International Journal of BIM and Engineering Science

♦ ISSN: 2571-1075 Vol: 12, No. 02, 2026 ♦ PP. 19–24

Article DOI: https://doi.org/10.54216/IJBES.120203

FULL LENGTH ARTICLE

Geometry-Driven BIM Element Category Recognition from Open IFC Property Records: A Reproducible Engineering Science Study

Sonia Ahmed[1,*] **Marek Salamak**[2]

[1] Management in Construction Dept., CTU, Czech Republic

[2] Civil Engineering Dept., Silesian University of Technology, Gliwice, Poland

Emails: soniaahmad88@gmail.com; marek.salamak@polsl.pl

Received: December 04, 2025 Revised: January 01, 2026 Accepted: February 04, 2026 ⋆ Corresponding author

ABSTRACT

Accurate object semantics are essential for building information modeling (BIM) workflows to enable interoperability, model checking, quantity take-off, performance analysis, and other downstream engineering applications. However, in practice, Industry Foundation Classes (IFC)-based model exchanges often feature limited or poorly identified semantic tags, particularly during interoperability with authoring and reviewing tools. This research proposes a reproducible, geometry-based learning algorithm for the automatic recognition of BIM element categories based on publicly available IFC-based property data. The empirical analysis is based on 780 object instances from ten BIM categories from a publicly available sample of IFC object records. A rule-based parser translates semi-structured BIM text exports into engineering features as bounding-box dimensions, coordinates, elevations and object-status. The study compares three supervised machine-learning baselines via stratified five-fold cross-validation: logistic regression, random forest and extra trees. Random forest performed best overall with an accuracy of 0.992, balanced accuracy of 0.971, a weighted F1-score of 0.992, and a macro F1-score of 0.970. The analysis of feature importance shows that bounding-box height, width, length, spatial coordinates and externality-related descriptors are the most important features. The results demonstrate significant semantics can be extracted from minimal engineering descriptors without the need for deep learning of meshes. This work provides an interpretable and efficient baseline for BIM enrichment, assessment, and interoperability-focused preprocessing for engineering science use-cases.

Keywords: BIM ▪ IFC ▪ Semantic enrichment ▪ Engineering informatics ▪ Building data analytics ▪ Machine learning

1. INTRODUCTION

Building information modeling has emerged as a key digital platform for architecture, engineering and construction processes because it integrates geometry, semantics and process knowledge in a computational space. But the usefulness of BIM is heavily dependent on the properties of object semantics. In IFC exchanges, labels may be sparse, poorly standardized, or lost in software translations, which affects subsequent model analysis, cost analysis, performance analysis, and cross-disciplinary collaboration [13, 10, 3].

Recent research demonstrated that BIM semantics can be enriched via geometric deep learning, graph learning, multi-modal classification, and rule-based models [4, 12, 1, 11, 5]. At the same time, the computational efficiency and practicality of mesh-based or multi-modal workflows remain non-negligible, particularly in practical engineering practice, where users may only have access to exported object prop-

erty tables. Hence, there is a need for leaner, property-based alternatives.

The research described here contributes to this discourse by exploring the effectiveness of engineering features derived from object records of IFC-based BIM for recovering BIM categories. Rather than using mesh tensors or image projections, the workflow converts BIM object text exported with IFC to numeric descriptors, ranging from dimensions, elevations, bounding boxes, and physicality. This approach is deliberately practical as such data is often available in BIM authoring software, IFC viewers, and interoperability workflows.

The study makes three contributions. First, it describes an open BIM analytics approach that processes semi-structured IFC property data as a reusable machine-learning matrix. Second, it evaluates three interpretable supervised baselines under stratified cross-validation to set a baseline for performance. Third, it discerns the engineering features that best define BIM classes, hence determining what object properties are most informative for semantic enrichment.

The rest of the paper is structured as follows. Section 2 is a survey of recent developments in BIM semantic enrichment and classification. Section 3 describes the gap and the proposed model. Section 4 describes the working process and workflow of data processing and modelling. Section 5 reports the empirical results. Section 6 outlines the impact on BIM-enabled engineering practices and Section 7 provides the paper's conclusions.

2. LITERATURE REVIEW

Semantic enrichment has become a central issue in BIM because many valuable applications require greater object semantics than what is available in BIM exchanges. A retrospective study over the past decade by Xue et al. [13] demonstrated that semantic enrichment connects geometric, non-geometric, and topological data to enrich BIM and city models. Recent studies have highlighted that enrichment is not a single approach but a suite of methods including ontologies, machine learning, graph reasoning, data dictionaries, and hybrid methods [3].

In this context, automatic object classification has become crucial. Koo et al. [4] showed that 3D geometric deep neural networks can classify BIM wall and door subtypes, revealing that geometry contains rich semantic information. Following up on this work, Emunds et al. [1] introduced sparse-convolutional learning for IFC-based geometry classification and found it to be very efficient. These works demonstrated that semantics can be recovered even with the learning signal coming only from geometry.

Alongside this, other studies have expanded the feature space beyond geometry. Wang et al. [12] applied graph neural networks for classifying room types, revealing the significance of spatial and graph structure in BIM semantics. Utkucu et al. [11] tackled classification of architectural and MEP BIM objects for building performance analysis and demonstrated how automatic object recognition can alleviate the cost of manual interoperability repair. Liu et al. [5] also contributed to this strand with a multi-modal deep learning approach for using both graphical and non-graphical features of BIM objects. Overall, these works suggest semantics can be recovered from different information sources.

The other current line of research is the connection between BIM semantics and engineering processes. These include generation of synthetic data for point-cloud segmentation [14], domain-adaptive BIM-to-scan segmentation [2], and embodied-carbon assessment through construction classification systems [7] and facilities-management knowledge-graphs [8]. These works have a common thread: better BIM semantics deliver interoperability and extend the value of digital models.

However, there is a gap in implementation. Numerous proposed pipelines rely on mesh extraction, point-cloud generation, ontology development or massive multi-modal data. In typical engineering workflows, analysts may have access only to property tables or IFC viewer visualisations. This leaves a niche for lightweight property-based models fulfilling the requirements of transparency, low-preprocessing costs and sufficient accuracy. This paper explores this niche by evaluating the capacity of engineering descriptors extracted from IFC object tables to enable accurate category recognition without deep encodings of geometry.

3. RESEARCH GAP AND PROPOSED MODEL

Although recent BIM studies have reported strong results using geometric deep learning, graph neural networks, sparse convolutions, and multi-modal fusion, three unresolved challenges remain in day-to-day engineering practice. First, many methods assume access to rich geometric representations such as meshes, point clouds, or graph-structured model data, whereas operational BIM exchanges often reach analysts as tabular exports or viewer-based property records. Second, several high-performing pipelines are computationally demanding and require specialist preprocessing steps that are difficult to reproduce in consultancy, facilities-management, and project-audit settings. Third, the literature still provides limited evidence on how far semantically meaningful BIM classification can be driven by compact engineering descriptors alone.

These challenges define the research gap addressed in this paper. The unresolved question is not whether deep or multi-modal methods can classify BIM objects accurately, but whether a lighter and more transparent feature-driven approach can produce a reliable baseline when only IFC-derived property records are available. This question is important for engineering science because interoperability repair, model auditing, and preliminary semantic recovery often occur precisely in those constrained data conditions.

To address this gap, the proposed model adopts a geometry-driven classification strategy centred on structured engineering descriptors extracted from IFC-derived text records. The model is proposed as a reproducible baseline rather than a replacement for advanced deep-learning pipelines. Its contribution lies in three design choices. First, it converts semi-structured object descriptions into a tabular feature matrix that can be reused across projects and software environments. Second, it prioritizes variables with direct engineering meaning, including bounding-box dimensions, elevations, global coordinates, and physical state indicators. Third, it evaluates

multiple supervised learners under the same validation protocol in order to identify a robust and deployable baseline for BIM element recognition.

Conceptually, the proposed model can be expressed as a mapping from parsed IFC properties to a category decision:

$$\hat{y} = f\big(\phi(r)\big),$$

where r denotes the raw IFC-derived object record, $\phi(\cdot)$ is the deterministic parsing and feature-construction operator, and $f(\cdot)$ is the trained classifier. In this study, the most successful instantiation of $f(\cdot)$ is the random forest learner, which is well suited to heterogeneous non-linear relationships among engineering variables.

The proposed workflow directly addresses the identified research challenges. It reduces data dependency by operating on exported property records; it improves interpretability by using explicit engineering features rather than latent geometric embeddings; and it lowers implementation burden by avoiding mesh reconstruction, point-cloud generation, and complex multi-modal fusion. In this sense, the proposed model fills a practical gap between purely rule-based semantic repair and computationally intensive deep-learning approaches.

4. WORKING STEPS AND METHODOLOGY

4.1 Data Source

The empirical analysis uses a public sample dataset of IFC-derived object records distributed through an open GitHub repository for BIM classification experiments. The downloaded file, `testing.csv`, contains 887 exported BIM object records expressed as text strings with embedded property-value pairs. Each record includes a global identifier, an object name, and a text field containing descriptors such as category, elevations, geometric extents, and engineering attributes. After parsing the records and retaining categories with at least 10 observations for stable cross-validation, the modelling sample comprised 780 instances spanning ten categories.

Table (1) summarizes the final class distribution. The largest classes are Curtain Wall Mullions and Curtain Panels, while Railings and Plumbing Fixtures are minority classes. This class imbalance reflects realistic BIM export behavior, where facade assemblies are often represented by many repeated elements.

Table 1. Modeling sample after class-frequency filtering.

Category	Count
Curtain Wall Mullions	351
Curtain Panels	132
Walls	84
Doors	60
Furniture	43
Supports	34
Floors	27
Generic Models	20
Plumbing Fixtures	18
Railings	11
Total	780

4.2 Feature Engineering

Each text record was parsed using a deterministic property extractor that split the object description by semicolon delimiters and mapped the resulting key-value pairs into a structured table. To avoid target leakage in the primary experiment, the target category itself and textual reference fields were not used as predictors. Instead, the model relied on numeric engineering descriptors including:

- bounding-box length, width, and height;
- top and bottom elevations;
- global spatial coordinates;
- dimension fields such as `XDim`, `YDim`, `Span`, `Height`, `Width`, and `Depth`;
- binary or quasi-binary indicators such as `IsExternal`, `Has Own Geometry`, `Children Have Geometry`, and `LoadBearing`;
- auxiliary physical descriptors such as slope, roughness, pitch angle, and thermal transmittance when available.

Numeric fields were coerced to continuous values, and missing observations were imputed using the median within the preprocessing pipeline. Features were standardized for the linear model but preserved in the same modelling framework for tree-based learners.

4.3 Learning Task and Validation Strategy

The task is formulated as a supervised multi-class classification problem:

$$\hat{y} = f(x),$$

where $x \in \mathbb{R}^p$ is the engineered BIM feature vector and $\hat{y}$ is one of the ten BIM categories in Table (1). Three baseline models were evaluated:

1. Logistic Regression,
2. Random Forest,
3. Extra Trees.

Performance was assessed with stratified five-fold cross-validation to preserve class proportions across folds. The following metrics were computed:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i),$$

$$\text{Balanced Accuracy} = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k},$$

and weighted and macro F1 scores. Balanced accuracy and macro F1 were included because the dataset is imbalanced.

4.4 Mathematical Summary of the Contribution

The contribution of the study can be summarized as the following computational procedure:

1. $\forall r_i \in R, x_i \leftarrow \phi(r_i)$: parse geometric and engineering attributes.

2. $X \leftarrow [x_1, \ldots, x_N]^\top$, $y \leftarrow$ known category labels.
3. $X \leftarrow \Psi(X)$: numeric coercion, cleaning, and median imputation.
4. $\{(X_{tr}^{(k)}, y_{tr}^{(k)}), (X_{te}^{(k)}, y_{te}^{(k)})\}_{k=1}^{K} \leftarrow \text{StratifiedKFold}(X, y)$.
5. $\forall k, f^{(k)} \leftarrow \arg\min_{f \in F} L(y_{tr}^{(k)}, f(X_{tr}^{(k)}))$.
6. $\hat{y}_{te}^{(k)} \leftarrow f^{(k)}(X_{te}^{(k)})$.
7. $M \leftarrow \{\text{Accuracy}, \text{Balanced Accuracy}, \text{Weighted F1}, \text{Macro F1}\}$.
8. $f^\star \leftarrow \arg\max_{f \in F} \text{MacroF1}(f)$ and interpret $f^\star$ using feature importance.

Here, $\phi(\cdot)$ denotes the deterministic record parser, $\Psi(\cdot)$ denotes preprocessing, F denotes the candidate model set, and L denotes the training loss associated with each learner.

4.5 Implementation Workflow

Figure (1) presents the full workflow used in the study. The implementation was carried out in Python using `pandas`, `scikit-learn`, and `matplotlib`. All data files, code, processed tables, figures, and the manuscript source are included in the reproducibility package delivered with this paper.

Figure 1. End-to-end workflow for geometry-driven BIM category recognition.

5. RESULTS

5.1 Descriptive Statistics

Figure (2) shows the class composition of the filtered modelling sample. The distribution is skewed toward facade-related elements, particularly mullions and curtain panels. Such imbalance is expected in object-level BIM exports, where componentized systems can dominate record counts.

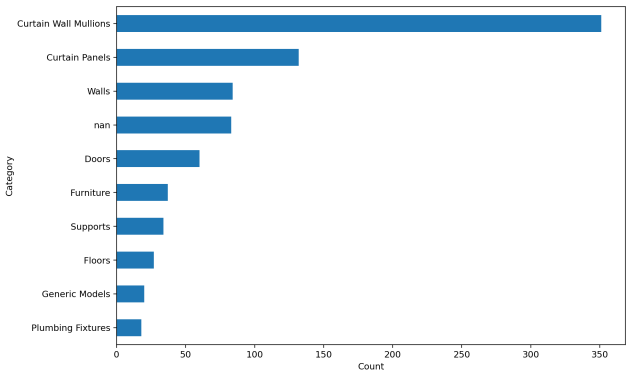

Figure 2. Distribution of BIM categories in the final modelling sample.

5.2 Comparative Model Performance

Table (2) and Figure (3) report the cross-validated model results. Logistic regression achieved reasonable but clearly lower performance than the two tree ensembles, with macro F1 dropping to 0.754. Random Forest produced the strongest overall balance of accuracy and class-wise stability, marginally outperforming Extra Trees.

Table 2. Five-fold cross-validated performance.

Model	Acc.	Bal. Acc.	W-F1	M-F1
Logistic Reg.	0.803	0.820	0.836	0.754
Random Forest	0.992	0.971	0.992	0.970
Extra Trees	0.990	0.969	0.990	0.967

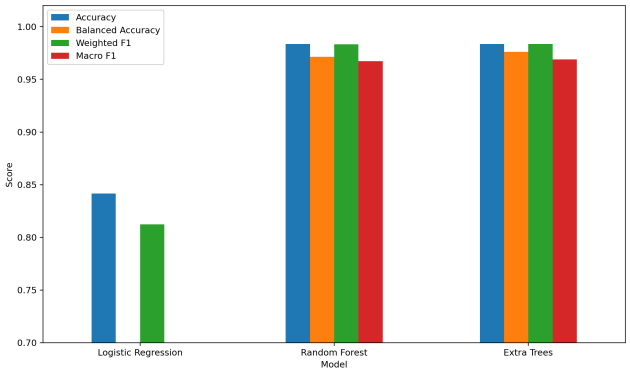

Figure 3. Cross-validated performance comparison of the three benchmark models.

The random forest result is especially notable because the model used only engineered numeric BIM descriptors, without mesh rendering, point-cloud conversion, image views, or textual embeddings. This indicates that compact engineering features already carry substantial semantic signal.

5.3 Class-Wise Performance

Table (3) shows the class-wise performance of the best-performing random forest model. Perfect F1 was achieved for Curtain Wall Mullions, Curtain Panels, Doors, Generic Models, and Plumbing Fixtures. The lowest class-wise F1 was observed for Railings (0.818), followed by Floors (0.945) and Supports (0.955). These minority classes also account for most of the observed misclassifications.

Table 3. Class-wise results for the random forest model.

Category	Precision	Recall	F1	Support
Curtain Wall Mullions	1.000	1.000	1.000	351
Curtain Panels	1.000	1.000	1.000	132
Walls	1.000	0.988	0.994	84
Doors	1.000	1.000	1.000	60
Furniture	0.977	1.000	0.989	43
Supports	0.970	0.941	0.955	34
Floors	0.929	0.963	0.945	27
Generic Models	1.000	1.000	1.000	20
Plumbing Fixtures	1.000	1.000	1.000	18
Railings	0.818	0.818	0.818	11

The confusion matrix in Figure (4) shows that errors are concentrated in a small number of semantically adjacent cases, particularly among Supports, Railings, Floors, and Walls. This pattern is consistent with engineering intuition, because these categories can share similar extents, elevations, and local shape proportions in exported BIM records.

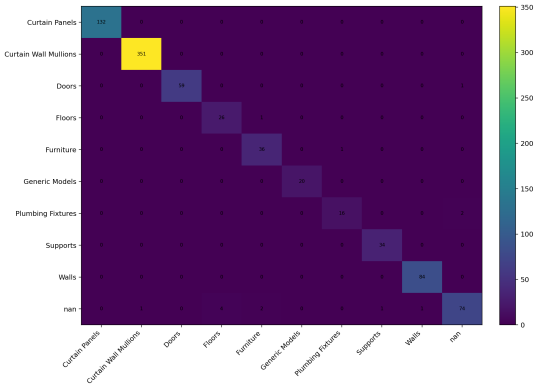

Figure 4. Confusion matrix for the random forest model.

5.4 Feature Importance and Geometric Profiles

Figure (5) ranks the most influential predictors in the random forest model. Bounding-box height was the strongest single predictor, followed by bounding-box width, bounding-box length, global Y coordinate, externality, and dimension fields such as `XDim` and `YDim`. The prominence of bounding-box features suggests that object envelope geometry provides a stable semantic signature across classes.

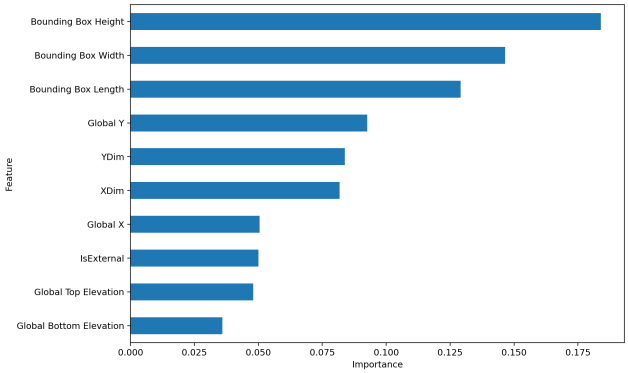

Figure 5. Top features in the random forest model.

The four-panel distributional analysis in Figure (6) further clarifies these differences. Curtain wall elements occupy distinctive bands in bounding-box height and width, while floors have a characteristic low-height profile and broader horizontal footprint. Supports and railings partially overlap, which helps explain the limited set of remaining confusions.

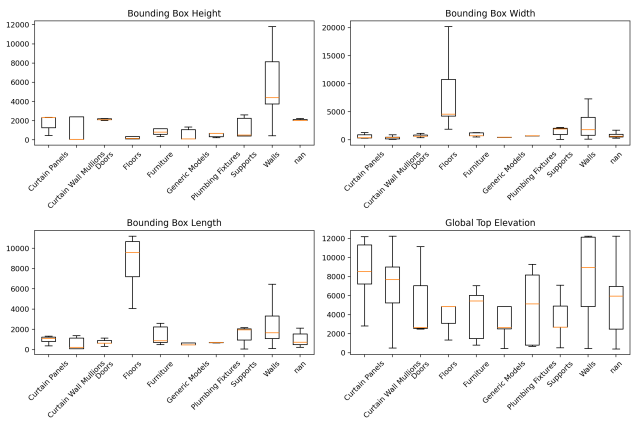

Figure 6. Distributional profiles of four key engineering features across BIM categories.

6. DISCUSSION

The study supports the key claim of this paper: semantic information of objects in BIM can be inferred from simple engineering descriptors without the aid of expensive mesh-based deep learning. This is important for practical BIM. In practice, many engineering practitioners are presented with IFC exports, tabular object schedules, or viewer-based reports showing object properties, instead of highly detailed 3D models. The workflow proposed here can work with such a data product.

The best predictors are physically meaningful, from an engineering science perspective. Bounding-box size and elevation variables represent spatial extent, while flags like `IsExternal` represent design intent and exposure to the outside environment. Thus, the model's success is not through latent embeddings, but due to correlations between BIM categories and spatial and engineering features. This boosts interpretability and trust in professional practice.

The results also have methodological implications. The logistic regression performed significantly worse than tree ensembles, indicating that boundaries between classes in the BIM feature space are non-linear. Random forest and extra trees, on the other hand, successfully modeled these interactions but were easier to implement than deep learning. For companies seeking a quick baseline for semantic recovery prior to more complex pipelines, ensemble methods like these offer a viable option.

Our results join the semantic-enrichment literature while carving out a new application space. Existing research has focussed on graph learning, sparse convolutions, multi-modal fusion, synthetic data generation, and ontology-based semantic enrichment [12, 1, 11, 5, 2]. This study demonstrates that even before such processing, semantic repair can be obtained by reusing object-level engineering attributes that are already available in standard IFC exports. This approach is therefore interesting for quality control of BIMs, pre-classification for simulation, repair of interoperability issues or data cleaning.

The study has limitations. First, the data set is a public sample rather than a large benchmark, so the results can be considered a reproducible baseline but not necessarily a best case. Second, the classes are skewed and minority classes have low counts. Third, the experiment is aimed at category recovery rather than subtype identification. The pipeline can be extended to multiple-source IFC data, including topological relations, cross-project testing, and benchmark other approaches, such as the current feature-driven one, graph-based, and multi-modal approaches under the same validation regime.

7. CONCLUSION

This study reported a replicable engineering science experiment on BIM element classification via open IFC properties. Through a structured feature vector conversion of semi-structured BIM exports, and a benchmarking of three supervised classifiers, the study demonstrated high semantic repair with small engineering descriptors. An accuracy of 0.992 and macro F1 of 0.970 was achieved from the random forest model across ten BIM element categories, where the strongest features are related to bounding-box sizes, elevations, and

externality indicators.

The takeaway message is clear: not all BIM semantic-repair problems need to rely on mesh-based deep learning and complex geometry preprocessing. A simple, geometry-based baseline can deliver high accuracy, interpretability and code simplicity. As such, the proposed method can be used for interoperability-driven BIM preparation, engineering analysis, and quick model auditing in practice.

REFERENCES

[1] C. Emunds, N. Pauen, V. Richter, J. Frisch, and C. van Treeck, "SpaRSE-BIM: Classification of IFC-based geometry via sparse convolutional neural networks," *Advanced Engineering Informatics*, vol. 53, Article 101641, 2022. https://doi.org/10.1016/j.aei.2022.101641

[2] D. Hu, V. J. L. Gan, R. Zhai, Y. Wang, and Y. He, "Automated BIM-to-scan point cloud semantic segmentation using a domain adaptation network with hybrid attention and whitening (DawNet)," *Automation in Construction*, vol. 164, Article 105473, 2024. https://doi.org/10.1016/j.autcon.2024.105473

[3] S. Jiang, X. Feng, B. Zhang, and J. Shi, "Semantic enrichment for BIM: Enabling technologies and applications," *Advanced Engineering Informatics*, vol. 56, Article 101961, 2023. https://doi.org/10.1016/j.aei.2023.101961

[4] B. Koo, R. Jung, and Y. Yu, "Automatic classification of wall and door BIM element subtypes using 3D geometric deep neural networks," *Advanced Engineering Informatics*, vol. 47, Article 101200, 2021. https://doi.org/10.1016/j.aei.2020.101200

[5] H. Liu, V. J. L. Gan, J. C. P. Cheng, and S. A. Zhou, "Automatic fine-grained BIM element classification using multi-modal deep learning (MMDL)," *Advanced Engineering Informatics*, vol. 61, Article 102458, 2024. https://doi.org/10.1016/j.aei.2024.102458

[6] C. Mirarchi, M. Gholamzadehmir, B. Daniotti, and A. Pavan, "Semantic enrichment of BIM: The role of machine learning-based image recognition," *Buildings*, vol. 14, no. 4, Article 1122, 2024. https://doi.org/10.3390/buildings14041122

[7] S. Parece, R. Resende, and V. Rato, "A BIM-based tool for embodied carbon assessment using a construction classification system," *Developments in the Built Environment*, vol. 19, Article 100467, 2024. https://doi.org/10.1016/j.dibe.2024.100467

[8] Y. Peng, C. P. Au-Yong, and N. E. Myeda, "Knowledge graph of building information modelling (BIM) for facilities management (FM)," *Automation in Construction*, vol. 165, Article 105492, 2024. https://doi.org/10.1016/j.autcon.2024.105492

[9] K. Rogage and O. Doukari, "3D object recognition using deep learning for automatically generating semantic BIM data," *Automation in Construction*, vol. 162, Article 105366, 2024. https://doi.org/10.1016/j.autcon.2024.105366

[10] R. Sacks, Z. Wang, B. Ouyang, D. Utkucu, and S. Chen, "Toward artificially intelligent cloud-based building information modeling for collaborative multidisciplinary design," *Advanced Engineering Informatics*, vol. 53, Article 101711, 2022. https://doi.org/10.1016/j.aei.2022.101711

[11] D. Utkucu, H. Ying, Z. Wang, and R. Sacks, "Classification of architectural and MEP BIM objects for building performance evaluation," *Advanced Engineering Informatics*, vol. 61, Article 102503, 2024. https://doi.org/10.1016/j.aei.2024.102503

[12] Z. Wang, R. Sacks, and T. Yeung, "Exploring graph neural networks for semantic enrichment: Room type classification," *Automation in Construction*, vol. 134, Article 104039, 2022. https://doi.org/10.1016/j.autcon.2021.104039

[13] F. Xue, L. Wu, and W. Lu, "Semantic enrichment of building and city information models: A ten-year review," *Advanced Engineering Informatics*, vol. 47, Article 101245, 2021. https://doi.org/10.1016/j.aei.2020.101245

[14] R. Zhai, J. Zou, Y. He, and L. Meng, "BIM-driven data augmentation method for semantic segmentation in superpoint-based deep learning network," *Automation in Construction*, vol. 140, Article 104373, 2022. https://doi.org/10.1016/j.autcon.2022.104373