World Electric Vehicle Journal

Article

Regression Machine Learning Models for the Short-Time Prediction of Genetic Algorithm Results in a Vehicle Routing Problem

Ivan Kristianto Singgih [1,2,3] and Moses Laksono Singgih [4,*]

[1] Department of Industrial Engineering, University of Surabaya, Surabaya 60293, Indonesia; ivanksinggih@staff.ubaya.ac.id
[2] The Indonesian Researcher Association in South Korea (APIK), Seoul 07342, Republic of Korea
[3] Kolaborasi Riset dan Inovasi Industri Kecerdasan Artifisial (KORIKA), Jakarta 10340, Indonesia
[4] Department of Industrial and Systems Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
* Correspondence: moseslsinggih@its.ac.id

Abstract: Machine learning techniques have advanced rapidly, leading to better prediction accuracy within a short computational time. Such advancement encourages various novel applications, including in the field of operations research. This study introduces a novel way to utilize regression machine learning models to predict the objectives of vehicle routing problems that are solved using a genetic algorithm. Previous studies have generally discussed how (1) operations research methods are used independently to generate optimized solutions and (2) machine learning techniques are used independently to predict values from a given dataset. Some studies have discussed the collaborations between operations research and machine learning fields as follows: (1) using machine learning techniques to generate input data for operations research problems, (2) using operations research techniques to optimize the hyper-parameters of machine learning models, and (3) using machine learning to improve the quality of operations research algorithms. This study differs from the types of collaborative studies listed above. This study focuses on the prediction of the objective of the vehicle routing problem directly given the input and output data, without optimizing the problem using operations research algorithms. This study introduces a straightforward framework that captures the input data characteristics for the vehicle routing problem. The proposed framework is applied by generating the input and output data using the genetic algorithm and then using regression machine learning models to predict the obtained objective values. The numerical experiments show that the best models are random forest regression, a generalized linear model with a Poisson distribution, and ridge regression with cross-validation.

Keywords: vehicle routing problem; genetic algorithm; prediction; regression machine learning; smart logistics

Citation: Singgih, I.K.; Singgih, M.L. Regression Machine Learning Models for the Short-Time Prediction of Genetic Algorithm Results in a Vehicle Routing Problem. *World Electr. Veh. J.* **2024**, *15*, 308. https://doi.org/10.3390/wevj15070308

Academic Editor: Grzegorz Sierpiński

Received: 29 June 2024
Revised: 4 July 2024
Accepted: 8 July 2024
Published: 14 July 2024

1. Introduction

In recent years, machine learning studies have advanced rapidly and encouraged collaboration with various research fields, including the operations research field. There are several main frameworks used when conducting research in both areas simultaneously. The first framework applies machine learning techniques to predict input data for operations research problems. One application is estimating the energy consumption of electric vehicles on different paths and routes before solving the routing problem [1]. Another application is clustering flights based on the similarity in the working crews before solving the flight connection optimization problem [2]. The last example is predicting the demand for cash transportation between bank branches based on historical data and calendar

information before determining the transportation schedules [3]. A review of this first framework can be observed in [4].

The second framework applies operations research techniques to optimize the machine learning method's results. Some examples are (1) using differential flower pollination metaheuristics to optimize the hyper-parameters of a support vector machine model for image-processing-based pavement condition observation [5] and (2) using the firefly algorithm to optimize the hyper-parameters in a support vector regression machine learning model used for the prediction of a building's energy consumption level [6]. A recent review of this type of study is presented in [7]. It shows that such research with such a framework is still rare.

The third framework applies machine learning models to improve the quality of operations research models. The first category in this framework is using machine learning methods (e.g., reinforcement learning) to find the best operator in metaheuristics, as described in a recent review [7]. The second category in this framework is using machine learning to improve the quality of operations research methods. Two examples are (1) using a decision tree to differentiate poor and good vehicle routing problem solutions [8] and (2) using machine learning techniques to select bins in a stochastic bin packing problem considering various features (the bin's capacity, the reduced cost, and variable values in the relaxed version of the optimization problem) [9].

Despite the continuous growth in machine learning studies in various fields and the development of numerous operations research techniques, collaboration between the machine learning and operations research fields is still in its initial phase. As mentioned in [10], most of the proposed machine learning methods have not yet been applied to solve vehicle routing problem variants, one of the most studied topics in operations research.

Machine learning was used by Arnold and Sörensen [8] to extract important features and develop a problem-specific decision tree. It opened up the opportunity to design heuristics with good knowledge of the studied problem. However, their approach still required the development and running of heuristics. Differing from the three frameworks mentioned above, this study introduces a more general framework that could be used to predict the results of a solution method given an operations research problem without running an operations research algorithm. Such a situation is encountered when the decision-makers need to predict the system's behavior without waiting for long periods of computational time. This prediction is important before making any related decisions. As an example, after the decision-makers predict the total travel times of trucks, they could measure how much energy (gasoline or electricity) is consumed for deliveries and possibly solve another follow-up optimization problem, e.g., (1) determining the number of energy supply centers to locate within the area and (2) allocating trucks to energy supply centers, to ensure that the trucks run smoothly.

The proposed framework could also be applied when data are generated based on (1) the decision-maker's knowledge or (2) historical data without any of the solution method's information. The broad implementation of the proposed framework is thus possible. Implementing the framework for such practical data could be beneficial when it is difficult for managers (as decision-makers) to install the required computational systems to run the optimization models [11]. For ease of understanding, this study demonstrates the proposed framework when solving the vehicle routing problem (VRP) using the genetic algorithm (GA) method.

A framework for the use of machine learning techniques to observe the behavior of operations research models (discussed in this study) has been suggested in recent studies [12,13]. De Bock et al. [12] listed the steps as follows: (1) data generation and preprocessing, (2) machine learning model selection for the processing of the data, and (3) the interpretation of the model running results, including a feature importance analysis and rule extraction. Although De Bock et al. [12] presented such a data generation framework, they [12] did not specify any details regarding (1) which method should be used for data generation and (2) how the data generation and the machine learning model's running

should be conducted when solving a specific case study. Different from [12], this study proposes a detailed framework for data generation using a specific operations research method and demonstrates how the proposed framework could be applied to solve a specific operations research problem. Different from [13], which proposed a classification-model-based prediction framework for a scheduling problem with several simple rules, this study proposes a framework for the routing problem that is solved using metaheuristics and predicted using regression machine learning models.

The structure of this study is as follows. Section 2 describes the proposed framework used to solve the operations research problem using regression machine learning techniques. Section 3 explains the case study: the VRP solved with the GA. Section 4 presents the numerical experiment's results. Section 5 explains managerial insights related to the implementation of the framework and lists possible applications of the proposed framework. Section 6 concludes the study.

2. Proposed Operations Research Problem Solving Using Machine Learning (OpReMaL) Framework

The solution of operations research problems is usually evaluated based on two performance indicators: (1) the solution quality regarding the best objective value and (2) the computational time. Although many solution methods are available, it is common practice to initially solve the problem using a mathematical model to obtain the optimal solutions for small-sized problems. The problems that arise when using a mathematical model to solve larger-sized problems are (1) the long computational time and (2) the possibility of not obtaining any feasible solution due to the complexity of the model. Therefore, various methods that obtain slightly worse solutions but during a much shorter computational time are applied, e.g., algorithms and metaheuristics [14].

In general, to obtain good-quality solutions, running any method for a longer computational time is necessary. Even though, in general, algorithms and metaheuristics require much less computational time than mathematical models, these methods might still need a long computational time to obtain solutions for larger-sized problems. As a consequence, it is necessary to develop a fast way to conduct real-time prediction and deal with the data generation process, which is costly [12]. This study resolves this long computational time issue by replacing the initial solution generation method, which is the operations research method (e.g., metaheuristics), with regression machine learning models. This study proposes the Operations Research Problem Solving Using Machine Learning (OpReMaL) framework, which is illustrated in Figure 1. The framework consists of (Step 1) the running of the operations research method to generate the input and output data and (Step 2) the running of the machine learning model to predict the generated output data based on the input data.

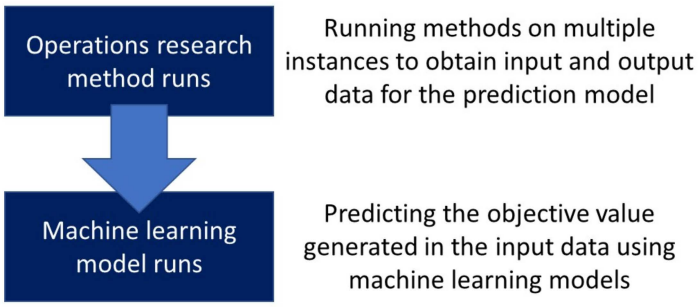

Figure 1. The proposed Operations Research Problem Solving Using Machine Learning (OpReMaL) framework.

In the first step of the OpReMaL framework (Figure 1), the operations research method is run for several instances to produce input and output data. An instance refers to a single operations research problem (with a set of problem parameters), generally solved to obtain a single best solution. The input is related to the characteristics of the problem;

meanwhile, the output refers to the objective value obtained after running a given solution method, e.g., metaheuristics. The OpReMaL framework could be considered a black box. The system represented by the black box is assumed to have a single means to generate the optimization solution. In practice, decision-makers could have (1) a method that generates a single solution for each instance or (2) a method that generates multiple solutions that would later be further evaluated for the final decision-making. This study considers the former. The generated input and output data will then be used to train the machine learning method.

In the second step of the framework (Figure 1), the machine learning model is trained and then used to predict the objective value in a much shorter time than when running operations research methods. The machine learning models predict the single objective value provided for each instance. When multiple objective values are considered, modifications could be performed with either of two options. The first involves preprocessing multiple objective values into a single weighted objective value. It includes the case in which the total costs are measured as a single value [15]. The second is applying the framework as many times as the number of objectives (and then selecting the best solution in the post-processing stage, e.g., using Pareto front analysis).

The OpReMaL framework starts with the operations research method, which is followed by the running of the machine learning model. During the real-time prediction process, the trained machine learning model is used directly within a very short prediction time. Such a situation occurs under the condition that the historical data size is already sufficiently large and there is no fundamental change in the problem parameters. On the contrary, when a new set of parameters is introduced into the problem, e.g., a new area of customers with different characteristics from the original ones, the whole framework (the operations research method and the machine learning model training) would need to be executed again.

3. Case Study: Vehicle Routing Problem Resolved Using Genetic Algorithm

3.1. Vehicle Routing Problem Resolved Using Genetic Algorithm (VRP-GA)

To show the effectiveness of the proposed OpReMaL framework, this study considers the case of the VRP that is resolved using the GA. The considered VRP is illustrated in Figure 2. Given the number of customers that must be visited and whose demand must be satisfied, multiple truck delivery routes are determined. The VRP considers that all homogeneous trucks (with the same capacity) start and end their travel from a single depot (node 0). The objective is to minimize the total travel times of all trucks. The mathematical model for this well-known VRP is presented in [16].

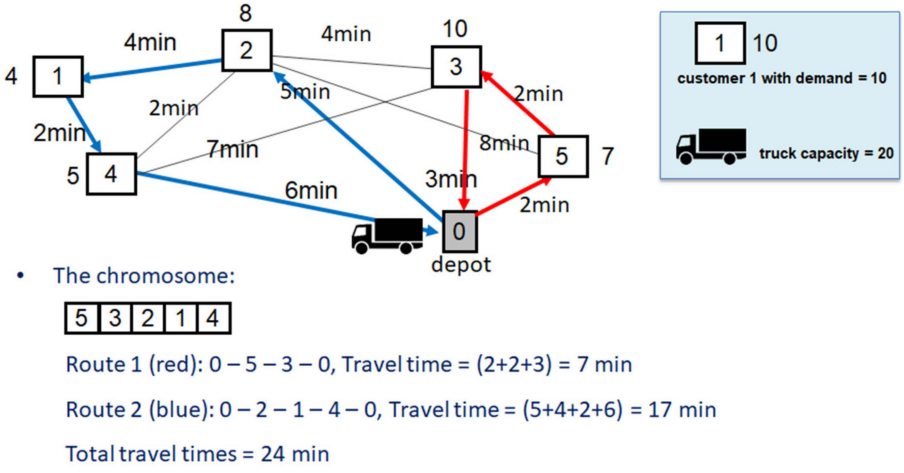

Figure 2. The considered vehicle routing problem.

The chromosome in Figure 2 represents a single solution. Given the sequence of customers to be visited by the truck, each route is constructed by subsequently adding

customers while calculating the total amount of items transported by the truck. Before the truck capacity is violated, a single truck route generation step is completed, and then another truck is scheduled to satisfy the next customer's demands. In the considered VRP, the truck starts and ends at the depot.

The GA used to solve the VRP is shown in Figure 3. The GA uses the solution representation shown in Figure 2. The algorithm starts by randomly generating *population_size* chromosomes in the initial population. The objective value of each chromosome is calculated. Next, new solutions are generated through crossover and mutation operations within *num_of_population* populations.

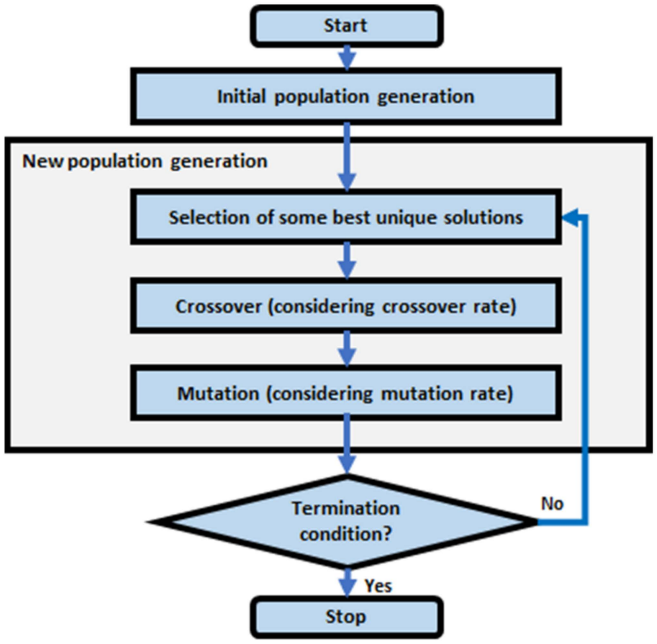

Figure 3. The genetic algorithm that is used to solve the vehicle routing problem.

In each population, the new solution generation process is conducted as follows (Figure 3): (1) selecting some best solutions (to be used as parent chromosomes during the crossover operation), (2) applying the crossover operator, and (3) applying the mutation operator. In part (1), a total of less than or equal to *num_of_selected_initial_chromosomes* unique best chromosomes are selected from the latest population. Next, each of the selected best chromosomes is given a selection probability (to be a parent chromosome) using the objective value conversion formula presented in Equation (1). The selection probability of each chromosome is then calculated using Equation (2). These equations set solutions with shorter travel times to have a larger selection probability.

$$inverted\_objective\_value = \frac{1}{objective\_value} \quad (1)$$

$$selection\_probability = \frac{inverted\_objective\_value}{\text{total } inverted\_objective\_value \text{ of all best chromosomes}} \quad (2)$$

When applying the crossover operator in part (2), two parent chromosomes are selected randomly based on the selection probabilities. After selecting these two parent chromosomes, a random number between 0 and 1 is generated. The crossover operator is applied if the random number is less than the *crossover_rate*. Otherwise, the parent chromosomes are stored as the result of the crossover operation. The outputs of the crossover operation are called child chromosomes. After applying the crossover operator, the two best chromosomes between two parent and two child chromosomes are selected to be stored in the new population. Given the *num_of_selected_initial_chromosomes* initial best chromosomes

that are already stored in the new population, more chromosomes are generated during the crossover operation until the new population is filled with *population_size* chromosomes.

Given the *population_size* chromosomes in the new population, in part (3), each chromosome in the new population is further processed by applying the mutation operator. After selecting one parent chromosome using the selection probability in Equation (2), a random number between 0 and 1 is generated. The mutation operator is applied if the random number is less than the *mutation_rate*. Otherwise, the parent chromosome is stored as the result of the mutation operation. After applying the mutation operator, the child chromosome is selected to be stored in the new population, even though its objective value is worse than the parent chromosome. Such a selection is allowed in order to ensure good exploration while generating new solutions. The convergence of the GA is encouraged by selecting the best chromosomes before applying the crossover operator.

The crossover and mutation operators are illustrated in Figure 4. In the crossover operation, two-point crossover is applied. Two cutting points are randomly selected within the parent chromosomes. The customer numbers at the middle of the cutting points are copied into the child chromosomes (the red and blue highlighted parts in Figure 4). The remaining parts in each child chromosome are copied from the other parent chromosome. The remaining customer numbers are copied from left to right until all customer numbers are inserted into the child chromosomes. The mutation operation is applied by randomly selecting two customer numbers and exchanging their positions to produce the child chromosome.

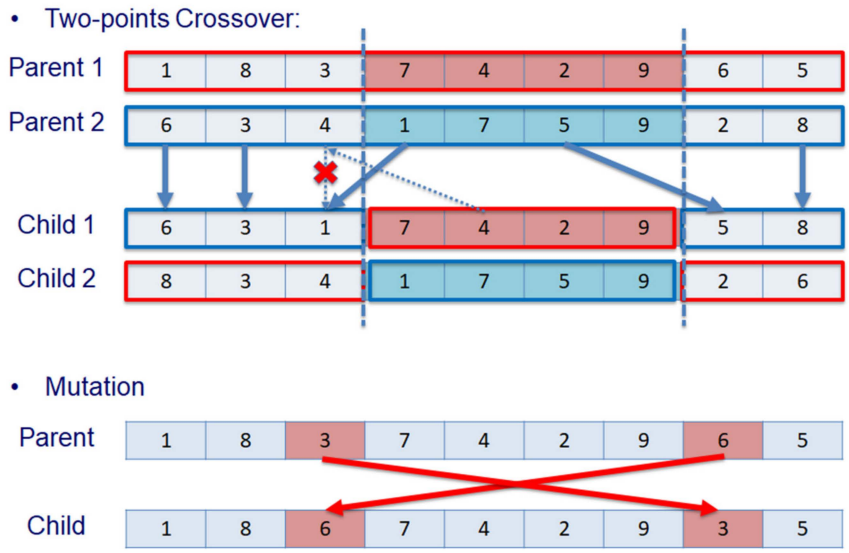

Figure 4. The crossover and mutation operators that are applied in this study.

3.2. Operations Research Problem Solving Using Machine Learning (OpReMaL) Framework Using Regression Machine Learning Models for the VRP-GA Case Study

In this study, regression machine learning models are used to predict the objective value of the VRP-GA (the total distances traveled by the trucks). The list of input and output data is shown in Figure 5. The input data represent the characteristics of each vehicle routing problem instance. Each input datum is presented as follows.

- *min_distance_depot*: the minimum distance between all customer–depot pairs;
- *average_distance_depot*: the average distance between all customer–depot pairs;
- *max_distance_depot*: the maximum distance between all customer–depot pairs;
- *min_distance_nondepot*: the minimum distance between all customer pairs, excluding the depot;
- *average_distance_nondepot*: the average distance between all customer pairs, excluding the depot;

- *max_distance_nondepot*: the maximum distance between all customer pairs, excluding the depot;
- *min_demand*: the minimum demand value among all customers;
- *average_demand*: the average demand value of all customers;
- *max_demand*: the maximum demand value among all customers;
- *num_customers*: the number of customers considered in the instance;
- *vehicle_capacity*: the capacity of homogeneous trucks.

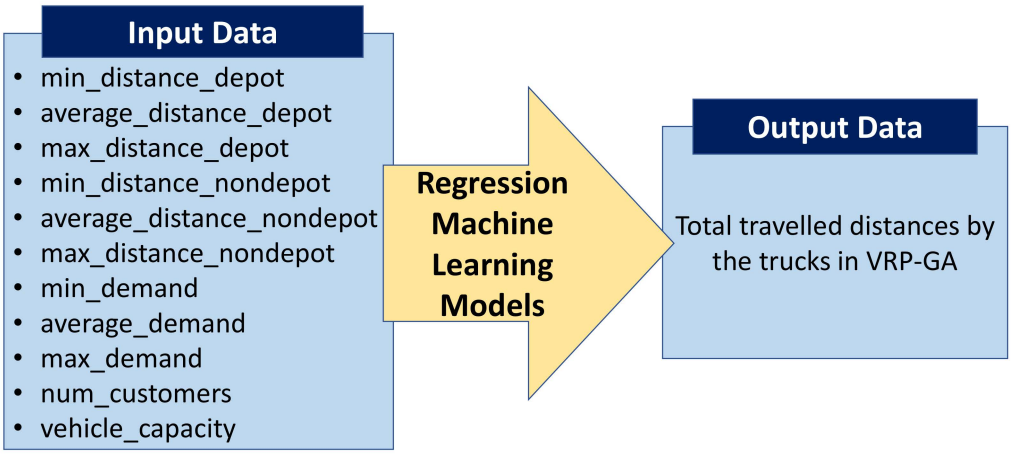

Figure 5. The input and output data used for the regression machine learning models.

Based on the best result obtained using the GA for each VRP instance, the regression machine learning models are applied to predict the total traveled distances of the trucks. The regression machine learning model is expected to predict the objective value of the VRP in a much shorter time than when using the GA.

4. Numerical Experiments

For the first step in the proposed framework (Figure 1), the data are generated using the GA. The characteristics of the problem instances and the GA parameter settings used in the numerical experiments are listed in Table 1. Other general characteristics and settings in all instances are listed in Table 2. Initially, a 1000 × 1000 map is generated, and then the customer coordinates in the x and y axes are determined randomly. The coordinates are then used to measure the Euclidean distances between the customers. A value for the capacity of the homogeneous trucks is selected randomly for each instance. The objective values of the VRP are calculated using the GA. The data can be accessed online at https://ubaya.id/vrp_ga_input_output (accessed on 2 December 2023). Considering various sets of instances for the machine learning prediction could offer a means to resolve the overfitting situation (which is caused by focusing only on one set of instances).

Table 1. The specific characteristics of the instances and the genetic algorithm parameter settings used in the numerical experiments.

Set of Instances	Number of Instances	Number of Customers [1]	*population_size*	*num_of_selected_initial_chromosomes*	*num_of_populations*	Average Computational Time Using the GA (s)
Set 1	2000	[10, 100]	50	30	50	4
Set 2	2000	[201, 300]	50	30	50	16
Set 3	500	[401, 500]	100	50	100	181
Set 4	50	[601, 700]	200	80	200	1811

[1] [minimum value, maximum value].

Table 2. The general characteristics of the instances and the genetic algorithm parameter settings used in the numerical experiments.

Characteristic or Parameter	Value
Map width (square area)	1000
Customer demand	[30, 100]
Homogeneous truck capacity	300, 400, or 500
crossover_rate	0.8
mutation_rate	0.2

For the second step in the proposed framework (Figure 1), the output data are predicted using regression machine learning models without solving the optimization problem again. Several regression machine learning models are tested, and then the best models are reported in this section. The following regression machine learning models from scikit-learn [17] are used for the predictions: (1) random forest regression, (2) linear regression, (3) RidgeCV, (4) ElasticNetCV, (5) LarsCV, (6) LassoCV, (7) LassoLarsCV, (8) OrthogonalMatchingPursuitCV, (9) ARDRegression, (10) BayesianRidge, (11) HuberRegressor, (12) RANSACRegressor, (13) TheilSenRegressor, (14) PoissonRegressor, (15) TweedieRegressor, (16) GammaRegressor, and (17) PassiveAggressiveRegressor. At the time of writing, scikit-learn has more than 90,000 citations based on Google Scholar. Its minimal dependencies and ease of use allow a high reproducibility rate in many studies. Each regression machine learning model is described in Table 3.

Table 3. Explanations of each regression machine learning model.

Regression Machine Learning Model	Explanation
(1) Random Forest Regression	An ensemble method consisting of some decision trees. It considers the decision trees' diversity when making decisions [18].
(2) Linear Regression	A linear equation used to represent relationships between variables, generated based on the observed data [19].
(3) RidgeCV	A multiple linear regression with a reduction in the weights of unimportant coefficients (ridge regression) and cross-validation. It allows the greater generalization of the prediction model [20,21].
(4) ElasticNetCV	A regularization method that eliminates the redundancy of variables. It has some penalty terms that are used as a compromise strategy between the LASSO and ridge regression techniques [22]. ElasticNetCV is an ElasticNet with cross-validation [21].
(5) LarsCV	A linear regression machine learning model that starts with all coefficients equal to 0 and then gradually updates the coefficients after identifying the most correlated input with the output data. It is very efficient because of its piecewise linear solution paths [23]. LarsCV applies cross-validation [21].
(6) LassoCV	A linear regression machine learning model with the least absolute shrinkage and selection operator (LASSO) that selects variables and determines regression coefficients simultaneously in one step [24]. LassoCV applies cross-validation [21].
(7) LassoLarsCV	A cross-validated LASSO, applied in the LARS algorithm [21].
(8) OrthogonalMatchingPursuitCV	A method that iteratively selects the feature that has the largest correlation with the current residual. Each selected feature will then be projected to the span of selected features. The iteration continues until K columns are selected [25]. It applies cross-validation [21].

Table 3. *Cont.*

Regression Machine Learning Model	Explanation
(9) ARDRegression	A Bayesian model with automatic relevance determination that prunes redundant features by estimating the parameters of the data distribution based on a maximum likelihood consideration [26].
(10) BayesianRidge	A Bayesian method that considers a common variance for all regression coefficients [27].
(11) HuberRegressor	A regression machine learning model that is robust to outlier data due to considering a linear loss for such outlier data [21].
(12) RANSACRegressor	An iterative algorithm that conducts the robust estimation of parameters based on inliers from the data after randomly extracting matching points. The inliers are determined based on a threshold [21,28].
(13) TheilSenRegressor	A median-based estimator that uses generalization in multiple dimensions, allowing it to be robust to multivariate outliers [21].
(14) PoissonRegressor	A generalized linear model that considers the dependent variables to be independent and random variables that follow a Poisson distribution [29].
(15) TweedieRegressor	A generalized linear model that considers the dependent variables to be independent and random variables that follow a Tweedie distribution [21].
(16) GammaRegressor	A generalized linear model that considers the dependent variables to be independent and random variables that follow a Gamma distribution [21].
(17) PassiveAggressiveRegressor	An online learning regression machine learning model that learns data that are added continuously [30]. It is suitable for large-scale learning [21].

The regression machine learning models are evaluated via the mean absolute error (MAE) values, as shown in Figure 6. The three best models with the lowest MAE values are random forest regression (2216.86), HuberRegressor (4940.09), and ARDRegression (5013.01). The average objective value of all instances is 98,576.

The regression machine learning models are evaluated via the mean squared error (MSE) values, as shown in Figure 7. The three best models with the lowest MSE values are random forest regression (12,204,263.19), ARDRegression (51,956,577.89), and TheilSenRegressor (58,285,906.56). The regression machine learning models are evaluated via the root mean squared error (RMSE) values, as shown in Figure 8. The three best models with the lowest RMSE values are random forest regression (3493.46), TheilSenRegressor (7634.52), and PassiveAggressiveRegressor (9610.48). A list of all MAE, MSE, and RMSE values for each model is shown in Table 4. The experiments with the MAE, MSE, and RMSE metrics show that the best model is random forest regression.

To further evaluate the proposed framework, the prediction is conducted when considering each set of instances (in Table 1) separately. The best MAE value, the average objective value, and the best regression machine learning model when considering each set of instances are shown in Table 5. The prediction quality is good because the deviation measured by the best MAE in contrast to the average objective value is less than 5% for each set of instances. When compared with the results in Figure 6, the best model is not always the same. It could be concluded that it is necessary to apply different regression machine learning models for each set of instances (Table 5). This would allow a better prediction to be produced, rather than considering instances from all sets simultaneously

(Figure 6). However, the decision-maker could still consider using the whole set of data to apply a more general method for the problem characteristics.

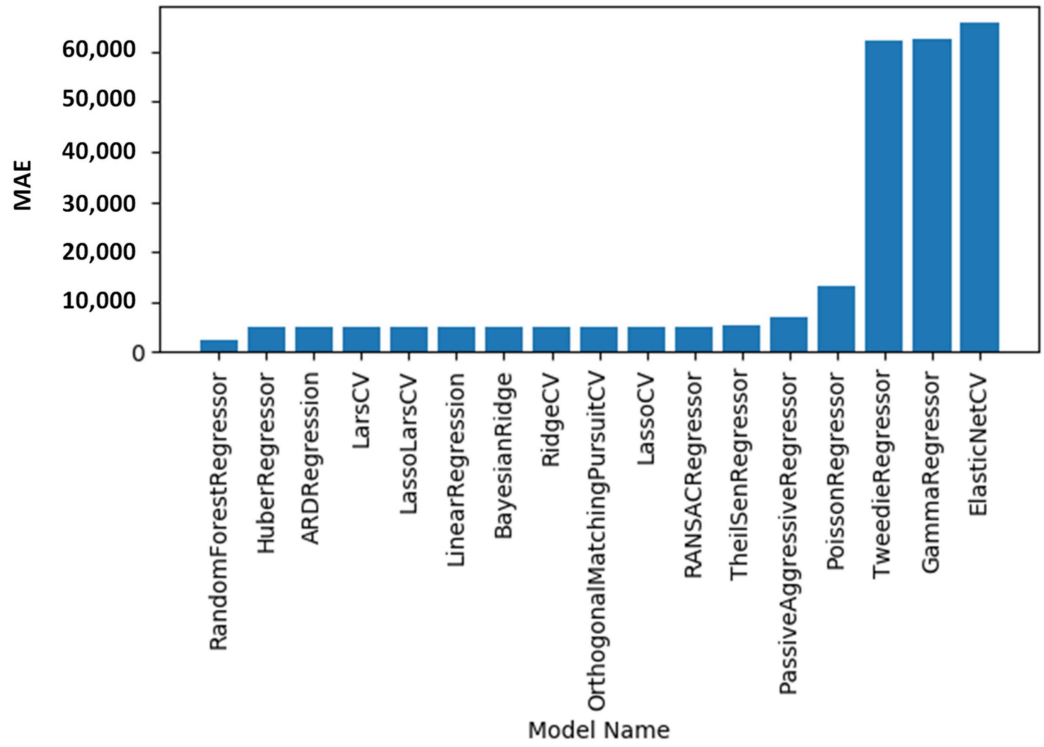

Figure 6. Performance of all regression machine learning models (with mean absolute error metric).

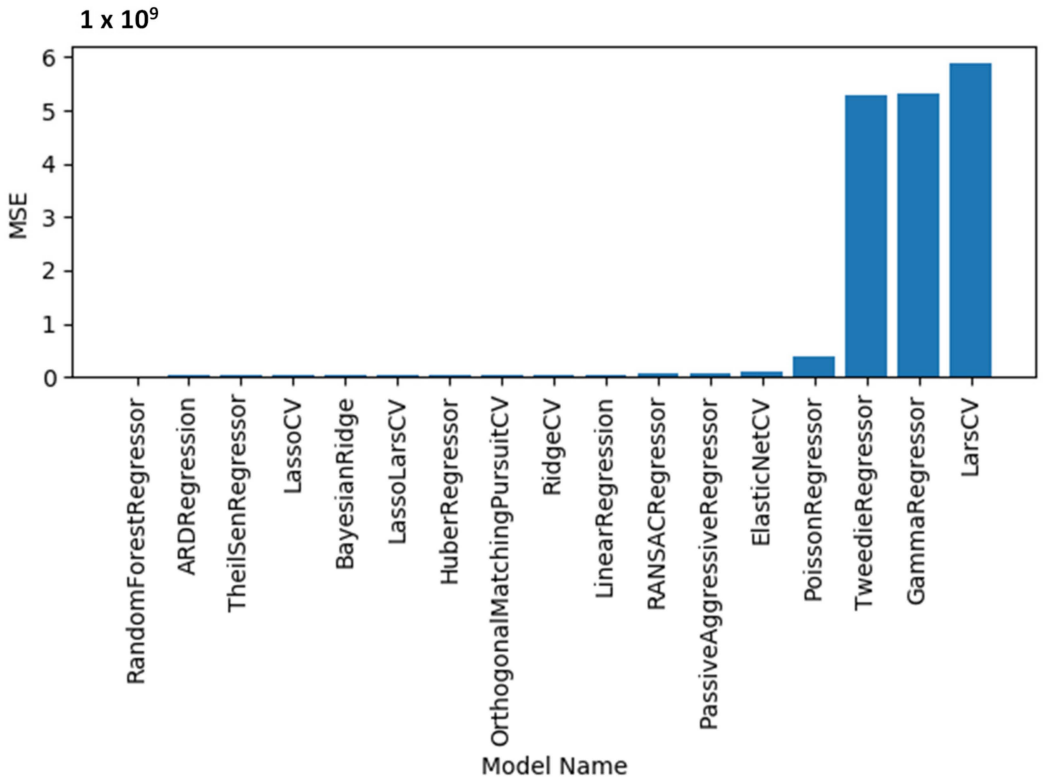

Figure 7. Performance of all regression machine learning models (with mean squared error metric).

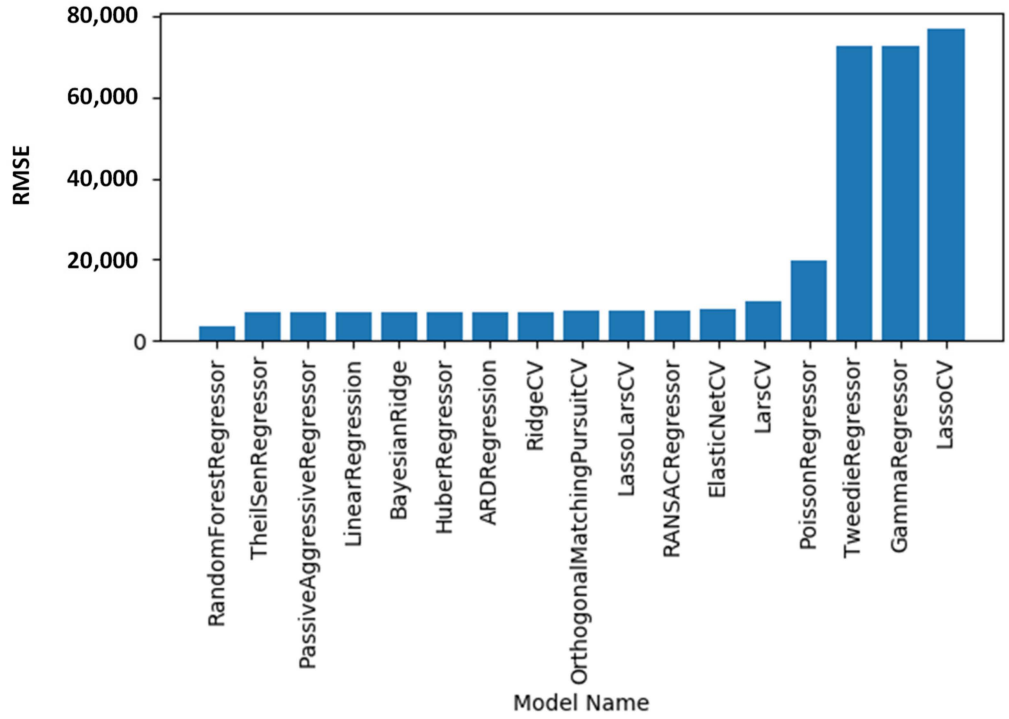

Figure 8. Performance of all regression machine learning models (with root mean squared error metric).

Table 4. Performance metrics for each regression machine learning model when considering all data instances.

Regression Machine Learning Model	MAE	MSE	RMSE
(1) Random Forest Regression	*** 2216**	*** 12,204,263**	*** 3493**
(2) Linear Regression	5013	51,871,442	7202
(3) RidgeCV	5015	51,800,637	7197
(4) ElasticNetCV	65,859	5,903,239,905	76,832
(5) LarsCV	5013	51,871,442	7202
(6) LassoCV	5038	52,427,706	7240
(7) LassoLarsCV	5013	51,871,442	7202
(8) OrthogonalMatchingPursuitCV	5037	52,412,692	7239
(9) ARDRegression	5013	51,956,577	7208
(10) BayesianRidge	5013	51,868,066	7201
(11) HuberRegressor	4940	56,948,824	7546
(12) RANSACRegressor	5043	51,806,962	7197
(13) TheilSenRegressor	5369	58,285,906	7634
(14) PoissonRegressor	13,100	397,203,436	19,929
(15) TweedieRegressor	62,313	5,299,130,027	72,795
(16) GammaRegressor	62,694	5,312,249,478	72,885
(17) PassiveAggressiveRegressor	6847	92,361,365	9610

*** Best performance value for each metric.**

Table 5. MAE values when each set of instances is considered separately.

Set of Instances	Best MAE	Average Objective Value	Best Model	Computational Time (s)
Set 1	932.58 (3.8%)	24,863	Random Forest Regression	2
Set 2	2142.51 (1.6%)	132,821	PoissonRegressor	<1
Set 3	3839.04 (1.6%)	234,490	PoissonRegressor	<1
Set 4	3452.44 (1.1%)	318,115	RidgeCV	<1

Each of the best regression machine learning models requires less than two seconds for its training. Using any regression machine learning model, the objective value prediction of any instance could be conducted within less than one second. When compared with the computational time presented in Table 1, the objective value prediction using the regression machine learning model is up to 1800 times faster than when the objective values are calculated using the GA, especially when dealing with large-sized problems. Further evaluations using each set of instances when using the MSE and RMSE are shown in Tables 6 and 7, respectively. The experiments with the MAE, MSE, and RMSE metrics show that the best models are random forest regression (for Set 1), PoissonRegressor (for Set 2), PoissonRegressor (for Set 3), and RidgeCV (for Set 4).

Table 6. MSE values when each set of instances is considered separately.

Set of Instances	Best MSE	Best Model	Computational Time (s)
Set 1	1,456,221.20	Random Forest Regression	2
Set 2	7,513,768.98	PoissonRegressor	<1
Set 3	23,010,249.50	PoissonRegressor	<1
Set 4	18,644,231.03	RidgeCV	<1

Table 7. RMSE values when each set of instances is considered separately.

Set of Instances	Best RMSE	Best Model	Computational Time (s)
Set 1	1206.74	Random Forest Regression	2
Set 2	2741.13	PoissonRegressor	<1
Set 3	4796.90	PoissonRegressor	<1
Set 4	4317.90	RidgeCV	<1

5. Managerial Insights and Potential Applications

The proposed OpReMaL framework was designed to predict the objective values of operations research problems. Different operations research problems have different problem characteristics (which would be considered as the input data). When generating the objective value as the output data, a specific operations research method would be applied to the set of input data. When solving each specific type of operations research problem, the best solution method could be different. This best solution method is selected through extensive numerical experiments [14,31]. Likewise, given different sets of input and output data, the best regression machine learning models should be tested. In this study, we test the effectiveness of the proposed framework by observing the VRP when solved using the GA.

In terms of computational time, it is shown that the prediction models could predict the objective values in a very short time (around 1 s). It is much shorter than the average time required to solve the VRP-GA for the largest-sized instance, which is around 1800 s (Table 1). We considered up to 700 customers in the numerical experiments, which was larger than the size of the real problem (e.g., 385 customers in [32]). This shows that the proposed method could deal with real-world problems effectively. It could be concluded that the proposed OpReMaL does not only predict the output of operations research problems well but also reduces the computational (prediction) time significantly. In the current big data era, it is strongly necessary to develop fast solution methods to ensure that good decisions are made based on recently collected data. It offers a huge opportunity to provide high-quality services and generate significantly larger profits for businesses and decision-makers.

In practice, the decision-makers simply need to run the prediction using regression machine learning when they need to observe the total traveled distances based on the given information of the VRP. However, when implementing the OpReMaL framework, it is necessary to understand when the prediction model needs to be tuned, which is when the characteristics of the VRP input differ from the ones considered before. The tuning starts with the addition of more input data by solving the VRP for the new data set using the GA,

using the updated input data to tune the regression machine learning models and then selecting the best one for the new predictions.

6. Conclusions

This study proposed the Operations Research Problem Solving Using Machine Learning (OpReMaL) framework to predict the objective values of a vehicle routing problem. The proposed framework requires a very short time without running an operations research algorithm, which might require a long computational time, especially for large-sized problems. The proposed framework (1) differs from most frameworks that combine operations research and machine learning methods and (2) is the first one that considers regression machine learning models to observe the characteristics of the vehicle routing problem solved using the genetic algorithm. The numerical experiment's results showed that the best models for all sets of instances were random forest regression, the generalized linear model with a Poisson distribution, and ridge regression with cross-validation.

The proposed OpReMaL framework predicts the behavior of data that belong to a specific operations research problem and solution method. For future studies, it would be interesting to observe more operations research problems (e.g., the location routing problem [33], routing problem for shared logistics [34], electric vehicle relocation problem [14], multi-altitude drone routing problem for post-disaster observation [32]) and more solution methods (e.g., beetle swarm optimization [35], hybrid metaheuristics [36,37]) and show how the proposed OpReMaL framework could also obtain good solutions. It is challenging to determine the appropriate input data selection and observe how different the prediction result would be when different operations research solution methods are implemented to solve the problem. Future studies could also consider testing more advanced machine learning techniques, e.g., ensemble machine learning models [38]. Another possible implementation of the proposed method is in predicting the features of the best solutions instead of the objective. The issue to resolve is how to deal with the limited capability of machine learning models to predict only a single value, while the features of the best solutions are much more complicated than a single value. Such a problem needs a great deal of further investigation.

Author Contributions: Conceptualization, I.K.S. and M.L.S.; methodology, I.K.S.; software, I.K.S.; validation, I.K.S.; formal analysis, I.K.S.; resources, I.K.S.; data curation, I.K.S.; writing—original draft preparation, I.K.S.; writing—review and editing, I.K.S. and M.L.S.; visualization, I.K.S.; supervision, M.L.S.; project administration, I.K.S.; funding acquisition, I.K.S. and M.L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The input and output data can be accessed online at https://ubaya.id/vrp_ga_input_output (accessed on 2 December 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Basso, R.; Kulcsár, B.; Sanchez-Diaz, I. Electric Vehicle Routing Problem with Machine Learning for Energy Prediction. *Transp. Res. Part B Methodol.* **2021**, *145*, 24–55. [CrossRef]
2. Archetti, C.; Cordeau, J.-F.; Desaulniers, G. Introduction to the Special Issue on Combining Optimization and Machine Learning: Application in Vehicle Routing, Network Design and Crew Scheduling. *EURO J. Transp. Logist.* **2020**, *9*, 100024. [CrossRef]
3. López Lázaro, J.; Barbero Jiménez, Á.; Takeda, A. Improving Cash Logistics in Bank Branches by Coupling Machine Learning and Robust Optimization. *Expert Syst. Appl.* **2018**, *92*, 236–255. [CrossRef]
4. Escioglu-Unver, N.; Yıldız Aktaş, M. Electric Vehicle Charging Service Operations: A Review of Machine Learning Applications for Infrastructure Planning, Control, Pricing and Routing. *Renew. Sustain. Energy Rev.* **2023**, *188*, 113873. [CrossRef]
5. Hoang, N.-D. Image Processing Based Automatic Recognition of Asphalt Pavement Patch Using a Metaheuristic Optimized Machine Learning Approach. *Adv. Eng. Inform.* **2019**, *40*, 110–120. [CrossRef]
6. Chou, J.-S.; Ngo, N.-T. Time Series Analytics Using Sliding Window Metaheuristic Optimization-Based Machine Learning System for Identifying Building Energy Consumption Patterns. *Appl. Energy* **2016**, *177*, 751–770. [CrossRef]

7. Karimi-Mamaghan, M.; Mohammadi, M.; Meyer, P.; Karimi-Mamaghan, A.M.; Talbi, E.-G. Machine Learning at the Service of Meta-Heuristics for Solving Combinatorial Optimization Problems: A State-of-the-Art. *Eur. J. Oper. Res.* **2022**, *296*, 393–422. [CrossRef]
8. Arnold, F.; Sörensen, K. What Makes a VRP Solution Good? The Generation of Problem-Specific Knowledge for Heuristics. *Comput. Oper. Res.* **2019**, *106*, 280–288. [CrossRef]
9. Bruni, M.E.; Fadda, E.; Fedorov, S.; Perboli, G. A Machine Learning Optimization Approach for Last-Mile Delivery and Third-Party Logistics. *Comput. Oper. Res.* **2023**, *157*, 106262. [CrossRef]
10. Accorsi, L.; Lodi, A.; Vigo, D. Guidelines for the Computational Testing of Machine Learning Approaches to Vehicle Routing Problems. *Oper. Res. Lett.* **2022**, *50*, 229–234. [CrossRef]
11. Liebchen, C.; Schülldorf, H. A Collection of Aspects Why Optimization Projects for Railway Companies Could Risk Not to Succeed—A Multi-Perspective Approach. *J. Rail Transp. Plan. Manag.* **2019**, *11*, 100149. [CrossRef]
12. De Bock, K.W.; Coussement, K.; Caigny, A.D.; Słowiński, R.; Baesens, B.; Boute, R.N.; Choi, T.-M.; Delen, D.; Kraus, M.; Lessmann, S.; et al. Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda. *Eur. J. Oper. Res.* **2023**, *317*, 249–272. [CrossRef]
13. Singgih, I.K. Production Flow Analysis in a Semiconductor Fab Using Machine Learning Techniques. *Processes* **2021**, *9*, 407. [CrossRef]
14. Singgih, I.K.; Kim, B.I. Multi-Type Electric Vehicle Relocation Problem Considering Required Battery-Charging Time. *Eur. J. Ind. Eng.* **2020**, *14*, 335. [CrossRef]
15. Budiyanto, M.A.; Singgih, I.K.; Riadi, A.; Putra, G.L. Study on the LNG Distribution to Mobile Power Plants Using a Small-Scale LNG Carrier for the Case of the Sulawesi Region of Indonesia. *Energy Rep.* **2022**, *8*, 374–380. [CrossRef]
16. Toth, P.; Vigo, D. Models, Relaxations and Exact Approaches for the Capacitated Vehicle Routing Problem. *Discret. Appl. Math.* **2002**, *123*, 487–512. [CrossRef]
17. Scikit-Learn 1. Supervised Learning. Available online: https://scikit-learn.org/stable/supervised_learning.html (accessed on 23 January 2023).
18. Kim, S.W.; Lee, Y.G.; Tama, B.A.; Lee, S. Reliability-Enhanced Camera Lens Module Classification Using Semi-Supervised Regression Method. *Appl. Sci.* **2020**, *10*, 3832. [CrossRef]
19. Abid Almubaidin, M.A.; Latif, S.D.; Balan, K.; Ahmed, A.N.; El-Shafie, A. Enhancing Sediment Transport Predictions through Machine Learning-Based Multi-Scenario Regression Models. *Results Eng.* **2023**, *20*, 101585. [CrossRef]
20. Wang, X.; Wang, X.; Ma, B.; Li, Q.; Wang, C.; Shi, Y. High-Performance Reversible Data Hiding Based on Ridge Regression Prediction Algorithm. *Signal Process.* **2023**, *204*, 108818. [CrossRef]
21. API Reference. Available online: https://scikit-learn.org/stable/api/index.html (accessed on 2 December 2023).
22. Yu, B.; Chen, C.; Wang, X.; Yu, Z.; Ma, A.; Liu, B. Prediction of Protein–Protein Interactions Based on Elastic Net and Deep Forest. *Expert Syst. Appl.* **2021**, *176*, 114876. [CrossRef]
23. Lee, S.; Jun, C.-H. Fast Incremental Learning of Logistic Model Tree Using Least Angle Regression. *Expert Syst. Appl.* **2018**, *97*, 137–145. [CrossRef]
24. Kramlinger, P.; Schneider, U.; Krivobokova, T. Uniformly Valid Inference Based on the Lasso in Linear Mixed Models. *J. Multivar. Anal.* **2023**, *198*, 105230. [CrossRef]
25. Polat, Ö.; Kayhan, S.K. High-Speed FPGA Implementation of Orthogonal Matching Pursuit for Compressive Sensing Signal Reconstruction. *Comput. Electr. Eng.* **2018**, *71*, 173–190. [CrossRef]
26. Sandhu, R.; Pettit, C.; Khalil, M.; Poirel, D.; Sarkar, A. Bayesian Model Selection Using Automatic Relevance Determination for Nonlinear Dynamical Systems. *Comput. Methods Appl. Mech. Eng.* **2017**, *320*, 237–260. [CrossRef]
27. Da Silva, F.A.; Viana, A.P.; Correa, C.C.G.; Santos, E.A.; de Oliveira, J.A.V.S.; Andrade, J.D.G.; Ribeiro, R.M.; Glória, L.S. Bayesian Ridge Regression Shows the Best Fit for SSR Markers in Psidium Guajava among Bayesian Models. *Sci. Rep.* **2021**, *11*, 13639. [CrossRef] [PubMed]
28. Zhu, F.; Li, H.; Li, J.; Zhu, B.; Lei, S. Unmanned Aerial Vehicle Remote Sensing Image Registration Based on an Improved Oriented FAST and Rotated BRIEF-Random Sample Consensus Algorithm. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106944. [CrossRef]
29. Gomes, F.J.S.; Oliveira, T.F.; Brazil, F.S.; Pamplona, A.R.; Farias, V.J.; Silveira, A.M. A Method for the Behavioral Analysis of Partial Discharges in Hydrogenerators by Generalized Linear Models. *Electr. Power Syst. Res.* **2016**, *140*, 284–287. [CrossRef]
30. Jorge, J.; Paredes, R. Passive-Aggressive Online Learning with Nonlinear Embeddings. *Pattern Recognit.* **2018**, *79*, 162–171. [CrossRef]
31. Singgih, I.K.; Ferdinand, F.N. Mathematical Modeling Education Using an Online Serious Game. In Proceedings of the 2019 5th International Conference on New Media Studies (CONMEDIA), Bali, Indonesia, 9–11 October 2019; pp. 184–188.
32. Singgih, I.K.; Lee, J.; Kim, B.-I. Node and Edge Drone Surveillance Problem with Consideration of Required Observation Quality and Battery Replacement. *IEEE Access* **2020**, *8*, 44125–44139. [CrossRef]
33. Yan, T.; Lu, F.; Wang, S.; Wang, L.; Bi, H. A Hybrid Metaheuristic Algorithm for the Multi-Objective Location-Routing Problem in the Early Post-Disaster Stage. *J. Ind. Manag. Optim.* **2023**, *19*, 4663–4691. [CrossRef]
34. Bi, H.; Zhu, X.; Lu, F.; Huang, M. The Meal Delivery Routing Problem in E-Commerce Platforms under the Shared Logistics Mode. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 1799–1819. [CrossRef]

35. Lu, F.; Chen, W.; Feng, W.; Bi, H. 4PL Routing Problem Using Hybrid Beetle Swarm Optimization. *Soft Comput.* **2023**, *27*, 17011–17024. [CrossRef]
36. Lu, F.; Bi, H.; Huang, M.; Duan, S. Simulated Annealing Genetic Algorithm Based Schedule Risk Management of IT Outsourcing Project. *Math. Probl. Eng.* **2017**, *2017*, e6916575. [CrossRef]
37. Lu, F.; Feng, W.; Gao, M.; Bi, H.; Wang, S. The Fourth-Party Logistics Routing Problem Using Ant Colony System-Improved Grey Wolf Optimization. *J. Adv. Transp.* **2020**, *2020*, e8831746. [CrossRef]
38. Tama, B.A.; Lim, S. Ensemble Learning for Intrusion Detection Systems: A Systematic Mapping Study and Cross-Benchmark Evaluation. *Comput. Sci. Rev.* **2021**, *39*, 100357. [CrossRef]