Proceedings of the ASME 2023
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2023
August 20-23, 2023, Boston, Massachusetts

IDETC2023- 114718

PREDICTING THE QUANTITY OF RECYCLED END-OF-LIFE PRODUCTS USING A HYBRID SVR-BASED MODEL

Hanbing Xia[1], Ji Han[2], Jelena Milisavljevic-Syed[1]

[1]Sustainable Manufacturing Systems Centre, Cranfield University, Cranfield, UK
[2]University of Exeter Business School, London, UK

ABSTRACT

End-of-life product recycling is crucial for achieving sustainability in circular supply chains and improving resource utilization. Forecasting the quantity of recycled end-of-life products is essential for planning and managing reverse supply chain operations. Decision-makers and practitioners can benefit from this information when designing reverse logistics networks, managing tactical disposal, planning capacity, and operational production. To address the challenge of small sample data with multiple factors influencing the recycling number, and to deal with the randomness and nonlinearity of the recycling quantity, a hybrid predictive model has been developed in this research. The model is based on k-nearest neighbor mega-trend diffusion (KNNMTD), particle swarm optimization (PSO), and support vector regression (SVR) using the data from the field of end-of-life vehicles as a case study. Unlike existing literature, this research incorporates the data augmentation method to build an SVR-based model for end-of-life product recycling. The study shows that developing the predictive model using artificial virtual samples supported by the KNNMTD method is feasible, the PSO algorithm effectively brings strong approximation ability to the SVR-based model, and the KNNMTD-PSO-SVR model perform well in predicting the recycled end-of-life products quantity. These research findings could be considered a fundamental component of the smart system for circular supply chains, which will enable the smart platform to achieve supply chain sustainability through resource allocation and regional industry deployment.

Keywords: Reverse Supply Chain; End-of-life Products; Machine Learning; Predictive Analysis; Circular Supply Chain; Sustainability

1. INTRODUCTION

The circular economy (CE) has brought about a shift towards circular supply chains (CSCs), which enable original equipment manufacturers (OEMs) to recover materials and end-of-life (EOL) products and remarket them, replacing conventional supply chains [1]. With the increasing population, global manufacturing competition, accelerated price development of scarce resources, growing levels of waste and pollution, and stringent regulations on EOL product recycling, reverse supply chain (RSC) management has become crucial [2]. RSC involves recovering parts and used products from customers or any other stage of the supply chain with the purpose of recycling, reusing, remanufacturing, or proper disposal. Many industries practice RSC, such as medical items, paper, automobiles, steel production, and electrical and electronic equipment (EEE) [3]. However, return uncertainty in quantity, quality, and timing poses a significant challenge on the supply side of RSC [4]. Return quantity is the most significant uncertainty in RSC, making quantity prediction crucial when designing RSC. This information helps industrial engineers make decisions on designing reverse logistics networks, managing tactical disposal, planning capacity, and operational production [5].

In recent years, the automotive industry has experienced remarkable expansion and has become one of the largest industries in most countries. According to the European Automobile Manufacturers Association, worldwide production of commercial vehicles reached 16.6 million in 2020, a 23.9% increase compared to 2009, resulting in an increase in the number of scrapped automobiles [6]. Accurately evaluating the quantity of end-of-life vehicles (ELVs) is crucial for

minimizing environmental pollution and supporting effective management of the automotive remanufacturing industry while conducting automotive RSC management research [7].

Current approaches [8] for predicting the quantity of recycled EOL products are ineffective in dealing with the nonlinear sequences in the recovery quantity of EOL products, leading to unstable prediction accuracy. Machine learning (ML) technology, as an Artificial Intelligence (AI) technology for analyzing large amounts of complicated data, can generate models and predict future data. However, due to the immaturity of the reverse recovery industry and the absence of standard regulations in some countries and areas, there is a difficulty in accumulating a large amount of data and effective empirical knowledge in EOL product recycling [9]. Therefore, the performance of ML models is generally unsatisfactory without a large amount of training data. Thus, it is necessary to improve the predictability of ML models using limited historical data and to evaluate the feasibility of ML predictive models in precisely predicting the quantity of recycled EOL products.

The proposed hybrid prediction model is aimed at addressing the gaps mentioned earlier, and the purpose of this research is to assess its feasibility in predicting recycling quantity using small samples of numerical data. To validate the effectiveness of the model, data from the field of ELVs will be used in the research. The model combines k-nearest neighbor mega-trend diffusion (KNNMTD), particle swarm optimization (PSO), and support vector regression (SVR), and is designed to improve the predictability of ML models using limited historical data. By evaluating the proposed model, this research will contribute to improving the accuracy and reliability of predicting the quantity of recycled EOL products, which is essential for effective management of RSC and circular supply chains.

The research is organized as follows: The research background is introduced in Section 1. A brief literature review related to the data augmentation methods for small sample numerical data and the ML-based models applied to predict the quantity of EOL products is shown in Section 2. The research methodology is introduced in Section 3, explaining the steps from feature selection, data augmentation, to data training and validation. The analysis of experimental results is shown in Section 4. Further discussion of the research results is analyzed in Section 5. Conclusions are provided in Section 6.

2. LITERATURE REVIEW

In this section, we will review data augmentation methods that can be used for small sample datasets, as well as predictive methods for accurately estimating the quantity of EOL products.

2.1 Data Augmentation Methods

In order to enhance the generalization capability of a ML-based prediction model, data augmentation techniques are employed on small datasets to increase the number of effective samples and reduce the distribution gap between the training samples and the actual samples. There are primarily four methods for augmenting small numerical datasets: data interpolation, noise injection, and virtual sample generation.

Data interpolation methods are commonly used to achieve sample expansion for small datasets. These methods include inverse distance weighted [10], kriging [11], natural neighbor [12], spline [13], synthetic minority oversampling technique (SMOTE) [14], Borderline-SMOTE [15] and adaptive synthetic (ADASYN) [16]. They construct an approximation function to solve for the function of unknown points by constructing approximation functions at several known discrete points in a certain interval. However, the interpolation method based on label space is limited by its inability to explore an unknown range of data and to consider the distributions of the entire minority samples.

Noise injection methods are effective for sample expansion, generating new samples by adding Gaussian noise with zero mean and fixed variance to the original training data set [17, 18].

Virtual sample generation (VSG) methods can be used to fill the information gap between real and hypothetical samples in the sample space due to insufficient data, improve the prediction ability of the model, and suppress model overfitting. Two main data augmentation methods based on feature space are kernel density estimation and domain extension. Kernel probability estimation, a typical method of probability density estimation, estimates the probability distribution of data and achieves the relative balance of data distribution. The domain extension method generates virtual samples in the whole value space by estimating the value range of data in the feature space. Mega trend diffusion (MTD), based on uniform distribution theory, is a representative domain extension method that generates virtual samples using a membership function from fuzzy theory while considering each of the attributes rather than the probability values [19]. A modified MTD approach, generalized time diffusion (GTD), has been proposed to capture the time dependency of sequential data by integrating successive time steps with small datasets [20].

In comparison to other data augmentation methods, KNNMTD has shown better performance. KNNMTD is a hybrid method that combines the k-nearest neighbors algorithm with MTD to identify the closest subsamples and estimate the domain ranges. [21]. Thus, this research applies the KNNMTD method to conduct data augmentation for building SVR-based models.

2.2 Predictive Methods for EOL Products Quantity

The ML methods applied to predict EOL product recycling include neural network (NN) [22], SVR [23], k-nearest neighbor (KNN) [24], decision tree (DT) [25], gradient boosting regression tree (GBRT) [26], extreme gradient boosting (XGBoost) [27], and random forest (RF) [28]. Among these ML models, the SVR algorithm has been shown to be superior in dealing with small sample datasets and can be used to predict the quantity of EOL products [29]. Additionally, this algorithm is well-suited to learn nonlinear behavior and to work with high-dimensional datasets.

SVR-based models are widely employed in RSC management [30], with emphasis on the automobile and EEE industries. Besides, the SVR algorithm has been utilized in conjunction with other ML algorithms to predict the recycled quantity of EOL products, partically municipal solid waste (MSW). For instance, an SVR model optimized by radical basis function (RBF) was proposed to predict annual MSW generation rates, showing an R^2 value of 0.97% [31]. Similarly, the integration of SVR and the autoregressive integrated moving average (ARIMA) approach were applied to develop the MSW generation prediction system [32]. Dai et al. (2020) used fuzzy information granulation (FIG) to predict the variables and then applied an SVR model optimized by a genetic algorithm (GA) to predict the MSW generation [23]. Moreover, a novel method was put forward that combined support vector machine (SVM) with wavelet transform (VT) for weekly prediction of MSW generation [33]. An SVM model integrated with principal component analysis (PCA) was also employed to forest weekly generated waste [34].

In conclusion, the above literature about SVR-based applications for predicting MSW generation indicates that they are a well-established tool for recycled quantity prediction. Thus, this research proposes an optimized SVR-based model to accurately predict the quantity of recycled EOL products.

3. METHODS

The methodology of this research is illustrated below. Initially, Pearson's correlation analysis is performed to select appropriate features for building the predictive model (Section 3.1). Subsequently, the original small sample dataset is split into a training set and a testing set, which are then processed using z-score standardization. The standardized training set is trained by the proposed hybrid predictive model (Sections 3.2-3.5). Finally, the original testing set is utilized to evaluate the prediction performance using four evaluation metrics, including R-squared (R^2), mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE) (Section 3.6). All the processing steps and tested regressors were implemented using Anaconda-based Python programming (Version 3.8).

3.1 Feature Selection

Pearson's correlation analysis is used to select appropriate features for building the predictive model. As a preliminary screening of features for ML modeling, linear correlation coefficients between the quantity of EOL products and their socio-economic variables can be computed using Pearson's correlation analysis, as shown in Eq. (1) [35].

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \quad (1)$$

$\rho_{X,Y}$ is the correlation coefficient between variables X and Y. μ_X and σ_X are the mean and standard deviation of variable X, respectively, and μ_Y and σ_Y are the mean and standard deviation of variable Y. To avoid multi-collinearity problems, features that are significantly correlated with each other are removed.

3.2 Support Vector Regression (SVR)

The SVR algorithm is a modified version of the support vector machine (SVM) that can obtain the global optimal solution by solving a convex quadratic optimization problem. The algorithm works by finding a regression plane using a kernel function to map input data $D=\{(x_1,y_1), (x_2,y_2),\ldots, (x_h,y_h)\}$ to a higher-order vector space and learns a linear model to predict y, as shown in Eq. (2).

$$f(x) = \boldsymbol{\omega}^T x + b \quad (2)$$

$\boldsymbol{\omega}^T$ is the weight vector, and b is the offset.

The insensitive loss coefficient ε is a crucial parameter in the SVR algorithm, which creates a gap of width 2ε around the linear function to make the model less sensitive to outliers. SVR calculates the distance between all sample points and f(x), and the optimal solution can be obtained with the shortest distance, as shown in Eqs. (3) and (4).

$$\min \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{d=1}^{h}\left(\delta_d, \hat{\delta}_d\right) \quad (3)$$

$$\text{s.t.} \begin{cases} f(x_d) - y_d \le \varepsilon + \delta_d \\ y_d - f(x_d) \le \varepsilon + \hat{\delta}_d \\ \delta_d \ge 0, \hat{\delta}_d \ge 0 \ (d = 1,2,\ldots,h) \end{cases} \quad (4)$$

C is the penalty factor. δ_d and $\hat{\delta}_d$ are relaxation factors at upper and lower boundaries.

To deal with nonlinearly separable data, a kernel function is used to map the features to a higher dimensional space. By solving the optimization problem with the kernel function, the optimal decision function can be obtained, as shown in Eqs. (5) and (6).

$$f(x) = \sum_{d=1}^{h}(\hat{a}_d - a_d)k(x, x_d) + b \quad (5)$$

$$k(x, x_d) = \exp\left\{-\frac{|x_d - x|^2}{2g^2}\right\} \quad (6)$$

$k(x, x_d)$ is the kernel function. a_d and $\hat{a}_d$ are Lagrangian function. g is the kernel parameter.

However, practical application of the SVR model is challenging due to the difficulty in selection of penalty factor C and kernel parameter g. A higher value of C can lead to overfitting, while the value of g affects the distribution of mapped data samples in the high-dimensional feature space and the number of support vectors.

3.3 K-nearest Neighbor Mega-trend Diffusion (KNNMTD) Method

In this research, we conducted data augmentation to generate data for predicting the quantity of EOL products based on small samples of numerical data. To achieve this, the artificial virtual samples are generated using the KNNMTD method based on the original training set.

Given a small dataset $\{X_{i,j} \mid i = 1, 2,\ldots, m; j = 1, 2,\ldots, n\}$. Consider an instance i with j attributes [21]. For each of the ith instance's j attributes, the nearest neighbors are discovered iteratively using the KNN method, until all attributes have been exhausted. Then, to obtain the subsample domain ranges, the diffusion coefficient is calculated as Eq. (7), of which k is the sample size.

$$h_{set}^{(i,j)} = \frac{\hat{s}_x^2}{k} = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_k)^2 / k-1}{k} \quad (7)$$

The estimated range of the diffused sample set is shown as Eqs. (8), (9), (10), (11), and (12).

$$a_{(i,j)} = u_{set}^{(i,j)} - Skew_L^{(i,j)} \times \sqrt{-2 \times \hat{s}_x^2 / N_L^{(i,j)} \times \ln(10^{-20})} \quad (8)$$

$$b_{(i,j)} = u_{set}^{(i,j)} + Skew_U^{(i,j)} \times \sqrt{-2 \times \hat{s}_x^2 / N_U^{(i,j)} \times \ln(10^{-20})} \quad (9)$$

$$u_{set}^{(i,j)} = \left(\min_{(i,j)} + \max_{(i,j)}\right)/2 \quad (10)$$

$$Skew_L^{(i,j)} = N_L^{(i,j)} \Big/ \left(N_L^{(i,j)} + N_U^{(i,j)}\right) \quad (11)$$

$$Skew_U^{(i,j)} = N_U^{(i,j)} \Big/ \left(N_L^{(i,j)} + N_U^{(i,j)}\right) \quad (12)$$

$N_L^{(i,j)}$ is the number of data points that are smaller than $u_{set}^{(i,j)}$, and $N_U^{(i,j)}$ is the number of data points that are larger than $u_{set}^{(i,j)}$.

When $\hat{s}_x^2 = 0$, the range are estimated as Eqs. (13) and (14).

$$a_{(i,j)} = \min_{(i,j)}/5 \quad (13)$$

$$b_{(i,j)} = \max_{(i,j)} \times 5 \quad (14)$$

When a and b exclude the minimum and maximum values, the lower bound (*LB*) and upper bound (*UB*) are calculated as Eqs. (15) and (16).

$$LB_{(i,j)} = \begin{cases} a_{(i,j)} & \text{if } a_{(i,j)} \le \min_{(i,j)} \\ \min_{(i,j)} & \text{if } a_{(i,j)} > \min_{(i,j)} \end{cases} \quad (15)$$

$$UB_{(i,j)} = \begin{cases} b_{(i,j)} & \text{if } b_{(i,j)} \ge \max_{(i,j)} \\ \max_{(i,j)} & \text{if } b_{(i,j)} < \max_{(i,j)} \end{cases} \quad (16)$$

The membership function (*MF*) is calculated as Eq. (17).

$$MF\left(x'_{(i,j)}\right) = \begin{cases} \frac{x'_{(i,j)} - LB_{(i,j)}}{u_{set}^{(i,j)} - LB_{(i,j)}} & \text{if } x'_{(i,j)} \le u_{set}^{(i,j)} \\ \frac{UB_{(i,j)} - x'_{(i,j)}}{UB_{(i,j)} - u_{set}^{(i,j)}} & \text{if } x'_{(i,j)} > u_{set}^{(i,j)} \end{cases} \quad (17)$$

The generated data were evaluated using the pairwise correlation difference (PCD) [36]. PCD measures the difference between Pearson correlation matrices (*corr*) of real data (X_r) and synthetic data (X_s) using the Frobenius norm, as shown in Eq. (18). A PCD value closer to 0 indicates that the artificial virtual data is more comparable to the original dataset.

$$PCD = \left\| corr(X_r) - corr(X_s) \right\|_F \quad (18)$$

3.4 Particle Swarm Optimization (PSO) Algorithm

PSO is a powerful global optimization algorithm that utilizes swarm intelligence. Each particle in the algorithm possesses two critical properties: velocity, which determines the speed of the particle, and position, which denotes a solution to the optimization problem. Additionally, each particle has an adaptation value, which is determined by the objective function. The flight of a particle is the process by which it searches for the optimal solution. The optimal solution searched by each particle individually is called the individual extreme value, while the optimal individual extreme value in the particle population is referred to as the current global optimal solution. The flight speed of a particle can be dynamically adjusted based on its historical optimal position and the optimal position of the population. This process continues iteratively until the algorithm obtains an optimal solution that satisfies the termination condition.

Suppose there is a cluster of N particles in a D-dimensional target search space where the position of the *i*th particle is denoted as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, the search speed of the *i*th particle is denoted as $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, the optimal position experienced by the *i*th particle is called the individual pole and is denoted as $P_g = (P_{g1}, P_{g2}, \ldots, P_{gD})$, and the optimal position over which the entire particle population economy passes is called the global extremum, denoted as $g_{best} = (g_1, g_2, \ldots, g_D)$. When finding these two optimal values, each particle updates its velocity and position according to the Eqs. (19) and (20).

$$v_{iD}(k+1) = w \cdot v_{iD}(k) + c_1 \cdot r_1 \cdot (p_{iD}(k) - x_{iD}(k)) + c_2 \cdot r_2 \cdot (p_{gD}(k) - x_{iD}(k)) \quad (19)$$

$$x_{iD}(k+1) = x_{iD}(k) + v_{iD}(k+1) \quad (20)$$

where *w* is the inertia weight, and its value affects the global search ability and local search ability of the PSO algorithm; the larger *w* is the stronger global search ability, and the smaller *w* is the stronger local search ability. c_1 and c_2 are called acceleration constants. c_1 is the individual learning factor of each particle, and c_2 is the global learning factor of each particle. r_1 and r_2 are random numbers between 0 and 1.

The PSO optimization algorithm is highly advantageous due to its remarkable search efficiency and outstanding convergence performance. As a result, this research employs the PSO optimization algorithm to optimize the parameters of the SVR model.

3.5 KNNMTD-PSO-SVR Model

In this section, a novel hybrid model is proposed to predict the recycled quantity of EOL products. This model combines the SVR model and PSO algorithm with artificial virtual samples.

Firstly, the KNNMTD method is utilized to generate an artificial virtual sample with varying k values based on different partition ratios. The PCD value is calculated between the generated artificial virtual samples and the original dataset, and an appropriate ratio and k value are selected to generate an artificial virtual sample. Subsequently, a new training set is formed by merging the original training set with the artificial virtual sample. Next, the PSO-based SVR model is trained using the new training set. The PSO algorithm is employed to determine the optimal values of the parameters *C* and *g*. Finally,

the trained SVR model with optimal parameters C and g is utilized to predict the recycled quantity of EOL products. The hybrid KNNMTD-PSO-SVR model flowchart is depicted in Figure 1.

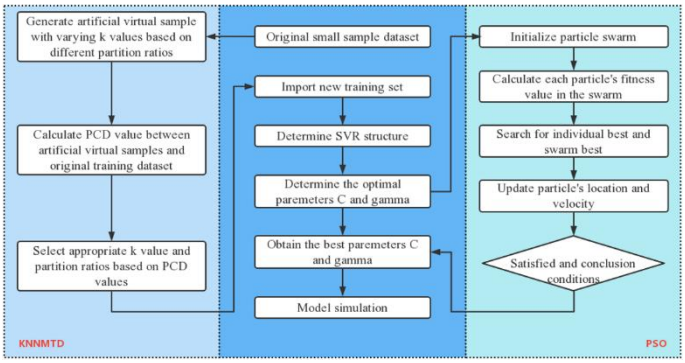

FIGURE 1: THE HYBRID KNNMTD-PSO-SVR MODEL FLOWCHART

3.6 Model Evaluation

The dataset is split into a training set and a testing set. The proposed model trains a new training set that combines the original training set with artificial virtual samples. The performance of the prediction is evaluated using four metrics, namely R^2, MAE, MSE, and MAPE, as shown in Eqs. (21), (22), (23), and (24).

These metrics are employed to analyze the prediction error and evaluate the effectiveness of the regression model. After the selection of an appropriate data processing method, the four performance evaluation metrics are chosen to assess the accuracy of the proposed model's predictions.

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (21)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (22)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (23)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (24)$$

y_i and $\bar{y}_i$ are the original value and average value of variable Y, respectively, and $\hat{y}_i$ represents the predicted value of variable Y.

4. EMPRICAL STUDY

4.1 Data Description

Several researchers have investigated the socioeconomic factors that influence the quantity of recycled ELVs [5, 37-41]. 15 socioeconomic factors were selected that affect the quantity of recycled ELVs [37-41]. These factors include auto production (AP), volume of auto sales (AS), vehicle ownership (VO), number of discarded vehicles (DV), recycled material price (RMP), passenger turnover (PT), number of ELVs industrial employees (IE), number of ELVs collection nodes (CN), number of ELVs enterprises (EP), highway freight turnover (HFT), gross domestic product (GDP), population (POP), number of vehicle drivers (VD), highway mileage (HM), and income per urban resident (IPUR). While most of these variables are sourced from available literature, the number of vehicle drivers is identified as an influential factor for the first time in this research. The necessary historical data for the study has been obtained from various agencies in China (see Table 1). This data is based on monthly observations from China's ELVs industry for the period between 2006 and 2020, with a total of 180 data points after data conditioning and integration.

TABLE 1: DATA SOURCES DURING THE 2006-2020 PERIOD

Data Sources	Historical Data
China Association of Automobile Manufacturers	China's automobile production number Volume of automobile sales in China ELVs industry employees in China China's ELVs industry collection node China's ELVs enterprises
China National Resources Recycling Association	ELVs recycled material price in China
China National Bureau of Statistics	Number of discarded vehicle in China China's vehicle ownership China's highway freight turnover China's passenger turnover China's GDP China's population (15-64 years old) Number of vehicle drivers in China China's highway mileage Income per urban resident in China

4.2 Feature Selection

While a variety of factors can influence the quantity of recycled ELVs, incorporating too many features in the data set can result in over-fitting. To build a prediction model, suitable features are selected using Pearson's correlation analysis. Figure 2 presents the results of the correlation analysis, which ranked the correlation between recycled ELVs quantities and its socio-economic variables. The p-value for a correlation coefficient of 0.8 or greater is smaller than 0.05. To avoid issues with multi-collinearity, highly correlated features (correlation coefficient > 0.8) are removed. More specifically, when two features have a correlation coefficient of 0.8 or greater, the feature with a lower correlation coefficient with the target variable is discarded. As a result, eight features are selected, including AP, PT, POP, VD, HM, IPUR, RMP, and EP. Further, Figure 2's ELVs column displays how recycled ELVs generation is associated with each feature.

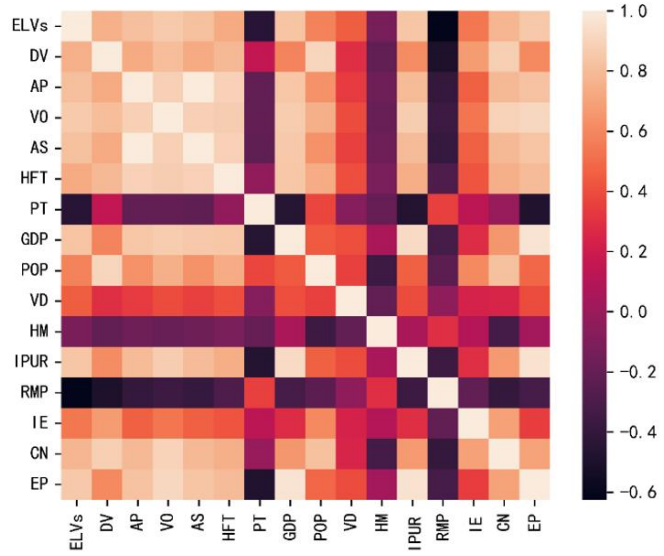

FIGURE 2: PAIRWISE PEARSON CORRELATION OF THE FEATURES TO RECYCLED ELVs

4.3 Analysis of Data Generated by KNNMTD-PSO-SVR

To prevent over-fitting ML-based algorithms due to poor performance with unknown data samples, it is essential to split the dataset into a training set and a testing set. This research evaluated four partition ratios, namely 90:10, 80:20, 70:30, and 60:40. The training set is used to develop the ML-based model, while the testing set is used to check for over-fitting. Artificial virtual samples are created using the KNNMTD approach, with k values varying from 3 to 10 based on the training set. In addition, the virtual sample size is set to 100 [42]. Increasing the virtual sample size improves the model's prediction performance; however, an unreasonable increase may result in many irrational virtual samples, reducing the prediction model's performance.

Table 2 shows the ratio that produces the minimum average PCD value. Figure 3 illustrates the PCD calculated between the generated artificial virtual samples and the original dataset of each partition ratio. It is evident that the appropriate ratio and k value selected for generating an artificial virtual sample are 70:30 and 4, respectively.

TABLE 2: MEAN PCD FOR EACH PARTITION RETIO

Partition ratio	**90:10**	**80:20**	**70:30**	**60:40**
Mean PCD	0.629	0.672	**0.505**	0.617

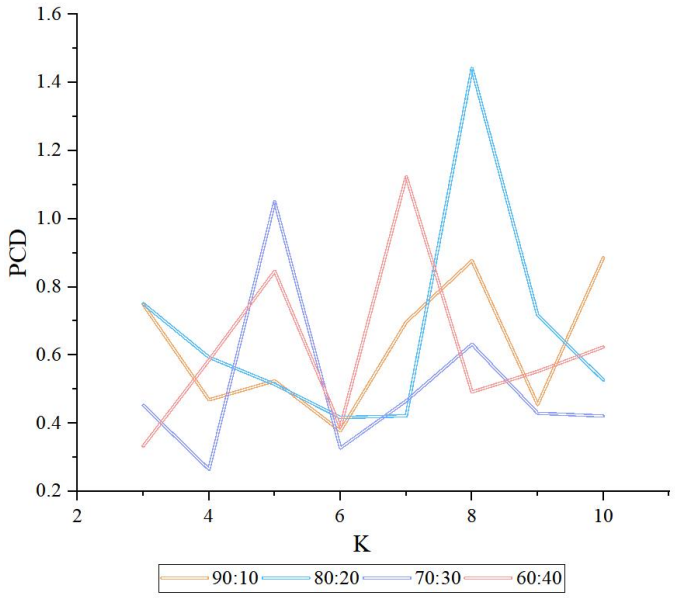

FIGURE 3: PCD WITH DIFFERENT K VALUES FOR EACH PARTITION RATIO

To enhance the prediction performance and reduce the impact of different data measurement units on the models, the z-score standardization method was employed to process the data in this research. During the modeling phase, the standardized original training set is first trained by the SVR model based on the selected ratio (see Figure 4).

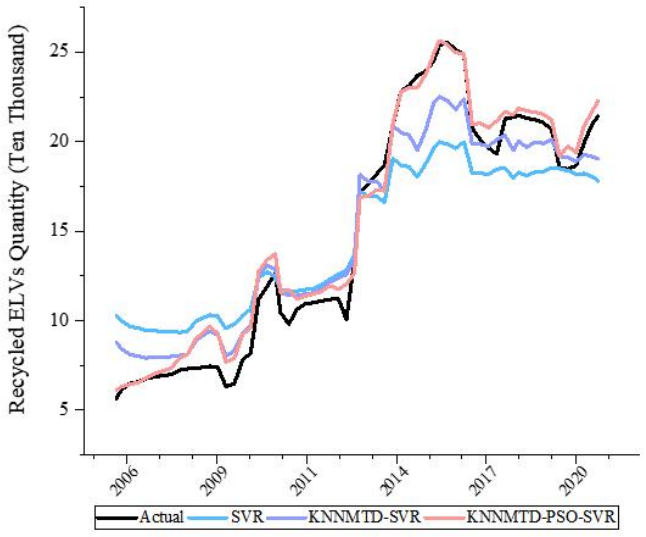

FIGURE 4: THE PREDICTED VALUES OF KNNMTD-PSO-SVR

To boost the performance of prediction results and augment small sample datasets, an artificial virtual sample is generated using the KNNMTD method. The original training set mixed with the artificial virtual sample forms a new training set. The new training set is then trained using the SVR model,

referred to as KNNMTD-SVR (see Figure 4). The original testing set is reserved for evaluating the prediction accuracy of the proposed model trained on this new training set.

TABLE 3: PSO OPTIMIZATION ALGORITHM PARAMETER SETTINGS

Parameter	Parameter setting	Parameter description
c_1	1.5	Local search capability
c_2	1.5	Global searching ability
maxgen	200	Maximum iteration steps
sizepop	20	Population size
Popcmin-popcmax	[0.1, 100]	Value range of parameter C
Popgmin-popgmax	[0.1, 1000]	Value range of parameter g
wmin-wmax	[0.4, 0.9]	Value range of inertia weight

To further improve the prediction accuracy of the SVR model, the PSO algorithm is employed as it offers significant advantages in terms of convergence speed and ability. In this research, the PSO algorithm is used to select the optimal kernel function parameters *g* and penalty factors *C* for the KNNMTD-SVR model. The parameter settings of the KNNMTD-PSO-SVR are provided in Table 3. The hybrid model is trained with the new training set before being tested with the original testing set. The prediction results of KNNMTD-PSO-SVR for the recycled quantity of ELVs in China during the 2006-2020 period (see Figure 4). Table 4 presents the evaluation results of R^2, MAE, MSE, and MAPE predicted by the SVR, KNNMTD-SVR, and KNNMTD-PSO-SVR models.

TABLE 4: THE PREDICTION PERFORMANCE OF KNNMTD-PSO-SVR

Index	SVR	KNNMTD-SVR	KNNMTD-PSO-SVR
R^2	0.8138	0.9357	0.9869
MAE	0.3726	0.2128	0.1136
MSE	0.1890	0.6420	0.2070
MAPE	0.6034	0.4679	0.3925

5. DISCUSSION

Forecasting the quantity of recycled EOL products is crucial for effective planning and operation of reverse supply chain management. The aim of this research is to develop a hybrid KNNMTD-PSO-SVR model (see Figure 1) for predicting the quantity of recycled EOL products using s small numerical data samples from the field of ELVs as a case study. The analysis of the results yielded the following key findings.

The first key finding of the study pertains to the selection of eight socioeconomic features related to ELVs using Pearson's correlation analysis, as shown in Figure 2. This method was used to avoid any problems with multi-collinearity. The selected features include auto production, passenger turnover, population, number of vehicle drivers, highway mileage, income per urban resident, and number of ELVs enterprises. Notably, the number of vehicle drivers, which was proposed as an influence factor in this study, has a positive relationship with ELV generation, thus confirming the assumption that the number of vehicle drivers can impact the quantity of recycled ELVs. The variables income per urban resident and auto production have the highest correlation coefficients of 0.839 and 0.815, respectively, indicating that an increase in residents' income and auto production could lead to higher quantities of recycled ELVs. Previous studies have also demonstrated that income per urban resident and auto production can affect ELV recycling [40, 43].

The second key finding of this research is that the SVR model developed for predicting the quantity of recycled ELVs performs well, with an R-squared value exceeding 0.8, as shown in Figure 4. This result is consistent with previous studies that have demonstrated the effectiveness of SVR for forecasting the generation of municipal solid waste, indicating that it is a reliable method for predicting the quantity of recycled EOL products [24, 33, 44].

Thirdly, the feasibility of predicting the quantity of recycled ELVs using artificial virtual samples supported by the KNNMTD method is demonstrated in this research (see Figure 4). Unlike previous studies on EOL products recycling [45, 46], this research proposes a KNNMTD-SVR model to improve the prediction performance for small sample datasets. Comparison between the KNNMTD-SVR model and the benchmark prediction model over the same period shows that the proposed KNNMTD-SVR model offers better performance (see Table 4). The model's prediction error rate decreases significantly after the addition of 100 virtual samples. Specifically, the average MAE, MSE, and MAPE values decreased by 0.16, 0.12, and 0.14, respectively. Moreover, the KNNMTD-SVR model is better at explaining variation, as indicated by R^2 values, while the SVR model performs poorly. Consequently, after applying artificial virtual samples, the model shows a significant improvement in learning accuracy, meaning that the prediction results of KNNMTD-SVR for recycled ELVs quantity during the 2006-2020 period are much closer to the actual values.

Fourth, the PSO algorithm effectively improves the approximation ability of the KNNMTD-SVR model (see Figure 4 and Table 4). Previous studies have proposed optimized SVR-based models for predicting solid waste generation, such as an SVR model combined with the principal component analysis (PCA) method [34], an SVR model optimized by genetic algorithm (GA) [23], and a hybrid model of wavelet transform (WT) and support vector machine (SVM)[33]. In this research, the PSO algorithm is utilized for the first time in combination with an SVR model to develop prediction models for estimating the quantity of EOL products. The PSO algorithm is applied to select the optimal kernel function parameters *g* and penalty factors *C*, thereby improving the prediction accuracy of the KNNMTD-SVR model. A comparison among the SVR, KNNMED-SVR, and KNNMTD-PSO-SVR models reveals that PSO enhances the prediction accuracy during the 2006-2020 period, making the predicted value of each period closer

to the actual value and contributing to the predicted value of each period in the SVR model.

Finally, the KNNMTD-PSO-SVR model outperforms the SVR and KNNMTD-SVR models in terms of reducing errors in MAE, MSE, and MAPE, indicating that it can effectively improve the prediction accuracy of recycled ELVs quantity. As the number of combined models increases, the prediction error continues to decrease. Notably, the KNNMTD-PSO-SVR model shows almost identical predicted values to the actual values over the analyzed period, which demonstrates the model's suitability for predicting the recycled quantity of not only ELVs but also other EOL products.

In conclusion, several conclusions can be drawn: (i) this research is the first to utilize SVR modelling to forecast the quantity of recycled ELVs, with a selection of appropriate explanatory variables; (ii) the KNNMTD method has been applied for the first time in forecasting the quantity of EOL products, by expanding the original sample set with artificial virtual samples; (iii) this research marks the first instance of combining the PSO algorithm with an SVR model to predict the quantity of EOL products; (iv) the PSO algorithm has been shown to have a positive impact on the predictive model, making it feasible to optimize the KNNMTD-SVR model for recycled EOL product quantity prediction; and (v) the hybrid KNNMTD-PSO-SVR model outperforms the KNNMTD-SVR and SVR models, demonstrating its potential for accurate and reliable forecasting of recycled EOL product quantities. Overall, these findings provide valuable insights for researchers and practitioners working in the field of EOL product recycling and sustainability.

This research makes significant theoretical contributions to prediction techniques and applications. The findings confirm the feasibility of developing the predictive model by applying artificial virtual samples, which provides an effective technique for predicting small sample data. Through numerical case studies and comparative analysis, the study validates the efficacy of the proposed model in predicting recycling quantity based on limited numerical data. The results indicate that the novel KNNMTD-PSO-SVR prediction model is suitable for EOL product recycling. Theoretical forecasting of EOL product recycling quantity, from the perspective of RSC management, not only provides valuable theoretical and practical assistance but also offers recommendations for industrial deployment and improving recycling utilization.

Furthermore, this research has significant implications for decision-makers and practitioners responsible for managing EOL products, a critical component of RSC management. The study assists manufacturers in fulfilling their extended producer responsibility (EPR) and aids industrial engineers in making informed decisions regarding reverse logistics network design, tactical disposal management, capacity planning, and operational production. Moreover, this research supports third-party remanufacturers in developing strategies, projecting profits, and assessing market opportunities. Additionally, the scientific prediction of recycled EOL product quantities can be viewed as a fundamental aspect of a smart RSC system, enabling the smart platform to promote supply chain sustainability through effective resource allocation and regional industry deployment.

However, this research has some limitations: (i) the data used in this research about the ELVs industry may be limited. To enhance the accuracy of the proposed predictive model, future research should incorporate more comprehensive and precise data on other EOL products; and (ii) this study only considered some socioeconomic factors and did not take into account relevant policy factors or customer willingness to return EOL products. To address these limitations, the authors recommend the following suggestions for future research: (i) exploring additional independent variables such as recycling laws and regulations, economic subsidies, and preferential policies for EOL product recycling, as well as individual factors such as recycling awareness, customer willingness, and educational level; and (ii) further investigating hybrid and ensemble ML-based models to enhance the prediction accuracy of EOL product quantity.

6. CONCLUSION

This research proposed a KNNMTD-PSO-SVR model to forecast EOL product recycling. The model exhibited fast training times of a few seconds and improved prediction accuracy compared to the SVR and KNNMTD-SVR models. The proposed hybrid model may be used to plan, design, and implement future integrated RSC management action strategies.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from Sustainable Manufacturing Systems Centre at Cranfield University.

REFERENCES

[1] Sudusinghe, J. I., and Seuring, S., 2021, "Supply chain collaboration and sustainability performance in circular economy: A systematic literature review," Int. J. Product. Econ., p. 108402.

[2] Tseng, M.-L., Bui, T.-D., Lim, M. K., Fujii, M., and Mishra, U., 2022, "Assessing data-driven sustainable supply chain management indicators for the textile industry under industrial disruption and ambidexterity," Int. J. Product. Econ., 245, p. 108401.

[3] Ayvaz, B., Bolturk, E., and Kaçtıoğlu, S., 2014, "A grey system for the forecasting of return product quantity in recycling network," International Journal of Supply Chain Management, 3(3).

[4] Das, D., and Dutta, P., 2022, "Product return management through promotional offers: The role of consumers' loss aversion," Int. J. Product. Econ., p. 108520.

[5] Xia, H., Han, J., and Milisavljevic-Syed, J., "Forecasting the Number of End-of-Life Vehicles: State of the Art Report," Proc. 16th International Design Conference (Online), Cambridge University Press.

[6] A., 2021, "The Automobile Industry Pocket Guide 2021-2022,".

[7] Ravi, V., and Shankar, R., 2017, "An ISM-based approach analyzing interactions among variables of reverse logistics in automobile industries," Journal of Modelling in Management.

[8] Lin, H.-T., Nakajima, K., Yamasue, E., and Ishihara, K., 2018, "Recycling of End-of-Life Vehicles in Small Islands: The Case of Kinmen, Taiwan," Sustainability, 10(12), pp. 4377-4390.

[9] Sakai, S.-i., Yoshida, H., Hiratsuka, J., Vandecasteele, C., Kohlmeyer, R., Rotter, V. S., Passarini, F., Santini, A., Peeler, M., and Li, J., 2014, "An international comparative study of end-of-life vehicle (ELV) recycling systems," Journal of Material Cycles Waste Management, 16(1), pp. 1-20.

[10] Lu, G. Y., and Wong, D. W., 2008, "An adaptive inverse-distance weighting spatial interpolation technique," Computers geosciences, 34(9), pp. 1044-1055.

[11] Van Beers, W. C., and Kleijnen, J. P., 2003, "Kriging for interpolation in random simulation," J. Oper. Res. Soc., 54(3), pp. 255-262.

[12] Fan, Q., Efrat, A., Koltun, V., Krishnan, S., and Venkatasubramanian, S., "Hardware-Assisted Natural Neighbor Interpolation," Proc. ALENEX/ANALCO, pp. 111-120.

[13] Habermann, C., and Kindermann, F., 2007, "Multidimensional spline interpolation: Theory and applications," Computational Econ., 30(2), pp. 153-169.

[14] Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P., "Smote for regression," Proc. Portuguese conference on artificial intelligence, Springer, pp. 378-389.

[15] Han, H., Wang, W.-Y., and Mao, B.-H., "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," Proc. International conference on intelligent computing, Springer, pp. 878-887.

[16] He, H., Bai, Y., Garcia, E. A., and Li, S., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proc. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp. 1322-1328.

[17] Ho, K. I.-J., Leung, C.-S., and Sum, J., 2010, "Convergence and objective functions of some fault/noise-injection-based online learning algorithms for RBF networks," ITNN, 21(6), pp. 938-947.

[18] Di Bella, A., Fortuna, L., Graziani, S., Napoli, G., and Xibilia, M., "Development of a Soft Sensor for a Thermal Cracking Unit using a small experimental data set," Proc. 2007 IEEE international symposium on intelligent signal processing, IEEE, pp. 1-6.

[19] Li, D. C., Wu, C. S., Tsai, T. I., and Lina, Y. S., 2007, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," Computers Operations Research, 34(4), pp. 966-982.

[20] Lin, Y. S., and Li, D. C., 2010, "The Generalized-Trend-Diffusion modeling algorithm for small data sets in the early stages of manufacturing systems," European Journal of Operational Research, 207(1), pp. 121-130.

[21] Sivakumar, J., Ramamurthy, K., Radhakrishnan, M., and Won, D., 2022, "Synthetic sampling from small datasets: A modified mega-trend diffusion approach using k-nearest neighbors," Knowledge-Based Systems, 236, p. 107687.

[22] Puntarić, E., Pezo, L., Zgorelec, Ž., Gunjača, J., Kučić Grgić, D., and Voća, N., 2022, "Prediction of the Production of Separated Municipal Solid Waste by Artificial Neural Networks in Croatia and the European Union," Sustainability, 14(16), p. 10133.

[23] Dai, F., Nie, G.-h., and Chen, Y., 2020, "The municipal solid waste generation distribution prediction system based on FIG–GA-SVR model," Journal of Material Cycles Waste Management, 22(5), pp. 1352-1369.

[24] Abbasi, M., and El Hanandeh, A., 2016, "Forecasting municipal solid waste generation using artificial intelligence modelling approaches," Waste Manage., 56, pp. 13-22.

[25] Kannangara, M., Dua, R., Ahmadi, L., and Bensebaa, F., 2018, "Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches," Waste Manage., 74, pp. 3-15.

[26] Lu, W., Huo, W., Gulina, H., and Pan, C., 2022, "Development of machine learning multi-city model for municipal solid waste generation prediction," Frontiers of Environmental Science Engineering, 16(9), pp. 1-10.

[27] Zhang, C., Dong, H., Geng, Y., Liang, H., and Liu, X., 2022, "Machine learning based prediction for China's municipal solid waste under the shared socioeconomic pathways," J. Environ. Manage., 312, p. 114918.

[28] Nguyen, X. C., Nguyen, T. T. H., La, D. D., Kumar, G., Rene, E. R., Nguyen, D. D., Chang, S. W., Chung, W. J., Nguyen, X. H., and Nguyen, V. K., 2021, "Development of machine learning-based models to forecast solid waste generation in residential areas: A case study from Vietnam," Resources, Conservation, Recycling, 167, p. 105381.

[29] Al-Anazi, A. F., and Gates, I. D., 2012, "Support vector regression to predict porosity and permeability: Effect of sample size," Computers geosciences, 39, pp. 64-76.

[30] Erkinay Ozdemir, M., Ali, Z., Subeshan, B., and Asmatulu, E., 2021, "Applying machine learning approach in recycling," Journal of Material Cycles Waste Management, 23(3), pp. 855-871.

[31] Jassim, M. S., Coskuner, G., and Zontul, M., 2022, "Comparative performance analysis of support vector regression and artificial neural network for prediction of municipal solid waste generation," Waste Manag. Res., 40(2), pp. 195-204.

[32] Yi, C., and Feng, D., 2020, "Integrating SVR and ARIMA Approach to Build the Municipal Solid Waste Generation Prediction System," Journal of Computers, 31(3), pp. 216-225.

[33] Abbasi, M., Abduli, M., Omidvar, B., and Baghvand, A., 2014, "Results uncertainty of support vector machine and hybrid of wavelet transform-support vector machine models for solid waste generation forecasting," Environmental Progress Sustainable Energy, 33(1), pp. 220-228.

[34] Noori, R., Abdoli, M., Ghasrodashti, A. A., and Jalili Ghazizade, M., 2009, "Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad," Environmental Progress Sustainable Energy, 28(2), pp. 249-258.

[35] Albright, S. C., Winston, W. L., Zappe, C. J., and Broadie, M. N., 2011, Data analysis and decision making, South-Western/Cengage Learning.

[36] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P., 2020, "Generation and evaluation of synthetic patient data," BMC Med. Res. Methodol., 20(1), pp. 1-40.

[37] Yano, J., Muroi, T., and Sakai, S.-i., 2015, "Rare earth element recovery potentials from end-of-life hybrid electric vehicle components in 2010–2030," J. Mater. Cycles Waste Manage., 18(4), pp. 655-664.

[38] Hu, S., and Kurasaka, H., 2013, "Projection of end-of-life vehicle (ELV) population at provincial level of China and analysis on the gap between the future requirements and the current situation of ELV treatment in China," J. Mater. Cycles Waste Manage., 15(2), pp. 154-170.

[39] Ochotnicky, P., Kacer, M., and Alexy, M., 2017, "Sustainability of the ELV processing system in the Slovak Republic and forecasting of waste streams from the operation of passenger motor vehicles," Waste Forum, 5, pp. 452-467.

[40] Xin, F., Ni, S., Li, H., and Zhou, X., 2018, "General Regression Neural Network and Artificial-Bee-Colony Based General Regression Neural Network Approaches to the Number of End-of-Life Vehicles in China," IEEE Access, 6, pp. 19278-19286.

[41] Hao, H., Zhang, Q., Wang, Z., and Zhang, J., 2018, "Forecasting the number of end-of-life vehicles using a hybrid model based on grey model and artificial neural network," Journal of Cleaner Production, 202, pp. 684-696.

[42] Li, D. C., Huang, W. T., Chen, C. C., and Chang, C. J., 2013, "Employing virtual samples to build early high-dimensional manufacturing models," IJPR, 51(11), pp. 3206-3224.

[43] Sokić, M., manojlović, V., Marković, B., and Štrbac, N., 2016, "Modeling and Prediction of the end of Life Vehicles Number Distribution in Serbia," Acta Polytechnica Hungarica, 13(4), pp. 159-172.

[44] Meza, J. K. S., Yepes, D. O., Rodrigo-Ilarri, J., and Cassiraga, E., 2019, "Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks," Heliyon, 5(11), p. e02810.

[45] Poschmann, H., Brüggemann, H., and Goldmann, D., 2021, "Fostering end-of-life utilization by information-driven robotic disassembly," Procedia CIRP, 98, pp. 282-287.

[46] Abou Baker, N., Szabo-Müller, P., and Handmann, U., 2021, "Transfer learning-based method for automated e-waste recycling in smart cities," EAI Endorsed Transactions on Smart Cities, 5(16), pp. e1-e1.