EXIOML: ECO-ECONOMIC DATASET FOR MACHINE LEARNING IN GLOBAL SECTORAL SUSTAINABILITY

Yanming Guo
School of Electrical & Computer Engineering
University of Sydney
yguo0337@uni.sydney.edu.au

Jin Ma
School of Electrical & Computer Engineering
University of Sydney
j.ma@sydney.edu.au

ABSTRACT

The Environmental Extended Multi-regional Input-Output analysis is the predominant Ecological Economic research framework for analysing the environmental impact of economic activities. This paper introduces the novel ExioML dataset as the first Machine Learning benchmark data in sustainability analysis. We open-sourced the ExioML data and development toolkit to lower barriers and accelerate the cooperation between Machine Learning and Ecological Economic research. A crucial greenhouse gas emission regression task evaluates the usability of the proposed dataset. We compared the performance of traditional shallow models against deep models by leveraging a diverse factor accounting table and incorporating multiple modalities of categorical and numerical features. Our findings reveal that deep and ensemble models achieve low mean square errors below 0.25 and serve as a future machine learning research baseline. Through ExioML, we aim to foster precise ML predictions and modelling to support climate actions and sustainable investment decisions. The data and codes are available: https://github.com/Yvnminc/ExioML

1 INTRODUCTION

The increase in Greenhouse Gas (GHG) emissions due to fossil-fuel-driven economic development has precipitated a global warming crisis. This concern led to establishing the Paris Agreement in 2015, aiming to limit long-term temperature rise to no more than 2 °C above pre-industrial levels (Schleussner et al., 2016). Concurrently, the Sustainable Development Goals (SDGs) were proposed to emphasize poverty eradication within the framework of emission reductions (Bruckner et al., 2022; Hubacek et al., 2017). To address the climate-trade dilemma, researchers from various disciplines strive to balance climate action with economic growth and human well-being (Rao et al., 2014; Jorgenson, 2014). Recently, the Machine Learning (ML) technique (LeCun et al., 2015) has emerged as a significant tool for accurate prediction to assist climate change decision-making (Rolnick et al., 2022). Specifically, ML algorithms have been explored in aiding nearly real-time global weather forecasting (Lam et al., 2023), land monitoring via satellite imagery (Stanimirova et al., 2023; He et al., 2016), and the prediction of disturbances in electric grids (Zheng et al., 2022).

The predominant Ecological Economic (EE) research framework, the Environmentally Extended Multiregional Input-Output (EE-MRIO) analysis, effectively models global economic interactions of sectors within a network structure (Leontief & Strout, 1963; Wang et al., 2019; Sun et al., 2020a; Steinberger et al., 2012; Jakob & Marschinski, 2013). However, a research gap exists between Ecological Economic and Machine Learning (EE-ML) cooperation, which can be summarised in three ways. The inaccessibility of the EE-MRIO dataset, the intensive data preprocessing required domain knowledge and the lack of Benchmark ML datasets and models.

We proposed ExioM as the first ML-ready benchmark data in EE research to fill these gaps. The ExioML is developed on top of the high-quality open-source EE-MRIO dataset ExioBase 3.8.2 with high spatiotemporal resolution, enabling tracking resource flow between international sectors for global sectoral sustainability assessment. The ExioML developing toolkit is open-sourced, provides EE-MRIO calculation tools with GPU acceleration, and contributes to the high flexibility of customising interested factors, which reduces the barrier for ML researchers. The usability of the ExioML dataset was validated using multimodality regression to quantify the sectoral GHG emission, and a low MSE result was achieved. This establishes a critical baseline for accuracy and efficiency in future ML research.

2 LITERATURE REVIEW

EE-MRIO describes the environmental footprint for global economic activities and has become the fundamental framework for EE research, supporting various studies such as Structure Decomposition Analysis (SDA) and Index Decomposition Analysis (IDA) to identify changes by decomposing key drivers (Hoekstra & Van den Bergh, 2003; Peters et al., 2017; Duan & Yan, 2019). Monitor embodied emissions with resource transfer and sustainability evaluation of supply chain in global trade (Meng et al., 2023; Tian et al., 2022; Wu et al., 2020; He & Hertwich, 2019; Li et al., 2020; Long et al., 2018; Sun et al., 2020a). ML algorithms have been applied with EE-MRIO for several applications, such as accurately identifying ecological hotspots and inefficiencies within the global supply chain to optimise logistic paths to decrease carbon emissions while considering cost-effectiveness (Akbari & Do, 2021). ML algorithms are naturally suitable for learning multi-dimensional patterns and are utilised for better sectoral sustainability assessment considering environmental, economic and social impacts (Abdella et al., 2020; Nilashi et al., 2019). Decomposition algorithms are leveraged to track the critical paths of products' energy, water, and emission footprints from giant trading networks for supporting sustainable investment decisions (Ding et al., 2022).

However, unlike the open-source culture of the mainstream ML community (Zheng et al., 2022), most EE studies haven't publicised the code and data. This leads to fragmented, independent, and inconsistent research leveraging ML techniques to explore the sustainability of different sectors and regions. Subsequent research cannot fairly compare the model performance, such as accuracy, robustness, and generalisation or reproduce the result based on previous research, leading to the slow EE-ML development (Zhu et al., 2023; Ballarin et al., 2023; Nangini et al., 2019).

No public benchmark EE-ML dataset is available to the best of our knowledge. The EE-MRIO data is a high-quality source for constructing ML-ready datasets. However, the existing EE-MRIO dataset faces limitations in terms of accessibility and resolution. Specifically, Eora offers high-quality data but is closed-source and requires high-cost purchasing (Lenzen et al., 2013). Global Trade Analysis Project (GTAP) covers only five reference time steps with non-free access (Chepeliev, 2023). The World Input-Output Database (WIOD) (Dietzenbacher et al., 2013) questioned the low resolution of temporal and spatial scope, whose latest temporal coverage is 2016. Therefore, uniform problem formulations and benchmark datasets need to be urgently developed.

3 EXIOML DATA AND DEVELOPMENT TOOLKIT

ExioML is constructed as the first benchmark dataset in EE-ML research. The architecture of the proposed dataset is illustrated in Figure 1. ExioML addressed the challenges of data limitation and simplified the EE-MRIO framework for ML researchers in a ready-to-use manner. ExioML offers high spatiotemporal resolution for 49 regions from 1995 to 2022 contains tabular and graphical formats, supporting diverse ML research, such as time-series emissions forecasting (Deb et al., 2017), factor analysis (Raihan et al., 2022; Zhang et al., 2019; Duan & Yan, 2019; Acheampong & Boateng, 2019; Khan et al., 2020), graph learning (Sun et al., 2020b) and clustering analysis (He et al., 2022; Kijewska & Bluszcz, 2016), under a uniform framework. The details of ExioML are summarised in Tablue 1.

Table 1: Summaries of details of ExioML dataset.

Characteristics	**Description**
Components	Factor accounting (Tabular data), Footprint Network (Graph data)
Time frame	Covers 28 annual time steps from 1995 to 2022
Geographical coverage	49 regions (44 countries and 5 rest of the world)
Sectors detail	200 products in PxP, 163 industries in IxI
Key factors	Value added, employment, energy consumption, and GHG emission

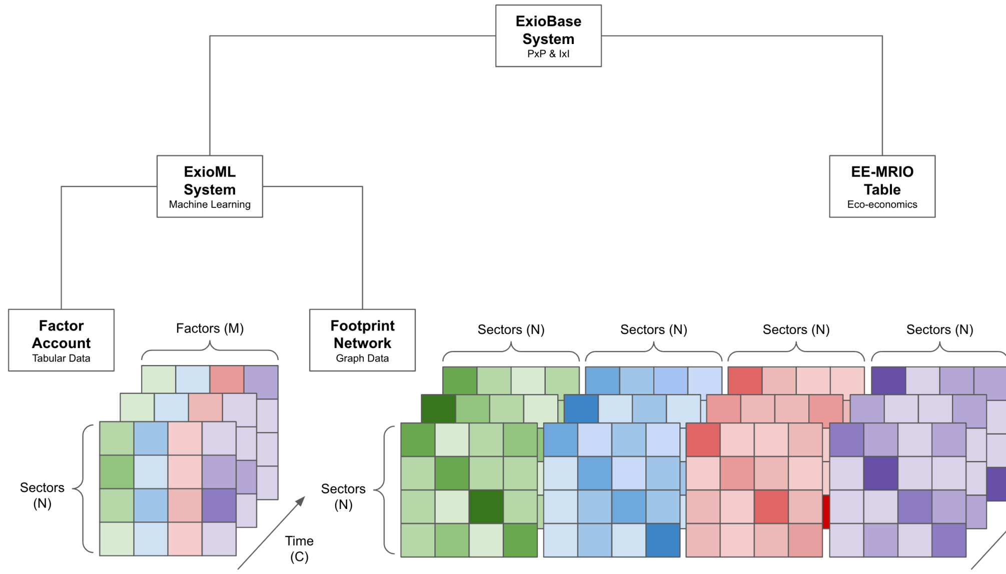

Figure 1: Architecture of ExioML system derived from the open-source EE-MRIO database, ExioBase 3.8.2. Each colour indicates an eco-economic factor: value added, employment, energy consumption and GHG emission. The system contains factor accounting data describing heterogeneous sector features. The footprint network models the global trading network tracking resource transfer within sectors. The data is presented into 2 categories: 200 products and 163 industries for 49 regions from 1995 to 2022 in the PxP and IxI datasets.

ExioML is derived from an open-source EE-MRIO dataset, ExioBase 3.8.2 (Stadler et al., 2018) and leveraged Pymrio (Stadler, 2021), a Python-based MRIO toolkit to facilitate data download and basic MRIO operations. ExioBase 3.8.2 is a comprehensive source database covering 417 emission categories and 662 material and resource categories occupying over 40 GB of storage. Intensive data preprocessing, such as heavy data cleaning, structure transformation, and footprint network calculation, is required to transfer the row data into a standardized ML paradigm. Here are the construction details of the ExioML database that utilises the MRIO framework. Initially, the column vector x symbolizes the global economy's output, with each element x^r representing the total output of region r. This could be determined by transaction matrix Z and demand matrix Y:

$$\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^m \end{pmatrix} = \begin{pmatrix} Z^{11} & Z^{12} & \cdots & Z^{1n} \\ Z^{21} & Z^{22} & \cdots & Z^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z^{m1} & Z^{m2} & \cdots & Z^{mn} \end{pmatrix} + \begin{pmatrix} \sum_s y^{1s} \\ \sum_s y^{2s} \\ \vdots \\ \sum_s y^{ms} \end{pmatrix}. \tag{1}$$

Z indicates the transaction matrix for n sectors in m regions, and Z^{rs} is the inter-regional requirement from region r to region s. y^{rs} is the final demand from region r to region s.

Then, the direct requirement matrix A, indicative of technological efficiency, is derived by multiplying the transaction matrix Z with the diagonalized, inverted vector $\hat{x}^{-1}$:

$$A = Z\hat{x}^{-1}. \tag{2}$$

Further, the economy's output vector X is expressible via the Leontief matrix L:

$$X = (I - A)^{-1}y = Ly. \tag{3}$$

This framework can incorporate environmental accounting, such as energy consumption and greenhouse gas (GHG) emissions, with the factor represented by F. The coefficient S normalizes this factor against the output x:

$$S = F\hat{x}^{-1}. \tag{4}$$

Finally, the footprint flow networks D could be determined by:

$$D = SLy. \tag{5}$$

We open-sourced the well-documented ExioML development toolkit to stimulate interdisciplinary corporations in three contributions. Firstly, the toolkit reduces the barrier for new ML researchers without EE domain knowledge by encapsulating a complex MRIO framework. Secondly, data storage and computational demands for the EE-MRIO dataset are substantial due to high-dimensional matrix operations. ExioML development toolkit utilises GPU accelerations and elegantly stores multi-dimensional networks in edge tables. Thirdly, the toolkit benefits later researchers with high-level flexibility to customise the factors of interest. MRIO features contain specific eco-economic meanings, such as scope 1, 2 and 3 emissions. Independent features should be carefully selected. To achieve that, four essential factors interested by the EE research community, which are emissions (F), population (P), GDP (G), and energy consumption (E) indicated by the Kaya Identity (González-Torres et al., 2021) are selected:

$$F = P \times \frac{G}{P} \times \frac{E}{G} \times \frac{F}{E}. \tag{6}$$

The factors included in ExioML are detailed in Table 2. ExioML comprises two pivotal elements. Firstly, factor accounting, presented in a tabular format, delineates heterogeneous characteristics of various sectors, a subset of the emission extension table with selected factors from ExioBase. Secondly, the footprint network, a multivariate time-series network, describes the transfer of the resource footprints among sectors. Each element is divided into two categories: The product-by-product (PxP) dataset encompasses 200 products, and the industry-by-industry (IxI) dataset covers 163 industries, crossing 49 regions spanning 1995 to 2022.

Table 2: Feature PxP and IxI dataset in factor account table.

Attribute name	**Type**	**Description**
Region	Categorical	Region code (e.g. AU, US, CN)
Sector	Categorical	Product or industry (e.g. biogasoline, construction)
Year	Numerical	Timestep (e.g. 1995, 2022)
Value added	Numerical	GDP expressed in millions of Euros
Employment	Numerical	Population engaged in thousands of persons
Energy carrier net total	Numerical	Sum of all energy carriers in Terajoules
GHG emission	Numerical	GHG emissions in kilograms of CO_2 equivalent

4 VALIDATION

Our study demonstrates a multimodel sectoral GHG emissions regression task leveraging categorical and numerical features to validate the usability of the proposed ExioML. The model aims to learn the underlying relationships of eco-economic factors to quantify sectoral GHG emission and assess sectoral sustainability using the factor accounting in the ExioML dataset (Sun & Huang, 2022).

The supervised regression task is to learn the mapping function $f_\theta : X \mapsto Y$ from feature vector $x_i \in X$ to labels $y_i \in Y$ on dataset $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$. In this case, ee utilized the PxP and IxI as dataset $\mathcal{D}$ of ExioML's factor accounting, comprising 221,307 and 179,185 instances after excluding missing values, respectively. The feature vector X is value-added, employment and energy consumption, and GHG emission is set as the label Y.

Regression models could be categorized into shallow and deep learning models. Deep learning in eco-economic analysis, particularly on tabular data, remains relatively underexplored (Gorishniy et al., 2021). We rigorously evaluated different algorithms on tabular learning algorithms, including tree-based models, MLP and Transformer architectures (Vaswani et al., 2017). Shallow models employ Leave-One-Out Text Encoding due to their inability to process categorical data natively (Zhang, 2016), whereas deep learning models learn categorical embeddings directly from raw data are implemented by PyTorch Tabular package (Joseph, 2021).

5 RESULTS

This analysis assessed each model based on its MSE and training time. To ensure unbiased evaluation and prevent leakage of target information, the dataset was partitioned into training, validation, and testing sets with proportions of 64%, 16% and 20%. Optimal model parameters were determined through a random search across 30 trials on the training and validation data. We conducted 10 experiments for each tuned hyperparameter with their performance on the test set. Detailed information regarding the model, implementation, and hyperparameters can be found in the Appendix.

Table 3 summarized the model's mean performance and standard deviation. The primary finding is that deep learning models exhibit marginally lower mean squared errors than ensemble counterparts. GANDALF shows the most effective performance in PxP and IxI data. Ensemble methods, GBDT and RF, emerge with competitive accuracy and less demand for computational resources. Although deep models benefit from GPU acceleration, their training time is significantly longer than shallow models. This extended training duration could potentially lead to increased emissions.

Table 3: Result for models with standard deviation for 10 runs and the top results are in **bold**.

	PxP		**IxI**	
Model	MSE	Time (s)	MSE	Time (s)
KNN	1.071 ± 0.010	0.035 ± 0.001	1.151 ± 0.018	0.026 ± 0.001
Ridge	2.265 ± 0.000	**0.005 ± 0.002**	2.514 ± 0.000	**0.004 ± 0.002**
DT	0.926 ± 0.051	0.316 ± 0.017	0.848 ± 0.070	0.254 ± 0.019
RF	0.356 ± 0.004	21.521 ± 0.157	0.302 ± 0.004	16.511 ± 0.060
GBDT	0.234 ± 0.006	30.276 ± 0.224	0.219 ± 0.007	32.847 ± 0.388
MLP	0.226 ± 0.007	219.218 ± 0.904	0.250 ± 0.092	205.051 ± 24.309
GANDALF	**0.204 ± 0.010**	352.756 ± 7.036	**0.189 ± 0.007**	383.119 ± 3.664
FTT	0.330 ± 0.007	330.578 ± 1.527	0.302 ± 0.023	468.911 ± 7.329

6 CONCLUSION

EE-MRIO framework is a powerful analysis tool in EE research and has been widely discussed for key factor analysis, sustainability assessment and global supply chain optimisation. ML algorithms are naturally fit for modelling non-linear complex relationships between multi-dimensional factors. However, the ML algorithms in the EE field are underdeveloped due to three major challenges. The inaccessibility of EE-MRIO data, the high domain knowledge requirement to utilise EE-MRIO framework and the absence of uniformed ML problem setting, benchmark data and algorithms.

This research introduces the first high-quality EE-ML benchmark dataset integrated with the EE-MRIO framework to address these research gaps. The proposed novel ExioML data and development toolkit addressed three significant ML borders in sustainability research. Firstly, ExioML ends data access limitations and provides high resolution based on the open-source EE-MRIO database, ExioBase 3.8.2. Secondly, we streamline the complexity of the EE-MRIO framework with the ExioML development toolkit. It provides MRIO computation with GPU accelerations for high-level customisation for later researchers. Finally, we demonstrated the evaluation of various machine learning models on factor accounting datasets focused on sectoral GHG emission prediction through regression tasks, serving as effective baselines with low mean square error performance. ExioML simulates the interdisciplinary EE-ML corporation and promotes promising research, such as supply chain optimization, suitability assessment, and footprint prediction, which contributes to formulating effective climate policies and sustainable investment decisions.

References

Galal M Abdella, Murat Kucukvar, Nuri Cihat Onat, Hussein M Al-Yafay, and Muhammet Enis Bulak. Sustainability assessment and modeling based on supervised machine learning techniques: The case for food consumption. *Journal of Cleaner Production*, 251:119661, 2020. doi: https://doi.org/10.1016/j.jclepro.2019.119661.

Alex O Acheampong and Emmanuel B Boateng. Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, 225:833–856, 2019. doi: https://doi.org/10.1016/j.jclepro.2019.03.352.

Mohammadreza Akbari and Thu Nguyen Anh Do. A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*, 28(10):2977–3005, 2021. doi: https://doi.org/10.1108/BIJ-10-2020-0514.

André Simões Ballarin, Jullian Souza Sone, Gabriela Chiquito Gesualdo, Dimaghi Schwamback, Alan Reis, André Almagro, and Edson Cezar Wendland. Climbra-climate change dataset for brazil. *Scientific Data*, 10(1):47, 2023. doi: https://doi.org/10.1038/s41597-023-01956-z.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.

Benedikt Bruckner, Klaus Hubacek, Yuli Shan, Honglin Zhong, and Kuishuang Feng. Impacts of poverty alleviation on national and global carbon emissions. *Nature Sustainability*, 5(4):311–320, 2022. doi: https://doi.org/10.1038/s41893-021-00842-z.

Maksym Chepeliev. Gtap-power data base: Version 11. *Journal of Global Economic Analysis*, 8(2), 2023. doi: https://doi.org/10.21642/JGEA.080203AF.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017. doi: https://doi.org/10.1016/j.rser.2017.02.085.

Erik Dietzenbacher, Bart Los, Robert Stehrer, Marcel Timmer, and Gaaitzen De Vries. The construction of world input–output tables in the wiod project. *Economic systems research*, 25(1):71–98, 2013. doi: https://doi.org/10.1080/09535314.2012.761180.

Yakui Ding, Yongping Li, Heran Zheng, Jing Meng, Jing Lv, and Guohe Huang. Identifying critical energy-water paths and clusters within the urban agglomeration using machine learning algorithm. *Energy*, 250:123880, 2022. doi: https://doi.org/10.1016/j.energy.2022.123880.

Yuwan Duan and Bingqian Yan. Economic gains and environmental losses from international trade: A decomposition of pollution intensity in china's value-added trade. *Energy economics*, 83:540–554, 2019. doi: https://doi.org/10.1016/j.eneco.2019.08.002.

John Foster. From simplistic to complex systems in economics. *Cambridge Journal of Economics*, 29(6):873–892, 2005. doi: https://doi.org/10.1093/cje/bei083.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

M González-Torres, L Pérez-Lombard, JF Coronel, and IR Maestre. Revisiting kaya identity to define an emissions indicators pyramid. *Journal of Cleaner Production*, 317:128328, 2021. doi: https://doi.org/10.1016/j.jclepro.2021.128328.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021. doi: https://doi.org/10.48550/arXiv.2106.11959.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kehan He and Edgar G Hertwich. The flow of embodied carbon through the economies of china, the european union, and the united states. *Resources, Conservation and Recycling*, 145:190–198, 2019. doi: https://doi.org/10.1016/j.resconrec.2019.02.016.

Ying He, Yuantong Xing, Xiancheng Zeng, Yijun Ji, Huimin Hou, Yang Zhang, and Zhe Zhu. Factors influencing carbon emissions from china's electricity industry: Analysis using the combination of lmdi and k-means clustering. *Environmental Impact Assessment Review*, 93:106724, 2022. doi: https://doi.org/10.1016/j.eiar.2021.106724.

Rutger Hoekstra and Jeroen CJM Van den Bergh. Comparing structural decomposition analysis and index. *Energy economics*, 25(1):39–64, 2003. doi: https://doi.org/10.1016/S0140-9883(02)00059-2.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: https://doi.org/10.1080/00401706.1970.10488634.

Klaus Hubacek, Giovanni Baiocchi, Kuishuang Feng, and Anand Patwardhan. Poverty eradication in a carbon constrained world. *Nature communications*, 8(1):912, 2017. doi: https://doi.org/10.1038/s41467-017-00919-4.

Michael Jakob and Robert Marschinski. Interpreting trade-related co2 emission transfers. *Nature Climate Change*, 3(1):19–23, 2013. doi: https://doi.org/10.1038/nclimate1630.

Andrew K Jorgenson. Economic development and the carbon intensity of human well-being. *Nature Climate Change*, 4(3):186–189, 2014. doi: https://doi.org/10.1038/nclimate2110.

Manu Joseph. Pytorch tabular: A framework for deep learning with tabular data. *arXiv preprint arXiv:2104.13638*, 2021. doi: https://doi.org/10.48550/arXiv.2104.13638.

Manu Joseph and Harsh Raj. Gate: Gated additive tree ensemble for tabular classification and regression. *arXiv preprint arXiv:2207.08548*, 2022. doi: https://doi.org/10.48550/arXiv.2207.08548.

Hayat Khan, Itbar Khan, and Truong Tien Binh. The heterogeneity of renewable energy consumption, carbon emission and financial development in the globe: a panel quantile regression approach. *Energy Reports*, 6:859–867, 2020. doi: https://doi.org/10.1016/j.egyr.2020.04.002.

Anna Kijewska and Anna Bluszcz. Research of varying levels of greenhouse gas emissions in european countries using the k-means method. *Atmospheric Pollution Research*, 7(5):935–944, 2016. doi: https://doi.org/10.1016/j.apr.2016.05.010.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: https://doi.org/10.1126/science.adi2336.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. doi: https://doi.org/10.1038/nature14539.

Manfred Lenzen, Daniel Moran, Keiichiro Kanemoto, and Arne Geschke. Building eora: a global multi-region input–output database at high country and sector resolution. *Economic Systems Research*, 25(1):20–49, 2013. doi: https://doi.org/10.1080/09535314.2013.769938.

Wassily Leontief and Alan Strout. Multiregional input-output analysis. In *Structural Interdependence and Economic Development: Proceedings of an International Conference on Input-Output Techniques, Geneva, September 1961*, pp. 119–150. Springer, 1963. doi: https://doi.org/10.1007/978-1-349-81634-7_8.

YL Li, B Chen, and GQ Chen. Carbon network embodied in international trade: Global structural evolution and its policy implications. *Energy Policy*, 139:111316, 2020. doi: https://doi.org/10.1016/j.enpol.2020.111316.

Ruyin Long, Jinqiu Li, Hong Chen, Linling Zhang, and Qianwen Li. Embodied carbon dioxide flow in international trade: A comparative analysis based on china and japan. *Journal of environmental management*, 209:371–381, 2018. doi: https://doi.org/10.1016/j.jenvman.2017.12.067.

Jing Meng, Jingwen Huo, Zengkai Zhang, Yu Liu, Zhifu Mi, Dabo Guan, and Kuishuang Feng. The narrowing gap in developed and developing country emission intensities reduces global trade's carbon leakage. *Nature Communications*, 14(1):3775, 2023. doi: https://doi.org/10.1038/s41467-023-39449-7.

Cathy Nangini, Anna Peregon, Philippe Ciais, Ulf Weddige, Felix Vogel, Jun Wang, François-Marie Bréon, Simeran Bachra, Yilong Wang, Kevin Gurney, et al. A global dataset of co2 emissions and ancillary data related to emissions for 343 cities. *Scientific data*, 6(1):1–29, 2019. doi: https://doi.org/10.1038/sdata.2018.280.

Mehrbakhsh Nilashi, Parveen Fatemeh Rupani, Mohammad Mobin Rupani, Hesam Kamyab, Weilan Shao, Hossein Ahmadi, Tarik A Rashid, and Nahla Aljojo. Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. *Journal of Cleaner Production*, 240:118162, 2019. doi: https://doi.org/10.1016/j.jclepro.2019.118162.

Glen P Peters, Robbie M Andrew, Josep G Canadell, Sabine Fuss, Robert B Jackson, Jan Ivar Korsbakken, Corinne Le Quéré, and Nebojsa Nakicenovic. Key indicators to track current progress and future ambition of the paris agreement. *Nature Climate Change*, 7(2):118–122, 2017. doi: https://doi.org/10.1038/nclimate3202.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986. doi: https://doi.org/10.1007/BF00116251.

Asif Raihan, Rawshan Ara Begum, Mohd Nizam Mohd Said, and Joy Jacqueline Pereira. Relationship between economic growth, renewable energy use, technological innovation, and carbon emission toward achieving malaysia's paris agreement. *Environment Systems and Decisions*, 42 (4):586–607, 2022. doi: https://doi.org/10.1007/s10669-022-09848-0.

Narasimha D Rao, Keywan Riahi, and Arnulf Grubler. Climate impacts of poverty eradication. *Nature Climate Change*, 4(9):749–751, 2014. doi: https://doi.org/10.1038/nclimate2340.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022. doi: https://doi.org/10.1145/3485128.

Carl-Friedrich Schleussner, Joeri Rogelj, Michiel Schaeffer, Tabea Lissner, Rachel Licker, Erich M Fischer, Reto Knutti, Anders Levermann, Katja Frieler, and William Hare. Science and policy characteristics of the paris agreement temperature goal. *Nature Climate Change*, 6(9):827–835, 2016. doi: https://doi.org/10.1038/nclimate3096.

Konstantin Stadler. Pymrio–a python based multi-regional input-output analysis toolbox. 2021. doi: https://doi.org/10.5334/jors.251.

Konstantin Stadler, Richard Wood, Tatyana Bulavskaya, Carl-Johan Södersten, Moana Simas, Sarah Schmidt, Arkaitz Usubiaga, José Acosta-Fernández, Jeroen Kuenen, Martin Bruckner, et al. Exiobase 3: Developing a time series of detailed environmentally extended multi-regional input-output tables. *Journal of Industrial Ecology*, 22(3):502–515, 2018. doi: https://doi.org/10.1111/jiec.12715.

Radost Stanimirova, Katelyn Tarrio, Konrad Turlej, Kristina McAvoy, Sophia Stonebrook, Kai-Ting Hu, Paulo Arévalo, Eric L Bullock, Yingtong Zhang, Curtis E Woodcock, et al. A global land cover training dataset from 1984 to 2020. *Scientific Data*, 10(1):879, 2023. doi: https://doi.org/10.1038/s41597-023-02798-5.

Julia K Steinberger, J Timmons Roberts, Glen P Peters, and Giovanni Baiocchi. Pathways of human development and carbon emissions embodied in trade. *Nature Climate Change*, 2(2):81–85, 2012. doi: https://doi.org/10.1038/nclimate1371.

Chuanwang Sun, Lanyun Chen, and Fan Zhang. Exploring the trading embodied co2 effect and low-carbon globalization from the international division perspective. *Environmental Impact Assessment Review*, 83:106414, 2020a. doi: https://doi.org/10.1016/j.eiar.2020.106414.

Licheng Sun, Lin Qin, Farhad Taghizadeh-Hesary, Jijian Zhang, Muhammad Mohsin, and Imran Sharif Chaudhry. Analyzing carbon emission transfer network structure among provinces in china: new evidence from social network analysis. *Environmental Science and Pollution Research*, 27:23281–23300, 2020b. doi: https://doi.org/10.1007/s11356-020-08911-0.

Wei Sun and Chenchen Huang. Predictions of carbon emission intensity based on factor analysis and an improved extreme learning machine from the perspective of carbon emission efficiency. *Journal of Cleaner Production*, 338:130414, 2022. doi: https://doi.org/10.1016/j.jclepro.2022.130414.

Kailan Tian, Yu Zhang, Yuze Li, Xi Ming, Shangrong Jiang, Hongbo Duan, Cuihong Yang, and Shouyang Wang. Regional trade agreement burdens global carbon emissions mitigation. *Nature communications*, 13(1):408, 2022. doi: https://doi.org/10.1038/s41467-022-28004-5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Song Wang, Yuhuan Zhao, and Thomas Wiedmann. Carbon emissions embodied in china–australia trade: A scenario analysis based on input–output analysis and panel regression models. *Journal of cleaner production*, 220:721–731, 2019. doi: https://doi.org/10.1016/j.jclepro.2019.02.071.

XD Wu, JL Guo, Chaohui Li, GQ Chen, and Xi Ji. Carbon emissions embodied in the global supply chain: Intermediate and final trade imbalances. *Science of the Total Environment*, 707:134670, 2020. doi: https://doi.org/10.1016/j.scitotenv.2019.134670.

Owen Zhang. Tips for data science competitions. `https://datascience.stackexchange.com/questions/10839`, 2016.

Zhonghua Zhang, Lu Xi, Su Bin, Zhao Yuhuan, Wang Song, Liu Ya, Li Hao, Zhang Yongfeng, Ahmad Ashfaq, and Shi Guang. Energy, co2 emissions, and value added flows embodied in the international trade of the brics group: A comprehensive assessment. *Renewable and Sustainable Energy Reviews*, 116:109432, 2019. doi: https://doi.org/10.1016/j.rser.2019.109432.

Xiangtian Zheng, Nan Xu, Loc Trinh, Dongqi Wu, Tong Huang, S Sivaranjani, Yan Liu, and Le Xie. A multi-scale time-series dataset with benchmark for machine learning in decarbonized energy grids. *Scientific Data*, 9(1):359, 2022. doi: https://doi.org/10.1038/s41597-022-01455-7.

Biqing Zhu, Zhu Deng, Xuanren Song, Wenli Zhao, Da Huo, Taochun Sun, Piyu Ke, Duo Cui, Chenxi Lu, Haiwang Zhong, et al. Carbonmonitor-power near-real-time monitoring of global power generation on hourly to daily scales. *Scientific Data*, 10(1):217, 2023. doi: https://doi.org/10.1038/s41597-023-02094-2.

7 Appendix

7.1 EE-MRIO Framework

EE-MRIO contains rich information that can be used to track resource and emission flow in global trading in matrix formation illustrated in Figure 2. EE-MRIO is developed on top of the MRIO framework with consideration of emission accounting, as known as factor accounting table. The MRIO table contains an input-output table recording the monetary resource input and output sectoral dependence, and a demand table describes the final monetary demand of products.

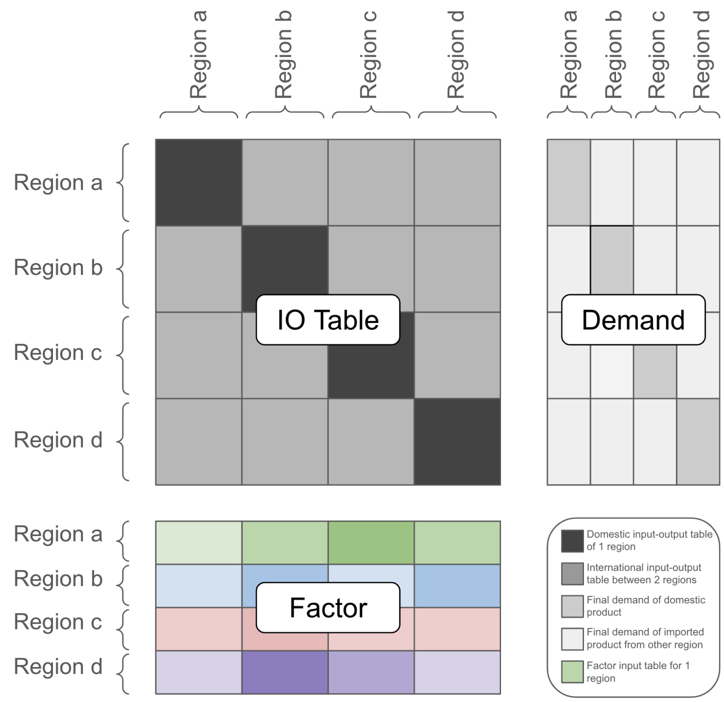

Figure 2: Description of EE-MRIO framework. Multi-regional input-output data tracks the monetary resource transfer of international sectors using the input-output table and demand table. Environmental Extended MRIO considers the environmental impact of economic activities using the factor accounting table. EE-MRIO data are represented in high-dimensional matrix formation. The footprint network could be calculated using an IO table, demand table and factor accounting table.

7.2 Data Structure of ExioML

ExioML contains factor accounting in a tabular format, and a footprint network in the graph structure shares four eco-economic drivers. Factor accounting measures sectoral resource usage, as discussed in the previous section. The footprint network tracks the resource transfer between sectors, and the details are illustrated in Table 4.

Table 4: Feature PxP and IxI dataset in footprint network.

Attribute name	Type	Description
Source	Categorical	Exporting sector
Target	Categorical	Importing sector
Year	Numerical	Timestep (e.g. 1995, 2022)
Value added	Numerical	GDP expressed in millions of Euros
Employment	Numerical	Population engaged in thousands of persons
Energy carrier net total	Numerical	Sum of all energy carriers in Terajoules
GHG emission	Numerical	GHG emissions in kilograms of CO_2 equivalent

The raw ExioBase 3.8.2 contains rich spatiotemporal information, covering 163 sectors in 49 regions for 28 years. Therefore, the number of connections for a single factor in one year is 63,792,169, which consumes vast storage, around 1 GB. ExioML footprint network is a scale-free network, and the weights of edges are followed by power-law distribution (Foster, 2005). Most of the connections are insignificant, around zero. The overall network is sparse, and the connections are concentrated to dominant nodes. To store the multi-dimensional network in a single file, we melt the graph adjacency matrix into an edge table in tabular format. We take the top 1,000,000 connections each year. Around 300,000 edges each year are left after performing inner-join with the selected factors. The total storage consumption for the PxP and IxI footprint network are 1.43 GB and 1.34 GB, respectively.

7.3 Distribution and Correlation Analysis for Factor Account

The inherent skewness in eco-economic data presents a significant challenge for analysis. To address this, employing log-transformation or scaling methods is a critical step in data preprocessing for effective machine learning modelling. Figure 3 illustrates the distribution of numerical features, while Figure 4 depicts the correlation coefficients between variables. An analysis of these figures reveals a strong positive linear Pearson correlation among value added, employment, energy, and emissions. In contrast, feature year demonstrates a lower positive correlation with the other features.

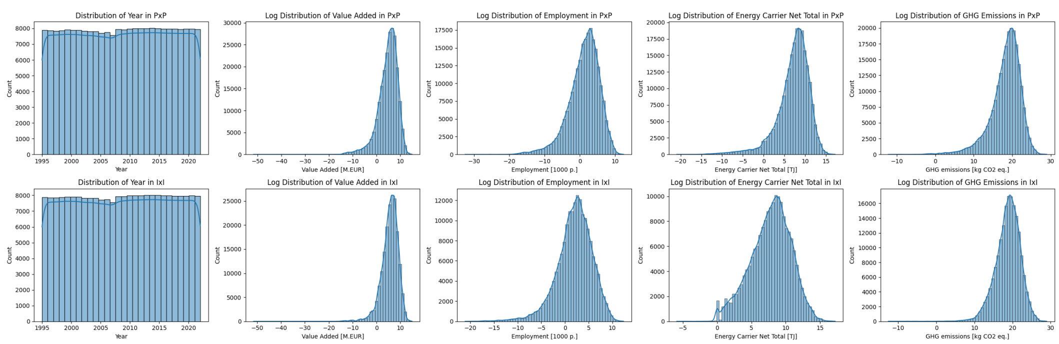

Figure 3: Distribution of factor accounting tabe in PxP and IxI.

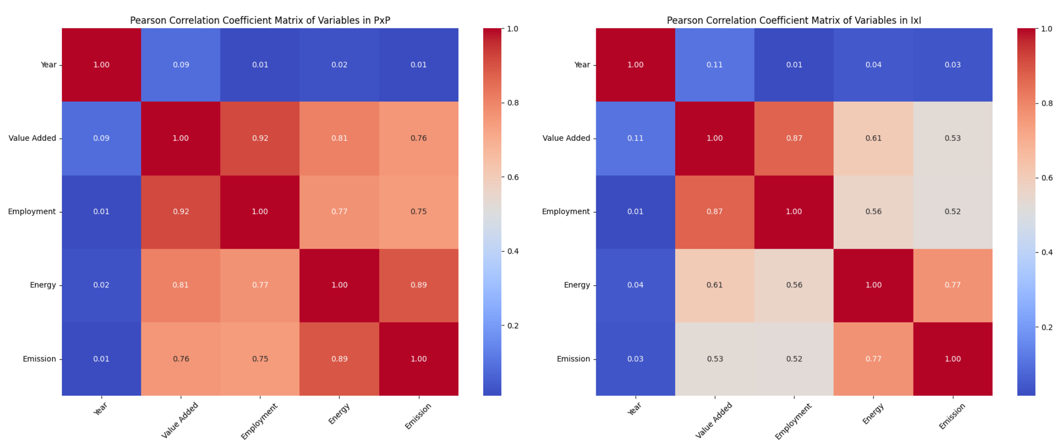

Figure 4: Pearson correlation matrix of factor accounting tabe in PxP and IxI.

7.4 Impletement Details

Algorithms included in GHG emission regression are as follows:

- **K-Nearnest-Neighbour (KNN):** This non-parametric method predicts based on the majority vote from its k nearest neighbours (Cover & Hart, 1967).

- **Ridge Regression:** A linear regression variant employing ℓ_2 regularization to prevent overfitting (Hoerl & Kennard, 1970).
- **Decision Tree (DT):** A feature splitting algorithm to tree flowchart structure for decision making (Quinlan, 1986).
- **Random Forest (RF):** Ensemble of multiple decision trees for robust prediction (Breiman, 2001).
- **Grandient Boost Decision Tree (DBDT):** Ensemble weak trees to improve the error of the previous tree sequentially (Friedman, 2001).
- **Multi-Layer Perceptron (MLP):** A deep artificial neural network employing backpropagation with the capability of non-linear function approximation (LeCun et al., 2015).
- **Gated Adaptive Network (GANDALF):** GANDALF leverages Gated Feature Learning Units (GFLU) for automated feature selection (Joseph & Raj, 2022).
- **Feature-Tokenizer Transformer (FTT):** Novel transformer architecture applies a tokenization mechanism on tabular features (Gorishniy et al., 2021).

The experimental procedures are outlined in Table 5. We divided the data into training, validation, and test sets and applied log transformation for distribution scaling. Shallow models used Leave-one-out Encoding for categorical features. Data normalization was employed for DL models to ensure stable training, although this did not significantly enhance performance. In terms of implementation, we used Sci-Kit Learn for shallow models and PyTorch Tabular for DL models. The MPS framework trained the deep models on Apple M3 Pro GPUs. Interestingly, based on our observations, the models performed better when trained on NVIDIA T4 GPUs with the CUDA framework.

Table 5: Impletement detail.

Data	**Description**	**Setting**	**Description**
# Train	141,635 (PxP), 114,678 (IxI)	Epoch	30
# Validation	35,409 (PxP), 28,670 (IxI)	Batch size	512
# Test	44,262 (PxP), 35,837 (IxI)	Shallow	Sci-kit Learn
Encoding	Shallow models only	Deep	Pytorch Tabular
Normalisation	DL models only	Version	Python 3.11.7
Scaling	Log-transformation	GPU	Apple M3 Pro 18 GB

7.5 Hyperparameter Tuning

A random search of hyperparameters was conducted over 30 trials within a defined parameter grid. The range of parameters tested and the best parameters identified are summarized in Table 6. Given our limited computational resources and time constraints, this grid was heuristically small and based on model characteristics and previous studies to optimize efficiency. The tuning process for each ensemble model took about 10 to 20 minutes per dataset, whereas for each DL model, it took roughly 2 to 3 hours. We plan to include implementing more efficient hyperparameter search algorithms such as Bayesian Optimisation and increasing the number of trials in future.

Table 6: Parameter grid for tuning and best parameter found for PxP and IxI.

Name	**Parameter range**	**Best PxP**	**Best IxI**
KNN			
# Neighbours	Range(1, 50, 2)	5	3
Weights	Uniform, Distance	Distance	Distance
Metric	Euclidean, Manhattan	Manhattan	Manhattan
Ridge			
Alpha	0, 0.001, 0.01, 0.1, 0, 1, 10, 100	0	100
DT			
Max depth	Range(1, 50, 2)	46	46
Min samples leaf	1, 2, 3	3	1
Min samples split	2, 4, 6, 8	4	6
Max features	Sqrt, Log2	Sqrt	Log2
RF			
Max depth	Range(1, 50, 2)	26	26
Min samples leaf	1, 2, 3	1	1
Min samples split	2, 4, 6, 8	2	2
Max features	Sqrt, Log2	Sqrt	Log2
# Estimators	50, 100, 150	100	100
GBDT			
Max depth	Range(1, 50, 2)	16	31
Min samples leaf	1, 2, 3	3	3
Min samples split	2, 4, 6, 8	8	2
Max features	Sqrt, Log2	Log2	Log2
# Estimators	50, 100, 150	100	100
Learning rate	0.01, 0.1, 1	0.1	0.1
MLP			
Layers	256-128-64, 128-64-32, 64-32-16	256-128-64	256-128-64
Dropout	0, 0.05, 0.1	0	0
Learning rate	0.001, 0.01	0.01	0.01
GANDALF			
GFLU stages	1,2,3,4,5	4	4
GFLU dropout	0, 0.05, 0.1	0	0
Feature sparsity	0, 0.05, 0.1	0	0
Learning rate	0.001, 0.01	0.01	0.01
FTT			
# Heads	4, 8	8	8
# Attention blocks	1, 2, 3, 4	2	3
# Multiplier	1, 2, 3, 4	4	1
Learning rate	0.001, 0.01	0.01	0.01