Resources, Conservation & Recycling 197 (2023) 107073

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

journal homepage: www.elsevier.com/locate/resconrec

Full length article

Predictive modeling for the quantity of recycled end-of-life products using optimized ensemble learners

Hanbing Xia [a], Ji Han [b], Jelena Milisavljevic-Syed [a,*]

[a] Sustainable Manufacturing Systems Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, MK43 0AL, UK
[b] University of Exeter Business School, London, SE1 7TY, UK

ARTICLE INFO

Keywords:
Sustainable reverse supply chain
End-of-life products
Machine learning
Predictive analysis
Ensemble model

ABSTRACT

The rapid development of machine learning algorithms provides new solutions for predicting the quantity of recycled end-of-life products. However, the Stacking ensemble model is less widely used in the field of predicting the quantity of recycled end-of-life products. To fill this gap, we propose a Stacking ensemble model that utilizes support vector regression, multi-layer perceptrons, and extreme gradient boosting algorithms as base models, and linear regression as the meta model. The k-nearest neighbor mega-trend diffusion method is applied to avoid overfitting problems caused by a small sample data set. The grid search and time series cross validation methods are utilized to optimize the proposed model. To verify and validate the proposed model, data related to China's end-of-life vehicles industry from 2006 to 2020 is used. The experimental results demonstrate that the proposed model achieves higher prediction accuracy and generalization ability in predicting the quantity of recycled end-of-life products.

1. Introduction

In recent years, traditional supply chain management has transformed into sustainable supply chain management due to growing ecological awareness and legal regulations (Masoumi et al., 2019). Efficient management of the reverse supply chain plays a significant role in sustainable management (Lenort et al., 2021). To design an effective system for reverse supply chain management, it is important to manage end-of-life (EOL) products (Rashid et al., 2021). EOL products can be considered a vital source of secondary raw materials through recycling, reusing, and remanufacturing (Numfor et al., 2021). In fact, the amount of EOL products generated each year is increasing at an alarming rate. For example, the world currently generates around 50 million tons of waste electrical and electronic equipment (WEEE) (Andeobu et al., 2021) and 2.01 billion tons of municipal solid waste (MSW) (Namoun et al., 2022) yearly. It is challenging to manage EOL products due to their uncertainty in quality, quantity, and return time (Hao et al., 2018). If the quantity of EOL products can be predicted in advance, accurate information will help decision-makers and practitioners effectively regulate the resources necessary for designing the reverse supply chain.

There have been considerable conventional methods designed for predicting the quantity of EOL products, including the population balance model (Lin et al., 2018), the market supply method (Jain and Sareen, 2006), the distribution delay method (Polák and Drápalová, 2012), the structural equation model (Agrawal and Singh, 2019), the time series model (Ochotnicky et al., 2017), the gray model (Ene and Öztürk, 2017), the graphical evaluation and review technique (Zhou et al., 2016), etc. Machine learning (ML) methods currently stand out for their superior accuracy in predicting EOL products due to their superior performance with unstable nonlinear data samples and large feature sizes (Ni et al., 2021). In the field of ML, ensemble models have attracted much interest and have proven to be highly predictive in a variety of applications (Cui et al., 2021). However, the Stacking ensemble model is relatively underutilized in predicting the quantity of recycled EOL products.

To address the gap, we propose a novel Stacking-based ensemble model to predict the quantity of EOL products. Our approach combines multiple machine learning algorithms to improve prediction accuracy and generalization ability, leading to better management of sustainable reverse supply chains and increased sustainability of recycling industry.

The paper is organized in the following way. Section 2 includes a detailed literature review of related research. Section 3 proposes a Stacking-based prediction model. Section 4 details the empirical study of the proposed model and the experimental results. Finally, the conclusion

* Correspondence author.
E-mail address: jelenams@cranfield.ac.uk (J. Milisavljevic-Syed).

https://doi.org/10.1016/j.resconrec.2023.107073
Received 4 January 2023; Received in revised form 12 May 2023; Accepted 31 May 2023
Available online 9 June 2023

and future research directions are shown in Section 5.

2. Literature review

In this section, the ML-based predictive methods for the quantity of EOL products are reviewed. The EOL products in this research refer to end-of-life vehicles (ELVs), medical waste (MW), MSW, and WEEE. Besides, the ML methods applied to predict EOL product recycling mainly include artificial neural network (ANN), support vector regression (SVR), k-nearest neighbor (KNN), decision tree (DT), gradient boosting regression tree (GBRT), extreme gradient boosting (XGBoost), and random forest (RF).

Artificial Neural Network (ANN) The NN algorithm consists of input neurons, middle neurons, and output neurons. The relationships between various input data can be built by the NN algorithm by simulating the human brain. This algorithm has a simple structure with a fast-training speed, but it is easy to fall into local optimality. ANN (Puntarić et al., 2022) is commonly employed in EOL product quantity prediction, which includes several algorithms such as feed forward neural network (FFNN) (Abbasi and El Hanandeh, 2016), back propagation neural network (BPNN) (Oguz-Ekim, 2021), deep neural network (DNN) (Nguyen et al., 2021), general regression neural network (GRNN) (Sodanil and Chatthong, 2014), adaptive network-based fuzzy inference system (ANFIS) (Golbaz et al., 2019), Elman neural network (ENN) (Meza et al., 2019), recurrent neural network (RNN) (Li and Ma, 2019), long short-term memory (LSTM) (Wang et al., 2022), and nonlinear autoregressive (NAR) (Kumar and Kumar, 2021). Multi-layer perceptrons (MLP) (Coskuner et al., 2021), which is a special type of FFNN with a more complex structure and stronger expressive power, have been put forward to predict EOL products. Additionally, a hybrid forecasting model based on autoregressive integrated moving average (ARIMA) methodology and ANN was proposed to predict ELVs in Brazil (de Souza et al., 2022). Further, several artificial intelligent algorithms were used to improve the performance of ANN models, including Bayesian optimization (BO), particle swarm optimization (PSO) (Elshaboury et al., 2021), artificial bee colony (ABC) (Xin et al., 2018), structural break (SB) analysis (Adamović et al., 2017), and genetic algorithm (GA) (Tian et al., 2013). Moreover, other methods such as principal component analysis (PCA) (Liu et al., 2022), gray model (GM) (Hao et al., 2018), triple exponential smoothing (TES) (Hao et al., 2021), and discrete wavelet theory (DWT) (Soni et al., 2019) were employed to assist the ANN models.

Support Vector Regression (SVR) The sample is mapped to high-dimensional space by the kernel function, and the hyperplane is used for regression. This algorithm has a high learning capacity for high-dimensional small sample data, but it is overly reliant on the kernel function (Zhang et al., 2022). A hybrid model of fuzzy information granulation (FIG), GA, and SVR was proposed to predict the MSW generation per capita for Hubei province in China (Dai et al., 2020). Further, the SVR optimized by the wavelet transform (WT) was used to forecast weekly MSW in Tehran and Mashhad (Abbasi et al., 2014).

K-Nearest Neighbor (KNN) As a nonparametric and instance-based lazy learning algorithm, KNN is known for its stability in the presence of noise. The weighted KNN algorithm has been developed and successfully applied to forecast MSW generation in Australia (Abbasi and El Hanandeh, 2016).

Decision Tree (DT) As a supervised ML algorithm, DT consists of root nodes, internal nodes, and leaf nodes. This algorithm shows great interpretability. DT has been employed to evaluate MSW generation in the city of Bogota, providing a possible decision-making strategy for waste disposal (Kannangara et al., 2018).

Gradient Boosting Regression Tree (GBRT) Based on the boosting strategy, the GBRT algorithm makes a joint decision by iterating multiple trees; that is, each tree gets its predicted value by learning the conclusions and residuals of all previous trees. This algorithm has strong robustness to outliers but is unsuitable for high-dimensional sparse data (Lu et al., 2022). The combination of GBRT and ANN was applied to predict building-level MSW generation in New York (Kontokosta et al., 2018).

Extreme Gradient Boosting (XGBoost) The XGBoost algorithm generates a tree according to feature splitting and continuously adds trees to fit the residual of the last prediction, so as to obtain new functions and improve model performance through gradual iteration. This algorithm can prevent overfitting effectively but is unsuitable for processing high-dimensional feature data and unstructured data (Zhang et al., 2022).

Random Forest (RF) RF is one of the classification and regression tree (CART) models based on Bagging integration. This algorithm has high accuracy in training results and good parallelism, but it performs poorly on small data sets (Nguyen et al., 2021).

Ensemble Model Ensemble learning is a type of hybrid ML model in which different or the same type of algorithm can be added multiple times to form a more powerful prediction model (Dasarathy and Sheela, 1979; Tan et al., 2019). Ensemble learning has three main ensemble models, namely Boosting, Bagging, and Stacking. Boosting has a strong dependence between individual learners and a serialization method that must be generated sequentially; that is, the next learner needs to delete a learner to learn, which cannot be parallelized (Freund and Schapire, 1997). Bagging is a parallelization method that can be generated simultaneously without strong dependence between individual learners (Breiman, 1996). Stacking is a parallel, phased ensemble method that adds a meta model layer to multiple heterogeneous base models and then outputs the prediction results. A decomposition-ensemble-based model integrating variational model decomposition (VMD), an exponential smoothing model (ESM), and GM was proposed for e-waste quantity prediction (Wang et al., 2021). Moreover, an ensemble voting regression algorithm based on RF, gradient boosting machine (GBM), and adaptive boosting (AdaBoost) was developed to predict the medical waste for Istanbul in Turkey (Erdebilli and Devrim-İçtenbaş, 2022).

Bagging and Boosting often choose the same model as the base models. The correlation between the models is greater, and the overfitting problem is easy to occur. In contrast, Stacking selects different models as base models to capture the correlation between the predicted results and the actual data more effectively. However, Stacking is less widely used in the field of sustainable reverse supply chains. Thus, to solve the disadvantage of a single model with weak generalization ability in the recycling field, this research proposes a novel Stacking ensemble model to predict the quantity of EOL products.

3. Method

The proposed method is described in this section. First, socioeconomic influence factors for EOL products are summarized from previous studies. The historical data for these variables is processed by z-score standardization and data augmentation (Section 3.1). Second, the proposed optimized Stacking ensemble model is developed (Section 3.2). Finally, three evaluation metrics, namely mean absolute error (MAE), mean square error (MSE), and R-squared (R^2), are used to evaluate the prediction performance of the proposed model (Section 3.3). Anaconda-based Python programming (version 3.8) is used to analyze data and build ML-based predictive models.

3.1. Data preprocessing

To eliminate the influence of the data's various attributes, the original values x and y will be standardized based on the mean (μ) and standard deviation (σ), as shown in Eqs. (1) and (2).

$$x' = \frac{x_i - \mu_x}{\sigma_x} = \frac{x_i - \frac{1}{n}\sum_{i=1}^{n} x_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2}} \quad (1)$$

$$y' = \frac{y_i - \mu_y}{\sigma_y} = \frac{y_i - \frac{1}{n}\sum_{i=1}^{n} y_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}} \quad (2)$$

To avoid overfitting of a small sample data set, data is augmented using the K-nearest neighbor mega-trend diffusion (KNNMTD) method (Sivakumar et al., 2022). By increasing the number of samples and expanding the data set, the limited data can be effectively utilized to the maximum extent, and the generalization of the ML model can be improved.

Consider the data point $X_{(i,j)}$, which means the instance i has j attributes. First, the KNN algorithm iteratively finds the nearest neighbors of $X_{(i,j)}$, which serves as the input for mega-trend diffusion (MTD). Then, to obtain the subsample domain ranges, the diffusion coefficient is calculated as Eq. (3).

$$h_{set}^{(i,j)} = \frac{\hat{s}_x^2}{k} = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_k)^2 \Big/ (k-1)}{k} \quad (3)$$

Where the superscript *(i,j)* represents the MTD parameter values that correspond to the *jth* attribute of the *ith* instance, $\hat{s}_x^2$ represents the sample variance, and k represents the sample size.

The estimated range of the diffused sample set is shown as Eqs. (4)–(8).

$$a_{(i,j)} = u_{set}^{(i,j)} - Skew_L^{(i,j)} \times \sqrt{-2 \times \hat{s}_x^2 \Big/ N_L^{(i,j)} \times ln\left(10^{-20}\right)} \quad (4)$$

$$b_{(i,j)} = u_{set}^{(i,j)} + Skew_U^{(i,j)} \times \sqrt{-2 \times \hat{s}_x^2 \Big/ N_U^{(i,j)} \times ln\left(10^{-20}\right)} \quad (5)$$

$$u_{set}^{(i,j)} = \left(min_{(i,j)} + max_{(i,j)}\right)/2 \quad (6)$$

$$Skew_L^{(i,j)} = N_L^{(i,j)} \Big/ \left(N_L^{(i,j)} + N_U^{(i,j)}\right) \quad (7)$$

$$Skew_U^{(i,j)} = N_U^{(i,j)} \Big/ \left(N_L^{(i,j)} + N_U^{(i,j)}\right) \quad (8)$$

Where $N_L^{(i,j)}$ is the number of data points that are smaller than $u_{set}^{(i,j)}$, $N_U^{(i,j)}$ is the number of data points that are larger than $u_{set}^{(i,j)}$, and the minimum and maximum value of the neighboring subsamples of *(i,j)*th instance are represented by $\min_{(i,j)}$ and $\max_{(i,j)}$, respectively.

When $\hat{s}_x^2 = 0$, the range are estimated as Eqs. (9) and (10).

$$a_{(i,j)} = min_{(i,j)}/5 \quad (9)$$

$$b_{(i,j)} = max_{(i,j)} \times 5 \quad (10)$$

When a and b exclude the minimum and maximum values, the lower bound (*LB*) and upper bound (*UB*) are calculated as Eqs. (11) and (12).

$$LB_{(i,j)} = \begin{cases} a_{(i,j)} \ if \ a_{(i,j)} \le min_{(i,j)} \\ min_{(i,j)} \ if \ a_{(i,j)} > min_{(i,j)} \end{cases} \quad (11)$$

$$UB_{(i,j)} = \begin{cases} b_{(i,j)} \ if \ b_{(i,j)} \ge max_{(i,j)} \\ max_{(i,j)} \ if \ b_{(i,j)} < max_{(i,j)} \end{cases} \quad (12)$$

The membership function (MF) is calculated as Eq. (13).

$$MF\left(x'_{(i,j)}\right) = \begin{cases} \dfrac{x'_{(i,j)} - LB_{(i,j)}}{u_{set}^{(i,j)} - LB_{(i,j)}} \ if \ x'_{(i,j)} \le u_{set}^{(i,j)} \\ \dfrac{UB_{(i,j)} - x'_{(i,j)}}{UB_{(i,j)} - u_{set}^{(i,j)}} \ if \ x'_{(i,j)} > u_{set}^{(i,j)} \end{cases} \quad (13)$$

To measure the performance between actual data and artificial virtual data, the pairwise correlation difference (PCD) is calculated using the Frobenius norm as Eq. (14).

$$PCD = \| corr(X_r) - corr(X_s) \|_F \quad (14)$$

Where X_r is the actual data matrices, X_s is the artificial virtual data matrices, and *corr* is the Pearson correlation matrices of X_r and X_s.

3.2. Model building

As a parallel ensemble learning strategy, Stacking contains multi-layer learning structures. It is essentially about training different ML algorithms on data from various data spaces and data structure perspectives. The Stacking ensemble model structure consists of two learning layers: the first one is the base model, comprising multiple heterogeneous ML models, while the second one is the meta model. The first layer employs the entire training set to train various base models and obtain the predicted values. On the other hand, the second layer trains the true values and predicted values obtained by the base models. Stacking can resolve the insufficient upper limits of a single model's learning ability, avoid the redundancy of the prediction model, and ensure prediction accuracy.

The goal of parameter optimization is to find a set of parameters that brings the model's generalization error as close to zero as possible. The generalization of a model can be negatively affected if it is too complex, resulting in overfitting and a high generalization error, and vice versa. In this research, a combination of time series cross validation and the grid search method is used to find the optimal parameter group of the Stacking model. Firstly, the grid search method enumerates all the model parameter combinations through the set of parameter values. Then, the model parameter combination with the highest average generalization ability score value is output by using time series cross validation.

(1) Grid search

The commonly ML-based method usually adjusts parameters use random search, Bayesian optimization, and grid search. Random search allows for manual control over the number of searches, but each search may yield different results. Bayesian optimization can record the previous search results for the next search, but it is easy to fall into the trap of local optimization instead of global optimization. In comparison, grid search, although the most time-consuming, can be exhaustive of all possible results, and the results are the same every time.

To improve the prediction accuracy, grid search method is selected to adjust parameter for Stacking model in this research since the experimental data are not very large. The steps of the grid search method are as follows:

Step 1: Initialize the mesh size, set the step distance, and define the parameter initial values;
Step 2: Loop through each set of parameter combinations;
Step 3: The parameter values of each parameter combination are used to train the Stacking model in combination with the time series cross validation, respectively. The R-squared value of the model is obtained, and the parameter value of the parameter combination is defined as the *best*;
Step 4: If a better combination of hyperparameters is found, replace the previous *best*;
Step 5: Combine the best hyperparameters to obtain the optimal parameter set, train the final model, and output the optimal Stacking model.

(2) Time series cross validation (TSCV)

To prevent overfitting and improve generalization ability, while also considering the temporal sequence of the dataset, the base model of the Stacking model should be trained by combining time series cross

validation (see Fig. 1), and then the output results will be used to train the meta model. The steps are as follows:

Step 1: Assume that the original data set is *(X, Y)*, the training set is $F = (X\_train, Y\_train)$, and the test set is $T = (X\_test, Y\_test)$. Firstly, the original training data set F was split into five consecutive and non-overlapping subsets: F_i $(i = 1,2,\ldots 5)$ based on time order;
Step 2: One of F_i is the test set, and the remaining four subsets are the training set for training the base model M_i. The trained model M_i is obtained, and the model M_i is used to predict the test set F_i to get the result P_{ii} $(i = 1,2\ldots 5)$, and the prediction result of the base model M_i to the original test set T is denoted as R_i $(i = 1,2\ldots 5)$;
Step 3: The obtained prediction results P_{ii} are then concatenated in chronological order to obtain the training data set P of the second layer meta model, which has the same number of samples as the original training data set F;
Step 4: To predict the result R_i and calculate the mean value to get the test set R of the meta model.

3.3. Statistical measures for model evaluation

This research uses the following three metrics to evaluate the effect and prediction error of the proposed model, which is mean absolute error (MAE), mean square error (MSE), and R-squared (R^2), as shown in Eqs. (15)–(17).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i| \tag{15}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2 \tag{16}$$

$$R^2(y_i, \widehat{y_i}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{17}$$

y_i and $\overline{y_i}$ are the original value and average value of variable Y, respectively, and $\widehat{y_i}$ represents the predicted value of variable Y.

4. Empirical study

To verify and validate the effectiveness of the proposed model, we use the data related to China's ELVs industry from 2006 to 2020. This research involves simulating recycled ELVs generation using seven different ML models, namely SVR, XGBoost, light gradient boosting machine (LGBM), RF, MLP, GBRT, and DT, in order to find the best predictive base models with less correlation. Before commencing the modeling process, we use grid search and time series cross validation to determine the best structure for each base model by obtaining model parameters. These parameters vary according to each model theory as discussed above. We develope Stacking ensemble models by combining the optimal base models with the meta model. Our main objective is to validate the prediction performance of our proposed Stacking model through empirical analysis, and compare it to other proposed models. The overall framework of this empirical study comprises five steps, as shown in Fig. 1.

4.1. Data collection

In general, to build a prediction model or make decisions for a problem, ML algorithms develop the relationships between input variables and output variables based on empirical data (Erkinay Ozdemir et al., 2021). Based on previous studies (Hao et al., 2018; Hu and Kurasaka, 2013; Ochotnicky et al., 2017; Xin et al., 2018; Yano et al., 2015), eight socioeconomic factors that influence the quantity of recycled ELVs are selected, including the number of auto production, passenger turnover, population, vehicle drivers, recycled material price, income of per urban resident, highway mileage, and the number of ELVs enterprise. These historical data on a monthly basis were extracted from the China Association of Automobile Manufacturers, the China National Resources Recycling Association, and the China National Bureau of Statistics.

4.2. Data augmentation

Augmented data are generated by integrating the original data set with artificial virtual samples in order to improve the generalization ability of ML models and prediction performance for small sample data sets. The artificial virtual samples are generated using the KNNMTD method. The artificial virtual sample size is set at 100 (Li et al., 2013). This is because an unreasonable increase in the artificial virtual sample size may lead to irrational virtual samples. The PCD with varying values of $k = [3, 10]$ is calculated, and the appropriate k value is 4. The evaluation results of MAE, MSE, and R^2 predicted by ML models with and without the use of KNNMTD method are presented in Table 1.

4.3. Stacking ensemble model

Based on the literature review analysis, the selected base models of Stacking mainly include SVR, XGBoost, LGBM, RF, MLP, GBRT, and DT. The optimal combination of parameters for these single ML algorithms with the use of the KNNMTD method is found by using grid search and time series cross validation, as shown in Table 2.

The Stacking ensemble model requires that the base model select heterogeneous single ML algorithms with excellent learning performance. This is because the smaller the correlation between the base models, the lower the variance of the Stacking model. The meta model of Stacking model requires strong robustness and generalization ability. To prevent model overfitting and improve prediction accuracy, the linear regression (LR) algorithm is selected as the meta model.

Besides, Table 1 shows that these single ML algorithms have a strong learning ability to predict the quantity of ELVs. Even though XGBoost, LGBM, GBRT, DT, and RF have different algorithm principles, they are all tree-based models with a similar data processing method. Thus, these tree-based models have a high correlation with each other. SVR and MLP are fundamentally different from these tree-based models, so the correlation between SVR, MLP, and other models is low. Therefore, the base model of the Stacking ensemble model is developed by SVR, MLP, and five other tree-based models, respectively. The final prediction performance of each Stacking model is shown in Table 3 and Fig. 2.

Considering the accuracy and difference in ML models, this research selects the Stacking 1 model, namely SVR, MLP, and XGBoost, as the base models and LR as the meta model to construct the Stacking ensemble model.

To further evaluate the performance of the proposed model, we utilize the learning curve to identify potential overfitting. In general, the learning curve plots the model's performance on both training data and testing data at different training set sizes. The learning curve usually consists of two lines representing loss of training data and loss of testing data, which is measured by the value of the MSE in our research. When drawing the learning curve, the training examples are set as the horizontal coordinate, and the MSE of the training set and verification are set as the vertical coordinate, as illustrated in Fig. 3. After data preprocessing, the generalization ability of our proposed Stacking model without using KNNMTD method is shown in Fig. 3(a), while the generalization ability of the proposed Stacking model after data augmentation is demonstrated in Fig. 3(b).

5. Discussion

The aim of this research is to develop a Stacking-based ensemble model to predict the quantity of recycled EOL products for a sustainable

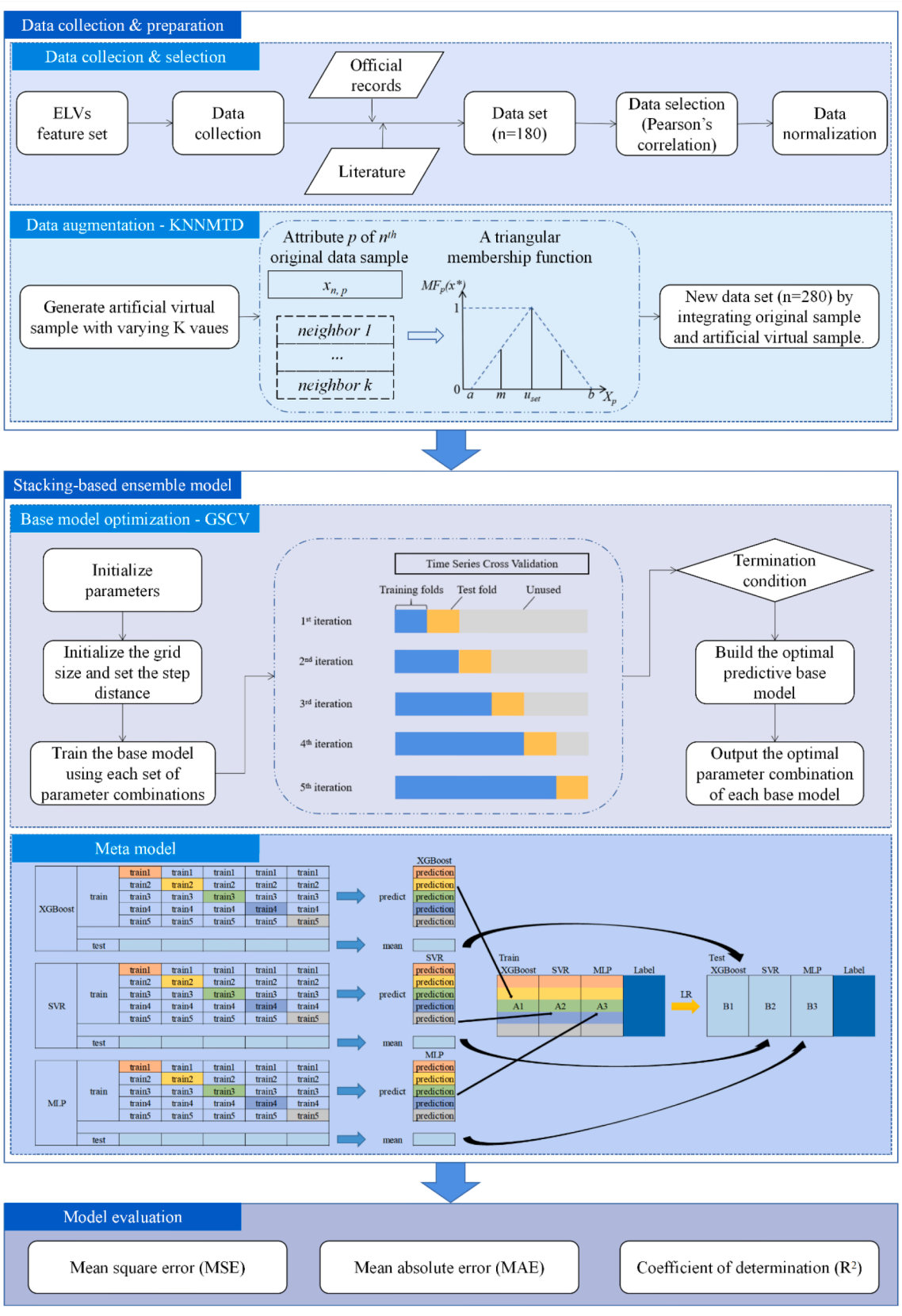

Fig. 1. The overall framework of the empirical study.

Table 1
Prediction results of single ML algorithms.

Model	MAE		MSE		R^2	
	$N = 180$	$N = 280$	$N = 180$	$N = 280$	$N = 180$	$N = 280$
SVR	0.1215	0.0776	0.0181	0.0066	0.4518	0.8651
XGBoost	0.0748	0.0620	0.0092	0.0054	0.7194	0.8901
LGBM	0.0782	0.0513	0.0101	0.0051	0.6931	0.8950
RF	0.0773	0.0556	0.0110	0.0054	0.6652	0.8896
MLP	0.1137	0.0528	0.0195	0.0042	0.4103	0.9138
GBRT	0.0962	0.0446	0.0129	0.0034	0.6086	0.9294
DT	0.0520	0.0371	0.0045	0.0051	0.8631	0.8952

Table 2
Hyperparameter description and optimization.

Model	Parameter	Description	Optimization
SVR	Kernel	kernel function	'RBF'
	gamma	coefficients of the kernel function	0.1
	C	regularization parameter	1.5
	epsilon	error tolerance of insensitive loss function	0.1
XGBoost	max_depth	maximum depth of each tree	5
	learning_rate	learning rate	0.03
	min_child_weight	minimum weight sum of child nodes	10
	gamma	threshold of controller node splitting	0
	colsample_bytree	subsample ratio of columns per tree	1
	subsample	subsample ratio of each tree	1
LGBM	n_estimators	number of boosting iterations	120
	learning_rate	learning rate	0.04
	max_depth	maximum depth of each tree	5
	colsample_bytree	subsample ratio of columns per tree	1
	min_split_gain	minimum gain required to split a node	0.01
	subsample	subsample ratio of each tree	0.5
	num_leaves	maximum number of leaf nodes on a tree	10
RF	n_estimators	the number of trees in the forest	20
	max_depth	maximum depth of each tree	5
	max_features	the number of features when fitting	3
	min_samples_leaf	minimum sample number of leaf node	1
	min_samples_split	minimum sample number to split a node	3
MLP	hidden_layer_sizes	number of neurons in the hidden layer	40
	alpha	L2 penalty parameter	0.001
	learning_rate_init	initial learning rate used	0.01
GBRT	n_estimators	number of boosting iterations	40
	max_depth	maximum depth of each estimator	5
	learning_rate	learning rate	0.09
	min_samples_leaf	minimum sample number of leaf node	5
	max_features	the number of features when fitting	0.5
	subsample	subsample ratio of each base learner	0.5
DT	max_depth	maximum depth of each tree	5
	min_samples_leaf	minimum sample number of leaf node	1
	min_samples_split	minimum sample number of leaf node	3

reverse supply chain. The data related to China's ELVs industry from 2006 to 2020 is used to validate the proposed optimized Stacking model. Thus, the main research results are analyzed as follows.

First, the performance evaluation of seven ML models, namely SVR, XGBoost, LGBM, RF, MLP, GBRT, and DT, shows that they achieved good results in predicting the quantity of EOL products after using the KNNMTD method and optimizing parameters through grid search and time series cross validation (see Table 1). Note that GBRT and MLP perform better, which indicates that the relationship between the quantity of EOL products and its socioeconomic variables tends to be complex and nonlinear.

Table 3
Prediction results of stacking models.

Stacking Model	MAE	MSE	R^2
Stacking 1 (SVR+ MLP+ XGBoost)	0.0305	0.0016	0.9515
Stacking 2 (SVR+ MLP+ LGBM)	0.0411	0.0027	0.9238
Stacking 3 (SVR+ MLP+ RF)	0.0441	0.0030	0.9024
Stacking 4 (SVR+ MLP+ DT)	0.0428	0.0028	0.8783
Stacking 5 (SVR+ MLP+ GBRT)	0.0332	0.0019	0.9300

Second, the Stacking 1 ensemble model, which uses SVR, MLP, and XGBoost as the base models and LR as the meta model, performs best in predicting the quantity of EOL products. Table 3 shows the predictive performance of the Stacking model under different base learner combinations. The MAE and MSE of the Stacking 1 model are 0.0305 and 0.0016, respectively, which are lower than the other Stacking models. Besides, the values predicted by the Stacking ensemble model overall are moving in the same direction as the real values, as demonstrated in Fig. 2. The predictions of some points with large fluctuations can also be accurately predicted.

Third, reduced error and R-squared, as seen in Tables 1 and 3, clearly advocate for the superiority of the proposed Stacking 1 model over a single base model. Compared with the worst single model SVR, the Stacking 1 model decreased MAE and MSE by about 0.0471 and 0.0050, respectively, and increased R^2 by 0.0864. This indicates that the proposed Stacking model integrates the strengths of single ML models to capture information, reducing the influence of a variable environment and multiple operating conditions and improving the overall prediction accuracy and generalization ability. Note that even with the introduction of SVR with slightly lower precision, the performance of the Stacking 1 model remains superior to other base models. There are three main reasons for this. Firstly, SVR has unique advantages in handling the regression problems with high dimensions and small samples. Secondly, XGBoost, as a single model, exhibits strong prediction performance, ensuring the prediction accuracy of the Stacking model. Finally, using algorithms with low correlation as the base models allows the Stacking 1 model to fully utilized the strengths of each algorithm, reducing the risk of falling into the local optimal, and providing robust prediction performance.

Fourth, for a small sample data set, data augmentation can help make more robust and accurate predictions. The learning curves in Fig. 3 demonstrate that the error value of the testing data set is higher than that of the training data set. However, the learning curve of the testing data set is far from that of the training data set, suggesting that the Stacking 1 model is slightly overfitting based on the original data set (see Fig. 3(a)). To address the issue, we use the KNNMTD method to generate artificial virtual data and integrate it with the original dataset to create a new expanded dataset for model training. After training on the expanded new data set, the training error and the testing error tend to converge and approach each other, indicating that the generalization ability of the ensemble learning prediction model is improved by the KNNMTD method, as shown in Fig. 3(b). This means data augmentation can effectively reduce the overfitting risk caused by small sample datasets.

In theory, this study contributes to prediction techniques for small sample data sets. This research proposes a novel Stacking ensemble model for predicting the quantity of EOL products, which uses SVR, MLP, and XGBoost as base models and LR as a meta model. This research conducts data augmentation using the KNNMTD method to avoid the overfitting caused by the small sample data set and improve the generalization ability of the proposed model.

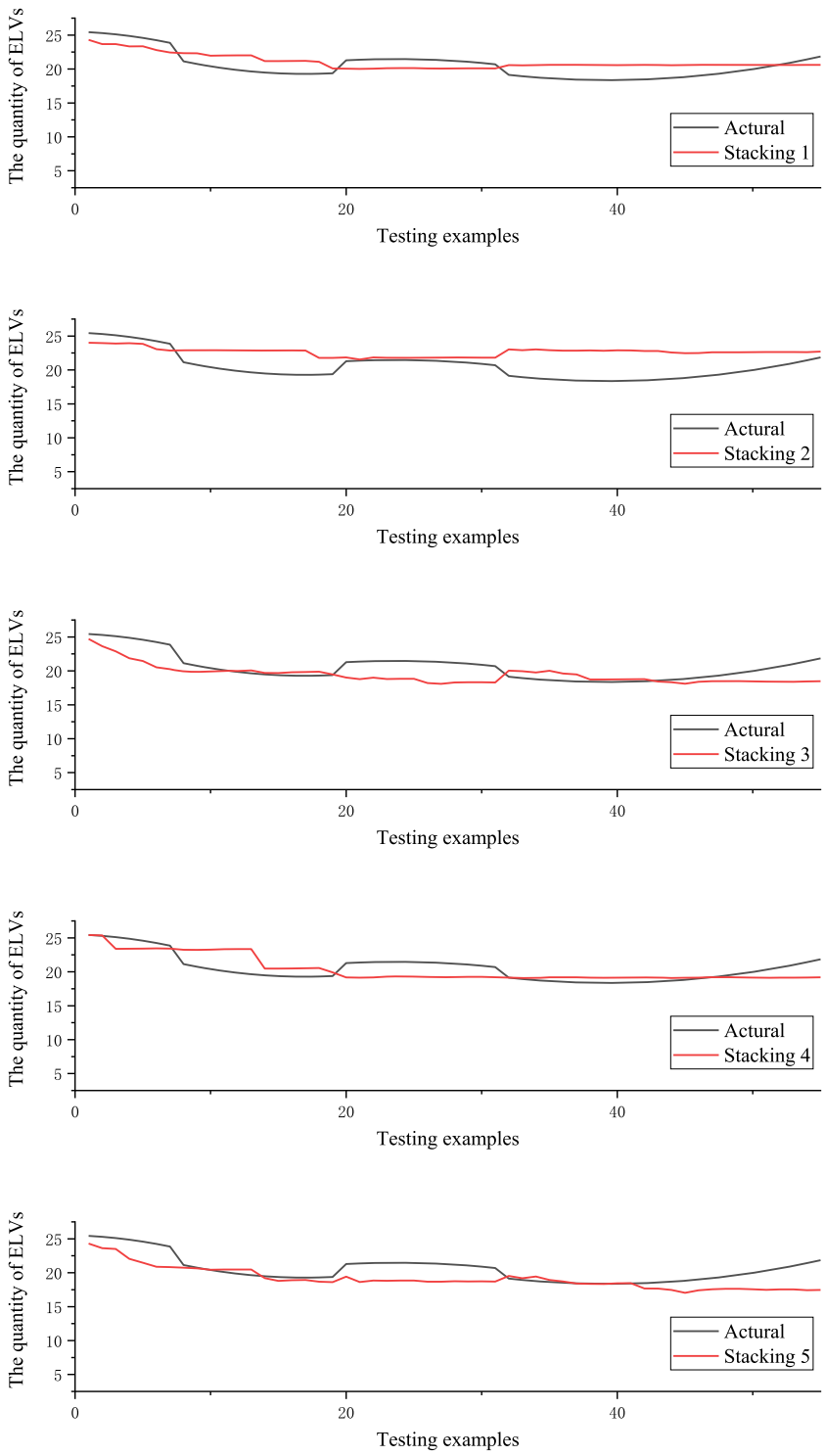

Fig. 2. The predicted values of stacking ensemble models.

Likewise, this research makes contributions to prediction applications for industry. Accurate predictions of recycling quantity can help decision-makers and practitioners design sustainable reverse supply chain and production plans, which can reduce environmental pollution and improve the competitiveness of enterprises. More specifically, the proposed Stacking ensemble model can achieve greater prediction accuracy and generalization ability than single ML models, making it a valuable tool for decision-makers in the recycling industry. When dealing with small sample data sets, the application of the KNNMTD method is particularly significant. Overall, the research findings have substantial practical implications for the recycling industry's management and sustainability.

In addition, the limitation of this research is that it is not efficient to try different base model combinations manually. In the future research direction, we will focus on the following aspects. Firstly, we plan to conduct more feature engineering on the input variables to generate problem-specific features. Secondly, we aim to establish a base model learning library and integrate it with the intelligent optimization algorithm to build a more intelligent prediction system. Thirdly, we will compare the performance of the proposed Stacking ensemble model

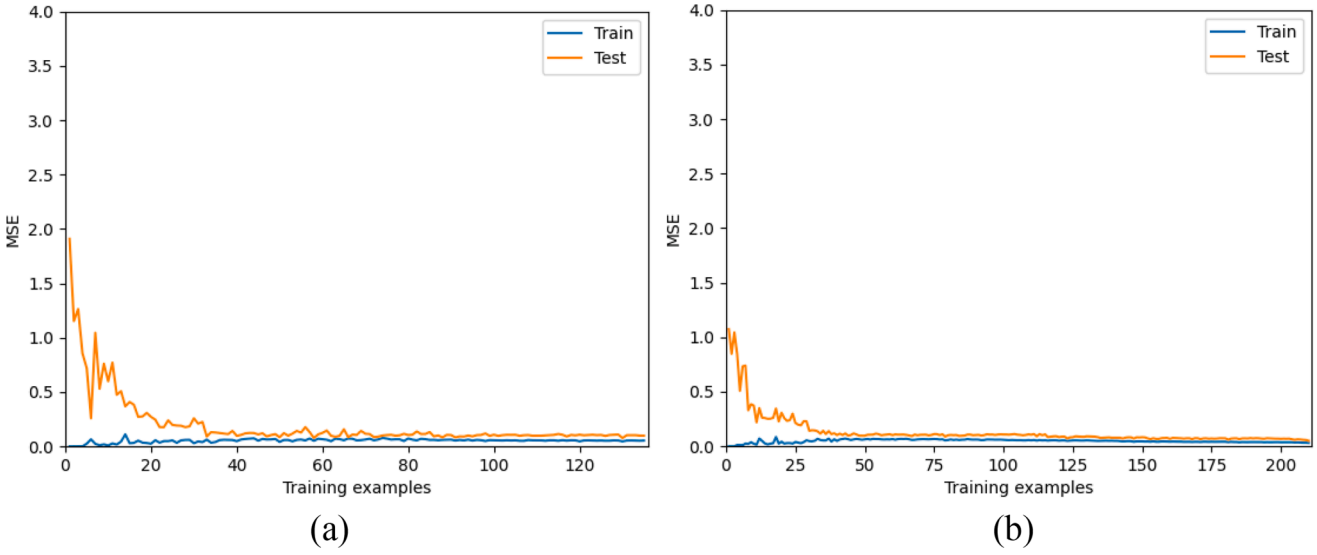

Fig. 3. Learning curve without (a) and with (b) data augmentation.

with other ensemble models. Fourthly, we intend to incorporate factors such as policy and consumer preferences into the prediction model as they influence the quantity of EOL products. Lastly, if the data set becomes larger in the future, we may adopt distributed computing methods to improve calculation speed.

6. Conclusions

The accurate prediction of the quantity of recycled EOL products is an important prerequisite for making effective short-term and long-term decisions and monitoring the overspill of hazardous waste. To improve the prediction accuracy and generalization ability, we propose a Stacking-based ensemble prediction model for the quantity of recycled EOL products. In the process of data preprocessing, data augmentation is used to avoid the overfitting problem caused by small sample data sets. In the process of model training, based on the correlation and prediction abilities of the base models, a combination strategy of base models is proposed. In addition, the proposed model is validated with relevant data from China's ELVs industry. The results indicate that, compared with other Stacking ensemble models and single ML models, the Stacking 1 model proposed in this research has better performance in prediction accuracy and stability.

The application of the proposed Stacking-based model can be expanded to a global scope to examine the EOL product generation trends in different countries and regions. From a sustainability point of view, this research can be used by practitioners and decision-makers as the basis for the development of recycling programs, the construction of processing facilities, the optimization of resource allocation, as well as the establishment of waste management systems and sustainable reverse supply chains. Consequently, this research not only provides a new direction for predicting EOL product recycling but also adds economic, technical, and social benefits to sustainable environmental conservation and the circular economy.

CRediT authorship contribution statement

Hanbing Xia: Conceptualization, Methodology, Data curation, Validation, Writing – original draft, Writing – review & editing. Ji Han: Visualization, Writing – original draft, Writing – review & editing. Jelena Milisavljevic-Syed: Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Abbasi, M., Abduli, M., Omidvar, B., Baghvand, A., 2014. Results uncertainty of support vector machine and hybrid of wavelet transform-support vector machine models for solid waste generation forecasting. Environ. Prog. Sustain. Energy 33 (1), 220–228. https://doi.org/10.1002/ep.11747.

Abbasi, M., El Hanandeh, A., 2016. Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste Manage. 56, 13–22. https://doi.org/10.1016/j.wasman.2016.05.018.

Adamović, V.M., Antanasijević, D.Z., Ristić, M.Đ., Perić-Grujić, A.A., Pocajt, V.V., 2017. Prediction of municipal solid waste generation using artificial neural network approach enhanced by structural break analysis. Environ. Sci. Pollut. Res. 24 (1), 299–311. https://doi.org/10.1007/s11356-016-7767-x.

Agrawal, S., Singh, R.K., 2019. Forecasting product returns and reverse logistics performance: structural equation modelling. Manage. Environ Qual. Int. J. https://doi.org/10.1108/meq-05-2019-0109.

Andeobu, L., Wibowo, S., Grandhi, S., 2021. A systematic review of e-waste generation and environmental management of Asia Pacific countries. Int. J. Environ. Res. Public Health 18 (17), 9051. https://doi.org/10.3390/ijerph18179051.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Coskuner, G., Jassim, M.S., Zontul, M., Karateke, S., 2021. Application of artificial intelligence neural network modeling to predict the generation of domestic, commercial and construction wastes. Waste Manage. Res. 39 (3), 499–507. https://doi.org/10.1177/0734242x20935181.

Cui, S., Yin, Y., Wang, D., Li, Z., Wang, Y., 2021. A stacking-based ensemble learning method for earthquake casualty prediction. Appl. Soft Comput. 101, 107038 https://doi.org/10.1016/j.asoc.2020.107038.

Dai, F., Nie, G.H., Chen, Y., 2020. The municipal solid waste generation distribution prediction system based on FIG–GA-SVR model. J. Mater. Cycles Waste Manage. 22 (5), 1352–1369. https://doi.org/10.1007/s10163-020-01022-5.

Dasarathy, B.V., Sheela, B.V., 1979. A composite classifier system design: concepts and methodology. Proc. IEEE 67 (5), 708–713. https://doi.org/10.1109/PROC.1979.11321.

de Souza, J.A.F., Silva, M.M., Rodrigues, S.G., Santos, S.M., 2022. A forecasting model based on ARIMA and artificial neural networks for end–OF–life vehicles. J. Environ. Manage. 318, 115616 https://doi.org/10.1016/j.jenvman.2022.115616.

Elshaboury, N., Mohammed Abdelkader, E., Al-Sakkaf, A., Alfalah, G., 2021. Predictive analysis of municipal solid waste generation using an optimized neural network model. Processes 9 (11), 2045. https://doi.org/10.3390/pr9112045.

Ene, S., Öztürk, N., 2017. Grey modelling based forecasting system for return flow of end-of-life vehicles. Technol. Forecast. Soc. Change 115, 155–166. https://doi.org/10.1016/j.techfore.2016.09.030.
Erdebilli, B., Devrim-İçtenbaş, B., 2022. Ensemble voting regression based on machine learning for predicting medical waste: a Case from Turkey. Mathematics 10 (14), 2466. https://doi.org/10.3390/math10142466.
Erkinay Ozdemir, M., Ali, Z., Subeshan, B., Asmatulu, E., 2021. Applying machine learning approach in recycling. J. Mater. Cycles Waste Manage. 23 (3), 855–871. https://doi.org/10.1007/s10163-021-01182-y.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comp. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.
Golbaz, S., Nabizadeh, R., Sajadi, H.S., 2019. Comparative study of predicting hospital solid waste generation using multiple linear regression and artificial intelligence. J. Environ. Health Sci. Eng. 17 (1), 41–51. https://doi.org/10.1007/s40201-018-00324-z.
Hao, H., Zhang, J., Zhang, Q., Yao, L., Sun, Y., 2021. Improved gray neural network model for healthcare waste recycling forecasting. J. Comb. Optim. 42 (4), 813–830. https://doi.org/10.1007/s10878-019-00482-2.
Hao, H., Zhang, Q., Wang, Z., Zhang, J., 2018. Forecasting the number of end-of-life vehicles using a hybrid model based on grey model and artificial neural network. J. Clean. Prod. 202, 684–696. https://doi.org/10.1016/j.jclepro.2018.08.176.
Hu, S., 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), 154–170. https://doi.org/10.1007/s10163-012-0102-9.
Jain, A., Sareen, R., 2006. E-waste assessment methodology and validation in India. J. Mater. Cycles Waste Manage. 8 (1), 40–45. https://doi.org/10.1007/s10163-005-0145-2.
Kannangara, M., Dua, R., Ahmadi, L., Bensebaa, F., 2018. Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. Waste Manage. 74, 3–15. https://doi.org/10.1016/j.wasman.2017.11.057.
Kontokosta, C.E., Hong, B., Johnson, N.E., Starobin, D.J., 2018. Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Comput. Environ. Urban Syst. 70, 151–162. https://doi.org/10.1016/j.compenvurbsys.2018.03.004.
Kumar, S., Kumar, R., 2021. Forecasting of municipal solid waste generation using non-linear autoregressive (NAR) neural models. Waste Manage. 121, 206–214. https://doi.org/10.1016/j.wasman.2020.12.011.
Lenort, R., Staš, D., Wicher, P., Straka, M., 2021. State of the art in the end-of-life vehicle recycling. Rocznik Ochrona Środowiska 23.
Li, D.C., Huang, W.T., Chen, C.C., Chang, C.J., 2013. Employing virtual samples to build early high-dimensional manufacturing models. Int. J. Prod. Res. 51 (11), 3206–3224. https://doi.org/10.1080/00207543.2012.746795.
Li, K., Ma, H., 2019. Prediction of municipal solid waste generation with Elman Neural Network——case study: shanghai City in China. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, pp. 1174–1177.
Lin, H.T., Nakajima, K., Yamasue, E., Ishihara, K., 2018. Recycling of end-of-life vehicles in small islands: the Case of Kinmen, Taiwan. Sustainability 10 (12), 4377–4390. https://doi.org/10.3390/su10124377.
Liu, H.M., Sun, H.H., Guo, R., Wang, D., Yu, H., Do Rosario Alves, D., Hong, W.M., 2022. Prediction of China's industrial solid waste generation based on the PCA-NARBP model. Sustainability 14 (7), 4294. https://doi.org/10.3390/su14074294.
Lu, W., Huo, W., Gulina, H., Pan, C., 2022. Development of machine learning multi-city model for municipal solid waste generation prediction. Front. Environ. Sci. Eng. 16 (9), 1–10. https://doi.org/10.1007/s11783-022-1551-6.
Masoumi, S.M., Kazemi, N., Abdul-Rashid, S.H., 2019. Sustainable supply chain management in the automotive industry: a process-oriented review. Sustainability 11 (14), 3945. https://doi.org/10.3390/su11143945.
Meza, J.K.S., Yepes, D.O., Rodrigo-Ilarri, J., 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), e02810. https://doi.org/10.1016/j.heliyon.2019.e02810.
Namoun, A., Tufail, A., Khan, M.Y., Alrehaili, A., Syed, T.A., BenRhouma, O., 2022. Solid waste generation and disposal using machine learning approaches: a survey of solutions and challenges. Sustainability 14 (20), 13578. https://doi.org/10.3390/su142013578.
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., Nguyen, V.K., 2021. Development of machine learning-based models to forecast solid waste generation in residential areas: a case study from Vietnam. Resour. Conserv. Recycl. 167, 105381 https://doi.org/10.1016/j.resconrec.2020.105381.
Ni, D., Xiao, Z., Lim, M.K., 2021. Machine learning in recycling business: an investigation of its practicality, benefits and future trends. Soft Comput. 25 (12), 7907–7927. https://doi.org/10.1007/s00500-021-05579-7.
Numfor, S.A., Omosa, G.B., Zhang, Z., Matsubae, K., 2021. A review of challenges and opportunities for end-of-life vehicle recycling in developing countries and emerging economies: a SWOT analysis. Sustainability 13 (9), 4918. https://doi.org/10.3390/su13094918.
Ochotnicky, P., Kacer, M., 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, 452–467.
Oguz-Ekim, P., 2021. Machine learning approaches for municipal solid waste generation forecasting. Environ. Eng. Sci. 38 (6), 489–499. https://doi.org/10.1089/ees.2020.0232.
Polák, M., Drápalová, L., 2012. Estimation of end of life mobile phones generation: the case study of the Czech Republic. Waste Manage. 32 (8), 1583–1591. https://doi.org/10.1016/j.wasman.2012.03.028.
Puntarić, E., Pezo, L., Zgorelec, Ž., Gunjača, J., Kučić Grgić, D., 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), 10133. https://doi.org/10.3390/su141610133.
Rashid, F.A.A., Hishamuddin, H., Radzi, M., 2021. Supply chain optimization for end-of-life vehicle recycling: a preliminary review. In: Proceedings of the 11th annual international conference on industrial engineering and operations management. Singapore, pp. 7–11.
Sivakumar, J., Ramamurthy, K., Radhakrishnan, M., Won, D., 2022. Synthetic sampling from small datasets: a modified mega-trend diffusion approach using k-nearest neighbors. Knowl. Based Syst. 236, 107687 https://doi.org/10.1016/j.knosys.2021.107687.
Sodanil, M., Chatthong, P., 2014. Artificial neural network-based time series analysis forecasting for the amount of solid waste in Bangkok. In: Ninth International conference on digital information management (ICDIM 2014). IEEE, pp. 16–20.
Soni, U., Roy, A., Verma, A., Jain, V., 2019. Forecasting municipal solid waste generation using artificial intelligence models—A case study in India. SN Appl. Sci. 1 (2), 1–10. https://doi.org/10.1007/s42452-018-0157-x.
Tan, M., Yuan, S., Li, S., Su, Y., Li, H., He, F., 2019. Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning. IEEE Trans. Power Syst. 35 (4), 2937–2948. https://doi.org/10.1109/TPWRS.2019.2963109.
Tian, G., Zhou, M., Chu, J., Wang, B., 2013. Prediction models of the number of end-of-life vehicles in China. In: 2013 International Conference on Advanced Mechatronic Systems: September 25-27, 2013, Luoyang, China. Luoyang. IEEE, pp. 357–362.
Wang, D., Yuan, Y.A., Ben, Y., Luo, H., Guo, H., 2022. Long short-term memory neural network and improved particle swarm optimization–based modeling and scenario analysis for municipal solid waste generation in Shanghai, China. Environ. Sci. Pollut. Res. 1–19. https://doi.org/10.1007/s11356-022-20438-0.
Wang, F., Yu, L., Wu, A., 2021. Forecasting the electronic waste quantity with a decomposition-ensemble approach. Waste Manage. 120, 828–838. https://doi.org/10.1016/j.wasman.2020.11.006.
Xin, F., Ni, S., Li, H., 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, 19278–19286. https://doi.org/10.1109/access.2018.2814054.
Yano, J., Muroi, T., 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), 655–664. https://doi.org/10.1007/s10163-015-0360-4.
Zhang, C., Dong, H., Geng, Y., Liang, H., Liu, X., 2022. Machine learning based prediction for China's municipal solid waste under the shared socioeconomic pathways. J. Environ. Manage. 312, 114918 https://doi.org/10.1016/j.jenvman.2022.114918.
Zhou, L., Xie, J., Gu, X., Lin, Y., Ieromonachou, P., Zhang, X., 2016. Forecasting return of used products for remanufacturing using Graphical Evaluation and Review Technique (GERT). Int. J. Prod. Econ. 181, 315–324. https://doi.org/10.1016/j.ijpe.2016.04.016.