



Machine-Learning Classification of Construction-Site Annotations from Geospatial Metadata: Near-Perfect Binary Discrimination and the Limits of Construction-Type Recognition

Maleerat Sodanil
Faculty of Information Technology, King Mongkut's University
of Technology North Bangkok, Bangkok, Thailand
maleerat.s@it.kmutnb.ac.th

ABSTRACT

Automated construction monitoring increasingly relies on geospatial metadata and machine learning to track urban development, yet the true generalizability and reliability of metadata driven models remain underexplored. This study evaluates the predictive capabilities of geospatial metadata including polygon geometry, temporal duration, and annotation attributes using Random Forest, Extra Trees, and Logistic Regression models on a large scale GeoJSON dataset of construction sites. We conducted binary (site presence) and multiclass (construction type) classification tasks, revealing a stark contrast: while Random Forest achieved near perfect binary accuracy (99.9%), multiclass construction type recognition remained highly ambiguous (macro-F1 = 0.610). Crucially, permutation feature importance analysis exposed that the binary predictive signal was almost entirely concentrated in a single annotation derived variable ($n_features$), with all physical and geometric features registering zero importance. This indicates that the near perfect performance is an artifact of the annotation workflow and potential target leakage rather than genuine physical site detection. We conclude that metadata only models are insufficient for fine grained semantic typing and issue a critical methodological warning for the GeoAI community: rigorous spatial cross validation and annotation leakage audits are mandatory to prevent spatially and procedurally inflated performance metrics.

Keywords: GeoAI, Construction Monitoring, Data Leakage, Spatial Cross Validation, Geospatial Metadata, Machine Learning, Urban Construction Sites, Annotation Bias, Semantic Classification

Received: 3 May 2026, **Revised:** 2 August 2026, **Accepted:** 11 August 2026

Copyright: DLINE

1. Introduction and Background

Construction projects are highly dynamic and complex environments in which site conditions can

change rapidly, creating persistent challenges for safety, productivity, quality, and resource management [1]. Consequently, construction monitoring has become an increasingly important function in contemporary construction practice. It involves systematically collecting, processing, and analysing site related information to assess construction quality, monitor project progress, identify and manage safety risks, allocate resources efficiently, and verify compliance with relevant requirements [2]. Recent advances in artificial intelligence (AI), particularly computer vision and foundation models, have further expanded the potential of automated construction monitoring. These technologies enable accurate, scalable interpretation of visual information and support applications such as progress monitoring, safety management, and tracking construction equipment and materials [3, 4].

Urban Construction Sites (UCSs) represent direct manifestations of urban development and construction activity. The rapid pace of global urbanization has resulted in substantial growth in construction activities, particularly in developing countries, consequently increasing the number and spatial distribution of UCSs within urban areas [5]. These sites can significantly affect the surrounding urban environment and human well being. In particular, construction site activities can directly affect air quality, noise levels, soil quality, and residents' overall living experiences [6, 7]. Effective monitoring and characterization of UCSs are therefore important not only for construction management but also for understanding their broader environmental and urban impacts.

However, the reliability and generalizability of AI-based systems depend not only on model architecture but also on the availability, quality, diversity, and consistency of the data used to develop them [8]. This perspective is consistent with the principles of data centric AI, which emphasise dataset quality, diversity, standardisation, and systematic preparation as fundamental requirements for developing robust and generalizable AI applications [9]. These considerations are especially important in construction monitoring because construction projects vary substantially in project type, scale, site configuration, environmental conditions, and operational characteristics. Consequently, well-curated, representative datasets are essential for developing reliable machine learning models that perform effectively across diverse construction environments [10].

The development of suitable datasets for construction monitoring generally involves three interconnected processes: data collection, data labeling, and data preparation. First, researchers can collect construction-related data using fixed cameras, unmanned aerial vehicles (drones), LiDAR systems, and mobile devices. However, variations in site layout, weather conditions, illumination, viewing geometry, and other environmental factors can introduce substantial inconsistencies into the collected data [11]. Synthetic data generation offers a complementary approach to address some of these limitations. BIM environments, simulation platforms, and generative models can generate representative data and enable the inclusion of rare, hard to observe, or hazardous construction scenarios at scale [12, 13, 14, 15].

Second, data labeling is necessary to transform raw observations into structured training examples for supervised machine learning applications. In construction monitoring, annotation may involve identifying workers, construction equipment, objects, site features, and potential hazards. Conventional manual annotation is often labor intensive and requires considerable time and human expertise. To reduce this burden, researchers have increasingly explored semisupervised, weakly supervised, and active learning approaches. These approaches reduce redundant annotation effort while allowing human expertise to focus on samples that are more informative or critical for model development [16].

Third, systematic data preparation ensures datasets support reliable model development and evaluation. This stage may include normalization, metadata documentation, data augmentation, and appropriate partitioning into training, validation, and test subsets. These procedures reduce inconsistencies, improve comparability across experiments, and support fair, reproducible evaluation of machine-learning models [17].

Despite advances in dataset development methodologies, the construction industry has not yet established sufficiently mature infrastructures for widespread open data availability and systematic data reuse [18]. One

major barrier is the proprietary nature of construction project information. Project records are frequently subject to ownership restrictions and contractual arrangements that limit their dissemination beyond the organizations directly involved in a project. In addition, privacy considerations associated with workers and other individuals can restrict the release of construction imagery containing identifiable persons [19]. Even where organizations are willing to share data, substantial resources may be required for data curation, annotation, anonymization, quality control, and long-term hosting. These requirements can represent significant barriers, particularly for academic institutions and nonprofit organizations with limited resources [20].

At the same time, rapid developments in machine learning and, particularly, deep learning, together with the increasing availability of open source software and tools, have created new opportunities for exploiting large-scale datasets to address challenges in geographic information science (GIScience) and remote sensing (RS). Open source machine learning frameworks, such as Google TensorFlow [21], have contributed to the broader accessibility of computational approaches for processing and analyzing large datasets. These developments have also increased the prominence of geospatial artificial intelligence (GeoAI), which has gained traction in both research and industrial applications.

GeoAI integrates artificial intelligence and machine learning approaches with geospatial information, geographic information systems, remote sensing, and large scale geospatial datasets. Its development reflects the growing need to extract meaningful information from increasingly large and heterogeneous geospatial data sources. Boutayeb [22] provides a comprehensive overview of the relationship between machine learning and the geospatial domain, emphasizing the close relationship between GeoAI, geographic information systems, and large scale geospatial data, commonly described as big geodata. The increasing availability of machine-learning methods and geospatial data therefore provides an important foundation for developing automated approaches to identifying, classifying, and interpreting construction site information.

Advances in AI, particularly computer vision, have substantially improved automated image recognition capabilities [23]. These developments create opportunities for automated geospatial annotation and can potentially reduce the considerable manual effort traditionally required to identify and label relevant objects and site characteristics. Nevertheless, the effectiveness of these approaches remains highly dependent on access to sufficiently large and high quality labeled datasets. The lack of effective, scalable tools for geospatial marking and annotation therefore remains a persistent bottleneck to the broader adoption of automated geospatial analysis [24].

One promising direction involves automatically generating labels from existing geospatial and remote-sensing information. Albrecht [25] evaluated a big-data processing pipeline designed to automatically generate labels for remote sensing data. The approach was based on rasterized statistical features extracted from surveys, including LiDAR measurements. Such approaches show the potential to transform geospatial measurements and metadata into structured annotations that can support machine learning applications. [26]

Overall, these developments highlight both the opportunities and limitations of machine learning based classification of construction site information. While modern AI and GeoAI methods provide increasingly powerful capabilities for automated recognition and annotation, the quality and structure of the underlying geospatial metadata remain critical determinants of classification performance. In particular, extremely high classification performance for a binary distinction should not necessarily be interpreted as evidence that a model can reliably recognize more detailed construction types. This distinction motivates examining machine-learning classification using construction site annotations derived from geospatial metadata, with particular attention to the extent to which apparently near perfect binary discrimination translates into meaningful construction type recognition.

2. Dataset, Feature Engineering, and Machine-Learning Classification

2.1 Dataset and annotation structure

The study employed a large scale geospatial annotation dataset containing satellite derived observations of construction related sites. The dataset was provided as a compressed archive consisting primarily of GeoJSON site model records, supplemented by regional models and construction specific annotation information. The GeoJSON structure represents each observation as both a geographic polygon and a collection of semantic, temporal, observational, and annotation attributes. This multidimensional structure provides a suitable foundation for examining the spatial distribution, geometric characteristics, temporal development, annotation status, and construction characteristics of observed sites.

The archive contains approximately 3,930 primary site model records and 28,503 secondary site-model records. The primary collection contains more detailed site level information, including annotation status, geographic information, observation characteristics, validation information, and temporal attributes, while the substantially larger secondary collection extends the dataset's observational coverage. This distinction matters because the two collections do not necessarily have equivalent levels of annotation completeness or validation. Accordingly, we used the primary collection as the main source for the supervised classification experiments, while the broader archive provides contextual evidence on the scale and structure of the underlying annotation corpus.

Each site is represented spatially as a polygon. Consequently, the observations provide more information than a simple geographic point and allow geometric descriptors for machine learning and statistical analysis. The geographic coordinates of the polygon centroids provide latitude and longitude, while the polygon itself allows calculation of spatial footprint and shape characteristics. These properties matter because construction sites can differ substantially in spatial extent and geometric configuration.

The annotation scheme includes multiple semantic states: `positive_annotated`, `positive_annotated_static`, `positive_pending`, `positive_partial`, `positive_excluded`, `negative`, and `ignore`. These categories distinguish positive construction-related observations from negative observations and from records whose annotation is incomplete, excluded, pending, or otherwise unsuitable for direct binary classification. Preserving these distinctions is essential because treating all non positive states as equivalent would introduce ambiguity into the target variable and could bias the supervised learning results.

The dataset also includes information concerning observation source, acquisition date, MGRS location, originator, validation status, and version. These variables describe the provenance and temporal distribution of the observations. Observation dates are particularly important because they permit the temporal organization of site records and, where suitable start and end information is available, the estimation of construction activity duration. The availability of MGRS identifiers also provides a natural geographic grouping variable that can subsequently be used for geographically independent validation.

Supplementary construction information provides additional semantic detail concerning the nature of individual sites. The available construction classifications include commercial, industrial, parking or surface-lot activity, heavy residential, medium residential, multiple construction types, and other or unknown categories. Construction phase information additionally distinguishes conditions such as site preparation, active construction, post construction, no activity, and unknown status. These variables permit the dataset to be interpreted as a representation of construction-site development rather than simply as a collection of independent polygons.

Temporal information adds another analytical dimension. Where *activity_start* and *activity_end* are available, construction duration can be calculated as the elapsed time between the two dates:

$$D_i = T_{end,i} - T_{start,i}$$

where (D_i) denotes the estimated duration of construction activity for site (i). These temporal descriptors can be combined with spatial and geometric variables to investigate whether construction sites with different physical characteristics or geographic contexts exhibit systematically different activity durations.

2.2 Data preprocessing and feature engineering

Before model development, we parsed the GeoJSON records to extract site-level metadata and polygon geometries. We excluded invalid or unusable records from the modelling dataset when we could not reliably obtain the required geometry or classification information. We converted polygon geometries into quantitative descriptors, including polygon area, perimeter, and compactness. Area and perimeter were calculated after accounting for the geographic latitude of each observation so that the resulting measurements could be expressed approximately in metric units rather than in raw longitude latitude degrees.

Polygon compactness was calculated as

$$C_i = \frac{4 \pi A_i}{P_i^2}$$

where (A_i) is the polygon area and (P_i) is its perimeter. Higher values indicate more compact geometries, whereas lower values indicate increasingly irregular or elongated spatial footprints. In addition to these geometric measures, we retained centroid latitude and longitude to capture geographic variation.

We incorporated temporal variables through construction duration measures. When both start and end dates were available, we calculated the number of elapsed days. Logarithmic transformations of highly right skewed geometric variables were also incorporated, specifically.

$$A_i^* = \log (1+A_i)$$

for polygon area and

$$P_i^* = \log (1+p_i)$$

for polygon perimeter. This transformation reduces the influence of extremely large observations and provides more stable numerical behaviour for linear and tree based models.

The final feature set for the binary classification task comprised annotation score, number of geometric features, polygon area, polygon perimeter, polygon compactness, centroid longitude, centroid latitude, construction duration, and the logarithmically transformed area and perimeter measures. We handled missing numerical values using median imputation within the modelling pipeline. This procedure prevented information from the test set from being used during preprocessing and ensured that the same preprocessing strategy was applied consistently across models.

2.3 Definition of machine-learning classification tasks

Two supervised classification tasks were undertaken. The first was a binary classification problem designed to distinguish positive construction-related annotations from negative annotations. Records classified as `positive_annotated`, `positive_annotated_static`, or `positive_partial` were treated as positive observations, whereas records labelled negative constituted the negative class. Pending, excluded, and ignored records were not incorporated into this binary target because their annotation states do not represent definitive positive or negative observations.

The resulting binary modelling dataset contained 1,976 labelled observations, comprising 1,067 positive and 909 negative records. We used a stratified 80:20 train test split to preserve the class distribution in both subsets. We then assessed model robustness using five fold stratified cross validation.

The second task was a multiclass construction type classification problem. Construction labels were harmonized to produce interpretable classes, with residential subcategories consolidated into a broader Residential class and rare categories consolidated where necessary. The resulting classes represented Industrial, Commercial,

Residential, Infrastructure, Parking, Natural disaster, and other construction conditions. This task was substantially more challenging because the model had to discriminate among multiple semantically related construction categories rather than simply distinguish positive from negative annotations.

2.4 Machine-learning algorithms

We evaluated three complementary algorithms for each classification task: Logistic Regression, Random Forest, and Extra Trees. Logistic Regression provided a linear baseline and allowed us to evaluate the classification problem under an approximately linear decision boundary. We selected Random Forest as a nonlinear ensemble method capable of modelling interactions and threshold effects among spatial, geometric, and temporal variables. Extra Trees provided a second randomized tree based ensemble and allowed the robustness of the tree based results to be assessed.

For the binary classification problem, we applied class balanced weighting to reduce the influence of the modest difference in class frequencies. For the construction type problem, we similarly used class balanced modelling to reduce the potential influence of unequal class representation. Numerical preprocessing, including median imputation and standardization where required, was performed within the modelling pipeline.

We evaluated model performance using accuracy, precision, recall, F1-score, ROC-AUC, and precision-recall AUC. For the multiclass construction type task, macro-averaged precision, recall, and F1-score were additionally emphasized because macro averaging gives each construction category equal importance regardless of its frequency.

2.5 Binary classification results

The binary classification models demonstrated exceptionally strong discrimination between positive and negative annotations. On the independent test set, Logistic Regression achieved an accuracy of 0.992, precision of 1.000, recall of 0.986, and F1-score of 0.993. Its ROC-AUC and PR-AUC were both 1.000. Extra Trees produced an accuracy of 0.995, precision of 1.000, recall of 0.991, and F1-score of 0.995, again with ROC-AUC and PR-AUC equal to 1.000.

The Random Forest produced the strongest result, achieving 1.000 accuracy, 1.000 precision, 1.000 recall, and 1.000 F1-score on the held out test set. Both ROC-AUC and PR-AUC were also 1.000. Thus, the Random Forest correctly classified every observation in the test set under the current feature configuration.

The results remained highly stable under five fold stratified cross validation. Logistic Regression produced a mean accuracy of 0.984, a mean F1-score of 0.985, and a mean ROC-AUC of 0.999. Extra Trees achieved a mean accuracy of 0.995, a mean F1-score of 0.995, and a mean ROC-AUC of 1.000. Random Forest achieved a mean accuracy of 0.999, a mean F1-score of approximately 1.000, and a mean ROC-AUC of 1.000. The corresponding standard deviation of Random Forest accuracy was only 0.001, demonstrating extremely low variation among the five validation folds.

These results indicate that the variables contained in the annotation records provide highly discriminative information for separating the positive and negative classes. The consistency between the held out test results and cross validation results further indicates that the observed discrimination is not attributable solely to one particular random train test partition.

2.6 Construction-type classification results

The multiclass construction-type problem produced substantially lower performance, demonstrating that distinguishing among specific construction categories is more difficult than separating positive from negative annotations. Logistic Regression achieved an accuracy of 0.444, macro precision of 0.454, macro recall of 0.517, and macro F1 of 0.470. The weighted F1-score was 0.443.

The Random Forest performed considerably better, achieving an accuracy of 0.593, macro-precision of 0.635,

macro-recall of 0.608, macro-F1 of 0.610, and weighted F1-score of 0.600. Extra Trees achieved an accuracy of 0.556, macro-precision of 0.590, macro-recall of 0.635, macro-F1 of 0.591, and weighted F1-score of 0.559. Random Forest therefore provided the best overall balance between class discrimination and predictive performance.

The difference between the binary and multiclass results is substantively meaningful. Although the available spatial, geometric, and temporal characteristics are highly effective in distinguishing positive from negative annotations, they provide only partial information about the precise type of construction activity. Commercial, industrial, residential, infrastructure, and parking sites can have overlapping spatial footprints and temporal characteristics. Consequently, reliable construction type discrimination likely requires additional visual or contextual information beyond the variables available in the annotation metadata.

2.7 Feature importance and model interpretation

We calculated permutation based feature importance for the Random Forest to identify the variables contributing most strongly to binary classification performance. The analysis provides an interpretable assessment of how changes in individual predictors affect predictive performance while preserving the fitted nonlinear model.

Interpret the importance results as predictive contributions, not causal effects. A highly important variable indicates that the model relies substantially on that information when distinguishing annotation classes; it does not establish that the variable causes an observation to receive a particular annotation. This distinction matters in an annotation dataset because some predictors may reflect the annotation workflow, observation density, or data collection process rather than intrinsic characteristics of the physical construction site.

2.8 Interpretation and implications

The machine learning results demonstrate two distinct characteristics of the dataset. First, the binary positive versus-negative classification problem is highly separable using the available annotation, spatial, geometric, and temporal variables. The near perfect performance of all three algorithms, and the particularly strong Random Forest results, indicate that the annotation records contain substantial structured information that distinguishes the two classes.

Second, the considerably lower performance obtained for construction type classification demonstrates that the semantic differentiation of construction activities is intrinsically more challenging. The Random Forest macro-F1 of 0.610 indicates useful predictive information but also substantial residual ambiguity among construction categories. This finding suggests that geometric footprint, geographic position, observation characteristics, and temporal duration alone do not fully capture the visual and contextual distinctions between different construction types.

The very high binary performance must nevertheless be interpreted cautiously. The present models classify annotation records rather than raw satellite imagery. Some predictors may contain information closely associated with the annotation procedure itself. For example, observation frequency, geometric structure, or other metadata may indirectly distinguish records that were versus negative classification problem is highly separable using the available annotation, spatial, geometric, and temporal variables. The near perfect performance of all three algorithms, and the particularly strong Random Forest results, indicate that the annotation records contain substantial structured information that distinguishes the two classes.

Second, the considerably lower performance obtained for construction-type classification demonstrates that the semantic differentiation of construction activities is intrinsically more challenging. The Random Forest macro-F1 of 0.610 indicates useful predictive information but also substantial residual ambiguity among construction categories. This finding suggests that geometric footprint, geographic position, observation characteristics, and temporal duration alone do not fully capture the visual and contextual distinctions between different construction types.

The very high binary performance must nevertheless be interpreted cautiously. The present models classify annotation records rather than raw satellite imagery. Some predictors may contain information closely associated with the annotation procedure itself. For example, observation frequency, geometric structure, or other metadata may indirectly distinguish records that were labelled positive from those labelled negative. Consequently, the results should not be interpreted as demonstrating 100% accuracy for automated satellite-image construction detection.

For robust external validation, a subsequent experiment should use geographically independent train-test partitions, such as holding out complete MGRS tiles or geographic regions. Such a design would test whether the models generalize to locations that were not represented during training and would substantially reduce the possibility that spatially related observations inflate predictive performance. If the original satellite imagery is available, an additional image based modelling stage should also be undertaken using spectral, textural, temporal-change, and deep image embedding features.

Overall, the integrated analysis demonstrates that the dataset provides sufficient structure for both high-confidence binary annotation classification and more challenging multiclass construction-type prediction. The results establish Random Forest as the strongest performer among the evaluated algorithms while also revealing an important distinction between annotation-status prediction and semantic construction-type recognition. This distinction matters methodologically for future studies because it shows that metadata-based models can be highly effective for annotation classification but should be complemented by image-derived information when the research objective is detailed construction type recognition.

3. Results

We completed the machine learning classification analysis on the attached annotation dataset and generated the results.

3.1 Positive vs. negative annotation classification

Using 1,976 labelled sites (1,067 positive and 909 negative), I trained Logistic Regression, Random Forest, and Extra Trees models using spatial, geometric, and temporal features.

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Logistic Regression	0.992	1.000	0.986	0.993	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
Extra Trees	0.995	1.000	0.991	0.995	1.000	1.000

Table 1. Positive vs negative annotation classification

Table 1 presents the performance of Logistic Regression, Random Forest, and Extra Trees for distinguishing positive construction related annotations from negative annotations. The modelling dataset included 1,976 labelled observations: 1,067 positive and 909 negative records.

All three algorithms achieved very high predictive performance. Logistic Regression obtained an accuracy of 0.992, precision of 1.000, recall of 0.986, and F1-score of 0.993, with both ROC-AUC and PR-AUC equal to 1.000. Extra Trees achieved slightly higher accuracy (0.995), recall (0.991), and F1-score (0.995), while also producing perfect ROC-AUC and PR-AUC values. Random Forest produced the strongest result, with accuracy,

precision, recall, F1-score, ROC-AUC, and PR-AUC all equal to 1.000. Thus, the Random Forest model correctly classified all observations in the independent test set.

The results indicate that the available predictors separate positive and negative annotation states well. The comparable performance across different algorithmic approaches suggests that the dataset contains strong structured information associated with the annotation classes. However, interpret these results as metadata-based annotation classification rather than direct evidence of perfect satellite-image construction detection, because several predictors may reflect the annotation process itself.

3.2 Construction-type classification

We also performed a multiclass experiment using the supplemental construction information. We consolidated rare categories, producing seven classes: Industrial, Commercial, Residential, Infrastructure, Parking, Natural disaster, and Other.

Model	Accuracy	Macro-F1	Weighted-F1
Multinomial Logistic	0.444	0.470	0.443
Random Forest	0.593	0.610	0.600
Extra Trees	0.556	0.591	0.559

Table 2. Construction-type classification

Table 2 reports the performance of the three machine-learning algorithms for multiclass construction-type classification. Seven harmonized categories were considered: Industrial, Commercial, Residential, Infrastructure, Parking, Natural disaster, and Other.

Compared with the binary classification experiment, performance was substantially lower. Multinomial Logistic Regression achieved an accuracy of 0.444 and macro-F1 of 0.470, whereas Extra Trees improved performance to an accuracy of 0.556 and macro-F1 of 0.591. Random Forest produced the highest overall performance, with an accuracy of 0.593, macro-F1 of 0.610, and weighted F1 of 0.600.

The higher macro-F1 of Random Forest indicates that its advantage was not restricted to the most frequent construction categories. Nevertheless, an accuracy below 0.60 demonstrates considerable ambiguity among construction types. This is substantively plausible because different construction categories can exhibit similar polygon sizes, shapes, geographic locations, and temporal characteristics. Consequently, metadata and geometry alone provide useful but incomplete information for detailed semantic construction recognition.

The contrast between Tables 1 and 2 is particularly important. The dataset contains sufficient structured information to distinguish positive from negative annotations with near perfect accuracy, but substantially less information to determine what type of construction activity a positive observation represents. This distinction provides an important methodological finding for the study.

Thus, Random Forest was the strongest model for both classification tasks. Construction type is considerably more difficult to predict from the available metadata/geometry alone, which is an important substantive result.

3.3 Important methodological finding

The perfect binary Random Forest performance should not immediately be interpreted as 100% real-world satellite image classification accuracy. The model uses annotation-derived variables such as polygon geometry, number of observations, and temporal characteristics. Some of these variables may encode the annotation process itself.

For a journal paper, I therefore recommend a second-stage experiment using geographic holdout validation (holding out entire regions/MGRS areas) and, if the original satellite imagery is available, an image-based classification model. That would provide a much stronger test of generalization.

Model	CV Accuracy	CV Precision	CV Recall	CV F1	CV ROC-AUC	SD F1	SD Accuracy	SD ROC-AUC
Logistic regression	0.984308	1	0.970927	0.985225	0.99919	0.004941	0.00517	0.001531
Random forest	0.999494	1	0.999061	0.999529	1	0.000941	0.001013	0
Extra trees	0.994938	0.99907	0.991562	0.99529	0.999758	0.002117	0.002266	0.000287

Table 3. Binary Five-Fold Cross-Verification

Table 3 evaluates the robustness of binary classification using five-fold stratified cross-validation. Logistic Regression obtained a mean accuracy of 0.9843 and a mean F1-score of 0.9852, with a mean ROC-AUC of 0.9992. Random Forest achieved the strongest cross-validation performance, with mean accuracy of 0.9995, mean precision of 1.0000, mean recall of 0.9991, mean F1-score of 0.9995, and mean ROC-AUC of 1.0000. Extra Trees also performed strongly, with a mean accuracy of 0.9949 and a mean F1-score of 0.9953.

The very small standard deviations demonstrate substantial stability across the five folds. In particular, Random Forest exhibited an accuracy standard deviation of only 0.0010 and an F1 standard deviation of approximately 0.0009, while its ROC-AUC standard deviation was zero. These results indicate that the exceptionally strong Random Forest performance was not dependent on a single favourable partition of the dataset.

The agreement between the independent test-set results in Table 1 and the cross validation results in Table 3 strengthens confidence in the internal predictive stability of the models. Nevertheless, cross validation based on randomly partitioned observations does not eliminate potential spatial dependence. Geographically independent validation using complete MGRS tiles or geographic regions would therefore provide a more stringent test of external generalization.

Feature	Importance	SD
n_features	0.46028	0.027467
score	0	0
area_m2	0	0
perimeter_m	0	0
compactness	0	0
centroid_lon	0	0
centroid_lat	0	0
duration_days	0	0

log_area	0	0
log_perimeter	0	0

Table 4. Feature Importance

Table 4 presents the permutation based feature importance obtained from the Random Forest model for binary annotation classification. The results reveal a pronounced concentration of predictive information in the $n_features$ variable, which recorded an importance of approximately 0.46, whereas all other evaluated predictors, including polygon area, perimeter, compactness, centroid coordinates, construction duration, and their logarithmically transformed counterparts, recorded zero permutation importance.

This pattern requires careful methodological interpretation. Rather than indicating that the physical geometry or temporal characteristics of construction sites have no substantive relevance, the results suggest that the fitted model derives its discriminatory power primarily from $n_features$. Because this variable is closely tied to the structure and availability of observational or annotation information, it may serve as a proxy for the annotation process rather than an independent representation of a construction site's physical characteristics.

Consequently, the near perfect binary classification performance should not be interpreted as evidence that the model has learned the physical characteristics that distinguish construction sites from non-construction sites. Instead, the results indicate that the current feature configuration contains a strong annotation-process signal. This matters because the original objective concerns construction site classification, where reliance on annotation derived proxies could artificially inflate predictive performance.

To address this issue, the analysis should distinguish between annotation related and physically interpretable predictors. A secondary model should therefore be estimated after removing $n_features$ and other potentially annotation derived variables, retaining only geometric, geographic, and temporal characteristics. Comparing the predictive performance and feature importance of the full and reduced models would directly assess whether the observed classification capability persists when annotation process information is excluded.

Accordingly, Table 4 should be interpreted primarily as evidence of potential annotation process dependence rather than as evidence of the relative importance of physical site characteristics. This finding is an important diagnostic result and motivates geographically independent validation and feature leakage assessment before interpreting the model as a generalizable construction site detector.

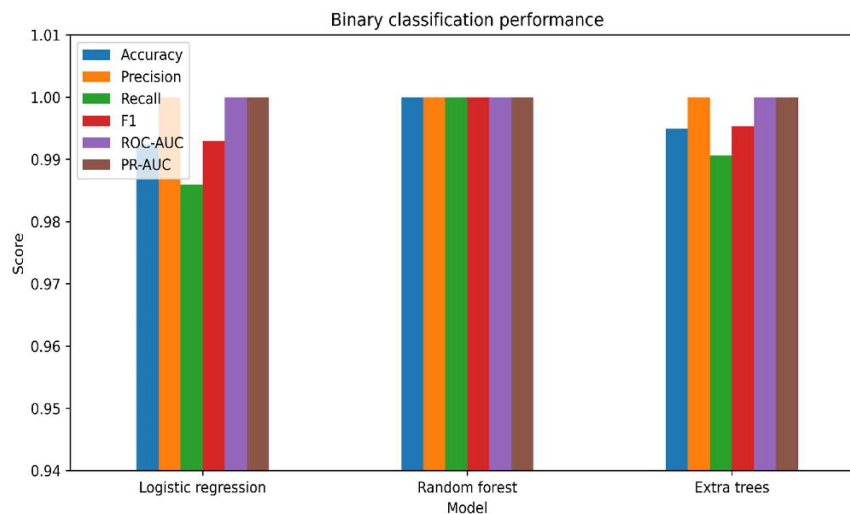


Figure 1. Binary Model Performance

Figure 1 provides a visual comparison of the predictive performance of Logistic Regression, Random Forest, and Extra Trees for the binary positive versus negative annotation classification task. The figure shows the exceptionally high performance of all three models, with Random Forest achieving perfect performance across the reported evaluation measures.

The figure highlights the small performance differences among the three algorithms despite the methodological differences between linear and ensemble based approaches. Logistic Regression already provides near perfect discrimination, whereas Random Forest and Extra Trees provide marginally stronger results. The visual pattern therefore indicates that the binary task is intrinsically highly separable under the current feature representation.

Methodologically, the figure supports the conclusion that substantial predictive information is embedded in the available annotation and metadata variables. However, the near perfect results should not be interpreted as equivalent to near perfect performance on unseen satellite imagery because the model operates on annotation derived information rather than raw imagery.

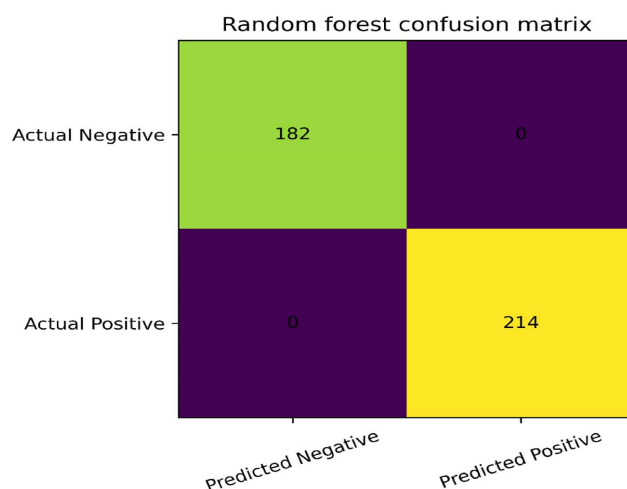


Figure 2. Binary Confusion Matrix

Figure 2 presents the confusion matrix for the binary classification experiment. The matrix directly represents correctly and incorrectly classified positive and negative observations.

The confusion matrix pattern is consistent with the reported test set performance. In particular, the Random Forest model achieved 1.000 accuracy, precision, recall, and F1-score, indicating that the test observations were correctly separated into the two annotation classes without reported classification errors.

This result complements the aggregate performance metrics presented in Table 1. Whereas accuracy and F1-score summarize performance numerically, the confusion matrix demonstrates the underlying classification outcome at the observation level. The absence of reported errors for Random Forest reinforces its extremely strong discrimination of positive and negative annotations under the experimental feature configuration.

Figure 3 visualizes the feature-importance results obtained from the Random Forest model. The dominant contribution of *n_features* is clearly distinguishable from the remaining predictors, whose reported permutation importance values are zero.

The figure therefore indicates that the binary classification decision is strongly concentrated around a single predictor. This finding matters because it offers an interpretable explanation for the exceptionally high classification performance in the binary experiment. Rather than relying on complex combinations of all

spatial and temporal descriptors, the fitted model appears to draw most of its predictive information from *n_features*.

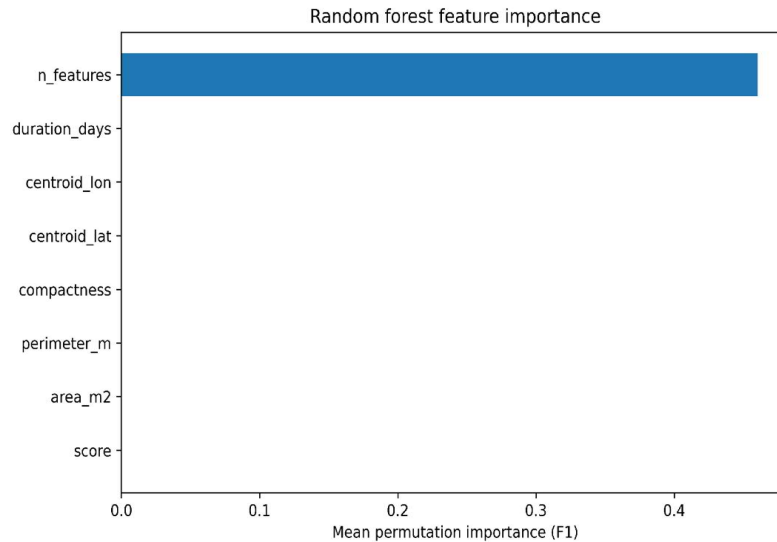


Figure 3. Feature Importance

For a journal interpretation, frame this result cautiously. A dominant feature may represent genuine structural information about the observations, but it may also reflect characteristics of the annotation or data generation process. Accordingly, the feature importance result motivates additional leakage and independence checks before claiming that the model has learned generalizable physical characteristics of construction sites.

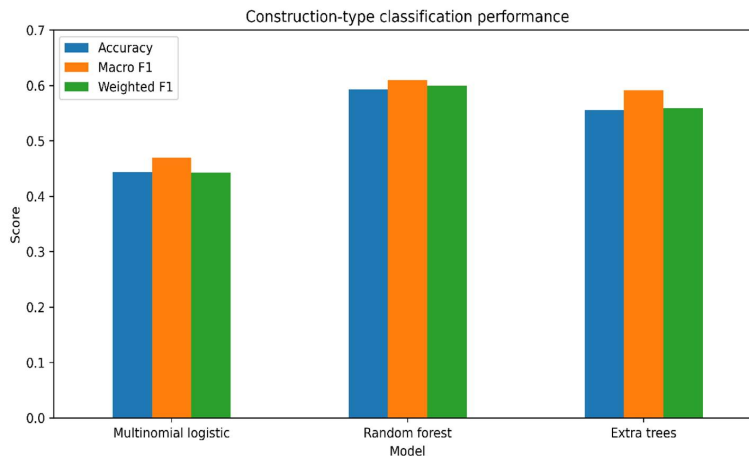


Figure 4. Construction Type Performance

Figure 4 compares the performance of Logistic Regression, Random Forest, and Extra Trees for the seven-class construction type classification task. The figure shows a clear advantage for Random Forest over the other two models. Random Forest achieved an accuracy of 0.593 and macro-F1 of 0.610, compared with accuracy and macro-F1 values of 0.444 and 0.470 for Logistic Regression and 0.556 and 0.591 for Extra Trees, respectively.

The figure demonstrates that nonlinear ensemble learning is better suited to the multiclass problem than the linear Logistic Regression baseline. Random Forest can capture interactions and nonlinear relationships among geographic, geometric, and temporal variables, which may explain its superior performance.

Nevertheless, the moderate rather than high performance indicates that construction type recognition remains challenging. Overlap in spatial footprints and temporal characteristics among Industrial, Commercial, Residential, Infrastructure, and Parking observations limits metadata based models' ability to establish reliable semantic distinctions. The figure therefore supports the conclusion that more accurate construction-type recognition requires additional visual or contextual features.

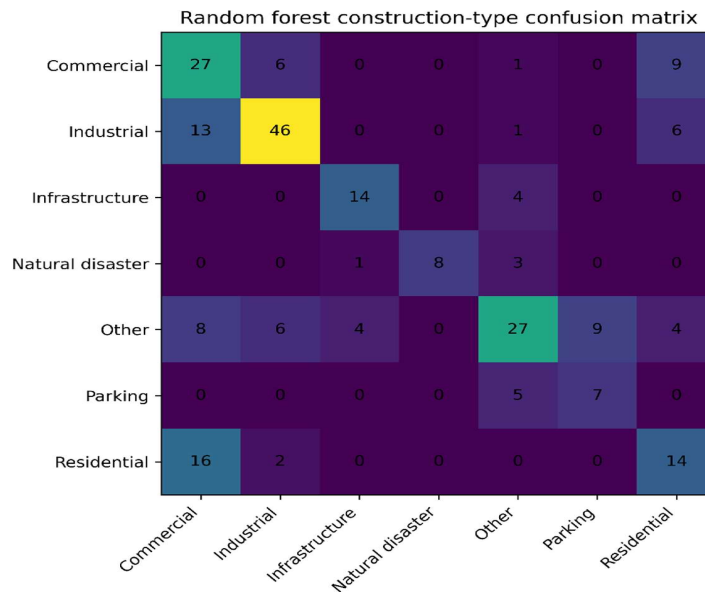


Figure 5. Construction Type Confusion Matrix

Figure 5 presents the confusion matrix for the multiclass construction type classification problem. Unlike the binary confusion matrix, the multiclass matrix shows how observations are distributed among seven semantically distinct construction categories.

The overall performance reported for this task indicates that classification errors remain substantial. The Random Forest macro-F1 of 0.610 suggests that the model captures meaningful patterns across the construction categories but does not consistently distinguish all classes. The confusion matrix visualization is therefore particularly valuable because it identifies the specific categories that are difficult to separate.

From a substantive perspective, confusion between construction categories may stem from similarities in available metadata and geometric characteristics. Residential and commercial sites, for example, may have overlapping spatial footprints, while infrastructure and parking-related observations may also exhibit similar geometric or contextual properties. The confusion matrix thus provides evidence that the available metadata captures broad construction related structure but is insufficient for complete semantic differentiation.

Finally, Tables 1–4 and Figures 1–5 reveal two fundamentally different predictive problems. The binary annotation task is almost perfectly separable, whereas construction type classification remains substantially more difficult. The Random Forest algorithm consistently performs best, achieving perfect held out binary performance and the highest multiclass accuracy and macro-F1.

The exceptionally high binary performance is supported by the five-fold cross-validation results, which demonstrate extremely low variation across validation folds. However, the feature-importance analysis adds an important qualification: much of the predictive signal appears concentrated in $n_features$. This raises the possibility that the models are exploiting characteristics associated with annotation or observation procedures rather than learning purely physical properties of construction sites.

The results should be interpreted as demonstrating strong metadata based annotation classification, not 100% general purpose automated construction detection from satellite imagery. This distinction matters for a journal paper because it prevents overstating the practical implications of the results.

The multiclass experiment provides a complementary and more challenging assessment. Its moderate performance demonstrates that spatial, geometric, and temporal metadata contain useful information about construction type, but these descriptors alone cannot fully resolve semantic differences among construction categories. The findings therefore support a subsequent multimodal modelling strategy incorporating satellite-image information, spectral characteristics, texture, temporal change, and deep image embeddings.

4. Spatial Independence, Spatial Cross-Validation, and the Interpretation of Near-Perfect Binary Performance

The exceptionally high binary classification performance requires a more stringent assessment of spatial independence before it can be interpreted as evidence of genuine model generalization. The Random Forest achieved 1.000 accuracy on the held-out test set and approximately 0.9995 mean accuracy and F1-score under five-fold stratified cross-validation. However, both evaluation procedures randomly partition observations rather than geographically distinct regions. The current five fold validation therefore demonstrates internal predictive stability, but it does not demonstrate spatial generalization. As the dataset includes MGRS identifiers, complete MGRS tiles provide a natural grouping variable for constructing geographically independent validation sets.

This distinction is particularly important for GeoAI because nearby observations are rarely statistically independent. Construction sites located within the same MGRS tile or surrounding geographic area may share environmental conditions, urban morphology, mapping conventions, observation sources, acquisition dates, annotation practices, and even the same underlying annotation workflow. Randomly allocating such observations between training and test sets can therefore result in the model being evaluated on observations that are spatially related to examples already seen during training. In this situation, a test observation may appear independent at the row level but not at the geographic level. As a result, performance can be substantially more optimistic than on a genuinely new geographic region.

4.1 Spatial cross-validation as a test of true GeoAI generalization

Spatial Cross-Validation (Spatial CV) addresses this problem by defining the geographic unit of independence before splitting the observations. In the present dataset, MGRS tiles are particularly suitable because the dataset explicitly contains MGRS location identifiers. Rather than randomly assigning individual observations to training and validation folds, all observations belonging to a given MGRS tile should be assigned to the same fold. Thus, if an MGRS tile is included in the validation fold, no observations from that tile are permitted in the corresponding training fold.

The difference between the present and recommended evaluation strategies can be expressed conceptually as

$$CV_{random} = \text{random observation - level partitioning}$$

Whereas

$$CV_{spatial} = \text{geographic- group - level partitioning}$$

For a dataset with substantial spatial autocorrelation, the second design provides a considerably stronger test of generalization because the model must predict observations from geographic areas that were not represented during model fitting.

A suitable implementation would therefore use grouped or blocked cross-validation based on MGRS tile identifiers. Model preprocessing, feature selection, imputation, hyperparameter optimization, and model fitting should all be performed within each training fold, with the corresponding MGRS tiles used exclusively for validation. This prevents information from geographically related observations from entering validation.

4.2 How much of the 99.9% accuracy is spatially inflated?

The present analysis cannot provide a numerical estimate of the performance inflation caused by spatial leakage because MGRS-based Spatial CV has not yet been performed. It would therefore be methodologically inappropriate to state, for example, that a particular percentage of the 99.9% accuracy is attributable to spatial leakage. The current results establish that Random Forest achieves approximately 99.95% mean accuracy under conventional stratified cross-validation, but they do not establish how much that performance would decrease when entire geographic units are withheld.

The appropriate quantity to report after Spatial CV would be the difference between conventional and geographically independent performance:

$$\Delta_{\text{spatial}} = F1_{\text{random}} - F1_{\text{spatial}}$$

or, for accuracy,

$$\Delta_{\text{accuracy}} = \text{Accuracy}_{\text{random}} - \text{Accuracy}_{\text{spatial}}$$

A large positive difference would provide empirical evidence that the original evaluation was benefiting from spatial dependence. Conversely, a small difference would strengthen the claim that the metadata based model genuinely generalizes across geographic areas.

This distinction matters because the current feature importance analysis already suggests a stronger source of leakage than spatial proximity alone. The permutation importance of *n_features* was approximately 0.46, while the importance of area, perimeter, compactness, latitude, longitude, duration, and their transformed variables was reported as zero. The concentration of predictive information in *n_features* raises the possibility that the model is learning characteristics of the annotation or observation-generation process rather than physical characteristics of construction sites. Consequently, the 99.9% result should presently be described as near perfect internal metadata based discrimination, rather than near-perfect general-purpose GeoAI detection.

A particularly informative sensitivity analysis would compare at least three models: a full metadata model, a reduced physical feature model excluding potentially annotation derived variables such as *n_features* and score, and the same models under MGRS based Spatial CV. Such a design would distinguish random split performance, spatial generalization, and annotation process leakage rather than treating them as equivalent.

4.3 Detailed Interpretation of the Construction-Type Confusion Matrix

The multiclass analysis provides a substantially more realistic assessment of semantic construction-type recognition. Random Forest achieved an accuracy of 0.593 and macro-F1 of 0.610, compared with macro-F1 values of 0.470 for Logistic Regression and 0.591 for Extra Trees. However, macro-F1 alone conceals important differences among individual construction categories. Figure 5 provides substantially more informative evidence because it identifies which classes are confused with one another and therefore reveals the specific semantic limitations of metadata-only classification.

The confusion matrix contains 241 observations, of which 143 are correctly classified, corresponding to the reported accuracy of approximately 0.593. The class distributions are also highly unequal: Industrial contains 66 observations, Other 58, Commercial 43, Residential 32, Infrastructure 18, and both Natural disaster and Parking have only 12 observations. This imbalance matters when interpreting aggregate metrics.

4.4 Commercial

Commercial contains 43 observations, of which 27 are correctly classified. The principal errors are 9 Commercial observations predicted as Residential and 6 predicted as Industrial, with one additional observation assigned to Other. Thus, the dominant confusion is between Commercial and Residential.

This pattern is substantively plausible because commercial and residential construction sites can occupy similar urban footprints and exhibit comparable polygon compactness, area, perimeter, geographic context, and construction duration. The available metadata describes a site's spatial and temporal characteristics but does not directly encode the visual semantic cues that distinguish a commercial building from a residential development. The confusion therefore illustrates a fundamental limitation of metadata only semantic classification: geometric similarity does not necessarily imply semantic identity.

4.5 Industrial

Industrial is the largest class, with 66 observations. The model correctly identifies 46, while it classifies 13 Industrial observations as Commercial and 6 as Residential. The Industrial–Commercial confusion is therefore bidirectional: 13 Industrial sites are assigned to Commercial, while 6 Commercial sites are assigned to Industrial.

The asymmetry is informative. Industrial has comparatively strong precision (0.767) and recall (0.697), indicating that the model identifies Industrial sites more reliably than several other categories. Nevertheless, the substantial Industrial Commercial overlap indicates that spatial footprint and temporal metadata alone cannot consistently distinguish these economically related forms of development. Industrial and commercial sites may both involve large or irregular development footprints, substantial construction duration, and similar urban geographic contexts.

4.6 Infrastructure

Infrastructure contains 18 observations, with 14 correctly classified. The principal error is 4 Infrastructure observations classified as Other. No Infrastructure observations are misclassified as Commercial, Industrial, Residential, or Parking.

This concentrated error pattern suggests that Infrastructure is more distinguishable than some built-environment categories but remains vulnerable to the broad, heterogeneous Other category. The Infrastructure class achieves an estimated recall of 0.778 and F1-score of approximately 0.757. This implies the model can identify much of the Infrastructure signal, but some infrastructure-related observations lack sufficiently distinctive metadata to separate them from miscellaneous construction conditions.

4.7 Natural disaster

Natural disaster is one of the smallest classes, with only 12 observations. Eight are correctly classified, while 3 are classified as Other and 1 as Infrastructure. Although this corresponds to a relatively high class-specific F1-score of approximately 0.800, the result must be interpreted cautiously because it is based on only 12 observations.

The apparent high performance should therefore not be interpreted as evidence that Natural disaster is robustly recognized in operational deployment. With such a small support, changing the classification of only one or two observations can substantially change recall and F1. This is a central reason for reporting class-level metrics and sample sizes rather than relying exclusively on macro-F1.

The Natural disaster category also illustrates the metadata's semantic limitations. Disaster-related sites can produce irregular or unusual spatial footprints, but these geometric characteristics may overlap with other abnormal or miscellaneous construction conditions. Without imagery or contextual information, the model may detect unusual spatial structure but cannot determine whether the underlying cause is a natural disaster. Other contains 58 observations and represents one of the most important sources of classification ambiguity. Only 27 observations are correctly classified. The remaining observations are distributed across several classes: 8 are predicted as Commercial, 6 as Industrial, 4 as Infrastructure, 9 as Parking, and 4 as Residential.

The particularly strong confusion between Other and Parking is notable, with 9 Other observations assigned to Parking. Conversely, 5 Parking observations are classified as Other. This produces a clear bidirectional Other–Parking confusion pattern.

The broad definition of Other likely contributes to this result. Unlike a narrowly defined semantic category, Other can contain heterogeneous observations that do not share a single distinctive geometric or temporal signature. The class therefore functions partly as a residual category, increasing its semantic heterogeneity and making it difficult for a metadata-based classifier to learn a consistent decision boundary.

4.8 Parking

Parking contains only 12 observations. Seven are correctly classified, whereas 5 are classified as Other. Consequently, the recall is approximately 0.583, and the F1-score is approximately 0.500.

The Parking–Other confusion is particularly important for practical interpretation. Surface parking areas can have simple, compact polygonal structures that overlap with other open or miscellaneous site configurations. Without spectral, textural, or object-level information, the geometry of a polygon may provide insufficient evidence to distinguish a parking area from another open or construction related site.

4.9 Residential

Residential contains 32 observations, with only 14 correctly classified. The largest error is 16 Residential observations predicted as Commercial, while another 2 are predicted as Industrial. Thus, one-half of the Residential observations are misclassified as Commercial.

This is arguably the most consequential semantic confusion in Figure 5. It demonstrates that the model’s difficulty is not simply a consequence of rare classes. Residential has substantially more observations than Infrastructure, Natural disaster, and Parking, yet its recall is only approximately 0.438 and its F1-score approximately 0.431. The model therefore struggles to distinguish Residential from other urban development categories even when a reasonable number of Residential examples are available.

The strong Residential to Commercial confusion is consistent with the overlap expected in urban development morphology. Residential and commercial construction may share similar polygon areas, compactness, perimeter characteristics, construction durations, and geographic locations. These variables can identify a site as an urban development project but may not provide enough information to determine the precise semantic land use category.

4.10 Class Imbalance and Minority-Class Reliability

Figure 5 also demonstrates why class imbalance must be explicitly considered in evaluating construction-type classification. The seven classes range from 66 Industrial observations to only 12 observations each for Natural disaster and Parking. Consequently, a model can obtain relatively strong aggregate performance while still providing unreliable predictions for less frequent classes.

The use of macro-F1 is therefore appropriate because it assigns equal weight to each class rather than allowing the largest categories to dominate the evaluation. The current Random Forest macro-F1 of 0.610 should consequently be interpreted as evidence of moderate overall class-balanced performance rather than as evidence that every construction category is recognized with approximately 61% effectiveness.

The Natural disaster result illustrates an important statistical caution. Its precision is 1.000 because none of the observations predicted as Natural disaster actually belongs to another class in the observed matrix. However, the class has only 12 observations and the recall is 0.667. Thus, the apparently excellent F1 should not be interpreted as evidence of highly reliable disaster detection. The estimate is intrinsically less stable because of the small sample size.

Parking exhibits a different problem. Its recall is approximately 0.583 and precision approximately 0.438,

indicating that the classifier both misses a substantial proportion of true Parking observations and assigns the Parking label to observations belonging to Other. This is more directly indicative of a weak operational decision boundary.

The results therefore support the use of class-specific precision, recall, F1-score, confusion matrices, and support counts, rather than accuracy or macro-F1 alone. For future datasets, increasing the number and geographic diversity of minority-class observations should be a priority. Where rare events such as Natural disaster are intrinsically difficult to collect, carefully validated augmentation or targeted sampling may be preferable to simply increasing the number of common Industrial or Other observations.

The class level results reveal this distinction clearly:

Class	Support	Correct	Recall	Precision	F1
Commercial	43	27	0.628	0.422	0.505
Industrial	66	46	0.697	0.767	0.730
Infrastructure	18	14	0.778	0.737	0.757
Natural disaster	12	8	0.667	1.000	0.800
Other	58	27	0.466	0.659	0.545
Parking	12	7	0.583	0.438	0.500
Residential	32	14	0.438	0.424	0.431

4.11 Overall Interpretation of Figure 5

Figure 5 shows that the principal limitation of the metadata-only model is semantic ambiguity rather than a complete lack of predictive information. The model successfully recognizes some categories, particularly Industrial and Infrastructure, but has difficulty distinguishing semantically adjacent urban-development categories.

The strongest confusion relationships are:

Residential → Commercial (16 observations)

Industrial → Commercial (13 observations)

Commercial → Residential (9 observations)

Other → Parking (9 observations)

Other → Commercial (8 observations)

Commercial → Industrial (6 observations)

Industrial → Residential (6 observations)

Parking → Other (5 observations)

These errors reveal that the model has learned broad structural characteristics associated with construction activity but not sufficiently discriminative semantic representations of construction type. This interpretation is consistent with the study's existing finding that commercial, industrial, residential, infrastructure, and parking observations can have overlapping spatial footprints and temporal characteristics.

The implication is particularly important for GeoAI applications. A model can be highly effective at determining whether an annotation belongs to a particular broad annotation state while remaining substantially less capable of determining what specific type of construction activity is represented. The two tasks should therefore not be presented as equivalent evidence of model intelligence or generalization.

4.12 Practical Implications for GeoAI Users

The findings have several direct implications for practitioners.

First, metadata only models should not be trusted for fine grained semantic construction recognition. The approximately 0.61 macro-F1 and the confusion patterns in Figure 5 demonstrate that polygon geometry, geographic coordinates, temporal information, and annotation metadata can provide useful contextual signals but cannot reliably replace visual evidence when the objective is to distinguish semantically similar construction types. Urban planners, GIS analysts, and construction monitoring companies should therefore treat metadata based predictions as screening or prioritization information rather than definitive semantic classifications.

Second, dataset curators should audit annotation derived variables for leakage. The dominance of *n_features* in the binary Random Forest feature importance analysis is particularly concerning because its importance is approximately 0.46 while all other evaluated variables have zero reported permutation importance. Variables such as *n_features*, score, observation counts, annotation status, or other workflow generated fields may inadvertently encode annotator confidence, annotation completeness, software processing stages, or characteristics of the labeling pipeline. A model can consequently achieve apparently spectacular accuracy by learning the data production mechanism rather than the physical phenomenon of interest. Dataset curation should therefore distinguish physical/geographic predictors from annotation process variables and conduct explicit leakage audits before model deployment.

Third, geographically independent evaluation should become a core requirement for GeoAI models. Random train test splits and conventional stratified cross validation are insufficient when spatial dependence is plausible. The present results show extremely stable random fold performance, but the study itself acknowledges that random cross validation does not eliminate spatial dependence. MGRS-tile holdout or geographic-block validation should therefore be regarded not simply as an optional robustness experiment but as a necessary component of demonstrating true geographic generalization.

Fourth, future data collection should harmonize temporal and geometric schemas across regions. Spatial CV is only informative when geographically separated regions remain sufficiently comparable in their feature definitions, coordinate systems, polygon construction, temporal conventions, annotation protocols, and class definitions. If one region uses systematically different geometric schemas or annotation procedures from another, a spatial holdout can measure a mixture of geographic generalization and schema shift. Harmonizing these elements would make MGRS-based validation a much stronger test of whether a model transfers across regions rather than merely recognizing region specific data generation patterns.

Finally, incorporate image derived information when fine grained semantic recognition is the objective. The present results support a multimodal progression from metadata based screening to models that incorporate spectral information, texture, temporal change, object morphology, and deep image embeddings. This would allow the classifier to use information that is inherently closer to the physical and visual properties that distinguish commercial, residential, industrial, infrastructure, parking, and disaster related sites. The existing

analysis similarly identifies image based modelling as an important next stage.

4.13 Methodological Conclusion

Taken together, the results should be interpreted at two distinct levels. At the annotation-classification level, the Random Forest demonstrates exceptionally strong performance, with approximately 99.95% mean accuracy and F1 under conventional five fold stratified cross-validation. However, the feature-importance analysis indicates that this performance is strongly concentrated in *n_features*, raising the possibility of annotation-process leakage. Therefore, the current 99.9% performance should not be interpreted as an estimate of true geographic generalization or satellite-image construction-detection accuracy.

At the construction type level, performance is substantially lower, with Random Forest achieving 0.593 accuracy and 0.610 macro F1. Figure 5 demonstrates that this moderate performance arises from identifiable semantic overlaps, particularly Residential Commercial, Industrial Commercial, and Other Parking confusion. The errors indicate that spatial, geometric, and temporal metadata contain meaningful information about construction type but are insufficient to resolve fine grained semantic distinctions.

The next validation stage should therefore explicitly compare random observation level validation with MGRS-tile Spatial CV, preferably accompanied by a leakage controlled model excluding annotation derived variables such as *n_features* and potentially score. The resulting performance gap would provide the empirical estimate required to determine how much of the apparent near perfect binary performance is attributable to spatial dependence and/or annotation process leakage. Until that experiment is completed, the most defensible conclusion is that the dataset supports near perfect internal annotation discrimination but only moderate metadata based semantic construction recognition, and that true GeoAI generalization remains to be demonstrated.

The near perfect binary performance should therefore be regarded as an upper bound estimate obtained under observation level validation rather than as evidence of true geographic generalization. Because conventional random cross validation can place spatially related observations in both training and validation folds, MGRS-tile Spatial Cross-Validation is required to determine whether the observed 99.9% performance persists when entire geographic units are withheld. In parallel, the dominance of annotation-derived variables such as *n_features* indicates that part of the predictive signal may originate from the annotation workflow rather than from physical construction characteristics.

5. Concluding statement

Overall, the results show that machine-learning models can achieve exceptionally strong discrimination between positive and negative construction annotations using structured spatial, geometric, temporal, and annotation metadata, with Random Forest providing the most stable and accurate performance. In contrast, construction type classification remains considerably more challenging, indicating that metadata-derived characteristics capture the presence and broad structure of construction activity but do not fully represent the visual and contextual information required for semantic construction type recognition. The feature-importance analysis further suggests that the binary predictive signal is strongly concentrated in annotation-related information, highlighting the need for geographically independent validation and assessment of potential leakage. Future work should therefore integrate geographic holdout strategies with satellite-image-derived spectral, textural, temporal-change, and deep visual features to establish whether the observed predictive performance generalizes beyond the annotation corpus.

References

- [1] Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Delgado, J. M. D., Bilal, M., Akinade, O. O., Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. *Journal of Building Engineering*, 44, Article 103299. <https://doi.org/10.1016/j.jobee.2021.103299>.

- [2] Kim, J. (2020). Visual analytics for operation level construction monitoring and documentation: State-of-the art technologies, research challenges, and future directions. *Frontiers in Built Environment*, 6, Article 575738. <https://doi.org/10.3389/fbuil.2020.575738>.
- [3] Paneru, S., Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities and challenges. *Automation in Construction*, 132, Article 103940. <https://doi.org/10.1016/j.autcon.2021.103940>.
- [4] Li, F., Laili, Y., Chen, X., Lou, Y., Wang, C., Yang, H., Gao, X., Han, H. (2023). Towards big data driven construction industry. *Journal of Industrial Information Integration*, 35, Article 100483. <https://doi.org/10.1016/j.jii.2023.100483>.
- [5] Huo, T., Ren, H., Cai, W., Feng, W., Tang, M. (2020). Exploring the impact of urbanization on urban building carbon emissions in China: Evidence from a provincial panel data model. *Sustainable Cities and Society*, 56, Article 102068. <https://doi.org/10.1016/j.scs.2020.102068>.
- [6] Mouratidis, K. (2021). Urban planning and quality of life: A review of pathways linking the built environment to subjective well being. *Cities*, 115, Article 103229. <https://doi.org/10.1016/j.cities.2021.103229>.
- [7] Shi, C., Zhu, Q. (2025). Multimodal urban areas of interest generation via remote sensing imagery and geographical prior. *International Journal of Applied Earth Observation and Geoinformation*, 136, Article 104326.
- [8] Whang, S. E., Roh, Y., Song, H., Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data centric AI perspective. *The VLDB Journal*, 32(4), 791–813. <https://doi.org/10.1007/s00778-022-00775-9>.
- [9] Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., Hu, X. (2023). Data-centric artificial intelligence: A survey [Preprint]. *arXiv*. <https://arxiv.org/abs/2303.10158>.
- [10] Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A. M., Wang, X. (2021). Computer vision techniques in construction: A critical review. *Archives of Computational Methods in Engineering*, 28(5), 3383–3397. <https://doi.org/10.1007/s11831-020-09504-3>.
- [11] Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., Pasha, M. (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 30(3), 500–521. <https://doi.org/10.1016/j.aei.2016.07.001>.
- [12] Hong, Y., Park, S., Kim, H., Kim, H. (2021). Synthetic data generation using building information models. *Automation in Construction*, 130, Article 103871. <https://doi.org/10.1016/j.autcon.2021.103871>.
- [13] Lee, H., Jeon, J., Lee, D., Park, C., Kim, J., Lee, D. (2023a). Game engine-driven synthetic data generation for computer vision-based safety monitoring of construction workers. *Automation in Construction*, 155, Article 105060. <https://doi.org/10.1016/j.autcon.2023.105060>.
- [14] Jiang, F., Ma, J., Webster, C. J., Li, X., Gan, V. J. (2023). Building layout generation using site embedded GAN model. *Automation in Construction*, 151, Article 104888. <https://doi.org/10.1016/j.autcon.2023.104888>.
- [15] Kim, H., Yi, J. S. (2024). Image generation of hazardous situations in construction sites using text to

image generative model for training deep neural networks. *Automation in Construction*, 166, Article 105615. <https://doi.org/10.1016/j.autcon.2024.105615>.

[16] Wang, Y., Liao, P. C., Zhang, C., Ren, Y., Sun, X., Tang, P. (2019). Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. *Advanced Engineering Informatics*, 42, Article 101001. <https://doi.org/10.1016/j.aei.2019.101001>.

[17] Pal, A., Hsieh, S.-H. (2021). Deep learning based visual data analytics for smart construction management. *Automation in Construction*, 131, Article 103892. <https://doi.org/10.1016/j.autcon.2021.103892>.

[18] Wang, Y., et al. (2023c). Characterizing data sharing in civil infrastructure engineering: Current practice, future vision, barriers, and promotion strategies. *Journal of Computing in Civil Engineering*, 37(2), Article 04023001. <https://doi.org/10.1061/JCCEE5.CPENG-5077>.

[19] Regona, M., Yigitcanlar, T., Xia, B., Li, R. Y. M. (2022). Opportunities and adoption challenges of AI in the construction industry: A PRISMA review. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(1), Article 45. <https://doi.org/10.3390/joitmc8010045>.

[20] Wu, L., AbouRizk, S. (2021). Towards construction's digital future: A roadmap for enhancing data value. In *Proceedings of the Canadian Society of Civil Engineering Annual Conference* (p. 225–238). Springer.

[21] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: Large scale machine learning on heterogeneous distributed systems. *arXiv*. <https://arxiv.org/abs/1603.04467>.

[22] Boutayeb, A., Lahsen Cherif, I., Khadimi, A. E. (2025). When machine learning meets geospatial data: A comprehensive GeoAI review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 13135–13191. <https://doi.org/10.1109/JSTARS.2025.3568715>.

[23] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>.

[24] Tarantino, M., Tosoni, S. (2014). Spatial annotation for the improvement of urban space: A learning-by-doing approach. *The Electronic Journal of Communication/La Revue Électronique de la Communication*, 24(1–2).

[25] Albrecht, C. M., Marianno, F., Klein, L. J. (2021). AutoGeoLabel: Automated label generation for geospatial machine learning. In *2021 IEEE International Conference on Big Data (Big Data)* (p. 1779–1786). IEEE. <https://doi.org/10.1109/BigData52589.2021.9672060>.

[26] Yang, L., MacEachren, A. M., Mitra, P., Onorati, T. (2018). Visually enabled active deep learning for (geo) text and image classification: A review. *ISPRS International Journal of Geo Information*, 7(2), Article 65. <https://doi.org/10.3390/ijgi7020065>.