2024 EC³
2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 14-17, 2024

# CAN MACHINE LEARNING AUTOMATE CARBON CLASSIFICATION OF MATERIALS WITHIN A BUILDING INFORMATION MODEL?

Abdulkadir Adeola Abdulrahman[1], Rogage Kay[2]
[1,2]Northumbria University, Newcastle, United Kingdom.

## Abstract

Buildings account for an estimated 35% of the UK's total greenhouse gas emissions. Assigning carbon data to building designs can aid contractors in sustainable material selection. Authoring inconsistencies in building data mean addition of environmental data is difficult and expensive. Machine learning enables classification of materials at a scale manual approaches cannot match. A machine learning approach is documented for classifying building products that enables automatic augmentation of environmental data from carbon databases. Our findings provide foundational research on automating data authoring, that can reduce costs and simplify processes associated with adding environmental assessment data to building designs.

## Introduction

In June 2019, UK parliament made a legally binding commitment to bring all Greenhouse Gas (GHG) emissions in the UK to Net Zero by 2050 (Department for Business, Energy & Industrial Strategy, 2019). The construction sector is a major contributor to the UK's carbon footprint and is responsible for 50% of all extracted materials and 35% of GHG emissions (Rogage et al., 2019). Construction activity accounts for around 50m tonnes of CO2 emissions, over half of this is linked to construction products and materials (Government Commercial Function, 2022). Modern approaches to collecting information about buildings across the supply chain offer unprecedented data resources for analysing and improving processes. Building Information Modelling (BIM) is an established term within academia and across the sector that is defined as a process for collaborative processing and management of design and construction information (Tallet et al., 2021). Whilst in its essence BIM is an information management approach, it is more specifically associated with standardised schemas and classification systems that enable interoperability and automated data processing. Cavalliere et al. (2019) stressed carbon management in construction is becoming an increasing priority as buildings make up such a high percentage of GHG emissions. Frameworks such as the Leadership in Energy and Environmental Design (LEED) (2024) and Building Research Establishment Environmental Assessment Method (BREEAM) (2024) certifications, enable contractors to calculate and certify embodied carbon impact when selecting building materials. Whilst calculating embodied carbon of buildings is not the aim of our research, our findings could link to existing frameworks such as LEED and BREEAM as part of the calculation method for the material selection process. Tools are needed that accurately classify materials within BIM and match them up to associated suitable carbon database items in order to optimise carbon management strategies (Cavalliere et al., 2019).

Whilst several classification systems exist, often product material data is unclassified or classified using non-standard approaches, making it difficult to automate assigning carbon information to products. This research explores a Machine Learning (ML) approach to address the problem of assigning carbon information to products. We seek to understand the potential for ML to accurately identify material descriptions in BIM data, so information from carbon databases can be automatically augmented to building products. Our developed approach uses text and sentiment analysis, and classification techniques in order to match BIM material descriptions with those found within carbon databases. A comprehensive literature review of classification algorithms using BIM data and carbon management provides the foundation for developing the ML approach. We then select and evaluate a number of algorithms for classifying BIM data against carbon impact categories in the Inventory of Carbon Energy (ICE) (Circular Ecology, 2020) database. We propose a novel application of ML techniques for predicting and classifying incomplete or inaccurate BIM data.

### Related Work

BIM has rapidly become an indispensable approach in AEC industries for creating digital representations of building designs (Basbagill, 2013). A BIM consists of any elements that represent assets within a building such as doors, windows, walls etc. Consistent and accurate classification of BIM elements including materials, remains an ongoing challenge for AEC professionals. Honic et al. (2019) found that inconsistent data and a lack of cooperation among different stakeholders, created a barrier to automation of data related to recycling of building materials. Manual classification of BIM elements can be time consuming, ineffective and create inconsistencies within data. An automated classification approach utilising ML could significantly enhance accuracy and efficiency when applied to element classification of BIM data.

Studies have highlighted ML's utility for accurately classifying heritage buildings (Bassier et al., 2017), maintenance issues (McArthur et al., 2018), wall and door elements (Koo et al., 2021) and quantifying environmental impact assessments (Cavalliere et al., 2019; Xu et al., 2022). Deep Learning (DL) is a subcategory of ML that uses historical experience as the basis for future predictions. DL specifically has shown promise as an accurate way of classifying complex BIM elements based on geometry (Koo et al., 2021; Rogage and Doukari, 2024). One area in which ML algorithms have been applied to BIM is in assessing embodied environmental impacts of building designs. Basbagill et al. (2013) and Cavalliere et al. (2019) offer examples of this application using Life Cycle Assessment (LCA), continuous BIM-based assessment and an automated LCA approach. Xu et al. (2022) proposed an innovative BIM-integrated LCA for prefabricated buildings as an automated way of performing the embodied carbon assessment process. All three studies demonstrate the potential of ML/BIM combination in mitigating environmental impacts while optimising building design performance. ML algorithms have also been successfully utilised with BIM for classification of maintenance issues and enhanced data collection, according to an approach devised by McCarthy et al. (2018) utilising ML visualisation techniques in BIM to classify issues while increasing data collection rates. Zabin et al. (2022) conducted an in-depth literature review pertaining to applications of ML to BIM projects and highlighted its capacity for automated quality control and error detection. Koo et al. (2021) introduced an automated approach for accurate classification of BIM elements using 3D geometric Deep Neural Networks (DNN) as walls and door elements. Rogage and Doukari (2024), further demonstrate use of 3D geometric DNN for classifying a further range of products. These studies demonstrate the promise of ML with BIM for improving accuracy and efficiency of classification of elements within BIM. However, consideration must be made regarding possible challenges or restrictions associated with its usage; including issues surrounding data quality, privacy concerns and interpretability (Zabin et al., 2022).

**Recurrent Neural Networks**

Here we consider ML models for classifying BIM data. Recurrent Neural Networks (RNN) are a type of artificial neural network that use layers to process data. RNN with Long Short-Term Memory (LSTM) architecture comprises three layers: input; recurrent; and output (Nosouhian et al., 2021). RNN input layers (shown on the left of figure 1) accept input sequences such as time series data or natural language text as vectors; then feed these vectors one at a time into their networks for processing. A type of RNN called LSTM units can store long-term memory, which enables them to comprehend text sequences better by capturing context and semantics (Kaur and Mohta, 2019). An RNN's recurrent layer serves to maintain its memory of previous inputs and generates outputs based on each cell taking an input and producing an output, along with updating their hidden state, each time step based on current and previous hidden states - thus giving rise to RNNs' ability to recall past data inputs (Kaur and Mohta, 2019). Within each layer are nodes or neurons, represented by circles in figure 1, which perform simple operations and pass signals to other neurons. Connections between neurons are represented by lines, the strength of connections determined by weights. Signals travel through the network from left to right via these weighted connections. As data passes through each layer, the network learns patterns and extracts meaningful features. An RNN's output layer (shown on the right of figure 1) takes the final hidden state from its recurrent layer and generates its output; it could take the form of either a simple feedforward layer, or it could employ complex features like softmax layers for classification tasks (Nosouhian et al., 2021). This output produces the network's predictions or classifications based on what it has learned. In many applications, neural networks contain multiple hidden layers to solve increasingly complex problems. The number and arrangement of layers and neurons can be modified to suit different tasks. One of the primary strengths of RNNs lies in their capacity to handle variable-length data sequences efficiently, due to using identical weights across time steps of sequence processing allowing it to process sequences of various lengths without disrupting operation of the network.

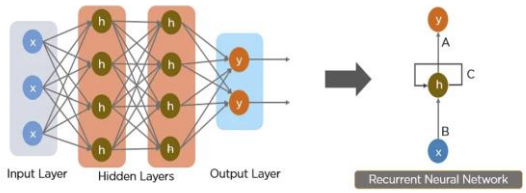


*Figure 1: Deep Learning architecture* (Badhan et al, 2024)

Figure 2 demonstrates how RNNs process sequential data. RNNs are commonly used for natural language processing tasks where the order of inputs is important, such as text classification. We can think of this grid as representing a sequence of inputs an RNN would read, from left to right and top to bottom. Each cell contains a token or word from the sequence. As the RNN processes each step, it considers both current input and information from the previous step, represented by the cell to the left. This looping, recurrent structure allows RNNs to connect previous information to later inputs in a sequence. The network develops an internal state capturing features of the entire sequence processed so far which is passed from each cell to the next. By reading the sequence from this grid in order, an RNN would build up understanding of word patterns and relationships over multiple time steps. This visualisation conveys how RNNs can process variable-length sequential data by sharing parameters across every step. The 2D array representation provides a simple way to conceptualise how RNNs flow information

through a network from start to end of a sequence, capturing context at each point to help classify or generate sequential outputs.

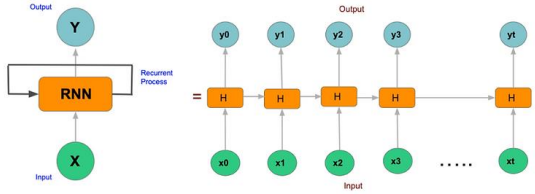


*Figure 2: RNN architecture* (Nagesh, 2020)

**Naive Bayes**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable (Qiang, 2010). Naive Bayes is a probabilistic classification algorithm that calculates the probability of an instance belonging to a class based on the probabilities of its features. It uses Bayes' theorem:

$$P(y|x) = \frac{(P(x|y)\ P(y))}{P(x)} \qquad (1)$$

Where:
P(y|x) is the probability of class y given the features x.
P(x|y) is the probability of features x given class y.
P(y) is the prior probability of class y.
P(x) is the probability of features x.

Naive Bayes estimates probabilities P(x|y) and P(y) from a training dataset. During prediction, it calculates the probability of each class label for a new instance and assigns it to the class with the highest probability. The model assumes features are independent, meaning the presence or absence of one feature does not affect the presence or absence of others. This assumption simplifies calculations and allows the model to handle high-dimensional feature spaces efficiently. Despite its simplicity, naive Bayes performs well in text classification tasks, such as spam detection and sentiment analysis. It is computationally efficient and can handle large datasets. However, it may not perform well when the independence assumption is violated or when there is a lack of sufficient training data.

**Random Forest**

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions (Wang and Wang, 2021). It randomly selects subsets of training data and features to train each tree. The final prediction is made by aggregating predictions of all trees, either through majority voting for classification or averaging for regression to improve prediction performance. The algorithm is illustrated by creating multiple decision trees, each trained on different subsets of data and features, and then combining their predictions. Random Forest is widely used for its effectiveness and ability to handle high-dimensional datasets.

Figure 3 shows three decision trees each making independent classifications. Branches of Decision Tree-1 and Decision Tree-2 represent two individual decision trees classifying an input sample and arrive at a "Result". These predictions are combined through majority voting. Majority voting is the process of having each tree "vote" for a class and selecting the class receiving the most votes overall. This aggregation helps to reduce variance and prevent overfitting compared to a single decision tree. By growing many decision trees on randomly sampled subsets of training data and combining their results, Random Forests are able to capture relationships any individual tree may miss. This ensemble approach typically yields better generalisation performance than a single estimator. This ensemble technique is effective for both classification and regression tasks.

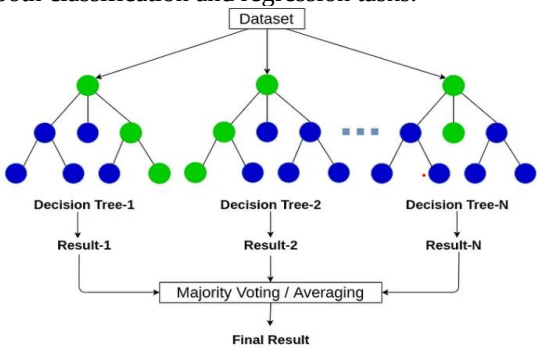


*Figure 3: Random Forest structure* (Zhu and Spachos, 2021)

## Simulation and Experiment

**Methodology**

The data science pipeline and methodology is shown in figure 4. This approach involves selecting data, then cleansing and preparing data using feature extraction with consideration to the model to be employed. Data exploration precedes model construction to comprehend the significance of various features, the relationships among features, data distribution patterns and hypotheses formulated about them. A model is then built with the aim of producing inferences or future event predictions from it or discovering a "root cause" behind an already observed event; its results are then assessed and presented upon evaluation.

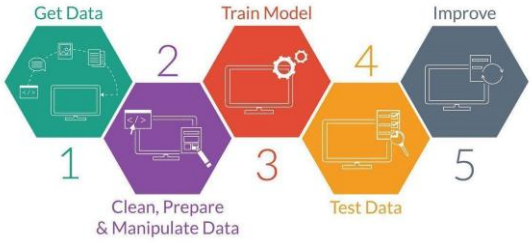


*Figure 4: Machine Learning workflow* (Muhammad, 2022)

**Data Sources**

The first data source contains approximately 10,000 material descriptions extracted from the ICE database (Circular Ecology, 2020). These include detailed information on embodied carbon and energy of various construction materials, reflecting their significant

influence in the construction industry. The second data source encompasses BIM data, including labels and classifications for different building products. The ICE database was selected for its comprehensive coverage and credibility in documenting embodied carbon and energy of construction products. This dataset is split into two parts: one containing detailed material descriptions, and the other consisting of material names or labels. Material descriptions provide extensive information about each building material, including its embodied carbon, which is crucial for the analysis.

**Exploring the Data**

Data exploration began with data cleaning and preparation. The BIM dataset comprised 11,488 rows with four columns covering Family, IFC Type, Material Types and categories such as "ICE Category" (this is the category the model will predict to automate the mapping of BIM data directly into the ICE database). Table 1 provides a dataset sample. After a preliminary study of the data, it became apparent two columns: Family and IFC Type; were not relevant because they are both classified in generic terms which are applicable to all material that falls under the general family representation respectively, so they were removed from the dataset entirely leaving only the Material and ICE Category columns.

*Table 1: Sample data from the BIM dataset*

| IFC Type | Material | ICE category | ICE Material |
|---|---|---|---|
| Air Terminal | Air Terminal - Diffuser Body | NaN | NaN |
| Air Terminal | Air Terminal - Perforated Plate | NaN | NaN |
| Air Terminal | Aluminium - Koolair | Aluminium | NaN |
| Air Terminal | Aluminium | Aluminium | NaN |
| Air Terminal | Anodised Aluminium - Koolair | Aluminium | NaN |

Further visualisation of the data (figure 5) revealed an uneven distribution of building materials. More specifically, certain samples had significantly more entries than others indicating an uneven representation in the data and potentially leading to biased predictions and reduced accuracy when training ML models. To address this challenge, data balancing was employed; an approach in which samples from each category are adjusted to create an even distribution for more accurate predictions through ML models trained on representative samples of data. There are various strategies available for balancing data, including oversampling and undersampling (Bonatti and Kirrane, 2019). Oversampling involves increasing the sample population from minority classes while undersampling involves decreasing it; both strategies aim to achieve an even distribution across categories. This study used oversampling, specifically the Synthetic Minority Over-sampling Technique (SMOTE), to balance data (Chu et al., 2016). SMOTE involves creating artificial samples in minority classes by interpolation among existing ones; thus ensuring each minority class receives enough samples, leading to more balanced distribution of data. By using SMOTE to balance data, the model was then trained on a more balanced distribution of materials within the dataset for more accurate predictions and vital insights for sustainable building practices.

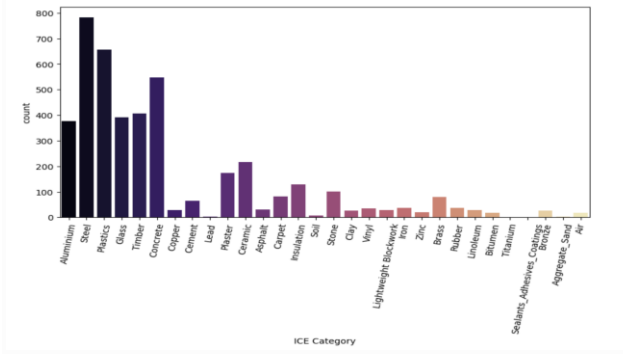


*Figure 5: Dataset Distribution*

A Bag of Words function was introduced to search and verify the presence of specific materials in the database, offering insights into their carbon footprints. This method represents each document as a vector of word frequencies, without regard to order or structure of words (Anello, 2021). This serves as indicators for classification of building materials, assisting in the ML process for accurate categorization and analysis. Use of these datasets allows for a detailed examination and understanding of the relationship between material properties and their environmental impact. This concluded the data preparation and manipulation stage ready for the final stage of training the model.

**Model Development**

The ML model was developed using three algorithms: RNN with LSTM, Random Forest, and naive Bayes. The combination of these algorithms offers a comprehensive approach to the classification of building materials, leveraging strengths of each to enhance overall effectiveness and reliability of the model in classifying building materials based on their carbon impact, as per the ICE recommendations.

**Model Training**

In this project, a common technique for training network parameters and performing classification was employed: splitting available data sets 80/20 on a train/test split. This strategy is known as train-test split and commonly utilised within ML to assess model performance. At this phase of training, the model was trained on its training set by using an optimization algorithm to adjust network parameters and minimise its loss function. Alpaydin (2020) defines loss function as measuring the difference between

predicted output of the model and actual output and an optimization algorithm can adjust these network parameters so as to decrease this difference as much as possible. Once the model is trained and evaluated on its test set, the next stage is to improve the model. Cross-validation is used for measuring generalisation ability, testing the model on new data allows for a more accurate representation of its performance on unseen sets of information and provides an opportunity to improve the model. Hyperparameter tuning was conducted to improve the Random Forest classifier model. The number of estimators was set to 400, and the criterion for measuring the quality of a split was chosen as "entropy". The random state was also fixed to ensure reproducibility of results. This careful selection of hyperparameters aims to optimise the model's performance by adjusting its complexity and learning capacity. For the Bidirectional LSTM model, specific architecture choices such as embedding dimension, number of LSTM units (16), and use of dropout for regularization indicate a tuning process aimed at balancing model capacity with the need to avoid overfitting. While the description doesn't detail an iterative search over a range of values, these choices reflect considerations typical in model tuning, focusing on optimising the network's structure for the task at hand. The metrics used to evaluate the accuracy and completeness of the model's predictions are Precision, Recall, and F1 Score.

**Precision** is the ratio of true positives (correct predictions) to the total number of positive predictions. It measures how precise the model is in identifying relevant cases. A high precision means that the model has a low rate of false positives (incorrect predictions).

**Recall** is the ratio of true positives to the total number of actual positive cases. It measures how well the model can recall or retrieve all relevant cases. A high recall means the model has a low rate of false negatives (missed predictions).

**F1 Score** is the harmonic mean of precision and recall. It balances both metrics and gives a single score that reflects overall performance of the model. A high F1 score means the model has both high precision and high recall. From table 2, Random Forest model has the highest scores across all metrics at 99%. This means the Random Forest model is the most accurate and complete in its predictions, compared to the other two models. Both naive Bayes and RNN have a Precision of 98%, Recall of 97%, and an F1 Score of 97%. These scores are slightly lower than the Random Forest model, but still very high and impressive.

## Discussion and result analysis

Multiple experiments using the RNN (LSTM), naive Bayes, and Random Forest models to classify building materials into various ICE categories were conducted. Training the model involved using datasets of building material data before testing its performance on test sets. A subset of building material data were manually classified with corresponding ICE categories before comparing accuracy between each classification method. Results of the experiment demonstrated high levels of accuracy on their test sets, achieving 97%, 98.38%, and 99.59% (See table 2) respectively. During data preprocessing, the dataset was simplified by removing columns ('IFC Type', 'Family', 'ICE Material') not directly relevant to the primary goal of predicting the 'ICE Category' from the 'Material' description. This decision represents a targeted form of feature selection, focusing the model's learning on the most relevant data available. The strategy adopted for imputation was straightforward yet effective: rows with missing 'ICE Category' values, which are critical for supervised learning, were dropped. This approach ensures the models are trained on complete records, enhancing the reliability of their predictions. However, it's important to note that this method of handling missing values might not be suitable for datasets where such omissions could lead to significant loss of valuable information. In those cases, more sophisticated imputation techniques might be necessary. It is crucial to note that these results, while promising, require further validation to ensure their reliability and applicability in real-world scenarios. The comparison with manual classification methods, although indicative of potential efficiency gains, should not overshadow the importance of accurate and safe classification of building materials. Table 2 shows how each model performed using different evaluation metrics.

*Table 2: Model comparison Between Random Forest, naive Bayes and RNN results*

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Random Forest | 99% | 99% | 99% |
| Naive Bayes | 98% | 97% | 98% |
| RNN | 98% | 97% | 97% |

Table 3 shows the first 5 rows of materials predicted using the RNN model. This means that the RNN model has successfully classified the building materials in the ICE Category based on what the model understands from the training dataset. This was the first model developed, while it successfully predicted the materials, the accuracy of its prediction could not be confirmed as this model does not suppose the confidence level measure.

*Table 3: RNN Prediction*

| Material | ICE Category |
|---|---|
| Air Terminal - Diffuser Body | Air |
| Air Terminal - Perforated Plate | Air |
| ConnectorSolidMaterial | Concrete |
| Duct_Afkast | Steel |
| Duct_Udsugning | Brass |

Table 4 shows the first 5 rows of predictions with their confidence level by the naive Bayes model. The results show the model has classified some materials such as “Air Terminal - Diffuser Body” and “Air Terminal - Perforated Plate” under the “Air” category correctly, with high confidence levels of 18.05 and 17.53 respectively. However, the model has made errors, such as classifying “Connector/Solid/Material” and “Duct_Lining/lining” as “Linoleum”, with low confidence levels of 3.43 each. These materials are likely made of metal or plastic, not linoleum, which is a type of flooring material. Similarly, “Duct_Airseal” is categorised as “Steel” with a confidence level of 9.05, which may not be accurate if the duct is made of another material.

In ML, confidence level is a measure of how likely a prediction is to be correct. It is usually expressed as a percentage between 0 and 100, where higher values indicate higher confidence. In the case of the naive Bayes prediction in table 4, where the model predicts that a building material belongs to the “Air” category with a confidence level of 18.05, it means that the model got 18.05% vote after splitting the total available percentage to the possible materials, materials with the highest vote gets predicted and the higher the percentage the better chance of the prediction being accurate (Srivastava et al., 2018).

*Table 4: Naive Bayes prediction*

| Material | ICE Category | Confidence Level |
|---|---|---|
| Air Terminal - Diffuser Body | Air | 18.05% |
| Air Terminal - Perforated Plate | Air | 17.53% |
| ConnectorSolidMaterial | Linoleum | 3.43% |
| Duct_Afkast | Steel | 9.05% |
| Duct_Udsugning | Linoleum | 3.43 |

Table 5 below also depicts the first 5 rows of material predictions and their confidence level by the Random Forest model. This model performs best although with 1% margin, and higher confidence level out of the models explored. The confidence level is calculated by taking the proportion of votes from the decision trees that agree on the same label. For example, if 100 trees are in the forest, and 96 of them vote for the “Air” category for the “Air Terminal - Perforated Plate” material, then the confidence level is 96/100 = 0.96 or 96%.

*Table 5: Random Forest prediction*

| Material | ICE Category | Confidence Level |
|---|---|---|
| Air Terminal - Diffuser Body | Air | 99.25% |
| Air Terminal - Perforated Plate | Air | 96.00% |
| ConnectorSolidMaterial | Concrete | 69.75% |
| Duct_Afkast | Steel | 89.00% |
| Duct_Udsugning | Steel | 77.75% |

The models demonstrated impressive performance because of the focused feature set. By concentrating on the 'Material' descriptions as the primary input feature and carefully preprocessing this text, the models could learn effectively from the most relevant data. Also, by addressing class imbalance through resampling techniques such as SMOTE or direct resampling helped mitigate biases and improve model performance across less represented categories. Finally, optimised model configurations including deliberate setting of hyperparameters and model architecture choices, such as number of estimators in the Random Forest and structure of the LSTM network, contributed to the fine-tuning of each model's ability to capture the nuances of the dataset.

## Conclusions

The contribution of this project lies in its exploration of ML techniques for automating classification of building data with carbon information. Whilst LCA tools and frameworks enable the calculation of embodied carbon, they rely on parametric modelling for environmental assessment, however, its application within BIM has been relatively underutilised and, more importantly, does not exist within commercial software packages used in BIM (Alwan et al., 2021). In the absence of this data, we offer an automated technique for classifying BIM products so that environmental assessment data within carbon inventory databases can be augmented to models, thus removing the need for authors to manually generate this data at design stages.

Our approach offers a foundation upon which further research can develop model accuracy and further validate the results. It demonstrates the potential of ML in automating classification tasks, which could lead to significant productivity gains in the construction industry. However, we acknowledge that our results are a preliminary step in this direction and that further research is needed to refine these techniques, improve accuracy, and fully realise their potential. Future research in this area should focus on enhancing the validation processes for ML models in construction, ensuring that they not only achieve high accuracy in controlled tests but also maintain this accuracy in practical applications. Further work is

also required to investigate the integration of additional variables and data sources to enrich the models and capture a more comprehensive range of factors affecting building material classification. By advancing these aspects, subsequent studies can build upon our findings, contributing further to the knowledge and application of ML in construction and ensuring that the transition towards automated processes prioritises safety, accuracy, and reliability.

**Limitations and Further Recommendations**

Although this project demonstrated the efficiency of using ML to classify building materials into different ICE categories, some limitations must be acknowledged within its research design. Performing the study on a larger dataset would improve generalisability and make networks more resilient against fluctuations. Additionally the quality of data used for training and testing the ML models could be improved. Accuracy depends heavily on input quality; any errors or inconsistencies could compromise network performance significantly.

This study concentrated on classifying building materials based on their carbon impacts according to ICE. While this approach is instrumental in understanding and mitigating the environmental footprint of construction materials, it's important to acknowledge that this focus on carbon impacts alone presents certain limitations. For instance, the classification does not account for other environmental aspects such as the life cycle impacts of materials beyond carbon emissions, which can also be significant in sustainable construction practices. Future studies might expand the scope to include a more holistic environmental assessment of building materials, encompassing a broader range of environmental metrics beyond just carbon impacts.

It is also recognised that our proof of concept prototype solution utilised the UK based ICE database for linking product data to carbon calculations via materials. The embodied carbon coefficients of building materials, vary from country to country. To address this, further validation is required to assess the generalisability of the solution on databases focusing on other countries such as Öekobaudat (Federal Institute for Building, Urban and Spatial Research, 2024), Gabi (Sphera, 2024) and Bauteilkatalog (2024).

**Future Work**

Future work could include expanding the study to cover other properties of building materials such as durability, strength and cost. Several other ML techniques can also be evaluated in order to determine which algorithm best predicts the materials. By employing ML techniques to predict these characteristics of materials used in construction projects, more informed decisions could be made regarding selection and usage for more sustainable and cost-effective outcomes. More research could also focus on creating ML models specifically tailored for sustainability in construction industry applications, by including sustainability criteria into ML model development processes.

## Acknowledgements

The datasets used within this study were provided by xbim Ltd who provided access to the ICE data alongside industry BIM data. The proprietary aspect of the ICE dataset means that it cannot be made available for further analysis by researchers outside the research team. Additionally, whilst the BIM data provided by xbim Ltd was anonymised, we were not authorised to make this dataset public. We would like to thank xbim Ltd for providing access to both datasets, and time to support the researchers in better understanding the data, without their support this project would not have been possible.

## References

Alpaydin, E. (2020). Introduction to machine learning. MIT Press, pp. 003-012.

Alwan, Z., Nawarathna, A., Ayman, R., Zhu, M., & ElGhazi, Y. (2021). Framework for parametric assessment of operational and embodied energy impacts utilising BIM. Journal of Building Engineering, 42, 102768.

Anello, E. (2021). A friendly guide to NLP: Bag-of-Words with Python example. Available at:https://www.analyticsvidhya.com/blog/2021/ 08/a-friendly-guide-to-nlp-bag-of-words-with- python-example/ (Accessed: 31 October 2023). Badhan, A.K., Bhattacherjee, A. and Roy, R., 2024. Deep Learning Techniques in Big Data Analytics. In Data Analytics and Machine Learning: Navigating the Big Data Landscape (pp. 171-193). Singapore: Springer Nature Singapore.

Bassier, M., Vergauwen, M. and Van Genechten, B., 2017. Automated classification of heritage buildings for as-built BIM using machine learning techniques. ISPRS Annals of the photogrammetry, remote sensing and spatial information sciences, 4(2W2), pp.25-30.

Basbagill, J., Flager, F., Lepech, M. and Fischer, M., 2013. Application of life-cycle assessment to early stage building design for reduced embodied environmental impacts. Building and Environment, 60, pp.81-92.

Bonatti, P.A. and Kirrane, S., 2019, July. Big Data and Analytics in the Age of the GDPR. In 2019 IEEE International Congress on Big Data (BigDataCongress) (pp. 7-16). IEEE.

Building Research Establishment Group (2024). BREEAM. Available at: https://bregroup.com/products/breeam/ (Accessed: 21 March 24).

Chu, X., Ilyas, I.F., Krishnan, S. and Wang, J., 2016, June. Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 international conference on management of data (pp. 2201-2206).

Circular Ecology (2020) Embodied Carbon Footprint Database. Available-at: https://circularecology.com/embodied-carbon-footprint-database.html (Accessed: 22 November 2023).

Department for Business, Energy & Industrial Strategy (2019) UK becomes first major economy to pass net zero emissions law. Available at: https://www.gov.uk/government/news/uk- becomes-first-major-economy-to-pass-net-zero-emissions-law (Accessed: 22 November 2023).

Federal Institute for Building, Urban and Spatial Research. (2024). Oekobaudat database. Available-at:https://www.oekobaudat.de/en.htm (Accessed: 21 March 2024).

Government Commercial Function, 2022, Promoting Net Zero Carbon and Sustainability in Construction. Available-at:https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1102389/20220901-Carbon-Net-Zero- Guidance-Note.pdf (Accessed: 22 November 2023).

Holliger Consult. (2002). bauteilkatalog.ch component catalogue. Available at: https://www.bauteilkatalog.ch/Home. (Accessed: 21 March 2024).

Honic, M., Kovacic, I., Sibenik, G. and Rechberger, H., 2019. Data-and stakeholder management framework for the implementation of BIM-based Material Passports. Journal of building engineering, 23, pp.341-350.

Kaur, M. and Mohta, A., 2019, November. A review of deep learning with recurrent neural networks. In 2019 International Conference on Smart Systems and Inventive Technology (ICCSIT) (pp. 460-465). IEEE.

Koo, B., Jung, R. and Yu, Y., 2021. Automatic classification of wall and door BIM element subtypes using 3D geometric deep neural networks. Advanced Engineering Informatics, 47, p.101200.

McArthur, J.J., Shahbazi, N., Fok, R., Raghubar, C., Bortoluzzi, B. and An, A., 2018. Machine learning and BIM visualisation for maintenance issue classification and enhanced data collection. Advanced Engineering Informatics, 38, pp.101-112.

Muhammad Mudasser Afzal., 2022. 5core steps to understand machine learning work flow–a guide for beginners. Medium, Page 1. https://medium.com/@mudasserch1/5-core- steps-to-understand-machine-learning-workflow-a-guide-for-beginners-737040850d9b (Accessed: 21st March 24).

Nosouhian, S., Nosouhian, F. and Khoshouei, A.K., 2021. A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU.

Qiang, G., 2010, May. An effective algorithm for improving the performance of Naïve Bayes for text classification. In 2010 Second international conference on computer research and development (pp. 699-701). IEEE.

Rogage, K., Clear, A., Alwan, Z., Lawrence, T. and Kelly, G., 2019. Assessing building performance in residential buildings using BIM and sensor data. International Journal of Building Pathology and Adaptation, 38(1), pp.176-191.

Rogage, K., & Doukari, O. (2024). 3D object recognition using deep learning for automatically generating semantic BIM data. Automation in Construction, 162, 105366.

Srivastava, S., Shukla, A. and Tiwari, R., 2018. Machine translation: from statistical to modern deep-learning practices. arXiv preprint arXiv:1812.04238.

Tallet, E., Gledson, B., Rogage, K., Thompson, A. and Wiggett, D., 2021. Digitally-Enabled Design Management. In Handbook of Research on Driving Transformational Change in the Digital Built Environment (pp. 63-89). IGI Global.

Nagesh. SC., 2020. Introduction to RNN and LSTM. The AI dream. Available at: https://www.theaidream.com/post/introduction-to-rnn-and-lstm. (p1) (Accessed: 21st March 24).

United States Green Building Council (2024) LEED rating system. Available at: https://www.usgbc.org/leed (Accessed: 21st March 24).

Wang, H. and Wang, G., 2021. Improving random forest algorithm by Lasso method. Journal of Statistical Computation and Simulation, 91(2), pp.353-367.

Zabin, A., González, V.A., Zou, Y. and Amor, R., 2022. Applications of machine learning to BIM: A systematic literature review. Advanced Engineering Informatics, 51, p.101474.

Zhu, L. and Spachos, P., 2021. Support vector machine and YOLO for a mobile food grading system. *Internet of Things*, *13*, p.100359.