Science of the Total Environment 912 (2024) 168969

Contents lists available at ScienceDirect

Science of the Total Environment

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

Review

A review of machine learning applications in life cycle assessment studies

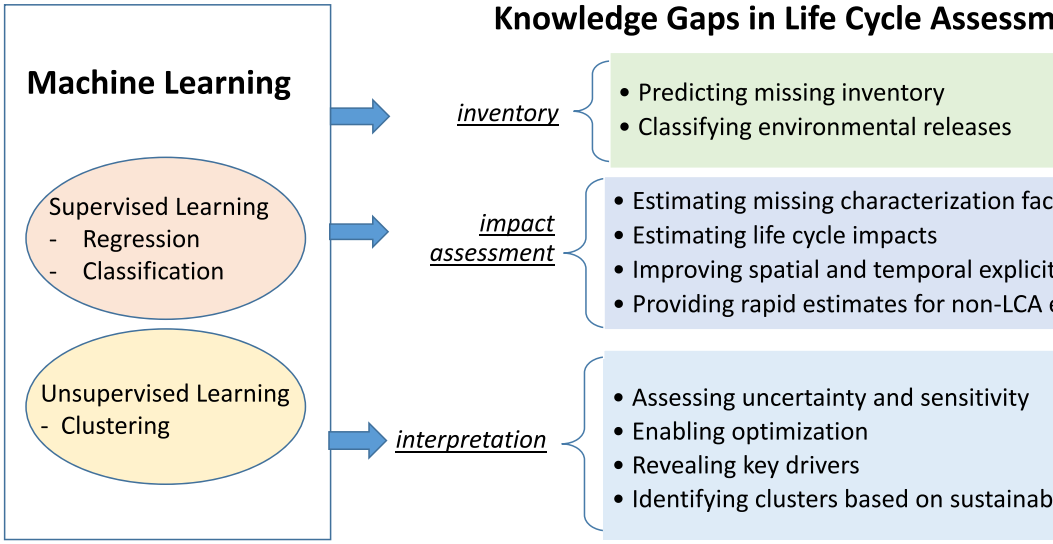
Xiaobo Xue Romeiko [a,*], Xuesong Zhang [b,*], Yulei Pang [c], Feng Gao [b], Ming Xu [d], Shao Lin [a], Callie Babbitt [e]

[a] Department of Environmental Health Sciences, University at Albany, State University of New York, United States of America
[b] Hydrology and Remote Sensing Laboratory, United States Department of Agriculture, United States of America
[c] Department of Math, Southern Connecticut State University, United States of America
[d] Dvision of Environmental Ecology, School of Environment, Tsinghua University, China
[e] Department of Sustainability, Golisano Institute for Sustainability, Rochester Institute of Technology, United States of America

HIGHLIGHTS

- MLs were used for life cycle inventory, impact assessment and interpretation stages.
- ML improved prediction accuracy, pattern discovery and computational efficiency for LCAs.
- Continuous data collection and compilation is needed to support more reliable ML for LCAs.
- Robust ML modeling selection and uncertainty evaluation are needed.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Deyi Hou

Keywords:
Life cycle assessment
Machine learning
Sustainability
Prediction
Model selection

ABSTRACT

Life Cycle Assessment (LCA) is a foundational method for quantitative assessment of sustainability. Increasing data availability and rapid development of machine learning (ML) approaches offer new opportunities to advance LCA. Here, we review current progress and knowledge gaps in applying ML techniques to support LCA, and identify future research directions for LCAs to better harness the power of ML. This review analyzes forty studies reporting quantitative assessment with a combination of LCA and ML methods. We found that ML approaches have been used for generating life cycle inventories, computing characterization factors, estimating life cycle impacts, and supporting life cycle interpretation. Most of the reviewed studies employed a single ML

* Corresponding authors.
E-mail addresses: xxue@albany.edu (X.X. Romeiko), xuesong.zhang@usda.gov (X. Zhang).

https://doi.org/10.1016/j.scitotenv.2023.168969
Received 16 April 2023; Received in revised form 24 November 2023; Accepted 27 November 2023
Available online 28 November 2023

method, with artificial neural networks (ANNs) as the most frequently applied approach. Both supervised and unsupervised ML techniques were used in LCA studies. For studies using supervised ML, training datasets were derived from diverse sources, such as literature, lab experiments, existing databases, and model simulations. Over 70 % of these reviewed studies trained ML models with less than 1500 sample datasets. Although these reviewed studies showed that ML approaches help improve prediction accuracy, pattern discovery and computational efficiency, multiple areas deserve further research. First, continuous data collection and compilation is needed to support more reliable ML and LCA modeling. Second, future studies should report sufficient details regarding the selection criteria for ML models and present model uncertainty analysis. Third, incorporating deep learning models into LCA holds promise to further improve life cycle inventory and impact assessment. Finally, the complexity of current environmental challenges calls for interdisciplinary collaborative research to achieve deep integration of ML into LCA to support sustainable development.

1. Introduction

Life Cycle Assessment (LCA) is the foundational method for quantitative sustainability assessment (Hellweg and Canals, 2014). LCA is a systematic assessment approach, capable of evaluating resource consumption and environmental impacts of products, as well as processes and services over their entire lifespan. Due to their comprehensive scope, LCA studies have been successfully applied to support technology development, policy analyses and green business marketing. Notably, the comprehensiveness of LCA entails extensive data collection across the supply chain and advanced data analytics. Collating diversely sourced big datasets over all upstream processes (i.e. resource extraction, production, and transport) as well as downstream processes (i.e. product use and disposal), ensuring high quality of all relevant datasets, and conducting prudent analyses are essential for credible LCAs.

Rapid developments in data generation, storage and analytics propel increasing interests in harnessing the power of big datasets (Cooper et al., 2013; Xu et al., 2015) and machine learning (ML) techniques to advance LCA (Romeiko et al., 2020a; Romeiko et al., 2020b; Xu et al., 2015). ML, a subfield of artificial intelligence, is the study of computer algorithms that improve automatically through experience (Mitchell, 1997). ML can decipher the complexity of datasets, enable prediction, and discover new knowledge and patterns hidden behind the datasets. ML methods are broadly categorized into supervised learning and unsupervised learning. Supervised learning identifies patterns that relate variables to measured outcomes and maximizes accuracy when predicting those outcomes (James et al., 2013). For example, linear, tree-based, distance-based, nature-inspired, neural network, and deep learning models are frequently supervised learning approaches (Hou et al., 2020; Naseri et al., 2020; Romeiko et al., 2020b; Slapnik et al., 2015a; Thilakarathna et al., 2020). Unsupervised learning exploits innate properties of the input datasets to detect trends and patterns without explicit designating the outcome of interest (Han et al., 2011). Unsupervised learning often includes clustering, association rules, and dimension reduction analyses (Abdella et al., 2020; Feng et al., 2019; Mao et al., 2019). Both learning approaches are widely and successfully applied in a variety of disciplines such as food (Saha and Manickavasagan, 2021), building (Fathi et al., 2020), climate (Rolnick et al., 2022), and public health (Santos et al., 2019).

Recent efforts have begun to explore utilizing ML to support LCA applications (Dick et al., 2015; Marvuglia et al., 2015b; Ramakrishnan et al., 2012; Slapnik et al., 2015b; Sousa et al., 2001; Sundaravaradan et al., 2011). For example, ML models have been used to predict missing life cycle inventory (Hou et al., 2020; Sundaravaradan et al., 2011), estimate the life cycle impacts of chemicals (Song et al., 2017), and assist in life cycle interpretation (Azari et al., 2016; Sharifa and Hammad, 2019). There are a few recent reviews summarizing the application of ML in LCAs (Barros and Ruschel, 2020; Ghoroghi et al., 2022). Although these reviews are valuable and insightful, they focus on either a single sector such as building (Barros and Ruschel, 2020) or applications at multiple scales such as building or cities/communities (Ghoroghi et al., 2022). An in-depth review of the current applications of ML in various LCA stages and associated merits and challenges is necessary. Additionally, a discussion of future research directions is needed to improve the integration of ML and LCA.

To fill in this knowledge gap, this article reviews existing publications that reported applications of ML approaches to support LCA. Specifically, this study identifies the purposes of applying ML approaches for LCA, examines the types of ML approaches applied to support LCA, and analyzes data sources used for ML development. Furthermore, this article discusses the strengths and limitations of current applications of ML in LCA, as well as future research directions for innovating LCA methods with ML approaches.

2. Methods

Guided by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Liberati et al., 2009), we followed the seven steps outlined in Fig. 1 to identify relevant articles and conduct the review. During Step 1, we identified original research articles published during years 2010–2020 by searching the "Web of Science" and "Science Direct" publication databases using different combinations of keywords, including "life cycle assessment", "life cycle analysis" and "machine learning". While the literature review requires a finite time frame, we recognize the potential limitation of excluding studies published after the review was complete in year 2021. For Step 2, we removed the repetitive studies from the previous step. In Step 3, we screened the articles based on their abstracts. If the abstract appeared to be irrelevant with either LCA or ML, we removed the articles from our review. Step 4 further screened the full text of the remaining articles and removed the articles that did not explicitly mention both LCA and ML. In step 5, those studies that did not conduct quantitative analyses with ML and LCA were removed. In Step 6, to expand the search, we tracked the articles, which have cited the articles identified in Step 5 via Google Scholar. Finally (step 7), we thoroughly evaluated these studies from both LCA and ML perspectives.

3. Results

3.1. Articles identified following the PRISMA guideline

Table 1 summarizes the identified publications under each combination of keywords and from different databases. With the keywords "life cycle assessment" and "machine learning", the search yielded 26 records from web of science and 203 records from science direct. Using the keywords "life cycle analysis" and "machine learning", five records were found on web of science and 93 records on science direct. Combining the publications resulting from different keywords and databases, our initial search found 327 records. After removing duplicate records, 305 articles were identified. After screening the abstracts of those 305 articles, 173 articles were removed due to irrelevance to the joint use of ML and LCA. We further screened the full manuscripts of the remaining 132 articles, and identified that 37 articles explicitly reported the application of both LCA and ML. Those articles excluded from this study applied either LCA or ML alone, but neither of them. Among the final 37 articles, 8 studies only qualitatively discussed the LCA and ML,

and the remaining 29 studies performed quantitative analysis with LCA and ML models. Additionally, by tracking the citations of these 29 studies, we included additional 11 articles, which also used quantitative LCA and ML models. As a result, 40 articles that conducted quantitative assessment with a combination of LCA and ML models were included in our final review and analysis (Tables S1 and S2). The Table S2 provides the details of the addressed knowledge gaps, applied LCA stages, ML model algorithms, ML model inputs, ML model outputs, data sources, ML validation procedures, model comparison, and study outcomes for each article.

3.2. Trends in ML & LCA publications and application areas

Most of the ML & LCA articles (27 out of 40) were published between 2019 and 2020 (Fig. 2). Prior to 2018, less than 4 articles/year were found to feature the combination of LCA and ML. These articles were published in 22 journals across different disciplines. Among those journals, Journal of Cleaner Production and Science of Total Environment are two major hubs for publishing LCA articles that used ML approaches. These articles cover various disciplines such as agriculture, buildings, chemicals, energies and manufacturing processes (Fig. 2).

Table 1
Number of articles found in the search databases.

Search #	Databases	Query	# of found articles	# of duplicated articles
1	Web of Science	"Life cycle assessment" and "machine learning"	26	0
2	Web of Science	"Life cycle analysis" and "machine learning"	5	1
3	Science Direct	"Life cycle assessment" and "machine learning"	203	16
4	Science Direct	"Life cycle analysis" and "machine learning"	93	5

Among these application areas, agriculture is the top focus area. Approximately 37 % of total articles discussed agricultural products or processes. Next, 20 % of total articles discussed buildings or building materials, and 15 % of total articles focused on chemical toxicity or green chemistry. Approximately 10 % of total articles analyzed the energy sector, and 7 % focused on manufacturing processes.

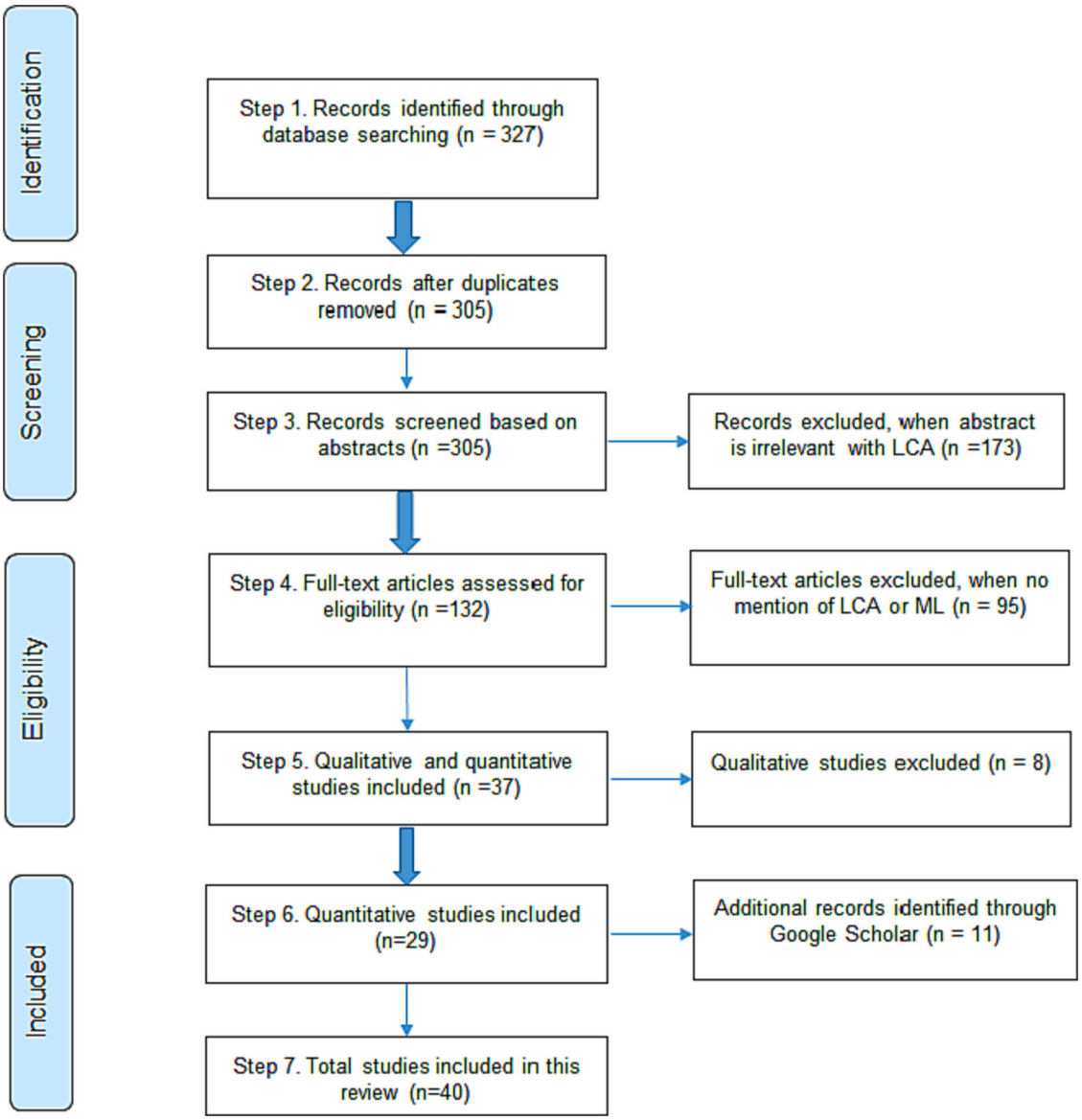


Fig. 1. The review process and associated articles based on the PRISMA guideline.

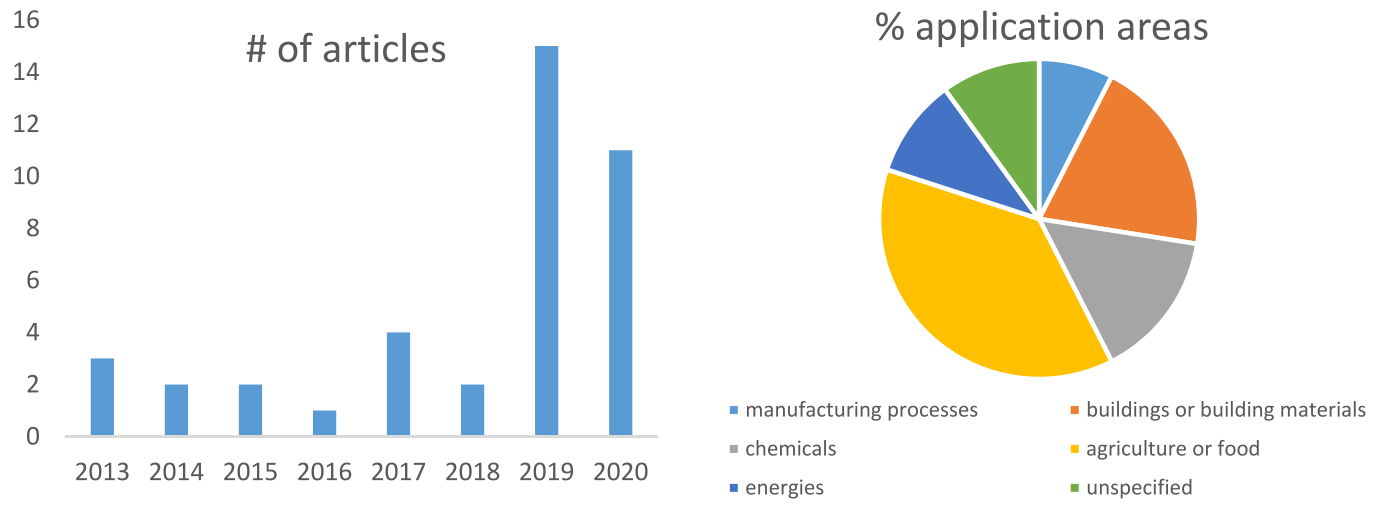


Fig. 2. The number of publications/year and application areas of the reviewed studies.

3.3. Addressed knowledge gaps by applying ML in LCA

The majority of the studies (35 studies in group 1 out of total 40) framed ML models as regression applications to predict values for the LCAs (Table 2). The first three subgroups (subgroups 1.1, 1.2, and 1.3), containing 23 studies, used ML models to fill in key data gaps of life cycle inventory, characterization factors, and life cycle impact, respectively. First, the life cycle inventory is lacking for emerging technologies and products due to unavailability of measurement datasets. Six studies utilized ML to fill in the data gaps in the life cycle inventories, as described in subgroup 1.1. Second, the characterization factors exist only for a small set of chemicals due to limited laboratory testing. Therefore, seven studies in subgroup 1.2 relied on ML to determine the missing characterization factors. Third, eight studies in subgroup 1.3 focused on applying ML models to compute life cycle impacts crop production in regions (such as Iran), where agricultural LCAs were less studied. Overall, ML models were applied to estimate key datasets to enable life cycle inventory and impact assessment in the first three subgroups.

ML models also were used to address method and application challenges of LCAs, as demonstrated in subgroups 1.4 through 1.7. First, LCA has been criticized as a spatially and temporally coarse approach, which limits its capability of supporting regional or time-sensitive decision making. To address this method challenge, two studies in subgroup 1.4 used ML to improve the spatial resolutions and temporal dynamics of agricultural life cycle impacts. Second, to aid designers in estimating the environmental impacts of various design options, MLs were developed as simplified and rapid surrogate models, which are more user friendly than traditional LCA models for designers and other non-LCA experts. Seven studies in subgroup 1.5 generated ML models based upon life cycle impact datasets with the goal of providing rapid and accurate life cycle assessment approaches for non-LCA experts. Third, like other quantitative approaches, the uncertainty and sensitivity of LCA models have been hot subjects. To contribute to this area, three studies in subgroup 1.6 applied MLs to quantify uncertainty of LCA models, and one additional study in the same subgroup assessed sensitivity of life cycle impacts. Finally, the subgroup 1.7, consisting of two studies, used ML to bridge LCA and optimization approaches, which eventually enabled determining the optimized product designs with the best product performances and lowest environmental impacts and costs.

Each of the remaining four groups include only one or two studies. (Marvuglia et al., 2015a) was the only study in group 2, whose aim was to reduce the input parameters for quantifying chemical characterization factors. (Romeiko et al., 2020a) and (Zhao et al., 2019) in group 3 used ML models for feature ranking in order to determine the relative importance of driving factors for agricultural life cycle impact estimates. Different from the aforementioned studies, (Abdella et al., 2020) in group 4 clustered the food sectors based on their sustainability performances. Additionally, (Tao et al., 2018) in group 5 framed their research questions as classification problems, and estimated environment releases from chemical use based upon the chemical classes.

3.4. ML applications in various stages of life cycle assessment

Based upon the relevant LCA stages for ML applications, this review categorized the total 40 studies into three groups (shown in Table 3). For the first group described in Section 3.4.1, ten studies focused on utilizing ML models in the life cycle inventory stage. For the second group described in Section 3.4.2, 22 studies applied ML in the life cycle impact assessment stage. For the last group described in Section 3.4.3, the remaining eight studies applied ML to support the life cycle interpretation stage.

3.4.1. ML for life cycle inventory

Depending on the scope of life cycle inventory, the ten studies in the first group were classified into two subgroups. For the first subgroup (subgroup 1.1), nine studies used ML to estimate the foreground life cycle inventory. Three studies among these nine studies used ML to estimate environmental emissions, which directly served as part of foreground life cycle inventories. Meng et al. (2019) used a linear regression model to predict greenhouse gas emissions, non-methane hydrocarbon and carbon monoxide from dual fuel diesel engines in oilfield operations. (Tao et al., 2018) estimated the release of organic chemicals from the use and post-use of chemical products with an artificial neural network (ANN) model. Nguyen et al. (2019) developed an ANN model as a surrogate model to estimate the greenhouse gas emissions and nutrient leaching of irrigated corn production systems at a high spatial resolution in eastern Colorado, USA.

The other six studies within the first subgroup used ML to estimate the product characteristics, which were fed into additional models to estimate foreground life cycle inventory. For example, Cheng et al. (2020a) used a random forest approach to predict biochar yields and characteristics, which were incorporated into the LCA framework for estimating energy use and greenhouse gas emissions of slow pyrolysis. Similarly, Cheng et al. (2020b) developed several ML models to predict yields and characteristics of biobased chemicals derived from hydrothermal treatment, and then calculated energy consumption and greenhouse gas emissions of hydrothermal treatment processes. Liao et al. (2020) estimated the yield of activated carbon from the biomass feedstocks, which was fed into Aspen simulation and LCA framework to

Table 2
Categories of knowledge gaps addressed in the reviewed studies.

Group #	ML applications	Addressed knowledge gaps	Studies
1	Value prediction	Subgroup 1.1: Predicting unknown life cycle inventory of emerging technologies or products	(Meng et al., 2019), (Nguyen et al., 2019), (Cheng et al., 2020a), (Thilakarathna et al., 2020), (Liao et al., 2020), (Cheng et al., 2020b), (Cornago et al., 2020), (Naseri et al., 2020), (Sharifa and Hammad, 2019)
		Subgroup 1.2:Estimating missing characterization factor	(Hou et al., 2020), (Slapnik et al., 2015a)
		Subgroup 1.3:Estimating life cycle impacts of emerging products or in understudied regions	(Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017), (Pishgar-Komleh et al., 2020b), (Duprez et al., 2019), (Vlontzosa and Pardalosb, 2017), (Zhu et al., 2020), (Ozbilen et al., 2013)
		Subgroup 1.4:Improving spatial and temporal explicitness of life cycle impacts of agriculture	(Lee et al., 2020), (Romeiko et al., 2020b)
		Subgroup 1.5:Providing rapid estimates of life cycle impacts for non-LCA experts	(Płoszaj-Mazurek et al., 2020), (Feng et al., 2019), (Mao et al., 2019), (D'Amico et al., 2019), (Asif et al., 2019), (Azari et al., 2016), (Song et al., 2017)
		Subgroup 1.6:Assessing uncertainty/sensitivity in life cycle impact estimates	(Feng et al., 2019), (Ziyadi and Al-Qadi, 2019), (Abokersha et al., 2020)
		Subgroup 1.7:Enabling optimization of product performance, cost and environmental impacts	(Azari et al., 2016), (Sharifa and Hammad, 2019)
2	Dimension reduction	Reducing the input parameters for quantifying characterization factors	(Marvuglia et al., 2015a)
3	Feature ranking	Revealing driving factors of life cycle impacts	(Romeiko et al., 2020a), (Zhao et al., 2019)
4	Classification	Estimating environmental releases from chemical use (value prediction)	(Tao et al., 2018)
5	Clustering	Clustering food sectors based on their sustainability performances (cluster)	(Abdella et al., 2020)

estimate energy use and carbon footprint of activated carbon production. Besides these three applications in chemical life cycle inventory conducted by Cheng et al. (2020a); Cheng et al. (2020b) and Liao et al. (2020), two studies by Thilakarathna et al. (2020) and Naseri et al. (2020) applied ML techniques to enable the quantification of life cycle inventories of concrete. (Thilakarathna et al., 2020) compared five ML models to predict the compressive strength of concrete, then calculated the embodied carbon footprint based on the developed ML model. Naseri et al. (2020) used an ANN model to predict the compressive

Table 3
Categories of machine learning applications in various LCA stages.

Group #	MLs in various LCA stages			Relevant articles
1	Life cycle inventory	Subgroup 1.1: Estimating foreground life cycle inventory	Estimating product properties	(Cheng et al., 2020a), (Thilakarathna et al., 2020), (Liao et al., 2020), (Cheng et al., 2020b), (Cornago et al., 2020), (Naseri et al., 2020)
			Estimating environmental releases	(Meng et al., 2019), (Tao et al., 2018), (Nguyen et al., 2019)
		Subgroup 1.2: Estimating overall life cycle inventory		(Sharifa and Hammad, 2019)
2	Life cycle impact assessment	Subgroup 2.1: Estimating life cycle impacts	Agriculture	(Lee et al., 2020), (Romeiko et al., 2020b), (Duprez et al., 2019), (Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017), (Pishgar-Komleh et al., 2020b), (Nabavi-Pelesaraei et al., 2018), (Vlontzosa and Pardalosb, 2017) (Płoszaj-Mazurek et al., 2020), (Feng et al., 2019), (Mao et al., 2019), (D'Amico et al., 2019), (Asif et al., 2019), (Azari et al., 2016)
			Building	
			Other areas such as chemicals and hydrogen	(Zhu et al., 2020), (Ozbilen et al., 2013), (Song et al., 2017) (Hou et al., 2020), (Marvuglia et al., 2015a), (Slapnik et al., 2015a)
		Subgroup 2.2: Estimating characterization factors		(Azari et al., 2016), (Sharifa and Hammad, 2019)
3	Life cycle impact interpretation	Subgroup 3.1:Supply chain optimization		(Romeiko et al., 2020a), (Zhao et al., 2019)
		Subgroup 3.2:Identifying the top influential factors		(Feng et al., 2019), (Ziyadi and Al-Qadi, 2019), (Abokersha et al., 2020)
		Subgroup 3.3: Uncertainty and sensitivity assessment		

(*continued on next page*)

Table 3 (*continued*)

Group #	MLs in various LCA stages	Relevant articles
	Subgroup 3.4: Assessing relationships between sustainability indicators and sustainability impacts	(Abdella et al., 2020)

strength of concrete, which was consequently used as a sustainability criteria along cost, energy consumption and life cycle carbon emissions for designing sustainable concrete mix. Additionally, Cornago et al. (2020) forecasted the hourly power generation for each energy source, which was then fed into the LCA model for estimating carbon emissions of energy mix.

For the second subgroup (subgroup 1.2), only one study was identified. Sharifa and Hammad (2019) developed a surrogate ANN for estimating cost and total life cycle inventory, which includes both foreground and background life cycle inventory.

3.4.2. ML for life cycle impact assessment

Out of the total 40 studies, 22 studies applied ML techniques in the life cycle impact assessment stage. Based upon their focus areas, these 22 studies can be classified into two subgroups. The first subgroup (subgroup 2.1), including most of the studies in this group (19 studies out of 22 studies), adopted ML approaches to estimate life cycle impacts of agricultural products, buildings or energy systems. More than half of these 19 studies focused on life cycle impacts of agriculture. For example, Kaab et al. (2019) employed both artificial neural networks and adaptive neuro fuzzy inference system models for predicting life cycle environmental impacts and output energy of sugarcane production in planted or ratoon farms. Similarly, two studies assessed life cycle environmental impacts of strawberry (Khoshnevisan et al., 2013a) and rice (Khoshnevisan et al., 2014b) with both artificial neural networks and adaptive neuro fuzzy inference system models. Nabavi-Pelesaraei et al. (2018) also used the same modeling approaches to predict energy output and environmental impacts of paddy production. Khanali et al. (2017) predicted the yield and life cycle environmental impacts in tea processing units in Guilan province of Iran with an artificial neural network model. Khoshnevisan et al. (2013a); Khoshnevisan et al. (2014a); Khoshnevisan et al. (2013b); Khoshnevisan et al. (2014b) assessed environmental impacts of potato, tomato and cucumber production with adaptive neuro fuzzy inference system models. Mousavi-Avvala et al. (2017) also used an adaptive neuro fuzzy inference system model to estimate energy use and environmental impacts of oilseed production. While the majority of these studies reported the spatially generic or coarse environmental impacts, two studies utilized ML techniques to compute spatially explicit environmental impacts. Romeiko et al. (2020b) compared six ML methods for predicting spatially explicit annual life cycle impacts of corn production in the US Midwest region from 2000 to 2008. Lee et al. (2020) used a boosted regression tree (BRT) model to project spatially explicit life cycle impacts of corn production in the US Midwest region under future climate scenarios.

The second largest focus area was the built environment, with four studies including Płoszaj-Mazurek et al. (2020), Mao et al. (2019), D'Amico et al. (2019) and Duprez et al. (2019) adopting ML methods to estimate life cycle impacts of buildings. For example, Płoszaj-Mazurek et al. (2020) applied three ML models to quickly estimate total carbon footprint of buildings and to enable the optimal architecture design during the early design phases. Mao et al. (2019) compared regression models for estimating life cycle carbon emissions during the building design stage. D'Amico et al. (2019) developed ANN models for rapidly estimating energy and life cycle environmental impacts of buildings during early design stage. Duprez et al. (2019) developed an ANN model to rapidly predict the global warming potential of new building design alternatives.

In addition to agriculture and building, ML has been utilized to estimate life cycle impacts of chemicals, energy and mining systems. Song et al. (2017) and Zhu et al. (2020) used ANN models to estimate life cycle impacts of chemicals. Ozbilen et al. (2013) built a ANN model to estimate global warming potential, acidification potential, and hydrogen plant efficiency of nuclear-based hydrogen production systems. Pishgar-Komleh et al. (2020b) applied ANN models to calculate life cycle energy use, greenhouse gas emission, and economic costs, which were then fed into a multi-objective optimization model. Asif et al. (2019) developed an ANN model to estimate the carbon footprint of a mining system.

The second subgroup of studies (subgroup 2.2), including Hou et al. (2020), Marvuglia et al. (2015a), and Slapnik et al. (2015a), utilized ML to predict characterization factors (CFs) of chemicals. Hou et al. (2020) developed ML models to estimate ecotoxicity hazardous concentrations 50 % (HC50) in USEtox to calculate chemicals' CFs. Marvuglia et al. (2015a) carried out a thorough exploratory data analysis to identify and select input parameters for predicting fate factors and intake fractions of chemicals. Slapnik et al. (2015a) computed chemical CFs in Slovenian and compared them with other European CFs.

3.4.3. ML for life cycle interpretation

ML techniques were applied to support life cycle interpretation in four different manners. First, two studies utilized ML models to solve optimization problems with the goal of minimizing life cycle environmental impacts and economic costs. For example, Sharifa and Hammad (2019) developed surrogate ANN for selecting near-optimal building energy renovation methods, which considers minimizing energy consumption, cost and environmental impacts as the multi-objectives. Azari et al. (2016) used a hybrid ANN and a genetic algorithm approach to enable optimization of building designs.

Second, two LCA studies used ML techniques to identify patterns and drivers of life cycle impacts. Romeiko et al. (2020b) used BRT to identify key contributors affecting the spatially and temporally explicit life cycle impacts of soybean production. Zhao et al. (2019) used random forest to identify the drivers for life cycle carbon footprints of herdsmen in the typical steppe region of Inner Mongolia, China.

Third, ML models were used to understand uncertainty and sensitivity of life cycle impacts in three studies. Feng et al. (2019) used an integrated fuzzy C-means clustering and extreme learning machine to assess the uncertainty of buildings' environmental impacts in early design stages. Ziyadi and Al-Qadi (2019) built an ANN surrogate model in conjunction with interval, Bayesian and model correction analysis methods to estimate input, parameter and model uncertainty. Abokersha et al. (2020) identified optimal integration of solar assisted district heating in urban communities by using ML incorporating global sensitivity analyses.

Fourth, ML models were used for classification and assessing relationships between indicators and sustainability impacts. Abdella et al. (2020) used a centroid-based clustering approach to classify the food industries, and used a logistic regression model to assess the relationship between the sustainability indicators and the total impacts of food industries.

3.5. Types of ML models in LCA studies

As shown in Table 4, the unsupervised machine learning approaches applied in LCA studies included linear models, tree-based models, neural networks, nature-inspired optimization algorithms, distance-based models and deep learning approaches. The linear models, such as linear regression, logistic regression, partial least regression and Gaussian process regression models, describe a continuous response variable as a function of one or more predictor variables. Tree-based models are a class of nonparametric algorithms that partition the feature space into a number of non-overlapping regions with similar

Table 4
Machine learning approaches used in the reviewed studies.

ML category	ML subcategory	ML algorithm	Studies
Supervised learning	Linear models	Linear regression	(Thilakarathna et al., 2020), (Romeiko et al., 2020b), (Cheng et al., 2020b), (Naseri et al., 2020), (Meng et al., 2019)
		Logistic regression	(Slapnik et al., 2015a)
		Partial least squares regression	(Marvuglia et al., 2015a)
		Gaussian process regression	(Thilakarathna et al., 2020)
	Tree-based models	Decision tree	(Thilakarathna et al., 2020), (Cheng et al., 2020b)
		Random forests	(Cheng et al., 2020a), (Hou et al., 2020),(Cheng et al., 2020b), (Mao et al., 2019), (Zhao et al., 2019)
		Adaptive boosting	(Hou et al., 2020)
		Gradient boosting	(Hou et al., 2020), (Płoszaj-Mazurek et al., 2020), (Romeiko et al., 2020a), (Lee et al., 2020), (Romeiko et al., 2020b)
		Extreme gradient boosting	(Romeiko et al., 2020b)
	Neural Networks	Artificial Neural Network	(Zhu et al., 2020), (Thilakarathna et al., 2020), (Romeiko et al., 2020b), (Hou et al., 2020), (Liao et al., 2020), (Sharifa and Hammad, 2019), (Ziyadi and Al-Qadi, 2019), (Cornago et al., 2020), (Naseri et al., 2020), (Abokersha et al., 2020), (Mao et al., 2019), (D'Amico et al., 2019),(Duprez et al., 2019), (Tao et al., 2018), (Nguyen et al., 2019), (Asif et al., 2019), (Azari et al., 2016), (Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Nabavi-Pelesaraei et al., 2018), (Ozbilen et al., 2013), (Pishgar-Komleh et al., 2020a), (Song et al., 2017), (Vlontzosa and Pardalosb, 2017)
		Adaptive neuro-fuzzy inference systems	(Kaab et al., 2019), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017), (Nabavi-Pelesaraei et al., 2018)
	Nature-inspired optimization algorithm	Water cycle algorithm	(Naseri et al., 2020)
		Soccer league competition algorithm	(Naseri et al., 2020)
	Distance-based Models	Support Vector Machine	(Thilakarathna et al., 2020), (Hou et al., 2020), (Naseri et al., 2020), (Mao et al., 2019)
		K nearest neighbor	(Hou et al., 2020)
	Deep Learning	Convolutional neural network	(Płoszaj-Mazurek et al., 2020)
Unsupervised learning	Clustering	K-means	(Abdella et al., 2020)
		Fuzzy C-means	(Feng et al., 2019)
	Feature Extraction	Principle component analysis	(Mao et al., 2019)

response values using a set of splitting rules. Nature-inspired algorithms are a set of novel problem-solving methodologies and approaches derived from natural processes. Distance-based Models classify queries by computing distances between these queries and a number of internally stored exemplars. Neural networks mimic the human brain in structure. Simple neural networks consist of an input, hidden, and output layer. Deep learning systems are neural networks consisting of several hidden layers arranged for convolution or recurrence. Supervised machine learning approaches applied in LCA studies included clustering and feature extraction. Cluster analysis is the art of finding groups in data, which is a branch of pattern recognition. Principle component analysis is a versatile statistical method for reducing a cases-by-variables data table to its essential features, called principle components. Principle component analysis is frequently used for reducing dimension and extracting features.

ANN was the most frequently applied ML approach, appearing in 26 studies (Table 4). Adaptive neuro-fuzzy inference systems, which was used by six studies, ranked as the second most frequently applied ML approach. Linear regression and random forest were adopted by five studies, respectively. Gradient boosting regression and support vector machine were utilized by four studies, respectively. Decision tree was employed by two studies. Each of the rest ML approaches only appeared in one study, respectively.

Most studies used a single ML method. Only 14 out of 40 used more than one ML method. Among these 14 studies, 10 studies compared various ML methods to determine the most accurate and rapid ML. Overall, these comparative studies suggest distinct ML methods perform the best for varied studies. Four studies, including Khoshnevisan et al. (2013a); Khoshnevisan et al. (2014a); Khoshnevisan et al. (2013b); Khoshnevisan et al. (2014b) and Mousavi-Avvala et al. (2017), found Adaptive neuro-fuzzy inference systems (ANFIS) had the highest predictive accuracy. In contrast, Thilakarathna et al. (2020), Duprez et al. (2019), Kaab et al. (2019) and Pishgar-Komleh et al. (2020b) reported that ANN provided the highest predictive accuracy instead. Two studies found random forest outperformed other models. For example, (Hou et al., 2020) found that random forest performed best among K nearest neighbor, support vector machine, neural network, random forest, adaptive boosting, and gradient boosting machine. Consistent with Hou et al. (2020), Cheng et al. (2020b) also found that random forest gave the best performance. In contrast, Romeiko et al. (2020b) identified gradient boosting regression tree as the most accurate and rapid option, compared with linter regression, support vector machine, ANN, random forest, and extreme gradient boosting. (Naseri et al., 2020) found that the water cycle algorithm performed better than the other five ML methods. (Mao et al., 2019) found that support vector machine ranked as the best performing model. Additionally, Thilakarathna et al. (2020), Romeiko et al. (2020b), Cheng et al. (2020b) and Naseri et al. (2020)

found that the linear model showed the poorest fit for the data.

3.6. Sources and sample sizes of datasets for ML models in LCA studies

The largest cluster of studies utilized LCA simulations as training sets for ML. Romeiko et al. (2020a) and Lee et al. (2020) used outputs from process-based LCA modeling along with climate, soil and farming practices information. D'Amico et al. (2019) and Feng et al. (2019) generated training datasets by coupling building information and process-based LCA modeling. (Zhao et al., 2019) used process-based LCA modeling datasets based on a survey. The publications by Kaab et al. (2019), Khanali et al. (2017), Mousavi-Avvala et al. (2017), and Nabavi-Pelesaraei et al. (2018), used life cycle modeling results along with survey information as training datasets. Pishgar-Komleh et al. (2020b) trained the ANN model with process-based LCA modeling results, which used questionnaires to compile life cycle inventory. Azari et al. (2016) and Ozbilen et al. (2013) used process-based LCA modeling results as training datasets. Additionally, Abdella et al. (2020) used outputs from input-output LCA modeling.

Besides using LCA simulated datasets, sector-specific, sensitivity and optimization models also provided training datasets for ML techniques. For example, Płoszaj-Mazurek et al. (2020) used Grasshopper scripts to generate training datasets for ML in the building sector. Abokersha et al. (2020) used the TRNSYS model simulations as training datasets in the energy sector. The training datasets are from the process-based DayCent agroecosystem simulation model in Nguyen et al. (2019)'s study. (Duprez et al., 2019) used datasets generated from sensitivity analysis for ML model training. Sharifa and Hammad (2019) generated training datasets from a multi-objective optimization model.

Existing LCA databases were frequently used as major sources of datasets for training ML models, as demonstrated by six studies, including Zhu et al. (2020), Hou et al. (2020), Marvuglia et al. (2015a), Slapnik et al. (2015a), Song et al. (2017) and Tao et al. (2018). For example, Zhu et al. (2020) supplied ecoinvent v3.5 database and a ReCiPe model to provide training datasets. Hou et al. (2020) and Marvuglia et al. (2015a) used the USEtox v2.11 database. Slapnik et al. (2015a) used the characterization factor database from the ReCiPe 1.08 model. Song et al. (2017) relied on the ecoinvent database. Tao et al. (2018) obtained training datasets from the European Union's specific environmental release categories, which included chemical release factors to environmental compartments (indoor air, outdoor air, wastewater and soil) for chemicals in different products.

The datasets used for training ML models can also originate from literature and lab/field experiments. Five studies, including Thilakarathna et al. (2020), Liao et al. (2020), Cheng et al. (2020a), Ziyadi and Al-Qadi (2019), and Naseri et al. (2020), used literature reported values. Moreover, Cheng et al. (2020b) relied on both literature values and lab experiments. Meng et al. (2019) used a combination of literature, field testing, and public datasets. Furthermore, Asif et al. (2019) collected field datasets from different equipment and mining activities. Some studies, including Vlontzosa and Pardalosb (2017), didn't specify the training datasets.

The sample size of the datasets used for these ML training ranged from 64 to 21,656. The median value of the sample size is 538. Only five studies had over 5000 samples. For example, Romeiko et al. (2020a); Romeiko et al. (2020b) had around 5000 samples for soybean LCAs. Romeiko et al. (2020b) and Lee et al. (2020) used around 8000 samples for corn LCAs. Duprez et al. (2019) used 5000, 10,000 and 15,000 samples for three different ML models, respectively. Nguyen et al. (2019) reported the largest sample size of 21,656. Overall, the majority of LCA studies (over 70 %) relied on a small number of datasets (less than 1500) for training the ML models.

3.7. ML model training and evaluation

3.7.1. ML model training and evaluation methods

Before applying a ML model to generate LCI, conduct impact assessment or assist with life cycle interpretation, it is critical to ensure the ML model can provide satisfactory prediction. Two model training and evaluation strategies are used in the reviewed studies (Table 5 and Fig. S1). The first strategy is the holdout method, in which the available datasets are divided into two groups: training and testing. A ML model is trained with the training dataset and tested on the testing dataset. The second strategy is cross-validation, which divides the entire dataset into two groups: (1) training and validation dataset and (2) testing dataset. The training and validation dataset is used to deterring the optimal ML model structure and parameter settings, while the testing dataset provides independent assessment of ML performance to decide if the training ML model is capable to provide satisfactory results.

One-fold cross validation was the most popular approach, which was used by twelve studies. The hold-out method was the second most popular approach, which was employed by nine studies. Following one-fold cross validation and hold-out methods, ten-fold cross validation was the third most popular approach, which was adopted by seven studies. Apart from ten-fold cross validation, other multi-folds validation approaches were also used, including three-fold and five-fold validation approaches. Three-fold validation approach was used by two studies, and five-fold cross validation was used by only one study. Moreover, three studies employed either Monte Carlo cross-validation or leave-one out validation approaches, which are variants of multi-fold cross validation approaches. For example, Meng et al. (2019) used the Monte Carlo cross-validation, in which a prescribed proportion of the training and validation dataset is randomly selected as training dataset and the rest is used for validation. Furthermore, two studies used the cross-validation method, but didn't explicitly mention the details such as the number of folds. Additionally, five studies didn't provide details

Table 5
Training datasets used for machine learning in the reviewed studies.

Groups	Training databases	Studies
1	LCA modeling outputs	(Romeiko et al., 2020a), (Lee et al., 2020), (Romeiko et al., 2020b), (D'Amico et al., 2019), (Feng et al., 2019), (Zhao et al., 2019) (Kaab et al., 2019), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Khanali et al., 2017), (Mousavi-Avvala et al., 2017), (Nabavi-Pelesaraei et al., 2018) (Pishgar-Komleh et al., 2020b), (Azari et al., 2016) (Ozbilen et al., 2013), (Abdella et al., 2020)
2	Other modeling outputs (i.e. sector-specific, sensitivity, optimization modeling outputs)	(Płoszaj-Mazurek et al., 2020), (Abokersha et al., 2020), (Nguyen et al., 2019), (Duprez et al., 2019), (Sharifa and Hammad, 2019)
3	Existing LCA databases (i.e. ecoinvent)	(Zhu et al., 2020), (Hou et al., 2020), (Marvuglia et al., 2015a), (Slapnik et al., 2015a), (Song et al., 2017), (Tao et al., 2018), (Zhu et al., 2020), (Hou et al., 2020), (Marvuglia et al., 2015a), (Slapnik et al., 2015a), (Song et al., 2017), (Tao et al., 2018)
4	Literature and lab experiment datasets	(Thilakarathna et al., 2020), (Liao et al., 2020), (Cheng et al., 2020a), (Ziyadi and Al-Qadi, 2019), (Naseri et al., 2020), (Cheng et al., 2020b), (Meng et al., 2019), (Asif et al., 2019), (Vlontzosa and Pardalosb, 2017)

regarding model validation, so it was impossible to classify these five studies.

The splits between training, validation, and testing datasets were often sampled randomly, and the corresponding shares varied across studies. The studies using the hold-out methods divided total datasets for training and validation purposes, whose training datasets represented 10 % to 85 % of total datasets. For example, (Abdella et al., 2020) had 65 % and 35 % for training and validation, respectively. Two studies, authored by Sharifa and Hammad (2019) and Azari et al. (2016) increased the training ratio to 70 %. Mousavi-Avvala et al. (2017) used 80 % and 20 % for training and validation ratios, respectively. D'Amico et al. (2019) and Tao et al. (2018) had the largest training ratio of 85 % among these hold-out studies. Additionally, Nguyen et al. (2019) varied the proportion from 10 % to 70 % with an increment of 20 %.

The studies using cross-validation approaches divided total datasets into training, validation and testing subsets. For the studies using cross-validation approaches, the most frequently used splitting ratio was 70 %, 15 % and 15 % for training, validation and testing, respectively. Nine studies used this popular splitting ratio including seven one-fold cross validation studies, and one three-fold cross validation and one study which didn't detail folds of cross-validation. Other cross-validation studies varied this popular splitting ratio by ±15 %. Song et al. (2017) allocated the highest fraction of the dataset for training (ca. 85 %) and the least fraction for testing (ca. 5 %). Mao et al. (2019) had the lowest fraction of the datasets for training (ca. 60 %) and highest fraction for testing (ca. 20 %). In general, the reported training and validation datasets combined represent 70 %–90 % of the total, and the testing dataset represents less or equal to 20 %. Additionally, it is worth noting that fourteen of the total forty reviewed studies did not report the split between training, validation and testing datasets.

3.7.2. ML model training and evaluation metrics

A wide range of metrics have been used to evaluate ML model performance (Table 6). As most of the ML applications aim to map an array of inputs to continuous response variables, metrics that measure the correlation or difference between predicted and observed response variables are most often used. First, the coefficient of determination (R^2), root mean square error (RMSE) and Mean absolute percentage error were found to be the three most frequently used metrics. R^2 ranked as the most frequently used metric, which appeared in 31 studies. RMSE used by 18 studies was the second most frequently used metric (Table 7). Mean absolute percentage error adopted by 12 studies was the third most frequently used metric. Moreover, the derivatives of these three most popular metrics were also used to evaluate the predictive accuracy, including mean square error, mean absolute error (MAE), correlation coefficient (R), coefficient of variance, root relative square error, normalized root mean square error (NRMSE), and percentage of data whose mean absolute percentage error is less than 30 % (E30). Slightly different from the aforementioned metrics, Feng et al. (2019) examined whether the observations fall within the ML model predicted 95 % interval. Abdella et al. (2020) used Akaike information criterion to evaluate the accuracy of predicted categorical membership. Furthermore, different metrics were used to evaluate model performance in classification problems. For example, Tao et al. (2018) used Precision, recall, and F1 score to assess the accuracy of distribution (in percent) of chemicals related to different endpoints. Last, two studies conducted by Ziyadi and Al-Qadi (2019) and Zhao et al. (2019) didn't mention the performance metrics.

The number of evaluation metrics used in these studies varied from one to six. The majority of studies used one metric (15 studies, 37.5 %) or two metrics (13 studies, 32.5 %). For example, studies conducted by Płoszaj-Mazurek et al. (2020), Romeiko et al. (2020a), Lee et al. (2020), Liao et al. (2020), Sharifa and Hammad (2019), Marvuglia et al. (2015a), Cornago et al. (2020), and Azari et al. (2016) used either R^2 or MSE. Eight studies used three metrics. Additionally, studies performed by Naseri et al. (2020) and Mao et al. (2019) used five or more metrics, including coefficient of determination (R^2), correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), normalized root mean square error (NRMSE), coefficient of variance of the RMSE, and percentage of data whose mean absolute percentage error is less than 30 % (E30).

Table 6
Model training and validation approaches employed by different studies.

Group #	Model training and validation approaches		Studies
1	Hold-out		(Abdella et al., 2020), (Liao et al., 2020), (Sharifa and Hammad, 2019), (D'Amico et al., 2019), (Tao et al., 2018), (Nguyen et al., 2019), (Azari et al., 2016), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017)
2	Cross-validation	One fold	(Zhu et al., 2020), (Feng et al., 2019), (Cornago et al., 2020), (Asif et al., 2019), (Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2013b), (Pishgar-Komleh et al., 2020b), (Pishgar-Komleh et al., 2020a), (Song et al., 2017), (Vlontzosa and Pardalosb, 2017)
		three folds	(Romeiko et al., 2020b), (Duprez et al., 2019)
		five folds	(Hou et al., 2020)
		ten folds	(Cheng et al., 2020a), (Thilakarathna et al., 2020), (Romeiko et al., 2020a), (Lee et al., 2020), (Cheng et al., 2020b), (Slapnik et al., 2015a), (Mao et al., 2019)
		Monte Carlo cross-validation	(Meng et al., 2019)
		leave-one out cross-validation	(Marvuglia et al., 2015a), (Naseri et al., 2020)
		Mentioned cross-validation, but no specific details	(Abokersha et al., 2020), (Khoshnevisan et al., 2013a)
3	No details regarding model validation		(Płoszaj-Mazurek et al., 2020), (Ziyadi and Al-Qadi, 2019), (Zhao et al., 2019), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b)

The reported ML model performance varies widely across studies or application purposes of the same ML model. For example, (Romeiko et al., 2020a) achieved R^2 from 0.64 to 0.78 for predicting life cycle global warming, eutrophication and acidification impacts. (Hou et al., 2020) achieved R^2 of 0.63 for predicting hazardous concentrations for ecotoxicity. Liao et al. (2020) reached 0.971 for predicting total AC yield. Even with the same inputs, the ML model trained by Mousavi-Avvala et al. (2017) had RMSE and MAE larger than 10 when applied to predict output energy, but had RMSE and MAE less than 1 for predicting the benefit to cost ratio. These results clearly show the substantial variability in the cases applying ML models to support LCA.

4. Discussion

4.1. Merits of applying ML in LCA

ML models present unique merits such as capabilities of enabling accurate prediction, discovering complex patterns, and efficiently analyzing large datasets. First, the majority of the reviewed LCA studies relied on ML's high predictive accuracy to fill in the missing values for life cycle inventories or impacts (Tables 2 and 3). Since the data characterizing these emerging products/technologies often doesn't exist, building predictive models based upon characteristics of existing

Table 7
Model performances and associated metrics.

Groups	Model performance metrics	Studies
Regression Related metrics	Coefficient of determination (R^2)	(Cheng et al., 2020a), (Płoszaj-Mazurek et al., 2020), (Zhu et al., 2020), (Thilakarathna et al., 2020), (Romeiko et al., 2020a), (Lee et al., 2020), (Abdella et al., 2020), (Romeiko et al., 2020b), (Hou et al., 2020), (Liao et al., 2020), (Cheng et al., 2020b), (Marvuglia et al., 2015a), (Naseri et al., 2020), (Abokersha et al., 2020), (Mao et al., 2019), (D'Amico et al., 2019), (Meng et al., 2019), (Duprez et al., 2019), (Nguyen et al., 2019), (Asif et al., 2019), (Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017), (Pishgar-Komleh et al., 2020b), (Ozbilen et al., 2013), (Pishgar-Komleh et al., 2020a), (Song et al., 2017), (Vlontzosa and Pardalosb, 2017)
	Correlation coefficient (R)	(Romeiko et al., 2020b), (Slapnik et al., 2015a), (Naseri et al., 2020), (Khoshnevisan et al., 2014a), (Vlontzosa and Pardalosb, 2017)
	Adjusted R-square (R-adj)	(Abokersha et al., 2020)
	Root mean square error (RMSE)	(Cheng et al., 2020a), (Zhu et al., 2020), (Thilakarathna et al., 2020), (Hou et al., 2020), (Cheng et al., 2020b), (Naseri et al., 2020), (D'Amico et al., 2019), (Meng et al., 2019), (Duprez et al., 2019), (Nguyen et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2013b), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Mousavi-Avvala et al., 2017), (Pishgar-Komleh et al., 2020b), (Pishgar-Komleh et al., 2020a)
	Mean absolute percentage error (MAPE)	(Cornago et al., 2020), (Abokersha et al., 2020), (Mao et al., 2019), (Meng et al., 2019), (Kaab et al., 2019), (Khanali et al., 2017), (Khoshnevisan et al., 2013a), (Khoshnevisan et al., 2014a), (Khoshnevisan et al., 2014b), (Pishgar-Komleh et al., 2020b), (Pishgar-Komleh et al., 2020a), (Song et al., 2017)
	Mean square error (MSE)	(Romeiko et al., 2020b), d(Sharifa and Hammad, 2019), (Naseri et al., 2020), (Asif et al., 2019), (Azari et al., 2016), (Vlontzosa and Pardalosb, 2017)
	Mean absolute error (MAE)	(Thilakarathna et al., 2020), (Naseri et al., 2020), (Mao et al., 2019), (Khoshnevisan et al., 2013b), (Mousavi-Avvala et al., 2017)
	Coefficient of variation (CV)	(Abokersha et al., 2020), (Mao et al., 2019)
	Akaike information criterion (AIC)	(Abdella et al., 2020)
	Whether the observation fall within the ML model predicted 95 % interval	(Feng et al., 2019)
	Percentage of data whose mean absolute percentage error is less than 30 % (E30)	(Naseri et al., 2020)
	Root relative square error (RRSE)	(Slapnik et al., 2015a)
	Normalized root mean square error (NRMSE)	(Mao et al., 2019)
Classification related metrics	Precision	(Tao et al., 2018)
	recall	(Tao et al., 2018)
	F1 score	(Tao et al., 2018)

products provides an alternative means for generating life cycle inventory. For example, Cheng et al. (2020a) applied ML to estimate biochar characteristics in order to compile a life cycle inventory. Meng et al. (2019) used ML approach to fill in data gaps for life cycle inventory of dual fuel technology. Meanwhile, the merit of high predictive accuracy also led to success of estimating life cycle impacts of alternative building designs and future agricultural production. For example, Płoszaj-Mazurek et al. (2020) and D'Amico et al. (2019) predicted environmental impacts of buildings with various design characteristics. Lee et al. (2020) used historical life cycle impacts to predict life cycle impacts of corn under future climate scenarios.

Second, ML provides novel insights towards drivers and patterns of environmental performances, which aids decision makers in forming solutions capable of improving environmental performances. For example, Hou et al. (2020) identified key influential factors for chemical toxicity. Romeiko et al. (2020b) ranked the importance of soil, weather and farming practices for spatially and temporally explicit life cycle impacts. Abdella et al. (2020) assessed quantitative and qualitative relationships between the sustainability indicators and the total sustainability impact of food industries. These findings reveal the underlying drivers causing environmental damages and assist in designing targeted intervention strategies capable of mitigating environmental damages.

Finally, compared with traditional process-based models, ML models showed faster execution, and flexible integration into other simulation platforms (i.e. optimization platforms). Such advantages allow ML models to rapidly complete a high number of simulation runs, therefore making ML models affordable for a range of computationally intensive tasks such as optimization, uncertainty and sensitivity assessment. Several studies used ML as surrogate models to generate life cycle impacts fed into optimization models, since ML surrogate models can rapidly provide accurate estimates, and be easily integrated into optimization platforms. For example, Sharifa and Hammad (2019) and Azari et al. (2016) utilized ML to support optimization tasks. Nguyen et al. (2019) used ML and optimization approaches to suggest management options for agricultural landscape. Similarly, uncertainty and sensitivity assessments are computationally intensive jobs too. ML surrogate models served as ideal approaches to efficiently and accurately carry out many simulations for uncertainty and sensitivity assessment, which can be difficult for traditional process-based models.

4.2. Challenges of applying ML in LCA

4.2.1. Lack of data

Lack of training datasets presents a major bottleneck for applying ML in LCA. Most of these reviewed studies used a small sample size, which is less than 1500 datasets for training ML models. Usually, a larger sample size rewards a more representative and accurate ML model. A few studies with over 5000 samples relied on simulation data generated by other models, often process-based models, to provide training datasets.

The computation burden for generating large training datasets could restrain the adoption of ML models, especially when the computing infrastructure is not available for expensive process-based modeling simulations. Additionally, although ecoinvent and other LCA databases are growing, their sizes remain smaller than the ML's databases in other disciplines such as earth and public health sciences. Ecoinvent v3.9, one of the largest LCI commercial databases, currently contains more than 18,000 industrial or agricultural processes covering around 3300 distinct products and is about 350 MB (The Ecoinvent Association). In comparison, the accumulated volume of remote sensing data obtained from satellite, airborne, unmanned aerial vehicles and ground-based instruments by 2020 has reached ~1.3 EB, and this number will keep increasing with the expansion of observation capacities and spatial temporal resolutions (Li et al., 2023). In healthcare, it was estimated that a single patient generates close to 80 megabytes/year in imaging and electronic medical record data (Suter-Crazzolara, 2018). Overall, the availability of high-quality training datasets remains a challenge for applying ML in LCAs.

4.2.2. Lack of detailed description about model selection and evaluation

The discussion of ML selection is limited. Most of these studies didn't mention sufficient details about the algorithm's selection and implementation. While most of these surveyed studies report results for a single ML algorithm, the readers would benefit from learning why the specific algorithm was chosen and how the choice was made.

There is also a lack of guidelines for model training. It's worth noting that tradeoffs between computational efficiency and accuracy exist among the choices of model training approaches. The holdout method is simpler than cross-validation and only requires one iteration of model training, making it computationally inexpensive relative to cross-validation. However, the holdout method could yield highly variable testing results, as the division of the dataset into training and testing samples are arbitrary. Although the one-fold cross-validation splits the entire dataset into three groups, which include the testing dataset that was not used for training and validation model performance, it shares the same weakness resulting from not fully using the data from the training and validating dataset. The computational efficiency of the hold-out and one-fold cross-validation is desired when the sample data size is large and a long time is required for model training. However, given that the sample size of the reviewed studies is generally less than 5000, the use of multi-fold cross-validation is preferred to fully employ the information from the available data. Despite the number of folds of cross-validation that could influence ML training and testing performance (Zhang et al., 2009), most of the studies did not examine the effects of the number of folds. It is suggested conducting multiple trials with cross-validation with different trials and reporting such information to help future applications to determine optimal number of folds of cross-validation to achieve reliable ML model predictions.

Despite the studies reporting a wide range of metrics, there is a lack of widely accepted criteria for determining whether a ML performance is satisfactory for regression problems. As values for RMSE, MSE, and MAE are highly dependent on the units of the response variables, it is difficult to directly compare those metrics across studies or set a uniform standard to determine if the performance of a ML model is satisfactory. As such, it is suggested that dimensionless metrics such as R, R^2, NRMSE, MRE, MAPE are more feasible for comparing ML performance across studies/applications. Note that, the use of dimensionless metrics does not guarantee a fair comparison between ML performance across problems, as different application cases have different inputs and require different levels of accuracy. Therefore, likely the criteria used to determine if a ML model performs satisfactorily vary case by case.

4.2.3. Model uncertainty

Most of the reviewed studies didn't consider the uncertainty embedded in ML training datasets and models. Many studies used ML as a surrogate model for prediction, optimization and uncertainty assessment. While these studies are valuable contributions, ML introduces additional uncertainty to the existing LCA model structure due to uncertainty of training datasets and diverse model choices (i.e. algorithms and training/validation procedures). Assessing the uncertainty of ML models is challenging and only was briefly discussed in two of these studies. For example, Romeiko et al. (2020b) focused on the uncertainty introduced by training algorithms. Nguyen et al. (2019) mentioned the uncertainty of ML algorithms, but didn't provide any quantitative assessment. None of these studies comprehensively analyzed the uncertainty associated with data, algorithm and model structure.

4.3. Future research directions in fusing ML and LCA

4.3.1. Build or access high-quality large datasets

Both ML and LCA approaches require extensive amounts of data for establishing the models. The lack of datasets is a major bottleneck for applying ML in LCA. It's worth noting that the LCA communities have been developing large databases under various initiatives. For example, The Life Cycle Initiative hosted by the United Nation Environmental Programme built the Global LCA Data Access network (GLAD), which is the largest directory of LCA datasets from independent LCA database providers around the world (UN Environment Programme, 2023). Although GLAD doesn't directly host databases, it supports LCA data accessibility by redirecting the users to the data providers and enables data interoperability. Federal LCA commons led by US governmental labs is positioned to serve as a central point of access to a collection of data repositories for LCA studies at no cost (Federal LCA Commons, 2022). European Commission's Life Cycle Data Network provides a globally usable infrastructure for the publication of quality assured LCA dataset from different organizations (European Commission, 2023). Despite these valuable efforts, further expansion of databases will be necessary to support the integration of ML and LCA.

Meanwhile, the big datasets generated by various industries provide new opportunities for ML and LCA. For example, the newly generated big data in agriculture can lead to innovative integration of ML and LCA in several stages of food supply chains. During the agricultural production stage, the adoption of sensors at farms provides large volume spatially and temporally explicit soil, climate, crop and emission datasets (Wolfert et al., 2017), which can serve as data foundation for the integrated ML and LCA for understanding the spatial and temporal heterogeneity of environmental impacts and designing mitigating strategies. During the transportation stage, the use of sensor monitoring humidity, temperature, light, and microbiological and product quality in transit is useful for the food industry in rescheduling, recalling, or redesigning supply chain logistics (Bhutta and Ahmad, 2021; Maksimović et al., 2015). During the food consumption stage, integration of AI and LCA may analyze food consumption behaviors and perceptions and elucidate environmental health impacts of food consumption at multiple scales (Samad et al., 2022). Overall, harnessing the existing big datasets in various industries and expanding the LCA databases for integrated AI and LCA may be fruitful research directions.

4.3.2. Robust modeling selection

Comparisons between two or more models are recommended in order to determine the most suitable ML models. Multiple-fold cross-validation is preferable in order to make full use of the available datasets for model training and validation, particularly when the sample size isn't too large. Additionally, stratified cross-validation is recommended to counteract the imbalance of available datasets. It is worth noting that imbalanced data is not unusual in ML. The distribution of the available data may not represent well the full spectrum of the true data. In those cases, the stratified cross-validation (Diamantidis et al., 2000; Zeng and Martinez, 2000) could ensure that each of the multiple folds data groups contains comparable fractions of the data, which belong to different target classes for classification problems, or ranges of response variables

for prediction problems.

Whereas it is difficult to establish a uniform guideline to assess ML model performance, it is still useful to have an accepted model evaluation reporting procedures (e.g., which metrics should be used and what are the cut-off thresholds for satisfactory performance). Such information provides a benchmark regarding the general expected performance of a ML model for a certain type of problems. This information will also help ML model users to decide if additional time and resources should be invested to train different ML models or collecting new data to improve model performance. As different metrics measure different aspects of the goodness of fit, it is suggested using multiple metrics, instead of only one. In addition to regression problems, for the studies that address distribution, uncertainty, and classification problems (Abdella et al., 2020; Feng et al., 2019; Tao et al., 2018), it is also suggested to use and report a common set of dimensionless metrics that could be compared across studies or application purposes.

4.3.3. More detailed information about ML applications and uncertainty evaluation

Future studies are recommended to report necessary details about ML model selection, evaluation and uncertainty. Model selection and evaluation are important aspects of ensuring reliable model outcomes to aid in solving real world problems. However, they often receive less attention than the ML algorithms themselves. It is important to consider the breadth of techniques and evaluate multiple techniques in order to select the right approach for the applications.

Additionally, caution should be paid to assess model uncertainty when applying ML to LCA. Understanding model uncertainty is necessary for understanding potential biases of modeling results and avoiding misinterpreting modeling results for decision making. However, most of the reviewed studies haven't evaluated model uncertainty. Meanwhile, many new ML methods have been developed to model uncertainty such as Bayesian deep learning, combination of fuzzy logic with neural networks, rough set theory and imprecise probability (Abdar et al., 2021). We recommend future studies to consider these methods to evaluate model uncertainty in integrated ML and LCA models.

4.3.4. Exploring new ML models

Future LCA studies should consider integration with new ML models. Specifically, deep learning may open new territories in LCA applications. The existing work has shown the initial success of integrating ML methods to improve life cycle inventory, impact assessment and interpretation. Most of these studies utilized ANN. In recent years, however, deep learning has elevated the potential for learning with ANN to new heights. Thus, deep learning methods may also be very fruitful within LCA. Deep learning is based on standard ANN algorithms but utilizes much larger and deeper networks trained on big datasets. The deep learning enables discovering the intricate structure in large datasets and disentangling complex features. Deep learning methods have been highly effective in areas such as image classification, speech recognition, anomalies detection, new material discovery, and other complex problems. Integrating LCA with deep learning may enable incorporating nontraditional data sources such as images to life cycle inventory and may aid in discovering new patterns of life cycle impacts.

5. Conclusions

This review analyzed forty peer-reviewed articles that reported the joint use of ML and LCA for quantitative sustainability assessment. ML has aided in advancing life cycle inventory, life cycle impact assessment and interpretation due to its capabilities of accurately predicting values, discovering hidden patterns and improving computational efficiency.

This review also revealed challenges of applying ML in LCAs. First, although ML training datasets were derived from diverse sources (primarily from model simulations), the size of training datasets is relatively small. Moreover, while a variety of ML models were used, there is still a lack of detailed model description and established guidelines regarding which metrics and the standards should be used to judge if the performance of a ML model is satisfactory. Furthermore, uncertainty analysis associated with ML predictions are rarely analyzed.

These findings led to the following suggestions including: (1) continuous data collection and compilation for supporting reliable ML and LCA modeling; (2) reporting sufficient details regarding the selection criteria for ML models and presenting model uncertainty analysis; and (3) exploring new ML models in LCA studies; and (4) deep integration of ML into various LCA stages to solve the complex environmental sustainability challenges.

CRediT authorship contribution statement

XXR: conception or design of the work
XXR: data collection
XXR, XZ and YP: data analyses and interpretation
XXR and XZ: drafting the manuscript
YP, FG, MX, SL and CB: critical revision
XXR and XZ: approval of the version to be submitted for publication
XXR, XZ and SL: project administration and funds procurement

CRediT authorship contribution statement

Xiaobo Xue Romeiko: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xuesong Zhang:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Validation, Writing – original draft, Writing – review & editing. **Yulei Pang:** Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Feng Gao:** Writing – review & editing. **Ming Xu:** Writing – review & editing. **Shao Lin:** Writing – review & editing. **Callie Babbitt:** 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

No data was used for the research described in the article.

Acknowledgements

This research was supported mainly by the U.S. Department of Agriculture, Agricultural Research Service (Cooperative Agreement 58-8042-1-040). USDA is an equal opportunity provider and employer. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. This research was also supported partially by NSF Award 2115405, NIA Award R01AG070949, and NASA 22-CMS22-0027.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.168969.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al., 2021. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf. Fusion 76, 243–297.

Abdella, G.M., Kucukvar, M., Nuri Ciha, O., Al-Yafay, H.M., Bulak, M.E., 2020. Sustainability assessment and modeling based on supervised machine learning techniques: the case for food consumption. J. Clean. Prod. 251, 119661.
Abokersha, M.H., Vallèsa, M., Cabezab, L.F., Boera, D., 2020. A framework for the optimal integration of solar assisted district heating in different urban sized communities: a robust machine learning approach incorporating global sensitivity analysis. Appl. Energy 267, 114903.
Asif, Z., Chen, Z., Zhu, Z.H., 2019. An integrated life cycle inventory and artificial neural network model for mining air pollution management. Int. J. Environ. Sci. Technol. 16, 1847–1856.
Azari, R., Garshasbi, S., Amini, P., Rashed-Ali, H., Mohammadi, Y., 2016. Multi-objective optimization of building envelope design for life cycleenvironmental performance. Energ. Build. 126, 524–534.
Barros, N.N., Ruschel, R.C., 2020. Machine learning for whole-building life cycle assessment: a systematic literature review. In: Santos, E.T., Scheer, S. (Eds.), Proceedings of the 18th International Conference on Computing in Civil and Building Engineering. Springer, Cham.
Bhutta, M.N.M., Ahmad, M., 2021. Secure identification, traceability and real-time tracking of agricultural food supply during transportation using internet of things. IEEE Access 9, 65660–65675.
Cheng, F., Luo, H., Colosi, L.M., 2020a. Slow pyrolysis as a platform for negative emissions technology: an integration of machine learning models, life cycle assessment, and economic analysis. Energy Convers. Manag. 223, 113258.
Cheng, F., Porter, M.D., Colosi, L.M., 2020b. Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? Energy Convers. Manag. 112252.
Cooper, J., Noon, M., Jones, C., Kahn, E., Arbuckle, P., 2013. Big data in life cycle assessment. J. Ind. Ecol. 17, 796–799.
Cornago, S., Vitali, A., Brondi, C., Sze, J., Low, C., 2020. Electricity technological mix forecasting for life cycle assessment aware scheduling. Procedia CIRP 90, 268–273.
D'Amico, A., Ciulla, G., Traverso, M., Lo Brano, V., Palumbo, E., 2019. Artificial neural networks to assess energy and environmental performance of buildings: an Italian case study. J. Clean. Prod. 239, 117993.
Diamantidis, N.A., Karlis, D., Giakoumakis, E.A., 2000. Unsupervised stratification of cross-validation for accuracy estimation. Artif. Intell. 116, 1–16.
Dick, M., Silva, Mad, Dewes, H., 2015. Mitigation of environmental impacts of beef cattle production in southern Brazil evaluation using farm-based life cycle assessment. J. Clean. Prod. 87, 58–67.
Duprez, S., Fouquet, M., Herreros, Q., Jusselme, T., 2019. Improving life cycle-based exploration methods by coupling sensitivity analysis and metamodels. Sustain. Cities Soc. 44, 70–84.
European Commission. European Platform on LCA. https://eplca.jrc.ec.europa.eu/.
Fathi, S., Srinivasan, R., Fenner, A., Fathi, S., 2020. Machine learning applications in urban building energy performance forecasting: a systematic review. Renew. Sust. Energ. Rev. 133, 110287.
Federal LCA Commons, 2022. Federal LCA Commons, a Central Point of Access to a Collection of Data Repoitories for Use in Life Cycle Assessment. https://www.lcacommons.gov/.
Feng, K., Lua, W., Wang, Y., 2019. Assessing environmental performance in early building design stage: an integrated parametric design and machine learning method. Sustain. Cities Soc. 50, 101596.
Ghoroghi, A., Rezgui, Y., Petri, I., Beach, T., 2022. Advances in application of machine learning to life cycle assessment: a literature review. Int. J. Life Cycle Assess. 27, 433–456.
Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques. Elsevier.
Hellweg, S., Canals, L.M.I., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. Science 344, 1109–1113.
Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int. 135, 105393.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: With Applications in R. Springer-Verlag.
Kaab, A., Sharifi, M., Mobli, H., Nabavi-Pelesaraei, A., Chauc, K.-w., 2019. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. Sci. Total Environ. 664, 1005–1019.
Khanali, M., Mobli, H., Hosseinzadeh-Bandbafha, H., 2017. Modeling of yield and environmental impact categories in tea processing units based on artificial neural networks. Environ. Sci. Pollut. Res. 24, 26324–26340.
Khoshnevisan, B., Rafie, S., Mousazadeh, H., 2013a. Environmental impact assessment of open field and greenhouse strawberry production. Eur. J. Agron. 50, 29–37.
Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., Sefeedpari, P., 2013b. Prognostication of environmental indices in potato production using artificial neural networks. J. Clean. Prod. 52, 402–409.
Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., Clark, S., 2014a. Environmental impact assessment of tomato and cucumber cultivation in greenhouses using life cycle assessment and adaptive neuro-fuzzy inference system. J. Clean. Prod. 73, 183–192.
Khoshnevisan, B., Rajaeifar, M.A., Clarkb, S., Shamahirband, S., Anuar, N.B., Shuib, N.L. M., et al., 2014b. Evaluation of traditional and consolidated rice farms in Guilan Province, Iran, using life cycle assessment and fuzzy modeling. Sci. Total Environ. 481, 242–251.
Lee EK, Zhang W-J, Zhang X, Adler PR, Lin S, Beth J. Feingolda, et al. 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. 2020; 714: 136697.
Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., et al., 2023. Big data in earth system science and progress towards a digital twin. Nat. Rev. Earth Environ. 4, 319–332.
Liao, M., Kelley, S., Yao, Y., 2020. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. ACS Sustain. Chem. Eng. 8, 1252–1261.
Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., et al., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 6, e1000100.
Maksimović, M., Vujović, V., Omanović-Mikličanin, E., 2015. Application of internet of things in food packaging and transportation. Int. J. Sustain. Agric. Manag. Inform. 1, 333–350.
Mao, X., Wang, L., Li, J., Quan, X., Wu, T., 2019. Comparison of regression models for estimation of carbon emissions during building's lifecycle using designing factors: a case study of residential buildings in Tianjin, China. Energy Build. 204, 109519.
Marvuglia, A., Kanevski, M., Benetto, E., 2015a. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. Environ. Int. 83, 72–85.
Marvuglia, A., Kanevski, M., Benetto, E., 2015b. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. Environ. Int. 83, 72–85.
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.
Mitchell, T.M., 1997. Machine Learning. McGraw-Hill Education.
Mousavi-Avvala, S.H., Rafieea, S., Sharifia, M., Hosseinpoura, S., Shahb, A., 2017. Combined application of life cycle assessment and adaptive neuro-fuzzy inference system for modeling energy and environmental emissions of oilseed production. Renew. Sust. Energ. Rev. 78, 807–820.
Nabavi-Pelesaraei, A., Rafiee, S., Mohtasebi, S.S., Hosseinzadeh-Bandbafha, H., Chau, K.-w., 2018. Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. Sci. Total Environ. 631–632.
Naseri, H., Jahanbakhsh, H., Hosseini, P., Nejad, F.M., 2020. Designing sustainable concrete mixture by developing a new machine learning technique. J. Clean. Prod. 258, 120578.
Nguyen, T.H., Nong, D., Paustian, K., 2019. Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks. Ecol. Model. 400, 1–13.
Ozbilen, A., Aydin, M., Dincer, I., Rosen, M.A., 2013. Life cycle assessment of nuclear-based hydrogen production via a copperechlorine cycle: a neural network approach. Int. J. Hydrog. Energy 38, 6314–6322.
Pishgar-Komleh, S.H., Akram, A., Keyhani, A., Sefeedpari, P., Shine, P., Brandao, M., 2020a. Integration of life cycle assessment, artificial neural networks, and metaheuristic optimization algorithms for optimization of tomato-based cropping systems in Iran. Int. J. Life Cycle Assess. 25, 620–632.
Pishgar-Komleh, S.H., Akram, A., Keyhani, A., Sefeedpari, P., Shine, P., Brandao, M., 2020b. Integration of life cycle assessment, artificial neural networks, and metaheuristic optimization algorithms for optimization of tomato-based cropping systems in Iran. Int. J. Life Cycle Assess. 25, 620–632.
Płoszaj-Mazurek, M., Ez, Ry 'nska, Grochulska-Salak, M., 2020. Methods to optimize carbon footprint of buildings in regenerative architectural design with the use of machine learning, convolutional neural network, and parametric design. Energies 13, 5289.
Ramakrishnan, N., Marwah, M., Shah, A., Patnaik, D., Hossain, M.S., Sundaravaradan, N., et al., 2012. Data mining solutions for sustainability problems. IEEE Potentials 31, 28–34.
Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., et al., 2022. Tackling climate change with machine learning. ACM Comput. Surv. 55, 1–96.
Romeiko, X.X., Lee, E.K., Sorunmu, Y., Zhang, X., 2020a. Spatially and temporally explicit life cycle environmental impacts of soybean production in the U.S. midwest. In: Environmental Science & Technology, 54, pp. 4758–4768.
Romeiko, X.X., Zhijia, G., Pang, Y., Lee, E.K., Zhang, X., 2020b. Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production. Sustainability 12, 1481.
Saha, D., Manickavasagan, A., 2021. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: a review. Curr. Res. Food Sci. 4, 28–44.
Samad, S., Ahmed, F., Naher, S., Kabir, M.A., Das, A., Amin, S., et al., 2022. Smartphone apps for tracking food consumption and recommendations: evaluating artificial intelligence-based functionalities, features and quality of current apps. Intell. Syst. Appl. 15, 200103.
Santos, Bsd, Steiner, M.T.A., Fenerich, A.T., RHP, Lima, 2019. Data mining and machine learning techniques applied to public health problems: a bibliometric analysis from 2009 to 2018. Comput. Ind. Eng. 138, 106120.
Sharifa, S.A., Hammad, A., 2019. Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA. J. Build. Eng. 25, 100790.
Slapnik, M., Istenic, D., Pintar, M., Udovc, A., 2015a. Extending life cycle assessment normalization factors and use of machine learning – a Slovenian case study. Ecol. Indic. 50, 161–172.
Slapnik, M., Istenič, D., Pintar, M., Udovč, A., 2015b. Extending life cycle assessment normalization factors and use of machine learning – a Slovenian case study. Ecol. Indic. 50, 161–172.
Song, R., Keller, A.A., Suh, S., 2017. Rapid life-cycle impact screening using artificial neural networks. Environ. Sci. Technol. 51, 10777–10785.

Sousa, I., Eisenhard, J.L., Wallace, D., 2001. Approximate life-cycle assessment of product concepts using learning systems. J. Ind. Ecol. 4, 61–81.
Sundaravaradan, N., Patnaik, D., Ramakrishnan, N., Marwah, M., Shah, A., 2011. Discovering Life Cycle Assessment Trees From Impact Factor Databases. In: Association for the Advancement of Artificial Intelligence Proceedings of the AAAI Conference on Artificial Intelligence, 25, pp. 1415–1420.
Suter-Crazzolara, C., 2018. Better patient outcomes through mining of biomedical big data. Front. ICT 5, 30.
Tao, M., Li, D., Song, R., Suh, S., Keller, A.A., 2018. OrganoRelease e a framework for modeling the release of organic chemicals from the use and post-use of consumer products. Environ. Pollut. 234, 751–761.
The Ecoinvent Association. Ecoinvent Database. https://ecoinvent.org/the-ecoinvent-database/, accessed June 2023.
Thilakarathna, P.S.M., Seo, S., Kristombu Baduge, K.S., Lee, H., Mendis, P., Foliente, G., 2020. Embodied carbon analysis and benchmarking emissions of high and ultra-high strength concrete using machine learning algorithms. J. Clean. Prod. 262, 12181.
UN Environment Programme, 2023. The Global LCA Data Access Network. https://www.unep.org/explore-topics/resource-efficiency/what-we-do/life-cycle-initiative/global-lca-data-access-network. (Accessed June 2023).
Vlontzosa, G., Pardalosb, P.M., 2017. Assess and prognosticate green house gas emissions from agricultural production of EU countries, by implementing, DEA window analysis and artificial neural networks. Renew. Sust. Energ. Rev. 76, 155–162.
Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J., 2017. Big data in smart farming – a review. Agric. Syst. 153, 69–80.
Xu, M., Cai, H., Liang, S., 2015. Big data and industrial ecology. J. Ind. Ecol. 19, 205–210.
Zeng, X., Martinez, T.R., 2000. Distribution-balanced stratified cross-validation for accuracy estimation. J. Exp. Theor. Artif. Intell. 12, 1–12.
Zhang, X., Srinivasan, R., Van Liew, M., 2009. Approximating SWAT model using artificial neural network and support vector machine. J. Am. Water Resour. Assoc. 45, 460–474.
Zhao, Y., Zhang, Q., Li, F.Y., 2019. Patterns and drivers of household carbon footprint of the herdsmen in the typical steppe region of inner Mongolia, China: a case study in Xilinhot City. J. Clean. Prod. 232, 408–416.
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, 11141–11151.
Ziyadi, M., Al-Qadi, I.L., 2019. Model uncertainty analysis using data analytics for life-cycle assessment (LCA) applications. Int. J. Life Cycle Assess. 24, 945–959.