Journal of Environmental Management 360 (2024) 121152

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Decision tree-based approach to extrapolate life cycle inventory data of manufacturing processes

Mohamed Saad, Yingzhong Zhang *, Jia Jia, Jinghai Tian

School of Mechanical Engineering, Dalian University of Technology, Dalian, 116024, China

ARTICLE INFO

Handling Editor: Lixiao Zhang

Keywords:
Life cycle assessment
Manufacturing process
LCI data gaps
Decision tree models
Greenhouse gas emission

ABSTRACT

Life cycle assessment (LCA) plays a crucial role in green manufacturing to uncover the critical aspects for alleviating the environmental burdens due to manufacturing processes. However, the scarcity of life cycle inventory (LCI) data for the manufacturing processes is a considerable challenge. This paper proposes a novel approach to extrapolate LCI data of manufacturing processes. Taking advantage of LCI data in the Ecoinvent datasets, decision tree-based supervised machine learning models, namely decision tree, random forest, gradient boosting, and adaptive boosting, have been developed to extrapolate the data of GHG emissions, i.e., carbon dioxide, nitrous oxide, methane, and water vapor. Initially, a correlation analysis was conducted to derive the most influential factors on GHG quantities resulting from manufacturing activities. First, the collected data have been preprocessed and split into train and test sets (70% and 30%, respectively). Second, a five-fold cross-validation method was applied to tune the hyperparameters of the models. Then, the models were re-trained using the best hyperparameters and evaluated using the test set. The results reveal that the Gradient Boosting model has a superior predictive performance for extrapolating the GHG emission data, with average coefficients of determination (R^2) on the test set <0.95. Moreover, the model predictions involve relatively low values of the average root mean squared error and an average mean percentage of error on the test set. The correlation and feature importance analyses emphasized that the workpiece material and manufacturing technology have a considerable effect on natural resource consumption, i.e., energy, material, and water inflows into the process. Meanwhile, energy consumption, water usage, and raw aluminum depletion were the most influential factors in GHG emissions. Eventually, a case study to extrapolate the inflows and the outflows for new manufacturing activities has been conducted using the validated models. The proposed GraBoost model provides a computational supplementary approach to estimate and extrapolate the GHG emissions for different manufacturing processes when LCI data are incomplete or don't exist within LCI databases.

1. Introduction

The US National Academies of Sciences, Engineering, and Medicine emphasized that the life cycle assessment (LCA) is a crucial tool to address the grand environmental issues in the 21st century (Environmental Engineering for the 21st Century: Addressing Grand Challenges, 2019). Consequently, LCA studies are essential for evaluating the environmental impacts of products and processes throughout their entire life cycle. LCA is a data-driven study; an LCA model of a product system (e.g., manufacturing system) is a sequential collection of interrelated unit processes represented by background and foreground data. Foreground data are a gate-to-gate inventory of a specific unit process incorporating the intermediate flows (flows of materials/energy between unit processes) and the elementary flows (natural resources from the environment and emission/waste released to the environment), normally collected by LCA practitioners. Unit processes would seldom represent the production of products that are used as intermediate products for a subsequent process. Therefore, unit processes are merged to generate value chains to evaluate the environmental impacts of the product system by aggregating the inventories for all unit processes within the product system. The result is a cradle-to-gate inventory containing only elementary flows of the product system, where all intermediate flows are traced back to the resource extraction. This aggregated life cycle inventory (LCI) represents the background data normally provided by the LCI databases as system processes. It is estimated that the background system usually covers up to 99 % of the unit processes in the

* Corresponding author.
E-mail address: zhangyz@dlut.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.jenvman.2024.121152
Received 2 March 2024; Received in revised form 25 April 2024; Accepted 10 May 2024
Available online 17 May 2024

product system (Wernet et al., 2016), and the life cycle impact assessment (LCIA) also relies on the cumulative inventory results. In conducting an LCA analysis of a product, LCA practitioners mainly focus on the activities in the foreground system of the analyzing product, and the environmental LCI data (or cumulative LCI data) generated by the foreground activity is usually provided by the background database.

Although LCI databases that can provide background data are constantly developing and a number of LCI databases have been developed, such as Ecoinvent, USLCI, and Quebec LCI database, the number of activities provided by background databases is still limited, a large number of production activities lacks the support of LCI background data, especially with the development of technology, some emerging production or service activities continue to emerge. As a result, inventory data gaps have always been a major issue affecting the effective implementation of LCA studies. On the one hand, the improvement of the background database requires a certain amount of time to carry out the LCI data collection of unit processes and cumulative calculations. On the other hand, in actual production activities, there are various types of activities, workpiece materials, production methods, etc., making it impossible to establish the LCI data support for each type of activity with different properties.

Due to the importance of the issue of LCI data gaps and the difficulty of data collection, the above issues have attracted a large amount of research. Researchers have been attempting to explore a computational solution to provide fast and cost-effective ways to bridge the LCI data gaps. Canals et al. (2011) introduced proxy and extrapolation approaches as alternative methods to overcome the issue of LCI data unavailability. On the one hand, proxy approaches either use direct substitutes of existing data sources for the target product with no changes, average data for a group of products that are assumed to be similar to the target product, or scale the existing data to fill the data gaps on the basis of quantities, production conditions, the composition of the product, etc. On the other hand, data extrapolation generates new data through machine learning (ML) techniques to bridge the data gaps by adapting the data source outside the range of their original validity to reflect the target scenarios appropriately. Meron et al. (2020) proposed a selection proxy methodology to select the most appropriate LCI dataset for a specific site from the available LCI background datasets. The developed methodology was applied to the LCA of coal-fired power stations. Zargar et al. (2022) reviewed the most common modeling approaches for LCI data estimation, including the major strategies to address data gap issues such as proxy selection and data-driven-based LCI generation. In the data-driven approaches, multiple endeavors paid attention to machine-learning techniques to estimate the LCI data and the environmental impacts for a wide range of products in different industries. For instance, Sun et al. (2023) developed ANN models for predicting the life cycle environmental impacts of chemicals. Zhu et al. (2020) developed an ANN model to predict the life cycle impacts of the sitagliptin production process and select green chemical substitutes based on their molecular descriptors to optimize the pharmaceutical manufacturing process. Lee et al. (2020) developed a boosted regression tree (BRT) model to estimate the future impacts of global warming and eutrophication for corn production in US Midwest counties for the years 2022–2100. Nevertheless, these studies were centered on estimating the LCI data and assessing the environmental impacts of particular products and industries, which required deep domain knowledge, involved a remarkable degree of uncertainty, and the validation of decisions was frequently limited.

Unit process data are the foundation of the cumulative LCI data and LCA results. In recent years, with the data science development, a few computational approaches on LCI data acquisition and completion for unit processes have been developed. Cashman et al. (2016) proposed to mine available data from the U.S. Environmental Protection Agency to support rapid LCI modeling of chemical manufacturing. Meng et al. (2019) proposed a data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology. Hou et al. (2018) proposed a similarity-based link prediction approach to estimating the LCI data of unit processes in Ecoinvent datasets. A reliable data estimation was possible when less than 5% of flows were missing in one unit process. Based on their works, Zhao et al. (2021) used a decision tree-based supervised learning technique (eXtreme Gradient Boosting) to estimate missing LCI data of unit processes. This approach enhanced the ability to predict flows when less than 20% of data is missing in a single unit process. Nevertheless, in the product system LCA, to obtain the LCI background data (cumulative LCI data, i.e., the cradle-to-gate elementary flow data) of an activity needs to combine with its upstream process's data and conduct product allocation and linear calculations. As a result, common users can't understand upstream process data and related calculation methods fully and rely only on background data. As mentioned above, the number of unit processes defined in the background database is limited, and the lack of unit process datasets often occurs. For example, in the latest Ecoinvent background dataset, there is a lack of grinding activity data. If a data-driven approach can extrapolate LCI data for a new process from similar processes in the existing datasets, it would be very helpful in conducting LCA studies effectively.

To our knowledge, there are currently no literature reports on extrapolating LCI data methods for similar processes from LCI cumulative datasets. In this study, we find that current LCI background databases commonly provide a very limited number of manufacturing processes with LCI data, far from meeting the LCA analysis requirements for mechanical products. Therefore, it is necessary to research and develop an approach to estimating and extrapolating LCI data of certain activities that are not in the background database. With this in mind, we take advantage of the highly reputable and reliable LCI data of the Ecoinvent database and develop a decision tree-based LCI data extrapolation approach for mechanical product manufacturing activities. In addition, due to the large number of elementary flow types (over 2100 types) in the Ecoinvent cumulative dataset, in this study, we chose greenhouse gas emission (GHG) flows that have serious environmental effects, such as global warming and climate change, as the research objects. This approach is based on the assumption that similar activities generate similar environmental impacts and the correlation analysis between the activity properties and elementary flows. The major contributions of this paper include: (i) a novel computational approach based on decision tree models to extrapolate LCI data of manufacturing processes is proposed; (ii) carry out the LCI data extrapolation for the missing manufacturing process in the background database; and (iii) analyze the effect of various manufacturing processes on GHG emissions and examine the key factors that contribute to the release of GHGs. The remainder of the paper is structured as follows: Section 2 addresses the material and methods. Section 3 presents the results and discussion. Finally, the paper's conclusion and future works are discussed.

2. Materials and methods

2.1. Theoretical analysis of the method

This study aims to provide an effective computational way to estimate and extrapolate the LCI data for manufacturing activities to bridge the data gaps. Extrapolation generates new data by adjusting the reliable data sources of manufacturing processes to reflect better a new target manufacturing activity. To achieve the above goals, this study bases the following scientific assumptions: (i) similar activities tend to have the same types of elementary flows; (ii) similar types of flows tend to have the same value for the same kind of activities; and (iii) there are correlations between the properties of manufacturing processes and their elementary flows.

Based on the above assumptions, to conduct extrapolation calculations, there are two issues to be addressed:

2.1.1. Classification

On the one hand, classification is the foundation for similarity

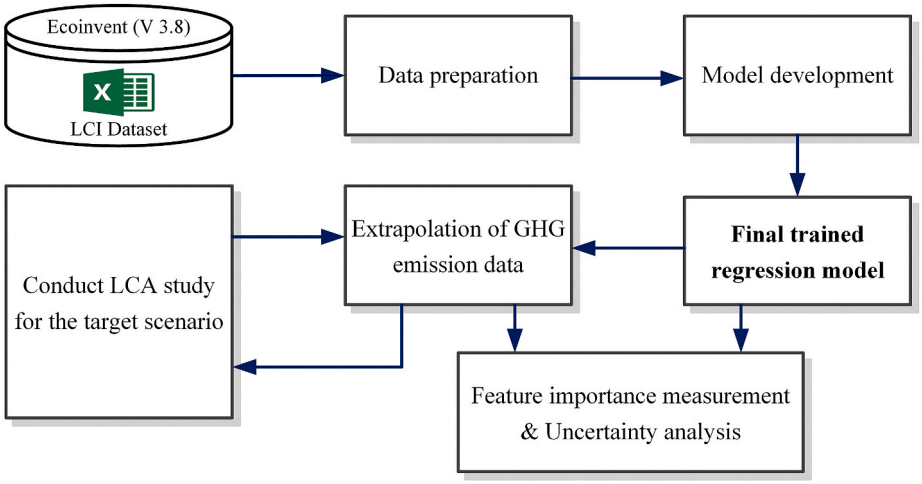

Fig. 1. Framework of the proposed approach.

comparison and processing. On the other hand, classification is a method of grouping objects with similar properties or features together. As a result, based on explicit knowledge, we can first classify manufacturing processes in the product lifecycle, which lays the foundation for discovering and clustering similar activities and LCI data.

At the same time, the manufacturing process of each classified subprocess can continue to be classified based on its properties, such as the processing method, workpiece material, control method, etc. The manufacturing processes with different properties produce different environmental impacts.

2.1.2. Regression

Regression is a mathematical method for quantitatively describing statistical relationships in data, which involves studying the relationship between the dependent variable (target) and the independent variables (predictors) to quantitatively calculate and predict numerical data. As mentioned above, extrapolation generates new data by adjusting the existing data to reflect better a new target manufacturing activity. As a result, we employ the regression computation method to solve the extrapolation of LCI numerical value. In fact, there are strong correlations between the properties of manufacturing processes and their LCI data, but most of them are complex nonlinear relationships.

Decision tree is a ML methodology that combines classification and regression. As a result, we propose to use decision tree-based models to model the relationships between the properties and the released emissions of manufacturing processes.

2.2. Methodological framework

According to above analysis and the requirement of the task, we proposed a methodological framework as shown in Fig. 1, which mainly includes the following four parts:

2.2.1. Data preparation

Regression model needs sample data to be trained. The data preparation work mainly includes data collection, data analysis, variable selections, and data preprocessing suitable for model training.

2.2.2. Model development

In the model development process, we selected four decision tree-based ML models, namely decision tree (DT), random forest (RF), gradient boosting (GraBoost), and adaptive boosting (AdaBoost), for modeling the relationships between the independent variables and the dependent variable as a regression problem.

The model development work includes model training based on the data provided by the first part work, model validation, and regression performance evaluation. The model with the best regression performance is selected as the regression extrapolating model.

2.2.3. Extrapolation of the GHG emission data for manufacturing processes

Based on the trained regression model, we need to implement two steps to conduct extrapolation of the GHG emission data for manufacturing processes. One step is to input property data of a new manufacturing process into the trained regression model, and a set of intermediate elementary flow data can be obtained. The other step is to

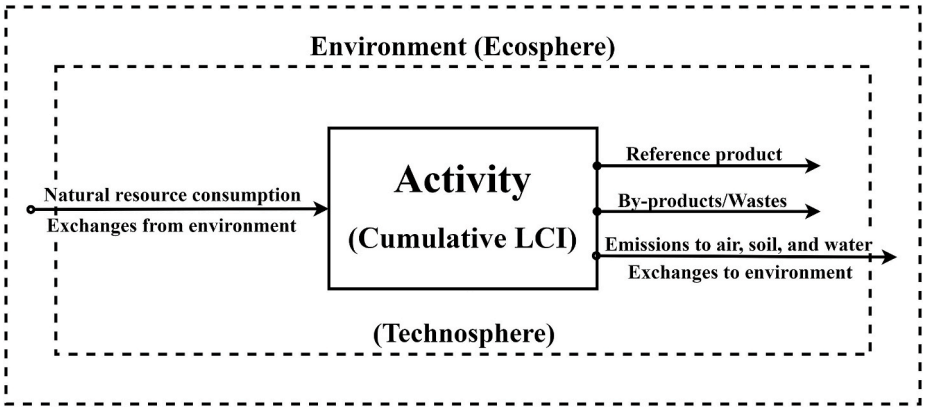

Fig. 2. Modeling of an LCI activity.

Table 1
LCI data structure in the Ecoinvent database (Wernet et al., 2016).

No.	Activity Name	Geo.	Reference Product	Functional unit	Elementary Flow		
					Compartment		
					Natural Resource	…	Air
					Water	…	Carbon dioxide
1	aluminum drilling, conventional	RER	drilled aluminum, conventional	1 Kg	0.0957573	…	7.03815
⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
315	impact extrusion of aluminum, 1 stroke	RoW	impact extrusion of aluminum, 1 stroke	1 Kg	0.007249	…	0.53895

input property data of a new manufacturing process and the obtained intermediate elementary flow data into another trained regression model, and a set of GHG emission data can be obtained.

2.2.4. Feature importance measurement and uncertainty analysis

The GHG-predicted values of the trained models enclose a degree of uncertainty due to the presence of data noise and model deviation. It is necessary to perform feature importance measurement and uncertainty analysis.

In the development of this extrapolation model, all classification and regression tasks have been implemented using Python programming language supporting the libraries scikit-learn, Pandas, Numpy, and Matplotlib on DELL workstation Intel(R) Core(TM) i7-12700H 2.30 GHz, 16 GB RAM.

2.3. Data preparation

2.3.1. LCI data structure

As analyzed above, this study focuses on bridging the data gaps of manufacturing activities in the background database, which stores cumulative LCI data. In the cumulative LCI data, there are no intermediate flows; all elementary flows are aggregated to the same type of elementary flows from different unit processes where all intermediate flows are traced back to the resource extraction. Fig. 2 presents LCI data modeling for an activity.

2.3.2. LCI dataset of the ecoinvent database

This work takes advantage of the Ecoinvent LCI database (v3.8), the most widely used; it's a repository for LCI data for different activities in multiple disciplines, e.g., industrial, agriculture, transportation, etc. As mentioned above, LCI data is a cumulative cradle-to-gate inventory. Every activity is represented as a row, and columns represent the values of the elementary flows of the activity, as shown in Table 1. In particular, a column is a record of the numerical value of a specific type of elementary flow associated with the unitary output (functional unit) of a specific activity (row), e.g., 7.03815 kg CO_2 emission to non-urban air due to the conventional drilling for 1 Kg of aluminum. 1 kg functional unit as a unitary output of all manufacturing processes puts all the activities in a comparative context.

From Table 1, we can see that the activity name contains rich semantics on the property of an activity. We can extract the properties of an activity from its name, such as the process type, workpiece material, process technology, control method, etc.

In this study, a total of 315 activities for mechanical manufacturing processes in the Ecoinvent dataset were gathered, incorporating traditional machining activities (turning, milling, and drilling), laser machining, deep drawing, impact extrusion, hot rolling, steel production, and aluminum production. Owing to the enormous number of elementary flows related to each manufacturing process, a correlation analysis has been conducted to investigate the most influential factors on GHGs quantities. Accordingly, based on the literature (Li et al., 2023; Malyan et al., 2022; Qin and Gong, 2022) and analysis for properties of manufacturing processes, an ensemble of nine quantitative and qualitative factors have been selected as the independent variables as shown in Table 2, the most influential factors on the quantities of GHGs, i.e., carbon dioxide (CO_2), nitrous oxide (NOx), methane (CH_4), and water vapor (WV). Table 3 reports the descriptive statistics for all variables in the datasets from the gathered dataset.

Table 2
The selected independent variables.

No.	Variable name	Meaning of variables	Unit
1	*EC*	Energy consumption	MJ
2	*WU*	Water usage	Kg
3	*AlD*	Aluminum depletion	Kg
4	*CuD*	Copper depletion	Kg
5	*FeD*	Iron depletion	Kg
6	*Geo*	Geography of the manufacturing activity	cateogry
7	*MP*	Manufacturing process type	cateogry
8	*WpM*	Workpiece material	cateogry
9	*PT*	Manufacturing process technology	cateogry

2.3.3. Data preprocessing and analysis

Initially, the predefined categorical variables for the manufacturing process, workpiece material, process technology, and geography need to be converted into continuous variables using the one-hot encoding method in Python, maintaining the original information of the data. Subsequently, the dataset underwent preprocessing procedures, including normalization and feature scaling, to guarantee its suitability for analysis. The variables were rescaled and standardized using log transformation to mitigate the skewed distributions of variables. Log transformation is commonly used to reduce skewness and improve the distributional properties of variables, making it more suitable for analysis and interpretation (West, 2022).

A correlation measurement analysis using the Pearson product-moment correlation (PPMC) method was used to measure the linear relationship between each pair of variables separately, where a correlation coefficient (ρ) indicates the correlation measure between every two variables.

2.4. Model development

In this study, we formularize the extrapolation of GHG emissions of manufacturing processes as regression tasks and figure out the relation between the independent variables and the dependent variables accurately using decision tree-based ML models.

In this study, we used four decision tree-based ML models, namely decision tree (DT), random forest (RF), gradient boosting (GraBoost), and adaptive boosting (AdaBoost), for modeling the relationships between the independent variables and the dependent variable as a regression problem. The model with the best regression performance is selected as the extrapolating model. The following briefly introduces the four models.

2.4.1. Decision tree model

Initially, the decision tree (DT) model generates a root node of all the training data $\{(x_i, y_i)\}_{i=1}^{p}$. Then, the classification algorithm splits the data into child nodes (N) based on mitigating the disparities (minimizing

Table 3
Sample distribution and descriptive statistics.

No.	Variable	Mean	Median	St.Dev	Maximum	Minimum	Type
1	*EC*	26.292	8.910	51.364	310.978	0.0003	Continuous input
2	*WU*	0.650	0.260	1.051	5.801	0.0003	Continuous input
3	*AlD*	0.168	0.009	0.362	1.295	0.0001	Continuous input
4	*CuD*	0.040	0.003	0.130	0.562	0.0001	Continuous input
5	*FeD*	**0.320**	**0.254**	**0.323**	**1.031**	**0.0001**	Continuous input
6	*Geo*	N/A	N/A	N/A	N/A	N/A	Categorical input
7	*MP*	N/A	N/A	N/A	N/A	N/A	Categorical input
8	*WpM*	N/A	N/A	N/A	N/A	N/A	Categorical input
9	*PT*	N/A	N/A	N/A	N/A	N/A	Categorical input
10	*CO2*	**7.280**	**3.601**	**9.968**	**56.255**	**0.008**	Output
11	*NOx*	**0.020**	**0.009**	**0.026**	**0.142**	**0.0001**	Output
12	*CH4*	**0.024**	**0.015**	**0.030**	**0.179**	**0.0001**	Output
13	WV	**0.190**	**0.055**	**0.395**	**2.496**	**0.0001**	Output

the deviations) in terms of the mean squared error (MSE). The splitting occurs over the predictors, $X_1, X_2, \ldots, X_p$ of all independent variables and possible cut points, $s_1, s_2, \ldots, s_N$, with values for the X_i feature either below or above a threshold s_j, until the predictions of the training subsets meet the optimal combination by maximizing the reduction of the MSE relative to the current node (Myles et al., 2004).

$$MSE_t = \frac{1}{N_t}\sum_{i\in D_t}(y_i - \hat{y}_t)^2 \quad (1)$$

This splitting procedure is an iterative process for each child node to split the data recursively until each node reaches the minimum node size (i.e., the number of training samples of the node) and becomes a terminal node (leaf). The traversal continues to a leaf node containing the output predicted value $\hat{y}_t$, representing the average of estimated values of the internal nodes (Vanneschi and Silva, 2023).

$$\hat{y}_{D_t} = \frac{1}{N_t}\sum_{i\in D_t} y_i \quad (2)$$

where MSE is the weighted mean square error of node i, N is the number of nodes in a decision tree D_t, y_i is the output value of a node i, and $\hat{y}_i$ is the average output of split nodes, and $\hat{y}_{D_t}$ is the final output prediction of the decision tree.

2.4.2. Random forest model

Random forest (RF) model is an ensemble learning algorithm that generates an ensemble of decision trees as weak learners and then applies the bagging technique (bootstrap and aggregation) to create multiple training datasets via the random resampling from the original training data $\{(x_i, y_i)\}_{i=1}^{p}$, with replacement K times (number of trees). Random bootstrap samples for random variables *sep* (p is the number of variables) are generated as base nodes to grow unbiased random trees T_k to eliminate the variance. Then, each node is split based on the best variable/split-point among the s, recursively repeating this procedure for each node until it reaches the minimum node size and becomes a leaf node. Each tree T_b is trained on X_b, and Y_b to fit the bootstrap samples repeatedly, and the final prediction is the average of all the predictions from all the individual trees (Biau, 2012).

$$\hat{y}_{rf}^{B}(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(x) \quad (3)$$

2.4.3. Gradient boosting model

Gradient boosting (GraBoost) model is also an ensemble learning algorithm based on a boosting technique that integrates multiple weak learners to build a strong learner. The purpose is to minimize the loss function $L(y, F(x))$ of the prediction. Initially, starting with a constant function $F_0(x)$, to minimize the loss function on the training set $\{(x_i, y_i)\}_{i=1}^{p}$, and incrementally add a weighted sum of M functions

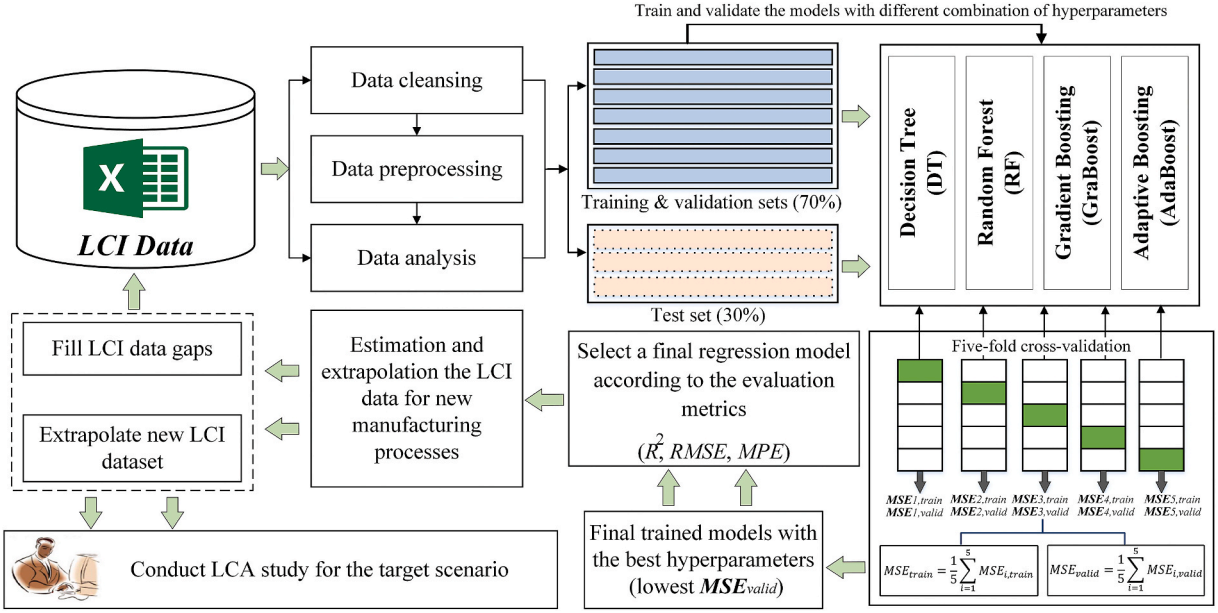

Fig. 3. Model training processes.

Table 4
The hyperparameters to be tuned for each model via grid search of cross-validation.

Model	Key hyperparameters tuned via cross-validation	hyperparameter range for grid search cross-validation
Decision tree	maximum depth of the tree	None, 1, 2, 3, 4, 5, 6, 7
	Min. leaf nodes	5, 10, 20, 30, 40, 50, 60
	Min. samples splits	2, 3, 4, 5, 6, 7
Random Forest	n_etimators (number of trees)	100, 300, 500, 700, 900, 1200, 1500, 1800
	maximum number of features	None, 1, 3, 4, 5, 6, 7, 8
Gradient Boosting	n_etimators (number of trees)	100, 300, 500, 700, 900, 1200, 1500, 1800
	Learning rate	0.05, 0.1, 0.2, 0.3, 0.4, 0.5
Adaptive Boosting	n_etimators (number of trees)	50, 100, 300, 500, 700, 900, 1200, 1800
	Learning rate	1.0, 1.5, 2.0, 2.5, 3.0

$h_m(x)\epsilon\mathscr{H}$ of weak learners.

For $m = 1,2,\ldots,M$, where $h_m\epsilon\mathscr{H}$ is a base learner function:

$$F_m(x) = F_{m-1}(x) - \gamma\sum_{i=1}^{p}\nabla_{F_{m-1}}L(y_i,F_{m-1}(x_i)) \tag{4}$$

Where $\gamma > 0$. For small γ, this implies that $L(y_i,F_m(x_i)) \le L(y_i,F_{m-1}(x_i))$.

The gradient descent is applied to find a local minimum of the loss function L by iterating on $F_{m-1}(x)$ at every step. Therefore, the final prediction of the model is the total of the ensemble with minimum error (Natekin and Knoll, 2013).

2.4.4. Adaptive boosting model

Adaptive boosting (AdaBoost) is another ensemble learning algorithm that builds a strong learner by iteratively adding weak learners. Initially, a weight $w_i = 1$ is assigned to each sample $i = 1,\ldots,N_1$ of the training set $\{(x_i,y_i)\}_{i=1}^{p}$ equal to the sampling error, which is used subsequently for training the new learner (Dietterich, 2000; Natekin and Knoll, 2013). The probability of training sample i in the training set is $w_i/\sum w_i$, and a regression machine m is constructed from that training set, where each machine makes a hypothesis: $h_m : x \to y$. Every member of the training set passes through the machine to obtain a prediction $\hat{\boldsymbol{y}}_i(\boldsymbol{x}_i)$ for $i = 1,\ldots,p$. The weights are updated, as it builds a new learner repeatedly with fine-tuned weights that give priority to data with a high error in prior iterations. Finally, for a particular input $\boldsymbol{x}_i$, each m machine has a prediction h_m, and the final prediction h_f is cumulative using the predictors of $m = 1,2,\ldots,M$.:

$$h_f(x) = inf\left\{y\epsilon Y : \sum_{m:h_t\le y}\log(1/\beta_m) \ge \sum_{m}\frac{1}{2}\log(1/\beta_m)\right\} \tag{5}$$

Where $L\epsilon[0,1]$ is the loss function, and $\beta = \frac{L}{1-L}$ is a measure of prediction confidence.

2.5. Model training

The training of regression models proceeds in the following steps, as shown in Fig. 3, as follows:

2.5.1. Data splitting

Initially, all dependent and independent variables except the categorical variables were normalized using log transformation. Then, the collected data (315 observations) have been split into a training and validation set (70% of data) and a test set (30% of data).

2.5.2. K-fold cross-validation

The Cross-Validation (Grid-SearchCV) method with five-fold was applied to the training and validation set for tuning the hyperparameters of each ML model by determining the best combination of hyperparameters. Cross-validation (CV) eliminates overfitting by averaging the performance measures over the k-folds, promoting the accurate estimation of the model. First, the data were divided into five folds randomly; four folds were used to train the model, and the last fold was used to validate the model. This procedure is repeated five times with each fold to get ten values of MSE for each set of training and validation sets, where model performance can be determined through the average MSEs of each set. The best hyperparameters are those that lead to the minimum average cross-validated MSE of the model, as shown in Fig. 3. The key hyperparameters of each model are presented in Table 4. Meanwhile, other parameters are set to the default values, e.g., the hyperparameter associated with the minimum samples in a leaf for the RF model is set to 1 as a default value to allow the trees to grow fully.

2.6. Model validation

After determining the hyperparameters, each model is re-trained on the entire training set (70% of data) for each output using the best combination of its hyperparameters. Then, the model performance is validated using the test set.

For model performance evaluation, we operate the coefficient of determination (R^2), root mean squared error (RMSE), and mean percentage error (MPE) as evaluation metrics (Steurer et al., 2021). R^2 estimates the variability proportion in the independent variable that the independent variables could explain; R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu)^2} \tag{6}$$

where n represents the number of samples of the test set, y_i is the output actual value, μ is the average output value, and $\hat{y}_i$ is the predicted value by the model.

Basically, a high R^2 value indicates a good fit for the data; however, it is not necessarily related to the prediction power of the model. Therefore, we also investigate the RMSE, which provides a comprehensive measure of the overall error of the model among predicted and actual values. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Finally, and for a concrete evaluation of the models, we also use the metric, the MPE, which enables comparison variances between differently scaled data where the output values have disparate magnitudes due to the diversity of manufacturing activities; MPE is calculated as:

$$MPE = \frac{|y_i - \hat{y}_i|}{y_i} \tag{8}$$

in order to get an unbiased evaluation of the model performance and generate the best hyperparameter combination, we repeat the previous steps ten times with ten different splits of data to calculate the average RMSE and MPE of the train and test sets. Eventually, the model with the best performance based on the evaluation criteria R^2, RMSE, and MPE been selected as a verified model to be used to estimate the GHG emissions for the manufacturing processes.

2.7. GHG emission data extrapolation

As aforementioned, the extrapolation task aims at generating new LCI data for a target scenario of machining processes when the data does not exist. This procedure refers to extrapolating the LCI data of the new manufacturing process, including the process energy/material inflows (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) and the outflows of GHG emissions. The extrapolation can be conducted in two major steps.

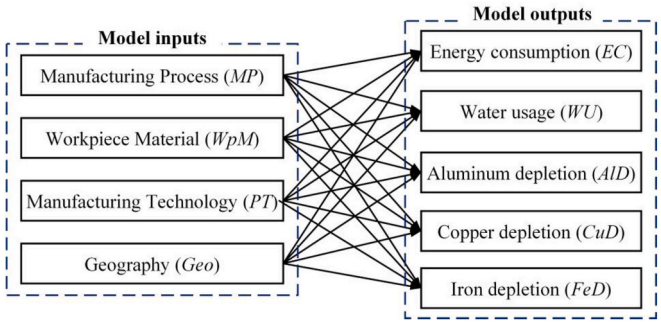

Fig. 4. Inputs and outputs for the extrapolation task of manufacturing process inflows.

2.7.1. Extrapolating elementary flow data of the new manufacturing process

Based on the property data of the new manufacturing process, i.e., the categorical variables *MP*, *WpM*, *PT*, and *Geo*, we train the ML models as mentioned in the previous section to extrapolate five elementary flow data of manufacturing processes: *EC*, *WU*, *AlD*, *CuD*, and *FeD*, as depicted in Fig. 4.

2.7.2. GHG emission data estimation

On the basis of the categorical variables (*MP*, *WpM*, *PT*, and *Geo*) of the new process and the extrapolated elementary flows (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) in the previous step. The verified model for GHG estimation can be applied to extrapolate the GHG emissions of the new activity as shown in Fig. 5.

2.8. Feature importance measurement

Intuitively, not all input variables have the same level of importance in the output predictions. The proposed ML models support the permutation feature importance (PFI), which provides a measurement of the relative importance of each independent variable on the model predictions. PFI measures the decrease in a model score when a single independent variable is randomly shuffled, breaking the relationship between the independent variable and the dependent variable. Hence, the drop in the model score reflects how much the model relies on the respective variable (Breiman, 2001). This aspect is beneficial due to the high dimensionality of LCI data, enabling determining the most influential input variables on the model output.

2.9. Uncertainty of the estimated GHG

Basically, the GHG-predicted values of the trained models enclose a degree of uncertainty due to the presence of data noise and model deviation. The bootstrap method is employed to estimate the uncertainty of the predicted values by giving a confidence interval. A confidence interval quantifies the potential range that the model predictions might fall in between when implementing the model on new data. The fundamental concept involves repeatedly resampling the initial data and using these samples to train the model respectively. This process generates several predictions, enabling the derivation of reasonably accurate confidence intervals of the predicted values. We trained the model that showed a high performance 100 times based on resampled training data to detect the 95% confidence interval of the predicted values for the GHGs. A 95% confidence interval of the extrapolated GHGs values can be expressed as:

$$CI(95\%) = \mu \pm 2\sigma \tag{9}$$

where:

CI = is the confidence interval of the model for the output predictions.

μ = is the mean value of the predictions.

σ = is the standard deviation of the predictions.

3. Results and discussion

3.1. Data analysis and exploration

The preliminary analysis reveals that the continuous variables are highly skewed (Fig. S1). Therefore, log transformation has been used to mitigate the skewed distributions, enhancing the data analysis and interpretability (Table S1). The distribution of log-transformed variables

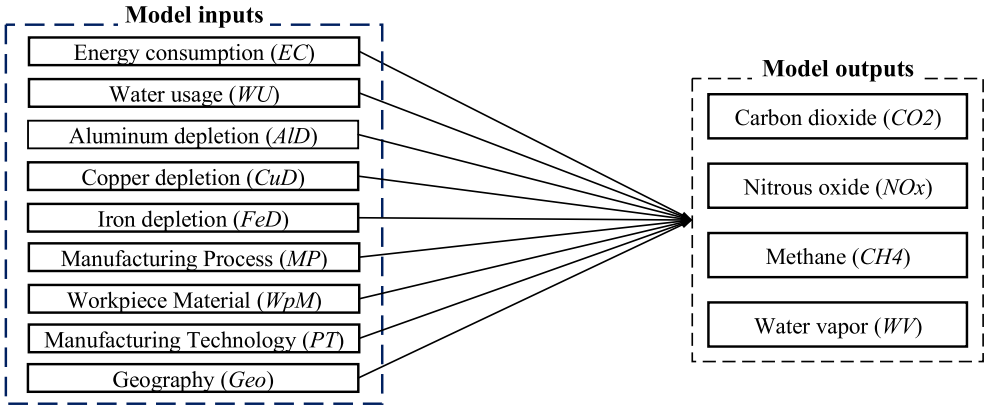

Fig. 5. Inputs and outputs for the estimation of manufacturing process outflows (GHG emissions).

Table 5
Pearson correlation coefficients (ρ) between each pair of input/output variables.

Input Output	*EC*	*WU*	*AlD*	*CuD*	*FeD*	*Geo*	*MP*	*WpM*	*PT*
CO2	0.76	0.84	0.44	0.20	0.33	0.14	0.27	−0.59	−0.10
NOx	0.59	0.80	0.31	0.36	0.098	0.12	0.20	−0.51	−0.084
CH4	0.52	0.73	0.26	0.11	0.23	0.19	0.26	−0.42	−0.16
WV	0.66	0.33	0.15	0.10	0.047	−0.53	0.27	−0.56	−0.14

Table 6
Correlation coefficients (ρ) between input/output variables for extrapolation task.

Input Output	*MP*	*WpM*	*PT*	*Geo*
EC	0.39	−0.72	−0.17	−0.27
WU	0.39	−0.60	−0.20	0.13
AlD	−0.43	−0.30	0.30	−0.13
CuD	0.14	−0.14	−0.028	0.076
FeD	0.42	0.042	−0.18	0.20

shows an approximate normal distribution for most of the variables (Fig. S2).

The PPMC method was utilized to measure the linear relationship between each pair of variables. For the GHG estimation task, Figs. S3–S6 display the correlation coefficient matrices, where off-diagonal elements indicate the correlation coefficient (ρ) between every pair of variables after data filtering. Table 5 presents the correlation coefficient (ρ) between each GHG variable (CO_2, NOx, CH_4, and WV) and each dependent variable involved in the estimation task.

Likewise, the PPMC analysis has been conducted to explore the relationship between the process variables and the output inflows for the extrapolation task. Fig. S11 to Fig. S14 shows the correlation matrices. The correlation (ρ) between each input variable (*MP*, *WpM*, *PT*, and *Geo*) and each target inflow (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) to be extrapolated is presented in Table 6.

3.2. Model performance

The hyperparameters for each model were tuned using five-fold cross-validation, and the best combination of hyperparameters was determined based on the minimum MSE of the validation set. The best hyperparameters may vary for different splits of data. Therefore, the CV procedure has been repeated for ten different splits of data, and the best hyperparameters have been selected for each split. Afterward, the ML models were re-trained on the entire training set (70% of data) using the relevant best hyperparameters. Eventually, the models are tested on the test set (30% of data) to evaluate their performance. Essentially, we embrace R^2, RMSE, and MPE as the criteria for best model selection for both estimation and extrapolation tasks, which are effective metrics that consider the extent of divergence of the predictive values from the actual

Table 7
Performance of the GraBoost model for extrapolation GHG emissions.

Output variable	Best parameters		Train set			Test set		
	n_estimators	Learning rate	R^2 (St.Dev)	*RMSE* (St.Dev)	*MPE* (%)	R^2 (St.Dev)	*RMSE* (St.Dev)	*MPE* (%)
CO_2	900	0.1	0.999 (0.001)	0.011 (0.018)	0.007	0.991 (0.001)	0.093 (0.005)	7.35
NOx	1200	0.2	0.999 (0.002)	0.0004 (0.0006)	18.91	0.972 (0.008)	0.004 (0.0006)	25.44
CH_4	500	0.05	0.999 (0.003)	0.0005 (0.001)	18.91	0.945 (0.014)	0.005 (0.001)	25.44
WV	500	0.2	1.000 (0.00)	0.001 (0.001)	6.86	0.973 (0.006)	0.042 (0.004)	16.60

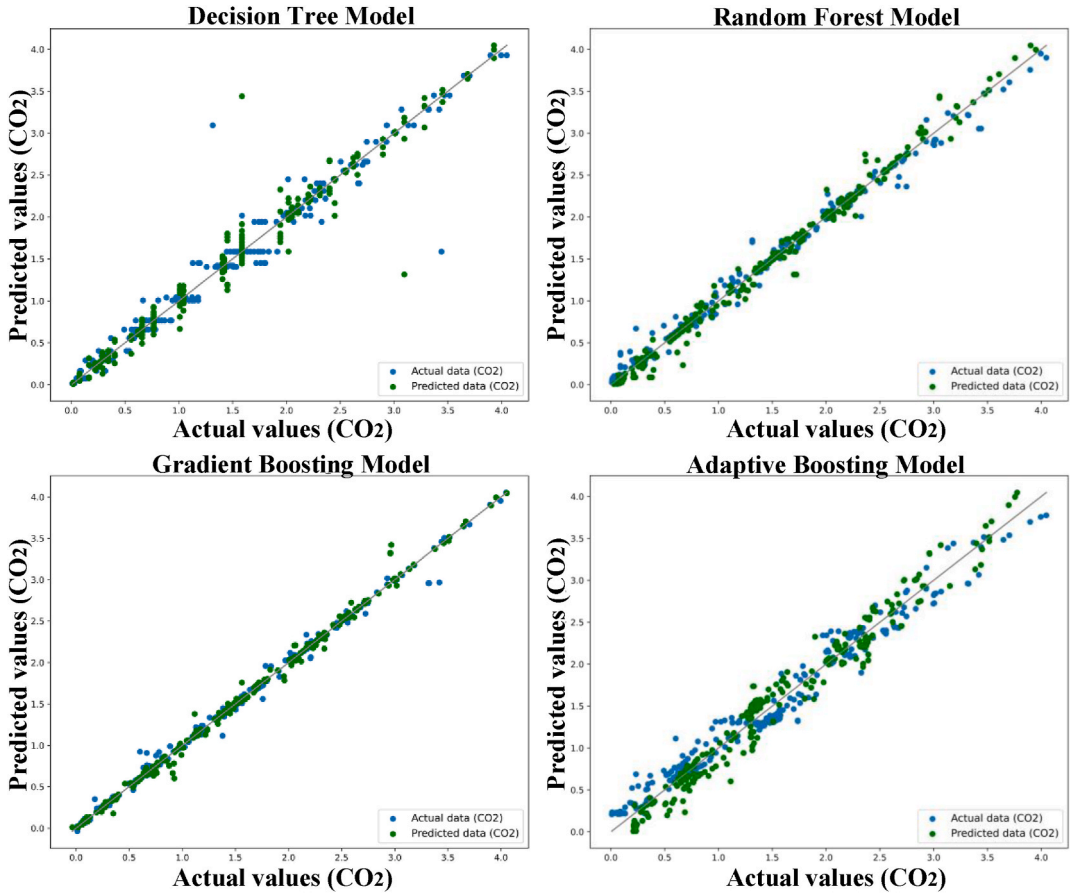

Fig. 6. Scatter plot of predicted and actual CO_2 values of the machine learning models.

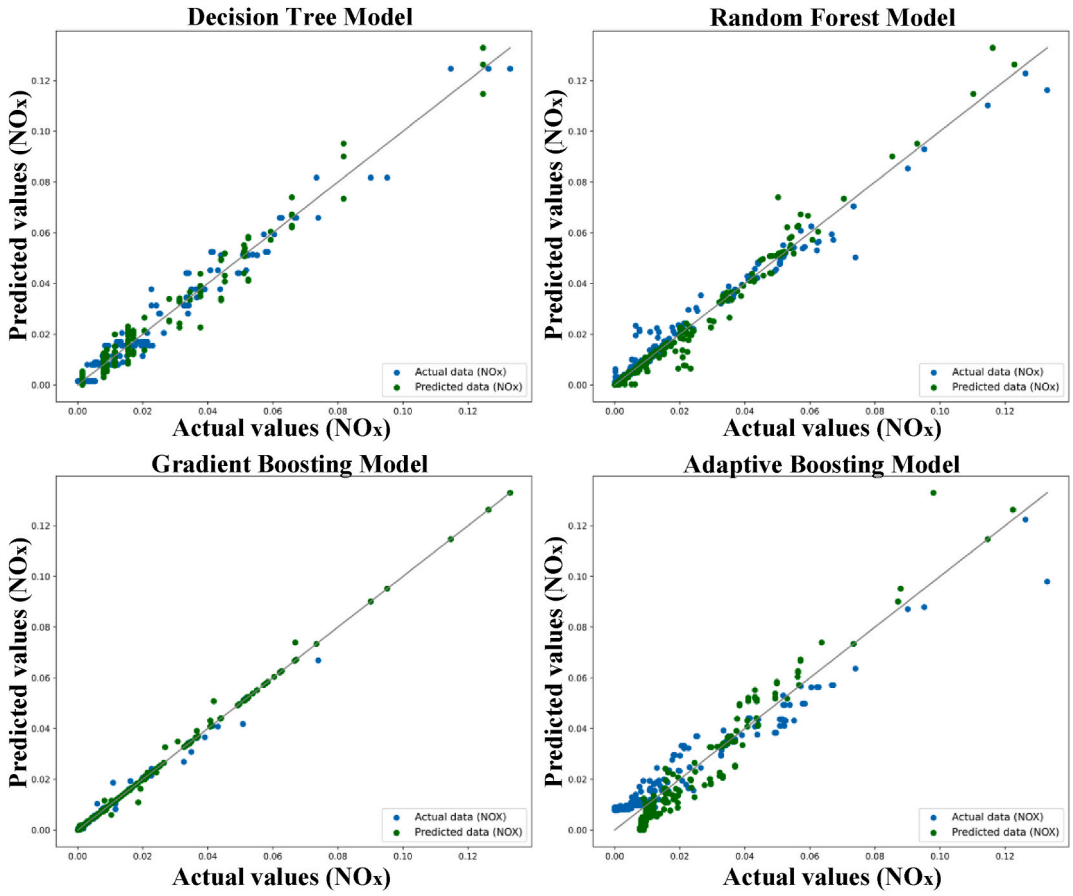

Fig. 7. Scatter plot of predicted and actual NOx values of the machine learning models.

value of the model.

3.2.1. *Model performance for GHG estimation task*

The detailed results, including R^2, RMSE, and MPE, for training the models on the whole training set for each GHG output using the best hyperparameters for ten splits of data and then the evaluation using the test set are provided (Tables S2–S17). The prediction performance of the ML models was compared, and the best model was selected based on the average of the evaluation metrics R^2, RMSE, and MPE of the model over the training and test sets.

The findings show that the GraBoost model has the best predictive performance on average to estimate the output GHG emissions. Table 7 demonstrates the average performance of GraBoost model to estimate GHG emissions due to manufacturing processes. For instance, GraBoost revealed the best performance for the estimation of CO_2 emissions, followed by RF, AdaBoost, and DT (Table S18). Fig. 6 displays the scatter plots of the actual values and the predicted values of CO_2 emissions for each model. Generally, the convergence of data points for actual and predicted values on the trendline indicates the accurate prediction of the output GHGs for the model. Evidently, the GraBoost model outperforms other models in predicting the data points over the range of CO_2 values with minimum MPE value for both training and test sets, where the model's predictions are, on average, off by 0.007% and 7.35% from the real values on training and test sets, respectively. Conversely, there are significant variations between the predicted and actual values for RF, AdaBoost, and DT algorithms represented by the dispersion among the values along the trendline.

Regarding the prediction of NOx emissions, GraBoost also shows high effectiveness, followed by RF, DT, and AdaBoost (Table S19). A high $R^2 = 0.999$ of the GraBoost model for the training set and a low RMSE = 0.0002 indicates a good fit with high predictive power. Furthermore, the lower MPE for both train and test sets mean accurate predictions across different magnitudes of NOx emissions. Although RF, DT, and AdaBoost models have high R^2 values, the models have relatively high values of RMSE and MPE, making the prediction inaccurate, as demonstrated in Fig. 7. Therefore, GraBoost generally has superior predictive accuracy on average.

Likewise, GraBoost has the best performance for estimating CH_4 and WV emissions compared to other models (Tables S20–S21). Evidently, the data points with logCH_4 values more than 0.05 are well predicted by GraBoost, while the other models are not able to fix those data points. Nevertheless, that does not prevail with all data points of CH_4, where there is a slight variation in the prediction of logCH_4 emission over the value 0.150, as shown in Fig. 8.

For WV prediction, the value $R^2 = 0.983$ for the RF model is higher than that of the GraBoost model $R^2 = 0.973$ on the test set; however, the GraBoost model still has a high R^2 value and low RMSE and MPE for the training set, which means a more precise estimation of WV emissions over the test set. Moreover, the MPE of the predictions is low for both training and testing sets, as shown in Fig. 9, which depicts the actual and predicted values of WV emission for all ML models.

3.2.2. *Model performance for LCI data extrapolation*

As mentioned above, the extrapolation task of LCI data of manufacturing processes incorporates two successive steps.

(1) The foremost is the extrapolation of the process inflows *EC*, *WU*, *AlD*, *CuD*, and *FeD* based on the generic data of the process in the Ecoinvent dataset (*MP*, *WpM*, *PT*, and *Geo*) using the ML models.

The models are trained the models on the whole training set for each process inflow using the best hyperparameters for ten splits of data, and

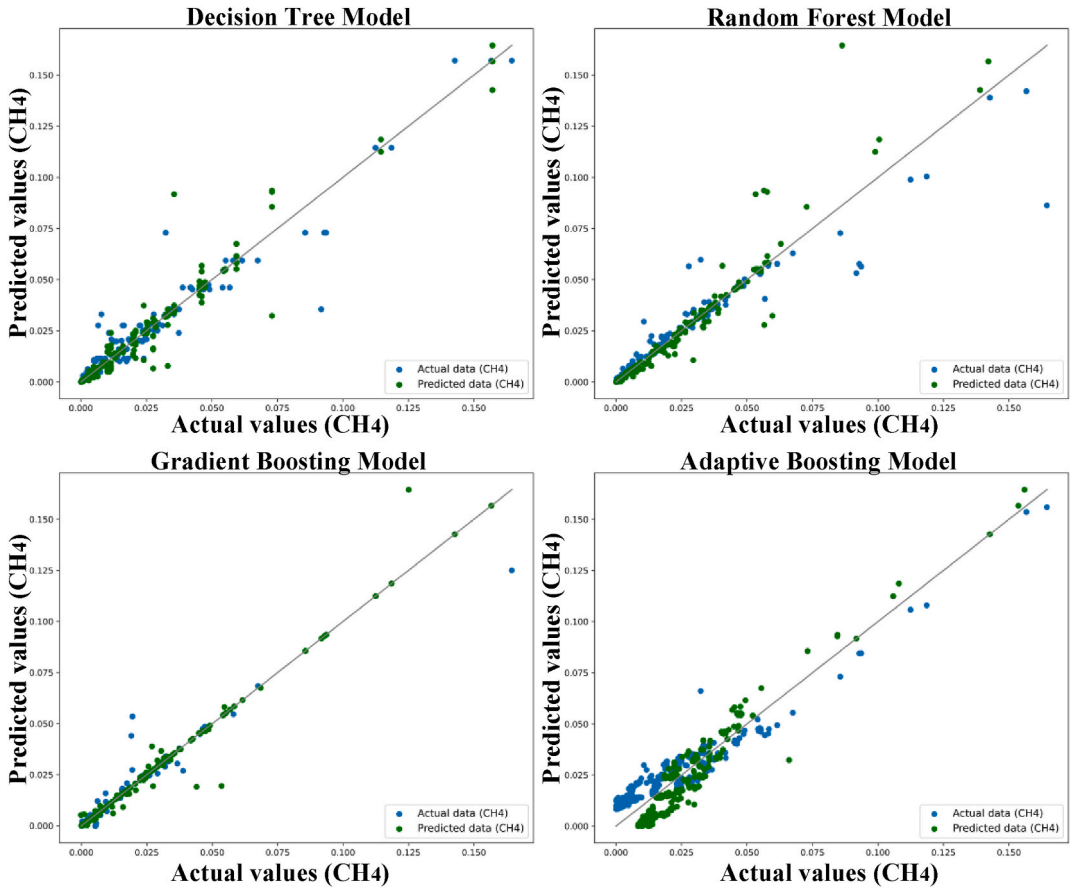

Fig. 8. Scatter plot of predicted and actual CH4 values of the machine learning models.

then the model is evaluated using the test set (Tables S27–S46). The prediction performance of the ML models was compared for selecting the best model for the extrapolation task. Likewise, the average models' performance over data splits is assessed based on the criteria R^2, RMSE, and MPE on the training and test sets (Tables S47–S51).

Obviously, the GraBoost model still outperforms the other ML models on average to extrapolate the process inflows accurately with lower RMSE and MPE. Table 8 demonstrates the average performance of the GraBoost model on training and test sets to extrapolate the process inflows due to manufacturing processes. Fig. 10 demonstrates the variation of data points between the actual values and predicted values for each process inflow.

(2) Second is the application of the developed GraBoost model as the best-verified model to extrapolate the process outflows (GHG emissions).

3.3. Model feature importance

The results emphasize the high performance of ML models in estimating and extrapolating LCI data of manufacturing processes. Nevertheless, the interpretation of the models could be ambiguous due to their complex non-linear behavior (Rudin, 2019; Muhlbacher et al., 2014). Tree-based models support PFI analysis to measure the relative importance of each feature in the model predictions, which could introduce plausible explanations of the mechanisms for estimation and extrapolation tasks. PFI analysis is derived through the contribution of each independent variable in reducing the model score through the change in mean squared error when the feature is shuffled randomly. First, for the estimation task, PFI analysis of each independent variable for GHG estimation is listed for all ML models (Tables S22–S25), and (Figs. S7–S10). Regardless of the type of output GHG, the PFI is relatively similar for all ML models. *PT*, *MP*, and *WpM* are the features with the least importance for most of the predictive models except the RF model. The reason is that these features are discrete variables where those variables with small unique values are typically less substantial in tree-based models (Palermo et al., 2009; Altmann et al., 2010). However, these variables have significant relevance in this study to distinguish the manufacturing activities, their related technology and the material being processed. Fig. 11 presents the relative importance of each feature (independent variable) on the GraBoost model predictions for every GHG output. Table S26 lists the scores associated with PFI in GraBoost model predictions for each GHG emission.

On the other hand, PFI analysis was performed to measure the influences of input variables on the prediction of each process inflows. Fig. 12 presents the relative PFI (independent variable) of the GraBoost model predictions for the extrapolation of every process inflow, i.e., *EC*, *WU*, *AlD*, *CuD*, and *FeD*.

3.4. Discussion

This study proposes the GraBoost model to bridge the data gap by estimating GHG emissions resulting from manufacturing processes and extrapolating LCI data in terms of the resources consumption of raw material and energy (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) and the released GHG to the environment (CO_2, NOx, CH_4, and WV). However, the GraBoost model has some shortcomings: (i) lacks interpretability, making it challenging to understand the model's decision-making process, and (ii) requires tuning of hyperparameters to avoid overfitting, which can be time-consuming and computationally expensive, especially for extremely large datasets. Generally, GraBoost is built to minimize the loss function (error) of the prediction. Therefore, the MSE of the training

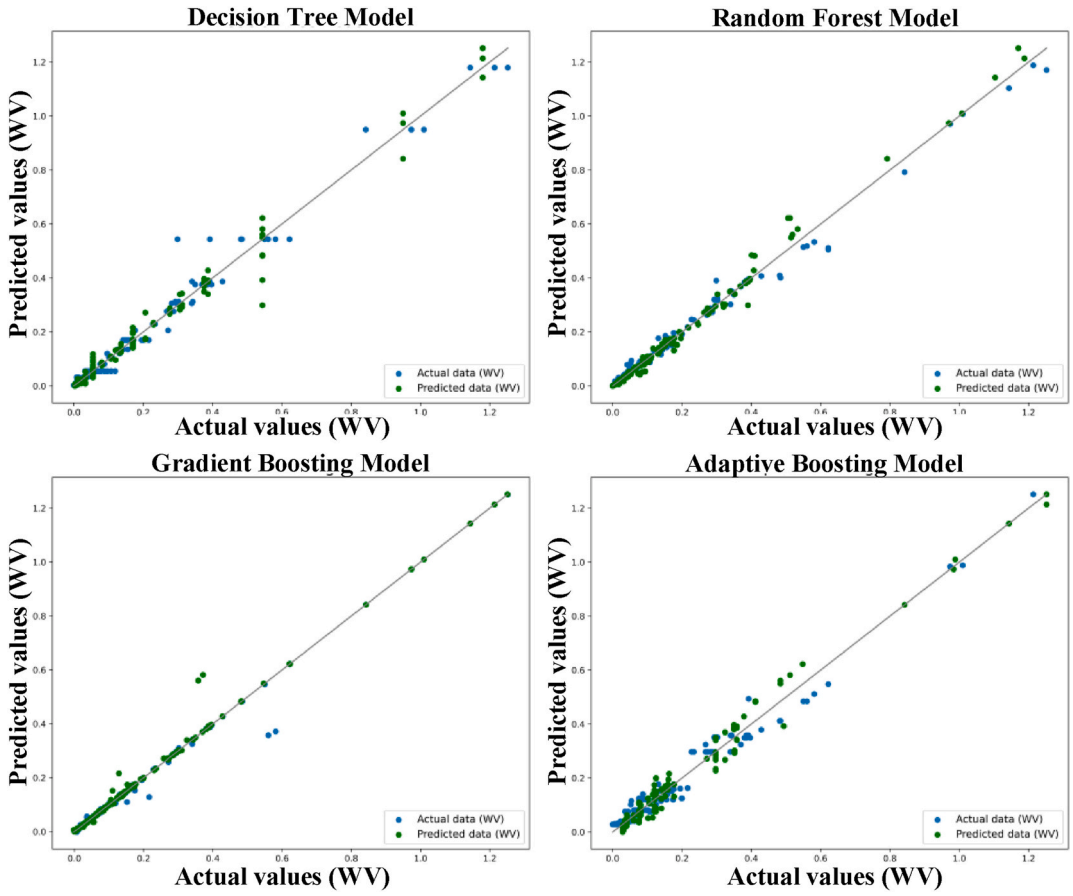

Fig. 9. Scatter plot of predicted and actual WV values of the machine learning models.

Table 8
Performance of the GraBoost model for extrapolation LCI data inflows.

Output variable	Best parameters		Training			Testing		
	n_estimators	Learning rate	R^2 (St.Dev)	*RMSE* (St.Dev)	*MPE*	R^2 (St.Dev)	*RMSE* (St.Dev)	*MPE*
EC	700	0.1	0.975 (0.011)	0.226 (0.041)	55.16	0.948 (0.008)	0.320 (0.023)	624.4
WU	700	0.3	0.992 (0.012)	0.028 (0.020)	76.19	0.939 (0.012)	0.103 (0.009)	116.7
AlD	700	0.3	0.999 (0.001)	0.005 (0.003)	132.3	0.969 (0.007)	0.040 (0.005)	128.1
CuD	900	0.05	1.000 (0.00)	0.001 (0.00)	44.71	1.000 (0.00)	0.070 (0.00)	182.6
FeD	100	0.3	0.983 (0.019)	0.028 (0.014)	60.22	0.903 (0.027)	0.067 (0.009)	65.6

set is often lower than other models. Initially, the model starts with a constant prediction equal to the average of the target variable. Then, it estimates the residual error between the actual values and the average. Based on the best hyperparameters, GraBoost builds a fixed-sized decision tree to lessen the previous error. Gradient descent is used to calculate the local minimum of the loss function at every step. Fig. 13 illustrates the decline of the loss function (deviance) with the boosting iterations of the training and test sets during the GHG estimation experiments. Consequently, the final prediction of the model is more accurate with less minimum error.

3.4.1. Model interpretation for estimation and extrapolation tasks

As aforementioned, due to the non-linear behavior of the GraBoost model, the relation between inputs and output predictions could be ambiguous. However, the PFI and PPMC analysis can provide a reasonable explanation of the model output predictions of inflows and outflows due to manufacturing processes.

Regarding GHG estimation, the correlation analysis has demonstrated highly significant correlations between the dependent variables and the released GHG from the manufacturing processes. For instance, CO_2 emission is correlated positively to *WU*, *EC*, and *AlD*, with correlation coefficients of ($\rho = 0.84$), ($\rho = 0.76$), and ($\rho = 0.44$), respectively (Panagiotopoulou et al., 2022). *WU* and *FeD* are relatively important variables and have a relatively high correlation with CO_2 ($\rho = 0.84$, $\rho = 0.33$), respectively. These could explain why higher *AlD* and *WU* are generally associated with higher CO_2 emissions.

Similarly, NOx emission has a positive correlation to *WU*, *EC*, *CuD*, and *AlD*, with correlation coefficients of ($\rho = 0.80$), ($\rho = 0.59$), ($\rho = 0.36$) and ($\rho = 0.31$), respectively. Meanwhile, the PFI analysis highlighted that *WU*, *AlD*, and *CuD* are the most influential features on the NOx predictions by the model.

For CH_4 emissions, *WU* and *AlD* are the most important features for GraBoost model predictions that originally have direct correlation coefficients ($\rho = 0.73$, $\rho = 0.26$), respectively, with CH_4. Ultimately, the GraBoost model relies on *EC* dramatically for the predictions of WV emissions. *EC* has a correlation coefficient ($\rho = 0.66$). Furthermore, there is a high negative correlation between the emissions of CO_2, NOx, CH_4, and WV and *WpM*, with correlation coefficients of ($\rho = -0.59$), ($\rho = -0.51$), ($\rho = -0.42$), and ($\rho = -0.56$), respectively. In addition, the negative correlation between *EC* and *WpM* ($\rho = -0.72$) indicates that

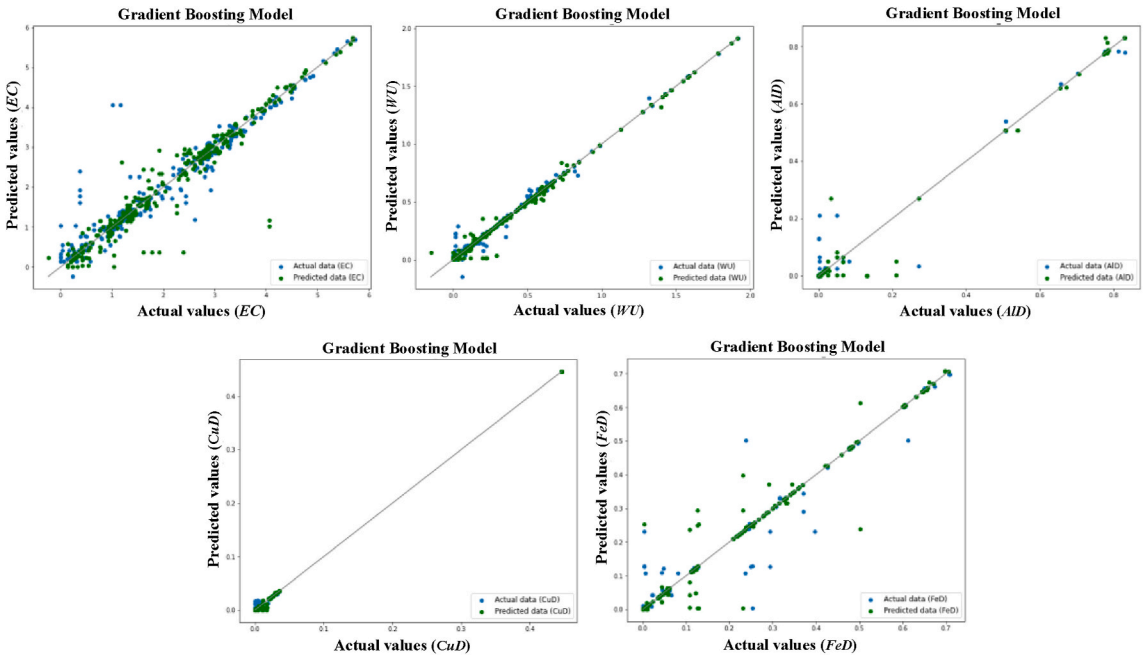

Fig. 10. Scatter plot of predicted and actual values of the process inflows (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) for the GraBoost model.

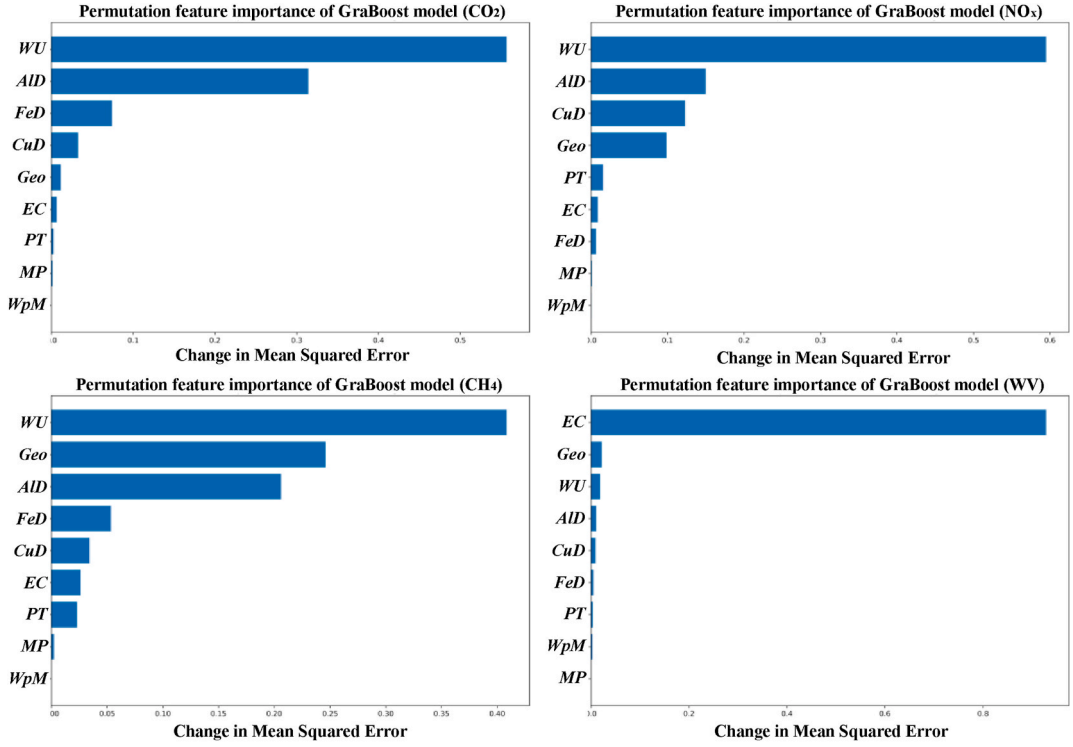

Fig. 11. PFI on the predictions of the GraBoost model to estimate GHG emissions.

different materials require disparate amounts of energy to be manufactured. This analysis elucidates that aluminum contributes to higher levels of emissions, whereas iron results in lower emissions. Furthermore, it reveals that the amounts of *WU* and *EC* are directly proportional to the emissions of CO_2, NOx, CH_4, and WV. Generally, LCA studies of the aluminum industry revealed high CO_2 emissions, methane, NOx, hydrofluorocarbons, and perfluorocarbons (Gautam et al., 2017).

In the case of the process inflow extrapolation task, *MP* and *WpM* showed a high correlation to *EC*, *WU*, *AlD*, *CuD*, and *FeD*, as presented in Table 6. At the same time, *MP* and *WpM* were the most influential features in the predictions of the GraBoost model.

3.4.2. Case study

To demonstrate the applicability and effectiveness of our proposed model, we select four new manufacturing activities, namely, iron pellet production, casting steel lost wax, aluminum oxide production, and casting aluminum lost wax to extrapolate LCI data (i.e., process inflows and GHG emissions) as case studies. Table S52 presents the activities and

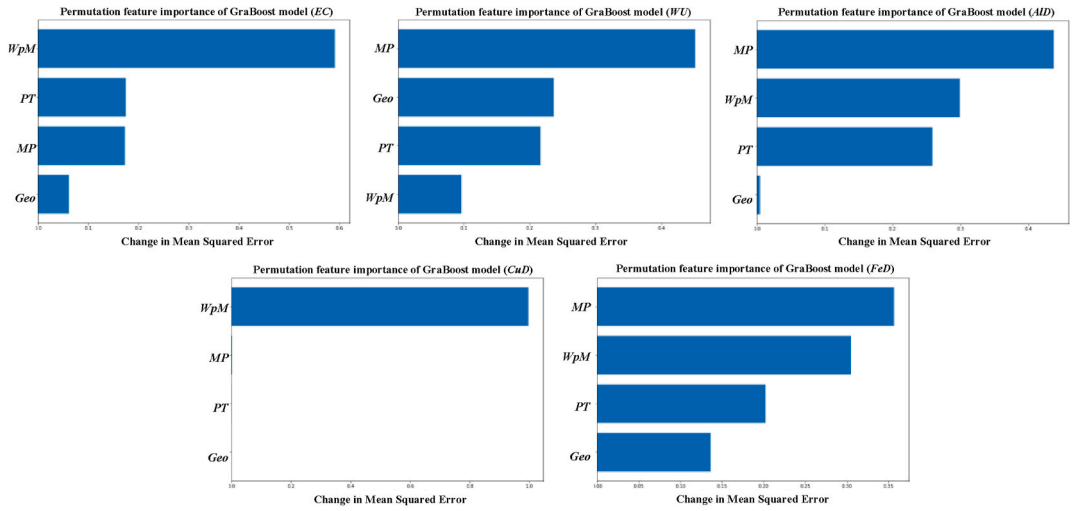

Fig. 12. PFI on the predictions of the GraBoost model to extrapolate the process inflows (*EC*, *WU*, *AlD*, *CuD*, and *FeD*).

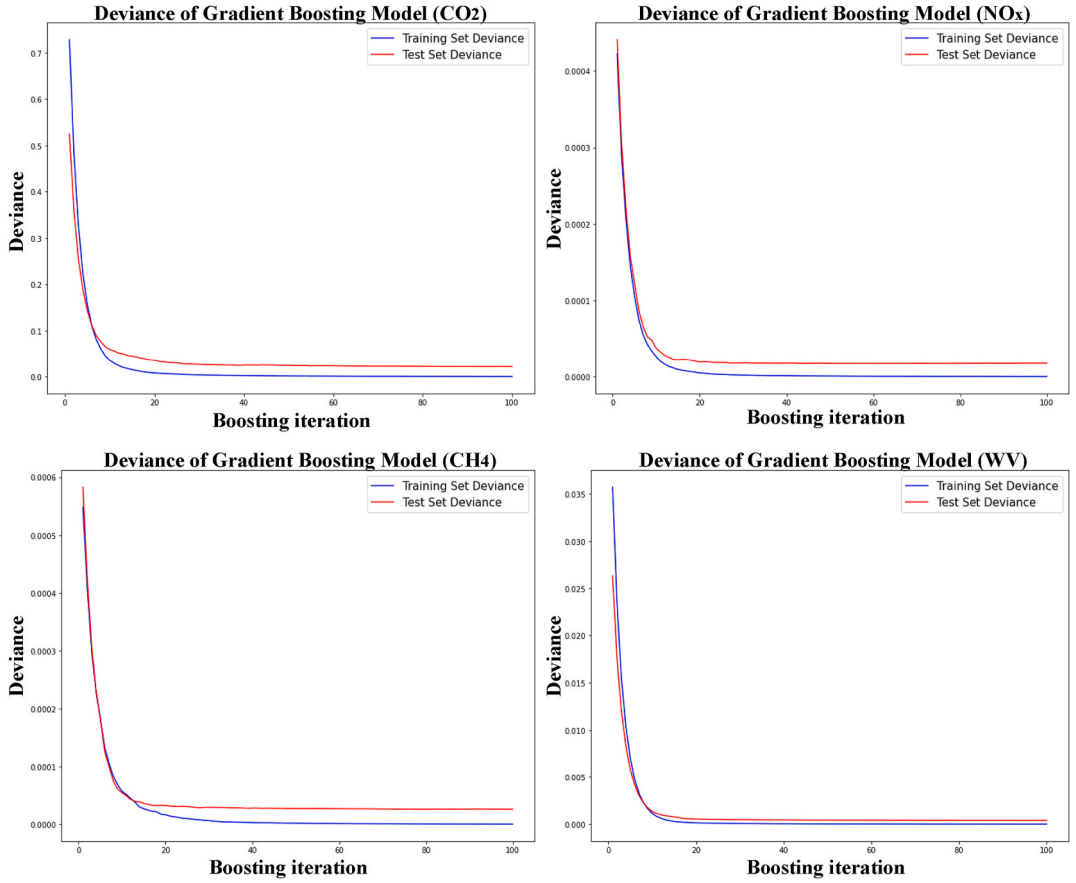

Fig. 13. Error minimization with iterations in the GraBoost model for each flow.

Table 9
95% confidence intervals of the extrapolated GHGs flows for the new activities.

Activity	Mean (μ)	St.Dev (σ)	Error margin (2σ)	Lower bound CI ($\mu - 2\sigma$)	Upper bound CI ($\mu + 2\sigma$)
CO_2	7.205	9.843	19.686	−12.481	26.891
NOx	0.0198	0.0259	0.0518	−0.032	0.0716
CH_4	0.0238	0.0302	0.0604	−0.0366	0.0842
WV	0.1893	0.3685	0.737	−0.5477	0.9263

their corresponding independent process variables. First, the trained GraBoost model was applied to extrapolate the process inflows (*EC*, *WU*, *AlD*, *CuD*, and *FeD*) (Table S53). Second, the developed GraBoost model was applied to estimate the GHGs emissions for each activity, based on the process variables and the extrapolated process inflows (Table S54). For the two steps, the extrapolated values were compared to the actual values of the flows. The RMSE and MPE have been estimated for all the extrapolated inflows and outflows for each activity, respectively (Tables S55–S56). The results reveal that our proposed approach can extrapolate new LCI data efficiently.

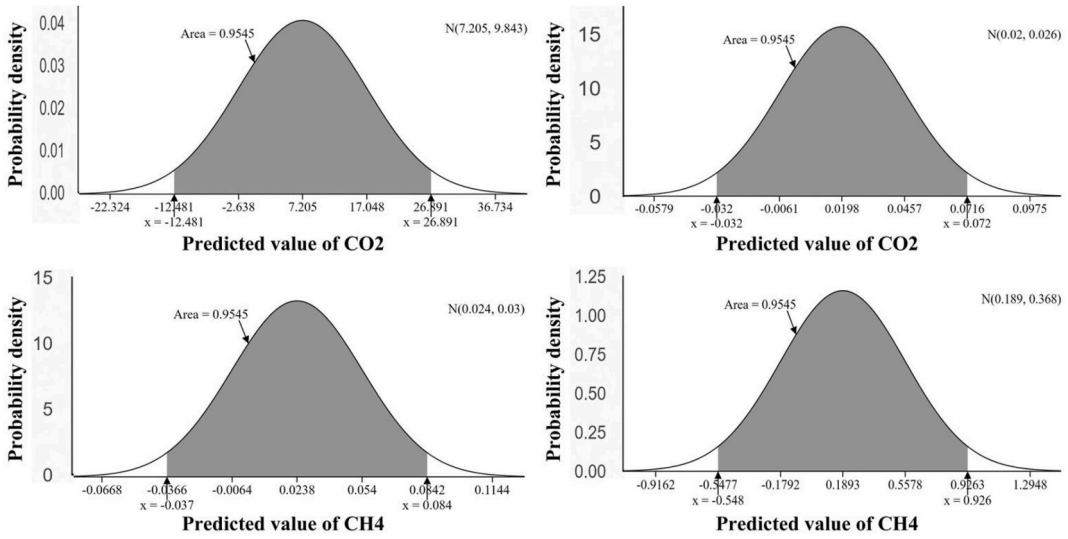

Fig. 14. The normal distribution of confidence intervals $CI(95\%) = \mu \pm 2\sigma$ GraBoost model output.

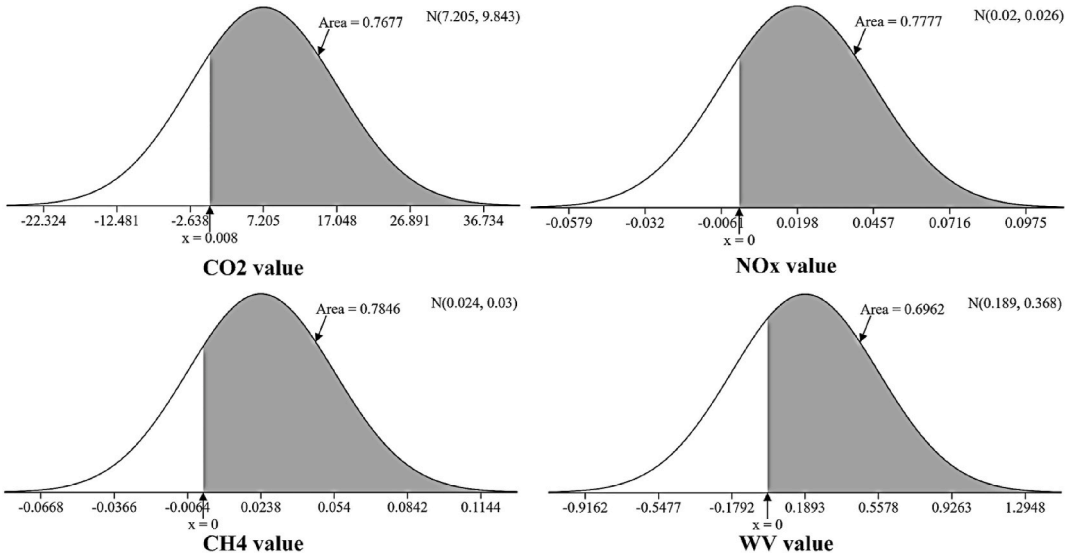

Fig. 15. The probability range for the extrapolated GHGs (area under the curve).

3.5. Uncertainty measurement for extrapolation task

The bootstrapping method was applied to determine the lower bound and upper bound of the 95% confidence interval, which indicate the potential range that the extrapolated values of GHGs fall in between when conducting the extrapolation task for new manufacturing activities using the GraBoost model. Table 9 illustrates the 95% confidence interval for each flow of the output GHGs. Fig. 14 presents the normal distribution for the predicted values of GHGs using the GraBoost model. Moreover, we measure the probability distribution indicated by the area under the normal curve between the minimum and maximum values of each flow, as depicted in Fig. 15. It reveals that our model can be used to extrapolate the values of GHGs of manufacturing activities over a wide range of values.

4. Conclusion

LCA studies for manufacturing processes are essential to grasp the hot spots for mitigating the environmental impact and promoting reliable, sustainable decision-making. Extrapolation of LCI data for new manufacturing activities is a considerable challenge due to the scarcity of data in the existing LCI databases. This paper proposed a novel computational approach based on ML models to extrapolate LCI data for manufacturing activities in terms of the inflows into the process (natural resource consumption of energy, water, and material) and the outflows into the environment (GHG emissions). Initially, LCI data of manufacturing activities (n = 315 observations) have been divided into two datasets for training and testing the models (70% and 30%, respectively). Then, a five-fold cross-validation has been applied for tuning the hyperparameters of the ML models. Second, the models were re-trained using the best hyperparameters and then evaluated using the test set. The findings reveal that the Gradient Boosting (GraBoost) model developed in this study outperforms the other models with the highest R^2 value and lowest RMSE and MPE on the test set. Even though the model structure could be complex to interpret, the GraBoost model provides a fast and efficient method to extrapolate energy consumption, water usage, and raw material depletion in manufacturing. Accordingly, it can estimate the quantities of CO_2, NOx, CH_4, and WV emissions due to manufacturing activities. The proposed approach can support quantifying carbon footprint for a wide range of manufacturing and

conducting LCA studies reliably.

CRediT authorship contribution statement

Mohamed Saad: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Yingzhong Zhang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Jia Jia:** Validation, Investigation, Formal analysis, Data curation. **Jinghai Tian:** Validation, Software, Investigation.

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

Data will be made available on request.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2020YFB1711601) and the National Natural Science Foundation of China (No. 51775081). The authors thank the editors and reviewers for their helpful suggestions on this study. Also, the authors want to express our sincere gratitude to PR'e Sustainability for granting the Ecoinvent database license.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2024.121152.

References

Altmann, A., Tološi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26. https://doi.org/10.1093/bioinformatics/btq134.

Biau, G., 2012. Analysis of a random forests model. J Machine Learning Res 13 (1), 1063–1095.

Breiman, L., 2001. Random forests. Mach. Learn. 45 https://doi.org/10.1023/A:1010933404324.

Canals, L.M.I., Azapagic, A., Doka, G., Jefferies, D., King, H., Mutel, C., Nemecek, T., Roches, A., Sim, S., Stichnothe, H., Thoma, G., Williams, A., 2011. Approaches for addressing life cycle assessment data gaps for bio-based products. J. Ind. Ecol. 15 https://doi.org/10.1111/j.1530-9290.2011.00369.x.

Cashman, S.A., Meyer, D.E., Edelen, A.N., Ingwersen, W.W., Abraham, J.P., Barrett, W. M., Gonzalez, M.A., Randall, P.M., Ruiz-Mercado, G., Smith, R.L., 2016. Mining available data from the United States environmental protection agency to support rapid life cycle inventory modeling of chemical manufacturing. Environ. Sci. Technol. 50, 9013–9025. https://doi.org/10.1021/ACS.EST.6B02160/SUPPL_FILE/ES6B02160_SI_001.PDF.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/3-540-45014-9_1.

Environmental Engineering for the 21st Century: Addressing Grand Challenges, 2019. Environmental Engineering for the 21st Century: Addressing Grand Challenges. https://doi.org/10.17226/25121.

Gautam, M., Pandey, B., Agrawal, M., 2017. Carbon footprint of aluminum production. In: Environmental Carbon Footprints: Industrial Case Studies. https://doi.org/10.1016/B978-0-12-812849-7.00008-8.

Hou, P., Cai, J., Qu, S., Xu, M., 2018. Estimating missing Unit process data in life cycle assessment using a similarity-based approach. Environ. Sci. Technol. 52, 5259–5267. https://doi.org/10.1021/acs.est.7b05366.

Lee, E.K., Zhang, W.J., Zhang, X., Adler, P.R., Lin, S., Feingold, B.J., Khwaja, H.A., Romeiko, X.X., 2020. Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach. Sci. Total Environ. 714 https://doi.org/10.1016/j.scitotenv.2020.136697.

Li, J., Irfan, M., Samad, S., Ali, B., Zhang, Y., Badulescu, D., Badulescu, A., 2023. The relationship between energy consumption, CO2 emissions, economic growth, and health indicators. Int J Environ Res Public Health 20. https://doi.org/10.3390/ijerph20032325.

Malyan, S.K., Singh, O., Kumar, A., Anand, G., Singh, R., Singh, S., Yu, Z., Kumar, J., Fagodiya, R.K., Kumar, A., 2022. Greenhouse Gases Trade-Off from Ponds: An Overview of Emission Process and Their Driving Factors. Water (Switzerland). https://doi.org/10.3390/w14060970.

Meng, F., LaFleur, C., Wijesinghe, A., Colvin, J., 2019. Data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology. Fuel 246, 187–195. https://doi.org/10.1016/J.FUEL.2019.02.124.

Meron, N., Blass, V., Thoma, G., 2020. Selection of the most appropriate life cycle inventory dataset: new selection proxy methodology and case study application. Int. J. Life Cycle Assess. 25 https://doi.org/10.1007/s11367-019-01721-8.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. 18, 275–285. https://doi.org/10.1002/CEM.873.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front Neurorobot 7, 63623. https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX.

Palermo, G., Piraino, P., Zucht, H.D., 2009. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. Comput. Biol. Chem. Adv. Appl. 2 https://doi.org/10.2147/aabc.s3619.

Panagiotopoulou, V.C., Stavropoulos, P., Chryssolouris, G., 2022. A critical review on the environmental impact of manufacturing: a holistic perspective. Int. J. Adv. Manuf. Technol. 118, 603–625. https://doi.org/10.1007/S00170-021-07980-W/TABLES/7.

Qin, J., Gong, N., 2022. The estimation of the carbon dioxide emission and driving factors in China based on machine learning methods. Sustain Prod Consum 33. doi:10.1016/j.spc.2022.06.027.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. https://doi.org/10.1038/s42256-019-0048-x.

Steurer, M., Hill, R.J., Pfeifer, N., 2021. Metrics for evaluating the performance of machine learning based automated valuation models. J. Property Res. 38 https://doi.org/10.1080/09599916.2020.1858937.

Sun, Y., Wang, X., Ren, N., Liu, Y., You, S., 2023. Improved machine learning models by data processing for predicting life-cycle environmental impacts of chemicals. Environ. Sci. Technol. 57 https://doi.org/10.1021/acs.est.2c04945.

Vanneschi, L., Silva, S., 2023. Decision tree learning. In: Natural Computing Series. https://doi.org/10.1007/978-3-031-17922-8_6.

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., 2016. The ecoinvent database version 3 (part I): overview and methodology. Int. J. Life Cycle Assess. 21 https://doi.org/10.1007/s11367-016-1087-8.

West, R.M., 2022. Best practice in statistics: the use of log transformation. Ann. Clin. Biochem. 59, 162–165. https://doi.org/10.1177/00045632211050531/ASSET/IMAGES/LARGE/10.1177_00045632211050531-FIG2.JPEG.

Zhao, B., Shuai, C., Hou, P., Qu, S., Xu, M., 2021. Estimation of unit process data for life cycle assessment using a decision tree-based approach. Environ. Sci. Technol. 55, 8439–8446. https://doi.org/10.1021/ACS.EST.0C07484/SUPPL_FILE/ES0C07484_SI_001.XLSX.

Zargar, S., Yao, Y., Tu, Q., 2022. A review of inventory modeling methods for missing data in life cycle assessment. J. Ind. Ecol. 26 https://doi.org/10.1111/jiec.13305.

Zhu, X., Ho, C.H., Wang, X., 2020. Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes. ACS Sustain Chem Eng 8. https://doi.org/10.1021/acssuschemeng.0c02211.