Building classification using random forest to develop a geodatabase for Probabilistic Hazard Information (PHI)

Jooho Kim[1*], Joshua J. Hatzis[2], Kim Klockow[3], and Patrick A. Campbell[4]

[1]Postdoc research associate, Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO) and NOAA National Severe Storms Laboratory, jooho.kim@noaa.gov, National Weather Center, 120 David L Boren Blvd, Norman, OK 73072, USA.

[2]Postdoc research associate, Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO), jjhatzis@ou.edu, National Weather Center, 120 David L Boren Blvd, Norman, OK 73072, USA.

[3]Research Scientist, Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO) and NOAA National Severe Storms Laboratory, kim.klockow@noaa.gov. National Weather Center, 120 David L Boren Blvd, Norman, OK 73072, USA.

[4]Research Scientist, Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO) and NOAA National Severe Storms Laboratory, patrick.campbell@noaa.gov,National Weather Center, 120 David L Boren Blvd, Norman, OK 73072, USA.

*Corresponding author: Jooho Kim / jooho.kim@noaa.gov

Abstract

To understand the community risk from severe weather threats, two components including weather information and community assets are crucial. Recently, Probabilistic Hazard Information (PHI) from the NOAA Forecasting a Continuum of Environmental Threats (FACETs) program has been developed to provide dynamic weather-related information between the watch and warning systems to weather forecasters, emergency management agencies and the public. To predict community physical risks on critical infrastructure and building properties using PHI, building type information is required. This study applied a machine-learning technique to predict building types using building footprint and city zoning data. We collected Oklahoma county building property data to train and test a random forest model. The result of this study showed that building footprint and city zoning data can be applied to classify multiple building types with an accuracy of 96%. The machine-learning based building classification contributed to the acquisition of building type data in the Oklahoma City metropolitan area. This geodatabase will be utilized to predict real-time critical infrastructure and building damage assessment using PHI. In addition to their importance to physical building damage assessment, the results can be utilized to develop post-disaster responses and planning.

Keywords: probabilistic hazard information (PHI), severe weather threats, building classification, random forest, building damage assessment

Introduction

Nearly 90% of all presidentially declared disasters are weather-related, such as tornadoes, hurricanes, storms, and floods. According to the National Oceanic and Atmospheric Administration (NOAA), a tornado is defined as "*a violently rotating column of air that extends from a thunderstorm and comes into contact with the ground*". While hurricanes, storms and floods have longer lead-times to prepare for community impacts from severe weather, tornadoes have a very short lead-time (average 13 mins) with high uncertainty (NOAA 2011). Around 1,000 tornadoes are reported annually, and the tornado intensity is measured on a scale of 0 through 5. It incorporates 28 different damage indicators, based on damage to a wide variety of structures ranging from trees to institutional buildings (McDonald and Mehta 2006a). In 2020, there were 1,218 tornadoes compared with 1,520 in 2019. In 2020, about 78 people perished compared with 41 in 2019. For example, in April 2020, 32 people perished in tornadoes in Georgia, Mississippi, South Carolina and Tennessee (Brackett and Childs 2020). The storms caused at least $3 billion in insured losses in the areas of Midwest and Mid-Atlantic states (Insurance Information Institute 2019). In March 2020, 24 people were killed in tornadoes in central Tennessee, including the city of Nashville. Tornado deaths in 2020 are the highest since 2011 (Gee et al. 2020).

While U.S. tornado-related fatalities are less than 100 per year on average (NWS, 2020), the average annual tornado-caused economic losses are between $4.7B and $7.2B (Simmons et al. 2013); less than 1–3% of the annual GDPs of the most-affected states. These remarkably low fatality numbers are due to advanced meteorological and engineering research, and the geolocational characteristics of tornadoes that did not directly affect larger cities or residential communities. Economic losses from tornadoes are still high. There has been $130 billion (CPI-adjusted estimate) in losses due to 54 major events involving tornadoes spawned by severe storms over the past 40 years. This figure may be lower than, but of a comparable order of magnitude to hurricane-related losses of nearly $1 trillion dollars over the period 1980–2019 from 45 hurricanes. These very costly tornado and hurricane events caused just 1,270 and 6,507 fatalities, respectively (NOAA/NCEI 2018).

Community risk from severe weather threats can be measured by two components: (1) severe weather-related data/information and (2) databases for community assets (e.g., critical infrastructure (assets, systems, and networks) and building property data (structure type, exterior materials, the number of stories) (see Fig. 1).

Recently, Probabilistic Hazard Information (PHI) under the NOAA Forecasting a Continuum of Environmental Threats (FACETs) program has been developed to provide dynamic weather-related information with weather forecasters, emergency management agencies, and the public (James et al. 2020; Karstens et al. 2015b; Rothfusz et al. 2018). FACETs is a concept being explored by NOAA to potentially shift the National Weather Service (NWS) from deterministic watch/warning products to high-resolution PHI products spanning from days (and longer) to within minutes of high-impact weather and water events.

To improve community resiliency using PHI from severe weather threats, it is critical to develop a geodatabase that includes a variety of community assets in terms of critical infrastructure, residential, commercial and industrial buildings. The goal of this study is to develop a machine learning model to create a geodatabase that can be utilized for community risk assessment using PHI. To this end, this study consists of three parts: (1) building property data collection and preparation for the training and testing datasets, (2) the development of a machine learning model to classify building type data and (3) predicting building types using the trained machine learning model. The findings support the development of a geodatabase for a better community risk assessment using PHI. Community risk assessment will support weather forecasters, emergency management agencies and facility managers to understand the impact of severe weather on communities and to develop a better emergency response and recovery.

The remainder of this study is organized as follows. Section 2 provides a systematic review of literature including community risk assessment, PHI and building classifications. Section 3 addresses a machine learning model and data collection/preparation to train and test a random forest model. In section 4, we review the results including model performance, factor importance and hyper-parameter tuning. Lastly, Section 5 provides our conclusions with a discussion of the limitations of this research and future research topics.

Literature review

In this section, we review the literature related to community damage assessment, the comparison of deterministic and probabilistic weather-related information systems, and building classification. While building classification is the primary focus of this paper it is important to understand how community damage assessment has been done in the past and how it may be implemented with developing weather information technologies to better understand the reasons for our methodology.

Community damage assessment

Damage assessment after storms is traditionally completed through on-site damage surveys (Vetrivel et al. 2016) which involves a team of surveyors going from building to building determining the individual damages (Burgess et al. 2014; Kuligowski et al. 2014). This technique yields a lot of detail on the amount of damage and the reasons for structure failure, however it is a labor-intensive process that can take days to weeks to complete for large damage areas (e.g., hurricanes) and so can slow down the clean-up and rebuilding processes after a disaster (Vetrivel et al. 2016; Zhou and Gong 2018).

The use of aircraft or satellite-based remote sensors to assess damages has become a common alternative in recent decades with optical, synthetic aperture radar (SAR), and light detection and ranging (LiDAR) among the most common sensors, each having their own strengths and weaknesses with respect to building damage assessment. Optically sensed images are easy to interpret and typically available from most satellites (Vetrivel et al. 2016); however, they are typically unavailable at night or under cloudy or inclement conditions (Cao and Choe 2020) and with most sensors nadir-facing it is difficult to assess damages when the roof remains intact (e.g., pancake collapse) (Plank 2014). SAR is available under nearly all weather conditions and can provide a side-view to identify damages to a building's structure or façade (Yun et al. 2015). When paired SAR scenes (with different radar phases) are used it is easy to identify damages to a building's structure; however, it is not available on all satellites (potentially increasing the length between consecutive images) (Yun et al. 2015), it can be more difficult to interpret by a layperson, and the resolution can be a little coarser (Cao and Choe 2020). LiDAR can easily capture finescale three-dimensional images and thus is effective at identifying cracks and damages to different structures within a building (Dai et al. 2018; Zhou and Gong 2018); however it is expensive and does not work under bad weather conditions (Pirasteh et al. 2019). One additional disadvantage common to all sensors is the requirement of having both pre-event and post-event imagery of the same quality and resolution for purposes of change detection (Dong and Shan 2013; Wu et al. 2016). Polarimetric SAR is capable of assessing damages from post-event imagery only, but not all SAR sensors are polarimetric and this method has proven slightly less accurate when compared to using pre-event images as well (Bai et al. 2017; Zhai and Huang 2016).

Many recent studies have used remotely sensed imagery along with artificial intelligence methods such as random forest and convolutional neural networks to produce damage maps after a disaster. Cao and Choe (2020) used optical satellite imagery and a convolutional neural network to identify damaged buildings in the Houston area after Hurricane Harvey with 97% accuracy. Yun et al. (2015) generated damage proxy maps from pairs (pre- and post-event) of interferometric SAR coherence maps for the 7.8 magnitude Gorkha earthquake in Nepal. These maps showed good correlation with an independent damage assessment and were shared with local stakeholders. Vetrivel et al. (2016) used oblique imagery to test the use of the Visual-Bag-of-Words (BoW) image classification technique in identifying structural damage to buildings after earthquakes. They found that BoW outperformed conventional global feature representation with an average accuracy of around 90%.

While post hoc damage assessment provides a useful set of information regarding a building's resistance to damage, it is also helpful to model potential damages to buildings in future disasters. Such models make assumptions about the amount of force (from a given hazard) required to cause certain degrees of damage to buildings and often include estimates of the economic cost of such damages (Khajwal and Noshadravan 2020). Building damage models have recently been used to study total hurricane (wind and storm surge) damage (Baradaranshoraka et al. 2017), wind damage (Ham et al. 2018; Khajwal and Noshadravan 2020), and flooding (Jamali et al. 2018; Nafari et al. 2016). The major strength of the modeling approach is that it allows for the projection of future damages. This would allow for projecting how climate change might influence building damages (Gettelman et al. 2018) or where in a region might see the greatest damages from a hurricane (Masoomi et al. 2019). The greatest disadvantage is that, as models, they are subject to many assumptions and simplifications and thus can suffer from greater uncertainty in their results (Mishra et al. 2017).

Forecasting a Continuum of Environmental Threats and probabilistic hazard information

In the mid-2000s, the National Weather Service (NWS) of NOAA commissioned the National Research Council (NRC) to recommend ways in which the NWS could more effectively estimate and communicate uncertainty in weather and climate forecasts so as to improve public safety, property protection, and economic viability (NRC 2006). Currently, the U.S. weather enterprise, its customers, and other NWS stakeholders and end-users have been accustomed to receiving and using deterministic products, i.e., the warning polygon, which indicates that a hazard is occurring or will occur within the polygon and it will not occur outside the polygon. While a deterministic weather forecast can be easily deployed to a variety of communication channels, it is limited in its ability to deliver the dynamics and uncertainty of severe weather threats (see Fig. 3).

Forecasting a Continuum of Environmental Threats (FACETs) is a framework for a next-generation severe weather watch and warning system that is modern, flexible and designed to communicate clear and simple hazardous weather information to serve the public (Rothfusz et al. 2018). The key advantage of FACETs is "*the ability of sophisticated recipients and value-adding enterprises to "mine" user-specific, actionable information from this high-resolution continuum of data so as to feed a wide variety of probabilistic and deterministic displays, formats, and applications for an equally wide variety of end-users*" (Rothfusz et al. 2018). Recent research also highlighted opportunities for including probabilistic information to enhance the deterministic watch-warning system. As a part of FACETs, Probabilistic Hazard Information (PHI) provides custom user-specific products that can be tailored to adapt to a variety of needs – for example, providing longer lead times, at lower confidence, for more vulnerable populations with a lower tolerance for risks under a grid-based probability map (see Fig. 3). Also, it can be updated in real-time to reflect the dynamics of severe weather threats and support weather forecasters, emergency management agencies, and the public for better decision-making.

An overarching goal of FACETs is to deliver a continuous, rapidly updated stream of PHI at high spatial resolutions from months to minutes prior to an event, filling gaps in the existing, deterministic system. PHI may not completely replace the existing deterministic watch and warning systems. Also, it is still at an early development stage and multiple research efforts have been conducted to prove the effectiveness of PHI in terms of the details and type of information, visualization, and user interface needed to effectively communicate the dynamics of weather conditions to weather forecasters, broadcasters, emergency management agencies and the public. Thus, a cautious evolution to communicating more specific (and probabilistic) information within watches and warnings is the preferred approach. To improve PHI products and understand the effectiveness of PHI-driven output in the FACETs paradigm, multiple research projects have been conducted in the domains of physical science (Karstens et al. 2015a, 2018; Stumpf and Gerard 2021), and social behavioral and economic science (Klockow-McClain et al. 2019; Miran et al. 2018; Shivers-Williams and Klockow-McClain 2021).

Building classifications

Building classifications have been studied in multiple research domains and applications such as spatial science, geography, urban planning, architecture, energy, and disaster/emergency management. Two key elements for better building classification are (1) ground truth datasets and (2) classification methods.

Ground truth datasets are a key to enhancing the accuracy of building classifications. General datasets used for classification are building footprints, city zoning ordinance maps, building property information, topology, and satellite images. Rule-based or data-driven approaches have achieved high accuracy rates in classification, but have been tested with small datasets such as parts of cities or counties (Beck et al. 2020; Hecht et al. 2018; Tardioli et al. 2018; Wurm et al. 2016). Also, data-driven approaches can be difficult to apply if there are any strong regional dependencies and they would be applicable only for cities with similar building structures or types (Hecht et al. 2015). Recent studies have proposed crowd-sourced data collection for predictive modeling (Hecht et al. 2015, 2018).

Rule-based and data-driven approaches have been used very frequently in building classification. Rule-based approaches are based on sets of rules in terms of building typology information, construction year, and building final use. While they are intuitive and easy to apply, they require complete datasets to classify. Hussain and Shan (2016) showed that a well-structured rule set can provide a high accuracy classification rate (over 91%) with a small representative dataset. However, a rule-based classification may not perform well when a set of rules are too complicated or a building dataset does not contain detailed building property data. These can be partially overcome by using data-driven approaches. Tardioli et al. (2018) applied a random forest model in building clustering with a high accuracy ranging from 89.6% to 97%. Park and Guldmann (2019) also applied a random forest model to classify buildings (commercial, residential, skyscraper, and small constructions) by using building footprints and LiDAR point clouds.

We reviewed the literature to 1) identify the gap of knowledge in community damage assessment, and 2) analyze advanced severe weather watch and warning systems (PHI) and building classification methods. Existing community damage assessment models are mainly focused on post-disaster damage assessment using emerging technologies. Thus, they have been limited in their ability to support emergency management agencies and the public by sharing possible damages to a community from severe weather threats. To predict community risks using PHI and wind building damage models in the enhanced EF scale, a geodatabase with community assets (e.g., critical infrastructure and building property) is a key component. Thus, this study proposed a machine-learning based building classification model to create a geodatabase with multiple types of buildings categorized in the enhanced EF scale. The scope of this research is limited to residential building type classification.

Methodology

This study applied the Random Forest machine learning approach to develop a building type classification model using building footprint, property and city zoning data. It consisted of three steps: (1) data preparation, (2) random forest model development including training, testing and validation, and (3) hyper-parameter tuning (see Fig. 4).

Data preparation

We collected building footprints for Oklahoma City from the Microsoft Building Footprint dataset (Microsoft 2020) and parcel zoning information from the City of Oklahoma City (Oklahoma City Planning Commission 2020). We also collected public records for residential properties and buildings in Oklahoma City, OK from the Oklahoma County Assessor's office (Oklahoma County Assessor 2019) (see Table 3). These records included information about the location (address) of the property, the building type (e.g., mobile home, duplex), the type of construction (e.g., single-family home, high-rise apartment), the number of stories, and the exterior type (e.g., concrete block, frame siding) (see Table 4). We chose these variables as the National Weather Service's damage assessment relies on size, shape and construction materials used to classify buildings into damage indicator types (McDonald and Mehta 2006b). Since one dataset only provided a subset of the relevant information in identifying building type we combined the zoning, footprint and assessor records into one large dataset. We removed any properties that were either missing data or were not located within Oklahoma City, leaving a total of 159,752 records. We then geolocated the addresses for each building using the Google Geocoding API and performed a spatial join to connect the geolocated building information with the building footprints and zoning information. During this process, we removed any records we could not geolocate and were left with a final total of 139,296 geocoded buildings (see Fig. 5 for an example of how closely the geolocated addresses match the building footprint locations). We then designed an algorithm (Fig. 6) to assign a damage indicator (DI; based on the DI definitions from the National Weather Service's (NWS) Enhanced Fujita scale documentation (McDonald and Mehta 2006a)) to each building based on the building type, type of construction, exterior type and number of stories. We tested the success of the building classification algorithm by manually inspecting a small, random subset of images from Google Earth (see Fig. 7 for an example verification) and found that the algorithm performed well under this subset of images. We are confident that the simplicity of our damage indicator algorithm and the availability of information about building type and construction materials from the county assessor data negates the need for a more rigorous verification. We conducted the small-scale verification to show that the algorithm works as expected.

Random Forest Model development and hyper-parameter selection

A random forest (RF) is a classification algorithm adapting an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using a random-selected subset of variables (Breiman 2001). Random forest classification employs an ensemble method to attain the outcome (see Fig. 8). Ensemble methods basically create multiple models (decision trees) and combine them to generated improved results. In a random forest classifier, the selection of the final output follows the majority voting system (also known as hard voting) - every individual classifier votes for a class, and the majority wins.

In this study, the random forest classifier in the scikit-learn python library was applied to develop a random forest model (Pedregosa et al. 2012). In raw data, the class imbalance problem or imbalanced distribution of value, is very common. Thus, it is critical to examine a raw dataset and the distribution of class values. Machine learning techniques attempt to deal with imbalanced data by focusing on minimizing the error rate for the majority class while ignoring the minority class (Tanha et al. 2020; Thabtah et al. 2020). Multiple methods have been used to handle the imbalanced data classification problem such as SMOTE, RUSBoost, MEBoost, AdaCost, AdaC1, AdaC2 and AdaC3. We applied the SMOTE technique to resolve the imbalanced data problem. Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique that generates synthetic samples, not oversampling by replacements, for the minority class (Chawla et al. 2002). It helps to overcome the overfitting problem from random oversampling that is designed to randomly duplicate examples in the minority class. SMOTE is designed to generate new instances (data points) on the feature space with the help of interpolation between positive instances that lie together. It still has a few drawbacks: (1) It can increase the overlapping of classes and make additional data noise; (2) it may not be effective in the high dimensional data setting (Blagus and Lusa 2013).

The overall process of random forest model development is described in Fig. 8. In the data preparation and dataset stages, we performed a data balancing process and split data into training and test datasets. We designed a random forest model to predict building types based on input parameters including *area, perimeter, the number of vertices and city zone.* We then measured model performance by accuracy, and conducted hyper-parameter tuning to improve the model performance. Finally, the model was evaluated by accuracy, confusion matrix and classification metrics. Default hyper-parameters were {max_depth: 2, max_features: 'log2', n_estimators: 100, min_samples_leaf: 1} for training and testing. Maximum depth represents the depth of each tree in the forest. If the setting is None, there is theoretically no limitation of the maximum depth of the tree, and each tree will take more information from the input data. Maximum features is the maximum number of the used features at each node. There is a high chance of overfitting if the number of maximum features is too large (Lee and Kim 2020). To avoid overfitting, there are a couple of methods including (1) stop growing when splitting data is not statistically significant, (2) acquire additional training data, and (3) remove irrelevant attributes. The number of estimators is the number of total trees involved in ensemble learning for building an RF model. Mean samples leaf is the minimum number of samples required to be a leaf note. Basically, a higher number of hyper-parameters requires more computational cost. However, performance improvements can be negligible after a certain number of hyper-parameters.

To measure model performance, we analyzed the accuracy score, confusion matrix and classification metrics (precision, recall and F1 score). While accuracy is one of the most intuitive performance measures, it is simply a ratio of correctly predicted observations to the total observations. However, accuracy score itself is not a great measure of classifier performance when the classes or datasets are imbalanced. So, multiple studies have performed confusion matrix and classification metrics to understand how well the model performed (Hecht et al. 2015; Kang and Ryu 2019; Lee et al. 2017; Pérez-González et al. 2019). Classification metrics, including precision, recall and F1 score, are measured with a combination of true positive, true negative, false positive and false negative (Lever et al. 2016). *Precision* is estimated by the ratio of correctly predicted positive observations to the total predicted positive observations. For example, in our work the question that this metric answers is: of all the buildings that are labeled as townhouses, how many of them are truly townhouses? Thus, a higher precision rate is related to a low false positive rate. *Recall* is based on the ratio of correctly predicted positive observations to all observations in the actual class. *F1 Score* is calculated by the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. *F1 score* is considered as more useful than accuracy, especially when a dataset has an uneven class distribution. Accuracy may be considered highly useful when dataset is symmetric, or false positives and false negatives have similar costs. By using these three classification performance metrics, we can identify (1) whether or not a trained model performs equally well for each class and (2) any pairs of classes it found hard to distinguish.

For machine learning models, all input and output variables should be numeric (Cerda et al. 2018; Chuang and Keiser 2018; Marasco and Kontokosta 2016). For example, Oklahoma City zones are classified as 28 different categories. Thus, it is required to encode the categorical data to numbers before plugging them into a random forest model. The two popular techniques are ordinal encoding and one-hot encoding. Ordinal encoding coverts string labels to integer values 1 though n. One-hot encoding creates one column for each value for comparison against all other values (see Table 5). For categorical variables where no ordinal relationship exists (e.g. city zoning type or color of vehicles), integer encoding may mislead the model. Also, (1) forcing an ordinal relationship via an ordinal encoding and (2) allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (Zheng and Casari 2016). In this study, we applied a one-hot encoding that removed integer encoded variables and created a column with a new binary variable (0 or 1) for each unique integer value in the variable. As the one-hot encoding approach can expand input feature space exponentially, feature reduction (or dimension reduction) may be required.

Feature importance was also measured by the machine-learning model for a better understanding of the model's logic and to improve the model performance by focusing on the important variables (Ahmad et al. 2017). It can be utilized to select or reduce input variables so that the model can have similar or improved performance using less computational effort.

Results

Model performance was evaluated by (1) accuracy score, (2) confusion matrix and (3) precision, recall and F-1 score. Accuracy score computes the fraction of the count of correct predictions. In multiclass classification, the function returns the subset accuracy.

$$accuracy\ score\ (y_i, \hat{y}_i) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}} 1(\hat{y}_i = y_i)$$

$$y_i = corresponding\ true\ value, \hat{y}_i = predicted\ value\ of\ the\ i_{th}\ sample$$

Initial accuracy score under the default hyper-parameters (max_depth = 2, n_estimators= 100, max_features='log2', min_samples_leaf =2) was 0.631. Then we searched for optimal hyperparameter values for the number of estimators, max features, and min samples leaf. The number of estimators is known to be an important factor in determining model performance in a random forest. In a comparison of results for which we varied the number of estimators, the model performance fluctuated when the number of estimators was lower than 1,000 and was stable for values of 1,000 or more (see Fig. 10). Testing different values of max features and min sample leaf, the model performance score then achieved over 0.9. Using our selected hyperparameters (n_estimators= 1,000, max_features=0.2, mix_samples_leaf =1), the accuracy score reached up to 0.965.

While accuracy score is a popular method to evaluate the overall model, it does not provide classifier performance for each class. Thus, confusion matrix is broadly applied to evaluate model performance for each class. It can be used for both binary and multiclass classification problems. In confusion matrix, diagonal values are true positive counts, while off diagonal values are false positive and false negative counts for each class against the other (see

Table 6).

The confusion matrix in Fig. 11 shows the performance of the output of our RF model. The rows represent true labels and columns represent predicted labels. Values on the diagonal (black-colored cells) represent the percentage of times where the predicted label matches the true label. Values in the other cells represent the percentage where the classifier mislabeled an observation; the column indicates what the classifier predicted, and the row indicates what the right label was.

Type 2 (*one or two family residence*) was correctly classified as type 2 98% of the time, with incorrect classification rates of: type 3 (0.16%), type 4 (0.13%) and type 5 (0.38%). Similarly, type 3 (*single wide mobile home)* was classified as type 2 (0.52%), type 3 (96%), type 4 (1.8%) and type 5 (2.2%). Type 4 (*double wide mobile home)*, was classified as type 2 (0.41%), type 3 (3.5%), type 4 (97%) and type 5 (0.24%). Type 5 (*Apt, condo, and townhouse)*, was classified as type 2 (0.78%), type 3 (0.62%) type 4 (1.1%) and type 5 (95%).

Table 7 summarizes the classification reports of the model. The model has an average precision of 97%, recall of 97%, and F1-score of 97%.

The result of feature importance is in Fig. 12. The values of feature importance derived from our RF are: Zone (56%), Area (22.1%), Perimeter (16.3%), and the number of vertices (4.7%).

The result of the building classification by the proposed random forest model is described below. Brown-colored buildings are *one or two-family residence*, green-colored buildings are *single-wide mobile home*, bright blue-colored buildings are *double-wide mobile homes*, and dark blue-colored buildings are *Apt, condos and townhouses*.

Conclusions

Paradigm shifts in weather forecast and severe weather warning systems, from deterministic watch–warning products to high-resolution, probabilistic hazard information (PHI) spanning periods from days to within minutes of high-impact weather and water events, will improve a community's preparedness for and response to severe weather threats. To leverage the benefits of PHI in community risk assessment, it is critical to build a geodatabase that includes multiple types of community assets including critical infrastructure and building property. The current state of research shows that only a few studies have explored building type classification for community risk prediction. Also, the need for multiple types of data has made the proposed methods for city- and state-level building type prediction difficult to apply. Thus, this study developed a Random Forest model using building footprint data and city zoning maps to classify buildings and conducted a case study using the dataset from the Oklahoma City metropolitan area. The results show that the random forest model predicted building types with an accuracy of about 96%.

This study showcased the ability of a Random Forest model to predict building types based on building footprints and city zoning data for the residential areas of Oklahoma City. Using the training model and additional datasets, including city zoning data and building footprints, we achieved building type data in Edmond, Yukon, Midwest City, Moore and Norman (see Fig. 14).

To expand to other areas in Oklahoma City and beyond there is a need for significantly more training and testing data. While both county assessor records and city zoning data are available for most municipalities and counties, this data is not available digitally for all locations. Additionally, when available, this data is not housed in a central location so each county or city must be contacted individually to request the data and there may be charges for access. Another limitation, as we increase the size of our training and testing datasets, is the need to geocode each address in the county assessor data to match each building footprint with a property record. Each geocoding API has limitations in terms of accuracy and processing speed. This project tested two APIs: Google versus Nominatum (OpenStreetMap) and found that only Google was accurate enough to match addresses to building footprint locations. Unfortunately, large batch requests (on the order of millions of records) to the Google Geocoding API can take significant time to process and be costly. In addition to needing more testing and training data the machine learning algorithm may require modification to include additional variables to account for regional differences in construction materials and building techniques (e.g., adobe structures in the southwest).

Another limitation was in identifying the building type (as per the EF scale damage indicators) from the provided property records. While the EF scale documentation provides typical construction information for each building type (McDonald and Mehta 2006a) some of the characteristics are not unique to one building type (e.g., both 'Apartments, Townhouses, and Condos' and 'Masonry Apartments and Motels' can have brick exteriors). Additionally, Apartments, Townhouses, and Condos are supposed to be limited to three stories or less, but some buildings with a building type of apartment, condo or townhouse had four stories and an exterior type that was not solid masonry. In this case we assumed that the number of stories requirement was not absolute and that only buildings with a solid masonry exterior would be classified as Masonry Apartments and Motels. Another example of a questionable building type classification was for buildings with a construction type of Modular. Typically, Modular Homes are lumped together with Mobile Homes, but since no width data was available for these buildings, we assumed them to be equivalent to Double-Wide Mobile Homes since manufactured homes tend to be on the wider side. Issues such as these are likely to continue to arise as the dataset increases in size and we add commercial and industrial buildings, so future work will likely include a consultation with a trained damage surveyor to refine our methodology.

Another limitation is that building type data is imbalanced (e.g., the number of high-rise apartments is relatively small compared to one- or two-story houses, apartments, condos and townhouses in our dataset), which can introduce errors when training a machine learning model. To overcome imbalances in their datasets, Lemaitre et al. (2015) implemented state-of-the-art methods that can be categorized into four groups: (1) under-sampling, (2) over-sampling, (3) combination of over- and under-sampling, and (4) ensemble learning methods. These approaches have been applied in multiple domains to handle imbalanced datasets (Buda et al. 2018; Douzas et al. 2018; Korkmaz 2020; Lee and Kim 2020).

Future work may involve testing the use of other property record datasets that are more centralized, such as Zillow's Assessor and Real Estate Database (ZTRAX) in place of the county assessor data to reduce the time required to collect all necessary data. The Zillow Assessor and Real Estate Database has been used in previous hazards research but mostly for economic studies regarding hazards such as floods (Murfin and Spiegel 2020; Pinter and Rees 2021) and fires (Garnache and Guilfoos 2019; Mietkiewicz et al. 2020). In addition to data on property value, the ZTRAX dataset also includes detailed descriptions of buildings, including size, construction type and exterior type, which can be used to identify the NWS damage indicator and thus would complement the dataset derived in this study.

Additionally, we may attempt to use remotely sensed optical images from various satellite, aerial and ground-based platforms (e.g., Google Street View) to aid in identifying building types as proposed by Kang et al. (2018).

Data Availability Statement

The following datasets are available from the author (jooho.kim@noaa.gov) upon reasonable request.

- Building Footprint
- City zoning data
- Building property data

Acknowledgement

Funding was provided by the National Oceanic and Atmospheric Administration (NOAA) Office of Oceanic and Atmospheric Research (OAR) and Office of Weather and Air Quality (OWAQ) (Federal Grant Number: NA18OAR4590386 - Implementing convective storm statistics from a large reanalysis of WSR-88D data for model verification and forecasting probabilistic uncertainty)

References

Ahmad, M. W., Mourshed, M., and Rezgui, Y. (2017). “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption.” *Energy and Buildings*, Elsevier B.V., 147, 77–89.

Bai, Y., Adriano, B., Mas, E., and Koshimura, S. (2017). “Building damage assessment in the 2015 Gorkha, Nepal, Earthquake using only post-event dual polarization synthetic apertuRe radar imagery.” *Earthquake Spectra*, 33(Special issue 1), S185–S195.

Baradaranshoraka, M., Pinelli, J. P., Gurley, K., Peng, X., and Zhao, M. (2017). “Hurricane Wind versus Storm Surge Damage in the Context of a Risk Prediction Model.” *Journal of Structural Engineering (United States)*, 143(9), 1–10.

Beck, A., Long, G., Boyd, D. S., Rosser, J. F., Morley, J., Duffield, R., Sanderson, M., and Robinson, D. (2020). “Automated classification metrics for energy modelling of residential buildings in the UK with open algorithms.” *Environment and Planning B: Urban Analytics and City Science*, 47(1), 45–64.

Blagus, R., and Lusa, L. (2013). “SMOTE for high-dimensional class-imbalanced data.” *BMC Bioinformatics*, 14.

Brackett, R., and Childs, J. W. (2020). “Tornado Outbreak Tears Across the South, Killing 33.” *The Weather Channel*, <https://weather.com/storms/tornado/news/2020-04-13-tornadoes-storms-impacts-south-carolina-georgia-mississippi-arkansas> (Mar. 2, 2021).

Breiman, L. (2001). “Random Forests.” *Machine Learning*, 45, 5–32.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). “A systematic study of the class imbalance problem in convolutional neural networks.” *Neural Networks*, Elsevier Ltd, 106, 249–259.

Burgess, D., Ortega, K., Stumpf, G., Garfield, G., Karstens, C., Meyer, T., Smith, B., Speheger, D., Ladue, J., Smith, R., and Marshall, T. (2014). “20 May 2013 Moore, Oklahoma, Tornado: Damage Survey and Analysis.” *Weather and Forecasting*, 29(5), 1229–1237.

Cao, Q. D., and Choe, Y. (2020). “Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks.” *Natural Hazards*, Springer, 103(3), 3357–3376.

Cerda, P., Varoquaux, G., and Kégl, B. (2018). “Similarity encoding for learning with dirty categorical variables.” *Machine Learning*, Springer US, 107(8–10), 1477–1494.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). “SMOTE: Synthetic Minority Over-sampling Technique Nitesh.” *Journal of Artificial Intelligence Research*, 16(Sept. 28), 321–357.

Chuang, K. V., and Keiser, M. J. (2018). “Comment on ‘Predicting reaction performance in C–N cross-coupling using machine learning.’” *Science*, 362(6416), 1–3.

Dai, K., Li, A., Zhang, H., Chen, S. E., and Pan, Y. (2018). “Surface damage quantification of postearthquake building based on terrestrial laser scan data.” *Structural Control and Health Monitoring*, 25(8), 1–18.

Dong, L., and Shan, J. (2013). “A comprehensive review of earthquake-induced building damage detection with remote sensing techniques.” *ISPRS Journal of Photogrammetry and Remote Sensing*, International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 84, 85–99.

Douzas, G., Bacao, F., and Last, F. (2018). “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE.” *Information Sciences*, 465, 1–20.

Garnache, C., and Guilfoos, T. (2019). “Does the Salience of Risk Affect Large, Risky Asset Purchases?” *SSRN Electronic Journal*.

Gee, B., Bellware, K., Kornfield, M., Bella, T., and Cappucci, M. (2020). “Tennessee tornadoes kill at least 24 people, including children, and leave a trail of destruction.” *The Washington Post*, <https://www.washingtonpost.com/nation/2020/03/03/nashville-tornado/> (Mar. 2, 2021).

Gettelman, A., Bresch, D. N., Chen, C. C., Truesdale, J. E., and Bacmeister, J. T. (2018).

"Projections of future tropical cyclone damage with a high-resolution global climate model." *Climatic Change*, Climatic Change, 146(3–4), 575–585.

Ham, H. J., Yun, W., Choi, S. H., Lee, S., and Kim, H. J. (2018). "Quantitative wind risk assessment for low and mid-rise apartment buildings based on a probabilistic model." *Journal of Asian Architecture and Building Engineering*, 17(2), 377–384.

Hartmann, A., Meinel, G., Hecht, R., and Behnisch, M. (2016). "A workflow for automatic quantification of structure and dynamic of the German building stock using official spatial data." *ISPRS International Journal of Geo-Information*, 5(8).

Hecht, R., Kalla, M., and Krüger, T. (2018). "Crowd-sourced data collection to support automatic classification of building footprint data." *Proceedings of the ICA*, 1, 1–7.

Hecht, R., Meinel, G., and Buchroithner, M. (2015). "Automatic identification of building types based on topographic databases – a comparison of different data sources." *International Journal of Cartography*, Taylor & Francis, 1(1), 18–31.

Hussain, E., and Shan, J. (2016). "Urban building extraction through object-based image classification assisted by digital surface model and zoning map." *International Journal of Image and Data Fusion*, Taylor & Francis, 7(1), 63–82.

Insurance Information Institute. (2019). *Facts + Statistics: Tornadoes and thunderstorms*.

Jamali, B., Löwe, R., Bach, P. M., Urich, C., Arnbjerg-Nielsen, K., and Deletic, A. (2018). "A rapid urban flood inundation and damage assessment model." *Journal of Hydrology*, Elsevier, 564(February), 1085–1098.

James, J. J. J., Ling, C., Karstens, C. D., Jr., J. C., Calhoun, K., Meyer, T., and LaDue, D. (2020). "Forecasters' Cognitive Task Analysis and Mental Workload Analysis of Issuing Probabilistic Hazard Information (PHI) during FACETs PHI Prototype Experiment." *Weather and Forecasting*, 35(4), 1505–1521.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., and Zhu, X. X. (2018). "Building instance classification using street view images." *ISPRS Journal of Photogrammetry and Remote Sensing*, The Author(s), 145, 44–59.

Kang, K., and Ryu, H. (2019). "Predicting types of occupational accidents at construction sites in Korea using random forest model." *Safety Science*, Elsevier, 120(July), 226–236.

Karstens, C. D., Correia, J., LaDue, D. S., Wolfe, J., Meyer, T. C., Harrison, D. R., Cintineo, J. L., Calhoun, K. M., Smith, T. M., Gerard, A. E., and Rothfusz, L. P. (2018). "Development of a human-machine mix for forecasting severe convective events." *Weather and Forecasting*, 33(3), 715–737.

Karstens, C. D., Stumpf, G., Ling, C., Hua, L., Kingfield, D., Smith, T. M., Correia, J., Calhoun, K., Ortega, K., Melick, C., and Rothfusz, L. P. (2015a). "Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 hazardous weather testbed." *Weather and Forecasting*, 30(6), 1551–1570.

Karstens, C. D., Stumpf, G., Ling, C., Hua, L., Kingfield, D., Smith, T. M., Correia, J., Calhoun, K., Ortega, K., Melick, C., and Rothfusz, L. P. (2015b). "Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed." *Weather and Forecasting*, 30(6), 1551–1570.

Khajwal, A. B., and Noshadravan, A. (2020). "Probabilistic Hurricane Wind-Induced Loss Model for Risk Assessment on a Regional Scale." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 6(2), 1–9.

Klockow-McClain, K. E., McPherson, R. A., and Thomas, R. P. (2019). "Cartographic Design for Improved Decision Making: Trade-Offs in Uncertainty Visualization for Tornado Threats." *Annals of the American Association of Geographers*, Routledge, 0(0), 1–20.

Korkmaz, S. (2020). "Deep learning-based imbalanced data classification for drug discovery." *Journal of Chemical Information and Modeling*, 60(9), 4180–4190.

Kuligowski, E. D., Lombardo, F. T., Phan, L. T., Levitan, M. L., and Jorgensen, D. P. (2014). *Technical investigation of the May 22, 2011 tornado in Joplin, Missouri*. Gaithersburg, Maryland.

Lee, J., Jang, H., Yang, J., and Yu, K. (2017). “Machine Learning classification of buildings for map generalization.” *ISPRS International Journal of Geo-Information*, 6(10).

Lee, S., and Kim, J. (2020). “Prediction of Nanofiltration and Reverse-Osmosis-Membrane Rejection of Organic Compounds Using Random Forest Model.” *Journal of Environmental Engineering*, 146(11), 04020127.

Lemaitre, G., Nogueira, F., and Aridas, C. K. (2015). “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.” *Journal of Machine Learning Research*, 40(2015), 1–5.

Lever, J., Krzywinski, M., and Altman, N. (2016). “Classification evaluation.” *Nature Methods*, 13(8), 603–604.

Marasco, D. E., and Kontokosta, C. E. (2016). “Applications of machine learning methods to identifying and predicting building retrofit opportunities.” *Energy and Buildings*, Elsevier B.V., 128, 431–441.

Masoomi, H., Van De Lindt, J. W., Ameri, M. R., Do, T. Q., and Webb, B. M. (2019). “Combined Wind-Wave-Surge Hurricane-Induced Damage Prediction for Buildings.” *Journal of Structural Engineering (United States)*, 145(1), 1–15.

McDonald, J. R., and Mehta, K. C. (2006a). *A Recommendation for an Enhanced Fujita Scale (EF-Scale)*. Texas Tech University.

McDonald, J. R., and Mehta, K. C. (2006b). *A recommendation for an Enhanced Fujita scale (EF-Scale)*. Wind Science and Engineering Center, Texas Tech University.

Microsoft. (2020). “Building Footprints.” *Microsoft*, <https://www.microsoft.com/en-us/maps/building-footprints> (Oct. 1, 2020).

Mietkiewicz, N., Balch, J. K., Schoennagel, T., Leyk, S., St. Denis, L. A., and Bradley, B. A. (2020). “In the Line of Fire: Consequences of Human-Ignited Wildfires to Homes in the U.S.

(1992–2015)." *Fire*, 3(3), 50.

Miran, S. M., Ling, C., Gerard, A., and Rothfusz, L. (2018). "The effect of providing probabilistic information about a tornado threat on people's protective actions." *Natural Hazards*, Springer Netherlands, 94(2), 743–758.

Mishra, S., Vanli, O. A., Alduse, B. P., and Jung, S. (2017). "Hurricane loss estimation in wood-frame buildings using Bayesian model updating: Assessing uncertainty in fragility and reliability analyses." *Engineering Structures*, 135, 81–94.

Murfin, J., and Spiegel, M. (2020). "Is the Risk of Sea Level Rise Capitalized in Residential Real Estate?" *The Review of Financial Studies*, (J. Scheinkman, ed.), 33(3), 1217–1255.

Nafari, R. H., Ngo, T., and Mendis, P. (2016). "An assessment of the effectiveness of tree-based models for multi-variate flood damage assessment in Australia." *Water (Switzerland)*, 8(7).

NOAA/NCEI. (2018). "U.S. Billion-Dollar Weather and Climate Disasters." *NOAA National Centers for Environmental Information*, <https://www.ncdc.noaa.gov/billions/> (Mar. 10, 2018).

NOAA. (2011). "Tornadoes 101." <https://www.noaa.gov/stories/tornadoes-101> (Apr. 6, 2021).

NRC. (2006). *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. *National Research Council*, National Academies Press, Washington, D.C.

NSSL. (2015). "FACETs: Forcasting a Continuum of Environmental Threats." <https://www.nssl.noaa.gov/news/factsheets/FACETs_2015.pdf>.

Oklahoma City Planning Commission. (2020). "City Zoning Map." <https://www.okc.gov/departments/planning/planning-commission> (Oct. 1, 2020).

Oklahoma County Assessor. (2019). "Public Record." <https://assessor.oklahomacounty.org/203/Search-Property-Records> (Jul. 14, 2020).

Park, Y., and Guldmann, J.-M. (2019). “Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach.” *Computers, Environment and Urban Systems*, 75(January), 76–89.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2012). “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, 12, 2825–2830.

Pérez-González, C. J., Colebrook, M., Roda-García, J. L., and Rosa-Remedios, C. B. (2019). “Developing a data analytics platform to support decision making in emergency and security management.” *Expert Systems with Applications*, 120, 167–184.

Pinter, N., and Rees, J. C. (2021). “Assessing managed flood retreat and community relocation in the Midwest USA.” *Natural Hazards*, Springer Netherlands, (0123456789).

Pirasteh, S., Rashidi, P., Rastiveis, H., Huang, S., Zhu, Q., Liu, G., Li, Y., Li, J., and Seydipour, E. (2019). “Developing an algorithm for buildings extraction and determining changes from airborne LiDAR, and comparing with R-CNN method from drone images.” *Remote Sensing*, 11(11).

Plank, S. (2014). *Rapid damage assessment by means of multi-temporal SAR-A comprehensive review and outlook to Sentinel-1. Remote Sensing*.

Rothfusz, L. P., Schneider, R., Novak, D., Klockow-McClain, K., Gerard, A. E., Karstens, C., Stumpf, G. J., and Smith, T. M. (2018). “FACETs: A Proposed Next-Generation Paradigm for High-Impact Weather Forecasting.” *Bulletin of the American Meteorological Society*, 99(10), 2025–2043.

Shivers-Williams, C. A., and Klockow-McClain, K. E. (2021). “Geographic scale and probabilistic forecasts: a trade-off for protective decisions?” *Natural Hazards*, Springer Netherlands, 105(2), 2283–2306.

Simmons, K. M., Sutter, D., and Pielke, R. (2013). “Normalized tornado damage in the United

States: 1950-2011." *Environmental Hazards*, 12(2), 132–147.

Smith, A. B. (2020). *U.S. Billion-dollar Weather and Climate Disasters, 1980 - present*. NOAA National Centers for Environmental Information.

Stumpf, G. J., and Gerard, A. E. (2021). "National Weather Service Severe Weather Warnings as Threats-in-Motion." *Weather and Forecasting*, 1–58.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., and Asadpour, M. (2020). "Boosting methods for multi-class imbalanced data classification: an experimental review." *Journal of Big Data*, Springer International Publishing, 7(1).

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., and Finn, D. P. (2018). "Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach." *Building and Environment*, Elsevier, 140(May), 90–106.

Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A. (2020). "Data imbalance in classification: Experimental evaluation." *Information Sciences*, 513, 429–441.

Vetrivel, A., Gerke, M., Kerle, N., and Vosselman, G. (2016). "Identification of structurally damaged areas in airborne oblique images using a Visual-Bag-of-Words approach." *Remote Sensing*, 8(3), 1–22.

Wu, F., Gong, L., Wang, C., Zhang, H., Zhang, B., and Xie, L. (2016). "Signature analysis of building damage with Terra SAR-X new staring SpotLight mode data." *IEEE Geoscience and Remote Sensing Letters*, IEEE, 13(11), 1696–1700.

Wurm, M., Schmitt, A., and Taubenbock, H. (2016). "Building Types' Classification Using Shape-Based Features and Linear Discriminant Functions." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(5), 1901–1912.

Yun, S. H., Hudnut, K., Owen, S., Webb, F., Simons, M., Sacco, P., Gurrola, E., Manipon, G., Liang, C., Fielding, E., Milillo, P., Hua, H., and Coletta, A. (2015). "Rapid damage mapping for the 2015 Mw 7.8 Gorkha Earthquake Using synthetic aperture radar data from COSMO-SkyMed and ALOS-2 satellites." *Seismological Research Letters*, 86(6), 1549–1556.

Zhai, W., and Huang, C. (2016). "Fast building damage mapping using a single post-earthquake PolSAR image: A case study of the 2010 Yushu earthquake 4. Seismology." *Earth, Planets and Space*, Springer Berlin Heidelberg, 68(1).

Zheng, A., and Casari, A. (2016). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

Zhou, Z., and Gong, J. (2018). "Automated Analysis of Mobile LiDAR Data for Component-Level Damage Assessment of Building Structures during Large Coastal Storm Events." *Computer-Aided Civil and Infrastructure Engineering*, 33(5), 373–392.

Table 1 Tropical Cyclone Billion-Dollar Disasters to affect the U.S. from 2015-2020 (Smith 2020)

Name	Date	*Total CPI-Adjusted Cost (Millions of Dollars)	Deaths
Hurricane Matthew	2016-10-08	11,000	49
Hurricane Harvey	2017-08-25	131,250	89
Hurricane Irma	2017-09-06	52,500	97
Hurricane Maria	2017-09-19	94,500	2,981
Hurricane Florence	2018-09-13	24,720	53
Hurricane Michael	2018-10-10	25,724	49
Hurricane Dorian	2019-08-28	1,626	10
Tropical Storm Imelda	2019-09-17	5,050	5
Hurricane Hanna	2020-07-25	1,075	0
Hurricane Isaias	2020-08-03	4,757	16
Hurricane Laura	2020-08-27	18,990	42
Hurricane Sally	2020-09-15	7,274	5
Hurricane Delta	2020-10-09	2,867	5
Hurricane Zeta	2020-10-28	3,482	6
Tropical Storm Eta	2020-11-08	1,460	12

* Consumer Price Index (CPI): measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.

Table 2 Summary of literature related to building classification and data creation

Topic	**Method**	**Data type**	**Limitations**	**References**
Classification for energy modeling of residential building	Building geometry analysis (building footprint)	Ordnance Survey Master Map as a source of 2D polygons representing building feature Address database	2D based building classification can affect the accuracy of energy modeling for residential building	(Beck et al. 2020)
Identification of representative buildings and building groups	Rule-based classification Random forest for predictive classification	Geo-referenced data of the building stock and urban infra-structure and territorial information	Limited datasets for rule-based clustering.	(Tardioli et al. 2018)
Creation of ground truth dataset	Crowd-sourced data collection	Residential buildings in the city of Dresden and Hamburg	Quality of ground truth data depends on the quality of annotation. Data acquisition time from crowd-sourcing	(Hecht et al. 2018)
Building type classification	Machine-learning classification algorithms including naive Bayes, decision tree, k-nearest neighbor, and support vector machine	Digital maps obtained from the National Geographic Information Institute	Limited input parameters in machine learning Limited map generalization process	(Lee et al. 2017)
Building type classification	Linear discriminant analysis (LDA)	67,734 buildings in Berlin and 92,447 buildings in Munich	Limited transferability of the classifiers due to strong regional dependencies. However, it could vary by the level of classification	(Wurm et al. 2016)
Quantification of the structure and dynamic of national building stocks	Workflow for data integration Rule-based classification	Building footprint Address data from the real estate cadaster Land use	Knowledge-based approach: A hierarchical rule set to classify the buildings into a set of predefined classes. Data driven approaches such as pattern recognition and machine learning technique could enhance existing classification models	(Hartmann et al. 2016)
Building type classification	Random forest classifier	Topographic raster maps, cadastral databases or digital landscape models)	Data acquisition for training and testing	(Hecht et al. 2015)

Table 3 Data sources

Data Source	Description	# of Records
Oklahoma County Assessor	Public Records for Buildings in Oklahoma County	328,271
City of Oklahoma City	Parcel Zoning Map for Oklahoma City	21,025
Microsoft	Building Footprints for Oklahoma	2,091,131

Table 4. Description of all fields in the combined building dataset. Field represents the name of the field with units in parentheses

Field	Definition	Type	Examples
Feature ID	Unique identifier for each building polygon	Identifier	4619; 6209
Area (sq. m)	Area of building polygon in square meters	Numeric	Mean=272.8; Max=27740.3
Perimeter (m)	Perimeter of building polygon in meters	Numeric	Mean=68.5; Max=1262.7
Zoning Ordinance	City zoning code for parcel where building polygon is located	Categorical [28]	R-2; R-3
Account Number	Account number for Oklahoma County Assessor record	Identifier	R211831080; R141245800
Account Type	Type of account for Oklahoma County Assessor record	Categorical [2]	Residential; Commercial
Physical Address	Physical address for parcel of land	Identifier	101XX Mantle Ct, Oklahoma City, OK
Building Number	Numeric identifier for the building (in case parcel has multiple buildings)	Numeric	Mean=1; Max=8
Parcel Number	Real Estate parcel ID number	Identifier	2709044044290; 1898143581260
Parcel Type	Type of parcel	Categorical [8]	Residential; Townhouse
Square Footage	Area of parcel in square feet	Numeric	Mean=1640; Max=347,168
Exterior Type	Type of exterior material for building	Categorical [36]	Frame Masonry Veneer; Hardboard Sheet
Number of Stories	Number of stories for building	Numeric	Mean=1; Max=24
Roof Type	Type of roof construction for building	Categorical [11]	Hip/Gable; Gable
Roof Covering	Type of roof covering material for building	Categorical [12]	Composition Shingle; Formed Seam Metal
Foundation Type	Type of foundation for building	Categorical [7]	Slab; Conventional
Construction Type	Type of building construction	Categorical [31]	Ranch 1 Story; 1½ Story Fin
Year Built	Four-digit year when building was constructed	Numeric	Mean=1966; Max=2019
Damage Indicator	Damage indicator number from EF scale documentation	Categorical [7]	2; 3
Damage Indicator Name	Name of damage indicator from EF scale documentation	Categorical [7]	One- and Two-Family Residences; Single-Wide Mobile Home
Damage Indicator Code	Code for damage indicator from EF scale documentation	Categorical [7]	FR12; MHSW
Number of Vertices	Number of vertices in building polygon	Numeric	Mean=5.7; Max=32

Note: Type is the data type with categorical variables containing the number of categories in brackets, Examples gives example values for each field with numeric fields containing the mean and maximum values. Example zoning ordinances include Medium-Low Density Residential (R-2) and Medium Density Residential (R-3).

Table 5 Data table of input building records after One-Hot encoding

	AREA	PERIM	NVERT	ZONE_AA	ZONE_R-1	ZONE_R-2	...	ZONE_C-1	ZONE_C-2
0	284.849593	86.602737	8	0	1	0	...	0	0
1	130.500664	45.696733	4	0	1	0	...	0	0
2	227.678254	61.497334	4	0	1	0	...	0	0
3	568.924061	98.021346	4	0	0	1	...	0	0
4	193.884134	59.667975	6	0	0	1	...	0	0
...	...	...	...	...	...	...	...	...	...

Table 6 Confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 7 Precision, recall and F1-score

	Precision	Recall	F1-score	Support
Type 2	0.99	0.98	0.99	40,237
Type 3	0.95	0.96	0.96	40,414
Type 4	0.94	0.97	0.95	40,379
Type 5	0.97	0.95	0.96	40,171
Weighted avg	0.97	0.97	0.97	161,171

Table of Figures

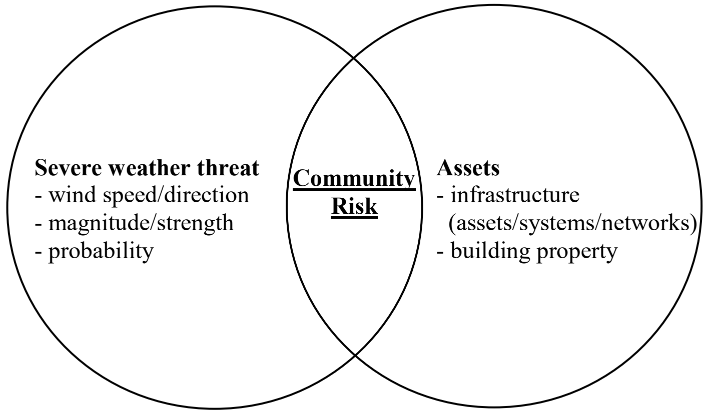
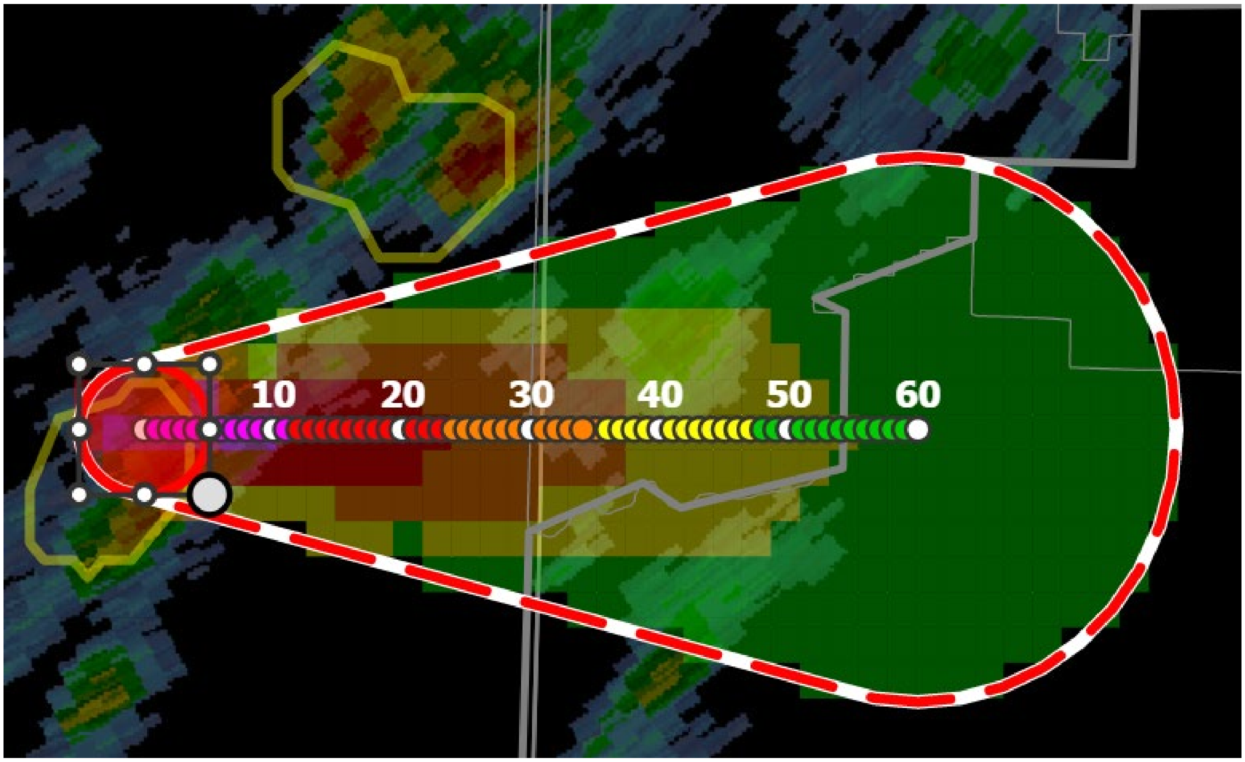
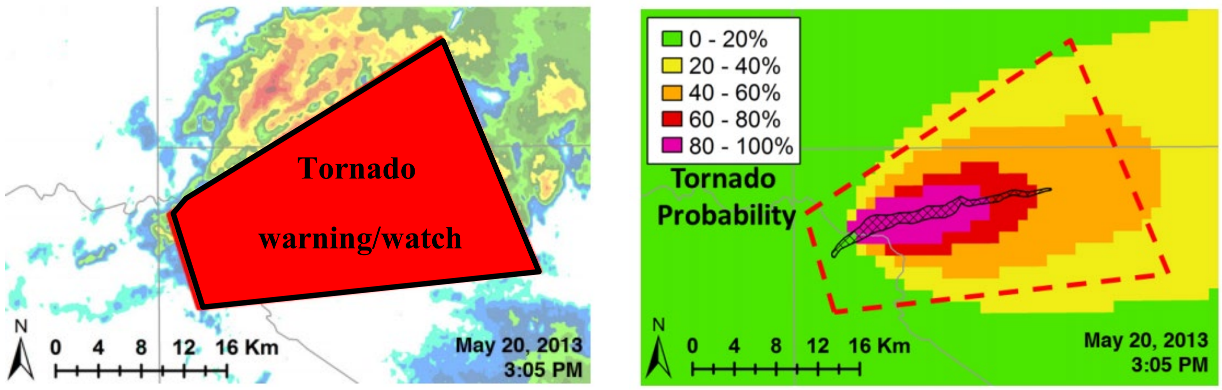


Fig. 1 Community risk from severe weather threats



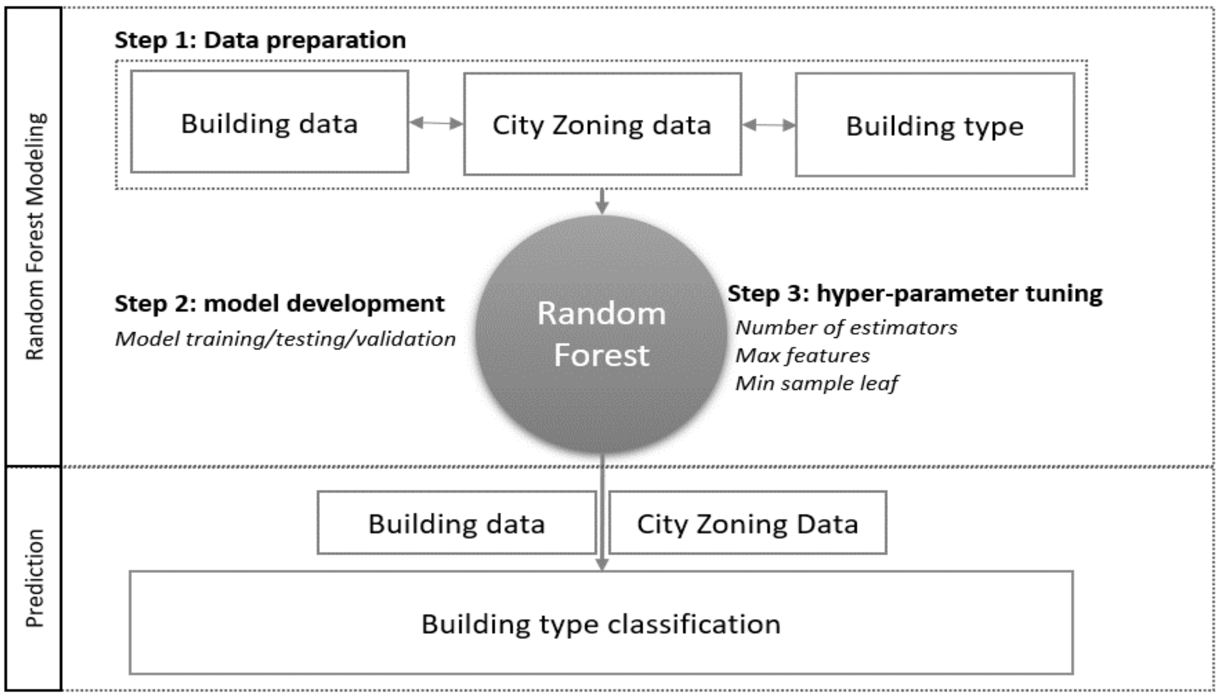
Note: PHI provides the risk probability on a grid map in real-time. The red-colored areas have higher probability of tornado risk (path) than the green-colored area

Fig. 2 Probabilistic Hazard Information (PHI)



(a) Deterministic method (b) Probabilistic method

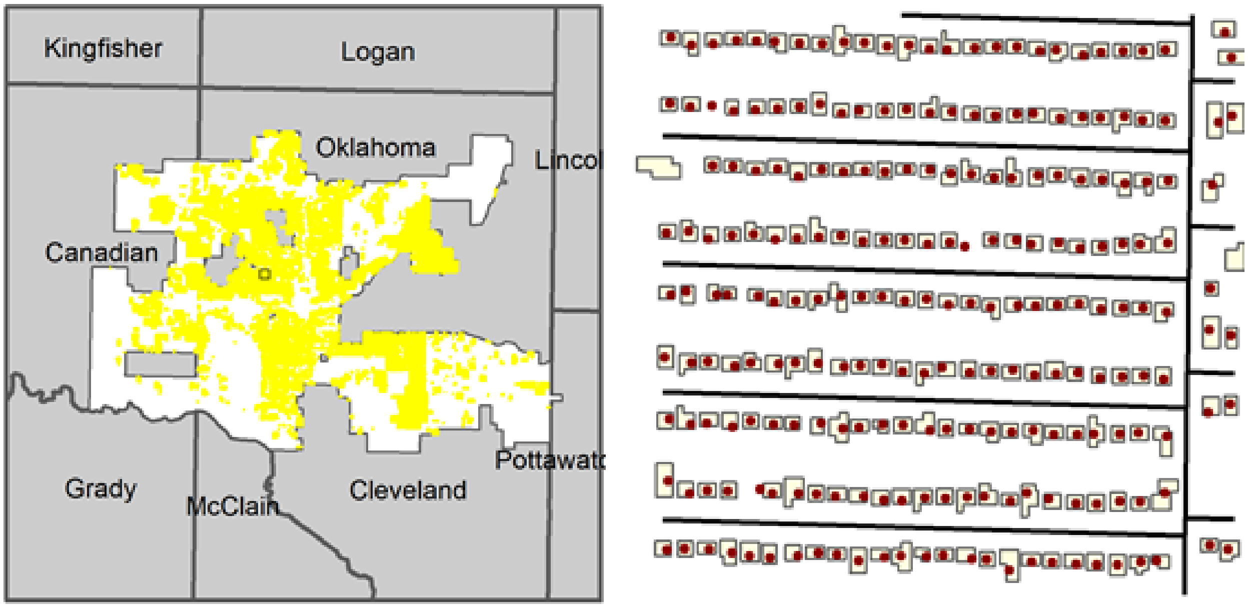
Fig. 3 Deterministic vs. probabilistic weather information (NSSL 2015)



341

342 Fig. 4 Framework of random forest model development

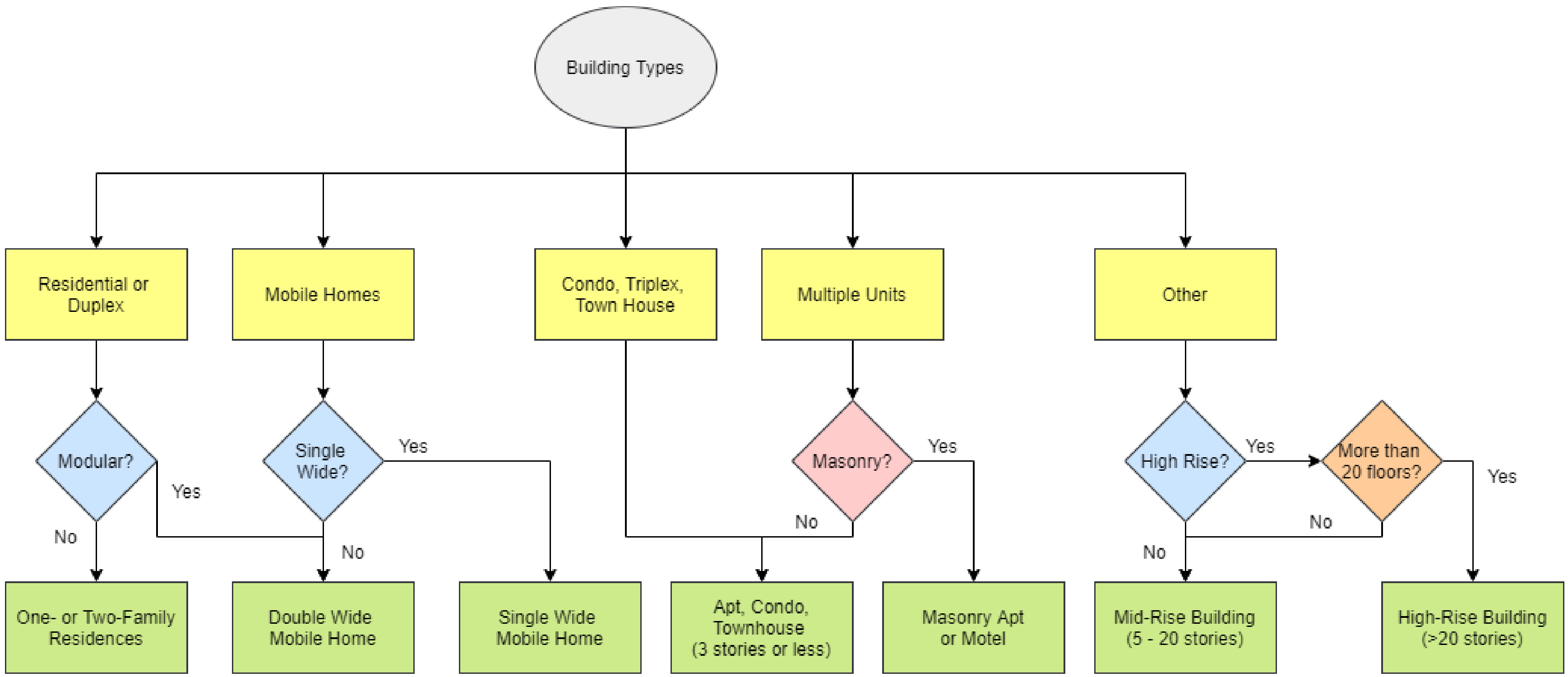
343



(a) Residential Zones in Oklahoma City (b) Sample Building Footprints with Geocoded Address Points

Note: Panel a shows all of the parcels that are zoned Residential in Oklahoma City in yellow while panel b shows a zoomed in view of the location marked by the box in a showing the local roads (black lines) and building footprints (yellow polygons) as well as the geolocated positions of the residential addresses, in the Oklahoma County Assessor's public records, (red dots). Note how each address (red dot) is closely associated with a single building footprint, showing the accuracy of the geolocated positions.

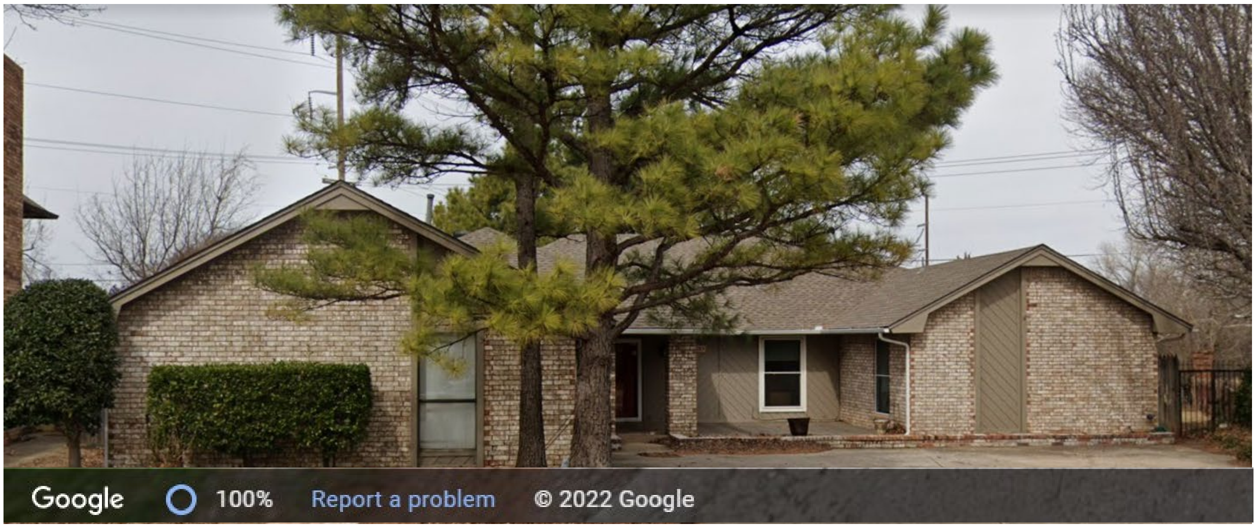
Fig. 5. Example of Geocoded Addresses in the Residential Areas of Oklahoma City



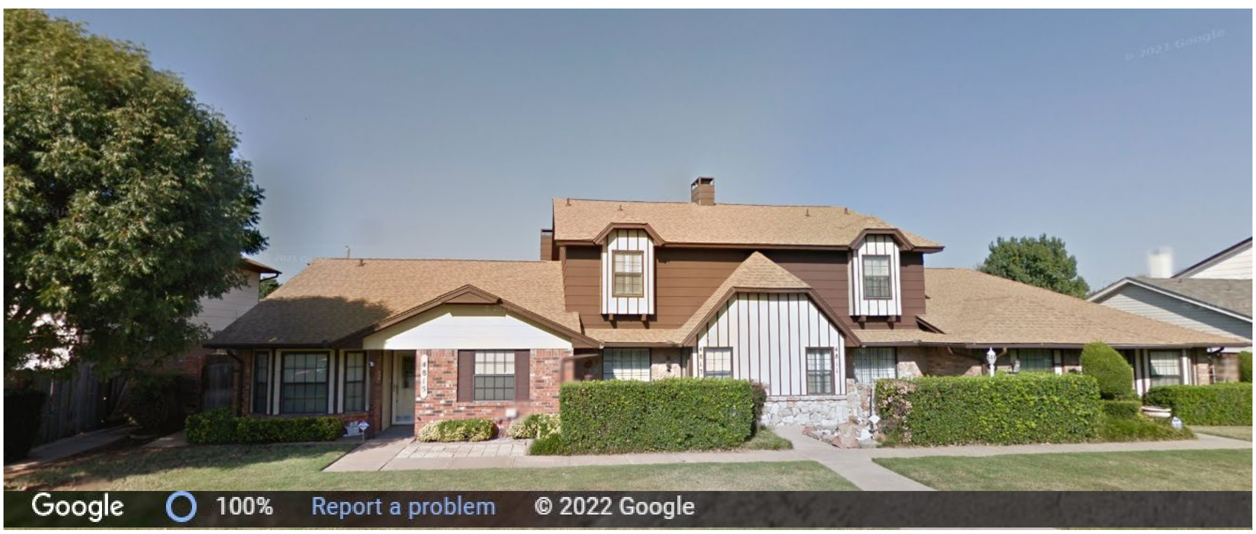
Note: The algorithm requires the following information about a building to determine its Damage Indicator: building type (yellow), type of construction (blue), exterior type (pink), and number of stories (orange)

Fig. 6 Building type classification algorithm.

(a) Example of One- and Two-Family Homes Category



(b)Example of Apartments, Condominiums, and Townhouses Category



Map data© 2022 Google

Note: Panel a is of a Single Family Home (classified as a One- to Two- Family Home (DI=2)) located at 4017 NW 70th Street, Oklahoma City, OK, panel b is of a Multi-Family Home (classified as an Apartment, Condominium or Townhouse (DI=5)) located at 4815 NW 72nd Street, Oklahoma City, OK

Fig. 7. Example images from Google Earth for verification of classification algorithm.

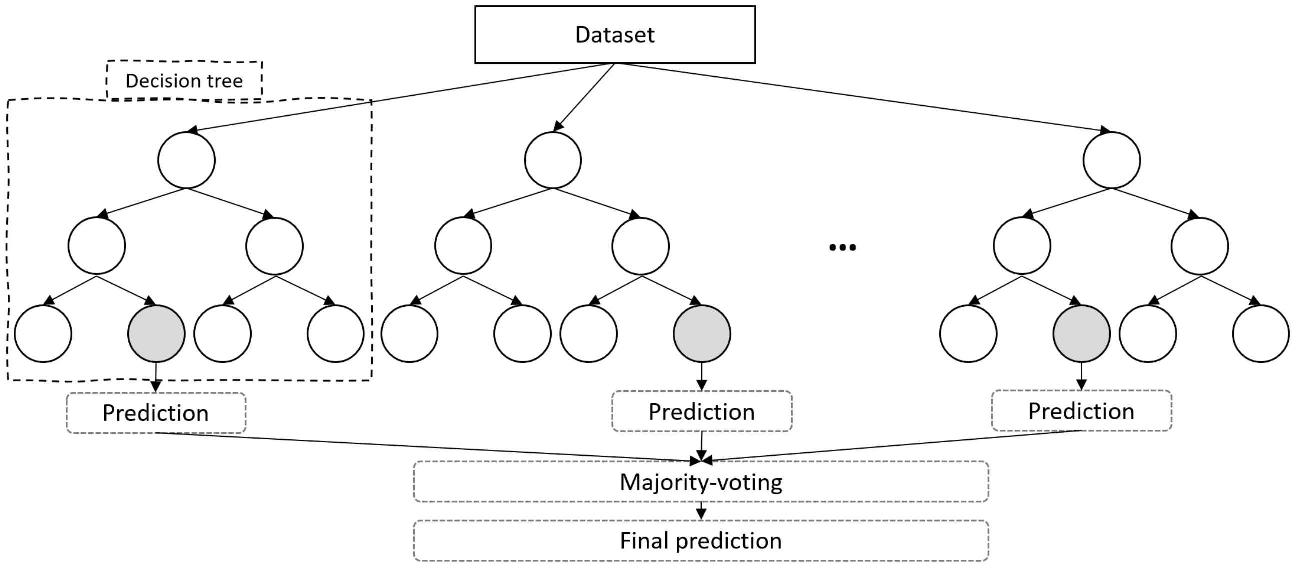


Fig. 8 Random Forest classifier

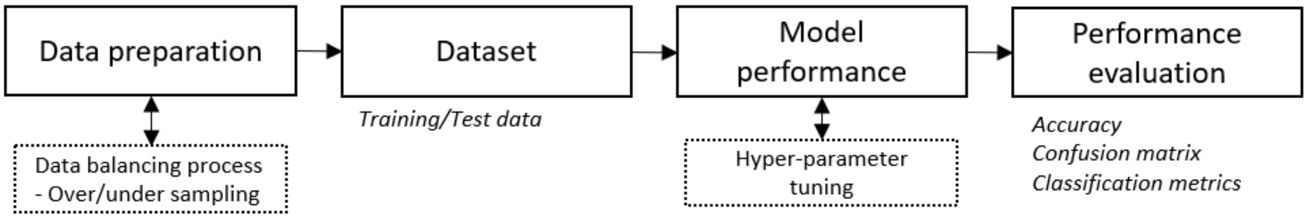
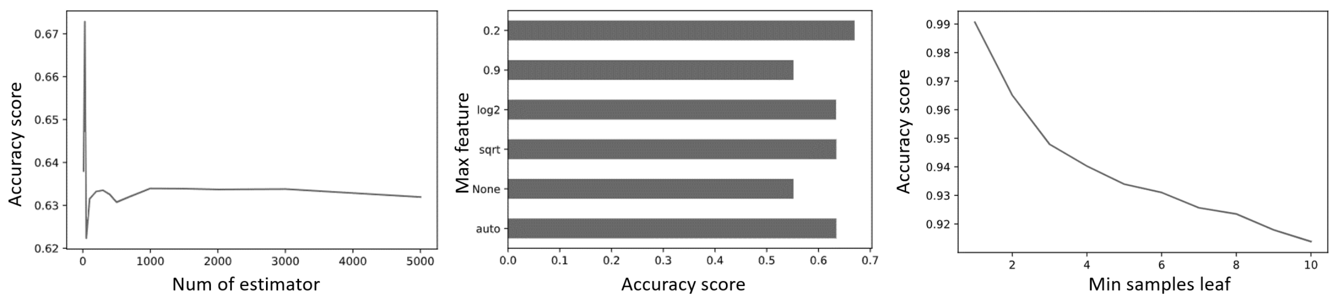


Fig. 9 Process of random model development



(a) Num of estimator (b) Max features (c) Min samples leaf

Fig. 10 Hyper-parameters and model performance

Predicted label

True label	***Type 2***	***Type 3***	***Type 4***	***Type 5***
Type 2	*0.98* *(39,553)*	*0.0052* *(208)*	*0.0041* *(163)*	*0.0078* *(313)*
Type 3	*0.0016* *(63)*	*0.96* *(38,707)*	*0.035* *(1,395)*	*0.0062* *(249)*
Type 4	*0.0013* *(52)*	*0.018* *(737)*	*0.97* *(39,131)*	*0.011* *(429)*
Type 5	*0.0038* *(154)*	*0.022* *(895)*	*0.024* *(973)*	*0.95* *(38,149)*

The actual number of data points for each cell is shown in parentheses

Note: type 2 (one- or two-family residence), type 3(single wide mobile home), type 4 (double wide mobile home) and type 5 (Apt, condo, townhouse). Actual amount of data is in parentheses

Fig. 11 Confusion matrix

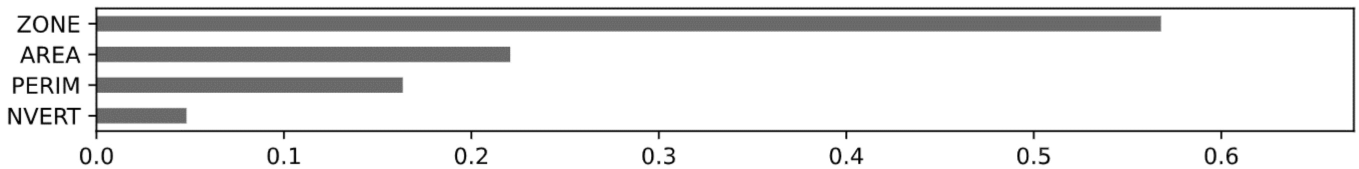


Fig. 12 Feature importance

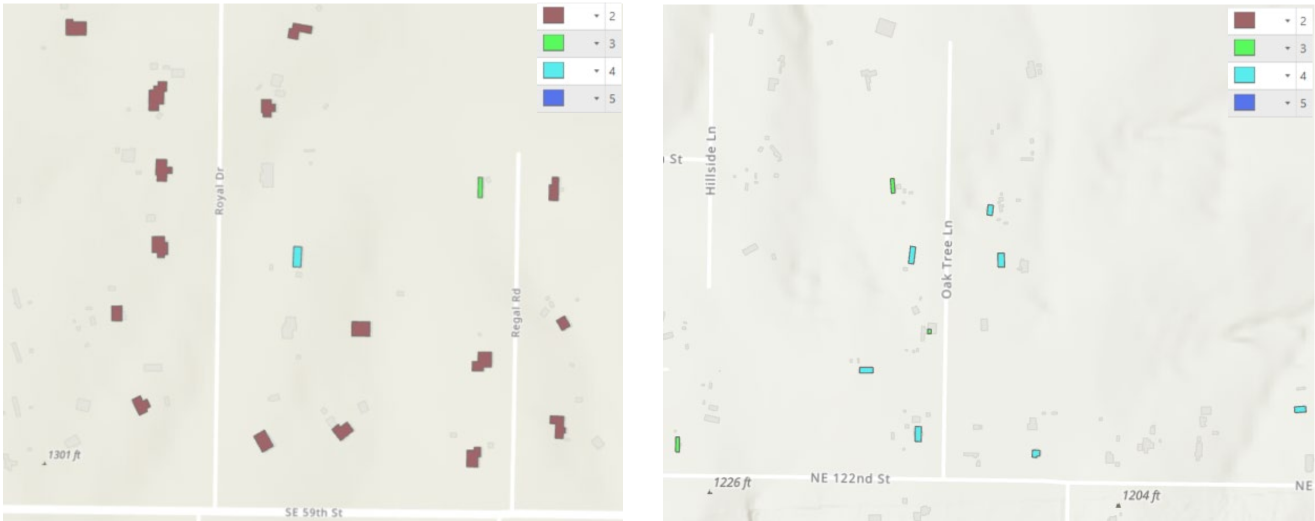


Fig. 13 Classified building footprints for a small section of Oklahoma county

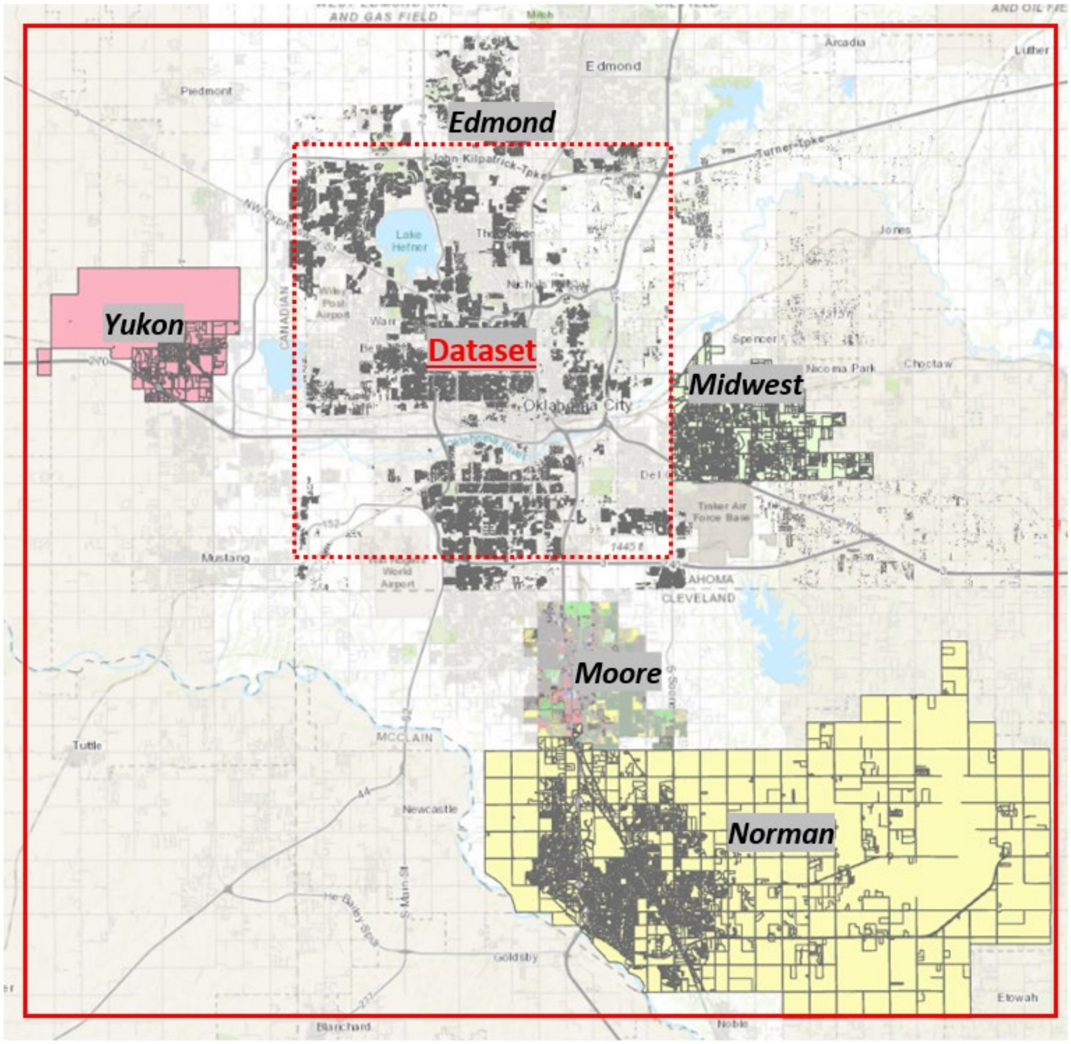


Fig. 14 Building type geodatabase creation using the trained ML model in the Oklahoma metropolitan city area