Towards Global Interpretability: Evaluating XAI Metrics in Building Footprint Extraction

Elif Ozlem Yilmaz[1], Taskin Kavzoglu[1]

[1] Department of Geomatics Engineering, Gebze Technical University, 41400, Gebze-Kocaeli, Turkiye

Keywords: Deep learning, Building footprint extraction, Explainable AI, GuidedGradCAM, GradientSHAP.

Abstract

Global population is projected to increase by about 70% by 2050, with a growing proportion of people living in urban areas. This trend highlights the importance of accurately assessing urban expansion. Automatic building detection from remotely sensed imagery using deep learning (DL) has demonstrated considerable potential for applications, including sustainable urban planning and infrastructure monitoring. However, the inherent black-box nature of DL models limits their transparency and reduces trust in model-driven decisions. Although various Explainable Artificial Intelligence (XAI) approaches have been proposed to highlight image regions influencing model predictions, qualitative visual inspection alone is insufficient for reliably evaluating the credibility of these explanations. This study evaluates several XAI techniques for building footprint extraction using a U-Net model trained on a refined Massachusetts Buildings Dataset. The segmentation model achieved precision, recall, F1-score, IoU, and overall accuracy values of 89.68%, 85.69%, 87.53%, 79.03%, and 94.35%, respectively. To investigate the model’s decision-making process, three explanation methods, namely Saliency, GradientSHAP, and GuidedGradCAM, were applied. The quality of the generated explanations was then quantitatively assessed using 16 evaluation metrics. Beyond single-image analysis, a dataset-level evaluation was conducted using 547 image patches containing building coverage greater than 20%. The results indicate that GuidedGradCAM produces more consistent and reliable explanations. Furthermore, dataset-level analysis using dense-building samples provides a statistically more robust representation of overall model behaviour compared to evaluations based on individual images. These findings highlight the importance of quantitative assessment in validating the interpretability of DL models for building footprint extraction.

1. Introduction

Currently, urbanization is one of the most significant global transformations, and more than half of the world’s population lives in cities, and this proportion is expected to increase to nearly 70% by 2050 (United Nations, 2019). Such accelerated growth places substantial pressure on urban environments, intensifying the need for effective spatial planning. A key requirement for addressing these planning challenges is the availability of accurate and up-to-date building footprint information, which supports infrastructure management, energy modelling, and disaster risk reduction. The importance of accurate building delineation extends directly to global sustainability and climate goals. Sustainable Development Goal (SDG) 11, “Sustainable Cities and Communities” underscores the need for reliable spatial information to monitor urban growth. Concurrently, SDG 13, “Climate Action” highlights the requirement to monitor infrastructure vulnerable to climate-related hazards. The rapid identification of structures exposed to risks like earthquakes or floods provides vital data for disaster risk management, linking precise building data directly to resilience and safety outcomes.

Remote sensing has become one of the main sources for generating building footprint maps. However, accurately identifying buildings from aerial or satellite imagery remains challenging (Maggiori et al., 2017; Ji et al., 2018). Buildings vary considerably in size, shape, and construction style, and environmental factors such as vegetation cover and shadows can further complicate the detection process. Recent developments in artificial intelligence, particularly deep learning, have opened new possibilities for addressing these difficulties. Semantic segmentation models are now widely used for building detection because they can capture complex spatial patterns and detect structures with different sizes and materials. As a result, these approaches often achieve considerably higher accuracy than traditional image processing techniques. Nevertheless, some challenges still remain. Variations in regional architecture and the presence of complex urban backgrounds can reduce model performance, which highlights the need for more robust and generalizable deep learning models (Chang et al., 2025).

Deep learning models excel in extracting multi-scale spatial features, yet their “black box” nature has raised concerns about the interpretability of their decisions (Yilmaz et al., 2025). For model predictions to be trusted by practitioners and policymakers, it is essential to understand the reasoning behind them. Explainable AI (XAI) techniques provide valuable tools for visualizing the regions or weighting the features, considering their influence on model output (Samek et al., 2017; Yilmaz and Kavzoglu, 2024). Methods such as Grad-CAM, Integrated Gradients, and SHAP are applied to visualize model decisions. However, most studies rely primarily on qualitative visual inspection of explanation maps, which makes it difficult to objectively assess the reliability of these explanations. In other words, visual explanations alone are insufficient; therefore, quantitative XAI metrics are needed to assess the faithfulness, relevance, and selectivity of explanations (Adebayo et al., 2018; Bhatt et al., 2020). By introducing quantitative evaluation, these measures transform explanations from subjective visual outputs into verifiable analytical results, thereby strengthening confidence in deep learning–based decision-making.

This study investigates the effectiveness and robustness of 16 XAI evaluation metrics applied to explanations produced by Saliency, GradientSHAP, and GuidedGradCAM. A U-Net model was trained using the refined Massachusetts Buildings Dataset proposed by Yilmaz and Kavzoglu (2025), and the interpretability of the resulting model was subsequently examined in a systematic manner. To comprehensively analyse the behaviour of the XAI approaches, both local and global explanations, whose importance were discussed in recent studies (e.g., Teke and Kavzoglu, 2024; Ozupek et al., 2025), were generated for the building footprint extraction task.

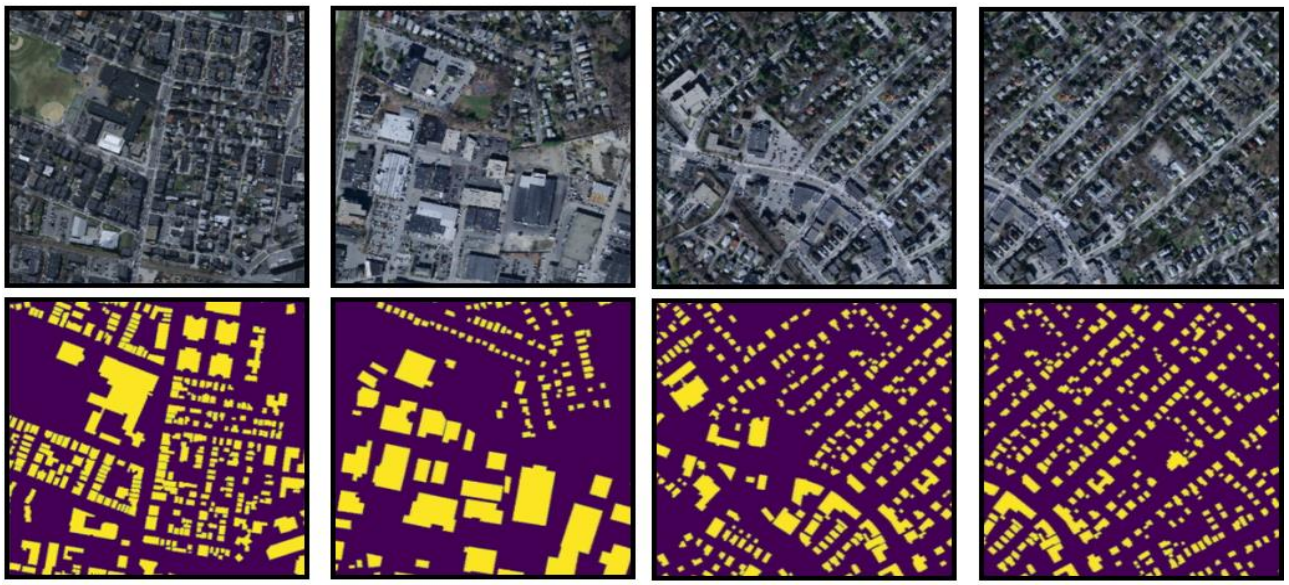

Figure 1. Examples of representative image-mask pairs from the dataset.

2. Dataset

The Massachusetts building dataset consists of 151 high-resolution aerial images of Boston from about 2011 (Figure 1). Due to errors observed in the original labels, such as incorrect labelling, non-building inclusion, missing annotations, and spatial misalignment, the Refined Massachusetts dataset, with these errors corrected, was used in this study. The size of the 151 images is 1500x1500 pixels. All images were divided into 512×512 semi-overlapping patches, and the resulting dataset was split into training (1495 patches), validation (320 patches), and test (320 patches) subsets (Yilmaz and Kavzoglu, 2025). In the global XAI assessment, images containing more than 20% building pixels in the entire dataset were used.

3. Methodology

3.1 U-Net Model

The U-Net architecture, which has an encoder–decoder structure, extracts high-level features in the encoder section and rescales them in the decoder section to produce pixel-level segmentation. Skip connections preserve fine detail information, enabling more accurate and sharp boundary segmentation. In this study, the AdamW optimizer was used, and the batch size was set to 8. The models were run for 200 epochs with a learning rate of 1×10^{-4}. ResNet50 was chosen as the backbone network for the U-Net architecture. Data augmentation techniques were also applied to improve generalization performance.

3.2 XAI Methods

To assess the interpretability of the proposed model, three explainability methods were employed: Saliency, GradientSHAP, and GuidedGradCam. Saliency provides a baseline derivative-based framework that quantifies the sensitivity of the model output with respect to each input pixel (Simonyan et al., 2014). It is computed in a single backward pass, which makes it a fast technique. It is computationally inexpensive, yet it produces noisy and sparse maps. GradientSHAP extends this approach by integrating gradient information with an expectation-based formulation, thereby generating more stable and noise-robust attribution maps (Lundberg and Lee, 2017; Springenberg et al., 2014). It was developed by combining gradient and Shapley value logic. In other words, it approximates Shapley-like citations using gradient integrals. Although it performs better than the previously mentioned method, its computational cost is quite high. GuidedGradCam fuses the class-discriminative localization capability of Grad-CAM with the fine-grained feature details obtained from Guided Backpropagation, resulting in high-resolution and class-specific explanation maps (Selvaraju et al., 2017). Specifically, it produces a class-selective coarse localization map by feeding the gradients of the target class into the final convolutional layer. It then generates a feature map by merging it with high-resolution gradient-based visualizations such as Guided Backprop.

3.3 XAI Metrics

Various metrics were employed to quantitatively assess the outputs produced by these explanation methods. In this context, metrics such as faithfulness estimate, faithfulness correlation, infidelity, AUC, pixel flipping, region perturbation, sparseness, complexity, relevance rank accuracy, relevance mass accuracy, IROF, maximum sensitivity, selectivity, TopKIntersection, average sensitivity, and efficient complexity were computed (Montavon et al., 2018; Samek et al., 2017; Yeh et al., 2019; Hooker et al., 2019). The descriptions of these metrics are summarized in Table 1. With the application of these metrics, each explainability method was evaluated not only through visual inspection but also through a comprehensive set of quantitative indicators that capture different dimensions of interpretability (Oblizanov et al., 2023).

4. Results and Discussion

The U-Net model with aforementioned setting and ResNet50 backbone was trained on the refined Massachusetts building dataset, and the accuracy assessment was conducted on the test data (Table 2). It can be seen that IoU of 79.03%, F-Score of 87.53% and accuracy of 94.35% were achieved. After the building footprint extraction process, the decision-making process of the U-Net model was analysed using XAI methods (Figure 2).

Metric	Description
AUC	shows the reduction of the area under the curve obtained in perturbation-based metrics to a single number
Average Sensitivity	indicates the average/distributional version of MaxSensitivity
Complexity	measures the uncertainty/spread of the explanation
Effective Complexity	measures how many important features you need to retain so that the model’s performance does not deteriorate beyond a certain tolerance
Faithfulness Correlation	calculates how linearly related the explanation scores are to the change in the model output when a specific feature subset is pulled to the baseline using Pearson correlation
Faithfulness Estimate	A similar approach to Faithfulness Correlation exists, but instead of cumulative subsets, it applies individual region/feature perturbations
IROF	splits the image into segments, sorts them according to their average value, masks the segments in order, and calculates the area of the output curve
Infidelity	The explanation vector measures how much the inner product of a perturbation vector differs from the actual output change of the model using the mean squared error
Maximum Sensitivity	The explanation measures how much it changes in the worst case within a certain radius
Pixel Flipping	measures how quickly the model output degrades by sequentially flipping/perturbing the most important pixels according to the attribution score
Region Perturbation	PixelFlipping’s region/patch version. Instead of pixels, local regions (patches) are sorted and perturbed
Relevance Mass Accuracy	measures how much of that positive attribution remains within the perceived truth
Relevance Rank Accuracy	tests whether the pixels in the ground truth drop to the top of the reference ranking
Selectivity	measures whether the attribution map is focused on the most important areas
Sparseness	quantifies how sparse the distribution is; sparseness is high if few regions have high values
TopKIntersection	measures the intersection of the highest K attribution index with the ground truth

Table 1. Summary of the XAI metrics employed in this study.

Model	Precision	Recall	IoU	F-score	Accuracy
U-Net	89.68	85.69	79.03	87.53	94.35

Table 2. Performance analysis of U-Net model (%).

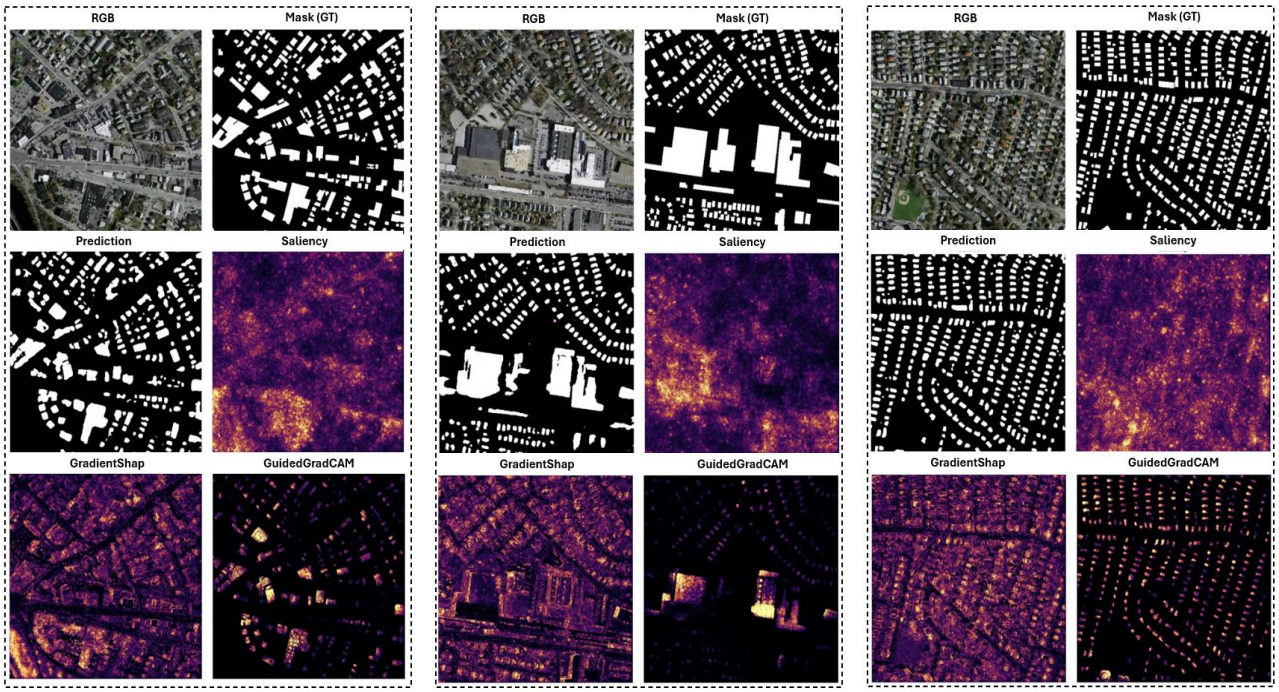

Figure 2. Some of the XAI feature map outputs.

XAI maps generated by the three methods for different types of urban structures, including industrial buildings and regular residential blocks, are presented in Figure 2. Model predictions are given alongside RGB images and their corresponding ground truth masks to evaluate the spatial alignment of explanations with model outputs. It was observed that the model

predictions were largely consistent with the location accuracy masks in all samples. High noise is exhibited in all Saliency maps. In other words, the explanations were not only concentrated in the building areas, but there was also high activation in the background objects. GradientSHAP outputs were more structured and localized than saliency. Specifically, building boundaries were clearer, and background activation was lower. On the other hand, GuidedGradCAM clearly identifies building shapes and boundaries among other methods. Industrial structures were highlighted very clearly.

As the main objective of the study, the explainability of the models for the three XAI methods (Saliency, GradientSHAP, GuidedGradCAM) was measured using 16 metrics for global and local assessments (Table 3). As mentioned earlier, samples containing 20% building pixels were used to conduct a global assessment. Local evaluation, on the other hand, measures explanation quality at the individual patch level. In this study, the first example, shown in Figure 2, is used as a representative sample (i.e., local explanation sample).

XAI Metrics	Global Assessment			Local Assessment		
	GradientSHAP	GuidedGradCAM	Saliency	GradientSHAP	GuidedGradCAM	Saliency
↑ **AUC**	0.5820	0.8885	0.5076	0.5317	0.9276	0.3685
↓ **Average Sensitivity**	0.5082	0.5082	0.5082	0.4323	0.4322	0.4322
↓ **Complexity**	12.0975	10.1122	12.2303	11.8660	10.8558	12.0830
↓ **Effective Complexity**	261894.23	55738.75	262117.13	260917.00	82452.00	262079.00
↑ **Faithfulness Correlation**	-0.0198	-0.0158	0.0016	-0.2623	0.1117	-0.0461
↑ **Faithfulness Estimate**	0.8313	0.4012	-0.0794	0.9858	0.5085	-0.9542
↑ **IROF**	5.5230	6.4616	4.9533	-7.3751	-6.7404	-6.7465
↓ **Infidelity**	6547370.10	353512.62	543054.09	21336.42	27243.60	69966.40
↓ **Maximum Sensitivity**	0.5856	0.5855	0.5857	0.5260	0.5258	0.5262
↑ **Pixel Flipping**	0.0044	0.0050	0.0045	0.0006	0.0007	0.0006
↑ **Region Perturbation**	0.0009	0.0011	0.0009	-0.0001	-0.0001	-0.0001
↑ **Relevance Mass Acc.**	0.3018	0.9323	0.2693	0.3416	0.9886	0.2795
↑ **Relevance Rank Acc.**	0.3227	0.8032	0.2729	0.3662	0.8704	0.2425
↑ **Selectivity**	0.0059	0.0056	0.0059	0.0009	0.0009	0.0009
↑ **Sparseness**	0.4743	0.9113	0.3779	0.5810	0.8491	0.4739
↑ **TopKIntersection**	1.2804	3.8517	1.2346	0.7246	2.7118	0.5701

Table 3. Quantitative comparison of XAI methods across global and local explanation metrics. Note that arrows indicate the desired optimization direction for each metric (↑ higher values are better, ↓ lower values are better)

XAI Metrics	Global Assessment			Local Assessment		
	GradientSHAP	GuidedGradCAM	Saliency	GradientSHAP	GuidedGradCAM	Saliency
AUC	-0.469	1.390	-0.921	-0.33	1.356	-1.026
Average Sensitivity	0.0005	0.0000	-0.0005	-1.226	0.003	1.223
Complexity	-0.637	1.412	-0.775	-0.494	1.395	-0.900
Effective Complexity	-0.706	1.414	-0.708	-0.700	1.414	-0.714
Faithfulness Correlation	-0.911	-0.481	1.392	-1.283	1.156	0.127
Faithfulness Estimate	1.201	0.045	-1.247	0.976	0.398	-1.374
IROF	-0.198	1.312	-1.114	-1.414	0.717	0.697
Infidelity	-1.414	0.740	0.674	0.839	0.566	-1.405
Maximum Sensitivity	1.224	0.000	-1.224	0.000	1.225	-1.225
Pixel Flipping	-0.889	1.397	-0.508	-0.707	1.414	-0.707
Region Perturbation	-0.807	1.409	-0.602	1.146	-1.290	0.144
Relevance Mass Acc.	-0.653	1.413	-0.760	-0.608	1.410	-0.802
Relevance Rank Acc.	-0.600	1.409	-0.809	-0.467	1.390	-0.923
Selectivity	0.825	-1.407	0.582	-0.056	1.253	-1.197
Sparseness	-0.489	1.394	-0.905	-0.340	1.359	-1.019
TopKIntersection	-0.688	1.414	-0.726	-0.626	1.411	-0.785

Table 4. Numerical comparisons of GradientSHAP, GuidedGradCAM, and Saliency based on global and local evaluation metrics. To enable a fair comparison between metrics with different range of scales, normalization was conducted using Z-score normalization.

Results revealed that GuidedGradCAM demonstrated the best performance in terms of most localization-related metrics, including AUC, RelevanceMassAccuracy, and sparseness. These metrics specifically measure properties related to

explanation localization, model fidelity, and reference sparsity. The metrics in which GradientSHAP shows the strongest performance are faithfulness estimate and maximum sensitivity. Similar results were also obtained in some perturbation metrics. Overall, the model’s decision is explainable, but spatial localization is weak. Except for the faithfulness correlation metric, most metrics derived from saliency showed poor performance because they tend to produce noisy attribution maps and lack clear spatial localization. Results also showed that the GuidedGradCAM method again demonstrated superior performance across most metrics and provided more consistent and reliable explanations at the sample level. For the selected sample patch, the GradientSHAP algorithm stands out with metrics such as faithfulness estimate and region perturbation. The saliency model performed poorly on many metrics.

When local and global evaluations are considered jointly, notable differences emerge between instance-level and dataset-level explanation behaviour. According to these metrics, certain XAI methods may exhibit stronger performance at the global scale while performing comparatively weaker at the local scale. Nevertheless, a consistent superiority is observed for the values derived through Z-score normalization (Table 4). For metrics such as complexity, lower values indicate favourable performance, whereas higher values reflect poorer outcomes. For metrics in which lower values correspond to better performance, the normalized scores were multiplied by −1 so that higher values consistently represent better performance across all metrics. Because XAI metrics exhibit different numerical ranges and optimization directions, Z-score normalization was applied to ensure a fair comparison and consistent visualization across metrics. When the value of a metric for a given method is equal to, or very close to, the mean of the compared methods, the corresponding normalized Z-score approaches zero. This indicates that the method exhibits average performance with respect to that metric.

Using normalized metric values, both local and global radar plots were created for the selected XAI methods (Figure 3). The radar plots highlight noticeable differences between the behavior of the methods. For instance, GuidedGradCAM tends to achieve higher scores across several explainability metrics in both global and local evaluations. This suggests that the method produces more spatially focused explanations, emphasizing relevant building structures while suppressing activations in background regions. GradientSHAP, on the other hand, shows stronger sensitivity to gradient-based features. Nevertheless, its ability to precisely localize building boundaries appears to be more limited compared to GuidedGradCAM. Saliency generally performs weaker than the other methods and frequently produces noisy and scattered attribution maps. As a result, the highlighted regions are often less aligned with meaningful structures in the image.

The radar graphs also revealed that the explanation methods exhibit different performance patterns under local and global XAI evaluations. In several metrics, a method may show higher performance in the global evaluation, but the same method may yield lower values in the local evaluation. For example, for the GradientSHAP method, the selectivity metric achieved a higher normalized value in global evaluation but showed a lower value in local evaluation, which show the dependency or subjectiveness on the selected image patch. Similarly, while some metrics may appear more advantageous in local analysis, this may not fully reflect the behavior of the methods across the entire dataset. In short, it indicates that local and global assessments reveal different dimensions rather than the same aspect of model interpretability. Since local evaluation is based on explanations pertaining to a single image patch, it is more influenced by example- or region-specific characteristics. In particular, factors such as structural density, building geometries, shadow effects, and background complexity in remote sensing images can significantly affect the explanation of the maps. This finding suggests that dataset-level evaluation provides a more reliable and stable basis for comparing XAI methods for all samples, as it reduces the influence of sample-specific variations in model assessment. Consequently, evaluating explanation reliability only at the instance level may lead to conclusions that are not fully representative. Considering dataset-level metrics together with local assessments allows a more comprehensive interpretation of explanation quality. To sum up, global evaluation offers a more robust analytical framework for comparing XAI methods, as it reflects the overall behaviour of the methods in a more representative manner.

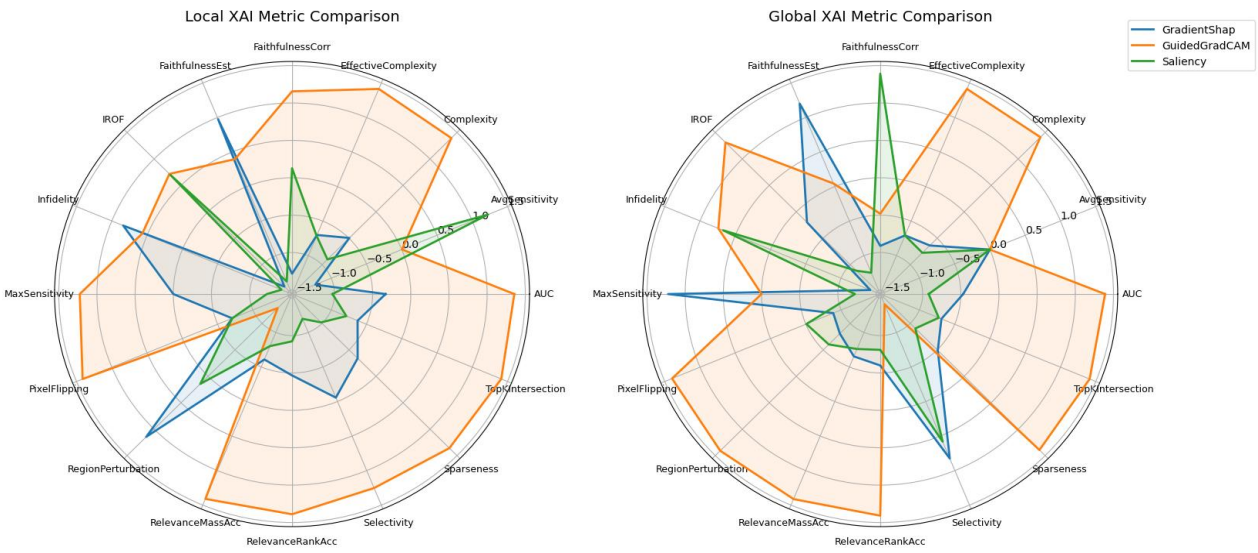

Figure 3. Radar charts comparing local and global explainability evaluation results for Saliency, GradientSHAP, and GuidedGradCAM based on 16 XAI metrics.

5. Conclusion

This study investigates the role of various XAI techniques in interpreting the decision-making processes of deep learning models used for building footprint extraction. Beyond qualitative visual inspection, the performance of the explanation methods was systematically evaluated using a comprehensive set of quantitative metrics at both local and global levels. The results indicate that the U-Net model achieves high accuracy in building footprint extraction when trained on the refined Massachusetts Building Dataset. To assess the interpretability of the trained model, Saliency, Gradient SHAP and Guided Grad-CAM techniques were employed. The results of the analysis indicate that visual interpretation alone may not be sufficient for drawing conclusions. Instead, quantitative analysis provides a more objective assessment of the reliability of the explanation techniques. 16 XAI metrics were employed to evaluate the explanation techniques from various perspectives. Considering both local and global evaluations, Whilst GuidedGradCAM demonstrated the most consistent and overall strongest performance across the evaluated XAI metrics, GradientSHAP performed well for certain faithfulness-related metrics. However, the spatial localisation capability of GradientSHAP was poor. On the other hand, the Saliency method produced noisy output for the attribution map, resulting in poor performance for most metrics.

Another important contribution of this study is the implementation of a global evaluation strategy that focuses on image patches containing a sufficient proportion of building pixels. This approach was adopted to better capture the model’s behaviour in relation to building footprints rather than responses triggered by unrelated features within the image patches. In addition to this dataset-level analysis, the performance of local explanations was examined using a representative image patch to explore instance-level behaviour. To enable meaningful comparison across metrics with different numerical scales and optimization directions, a direction-aware Z-score normalization scheme was applied. The results indicate also that global explainability analysis provides a more comprehensive understanding of the model’s decision-making behaviour. Among the evaluated approaches, the GuidedGradCAM method produced more stable and consistent explanations, particularly for the task of building footprint extraction. These findings further underscore the importance of incorporating dataset-level evaluation when comparing XAI techniques. In this regard, a global evaluation framework offers a more robust and representative analytical perspective, as it reflects the overall behaviour of explanation methods across the dataset while reducing the influence of variations arising from individual samples.

Acknowledgements

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Project No. 124Y240, titled *“Comparative Analysis of Explainable Artificial Intelligence Algorithms in Deep Learning-Based Building Footprint Extraction Applications.”*

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31

Bhatt, U., Weller, A., Moura, J.M.F., 2020. Evaluating and aggregating feature-based model explanations. *arXiv preprint* arXiv:2005.00631.

Chang, J., He, X., Song, D., Li, P., Qiao, M., Cheng, X., 2025. A multi-scale attention network for building extraction from high-resolution remote sensing images. *Scientific Reports*, 15, 24938.

Hooker, S., Erhan, D., Kindermans, P. J., Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, *32*.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery dataset. *IEEE Transactions on Geoscience and Remote Sensing,* 57(1), 574-586

Lundberg, S. M., Lee, S. I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems,* 30, 4765–4774.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. High-resolution aerial image labeling with convolutional neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, 55(12), 7092-7103.

Montavon, G., Samek, W., Müller, K. R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1-15.

Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., Dorofeeva, A., 2023. Evaluation metrics research for explainable artificial intelligence global methods using synthetic data. *Applied System Innovation*, 6(1), 26.

Ozupek, E., Teke, A., Celik, N., Kavzoglu, T. 2025. Explainable artificial intelligence to explore the intrinsic characteristics of climatic parameters governing meteorological drought forecasting: opening the black box. *Stochastic Environmental Research and Risk Assessment*, 39, 3201-3222.

Samek, W., Wiegand, T., Müller, K.R., 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint* arXiv:1708.08296.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *In: Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint* arXiv:1312.6034.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. *arXiv preprint* arXiv:1412.6806.

Teke, A., Kavzoglu, T., 2024. Exploring the decision-making process of ensemble learning algorithms in landslide susceptibility mapping: Insights from local and global explainable AI analyses. *Advances in Space Research*, 74(8), 3765–3785.

United Nations, 2019. World Urbanization Prospects: The 2018 Revision. United Nations, Department of Economic and Social Affairs, Population Division.

Yeh, C. K., Hsieh, C. Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, *32*.

Yilmaz, E.O., Kavzoglu, T., 2024. Burned area detection with Sentinel-2A data: Using deep learning techniques with explainable artificial intelligence. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-5-2024, 251–257.

Yilmaz, E.O., Teke, A., Kavzoglu, T., 2025. A performance analysis of U-Net and U-Net++ in building footprint extraction using XAI. *In: Proceedings of the 33rd Signal Processing and Communications Applications Conference (SIU*), IEEE, 1–4.

Yilmaz, E.O., Kavzoglu, T., 2025. DeepSwinLite: A Swin Transformer-based light deep learning model for building extraction using VHR aerial imagery. *Remote Sensing*, 17(18), 3146.