'Università e della Ricerca (MUR) Project under Grant PE0000013-FAIR.

ABSTRACT The vehicle routing problem (VRP) and its variants have been intensively studied by the operational research community. The existing surveys and the majority of the published articles tackle traditional solutions, including exact methods, heuristics, and meta-heuristics. Recently, machine learning (ML)-based methods have been applied to a variety of combinatorial optimization problems, specifically VRPs. The strong trend of using ML in VRPs and the gap in the literature motivated us to review the state-of-the-art. To provide a clear understanding of the ML-VRP landscape, we categorize the related studies based on their applications/constraints and technical details. We mainly focus on reinforcement learning (RL)-based approaches because of their importance in the literature, while we also address non RL-based methods. We cover both theoretical and practical aspects by clearly addressing the existing trends, research gap, and limitations and advantages of ML-based methods. We also discuss some of the potential future research directions.

INDEX TERMS Vehicle routing problem (VRP), machine learning, reinforcement learning, deep learning, combinatorial optimization.

I. INTRODUCTION

The VRP is one of the most challenging and studied problems in the operations research (OR) field. As can be seen in Figure 1, the number of published papers over the years shows an exponential increase. Almost 50% of the studies were published during and after 2021, indicating the importance of the subject.

The vehicle routing problem (VRP) [1] is a combinatorial optimization problem belonging to the NP-hard class [2].

As shown in Figure 2, the VRP can be defined as an optimization problem with a set of scattered customers with stochastic or dynamic demands and a number of vehicles that could be homogeneous or heterogeneous. The goal is to find the lowest possible trip cost (distance/time) such that all customers are visited. Several variants of the VRP have been defined by introducing constraints to the problem. For instance, capacitated VRP (CVRP) introduces limited capacity for goods, or VRP with time windows (VRPTW) introduces specific time frames [3].

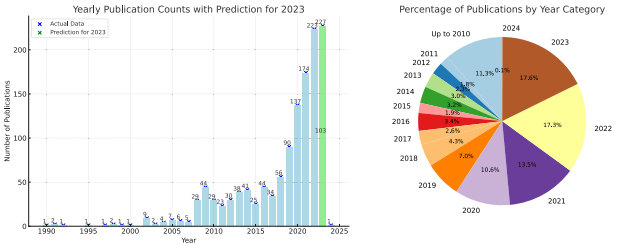


FIGURE 1. The number of VRP-related papers indexed in Scopus database with the keywords "Vehicle Routing Problem" and "Learning".

The existing literature mainly concentrates on exact approaches, heuristics, and meta-heuristic algorithms to solve the VRPs [4]. Recently, machine learning (ML)-based

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

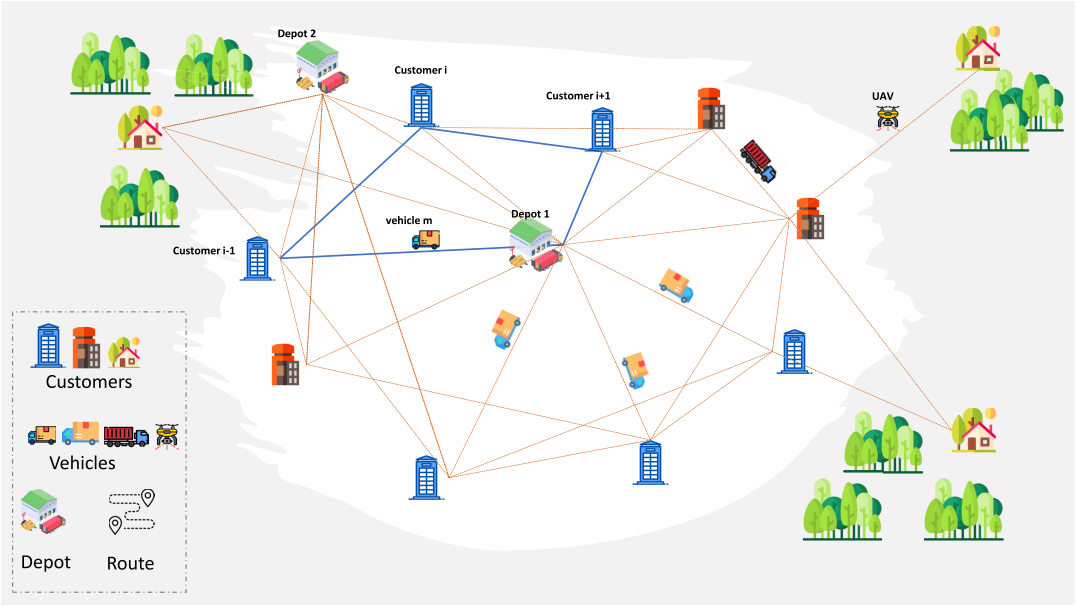


FIGURE 2. The schematic of the VRP where the vehicles (homogeneous or heterogeneous), with single or multi-depots. The dashed lines show the feasible routes, while the solid line shows the selected route for vehicle m that visits the customer $(i-1, i, i+1)$ and returns to depot 1. The objective varies depending on the introduced constrains. However, it usually includes the minimization of total travel time and/or energy consumption while the time windows, capacity, or other constraints are satisfied.

algorithms have gained the interest of numerous researchers for solving the VRPs [5]. Considering the importance of VRP, there are many surveys in the current literature. However, there is still no comprehensive review on ML-based solutions for VRPs, that discusses the problem from a technical point of view.

In this paper, we focus on the rapidly expanding ML-based VRP research and provide a comprehensive review and categorization from both a technical and applied standpoint. We categorize the related works, address new research directions, and present an overview of practical implementation guidelines, including existing benchmark datasets and software solutions (see Sec. I-C for the contributions of this paper).

A. EXISTING SURVEYS

We examined over 20 related surveys to find out if a review on ML-based methods has been presented in the VRP-related literature [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. There are only two related surveys that partially address ML-based methods. Bai et al. [27] incorporate ML and review hybrid solutions for VRPs that integrate analytical techniques with ML tools. They divide integration efforts into three main categories: ML-assisted VRP modeling, ML-assisted offline, and ML-assisted online optimizations. The study focuses on providing a high-level overview rather than specific features of the publications, such as mathematical formulations, solution methods, datasets, and future technology directions. In [30] the authors very briefly address the learning-based approaches to the VRPs by almost ignoring all the related studies. This gap in the literature motivated us to perform this review on ML-based approaches for VRPs.

B. METHODOLOGY

We searched the "Scopus", "Web of Science" and "arXiv" databases to cover relevant literature on the use of ML-based methods for VRPs. We focused on the studies published between 2020 and 2023, although some older studies are addressed based on their importance. Our initial search inputs were around 300 publications, which were reduced after a critical assessment. We attempt to investigate all available relevant studies and encompass over 20 published related surveys. The majority (61%) of papers submitted for critical evaluation are drawn from trustworthy journals (see Figure 3), and 39% of our covered studies are published in conferences or archive pre-prints. As can be seen in Figure 3, 64% and 18% of the journals are ranked Q1 and Q2, respectively.

C. OUR CONTRIBUTIONS

In this paper, we aim to address the following four important questions:

1) How ML-based solutions are defined for VRPs?
2) How to categorize the ML-based methods for VRPs?
3) What are the limitations, advantages, and possible future directions in the integration of ML into VRPs?
4) What are the practical considerations for using or designing ML-based solutions for VRPs?

Considering the above-mentioned questions, the structure of this research is presented in Figure 4. In this survey,

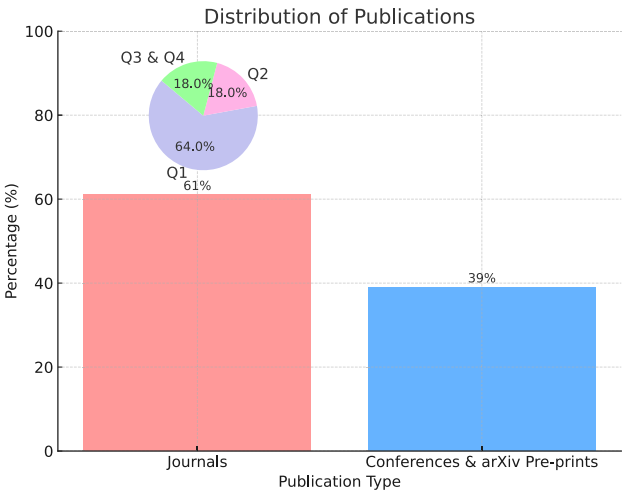


FIGURE 3. Categorization of the recent studies (2020-2023) covered in this paper based on the publication type and the journals' scoring (Q1-Q4) in Scopus.

we provide several novel contributions to existing literature. We first present a classification of ML-based methods, with a focus on reinforcement learning approaches. To the best of our knowledge, this is the first time the ML-based methods have been investigated comprehensively for VRPs. Second, we discuss the practical challenges of implementing these strategies in real-world scenarios. Finally, we identify present research gaps and suggest future research areas to further progress the discipline.

The main contributions of this paper are presented as follows:

- We provide a thorough review of ML-based solutions for VRPs. We cover the publications between 2020 and 2023. Some notable research papers conducted before 2020 are also included. We present and discuss the implemented models of RL, including the sequence-to-sequence, attention mechanism, and Markov decision process.
- We categorize the covered studies from a variety of perspectives, including their applications (six categories) and proposed algorithms (five categories).
- We perform a high-level classification and illustrate how ML-based methods can complement existing exact or heuristic approaches, improving the overall system performance and enabling the resolution of large-scale VRPs.
- We discuss the existing limits of the research and discuss future research directions. We attempt to bridge the gap between the OR and ML communities by highlighting possible collaboration directions.
- We address practical considerations by providing implementation guides, benchmark instances, and open-source solvers.

The remainder of this paper is structured as follows: The background information on solution approaches for the VRP and their different forms are given in Section II. In Section II-C, we look at studies that use ML to solve VRPs, focusing on RL-based solution approaches. Section IV presents a discussion, summarizing the knowledge of the considered scientific literature. Section V gives practical suggestions on accessible benchmark datasets and open-source software packages. Finally, Section VI concludes and makes recommendations for future research. The abbreviations used in this paper are presented in Table 1.

II. SOLUTION APPROACHES

In the scientific literature, typically the vehicles are assumed to be homogeneous (same characteristics), and the customer set is defined a-priori, as well as information on the transportation network and cost functions. These assumptions help to approximate the problems, while the real world demands that we handle more dynamic behavior by addressing complex multi-depot stochastic systems [31]. This is why solution methods play a critical role in enabling us to integrate more real-world needs into VRPs. As we discuss in this paper, the main advantage of ML-based methods lies in their potential to handle dynamic demands, uncertainty, and large-scale instances. In this section, we will look deeper into the various ML-based approaches, evaluating their strengths and weaknesses. For instance, while RL-based approaches provide flexibility in dealing with dynamic demands, they frequently need large computational resources. Non-RL approaches, on the other hand, are often easier to implement, but they may not perform as well in complicated, real-time contexts.

VRP solution strategies can be classified according to different features, and they are usually broken down into three main groups: exact methods, heuristics, and metaheuristics [32]. However, with the new advancements, there is a new categorization, as we address in this survey, by ML-based approaches. In what follows, we briefly describe the aforementioned solution strategies.

A. EXACT METHODS

Deterministic VRPs are typically formulated as integer linear programs (ILP) [33] or constraint optimization programs [34]. The goal is to identify a binary decision variable assignment that minimizes the objective while satisfying operational constraints. Most algorithms use a divide-and-conquer strategy to divide the initial solution space into smaller sets. Branch-and-bound [35] is a simple algorithmic structure implementation. The branch-and-bound method is considered in a variety of research papers [36].

Exact methods are computationally expensive and cannot provide an optimal solution for large-scale problems. However, exact approaches could be combined with heuristic algorithms to filter out subsets of the solution space that cannot improve the cost function. One of the most straightforward ways to apply the lower-bound heuristic is to linearly relax the ILP into a convex linear program, which can be solved in $O(n^3)$ by many algorithms, such as the simplex method.

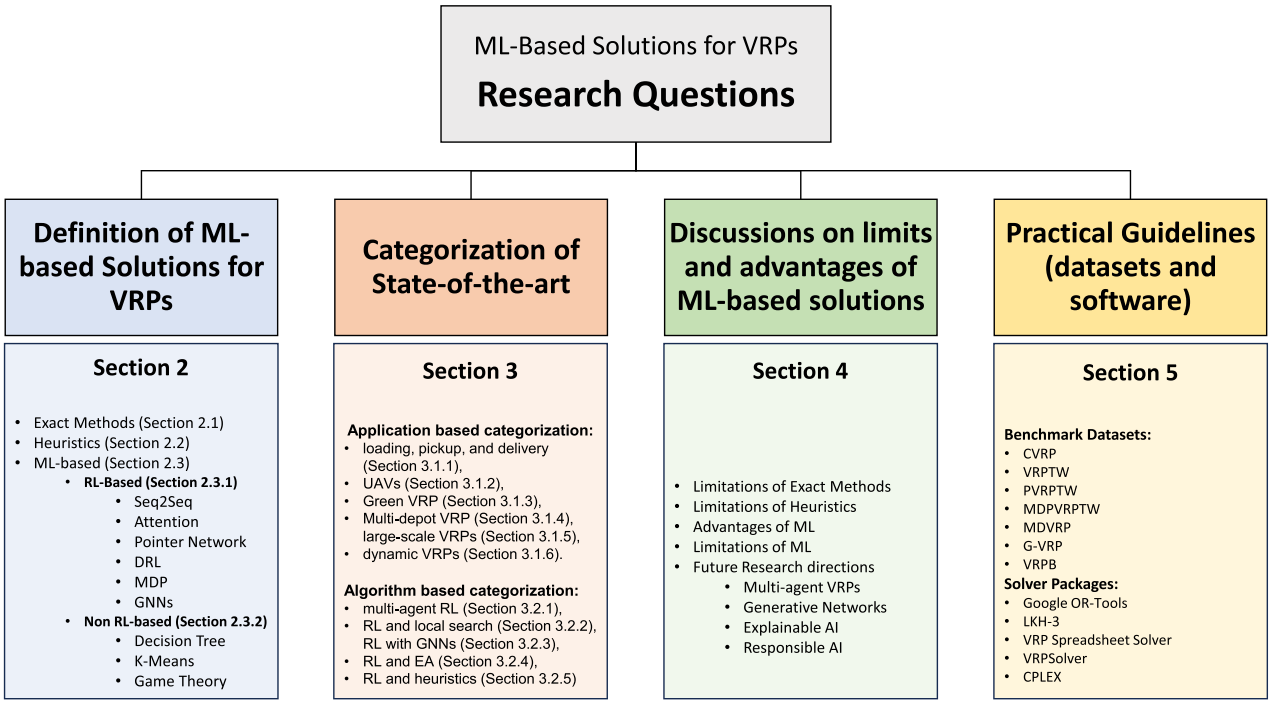


FIGURE 4. The structure of this survey by our contribution considering the raised questions.

B. HEURISTICS AND META-HEURISTICS

Heuristic approaches are widely employed to quickly find the feasible solutions on large instances [37].

Meta-heuristics are more advanced procedures that guide the search process to explore the solution space and find near-optimal solutions efficiently [8]. Meta-heuristics can be classified into two main classes: single-solution-based and population-based approaches.

1) SINGLE-BASED SOLUTION

Among the most known single-based methods, we mention the variable neighborhood search (VNS) [38], [39], [40], the large neighborhood search (LNS) methods [41], the tabu search (TS), and the simulated annealing (SA) [42], [43].

2) POPULATION-BASED SOLUTION

Population-based approaches try to find a solution (from a group of solutions), either by pairing and combining existing ones or by teaching them to work together through a learning process [44]. Elsaher and Awad [8] underline that population-based approaches can be broken down into two main categories: approaches that use evolutionary computation (EC) and approaches that use swarm intelligence (SI). A lot of hybrid algorithms use meta-heuristics along with ML to find improved solutions [45], [46], [47], [48], [49].

Heuristics and meta-heuristics might become trapped in local optima, especially in complex or large-scale problems. Usually problem-specific heuristics are not flexible enough to adjust to dynamic constraints and settings that yield unsatisfactory solutions. Although meta-heuristics offer greater flexibility and can explore a wider range of solutions, they can be computationally demanding and necessitate precise parameter adjustment.

C. ML-BASED APPROACHES

From one perspective, we categorize the ML-based approaches into reinforcement learning (RL)-based and non-RL-based. The majority of the related studies focus on RL-based approaches. ML can help address VRPs in two main directions: describing the problem as a Seq-2-Seq or Markov decision process and applying ML-based techniques to solve the problem. The majority of covered papers in the literature are based on RL and its variants. ML approaches are generally applied to single-agent environments with stationary environments. However, solving VRP with ML brings significant issues, such as the training process taking time, the performance of a ML-based algorithm being heavily dependent on the training dataset, and convergence and a near-optimal solution not being guaranteed. In continuing, we provide a more technical presentation of both RL-based and non RL-based methods.

1) RL-BASED METHODS

The first efforts at using RL for addressing the VRPs are represented by the Sequence-to-Sequence (Seq2Seq) models. They are a sort of neural network (NN) architecture that maps a variable-length input sequence to a variable-length output sequence. These models have been utilized in natural language processing and were adapted for VRPs. As can be

TABLE 1. Abbreviations used in this study; presented by Alphabetical order.

Actor-Critic	AC	evolutionary computation	EC	Mean-field theory	MFT	reinforcement learning	RL
attention mechanism	AM	first-generated first-served	FGFS	mixed-integer linear programming	MILP	SDVRP with time windows	SDVRPTW
Attention-to-Attention mechanism	AtAM	Game theory	GT	multi depot VRP	MDVRP	Sequence to Sequence	Seq2Seq
autonomous guided vehicles	AGVs	Generative adversarial networks	GANs	multi-agent RL	MARL	simulated annealing	SA
capacitated VRP	CVRP	graph convolutional cooperative multi-agent reinforcement learning	GCC-MARL	multi-decoder Attention model	MDAM	site-dependent VRP	SDVRP
convolutional neural network	CNN	graph convolutional network	GCN	multi-depot periodic VRP with time windows	MDVRPTW	stochastic integer linear programming	SILP
Coordinated Decision of Loading and Routing	CDLR	graph neural networks	GNNs	multi-depot VRP with time windows	MDVRPTW	swap-body VRP	SBVRP
deep policy dynamic programming	DPDP	green VRP	GVRP	multi-head Attention	MHA	tabu search	TS
deep Q-network	DQN	hybrid genetic search	HGS	multi-objective discrete learnable evolution model	MODLEM	Unmanned Aerial Vehicles	UAVs
Deep reinforcement learning	DRL	integer linear programs	ILP	Neural Network	NN	variable neighborhood search	VNS
double-deep Q-learning network	DDQN	Lin-Kernighan heuristic	LKH	operations research	OR	vehicle routing problem	VRP
dynamic and stochastic inventory routing problem	DSIRP	machine learning	ML	pattern injection local search	PILS	VRP with backhauls	VRPB
dynamic and uncertain VRP	DU-VRP	Markov Decision Process	MDP	periodic VRP with time windows	PVRPTW	VRP with simultaneous delivery and pickup with time window	VRPSDPTW
dynamic stochastic electric VRP	DS-EVRP	Markov routing game	MRG	Recurrent Neural Network	RNN	VRP with time windows	VRPTW

seen in Figure 5, considering a small VRP instance with four customer locations (A, B, C, D), the Seq2Seq model takes the locations and demands as input sequence and learns to predict the optimal visiting order as output sequence. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the input sequence, where each x_i is a vector representing the location of customer i. The Seq2Seq model aims to determine a sequence $Y = \{y_1, y_2, \ldots, y_n\}$ where y_i represents the i-th location the vehicle should visit. The model consists of two main components: the encoder and the decoder. The encoder processes the input sequence and compresses the information into a context vector, while the decoder takes the context vector and generates the output sequence one element at a time.

In summary, the Seq2Seq model for VRP includes the following key elements:

- **Input sequence**: A set of delivery locations represented as a sequence of vectors. Let $X = x_1, x_2, \ldots, x_n$ be the input sequence, where x_i represents the vector representation of the i-th delivery location.
- **Encoder**: Takes the sequence X as input and determines a vector representation of the input sequence. Let $h = f_{encoder}(X)$ be the output of the encoder, where $f_{encoder}$ is the encoder NN.
- **Decoder**: Takes the output of the encoder, h, as input and generates an output sequence of the optimal routes. The decoder is a Recurrent Neural Network (RNN) that generates the output sequence one element at a time, conditioned on the previous elements. Let $Y = y_1, y_2, \ldots, y_m$ be the output sequence, where y_i represents the i-th element. Let d_i be the decoder input at time step i, defined as $d_i = [h, y_i - 1]$. Let $p(y_i|d_i)$ be the probability distribution over the possible next elements of the output sequence at time step i, given the decoder input d_i.
- **Training**: Using a supervised learning approach, where the optimal routes are the target output for each input, the model is trained to minimize the distance traveled by vehicles.

The Seq2Seq models cannot incorporate complex constraints or long-term dependencies [50]. To overcome this issue, the Attention mechanism (AM) was introduced. AM enables the model to focus on different parts of the input (e.g., customers' locations) when predicting the route, leading to more effective solutions [51]. As illustrated in Figure 6, considering the same VRP instance with customer locations (A, B, C, D), AM computes a set of attention weights, indicating the importance of each input element for each output step [52]. First, the list of customer locations and needs (X) is sent through an embedding layer to turn each customer's raw features into a dense vector representation. This step enhances the model's ability to capture complex relationships within the data. Next, the embedded representations are fed into the attention layer during the decoding phase.

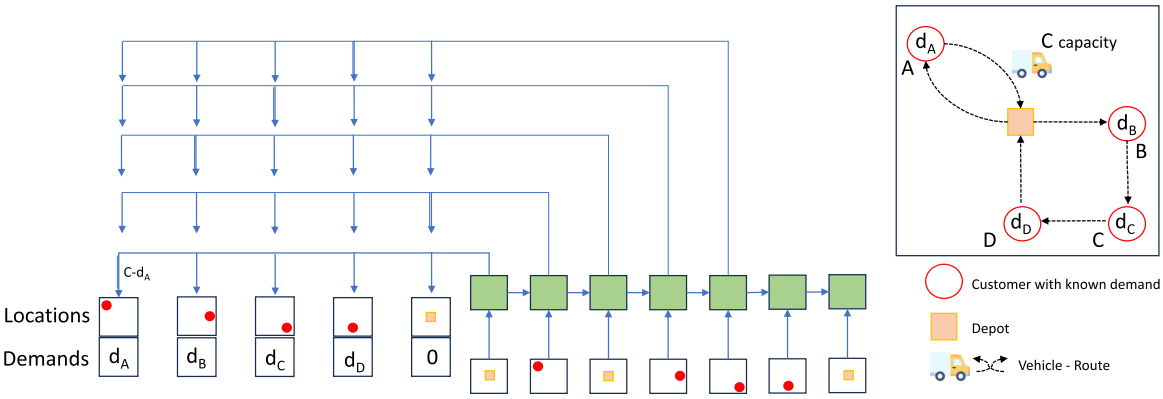


FIGURE 5. The schematic of the seq2seq model. In this example, we consider one depot and four customers (A, B, C, D). We assume a capacity availability that permits the existence of two optimal routes. For instance, vehicle capacity is $C = 10$ while $d_A = 9$, $d_B = 2$, $d_C = 3$, and $d_D = 9$.

At each decoding step i, AM computes a context vector c_i as a weighted sum of the embedded inputs, where the weights represent the relevance of each input to the current decoder state. In particular,

$$c_i = \sum_{j=1}^{n} \alpha_{ij} E_j, \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})}$$

where α_{ij} is the attention weight, e_{ij} is the alignment score between the input at position j and output position i and E_j is the embedded representation of the j-th input. In addition, $y_i = g(c_i, y_{i-1})$ where the function g generates the i-th output based on the context vector c_i and the previously generated output y_{i-1}.

Attention-based methods have been very popular. Deep attention with dimension reduction is studied in [53] and a dynamic Attention-based approach with mixed-instance training methods is presented in [54].

Pointer networks are a type of Seq2Seq model that creates chains of discrete tokens by pointing to places in the input chain. This was particularly suitable for VRP, where the output route is a reordering of the input locations [55]. It should be noted that the combination of these methods is also addressed in the literature. For instance, Vinyals et al. [56] propose a Seq2Seq Pointer Network based on an RNN that uses an AM in a supervised manner. Both attention-based models and pointer networks are difficult to scale due to their high computational complexity and memory usage. They typically require a large amount of training data, which makes them potentially poor in generalization and integrating complex constraints.

Deep reinforcement learning methods have been extensively applied to VRPs, typically within the Markov Decision Process (MDP) framework. Deep reinforcement learning (DRL) methods combine RL and deep learning [57] (see Figure 7). DRL allows vehicles to make decisions based on unstructured input data without requiring manual state space engineering. DRL algorithms are capable of processing massive volumes of data and determining which actions to take. DRL also struggles with large computing cost and training time, since learning efficient policies requires several interactions with the environment. While DRL provides a more flexible framework for policy learning than pointer networks, it frequently necessitates more fine-tuning and processing resources. DRL offer more adaptability to dynamic settings in compared to attention-based models and pointer networks.

MDPs provide a mathematical framework for modeling decision-making with partially random outcomes that are reliant on earlier actions. RL is used to develop a policy that maps states to actions in such a way that the predicted cumulative reward over time is maximized. The key distinction between Seq2Seq models and MDPs is that the former construct optimal routes based only on the input delivery locations, without taking into account the present status of the system or the actions made by the vehicles. MDPs, on the other hand, explicitly model the state of the system as well as the actions of the delivery vehicles and use RL to establish a policy that maximizes the predicted cumulative reward over time. Therefore, MDPs are better models for stochastic and dynamic VRPs.

MDP is represented by a set of states, actions, rewards, and the corresponding transition probability function. At the time t, given the current state of environment S_t, an agent selects action A_t that yields a reward in the next step, denoted by R_{t+1}, and a new state S_{t+1}. The probability function is defined in Eq. (1) in which the next state is determined by the current state and action.

$$p(s'|s, a) = P(S_{t+1} = s'|S_t = s, A_t = a) \tag{1}$$

The reward function denoted as $r(s, a, s')$, defines the reward for the taken action a at state s with the consecutive state s', as shown in Eq. (2).

$$r(s, a, s') = \mathbb{E}[R_{t+1}|S_t = s, A_t = a, S_{t+1} = s'] \tag{2}$$

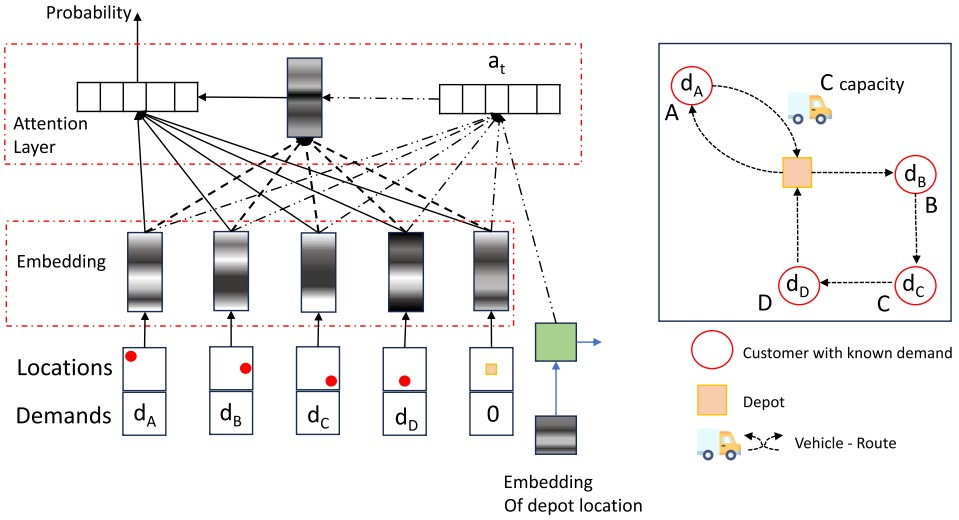


FIGURE 6. The schematic of the Attention Mechanism. In this example, we consider one depot and four customers (A, B, C, D). We assume a capacity availability that permits the existence of two optimal routes. For instance, vehicle capacity is $C = 10$ while $d_A = 9$, $d_B = 2$, $d_C = 3$, and $d_D = 9$.

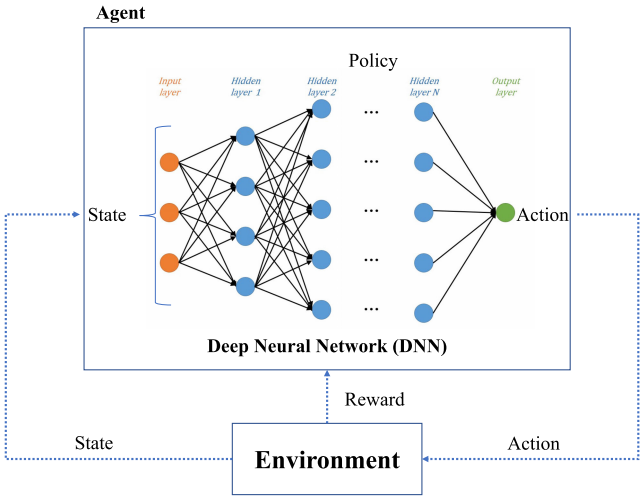


FIGURE 7. A high-level overview of deep reinforcement learning (DRL), which is used to solve MDP models in VRPs.

The policy denoted by $\pi(a|s)$, maps the states to the selected actions as presented in Eq. (3).

$$\pi(a|s) = p(A_{t+1} = a|S_t = s) \tag{3}$$

The value function denoted as $v(s)$ presents the long-term rewards. A value function with a certain policy π is denoted by $v_\pi(s)$ and defined as the total reward from state s with planning horizon T and discount fActor γ as presented in Eq. (4).

$$v_\pi(s) = \mathbb{E}[\sum_{k=0}^{T-1} \gamma^k R_{t+k+1}|S_t = s] \tag{4}$$

The action value function is denoted by $q_\pi(s, a)$ and shows the reward when taking action a at state s under policy π as given in Eq. (5).

$$q_\pi(s, a) = \mathbb{E}[\sum_{k=0}^{T-1} \gamma^k R_{t+k+1}|S_t = s, A_t = a] \tag{5}$$

The common objective is to find a policy, that leads to maximum total rewards by applying the Bellman optimality conditions [58]. Each problem instance includes a set of N_c customers, a set of N_v vehicles, and a depot. Each customer has several characteristics, including location (v), demand (d), service time (s), time windows (e, l) so that they could be represented as a tuple $c_i = (v_i, d_i, s_i, e_i, l_i)$, $i = 1, 2, \ldots, N_c$. Vehicles are assumed to have a fixed capacity (Q) and speed (E) [59].

The vehicles start at the depot and meet the demands of customers sequentially. Vehicles need to comply with the following rules (single depot scenario):

- The vehicles depart from the depot and return to the depot after visiting several customers.
- The sequence of customers visited constitutes the routes of the vehicles.
- Routes begin at and end at the depot.
- The depot can only appear at both ends of a route.
- Once back at the depot, vehicles are not allowed to visit customers again.
- Vehicles have a maximum load, i.e. capacity Q. The sum of the demands of customers on a route should not exceed capacity Q.
- Customers' time windows could be considered *soft*, which implies that vehicles can visit them before or after the time window. Even yet, there will be a penalty if a customer is visited outside of the time window. The penalty is specified as a linear function of the arrival time and the time window. Vehicles are not permitted to stop at a customer node. When vehicles arrive at a customer

node, they should promptly serve the customer and then leave.

The solution of a VRP instance could be represented as $R = (r_1, r_2, \ldots, r_v)$, where r_i is the i-th vehicle's route. The goal is to find a solution with minimal cost. For an instance P and a solution R, the cost could be computed as follows:

$$Length(r_i|P) = \sum_{j=1}^{|r_i|-1} \|v_{r_i}(j), v_{r_i}(j+1)\|_2 \quad (6)$$

$$Penalty(r_i|P) = \sum_{j \in r_i} ((e_j - t_{i,j})\,\alpha\,k_e + (t_{i,j} - l_j)\,\beta\,k_l) \quad (7)$$

$$Cost(R|P) = sum_{i=1}^{N_v}(Length(r_i|G) + Penalty(r_i|G)) \quad (8)$$

where $Length(r_i|P)$ is the total route length of vehicle i, $\|.\|_2$ is l_2 norm, $r_i(j)$ is the j-th customer in the path of vehicle i. $Penalty(r_i|P)$ represents the early and late penalty of route i, t_i is the time when vehicle i arrives at customer j, the travel time between two customers equals the distance of them divided by the speed E of the vehicle, the arrival time of the current customer is equal to the arrival time of the previous customer plus the service time of the previous customer and the travel time between the two customers. α and β are early and late arrival penalty coefficients, respectively, whereas k_e and k_l indicate early or late arrival; indeed, it is equal to 1 when arrival time is earlier or later than the time window; otherwise, 0.

RL-based methods aim at learning a policy that maps states to actions in such a way that the expected cumulative reward, R, over time is maximized. From Figure 8, it is evident that different model-free and model-based methods have been proposed [60]. The model-based RL is used to improve uncertainty handling, and it also enables the proposed method to construct and store known transport events. To learn the policy, algorithms such as *Q-learning* [61], [62] and *policy gradient* are very common in the literature. In Q-learning, the vehicles estimate the expected cumulative reward of taking action in a given state and following the optimal policy thereafter by updating their estimates based on the observed rewards and transitions. The policy can then be obtained by selecting the action with the highest estimated action-value in each state. Q-Learning is a model-free algorithm in which the Q-values can be learned using a table or a neural network, and the policy can be derived from the Q-values using an "ϵ-greedy" policy or a "softmax" policy [63]. Some deep learning models, such as [50], [64], and [65], use auto-regressive decoding to create the solution to the VRP incrementally. In these studies, the RL is used to train a policy that selects the next node in the solution based on a reward function generated at each step.

Model-based RL makes use of simulated interactions, it can learn optimal policies more quickly and with more sample efficiency. Model-based RL makes use of an environment model for planning and decision-making. However, in complicated or dynamic contexts, it may struggle with model inaccuracies, resulting in sub-optimal solutions. Model-free RL can adapt better to complex real-world conditions while it often requires more training data and computational resources. As a result, model-free RL needs longer training times but provides more robust performance.

The Actor-Critic (AC) methods provide a framework for learning policies (see Figure 8). For instance, Bello et al. [66] apply the AC for unsupervised learning. In contrast to AC, AM and Pointer networks provide architectural tools to handle the sequential and combinatorial aspects of VRP effectively. The AC is also combined with DRL. In particular, Zhao et al. [67] propose two DRL-based algorithms. At first, they utilize the AC method and combine their proposed DRL with a local search method to improve the solution quality. They use the output of the DRL-based method as the initial solution of the local search algorithm. They show that the DRL-based solution overcomes the existing state-of-the-art approach, while the combined DRL and local search method overcomes their own proposed DRL-based algorithm.

VRPs have graph-based structures inherently. The customers or depots could be considered as nodes of this graph, while the possible routes could be modeled as the edges with various constraints and objectives. Therefore, graph neural networks (GNNs) are also utilized in the literature to provide a solution for VRPs [68]. The GNNs are integrated with RL and serve as a function approximator for the policy estimation. GNNs are also combined with deep Q-learning for model-free policy-based solutions [69]. The combination of GNNs with heuristics such as hybrid genetic search (HGS) is also presented in the literature [70]. In general two types of single-head and multi-head attentions are introduced for VRPs. Based on the experimental results on VRPs it is shown that single-Head Attention works better with larger instances compared to the multi-Head Attention [71].

RL-based models need a large number of interactions with the environment in order to develop efficient routing policies. This means a high computing cost and lengthy training times. Attention-based models and pointer networks can produce faster initial solutions in compared to the RL-based methods; however, they compromise the quality of long-term optimization. RL-based methods require less computational resources and fine-tuning in compared to the attention-based models and pointer networks.

D. NON RL-BASED METHODS

Although the majority of the existing ML-based solution approaches are based on RL, other machine learning-based methods such as decision trees, game theory, and k-means are presented in the literature. Typically non RL-based methods are proposed in hybrid solutions.

In [72], a decision tree is used to figure out the direction of the evolution process in a multi-objective learnable evolution model. The authors use heuristics and a sequential search method to solve the problem that has three goals: travel distance, driver pay, and the number of vehicles. Their experimental results suggest that their model is capable of

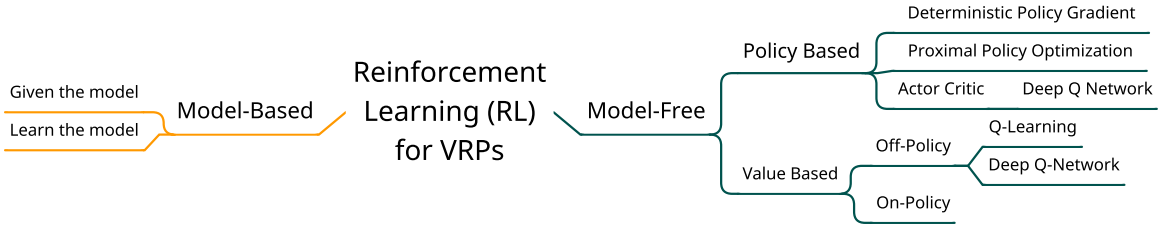


FIGURE 8. Categorization of RL-based methods in the literature to solve the VRPs. Model-free RL provides resilience and adaptability, whereas model-based RL offers efficiency and speed.

finding better Pareto-front solutions than other evolutionary algorithms.

In [73], the authors use machine learning to predict the value of variables that can only be one of two options in the best solution. They also say that this framework can be used to predict branching scores for variables that are fractional based on complete strong branching. When the predicted decision variables are added to a node selection strategy, the expected branching score is then used on the variable selection policy.

Game theory (GT) in conjunction with ant colony optimization is studied in [74]. The authors investigate dynamic and stochastic ways to tackle complicated dynamic vehicle routing problems, minimize total travel time, find the shortest routing path, and consider dynamic demands. They use GT to help with decision-making and discover optimal solutions to conflict and cooperation situations.

Wang et al. [75] use the k-means for collaborative multi-center VRP with time windows and mixed deliveries and pickups. The authors describe the issue as a mixed-integer programming model to lower operational costs. They then present a two-step hybrid method that combines customer clustering with vehicle routing optimization. The authors specifically talked about a 3D k-means clustering method based on space-time distances for delivery and pickup constraints at the same time in a vehicle routing problem with a time window. They also combine the genetic algorithm with particle swarm optimization to describe a hybrid heuristic technique for optimizing vehicle routes. They use 30 small-scale instances in the CPLEX solver. Villalba and Rotta also studied K-means to investigate the VRP with a time windows (R20-villalba2022clustering). They provide a clustering-based algorithm that uses K-means and optics clustering approaches, as well as nearest neighbor and local search 2-opt heuristics.

The total traveled distance and the total waiting time of drivers are two typical objective functions in VRPs that combine the multi-objective vehicle routing problem with the time frame. The decomposition-based multi-objective evolutionary algorithm (MOEA/D) has been applied to single-objective optimization issues. A method called MOPILS was developed by the authors in [76] to solve the multi-objective vehicle routing problem. It is based on pattern injection local search (PILS) and combines MOEA/D.

Column generation is an iterative approach for solving a variety of optimization problems that decomposes the problem into primary and multi-pricing problems [77]. Morabit et al. [77] propose a new heuristic pricing algorithm that combines ML and heuristic pricing. The authors leverage data from previous executions to minimize the network size. On the training data, the authors use a random forest model, and the hyper-parameters are tweaked using a cross-validation approach. The authors conducted some experiments, and the results reveal that this strategy reduces the execution time for solving VRPS with time windows constraints, by roughly 40%.

In the study by Mandi et al. [78], preferences like travel time or fuel use are modeled along with the arc probabilities so that they can be used in the optimization process. The authors look at the data that is available to pick out features, neural architectures, and loss functions. They then use a neural network model to guess the chances that these things will happen. The use of ML-based state-of-charge estimation for electric vehicles by Adaboost and XGBoost algorithms is proposed in [79].

III. STATE-OF-THE-ART: CATEGORIZATION OF RL-BASED METHODS

In this section, we review and categorize the covered studies based on different perspectives. First, we consider the applications, followed by the categorization based on their algorithms. Considering the applications, we classify the studies into six categories: 1) loading, pickup, and delivery (Section III-A1), 2) UAVs (Section III-A2), 3) Green VRP (Section III-A3), 4) Multi-depot VRP (Section III-A4), 5) large-scale VRPs (Section III-A5), and 6) dynamic VRPs (Section III-A6). We further categorize the studies based on their proposed algorithms into five categories: 1) multi-agent RL (Section III-B1), 2) RL and local search (Section III-B2), 3) RL with GNNs (Section III-B3), 4) RL and evolutionary algorithms (Section III-B4), and 5) RL and heuristics (Section III-B5).

A. APPLICATION-BASED CATEGORIZATION

In this section, we categorize the covered studies based on the constraints imposed by the specific application.

1) LOADING, PICKUP AND DELIVERY CONSTRAINTS

Loading constraints may include considerations such as loading and unloading times, the arrangement of items within the vehicle, and the type of items being loaded. Pickup constraints refer to the requirements associated with picking up goods from specified locations before they are delivered to their destinations. This version of the problem can also include time windows within which pickups must occur, specific sequences in which pickups must be made, or even constraints on the vehicles that can perform pickups. Delivery constraints may include considerations like customer availability, confirmation requirements, and access restrictions to delivery locations. These constraints make the solution to the VRP more complex.

Hansuwa et al. [80] analyze VRP with simultaneous pickup and delivery and a time windows constraint. The authors use practical applications with data uncertainties in both delivery and pickup to model the problem. The authors investigate a sequence of V vehicles, with the cost of a central depot proportional to travel time. They assume that vehicles can pick up and drop off at either the depot or the customers. Vehicle capacity, customer service time windows length, and depot opening and closing times are some of the requirements that the solution must meet. Even though the authors look at different vehicles, the fleet's capacity and cost are comparable. Some additional constraints are investigated, such as permitting vehicles to travel to new destinations, prohibiting customer revisits, and allowing the vehicle to idle in the depot after the assignment. The goal is to identify routes that reduce vehicle dispatch and travel costs. The authors investigate several ways to solve the capacitated simultaneous delivery and pickup with time windows (VRPSDPTW) problem, which involves unknowns like travel time, service time, delivery quantities, and pickup quantities. These include RL, the ellipsoidal uncertainty robust counterpart model, and linear programming of the robust counterpart of mixed-integer linear programming with box uncertainty. The authors apply DRL's power to address real-world VRPSDPTW situations with undetermined data quality. In summary, the authors train and handle feature data using the AM-based DRL technique, both in terms of scale and uncertainty. The authors show that their techniques generate acceptable responses for small or medium-sized datasets, but suboptimal solutions for larger ones. In contrast, AM-based DRL produces acceptable quality outputs for large-scale practical data sets used in real-world applications [80].

In [81], the authors address VRP with pickup and delivery, where DRL is used to handle the scale-up challenge. They present a DRL-based technique that collaborates with a heterogeneous AM. They develop six types of AMs for policy networks within DRL, employing an encoder-decoder structure. Among them, three types of AMs are used to learn the relationship between each pickup point and the points of other roles, while the remaining three types focus on learning the relationship between each delivery point and the points of other roles. In their proposed method, adaptive masks are used to reject invalid locations, ensuring feasibility. Because of this heterogeneous Attention strategy, the policy network can learn the pairing and precedence links. The authors train the policy network using an RL methodology with a roll-out baseline, and the policy gradient method is distinguished by an Actor-network and a self-critic network. The supplied incentive is calculated using a roll-out baseline with a comparable structure to the Actor-network. After receiving the reward, the RL algorithm adjusts the parameters of two networks using the policy gradient approach. The authors compare the proposed RL-based method's performance to heuristics like simulated annealing, OR-Tools, and an earlier version of the AM and show that their techniques outperform, in terms of effectiveness and efficiency, the state-of-the-art approaches. The article does not account for uncertainties.

Chen et al. [82] study the combined delivery and pickup demands and propose a framework with encoder-decoder architecture. They utilize a GNN encoder to extract the feature and use the AM in the decoder. They also propose a Coordinated Decision of Loading and Routing (CDLR) mechanism to determine the loading rate. They show that the combined GNN encoder and CDLR simultaneously can better handle dynamic demands.

Qiu et al. [83] investigate the home delivery and installation routing problem with synchronization limitations imposed by a home industry company. They assume that the goods need to be delivered and installed in the customer's home. To shorten the overall travel distance of delivery and service routes, the authors propose employing DRL within an encoder-decoder and multi-head AM in conjunction with a beam search strategy. The authors conduct experiments and compare the suggested DRL-based technique to the Lin-Kernighan heuristic (LKH), adaptive large neighborhood search, and step-wise transformer AM. The findings of their evaluation show that the proposed DRL-based solutions outperform various established tactics and also provide some management implications.

The main features of the analyzed papers are reported in Table 2. Several earlier studies address the loading, delivery, and pickup constraints [84], [85], [86].

2) UNMANNED AERIAL VEHICLES (UAVS)

The UAV-based VRP is a variant that incorporates the use of UAVs (drones) in the routing and delivery process. The goal is to find the optimal routes for a fleet of vehicles in various locations in the most efficient manner. In UAV-VRP, the problem includes both ground vehicles and UAVs.

In a recent paper, Chen et al. [87] present same-day delivery with combined vehicles and drones. They tackle the problem, assuming that vehicles and drones can transport goods from a single depot. They take into account that 1) vehicles and UAVs have varying capacities, speeds, and features (including battery swaps for UAVs), 2) the requests from the customers are limited in time (time-windows), and

TABLE 2. ML-based solutions for the VRP with loading, pickup, and delivery constraints.

Index	Ref.	Authors	Year	Solution	Focus Area
1	[81]	Jingwen Li et al.	2021	ML (RL)	heterogeneous Attention mechanism to empower the policy in deep reinforcement learning
2	[80]	Hansuwa et al.	2022	ML (RL)	mixed-integer linear programming (MILP) and Attention Model (AM) based deep reinforcement (DRL) learning
3	[82]	Chen et al.	2023	ML (RL)	Encoder-Decoder Framework and Attention Model (AM) based deep reinforcement learning (DRL)
4	[83]	Qiu et al.	2022	ML (Deep RL)	Encoder-Decoder Framework and multi-head Attention mechanism (DRL)
5	[75]	Wang et al.	2022	ML (K-means and heuristics)	k-means clustering combined with genetic algorithm and particle swarm optimization
6	[85]	Göçmen and Erol	2020	ML (Neural Network)	three-dimensional loading constraint with machine learning-based hybrid solution

3) the system operator (human or computer) must have sufficient information to make decisions (accept or reject the request).

They aim to maximize the number of served customers considering the above-mentioned constraints. They propose a deep Q-learning technique. Deep Q-learning uses deep neural networks as an approximation architecture to learn the value of state-action pairings. Because NNs can generally be trained offline, the suggested method should be suitable for real-time decision-making. Time, fleet, and actions are three characteristics of actions and state space investigated by the authors. All of these features are supplied into the NN, which is then normalized using min-max. The authors evaluate heterogeneous fleets. They evaluate the proposed deep Q-learning method's performance when customers are distant in time, space, or both, and the number of vehicles ranges from 2 to 4, with 10 to 15 drones available.

Because the training phase is designed to be offline, the robustness of the machine learning technique is heavily dependent on the training data. In this circumstance, training the system with large-scale real-world datasets with controlled uncertainty data can obviously improve the performance of machine learning algorithms.

In another study, Delamer and Givigi [88] address dynamic settings for UAVs and propose a solution to the dynamic traveling repairman problem. The authors consider UAVs with no mobility constraints and unlimited sensing, that must serve several targets while attempting to reduce the waiting time for each target (customer). They combine RL with proximal policy optimization. The authors model the problem as a MDP with binomial target distributions (customers). The proposed proximal policy optimization is built with two NN architectures: a feed-forward network and a convolutional neural network (CNN). To assess the performance of the defined techniques, the authors employ nearest-first (NF) and first-generated first-served (FGFS) as baseline heuristic policies. The NF (customer) prioritizes the nearest target. It is considered that this strategy approximates the waiting time and the number of services. The FGFS caters to the target (customer) who has waited for the longest. Their evaluation results indicate that as the size of the state space increases, a feed-forward network outperforms heuristic techniques. The CNN outperforms the heuristic policy FGFS but not the heuristic NF.

The study in [88] fails to account for various practical elements and uncertainties, and it lacks comprehensive evaluations. The authors of [88] consider restricted sensor capabilities and multi-agent RL to delve deeper into practical environments.

Wang et al. [89] study the problem of autonomous ground vehicle routing and propose a DRL strategy for tactical driving in complex highways, while accounting for real-time traffic dynamics. They use a deep Q-network, which takes in account dynamic traffic data and generates typical tactical driving decisions as action. They aim to design the solution while considering the successful highway exit, average driving speed, and driving safety and comfort. They also build a CNN to extract traffic characteristics that help Q-learning decision-making.

Wang et al. [90] present a multi-resolution, multi-agent, mean-field RL algorithm (3M-RL). The author's purpose is to organize the flight paths of the UAVs so that they do not collide with each other and arrive at their destinations safely. Mean-field theory (MFT) approximates the influence of other agents on a single agent by a single averaged effect known as the mean-field. A multi-body problem is reduced to a single-body problem using the MFT. The authors employ the MFT to replace interactions with the agent with average interactions. This is accomplished through RL, which is accomplished through the actions of other agents based on their surroundings. The problem can be modeled as a single-agent RL problem, with the state space remaining constant as the total number of agents increases. When applying this MFT to UAVs, extra considerations,

TABLE 3. RL-based solutions for the VRP with hybrid UAV/vehicles.

Index	Ref.	Authors	Year	Vehicle Type	ML Method	Focus Area
1	[87]	Chen et al.	2022	Vehicle/UAV	Reinforcement Learning	Single depot capacitated with time windows
2	[88]	Delamer and Givigi	2022	UAVs	RL with proximal policy optimization	UAVs with no mobility constraints and unlimited sensing
3	[89]	Wang et al.	2021	UAVs	Reinforcement Learning	Multi-resolution, multi-agent, mean-field and a Actor-Critic neural network

like the safe distance between UAVs, must be taken into consideration. Each UAV makes judgments based on local observations and does not communicate with other UAVs, according to the study's proposed technique. Using an Actor-Critic NN, the approach trains a routing policy with multi-resolution observations, combining specific local information and aggregated global information based on mean-field. In Table 3 the main features of the analyzed papers are given.

3) GREEN VRP

The operational research community has devoted high attention to environmental concerns as well as to the management of electric vehicles and their limitations. In most cases, the primary purpose is to reduce energy usage or pollution. These concerns are related to the uncertain parameters, that have been considered in the stochastic VRPs.

Alqahtani and Hu [91] investigated the VRP for electric vehicles (EVs). By reducing reliance on power from the main power grid, the authors were able to reduce energy consumption while omitting other costs. The authors assume that the EV is equipped with an energy storage unit, that is a photovoltaic panel, which provides power to customers in various places to meet their energy demand at lower energy costs. In this problem, which is modeled as an MDP, the system state is a tuple of four variables relating to vehicle position, battery state of charge, solar irradiance, and energy load. The actions are considered to be vehicle movement and energy transition. Vehicles can move in four directions (up, down, left, and right) or not move at all, while energy transactions include charging, discharging, and idling. Each vehicle's movement is limited to the zones around it (limited in the grid). The authors describe an RL paradigm in which a deep Q-network (DQN) evaluates the Q-value for each action the EV takes. A case study of four EVs and twenty users is used to evaluate the selected DQN. The experiments show that the RL algorithms are better at saving energy than the genetic algorithm, particle swarm optimization, and artificial fish swarm algorithms. When compared to the previously described baseline algorithms, the DQN algorithm can produce a near-optimal solution in a reasonable execution time.

Logistics' rising carbon emissions will significantly harm the environment. To reduce carbon emissions in logistics, Zou et al. [51] address the low-carbon multi-depot vehicle routing problem. The authors propose a DRL with an improved transformer model (TAOA) that includes both a multi-head Attention mechanism (MHA) and an Attention-to-Attention mechanism (AtAM) to address issues caused by recurrent neural networks and AMs in encoders and decoders, such as the long-distance dependence problem and the neglected correlation between query vectors. The MHA is applied to process different sections of the input sequence, and the AtAM is used to compensate for the MHA's lack of correlation between query outcomes and query vectors. The training of network parameters with TAOA takes time, but the following prediction results are obtained rather quickly. The authors employ the advantage Actor-Critic technique from intensive learning to train the model. Their proposed model outperforms the traditional transformer model (Kools), the genetic algorithm, and Google OR-Tools.

Basso et al. [92] consider energy usage and random customers as uncertainties in the dynamic stochastic electric vehicle routing problem (DS-EVRP). The DS-EVRP is presented as an MDP with a set of possible options and a state transition function. Since the authors assume that the vehicle fleet is homogeneous, meaning all vehicles have the same features, the problem is addressed for a single vehicle. The goal is to create a channel for a single EV to service the customer at a given time slot. This problem is viewed as a single depot that receives both predictable and unpredictable customer requests. The deterministic requests are received with a predetermined probability before the EV leaves the depot. The stochastic requests are received with an unknown probability after the EV leaves the depot. All customers' addresses are assumed to be known, and all requests are granted. Since it is assumed that battery capacity is limited, it may be important to prepare for charging on the fly. The authors use tabu search techniques to solve the static EVRP, which also acts as a comparison benchmark. Stochastic EVRP users must anticipate future demand and energy consumption. At the same time, certain aspects are known in advance; some are unknown with known probabilities, and some are unknown with unknown probabilities. As a technique to reduce energy consumption and the risk of battery depletion, the authors propose a safe RL algorithm. Through Monte Carlo simulations, their proposed Q-learning-based system learns about random user needs and energy use and then guesses the path. For offline policy training, they employ a ϵ-greedy policy. The authors run simulations with a medium-duty truck weighing 10,700 kg (with the battery) and carrying a maximum

cargo of 16,000 kg. The numerical results demonstrate that a RL-based solution approach is more effective and reliable than existing heuristics. The study addresses dynamic requests in EVRP. However, numerous concerns are worth investigating, especially the heterogeneous fleet. The training step is also offline and heavily reliant on the training data.

In [93] the authors investigate public transportation, while accounting for dynamic changes in demand, travel times, and traffic. They propose combining RL with a dynamic scheduling strategy of EVs, such as vehicle-to-grid transaction capabilities or dynamic charging strategies. The trained model includes the number of requests, drop-off requests, battery level, time, and current location. The authors claim that their method supports dynamic vehicles (agents) and outperforms a fixed scheduling method.

The authors of [94] look into the issue of an unbalanced inventory of electric motors in public transportation and suggest a way to manage the fleet so that there is the right number of vehicles at each station, while taking into account the cost of moving tasks and rental opportunities. DRL is used as a decision-making function to find the optimal fleet allocation action based on the most recent status of the number of automobiles at each station.

Basso et al. [95] examine time-dependent electric VRP with chance restrictions (EVRP-CC) and partial recharge. The goal is to anticipate the probability distribution of energy consumption for each network road link. The authors propose a probabilistic Bayesian ML approach for forecasting the estimated energy consumption and variation for road linkages, paths, and routes. The proposed approach method is divided into two stages: discovering the best paths and optimizing the routes. To account for the unpredictability of energy consumption, they anticipate charging within a confidence interval. They use Bayesian regression to find a solution. The prior is based on a model of how vehicles move, and the posterior is improved by gathering more data. This allows the system to estimate without any training data, by using a prior computed probabilistic speed profile derived from map data and a simplified vehicle model. While the vehicles are on the road, the prediction precision can rapidly improve with the posterior real data. As effective energy consumption criteria, the authors consider friction and drag, battery temperature regulation, cabin equipment, and external auxiliaries. To test the performance of their proposed technique, the authors run a series of experiments using real-world traffic data on a global simulation platform. They compare their proposed ML-based technique with a deterministic formulation and show that it is more accurate in terms of energy prediction as well as energy savings.

The defined method has the advantage of being less reliant on training data and being able to be updated with real data [95]. Although this study might be extended to incorporate alternate vehicle layouts, energy use correlation for adjacent road connectors, and congestion mitigation.

Aljohani et al. [96] examine metadata-driven routing optimization for Evs to lower energy usage. The authors propose a real-time data-driven electric vehicle routing optimization to reduce energy consumption. As an agent, they use a double-deep Q-learning network (DDQN) to understand the EV's maximum travel policy. The policy model is taught to guess what the agent should do next by looking at the reward signals and Q-values it receives, which show possible routes. The Markov chain model is used to calculate the agent's energy requirements on the road. Authors use the Geocoding API to translate physical locations into geographical coordinates. The EV can travel in one of eight ways. The authors perform two tests with two routes, that are comparable in length but have different geographic characteristics. Both investigations were conducted at a certain time and date with a limited battery. Their evaluation results indicate that their DDQN framework consumes less energy than conventional techniques.

The user choice on multiple criterion route suggestion is examined in [97]. The authors consider a multi-objective route suggestion system that takes into account three factors: fuel consumption, travel time, and air quality. They use the Q-learning-based RL method promptly. The authors employ OpenStreetMap to generate a road network graph, which they then update regularly using existing predictors for air quality, travel time, and fuel usage.

The main features of the analyzed studies are summarized in Table 4.

4) MULTI-DEPOT VRP

The number of available depots significantly impacts the complexity of the VRP and the solution strategies. Multi-Depot VRP provides a more realistic representation of many real-world logistics and transportation problems, but it also requires more sophisticated solution methods to tackle the added complexity.

The authors in [98] study the CVRP and propose an augmented state representation. They use Q-Learning for auto-regressive techniques, that generate solutions by inserting nodes progressively. They also show how RL may be used to solve the multiple depot VRP using the proposed technique. Based on the collected computational results, the proposed model does not outperform the solvers (i.e., LKH3); however, it is useful to find the initial solutions.

In [99], the authors examine MDP for the multi-depot dynamic VRP with stochastic road capacity and propose a solution approach based on a simplified two-stage stochastic integer linear programming (SILP) model. The proposed method is suitable for obtaining a policy that is dynamically developed on the fly throughout the roll-out process. The roll-out method is part of the approximate dynamic programming look-ahead solution strategy.

The summary of the reviewed paper is presented in Table 5.

TABLE 4. ML-based solutions for the Green VRP and Electric Vehicles.

Index	Ref.	Authors	Year	Objective	ML Method	Focus Area
1	[96]	Aljohani et al.	2021	lowering energy usage	RL (double deep Q-learning)	metadata-driven routing optimization
2	[95]	Basso et al.	2021	anticipate energy consumption	Bayesian machine learning	A two-stage time-dependent EVRP
3	[94]	Anchuen et al.	2022	cost of moving tasks and rental opportunities	Deep reinforcement learning	unbalanced electric motor inventory problem
4	[93]	Eshkevari et al.	2022	Dynamic changes	RL with dynamic scheduling	public transportation with EVs
5	[92]	Basso et al.	2022	uncertainties and dynamic Changes	Safe RL (Q-learning)	energy usage and random customers as uncertainties
6	[51]	Zou et al.	2022	low-carbon multi-depot VRP	deep RL with transformer	multi-head Attention and Attention-to-Attention mechanism
7	[91]	Alqahtani and Hu	2022	Low energy and less dependency	RL (deep Q-network)	reduce energy consumption by reducing dependency on power supply
8	[97]	Sarker et al.	2020	Multiple Criterion	RL (Q-learning)	Consider fuel consumption, travel time, and air quality

TABLE 5. ML-based solutions for the Multi-depot VRPs.

Index	Ref.	Authors	Year	Objective	ML Method	Focus Area
1	[98]	Bdeir et al.	2021	Travel distance	RL (Q-learning)	Q-Learning for auto-regressive
2	[99]	Anuar et al.	2021	stochastic road capacity	ML (dynamic Policy)	stochastic integer linear programming (SILP)

5) LARGE-SCALE VRPS

Poullet [37] proposes a two-stage technique to solve large-scale VRP with time windows constraints. In the first stage, a clustering algorithm based on optimal classification trees is created to segment customers into smaller subsets. In the second stage, an Actor-Critic RL strategy on these smaller customer clusters is defined.

Gupta et al. [100] seek a quick solution for the capacitated VRP with time windows, to provide a ML-based solution for large-scale problems. The authors propose a deep Q-network with an encoder-decoder-based RL technique. The encoder is a technique for Attention, whereas the decoder is a fully connected NN. They outperform heuristics, a meta-heuristic algorithm, and a multi-agent RL framework, as indicated by their results.

The goal of Liu et al. [101] is to tackle the large-scale VRP. The authors propose a pre-training mechanism for online shared networks and use the multi-head Attention mechanism (MHA) to train the graph Pointer Network in dual-network RL. When the time windows constraints are taken into account, the authors claim that their proposed approach can be used for large-scale VRPs with 100/300/500 customers. According to their evaluations, their proposed method provides appropriate solution quality and offline solution efficiency.

Authors in [102] investigate the routing of autonomous guided vehicles (AGVs). The authors show that the problem cannot be solved using meta-heuristic methods due to the high real-time demands for AGVs. Oversimplification of large-scale AGV systems typically results in unsatisfactory solutions. The authors describe a DRL technique to address the AGVs routing problem by defining the problem as a MDP. Asynchronous deep Q-network is also used as the basic architecture for RL.

Jiang et al. [103] offers a new coding strategy for solving the distribution task in the multi-distribution center scenario VRP with capacity restrictions. They use improved RL with a MhAM during the encoding phase. They make use of correlation information between the nodes of the encoding output distribution. They conduct computer simulations with Google OR-Tools, and the results show that the suggested method outperforms the ant colony and the simulated annealing algorithms.

The main characteristics of the considered papers are presented in Table 6.

6) DYNAMIC VRPS

In dynamic VRP, certain elements of the problem can vary over time or are not known with certainty, unlike static VRP, where all data are known in advance.

Pan and Liu [63] study dynamic real-world logistics and present a novel DRL framework for tackling a dynamic and uncertain VRP (DU-VRP). In a changing environment, the purpose is to fulfill customers' unpredictable demands. In this problem, given ambiguous knowledge about customer desires, the partial observation MDP is used to detect changes in customer requests in a real-time decision support system consisting of a deep neural network with a dynamic AM. RL is used to control the objective function of the DU-VRP to better train the routing process dynamics.

The authors of [104] use a route-based MDP to examine the dynamic VRP with time windows. They assume that the customers could be known in advance or stochastic. To approximate the objective function, the authors describe a solution technique that combines DRL and a routing heuristic. DRLSA is a proposed system that is based on simulated annealing and allows for optimal re-routing decisions. The authors demonstrate how the cost of the remaining vehicle routes can be used as a proxy for the required route sequence and time window. The authors assess the performance of the proposed DRLSA, comparing their results to those achieved using approximate value iteration and the many scenario techniques.

Peng et al. [105] developed the dynamic AM. To solve VRPs, the authors use a dynamic encoder-decoder architecture with a RL AM. The nodes in the dynamic design are integrated instantly as the vehicles return to the depot,

TABLE 6. RL-based solutions for large-scale VRPs.

Index	Ref.	Authors	Year	Objective	ML Method	Focus Area
1	[37]	Poullet et al.	2020	CVRP with time windows	RL (Actor-Critic)	Two-stage solutions (clustering + RL)
2	[100]	Gupta et al.	2022	CVRP	RL (Attention mechanism)	encoder-decode based RL
3	[101]	Liu et al.	2022	CVRP with time windows	RL (graph pointer network)	multi-head Attention mechanism
4	[102]	Lu et al.	2020	multi-distribution center	RL (Deep Q-Network)	Asynchronous deep Q-network
5	[103]	Jiang et al.	2021	stochastic road capacity	RL (multi-headed Attention)	solving the distribution task

as opposed to the old design. This embedding is fixed in the standard AM and represents the beginning state of the input instance. They propose two methods: sample roll-out and greedy roll-out. The first is stochastic and selects a node through sampling, whereas the second is deterministic and selects a node through maximum probability. In the AM, node features describe an instance's status. Unlike previous research that looked at fixed node attributes across time, the authors look at changing node properties and updating them depending on model decisions at various development stages. The authors describe a dynamic AM that uses hidden structure information during the development process and is based on a dynamic encoder-decoder architecture with dynamic node properties. To demonstrate the success of the suggested technique, the authors perform some experiments on the models and analyze instances with 50 and 100 customers.

In [106], authors investigate RL for VRPs using Attention-based RL models rather than earlier recurrent NN-based techniques. The authors explore dynamic network typologies and build a new Attention-based RL model that delivers increased node embedding via batch normalization reordering and gate aggregation, as well as dynamic-aware context embedding on multiple relational structures via an attentive aggregation module. The authors evaluate their proposed algorithm using the CVRP.

In [107], the authors focus on urban mobility for people and products to lower transportation system operational costs and negative externalities, and they propose merging passenger transportation with commodities delivery to improve vehicle-based transportation. Their proposed distributed model-free DRL system (FlexPool) learns optimal dispatch policies through interaction with the environment, allowing passengers to be pooled for ride-sharing and items to be delivered via a multi-hop transit strategy.

The authors in [108] investigate real-time intelligent vehicle routing systems and offer a DRL method for tackling the problem as a series of decisions. The authors investigate the proposed technique using the SUMO simulator and nine traffic scenarios, as well as the Wilcoxon test, to validate the results.

The trajectory data might be used to generate a high-resolution, uncertain road-network graph [109]. The authors of [109] investigate probabilistic budget routing and attempt to discover the path with the highest arriving probability while time and budget are limited. The authors suggest a hybrid strategy that blends convolution with ML-based estimates. To improve accuracy, the authors explore distribution dependencies.

Yuan et al. [110] provide an RL and self-supervised learning technique that uses an AM to learn a policy for solution generation and integrates a contrastive self-supervised learning model to learn the Attention encoder node by node. They use a two-phase learning technique that comprises node-wise learning and solution-wise learning to train both the AM and the self-supervised model. They conduct numerical experiments to evaluate the effectiveness of the proposed technique. The summary of the reviewd papers is presented in Talbe 7.

B. ALGORITHM-BASED CATEGORIZATION

The reviewed studies, categorized in Section III-A, use the RL method as their main solution algorithm. However, heuristic and meta-heuristic methods have been frequently used in the past, and ML-based algorithms can be applied to improve the performance of these heuristics. In this section, we categorize the covered studies with this perspective.

1) MULTI-AGENT REINFORCEMENT LEARNING

The standard RL learning process involves a single agent (vehicle), while the multi-agent RL (MARL) extends this to a multi-agent environment where vehicles learn and make decisions collectively.

Ren et al. [59] propose a MARL solution approach, using road recorders. Their proposed method, which includes management and strategy modules, can determine the optimal number of vehicles for VRP with pre-defined constraints. The interaction environment for DRL is provided by the management module. The strategy module is made up of an encoder, many route recorders, and a decoder. Based on data from the management module, it produces a route for each vehicle. During training, the management module generates problem examples for the strategy module to solve. The strategy module first processes the customer information through the encoder then constructs the state of the agents (vehicles) based on the encoder and route recorder output, and lastly, outputs the vehicle's route step by step. Finally, update the neural network parameters in the strategy module based on the output solution and its corresponding cost value. DRL's environment is provided via the management module. When the strategy module outputs the vehicle's next destination and the vehicle fleet travels, it should update the problem status information

TABLE 7. RL-based solutions for dynamic VRPs.

Index	Ref.	Authors	Year	Objective	ML Method	Focus Area
1	[63]	Pan and Liu	2022	unpredictable demands	RL (Attention mechanism)	DRL for dynamic and uncertain VRP
2	[104]	Joe et al.	2020	Unknown customers	Deep RL	combined DRL and routing heuristic
3	[105]	Peng et al.	2019	dynamic VRP	RL (Attention mechanism)	dynamic encoder-decoder architecture
4	[106]	Xu et al.	2021	dynamic VRP	RL (Attention mechanism)	dynamic network typologies
5	[107]	Manchella et al.	2021	Urban transport	Deep RL	distributed model-free deep RL
6	[108]	Koh et al.	2020	Real-time VRP	Deep RL	series of decisions with RL
7	[109]	Pedersen et al.	2020	uncertain road-network graph	Hybrid ML (convolutional)	probabilistic budget routing
8	[110]	Yuan et al.	2021	self-supervised VRP	RL (Attention mechanism)	node by node Attention encoder

in addition to generating problem instances, making the management module critical for possible remedies. Based on a predetermined distribution, the management module generates many instances for training and validation. As the vehicle fleet travels, this module updates the status of customers and vehicles. The status is comprised of the current location of the vehicles as well as the customers visited. Once the strategy module has provided the next destination for a vehicle, the vehicle's location is changed to the chosen customer and the remanining load is updated. To avoid infeasible routes, a masking approach must be used to exclude customers who are not available at the current time step, such as customers who have been visited previously or whose demand exceeds the vehicle's remaining load. The strategy module receives problem data and mask information from the management module and turns them into the route sequence for each vehicle. The authors conduct various experiments and compare the performance of OR-Tools and a genetic algorithm on over 1000 instances. The authors show that while performance is essentially comparable in small-scale VRPs, the novel technique outperforms earlier heuristics in large-scale VRPs (for example, two vehicles with 100 customers).

Shou et al. [111] model the VRP as a Markov routing game (MRG), in which each vehicle (agent) learns and updates an en-route path choice strategy while interacting with other vehicles in the network. The authors combine MARL and a mean-field multi-agent deep Q-learning approach to solve MRG effectively. The authors perform several experiments to demonstrate the effectiveness of the proposed technique.

Authors in [112] explore the VRP in dynamic and uncertain environments and provide a learning policy, that gives decision criteria for generating routes based on online measurements of the environment state, including the customers' setup. Deep neural networks are employed to implement the learning policy in the proposed method, which is known as multi-agent routing with deep AM. Using a sequential multi-agent decision-making model, the authors formalize the description and temporal evolution of a dynamic and stochastic VRP. They use deep neural networks with AM to learn generalizable state representations and formulate online decision rules for dynamic and stochastic data. The authors run numerous experiments on both stochastic and deterministic capacitated VRP with 20 customers and four vehicles.

Authors in [52] investigate the multi-VRP with soft time windows. The authors present a multi-agent AM, a novel RL method that solves routing issues in real-time, while taking advantage of considerable offline training. The proposed method employs an encoder-decoder structure with Attention layers to generate tours of several vehicles repeatedly. For training, the authors use a multi-agent RL method with an unsupervised auxiliary network. The numerical results show that the proposed method outperforms Google OR-Tools and traditional methods while using the least amount of processing effort.

Ding et al. [113] investigate vehicular crowd sensing, which collects data from sensor-equipped urban cars. The authors consider hiring vehicles and propose a novel graph convolutional cooperative multi-agent reinforcement learning (GCC-MARL) framework for distributed routing decisions. The authors undertake trials and show that their proposed approach outperforms state-of-the-art algorithms in terms of both gross merchandise volume and data usefulness.

The authors of [114] look into large-scale situations involving multiple vehicles and time constraints. They then come up with a new multi-agent RL model capable of concurrently optimizing both route length and vehicle arrival times. The encoder-decoder structure acts as the model's foundation. The encoder mines the relationship between the problem's customer nodes, and the decoder produces each vehicle's path repeatedly. Their results show that their proposed model outperforms heuristic strategies in terms of both performance and computing time.

The authors of [115] suggest a new way to solve hierarchical problems called ''learning collaborative policies''. This method uses DRL as the ''seeder'' policy and ''reviser'' policy to find the almost perfect solution. The seeder generates as many diverse candidate solutions as possible and investigates action space, i.e., the sequence of assignment actions. They use an entropy regularization incentive to train the seeder's policy, while the reviser alters each candidate solution and divides the entire route.

The authors of [116] look into real-world CVRP where the vehicle fleet is not a clone of a single vehicle. Taking into account the heterogeneous CVRP with varied characteristics that affect their capacity (or travel speed), seek to reduce the fleet's vehicle's longest or total travel duration. To solve the issue, the authors suggest a DRL method based on the AM that learns to build a solution by picking both a vehicle and a node for this vehicle at each step. This is done with a vehicle selection decoder taking into account the different types of vehicles and a node selection decoder taking into account the building of routes.

In one of the very recent studies, the authors discuss the collaborative VRP and propose a deep multi-agent RL approach [117]. The authors state that characteristic functions scale exponentially with the number of agents while in their proposed method, the agents do not require access to the characteristic function, thus significantly reducing run-time.

2) RL WITH LOCAL SEARCH

In recent work, Pugliese et al. [118] investigate crowd-shipping distribution using ordinary people who normally travel on the roads in their vehicles. According to the authors, these cars are extra resources for the delivery company's trucks. These drivers are considered by the authors, who design a heuristic technique based on variable neighborhood search. They use ML-based techniques to explore the search space. The authors, in particular, use a learning technique to guide the selection of local search moves during the intensification phase, which can be thought of as an adaption of the Q-learning algorithm. Their findings show that the proposed framework saves CPU time, especially when there are a large number of customers.

Zhao et al. [67] propose a DRL model comprised of an Actor, an adaptive critic, and a routing simulator combined with local search to address the VRP. Based on the AM, the Actor is supposed to generate routing strategies, and the adaptive critic alters the network architecture to boost the convergence rate. These completed actions improve the solution quality of the training phase. The routing simulator provides graph information and motivation via the Actor and adaptive critic. The combination of DRL and local search is used to improve solution quality even further. The authors conduct experiments with 20, 50, and 100 vehicles, and the results show that the DRL model outperforms Google OR-Tools, the LNS algorithm, and previous DRL strategies on its while combining the DRL model with various local search methods yields promising results. Their approach does not account for several dynamic real-life aspects, and performance can also be improved with recent advances in RL, such as multi-header AM.

Wu et al. [119] define heuristic algorithms to solve VRPs that necessitate extensive trial-and-error. They use ML and present a framework for directly learning from improvement heuristics rather than learning from construction heuristics. Hence, they provide a framework for DRL for learning routing problem improvement solutions. The authors design a self-Attention-based deep architecture to serve as the policy network that guides solution selection. The technique is improved by doing neighborhood searches repeatedly in the direction of quality improvement. The authors apply DRL to automatically identify improvement policies, whereas previous heuristic methods require domain knowledge to identify the policies, and usually, their improvements are limited. The authors apply their method to solve the CVRP, showing its effectiveness.

Chen et al. [120] analyze the periodic VRP with time windows and open routes, assuming that trucks do not return to the depot after every single delivery, but rather at the end of every two shifts. They show that using existing mathematical models to solve the problem is impracticable, and those exact search methods cannot be used to tackle large-scale instances. Then, the authors propose VNS-RLS, a variable neighborhood search method that incorporates RL. The authors use RL to drive the search during the local search improvement phase.

In another study on the application of ML to improve the performance of heuristics on solving VRPs, Hottung, and Tierney [65] provide a novel large neighborhood search framework that incorporates the learned heuristics for producing new solutions. The learning of their suggested method is based on a deep neural network with an AM. The authors evaluate their proposed technique on CVRP with up to 297 customers and on the split delivery vehicle routing problem. They use batch and single instance searches to show that their proposed method outperforms heuristic solvers.

Paulo et al. [121] propose to use DRL to train a local search heuristic based on 2-opt. They create a policy neural network that uses the gradient technique to train a stochastic policy that selects 2-opt operations given a current solution. The authors conduct experiments considering 10,000 occurrences. They compare their proposed method to heuristics for nearest, random, and farthest insertion constructs, as well as OR-Tools with 2-opt and LKH as augmentation heuristics. Their novel strategy, according to the results, improves near-optimal solutions, showing it is faster than previous learning methods.

In [122], authors present a unique DRL method to learn construction heuristics and apply a multi-decoder Attention model (MDAM) to train many separate policies. This method raises the chances of finding an acceptable answer. To capitalize on the diversity of their proposed MDAM, the authors additionally use a customized beam search technique. The authors run a series of experiments to evaluate the performance of the proposed MDAM. They report on the run-times for solving 10,000 test instances to decrease the cost. The results show that their suggested MDAM outperforms the existing baseline techniques.

In [123], the authors propose a framework for value-function-based SRL using a combinatorial action space, in which the action selection issue is explicitly described as a mixed-integer optimization problem. The proposed framework is applied to the CVRP. It models an action as the creation of a single route using a deterministic policy that is improved using a simple policy iteration method. In medium-sized instances, the authors conduct simulations whose results outperform state-of-the-art OR techniques.

3) RL WITH GRAPH NEURAL NETWORKS

Duan et al. [124] were motivated to improve the performance of VRP systems, particularly when compared to OR-Tools. They propose a method for training the model

that incorporates RL (supervised). The authors use a graph convolutional network (GCN) with node attributes to describe cooperation and demand. The edge feature, as input to their proposed technique, models the actual distance between nodes. The authors use independent decoders to represent these two node and edge features. One decoder's output is supervised by the output of the other decoder. They show that the edge feature is significant enough to warrant explicit inclusion in the model and that the joint learning technique can accelerate training convergence and improve solution quality.

The deep learning architecture proposed by Hagström [68], combines a graph neural network with beam search for solving VRP, called VRPNet. The VRPNet network is based on the recurrent relational network architecture and learns a probabilistic representation of the solution space. VRPNet computes messages between nodes iteratively and updates the hidden states. The nodes broadcast the current state to their neighbors throughout each iteration (parallel computation). As a result, nodes can update the hidden state, and the cycle continues. The model produces promising results on small-scale problems.

4) RL COMBINED WITH EVOLUTIONARY METHODS

Achamrah et al. [125] examine the dynamic and stochastic inventory routing problem (DSIRP) and propose a novel solution based on a hybridization of mathematical modeling, genetic algorithms, and DRL. They test the proposed technique on 150 single-vehicle DSIRP benchmark examples and 450 multi-product DSIRP benchmark instances.

Moradi [36] addresses the VRP with time windows and proposes a multi-objective discrete learnable evolution model (MODLEM) that avoids undefined search by using a ML technique such as decision trees. To control the approach's multi-objective characteristic, they include a robust strength Pareto evolutionary algorithm in the learnable evolution model. The suggested MODLEM performance is evaluated using Solomon VRPTW [36]. The results show that MODLEM behaves the same as the state-of-the-art approaches in terms of solution quality and computing time in both areas.

Zhou et al. [126] apply ML to improve the performance of adaptive ant colony optimization. They provide an approach that is based on adaptive gradient descent (ADACO) theory and define the transition probability as a policy. In the proposed constraint-aware policy optimization for VRPTW, vehicles (agents) learn the constraints as a representation of the full environment to boost the generalization of RL methodologies. The authors evaluate the performance of their proposed ADACO technique on a range of instances spanning from 51 to 4,461 nodes, showing that ADACO is stable and less hyper-parameter sensitive and that its performance is comparable to state-of-the-art algorithms.

5) RL WITH HEURISTICS

The authors of [127] analyze a real-world VRP situation, taking into account practical constraints. Their proposed framework is based on DRL and a greedy heuristic. They use a self-attention framework in conjunction with heuristics to ensure that the self-attention framework only embeds the graph once. The self-attention or heuristic function manages various constraints, such as visiting each customer once. The authors analyze the proposed technique for the traveling salesman problem with time windows and rejection. The rejection happens when the customers cannot be served in the allocated time windows. The results show that their proposed technique outperforms the tabu search heuristic in terms of solution quality and computational time. The authors assume that the customers and their requirements are known in advance.

Kool et al. [128] study end-to-end DRL algorithms and their ability to improve the performance of approximation solution heuristics. The authors propose deep policy dynamic programming (DPDP) combined with heuristics. In DPDP, the deep neural network policy is used to prioritize and constrain the dynamic programming state space. The authors run a series of experiments. The results on 100 nodes show that the neural policy improves the performance of DP algorithms with limited constraints in the same way that LKH does.

Qin et al. [129] study practical heterogeneous VRPs with a specified fleet of vehicles of varied capacities. The purpose is to minimize the maximum routing time for the fleet. The authors mixed-integer linear programming model can use to identify optimal solutions for small-scale problems, while RL-based hyper-heuristics and DRL tackle large-scale problems. They assess two high-level policies and review meta-heuristics such as the artificial bee colony algorithm, ant colony optimization, cuckoo search, genetic algorithm, particle swarm optimization, and simulated annealing to develop the low-level policy. They use a policy-based RL technique to improve the effectiveness of the hyper-heuristic framework. The authors extract hidden patterns from the gathered data to combine the benefits of low-level heuristics. On large-scale problems, the numerical results show that the suggested approach outperforms existing meta-heuristic algorithms and the MILP solutions.

IV. DISCUSSION AND FUTURE RESEARCH DIRECTION

The key question that this survey attempts to answer is why and how ML can help to solve the VRPs or improve the performance of the existing solutions. As our study demonstrates, the bulk of available studies and reviews in the literature considered exact methods or heuristics. In small-size instances, both optimal approaches and heuristics perform well. Heuristics are much faster than traditional optimization procedures, but there is no guarantee that they will lead to the optimal solution. As a result, heuristic-based solutions are preferred in real-world VRPs [36]. The following summarizes the learned lessons for conventional non-ML solutions.

- Exact approaches provide the best or close to the best solution for any occurrence of the vehicle routing problem.
- VRPs are NP-hard problems; hence, exact approaches are only relevant in particular cases and for small-scale applications. For instance, CPLEX solver can not be trusted with more than 25 customers [118].
- For any case of the vehicle routing problem, approximate approaches provide a sub-optimal solution.
- Typically, the estimated approach distance from the optimal outcome is known (could be calculated). As a result, approximation approaches are sometimes used as a bound for other methods.
- Heuristic approaches may fail in some unusual circumstances. However, there has been no disagreement over the convergence requirements of heuristic approaches.
- When compared to optimal approaches, heuristic methods are noted for their speed and cheaper processing costs.
- When compared to optimal procedures, heuristic methods can usually be used to tackle bigger-scale vehicle routing problems.
- Meta-heuristic approaches used to solve the VRPs might be single-point or population-based.
- To create hybrid methodologies, population-based meta-heuristics can be combined with existing heuristics or machine learning-based methods. For instance, meta-heuristics, such as genetic algorithms or ant colony optimization, are combined with heuristics, such as big neighborhood search.

ML can help to address VRPs in two main directions. The first approach is to describe the vehicle routing problem as a Seq-2-Seq or MDP and then apply a ML-based technique to solve the problem. The majority of covered papers in the literature are based on RL and its variants. The second way that ML-based methods can be used is to make existing heuristics work better by teaching them a policy or making the search space smaller. In vehicle routing problems, for example, it is well known that the quality of heuristic solutions is dependent on the features of the problems to be solved.

Another finding is that ML approaches are generally applied to single-agent environments with stationary environments. Vehicles (agents) must know the past condition of the environment due to stationary assumptions that are not compatible with real-world needs.

DRL is commonly utilized to solve the VRPs. DRL provides a framework that could be used for solving decision-making problems in dynamic contexts. However, solving VRP with ML, particularly DRL, brings significant issues, which we mention below:

- DRL procedures are quick to implement. However, the training process takes time.
- The performance of a ML-based algorithm is typically heavily dependent on the training dataset. As a result, a high number of instances are needed to train the DRL model.
- Although the DRL approaches perform well in computer experiments, convergence and having a near-optimal solution is not guaranteed.
- In small or medium-scale VRPs, cutting-edge heuristics outperform DRL. However, as the number of vehicles increases, so does the computing complexity, DRL shows a very good performance in handling such large-scale problems.

ML-based methods have been applied to a wide range of vehicle routing problems, but there is still much more to be done to bridge the gap between the two OR and ML communities. The majority of existing studies are based on the deterministic assumption, and the few stochastic studies are limited in the restrictions they address. In the following, we thoroughly explain some of the fundamental concepts that ML approaches can address.

- Almost all of the studies presume that the fleet of vehicles is homogeneous (identical vehicles). Vehicles with a wide range of characteristics may better represent real-world applications. This is one of the issues that could be best addressed by ML-based approaches.
- In almost all the studies, the number of available vehicles is known in advance, whereas determining the optimal number of vehicles is another issue to be considered.
- While multitasking optimization is supposed to give excellent efficiency and accuracy, it is rarely at the forefront of VRPs [114]. In the VRPs, dynamic multitask optimization methods could potentially be applied in future research.
- The cases with uncertain and stochastic needs are addressed, but there is still room for additional realistic and plausible uncertainties to be considered. There are many examples, especially for electric self-driving scenarios.
- There is still a gap in studying the convergence of hybrid approaches or the conditions under which the solution is likely to be achievable.
- Taking into account integrated information, such as delivery history data, road network information, and various constraints, into online decision-making could be a viable approach for large-scale problems.
- Although ML-based methods are used to increase heuristic performance, there is still a need for more efficient ML-based methods for optimal hyper-parameters or learning strategies.
- Multi-agent VRP can capture more complex and realistic scenarios, especially when there are interactions or dependencies among different entities in the system. This subject can still be better addressed in future studies.
- Generative adversarial networks (GANs) have recently been at the center of attention for many research studies. While there are very few addressing the VRPs [130].

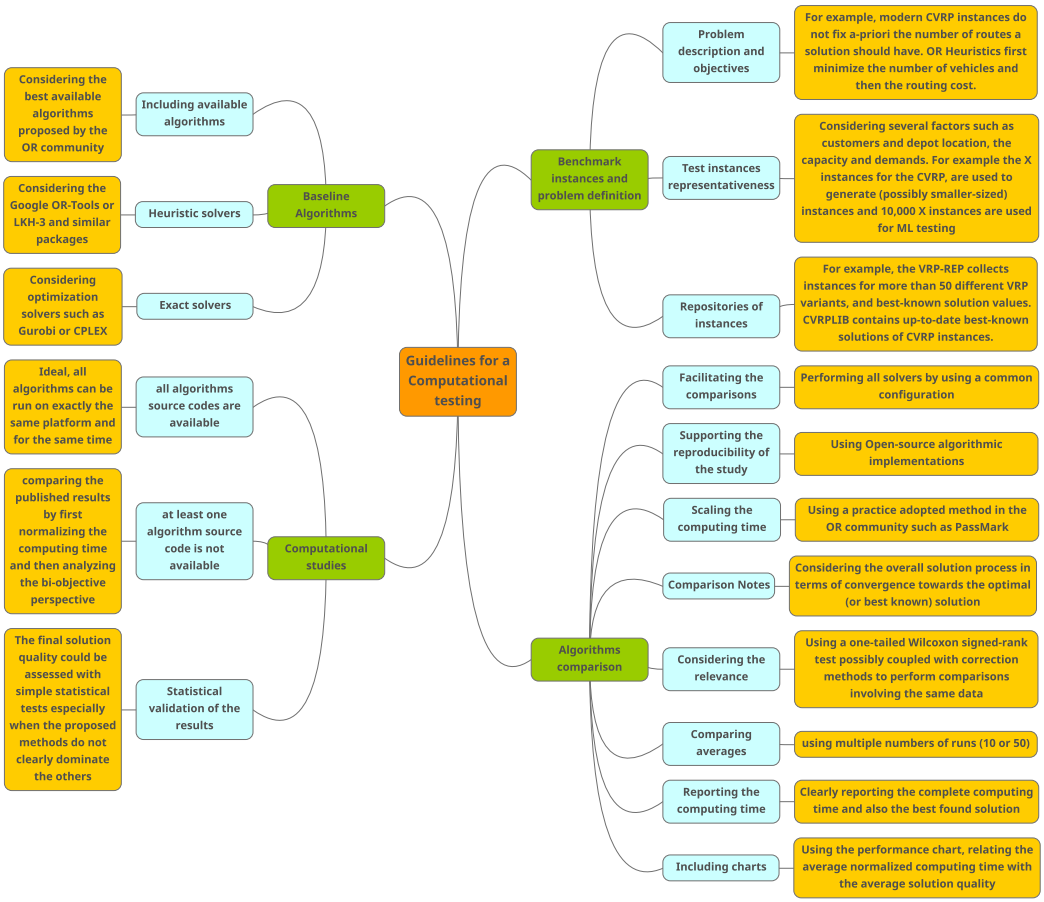


FIGURE 9. The overview of the practical comments that need to be taken into account in evaluating the ML-based solutions for VRPs.

TABLE 8. The dataset used for CVRP benchmarks.

Index	Reference	Variant	Dataset Benchmark	No. of instances
1		CVRP	Augerat 1995 - Set A	27
2		CVRP	Augerat 1995 - Set B	23
3		CVRP	Christofides et al. 1979 - CMT	14
4		CVRP	Christofides et al. 1979 - Set M	5
5		CVRP	Golden et al. 1998 — Set 1	20
6		CVRP	Li et al. 2005	12
7		CVRP	Augerat 1995 - Set P	24
8		CVRP	Christofides and Eilon 1969 - Set E	13
9		CVRP	Fisher 1994 - Set F	3
10		CVRP	Geoffrey De Smet 2014 - Belgium CVRP	23
11		CVRP	VeRoLog Members VRP	1
12		CVRP	VeRoLog Members VRP 2016	1
13		CVRP	Uchoa et al. 2014	100
14		CVRP	De Smet 2017 - Belgium road-km/road-time/air - 50-2750 visits	50
15		CVRP	Letchford and Salazar-Gonzalez 2018	240

- Transfer learning is another technique that is being utilized in many ML-based methods, but it has not been well studied in VRPs [131].
- New concepts could be considered in combination with RL-based methods. For instance, the concepts of meta-learner [132] and ML-quantum computing [133].
- The new paradigms of responsible AI or explainable AI have not yet been addressed in the literature. This might be important, especially considering the new rules on minimizing the risk of AI in real-world applications.

To summarize, we believe that ML techniques can improve solutions to VRPs, particularly when dealing with large-scale problems. They could be the only reasonable solution when the environment and demands introduce many uncertainties and the problem deals with a heterogeneous (non-identical) fleet of vehicles.

V. PRACTICAL GUIDELINES

When dealing with medium- or large-scale problems, ML-based solutions are most effective. It worth mentioning

TABLE 9. The dataset used for VRP with time windows constraints.

Index	Reference	Variant	Dataset Benchmark	No. of instances
1		VRPTW	Solomon 1987 - C1	27
2		VRPTW	Solomon 1987 - C2	24
3		VRPTW	Solomon 1987 - R1	36
4		VRPTW	Solomon 1987 - R2	33
5		VRPTW	Solomon 1987 - RC1	24
6		VRPTW	Solomon 1987 - RC2	24
7		VRPTW	Gehring and Homberger 1999 - C1	50
8		VRPTW	Gehring and Homberger 1999 - C2	50
9		VRPTW	Gehring and Homberger 1999 - R1	50
10		VRPTW	Gehring and Homberger 1999 - R2	50
11		VRPTW	Gehring and Homberger 1999 - RC1	50
12		VRPTW	Gehring and Homberger 1999 - RC2	50
13		VRPTW	De Smet 2017 - Belgium road-km/road-time/air - 50-2750 visits	50
14		VRPTW	goeke 2018	92
15		PVRPTW	Cordeau_al_2001_PVRPTW	20
16		PVRPTW	Vidal_al_2013_PVRPTW	28
17		PVRPTW	Pirkwieser and Raidl 2009	45
18		MDVRPTW	Cordeau_al_2001_MDVRPTW	20
19		MDVRPTW	Vidal_al_2013_MDVRPTW	28
20		MDVRPTW	Aliahmadi et al. 2021 — Set 1	6
21		MDPVRPTW	Vidal_al_2014_MDPVRPTW	40

TABLE 10. Benchmarks for GVRP, MDVRP, SBVRP, SDVRP, SDVRPTW, and VRPB.

Index	Reference	Variant	Dataset Benchmark	No. of instances
1		G-VRP	AB	40
2		G-VRP	Christofides et al. 1979	42
3		G-VRP	Golden et al. 1998 — Set 2	25
4		G-VRP	goeke 2018	92
5		G-VRP	Erdogan and Miller-Hooks 2012	52
6		G-VRP	Koç and Karaoglan 2016	52
7		MDVRP	De Smet 2017 - Belgium road-km/road-time/air - 50-2750 visits	50
8		MDVRP	Cordeau_al_1997_MDVRP	33
9		MDVRP	Kancharla_Ramadurai_SCS	24
10		SBVRP	PTV 2014	12
11		SBVRP	Absi et al. 2015 — Set 1	20
12		SDVRP (Site-Dependent)	Cordeau_al_1997_SDVRP	35
13		SDVRPTW (Site-Dependent)	Cordeau_al_2001_SDVRPTW	24
14		SDVRPTW (Site-Dependent)	Vidal_al_2013_SDVRPTW	28
15		VRPB	Goetschalckx and Jacobs-Blecha 1989	62
16		VRPB	Toth and Vigo 1999	40
17		VRPB	Queiroga et al. 2019	300

that the efficacy of ML-based approaches is strongly dependent on the quality of the data used for training. The robustness and validity of the data gathering process are important consideration that are sometimes disregarded in ML-based VRP techniques. Recent research, such as what is performed in [134] shows that seemingly random sampling approaches might introduce potential biases and errors during data collection. As a result, evaluating the data gathering strategies employed in existing research is critical to ensuring the reliability and accuracy of machine learning models. Validation of ML-based solutions is critical to guaranteeing their practical application. Cross-validation, simulation, and real-world testing should be used to evaluate model performance. Furthermore, case studies on successful deployments of ML-based VRP systems can provide significant insights into the actual problems and benefits of these technologies.

In the following sections, we will focus on two interlinked key elements: benchmark datasets and available solvers, which are often overlooked in the literature. Figure 9 provides an overview of the system studied in the experiments along with the practical guideline [135].

A. BENCHMARK DATASETS

The computational analysis to evaluate the performance of the proposed solutions to the VRPs is an important part of the related research. The datasets containing the instances to be tested are mentioned in the literature. This section provides a quick overview of existing benchmarks for various kinds of vehicle routing problems.[1]

We begin with the original CVRP, for which benchmarks are provided from 1995 to 2018, ranging from one to 240 instances. Table 8 provides a summary of these instances.

Another feature that is taken into account in various VRP variants is the time window. The covered variants are

[1]We believe that all the frequently addressed datasets are covered; however, there might be still some benchmarks that have been missed

TABLE 11. Benchmarks for other variants of the VRP.

Index	Reference	Variant	Dataset Benchmark	No. of instances
1		ACVRP Asymmetric Capacitated Vehicle Routing Problem	De Smet 2017 - Belgium road-km/road-time/air - 50-2750 visits	50
2		ACVRP Asymmetric Capacitated Vehicle Routing Problem	FTV1994	8
3		CARP-TP Capacitated Arc Routing Problem with Turn Penalties	Vidal_2017_CARP-TP	42
4		ColSVRP A multi-partner variant of the Selective Vehicle Routing Problem	Defryn et al. 2016 — clustered	390
5		ColSVRP A multi-partner variant of the Selective Vehicle Routing Problem	Defryn et al. 2016 — distance	390
6		ColSVRP A multi-partner variant of the Selective Vehicle Routing Problem	Defryn et al. 2016 — random	390
7		conMCVRP Consistent Multi-Compartment Vehicle Routing Problem	Martins et al. 2018	27
8		ConVRP Consistent Vehicle Routing Problem	ConVRP_small	10
9		ConVRP Consistent Vehicle Routing Problem	ConVRP_0.7	12
10		ConVRP Consistent Vehicle Routing Problem	ConVRP_extended	24
11		COP Clustered Orienteering Problem	Angelelli et al. 2014	924
12		CTTRP Capacitated truck-and-trailer routing problem	Bartolini and Schneider 2018	168
13		CVRPTWUPI Capacitated Vehicle Routing Problem with Time Windows, Unmatched pickups and deliveries	Gromicho et al. 2015	6
14		CVTSP Carrier-Vehicle Traveling Salesman Problem	Gambella et al. 2016	72
15		DOPLRP Doubly Open Park-and-loop Routing Problem	Cabrera et al. 2021	60
16		E-VRP-NL Electric vehicle routing problem with non-linear charging function	Montoya et al. 2017	120
17		FSM-DARP-RC fleet size and mix dial-a-ride problem with reconfigurable vehicle capacity	Tellez et al. 2018	15
18		GenConVRP Generalized Consistent Vehicle Routing Problem	GenConVRP	46
19		GVRP-MTPR Green Vehicle Routing Problem with Multiple Technologies and Partial Recharges	Felipe et al. 2014	41
20		H^{2}E-FTW Heterogeneous Electric Fleet routing problem with Time Windows and recharging stations	Hiermann et al. 2017 — Set 1	56
21		MC-VRPSD Multicompartment vehicle routing problem with stochastic demands	Mendoza et al. 2010	180
22		MCGRP-TP Mixed Capacitated General Routing Problem with Turn Penalties	Vidal_2017_MCGRP-TP	28
23		MD-TEVRP-DO Multi-Depot Two-Echelon Vehicle Routing Problem with Delivery Options	Zhou et al. 2017	36
24		MDPVRP Multi-Depot Periodic Vehicle Routing Problem	Vidal_al_2012_MDPVRP	10
25		MOGenConVRP multi-objective generalized consistent vehicle routing problem	MOGenConVRP	46
26		MOSSP Multi-objective shortest path problem	Kergosien et al. 2021	18
27		MTMDVRPTW-DA Multi-Trip Multi-Depot Vehicle Routing Problem with Time Windows and Driver Assignment	Haneyah et al. 2016	3
28		PDPTW-EV Pickup and delivery problem with time windows and electric vehicles	goeke 2018	92
29		POP Probabilistic Orienteering Problem	Angelelli et al. 2017	264
30		PRP Pollution Routing Problem	Kramer et al. 2015	360
31		PRP Pollution Routing Problem	Demir et al. 2012	180
32		PTP Prisoner Transportation Problem	Christiaens et al. 2020	120
33		PVRP Periodic Vehicle Routing Problem	Cordeau_al_1997_PVRP	42
34		TBRD Truck-Based Robots with Depots	Ostermeier et al. 2020b	160
35		TOP Team Orienteering Problem	Archetti et al. 2007	387
36		TRSP Technician Routing and Scheduling Problem	Pillac et al. 2013	57
37		TSPPDDL Travelling Salesman Problem with Pickup, Delivery and Draught Limits	Malaguti et al. 2016	936
38		TWAVRP time windows Assignment Vehicle Routing Problem	Dalmeijer and Spliet 2018	90
39		TWAVRP time windows Assignment Vehicle Routing Problem	Dalmeijer and Desaulniers 2020	190
40		VRP-SL Vehicle Routing Problem with Service Level Constraints	Bulhões et al. 2017	180
41		VRPFT Vehicle Routing Problem with Floating Targets	Gambella et al. 2016c — Set 1	18
42		VRPFT with fixed line moving directions	Gambella et al. 2016c — Set 2	18
43		VRPSD Vehicle routing problem with stochastic demands	Christiansen and Lysgaard 2007	40
44		VRPSD-DC (CC) Vehicle routing problem with stochastic demands and duration constraints (chance constraint formulation)	Mendoza et al. 2016	39
45		VRPSD-DC (LP) Vehicle routing problem with stochastic demands and duration constraints (linear penalty formulation)	Mendoza et al. 2016	39
46		VRPSD-DC (PP) Vehicle routing problem with stochastic demands and duration constraints (piecewise linear penalty formulation)	Mendoza et al. 2016	39
47		VRPSD-DC (QP) Vehicle routing problem with stochastic demands and duration constraints (quadratic penalty formulation)	Mendoza et al. 2016	39

the vehicle routing problem with time windows (VRPTW), periodic vehicle routing problem with time windows (PVRPTW), multi-depot vehicle routing problem with time windows (MDVRPTW), and multi-depot periodic vehicle

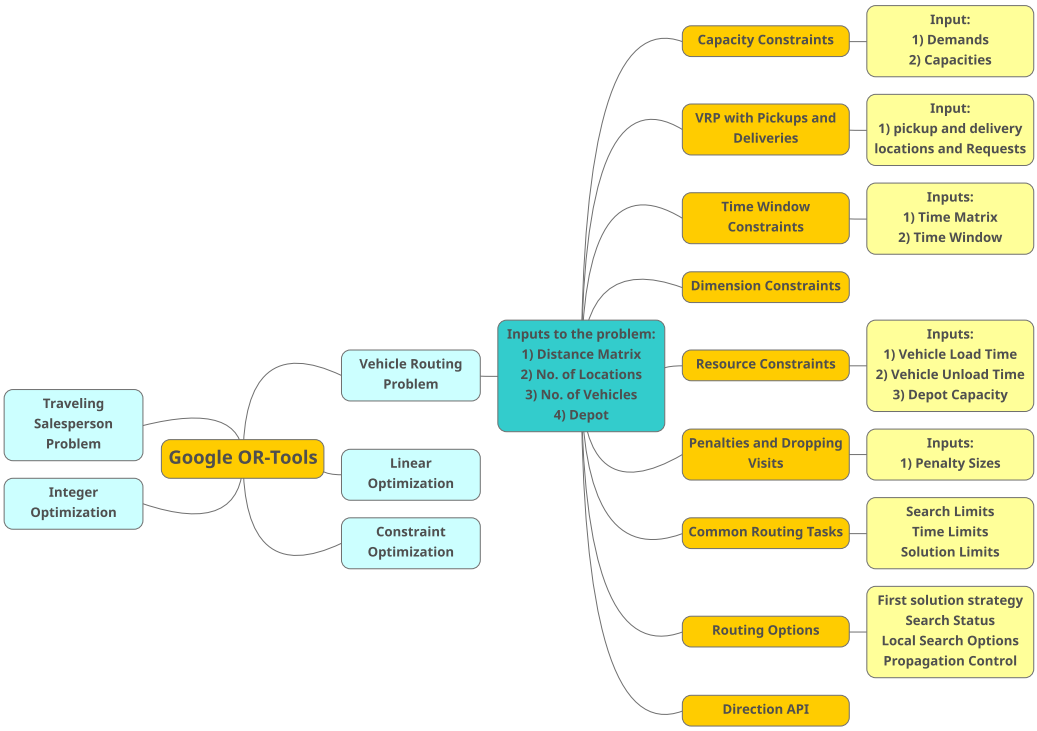


FIGURE 10. The capabilities of Google OR-Tools, which are widely utilized in VRP-related research.

routing problem with time windows (MDVRPTW), which are summarized in Table 9.

Table 10 displays the dataset benchmarks for the green vehicle routing problem (GVRP) variant, multi depot vehicle routing problem (MDVRP), vehicle routing problem with backhauls (VRPB), swap-body vehicle routing problem (SBVRP), site-dependent vehicle routing problem (SDVRP), and SDVRP with time windows (SDVRPTW).

Table 11 summarizes the remaining available benchmarks in the literature.

B. SOLVER PACKAGES

In what follows, we briefly describe some of the most well-known software products used for the VRPs. Almost the majority of them are based on heuristics, because VRP is NP-Hard and no exact method can be guaranteed to identify optimal solutions in reasonable computational time, particularly when dealing with large-scale problems. The majority of studies in the literature focus on open-source software products.

1) GOOGLE OR-TOOLS

It is the most commonly used software solution [136], particularly in ML-based analyses of VRPs. OR-Tools is a free and open-source software that includes solvers for constraint programming, linear and mixed-integer programming, vehicle routing, and graph algorithms. The vehicle routing package includes solvers for the traveling salesperson problem, the vehicle routing problem, the VRP with capacity constraints, the VRP with time windows, the VRP with resource constraints (such as space or personnel to load and unload vehicles at the depot), and the VRP with dropped visits, in which the vehicles are not required to visit all locations but must pay a penalty for each visit that is dropped. The codes are provided in Python, C++, Java, and C-sharp. It is possible to modify the codes, such as the search method. Sample solutions could be used as a benchmark for the aforementioned VRPs. Figure 10 depicts the capabilities of the OR-Tool.

2) LKH-3

It is a VRP solver recognized in the ML community [137]. The broad application of benchmarks from the literature demonstrates the effectiveness of this solver. The LKH-3 is an open-source, C programming language implementation that is portable across a variety of computer platforms and supports many VRPs, as shown in Figure 11.

3) VRPSOLVER

It uses the exact methods of branch-cut-and-price. This VRPSolver interface is available in Julia v1.4.2 and can be used in the following scenarios [138]:

- Comparing the heuristic algorithms with the lower bound/optimal solution.
- Using the Exact algorithms as a benchmark.
- Developing and testing effective models for novel vehicle routing problems.

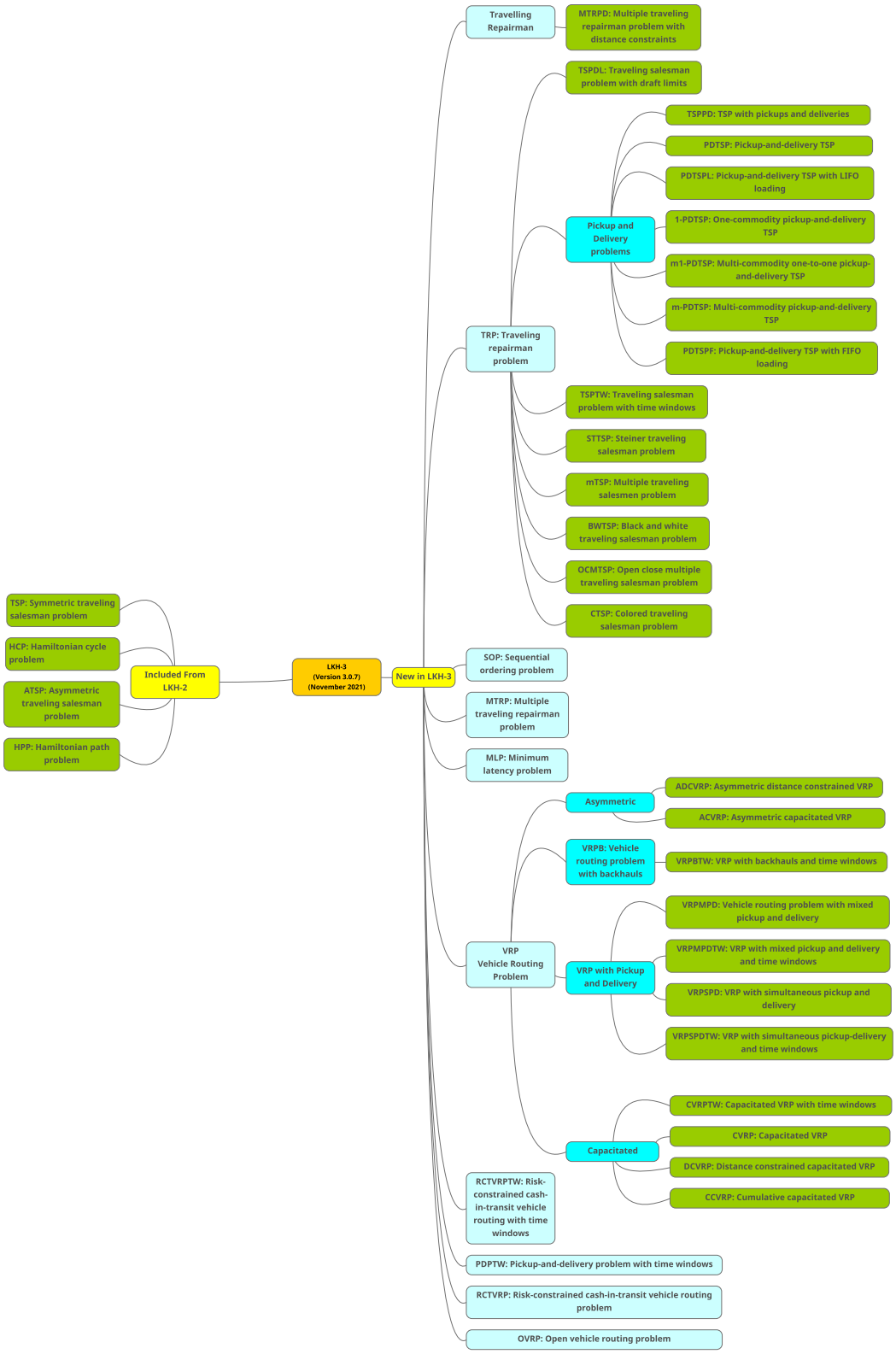


FIGURE 11. The LKH-3 capabilities that are widely employed in VRP and TSP research [137].

The VRPSolver license is only for academic use. The associated software package is offered as a Docker image, which may be run on MacOS, Ubuntu, or Windows. Docker Toolbox can be used by users of MacOS and Windows computers that do not meet the Docker requirements. VRPSolver makes use of BaPCod, a C++ package that implements generic

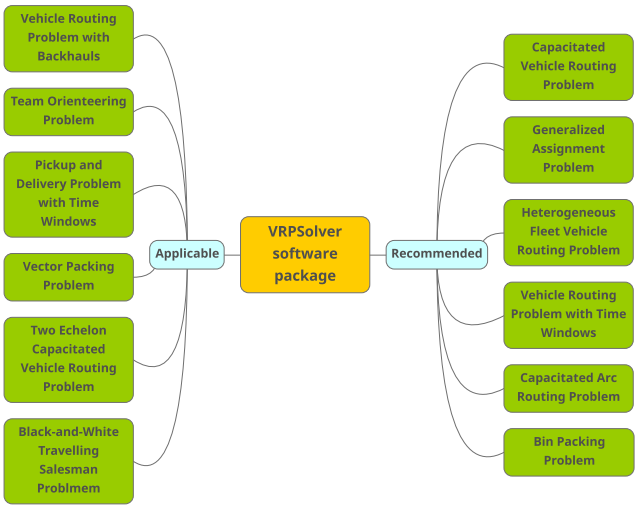


FIGURE 12. The capabilities of the VRPSolver package used for solving the VRPs [138].

Branch-Cut-and-Price operations. Figure 12 summarizes some of the VRPSolver versions that have been solved. Figure 12 offers an overview of several variants of the VRP addressed by using the VRPSolver.

4) VRP SPREADSHEET SOLVER—MICROSOFT

It is an open-source tool for solving and visualizing the outcomes of VRPs. The VRP Spreadsheet solver takes advantage of publicly available GIS and meta-heuristics to solve the original VRP with up to 200 customers [139]. A Windows-based software known as the Larry Snyder VRP solver is also provided, which implements a randomized version of the Clarke-Wright savings algorithm and uses a list of customer locations (latitude and longitude) and demand as input to display the results in graphical (map) form.

While the literature lists numerous software solutions, few are considered in our examined literature, such as the ''ArcGIS Network Analyst extension'', TSP Solve by Gurobi, and ''Wolfram Demonstrations Project''.

Beyond the solvers detailed in this work, there exist distinctive software packages within the Git community specifically designed for addressing VRPs and their variations. Our GitHub search has generated over 1000 repository results.

We believe that existing software package solutions cannot be directly applied to new research areas, but can serve as benchmarks for ML-based solution evaluations.

VI. CONCLUSION

The purpose of this study was to determine why and how machine learning can help solve vehicle routing problems or improve existing solutions. As discussed, majority of the existing research use exact methodologies or heuristics, which are preferable in small-scale scenarios. Heuristic-based solutions are known for their speed and low processing costs, but there is no guarantee that they will produce the best solution. We investigated the application of ML methods for solving VRPs. We reviewed the existing related survey publications (22 papers). We analyzed the associated literature from 2020 to 2023, as well as some important studies from 2015 to 2020. We presented a mathematical overview, including the fundamentals for reinforcement learning, the Markov decision process, Attention mode, Pointer Networks, and graph neural networks. We further categorized the studies covered from various perspectives. We divided machine learning-based methods into two broad categories: RL-based methods and non-RL-based methods. We considered supplementary categorization in RL-based approaches based on the characteristic of the problem in which such technology is used, i.e., VRP with loading, delivery, and pickup constraints, UAVs, Green and electric vehicles, large-scale VRPs, and dynamic VRPs. We also categorized the studies based on their implemented algorithms and showed that the majority of the studies utilize RL. We also included a comprehensive section on implementation guidelines, which include both the benchmark dataset and open-source solvers. The key advantage of ML-based approaches is their ability to deliver solutions for large-scale VRP instances. However, these strategies are dependent on the quality and availability of data. As a result, scalable strategies that might improve data collecting and hence boost model robustness remain as a future research path. Furthermore, investigating hybrid approaches that combine ML and conventional optimization techniques may result in significant advancements in solving complex VRPs, particularly in terms of computational complexity. Some fundamental concepts that ML approaches can address in future include addressing homogeneous vehicles, determining the optimal number of vehicles, incorporating integrated information into online decision-making, increasing heuristic performance, capturing more complex and realistic scenarios, using generative adversarial networks (GANs), transfer learning, new concepts in combination with RL-based methods, and new paradigms of responsible AI or explainable AI.

REFERENCES

[1] G. B. Dantzig and J. H. Ramser, ''The truck dispatching problem,'' *Manag. Sci.*, vol. 6, no. 1, pp. 80–91, Oct. 1959.

[2] J. K. Lenstra and A. H. G. R. Kan, ''Complexity of vehicle routing and scheduling problems,'' *Networks*, vol. 11, no. 2, pp. 221–227, Jun. 1981.

[3] O. Bräysy and M. Gendreau, ''Vehicle routing problem with time windows, Part I: Route construction and local search algorithms,'' *Transp. Sci.*, vol. 39, no. 1, pp. 104–118, Feb. 2005.

[4] N. Giuffrida, J. Fajardo-Calderin, A. D. Masegosa, F. Werner, M. Steudter, and F. Pilla, ''Optimization and machine learning applied to last-mile logistics: A review,'' *Sustainability*, vol. 14, no. 9, p. 5329, 2022.

[5] J. Fellers, J. Quevedo, M. Abdelatti, M. Steinhaus, and M. Sodhi, ''Selecting between evolutionary and classical algorithms for the CVRP using machine learning: Optimization of vehicle routing problems,'' in *Proc. Genetic Evol. Comput. Conf. Companion*, Jul. 2021, pp. 127–128.

[6] H. Pollaris, K. Braekers, A. Caris, G. K. Janssens, and S. Limbourg, ''Vehicle routing problems with loading constraints: State-of-the-art and future directions,'' *OR Spectr.*, vol. 37, no. 2, pp. 297–330, Mar. 2015.

[7] I. Khoufi, A. Laouiti, and C. Adjih, ''A survey of recent extended variants of the traveling salesman and vehicle routing problems for unmanned aerial vehicles,'' *Drones*, vol. 3, pp. 1–30, Sep. 2019.

[8] R. Elshaer and H. Awad, "A taxonomic review of metaheuristic algorithms for solving the vehicle routing problem and its variants," *Comput. Ind. Eng.*, vol. 140, Feb. 2020, Art. no. 106242.

[9] T. Vidal, G. Laporte, and P. Matl, "A concise guide to existing and emerging vehicle routing problem variants," *Eur. J. Oper. Res.*, vol. 286, no. 2, pp. 401–416, Oct. 2020.

[10] M. Sánchez, J. M. Cruz-Duarte, J. C. Ortíz-Bayliss, H. Ceballos, H. Terashima-Marin, and I. Amaya, "A systematic review of hyper-heuristics on combinatorial optimization problems," *IEEE Access*, vol. 8, pp. 128068–128095, 2020.

[11] A. Thibbotuwawa, G. Bocewicz, P. Nielsen, and Z. Banaszak, "Unmanned aerial vehicle routing problems: A literature review," *Appl. Sci.*, vol. 10, no. 13, p. 4504, Jun. 2020.

[12] E. Ghorbani, M. Alinaghian, G. B. Gharehpetian, S. Mohammadi, and G. Perboli, "A survey on environmentally friendly vehicle routing problem and a proposal of its classification," *Sustainability*, vol. 12, no. 21, p. 9079, Oct. 2020.

[13] W. K. Anuar, L. S. Lee, S. Pickl, and H.-V. Seow, "Vehicle routing optimisation in humanitarian operations: A survey on modelling and optimisation approaches," *Appl. Sci.*, vol. 11, no. 2, p. 667, Jan. 2021.

[14] M. Ostermeier, T. Henke, A. Hübner, and G. Wäscher, "Multi-compartment vehicle routing problems: State-of-the-art, modeling framework and future directions," *Eur. J. Oper. Res.*, vol. 292, no. 3, pp. 799–817, Aug. 2021.

[15] S. Y. Tan and W. C. Yeh, "The vehicle routing problem: State-of-the-art classification and review," *Appl. Sci.*, vol. 11, p. 10295, Nov. 2021.

[16] M. Asghari and S. M. J. Mirzapour Al-E-Hashem, "Green vehicle routing problem: A state-of-the-art review," *Int. J. Prod. Econ.*, vol. 231, Jan. 2021, Art. no. 107899.

[17] Y. Xiao, Y. Zhang, I. Kaku, R. Kang, and X. Pan, "Electric vehicle routing problem: A systematic review and a new comprehensive model with nonlinear energy recharging and consumption," *Renew. Sustain. Energy Rev.*, vol. 151, Nov. 2021, Art. no. 111567.

[18] A. Gutiérrez-Sánchez and L. B. Rocha-Medina, "VRP variants applicable to collecting donations and similar problems: A taxonomic review," *Comput. Ind. Eng.*, vol. 164, Feb. 2022, Art. no. 107887.

[19] N. Soeffker, M. W. Ulmer, and D. C. Mattfeld, "Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review," *Eur. J. Oper. Res.*, vol. 298, no. 3, pp. 801–820, May 2022.

[20] N. Sluijk, A. M. Florio, J. Kinable, N. Dellaert, and T. Van Woensel, "Two-echelon vehicle routing problems: A literature review," *Eur. J. Oper. Res.*, vol. 304, no. 3, pp. 865–886, Feb. 2023.

[21] D. Fleckenstein, R. Klein, and C. Steinhardt, "Recent advances in integrating demand management and vehicle routing: A methodological review," *Eur. J. Oper. Res.*, vol. 306, no. 2, pp. 499–518, 2022.

[22] G. E. A. Fröhlich, M. Gansterer, and K. F. Doerner, "Safe and secure vehicle routing: A survey on minimization of risk exposure," *Int. Trans. Oper. Res.*, vol. 30, no. 6, pp. 3087–3121, 2022.

[23] A. Mor and M. G. Speranza, "Vehicle routing problems over time: A survey," *Ann. Oper. Res.*, vol. 314, pp. 255–275, Jul. 2022.

[24] Y.-J. Liang and Z.-X. Luo, "A survey of Truck–Drone routing problem: Literature review and research prospects," *J. Oper. Res. Soc. China*, vol. 10, no. 2, pp. 343–377, Jun. 2022.

[25] K. Corona-Gutiérrez, S. Nucamendi-Guillén, and E. Lalla-Ruiz, "Vehicle routing with cumulative objectives: A state of the art and analysis," *Comput. Ind. Eng.*, vol. 169, Jul. 2022, Art. no. 108054.

[26] G. Macrina, L. D. P. Pugliese, and F. Guerriero, "The green-vehicle routing problem: A survey," in *Modeling and Optimization in Green Logistics*. Berlin, Germany: Springer, 2020, pp. 1–26.

[27] R. Bai, X. Chen, Z.-L. Chen, T. Cui, S. Gong, W. He, X. Jiang, H. Jin, J. Jin, G. Kendall, J. Li, Z. Lu, J. Ren, P. Weng, N. Xue, and H. Zhang, "Analytics and machine learning in vehicle routing research," *Int. J. Prod. Res.*, vol. 61, no. 1, pp. 4–30, Jan. 2023.

[28] G. Macrina, L. D. P. Pugliese, F. Guerriero, and G. Laporte, "Drone-aided routing: A literature review," *Transp. Res. C, Emerg. Technol.*, vol. 120, Nov. 2020, Art. no. 102762.

[29] S. Elatar, K. Abouelmehdi, and M. E. Riffi, "The vehicle routing problem in the last decade: Variants, taxonomy and metaheuristics," *Proc. Comput. Sci.*, vol. 220, pp. 398–404, Jan. 2023.

[30] R. Shi and L. Niu, "A brief survey on learning based methods for vehicle routing problems," *Proc. Comput. Sci.*, vol. 221, pp. 773–780, Jan. 2023.

[31] J. Oyola, H. Arntzen, and D. L. Woodruff, "The stochastic vehicle routing problem, a literature review, Part I: Models," *EURO J. Transp. Logistics*, vol. 7, no. 3, pp. 193–221, Sep. 2018.

[32] K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse, "The vehicle routing problem: State of the art classification and review," *Comput. Ind. Eng.*, vol. 99, pp. 300–313, Sep. 2016.

[33] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. North Chelmsford, MA, USA: Courier Corporation, 1998.

[34] C. Solnon, *Ant Colony Optimization and Constraint Programming*. Hoboken, NJ, USA: Wiley, 2013.

[35] A. Abdelshafie, M. Salah, T. Kramberger, and D. Dragan, "Repositioning and optimal re-allocation of empty containers: A review of methods, models, and applications," *Sustainability*, vol. 14, no. 11, p. 6655, May 2022.

[36] B. Moradi, "The new optimization algorithm for the vehicle routing problem with time windows using multi-objective discrete learnable evolution model," *Soft Comput.*, vol. 24, no. 9, pp. 6741–6769, May 2020.

[37] J. Poullet, "Leveraging machine learning to solve the vehicle routing problem with time windows," Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2020.

[38] N. Mladenović and P. Hansen, "Variable neighborhood search," *Comput. Oper. Res.*, vol. 24, no. 11, pp. 1097–1100, Nov. 1997.

[39] V. C. Hemmelmayr, K. F. Doerner, and R. F. Hartl, "A variable neighborhood search heuristic for periodic routing problems," *Eur. J. Oper. Res.*, vol. 195, no. 3, pp. 791–802, Jun. 2009.

[40] M. Polacek, R. F. Hartl, K. Doerner, and M. Reimann, "A variable neighborhood search for the multi depot vehicle routing problem with time windows," *J. Heuristics*, vol. 10, no. 6, pp. 613–627, Dec. 2004.

[41] P. Shaw, "Using constraint programming and local search methods to solve vehicle routing problems," in *Proc. 4th Int. Conf. Princ. Pract. Constraint Programming* (Lecture Notes in Computer Science), vol. 1520, 1991, pp. 417–431.

[42] A. Arockia, M. Lochbrunner, T. Hanne, and R. Dornberger, "Benchmarking Tabu search and simulated annealing for the capacitated vehicle routing problem," in *Proc. 4th Int. Conf. Comput. Manag. Bus.*, Jan. 2021, pp. 118–124.

[43] U. Ritzinger, J. Puchinger, and R. F. Hartl, "A survey on dynamic and stochastic vehicle routing problems," *Int. J. Prod. Res.*, vol. 54, no. 1, pp. 215–231, Jan. 2016.

[44] G. D. Konstantakopoulos, S. P. Gayialis, and E. P. Kechagias, "Vehicle routing problem and related algorithms for logistics distribution: A literature review and classification," *Oper. Res.*, vol. 22, no. 3, pp. 2033–2062, Jul. 2022.

[45] J. G. Cavalcanti Costa, Y. Mei, and M. Zhang, "Learning to select initialisation heuristic for vehicle routing problems," in *Proc. Genetic Evol. Comput. Conf.*, Jul. 2023, pp. 266–274.

[46] P. Yue, S. Liu, and Y. Jin, "Graph Q-learning assisted ant colony optimization for vehicle routing problems with time windows," in *Proc. Companion Conf. Genetic Evol. Comput.*, Jul. 2023, pp. 7–8.

[47] P. Mukherjee and S. Dey, "Efficient vehicle routing problem: A machine learning and evolutionary computation approach," in *Proc. Companion Conf. Genetic Evol. Comput.*, Jul. 2023, pp. 3–4.

[48] W. Yi, R. Qu, L. Jiao, and B. Niu, "Automated design of metaheuristics using reinforcement learning within a novel general search framework," *IEEE Trans. Evol. Comput.*, vol. 27, no. 4, pp. 1072–1084, Aug. 2023.

[49] N. Frías, F. Johnson, and C. Valle, "Hybrid algorithms for energy minimizing vehicle routing problem: Integrating clusterization and ant colony optimization," *IEEE Access*, vol. 11, pp. 125800–125821, 2023.

[50] M. Nazari, A. Oroojlooy, L. Snyder, and M. Takác, "Reinforcement learning for solving the vehicle routing problem," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[51] Y. Zou, H. Wu, Y. Yin, L. Dhamotharan, D. Chen, and A. K. Tiwari, "An improved transformer model with multi-head attention and attention to attention for low-carbon multi-depot vehicle routing problem," *Ann. Oper. Res.*, Jun. 2022, doi: 10.1007/s10479-022-04788-z.

[52] K. Zhang, F. He, Z. Zhang, X. Lin, and M. Li, "Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach," *Transp. Res. C, Emerg. Technol.*, vol. 121, Dec. 2020, Art. no. 102861.

[53] F. Guo, Q. Wei, M. Wang, Z. Guo, and S. W. Wallace, "Deep attention models with dimension-reduction and gate mechanisms for solving practical time-dependent vehicle routing problems," *Transp. Res. E, Logistics Transp. Rev.*, vol. 173, May 2023, Art. no. 103095.

[54] A. Bdeir, J. K. Falkner, and L. Schmidt-Thieme, "Attention, filling in the gaps for generalization in routing problems," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 13718, M. R. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, and G. Tsoumakas, Eds., Cham, Switzerland: Springer, 2023, pp. 505–520, doi: 10.1007/978-3-031-26422-1_31.

[55] J. J. Q. Yu, W. Yu, and J. Gu, "Online vehicle routing with neural combinatorial optimization and deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3806–3817, Oct. 2019.

[56] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[57] H. Huang, Y.-S. Lin, J.-R. Kang, and C.-C. Lin, "A deep reinforcement learning approach for crowdshipping vehicle routing problem," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manag. (IEEM)*, Dec. 2022, pp. 0598–0599.

[58] J. Rubin, O. Shamir, and N. Tishby, "Trading value and information in MDPs," in *Decision Making With Imperfect Decision Makers*. Berlin, Germany: Springer, 2012, pp. 57–74.

[59] L. Ren, X. Fan, J. Cui, Z. Shen, Y. Lv, and G. Xiong, "A multi-agent reinforcement learning method with route recorders for vehicle routing in supply chain management," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16410–16420, Sep. 2022.

[60] T. Phiboonbanakit, T. Horanont, V.-N. Huynh, and T. Supnithi, "A hybrid reinforcement learning-based model for the vehicle routing problem in transportation logistics," *IEEE Access*, vol. 9, pp. 163325–163347, 2021.

[61] D. Tamagawa, E. Taniguchi, and T. Yamada, "Evaluating city logistics measures using a multi-agent model," *Proc.-Social Behav. Sci.*, vol. 2, no. 3, pp. 6002–6012, 2010.

[62] O. Bouhamed, H. Ghazzai, H. Besbes, and Y. Massoud, "Q-learning based routing scheduling for a multi-task autonomous agent," in *Proc. IEEE 62nd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2019, pp. 634–637.

[63] W. Pan and S. Q. Liu, "Deep reinforcement learning for the dynamic and uncertain vehicle routing problem," *Int. J. Speech Technol.*, vol. 53, no. 1, pp. 405–422, Jan. 2023.

[64] W. Kool, H. van Hoof, and M. Welling, "Attention, learn to solve routing problems!" 2018, *arXiv:1803.08475*.

[65] A. Hottung and K. Tierney, "Neural large neighborhood search for the capacitated vehicle routing problem," 2019, *arXiv:1911.09539*.

[66] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, "Neural combinatorial optimization with reinforcement learning," 2016, *arXiv:1611.09940*.

[67] J. Zhao, M. Mao, X. Zhao, and J. Zou, "A hybrid of deep reinforcement learning and local search for the vehicle routing problems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7208–7218, Nov. 2021.

[68] F. Hagström, "Finding solutions to the vehicle routing problem using a graph neural network," Ph.D. thesis, School Sci., Aalto Univ., Espoo, Finland, 2022.

[69] Q. Cappart, D. Chételat, E. Khalil, A. Lodi, C. Morris, and P. Veličković, "Combinatorial optimization and reasoning with graph neural networks," 2021, *arXiv:2102.09544*.

[70] Í. Santana, A. Lodi, and T. Vidal, "Neural networks for local search and crossover in vehicle routing: A possible overkill?" in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (Lecture Notes in Computer Science), vol. 13884, A. A. Cire, Ed., Cham, Switzerland: Springer, 2023, pp. 184–199, doi: 10.1007/978-3-031-33271-5_13.

[71] V. S. K. M. Vamsi, Y. R. Telukuntla, P. S. Kumar, and G. Gutjahr, "Comparison of attention mechanisms in machine learning models for vehicle routing problems," in *Machine Learning and Computational Intelligence Techniques for Data Engineering* (Lecture Notes in Electrical Engineering), vol. 998, P. Singh, D. Singh, V. Tiwari, and S. Misra, Eds., Singapore: Springer, 2023, pp. 629–638, doi: 10.1007/978-981-99-0047-3_53.

[72] Y. Niu, D. Kong, R. Wen, Z. Cao, and J. Xiao, "An improved learnable evolution model for solving multi-objective vehicle routing problem with stochastic demand," *Knowl.-Based Syst.*, vol. 230, Oct. 2021, Art. no. 107378.

[73] N. Furian, M. O'Sullivan, C. Walker, and E. Çela, "A machine learning-based branch and price algorithm for a sampled vehicle routing problem," *OR Spectr.*, vol. 43, no. 3, pp. 693–732, Sep. 2021.

[74] S. M. Darwish and B. E. Abdel-Samee, "Game theory based solver for dynamic vehicle routing problem," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl. (AMLTA)*, in Advances in Intelligent Systems and Computing, vol. 921, A. Hassanien, A. Azar, T. Gaber, R. Bhatnagar, and M. F. Tolba, Eds., Cham, Switzerland: Springer, 2020, pp. 133–142, doi: 10.1007/978-3-030-14118-9_14.

[75] Y. Wang, L. Ran, X. Guan, J. Fan, Y. Sun, and H. Wang, "Collaborative multicenter vehicle routing problem with time windows and mixed deliveries and pickups," *Exp. Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116690.

[76] C. Legrand, L. Jourdan, M.-E. Kessaci, D. Cattaruzza, "Machine learning for multi-objective problems," in *Proc. 23ème Congrès Annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF)*. Villeurbanne, France: INSA Lyon, Feb. 2022.

[77] M. Morabit, G. Desaulniers, and A. Lodi, "Machine-learning-based arc selection for constrained shortest path problems in column generation," 2022, *arXiv:2201.02535*.

[78] J. Mandi, R. Canoy, V. Bucarey, and T. Guns, "Data driven VRP: A neural network model to learn hidden preferences for VRP," 2021, *arXiv:2108.04578*.

[79] S. Deb, A. K. Goswami, B. G. Hajra, R. L. Chetri, M. Roy, and R. Roy, "Development of machine learning based state-of-charge prediction model for plug-in electric vehicle's relocation in sharing system," *IEEE Trans. Ind. Appl.*, vol. 59, no. 4, pp. 4662–4672, Jul. 2023.

[80] S. Hansuwa, M. R. Velayudhan Kumar, and R. Chandrasekharan, "Analysis of box and ellipsoidal robust optimization, and attention model based reinforcement learning for a robust vehicle routing problem," *Sādhanā*, vol. 47, no. 2, pp. 1–23, Jun. 2022.

[81] J. Li, L. Xin, Z. Cao, A. Lim, W. Song, and J. Zhang, "Heterogeneous attentions for solving pickup and delivery problem via deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2306–2315, Mar. 2022.

[82] J. Chen, Z. Zong, Y. Zhuang, H. Yan, D. Jin, and Y. Li, "Reinforcement learning for practical express systems with mixed deliveries and pickups," *ACM Trans. Knowl. Discovery Data*, vol. 17, no. 3, pp. 1–19, Feb. 2023.

[83] H. Qiu, S. Wang, Y. Yin, D. Wang, and Y. Wang, "A deep reinforcement learning-based approach for the home delivery and installation routing problem," *Int. J. Prod. Econ.*, vol. 244, Feb. 2022, Art. no. 108362.

[84] A. Bortfeldt, T. Hahn, D. Männel, and L. Mönch, "Hybrid algorithms for the vehicle routing problem with clustered backhauls and 3D loading constraints," *Eur. J. Oper. Res.*, vol. 243, no. 1, pp. 82–96, May 2015.

[85] E. Göçmen and R. Erol, "Transportation problems for intermodal networks: Mathematical models, exact and heuristic algorithms, and machine learning," *Exp. Syst. Appl.*, vol. 135, pp. 374–387, Nov. 2019.

[86] S. Reil, A. Bortfeldt, and L. Mönch, "Heuristics for vehicle routing problems with backhauls, time windows, and 3D loading constraints," *Eur. J. Oper. Res.*, vol. 266, no. 3, pp. 877–894, May 2018.

[87] X. Chen, M. W. Ulmer, and B. W. Thomas, "Deep Q-learning for same-day delivery with vehicles and drones," *Eur. J. Oper. Res.*, vol. 298, no. 3, pp. 939–952, May 2022.

[88] J.-A. Delamer and S. Givigi, "A system based on deep-learning for dynamic routing problems," in *Proc. IEEE Int. Syst. Conf. (SysCon)*, Apr. 2022, pp. 1–7.

[89] H. Wang, S. Yuan, M. Guo, C.-Y. Chan, X. Li, and W. Lan, "Tactical driving decisions of unmanned ground vehicles in complex highway environments: A deep reinforcement learning approach," *Proc. Inst. Mech. Eng., D, J. Automobile Eng.*, vol. 235, no. 4, pp. 1113–1127, Mar. 2021.

[90] W. Wang, Y. Liu, R. Srikant, and L. Ying, "3M-RL: Multi-resolution, multi-agent, mean-field reinforcement learning for autonomous UAV routing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8985–8996, Jul. 2022.

[91] M. Alqahtani and M. Hu, "Dynamic energy scheduling and routing of multiple electric vehicles using deep reinforcement learning," *Energy*, vol. 244, Apr. 2022, Art. no. 122626.

[92] R. Basso, B. Kulcsár, I. Sanchez-Diaz, and X. Qu, "Dynamic stochastic electric vehicle routing with safe reinforcement learning," *Transp. Res. E, Logistics Transp. Rev.*, vol. 157, Jan. 2022, Art. no. 102496.

[93] S. S. Eshkevari, S. S. Eshkevari, S. N. Pakzad, H. Muñoz-Avila, and S. Kishore, "Routing of public and electric transportation systems using reinforcement learning," in *Data Science in Engineering, Volume 9*. Berlin, Germany: Springer, 2022, pp. 263–273.

[94] F. N. Al-Wesabi, A. A. Albraikan, A. M. Hilal, M. M. Eltahir, M. A. Hamza, and A. S. Zamani, "Fleet optimization of smart electric motorcycle system using deep reinforcement learning," *Comput., Mater. Continua*, vol. 71, no. 1, pp. 1925–1943, 2022.

[95] R. Basso, B. Kulcsár, and I. Sanchez-Diaz, "Electric vehicle routing problem with machine learning for energy prediction," *Transp. Res. B, Methodol.*, vol. 145, pp. 24–55, Mar. 2021.

[96] T. M. Aljohani, A. Ebrahim, and O. Mohammed, "Real-time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model," *Electric Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106962.

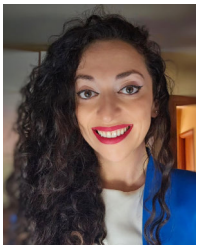
[97] A. Sarker, H. Shen, and K. Kowsari, "A data-driven reinforcement learning based multi-objective route recommendation system," in *Proc. IEEE 17th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Dec. 2020, pp. 103–111.
[98] A. Bdeir, S. Boeder, T. Dernedde, K. Tkachuk, J. K. Falkner, and L. Schmidt-Thieme, "RP-DQN: An application of Q-learning to vehicle routing problems," in *Proc. German Conf. Artif. Intell.*, 2021, pp. 3–16.
[99] W. K. Anuar, L. S. Lee, H.-V. Seow, and S. Pickl, "A multi-depot vehicle routing problem with stochastic road capacity and reduced two-stage stochastic integer linear programming models for rollout algorithm," *Mathematics*, vol. 9, no. 13, p. 1572, Jul. 2021.
[100] A. Gupta, S. Ghosh, and A. Dhara, "Deep reinforcement learning algorithm for fast solutions to vehicle routing problem with time-windows," in *Proc. 5th Joint Int. Conf. Data Sci. Manag. Data*, Jan. 2022, pp. 236–240.
[101] M. Liu, Z. Wang, and J. Li, "A deep reinforcement learning algorithm for large-scale vehicle routing problems," *Proc. SPIE*, vol. 12254, pp. 824–829, May 2022.
[102] C. Lu et al., "Deep reinforcement learning for solving AGVs routing problem," in *Verification and Evaluation of Computer and Communication Systems* (Lecture Notes in Computer Science), vol. 12519, B. B. Hedia, Y. F. Chen, G. Liu, and Z. Yu, Eds., Cham, Switzerland: Springer, 2020, pp. 222–236, doi: 10.1007/978-3-030-65955-4_16.
[103] J. Jiang, R. Ma, and G. Shen, "Research on parallel distribution routing optimization based on improved reinforcement learning algorithm," in *Proc. IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. (AEECA)*, Aug. 2021, pp. 183–186.
[104] W. Joe and H. C. Lau, "Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers," in *Proc. Int. Conf. Automated Planning Scheduling*, vol. 30, 2020, pp. 394–402.
[105] B. Peng, J. Wang, and Z. Zhang, "A deep reinforcement learning algorithm using dynamic attention model for vehicle routing problems," in *Artificial Intelligence Algorithms and Applications* (Communications in Computer and Information Science), vol. 1205, K. Li, W. Li, H. Wang, and Y. Liu, Eds., Singapore: Springer, 2020, pp. 636–650, doi: 10.1007/978-981-15-5577-0_51.
[106] Y. Xu, M. Fang, L. Chen, G. Xu, Y. Du, and C. Zhang, "Reinforcement learning with multiple relational attention for solving vehicle routing problems," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 11107–11120, Oct. 2022.
[107] K. Manchella, A. K. Umrawal, and V. Aggarwal, "FlexPool: A distributed model-free deep reinforcement learning algorithm for joint passengers and goods transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2035–2047, Apr. 2021.
[108] S. Koh, B. Zhou, H. Fang, P. Yang, Z. Yang, Q. Yang, L. Guan, and Z. Ji, "Real-time deep reinforcement learning based vehicle navigation," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106694.
[109] S. A. Pedersen, B. Yang, and C. S. Jensen, "Anytime stochastic routing with hybrid learning," *Proc. VLDB Endowment*, vol. 13, no. 9, pp. 1555–1567, May 2020.
[110] Z. Yuan, G. Li, Z. Wang, J. Sun, and R. Cheng, "RL-CSL: A combinatorial optimization method using reinforcement learning and contrastive self-supervised learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1010–1024, Aug. 2023.
[111] Z. Shou, X. Chen, Y. Fu, and X. Di, "Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment," *Transp. Res. C, Emerg. Technol.*, vol. 137, Apr. 2022, Art. no. 103560.
[112] G. Bono, J. S. Dibangoye, O. Simonin, L. Matignon, and F. Pereyron, "Solving multi-agent routing problems using deep attention mechanisms," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7804–7813, Dec. 2021.
[113] R. Ding, Z. Yang, Y. Wei, H. Jin, and X. Wang, "Multi-agent reinforcement learning for urban crowd sensing with for-hire vehicles," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2021, pp. 1–10.
[114] K. Ren, F.-X. Xiao, and H.-G. Han, "Dynamic multitask optimization with improved knowledge transfer mechanism," *Int. J. Speech Technol.*, vol. 53, no. 2, pp. 1666–1682, Jan. 2023.
[115] M. Kim and J. Park, "Learning collaborative policies to solve NP-hard routing problems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10418–10430.
[116] J. Li, Y. Ma, R. Gao, Z. Cao, A. Lim, W. Song, and J. Zhang, "Deep reinforcement learning for solving the heterogeneous capacitated vehicle routing problem," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13572–13585, Dec. 2022.
[117] S. Mak, L. Xu, T. Pearce, M. Ostroumov, and A. Brintrup, "Fair collaborative vehicle routing: A deep multi-agent reinforcement learning approach," *Transp. Res. C, Emerg. Technol.*, vol. 157, Dec. 2023, Art. no. 104376.
[118] L. D. P. Pugliese, D. Ferone, P. Festa, F. Guerriero, and G. Macrina, "Combining variable neighborhood search and machine learning to solve the vehicle routing problem with crowd-shipping," *Optim. Lett.*, vol. 17, no. 9, pp. 1981–2003, Dec. 2023.
[119] Y. Wu, W. Song, Z. Cao, J. Zhang, and A. Lim, "Learning improvement heuristics for solving routing problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 5057–5069, Sep. 2022.
[120] B. Chen, R. Qu, R. Bai, and W. Laesanklang, "A variable neighborhood search algorithm with reinforcement learning for a real-life periodic vehicle routing problem with time windows and open routes," *RAIRO-Oper. Res.*, vol. 54, no. 5, pp. 1467–1494, Sep. 2020.
[121] P. da Costa, J. Rhuggenaath, Y. Zhang, A. Akcay, and U. Kaymak, "Learning 2-opt heuristics for routing problems via deep reinforcement learning," *Social Netw. Comput. Sci.*, vol. 2, no. 5, p. 388, Sep. 2021.
[122] L. Xin, W. Song, Z. Cao, and J. Zhang, "Multi-decoder attention model with embedding glimpse for solving vehicle routing problems," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 12042–12049.
[123] A. Delarue, R. Anderson, and C. Tjandraatmadja, "Reinforcement learning with combinatorial actions: An application to vehicle routing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 609–620.
[124] L. Duan, Y. Zhan, H. Hu, Y. Gong, J. Wei, X. Zhang, and Y. Xu, "Efficiently solving the practical vehicle routing problem: A novel joint learning approach," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3054–3063.
[125] F. E. Achamrah, F. Riane, and S. Limbourg, "Solving inventory routing with transshipment and substitution under dynamic and stochastic demands using genetic algorithm and deep reinforcement learning," *Int. J. Prod. Res.*, vol. 60, no. 20, pp. 6187–6204, Oct. 2022.
[126] Y. Zhou, W. Li, X. Wang, Y. Qiu, and W. Shen, "Adaptive gradient descent enabled ant colony optimization for routing problems," *Swarm Evol. Comput.*, vol. 70, Apr. 2022, Art. no. 101046.
[127] R. Zhang, A. Prokhorchuk, and J. Dauwels, "Deep reinforcement learning for traveling salesman problem with time windows and rejections," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.
[128] W. Kool, H. van Hoof, J. Gromicho, and M. Welling, "Deep policy dynamic programming for vehicle routing problems," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (Lecture Notes in Computer Science), vol. 13292, P. Schaus, Ed., Cham, Switzerland: Springer, 2022, pp. 190–213, doi: 10.1007/978-3-031-08011-1_14.
[129] W. Qin, Z. Zhuang, Z. Huang, and H. Huang, "A novel reinforcement learning-based hyper-heuristic for heterogeneous vehicle routing problem," *Comput. Ind. Eng.*, vol. 156, Jun. 2021, Art. no. 107252.
[130] Q. Wang, "VARL: A variational autoencoder-based reinforcement learning framework for vehicle routing problems," *Int. J. Speech Technol.*, vol. 52, no. 8, pp. 8910–8923, Jun. 2022.
[131] A. Yaddaden, S. Harispe, and M. Vasquez, "Is transfer learning helpful for neural combinatorial optimization applied to vehicle routing problems?" *Comput. Informat.*, vol. 41, no. 1, pp. 172–190, 2022.
[132] E. Díaz de León-Hicks, S. E. Conant-Pablos, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "Addressing the algorithm selection problem through an attention-based meta-learner approach," *Appl. Sci.*, vol. 13, no. 7, p. 4601, Apr. 2023.
[133] J. Dornemann, "Solving the capacitated vehicle routing problem with time windows via graph convolutional network assisted tree search and quantum-inspired computing," *Frontiers Appl. Math. Statist.*, vol. 9, Jun. 2023, Art. no. 1155356.
[134] Y. Shang, "Subgraph robustness of complex networks under attacks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 4, pp. 821–832, Apr. 2019.
[135] L. Accorsi, A. Lodi, and D. Vigo, "Guidelines for the computational testing of machine learning approaches to vehicle routing problems," *Oper. Res. Lett.*, vol. 50, no. 2, pp. 229–234, Mar. 2022.
[136] L. Perron and V. Furnon. (2019). *Or-Tools*. [Online]. Available: https://developers.google.com/optimization
[137] K. Helsgaun, "An extension of the Lin–Kernighan–Helsgaun TSP solver for constrained traveling salesman and vehicle routing problems," *Roskilde, Roskilde Univ.*, vol. 12, pp. 24–50, Dec. 2017.

[138] A. Pessoa, R. Sadykov, and E. Uchoa, "Solving bin packing problems using VRPSolver models," in *Operations Research Forum*, vol. 2. Berlin, Germany: Springer, 2021, pp. 1–25.
[139] G. Erdogan, "An open source spreadsheet solver for vehicle routing problems," *Comput. Oper. Res.*, vol. 84, pp. 62–72, Aug. 2017.

REZA SHAHBAZIAN (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Calabria, Italy. He has been a Postdoctoral Research Assistant with Shahid Beheshti University, Iran, and the University of Calabria. He is currently an Assistant Professor with the University of Calabria. His research interests include machine learning applications, neuro-symbolic AI, optimization, and signal processing, with more than 30 publications in journals and conference proceedings.

LUIGI DI PUGLIA PUGLIESE received the Ph.D. degree in operations research from the University of Calabria, Italy. His Ph.D. thesis entitled "Models and Methods for the Constrained Shortest Path Problem and Its Variants." He is currently a Researcher with the Istituto di Calcolo e Reti ad Alte Prestazioni, Consiglio Nazionale delle Ricerche, Italy. His research interests include network optimization, logistics, project scheduling, combinatorial optimization, multi-objective optimization, and robust optimization.

FRANCESCA GUERRIERO (Senior Member, IEEE) received the degree (Hons.) in management engineering and the Ph.D. degree in system engineering and computer science from the University of Calabria (UNICAL). She was a Visiting Research Fellow with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. She is currently a Full Professor of operation research and the Dean of the Department of Mechanical, Energy and Management Engineering, University of Calabria. She is the Vice-President of AIRO, the Italian Association of Operations Research. She is the co-author of more than 150 articles published in prestigious journals in the operations research field. She supervises the master's and Ph.D. research projects. Her research interests include network optimization, logistics and distribution, revenue management, optimization, and big data. She has been and is a member of the scientific committee of several international conferences. She acts as a referee for numerous international scientific journals and is a member of the editorial board of several scientific journals.



GIUSY MACRINA received the degree (Hons.) in management engineering and the Ph.D. degree in mathematics and computer science from the University of Calabria (UNICAL), in 2013 and 2018, respectively. She was a Visiting Student with the Centre interuniversitaire de recherche sur les reseaux d'entreprise (CIRRELT), Montreal. She is currently an Assistant Professor with the Department of Mechanical Energy and Management Engineering, UNICAL. Her main area of scientific activity is operations research, she focuses on logistics and distribution problems. She is the author of more than 20 articles published in international peer-reviewed journals. She is a Member of the Board of AIRO Young, the Young Chapter of Italian Operations Research Society (AIRO).

Open Access funding provided by 'Università della Calabria' within the CRUI CARE Agreement