MD Tarek Shikdar

Interpretable Machine Learning Framework for Embodied Carbon Estimation in Reinforced Concrete Wall Elements

School of Technology and Innovations
Sustainable and Autonomous Systems

Vaasa 2026

UNIVERSITY OF VAASA	
School of Technology and Innovations	
Author:	MD Tarek Shikdar
Title of the thesis:	Interpretable Machine Learning Framework for Embodied Carbon Estimation in Reinforced Concrete Wall Elements.
Degree:	Master of Science.
Degree Programme:	Sustainable and Autonomous Systems.
Supervisor:	Mohammed Elmusrati.
Co Supervisor:	Mohamed Noureldin (Assistant Professor Department of Civil Engineering, Structures Group Aalto University, School of Engineering)
Year:	2026 **Pages:** 66

ABSTRACT:

Buildings represent a significant proportion of total anthropogenic greenhouse gas emissions, as operational carbon is gradually reduced by energy efficiency in design. Therefore, embodied carbon produced during the extraction of materials, transportation, manufacture, assembly, disassembly, and disposal of structural components is receiving increasing attention in literature. Conventional Life-Cycle Assessment (LCA) techniques, while comprehensive, have proven to be slow and tedious and cannot be applied repeatedly during early-stage design. This thesis addresses the above gap through the development of a machine learning model capable of predicting the embodied carbon intensity of Reinforced Concrete (RC) walls based on early-stage design inputs.

Synthetic data of a considerable size was created by systematically permuting ten design variables related to wall configuration (geometrical properties, type and strength of concrete used, ratio of reinforcement), and transport distances, along with the results of life cycle carbon calculation using the Inventory of Carbon and Energy (ICE) emission factors and cradle-to-grave system boundary defined by EN 15804.

The dataset includes a variety of structural configurations that can be considered realistic for low-to mid-rise structures in Finland. An XGBoost regression model was trained on the dataset and assessed using regression metrics and five-fold cross-validation. The model interpretability was analyzed using various methods of explanation, namely, Global and Local SHAP, LIME-based local interpretation, Partial Dependence Plots (PDP), Individual Conditional Expectation curves (ICE), and Two-Way PDPs. The above methods were selected to achieve sufficient model interpretability and avoid black-box estimation.

As a result, an XGBoost model was obtained with nearly perfect predictive performance demonstrated across all cross-validation folds. Model interpretability revealed wall thickness to be the most influential design variable, followed by the compressive strength of concrete and reinforcement ratio. This was expected since the above variables define material volumes, as well as associated carbon emission rates and their domination was observed across all methods of model explainability. Transport-related variables exhibited systematic impact to the extent lower than material-related, while wall length and wall height proved to be relatively unimportant for predicting embodied carbon in units of area.

This study proves the applicability of an interpretable ML model for rapid evaluation of multiple alternative configurations at an early stage of design, without the need for complete LCA calculation for each. The current study contributes to the body of literature, as it focuses on predicting embodied carbon on a structural element level rather than material-level and whole-building prediction. For future research, validation of the developed model against empirically collected data and expansion of the framework for predicting the embodied carbon of other RC elements are recommended.

KEYWORDS: embodied carbon, reinforced concrete walls, machine learning, XGBoost, life-cycle assessment, early-stage design, SHAP, interpretability

Contents

Figures

Tables

Abbreviations

BIM	Building Information Modelling
cICE	Centred Individual Conditional Expectation

EN 15804	European Standard for Environmental Product Declarations
EOL	End of Life
ICE	Inventory of Carbon and Energy / Individual Conditional Expectation
LCA	Life-Cycle Assessment
LIME	Local Interpretable Model-agnostic Explanations
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PDP	Partial Dependence Plot
RC	Reinforced Concrete
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
XGBoost	Extreme Gradient Boosting

1 Introduction

Buildings make up a large share of the world's energy consumption and GHG emissions (Khan et al., 2022;Ma et al., 2024). In recent years, much effort has been put into minimizing energy use in building operations, achieving some success. As operational emissions start to be reduced, more emphasis has been placed on embodied emissions from material production and transport to construction sites, as well as demolition once the structure is no longer needed. Embodied emissions are currently seen as an essential element of a building's total GHG footprint and have become an area of focus, especially when it comes to modern high-performance buildings (Myint & Shafique, 2024). Concrete and steel are the largest sources of embodied carbon due to their significant energy consumption during production (Hammond et al., 2011; Khan et al., 2022). The problem with estimating embodied carbon, which relies on established LCA approaches, is that such calculations can be resource-intensive and cannot be reused during various stages of design when crucial choices need to be made. There has been some promising work done in machine learning in this regard (Ghoroghi et al., 2022; El Hafdaoui et al., 2023; Su et al., 2024), and that is the approach this thesis takes, specifically for reinforced concrete wall elements.

1.1 Background

In recent times, carbon emission reduction in the built environment is one of the main goals of sustainable construction. Previously, focus was mainly on operational energy, referring to the energy use by electricity and heating. It was logical since operational emissions were predominant. However, with better insulation and greater use of renewable energy, the role of embodied carbon emissions is gaining more weight. Research has shown that the contribution of embodied carbon to a building's lifetime carbon emissions can be considerable, especially in those buildings with a relatively low level of operational emissions (Ma et al., 2024; Myint & Shafique, 2024).

At this point, concrete and steel play an important part. Both are the most commonly applied structural elements in construction, but the manufacturing of both – especially of concrete (due to cement) and primary steel – is characterized by intensive energy consumption, which produces significant levels of CO_2 (Hammond et al., 2011; Khan et al., 2022). Therefore, decisions regarding the structure will directly affect the amount of embodied carbon emissions due to the amounts of concrete and steel required. Evaluating these decisions on their carbon impacts at the very beginning increases the chances of success.

1.2 Problem Statement

One of the established ways to estimate embodied carbon is life-cycle assessment (LCA). The technique is efficient and reliable but, at the same time, quite slow and requires plenty of comprehensive information (Gervasio & Dimova, 2018; Khan et al., 2022). When designing a structure during its initial stage, when designers are considering all possibilities and choosing between different options, performing LCA for each of them is unrealistic. Therefore, many decisions related to embodied carbon are taken without any sufficient quantitative information to base those choices upon.

The problem is especially relevant for reinforced concrete structures, since at an early stage of design decisions about thickness of walls, strength of concrete, and number of reinforcements are often irreversible, but cannot be estimated in a fast manner (Hammond et al., 2011; Ma et al., 2024). One of the ways to solve the problem is machine learning: it allows to predict the effect of specific design features on the resulting embodied carbon, making predictions available almost instantly (Ghoroghi et al., 2022; El Hafdaoui et al., 2023; Su et al., 2024). However, while there are studies related to LCA using ML algorithms, they mostly deal with buildings as a whole or materials. Wall elements of reinforced concrete structures are not widely studied.

1.3 Research Aim and Objectives

1.3.1 Aim

The goal of this thesis is to develop a machine learning model for forecasting the carbon intensity embodied in reinforced concrete walls, using input data that is obtainable from the early stages of structural design.

1.3.2 Objectives

To achieve this aim, the thesis pursues the following objectives:

- To review existing literature on LCA, embodied carbon in buildings, reinforced concrete elements, and machine learning applications in carbon prediction.
- To generate a dataset of RC wall configurations with corresponding embodied carbon values using established emission factors and life-cycle calculation equations.
- To develop and train a machine learning model for predicting the embodied carbon intensity of RC wall elements.
- To evaluate the predictive performance, accuracy, and interpretability of the selected model.
- To discuss what the results mean for early-stage structural design decision-making.

1.3.3 Research Questions

To guide the research, the following questions are addressed:

- Which RC wall design parameters are most influential in determining embodied carbon intensity?

- Can machine learning predict RC wall embodied carbon intensity with high accuracy using basic early-stage inputs?
- How can interpretability methods help explain the influence of RC wall parameters on embodied carbon predictions?
- How can the proposed model support early-stage structural decision-making for RC walls?

1.3.4 Scope and Limitations

The thesis focuses on RC walls. The model uses early design parameters such as wall geometry, concrete strength, steel content, transportation distance, and disposal efficiency to estimate the embodied carbon intensity of structures. The data is artificially produced based on parametric values of these parameters and life cycle calculations, suggesting that the model is more likely a comparative design tool than a highly accurate estimation tool. The primary constraints include assumptions used in some life cycle stages and lack of empirical data from reality.

1.3.5 Structure of the thesis

Chapter 1 - Introduction: This chapter presents an overview of the topic, problem statement, purpose, goals, research questions, scope, and structure of the thesis.

Chapter 2 - Literature Review: This chapter reviews the literature on life cycle assessment, carbon embodied within buildings, structural design parameters of reinforced concrete structures, and use of machine learning for carbon prediction.

Chapter 3 - Methodology: This chapter describes the dataset used, system boundaries, functional unit, input parameters, modeling process, model training, and results evaluation.

Chapter 4 - Results: This chapter presents the results for model accuracy, interpretation of results, sensitivity analysis, and interaction analysis.

Chapter 5 - Discussion: This chapter discusses the results, their relevance to structural design at an early stage, positions the results considering previous literature, and outlines the limitations of the study.

Chapter 6 - Conclusion: This chapter concludes the thesis by highlighting the main results, contributions, limitations, and suggestions for future research.

2 LITERATURE REVIEW

2.1 Introduction to the Literature Review

The current chapter explores the existing literature in relation to life-cycle assessment, embodied carbon in buildings, design parameters of reinforced concrete members, and machine learning for embodied carbon estimation. This review aims at a targeted examination of a focused set of issues within the field of sustainable construction. The specific combination of issues discussed herein lies behind the motivation for this study. The importance of embodied carbon in reinforced concrete (RC) wall members, existing approaches, and the opportunity to employ machine learning in order to provide additional value form the focus of the review. Concrete and steel form some of the most prominent sources of embodied carbon in today's construction industry (Hammond et al., 2011; Khan et al., 2022; Ma et al., 2024). Consequently, RC wall members serve as a meaningful unit of analysis.

2.2 Life-Cycle Assessment (LCA) in Buildings

The life cycle assessment (LCA) is the established procedure for assessing the environmental impacts of building through the entirety of its life cycle. As noted by Gervasio and Dimova (2018), the European Standards approach divides this procedure into four modules, including the production stage (A1–A3), the construction stage (A4–A5), the use stage (B1–B7), the end-of-life stage (C1–C4), as well as the Module D with extra benefits outside the system boundaries.

Several LCA methods exist. According to Khan et al. (2022), process-based LCA, hybrid LCA, and input–output LCA serve as the main methodologies used within embodied carbon analysis. The most used approach is the process-based LCA due to the possibility to receive material-level information, although it might overlook the upstream impacts. Hybrid methods attempt to compensate for this limitation by combining process and economic modeling. In any case, there are considerable differences between the results provided by different research due to different system boundaries, databases, and impact methods (Gervasio & Dimova, 2018; (Khan et al., 2022).

However, the main problem with using the LCA is its data intensity. Although this aspect can be appropriate for the verification of already constructed buildings, it becomes a considerable problem for architects at the preliminary stages. At these stages, the assessment of multiple alternatives must be completed quickly without having access to sufficient information. Moreover, according to Pomponi & Moncaster (2016), more than 90% of all research in the systematic analysis of 102 LCA studies were limited only to the manufacturing stage (e.g. A1-A3); thus, their results had significantly lower values of embodied carbon when compared to other studies with more extended boundaries. In this way, the comparison of the research results poses additional problems.

2.3 Embodied Carbon in Buildings

While energy efficiency has increased, the percentage of emissions attributed to the actual construction of buildings versus their operation has greatly increased. The term embodied carbon refers to greenhouse gas emissions related to producing, delivering, maintaining, and eventually tearing down the building material (Khan et al., 2022). Historically, embodied carbon has been viewed as a secondary issue compared to operational energy use but is now considered differently. Röck et al. (2020), conducting a comprehensive review of over 650 building LCA studies, discovered that while GHG emissions across life cycles decrease because of improvements in operational performance, embodied emissions have increased both absolutely and relatively. Now, in highly energy-efficient buildings, embodied carbon represents 45-50% of total life cycle emissions and can reach above 90% in some instances (Röck et al., 2020). This makes embodied carbon one of the most critical issues facing buildings today from a climate perspective.

Khan et al. (2022) explain that the embodied carbon can represent a significant proportion of the total life cycle emissions of a building, especially for highly efficient or low-energy ones. Similarly, Myint and Shafique (2024) and Ma et al. (2024) illustrate with case studies how structural and envelope materials can result in carbon footprints equal to or higher than the operational emissions, depending on various assumptions made for the building design and energy source mixes in regions.

Another issue about embodied carbon is its variability from study to study, resulting from variations in system boundary, database, type of buildings under investigation, and other model assumptions (Gervasio & Dimova, 2018; Khan et al., 2022). Pomponi and Moncaster (2016), conducting a systematic literature review on mitigation strategies for embodied carbon, found 17 such different strategies and concluded that none of them was adequate individually. Mitigating embodied carbon would require a holistic approach combining design improvement, low-carbon materials and reuse, and strong policy drivers.

Concrete and steel are two major materials that play a crucial role in the problem of embodied carbon. These are among the most common structural components and are characterized by a very energy-intensive manufacturing process and release of a large amount of CO_2 (Hammond & Jones, 2011; Khan et al., 2022). According to Ma et al. (2024), in a high-rise building case study, concrete alone represented 74% of total embodied carbon emissions across the life cycle. Emissions factors for these materials are well-known and available, for instance, in ICE database (Hammond & Jones, 2011).

A few authors stress that the best time to affect the embodied carbon issue is when the building is still under design development. Indeed, according to Ma et al. (2024) and Myint and Shafique (2024), after selecting certain structural components, the possibilities for lowering embodied carbon become limited. It emphasizes the importance of quick and reliable embodied carbon evaluation tools for the design phase, specifically for reinforced concrete wall elements.

2.4 Structural Design Parameters and Embodied Carbon in Reinforced Concrete Elements

First, the decisions made about structural design play a key role in determining embodied carbon because they determine material quantities, transportation needs, and disposal strategies. The impact of this variable is well supported in scientific literature. For example, comparing different structural systems, Moussavi Nadoushani and Akbarnezhad (2015) found that due to variations in material consumption and design factors, the carbon footprint could be very different and confirmed the dependence of embodied carbon not only on material but also structural design.

In addition, the emergence of digital design solutions has made it possible to perform such evaluations at an earlier stage. Specifically, Hunt and Osorio-Sandoval (2023) introduced a BIM-based prototype to evaluate the embodied carbon of structural designs and showed how to identify carbon hotspots and assess the effects of design changes during preliminary evaluation phases. Likewise, Ma et al. (2024) conducted a BIM-based life-cycle analysis of the embodied carbon in a high-rise structure and concluded that embodied carbon in terms of structural systems was driven by material consumption, particularly the use of concrete. Thus, all studies conclude that structural design plays a key role in shaping embodied carbon, and evaluating its effects early brings more benefits.

Regarding the factors affecting the embodiment of carbon in reinforced concrete structures, the main influencing factors can be geometry, the strength of concrete, the amount of reinforcement, and transportation assumptions. All these variables are directly related to the consumption of materials used, i.e., they define the amount of concrete and steel necessary for structural systems, which is the largest share of embodied carbon across all lifecycle stages. In the case of reinforced concrete walls, the following parameters, in particular, wall thickness, geometrical parameters, reinforcement ratio, and concrete strength will play a crucial role.

At the same time, the existing research is mainly devoted to comparing the carbon footprints of buildings, conducting system analysis, or performing BIM calculations of projects (Hunt & Osorio-Sandoval, 2023; Moussavi Nadoushani & Akbarnezhad, 2015; Ma et al., 2024). Generalizable and interpretable models, capable of predicting carbon embodied in reinforced concrete walls as separate objects, remain under-researched. This constitutes a notable deficiency considering the wide spread of these types of walls in low- and medium-rise buildings.

2.5 Machine Learning in Life-Cycle Assessment

Machine learning algorithms have been increasingly used for life cycle assessment and embodied carbon calculations, especially in situations involving extensive and diverse datasets. According to a review conducted by Ghoroghi et al. (2022), machine learning algorithms can decrease computational burden and allow fast prediction of environmental indicators, including global warming potential. El Hafdaoui et al. (2023) and Su et al. (2024) highlight another advantage associated with machine learning approaches, namely, the ability of algorithms to discover nonlinear dependencies between design variables and embodied carbon levels, thus outperforming linear regression.

However, there is one more aspect that should be discussed. Machine learning approaches demonstrate excellent predictive capabilities, but it can be quite challenging to interpret results from the engineering perspective because models are hard to explain (El Hafdaoui et al., 2023; Ghoroghi et al., 2022). Hence, interpretability is another essential characteristic of ML algorithms for embodied carbon predictions, especially for cases when results are used in early design decision making. Finally, it is important to note the benefits of artificial datasets in ML-related studies.

(Iqbal & Noureldin, 2025)) proposed a machine learning framework for estimating embodied carbon in reinforced concrete structural systems based on a large artificial dataset produced using life cycle assessment equations. The authors found out that algorithms based on gradient boosting, such as XGBoost, could achieve a very high level of predictive accuracy on this type of dataset, thus demonstrating that parametric LCA combined with machine learning can be effectively used for embodied carbon assessment in early stages of design.

2.6 Research Gaps and Opportunities

Several gaps have been found from the literature search performed for this dissertation.

The first gap is that while LCA is a proven approach, it is not only very time-consuming and data-intensive but also involves a comparison of several alternatives (which is common during early stage of building design) making it impractical for repeated application in early stages of design (Gervasio & Dimova, 2018; Khan et al., 2022). Therefore, the need for alternative approaches emerges that would allow for faster prediction of embodied carbon, but without replacing LCA altogether.

The second gap relates to the focus of existing studies. While more studies recognize embodied carbon as a major issue, they often focus on whole buildings, building systems, or materials themselves and seldom focus specifically on the structural parts of a building or individual building elements. According to Pomponi and Moncaster (2016), the great majority of LCA literature related to the built environment deals with the building itself, or a single system or element, but does not include individual structural components as separate units used for prediction. Moreover, there are not many studies focusing specifically on concrete wall elements and the role played by early design decisions in their embodied carbon (Hunt & Osorio-Sandoval, 2023; (Ma et al., 2024; Moussavi Nadoushani & Akbarnezhad, 2015; Ma et al., 2024).

A third important gap relates to predictability and transparency of machine learning models used in LCA studies. Despite high accuracy, many studies produce black-box models with poor explanation (El Hafdaoui et al., 2023; Ghoroghi et al., 2022; Su et al., 2024). In other words, while highly accurate, such models lack explainability which makes them hard to understand and use in practice. Thus, there is a need to create models that not only have high accuracy but are also easy to understand and interpret.

Finally, one may point out the limitations associated with lack of relevant data. Most existing studies face problems with generalizing their results due to small sample size and narrow contexts in which data was collected (Ma et al., 2024; Myint & Shafique, 2024). This highlights the need for approaches that allow creating large synthetic databases based on parametric data generation.

Thus, the current dissertation seeks to address these four gaps with a machine learning-based approach for predicting embodied carbon intensity of reinforced concrete wall elements based on parametrically generated data.

3 Methodology

3.1 Overall Research Workflow

In this section, an ML framework for predicting the life-cycle embodied carbon intensity of reinforced concrete (RC) wall elements is proposed. This involves combining traditional LCA modelling with ML approaches for estimating the environmental impact related to structural design variations.

This study involves several research activities. First, the structural design variables of RC wall elements are identified, which include wall geometry, materials properties, and transportation distances. A dataset containing different types of wall elements is generated by performing permutations of the above-mentioned variables in Python.

Next, life-cycle embodied carbon is calculated for each type of wall element in the dataset based on the corresponding LCA equations covering material production, transportation, construction and deconstruction phases. After that, an ML model is built using the obtained dataset for predicting the life-cycle embodied carbon of the given inputs. The selected ML model (XGBoost) will be assessed using different metrics to evaluate its performance and interpret the effects of each variable on predicted carbon values. The general flow of this research can be summarized using Figure 3.1.

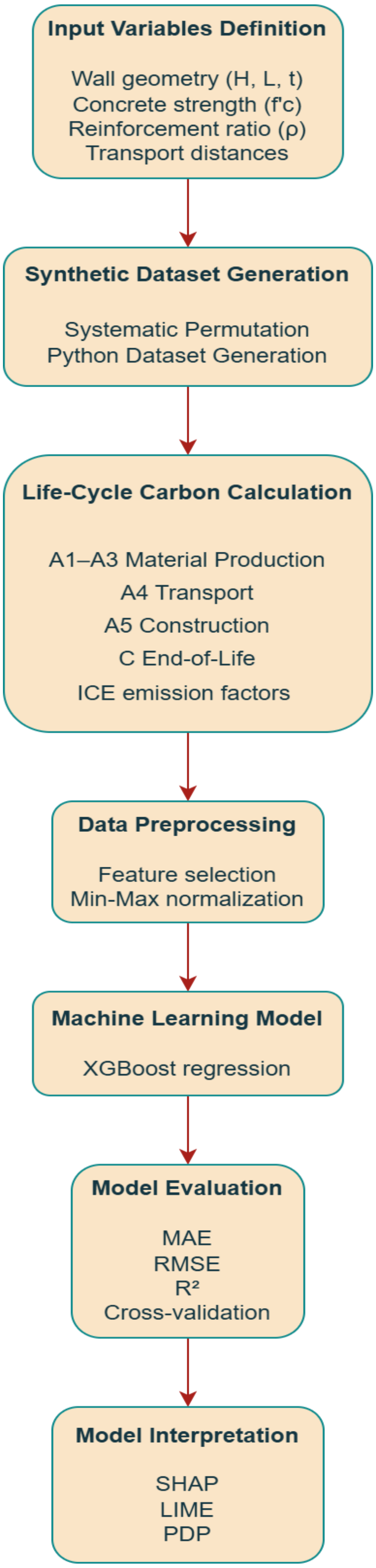

Figure 3.1 Overall research workflow for predicting the embodied carbon intensity of reinforced concrete wall elements using life-cycle assessment and machine learning.

3.2 Data Sources and Dataset Description

In this part, information about the data source and dataset creation processes will be provided. The proposed machine learning model for prediction of carbon footprint of the embodied carbon intensity of reinforced concrete walls will use a dataset that will include structural design parameters combined with life cycle emissions estimation based on reputable emission factor databases.

The dataset will represent reinforced concrete (RC) wall elements that may be utilized in low-rise to mid-rise buildings in Finland and Europe in general. Concrete and reinforcing steel represent the main constituents of such structures, whose production contributes to carbon footprint of building materials significantly. To create a relevant dataset for the purpose of machine learning, a synthetic approach was chosen, since life cycle assessment datasets on RC wall elements do not exist in sufficient quantity for such application.

A data synthesis approach is used in another research works as well. One of them is the study by Iqbal and Noureldin (2025). In that work, the two authors generated a tremendous amount of synthesized data for use in applying machine learning to analysis of reinforced concrete structures.

3.2.1 Dataset Scope and Boundaries

The database includes RC wall models reflecting ordinary buildings' construction. The walls' heights are between 2.6 and 3.6 meters, lengths are 2 to 8 meters, and thickness are from 150 to 300 millimeters. The reinforcement is between 80 and 180 kilograms per cubic meter, whereas the concrete compressive strength is between 20 and 50 MPa. All these numbers reflect common structural design data in Finland and Europe.

The system boundaries selected for this research are cradle-to-grave based and compatible with EN 15804 guidelines. The life cycle stages involved in the calculations include:

- A1–A3 (Product stage): Raw material extraction and manufacturing of concrete and steel reinforcement.
- A4 (Transport stage): Transportation of materials from production plants to the construction site.
- A5 (Construction stage): Construction activities during wall installation.
- C (End-of-life stage): Demolition processes and transportation of waste materials.

Emission factors for material production are obtained from the Inventory of Carbon and Energy (ICE) database (Hammond & Jones, 2011). Concrete emission factors are assigned based on compressive strength classes, while steel reinforcement emissions are based on typical rebar production values. Transport emissions are estimated using representative heavy goods vehicle emission factors commonly applied in European LCA studies.

3.2.2 Dataset Generation

All these parameters mentioned in section 3.4 varied in a systematic way using the itertools.product method from python. Based on this approach, a synthesized dataset consisting of 2,903,040 combinations of the reinforced concrete wall is generated. Embodied carbon estimation in all stages of the lifecycle of the reinforced concrete wall is done for each combination separately.

Calculation of embodied carbon at all lifecycle stages is done using basic lifecycle assessment theory and emission factors derived from the Inventory of Carbon and Energy (ICE) database (Hammond & Jones, 2011). Emissions at each stage are calculated by:

A1–A3 (Material Production):

Concrete mass (kg) is calculated as:

$$\boldsymbol{m_{conc} = V \times 2350} \tag{1}$$

Steel mass (kg) is calculated as:

$$m_{steel} = V \times \rho \tag{2}$$

The embodied carbon from material production is calculated as:

$$EC_{A1-A3} = \left(m_{conc} \times EF_{conc} \times \left(1 + \frac{w_{conc}}{100}\right) + \left(m_{steel} \times 0.77 \times \left(1 + \frac{w_{steel}}{100}\right)\right)\right. \tag{3}$$

where:

m_{conc} = mass of concrete (kg)

m_{steel} = mass of steel reinforcement (kg)

V = volume of the wall (m^3)

ρ = reinforcement ratio (kg/m^3)

EF_{conc} = emission factor of concrete ($kgCO_2e/kg$)

w_{conc} = concrete waste rate (%)

w_{steel} = steel waste rate (%)

A4 (Transport):

Transport emissions are calculated as:

$$EC_{A4} = \left(\frac{(m_{conc} + m_{steel})}{1000}\right) \times D_{conc} \times 0.11 \tag{4}$$

where:

D_{conc} = transport distance from production plant to construction site (km)

0.11 = emission factor for heavy goods transport ($kgCO_2e/tkm$)

A5 (Construction):

Construction emissions are estimated as:

$$EC_{A5} = V \times 15.0 \times SF \tag{5}$$

The slump factor is defined as:

$$SF = 1.1 - (\frac{S - 12}{9}) \times 0.2 \tag{6}$$

where:

V = volume of the wall (m³)

SF = slump factor (–)

S = slump value (cm)

The slump factor ranges from 1.1 at S = 12 cm to 0.9 at S = 21 cm, reflecting that higher slump concrete requires less mechanical compaction effort, which reduces construction-phase equipment emissions.

C (End-of-Life):

End-of-life emissions are calculated as:

$$EC_C = (V \times 5.0) + \left(\frac{(m_{conc} + m_{steel})}{1000}\right) \times D_{EOL} \times 0.11 \tag{7}$$

where:

V = volume of the wall (m³)

m_{conc} = mass of concrete (kg)

m_{steel} = mass of steel reinforcement (kg)

D_{EOL} = transport distance for demolished materials (km)

The 5.0 $kgCO_2e/m^3$ value represents a conservative proxy for demolition energy consumption.

Through this procedure, the synthetic dataset captures a wide range of realistic reinforced concrete wall configurations and associated life-cycle carbon emissions, providing a consistent and structured dataset for training the machine learning model.

3.2.3 Data Characteristics

The generated dataset consists of 2,903,040 samples where each of them corresponds to one specific configuration of a reinforced concrete wall characterized by different combinations of input features described in section 3.4. Life-cycle embodied carbon of each configuration is calculated for each life-cycle phase under consideration and converted to the value of embodied carbon intensity measured in $kgCO_2e$ per square meter of wall area.

Statistical analysis shows that the mean value of embodied carbon intensity equals approximately 102.6 $kgCO_2e/m^2$ and varies from 50.3 $kgCO_2e/m^2$ to 182.8 $kgCO_2e/m^2$. The average value of embodied carbon measured in total amount for one wall configuration equals approximately 1,591 $kgCO_2e$. Also, the mean value of concrete compressive strength equal to 35 MPa (which is located precisely in the middle of a strength interval of 20–50 MPa) and the mean transport distance to the site of construction equal to 65 km are determined.

Thus, it is clear that the presented dataset includes a large number of realistic structural configurations, including variation in wall dimensions, characteristics of materials used, ratios of reinforcement, as well as transport distances required for construction. This allows us to say that the dataset can be efficiently applied in machine learning algorithms due to its diversity.

Finally, it should be mentioned that the relationship between input features and output results is determined directly from life-cycle equations, not empirically.

3.3 System Boundaries and Functional Unit

In order to achieve comparability and make sure that the analysis is relevant in terms of informing structural design decision-making at an early stage, the system boundary considered in this study includes cradle-to-grave carbon emissions of reinforced concrete (RC) wall elements. In addition, operational emissions related to building use phase (B-stage) are not considered, as they are heavily dependent on building operations and cannot be influenced by structural material choice (Iqbal & Noureldin, 2025). Life-cycle phases that have been taken into account are the following:

- A1–A3: Raw materials extraction, processing, and manufacturing of concrete and steel reinforcements.
- A4: Transportation of materials from the manufacturing facilities to the construction site.
- A5: Construction phase activities performed on-site, including placement of materials and operation of machinery, modeled using a proxy based on concrete slump.
- C: End-of-life phase, including demolition and transport of demolition waste to landfill/recycling facilities.

The system boundary used here is compliant with the life cycle phases specified in EN 15804:2012+A2:2019 and is consistent with the phases included in this thesis. Use phase (B-stage) and maintenance emissions are not considered, since the aim of this study is to assess the impact of embodied carbon emissions caused by structural material choice/design considerations.

The functional unit used in this study is the embodied carbon intensity per square meter of wall surface area, expressed in $kgCO_2e/m^2$. The value is calculated as:

$$EC\_intensity = \frac{Total\ Embodied\ Carbon}{Wall\ Area} \tag{8}$$

The wall area is defined as:

$$Wall\ Area = H \times L \tag{9}$$

where:

H = wall height (m)

L = wall length (m)

The functional unit employed in this study allows comparing different wall designs irrespective of their size. This means that variations in geometry, material strength, or reinforcement ratio can be assessed consistently at the early stage of structural design.

3.4 Input Variables for Reinforced Concrete Wall Elements

Input variables used by the machine learning model to be developed in this study describe important design parameters of reinforced concrete (RC) wall elements. Such input variables will reflect the geometric and material properties of the structure, and transportation parameters of RC walls that affect the structure's embodied carbon. Variable selection will follow current practices in structural design, and life-cycle assessments (LCA).

As can be seen in Table 3.1, the selected input variables fall into four major categories: geometrical parameters; parameters describing material properties of the structural element; parameters reflecting transport distances; and parameters related to construction processes. Parameters describing wall geometry include wall height, length, and thickness. Material parameters include concrete compressive strength and reinforcement ratio, which determine the volume of material used. Parameters related to transportation include distances between production plants and construction sites, which define emissions resulting from the transport process at the A4 life-cycle phase. Finally, construction-related parameters include concrete slump and material waste rates.

Table 3.1 summarizes the list of variables used to generate the dataset, the corresponding symbols, units, and value ranges.

Table 3.1 Input variables used for reinforced concrete wall configurations and their ranges applied in the dataset generation process.

Category	Variable	Symbol/Unit	Range / Values
Design/Geometry	Wall height	H (m)	2.6–3.6, step 0.2 m (6 values)
	Wall length	L (m)	2–8, step 1 m (7 values)
	Wall thickness	t (mm)	150, 200, 250, 300 (4 values)
Concrete	Compressive strength	f'_c (MPa)	20, 30, 40, 50 (4 values)
	Slump value	S (cm)	12, 15, 18, 21 (4 values)
	Waste rate	w_{conc} (%)	0, 2, 5 (3 values)
Reinforcement	Ratio	ρ (kg/m³)	80–180, step 20 kg/m³ (6 values)
	Waste rate	w_{steel} (%)	0, 2, 5 (3 values)
Transport (A4)	Plant-to-site distance	D_{conc} (km)	5, 20, 50, 100, 150 (5 values)
End-of-Life (C)	Demolition transport distance	D_{EOL} (km)	10, 30, 60, 100 (4 values)

As can be seen in Table 3.1, all values assumed by the variables in the problem are within realistic value ranges. For instance, wall thickness ranges from 150 to 300 mm and reflect typical RC wall sections used in low-rise to mid-rise constructions. Typical compressive strengths of structural concrete are included in the range considered (from 20 MPa to 50 MPa). Moreover, reinforcement ratio ranges from 1% to 6%, corresponding to typical reinforcement in RC structural walls.

It should be noted that some parameters, which are necessary for the calculation of carbon emissions, are calculated from input variables and are not considered input features themselves. For instance, concrete volume in the wall is calculated as follows:

$$V = H \times L \times \left(\frac{t}{1000}\right) \tag{10}$$

Moreover, the mass of reinforcement steel used can be found through the equation that includes the reinforcement ratio.

3.5 Machine Learning Model Selection

The problem of estimating the life-cycle embodied carbon intensity of reinforced concrete wall elements is formulated as a regression problem. Therefore, a supervised machine learning approach can be used to predict the target variable depending on structural parameters. The main task of the model is to identify the correlation between input variables and the embodied carbon intensity.

Among all machine learning algorithms capable of solving regression problems, XGBoost is chosen as the baseline model for this work. XGBoost is an ensemble-based regression approach known as extreme gradient boosting. This algorithm combines a sequence of decision trees, where each new one reduces the prediction error of previous iterations. Thus, XGBoost can learn complex relationships between input features, which is critical for engineering applications where multiple parameters may interact with each other.

Another benefit of XGBoost is that this method works efficiently with large and structured datasets. It is beneficial in this case since the artificial dataset created for this study has many rows. Moreover, the algorithm includes several regularization techniques that prevent overfitting and allow obtaining more generalized models. Finally, XGBoost allows measuring feature importances, which contributes to interpretability, which is necessary for engineering purposes. These characteristics of XGBoost were among the reasons why it was utilized successfully in numerous studies combining machine learning with life-cycle assessment of construction systems (El Hafdaoui et al., 2023; Su et al., 2024).

Thus, in the current work, the input variables specified in Section 3.4 are used to train a regression model based on XGBoost. The output will be the life-cycle embodied carbon intensity of reinforced concrete wall elements expressed in $kgCO_2e/m^2$. The target variable equals the value of the corresponding parameter obtained while generating the dataset described in Section 3.2. The applicability of the algorithm compared to other options is further confirmed by a comparative study carried out in Section 4.3.

3.6 Model Training and Validation Approach

This section explains the processes that were utilized in data pre-processing, data splitting, training of the machine learning model, and its evaluation, which was done for predicting the embodied carbon intensity of RC wall elements. The modelling procedure is conducted in line with the framework provided by Iqbal and Noureldin (2025). The entire process will be performed using Python programming language and scikit-learn and XGBoost libraries.

The main purpose of the process is to make sure that the machine learning model is tested on the basis of representative data. This is important because the dataset used in this study is artificially constructed. Therefore, special emphasis should be placed on ensuring consistent distributions in both training and testing datasets.

The steps involved in the modelling process are discussed in more detail below.

3.6.1 Data Preprocessing and Normalisation

Before training the model, all input variables and the target variable are scaled to the interval [0, 1] using Min–Max normalisation. The Min–Max transformation is defined as:

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{11}$$

where:

x = original value of the feature
x_{min} = minimum value of the feature in the dataset
x_{max} = maximum value of the feature in the dataset
x' = normalised value of the feature

This step makes sure that the features with different units and magnitudes are scaled properly to prevent the domination of those features having large numeric scales during training. This normalization approach is also employed within the modeling framework proposed by Iqbal and Noureldin (2025).

In order not to have any data leakage during modeling, only the original input design variables should be considered as input model features. Calculated values like concrete volume, wall area, material masses, phase-wise CO2 emissions, and total embodied carbon should not be considered model features since they can be computed based on the target variable value. Thus, the inclusion of such quantities would give the model an indirect way to know the right solution, thereby increasing the prediction accuracy of the model artificially.

Hence, the ten model input features are: concrete compressive strength (f'_c, MPa); wall height (H, m); wall length (L, m); wall thickness (t, mm); reinforcement ratio (ρ, kg/m^3); concrete slump (S, cm); transport distance to construction site (D_{conc}, km); concrete waste rate (w_{conc}, %); steel waste rate (w_{steel}, %); and demolition transportation distance (D_{EOL}, km).

3.6.2 Train–Test Split

The entire data, comprising of 2,903,040 observations, is split between the training and testing set in the ratio of 70/30. A predetermined value of random state (random state = 0) is used to maintain consistency in the results. This leads to 2,032,128 observations being utilized for training purposes while 870,912 observations are left aside for testing. The same method is used by Iqbal and Noureldin (2025) and provides enough observations in the testing data for conducting an appropriate analysis based on its size and structure.

3.6.3 Model Training

The process of model training on the training dataset takes place using the XGBoost regression model with default hyperparameters. In this connection, the random state parameter is equal to zero (0). No optimization of parameters and early stopping strategy is applied at this point, and that follows the baseline model building approach proposed by Iqbal and Noureldin (2025).

It should be noted that training of the model proceeded without difficulties with convergence. This means that the default configuration of the model fits the dataset.

3.6.4 Cross-Validation

To test the robustness of the model and to prove that the result of training is not affected by train-test splitting, five-fold cross-validation is carried out. This technique implies dividing the data set into five equally sized folds and training the model using four folds with testing on the fifth one; repeating this step five times so that each time each fold is tested once.

R^2 values received for all folds during the cross-validation procedure are 0.99995, 0.99995, 0.99994, 0.99997, and 0.99995 respectively, which leads to the final cross-validation value of $R^2 = 0.999951 \pm 0.000009$. As can be seen from the standard deviation, the result obtained does not vary greatly from fold to fold. It can be considered that model performance is stable.

In this case the reason for such high consistency in results is the deterministic way of generating data with explicit relations between the input variables and the target one.

3.7 Performance and Interpretability Metrics

This section presents the quantitative performance metrics and interpretability methods used to evaluate and explain the trained XGBoost model. Predictive performance is assessed using standard regression metrics calculated on the held-out test dataset. Model interpretability is examined using four complementary approaches: XGBoost built-in feature importance, SHAP-based global and local explanations, LIME for local interpretation, and Partial Dependence Plots (PDP), including Individual Conditional Expectation (ICE) curves and selected two-way interaction plots. The methods described in this section are used to generate the evaluation and interpretability results reported in Chapter 4.

3.7.1 Regression Performance Metrics

The prediction results are assessed using the four widely accepted metrics used for measuring regression performance, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The computation of all performance results uses normalized target variables.

These performance metrics can give additional insights into the performance of a machine learning model. MAE represents the average difference between the real and predicted values of the target variable, MSE assigns more weights to prediction errors, RMSE presents the average value of prediction errors on the scale of normalized target variable, and finally, the coefficient of determination shows how much of variance in the target variable is described by the proposed machine learning model.

3.7.2 XGBoost Feature Importance

XGBoost includes a built-in feature importance measure that evaluates the contribution of each input variable based on the total gain produced by that feature across all decision tree splits. In this study, feature importance is used as an initial interpretability tool to identify which variables contribute most strongly to the model predictions.

The gain-based importance measure provides a global overview of variable influence within the trained model. This makes it useful for examining how the model distributes predictive importance across structural design parameters and for identifying variables that have the greatest overall contribution to embodied carbon prediction.

3.7.3 SHAP Global Interpretability

The SHAP (SHapley Additive exPlanations) values are applied to provide a sound theoretical basis for the contribution made by each of the variables to predictions made by the model. The idea of SHAP is that each variable is treated as a contributor to the result in line with game theory principles. For this analysis, the SHAP values were calculated by TreeExplainer as exact methods for decision tree models.

To achieve computational efficiency without compromising on the sample's representativeness, the SHAP analysis was carried out on a subsample of 50,000 records from the dataset. The global SHAP analysis was conducted to explore the importance and the direction of impact made by the input variables throughout the dataset. Specifically, the SHAP beeswarm plot allows us to observe the contribution made by the features towards the predicted values of embodied carbon intensity.

3.7.4 SHAP Local Interpretability

Apart from the global SHAP method analysis, another form of SHAP analysis that is applied in the examination of the contribution of individual variables to model prediction is the local SHAP method. Local SHAP waterfall plots are plotted to carry out the analysis using samples. In this analysis, model prediction is compared to a base value, and each input contributes to the prediction either positively or negatively.

The advantage of local SHAP analysis in predicting the model is that it helps in examining specific cases because it allows the examination of the joint effect of different input features on the prediction. This contrasts with the global SHAP analysis explained in Section 3.7.3.

In this way, model predictions can be analyzed based on the contribution of combinations of input variables.

3.7.5 LIME Local Interpretability

To complement the SHAP-based local analysis described in Section 3.7.4, Local Interpretable Model-agnostic Explanations (LIME) is used as an additional local interpretability method. LIME explains an individual prediction by fitting a simple surrogate model in the neighbourhood of that prediction, thereby approximating the local behaviour of the trained model.

In this study, LIME is applied to representative test samples to examine how individual input variables contribute to a specific predicted embodied carbon intensity value. The method produces a local feature-weight explanation, indicating whether each variable contributes positively or negatively relative to the local baseline prediction.

The use of LIME provides an additional model-agnostic perspective on local prediction behaviour and supports comparison with the SHAP-based local explanations presented in the previous subsection. In this way, local interpretability is examined through two complementary approaches, enabling a more robust assessment of how individual predictions are formed.

3.7.6 Partial Dependence Plots and ICE Curves

Apart from investigating model behavior using measures of global and local feature importance, the Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) curves techniques are implemented in the paper. The Partial Dependence Plots are applied to plot the average marginal impact of the input predictor on the predicted Embodied Carbon Intensity of buildings in the dataset, while other variables are kept constant. In this paper, the univariate Partial Dependence Plots are created for all the ten input features to explore the changes in Embodied Carbon Intensity as a function of each input variable separately.

As a complement to one-way PDP plots, Individual Conditional Expectation (ICE) curves are plotted for the most impactful input features revealed via feature importance analysis. In contrast to PDPs, which give average responses across all the observations, ICE curves represent the prediction response for each individual observation when a certain feature is changed. Thus, the possible variability in model behavior related to heterogeneity and interactions can be studied at an individual level per wall configuration.

ICE curves are visualized in the centered form (cICE), as it facilitates comparison of the response behavior and demonstrates the change relative to the baseline value. Consequently, the PDP and ICE analyses are helpful as an interpretational tool providing an intermediate level of information about the model, between global importance and local explanations.

3.7.7 Two-Way Partial Dependence Plots

As a continuation of the sensitivity analysis performed in Section 3.7.6, the approach is extended using two-way Partial Dependence Plots (PDP). These allow assessing whether the effect of two variables on the prediction of embodied carbon intensity is linear or whether there are interactions between variables in the model response surface.

Two-way PDP analysis is conducted in this research on feature pairs with the greatest importance for the models, as identified in Sections 3.7.1-3.7.4 and in Section 3.7.6 in relation to sensitivity analysis.

Using two-way PDP is justified in this case since the embodied carbon intensity prediction depends on several interacting design variables – structural design parameters of geometrical, material and transport types. It should be noted that two-way PDP analysis is a logical extension of global model sensitivity analysis complementing one-way PDP and ICE analysis of Section 3.7.6.

It should also be noted that this subsection maintains the integrity of the method chain introduced in this chapter, which considers global interpretability in terms of feature importance and SHAP analysis, local interpretability with the help of SHAP and LIME, and sensitivity/interaction behaviour through one-way and two-way PDP analysis.

4 Results

This chapter discusses the findings generated by the trained XGBoost model developed for the prediction of the embodied carbon intensity of RC wall members. As per the methodological approach stated in Chapter 3, findings generated through this model will be discussed under the categories of prediction, errors, feature importance, and interpretability.

4.1 Description of the Trained Model

The results presented in this chapter are based on the XGBoost regression model described in Chapter 3. The model is trained using the synthetic dataset generated from the defined RC wall input variables and corresponding life-cycle embodied carbon calculations. Following the adopted train–test split procedure, the model is trained on 2,032,128 records and evaluated on a held-out test set of 870,912 records.

The trained model uses ten input variables representing geometric, material, transport, and construction-related parameters of RC wall elements. The target variable is embodied carbon intensity, expressed in $kgCO_2e/m^2$. Model evaluation is based on standard regression performance metrics together with a set of interpretability and sensitivity methods, including XGBoost built-in feature importance, SHAP, LIME, Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) curves, and two-way PDP analysis.

4.2 Prediction Accuracy and Error Analysis

The predictive performance of the trained model is presented in Table 4.1. The evaluation includes Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and the mean cross-validation R^2.

Table 4.1 Regression performance metrics for the XGBoost model evaluated on the held-out test dataset.

Metric	Value	Interpretation
MAE	0.0012	The average absolute prediction error equals 0.12% of the normalized range.
MSE	0.0000024	The near-zero squared error indicates extremely small prediction deviations.
RMSE	0.0016	The typical prediction deviation equals 0.16% of the normalized scale.
R^2	0.9999	The model explains 99.99% of the variance in embodied carbon intensity
Mean CV R^2 (5-fold)	0.999951 ± 0.000009	Cross-validation confirms highly stable generalization.

The results indicate very low prediction error across all reported metrics. The model achieves an MAE of 0.0012 and an RMSE of 0.0016 on the normalized target scale, while the corresponding MSE remains close to zero. The coefficient of determination reaches 0.9999, indicating that the trained model explains nearly all variance in the target variable. In addition, the five-fold cross-validation results show a mean R^2 of 0.999951 with a very small standard deviation, indicating highly consistent predictive performance across different data partitions.

The relationship between predicted and actual embodied carbon intensity values is illustrated in Figure 4.1. The figure shows a very close alignment between the predicted and observed values across the held-out test set.

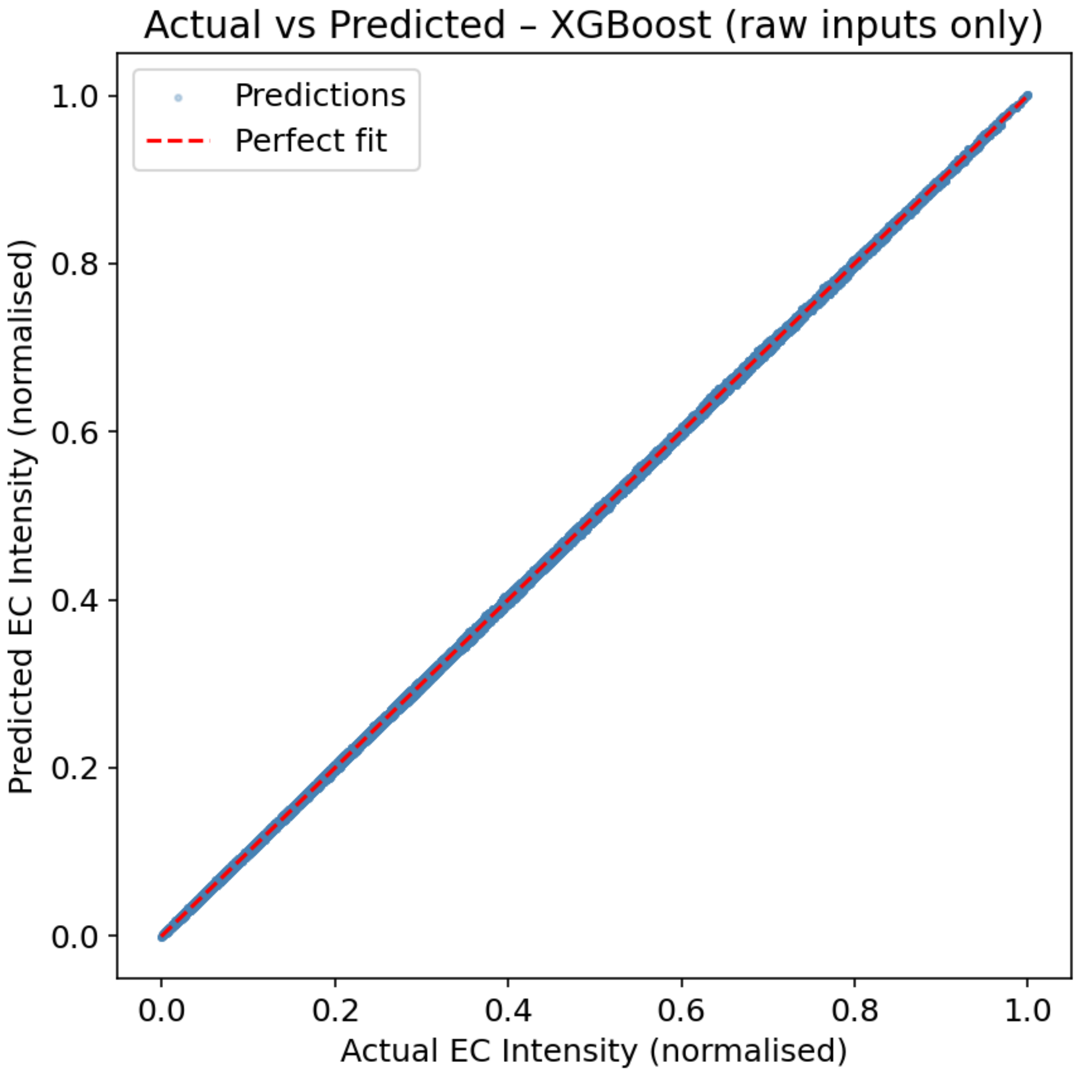

Figure 4.1 Correlation between actual and predicted embodied carbon intensity values. The near-perfect alignment reflects the surrogate nature of the machine learning model rather than predictive capability over real-world data.

The performance achieved is very much in line with the nature of the data set. Since the data set is generated based on the deterministic life cycle equations, there exists an inherent mathematical relationship between independent and dependent variables. Therefore, the role of the machine learning model used is not limited to being just a predictor; it serves as an approximate model for analysis that offers value through feature importance assessment, sensitivity analysis, and parameter impact analysis.

Consequently, the interpretability techniques utilized throughout this research – SHAP, LIME, PDP, and ICE plots – become crucial for contributing to the thesis.

4.3 Comparison between models

To validate the choice of XGBoost as the main predictive model, two other algorithms were considered with the same dataset used for training and testing purposes: Linear Regression and Random Forest. The models were all trained with the same set of inputs and normalization as mentioned in Section 3.6.

Table 4.2 Comparison of model performance on the held-out test dataset.

Model	MAE	RMSE	R^2
Linear Regression	0.0209	0.0267	0.9844
Random Forest	0.0000	0.0000	1.0000
XGBoost	0.0012	0.0016	0.9999

From the output, we can see that the R^2 for the Linear Regression is 0.9844, which shows that there are some non-linear relationships in the data and hence cannot be modeled using the linear approach. The R^2 value for Random Forest is 1.0000 with almost no error, and this implies that the model is over-fitting the deterministic data structure. The XGBoost model generated an R^2 value of 0.9999 with non-zero errors, indicating very accurate predictions. For this reason, XGBoost was chosen as the best fitting algorithm since it does not exhibit the overfitting problem.

4.4 Feature Importance Results

The relevance of each input variable as compared to others in the prediction performance of the model is initially evaluated by employing the inbuilt XGBoost variable importance metric based on total gain. This variable ranking is demonstrated in Figure 4.2.

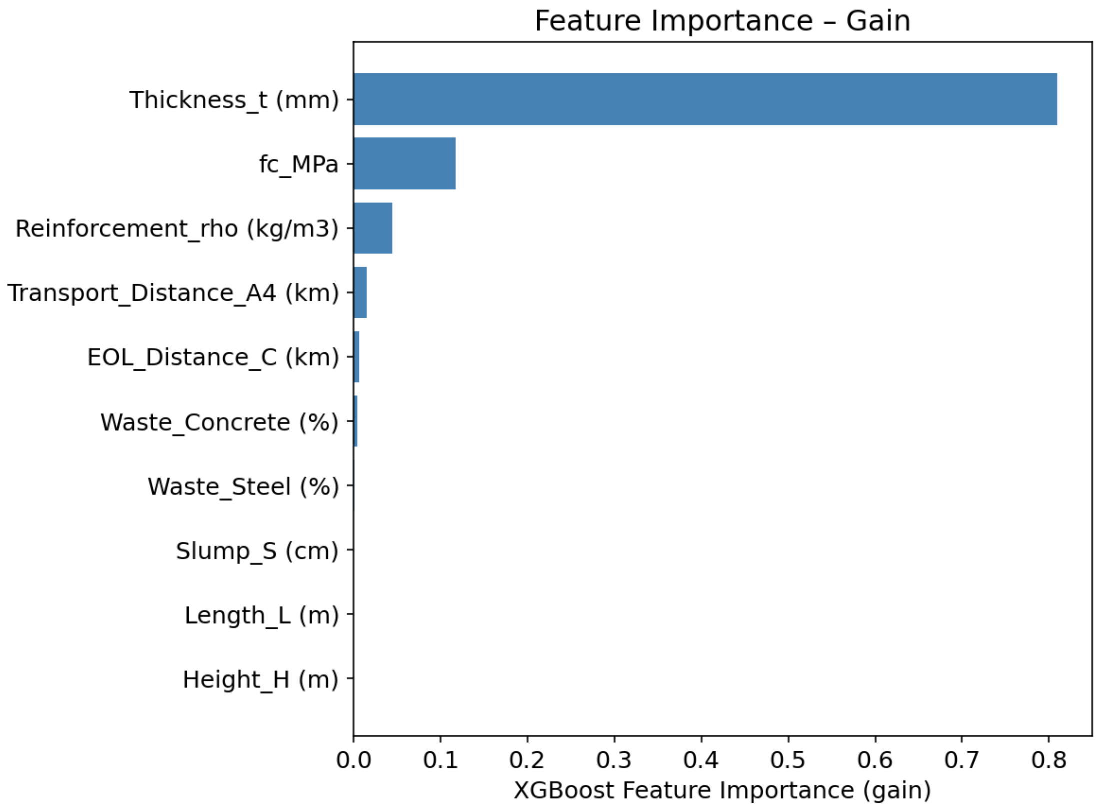

Figure 4.2 XGBoost feature importance based on total gain across all decision tree splits. Wall thickness, concrete compressive strength, and reinforcement ratio emerge as the most influential variables.

Considering the findings of the study, wall thickness is found to be the highest influential parameter when compared to all the other parameters used in the model. Other highly influential parameters include concrete compressive strength, reinforcement ratio, transportation distance to the construction site, and demolition transportation distance. From the findings of the study, it is evident that the model is successful in grouping the parameters that are used in calculating carbon emissions. Physically, the findings of the model are consistent with the laws of engineering. Wall thickness is the most influential parameter because it affects the quantity of concrete and steel contained in the wall element, which has relatively high amounts of carbon emission because of the intensive use of energy during its production process. High reinforcement ratio is the second most influential parameter because the higher the amount of steel required in wall units, the more the production of primary steel. Higher production of primary steel leads to more energy consumption during the production process. Transport distance to the construction site also contributes to embodied carbon because heavier and more voluminous wall elements require greater fuel consumption to deliver materials from production plants to the site.

4.5 SHAP Global Results

To complement the XGBoost gain-based importance measure, global SHAP analysis is used to examine both the magnitude and direction of feature contributions across the dataset. The overall SHAP feature importance ranking is presented in Figure 4.3, while the corresponding SHAP beeswarm plot is shown in Figure 4.4.

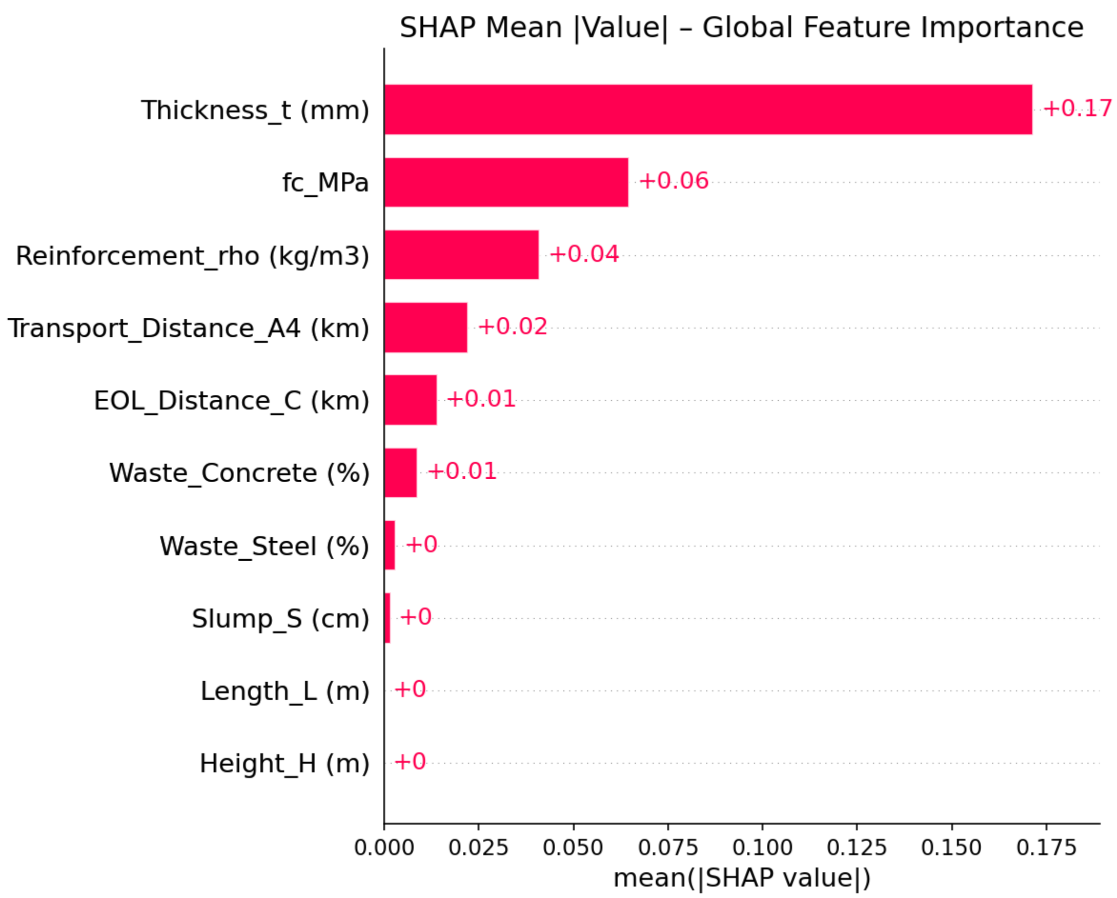

Figure 4.3 Global SHAP features importance showing the relative contribution of each input variable to the predicted embodied carbon intensity. Wall thickness dominates the global importance ranking, consistent with feature importance results.

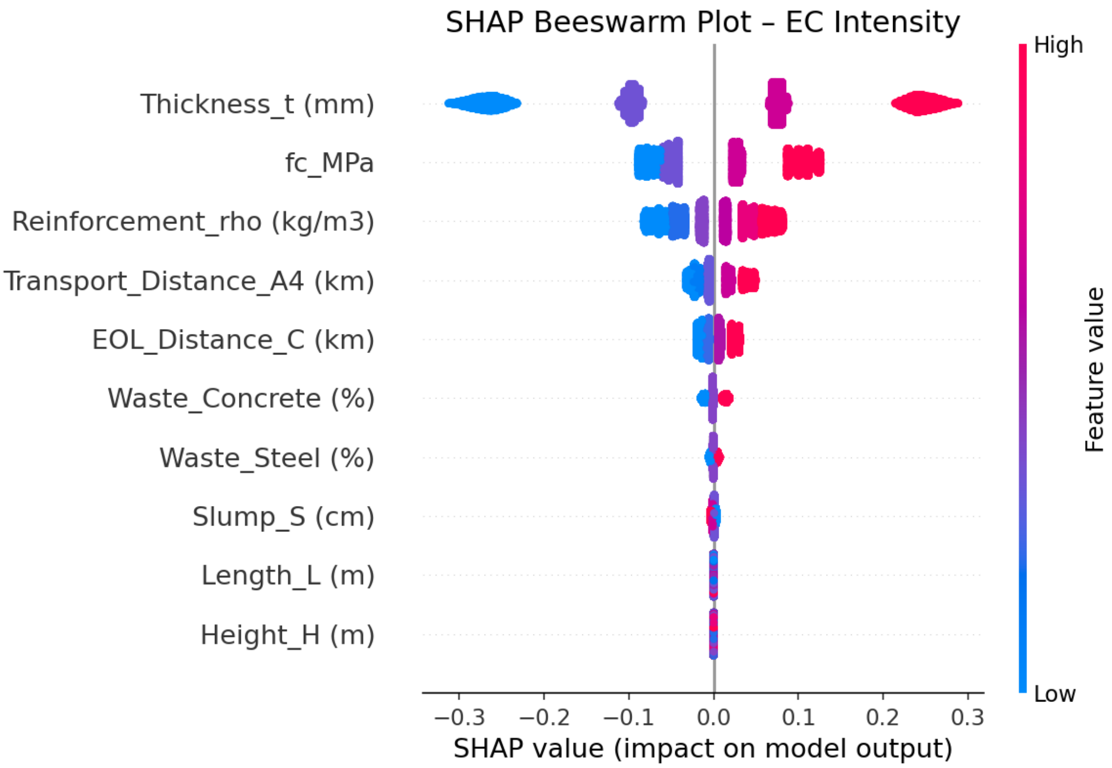

Figure 4.4 SHAP beeswarm plot showing the global contribution of input variables to the predicted embodied carbon intensity. The directional patterns confirm that higher thickness, strength, and reinforcement ratio consistently increase predicted embodied carbon intensity.

The SHAP results are consistent with the feature-importance ranking reported in Figure 4.2. Wall thickness is the dominant predictor of embodied carbon intensity, followed by concrete compressive strength and reinforcement ratio. The SHAP beeswarm plot further shows the directional contribution of each variable. Larger thickness values contribute positively to predicted embodied carbon intensity, while lower thickness values reduce it. Concrete compressive strength also exhibits a clear directional pattern, where lower strength classes contribute to lower predicted intensity values.

Reinforcement ratio shows a similarly consistent effect, with higher reinforcement levels associated with increased predicted intensity.

The transport-related variables, namely plant-to-site distance and end-of-life transport distance, also contribute positively when their values increase. Concrete and steel waste rates exhibit smaller but still consistent positive effects. By contrast, concrete slump shows a weaker negative contribution at higher values, consistent with the lower construction-phase emissions represented in the A5 proxy.

An important result of the SHAP analysis is that wall height and wall length show near-zero contributions across the dataset. This indicates that the trained model attributes little predictive importance to wall footprint dimensions when embodied carbon is expressed on a per-unit-area basis.

4.6 SHAP Local Result

To examine individual predictions in more detail, local SHAP analysis is applied to representative test samples. A representative SHAP waterfall plot is presented in Figure 4.5.

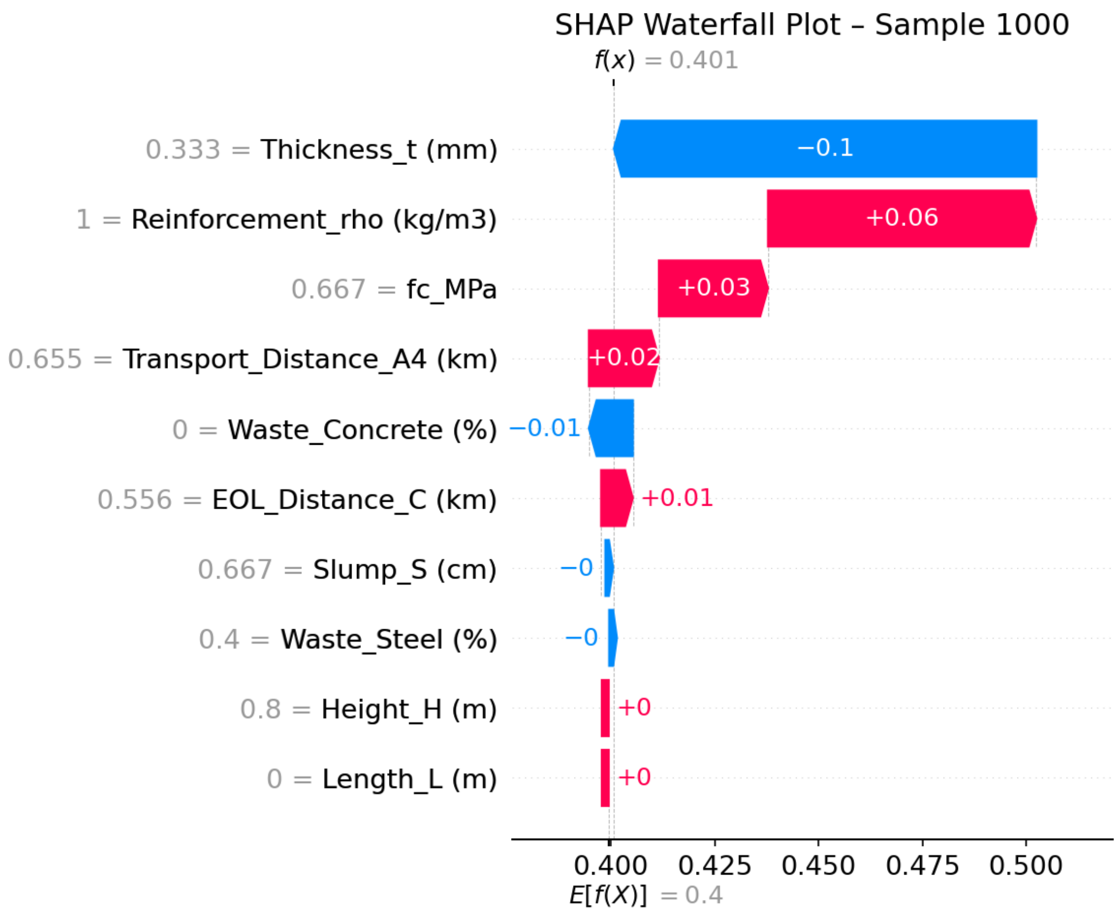

Figure 4.5 SHAP waterfall plot illustrates the local contribution of input variables to the predicted embodied carbon intensity for a representative RC wall configuration. The waterfall plot shows how individual variable values push the prediction above or below the baseline for a specific wall configuration.

The waterfall plot decomposes an individual prediction relative to the baseline value and shows how each input variable increases or decreases the final predicted embodied carbon intensity. The local SHAP result indicates that the direction and relative magnitude of feature contributions are consistent with the global SHAP patterns reported earlier. In particular, the dominant contributions are associated with the key material and geometry-related variables identified in the global analysis.

4.7 LIME Local Results

As an additional local interpretability method, LIME is applied to representative test samples. A representative LIME explanation is shown in Figure 4.6.

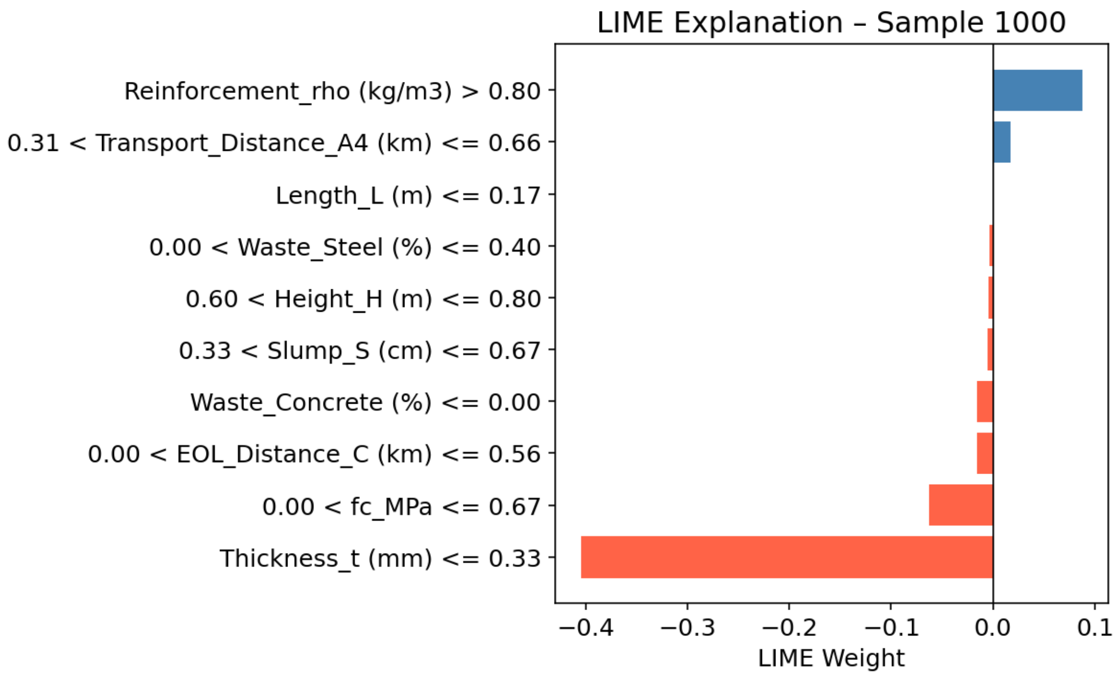

Figure 4.6 LIME explanation showing the local feature contributions to the predicted embodied carbon intensity for a representative reinforced concrete wall configuration. The LIME weights are consistent with SHAP-based local explanations, confirming stable local model behavior.

The LIME output identifies the local contribution of input variables relative to the baseline prediction. The resulting feature-weight pattern is consistent with the SHAP-based local explanation, with wall thickness, concrete compressive strength, and reinforcement ratio appearing among the most influential local drivers of prediction. The agreement between SHAP and LIME indicates that the local behavior of the trained model is stable across two different interpretability frameworks.

4.8 Feature Sensitivity Results

The sensitivity of the trained model to variations in individual input variables is examined using one-way Partial Dependence Plots and centered Individual Conditional Expectation curves. A representative result for the top seven input variables is shown in Figure 4.7.

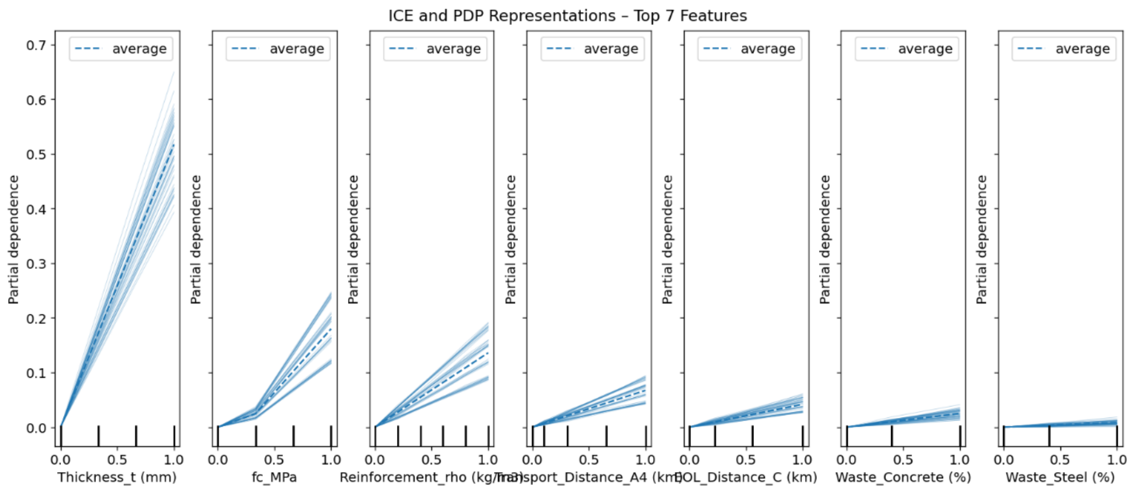

Figure 4.7 Partial dependence and centered ICE curves for the top seven input variables influencing predicted embodied carbon intensity. Monotonic response patterns are observed for the dominant variables, while wall height and length show minimal sensitivity.

The one-way PDP and centered ICE curves indicate that predicted embodied carbon intensity increases as wall thickness increases. Similar monotonic response patterns are observed for concrete compressive strength, reinforcement ratio, and both transport-related variables. By contrast, wall height and wall length show limited variation in the one-way response curves, indicating relatively low sensitivity of model predictions to those dimensions. Waste-rate variables show smaller positive effects, while slump exhibits a weaker inverse relationship with predicted intensity.

4.9 Interaction Results

To examine combined variable effects, two-way Partial Dependence Plots are used to assess interactions between the most influential feature pairs. A representative two-way PDP result is shown in Figure 4.8.

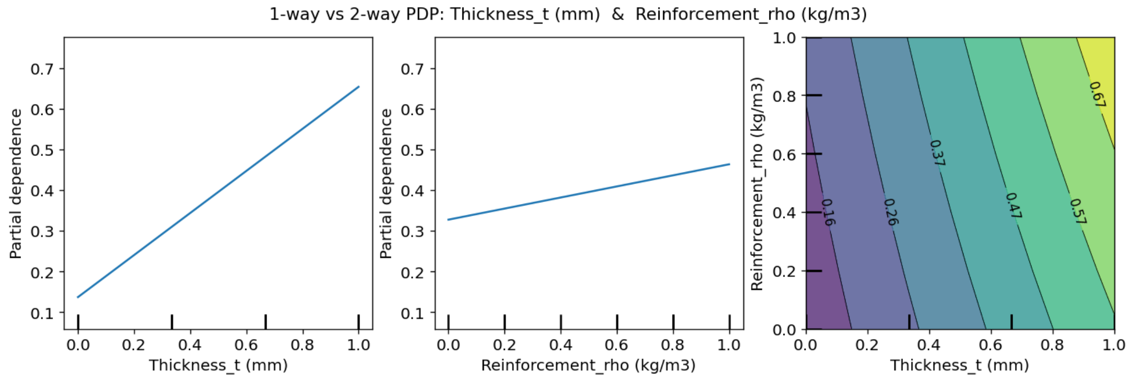

Figure 4.8 Two-way partial dependence plot illustrates the interaction between wall thickness and reinforcement ratio on predicted embodied carbon intensity. The combined increase of both variables produces a stronger carbon response than either variable acting alone, indicating interaction effects.

The interaction plot shows that the combined increase of wall thickness and reinforcement ratio is associated with a stronger increase in predicted embodied carbon intensity than variation in either variable alone. Similar interaction behavior is also observed in the two-way analysis of wall thickness and concrete compressive strength. These results indicate that the trained model captures not only the individual influence of key design variables, but also their joint effect on embodied carbon intensity.

5 Discussion

In this chapter, the primary outcomes of this research in terms of research questions and previous literature discussed in Chapter 2 are analyzed. This chapter focuses on four aspects: result interpretation, literature comparison, implications for early structural design, and limitations of the suggested method. Although Chapter 4 outlined the results of the model and their interpretability, Chapter 5 will further discuss their importance and relevance to the larger scope of embodied carbon analysis and machine learning-enabled decision support.

5.1 Interpretation of the Results

It is worth noting the extremely accurate prediction ability shown by the trained model in case of the test dataset in terms of an impressive $R^2 = 0.9999$. Nevertheless, such impressive accuracy should be interpreted cautiously. The fact is that the current dataset has been created based on deterministic life cycle equations so that almost perfect predictions were predictable. Therefore, it should not be regarded as something extraordinary, but just as confirmation of the success of reproducing the equations used to generate data.

However, it is important to mention that this kind of accuracy does not mean equal levels of applicability to practical use. Embodied carbon figures can be influenced by geographical variations, unique contractor characteristics, uncertainties in measurements, and poor data quality that do not exist in the deterministic simulation dataset. It should be considered as the approximation of the parametric LCA model, which has been employed to develop this dataset.

The results obtained during the analysis show that out of all variables used, the wall thickness has the greatest effect on the target variable followed by concrete compressive strength and reinforcement ratio. These findings are logically consistent with the nature of carbon calculation. The wall thickness variable determines the required volume of concrete that will be used for the creation of the wall in question, thus determining the weight of this component. At the same time, the weight of the wall determines the volume of steel incorporated into the structure. Concrete compressive strength is responsible for the emission factor of the concrete itself whereas reinforcement ratio determines the volume of steel involved in the construction project. Material-related variables have the highest impact on carbon emissions during the production stage and transportation stage, respectively.

From an engineering perspective, these findings appear physically plausible and are well explained. As mentioned above, wall thickness is a dominating variable in question due to a simple reason. An increase in this parameter leads to an increase in concrete volume, which, consequently, results in higher volumes of both concrete mass and embedded steel mass. Given that cement production requires a considerable amount of energy in terms of kiln firing operations and CO_2 release because of calcination processes, the effect of wall thickness becomes obvious from the practical perspective – structural engineers who want to minimize embodied carbon should be careful about wall thickness.

Concrete compressive strength comes second because different strength classes involve different emission factors of concrete as different types of cement are required to achieve a certain compressive strength. If the structural calculations permit, it would make sense to opt for a lower strength class to achieve the desired reduction in carbon emissions.

The same applies to reinforcement ratio since an increase in steel per unit volume of wall corresponds to an increased need for producing primary steel, which is an energy-intensive process. All this explains why this variable shows positive relationships with the target variable in question consistently.

At the same time, transport distance is another parameter that can contribute to embodied carbon intensity. It may be noted that more massive and voluminous structures require additional resources to deliver necessary construction materials to the site. Therefore, this variable plays a secondary role in determining the target but still plays this role consistently as demonstrated by results obtained.

It is important to state that although transport-related variables affect embodied carbon intensity, they are far inferior to other variables analyzed in terms of the significance of this effect. Namely, the distance related to the transportation of construction materials from the plant to the site and vice versa can affect embodied carbon intensity but are still secondary to the factors discussed previously. Moreover, waste rates for concrete and steel show positive impacts whereas slump demonstrates only slightly negative impact through a construction-related variable A5.

The complete lack of correlations between the wall height/length and the target variable may serve as the last interesting example. This fact is justified by the very definition of the target variable, namely, embodied carbon intensity expressed as $kgCO_2e/m^2$ implies that any size-related variables do not affect the target significantly.

Overall, it appears that the trained XGBoost model captures physical relationships between structural parameters and intensity of embodied carbon quite reasonably. Importantly, the agreement between features of importance obtained via various methods proves this conclusion.

5.2 Comparison with Previous Literature

The results obtained from this research appear to align well with previous literature on the topic of embodied carbon in construction. First and foremost, the identification of wall thickness, concrete compressive strength, and reinforcement ratio as the most influential factors in embodied carbon intensity is consistent with earlier studies that indicated these materials as the primary drivers of embodied carbon in buildings

(Hammond & Jones, 2011; Ma et al., 2024). The findings of the current study prove this statement by revealing that the critical parameters include those that determine the volume of the concrete and emissions factor, as well as steel mass. In this study, machine learning models function primarily as interpretable surrogate models rather than purely predictive tools, since the dataset is derived from deterministic life-cycle relationships.

Moreover, the growing significance of embodied carbon when compared with operational emissions in the context of sustainable construction is another piece of information to bear in mind while interpreting the obtained results. Röck et al. (2020) examined 650+ building LCA cases and found that embodied emissions accounted for 45–50% of the GHG emissions of highly efficient buildings. This figure shows how essential tools for assessing embodied carbon quickly become, and the current research responds to this issue at the structural element level.

In addition, the identification of early-stage design variables as being sufficient for obtaining reliable and valuable embodied carbon predictions also appears to be aligned with other literature. Ma et al. (2024) and Myint and Shafique (2024) state that the greatest opportunities for decreasing carbon footprint are during the planning phase when no final design decisions have been taken yet. In line with Pomponi & Moncaster's (2016) systematic literature review, the improvement of design process was one of the most frequent suggestions when dealing with embodied carbon issues.

Lastly, the relative insignificance of the influence of transport-related variables as compared to wall thickness, concrete strength, and reinforcement ratio also seems to be in line with the existing studies. Transport activities were identified as contributors to the overall embodied carbon in some works; however, material production remains the leading cause (Ma et al., 2024; Myint & Shafique, 2024). While this statement is also confirmed by the current study, its findings do not deny the contribution of transport to the overall embodied carbon content.

The machine learning aspect of the current research can also be regarded in the light of the previously published literature. As mentioned above, machine learning becomes increasingly popular to achieve fast assessment, when traditional LCA cannot be performed repeatedly owing to its time-consuming nature (Ghoroghi et al., 2022; El Hafdaoui et al., 2023; Su et al., 2024). However, the current study distinguishes itself by predicting the values of embodied carbon for specific reinforced concrete wall elements rather than entire buildings or even certain materials.

It should also be noted that using various interpretability techniques to reveal the working principles of the machine learning model is justified as per the previous literature as well. As stated by Ghoroghi et al. (2022) and El Hafdaoui et al. (2023), it is crucial for the model to be able to explain its predictions in terms of engineering science because otherwise, such a tool may be considered only academically interesting. Using SHAP, LIME, PDP, ICE, and 2-way PDP analysis fulfills precisely this purpose.

One must admit, however, that the numerical comparison of the findings obtained in the current paper and those provided by others is not possible because of the specificity of the subject matter. Most previous papers either estimate the embodied carbon of buildings or assess the efficiency of some measures in reducing the carbon footprint; thus, direct comparison cannot be conducted.

5.3 Implications for Early-Stage Structural Design

These findings present significant practical implications in relation to early-stage decisions within the structural design process. Firstly, the results prove that a small number of early-stage variables provide valuable predictive information on embodied carbon intensity. Consequently, embodied carbon evaluation does not need to wait until later stages of the design process, which are associated with detailed and accurate project data. Rather, preliminary comparative analysis of different reinforced concrete (RC) walls may be conducted based on the data available in the early design phase.

From the point of view of the design process, it can be observed that thickness is an important variable that should be considered carefully. As thickness impacts both quantities of concrete and steel materials used, it becomes a key factor to predict embodied carbon. At the same time, concrete strength class and reinforcement ratio prove to be essential design variables, the choice of which influences considerably embodied carbon intensity. Overall, it means that at the first stage of designing, the low-carbon optimization of thickness, concrete strength class, and reinforcement ratio should come prior to other design variables.

Transport variables can also be seen as practically meaningful. Despite not being a central component within the model, results reveal that transport makes a systematic contribution to predicting embodied carbon intensity. This implies that early-stage embodied carbon estimation may consider transport and logistics, especially in the case where alternative scenarios assume different material sources.

Finally, global and local interpretability of the results is also worth considering. Methods of global interpretability allow revealing important variables; methods of local interpretability provide information on how design decisions contribute to obtaining a particular result. Partial dependence plot (PDP), individual conditional expectation (ICE) and two-way PDP plots help better understand the marginal effects and the role of variables within the model. Practically speaking, interpretability is the key to making decisions with explanations based on predictions made with the machine learning model.

In general, the study shows that an interpretable ML approach can serve as a tool for comparative decision support in structural design. The machine learning model presented above does not substitute a project-specific LCA, which should always be performed after design completion. Rather, its contribution consists of fast evaluation of alternative designs of RC walls at the early stage of structural designing.

5.4 Limitations of the Proposed Approach

Despite positive outcomes, there are several important limitations associated with the proposed method. The first key limitation of the paper is associated with empirical validation. As stated above, all data used in the research is synthetic and was generated based on life-cycle equations without the involvement of any real data obtained during projects implementation. Consequently, it cannot be claimed that the developed predictive model can help estimate embodied carbon values in practice since its output cannot be treated as a precise prediction but rather an approximation of the calculation method used as data source. Indeed, in case of real designs, embodied carbon depends on a variety of factors, including specifics of the project itself, regional characteristics associated with material production, contractor's actions, and measurement inaccuracies. Thus, empirical validation based on real projects' data should be conducted to increase external validity of this study.

The second limitation refers to the modeling assumptions made for life cycle calculation process. It should be stated that the used system boundaries can be regarded as reasonable for the current purpose, however, A5 phase, as well as the end-of-life phase, is simplified and based on proxy modeling rather than detailed process modeling. While such approach seems to be acceptable in case of comparative evaluation and is even preferred in case of early-stage prediction, the degree to which the results can be considered should still be limited.

As the next limitation, one should speak about the exclusion of B-phase or use phase from consideration, as stated above. Such limitation is justifiable in this specific case as the focus of the research lies with estimation of embodied carbon of wall elements. Nevertheless, it should be noted that the current model describes embodied carbon of structural material in certain life cycles phases only and cannot provide the results of the whole life cycle assessment.

The fourth limitation relates to the selection of the structural element that will serve as the basis for predicting. The reinforced concrete wall element is a clearly defined and practical structural unit, which makes it a good candidate for modelling. Nonetheless, this selection does not consider the entire range of structural systems and building elements used in practice. The results are applicable to the design of RC walls and cannot be immediately applied to other elements such as slabs, columns, and beams, nor to other structural systems like steel framing, composite, or wood structures. The design process, materials used, and fabrication processes involved in these other structural systems vary significantly and would necessitate a different model. In addition, there is no consideration for the variability of structural elements in a complete building. In actual application, the combined effects of all structural units in the building and the entire design process cannot be accurately predicted by the current model.

Lastly, while the authors compared performance of three regression models (Linear Regression, Random Forest, and XGBoost) for prediction purposes, this comparison cannot be referred to as exhaustive benchmarking, since it did not cover all available regression models. Nevertheless, it is possible to state that according to the results obtained XGBoost algorithm has demonstrated the best accuracy and generalizability among the options considered.

Consequently, the discussed limitations of the study do not reduce its value; rather, they determine its scope.

6 Conclusions and Future Work

The main goal of this thesis is to create a machine learning model that can predict the carbon intensity of the reinforced concrete (RC) wall elements based on early-design input parameters. The idea comes from the rising importance of carbon impact in our built world and the lack of rapid and interpretable tools used for comparisons between different RC walls during early design stages. The proposed approach utilizes the combination of life-cycle assessment based synthetic data generation and XGBoost regression model for RC wall elements.

6.1 Responses to Research Questions

RQ1: Which RC wall design parameters are most influential in determining embodied carbon intensity?
The wall thickness proved to be the most important factor, followed by concrete compressive strength and reinforcement ratio. These are directly related to material amounts and emission rates, and this importance was repeatedly confirmed using the XGBoost model feature importance measure, SHAP, and PDP plots.

RQ2: Can machine learning predict RC wall embodied carbon intensity with high accuracy using basic early-stage inputs?
The XGBoost model training produced an R^2 value of 0.9999, while the mean cross validation R^2 was 0.999951, which indicates the prediction of embodied carbon intensity with extremely high accuracy, using just ten initial design variables.

RQ3: How can interpretability methods help explain the influence of RC wall parameters on embodied carbon predictions?
The SHAP approach was effective in explaining the impact as well as the direction of influence of each input parameter on a global as well as a local scale. LIME provided additional explanation for local impact, while PDP and ICE were used to show marginal sensitivities.

RQ4: How can the proposed model support early-stage structural decision-making for RC walls?

With the help of this tool, a quick analysis is possible on a variety of RC wall designs using parameters that are commonly available before a final design decision is made. This tool helps focus on the parameters having high impact on the embodied carbon emissions, which can be seen from its ability to determine the wall thickness, concrete grade, and steel content as the main contributors to the embodied carbon.

6.2 Summary of Main Findings

The findings show that the embodied carbon intensity of reinforced concrete wall components can be predicted with extremely high accuracy based on a relatively small number of initial design variables. The trained model showed extremely low error rates and demonstrated outstandingly stable performance during cross-validation folds, implying that it could learn the relationships contained in the artificial dataset. Therefore, one can conclude that the results of the study indicate the applicability of the proposed model as a surrogate for the actual calculation procedure used in this paper.

According to the results obtained, the thickness of walls has the greatest impact on the value of embodied carbon intensity, followed by the compressive strength of concrete and the reinforcement ratio. The variables associated with transportation also play a role in calculations, although less prominent than material parameters. Conversely, wall height and wall length are not significant determinants of the calculated intensity measure when stated per square meter. This implies that, in the chosen functional unit, the proposed model is most sensitive to parameters that have an impact on the quantity and quality of materials and transportation costs.

In addition, the interpretability analysis supports the above conclusions. All the methods used to analyze the behavior of the predictive model identified the same set of dominating variables and revealed that its predictions depend on reasonable engineering parameters. In this case, the study not only provides a model for predicting carbon intensity but also shows that the behavior of this model is comprehensible from an engineering perspective.

6.3 Contributions of the Thesis

The key contributions of the thesis consist of four major aspects.

Firstly, it proposes a machine learning approach to estimate the carbon intensity embodied in RC wall elements based on early design stage information. As such, this paper makes an essential contribution to the state of art by focusing on an element level prediction unit instead of the traditional material level or building level approaches.

Secondly, it provides a rich synthetic data set of RC wall designs calculated based on equations of life cycle carbon as well as a defined range of input variables.

Thirdly, it shows that interpretability can be achieved by choosing a tree-based model with high prediction accuracy. The use of the XGBoost model combined with such methods as SHAP, LIME, PDP, ICE and two-way PDP make it possible.

Fourthly, this thesis gives an engineering-oriented framework for comparing embodied carbon intensity between alternative RC walls at an early design stage. Unlike the black box prediction approach, the model enables an engineer to understand which design decisions result in higher or lower values of the output variable.

6.4 Practical Relevance

The applied value of the research is associated with its possible application as a decision-making tool at an early stage. It allows for quick prediction of embodied carbon intensity based on parameters such as wall thickness, concrete strength, reinforcement percentage, and transportation assumptions that can be obtained during the early stages of project development. Thus, it is possible to evaluate different RC wall designs before obtaining all required information about the project.

For structural designers and engineers, the results mean that the most significant actions that will affect the embodied carbon intensity in the first place will be optimizing wall thickness, selecting the appropriate concrete strength, and controlling the reinforcement percentage. In this way, the developed model provides a more carbon-aware method of designing RC walls without a full conventional life-cycle assessment of each option. At the same time, interpretability approaches increase the practical usefulness of the tool since they allow explaining the relationships between input parameters and the predicted output. Such an understanding is essential to ensure that machine-learning models can be trusted in structural design.

6.5 Recommendations for Future Research

There are many areas of potential future research arising from the performed study.

To begin with, additional investigation might focus on validating the presented modelling approach based on empirical data related to specific projects' designs. Namely, at that, the model will be validated, and its performance evaluated with respect to the current industrial conditions. It will also be possible to perform further analysis regarding the adequacy of the dataset structure to real-life situations related to the design of RC walls.

Next, the scope of applicability of the proposed model may be further expanded in terms of the structures investigated, namely, other structural members (slabs, columns, etc.) may be taken as a target. Hence, it will be possible to test whether the same modelling strategy may be applied to other types of structures made of reinforced concrete.

The life cycle model of the current dataset can be expanded in terms of its detail – in particular, the construction process, and the disposal phase at the end of the lifetime of the RC structures may be more elaborated to achieve better results.

Moreover, future research may focus on comparing other regression algorithms to XGBoost in terms of their performance and suitability for solving similar problems with the current dataset.

Finally, further research should be dedicated to applying the current prediction framework to building information modeling environment to enhance the applicability of the model. In addition, scaling the methodology up to the entire structures made of different elements (e.g., RC walls, slabs, columns, and foundation) will make the tool much more effective with respect to BIM application to estimating the embodied carbon during preliminary design.

References

El Hafdaoui, H., Khallaayoun, A., Bouarfa, I., & Ouazzani, K. (2023). Machine learning for embodied carbon life cycle assessment of buildings. *Journal of Umm Al-Qura University for Engineering and Architecture*, *14*(3), 188–200. https://doi.org/10.1007/s43995-023-00028-y

Gervasio, H., & Dimova, S. (2018). *Model for Life Cycle Assessment (LCA) of buildings EFIResources: Resource Efficient Construction towards Sustainable Design*. https://doi.org/10.2760/10016

Ghoroghi, A., Rezgui, Y., Petri, I., & Beach, T. (2022). Advances in application of machine learning to life cycle assessment: a literature review. *International Journal of Life Cycle Assessment*, *27*(3), 433–456. https://doi.org/10.1007/s11367-022-02030-3

Hammond, Geoffrey., Jones, Craig., Lowrie, Fiona., & Tse, Peter. (2011). *Embodied carbon : the Inventory of Carbon and Energy (ICE)*. BSRIA.

Hunt, J., & Osorio-Sandoval, C. A. (2023). Assessing Embodied Carbon in Structural Models: A Building Information Modelling-Based Approach. *Buildings*, *13*(7). https://doi.org/10.3390/buildings13071679

Iqbal, N., & Noureldin, M. (2025). Automated methodological approach using artificial intelligence and fuzzy inference for qualitative sustainable and economic assessment of buildings. *Scientific Reports*, *16*(1), 2073. https://doi.org/10.1038/s41598-025-31913-2

Khan, S. A., Alam, T., Khan, M. S., Blecich, P., Kamal, M. A., Gupta, N. K., & Yadav, A. S. (2022). Life Cycle Assessment of Embodied Carbon in Buildings: Background, Approaches and Advancements. In *Buildings* (Vol. 12, Number 11). MDPI. https://doi.org/10.3390/buildings12111944

Ma, L., Azari, R., & Elnimeiri, M. (2024). A Building Information Modeling-Based Life Cycle Assessment of the Embodied Carbon and Environmental Impacts of High-Rise Building Structures: A Case Study. *Sustainability (Switzerland)*, *16*(2). https://doi.org/10.3390/su16020569

Moussavi Nadoushani, Z. S., & Akbarnezhad, A. (2015). Effects of structural system on the life cycle carbon footprint of buildings. *Energy and Buildings*, *102*, 337–346. https://doi.org/10.1016/j.enbuild.2015.05.044

Myint, N. N., & Shafique, M. (2024). Embodied carbon emissions of buildings: Taking a step towards net zero buildings. *Case Studies in Construction Materials*, *20*. https://doi.org/10.1016/j.cscm.2024.e03024

Pomponi, F., & Moncaster, A. (2016). Embodied carbon mitigation and reduction in the built environment – What does the evidence say? *Journal of Environmental Management*, *181*, 687–700. https://doi.org/10.1016/j.jenvman.2016.08.036

Röck, M., Saade, M. R. M., Balouktsi, M., Rasmussen, F. N., Birgisdottir, H., Frischknecht, R., Habert, G., Lützkendorf, T., & Passer, A. (2020). Embodied GHG emissions of buildings – The hidden challenge for effective climate change mitigation. *Applied Energy*, *258*, 114107. https://doi.org/10.1016/j.apenergy.2019.114107

Su, S., Zang, Z., Yuan, J., Pan, X., & Shan, M. (2024). Considering critical building materials for embodied carbon emissions in buildings: A machine learning-based prediction model and tool. *Case Studies in Construction Materials*, *20*. https://doi.org/10.1016/j.cscm.2024.e02887