2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

MULTIMODAL DATA PROCESSING FOR BUILDING MATERIAL PROPERTY PREDICTIONS

Julia Kaltenegger[1], Ekaterina Petrova[1], André Borrmann[2], and Pieter Pauwels[1]
[1]Eindhoven University of Technology, Netherlands
[2]Technical University Munich, Germany

Abstract

Building materials significantly impact a building's environmental footprint. Material inventories promote building element reuse, yet documentation of material property data for specific buildings is scarce and often confined to closed repositories. Effective material-related performance assessments demand knowledge of building characteristics, including material types, properties, and environmental indicators. Traditional (unimodal) prediction models predict material properties but overlook the complexities of real-world applications, lacking contextual insights from diverse data sources. This study introduces a multimodal data processing method that incorporates machine learning for material property prediction, integrating knowledge engineering to extract building characteristics as contextual information, and validates it on a residential brick facade.

Introduction

The built environment accounts for approximately 50% of global material extraction, making waste reduction a key component of the EU's climate goals. Renovation and circular design strategies promote the reuse of materials across their entire life cycle (Moschen-Schimek et al., 2023). Reclamation audits assess the reuse potential of building elements by documenting their characteristics, including dimensions, quantities, condition, environmental impacts, and recommendations for disassembly. For example, quantifying the carbon footprint can ultimately help reduce the need for new production, thereby promoting material reuse (Smeyers et al., 2022). Life Cycle Assessment (LCA) quantifies carbon footprints using material data, environmental indicators, and element quantities. Material inventories also support deconstruction and inform the Material Circularity Index (MCI). However, data for such computations is scarce, requiring case-by-case collection for each application. Material properties are neither documented for existing buildings nor easily determined from available datasets without manually searching through scattered material databases (Smeyers et al., 2022). Furthermore, available databases reveal numerous missing values, presenting opportunities to utilise Machine Learning (ML) to predict missing information.

Recent research is increasingly highlighting the potential of ML to predict material properties. The material and building domain addressed diverse input and output variables for ML models. In material research, independent parameters such as the fibre composition and the elemental and chemical structure serve as inputs to predict the thermal, physical and mechanical properties of materials. In the building domain, material property predictions typically require input related to the building elements and their constituent materials. Therefore, existing material property data is used to predict other properties, such as thermal conductivity, based on moisture content and density (Stergiou et al., 2023). Building-specific material property prediction requires information that accurately determines the material types, properties, and performance for each building and its elements. As suggested by Fenton et al. (2024), building characteristics can be embedded as qualitative variables in the ML model (Fenton et al., 2024). However, including building-specific characteristics as features directly in the ML model limits their generalisation, scalability and reusability across different buildings.

Accurate data-driven approaches require enriched real-world contextual information drawn from multiple sources, including textual, tabular, and imagery data. Traditionally, multimodal ML models integrate these modalities via embeddings that convert data from various modalities into vector spaces. Such embeddings must be aligned in a shared vector space so that similar entities across different data types are grouped and thus can share meaningful information. For building material data sources representing existing buildings, hardly any dataset contains a grouped representation of similar concepts across different modalities. State-of-the-art knowledge extraction methods can overcome these challenges by creating relationships between datasets and cross-referencing contextual information, which is crucial for predicting material properties.

This study proposes a multimodal data processing method for material property prediction, combining textual and numerical data from various data source, to improve

material performance assessments in the context of existing buildings. Currently, key metrics such as MCI and embodied carbon, as part of the LCA, rely on the mass of building elements. However, in existing buildings, element mass is often unknown and roughly estimated, leading to high uncertainty. To address this gap, data from various sources can provide information on building archetypes and construction period, leading to approximate values for construction thickness, element surfaces, thermal resistance and façade material types. This information enables more precise input data for accurately predicting material density (kg/m³) and estimating element mass (kg). The resulting mass estimates can then be used to calculate MCI. Additionally, the approach supports the prediction of GWP ($kgCO_2eq$), which, when combined with element mass, allows for the calculation of embodied carbon ($kgCO_2eq/BE$).

By generating these missing material properties, the method enables more accurate assessments of material performance and supports better decision-making for the end-of-life phase of buildings. The framework utilises open data to train ML models for density and GWP prediction through both traditional and automated ML pipelines. Tested on a Dutch brick facade, it extracts building data, identifies materials, predicts properties, and computes MCI and embodied carbon. The following sections discuss the state-of-the-art in ML-based material property prediction and multimodal ML frameworks, followed by the research methodology, proposed approach, results, and conclusions.

State-of-the-art

Machine Learning for material property prediction

ML for material property prediction has gained attention in both material science and the building domain. ML models for material property prediction and performance optimisation study the thermal, mechanical, and optical properties to optimise the performance of, e.g., phase change materials (PCM), insulation components, waste, and recycled materials (Stergiou et al., 2023). Most common ML prediction tasks rely on supervised learning to model the function that best describes the relationships between the material properties, given the pairs of examples (x_i, y_i), and trying to learn the relation between the input variables x and the target value y. Common models include linear and multiple regression, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The selection of an appropriate model depends on the type of problem, dataset size, data dimensionality, degree of non-linearity and data type.

For datasets with high dimensionality and non-linearity, preprocessing through feature engineering and resampling is crucial to reduce the number of irrelevant or redundant features, thereby making the ML model more efficient and less complex. Additionally, low-volume or low-dimensional datasets require feature and missing data imputation to maintain data integrity and enhance the prediction accuracy. Relevant features are selected to capture trends in dependent variables across materials. Techniques such as the filter method rank features by relevance (e.g., variance), while wrapper methods use model performance as the selection criterion (Salcedo-Sanz et al., 2018). Bootstrapping and cross-validation assess performance on unseen data, preventing overfitting (Fenton et al., 2024). Understanding material property relationships is crucial for ML tasks that feed into thermal computations, LCA and MCI. Each application requires distinct model assumptions to account for the specific dependencies between material attributes. Significant correlations have been established between material or element mass and the MCI index, demonstrating that higher mass significantly impacts the MCI score (Khadim et al., 2025). For thermal performance assessments, predictions often rely on the assumption of a linear relationship between thermal conductivity and density (Latapie et al., 2024). While this simplification works for conventional materials, it fails to capture the behaviour of innovative materials, such as aerated concrete, which is lightweight (low density) but has high thermal conductivity. The service lifetime, disassembly potential, construction waste generation, use, and end-of-life (EOL) phases further influence the MCI score. Training ML models on synthetic data and relying on simplified virtual buildings often misrepresents reality and neglects contextual information of existing buildings.

Besides the quantitative data, qualitative building characteristics are equally important for accurately predicting material properties. As such, research on data-driven material predictions incorporates contextual building information as soft features during model training, thereby relying on case-by-case model parameter definitions (Fenton et al., 2024). Previous studies reviewed tools and techniques that integrate multiple data sources for building material stock and urban-scale LCA (Pei et al. 2024; Byers et al. 2024). They highlight that the complexity of integrating diverse sources leads to challenges when handling different data types and the related computations. Integrating ML models guided by contextual information such as building type, archetype, and construction period is essential as it helps merge data and thus combine information from different scales. The multi-scalar data sources analysed in the current study, as well as the information they provide, are shown in Fig. 1.

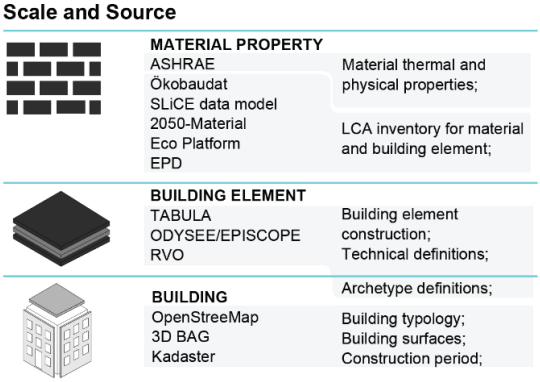

Figure 1: Multi-scalar data for material property prediction

Multimodal Machine Learning frameworks

Incorporating multiple data sources and modalities is crucial for addressing the complexity of real-world prediction tasks. The term modalities refers to the diversity of data types, e.g., text, tabular, numeric, imagery and videos. Multimodal learning facilitates the exploration of intermodal connections, leading to a deeper understanding of natural phenomena (Ma et al., 2022). However, data heterogeneity presents challenges related to representation, translation, alignment, fusion, and co-learning. Recent advancements, such as Automated Machine Learning (AutoML), simplify the creation of complex pipelines for uni- and multimodal integration (Tang et al., 2024). AutoML makes ML more accessible by automating intricate processes such as model selection, hyperparameter optimisation, and feature engineering. Its goal is to enable non-experts to build ML models without writing complex, low-level code (Salehin et al., 2024).

Unimodal AutoML approaches utilise end-to-end ML pipelines, providing a systematic and automated mechanism for constructing, executing, and deploying initial ML models across various application domains. Multimodal AutoML combines self-distinct ML pipelines to facilitate the orchestration and automation of workflows, particularly when handling diverse data modalities (Tang et al., 2024). Unimodal AutoML frameworks such as Auto-Sklearn (Feurer et al., 2015) build upon the sci-kit library and leverage Bayesian optimisation and ensemble construction to provide a robust solution for classification and regression tasks. Ensemble models are commonly used for unimodal AutoML. Ensemble models use a finite set of alternative models (base models or weak learners), thereby combining multiple hypotheses in a single model to reduce the generalisation error of the predictions (e.g., Light Gradient-Boosting Machine (Feurer et al., 2015).

Multimodal AutoML frameworks such as AutoKeras (Google) (Jin et al., 2023) and AutoMM from AutoGluon (Amazon) (Tang et al., 2024) are equipped to process multimodal data leveraging deep learning, utilising foundational models (provided via HuggingFace, TIMM, and MMDetection) that are extensively trained on large datasets. The multimodal predictor trains a single neural network that can jointly process multiple feature types, including both categorical and numerical data (Tang et al., 2024). ML models for tasks such as regression, object detection, and image classification embed images, text, and tabular data across modalities to enable seamless integration and prediction. Thereby, relationships across modalities are identified through specific preprocessing for data fusion and integration into a unified dataset. Meta-learning techniques, inspired by the ability to generalise experiences, infer connections between modalities by leveraging complementary information from diverse data sources (Ma et al., 2022; Pio et al., 2024). For example, insights from a numerical dataset can be combined with an image dataset by linking each image to corresponding numerical attributes. Therefore, advanced and often manual data manipulation, as well as explicit dataset alignment, are required, which makes the process labour-intensive. In material property prediction, this approach is impractical, often involving federated data sources with no connections.

Methods

This study presents a data processing pipeline for material property prediction using multimodal data (Fig. 2), comprising three key stages. First, it integrates open-access material property datasets (e.g., ASHRAE, SliCE) and building-related data sources (e.g., 3D BAG, OSM, TABULA, ARCHETYPE). Second, it utilises data extraction, transformation, and loading processes to train ML models and derive contextual information. Third, the

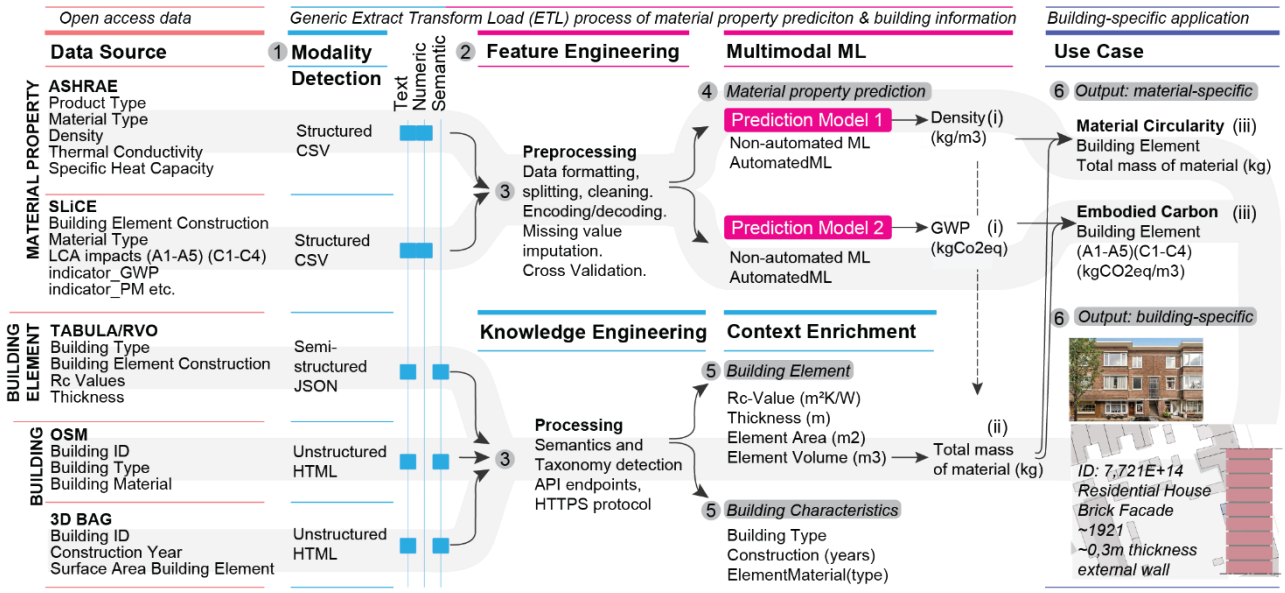

Figure 2: A multimodal pipeline for material property prediction. Open-access data for model training and testing, extract, transform, and load data to both the prediction model and context retrieval. Building-specific information retrieval is tested.

pipeline is applied to a residential building with a brick façade, where building characteristics are extracted, material properties are predicted, and the MCI and embodied carbon are calculated. Given the heterogeneity of data modalities, the framework incorporates specialised extraction and transformation pipelines for structured, unstructured, and semi-structured data (e.g., JSON, text, CSV).

The methodological workflow consists of (1) detecting data modalities and types, (2) extracting relevant features through feature and knowledge engineering, (3) transforming raw data based on modality-specific preprocessing requirements, (4) loading quantitative data into prediction models, (5) extracting contextual information related to the building elements' characteristics and (6) applying predictions using a specific use case as example. The study focuses on predicting key material properties (i) (i.e., density in kg/m³, and GWP in $kgCO_2eq$) through two ML pipelines per task: traditional and automated ML. Supervised learning methods are used to minimise prediction errors between predicted and observed values. Consequentially, (ii) the material mass (kg) can be computed. Contextual qualitative and semi-structured building data support the inference of material properties at both the building and element levels. Finally, the predicted material properties are used to compute (iii) MCI and embodied carbon, thereby enabling a more comprehensive assessment of material performance.

Material Density Prediction

Modality detection and feature engineering

The ASHRAE database (ASHRAE, 2021) provides physical and thermal properties, which are used to train, test and validate the ML model pipeline. This dataset includes load-bearing, non-load-bearing, and insulating materials such as masonry, concrete, and fiber insulation, listing key properties: material density (ρ) (kg/m³), thermal conductivity (λ) (W/m·K), and specific heat capacity (J/kg·K). The dataset exhibits low dimensionality, a limited sample size (n = 250), and two modalities: text data (material names as categorical variables) and numerical data (property values as continuous variables). The goal is to develop a predictive model that estimates material density using the material name, thermal conductivity, and specific heat capacity as input features. As such, the conceptual model (Fig. 3) defines the material name (x_1), thermal conductivity (x_2), and specific heat capacity (x_3) as independent variables, while material density (y) is the dependent variable. The prediction task compares traditional ML and AutoML pipelines: The former employs ridge-regularised linear regression (L2) and an extra tree regressor, while the latter utilises an ensemble model.

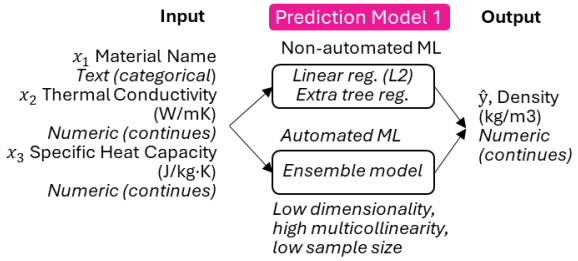

Figure 3: Conceptual model for material density prediction

Data preprocessing

The non-automated ML approach involves labour-intensive data processing and encoding, whereas the automated ML ensemble model requires less processing, utilising non-encoded data and multimodality as hyperparameter. Numerical data is split, mean values computed, and units converted to the metric system. Descriptive analysis reveals a positive linear relationship between density and conductivity, consistent with the findings of Latapie et al. (2024), who also reported high multicollinearity. There is a weak negative correlation between density and heat capacity. Missing value analysis reveals that values are missing not at random (MNAR) between density and conductivity, and missing at random (MAR) between density and heat capacity, as well as between conductivity and heat capacity. Uni- and multivariate imputation methods, evaluated with Mean Squared Error (MSE), prevent sample size reduction. The simple imputer (univariate) uses KNN imputation based on Euclidean distance, averaging values of nearest neighbours. The iterative imputer (multivariate) models missing features in a round-robin fashion using Bayesian ridge and extra tree regressors. For non-automated ML, the text data is encoded into dummy variables, with feature reduction applied during training to prevent overfitting.

Machine Learning model and training

The linear regression utilises the regression model (1) to test the hypothesis (H_0) function as a weighted linear combination of features. The loss function (2) is the sum of the squared distance between the observed and predicted values. To avoid overfitting, the ridge regression (L2 regularisation) introduces a penalty term, the square sum of the coefficients, to the loss function. The penalty term penalises high-value coefficients by slightly increasing the training error, thus restricting fitting the training data too closely. Ridge regression reduces coefficients close to zero, which helps distribute the effects of correlated variables evenly across features. In other words, it penalises large coefficients and reduces the influence of less valuable features.

$$y_i = \beta_0 + \beta_1 x_1 .. \beta_n x_n + \varepsilon \quad (1)$$

$$Loss\ Function: \sum_{i=1}^{n} (r_j^{\,2}) + \mathrm{T}$$

$$L2\ penalty\ T: \alpha \sum_{j=1}^{n} r_j^{\,2} \quad (2)$$

The Extra Trees Regressor (ETR) derives from the idea of an ensemble ML algorithm that uses a collection of decision trees to perform regression tasks. It can increase training time and prediction accuracy by introducing more randomness, which reduces variance and diminishes collinearity between features. A meta-estimator is implemented that fits several randomised decision trees (a.k.a. extra-trees) on various data samples. Averaging the predictions aims to improve the predictive accuracy and control overfitting. Finally, the ensemble model is applied, using "multimodal" as hyperparameter optimisation. The model combines the XGBoost, LightGBM, CatBoost, and NeuralNetTorch. It uses the tabular predictor as a method and regression with multimodal data input as a hyperparameter.

The data is split using a K-fold and shuffle split to generate the training, testing, and validation datasets, with proportions of 70%, 15%, and 15%, respectively. The model selected is identified using cross-validation, with evaluation metrics including the coefficient of determination (R^2) in (3) and the Mean Squared Error (MAE) in (4). The R^2 is a statistical measure in regression analysis that studies the proportion of variance in the dependent variable explained by the independent variable(s). It is calculated as the ratio of the residual sum of squares (RSS) to the total sum of squares (TSS), indicating the goodness of fit on a scale from 0 to 1, where 1 indicates a perfect fit.

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (3)$$

The MAE is a measure used to evaluate model quality by calculating the average absolute difference between actual and predicted values. It tells how far predictions are from the actual values on average. A lower MAE indicates better accuracy. It is calculated by dividing the Sum of Squared Errors (the differences between observed and predicted values) by the number of data points.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}^2| \qquad (4)$$

The prediction accuracy for the observed density (y), and the predicted density ($\hat{y}$) per model is represented in Fig. 4. The R^2 and the MAE show that the ETR and the ensemble model perform best (Table 1).

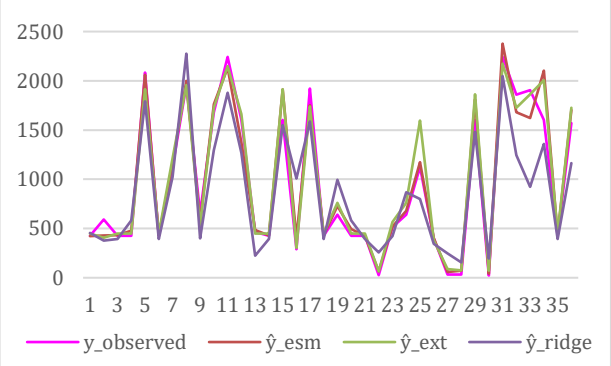

Figure 4: Material Density Property Prediction (kg/m³). Comparison between observed density (y) and predicted density (ŷ) from the ensemble model (esm), the extra tree regressor (ext), and the ridge regression (ridge)

Table 1: Model validation using R^2 and MAE of studied ML models for density property prediction

	Ridge Regression	Extra Tree Regressor	Ensemble Model
R^2	0.689	0.940	0.965
MAE	241.66	106.82	77.61

Material Global Warming Potential Prediction

Modality detection and feature engineering

The environmental impact data, sourced from the SLiCE dataset (Röck et al., 2023), includes 25 impact indicators across life cycle stages (A1-3, construction installation A4-5, and end-of-life stage C1-C4). It covers impacts like fossil fuel depletion, eco-toxicity, and land use. The dataset has high dimensionality, non-linearity, and 6,194 samples, with two modalities: categorical text data listing building elements (nominal/categorical data) and materials and numerical environmental impact indicators (numerical/continuous data).

The conceptual model shows the material name, building element, life cycle stage, and environmental indicator as independent variables ($x_1 - x_{13}$). The GWP is studied as the dependent variable (y), see Fig. 5. The GWP contains greenhouse gas emissions, CO_2, CH_4, N_2O, and CO, and considers 'fossil', 'from soil or biomass stock', and 'non-fossil' resources (Röck et al., 2023). The prediction tasks use non-automated ML and two automated ML methods. The former introduces SVM and an ANN with a feedforward architecture. The latter uses an ensemble model and AutoML for multimodality. As such, multi-label predictions can also be enabled to predict multiple environmental indicators based on material type definitions, yet these are out of scope for this study.

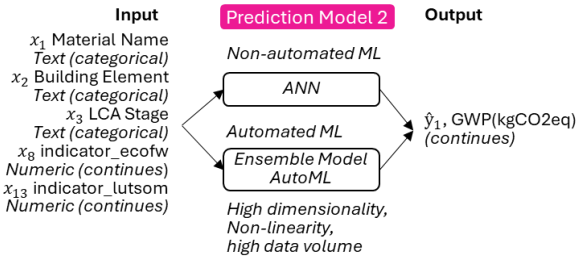

Figure 5: Conceptual model for material GWP prediction

Data preprocessing

The nominal (text) variable (x_1–x_6) showing information on material types and life cycle stage-related information are encoded, transforming the strings to dummy code to a multi-dimensional array. The data retains correlating features and supports a more significant model fit. Correlation matrix identified high correlating features, which enabled to reduce the input targets. The null values are eliminated, reducing to a dataset n=4181. The dataset is split 70% for training, 15% for testing and 15% for validation.

Machine Learning model and training

The ANN with multilayer perception topology uses a regression model to predict the GWP. The multilayer perceptron (MLP*)* neural network is a nonlinear regression model where the perceptron forms the basic unit of a neural network (5). The x represents the input vector to the neural network, ω represents the weights associated, and φ is the activation function. As an activation function (6), the Rectified Linear Unit (ReLU) $\varphi(z_i) = \max(0, x)$ is chosen. The Sigmond and Tanh Functions were tested yet provided less significant results.

$$f(x) = \varphi(\mathrm{w}, \mathrm{x}) \quad (5)$$

$$a_i = \varphi(z_i) \quad (6)$$

Neural networks are trained using feedforward propagation and backpropagation. Feedforward moves information forward, while backpropagation adjusts weights and biases based on errors, optimising performance with gradient descent. Gradients of the loss function (∇L) guide updates to the network's parameters. The MLP consists of an input layer (100 neurons), one hidden layer (200 neurons), and an output layer. Keras and TensorFlow are used, with each neuron corresponding to an input variable. The input layer is connected to the hidden layer via 20,000 weights (100×200 connections).

The second model is the SVR, which extends SVM by using an ε-insensitive loss function, which defines a "tube" around the hypothesis function that ignores minor errors. The aim is to size the tube to find the best trade-off between the continuous function while minimising the prediction error, the difference between the predicted and the true value. Different loss functions, including linear, quadratic, and radial basis kernels, are tested based on linear and non-linear mappings. The radial-based kernel performs best for non-linear data.

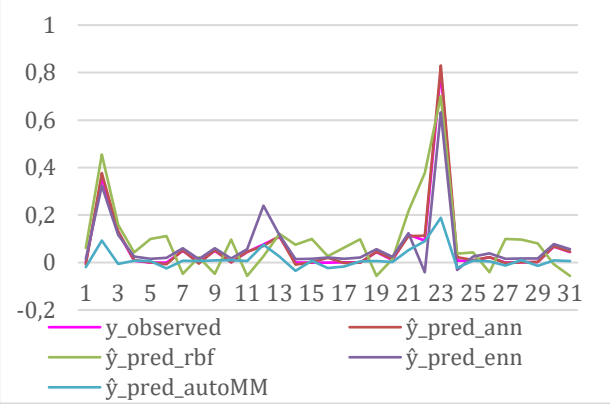

Figure 6: Material GWP prediction (kgCO₂eq). Compare observed GWP (y), with the predicted GWP (ŷ) from the ANN (ann), the SVM with the radial base function as kernel (rbf), the ensemble model (esm), and the AutoML model (AutoMM)

Finally, the ensemble model and AutoML are implemented using the Autogluon library. The tabular data containing multimodal data types is used as input without encoding and fed into the neural network. The results of the observed and predicted GWP are presented in Fig. 6. The R^2 and the MAE in Table 2 indicate that the ANN and AutoML perform best.

Machine Learning model performance evaluation

Overfitting and underfitting are studied to assess how well a model learns from input data and generalises to make accurate predictions on unseen data. The loss function, using the MAE metric, helps identify overfitting and underfitting by comparing the training and validation losses. A higher bias (underfitting) results in lower variance in parameter estimation, while lower bias (overfitting) leads to higher variance. Hyperparameter tuning adjusts the model's bias, which reflects the closeness of predictions to the target and variance. This involves initialising weights, normalising activations, optimising the learning rate, and early stopping.

Table 2: Model validation using R2 and MAE of studied ML models for global warming potential prediction

	ANN	SVM (RBF)	Ensemble Model	AutoML
R^2	0.967	0.898	0.689	0.946
MAE	0.07	0.16	241.66	0.122

Building type and archetype definition

OSM and 3DBAG are data sources that hold (semantic) information representing the existing building stock. Both platforms hold unique identifiers (ID) (ref: bag) per building; thus, semantic information can be cross-referenced. Building-specific information is derived through API requests to the Overpass API and 3D BAG API. From the OSM, we retrieve building typology, construction date and building material type. From the 3DBAG, the area and volume of building elements, such as exterior walls, roof, and ground floor, are retrieved (Listing 1 and Listing 2, respectively).

```
Endpoint OSM:
https://overpass-api.de/api/interpreter
HTTP Method: POST
HTTP Headers: Content-Type: application/json
Body:                                    [out:json];
way["ref:bag"="772100000297133"]; out tags;
Response: “Building Material:brick” “Building
Type:house” “ConstructionYear:1921”
```

Listing 1 Overpass API POST endpoint to retrieve building type, building material, and construction year

```
Endpoint 3DBAG:
https://api.3dbag.nl/collections/pand/items/NL.
IMBAG.Pand.0772100000297133
HTTP Method: GET
HTTP Headers: Content-Type: application/json
Response: “External Wall Surface:123.5” “In
Between Wall Surface:89.13” “Finish Wall
Surface:34.37”
```

Listing 2 3D BAG API GET endpoint to retrieve building elements, external wall surface.

Building typologies can be categorised into archetypes. In a Dutch context, building types such as residential single-family and multifamily houses are further categorised into

detached, semi-detached, terrace, gallery and apartment houses. An archetype represents statistically derived average values for occupation area, thickness of building elements, construction and material types, and thermal resistance values. The TABULA representing building typologies in Europe (Loga et al., 2016) and the reference type of existing residentials in the Netherlands (RVO, 2022) are used as data sources.

Material Circularity and Embodied Carbon Computation

MCI utilises the Linear Flow Index (LFI) and the product's Utility (U) (7). The LFI is the product of the amounts of total mass (M′), Virgin input (V), and material Wastage (W) (8). The U refers to the lifetime of the product L_{brand} and the minimum technical (TL) or functional lifetime (FL) (9) (Khadim et al., 2025). We assume 60 years for L_{brand} and 100 years for TL. The LFI is calculated using the mass (kg) derived from the density prediction (kg/m^3) and the volume (m^3).

$$MCI = Max\ (0{,}1 - UxLFI) \quad (7)$$

$$LFI = \frac{V+W}{2M'} \quad (8)$$

$$U = \frac{0.9}{\min(FL,TL)/L_{brand}} \quad (9)$$

Recent developments show a direct link between GWP impact (via LCA, ISO 14040, EN 15978) and MCI. Material mass (kg) associated with extraction (A1), production (A3), transport (A2, A4), construction (A5), and EOL (C1-C4) affects GWP distribution across the life cycle. However, only by knowing the building's and elements' technical lifetimes can the MCI be accurately calculated. The embodied carbon impact (10) is derived by multiplying the predicted GWP by the element's mass.

$$EC_{BE}(kgCO2eq) = \sum(GWP\ (\ kgCO2eq) \ast mass\ (kg)) \quad (10)$$

Experiment and Results

The contextual information for a specific Dutch residential house (ID: 7,721E+14) is extracted (see Table 3). It contains building type, construction year, surface area, and building materials (as derived from OSM and 3DBAG). The construction type and estimated thicknesses of the building elements are derived from the archetype definitions (RVO and TABULA). The material property predictions for density and GWP are performed for the corresponding brick material, specifically the clay facing brick in a cement mortar joint. The material density yields a predicted value of 1672.55 kg/m³ compared to the observed value of 1737.69 kg/m³. The predicted GWP potential is 0.127 kg CO2eq, and the observed value is 0.118 kg CO2eq, as shown in Fig. 7. The performance assessment for MCI and the embodied carbon impact is 0.0923 (where 1 represents a fully circular system and 0 represents a linear system). The embodied carbon impact yields 2,190.63 kg CO2eq for a 30 cm brick external wall.

Table 3: Terrace House as-built characteristics (ID: 7,721E+14)

Data source	Terminology	Instance
3D BAG	Building Type	Residential
RVO	Archetype	Terrace House
BAG	Construction Period	1921
RVO	Façade Wall Surface	34.37 m^2
RVO	Façade Wall Thickness	0,3 m
TABULA	External Wall	Monolith no insulation
Prediction	Density Predicted	1672,55 kg/m^3
Prediction	GWP Predicted	0,127kgCO_2eq
Computed	Mass per BE	17245,69kg/BE
Computed	MCI per BE	0,0923 index
Computed	Embodied carbon per BE	60,75kgCO_2eq

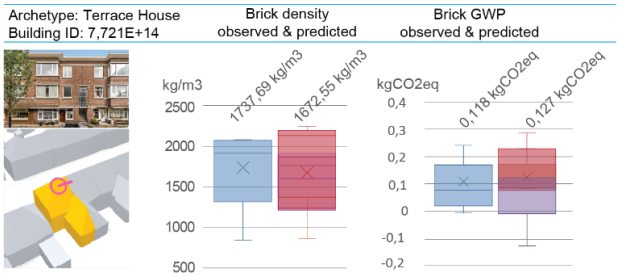

Figure 7: Selected residential house (ID: 7,721E+14), archetype, predicted density property and global warming potential

Conclusions

This study presents a multimodal pipeline for predicting material properties in the context of existing buildings. Both non-automated and automated ML models are tested to predict material density and GWP. Contextual information from a Dutch residential building is extracted via API endpoints to enrich the ML pipelines. The predicted material properties estimate the MCI and embodied carbon of a building's exterior wall in a use case. The non-automated ML pipelines are labour-intensive, yet offer greater flexibility and adaptability in feature engineering informed by domain knowledge. In contrast, automated ML significantly outperforms in multimodal data processing, requiring minimal effort and offering reusable models with few lines of code.

The proposed integration pipeline excels in its use-case-specific adaptability and flexibility, enabling more accurate material property estimation through contextualisation. Instead of relying on average material properties, the prediction model surpasses conventional methods, supporting more informed end-of-life decision-making for materials. As a limitation, this study is focused on the Dutch context, and additional expert knowledge would enhance ML methods and knowledge extraction. Future work includes specifying material subtypes, predicting additional material properties, automating ML using more datasets, incorporating extra modalities (e.g., imagery), and investigating multi-label predictions. Material type matching will evolve from manual to semantic similarity matching, utilising natural language processing and ontological concepts.

References

American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. (ASHRAE). (2021). 2021 ASHRAE® Handbook - Fundamentals (SI Edition).

Byers, B. S., Raghu, D., Olumo, A., de Wolf, C., & Haas, C. (2024). From research to practice: A review on technologies for addressing the information gap for building material reuse in circular construction. In *Sustainable Production and Consumption* https://doi.org/10.1016/j.spc.2023.12.017

de Rijksdienst voor Ondernemend Nederland (RVO). (2022). *Voorbeeldwoningen 2022 bestaande bouw In opdracht*.

Fenton, S. K., Munteanu, A., de Rycke, K., & de Laet, L. (2024). Embodied greenhouse gas emissions of buildings—Machine learning approach for early stage prediction. *Building and Environment*, *257*. https://doi.org/10.1016/j.buildenv.2024.111523

Feurer, M., Klein, A., Jost, K. E., Springenberg, T., Blum, M., & Hutter, F. (2015). *Efficient and Robust Automated Machine Learning*. http://automl.org

Jin, H., Chollet, F., Song, Q., & Hu, X. (2023). AutoKeras: An AutoML Library for Deep Learning. In *Journal of Machine Learning Research* (Vol. 24). http://jmlr.org/papers/v24/20-1355.html.

Khadim, N., Agliata, R., Han, Q., & Mollo, L. (2025). From circularity to sustainability: Advancing the whole building circularity indicator with Life Cycle Assessment (WBCI-LCA). *Building and Environment* https://doi.org/10.1016/j.buildenv.2024.112413

Loga, T., Stein, B., & Diefenbach, N. (2016). TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable. *Energy and Buildings*,*132*, 4–12.

Ma, Y., Zhao, S., Wang, W., Li, Y., & King, I. (2022). Multimodality in meta-learning: A comprehensive survey. *Knowledge-Based Systems*, *250*. https://doi.org/10.1016/j.knosys.2022.108976

Moschen-Schimek, J., Kasper, T., & Huber-Humer, M. (2023). Critical review of the recovery rates of construction and demolition waste in the European Union – An analysis of influencing factors in selected EU countries. *Waste Management*, *167*, 150–164. https://doi.org/10.1016/j.wasman.2023.05.020

Pio, P. B., Rivolli, A., Carvalho, A. C. P. L. F. de, & Garcia, L. P. F. (2024). A review on preprocessing algorithm selection with meta-learning. In *Knowledge and Information Systems* (Vol. 66, Issue 1, pp. 1–28). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s10115-023-01970-y

Röck, M., Passer, A., & Allacker, K. (2023). Slice: An Open Data Model for Scalable High-Definition Life Cycle Engineering, Hotspot Analysis and Dynamic Assessment of Buildings. Zenodo. https://doi.org/10.5281/zenodo.8369244

Rosa Latapie, S., Sabathier, V., & Abou-Chakra, A. (2024). Bio-based building materials: A prediction of insulating properties for a wide range of agricultural by-products. *Journal of Building Engineering*, *86*, 108867. https://doi.org/10.1016/j.jobe.2024.108867

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., & García-Herrera, R. (2018). Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*, *90* (December 2016), 728–741. https://doi.org/10.1016/j.rser.2018.04.008

Salehin, I., Islam, Md. S., Saha, P., Noman, S. M., Tuni, A., Hasan, Md. M., & Baten, Md. A. (2024). AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, *2*(1), 52–81. https://doi.org/10.1016/j.jiixd.2023.10.002

Stergiou, K., Ntakolia, C., Varytis, P., Koumoulos, E., Karlsson, P., & Moustakidis, S. (2023). Enhancing property prediction and process optimization in building materials through machine learning: A review. In *Computational Materials Science* (Vol. 220). https://doi.org/10.1016/j.commatsci.2023.112031

Smeyers, T. & Mertens (2022). *REUSE TOOLKIT THE RECLAMATION AUDIT*. http://www.nweurope.eu/fcrbe

Tang, Z., Fang, H., Zhou, S., Yang, T., Zhong, Z., Hu, T., Kirchhoff, K., & Karypis, G. (2024). AutoGluon-Multimodal (AutoMM): Supercharging Multimodal AutoML with Foundation Models. https://doi.org/10.48550/arXiv.2404.16233