

Journal Pre-proof

A BIM and Machine Learning Integration Framework for Automated Property Valuation

Tengxiang Su, Haijiang Li, Yi An



PII: S2352-7102(21)00494-0

DOI: <https://doi.org/10.1016/j.jobe.2021.102636>

Reference: JOBE 102636

To appear in: *Journal of Building Engineering*

Received Date: 20 January 2021

Revised Date: 23 April 2021

Accepted Date: 26 April 2021

Please cite this article as: T. Su, H. Li, Y. An, A BIM and Machine Learning Integration Framework for Automated Property Valuation, *Journal of Building Engineering*, <https://doi.org/10.1016/j.jobe.2021.102636>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Crown Copyright © 2021 Published by Elsevier Ltd.

A BIM and Machine Learning Integration Framework for Automated Property Valuation

Tengxiang Su, Haijiang Li *, Yi An

BIM for Smart Engineering Centre, Cardiff School of Engineering, Cardiff University, Queen's Buildings,
Cardiff CF24 3AA, Wales, Cardiff CF24 3AA, Wales, UK

*Correspondence author: LiH@cardiff.ac.uk (H.L.); First author: SuT2@cardiff.ac.uk (T.S.)

Abstract

Property valuation contributes significantly to market economic activities, while it has been continuously questioned on its low transparency, inaccuracy and inefficiency. With Big Data applications in real estate domain growing fast, computer-aided valuation systems such as AI-enhanced automated valuation models (AVMs) have the potential to address these issues. While a plethora of research has focused on improving predictive performance of AVMs, little effort has been made on information requirements for valuation models. As the amount of data in BIM is rising exponentially, the value-relevant design information has not been widely utilized for property valuation. This paper presents a system that leverages a holistic data interpretation, improves information exchange between AEC projects and property valuation, and automates specific workflows for property valuation. A mixed research method was adopted combining the archival literature research, qualitative and quantitative data analysis. A BIM and Machine learning (ML) integration framework for automated property valuation was proposed which contains a fundamental database interpretation, an IFC-based information extraction and an automated valuation model based on genetic algorithm optimized machine learning (GA-GBR). The main findings indicated: (1) Partial information requirements can be extracted from BIM models, (2) Property valuation can be performed in a more accurate and efficient way. This research contributes to managing information exchange between AEC projects and property valuation and supporting automated property valuation. It was suggested that the infusion of BIM, ML and other emerging digital technologies might add values to property valuation and the construction industry.

Keywords: Property valuation, Information exchange, Building information modelling (BIM), Industry foundation class (IFC), Machine learning (ML).

1. Introduction

Property valuation, also known as real estate appraisal, plays a fundamental role in a nation's economy and financial stability. According to Taffese [1], financial and economic decisions are based on the accuracy of property valuation results. For example, the housing market bubbles can cause serious financial risks such as the subprime mortgage crisis in the Great Recession 2008. Various stakeholders ask for property valuation for several objectives: banks and insurance company use it for mortgage release, traders use it for house transactions, property developers use it for house investment, local authorities use it for house taxation. There are three main conventional methods for property valuation practice, namely the sales comparison method, the income approach and the cost approach [2]. According to Su and Li

[3], property valuation is affected by various parameters which are related to environmental quality (pollution, land use and sustainable resource), social and economic quality (vacancy rate, rental growth potential), technical and functional quality (structure, age, size, construction materials, indoor air quality, flexibility and adaptability), process quality (quality control during construction) and site quality (transport access, amenities). The dynamics of some parameters and the low transparency of real estate market make it hard for property valuation professionals to perform an accurate and objective valuation of property price.

In the last two decades, due to the exponentially increasing amount of data in various domains, artificial intelligence has been widely applied in chatbots, healthcare, finance and economics (prediction and risk management), human resource management, logistics and supply chain etc. Similarly, there are a large amount of research on artificial intelligence in property valuation, which is detailed in literature review (Section 2.1). For property valuation which influenced by many objective and subjective factors, artificial intelligence models have several advantages: to efficiently assess information from big data; to identify non-linear relationships between house characters, market factors and property price; to be more objective about the selection of input attributes [4–6]. Researchers have tried different mathematic models for property valuation such as hedonic regression, ridge regression, support vector regression, ensemble learning, neural networks [7–11]. Many research experiments have concluded that artificial intelligence models can be an improved alternative for house price prediction. For example, Graczyk et al. [10] conducted an experiment that testing different ensemble models for property valuation and concluded that all ensemble classifiers with additive regression produced significant error reduction compared to original models. Liu et al. [11] concluded that BP neural network models had good performances on house price estimation. Fan et al. [12] found that tree-based approaches played an important role in statistical pattern recognition of the relationship between the input attributes and property price. However, research on the fundamental database and information requirements for property valuation is limited. As robust and reliable data is a prerequisite for training automated valuation models (GIGO-Garbage in, Garbage out), the value-relevant design information has not been widely used for property valuation, due to the gap of knowledge and digital skills between property valuation and AEC professionals.

For property valuation professionals, it is essential for them to access to and use lifecycle building performance information (commissioning, design, project execution, operations and maintenance, and recommissioning) from reliable data sources. For example, a plethora of research has shown positive relationships between buildings' sustainable variables and observed property market values. After the analysis of over 1200 green-rated buildings, CoStar Group concluded that the average LEED impact and Energy Star impact on sales price per square foot showed a positive 9.94% and 5.76% respectively [13]. Similar studies in UK and Switzerland concluded that green buildings have a premium market price [14,15]. While researchers and practitioners are trying to perform sustainable property valuation, the sustainability-related data upon property values is not available from real estate market. On the one hand, information losses and misunderstandings among different market actors happen inevitably when using different descriptive ways for data interpretations. According to Ventolo [16], there are about 45 data sources involved in the traditional building survey: government councils, professional journals, local material suppliers, building and architectural plans etc. Each market actor in real estate market uses raw data for property valuation according to their own benefits, or they collect and process information from other data source suppliers. Different market actors use various descriptive ways interpreting information in different formats, which causes information losses and misunderstandings during information exchange processes.

Building information modeling (BIM), as an innovative information modelling technology, has been widely developed by a large group of researchers and industrial professionals for project information exchange and management in the Architecture, Engineering, Construction, Operation and Maintenance (AECOM) domain [17]. For example, Marmo et al. [18] proposed to extend the current IFC schema to support building performance assessment and maintenance management. Artus et al. [19] tried modeling damage information using existing IFC schema to support mixed reality inspections and maintenance. Zhiliang et al. [20] proposed an IFC extension to manage information about construction cost estimating for tendering in China. Currently, since no robust standards define the specific requirements for information exchange among different market actors, property valuation professionals have to acquire related information manually. This time-consuming process can be partly automated by using BIM related technologies and concepts: Industrial Foundation Class (IFC) standards, Information Delivery Manual (IDM) and the domain-specific Model View Definition (MVD). In addition, the application of BIM models and related standards in nature have the capability to define, collect, store, manage and exchange related information including sustainability-related data for property valuation in an interoperable and reusable way. Moreover, the ‘BIM according to the ISO 19650 series’ provides an international standard that defines the information management principles and requirements for an improved collaborative working environment in the construction industry [21]. Therefore, property valuation professionals concluded that there was great potential to expand the current BIM data for property valuation use, such as linking data with Building Management Systems [22].

To facilitate the application of BIM in property valuation and support automated property valuation, this paper proposed an integration framework that using BIM and Machine learning (ML) technologies for property valuation which contains a holistic database interpretation, an IFC-based information extraction and an advanced valuation model (GA-GBR) based on machine learning. It is expected that the fundamental database interpretation and IFC-based information exchange could make it easier for valuation professionals to use valued-related design information at various stages within AEC projects. The BIM-ML framework could benefit both AEC and the appraisal professionals at different aspects. The rest of this paper is divided as follows. Section 2 introduces the background of property valuation, BIM, machine learning and the *Construction 4.0*. Section 3 informs the research methodology throughout this paper. Section 4 presents the information exchange between AEC projects and property valuation, during which the fundamental database interpretation and an IFC-based information extraction algorithm are developed. Section 5 explains the methodology and framework for the proposed GA-GBR model. In section 6, both GBR and GA-GBR model are experimented with traded property data from America, and model accuracy and predictive performance are compared. After that, a case study is provided to test the BIM-ML framework using a BIM model, an IFC-based information extraction algorithm and the trained GA-GBR model. Section 7 and Section 8 give the discussion and conclusion of this research.

2. Related Work

2.1 Property valuation

For property valuation, an accurate analysis and estimation of the market price of properties or recent property transactions should be a representation of the attributes of properties, which reflect the underlying fundamentals of market culture and geographical locations [2]. Sales comparison method is currently the most widely used method for property valuation when

similar property transactions in the same market area are available. The valuation process involves firstly comparing the attribute differences between the subject property and transacted properties with similar features, and then adjusting the selling price based on so called “distance” [3]. The “distance”, D is calculated as follows:

$$D = \sqrt[\lambda]{\sum_i [A_i (X_i - X_{si})]^\lambda + \sum_j [A_j \delta (X_j, X_{sj})]^\lambda} \quad (1)$$

where λ = Minkowski exponent lambda; A_i = weight associated with the i th continuous characteristic; A_j = weight associated with the j th categorical characteristic; X_i = value of the i th characteristic in the sale property; X_j = value of the j th characteristic in the sale property; X_{si} = value of the i th characteristic in subject property; X_{sj} = value of the j th characteristic in subject property; \sum_i = summation of terms of i th characteristics; \sum_j = summation of terms of j characteristics; $\delta(a, b)$ = inverse delta function .

Literature analysis of property valuation has concluded three popular research trends: research focusing on sustainability assessment in property valuation, the hedonic pricing models and artificial intelligence models for property valuation [23–25]. For sustainable property valuation, researchers and practitioners are focusing on the social responsibilities, financial benefits and potential risk reductions that sustainable developments may bring into property valuation. The generally accepted environmental, social and economic benefits of green buildings are recognized as energy efficiency, increased health comfort of tenants, low lifecycle energy cost etc [26]. Researchers tried to quantify the direct or indirect financial gains or reduced property risks that sustainable features might bring into property values by referring to internationally accepted green rating systems such as LEED, CASBEE and BREEAM [13,27]. However, the application of sustainability assessment in property valuation is still limited, due to the limited sustainability-related data on real estate market or the valuation professionals lack of knowledge and skills of sustainability assessment.

With respect to hedonic pricing models and artificial intelligence in property valuation, although some researchers believe that hedonic pricing models are more reliable over artificial intelligence models, there are potential limitations of hedonic pricing models. For instance, when it comes to fundamental model assumptions, hedonic pricing models seem unable to effectively deal with the identification of macro-economic influences, the selection of suitable predictor variables and the choice of hedonic equations [28,29]. With the coming of big data era and improved machine assisted computation techniques, using artificial intelligence models for improving the accuracy and efficiency of property valuation has attracted a big number of researchers and professionals. To get a comprehensive understanding of artificial intelligence in property valuation, this research did a systematic scientific search from four main academic databases namely Web of Science, Google Scholar, Science Direct and Scopus. The search criterion was designed as using two groups of keywords: (Property Valuation **or** Mass appraisals **or** House Price **or** Real Estate Appraisal) **and** (Artificial Intelligence **or** Machine Learning) within (Title or Keywords). After the removal of duplicates and manually checked by the author, 45 documents were selected as relevant with artificial intelligence in property valuation (Table 1).

Neural Networks	Genetic Algorithm	GA Optimization	Ensemble	Other
20	5	10	4	6

Table 1: Statistics of collected literature on AI for property valuation from 1997-2020.

Table 1 illustrates that neural networks are the most popular statistic models for property valuation. Many studies have concluded that neural network models provide good performances in house price predictions with small size of experiment data [11,30,31]. However, Rafiei and Adeli [32] argued that BP neural network applications on a limited number of factors have potential limitations that BP requires millions of iterations to converge and cannot deal with complex problems in a reasonable computing time. There are only several studies [8,33,34] on genetic algorithm (GA) applications for property valuation. Ahn et al. [8] used ridge regression coupled with genetic algorithm to enhance real estate price prediction, in which genetic algorithm helps find suitable predictor variables for property valuation. Giudice et al. [34] concluded that genetic algorithms show little improvement for property valuation and the superiority to interpret the real estate markets. However, the integrations of genetic algorithms and other statistic models are quite successful, especially with neural networks. The GA optimized neural network models have the ability to absorb the advantages and get rid of the limitations of both GA and neural networks. For example, Rafiei and Adeli [32] designed a comprehensive model based on the integration of deep restricted Boltzmann machine (DRBM) neural networks and genetic algorithm to predict house price at the design stage or the beginning of the construction. In this research, GA was developed to determine the most influential set of input attributes which generating the best results. Sun [35] used genetic algorithm to optimize the connection weight and threshold of BP neural network for property valuation and reported better performance than traditional BP neural network. On the other hand, in the last two decades, ensemble learning models have attracted many researchers' attention because they usually have good performances for different applications and are flexible to be extended. For instance, Graczyk et al. [10] conducted an experiment that testing different ensembles for real estate appraisal on Weka. It was concluded that in terms of MAPE no single algorithm generated the best ensembles. The ensembles were tested with only four input parameters, which normally lead to the unstable performance of machine learning algorithms. It is worth noticing that all ensemble classifiers with additive regression produced significant error reduction compared to original models.

In summary, it is concluded from the literature that the individual neural network or genetic algorithm has not achieved satisfactory results, but GA optimized neural networks have achieved good performance for real estate appraisals. While ensemble learning has been reported good performance in various domains, but the application of ensemble learning in property valuation is quite limited. Compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility.

2.2 Building information modelling (BIM)

Building Information Modeling (BIM), as a digital and computable representation of a building and its related lifecycle information, has significantly improved information flow among stakeholders involved at various stages-from the early design to the construction and long operation stage [36]. To support information communication, digital technologies such as databases, model servers and project platforms are often employed in a comprehensive manner. A set of common BIM features were concluded by Vanlande et al. [37] such as the ability to store, share and exchange data, the capability to define building information in 3D dimensions, the extensible ability to cover unimplemented information domains, lifecycle information management and the ability to cover all physical and functional features of a building. According to Lindblad [38], the benefits of BIM adoption involve more efficient data exchange, less data input and transfer errors, increased productivity, streamlined construction processes, automated workflow, improved product quality and building performance. Due to its many advantages, BIM has already been adopted in different

construction projects all over the world. Regard on country-wise BIM research, the highest number of publications is in UK with most of the research is done between 2013-2019, followed by South Korea, Australia and China [39]. There are also challenges and barriers for BIM adoptions. The BIM adoption issues are listed as habits of 2D-based work, limited higher education BIM training, the possible reluctance of specialists to holistic planning approaches, lack of fee structures for BIM-specific services, inconsistency among countries regarding the acceptance and adoption of technologies [40]. To address these BIM adoption issues, multiple sources of data collection should be used such as data sources from environmental factors, the perceptions of technology adopters, cross-cultural studies, economic factors and joint BIM implementations with Green building, clouding computing, IoT and Data Science [39].

The Industry Foundation Classes (IFC), contains geometric information and semantic information, is firstly developed by buildingSMART in 1997 as a non-proprietary exchange format of building information to facilitate data sharing and exchange across IFC-compatible applications [41]. After IFC4 released by ISO 16739-21 in 2013, it is accepted as a standardized data format to support building information modelling, information exchange and a variety of analysis based on BIM models, such as quantity-takeoff, cost estimating, damage inspection, energy simulation etc [42]. Beyond this, an information delivery manual (IDM) [43] is defined at project levels to formally specify the user requirements and ensure that the final model would be semantically meaningful to provide most of the information needed for specific business processes. Based on the IDM, a Model View Definition (MVD) was then defined as information concepts needed and proposed as a binding to the IFC standard for exchange of BIM models [44].

On the other hand, under recent concept *Construction 4.0*, the engineering and construction industry is looking for a higher level of digitalization that involves the digitalization of the construction industry in terms of the consequential need for increased data management and potential greater automation [45]. In the context of project management, due to different agendas and incompatible priorities of multiple companies and stakeholders in the construction industry, the organization of different information requirements becomes a key factor in supporting informed decision making throughout the whole project life cycle [36]. In this respect, based on the UK's previous standards for BIM known as PAS 1192 series, BS EN ISO 19650 series of standards (ISO 19650 series) have been released by the BSI, which provides an internationally accepted framework for digital information management processes in a project [46]. According to the ISO 19650-Guidance from UK BIM Alliance [21], the BS EN ISO 19650 series is an international standard of good practice that defines information management principles and requirements within a broader context of digital transformation in the construction and asset management industries. The aim of 'BIM according to the ISO 1950 series' is to get benefit from clearly defined specification and delivery of just the right amount of information (including components, buildings, infrastructure and systems) concerning key decision points within the lifecycle of an asset. A successful implementation of information management that complying with 'BIM according to the ISO 1950 series' can be characterized as: (1) clear interpretation for the information needed by the appointing party and lead appointed party, and for the standards, methods, processed, deadlines and protocols that associated with the production (2) the quantity and quality of information produced being just sufficient to meet the defined information requirements (3) efficient and effective transfer of information among different parties (4) informed and timely decision making [21]. At a project level, there are various types of information requirements and information models from different project stages, such as organizational information requirements (OIR), asset information requirements (AIR), project information requirements (PIR), exchange

information requirements (EIR), asset information model (AIM) and project information model (PIR). The hierarchy and relationships between information requirements and information models are defined in Figure 1 below from ISO 19650-1 [47]. In the context of information requirements between property valuation and AEC projects, at the start of a project delivery process, an AIM can be developed based on AIR considering the managerial, commercial and technical aspects that related to property valuation. Based on relevant information requirements that transferred from AIM to PIM, progressive development of the design model into virtual construction model will be made in the AEC projects. At the end of the project delivery process, the information requirements are transferred from PIM to AIM which can support informed property valuation.

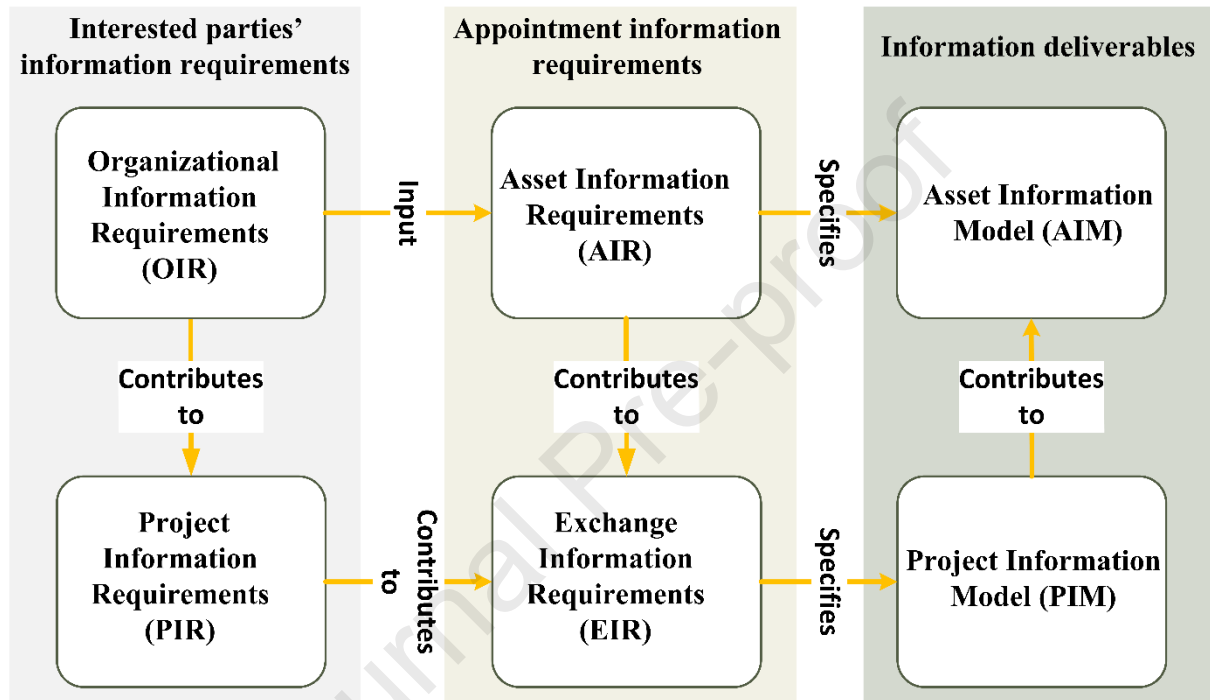


Figure 1: Hierarchy of information requirements (reproduced according to Figure 2 in ISO 19650-1).

2.3 Machine learning

Machine learning (ML): a sub-field of Artificial Intelligence, is defined as ‘A computer program is said to learn from experience E with respect to some class of task T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ’ [48]. Machine learning focuses on creating algorithms or programs that enable computational systems to learn from data on their own [49]. A machine learning system typically has three major components – data, models, and learning. The core of the process is to fit data to a model and train a function approximation algorithm (Hypothesis) based on certain performance criteria. According to Mitchell [48], the basic design of a machine learning system can be classified into 4 main steps: (1) choosing the training experience; (2) choosing the target function; (3) choosing a representation for the target function; (4) choosing a function approximation algorithm. The typical ML process is illustrated in Figure 2 below, in which the training data provides the training experience that the ML system will learn from. The model performance referring to the target function that determining exactly what type of knowledge will be learned and how this will be used by the performance program. After the definition of target function (Model performance), a model representation

(Learning algorithm) will be proposed to describe the target function. Finally, a function approximation algorithm (Hypothesis) will be learned from the training examples based on a specific performance criteria. ML tasks can be classified as: supervised learning (classification or regression), unsupervised learning (clustering or association) and reinforcement learning (reward based). Typical ML applications are well known as follows: Self driving cars, Web search, recommender system, dynamic pricing, fraud detection, spam filters, ad placement.

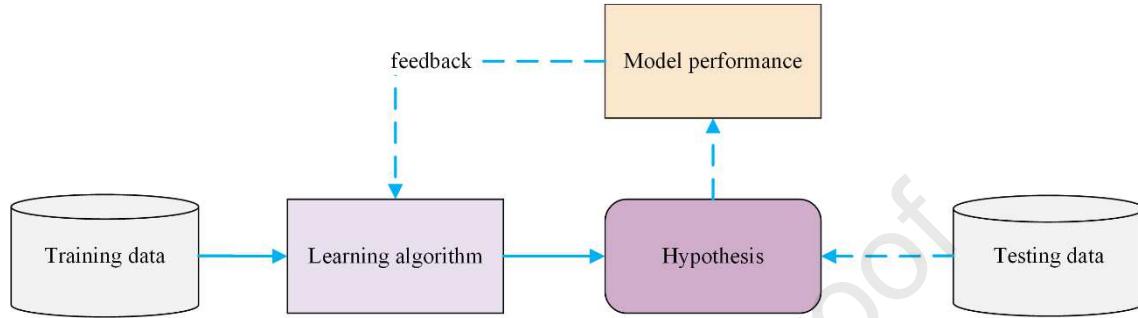


Figure 2: Typical Machine Learning process.

Ensemble learning is a machine learning method which can obtain high performance for testing, of which multiple base learners are trained and combined for the same task. It has been widely used in various applications such as gene expression analysis, text categorization, bankrupt prediction etc. According to Yun [50], there are three fundamentals of ensemble learning: (1) the strategies to train each of base learners; (2) the combining methods of multiple base learners; (3) the critical factors to value the success of ensemble learning model (the bias-variance decomposition). Bagging and Boosting are two representative ensemble methods. In bagging, base learners are trained in a parallel manner. Each base learner is independently trained on resampled training set, which is randomly chosen from the original training set [10]. The data resampling technique ensures the uniqueness of each base learner, which gives bagging ensembles the ability to significantly reduce the variance. Therefore, bagging ensembles are devoted to unstable learners which suffered from large variance such as neural networks [51]. In boosting, base learners are sequentially trained where the first one is trained on the whole training set and the following one is trained based on the performance of the previous one [10]. Boosting ensembles can reduce the bias significantly. Difference between bagging and boosting methods is explained in Figure 3 and Figure 4.

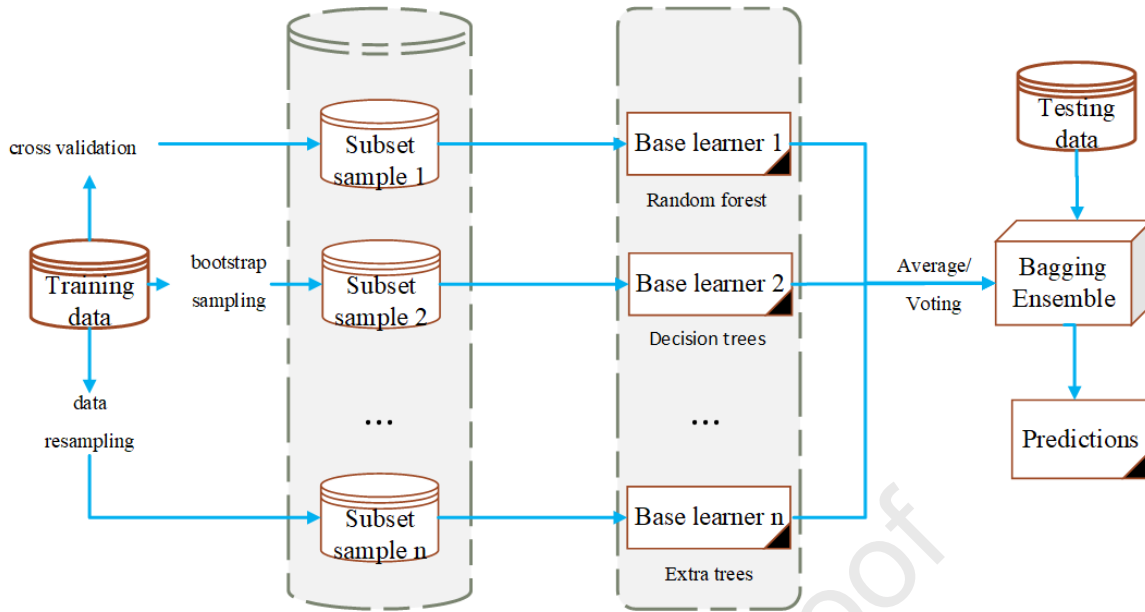


Figure 3: Bagging method.

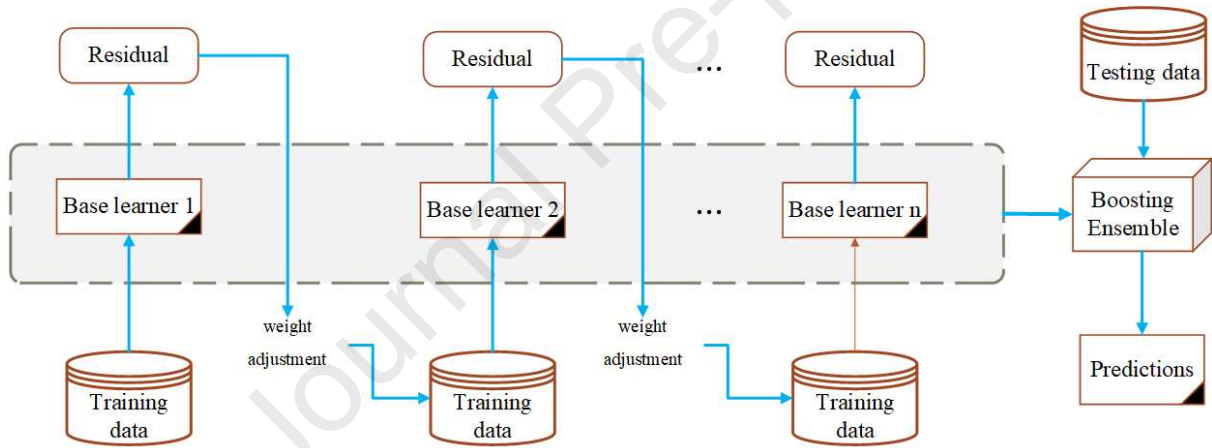


Figure 4: Boosting method.

The gradient boosting regression (GBR) model for property valuation used in this paper belongs to boosting ensemble method. Typically, a gradient boosting regression (GBR) model is constructed in two steps, the details are explained in Table 2 below. First, weak learners $f_0(x)$ are initialized and generated in a sequential style where a new weak learner is trained based on the error of the whole ensemble learnt so far. The logic behind is to produce the new estimators to be maximally correlated with the negative gradient (r_{im}) of the whole ensemble's loss function. Second, the base learners are combined as $\hat{f}(x)$ to do prediction through weighted averaging method. GBR machines have shown great success in various domains (computer-aided medical diagnosis, energy prediction and face recognition) as they often provide high predictive accuracies. The main advantage of GBR ensemble is its flexibility and extensibility, since researchers can choose different classifiers (linear models, decision trees, instance-based, Bayesian or rule-based learners) to train the base learners and customize their loss functions with regard to specific tasks. Besides, the ensemble provides several hyperparameter tuning options (the number of boosting iterations, learning rate, the

maximum depth of individual estimators) that make it flexible to use. However, GBR model is essentially a greedy algorithm and can overfit a training dataset quickly with the number of base learners increasing. Therefore, it is necessary to use regularization to ensure the model's generalization capability.

Inputs:

Training dataset: $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$; Loss function: $L(y, f(x))$

Algorithm:

1. Initialize $f_0(x) = \arg \min_Y \sum_{i=1}^N L(y_i, Y)$

2. For $m = 1$ to M :

a) For $i = 1, 2, \dots, N$ compute the negative gradient:

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$

b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$

c) For $j = 1, 2, \dots, J_m$ compute:

$$Y_{jm} = \arg \min_Y \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + Y)$$

d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} Y_{jm} I(x_i \in R_{jm})$

3. Output $\hat{f}(x) = f_M(x) = \sum_{m=1}^M \sum_{j=1}^{J_m} Y_{jm} I(x_i \in R_{jm})$

Table 2: Gradient tree boosting algorithm.

Genetic algorithm (GA), which simulates the evolutionary process of Darwin's biological evolution theory, has been widely used for multi-parameter optimization problems and non-linearization problems [52]. GA is essentially a heuristic search algorithm which typically performs the search process in four steps (as shown in Figure 5): (1) set initial population for real problems; (2) check the fitness criterion with each member of the population; (3) parents selection with higher fitness values; (4) perform crossover and mutation operators to product new offspring for the solution [53]. One critical element of GA is to choose the right fitness function which defines the ability of each chromosome to solve specific real problems. In this paper, GA is used for searching the optimal tradeoff between diversity and accuracy of GBR ensemble model, as the high model complexity and diversity usually lead to overfitting and the low one lack of accuracy.

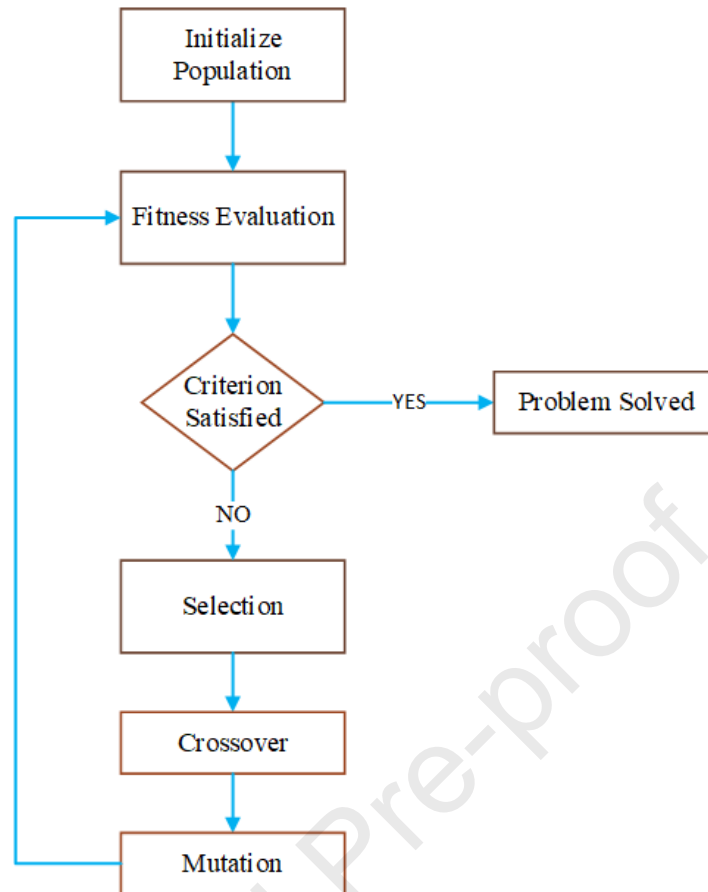


Figure 5: Diagram of simple genetic algorithm.

2.4 Construction 4.0

While there is still no agreement on a universal definition for *Industry 4.0*, it was described as ‘a shift in the manufacturing logic towards an increasingly decentralized, self-regulating approach of value creation, enabled by concepts and technologies such as CPS, Internet of Things (IoT), Internet of Services (IoS), cloud computing or additive manufacturing and smart factories’ [54]. Embracing the current trend of automation and data exchange in manufacturing technologies, the digital transformation under *Industry 4.0* is radically transforming industry and production value chains, with the aim of achieving efficiency, safety, transparency, cost reduction and productivity increases through automation, integration and computer-supported collaborative working [55]. The term *Construction 4.0*, derived from the broader concept - *Industry 4.0*, mainly focuses on the application of computer and cyber-physical systems (CPS) technologies in construction industry [56]. It requires to transform the construction industry towards the 4th industrial revolution that involves digitization of the construction industry and industrialization of construction processes [57,58].

While construction professionals showing increased awareness and willingness to embrace this digital revolution, there are still challenges such as the complexity and heterogeneous feature of construction projects, the uncertainty over tangible and intangible constraints in the individual projects, a highly fragmented supply chain and low-efficient information exchange process caused by the isolated information ‘island’ among different stakeholders and participants [59]. Digital technologies (Figure 6) such as IoT, cloud computing, 3D scanning and augmented reality, BIM and Machine learning play a significant part in addressing these

issues. Leveraging fusion of various technologies can provide significant improvement in the productivity and profitability, which construction industry seeks the most [60]. For example, the integration of BIM and cloud computing has been recognized as the second generation of building information management development and is expected to achieve a greater level of digitalization and collaboration. First, cloud-BIM data can be accessed using various mobile devices such as laptops, tablets and smartphones anytime and anywhere, enabling timely access to updated information, improving decision making and ensuring project delivery [61]. Second, cloud computing and BIM technologies, along with data stored on the Cloud, provides a real-time and collaborative environment for various project stakeholders from different locations [62]. The integration of BIM and augmented reality (AR) is extremely helpful for supporting complex construction tasks and facilitating decision making. For instance, the merge of BIM and AR could provide a vivid presentation of geometric information for operational and managerial tasks, allowing one to visualize how the design fits on-site before construction takes place, managing conflicts and checking safety problems during construction [57,63]. Besides, the integrated BIM and AR could also provide non-geometric information (material information, rigging orders, construction schedules) on tasks and relevant building components, and therefore enhance the quality of construction works [64,65].

Cyber physical systems (CPS), which provide close communication and interaction between cyber and physical components, is expected to play an important part in the design and development of *Construction 4.0*. According to Boyes [66], cyber physical systems (CPS) is defined as 'A system comprising a set of interacting physical and digital components, which may be centralized or distributed, that provides a combination of sensing, control, computation and networking functions, to influence outcomes in the real world through physical processes'. While both the conventional information and communications system (ICT) and CPS focus on processing data between digital and physical components, CPS pay particular attention on the control of physical process to produce positive outcomes [56]. BIM, as a mutual channel for information exchange among operators during the lifecycle of a building, can serve as a complement to CPS that support the increased digitalization requirements in CPS [67]. The

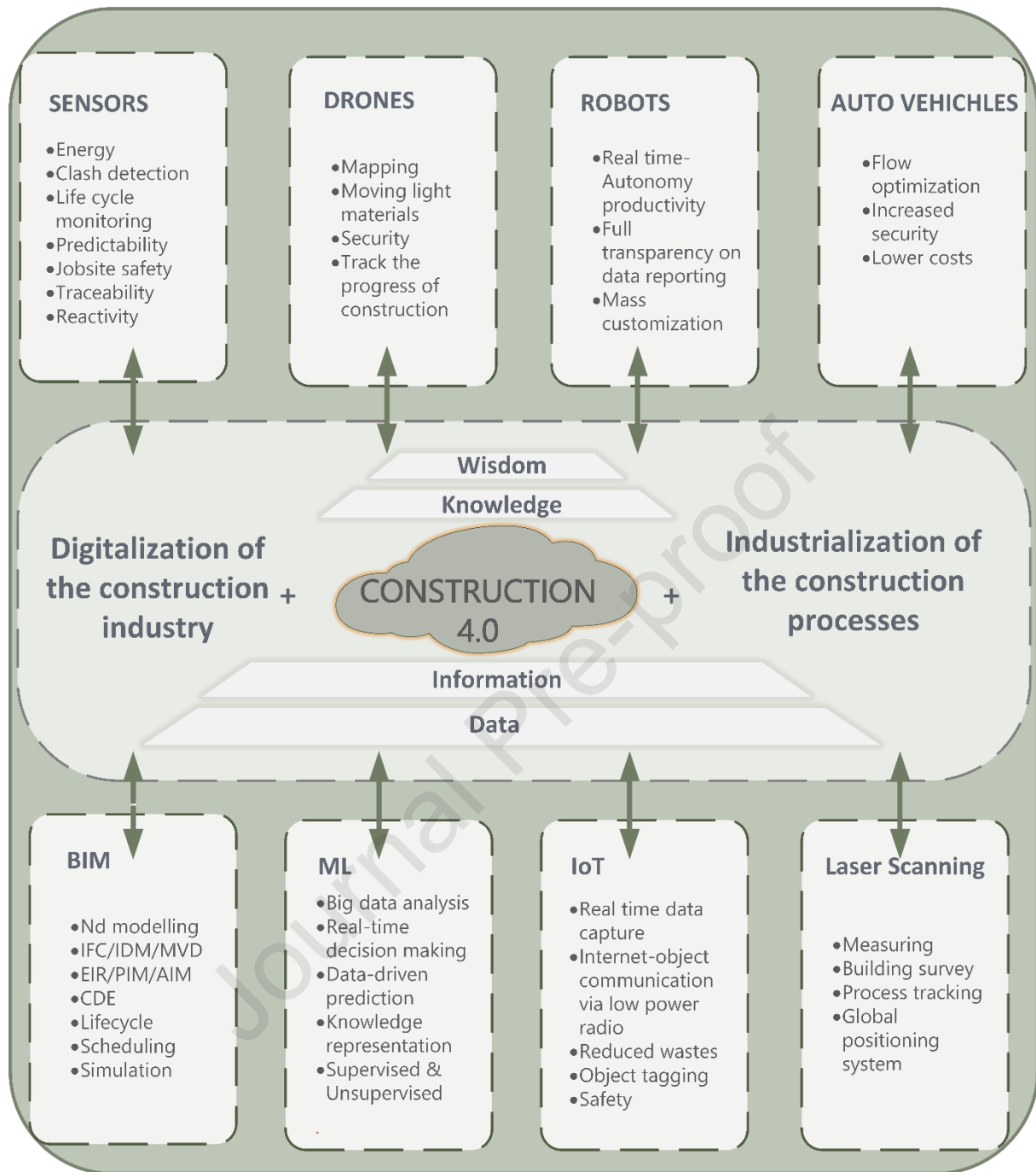


Figure 6: Digital technologies support the ‘Construction 4.0’.

importance and benefits of integrating BIM and CPS have been stressed by several authors. For instance, Ying et al. [68] proposed a cyber-physical based intelligent structural disaster prevention system based on BIM platform, in which BIM and IoT technologies are adopted for constructing the cloud architecture of CPS. With the development of 5G technology in terms of stable, reliable, real-time and secured network communication, the proposed comprehensive system integration is expected to achieve a high degree of integration of monitoring, identification and control, and therefore, to improve the real-time and accurate intelligent monitoring of structural disaster prevention. Bonci et al. [67] proposed a BIM and CPS integration for automatic building efficiency monitoring, in which digital models

developed as BIM which serves as the mirror of the physical system and stores the actual performance recorded by the building during operation phase. As a result, BIM works as the repository of information over several phases of the building lifecycle and keeps the facility at high performance levels and support informed decision making.

The emergence of IoT, Cloud computing, BIM and CPS bring an exponential increase of data (structured or unstructured) in the construction industry, including texts, geometrics, images, videos and sounds, which needs to be processed by Big Data technology (including ML) [69]. With the technological advancements focusing on information modelling, the advent of Big Data era has encouraged a large amount of research on Big Data applications in construction industry. Innovation use of data and Big Data analytical methods have played significantly positive effects on knowledge discovery in construction by automatically discovering hidden knowledge from Big Data repositories. Typical data mining methods or algorithms for prediction tasks are NN-based (ANN, RNN, LSTM), regression (SVM, MLR), DT-based (boosting tree, decision tree and random forest), deep learning (CNN, DBM-SoftMax) and others. While most studies focused on one or more mature data mining methods, only a small group of studies developed improved methods by combining several methods and ensemble models generally produce better performance than individual methods [70]. For example, Cheng et al. [71] designed a multilevel Apriori algorithm based on genetic algorithms, which combines the two algorithms to extract the association rules of construction defects. Yu and Lin [72] introduced a variable-attribute fuzzy adaptive logic control network (VaFALCON) to solve issues of mining incomplete construction data. According to Yan et al. [70], the application of Big Data in construction industry has drawn international attention of both academics and practitioners. Data mining (DM) applications in construction industry are classified into 9 main fields such as building energy consumption, cost estimation, safety management, building design, framework establishment and others. In the 119 selected articles, DM applications on building energy is ranked at the top with 33 articles related, only 9 are related to framework establishment. It indicated that there is no framework establishment research related to the integration of BIM and machine learning. As the amount of data in BIM is rising exponentially, data analytics concepts and tools integrated with BIM might bring added value and produce revolutionary influence on industrial practices, but it exists a significant gap in the property valuation.

3. Research Methodology

In this section the general methodology through this research will be described. While it contains certain elements of both positivism and interpretivism, the pragmatism research philosophy has carried out throughout the lifecycle of this research. For pragmatists, the nature of the research question, the research background and likely research consequences are the main driving forces determining the methodological choice [73]. Generally, both quantitative and qualitative research methods are valued by pragmatists. According to Saunders Mark and Phiip [74], this study uses a partially integrated mixed research methods that contains both quantitative and qualitative methods at the fundamental database collection stage and quantitative methods through the data analysis stage. With regard to research strategies, the archival research, experiment and case study are selected. The archival research is deployed by using the existing information and archived research documents, projects and industrial standards. The experiment method is used to compare the predictive performances between the traditional machine learning model and the proposed GA-GBR model. To evaluate the proposed BIM-ML framework, the case study method is proceeded.

To improve information exchange between property valuation and AEC projects and support automated property valuation, the research design (Figure 7) follows the principles of design science research (DSR) methodology [75], which suggests a series approach:

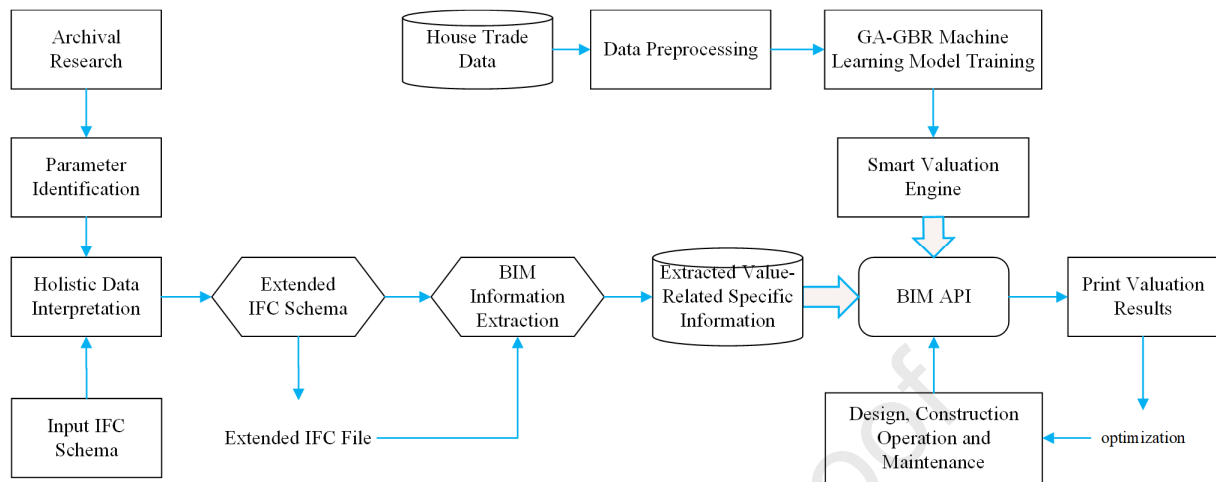


Figure 7: A BIM-ML framework for automated property valuation.

1) Problem explication

The related problems have been explained in the introduction and literature review section which illustrated that the value-relevant design information existed in AEC projects has not been widely used for property valuation, and there is a need to improve information exchange between AEC projects and property valuation. As the volume of data in BIM is rising exponentially, the use of BIM and artificial intelligence information technologies has been recognized as a revolution in construction industry, but it exists a significant gap in the property valuation field.

2) Requirement definition

To identify the required information within the mentioned context, according to Building infrastructure Axiology, the building object hierarchy related to property valuation was provided. The value-relevant attributes or properties for property valuation were achieved through quantitative analysis of 174 achival documents including research papers, practical projects and industry standards.

3) Design and development

IFC extensions including related building object entities and properties were developed based on the required information for property valuation. After that, the required value-specific design information were extracted automatically from an IFC-based BIM instance model to support property valuation. In the meantime, the GBR and GA-GBR model were experimented with traded property data in American real estate market and model accuracy and predictive performance are compared. The advanced valuation model based on genetic algorithm optimized machine learning serves as an automated valuation engine for property valuation.

4) Evaluation

The proposed BIM-ML integration framework for automated property valuation was tested on a case study using a BIM model of a duplex house to see if it meets the requirement defined in this paper.

4. Automatic Information Exchange between AEC Projects and Property Valuation

In this section, the target information will be identified including building object entities and properties. IFC extensions based on the target information will be developed. After that, the required value-specific design information is extracted automatically from an IFC-based BIM instance model using an information extraction algorithm. The workflow of this section is displayed in Figure 8 below.

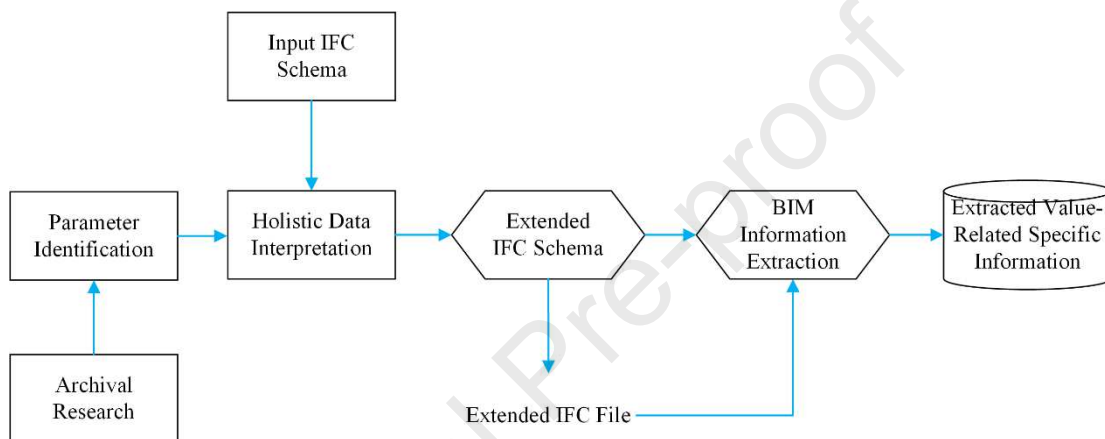


Figure 8: Workflow for IFC extension development and required information extraction.

4.1 The fundamental database interpretation using IFC concept

To provide an overview of the value-related design information for property valuation, a partial building object hierarchy in the Building infrastructure Axiology (Build-Infra-Axio) was displayed in Figure 9 [76]. The Build-Infra-Axio is theoretically founded on the basis of Hartman's formal axiology, which introduced a way to quantify the value of an object based on its properties [77]. For example, a 'good' door should have all the properties (fire rating, security rating, infiltration, self-closing etc) connected to the concept 'door'.

According to our previous archival research of 174 research documents including research paper, projects and industrial valuation standards, the list of value-relevant building properties was explained in Table 3. The required information has been classified into 6 different types of information related to property valuation: information related to environmental quality, social and economic quality, functional quality, process quality, technical quality and site quality [3]. Information included in traditional building survey is compared with information required for sustainability assessment and information achievable within BIM related processes. The column A in yellow color stands for information needed for traditional property valuation. The column B in green color means information needed for green assessment. The column C in dark red stands for information can be defined and developed in the BIM related platform (design, planning, operation and maintenance phases), which is the core for automated partial information exchange for property valuation. The column D in light red means information needed by both property valuation and green assessment.

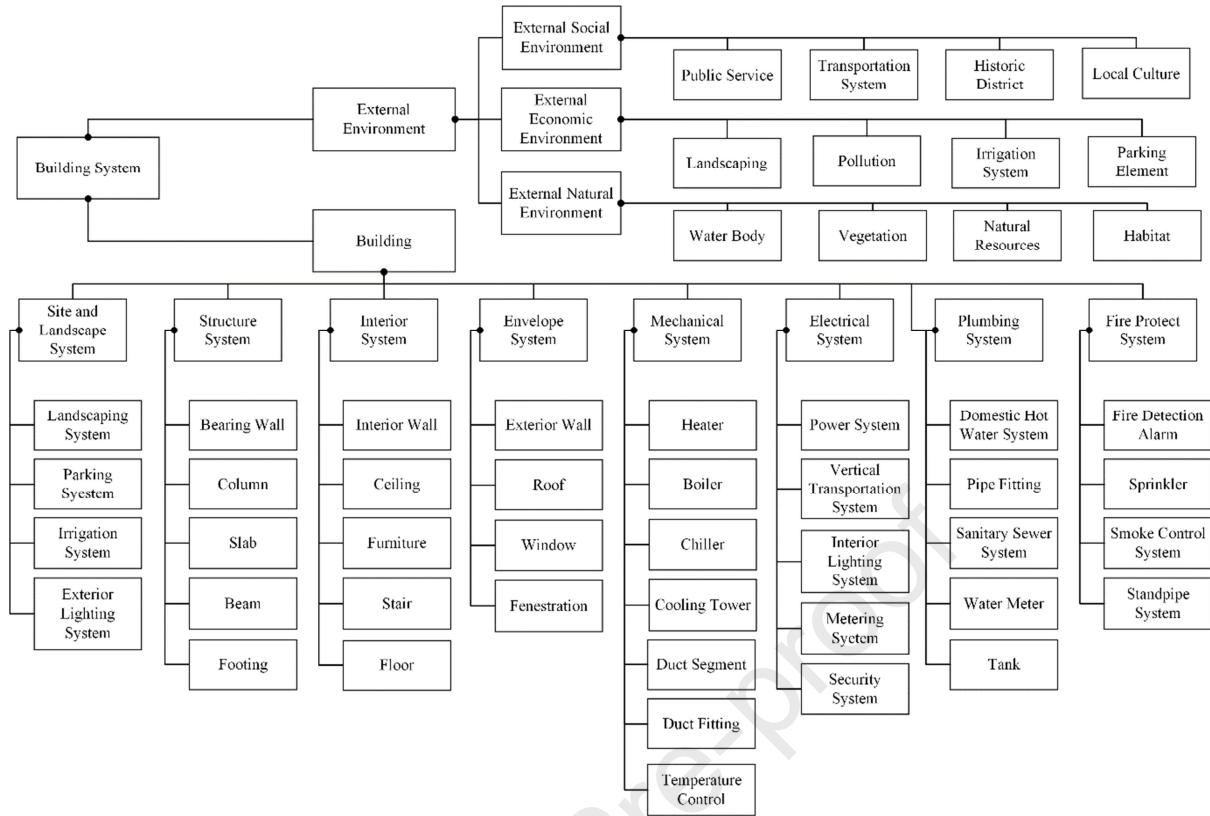


Figure 9: Partial building object hierarchy related to property valuation referring to the Building infrastructure Axiology.

Type of Information	Subtype	Performance indicator and attribute	A	B	C	D
Environmental Quality	Local Environmental Impact	Climate Change				
	Pollution	Noise from transport service and building service equipment, water pollution, land contamination, electromagnetic pollution				
	Land Use	Soil Characteristics				
		Layout, size, inclination, topography				
	Sustainable Resource	Rainwater use				
		Green area				
		Sunlight/Shading				
Waste Water Volume	Waste water disposal					
Social and Economic Quality	Commercial Viability	Policy and economic situation				
		Demographic structure and development				
		Purchasing power, letting prospects, expected rates of return				
		Rental growth potential, inflation				

		expectations, rental payments, other payments	Yellow	White	White	White
		Payments for construction, acquisition, disposal, payments for operating costs, marketing / letting fee, payments for revitalization	Yellow	Green	Red	Pink
		Number of tenants, Duration and structure of rental contracts	Yellow	White	White	White
		Vacancy rate, tenant fluctuation	Yellow	White	White	White
	Safety and Security	Location regrading natural hazards (risk of floods, landslides, collapse)	Yellow	Green	White	Pink
	Lifecycle Cost	Water demand and price, energy demand and price	Yellow	Green	White	Pink
Functional Quality	Indoor Air Quality	Low emitting materials	Yellow	Green	Red	Pink
	Acoustic Comfort	Noise reduction	Yellow	Green	Red	Pink
	Visual Comfort	Sufficient natural light	Yellow	Green	Red	Pink
	Flexibility and Adaptability	Flexibility of use (residential, office, medical practice), adaptability to users	Yellow	Green	Red	Pink
		Wheelchair accessibility	Yellow	Green	Red	Pink
		Wheelchair accessible washrooms	White	Green	Red	White
		Usability of outside space	Yellow	Green	Red	Pink
		Elevators (for all stories or not)	Yellow	Green	Red	Pink
		Wide doors and wide halls	Yellow	Green	Red	Pink
	Brand Value	Floor plan, storey height	Yellow	Green	Red	Pink
		Green certification	White	Green	White	White
		Famous designer	Yellow	White	Red	White
	User Control	Individual temperature controls	White	Green	White	White
Design/Aesthetic Quality	Architectural quality, Holistic monument	Yellow	Green	Red	Pink	
Process Quality	Sustainability Aspects in Tender Phase	Ecological construction materials, risks and impacts for the local environment and residence	White	Green	Red	White
	Documentation for Sustainable Management	Documented maintenance and servicing activities	Yellow	Green	Red	Pink
	Urban Planning and Design Procedure	Public accessibility, quality of layout,	Yellow	Green	Red	Pink
	Construction Process/Site	Quality control during construction (air-tightness, thermography, sound insulation)	White	Green	Red	White
	FM-compliant Planning	Maintenance management	White	Green	Red	White
Basic Information		Structure, age, size, construction type, main construction materials	Yellow	White	Red	White
		Availability of green roofs/green	Yellow	Green	Red	Pink

Technical Quality		facades				
		Degree of revitalization				
		Building equipment and appliances				
	Sound Insulation	Noise protection techniques and components				
	Quality of the Building Envelope	Heat insulation				
		Moisture proofing of the thermal building envelope				
	Ease of Cleaning Building Components	Ease of conducting cleaning, building services and maintenance works				
	Recyclability and Energy efficiency	Ease of recovery and recycling, efficiency of heating ventilation, air conditioning, rainwater use				
	Immission Control	External and internal accessibility				
	Infrastructure	Fitness				
	Quality of Indoor and Outdoor Spaces	Balcony, storage space				
	Safety and Security	Clear arrange routes for escape				
		Protection against burglary				
		Fire Protection				
Quality of sanitary and electronic fixtures						
Structural safety						
Durability of building components						
Site Quality	Local Environment and Policy	Visual context, building permission and planning regulations				
	Transport Access	Public transport, parking				
	Amenities	Area and distance to facilities (shopping, social and medical)				

Table 3: Holistic analysis of value-relevant attributes for property valuation.

After the required information identified, IFC extension for property valuation was developed to cover additional value-related concepts. To improve information exchange among different software sources between AEC projects, industry foundation classes (IFC) was created by buildingSMART as a dominant non-proprietary exchange format [78]. Examples of research and projects to develop IFC extensions were published in the literature: IFC for roads, bridges, tunnels, tendering, design change management and integration with GIS [20,79–83]. The main two steps for developing IFC extensions involves: (1) identifying and analyzing the building object entities and property sets that support property valuation in the existing IFC schema (IFC4-Addendum 2); (2) searching and adding the missing, but necessary, value-relevant entities and property sets to IFC4. Currently, the most updated official IFC schema is IFC4-Addendum 2 from ISO 16739-1:2018 [84]. The existing IFC schema contains building objects (*IfcWall*, *IfcWindow*, *IfcStair*, *IfcSlab*, *IfcDoor*, *IfcCuirainWall* etc) and the property sets of each building object (*Pset_WallCommon*, *Pset_ConcreteElementGeneral*,

Pset_WindowCommon, *Pset_EnvironmentalImpactValues* etc). Some of these property sets can be used for property valuation. For instance, *WaterConsumption*, *RenewableEnergyConsumption* and *ClimateChange* in *Pset_EnvironmentalImpactValues* of *IfcWindow* and *IfcWall* can be used to perform property valuation from sustainability's perspective. *AcousticRating*, *FireRating* and *ThermalTransmittance* in *Pset_WallCommon* can be used to perform property valuation about building functional quality. However, to support a holistic property valuation including sustainability assessment, there is a need to extend the IFC schema including additional entities (water supply, soil, parking) and additional properties of each entity. Therefore, the building system should not only cover the basic building objects in *IfcProduct* schema, but also includes a variety of external environmental elements and sustainable features. The partial proposed entities to be added in the IFC extension are displayed in an entity inheritance diagram in Figure 10. For example, the existing entity *IfcGeographicElement* is developed to include subtypes such as *IfcSoilElement*, *IfcLandscapeElement*, *IfcWaterBodyElement* and *IfcHabitatElement* to support assess the environmental quality surround the building. *IfcSiteElement* is a new entity added to the IFC schema to represent facilities outside the building that support several activities such as parking and waste collection.

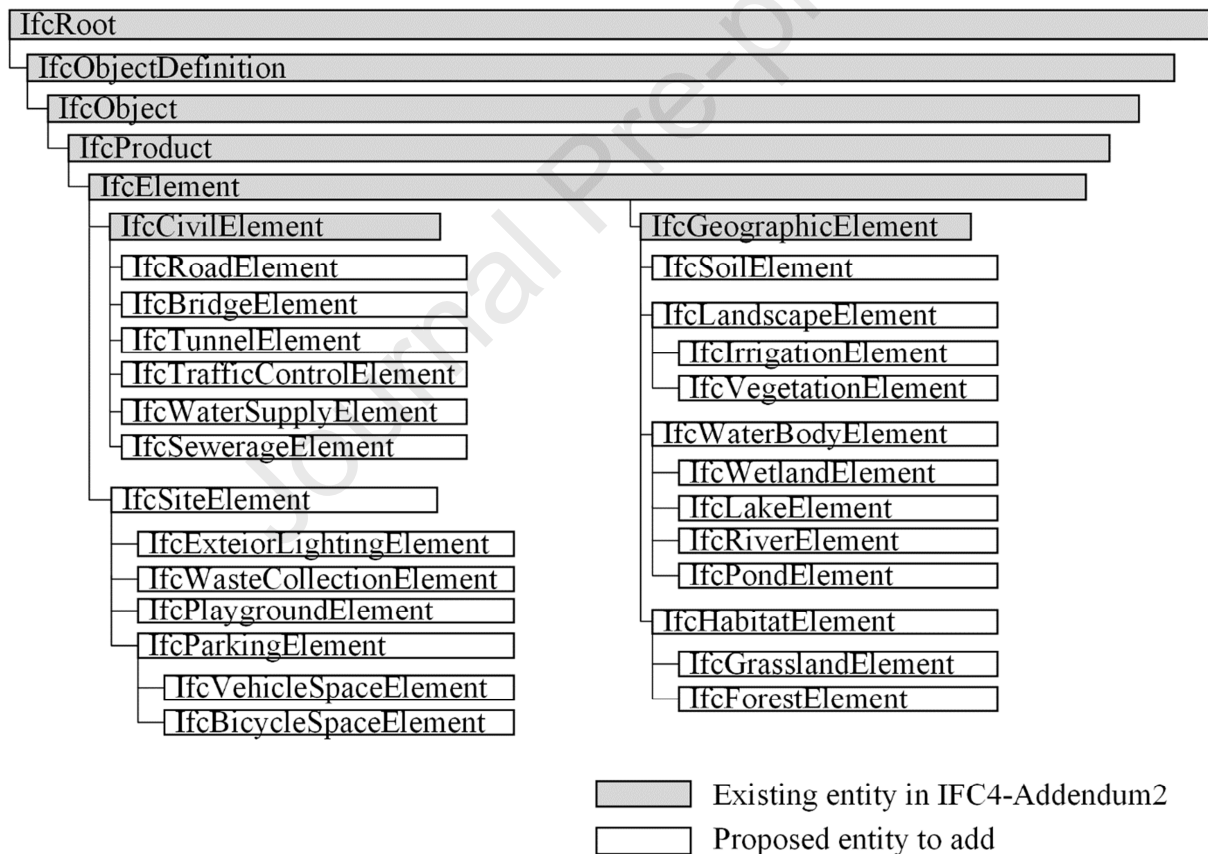


Figure 10: Proposed entities to add in the IFC schema.

A set of property sets are proposed to be added to each of the building objects to support sustainability assessment in property valuation. For instance, the property of low emitting material needs to be added in the *Pset_WallCommon* to evaluate the indoor air quality of a building, which might have an influence on the occupant well-being and work productivity. Examples of properties proposed are shown in Table 4. The classification of the information type of each property is referring to the developed data template in Table 3.

Property name	Value	Template	Description	Information type
<i>RecycledMaterial</i>	<i>IfcPositiveRatioMeasure</i>	Single Value	The ratio of the cost of materials with recycled content, which equals to the sum of postconsumer recycled content, to the total cost of materials of the objects [85].	Process Quality
<i>AirInfiltration</i>	<i>IfcReal</i>	Single Value	The volumetric flow rate of unintentional or accidental introduction of outside air into a building [86].	Functional Quality
<i>SoundInsulation</i>	<i>IfcReal</i>	Single Value	The effectiveness of a surface material in absorbing sound [87]	Environmental Quality
<i>Cost</i>	<i>IfcReal</i>	Single Value	Recycled non-hazardous construction and demolition debris to the total volume (or weight) of materials of this object. It expresses the total estimated construction cost of the object [85].	Social and Economic Quality
<i>LowEmittingMaterial</i>	<i>IfcPositiveRatioMeasure</i>	Single Value	The ratio of the volume or weight of materials with low VOC emissions into indoor air, as within the VOC limit in applicable regulations, to the total volume or weight of materials of this object [85].	Technical Quality

Table 4: Partial set of properties added in the *Pset_WallCommon* for property valuation.

4.2 IFC-based information extraction development

To extract required information elements about building objects and their attributes for property valuation, information exchange between AEC projects and property valuation is delivered through an IFC-based information extraction algorithm. In an IFC-based BIM instance model, the required information defined in the IFC data structure is the globally unique identifier number (GUID), the attributes of building objects including building object names and types, the property sets names and their nominal values of building objects. To develop an IFC-based information extraction algorithm, it is necessary to understand the types of information elements and their relationships between *IfcObject* and *IfcProperty* in an IFC-based instance model. Referring to IFC4-ADD2 schema, *IfcObject* and *IfcProperty* are linked directly and indirectly. The relationships between them are displayed in Figure 11 and Figure 12 [84]. On the one hand, they are directly linked through *IfcRelDefinesByProperties* (Figure 11) – an objectified relationship that defines the relation between objects and property sets. On the other hand, they are indirectly linked through *IfcRelDefinesByType* (Figure 12). *IfcRelDefinesByType* defines the relationship between an object type and object occurrences which can leverage a one-to-N relationships.

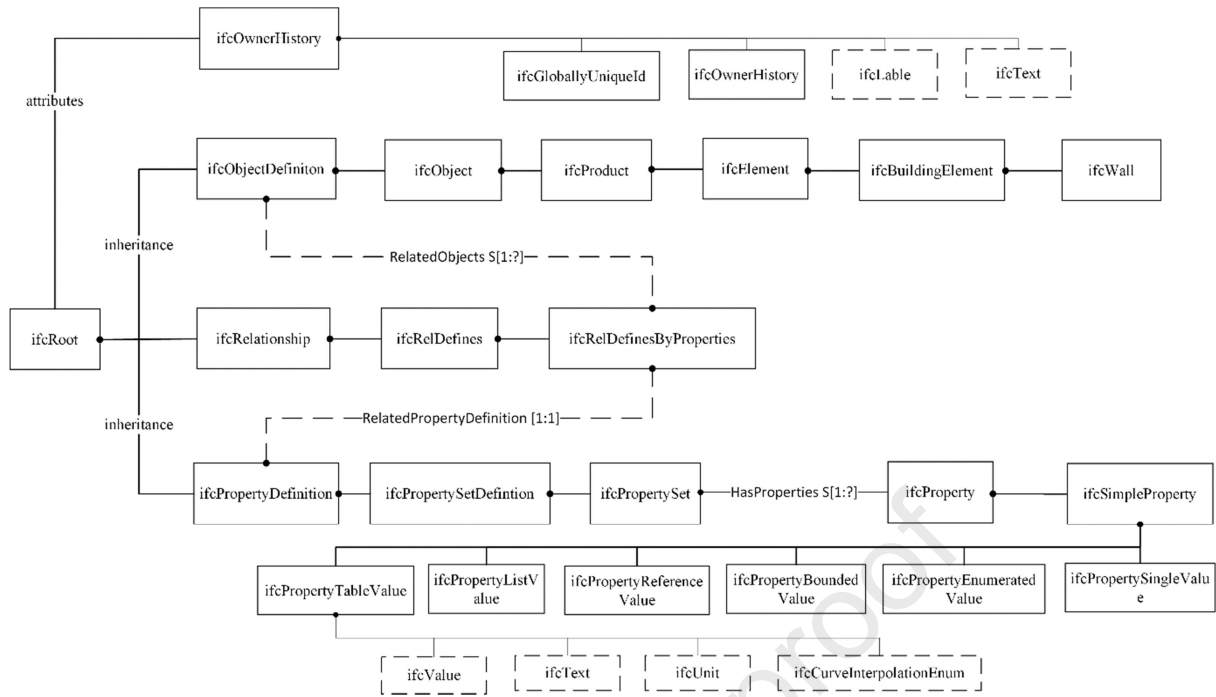


Figure 11: Direct relationships between *ifcObject* and *ifcProperty* in IFC4-Addendum 2.

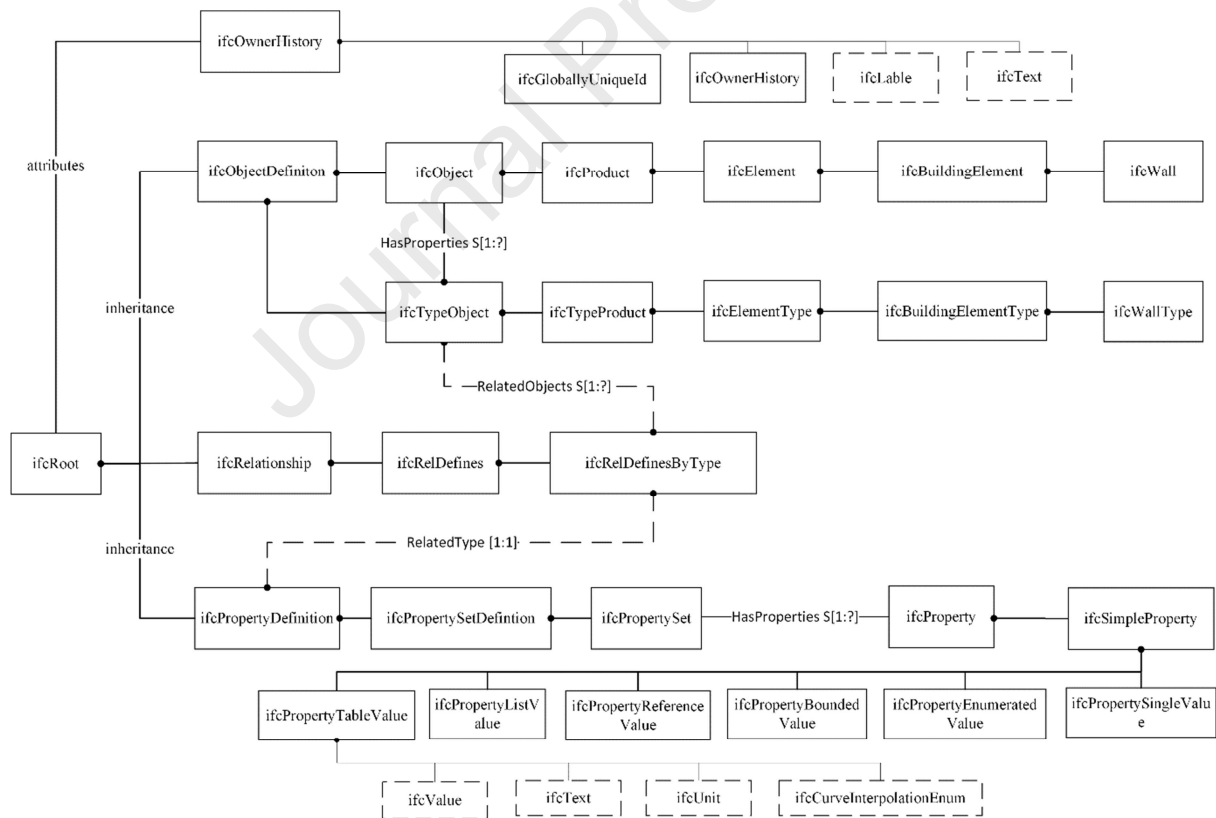


Figure 12: Indirect relationships between *ifcObject* and *ifcProperty* in IFC4-Addendum 2.

The IFC-based information extraction algorithm was developed on python 3.7 using *IfcOpenshell-python* module on *Pycharm* software [88]. The flowchart of the IFC-based information extraction algorithm that extracting the required building entities and properties directly and indirectly is illustrated in Figure 13. The extraction process can be classified into

eight steps: (1) it will iteratively go through the *IfcRelDefinesByProperties* and *IfcRelDefinesByType* instances until all required data instances are extracted; (2) it will extract all the ID numbers of *IfcObject* and *IfcPropertySet* or *IfcTypeObject* from instances extracted in step 1; (3) it will find instances of *IfcObject*, *IfcPropertySet* and *IfcTypeObject* based on ID numbers of instances extracted in step 2; (4) it will extract ID numbers of *IfcPropertySet* and *IfcProperty* instances from step 3; (5) it will find instances of *IfcProperty* and *IfcPropertySet* based on ID numbers extracted in step 4; (6) it will extract ID numbers of *IfcProperty* instances from the extracted *IfcPropertySet* instances and find instances of *IfcProperty* based on ID numbers of *IfcProperty* instances; (7) it will extract object names, object types, property names and property nominal values from step 3, step 5 and step 6; (8) the duplicates of the extracted object names, object types, property names and property nominal values will be compared and removed.

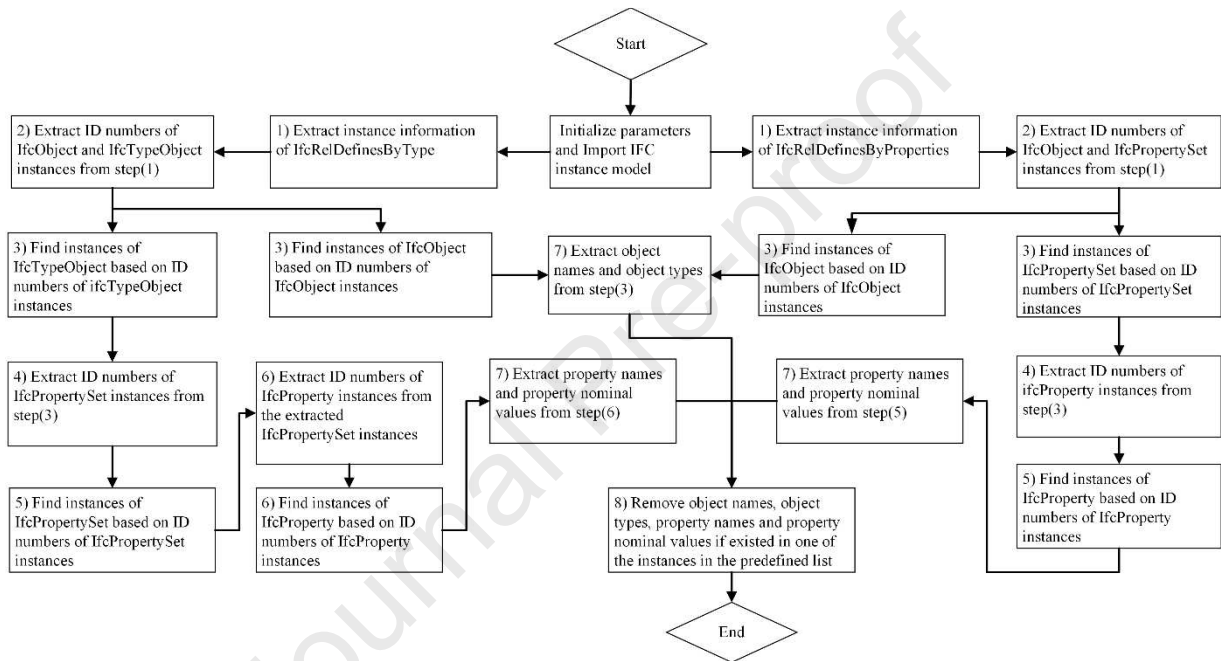


Figure 13: Flowchart of the IFC-based information extraction algorithm.

An example of the application of the developed IFC-based information extraction algorithm to extract required information through the *IfcRelDefinesByProperties* and *IfcRelDefinesByType* instances is explained in Figure 14. First, an *IfcRelDefinesByType* instance with the ID number of #442681 and an *IfcRelDefinesByProperties* instance with the ID number of #563 were extracted. Subsequently, the ID numbers of an *IfcObject* instance (#215), an *IfcPropertySet* instance (#558) and an *IfcTypeObject* (#551) were extracted from instance (#442681) and instance (563). Second, an *IfcWall* instance (#215), an *IfcPropertySet* instance (#558) and an *IfcWallType* instance (#551) were found by the algorithm. Subsequently, the ID numbers of two *IfcProperty* instances were extracted from the *IfcWallType* instance (#551) and the *IfcPropertySet* instance (#558) respectively. Third, an *IfcProperty* instances (#554) and an *IfcProperty* instance (#557) were found by the algorithm. After that, the object name (Basic Wall:300_22_wand_HSBwand_12-140-12:7326535), object type (Basic Wall:300_22_wand_HSBwand_12-140-12:7011920), property names (*FireRating* and *LoadBearing*) and property nominal value (*IfcLabel('60')* and *IfcBoolean(F.)*) were extracted with removed duplicated data from the *IfcObject* instance (#215), the *IfcProperty* instances (#554) and the *IfcProperty* instance (#557).

(1) extraction based on direct relationships

```

{An IfcProperty instance}:
#557=IfcPropertySingleValue(FireRating,IfcLabel('60'),$)
                                     Property Name      Property Nominal Value

{An IfcWall instance}:
#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw',#41,Basic Wall:300 22 wand HSBwand 12-140-12:7326535,$,'Basic
Wall:300-22_wand_HSBwand_12-140-12:7011920',#187,#213,'7326535')
                                                                                                     Object Name

{An IfcPropSet instance}:
#558=IfcPropertySet('1MhDg8jdf98g77kSIHoeHw',#41,'Pset_WallCommon',$(#553,#554,#555,#556,#557))

{An IfcRelDefinesByProperties instance}:
#563=IfcRelDefinesByProperties('33prtgrUP4ZeaQc3XIaZwF',#41,$,$(#215),#558)
                                                                                                     RelatedProperty
                                                                                                     Definition
                                                                                                     RelatedObjects

```

(2) extraction based on indirect relationships

```

{An IfcProperty instance}:
#554=IfcPropertySingleValue(LoadBearing,IfcBoolean(F),$)
                                     Property Name      Property Nominal Value

{An IfcPropertySet instance}:
#558=IfcPropertySet('1MhDg8jdf98g77kSIHoeHw',#41,'Pset_WallCommon',$(#553,#554,#555,#556,#557))

{An IfcWallType instance}:
#551=IfcWallType('1iAXenQ2f67eN2WTRSSr6B',#41,'Basic Wall:300_22_wand_HSBwand_12-140-
12',$,558,$,$,'7011920',$,$,STANDARD.)
                                     RelatingType

{An IfcWall instance}:
#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw',#41,Basic Wall:300 22 wand HSBwand 12-140-12:7326535,$,'Basic
Wall:300_22_wand_HSBwand_12-140-12:7011920',#187,#213,'7326535')
                                                                                                     Object Name

{An IfcRelDefinesByType instance}:
#442681=IfcRelDefinesByType('2iM7Gcqlv00x8z3DIHqa85',#41,$,$,
(#215,#654,#1048,#1470,#1895,#2402,#2838,#3129,#3407,#3451,#3495,#3539,#3596,#3965,#5125,#5345,#5478,#5711,#584
4,#6077,#6221,#6441,#6666,#6710,#6754,#6798,#8090,#8446,#8811,#42471,#42654,#43151,#413124),
#551)
                                     Object Type
                                     RelatedObjects

```

Figure 14: An example for using the IFC-based information extraction algorithm.

An example of an extracted information item for required information extraction is displayed in Table 5.

Instance GUID	Object name	Object type	Property name	Property nominal value
215	Basic Wall: 300 22 wand HSBwand 12-140-12: 7326535	Single Value	FireRating	IfcLabel('60')

Table 5: An example of an extracted information item.

5. Genetic Algorithm Optimized Gradient Boosting Regression Ensemble Model (GA-GBR)

In this section the methodology of GA-GBR model will be described. As explained earlier, the gradient boosting regression (GBR) ensemble model belongs to a boosting ensemble machine learning. The proposed GA-GBR model creates a number of base learners (the weak classifiers) and their associated sequence of combination using the genetic algorithm [89]. The GA uses fitness function to limits the number of base learners, which gives the GA-GBR

model a better interpretability over other boosting ensemble algorithms. It has been recognized that a good ensemble depends on the individual learners being as accurate and diverse as possible. However, generating diverse individual learners is quite challenging as the individual base learners are usually highly correlated for dealing with the same training data. This means the more accurate individual learners are, the less diverse they are. Therefore, the success of a GBR ensemble is essentially about getting a good tradeoff between the accuracy and diversity of individual base learners. To encourage the diversity of an ensemble, the basic logic is to inject some heuristic mechanisms into the learning process. There are four typical effective methods: **manipulation of training data**, **manipulation of input features** (the random subspace method), **manipulation of learning parameters and manipulation of output representation** [90]. The optimal method of GBR diversity in this research belongs to manipulation of input features using genetic search. For data with a big number of input features, input feature manipulation method often gives a good result.

Implementation of the proposed GA-GBR model has three phases (as shown in Figure 15): base learner initialization, problem encoding and genetic search. Details about each phase will be described next.

5.1 Base learner initialization

The objective function of an individual base learner is to learn a mapping $f(x)$ between the input feature vector and the prediction output (house price). GBR ensemble combines multiple base learners $f_m(x_i)$ to generate a strong model $\hat{f}(x)$, which is displayed below. There are three common base-learner models: linear models, smooth models and decision trees. The base learners in this research use decision trees of same sizes which are good at handling mixed types of data and modeling complex functions.

$$\hat{f}(x) = \sum_{m=1}^M f_m(x_i) \quad (2)$$

Typical loss functions for regression model are Gaussian L_2 loss function, Laplace L_1 loss function, Huber loss function and Quantile loss function [91]. After initial testing on the GBR model, it had the best prediction accuracy with Huber loss function. The Huber loss function used in this study comprises two parts of L_1 and L_2 loss function, which is illustrated as follows:

$$L(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & |y - f(x)| \leq \delta \\ \delta(|y - f(x)| - \frac{\delta}{2}) & |y - f(x)| > \delta \end{cases} \quad (3)$$

where $f(x)$ are base learners, δ is used to define the robustification effect of the loss function.

The genetic algorithm makes use of a pool of trained base learners which are trained a priori, to avoid regenerating them in successive generations of GA optimization [92].

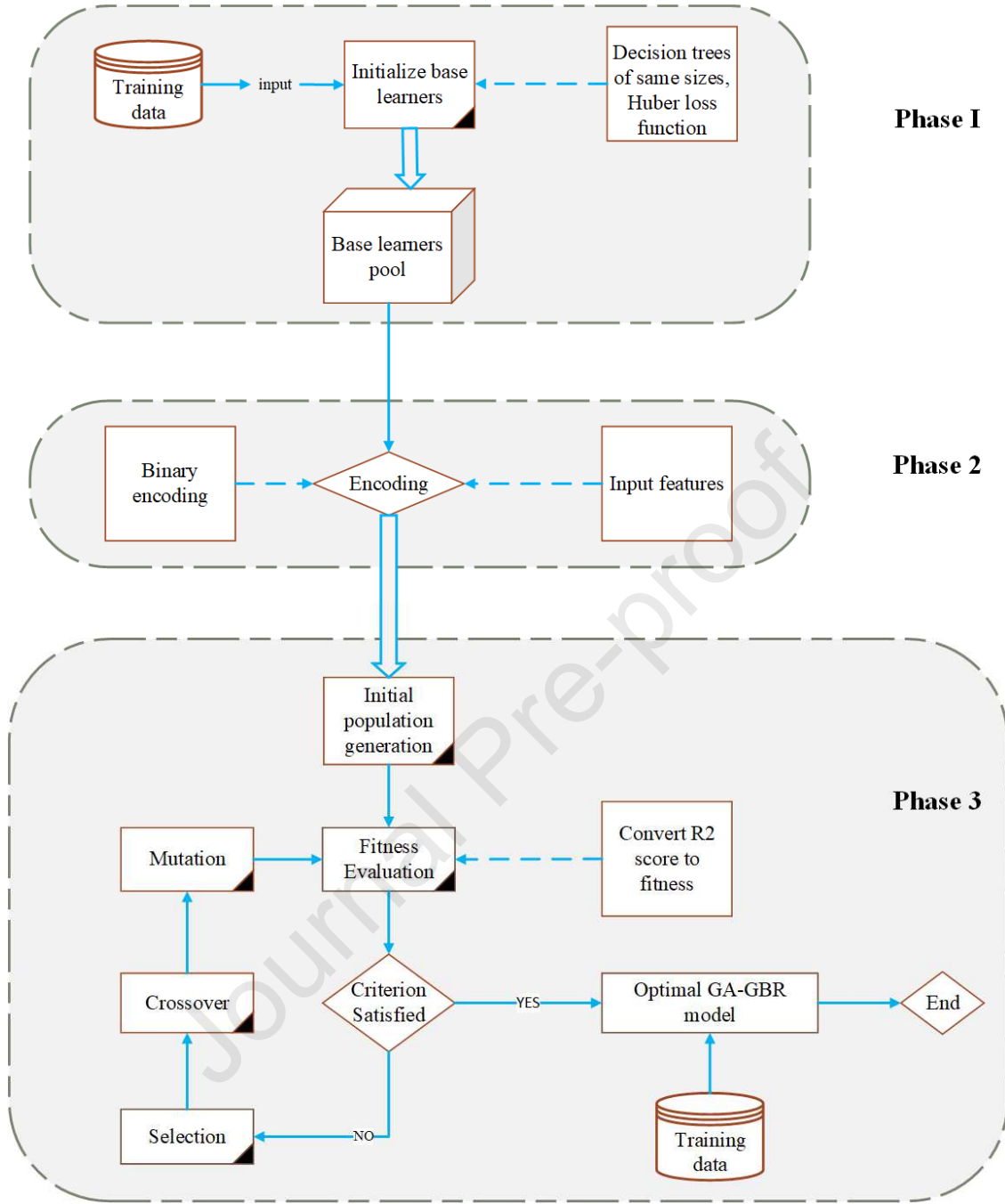


Figure 15: Framework of the GA-GBR model.

5.2 Problem encoding

There are several ways to encoding a problem in genetic algorithm, such as binary codification, decimal, hexadecimal and so on. This research uses binary encoding to represent the solutions. As mentioned earlier, to encourage the diversity of base learners without losing accuracy, input feature manipulation method often gives a good result when the training datasets have a big number of input features. The collection of experimental datasets in this research has 56 input parameters (details will be explained in Section 6.1).

In the GA-GBR model training, each chromosome in the population represents an individual which has all the input features of the training data (shown in Figure 16). The size of the

chromosome is decided by the 56 input features, for examples, parameter A represents house size and parameter B represents central heating.

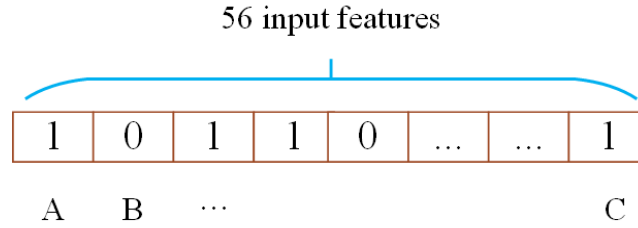


Figure 16: Binary encoding of chromosome.

5.3 Genetic search

As mentioned in section 2.3, Genetic search consists of three main steps: (1) initial population generation; (2) fitness function definition; (3) selection, cross over and mutation.

1) Chromosome

When dealing with a real search problem, the search parameters need to be encoded and represent the problem as a function objective. In this case, the chromosome is represented by the 56 input parameters with each feature has two different values 0 and 1 (0 represents false, 1 represents true). While the big size of population often increases the diversity of the population, but also harms the speed of genetic search.

2) Fitness function definition

Fitness function is defined in GA to assess the ability of a chromosome to solve the problem. In this research, the regression accuracy measure-coefficient of determination (R^2) is set as the fitness function of the genetic algorithm. It represents the proportion of variance that produced by the independent variables in a machine learning model and explains how well the model fits the training data and the generalization of the model. R^2 function is defined as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where \hat{y}_i represents the predicted value of the i^{th} sample, y_i is the corresponding true value, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$.

3) Selection, crossover and mutation

Parents are selected based on their fitness values: chromosomes with higher R^2 are more likely to be parents of new generations. Tournament selection and roulette wheel selection are two popular selection methods. This study uses the roulette wheel selection method.

After parent selection, 2-point crossover method is performed between two chromosomes. Each crossover point is set on the boundary between the two chromosomes below. For example, binary string between the two crossover points of chromosome 1 and the rest part of chromosome 2 are copied to make a new offspring (shown in Figure 17).

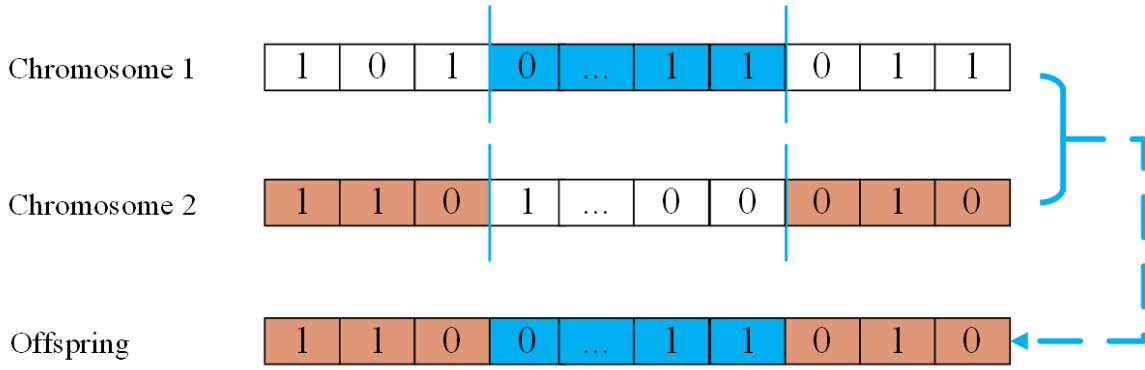


Figure 17: 2-point crossover method.

The last operator of genetic algorithm is mutation which randomly chooses a point in one bit-string and inverts the value of selected bit (from 1 to 0 or from 0 to 1). Mutation can occur at each point of chromosomes with small possibility, mutation rate between 0.5%-1% normally gives better performance of GA search [93]. Operators with big mutation rate often have strong ability to generate new offspring and prevent premature convergence of genetic search, but also may harm the stability of population structure.

The parameters in GA in this research are explained in Section 6.2 (GA-GBR model training).

6. Empirical Studies

6.1 Data preparation

The experimental data was collected from public machine learning datasets collection on *GitHub* website, 18597 traded houses were selected from January 2000 to December 2017 in 40 different cities of USA. Each property data contains 56 property attributes including total area, living area, garage type, garage size, pool, bathroom, fireplace, number of bedrooms, carport area, built year, stories, central cooling, central heating and address (different cities). The details about the property attributes were illustrated in Table 6. Continuous, categorical and binary attributes were applied. For example, continuous attributes contained total area, living area, garage area, built year and house price. Fireplace, pool, central heating and central cooling used binary variable, garage type had three categorical types: attached, detached and none.

Attributes	Description
Total area	Size of property in square meters (m ²)
Living area	Size of living area in square meters (m ²)
Garage area	Size of garage in square meters (m ²)
Pool	Whether have a pool or not
Attached garage	Whether have attached garage or not
Full bathroom	The number of full bathrooms

Fireplace	Whether have a fireplace
No garage	Have no garage
Bedrooms	The number of bedrooms
Carport area	Size of carport area in square meters (m ²)
Built year	The built year of property
Stories	The total number of stories
Half bathroom	The number of half bathrooms
Central cooling	Whether have central cooling system or not
Central heating	Whether have central heating system or not
Detached garage	Whether have detached garage or not
City	The name of cities

Table 6: Property attributes.

The descriptive statistics (Table 7) displays the variability within the data. The average property size is 2136.8 m² with three bedrooms. The average house price is \$412567.01 and standard deviation of house price is \$343382.92.

Attributes	Minimum	Maximum	Mean	Standard Deviation
Total area	128	12448	2136.8	928.39
Living area	124	9806	1994.84	850.66
Garage area	0	8318	455.24	244.12
Carport area	0	7201	41.29	170.23
Bedrooms	0	7	3	0.98
House price	626	21041998	412567.01	343382.92

Table 7: Descriptive statistics of attributes.

6.2 GBR and GA-GBR model training

1) Experimental setup

The training of the GBR and GA-GBR models was performed on the Python 3.7 using scikit-learn library on *Pycharm* which was an integrated development environment using python language for machine learning. The experimental dataset was divided into 20 groups with 1000 houses data in each of them. To find the best performance of the GBR model, model hyperparameters such as number of estimators, learning rate, maximum depth of decision trees, minimum sample leaf and loss function were tested using grid search optimization algorithm. The hyperparameters tested with grid search algorithm were displayed as follows:

- Number of estimators: 100, 150, 250, 500
- Learning rate: 0.01, 0.05, 0.1, 0.2
- Maximum depth: 4, 6, 8, 10
- Minimum sample leaf: 3, 5, 9
- Maximum features: 0.1, 0.3, 1
- Loss function: Ls, Lad, Huber

2) GBR model training

After the dataset was qualified, inspected and prepared, the dataset was randomly split into training set (70%) and testing set (30%). It was found that the experimental error of the GBR model was the smallest when model hyperparameters were set as the number of estimators (150), learning rate (0.05), maximum depth (8), minimum sample leaf (5), maximum features (0.1), loss function (Huber). For feature selection, using the GBR classifiers, the input features were sorted based on the feature importance ranking from the model. The input features and their weights of feature importance by percentage were displayed in Table 8. The 56 input features were further used for problem encoding in the next GA-GBR training section.

Attributes	Weights	Attributes	Weights	Attributes	Weights
Total area	25.92%	No garage	2.77%	Central heating	0.37%
Living area	19.25%	Bedrooms	2.48%	Detached garage	0.11%
Garage area	7.87%	Carport area	1.89%	City Wendybury	0.06%
Pool	3.81%	Built year	0.97%	City Toddshire	0.05%
Attached garage	3.56%	Stories	0.97%	City Amystad	0.03%
Full bathroom	3.48%	Half bathroom	0.78%	City Leahview	0.01%
Fireplace	3.28%	Central cooling	0.38%	...	

Table 8: Weights of attributes by percentage.

3) GA-GBR model training

Generally, the GA-GBR model was trained based on the framework designed in Section 5. The initial population was randomly generated with N solutions, with the number of base learners and their associated combination method. After N solutions randomly generated, the next generation of solutions was generated through genetic search operations. The fitness of each chromosome of the new generation was evaluated according to the fitness function. The chromosomes with higher R^2 scores than the GBR model were selected.

The parameters in GA were tested using trial-and-error method using gradient boosting regression ensemble library in scikit-learn and the specifically designed genetic algorithm.

Through repeated tests, it was found the best performance of GA-GBR model when the parameters in GA were set as follows:

- Population size: 600
- Generations: 32
- Crossover probability: 0.5
- Mutation rate: 0.1

For one of the 20 groups, in Figure 18 the blue line showed the coefficient of determination (R^2) of the GBR model trained with the best performance when hyperparameters were set as the number of estimators (150), learning rate (0.05), maximum depth (8), minimum sample leaf (5), maximum features (0.1), loss function (Huber). The red line explained the dynamics of R^2 of the best chromosome in each generation during genetic search process in the GA-GBR model. It was found that 21 new GA generations (out of 32 generations) had a higher R^2 over the conventional GBR model, which made the GA-GBR model self-adaptive and generalized to the complex property price prediction. Finally, in each group, the best chromosome with the highest R^2 was selected for testing model predictive accuracy. The detailed experimental results for testing will be displayed in next section.

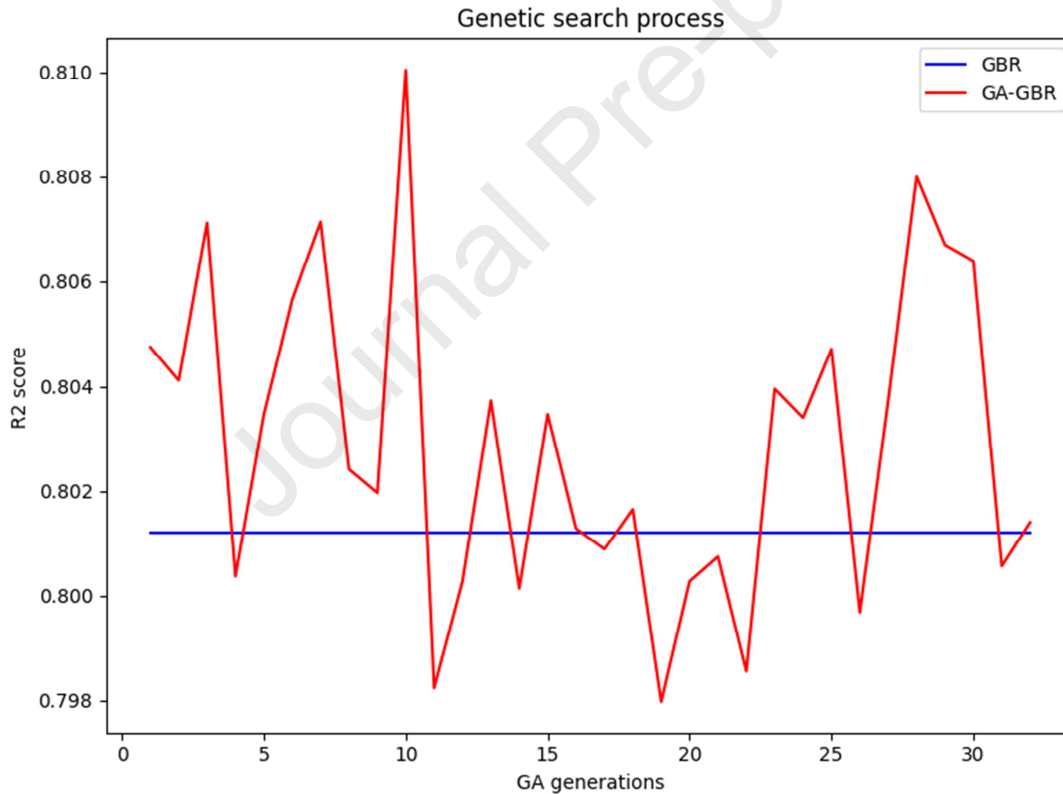


Figure 18: Coefficient of determination (R^2) changes during genetic search process.

6.3 Experimental results and performance comparison

The predictive accuracy and performance of both the GBR and GA-GBR models in the 20 groups were tested based on coefficient of determination (R^2) and mean squared error (MSE). R^2 and MSE is computed as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

where \hat{y}_i represents the predicted value of the i^{th} sample, y_i is the corresponding true value, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$.

$$MSE = \frac{1}{n} \sum_{t=1}^n (A_t - F_t)^2 \quad (6)$$

where A_t is the actual value and F_t is the predicted value.

The performance of the trained GBR model and GA-GBR model was evaluated on test sets, which were randomly chosen from the whole dataset to remove bias. In Figure 19, for all the 20 groups, the mean of R^2 in GA-GBR had an advantage of 1.9% over the GBR model, with 79.5% for GA-GBR and 77.45% for GBR respectively. The maximum R^2 of the two models were: 89.4% for GA-GBR and 88.6% for GBR. According to McCluskey's research on prediction accuracy of different machine learning models for property valuation, in terms of R^2 , showing 78% for model regression modelling (MRM), 82.3% for ANN, 88.7% for spatial simultaneous autoregressive (SAR), 87.9% for geographically weighted regression [94]. Compared with these statistic models, the maximum R^2 of the GA-GBR model in this research is 89.4%, which shows the superiority. For MSE which computed risk corresponding to the expected value of the squared loss, the mean MSE of GA-GBR models in the 20 groups was 135460 which was lower than 143016 in the GBR models for testing sets. The minimum MSE of the two models were: 87029 for GA-GBR and 88026 for GBR respectively (Figure 20).

It was found from Figure 21 that there was a large difference between the actual price and the predicted price by GBR model, and the overall fitting degree was low. While the difference between the actual price and predicted price by the GA-GBR was small, with a high overall fitting degree. Therefore, it was concluded that the GA-GBR model was more accurate for house price prediction and had a higher overall fitting degree. Besides, it was found from Figure 22 that the convergence of the training of the GA-GBR was much faster than that of the GBR model. The reason was that the GA-GBR model had searched the most appropriate number of base learners and their combination in terms of sequence, with increased model diversity and reduced model dimensionality.

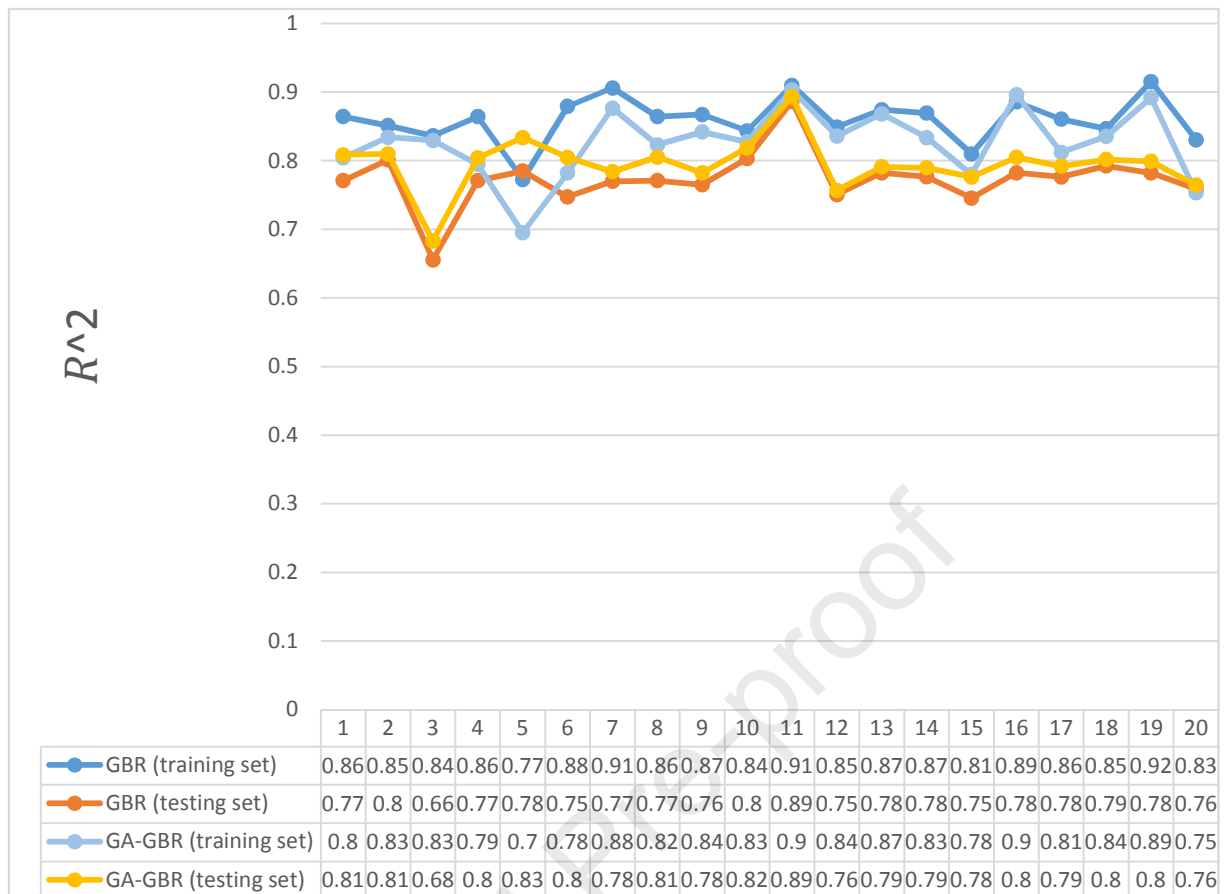


Figure 19: Comparison of the Coefficient of determination (R^2) of two models.

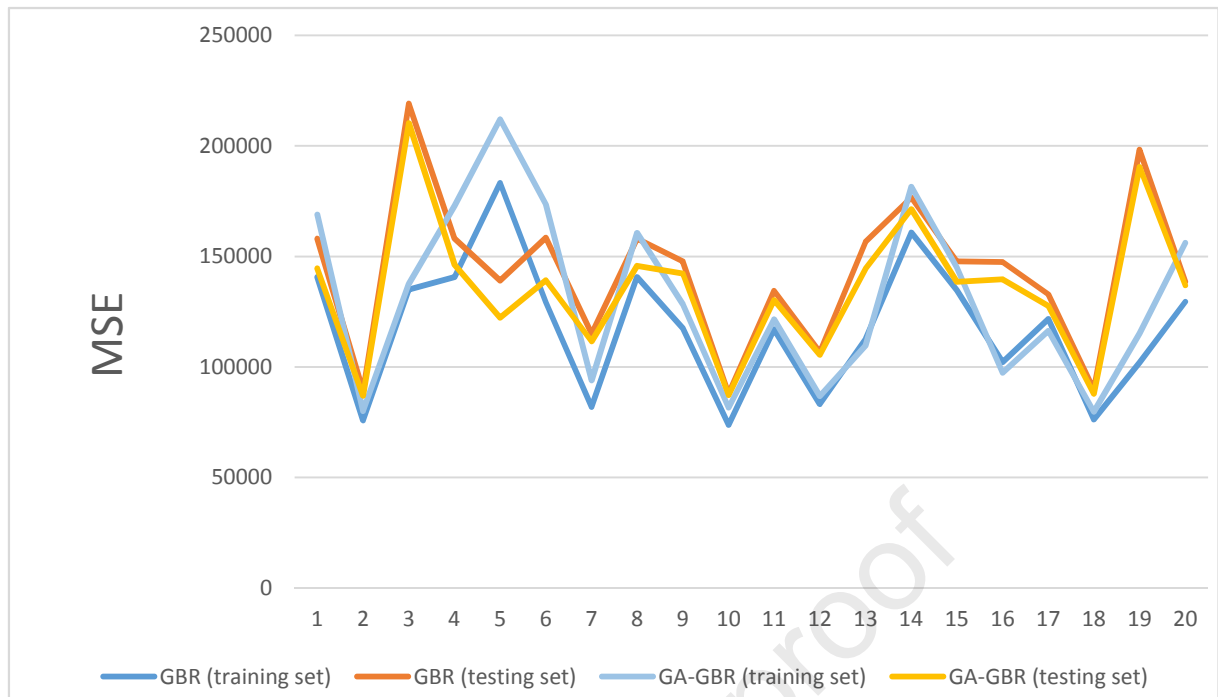


Figure 20: Comparison of MSE of two models.

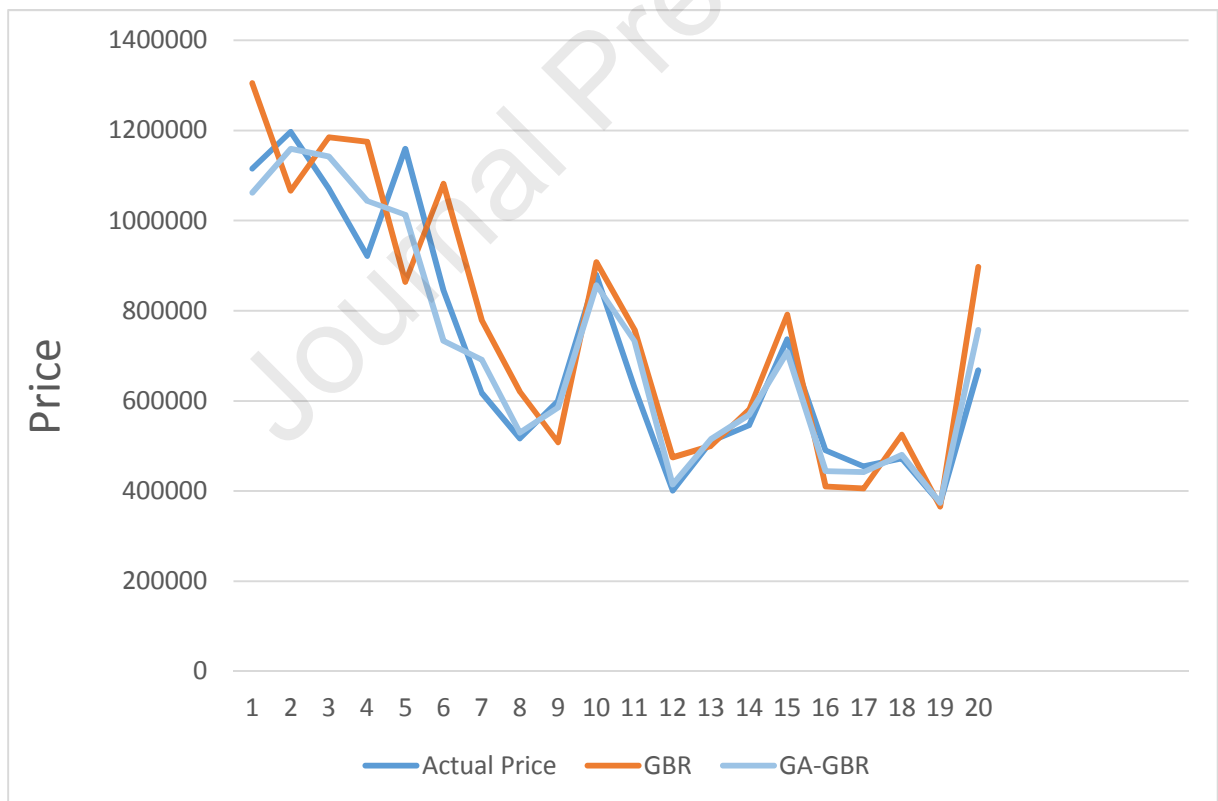


Figure 21: Comparison of predicted value and actual value by the two models.

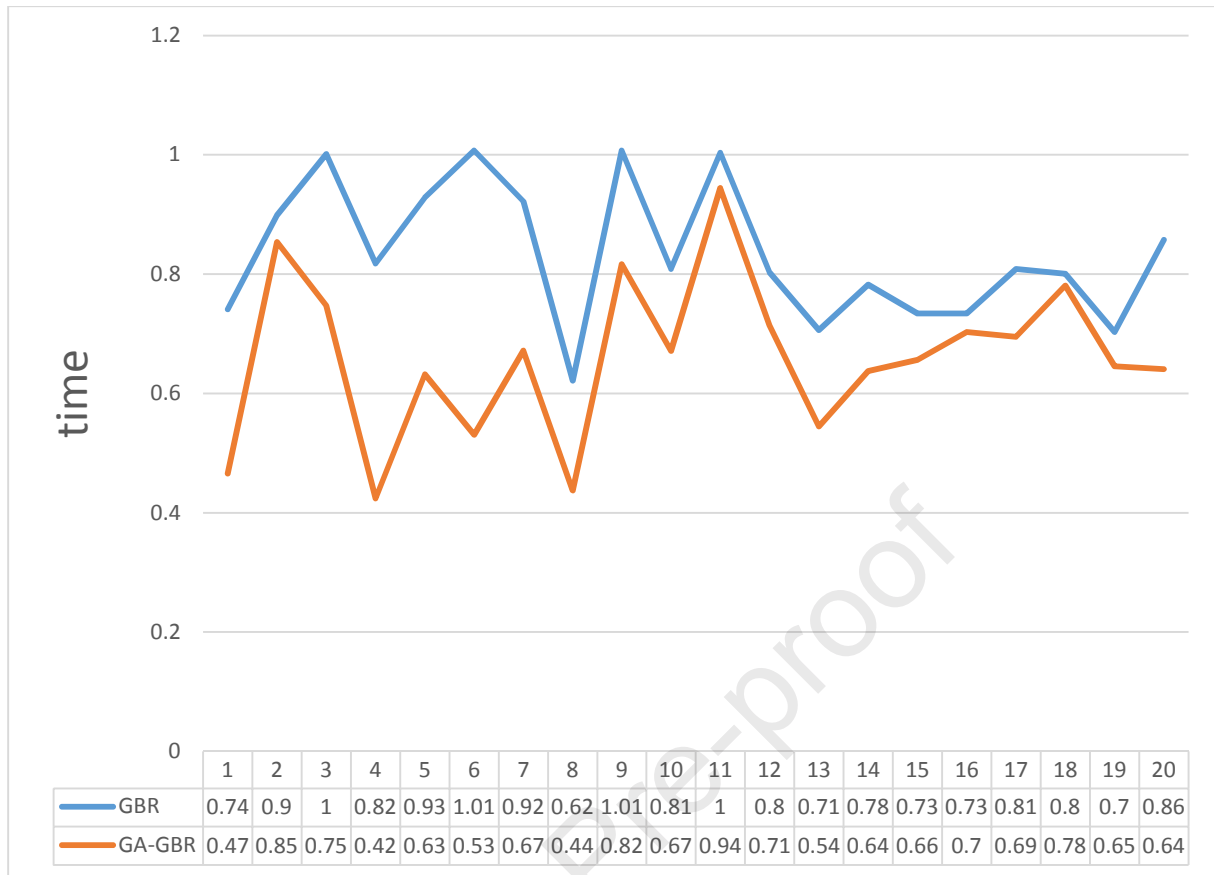


Figure 22: Comparison of training time of two models.

6.4 A case study for the BIM-Machine learning integration framework

In this section, an example of a duplex house Revit model for testing the BIM-Machine learning framework. An IFC file is firstly exported from the Revit model for the target information extraction. The IFC instance model has 1085 data instances including core building objects such as walls, doors, slabs, windows, roofs, stairs and furniture. Each object has its specific properties, such as fire rating, combustible and surface spread of flame, etc. After consulting with architectural designers, additional property information was added in the BIM model based on the extended IFC schema. For example, the low emitting ratio of the exterior wall of the duplex house was added as 85.3% based on the materials and structure of the walls. The syntactic and semantic validation are performed on Solibri Model Checker referring to the ISO 10303-11 [95], with no missing mandatory entities or incorrect data structure. Partial extracted information is displayed in Table 9.

Object name	Object type	Property name	Property nominal value	Adapted nominal value
Basic Wall: 300 22 wand HSBwand 12-140-12: 7326535	Single Value	Fire Rating	IFCLabel('60')	60 min
	Single Value	Load bearing	IFCBoolean(.F.)	Not load bearing
	Single Value	Low emitting material	IFCPositiveRatioMeasure(85.3)	85.3%

Living	Single Value	Gross planned area	IFCAreaMeasure(575.3)	575.3
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Patio	Single Value	Gross planned area	IFCAreaMeasure(144.4)	144.4
	Single Value	Is internal	IFCBoolean(.F.)	External element
Kitchen	Single Value	Gross planned area	IFCAreaMeasure(287.2)	287.2
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Bedroom 1	Single Value	Gross planned area	IFCAreaMeasure(203.2)	203.2
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Bedroom 2	Single Value	Gross planned area	IFCAreaMeasure(334.6)	334.6
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Bathroom 1	Single Value	Gross planned area	IFCAreaMeasure(108.3)	108.3
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Bathroom 2	Single Value	Gross planned area	IFCAreaMeasure(84.5)	84.5
	Single Value	Is internal	IFCBoolean(.T.)	Internal element
Bathroom 3	Single Value	Gross planned area	IFCAreaMeasure(88.2)	88.2
	Single Value	Is internal	IFCBoolean(.T.)	Internal element

Table 9: Partial extracted information item from the IFC instance of the duplex house.

The extracted building object and space information can be used for property valuation. For example, fire rating (IFCLabel('60')) means that the wall has a fire rating property with 60 minutes, which can be used for quantify the safety value. Low emitting material (IFCPositiveRatioMeasure(85.3)) means that the ratio of the weight of materials with low VOC emissions into indoor air to the total weight of the wall. The extracted space information (from IFCSpace instance) such as living area, patio, kitchen, bedroom and bathroom can be used for the evaluation of indoor air quality on property values.

Second, information extracted from the IFC file is processed in python 3.7 and tested with the GA-GBR model, with the prediction value of \$587091.02, testing mean absolute error (MAE) at \$59225.13, testing mean absolute percentage error (MAPE) at 10.08%. The MAPE, a measure of accuracy in a series value and usually expresses accuracy as a percentage, is calculated as:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \quad (7)$$

where A_t is the actual value and F_t is the forecast value.

7. Discussion

In the last two decades, data has become the power for industrial decision-making processes. While it is still in its infancy, big data and artificial intelligence have an important impact on valuation practice now. As digital technology advances fast, many stakeholders including

investors, banks, public authorities and real estate companies expect to benefit from the full potential of automated valuation services that can perform the valuation fast, improve the transparency of real estate market, reduce inaccuracies from reliance on human judgement and attendant bias [96]. In this context, the current valuation practice is facing challenges on several aspects including data collection and exchange, valuation method and the role of valuer. In the future, data collection is expected to become a more specialized profession or a more automated one, with the technological developments such as inspection with drones, the IoT, smart buildings and BIM.

Literature review indicated that there is potential for information contained in the AEC projects to be of use to property valuers at various stages of the property lifecycle, especially for building-related performance information such as energy cost, acoustics, air quality, safety and security, environmental and health impacts. This information cannot be solely acquired by a licensed property valuer through a building inspection, but requires access to and analysis of other reliable sources of information, for instance, information provided by facility managers and information founded in documentation of the design and planning process [97]. However, the value-relevant design information has not been widely utilized for property valuation due to the gap of knowledge and digital skills between property valuation and Architecture, Engineering, Construction (AEC) professionals. In the context of this research, the list of value-related building properties has been classified into 6 different types: information related to environmental quality, social and economic quality, functional quality, process quality, technical quality and site quality. Based on the information requirements, an IFC extension for property valuation has been developed. The extended IFC schema not only fills a knowledge gap that considering building entities and properties for property valuation including sustainability perspective, but also helps property valuation professionals who lack of BIM knowledge and digital skills to acquire value-specific information from AEC projects. In the long term, sustainability-related building features in the extended IFC schema can be an information enrichment for sustainable property valuation.

In respect to automated valuation methods, some advocators hold the opinion that the majority of valuations will be carried out by AI systems and AI enhanced AVMs will replace the valuer, considering the fact that AVMs are undertaking mass valuation work performed for banks. Others believe AI enhanced AVMs will change the valuation process and help the valuer in many aspects, but it will not replace some part of valuation where the valuer interprets data and makes judgements on the impact of that data on value [96]. This research presented a study on an AVM based on genetic algorithm optimized gradient descent regression ensemble machine learning for property valuation. To solve the conflicts between the accuracy of individual weak learners and the diversity among them, the GA-GBR model was proposed for property valuation. Since the complex GA-GBR model was related to 56 input variables, data dimensionality reductions strategies are considered for efficient execution. The experimental outcomes show several advantages of the proposed GA-GBR model. First, the base learners in gradient boosting ensembles are trained with sequential methods, in which the dependence between the base learners can be exploited. The base learners are then combined using a weighted sum method to convert weak learners to strong ensembles. The GBR ensemble can significantly reduce the bias without increasing the variance. Second, since there are 56 input attributes for property valuation, some of the non-important attributes can have significant influences on the whole model. There are numerous combinations of those attributes, and the different sequences of training them make the task even more complex. Genetic algorithm with its macro-exploration and global optimization capability, can help select the most useful input attributes and their associated best sequences without losing any accuracy. Lastly, genetic algorithm as a natural random search algorithm is

good at solving non-linearization problems. The non-linear relationships among the input attributes and property values usually make the prediction task hard to interpret. From this perspective, the GA-GBR model, which absorbs the advantages of both genetic algorithms and boosting ensembles, is more interpretable than the genetic algorithm optimized neural networks which are recognized as ‘black boxes’.

In summary, the use of innovative information BIM and artificial intelligence could be a revolution for data-driven applications in construction industry, but there is a significant gap in property valuation field. To fill the gap, this research studied the technology infusion of BIM and machine learning for automated property valuation, in which partial information requirements were extracted automatically from BIM models and ML-enhanced AVMs served as an automated data analysis engine. The BIM-ML framework not only helps property valuation professionals who normally are not familiar with BIM language to use value-specific information existed in AEC projects, but may also benefit the AEC professionals in terms of selecting the design alternatives that offering the long-term value to human beings. The real-time valuation results from the automated valuation engine (GA-GBR model) can be treated as constraints to optimize design, construction and operation strategies, which can be further developed as a decision-making tool for construction companies or property investors. In addition, the proposed BIM-ML framework has the potential to be applied to other applications such as building energy prediction, structure damage prediction and supply chain management etc. There are some deficiencies in this study. While the GA-GBR model is tested with improved prediction accuracy and predictive performance, it is only tested using data from US and the size of the sample is still not big enough. For the GA-GBR model, apart from genetic algorithms, there are many other heuristic algorithms such as particle swarm optimization (PSO), ant colony optimization (ACO) and firefly algorithm (FA) [98–100]. The integrations of these heuristic algorithms and hybrid machine learning models might add values for property valuation.

8. Conclusions and future work

To facilitate information exchange between AEC projects and property valuation and support automated property valuation, this paper presented a BIM and Machine learning integration framework for property valuation, which contains a fundamental database interpretation, an IFC-based information extraction and an advanced valuation model (GA-GBR) based on genetic algorithm optimized machine learning. Along with the developments of the three components in the proposed BIM-ML system, this research contributes to as follows:

First, this research contributes to the knowledge development of an extended IFC schema for property valuation. Among 95 variables reviewed in the literature, 62 of them are identified as relevant to this research, which is further used for the definition of IFC extension. The extended IFC schema (IFC4 – Addendum 2) included new entities (*IfcGeographicElement*, *IfcCivilElement* and *IfcSiteElement*) of building objects and new properties (*RecycledMaterial*, *AirInfiltration*, *SoundInsulation*, *Cost*, *LowEmittingMaterial*) of the property sets.

Second, while research on BIM information modelling and information extraction has focused on cost estimating, code compliance checking and clash detection, it is limited in supporting extracting value-related design information for property valuation. After the IFC extension development and validation, the required value-specific design information is extracted automatically from an IFC-based BIM instance model using an information extraction algorithm, which is developed based on the open-source BIM information extraction library-

IfcOpenShell. The extraction process is divided into eight main steps, during which the value of required variables are extracted from an IFC instance model automatically. After that, an example of the application of the developed IFC-based information extraction algorithm is provided.

Last, a genetic algorithm optimized machine learning model (GA-GBR) is firstly applied to automated property valuation. Since the model is comprehensive and includes a large number of variables, the genetic algorithm is designed not only for finding the tradeoff between the accuracy and diversity of the ensemble model, but also for data dimensionality reductions – find the best combination of input variables and the sequence of them. Compared to the prediction accuracy of previous research on AVMs which showing the maximum at 87.9%, in terms of R^2 , the proposed GA-GBR model in this research shows an advantage at 89.4%.

Several main findings have been identified in this research:

- In the context of climate change, sustainability-related building features such as internal natural light distribution, natural ventilation, ecological construction materials in the extended IFC schema can serve as a valuable information source for sustainable property valuation.
- Partial value-related design information can be extracted from BIM models in a more accurate and efficient way.
- The accuracy of property valuation could be improved by AVMs based on genetic algorithm optimized gradient descent regression ensemble machine learning model.
- The proposed GA-GBR model has the advantage of recognizing non-linear relationships between property variables, market factors and real estate values, efficiently dealing with the market changes and reducing inaccuracies that might come from human judgements and attendant biases.

While BIM brings many benefits, it cannot solve all the issues regards information modelling and information management in building and construction industry. In the future, technology infusion of BIM, Machine Learning and other emerging digital technologies (IOT, digital twin systems, block chain and cloud computing) is worth exploring for property valuation and construction industry.

Acknowledgement

Table 3 is based on the previous publication in EG-ICE 2020 conference.

References

- [1] W.Z. Taffese, Case-based reasoning and neural networks for real estate valuation, Proc. IASTED Int. Conf. Artif. Intell. Appl. AIA 2007. (2007) 84–89.
- [2] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, N. French, Real estate appraisal: a review of valuation methods, J. Prop. Invest. Financ. 21 (2003) 383–401. <https://doi.org/10.1108/14635780310483656>.
- [3] T. Su, H. Li, Data exchange analysis for property valuation on sustainability perspective, in: EG-ICE 2020 Work. Intell. Comput. Eng. Proc., Universitätsverlag der

- TU, Berlin, 2020: pp. 373–382. %0910.14279/depositonce-9977.
- [4] T. Dimopoulos, N. Bakas, An artificial intelligence algorithm analyzing 30 years of research in mass appraisals, *Rel. Int. J. Real Estate L. Plan.* 2 (2019) 10–27. <https://doi.org/https://doi.org/10.26262/reland.v2i0.6749>.
- [5] V. Kontrimas, A. Verikas, The mass appraisal of the real estate by computational intelligence, *Appl. Soft Comput. J.* 11 (2011) 443–448. <https://doi.org/10.1016/j.asoc.2009.12.003>.
- [6] B. Park, J. Kwon Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, *Expert Syst. Appl.* 42 (2015) 2928–2934. <https://doi.org/10.1016/j.eswa.2014.11.040>.
- [7] Z. Aladwan, M.S.S. Ahamad, Hedonic Pricing Model for Real Property Valuation via GIS - A Review, *Civ. Environ. Eng. Reports.* 29 (2019) 34–47. <https://doi.org/10.2478/ceer-2019-0022>.
- [8] J.J. Ahn, H.W. Byun, K.J. Oh, T.Y. Kim, Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting, *Expert Syst. Appl.* 39 (2012) 8369–8379. <https://doi.org/10.1016/j.eswa.2012.01.183>.
- [9] J.Y. Wu, Housing Price prediction Using Support Vector Regression, Master's Proj. (2017) 540. https://scholarworks.sjsu.edu/etd_projects/540/.
- [10] M. Graczyk, T. Lasota, B. Trawiński, K. Trawiński, Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal, *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 5991 LNAI (2010) 340–350. https://doi.org/10.1007/978-3-642-12101-2_35.
- [11] X.S. Liu, Z. Deng, T.L. Wang, Real estate appraisal system based on GIS and BP neural network, *Trans. Nonferrous Met. Soc. China (English Ed.)* 21 (2011) s626–s630. [https://doi.org/10.1016/S1003-6326\(12\)61652-5](https://doi.org/10.1016/S1003-6326(12)61652-5).
- [12] G.Z. Fan, S.E. Ong, H.C. Koh, Determinants of house price: A decision tree approach, *Urban Stud.* 43 (2006) 2301–2316. <https://doi.org/10.1080/00420980600990928>.
- [13] N. Miller, J. Spivey, A. Florance, Does green pay off?, *J. Real Estate Portf. Manag.* 14 (2008) 385–399.
- [14] R. Salvi, M., Horejajova, A. and Muri, *Minergie machtsich bezahlt*, (2008).
- [15] RICS, *Sustainability and commercial property valuation: RICS Professional Guidance*, Global, 2013.
- [16] W.L. Ventolo, *Fundamentals of real estate appraisal*, 12th ed., DF Institute, Inc, La Crosse, 2015.
- [17] C. Eastman, K. Liston, R. Sacks, K. Liston, *BIM Handbook* Paul Teicholz Rafael Sacks, 2008. <https://doi.org/2007029306>.
- [18] R. Marmo, F. Polverino, M. Nicolella, A. Tibaut, Building performance and maintenance information model based on IFC schema, *Autom. Constr.* 118 (2020). <https://doi.org/10.1016/j.autcon.2020.103275>.

- [19] M. Artus, C. Koch, *Modeling Physical Damages Using the Industry Foundation Classes – A Software Evaluation*, Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-51295-8_36.
- [20] M. Zhiliang, W. Zhenhua, S. Wu, L. Zhe, Application and extension of the IFC standard in construction cost estimating for tendering in China, *Autom. Constr.* 20 (2011) 196–204. <https://doi.org/10.1016/j.autcon.2010.09.017>.
- [21] UK BIM Alliance, *Information Management According to BS EN ISO 19650 – Guidance Part 1: Concepts*, 2nd ed., London, 2019. https://www.ukbimalliance.org/wp-content/uploads/2019/04/Information-Management-according-to-BS-EN-ISO-19650_-Guidance-Part-1_Concepts_2ndEdition.pdf (accessed March 9, 2021).
- [22] S.J. Wilkinson, J.R. Jupp, Exploring the value of BIM for corporate real estate, *J. Corp. Real Estate.* 18 (2016) 254–269. <https://doi.org/10.1108/JCRE-11-2015-0040>.
- [23] P.T. de La Paz, V.R. Perez-Sanchez, R.T. Mora-Garcia, J.C. Perez-Sanchez, Green premium evidence from climatic areas: A case in Southern Europe, Alicante (Spain), *Sustain.* 11 (2019). <https://doi.org/10.3390/su11030686>.
- [24] R.B. Abidoye, A.P.C. Chan, Critical review of hedonic pricing model application in property price appraisal: A case of Nigeria, *Int. J. Sustain. Built Environ.* 6 (2017) 250–259. <https://doi.org/10.1016/j.ijbsbe.2017.02.007>.
- [25] R.B. Abidoye, A.P.C. Chan, Artificial neural network in property valuation: application framework and research trend, *Prop. Manag.* 35 (2017) 554–571. <https://doi.org/10.1108/PM-06-2016-0027>.
- [26] D. Lorenz, T. Lützkendorf, Sustainability in property valuation: Theory and practice, *J. Prop. Invest. Financ.* 26 (2008) 482–521. <https://doi.org/10.1108/14635780810908361>.
- [27] T. Rogmans, M. Ghunaim, A framework for evaluating sustainability indicators in the real estate industry, *Ecol. Indic.* 66 (2016) 603–611. <https://doi.org/10.1016/j.ecolind.2016.01.058>.
- [28] A.S. Adair, J.N. Berry, W.S. McGreal, Hedonic modelling, housing submarkets and residential valuation, *J. Prop. Res.* 13 (1996) 67–83. <https://doi.org/10.1080/095999196368899>.
- [29] R.B. Abidoye, A.P.C. Chan, Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network, *Pacific Rim Prop. Res. J.* 24 (2018) 71–83. <https://doi.org/10.1080/14445921.2018.1436306>.
- [30] O.M. Lewis, J.A. Ware, D. Jenkins, A novel neural network technique for the valuation of residential property, *Neural Comput. Appl.* 5 (1997) 224–229. <https://doi.org/10.1007/BF01424227>.
- [31] P. Morano, F. Tajani, C.M. Torre, Artificial intelligence in property valuations An application of artificial neural networks to housing appraisal PIERLUIGI, *Adv. Environ. Sci. Energy Plan.* (2015) 23–29.
- [32] M.H. Rafiei, H. Adeli, A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units, *J. Constr. Eng. Manag.* 142 (2016) 04015066. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001047](https://doi.org/10.1061/(asce)co.1943-7862.0001047).

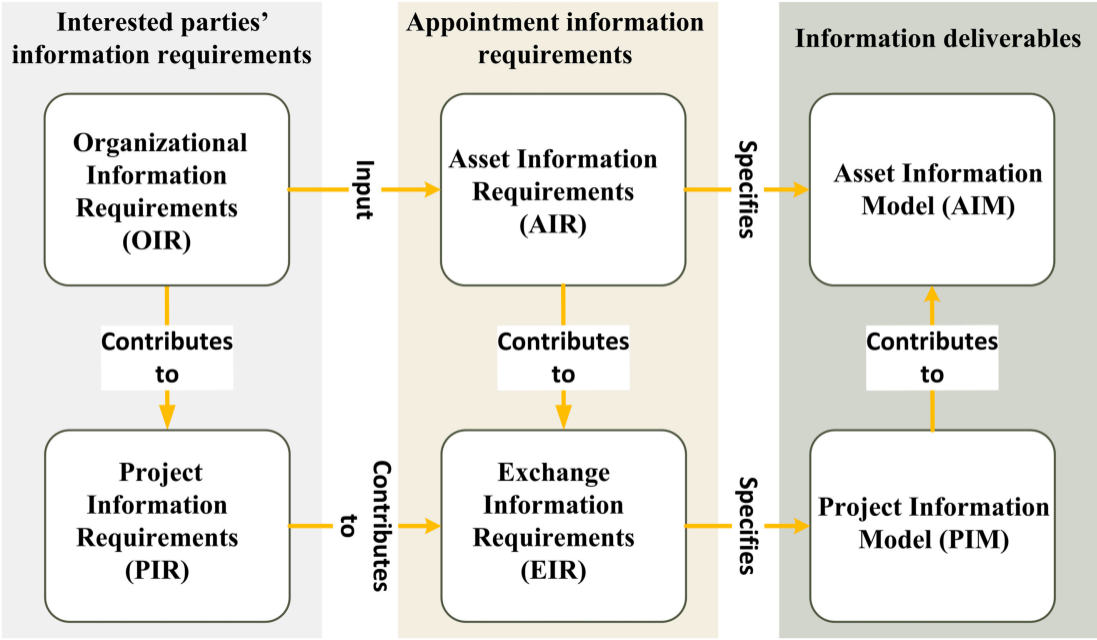
- [33] Benedetto Manganelli, Gianluigi De Mare, Antonio Nesticò, Using Genetic Algorithms in the Housing Market Analysis, in: *Comput. Sci. Its Appl. -- ICCSA 2015*, 2015: pp. 36–45. https://doi.org/https://link.springer.com/chapter/10.1007/978-3-319-21470-2_3.
- [34] V. Del Giudice, P. De Paola, F. Forte, Using genetic algorithms for real estate appraisals, *Buildings*. 7 (2017) 1–12. <https://doi.org/10.3390/buildings7020031>.
- [35] Y. Sun, Real estate evaluation model based on genetic algorithm optimized neural network, *Data Sci. J.* 18 (2019) 1–9. <https://doi.org/10.5334/dsj-2019-036>.
- [36] A. Borrmann, M. König, C. Koch, J. Beetz, *Building Information Modeling: Technology Foundations and Industry Practice*, 2018. https://doi.org/10.1007/978-3-319-92862-3_13.
- [37] R. Vanlande, C. Nicolle, C. Cruz, IFC and building lifecycle management, *Autom. Constr.* 18 (2008) 70–78. <https://doi.org/10.1016/j.autcon.2008.05.001>.
- [38] H. Lindblad, Study of the implementation process of BIM in construction projects: Analysis of the barriers limiting BIM adoption in the AEC industry, MSc Thesis. (2013) 64. <http://kth.diva-portal.org/smash/get/diva2:633132/FULLTEXT01>.
- [39] H.M. Faisal Shehzad, R. Binti Ibrahim, A.F. Yusof, K.A. Mohamed khaidzir, S. Shawkat, S. Ahmad, Recent developments of BIM adoption based on categorization, identification and factors: a systematic literature review, *Int. J. Constr. Manag.* 0 (2020) 1–13. <https://doi.org/10.1080/15623599.2020.1837719>.
- [40] C.M. Herr, T. Fischer, BIM adoption across the Chinese AEC industries: An extended BIM adoption model, *J. Comput. Des. Eng.* 6 (2019) 173–178. <https://doi.org/10.1016/j.jcde.2018.06.001>.
- [41] T. Volker, Industry Foundation Classes (IFC) - BIM Interoperability through a vendor-independent file format, A Bentley. (2011) 279–318.
- [42] R. Volk, J. Stengel, F. Schultmann, Building Information Modeling (BIM) for existing buildings - Literature review and future needs, *Autom. Constr.* 38 (2014) 109–127. <https://doi.org/10.1016/j.autcon.2013.10.023>.
- [43] BuildingSmart, *Information Delivery Manuals*, (2017).
- [44] R. Sacks, A. Kedar, A. Borrmann, L. Ma, I. Brilakis, P. Hüthwohl, S. Daum, U. Kattel, R. Yosef, T. Liebich, B.E. Barutcu, S. Muhic, SeeBridge as next generation bridge inspection: Overview, *Information Delivery Manual and Model View Definition*, *Autom. Constr.* 90 (2018) 134–145. <https://doi.org/10.1016/j.autcon.2018.02.033>.
- [45] Botton C and Forgues D, *Construction 4.0 The next revolution in the construction industry*, (2020). <https://www.canbim.com/articles/construction-4-0> (accessed March 9, 2021).
- [46] M. Winfield, Construction 4.0 and ISO 19650: A panacea for the digital revolution?, *Proc. Inst. Civ. Eng. Manag. Procure. Law.* 173 (2020) 175–181. <https://doi.org/10.1680/jmapl.19.00051>.
- [47] BSI, *BS EN ISO 19650-1:2018 Organization of information about construction works — Information management using building information modelling — Part 1: Concepts*

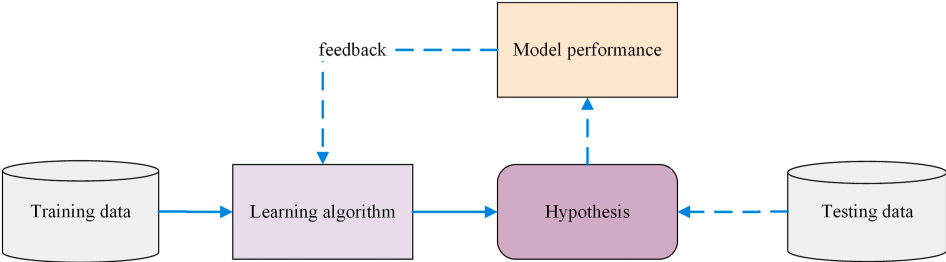
- and Principles, (2018).
https://shop.bsigroup.com/ProductDetail?pid=000000000030333754&_ga=2.135734898.2106602755.1615403568-1350352573.1614959933 (accessed March 10, 2021).
- [48] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
- [49] C.M. Bishop, *Pattern recognition and machine learning*, Springer Science & Business Media, 2006.
- [50] Yang Yun, *Temporal data mining via unsupervised ensemble learning*, Elsevier, 2016.
- [51] Zhou Zhi-Hua, *Ensemble Learning*, (2012).
<https://cs.nju.edu.cn/zhoush/zhoush.files/publication/springerEBR09.pdf>.
- [52] E. Sevinç, A. Coşar, An evolutionary genetic algorithm for optimization of distributed database queries, *Comput. J.* 54 (2011) 717–725.
<https://doi.org/10.1093/comjnl/bxp130>.
- [53] C. Wong, F., & Tan, *Hybrid neural, genetic and fuzzy systems*, Wiley, New York, 1994.
- [54] E. Hofmann, M. Rüsçh, Industry 4.0 and the current status as well as future prospects on logistics, *Comput. Ind.* 89 (2017) 23–34.
<https://doi.org/10.1016/j.compind.2017.04.002>.
- [55] V. Villani, F. Pini, F. Leali, C. Secchi, Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications, *Mechatronics*. 55 (2018) 248–266. <https://doi.org/10.1016/j.mechatronics.2018.02.009>.
- [56] H. Boyes, B. Hallaq, J. Cunningham, T. Watson, The industrial internet of things (IIoT): An analysis framework, *Comput. Ind.* 101 (2018) 1–12.
<https://doi.org/10.1016/j.compind.2018.04.015>.
- [57] F. Craveiro, J.P. Duarte, H. Bartolo, P.J. Bartolo, Additive manufacturing as an enabling technology for digital construction: A perspective on Construction 4.0, *Autom. Constr.* 103 (2019) 251–267. <https://doi.org/10.1016/j.autcon.2019.03.011>.
- [58] E. Forcael, I. Ferrari, A. Opazo-Vega, J.A. Pulido-Arcas, Construction 4.0: A Literature Review, *Sustainability*. 12 (2020) 9755. <https://doi.org/10.3390/su12229755>.
- [59] O. Noran, D. Romero, S. Burchiu, Exploring the Path Towards Construction 4.0: Collaborative Networks and Enterprise Architecture Views, in: *IFIP Adv. Inf. Commun. Technol.*, Springer, 2020: pp. 547–556. https://doi.org/10.1007/978-3-030-57997-5_63.
- [60] S. Rastogi, CONSTRUCTION 4.0: THE 4 th GENERATION REVOLUTION, in: *Indian Lean Constr. Conf.*, 2017: pp. 1–11.
https://www.researchgate.net/publication/331131645_CONSTRUCTION_40_THE_4_th_GENERATION_REVOLUTION (accessed March 5, 2021).
- [61] J. Matthews, P.E.D. Love, S. Heinemann, R. Chandler, C. Rumsey, O. Olatunj, Real time progress management: Re-engineering processes for cloud-based BIM in construction, *Autom. Constr.* 58 (2015) 38–47.
<https://doi.org/10.1016/j.autcon.2015.07.004>.

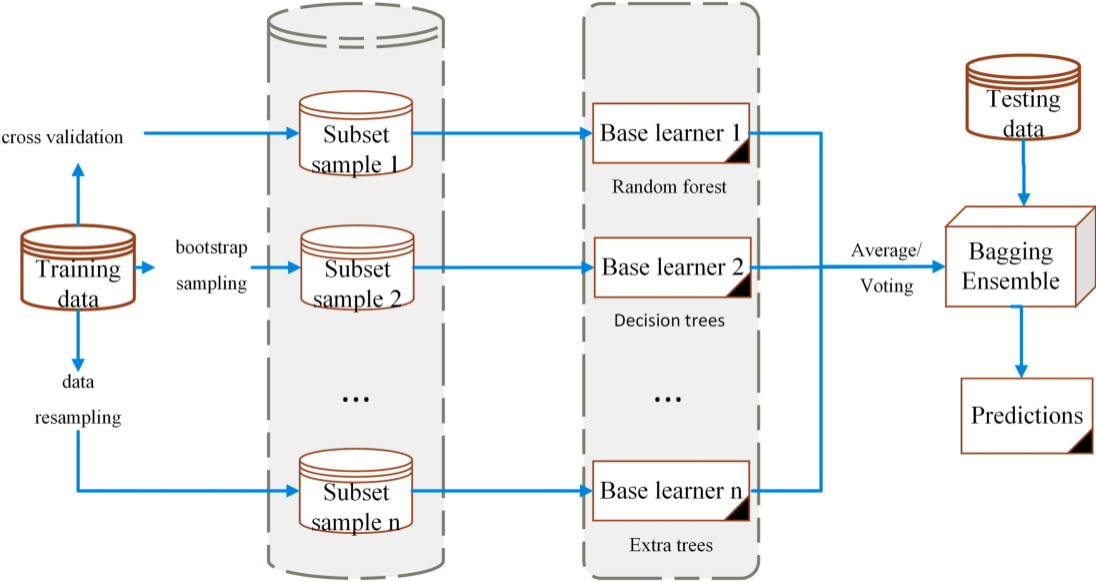
- [62] M.N. Birje, P.S. Challagidad, R.H. Goudar, M.T. Tapale, Cloud computing review: Concepts, technology, challenges and security, *Int. J. Cloud Comput.* 6 (2017) 32–57. <https://doi.org/10.1504/IJCC.2017.083905>.
- [63] K. Chen, F. Xue, The renaissance of augmented reality in construction: history, present status and future directions, *Smart Sustain. Built Environ.* (2020). <https://doi.org/10.1108/SASBE-08-2020-0124>.
- [64] J. Ratajczak, M. Riedl, D. Matt, BIM-based and AR Application Combined with Location-Based Management System for the Improvement of the Construction Performance, *Buildings.* 9 (2019) 118. <https://doi.org/10.3390/buildings9050118>.
- [65] Y.-C. Chen, H.-L. Chi, S.-C. Kang, S.-H. Hsieh, Attention-Based User Interface Design for a Tele-Operated Crane, *J. Comput. Civ. Eng.* 30 (2016) 04015030. [https://doi.org/10.1061/\(asce\)cp.1943-5487.0000489](https://doi.org/10.1061/(asce)cp.1943-5487.0000489).
- [66] H. Boyes, *A security framework for cyber-physical systems*, Coventry, 2017.
- [67] A. Bonci, A. Carbonari, A. Cucchiarelli, L. Messi, M. Pirani, M. Vaccarini, A cyber-physical system approach for building efficiency monitoring, *Autom. Constr.* 102 (2019) 68–85. <https://doi.org/10.1016/j.autcon.2019.02.010>.
- [68] Y. Lei, Y. Rao, J. Wu, C.H. Lin, BIM based cyber-physical systems for intelligent disaster prevention, *J. Ind. Inf. Integr.* 20 (2020) 100171. <https://doi.org/10.1016/j.jii.2020.100171>.
- [69] M. Bilal, L.O. Oyedele, J. Qadir, K. Munir, S.O. Ajayi, O.O. Akinade, H.A. Owolabi, H.A. Alaka, M. Pasha, Big Data in the construction industry: A review of present status, opportunities, and future trends, *Adv. Eng. Informatics.* 30 (2016) 500–521. <https://doi.org/10.1016/j.aei.2016.07.001>.
- [70] H. Yan, N. Yang, Y. Peng, Y. Ren, Data mining in the construction industry: Present status, opportunities, and future trends, *Autom. Constr.* 119 (2020) 103331. <https://doi.org/10.1016/j.autcon.2020.103331>.
- [71] Y. Cheng, W. Der Yu, Q. Li, GA-based multi-level association rule mining approach for defect analysis in the construction industry, *Autom. Constr.* 51 (2015) 78–91. <https://doi.org/10.1016/j.autcon.2014.12.016>.
- [72] W. Der Yu, H.W. Lin, A VaFALCON neuro-fuzzy system for mining of incomplete construction databases, *Autom. Constr.* 15 (2006) 20–32. <https://doi.org/10.1016/j.autcon.2005.01.006>.
- [73] A. Tashakkori, C. Teddlie, B.K. Nastasi, J.H. Hitchcock, L.M. Brown, *An Inclusive Framework for Conceptualizing Mixed Methods Design Typologies: Moving Toward Fully Integrated Synergistic Research Models*, 2015. <https://doi.org/10.4135/9781506335193.n13>.
- [74] T.A. Saunders Mark, Lewis Phiip, *Research Methods for Business Students*, 7th ed., Pearson Education Limited, England, 2016. <https://doi.org/10.1108/qmr.2000.3.4.215.2>.
- [75] K. Peffers, T. Tuunanen, M.A. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, *J. Manag. Inf. Syst.* 24 (2007) 45–77.

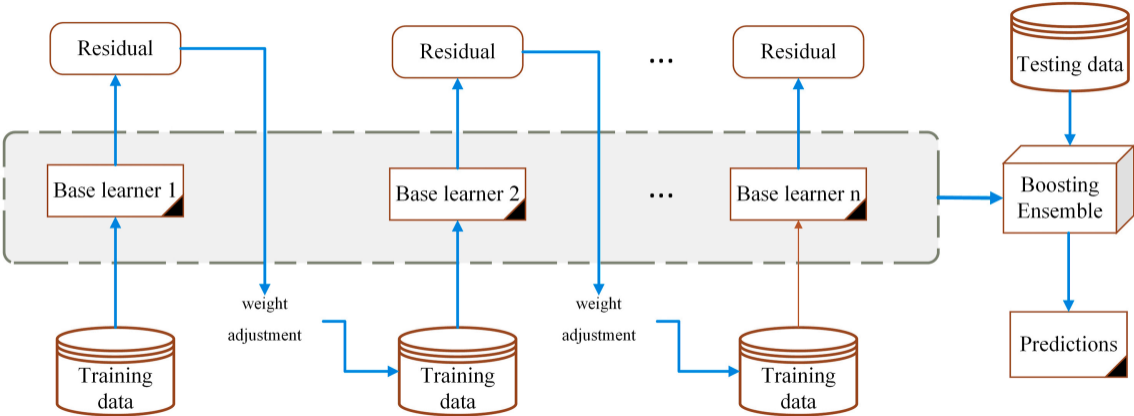
- <https://doi.org/10.2753/MIS0742-1222240302>.
- [76] L. Zhang, N.M. El-Gohary, Human-Centered Value Analysis: Building-Value Aggregation Based on Human Values and Building-System Integration, *J. Constr. Eng. Manag.* 143 (2017) 04017040. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001252](https://doi.org/10.1061/(asce)co.1943-7862.0001252).
- [77] H. R, *The structure of value*, Southern Illinois University Press, Carbondale, 1967.
- [78] R. Volk, J. Stengel, F. Schultmann, Building Information Modeling (BIM) for existing buildings - Literature review and future needs, *Autom. Constr.* 38 (2014) 109–127. <https://doi.org/10.1016/j.autcon.2013.10.023>.
- [79] J.C.P. Cheng, Y. Deng, Linking BIM and GIS Models in Infrastructure by Example of IFC and CityGML, *Comput. Civ. Eng.* 2015. (2015) 667–674. <http://ascelibrary.org/doi/10.1061/9780784479247.083>.
- [80] J. Amann, A. Borrmann, F. Hegemann, J. Jubierre, M. Flurl, C. Koch, M. König, A Refined Product Model for Shield Tunnels Based on a Generalized Approach for Alignment Representation, *Proc. ICCBEI 2013*. (2013) 8.
- [81] A. Borrmann, S. Muhic, J. Hyvärinen, T. Chipman, S. Jaud, C. Castaing, C. Dumoulin, T. Liebich, L. Mol, The IFC-Bridge project – Extending the IFC standard to enable high-quality exchange of bridge information models, *Proc. 2019 Eur. Conf. Comput. Constr.* 1 (2019) 377–386. <https://doi.org/10.35490/ec3.2019.193>.
- [82] A. Jaly-zada, C. Koch, W. Tizani, Ifc extension for design change management, *Proc. 32nd CIB W78 Conf. 2015, 27th-29th Oct. 2015, Eindhoven, Netherlands*. (2015) 327–335.
- [83] S.H. Lee, B.G. Kim, IFC extension for road structures and digital modeling, *Procedia Eng.* 14 (2011) 1037–1042. <https://doi.org/10.1016/j.proeng.2011.07.130>.
- [84] ISO 16739-1:2018 Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries, (2018).
- [85] US Green Building Council, LEED reference guide for design and construction, (2009).
- [86] R. and A.E. American Society of Heating, ASHRAR handbook - fundamentals, (n.d.).
- [87] Audimute, NRC ratings: what is a noise reduction coefficient?, (2018). <https://www.audimute.com/nrc-ratings-noise-reduction-coefficient>.
- [88] IfcOpenShell, IfcOpenShell-python: A python module based on IfcOpenShell, (2018). <http://ifcopenshell.org/python>.
- [89] Altincay Hakan, Optimal resampling and classifier prototype selection in classifier ensembles using genetic algorithms, in: *Pattern Anal. Appl.*, 2004: pp. 285–295. <https://doi.org/10.1007/BF02683994>.
- [90] Zhou Zhi-Hua, *Ensemble learning: foundations and algorithms*, Chapman & Hall/CRC, New York, 2012.
- [91] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, *Front. Neurobot.* 7 (2013). <https://doi.org/10.3389/fnbot.2013.00021>.

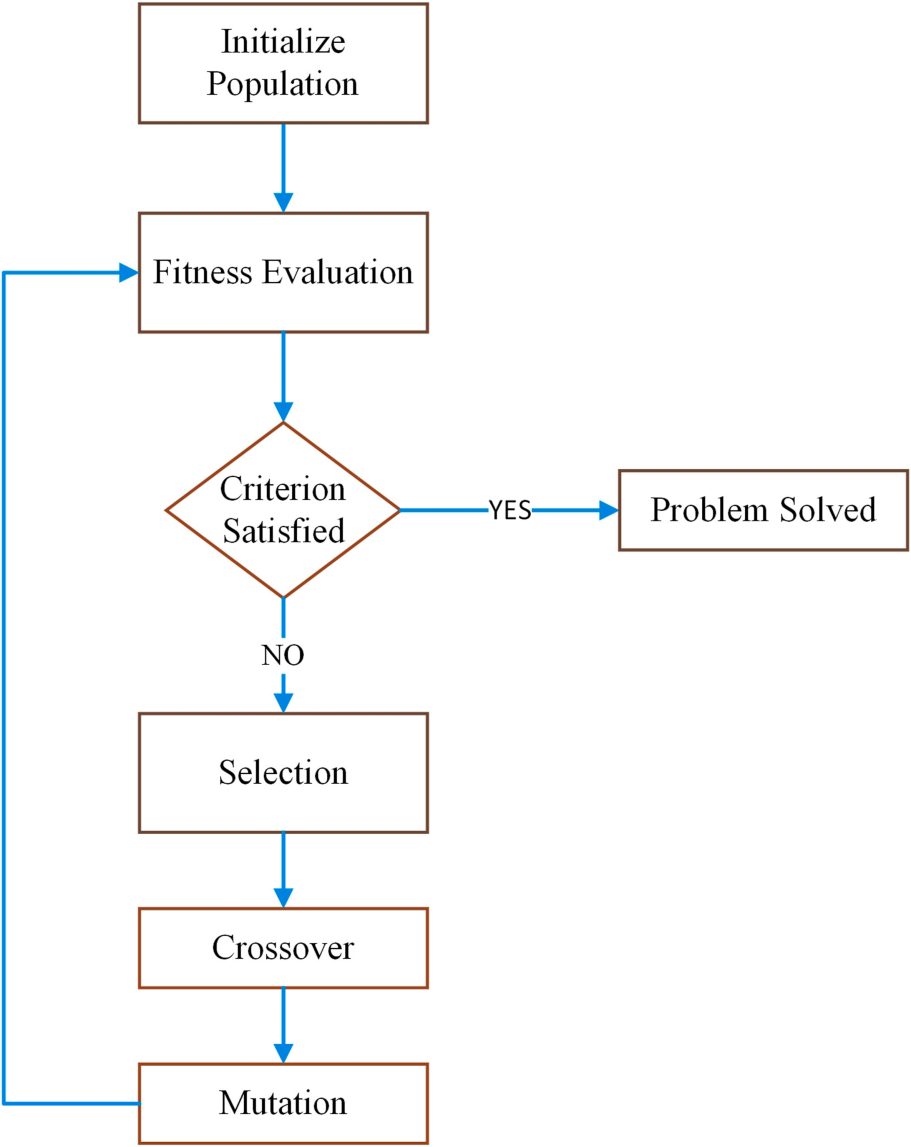
- [92] Zhou Zhi-Hua, Genetic algorithm based selective neural network ensemble, in: IJCAI'01: Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2, 2001: pp. 797–802.
- [93] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, V.B.S. Prasath, Choosing mutation and crossover ratios for genetic algorithms-a review with a new dynamic approach, *Inf. 10* (2019). <https://doi.org/10.3390/info10120390>.
- [94] W.J. McCluskey, M. McCord, P.T. Davis, M. Haran, D. McIlhatton, Prediction accuracy in mass appraisal: A comparison of modern approaches, *J. Prop. Res.* 30 (2013) 239–265. <https://doi.org/10.1080/09599916.2013.781204>.
- [95] ISO, Industrial automation systems and integration — Product data representation and exchange — Part 11: Description methods: The EXPRESS language reference manual, (2014).
- [96] RICS, the Future of Valuations, 2017.
- [97] T. Lützkendorf, D. Lorenz, Capturing sustainability-related information for property valuation, *Build. Res. Inf.* 39 (2011) 256–273. <https://doi.org/10.1080/09613218.2011.563929>.
- [98] W. Deng, J. Xu, H. Zhao, An Improved Ant Colony Optimization Algorithm Based on Hybrid Strategies for Scheduling Problem, *IEEE Access.* 7 (2019) 20281–20292. <https://doi.org/10.1109/ACCESS.2019.2897580>.
- [99] D. Tian, Z. Shi, MPSO: Modified particle swarm optimization and its applications, *Swarm Evol. Comput.* 41 (2018) 49–68. <https://doi.org/10.1016/j.swevo.2018.01.011>.
- [100] L. Zhang, L. Liu, X.S. Yang, Y. Dai, A novel hybrid firefly algorithm for global optimization, *PLoS One.* 11 (2016) 1–17. <https://doi.org/10.1371/journal.pone.0163230>.











Initialize
Population

Fitness Evaluation

Criterion
Satisfied

YES

Problem Solved

NO

Selection

Crossover

Mutation

SENSORS

- Energy
- Clash detection
- Life cycle monitoring
- Predictability
- Jobsite safety
- Traceability
- Reactivity

DRONES

- Mapping
- Moving light materials
- Security
- Track the progress of construction

ROBOTS

- Real time-Autonomy productivity
- Full transparency on data reporting
- Mass customization

AUTO VEHICLES

- Flow optimization
- Increased security
- Lower costs

Digitalization of the construction industry +

CONSTRUCTION 4.0

+ **Industrialization of the construction processes**

Wisdom

Knowledge

Information

Data

BIM

- Nd modelling
- IFC/IDM/MVD
- EIR/PIM/AIM
- CDE
- Lifecycle
- Scheduling
- Simulation

ML

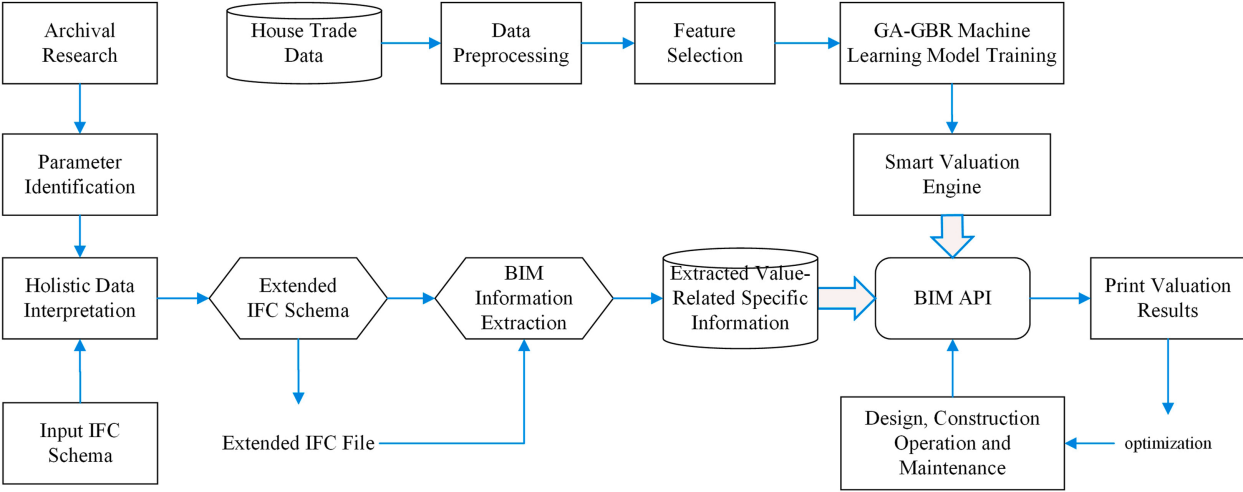
- Big data analysis
- Real-time decision making
- Data-driven prediction
- Knowledge representation
- Supervised & Unsupervised

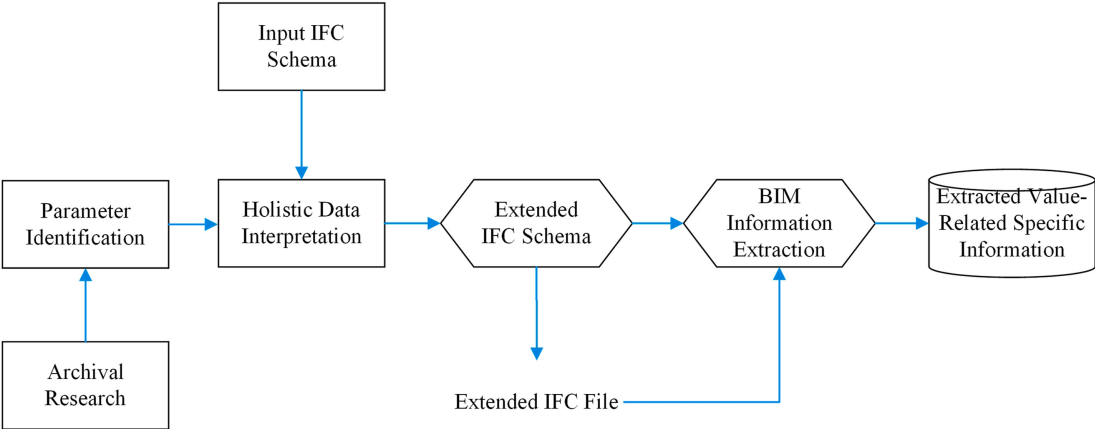
IoT

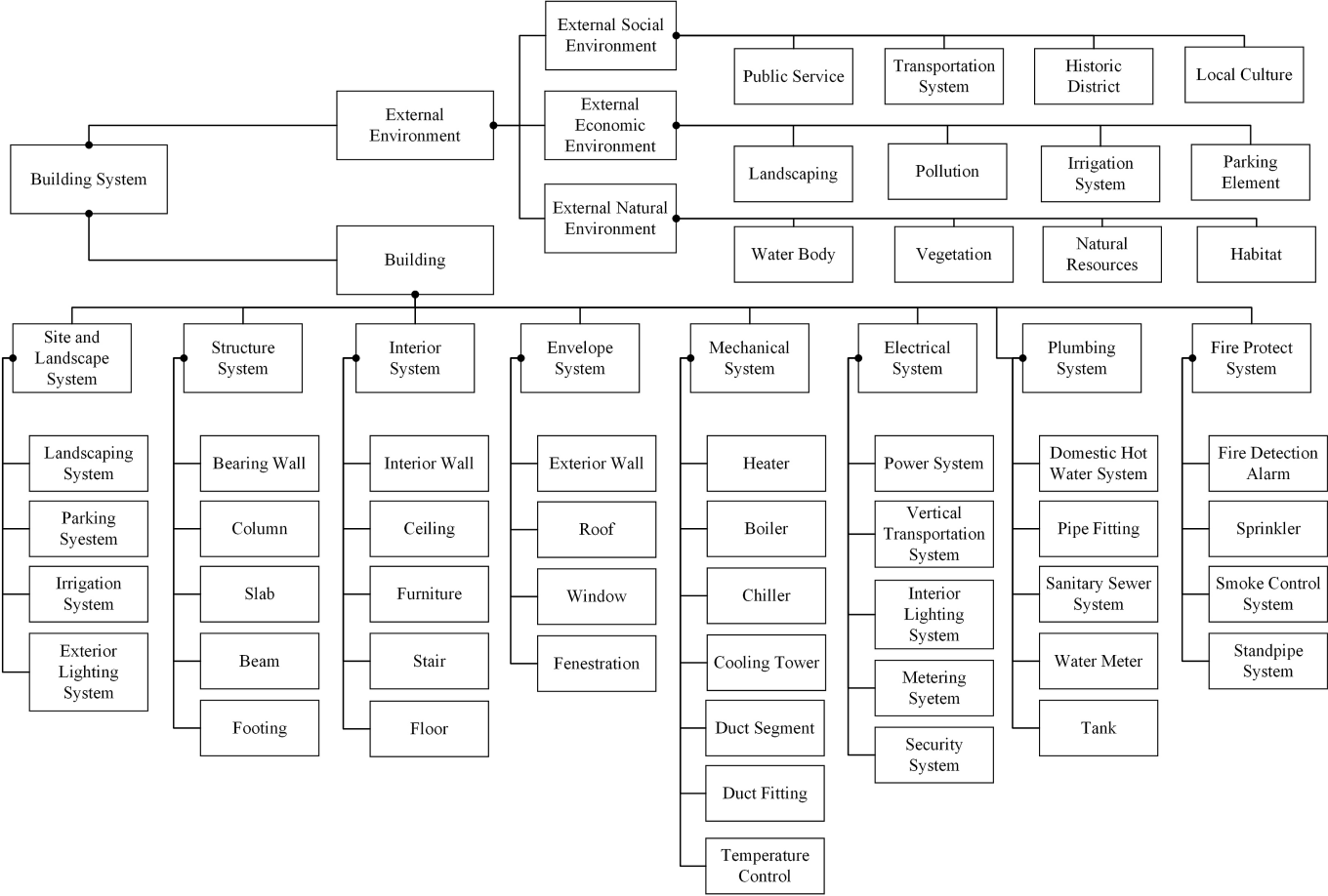
- Real time data capture
- Internet-object communication via low power radio
- Reduced wastes
- Object tagging
- Safety

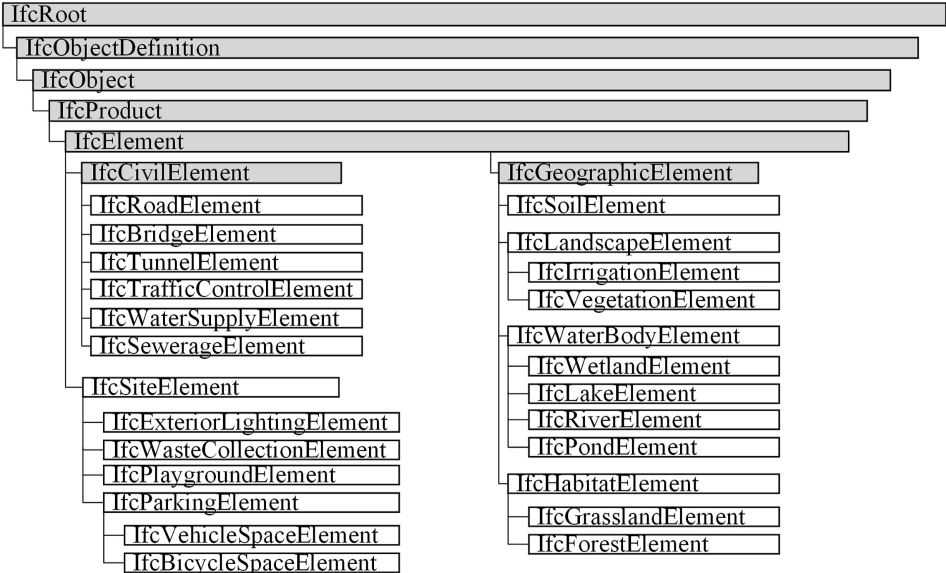
Laser Scanning

- Measuring
- Building survey
- Process tracking
- Global positioning system



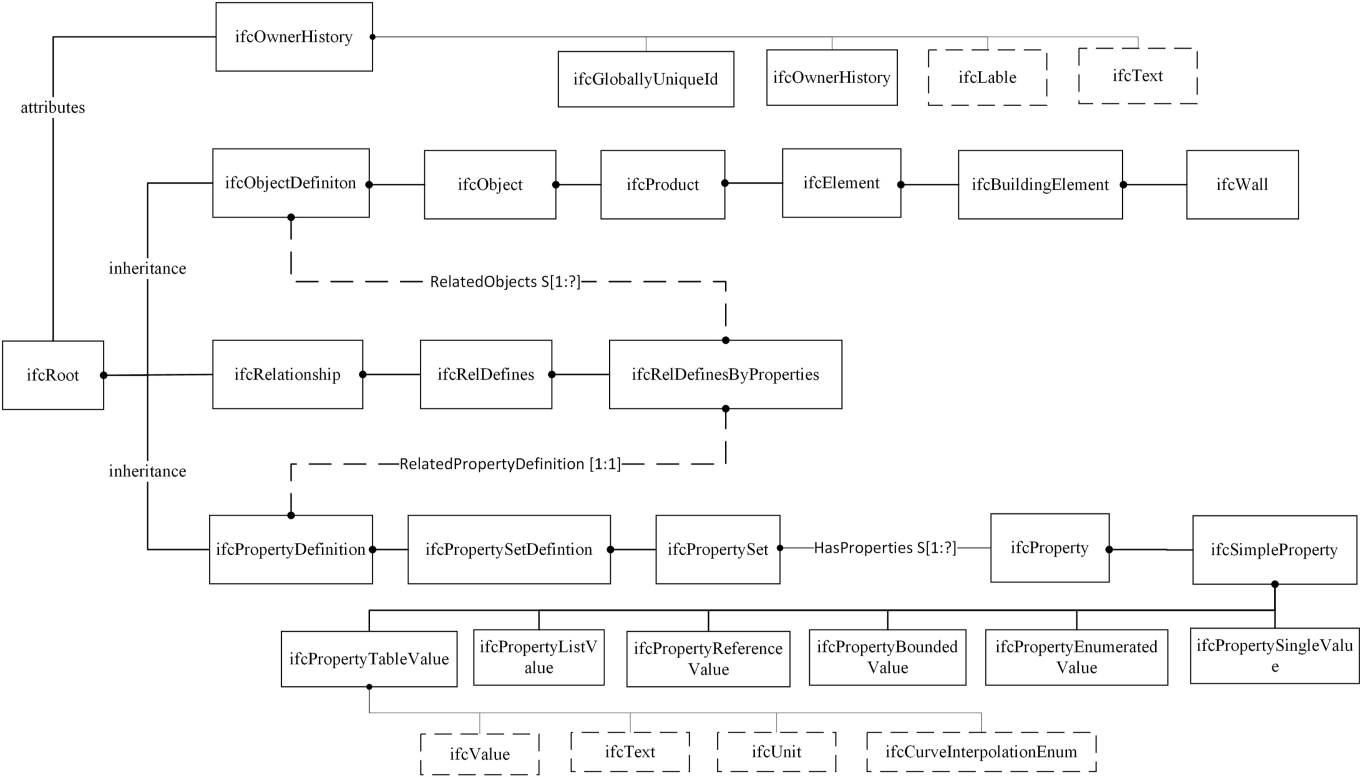


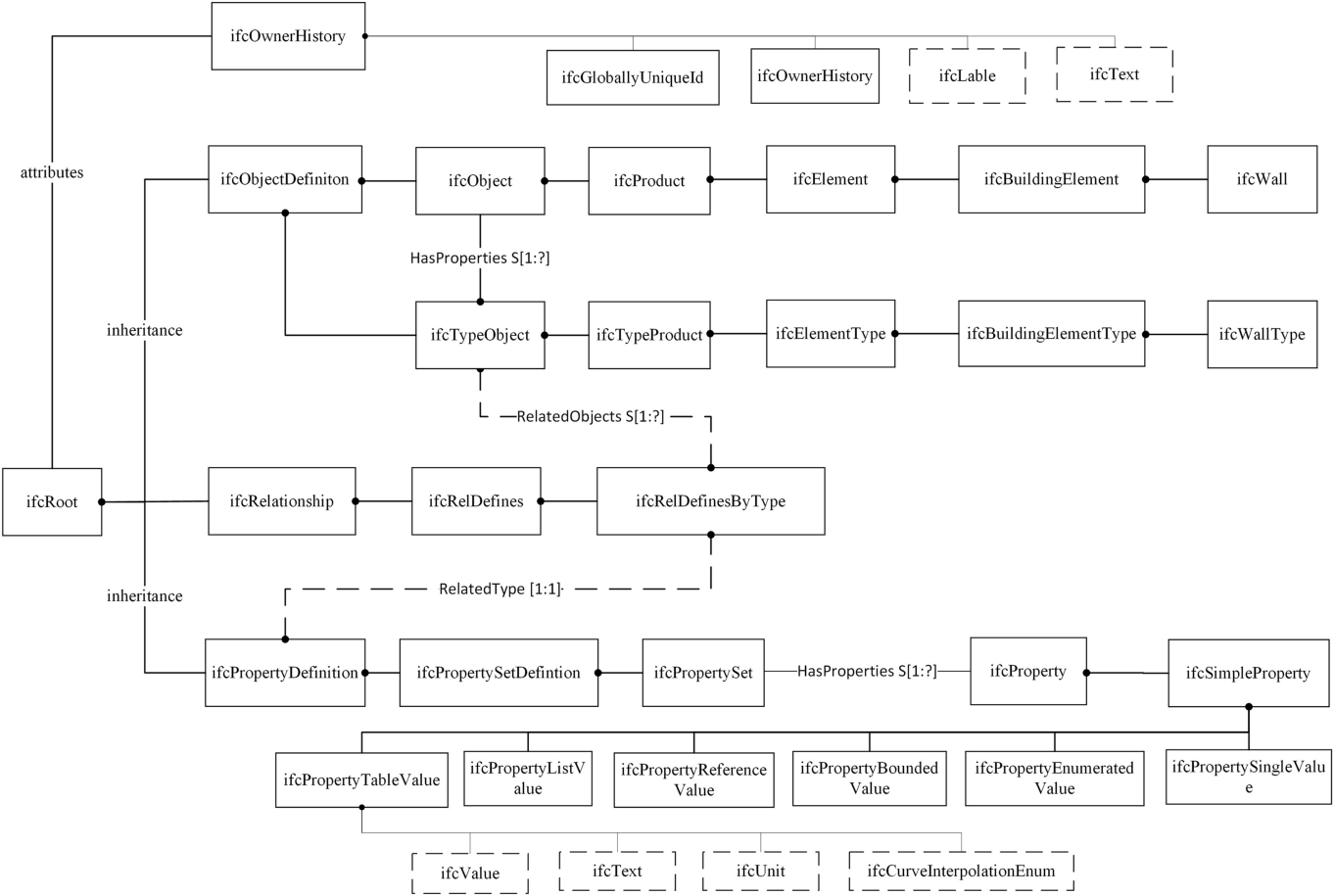


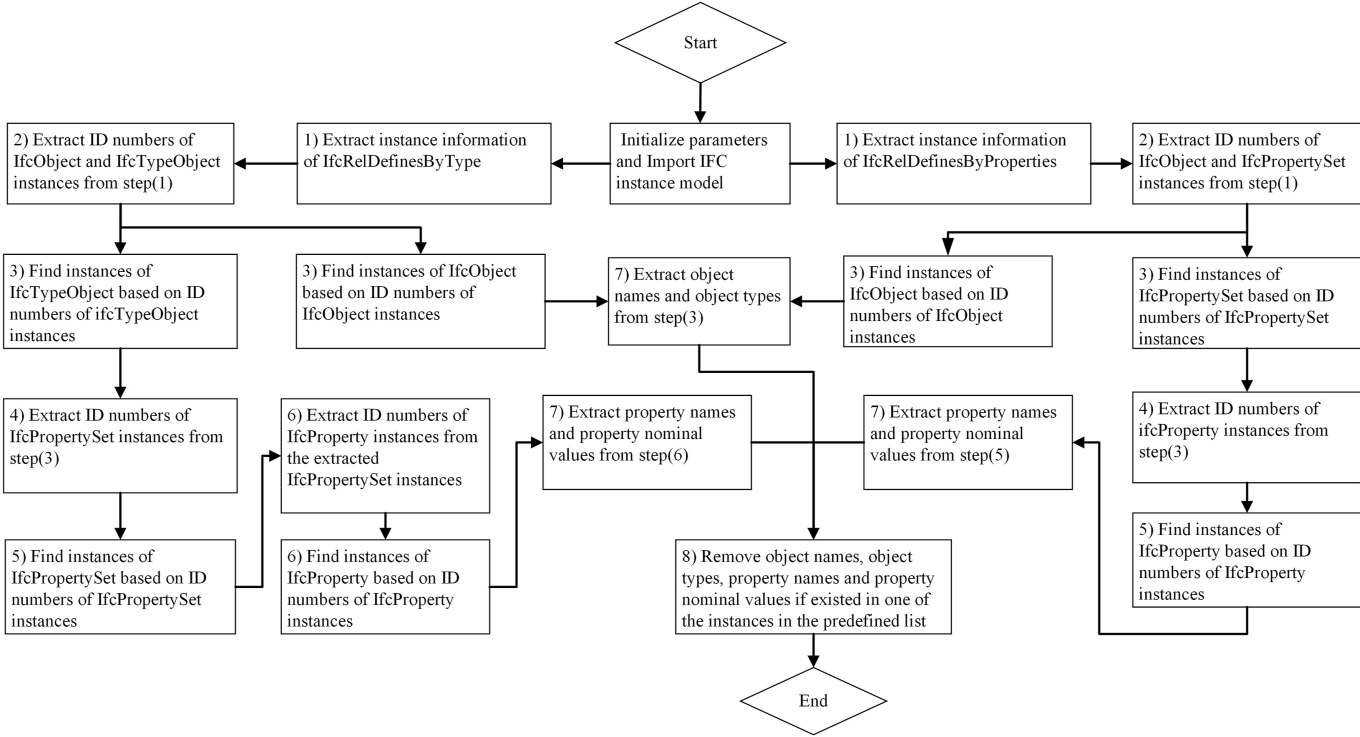


Existing entity in IFC4-Addendum2

Proposed entity to add







(1) extraction based on direct relationships

{An IfcProperty instance}:

#557=IfcPropertySingleValue('FireRating', \$[IfcLabel('60'), \$)

Property Name

Property Nominal Value

{An IfcWall instance}:

#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw', #41, 'Basic Wall:300 22 wand HSBwand 12-140-12:7326535', \$, 'Basic Wall:300_22_wand_HSBwand_12-140-12:7011920', #187, #213, '7326535')

Object Name

{An IfcPropetSet instance}:

#558=IfcPropertySet('1MhDg8jdf98g77kSIHoeHw', #41, 'Pset_WallCommon', \$, (#553, #554, #555, #556, #557))

{An IfcRelDefinesByPropeties instance}:

#563=IfcRelDefinesByProperties('33prtgrfUP4ZeaQc3XIaZwF', #41, \$, \$, (#215), #558)

*RelatedProperty
Definition*

RelatedObjects

(2) extraction based on indirect relationships

{An IfcProperty instance}:

Property Name

Property Nominal Value

#554=IfcPropertySingleValue('LoadBearing', \$[IfcBoolean(F), \$)

{An IfcPropertySet instance}:

#558=IfcPropertySet('1MhDg8jdf98g77kSIHoeHw', #41, 'Pset_WallCommon', \$, (#553, #554, #555, #556, #557))

{An IfcWallType instance}:

#551=IfcWallType('1iAXenQ2f67eN2WTRS\$r6B', #41, 'Basic Wall:300_22_wand_HSBwand_12-140-12', \$, #558, \$, \$, '7011920', \$, .STANDARD.)

{An IfcWall instance}:

#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw', #41, 'Basic Wall:300 22 wand HSBwand 12-140-12:7326535', \$, 'Basic Wall:300_22_wand_HSBwand_12-140-12:7011920', #187, #213, '7326535')

Object Name

{An IfcRelDefinesByType instance}:

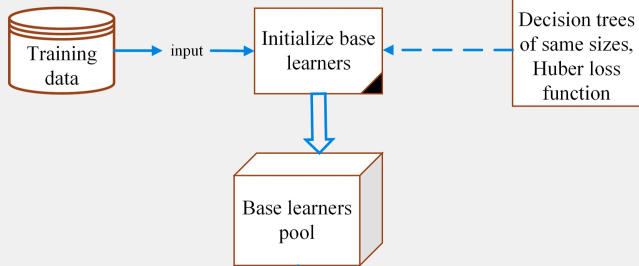
#442681=IfcRelDefinesByType('2iM7Gcqlv00x8z3DIHqa85', #41, \$, \$,

(#215, #654, #1048, #1470, #1895, #2402, #2838, #3129, #3407, #3451, #3495, #3539, #3596, #3965, #5125, #5345, #5478, #5711, #5844, #6077, #6221, #6441, #6666, #6710, #6754, #6798, #8090, #8446, #8811, #42471, #42654, #43151, #413124), #551)

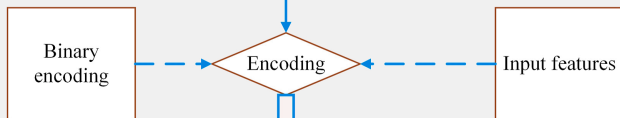
RelatedObjects

RelatingType

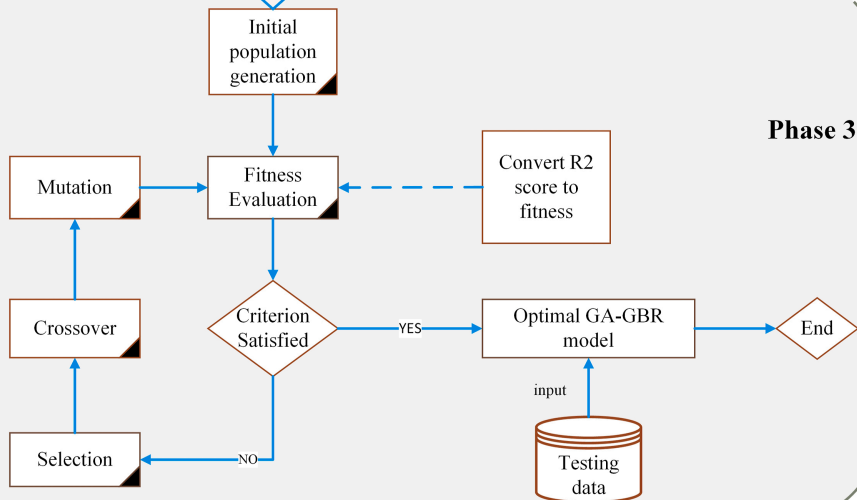
Phase I



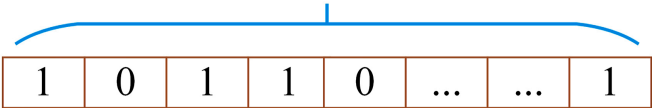
Phase 2



Phase 3



56 input features



A

B

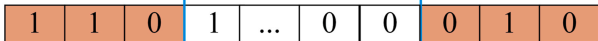
...

C

Chromosome 1



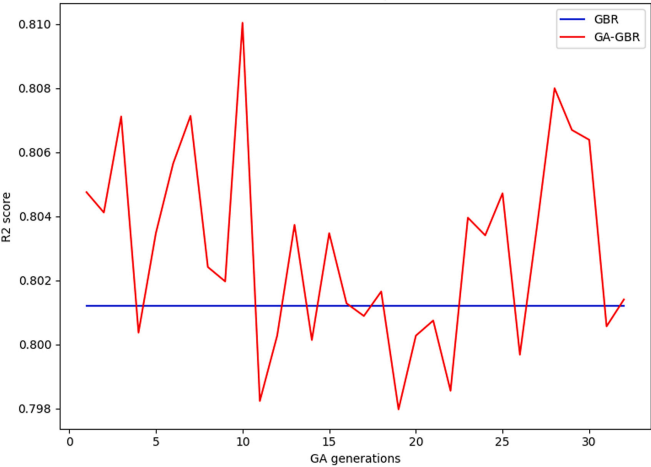
Chromosome 2



Offspring



Genetic search process



Author contributions

Use this form to specify the contribution of each author of your manuscript. A distinction is made between five types of contributions: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

For each author of your manuscript, please indicate the types of contributions the author has made. An author may have made more than one type of contribution. Optionally, for each contribution type, you may specify the contribution of an author in more detail by providing a one-sentence statement in which the contribution is summarized. In the case of an author who contributed to performing the analysis, the author's contribution for instance could be specified in more detail as 'Performed the computer simulations', 'Performed the statistical analysis', or 'Performed the text mining analysis'.

If an author has made a contribution that is not covered by the five pre-defined contribution types, then please choose 'Other contribution' and provide a one-sentence statement summarizing the author's contribution.

Manuscript title: **A BIM and Machine Learning Integration Framework for Automated Property Valuation**

Author 1: Tengxiang Su

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 2: Haijiang Li

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
supervision

Author 3: Yi An

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 4: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 5: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 6: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 7: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 8: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 9: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

Author 10: Enter author name

- Conceived and designed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Collected the data**
Specify contribution in more detail (optional; no more than one sentence)
- Contributed data or analysis tools**
Specify contribution in more detail (optional; no more than one sentence)
- Performed the analysis**
Specify contribution in more detail (optional; no more than one sentence)
- Wrote the paper**
Specify contribution in more detail (optional; no more than one sentence)
- Other contribution**
Specify contribution in more detail (required; no more than one sentence)

- Automatic information exchange between AEC projects and property valuation
- A genetic algorithm optimized machine learning model serves as a smart valuation engine to support automated property valuation
- A BIM and Machine learning integration framework.
- Developed IFC Extension for property valuation and an IFC-based information extraction algorithm

Journal Pre-proof

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof