Environmental Data Science (2025), 4: e10, 1–15
doi:10.1017/eds.2025.2

APPLICATION PAPER

Building occupancy type classification and uncertainty estimation using machine learning and open data

Tom Narock[1], J. Michael Johnson[2], Justin Singh-Mohudpur[2] and Arash Modaresi Rad[3]

[1]Center for Natural, Computer, and Data Sciences, Goucher College, Baltimore, MD, USA
[2]Lynker, Boulder, CO, USA
[3]School of Computing, Boise State University, Boise, ID, USA
Corresponding author: Tom Narock; Email: thomas.narock@goucher.edu

Received: 02 June 2023; **Revised:** 12 December 2024; **Accepted:** 02 January 2025

Keywords: Bayesian neural network; building type classification; flood risk; machine learning; open data

Abstract

Federal and local agencies have identified a need to create building databases to help ensure that critical infrastructure and residential buildings are accounted for in disaster preparedness and to aid the decision-making processes in subsequent recovery efforts. To respond effectively, we need to understand the built environment—where people live, work, and the critical infrastructure they rely on. Yet, a major discrepancy exists in the way data about buildings are collected across the United SStates There is no harmonization in what data are recorded by city, county, or state governments, let alone at the national scale. We demonstrate how existing open-source datasets can be spatially integrated and subsequently used as training for machine learning (ML) models to predict building occupancy type, a major component needed for disaster preparedness and decision -making. Multiple ML algorithms are compared. We address strategies to handle significant class imbalance and introduce Bayesian neural networks to handle prediction uncertainty. The 100-year flood in North Carolina is provided as a practical application in disaster preparedness.

Impact Statement

In this study, we show how machine learning (ML) can be used to predict a building's occupancy type. We identify several features from open datasets capable of predicting building occupancy type while also evaluating multiple ML approaches. Residential buildings significantly outnumber commercial buildings, with commercial buildings significantly outnumbering other types of buildings. We evaluate multiple strategies for handling this class imbalance and show how Bayesian neural networks can assist in uncertainty quantification. Our results suggest strategies for utilizing ML-based approaches across the continental United States.

1. Introduction

As cities become more complex and integrative we have an increasing need to understand where urban features are, and what their primary purpose is. Recognizing the interconnectedness of systems such as buildings, transportation, power, and water is critical to the success, safety, and sustainability of urban regions. Yet, there is a major discrepancy in the way building data is collected across the United States. There is no harmonization in what data are recorded by city, county, or state governments let alone at a

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

national scale. Most previous efforts to harmonize the urban environment are proprietary, owned by groups such as Google, Yelp, and Zillow, and utilized solely in their respective apps. For example, in the area of urban flooding, which causes billions of dollars in damages annually, a reliable building inventory is critical for assessing the number of people affected, the propagation of shocks throughout the economy, and forecasting detailed socioeconomic risk from flooding. Publicly available open-source building inventory data sets exist; yet, they are not spatially aligned nor do many of them contain a critical element necessary for disaster assessment—building occupancy type. Knowing a building's occupancy type (residential, commercial, educational, etc.) plays a critical role in analyzing risk, assessing damage, performing mitigation efforts, and coordinating search and rescue.

In this work, we identify available open datasets and compare multiple (ML) algorithms in an attempt to accurately predict building occupancy type. We use disaster assessment to demonstrate how multiple use cases exist for the application of building occupancy type. Yet, disaster assessment is not the only use of a publicly available building inventory. There are many applications across disaster response, economic modeling and forecasting, and urban planning that benefit from a public dataset containing building-type information. The user base spans individual citizens to local and state governments to international collaboration. As a result, there are a number of groups currently attempting to produce public datasets on par with those datasets held by private companies. The US Geological Survey,[1] for example, is collaborating on an international effort to create a building inventory for earthquake loss assessment. Google AI[2] is working to produce a public building inventory with the multi-pronged goals of assessing population estimates when census information is not up to date, coordinating the humanitarian response in case of emergencies, measuring the consequences of natural disasters by estimating the number of buildings or households affected, understanding the human impact on the natural environment, assist vaccination planning by knowing the density of the population and settlements, and calculating statistical indicators in conjunction with housing locations to predict the mean travel time to nearby health care. On the games and public information side, the video game company Ubisoft created an online environment based on public building inventories upon which the online action game 'Tom Clancy's The Division' (for Windows, Playstation, and Xbox) is based. The public building inventories in the game enable realistic societal infrastructure and transportation simulations. Additionally, numerous YouTube videos and educational materials have been developed around public building inventories (Foody et al., 2017).

These multiple use cases require multiple ML approaches, which we evaluate and compare. Moreover, an imbalance in building occupancy types—states have significantly more residential buildings than other types of buildings—means that additional ML strategies must be employed. We present strategies for handling class imbalance while also introducing methods for assessing uncertainty in building occupancy-type predictions. The utility of our ML models is demonstrated in flood assessment applications in the state of North Carolina. Alabama is used to explore the transferability of the prediction model to other states. Ultimately, we aim to assess the accuracy of building occupancy type classification techniques using only open data that is available on the scale of the continental United States (CONUS). Related research has shown the predictive power of various data in determining building occupancy type. Yet, many highly predictive data sources (e.g., LIDAR (Lu et al., 2014)) are not freely and publicly available at the continental scale. This dramatically limits their utility in any open CONUS-scale applications.

2. Related work and the need for ML

OpenStreetMap[3] (OSM, OpenStreetMap, 2022) is a crowd-sourced initiative that aims to provide free and open access to spatial data on a global scale. OSM's representations of streets, natural landmarks, and

[1] https://pubs.usgs.gov/of/2008/1160/downloads/OF08-1160.pdf
[2] https://data.europa.eu/en/news-events/news/discover-open-buildings-dataset
[3] OpenStreetMap, https://www.openstreetmap.org/

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

building outlines are often more comprehensive and accurate than traditional data sources such as the CIA World Factbook and United States Census Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line data (Fan et al., 2014; Boeing, 2017; Jacobs and Mitchell, 2020; Moradi et al., 2022). Yet, even in areas with abundant data, the free-form and optional nature of OSM's building occupancy type attribute results in most mapped buildings having no reliable occupancy type. The lack of reliable building occupancy type has led several researchers to explore using machine learning to predict this value. For instance, (Atwal et al., 2022) achieved 98% accuracy in the binary classification task of predicting residential versus non-residential occupancy type. To achieve this, the authors used other features within the OSM dataset to train a decision tree. While the accuracy result is impressive, one of the required inputs to the decision tree is the so-called OSM building tag. There are two major challenges with this requirement. First, OSM building tags are rarely available and when they are they can be ambiguous. As noted by (Hoffmann et al., 2019) "It is important to note that, apart from the vocabulary difference and spelling error in the building tag, OSM also faces ambiguities in their finer classification scheme that is defined in the OSM Wiki." Such a decision tree could not be applied to tens of millions of buildings when used at CONUS scale. Second, having access to a highly informative tag such as "house" or "apartment" would seem to negate the need for ML.

Hoffmann et al. (Hoffmann et al., 2019) used a deep learning approach that predicted building type from a combination of aerial and street view images. They designed a multiclass prediction system in which the available categories were: commercial, residential, public, and industrial. (Hoffmann et al., 2019) concludes that with a multiclass setup and the challenges of OSM, a classification accuracy of about 60–80% on average would be a realistic expectation.

Our goals in this research are fourfold. First, to demonstrate the need for both binary and multiclass classification. Second, to see if the classification accuracy laid out by (Hoffmann et al., 2019) can be achieved using publicly available data and without the need for proprietary APIs and commercial image collections. Third, to determine which building-related features openly available at CONUS scale have the most predictive power. And, fourth, to provide a method of capturing uncertainty in predictions. If the challenges of open datasets like OSM coupled with large class imbalances limit prediction accuracy, then we need methods for determining when ML algorithms are making reliable predictions.

It is worth noting that our building occupancy type classification is fundamentally different from the similarly sounding building occupancy classification. Numerous studies in the literature (e.g., (Kanthila et al., 2022; Sayed et al., 2022)) are focused on the occupants' presence, power usage habits, and interior conditions with the goal of enhancing building operation and management. Their focus is on curtailing energy consumption and reducing costs. Our work focuses on determining building type (i.e. residential, commercial, industrial, educational, etc.). We are interested in the type of structure and not how its occupants are utilizing power, heating, and cooling. This difference in application results in a need for different types of data, which results in different machine-learning needs and challenges.

3. ML dataset

No CONUS-scale ML-ready dataset for building occupancy type currently exists. It is not even apparent which building-level features available openly at the CONUS scale have the most predictive power. We do not claim that the features utilized in this study are the most predictive. Rather, we gather a sampling of different types of open data available at the building level and perform exploratory analysis. Our goal is to predict a building's type (e.g., residential/commercial/industrial) based on known features of the building (area, location, etc.). To accomplish this, we gather a number of building features from open datasets. Our sources of building features are the Open Street Map (OSM), the Multi-Resolution Land Characteristics Consortium (MRLC),[4] the United States Census' County Business Patterns (CBP),[5] and the United States

[4] https://www.mrlc.gov/data/type/urban-imperviousness
[5] https://www.census.gov/programs-surveys/cbp/data/tables.html

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

Table 1. *Features utilized in our machine learning training data*

Column	Description	Data Source
X	X coordinate of the building in the EPGS:5070 system	OSM
Y	Y coordinate of the building in the EPGS:5070 system	OSM
Area	Area of building in square meters	OSM
MedianIncomeCounty	Median income of the county in which the building resides	ACS
HousingUnitsCounty	Number of housing units in the county in which the building resides	ACS
HousingDensityCounty	Number of housing units in the county divided by the number of people residing in the county where the building resides	ACS
Impervious	Percentage of the area surrounding the building that is comprised of impervious surfaces such as roads and other paved surfaces. The value provided is the mean area-weighted average of imperviousness underneath the building footprint	MRLC
AgCount	Number of agricultural businesses in the county in which the building resides	CBP
CmCount	Number of commercial businesses in the county in which the building resides	CBP
GvCount	Number of government buildings in the county in which the building resides	CBP
EdCount	Number of educational buildings in the county in which the building resides	CBP
InCount	Number of industrial buildings in the county in which the building resides	CBP
OsmNearestRoad	Type of nearest road to the building	OSM
OccupancyType	Building occupancy classification	USA Structures

Census' American Community Survey (ACS)[6]. We have spatially aligned all features by using OSM longitude and latitude to determine which county the building resides in. We have then gone to the ACS and looked up socio-economic county data such as median income and housing density. Imperviousness values are taken from the MRLC. The building area polygons from OSM are overlayed with 30-meter resolution imperviousness data and a weighted average of the imperviousness underneath the building footprint was computed. The resulting ML-ready data have the features listed in Table 1.

Assessing the accuracy of building occupancy-type predictions involves a source of "ground-truth" data. We utilize the US structures[7] dataset maintained by FEMA and created in conjunction with DHS Science and Technology and Oak Ridge National Laboratory. US structures was created by extracting building outlines via commercially available satellite imagery. Building occupancy type (e.g., residential, commercial, industrial) was then determined from a combination of local governments who agreed to share it, open data from the National Geospatial-Intelligence Agency (NGA), Census housing data, and parcel data. These US Structures building types are what we are trying to predict. The US Structures 'OccupancyType' is the final column in our training data and our source of "ground-truth" values. We note that the US Structures dataset is smaller than OSM and currently only available in Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Texas, and Virginia. Thus, while a good source of ground-truth data, US Structures does not preclude the need for machine learning. Utilizing

[6] https://www.census.gov/programs-surveys/acs/data.html
[7] https://gis-fema.hub.arcgis.com/pages/usa-structures

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

Table 2. *Summary statistics of building occupancy types in North Carolina*

Building occupancy type	Number of buildings	Percentage of dataset (%)
Residential	976,690	90.6
Commercial	64,029	5.9
Industrial	16,722	1.6
Assembly	7323	0.68
Education	6457	0.60
Government	4910	0.46
Agriculture	1651	0.15
Utility and misc.	362	0.03

North Carolina as a test site, we have created an ML-ready dataset of 1,078,144 buildings with known building occupancy types. We use Alabama US Structures as a test of transferability. An ML model trained in North Carolina is applied to Alabama to asses if CONUS-scale prediction can be done with a single ML model or requires state-specific models. US structures recognizes eight building occupancy types and the occurrence of these types in North Carolina is shown in Table 2.

Evident from Table 2 is that we have a significant class imbalance. The residential buildings, for example, outnumber the commercial buildings by 15 to 1. The residential buildings outnumber the agricultural buildings by nearly 600 to 1. This can be a challenge for machine learning classification algorithms. The algorithm may not "learn," but rather obtain high accuracy simply by picking the majority class. We look at techniques for dealing with such imbalance.

4. Machine learning algorithms and evaluation metrics

Applications using building-type information demand multiple types of classification. For example, during emergency response, first responders may only be interested in identifying residential buildings. In this use case, binary classification (residential vs. non-residential) is sufficient. However, for economic forecasting applications, a finer-grained distinction is needed (i.e., residential vs. commercial vs. industrial). In such a scenario, multiclass classification is needed to determine a comprehensive economic assessment. We provide examples of both scenarios. Specifically, we design a set of supervised learning models in which building features (location, square footage, proximity to other resources, etc.) are used to predict building occupancy type. We assess the accuracy, precision, and recall of two machine learning algorithms while also investigating techniques for dealing with unbalanced classes.

For the binary classification scenario, all non-residential building types are changed to "Other." In the multiclass scenario, we attempted to accurately classify all eight building occupancy types recognized by US structures. As detailed in subsequent sections, the accuracy was too low to be practically useful. We settled on a three-class model of Residential, Commercial, or "Other."

In both binary and multiclass scenarios, the OsmNearestRoad and OccupancyType, which are initially text, are encoded to numbers and we scale the data (for neural networks) based on standard scores. Our ML dataset is split into training and testing portions using the standard 80/20 training/testing split. In other words, 80% of the buildings in North Carolina are used to train the various ML algorithms and the remaining 20% is held aside to evaluate the trained algorithms. The training data are applied to multiple variations of random forest and neural network approaches. We include the standard random forest approach as a baseline and then compare variations designed to better handle class imbalance.

Trained ML models are then evaluated on the test set with the following metrics, which are based on the number of true positive (TP), false negative (FN), and false positive (FP) predictions.

Balanced accuracy—a variation of accuracy used in classification problems with imbalanced datasets defined as the average of recall obtained on each class. Recall of each class is found from

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

We also report precision, which is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The F1 score is the harmonic mean of precision and recall given by

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Precision, recall, and F1 scores exist for each class. In the binary classification scenario, for example, there are precision, recall, and F1 values for Residential predictions as well as precision, recall, and F1 values for non-residential Predictions. We provide average precision, recall, and F1 values when summarizing our results in tables. The reported averages are the so-called macro average, which is computed by taking the arithmetic mean (unweighted) across the classes. This approach treats each class with equal significance.

4.1. Binary classification

Table 3 shows the distribution of building occupancy types after non-residential buildings are converted to "Other." There still exists a significant imbalance in the building occupancy types. We trained a random forest model using 5-fold cross-validation to ensure accuracy was not dependent on how the data was split.

Undersampling refers to a group of techniques designed to balance the class distribution. Undersampling techniques remove examples from the training dataset that belong to the majority class in order to better balance the class distribution. This is in contrast to oversampling, which involves adding examples to the minority class in an effort to reduce the class distribution imbalance. Near Miss (Zhang and Mani, 2003) refers to a collection of undersampling methods that select examples based on the distance of majority class examples to minority class examples. Distance is determined in feature space using Euclidean distance. NearMiss allows us to keep majority class examples that are on the decision boundary leaving 101,453 residential buildings and the same number of non-residential buildings.

Perhaps the simplest approach to handle a severe class imbalance is to change the decision threshold. A random forest model is comprised of multiple decision trees. We can obtain from the model the proportion of decision trees that voted "residential" and the proportion of decision trees that voted "non-residential" for a given building. An obvious default approach is to set the threshold at 0.5. If the proportion of decision trees voting "residential" is above 0.5 then the algorithm predicts "residential." Otherwise, the algorithm predicts "non-residential." Threshold moving simply moves this threshold attempting to achieve higher accuracy. For example, we may require a proportion of 0.6 or greater to predict "residential." Threshold moving can be effective for class imbalance because it helps avoid simply predicting the majority class. In our case, more evidence is needed before predicting "residential."

The "best" threshold to use is determined by finding the threshold with the maximum F1 score, a metric which is the harmonic mean of precision and recall. Figure 1 illustrates the results of threshold moving applied to our binary classification random forest model. The points connected by a solid line

Table 3. *Building occupancy type distribution for binary classification*

Building occupancy type	Number of buildings	Ratio to residential
Residential	976,689	–
Other	101,453	1:10

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

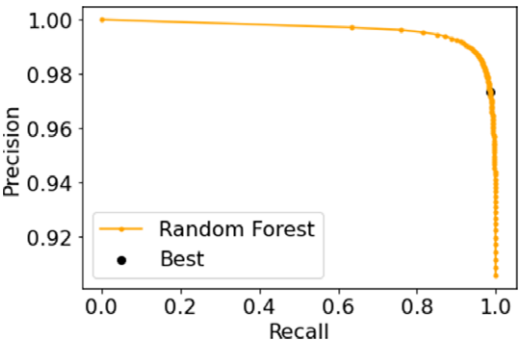

Figure 1. *Precision recall curve for random forest threshold moving.*

show the change in precision and recall as the threshold is modified. The large solid black dot highlights the location where precision and recall combine to maximize F1. Our experiments revealed an optimal threshold of 0.55.

We also trained a binary classification neural network to compare to the random forest approaches. Because of the class imbalance, we want the neural network to pay more attention to the fewer examples of non-residential buildings. A common technique for achieving this is to weight the classes using:

$$\text{weight} = (1/n_c) \times (N/2) \quad (4.1)$$

where n_c is the number of buildings in the class and N is the total number of buildings.

We arrive at a residential weighting of 0.55 and a non-residential weighting of 5.31. The neural network itself has two hidden layers of size 30 and 15, respectively. The network was set to train for 150 epochs. However, early stopping was applied with the training set to stop if accuracy did not improve for seven consecutive epochs. After 36 epochs, training stopped due to no improvement in accuracy.

4.2. Multiclass classification

US structures recognizes eight building occupancy types. We first repeated the machine learning process using all eight classes in a multiclass classification scenario. We found class imbalance to be so severe we could not reliably predict more than the three most frequently occurring building types. The classification accuracy for all eight occupancy types was around 30%. When Assembly, Education, Government, Agriculture, and Utility and Misc. were combined into a single "Other" class, the accuracy only improved to about 50%. We limit our discussion here to machine learning models designed to predict "residential," "commercial," or "other."

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from a single algorithm. One-vs-rest and one-vs-one have emerged as two popular ensemble techniques for classification. While not specifically designed for addressing class imbalance, decomposing the classification problem into a set of binary classification problems can sometimes help. The one-vs-one and one-vs-rest techniques in particular have led to promising results in empirical studies involving imbalanced classes (Fernandez et al., 2013; Krawczyk, 2016; Zhang et al., 2016).

The one-vs-one strategy splits a multi-class classification into one binary classification problem per each pair of classes, for example, residential vs. commercial, residential vs. other, commercial versus other. The final class assignment is determined by aggregating the results of the binary classifiers. The one-vs-rest strategy is a related approach, however, fewer models are created. There is one binary classification problem per class, for example, residential vs. rest, commercial vs. rest, and other vs. rest. We applied both approaches to our 3-class classification.

A multiclass neural network was implemented with similar characteristics to those utilized in the neural network binary classification. The resulting class weights were 0.55 for residential buildings, 8.42 for commercial buildings, and 14.40 for "other" buildings. In addition to class weights, focal loss (Lin et al.,

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

2017) is an approach for dealing with imbalanced classes in neural networks. Focal loss addresses class imbalance during training by applying a modulating term to the neural network loss function. The intention is to focus learning on difficult misclassified examples. This modulating term is dynamically scaled meaning it decays to zero as confidence in the correct class increases. Intuitively, this modulating term decreases the contribution of easily classified examples during training while increasing the contribution of more difficult classification examples. We implemented focal loss using the TensorFlow software library. Both binary and multiclass focal loss were tried. The default values of the software for setting the modulating term were used in both applications.

5. Results

5.1. Binary classification

Results from the binary classification experiments are shown in Table 4. We found fairly consistent metrics across all four algorithms with the class-balanced neural network having the highest accuracy at 89%. Class balancing strategies for random forests produced mixed results. Threshold moving led to a slight improvement in accuracy over the default random forest while near-miss performed worse. The confusion matrix for neural network classification is shown in Figure 2 where we can see precision and recall in more detail. The neural network does well with residential buildings, correctly classifying 92% of the test set. Non-residential buildings in the test set are correctly classified 85% of the time with 15% of non-residential buildings being misclassified as residential.

Feature importance is a technique for assessing how important each of our input features is to making accurate predictions. Features with low importance do not contribute much (do not have much weight) to prediction accuracy. Low-importance features can be ignored creating simpler more scalable machine learning models. We use a technique called Permutation Importance in which feature values are randomly shuffled. The resulting mean accuracy decrease plot expresses how much accuracy the model loses through feature permutation. The more the accuracy suffers, the more important the variable is for successful classification. Feature importance was carried out on both the neural network and random forest models. We found the same feature ranking across all models so, for brevity, we show only the mean accuracy decrease plot from the neural network. This is displayed in Figure 3.

We find that imperviousness has the most predictive power followed by a building's area, the type of road it is nearest to, and the building's longitude and latitude. We had hypothesized that Census data on the number of agricultural, commercial, industrial, and governmental businesses in the area would be helpful. However, these data contribute little predictive power. One feature from the US Census, mean income in the county, does help somewhat. Residential buildings vary in size across the state and we suspect the correlation between mean income and home size in a county is being utilized.

5.2. Multiclass classification

Results from the multiclass experiments are shown in Table 5. When a random forest model is applied to this three-class classification problem, we achieve a balanced accuracy of 67%. The confusion matrix for

Table 4. *Results from binary classification experiments*

Method	Balanced accuracy (%)	Macro F1 (%)	Macro precision (%)	Macro recall (%)
NN	89%	80	76	89
RF threshold	87	88	90	87
Random forest	86	88	91	85
RF near-miss	84	84	84	84
Focal loss	81	84	89	81

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

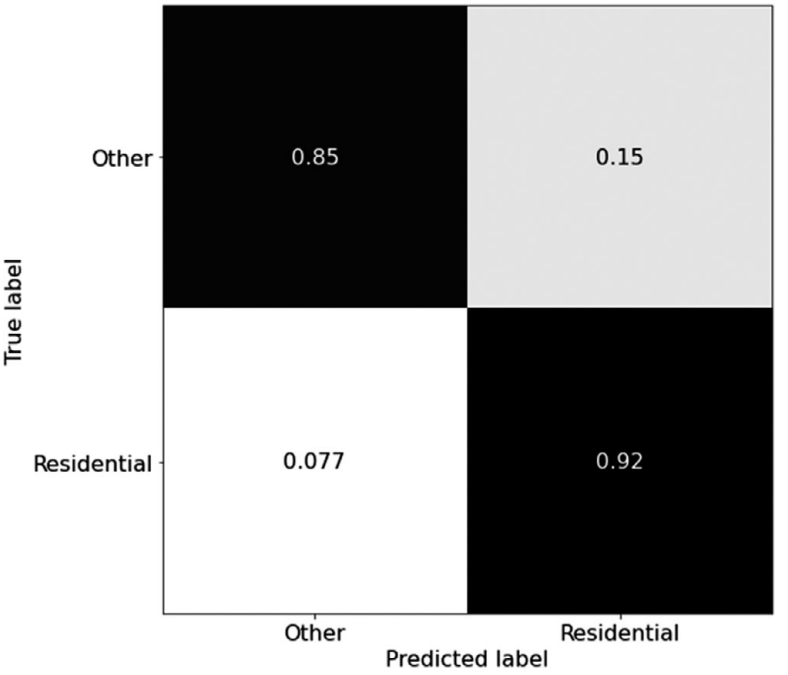

Figure 2. *Confusion matrix for neural network binary classification. Values are listed as proportions.*

this approach is shown in Figure 4. Feature importance analysis was repeated using the multiclass random forest model. Results did not differ from those shown in Figure 3.

The severe class imbalance limits the practical application of multiclass classification with three classes. Looking deeper into our ML models trained with more classes, we find predictive "confusion" is correlated with the class imbalance. In other words, the less frequently an occupancy type appears in our dataset, the more likely it is to be confused with other occupancy types. Industrial buildings, which are outnumbered by residential buildings nearly 60–1, are confused for commercial buildings 33% of the time and misclassified as residential buildings 23% of the time. Additional improvements in accuracy, precision, and recall have been found through the addition of aerial and street view data (Hoffmann et al., 2019). Yet, adding these additional datasets adds a reliance on proprietary APIs and commercial image collections. In the next section, we explore Bayesian neural networks as a means of addressing prediction uncertainty.

6. Bayesian neural networks for prediction uncertainty

The traditional approach to machine learning is to optimize a loss function to obtain an optimal setting of the model parameters. Numerous iterations over the training data allow the model parameters to be fine-tuned. The machine learning training process results in a set of optimal model parameters that are then used in subsequent predictions.

A Bayesian approach, by contrast, does not seek fixed model parameters. In a Bayesian neural network (Wilson and Izmailov, 2020) the network weights are distributions. Passing the same input through the trained network will result in different predictions as different weights are chosen from the distributions on each pass. Getting a distribution of predictions for each input is useful as the ensemble of predictions can be used to quantify uncertainty in neural network weights—commonly referred to as epistemic uncertainty—and the probability of being in each class leads to a determination of prediction uncertainty—commonly referred to as aleatoric uncertainty.

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

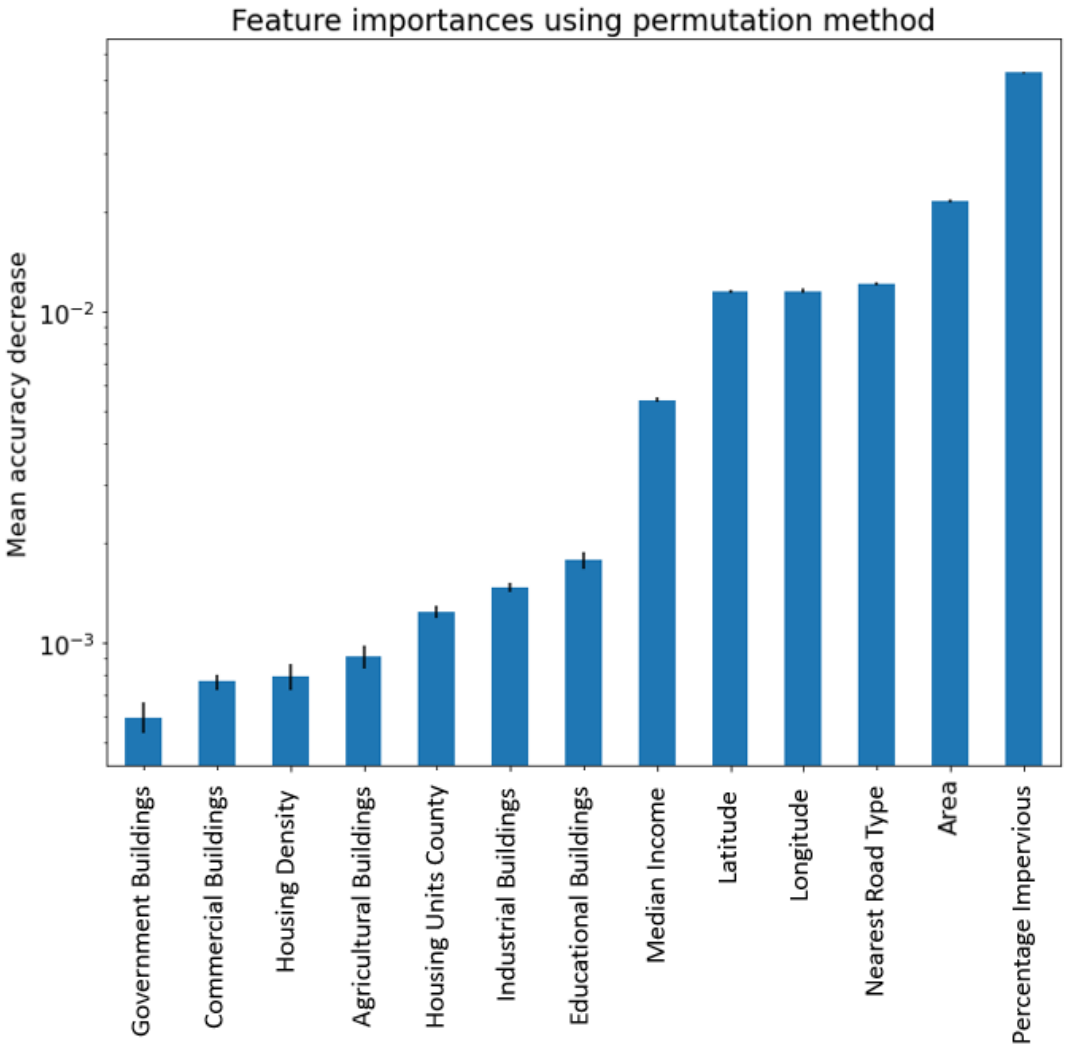

Figure 3. *Permutation feature importance. Features are shown in ascending order of importance. The bars show the mean accuracy decrease after 10 repetitions with error bars displaying the standard deviation of the 10 repetitions. Building types and income values are county aggregates obtained from the US Census. Percent impervious is the percentage of area surrounding the unknown building comprised of impervious materials.*

Table 5. *Results from multiclass classification experiments*

Method	Balanced accuracy (%)	Macro F1 (%)	Macro precision (%)	Macro recall (%)
Random forest	67	72	78	67
One-vs-one	67	71	78	67
One-vs-rest	67	73	79	66
Focal loss	58	62	71	58
NN	37	87	85	89

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

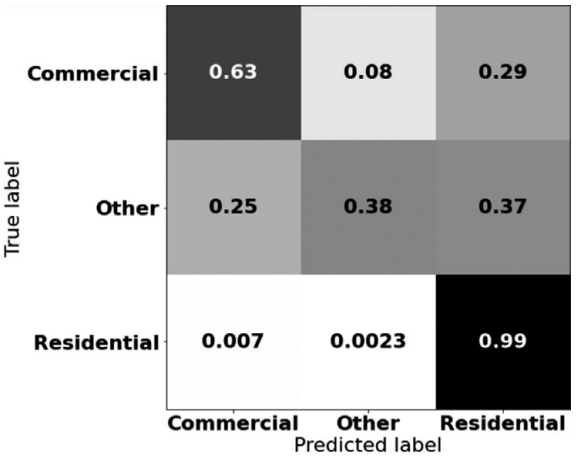

Figure 4. *Confusion matrix for multiclass random forest.*

Our Bayesian neural network is implemented using the TensorFlow Probability software library (Dillon et al., 2017). The other layers in our neural networks are standard Dense layers from the Keras library (Chollet and others, 2015). Aleatoric uncertainty is quantified using Shannon entropy, which is common for classification problems (Hüllermeier and Waegeman, 2021). Shannon entropy is defined as

$$H = -\sum_{i=1}^{n} p_i \times ln(p_i)$$

where n is the number of possible output categories (e.g., residential, commercial, and other) and p_i is the predicted likelihood of being in each of those categories. Entropy in information theory is analogous to entropy in statistical thermodynamics. Larger values of H correspond to the probabilities being spread more evenly across the possible classes. Thus, larger values of H correspond to the neural network being more uncertain in its prediction.

We create a Bayesian neural network for the binary classification problem and a second Bayesian neural network for the 3-class (residential, commercial, and other) problem. After training the networks, each building in the test set was passed through the network 100 times. The mean of the 100 predictions was computed and Shannon entropy was used to determine prediction uncertainty. This approach enables a user-defined uncertainty threshold. Predictions with uncertainty above the threshold are ignored due to the neural network having little confidence in the prediction. In this manner, the user is trading off number of predictions for higher prediction accuracy. Low values of the uncertain threshold—implying the neural network must have high confidence in the prediction—leads to fewer predictions being made; however, those predictions are of higher accuracy.

Figure 5 compares confusion matrices for the multiclass Bayesian neural network. The left panel shows the results when there is no uncertainty threshold, that is the neural network is forced to make a prediction for every building. The right panel shows the results for an example uncertainty threshold of 0.4 We are making about 50% fewer predictions in the right panel; yet, the neural network is much more confident—and correct—in those predictions.

7. Application and transferability

As a simple demonstration of how building occupancy type prediction can be used, we identified 219,054 buildings in the OSM North Carolina dataset not used in our train/test data that had unknown building occupancy types. Further, we obtained the 100-year flood map from the FEMA National Flood Hazards Layer. The phrase "100-year flood" is used to describe the extent of a flood that statistically has a 1-percent

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

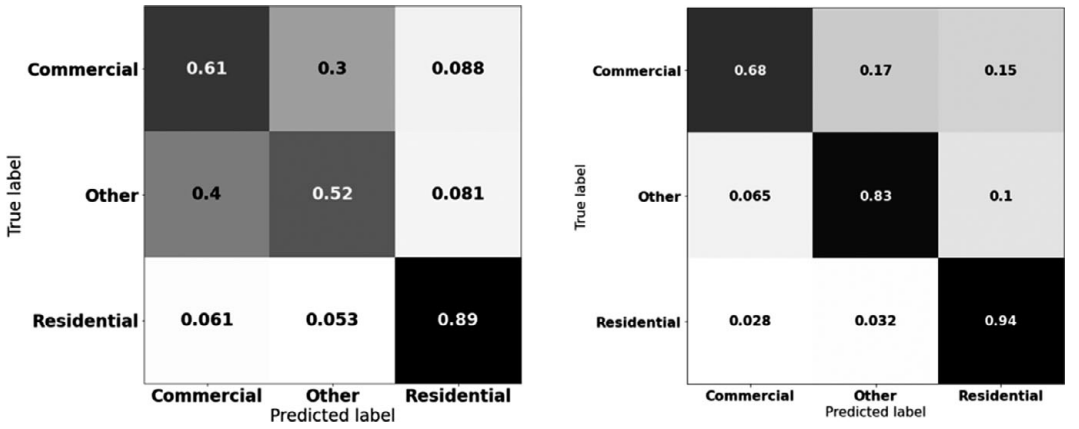

Figure 5. *Confusion matrices from the multiclass Bayesian Neural Network. On the left, results when all predictions are kept. On the right, results when prediction uncertainty is below 0.4.*

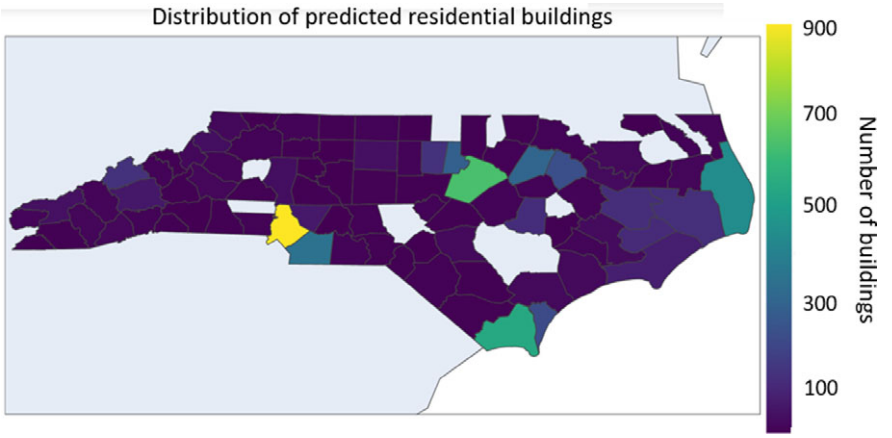

Figure 6. *Predicted residential flooding for buildings with previously unknown occupancy type.*

chance of occurring in any given year (Maidment, 2009). Here, it is used for illustrative purposes of a disaster preparedness application. Flood maps indicated that 5838 buildings out of the 219,054 unknown North Carolina buildings would be impacted by a 100-year flood. Our model predicted these to be 5452 residential, 349 commercial, and 37 "other." The spatial distribution of the predicted residential and commercial buildings is shown in Figures 6 and 7, respectively. This simple example highlights two important points. First, despite open datasets such as OSM and US structures, there still exist hundreds of thousands of buildings in North Carolina with unknown occupancy types. Public safety assessments and economic forecasting of a natural disaster, such as a 100-year flood, are limited in their ability to fully account for the entire state. A machine learning approach, such as the one described here, can quickly and reliably predict those missing occupancy types.

We obtained US structures data from Alabama to evaluate the transferability of our models. As mentioned previously, US structures has ground truth building occupancy types for nine states. Our goal was to assess if an ML model trained in one state could be used in another state. Unfortunately, this was not the case. The Alabama dataset contained 2.2 million buildings with known occupancy types. The multiclass model trained on North Carolina data was only able to accurately classify 48% of Alabama buildings by default. Accuracy increased using the Bayesian approach and uncertainty thresholds.

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

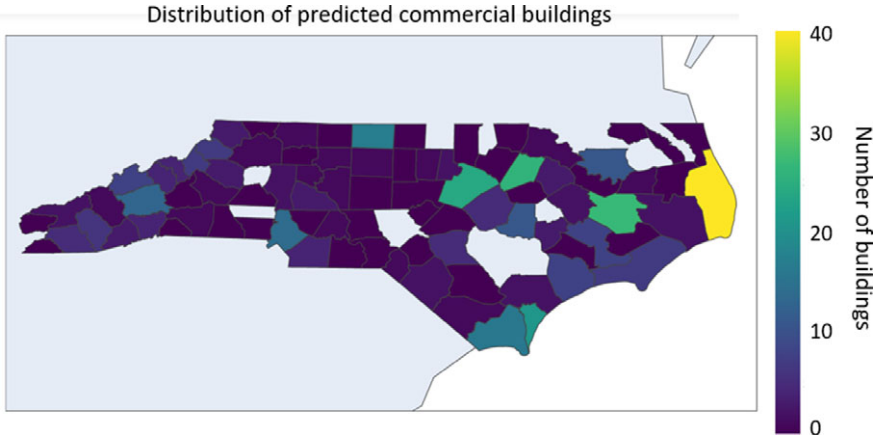

Figure 7. *Predicted commercial flooding for buildings with previously unknown occupancy type.*

However, the improved accuracy is not significant enough to think that one neural network model could be applicable across the continental United States as the distributions of the selected features vary too much between states. Put another way, while the selected features are useful for predicting building occupancy type, the weighting of each feature varies from state to state. Practitioners may need to train multiple state-specific neural networks. Alternatively, it may be feasible to train regional neural network models; yet, it is unclear to us at present how to optimally construct such training datasets. More research is needed into multi-state neural networks and the specific impacts of geographic features (that is imperviousness) and socio-economic features (that is median income and median area). We leave this for subsequent research.

8. Conclusion

We have identified features, freely and publicly available at the continental scale, that are useful in building occupancy type classification. These features can become part of a CONUS-scale building classification workflow with accuracy values comparable to existing ML systems (e.g., (Hoffmann et al., 2019)) and without the reliance on proprietary APIs and commercial image collections. Specifically, with the current set of features, we achieve 87% accuracy with binary classification and 74% accuracy on the three-class (residential/commercial/other) classification problems using Bayesian neural networks and uncertainty thresholds. As a specific comparison, our technique correctly classifies 62% of commercial buildings when predictions are made on all buildings and 74% of commercial buildings when predictions are limited to most certain commercial buildings. This is compared to (Hoffmann et al., 2019)'s 64% for commercial buildings. Additionally, we have added a Bayesian component that flags uncertain predictions, which can be helpful in disaster response and economic forecasting applications.

We have also found that our ML models, trained on North Carolina buildings, do not transfer well to another state. This suggests that CONUS efforts may require regional models or even state-specific models. Our future research will compare the accuracy of regional models (ML models trained on data from multiple states) to the accuracy of individual state models. We will also explore more building features from open datasets that can improve prediction accuracy.

Acknowledgements. We are grateful for the support provided by the GeoScience MAchine Learning Resources and Training (GeoSMART) program (NSF grant 2117834), which allowed us to create a machine-learning tutorial from this research. Tutorial creation was further enabled by resources and materials provided by the ESIP Lab with support from the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the United States Geologic Survey (USGS).

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

Author contributions. Conceptualization: T.N.;J.M.J.;J.S-M Methodology: T.N.;J.M.J.;J.S-M;A.M.R. Data curation: T.N.;J.M.J. Data visualisation: T.N. Writing original draft: T.N.;J.M.J.;A.M.R. Revisions: T.N.;J.M.J;J.S-M; All authors approved the final submitted draft.

Competing interest. None declared.

Data availability statement. We thank the GeoScience MAchine Learning Resources and Training (GeoSMART) program for its interest in turning our research into a community tutorial. Preprocessing routines, a Jupyter Notebook demonstrating our machine learning training and testing, and our ML-ready data (Narock and Johnson, 2023b) are available via the tutorial at https://doi.org/10.5281/zenodo.7871800. The integrated spatially aligned data used to create our ML-ready data (Narock and Johnson, 2023a) are freely available via https://doi.org/10.5281/zenodo.7696864. The Bayesian component and focal loss were not part of the original tutorial and were added during the revision phase of this paper. We have since added an 'EDS_paper' folder to (Narock and Johnson, 2023b) and host those notebooks at https://github.com/geo-smart/flood-risk-ml-tutorial.

Ethical statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Atwal KS, **Anderson T**, **Pfoser D** et al. (2022) Predicting building types using OpenStreetMap. *Scientific Reports 12*, 19976 https://doi.org/10.1038/s41598-022-24263-w

Boeing GO (2017) New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems 65*, 126–139.

Chollet F and **others** (2015) *Keras*. GitHub. Available at https://github.com/fchollet/keras

Dillon JV, **Langmore I**, **Tran D**, **Brevdo E**, **Vasudevan S**, **Moore D**, **Patton B**, **Alemi A**, **Hoffman M and Saurous RA** (2017) TensorFlow distributions. *arXiv preprint* arXiv:1711.10604

Fan H, **Zipf A**, **Fu Q and Neis P** (2014) Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science 28*, 700–719.

Fernandez A, **Lopez V**, **Galar M**, **del Jesus MJ and Herrera F** (2013) Analyzing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. *Knowledge-based Systems 42*, 97–110.

Foody G, **See L**, **Fritz S**, **Mooney P**, **Olteanu-Raimond A-M**, **Fonte CC and Antoniou V** (eds.) (2017) *Mapping and the Citizen Sensor*. Ubiquity Press. Available at http://www.jstor.org/stable/j.ctv3t5qzc

Hoffmann EJ, **Wang Y**, **Werner M**, **Kang J and Zhu XX** (2019) Model fusion for building type classification from aerial and street view images. *Remote Sensing 11*, 1259. https://doi.org/10.3390/rs11111259

Hüllermeier E and Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning 110*, 457–506 https://doi.org/10.1007/s10994-021-05946-3

Jacobs KT and Mitchell SW (2020) OpenStreetMap quality assessment using unsupervised machine learning methods. *Transactions in GIS 24*, 1280–1298

Kanthila C, **Boodi A**, **Beddiar K**, **Amirat Y and Benbouzid M** (2022) Building occupancy detection using machine learning-based approaches: evaluation and comparison. In *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society, Brussels, Belgium*, pp. 1–6 https://doi.org/10.1109/IECON49645.2022.9968768.

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*, 221–232

Lin T-Y, **Goyal P**, **Girshick R**, **He K and Dollár P** (2017) Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988.

Lu Z, **Im J**, **Rhee J and Hodgson M** (2014) Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning 130*, 134–148.

Maidment D (2009) FEMA flood map accuracy. In *World Environmental and Water Resources Congress 2009: Great Rivers*. https://doi.org/10.1061/41036(342)492 Published online: April 26, 2012

Moradi M, **Roche S and Mostafavi MA** (2022) Exploring five indicators for the quality of OpenStreetMap road networks: A case study of Québec, Canada. *Geomatica*, Volume *75*, Issue 4, December 2021, Pages 1–31.

Narock T and Johnson JM (2023a) *North Carolina Building Features (1.0.0) [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.7696864

Narock T and Johnson JM (2023b) *geo-smart/flood-risk-ml-tutorial: Initial Release v1.0.0 (v1.0.0).* Zenodo. https://doi.org/10.5281/zenodo.7871800

OpenStreetMap (2022) OpenStreetMap data. Available at https://www.openstreetmap.org/

Sayed AN, **Himeur Y and Bensaali F** (2022) Deep and transfer learning for building occupancy detection: a review and comparative analysis. *Engineering Applications of Artificial Intelligence 115*, 105254, 0952–1976. https://doi.org/10.1016/j.engappai.2022.105254

Wilson AG and Izmailov P (2020) Bayesian deep learning and a probabilistic perspective of generalization. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada. Available at https://proceedings.neurips.cc/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf

https://doi.org/10.1017/eds.2025.2 Published online by Cambridge University Press

Zhang J and Mani I (2003) KNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of the Workshop on Learning From Imbalanced Data II International Conference on Machine Learning*, Washington, DC. Available at https://www.site.uottawa.ca/nat/Workshop2003/jzhang.pdf

Zhang Z, **Krawczyk B**, **Garcia S**, **Rosales-Perez A and Herrera F** (2016) Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data *Knowledge-Based Systems 106*, 251–263.

Cite this article: Narock T, Johnson J.M, Singh-Mohudpur J and Modaresi Rad A (2025). Building occupancy type classification and uncertainty estimation using machine learning and open data. *Environmental Data Science*, 4: e10. doi:10.1017/eds.2025.2