

Intelligent Analysis of Building Information Models Based on Machine Learning Promotes Sustainable Building Design

Qiyue Shi

*College of Human Ecology, Yonsei University, Seoul, Korea
17yue80s@gmail.com*

Abstract. As the global construction industry moves toward sustainable development, the use of Building Information Modeling (BIM) to promote green building design is gaining attention. However, existing research primarily focuses on single-dimensional performance prediction, lacking comprehensive optimization across multiple dimensions and the entire lifecycle. This study aims to construct a machine learning-based BIM intelligent analysis framework that enhances multi-dimensional feature extraction and fusion to support sustainability in architectural design. Based on a case study of the teaching building at Suzhou Science and Technology City Experimental Middle School, results show that the framework significantly improves sustainability indicators such as energy efficiency and environmental comfort. This study offers a new technological path for intelligent sustainable building design and holds important academic value.

Keywords: Machine learning, BIM, Sustainable architectural design, Intelligent analysis framework

1. Introduction

Driven by the deepening of global climate governance, the construction industry, as a major energy consumer [1], faces growing transformation pressure [2]. According to the United Nations Environment Programme, carbon emissions throughout the building life cycle remain high, making sustainable transformation essential to achieving the Paris Agreement's temperature targets [3]. In this context, building a sustainable system with resource-efficient circulation, ecosystem synergy, and climate resilience has become a necessary response to global challenges [4].

However, sustainable design practices still face key bottlenecks. First, design decisions rely heavily on empirical paradigms, limiting compliance with refined sustainability indicators [5]. Second, performance analysis often lags behind design, missing the window for early-stage optimization [6]. Third, multi-objective balancing lacks systematic methods, often resulting in local rather than global optima across economic, environmental, and social goals [7].

The evolution of Building Information Modeling (BIM) provides a technological foundation to address these challenges. With parametric modeling at its core, BIM integrates geometry, physical properties, performance data, and lifecycle management, enabling data-driven sustainable design [8,9]. Yet, BIM still lacks deep analysis and proactive decision-making. Its use often remains at visualization and basic statistics, failing to uncover deeper patterns. Additionally, BIM lacks self-

learning capabilities to solve complex, multi-constraint optimization problems or offer forward-looking recommendations.

This study introduces machine learning to enhance BIM intelligence in design. By optimizing multidimensional sustainability indicators, it supports accurate prediction and performance improvement, providing a new path for green buildings and advancing low-carbon, efficient development.

2. Theoretical foundations and technical integration

2.1. Structured extraction of sustainability information

BIM, as the core carrier of digital building assets, has expanded its value from geometric expression to full lifecycle data integration. This study focuses on the structured extraction of semantic information strongly related to sustainability performance in BIM models.

In the topological dimension of geometric space, the IFC standard mode is used to automatically extract building volume, window to wall ratio (WWR), spatial zoning, and geometric attributes of shading components, providing accurate geometric basis for spatial energy efficiency evaluation and sunshine simulation in subsequent sustainability analysis [10]; At the level of materials and construction library, deeply analyze the material data at the component level and associate it with the physical attribute database in Appendix A Table A.1, covering key parameters such as thermal conductivity (U-Value), solar heat gain coefficient (SHGC), visible light transmittance (VLT), and internal energy (EE) and global warming potential (GWP) in the environmental database, laying a solid data foundation for the quantitative evaluation of building sustainability performance; In terms of automated extraction technology, an efficient parsing pipeline is built based on Revit API and IfcOpenShell library to achieve precise mapping from discrete parameters to feature vectors. The expression is as follows:

$$F_{BIM} = \Phi(\text{MIFC}), \quad \Phi : \text{IFC entity} \rightarrow \mathbb{R}^d \quad (1)$$

Among them, F_{BIM} is a d -dimensional feature vector, and Φ is an extraction function based on pattern rules.

2.2. Algorithm fusion framework for sustainability prediction

Machine learning models high-dimensional nonlinear systems through data-driven methods, overcoming the computational bottlenecks of traditional physical simulations such as EnergyPlus [11]. This study develops a hybrid algorithm framework to predict sustainability indicator demand.

The task adaptation model selection adopts LightGBM [12] under the gradient boosting decision tree framework for regression tasks. Its histogram algorithm accelerates feature splitting, supports native processing of category features, and the Leaf wise growth strategy improves accuracy with the assistance of multi-layer perceptrons to capture deep nonlinearity [13]. The expression is as follows:

$$\hat{y} = \sigma(W_k \cdot \text{ReLU}(W_{k-1} \cdots \text{ReLU}(W_1 F_{BIM} + b_1) \cdots) + b_k) \quad (2)$$

Where σ is the activation function of the output layer.

The classification task uses a support vector machine with radial basis kernel for nonlinear decision boundaries.

In terms of SHAP value analysis, this article draws on existing practices [14] and uses the following expression to quantify feature contribution:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f(S \cup \{i\}) - f(S)] \quad (3)$$

Among them, M is the feature set, f is the model, output ϕ_i is the Shapley value of the i -th feature.

2.3. Quantitative benchmark

The core indicators of this study are shown in Table A.2, and their quantification method constitutes the supervisory signal of the ML model.

2.4. Technology integration framework

The technology integration framework of this article can be seen in Figure B.1. This article is based on the BIM based intelligent sustainability assessment system to build a technology integration framework. The core process first uses a feature extraction engine to analyze the three major feature dimensions of geometric topology, material lifecycle assessment, and environmental correlation from the original BIM model, and integrates them to form a comprehensive feature vector; Subsequently, a multi task machine learning model was used to predict building performance; Finally, the design constraints are verified through the sustainable indicator integration module. If the requirements are met, a certification report is generated. If the requirements are not met, sensitivity analysis is triggered to generate optimization suggestions, forming a closed-loop feedback mechanism.

3. Research design

3.1. Machine learning training strategy

The training strategy in this article adopts hierarchical data partitioning combined with Bayesian optimization, as shown in Figure 1. The process first divides the original dataset into training, validation, and test sets in a 70%–15%–15% ratio. Then, a 5-fold cross-validation is performed on the training set, combined with Bayesian optimization for hyperparameter search. The optimized hyperparameters are used to train the final model, with the validation set integrated for monitoring. Model performance is finally evaluated on a completely independent test set.

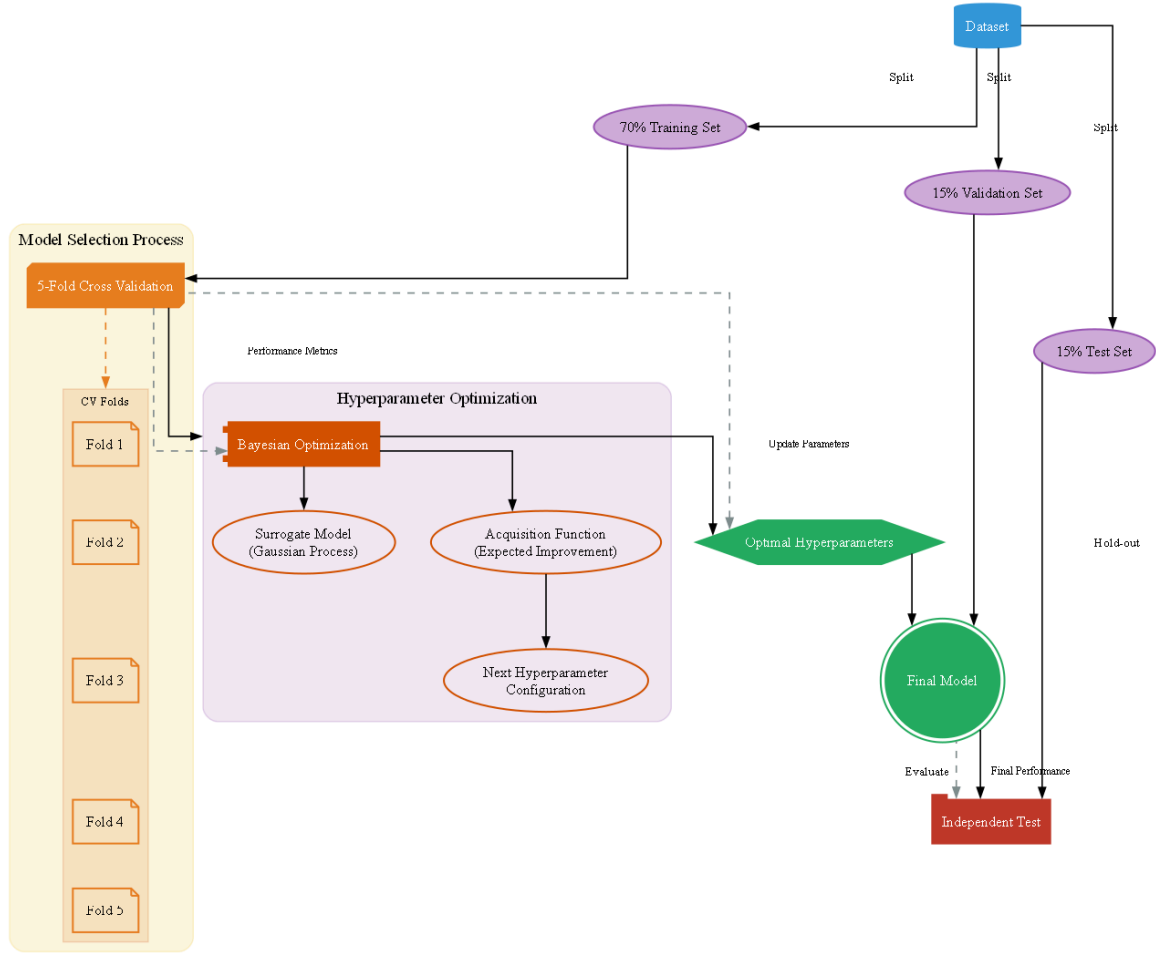


Figure 1. Schematic diagram of machine learning training strategy

The Bayesian optimization objective function is expressed as follows:

$$\operatorname{argmin}_{\theta} \frac{1}{K} \sum_{k=1}^K \operatorname{RMSE}_k(\theta) + 0.2 \cdot \operatorname{TrainingTime}(\theta) \quad (4)$$

3.2. Model deployment

This article chooses to deploy lightweight models using the Revit Dynamo integrated architecture, as shown in Figure B.2.

4. Performance evaluation

4.1. Background of the case

This study selected the teaching building of Suzhou Science and Technology City Experimental Middle School as a verification case. The project is a public education building that meets the Class A standard of GB 50189-2015, with a building area of 18640 square meters; Located in China's Thermal Power Zone III, with hot summers and cold winters, the HDD_{18} is 1896 and CDD_{26} is 282,

and it includes typical educational building spatial features such as classrooms, laboratories, and auditoriums.

4.2. Data collection

The data sources and corresponding processing methods for this article are shown in Table 1.

Table 1. Data collection

Indicator	Data Source	Processing Method
GWP	Bill of building materials × CECS 374:2020 emission factors	Calculated by structural/envelope/finishing divisions
EUI	Annual data from smart meters (2023.06–2024.05)	Outlier removal for holidays using 3σ criterion
sDA	Radiance simulation (CIE overcast sky model)	0.6 m grid division per IES LM-83 standard
PMV	Indoor sensor network (temperature/humidity/black globe temperature)	Calibrated per ISO 7726, calculated via ISO 7730 formula
WUI	Water utility metering data + cooling tower makeup water records	Zoned metering data fusion

Afterwards, stratified sampling was conducted according to the functional zoning of the building, and the dataset structure is shown in Table 2.

Table 2. Dataset structure

Data Type	Training Set	Validation Set	Test Set	Total
Spatial Units (Sample Count)	127	32	29	188
Feature Dimensions	327	327	327	327

4.3. Performance results

The comparison of indicator prediction accuracy is shown in Table 3. The prediction errors of the five major sustainability indicators are all controlled within the allowable error range of the industry, and the accurate and stable prediction of key continuous indicators confirms that the model effectively captures the complex nonlinear relationship between building physical characteristics and carbon emissions.

Table 3. Comparison of indicator prediction accuracy

Indicator	ML Prediction	Benchmark Value	Absolute Error	Allowable Tolerance
GWP	142 kgCO ₂ e/m ²	151 kgCO ₂ e/m ²	-5.9%	±10%
EUI	152 kWh/m ² /yr	146 kWh/m ² /yr	+4.1%	±8%
sDA	62% compliant	58% compliant	+6.9%	—
PMV	-0.32	-0.41	+0.09	±0.15
WUI	0.48 m ³ /m ² /yr	0.52 m ³ /m ² /yr	-7.7%	±10%

In terms of design optimization iteration analysis, the design strategy evolution was carried out in combination with the model. This article went through a total of 8 rounds of iteration process, and

the specific results are shown in Table A.3. Intuitively, as shown in Figure 2, significant improvements have been achieved in various indicators with the help of the model's decision support.

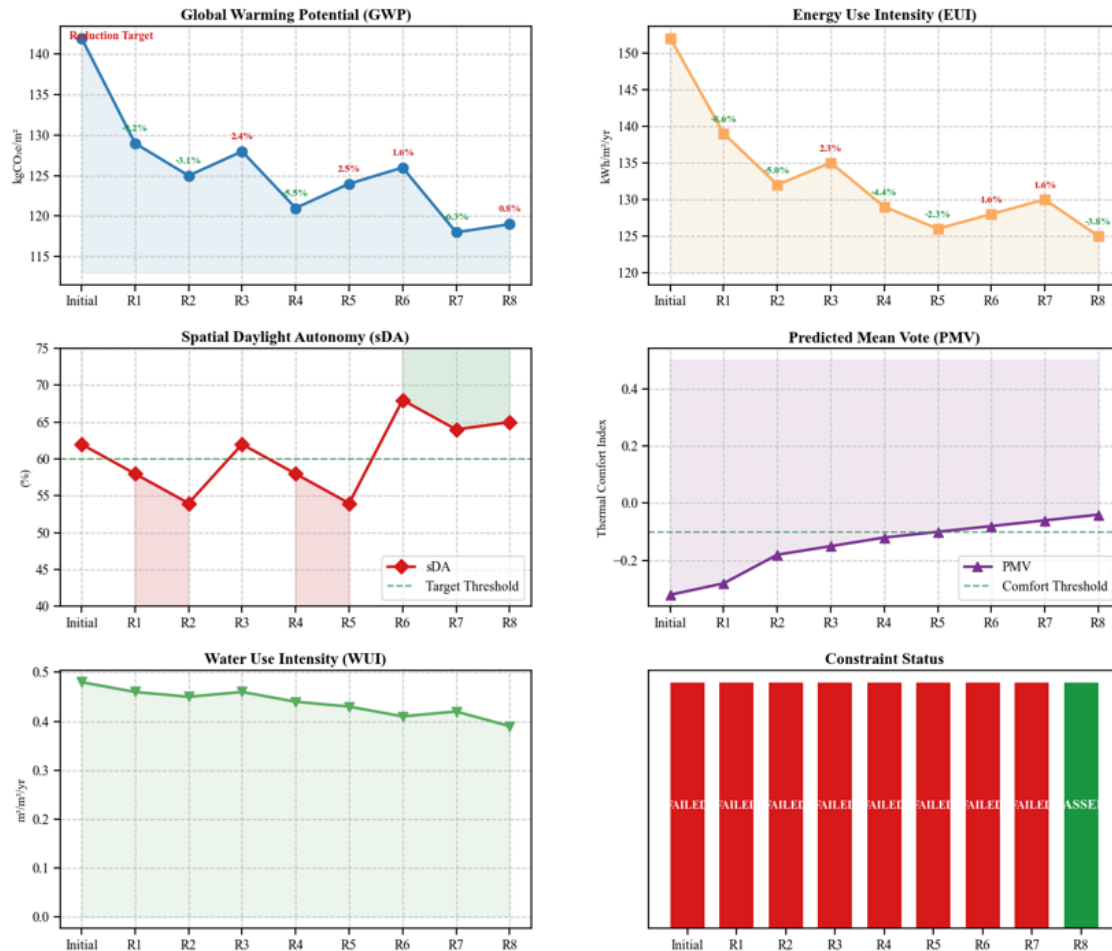


Figure 2. Design optimization iteration analysis

5. Conclusion

The machine learning based BIM intelligent analysis method proposed in this study effectively improves the prediction accuracy of sustainability indicators in architectural design. By constructing a multidimensional feature extraction framework and combining it with machine learning algorithms, intelligent decision support for performance optimization in the architectural design process has been successfully achieved. The experimental results show that the application of this method in the teaching building of Suzhou Science and Technology City Experimental Middle School significantly improves multiple sustainability indicators such as building energy efficiency and environmental comfort, achieving the expected optimization goals. This study provides a new technological path for the sustainable development of the construction industry, promotes the popularization of intelligent and data-driven decision-making in architectural design, and has important reference value for technological innovation and application practices in related fields.

References

- [1] El-Hakim, Y., & AbouZeid, M. N. (2024). Towards mitigating climate change negative impact: The role of regulations and governance in the construction industry. *Sustainability*, 16(16), 6822.<https://doi.org/10.3390/su16166822>
- [2] Pfnür, A., & Wagner, B. (2020). Transformation of the real estate and construction industry: empirical findings from Germany. *Journal of Business Economics*, 90(7), 975-1019.<https://doi.org/10.1007/s11573-020-00972-4>
- [3] Huang, M. T., & Zhai, P. M. (2021). Achieving Paris Agreement temperature goals requires carbon neutrality by middle century with far-reaching transitions in the whole society. *Advances in Climate Change Research*, 12(2), 281-6.<https://doi.org/10.1016/j.accre.2021.03.004>
- [4] Akadiri, P. O., Chinyio, E. A., & Olomolaiye, P. O. (2012). Design of a sustainable building: A conceptual framework for implementing sustainability in the building sector. *Buildings*, 2(2), 126-52.<https://doi.org/10.3390/buildings2020126>
- [5] Pan, W., & Ning, Y. (2014). Dialectics of sustainable building: Evidence from empirical studies 1987–2013. *Building and environment*, 82, 666-74.<https://doi.org/10.1016/j.buildenv.2014.10.008>
- [6] Oduyemi, O., & Okoroh, M. (2016). Building performance modelling for sustainable building design. *International Journal of Sustainable Built Environment*, 5(2), 461-9.<https://doi.org/10.1016/j.ijbsbe.2016.05.004>
- [7] Brunelli, C., Castellani, F., Garinei, A., Biondi, L., & Marconi, M. (2016). A procedure to perform multi-objective optimization for sustainable design of buildings. *Energies*, 9(11), 915.<https://doi.org/10.3390/en9110915>
- [8] Wong, K. D., & Fan, Q. (2013). Building information modelling (BIM) for sustainable building design. *Facilities*, 31(3/4), 138-57.<https://doi.org/10.1108/02632771311299412>
- [9] Fadeyi, M. O. (2017). The role of building information modeling (BIM) in delivering the sustainable building value. *International Journal of Sustainable Built Environment*, 6(2), 711-22.<https://doi.org/10.1016/j.ijbsbe.2017.08.003>
- [10] Liu, H., Zhang, Z., Ma, X., Lu, W., Li, D., & Kojima, S. (2021). Optimization Analysis of the Residential Window-to-Wall Ratio Based on Numerical Calculation of Energy Consumption in the Hot-Summer and Cold-Winter Zone of China. *Sustainability* (2071-1050), 13(11).<https://doi.org/10.3390/su13116138>
- [11] Wang, H., Lu, N., Wu, F., & Zhai, J. (2023). Coupling Computational Fluid Dynamics and EnergyPlus to Optimize Energy Consumption and Comfort in Air Column Ventilation at a Tall High-Speed Rail Station. *Sustainability*, 15(17), 12948.<https://doi.org/10.3390/su151712948>
- [12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.<https://doi.org/10.5555/3294996.3295074>
- [13] Negi, G. C. S., & Singh, S. P. (1992). Leaf growth pattern in evergreen and deciduous species of the Central Himalaya, India. *International journal of biometeorology*, 36(4), 233-242.<https://doi.org/10.1007/BF02726404>
- [14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.<https://doi.org/10.5555/3295222.3295230>

Appendix

Appendix A

Table A.1 Core sustainability characteristics extracted by BIM

Feature Category	Parameter Examples	Data Source
Geometric Topology	Window-Wall Ratio (WWR)	IFC Entity Parsing
	Building Azimuth Angle (Azimuth)	IfcBuilding Global Coordinate System Conversion
	Shape Factor (S/V Ratio)	IfcBuilding Bounding Box Volume/Surface Area Calculation
	Floor Height	IfcBuildingStorey Geometric Properties
Envelope Structure	Wall Heat Transfer Coefficient (U-Value)	Physical Property Database Associated with Material Library
	Glass Solar Heat Gain Coefficient (SHGC)	Optical Properties of IfcWindow Material Layers
	Glass Visible Light Transmittance (VLT)	Optical Properties of IfcWindow Material Layers
	Roof Thermal Inertia Index	Calculation of Product of Material Density and Specific Heat Capacity
Environment-Related	Geographic Location (Longitude/Latitude)	IfcSite Entity Properties
	Climate Zone	Meteorological Database Mapping
	Shading Coefficient (SC) of Shading Devices	IfcShadingDevice Geometric Projection Analysis
Resource and Environmental Load	Material Global Warming Potential (GWP)	Ecoinvent v3.8 LCA Database
	Material Embodied Energy (EE)	ICE v3.0 Database
Spatial Organization	Functional Zone Connectivity (Space Adjacency)	IfcSpace Topological Relationship Analysis
	Room Depth/Width Ratio	IfcSpace Geometric Boundary Calculation

Table A.2 Definition of sustainability indicators

Indicator	Definition	Benchmark Method	ML Task Type
Embodied Carbon (GWP)	Total carbon dioxide equivalent emissions from building materials throughout their life cycle	Ecoinvent v3.8 LCA Database	Regression
Operational Energy Use Intensity (EUI)	Annual total energy consumption per unit area of a building	EnergyPlus v23.1 Dynamic Simulation	Regression
Daylight Compliance Rate (sDA)	Percentage of grid points in a building that meet 300 lux illuminance and exceed 50% of annual occupied hours	Radiance/Daysim Radiance Simulation	Binary Classification
Thermal Comfort (PMV)	Predicted Mean Vote value for estimating population satisfaction with the thermal environment	ISO 7730:2005 Steady-State Model Calculation	Regression
Water Use Efficiency (WUI)	Annual total freshwater consumption per unit area of a building	EPA WaterSense Benchmark Database	Regression

The expression for GWP is as follows:

$$GWP = \sum_{i=1}^n (m_i \times CF_i) \quad (A.1)$$

M_i is mass of material i (kg); CF_i is GWP characterization factor of material i (kgCO₂e/kg); n is number of material types.

The expression for EUI is as follows:

$$EUI = \frac{\sum_{j=1}^4 (a_j \times Q_j^{EP})}{\text{Building}} \quad (A.2)$$

Q_j^{EP} is annual energy demand output by EnergyPlus simulation (kWh); $j=1$ means Heating; $j=2$ means Cooling; $j=3$ means Ventilation; $j=4$ means Lighting + Equipment; a_j is primary energy conversion factors.

The expression for sDA is as follows:

$$sDA = \frac{|\{p \in \mathbb{P} | DAP_p \geq 50\% \}|}{|\mathbb{P}|} \times 100\% \quad (A.3)$$

P is sensor point grid; DAP_p is percentage of time point exceeds 300 lux.

The expression for PMV is as follows:

$$PMV = f(T_a, T_r, v, RH, M, I_{cl}) \quad (A.4)$$

T_a is air temperature (°C); T_r is mean radiant temperature (°C); v is relative air velocity (m/s); RH is relative humidity (%); M is metabolic rate (W/m²); I_{cl} is clothing insulation (clo).

The expression for WUI is as follows:

$$WUI = \frac{\sum_{k=1}^3 (\beta_k \times V_k)}{\text{Building}} \quad (\text{m}^3/\text{m}^2/\text{yr}) \quad (A.5)$$

V_k is annual water volume (m³); β_k is regional water stress weighting factor.

Table A.3 Design optimization iteration analysis table

Iteration Round	GWP (kgCO ₂ e/m ²)	EUI (kWh/m ² /yr)	sDA (%)	PMV	WUI (m ³ /m ² /yr)	Constraint Status
Initial Scheme	142	152	62	-0.32	0.48	✗ (4/5)
Round 1	129 (-9.2%)	139 (-8.6%)	58 (-6.5%)	-0.28↑	0.46 (-4.2%)	✗ (Insufficient sDA)
Round 2	125 (-3.1%)	132 (-5.0%)	54 (-6.9%)	-0.18↑	0.45 (-2.2%)	✗ (Insufficient sDA)
Round 3	128 (+2.4%)	135 (+2.3%)	62 (+14.8%)	-0.15↑	0.46 (+2.2%)	✗ (GWP Rebound)
Round 4	121 (-5.5%)	129 (-4.4%)	58 (-6.5%)	-0.12↑	0.44 (-4.3%)	✗ (Insufficient sDA)
Round 5	124 (+2.5%)	126 (-2.3%)	54 (-6.9%)	-0.10↑	0.43 (-2.3%)	✗ (Insufficient sDA)
Round 6	126 (+1.6%)	128 (+1.6%)	68↑ (+25.9%)	-0.08↑	0.41↓ (-4.7%)	✗ (GWP)
Round 7	118↓ (-6.3%)	130 (+1.6%)	64 (-5.9%)	-0.06↑	0.42 (+2.4%)	✗ (WUI)
Round 8	119↑ (+0.8%)	125↓ (-3.8%)	65↑ (+1.6%)	-0.04↑	0.39↓ (-7.1%)	✓ (5/5)

Appendix B

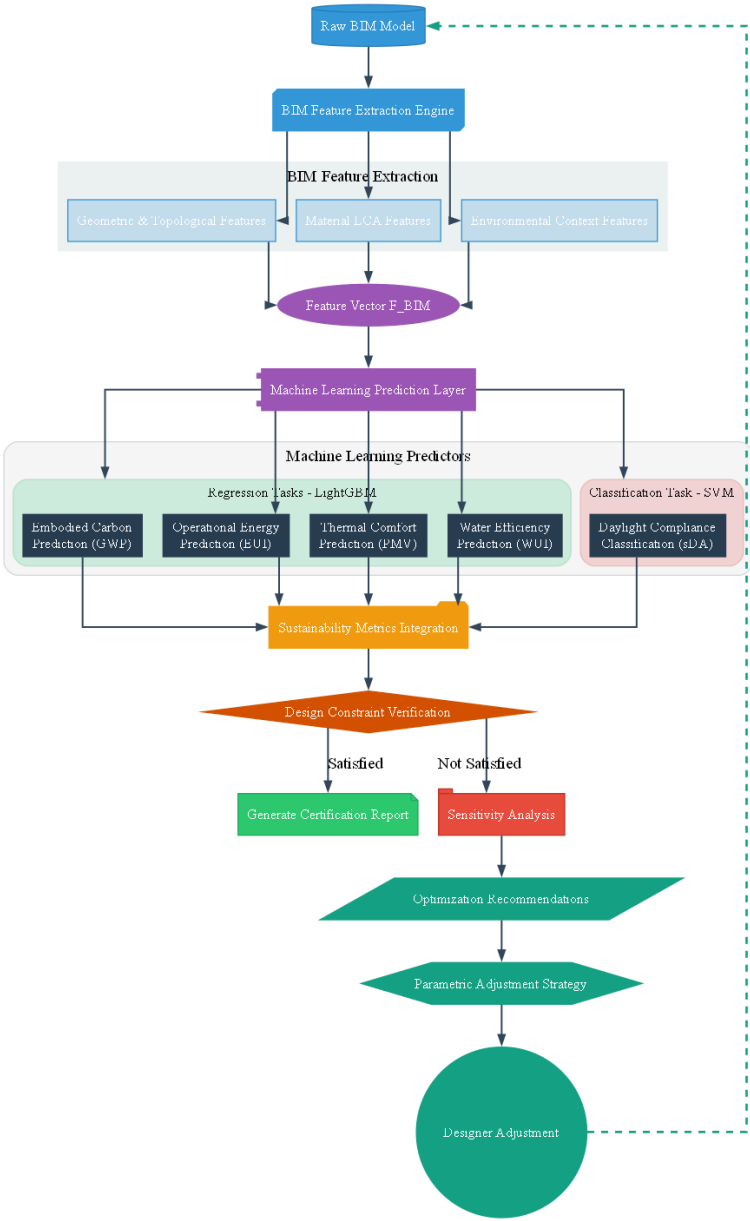


Figure B.1 Schematic diagram of technology integration framework

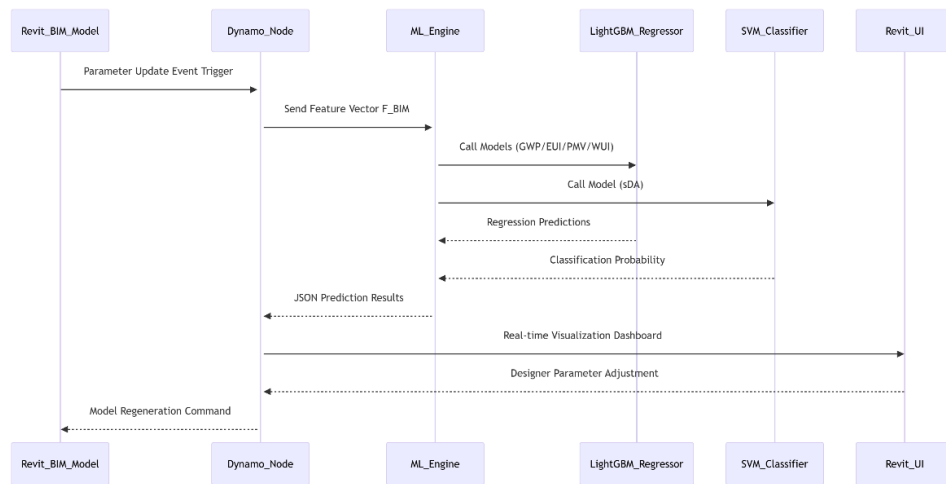


Figure B.2 Schematic diagram of Revit Dynamo integration architecture