

Structural condition assessment of existing buildings using machine learning-The role of SMOTE-based data augmentation

Jie Liu¹, Neng Wang² and Guiwen Liu³

¹ PhD Student, School of Management Science and Real Estate, Chongqing University, China

² Faculty, School of Management Science and Real Estate, Chongqing University, China

³ Vice-President, Chongqing University, China

Corresponding author's E-mail: gwliu@cqu.edu.cn

Abstract

To ensure a safe environment for occupants, analyzing the structural safety of existing buildings is essential. The main objective of this paper is to combine machine learning (ML) algorithms with the Synthetic Minority Over Sampling Technique (SMOTE) to establish a comprehensive condition assessment model for structural safety in existing buildings and provide an interpretation of results. Firstly, a raw dataset comprising 18,090 existing buildings in a region of Southwest China was assembled, containing fundamental information about each building. This dataset was then pre-processed and augmented using the SMOTE method. Subsequently, the analysis was performed using four different ML algorithms, including artificial neural network (ANN), decision tree (DT), random forests (RF), and Adaptive Boosting (AdaBoost). Hyperparameters for these models were optimized using a grid search method with 5-fold cross-validation. Finally, a feature importance analysis was conducted based on the best-performing algorithm. The SMOTE-based RF model demonstrated the highest performance, with the evaluation metric G-mean reaching 96.34%. Among all input features, Service Life, Function, and Location were identified as the three most important factors influencing the structural condition of existing buildings. This study represents a promising approach for assisting government regulators in making critical maintenance decisions more effectively and efficiently through a rapid screening model for identifying buildings with potential structural issues.

Keywords: Building condition assessment, Structural safety, Synthetic Minority Over Sampling Technique, Machine learning

1. INTRODUCTION

Existing buildings are those environments already built for the use of occupants. Considering that the safety of existing buildings plays a key role in public benefits, ecological environment and urban image, it has received wide public attention. Assessing the structural safety condition of existing buildings is vital for identifying and mitigating potential hazards, thereby preventing or minimizing

36 accident losses. To ensure the safety and health of the built environment, it is necessary to establish an
37 efficient and effective assessment method to formulate maintenance strategies.

38 **2. LIRERATURE RIVIEW**

39 As a representative of data-driven methods, machine learning method has been applied in damage
40 classification and strength estimation tasks (Ji et al., 2021), for its advantages in data synthesis
41 analysis (Zhang et al., 2021). Regarding the structural condition assessment of buildings, most current
42 researches are about the structural elements, one building only or nearby buildings in specific region,
43 while with relatively little focus on a mass of building groups. Machine learning approaches are
44 adopted to evaluate the condition of structural elements in buildings from plenty of researches (Shen et
45 al., 2022), with data usually from physical experiments (Mangalathu et al., 2020). And when studying
46 the structural condition of nearby buildings in a specific region, some researches apply convolutional
47 neural network with data from measurement devices (Oh and Park, 2022), such as Internet of Things
48 (IoT) and reality capture sensors (Einizinab et al., 2023).

49 The performance of building condition assessment using the above method can be guaranteed to some
50 extent, while at the cost of economy and efficiency. Under this circumstance, the management of
51 large-scale existing buildings is not applicable. Being able to identify potentially unsafe buildings in
52 advance will lead to benefits such as reduced costs and clear targets. Consequently, an economical and
53 efficient tool is greatly needed for preliminary screening anomalous safety condition among a mass of
54 existing buildings, so that the contributions of this paper include :1) establishing a reliable existing
55 building structural condition assessment model by combining machine learning models with SMOTE
56 algorithm; 2) Revealing the main features and internal mechanisms, which can be a reference for the
57 maintenance priority decision of existing buildings.

58 **3. METHODOLOGY**

59 **3.1. Data compilation**

60 For the purposes of this study, report records of 18,090 existing buildings were obtained with the
61 assistance of the official government. These records cover basic information about buildings in a
62 mountainous and hilly region of Southwest China. The selection of this research backdrop is due to
63 the reliability and availability of survey reports above (Ji et al., 2021), as well as the representative of
64 samples exhibiting complexity and diversity. Various input features were selected from the existing
65 information in the report as independent variables. Finally, 14 features related to structural condition
66 were extracted, including structure type, service life, location, area and others. A detailed description
67 of the input features and output condition are represented in Table 1.

68

69

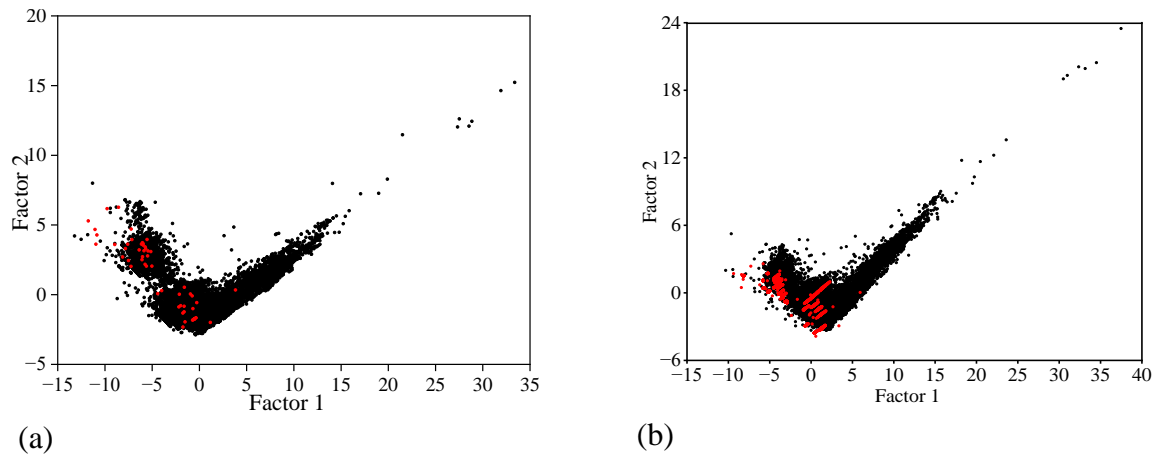
70

Table 1. Detailed description of variables.

Variable	Description
<i>Y</i>	The structural condition of a building
<i>Loc</i>	Street or town where building is located
<i>OFL</i>	Number of stories above ground of a building
<i>UFL</i>	Number of stories underground of a building
<i>Area</i>	Total square footage of all floors across all buildings situated on the site
<i>HT</i>	The vertical height of a building
<i>LS</i>	The service year of a building
<i>Str</i>	Principal architectural frameworks comprising a building
<i>Fn</i>	Major use of building
<i>Prot</i>	Classification of protective buildings
<i>Dsg</i>	Whether to obtain professional design permission
<i>CP</i>	Whether the construction permit is available
<i>Rec</i>	Whether to renovate or reconstruct a building
<i>Sei</i>	Whether to perform seismic reinforcement
<i>Mgmt</i>	Whether property management is available

71 3.2. Data augmentation with SMOTE

72 The imbalance in class distribution is a common occurrence across diverse real-world
73 applications, such as the frequency of building collapse is much lower than its safe condition.
74 In that case, classification algorithms tend to pay unequal attention to the majority samples,
75 overshadowing the minority samples and affecting their representation (Zheng et al., 2023).
76 The Synthetic Minority Over Sampling Technique (SMOTE) stands as a widely adopted
77 approach in addressing the issue of class imbalance through data resampling techniques, the
78 fundamental principle is that features in proximity to each other in the feature space exhibit
79 similar characteristics. New synthetic samples are generated by interpolating between the
80 original minority instance and the chosen neighbor. Based on previous study, the value of *k* in
81 SMOTE is set to 5 (Zhang et al., 2022). In order to reflect the process of data resampling with
82 SMOTE, it is visually displayed in the form of a two-dimensional graph using Principal
83 Component Analysis (PCA), as shown in Fig.1. And the black scatter represents a safe
84 building, while the red represents an unsafe building. Based on the SMOTE technique, the
85 number of safe and unsafe samples is consistent.



86 **Figure 1. Data distribution before and after processing by SMOTE:**
 87 **(a) Original dataset, (b) Synthetic dataset**

88 3.3. Model development

89 The assessment method employs four machine learning classification algorithms, which are artificial
 90 neural network (ANN), Decision tree (DT), random forests (RF) and Adaptive Boosting (AdaBoost).
 91 To avoid overfitting and underfitting the model, these data are randomly divided into the training set
 92 (70% of the data) and the test set (30% of the data) (Chung et al., 2021). Additionally, to select the
 93 optimal hyperparameters, a five-fold cross-validation technique is employed to further enhance the
 94 model's reliability and robustness. The tuning parameter values are shown in Table 2.

Table 2. The tuning parameter values of SMOTE-based machine learning algorithms.

Algorithms	Hyperparameters	Definition	Value
ANN	<i>size</i>	The number of nodes in the hidden layer	5
	<i>layer</i>	layer in binary classification.	1
	<i>fuction</i>	activation function	sigmoid
DT	<i>minsplitt</i>	The minimum number of observations in a node	13
AdaBoost	<i>size</i>	The number of neurons	5
	<i>n_estimators</i>	The number of weak classifiers (the maximum iteration number)	10
RF	<i>ntree</i>	The number of trees	100
	<i>minsplitt</i>	The minimum number of observations in a node	13

95 4. RESULTS AND DISCUSSION

96 4.1. Model performance

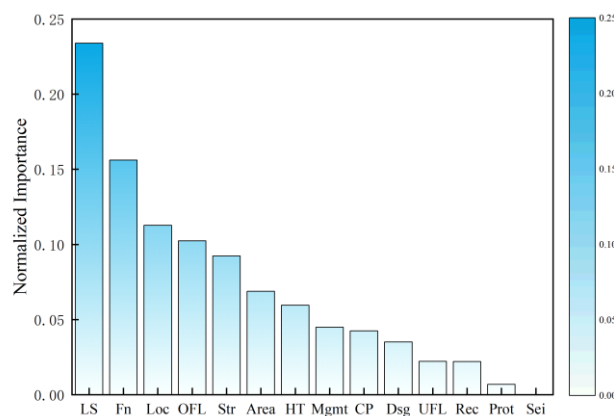
97 Table 3 indicate that SMOTE-based machine learning methods, especially ensemble learning models
 98 show extremely high performance. In addressing data imbalance issues, studies have integrated
 99 sampling methods with machine learning algorithms, leading to noteworthy enhancements in model
 100 accuracy. The results also prove that RF algorithm has better prediction ability than any other models
 101 mentioned. This rarely high accuracy of the RF is in agreement with previous studies in classification
 102 (Chung et al., 2021), when there are more complex relationships among the features (Park and Park,
 103 2021), as in this study. Thus, the ensemble learning algorithms based on data augmentation can be
 104 used as a basic approach for assessing the condition of large-scale existing buildings. This can be
 105 attributed to RF algorithm integration through voting, which requires lower quality of training samples.

106 **Table 3. Results of structural condition of test set using ANN, DT, AdaBoost and RF.**

Models	Recall		Precision		F1		G-mean	Accuracy
	safe	unsafe	safe	unsafe	safe	unsafe		
ANN	91.84%	94.12%	93.98%	92.02%	92.90%	93.06%	92.97%	92.98%
DT	98.75%	65.66%	74.20%	98.13%	84.73%	78.68%	80.52%	82.20%
AdaBoost	96.19%	94.91%	94.98%	96.14%	95.58%	95.52%	95.55%	95.55%
RF	98.48%	94.26%	94.49%	98.41%	96.44%	96.29%	96.34%	96.37%

107 4.2. Feature importance analysis

108 Feature importance ranking method can be employed to assess the importance of input parameters on
 109 condition. The optimal RF model provides the core parameter identification and importance ranking
 110 for the structural condition assessment. The three most important features for the assessment model
 111 are *LS*, *Fn* and *Loc*. Among the input features, and *Sei* is the least important feature (refer to Figure 2).



112 **Figure 2. Contribution of each input variable on the structural condition.**

113 5. CONCLUSIONS

114 The condition assessment of a mass of existing buildings is crucial for the study of global
115 environmental issues caused by building collapsing. In this study, four ML algorithms based on
116 SMOTE data augmentation are utilized to assess the structural condition of existing buildings. When
117 assess structural condition of existing buildings, RF model performs the best. The three most
118 important input features are *LS*, *Fn*, and *Loc*, and the least important feature is *Sei*.

119 6. ACKNOWLEDGMENTS

120 This study was financially supported by the Chongqing Housing and Urban Rural Construction
121 Commission (Grant No. 2022:6-4), the National Natural Science Foundation of China (Grant
122 No.72271035).

123 7. REFERENCES

- 124 CHUNG, L. C. H., XIE, J. & REN, C. 2021. Improved machine-learning mapping of local climate zones in
125 metropolitan areas using composite Earth observation data in Google Earth Engine. *BUILDING AND*
126 *ENVIRONMENT*, 199.
- 127 EINIZINAB, S., KHOSHELHAM, K., WINTER, S., CHRISTOPHER, P., FANG, Y., WINDHOLZ, E.,
128 RADANOVIC, M. & HU, S. 2023. Enabling technologies for remote and virtual inspection of building
129 work. *Automation in Construction*, 156.
- 130 JI, S., LEE, B. & YI, M. Y. 2021. Building life-span prediction for life cycle assessment and life cycle cost using
131 machine learning: A big data approach. *BUILDING AND ENVIRONMENT*, 205.
- 132 MANGALATHU, S., JANG, H., HWANG, S.-H. & JEON, J.-S. 2020. Data-driven machine-learning-based
133 seismic failure mode identification of reinforced concrete shear walls. *Engineering Structures*, 208,
134 110331.
- 135 OH, B. K. & PARK, H. S. 2022. Urban safety network for long-term structural health monitoring of buildings
136 using convolutional neural network. *Automation in Construction*, 137.
- 137 PARK, H. & PARK, D. Y. 2021. Comparative analysis on predictability of natural ventilation rate based on
138 machine learning algorithms. *BUILDING AND ENVIRONMENT*, 195.
- 139 SHEN, Y., WU, L. & LIANG, S. 2022. Explainable machine learning-based model for failure mode
140 identification of RC flat slabs without transverse reinforcement. *Engineering Failure Analysis*, 141.
- 141 ZHANG, A., YU, H., HUAN, Z., YANG, X., ZHENG, S. & GAO, S. 2022. SMOTE-RkNN: A hybrid re-
142 sampling method based on SMOTE and reverse k-nearest neighbors. *Information Sciences*, 595, 70-88.
- 143 ZHANG, L., WEN, J., LI, Y. F., CHEN, J. L., YE, Y. Y., FU, Y. Y. & LIVINGOOD, W. 2021. A review of
144 machine learning in building load prediction. *APPLIED ENERGY*, 285.
- 145 ZHENG, G. Z., ZHANG, Y. Q., YUE, X. H. & LI, K. 2023. Interpretable prediction of thermal sensation for
146 elderly people based on data sampling, machine learning and SHapley Additive exPlanations (SHAP).
147 *BUILDING AND ENVIRONMENT*, 242.

148