Modeling Earth Systems and Environment (2023) 9:2371–2384
https://doi.org/10.1007/s40808-022-01627-9

ORIGINAL ARTICLE

Canopy classification using LiDAR: a generalizable machine learning approach

R. Sky Jones[1] · Racha Elkadiri[1] · Henrique Momm[1]

Received: 9 August 2022 / Accepted: 1 December 2022 / Published online: 18 December 2022

Abstract

The integration of machine learning algorithms with LiDAR-derived datasets has long been used in the development of canopy classification models with the main objective of differentiating between tree canopies and other types of land cover. However, these integrated canopy classification models often require investigators to provide training reference data, use vendor classification codes that vary in quality and availability, and/or rely on site-specific information to achieve the required accuracy. In this study, a generalizable canopy classification model based solely on LiDAR-derived datasets is proposed and evaluated. Ten watersheds that are located in different regions in the continental USA were selected to represent a wide range of physiography, topography, climate, tree diversity, and LiDAR data characteristics, ensuring model applicability to different environments. Three canopy classification model development strategies were considered: general, specific, and single. The final decision tree-based general canopy classification model contains five datasets with the roughness of the filtered DHM yielding the highest normalized feature importance of 0.9. The developed general canopy detection model accuracy was comparable to of the specific/single models and it generated an average testing kappa statistic of 0.90 and 0.96 when applied to training/testing and testing datasets, respectively. This study demonstrated the existence of a consistent canopy signal in LiDAR datasets across the contiguous US that can be used to create a general canopy classification models that are functional regardless of the study area or LiDAR quality.

Keywords Canopy classification · Generalizable classification model · LiDAR · Machine learning · Point cloud · Decision trees

Introduction

Canopy detection models, which seek to differentiate tree canopy cover from all other land cover types, are used to track reforestation/deforestation efforts, to assess wildfire risk, quantify potential carbon capture, and support air pollution mitigation efforts among other applications and research inquiries (de Sosa et al. 2018; Herzfeld et al. 2014; Derwin et al. 2020). Land cover classification is one of the oldest and most widely used applications of geospatial technologies (Giles 2002). Traditional classification models have utilized aerial or satellite-based imagery, particularly multi-spectral imagery, in conjunction with machine learning algorithms, to quantify land cover and assess their changes over time (Lary et al. 2016; Zhang et al. 2016; Kussul et al. 2017; Lei et al. 2017; Maxwell et al. 2018; Jamali 2020). Advancements in light detection and ranging (LiDAR) technology and significant increases in the availability of LiDAR-derived datasets over the last two decades has allowed the use of LiDAR for land cover classification (Dubayah and Drake 2000; O'Neil-Dunne et al. 2013). LiDAR-derived datasets were used either in conjunction with other remotely sensed imagery or as the sole source of data. Raw LiDAR data consists of a set of irregularly spaced three-dimensional georeferenced points. Each point is described by a set of attributes, such as the intensity of the backscattered signal, information about the number of returned pulses, and the scan angle. Rasterization of the attribute data encoded in a point cloud is used to overcome the irregular sampling characteristic of LiDAR point clouds and to produce continuous raster grid datasets depicting various aspects of the Earth's surface, depending on the parameters used to produce the raster (Guiotte et al. 2020).

✉ Henrique Momm
henrique.momm@mtsu.edu

[1] Department of Geosciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA

LiDAR-supplemented classification models have achieved remarkable classification quality, even with very granular schemes that attempt to recognize many feature classes in complex systems (Goodwin et al. 2009; MacFaden et al. 2012; Muftah et al. 2022). The fact that LiDAR captures the physical structure of sensed objects has extended the limits of classification studies by enabling tree crown delineation (Holmgren and Lindberg 2019; Zhen et al. 2016) or even using tree structure for tree species identification (Heinzel and Koch 2011; Wang et al. 2019; You et al. 2021). Furthermore, the fusion of LiDAR-derived datasets and hyperspectral imagery represents the state of the art for land cover classification by allowing the integration of spectral information and surface physical characteristics into advanced machine learning algorithms (Ghamisi et al. 2015; Kuras et al. 2021; Zhang and Lin 2017; Luo et al. 2015).

Though custom-trained machine learning models are extremely accurate, they are limited by two factors. First, machine learning algorithms require the input of training data (reference data) that can be time-consuming to produce and also be prone to sampling bias or training error, particularly when they are created by inexperienced users. Second, the trained model is commonly only applicable to the dataset and location on which the model was developed (i.e., the model is specialized to those conditions). Specifically, when working with LiDAR datasets, a canopy classification model may not even provide satisfactory results if a different dataset of the same geographic extent, but collected by a different vendor is used. Differences in collection methodologies such as flight pattern and/or altitude, sensor specifications and characteristics, collection date and time, and post-processing of raw data can alter the sensed characteristics of ground features (García-Quijano et al. 2008). Datasets from different vendors can differ in the sampling intensity (number of points per unit of area), number of returns, pulse intensity, pulse classification, and overall accuracy of coordinates and elevation measurements. Similarly, inherent characteristics of sites can pose challenges for model's generalization capability, as canopy classification models developed in one site may not perform in another site with different physical properties. A model trained, for example, in an urban area, will likely not produce satisfactory results in an agricultural or forested area or even in a different urban area.

Thus, these site- and dataset-specific models can yield accurate results, but require expert knowledge to generate representative training datasets and a new model must be created for each new study area or dataset. A generalizable model is one that could be applied to any dataset in any study area by using land cover signals that are consistent regardless of site characteristics and/or data collection technology. Therefore, such robust classification model could be used without the need to re-generate and use additional training data. For example, LiDAR data such as pulse return intensity must be excluded from a generalizable classification model because it is a function of the sensor's signal strength and recording sensitivity and flight properties rather than inherent land cover physical characteristics. It is possible that there would be limited correlation between land cover type and backscatter intensity across different vendor even though a correlation might exist within a single vendor (Hopkinson 2007). A potential challenge, however, is excluding some of the LiDAR-derived datasets to support the overall generalization ability of the classification model, but inducing the elimination of valuable information limiting accuracy and granularity.

In this study, a generalizable canopy detection model based on LiDAR datasets is proposed and evaluated. Specifically, the objectives of this investigation are: (1) develop a generalizable canopy detection model based solely on LiDAR data, (2) identify key LiDAR primary and secondary datasets that are agnostic to site- and dataset-specific attributes, and (3) evaluate and quantify the proposed general model performance in different conditions and compare to site-specific models.

Materials and methods

Study sites and primary datasets

A total of ten watersheds across the contiguous US were selected for this study (Fig. 1; Table 1). Primary datasets consist of LiDAR three-dimensional point clouds and high-resolution aerial imagery. The LiDAR datasets were provided in LAS file format and downloaded from the United States Geological Survey (USGS) websites EarthExplorer (https://earthexplorer.usgs.gov) and The National Map (https://nationalmap.gov). LiDAR datasets were generated by distinctive vendors at different periods of time (Table 1).

Aerial orthoimages were acquired from the National Agricultural Imagery Program (NAIP) within 4 years prior of each LiDAR mission (with the exception of HUC 080902030201, for which aerial imagery was collected 1 year after the LiDAR mission). These images were used as the primary source of information to generate reference datasets for each of the ten watersheds (Table 1). Canopy locations were manually digitized by a trained image analyst, who worked in all ten watersheds to minimize the potential bias inherent from work generated by multiple analysts. The decision-making process was based on the use of NAIP imagery and complemented by multi-temporal images available in the Google Earth platform and from review of LiDAR-derived raster grids. The generated GIS layers were quality controlled by the co-authors of this study. The selected study sites are described in the following sections.

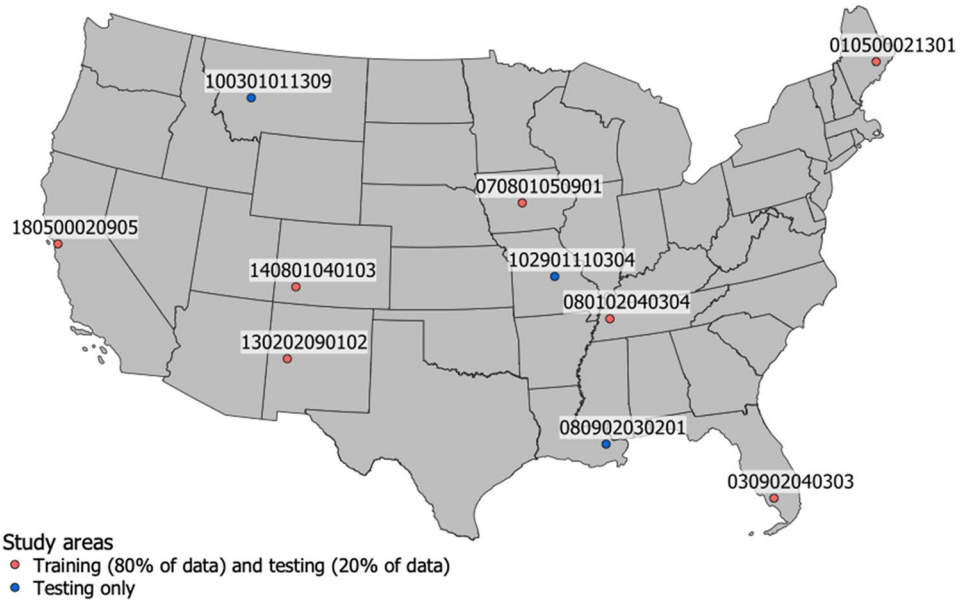

Fig. 1 Spatial distribution of selected watersheds with respective US Geological Survey 12-digit hydrologic unit code. Seven watersheds were used for training and testing and three for testing only

Table 1 Selected training and testing watersheds and respective LiDAR characteristics

HUC12	Location	Dominant land cover type	Size km^2	Digitizing		Nominal spacing, m	Vendor	End of mission
				km^2	%			
Watersheds used for model training and testing								
010500021301	Penobscot, ME	Rugged coastal	80.0	10.4	13.0%	0.70	Quantum Spatial	2015
030902040303	Naples, FL	Flat coastal	74.7	8.1	10.8%	0.67	Woolpert	2008
070801050901	Ames, IA	Agricultural	50.7	8.8	17.3%	0.88	Sanborn	2010
080102040304	Trenton, TN	Agricultural	161.1	10.3	6.4%	0.70	Laser Mapping Specialists	2011
130202090102	Datil, NM	Mountain/desert	66.1	7.7	11.6%	0.71	Merrick-Surdex	2018
140801040103	Telluride, CO	Mountain	136.8	13.2	9.6%	0.35	Woolpert	2017
180500020905	San Francisco, CA	Urban	25.3	1.3	5.1%	0.60	Earth Eye	2010
Watersheds used for model testing								
080902030201	New Orleans, LA	Urban	61.2	4.0	6.6%	0.70	Fugro Earthdata	2017
100301011309	Helena, MT	Urban/arid	55.5	3.2	5.8%	1.46	Sanborn	2012
102901110304	Freeburg, MO	Agricultural	77.8	6.1	7.8%	1.33	Photo Science	2010

Penobscot, Maine

Branch Lake watershed (USGS hydrologic unit code: 010500021301) is located approximately 10 miles to the northeast of the coastal town of Penobscot, Maine, and drains to the eponymous Branch Lake. The watershed has a total area of 80.0 km^2. Most of the watershed consists of coniferous forest and open water. Limited agricultural land, open fields and exposed rock are present as well. Overall development in the watershed is low. Relief within the watershed is moderate. Approximately 10.4 km^2, or 13.0%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 4/19/2015 and 5/12/2015 with a nominal pulse spacing of 0.7 m.

Naples, Florida

Strand State Preserve watershed (USGS hydrologic unit code: 030902040303) is located approximately 20 miles to the east of the town of Naples in southwestern Florida. The watershed is named for the Fakahatchee Strand State Preserve and is a Florida State Park. The watershed has a total area of 74.7 km^2. The watershed is dominated by herbaceous forest, shrubs and wetlands. There is very limited development and relief within the watershed. Approximately, 8.1

km^2, or 10.8%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 6/12/2007 and 3/8/2008 with a nominal pulse spacing of 0.67 m.

Ames, Iowa

Walnut Creek watershed (USGS hydrologic unit code: 0070801050901) is located approximately 5 miles to the south of the city of Ames, Iowa. The watershed has a total area of 50.7 km^2. The watershed is dominated by agricultural land, with sporadic deciduous tree clusters and riparian buffers. Development is limited to sporadic housing and connecting roads. Relief within the watershed is very limited. Approximately 8.8 km^2, or 17.3%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 4/7/2007 and 5/27/2010. The LiDAR has a nominal pulse spacing of 0.88 m.

Trenton, Tennessee

North Fork Forked Deer River Middle watershed (USGS hydrologic unit code: 080102040304) partially encompasses the town of Trenton in western Tennessee. The watershed has a total area of 161.1 km^2. Land cover in the watershed consists primarily of agricultural fields and thick deciduous-dominated riparian buffers. There are sporadic townships throughout the watershed. Relief within the watershed is modest. Approximately 10.3 km^2, or 6.4%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 12/2/2011 and 1/4/2012 with a nominal pulse spacing of 0.70 m.

Datil, New Mexico

Ox Spring Canyon watershed (USGS hydrologic unit code: 130202090102) is located approximately 10 miles north of the town of Datil, New Mexico. The watershed drains a portion of the northern face of Madre Mountain, part of the Datil mountain range. The watershed has a total area of 66.1 km^2. Land cover in the watershed is a mix of sand and sandy loam, including dry streambeds, and sparse, shrubby ligneous cover at lower elevations and relatively denser mixed deciduous and coniferous forest at higher elevations. During the winter, snow may be present at higher elevations. Development in the watershed is nonexistent apart from occasional dirt roads. Relief within the watershed is significant as the watershed encompasses an area that includes mountain peaks at its head and flat scrubland near its terminal point. Approximately, 7.7 km^2, or 11.6%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 11/2/2017 and 2/8/2018 with a nominal pulse spacing of 0.71 m.

Telluride, Colorado

Mineral Creek watershed (USGS hydrologic unit code: 140801040103) is located approximately 5 miles south of the town of Telluride, Colorado. The watershed is situated within the Rocky Mountains and has a total area of 138.6 km^2. Land cover in the watershed is diverse and consists of bare rock, grass fields, alluvial deposits, melt-fed lakes and mixed coniferous–deciduous forest. Some amount of snow and ice is typically present perennially, though snow cover can be ubiquitous during the winter. Development in the watershed is limited to footpaths. Relief within the watershed is significant. Approximately 13.2 km^2, or 9.6%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 9/6/2017 and 10/24/2017 with a nominal pulse spacing of 0.35 m.

San Francisco, California

Lobos Creek-Frontal San Francisco Bay Estuaries watershed (USGS hydrologic unit code: 180500020905) encompasses part of the city of San Francisco, California. The watershed has a total area of 25.3 km^2. Land cover in the watershed largely consists of high-density urban development, including some high-rise buildings. However, a significant amount of plant cover is present as the watershed encompasses Presidio Park and part of the Golden Gate Park in addition to numerous smaller green spaces. A significant number of individual trees are also present along streets throughout the watershed. Thus, herbaceous and woody plant cover comprise an appreciable portion of the land cover in the watershed. San Francisco is a famously hilly city, and relief within the watershed is moderate relatively to the drastic relief in some of the other watersheds included in this study. Approximately, 1.3 km^2, or 5.1%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 4/23/2010 and 7/14/2010 with a nominal pulse spacing of 0.6 m.

New Orleans, Louisiana

Morrison Canal watershed (USGS hydrologic unit code: 080902030201) encompasses part of the city of New Orleans, Louisiana, and has a total area of 61.2 km^2. Land cover in the watershed consists of medium density urban development, industrial yards, mowed fields, wetlands and open water. Limited agglomerations of trees are also present. Relief within the watershed is minimal. Approximately, 4.0 km^2, or 6.6%, of the watershed was manually classified. The used LiDAR data for this watershed was

collected between 1/23/2017 and 4/24/2017 with a nominal pulse spacing of 0.70 m.

Helena, Montana

Last Chance Gulch watershed (USGS hydrologic unit code: 100301011309) encompasses part of the city of Helena, Montana. The watershed has a total area of 55.5 km^2. Though a portion of the watershed is within the city limits of Helena and thus partially comprises low to medium density urban development, most of the watershed is within a rugged, undeveloped mountain-side topography draining to the southwest. Land cover in this drainage consists of open grassy areas punctuated by sparse shrub cover on south-facing slopes and dense mixed coniferous–deciduous tree cover on northerly slopes. Relief within the watershed is significant. Approximately, 3.2 km^2, or 5.8%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 5/8/2012 and 5/9/2012 with a nominal pulse spacing of 1.46 m.

Freeburg, Missouri

Loose Creek–Maries River watershed (USGS hydrologic unit code: 102901110304) encompasses part of the town of Freeburg, Missouri. The watershed has a total area of 77.8 km^2. The watershed's land cover is dominated by agricultural land and open fields punctuated by relatively large stands of deciduous trees. Sporadic rural development is also present. Relief within the watershed is moderate. Approximately, 6.1 km^2, or 6.8%, of the watershed was manually classified. The used LiDAR data for this watershed was collected between 4/14/2010 and 5/7/2010 with a nominal pulse spacing of 1.33 m.

Derived datasets

A total of 26 raster grid products were developed from rasterizing LiDAR point clouds using their respective attributes. Custom Python scripts were developed utilizing a suit of existing libraries (WhiteboxTools, laspy, GDAL, Orfeo, and scipy) to process the original LiDAR datasets in LAS file format representing irregular point distribution and generate a wide range of raster grid products in regular spatial arrangement (Table 2). Because the desired classification model is intended to be generalizable, no generated raster grid products made use of proprietary tools/algorithms (i.e., classification). For raster grids created by interpolating LiDAR point cloud values, a triangular irregular network (TIN) algorithm was used. All raster grids created used a 1-m spatial resolution, regardless of the original LiDAR point sampling density. Illustrations of some of the raster grids are provided in Fig. 2.

Elevation-based features

A digital surface model (DSM) is a raster grid created by interpolating the elevation values from only the LiDAR pulses representing first returns. Thus, the DSM represents the elevation of the Earth's surface including aboveground structures such as buildings and tree canopies.

A digital elevation model (DEM) is a raster grid similar to a DSM, but is created by interpolating the elevation values of LiDAR pulses identified as the last returns. Thus, the DEM represents the elevation of the Earth's surface

Table 2 Summary of LiDAR-derived raster grids used in the canopy model development

Feature type	Raster grid name	Used to train model	Appears in decision tree
Elevation-based features	Digital surface model	No	No
	Digital elevation model	No	No
	Digital height model	No	No
	Filtered digital height model	Yes	No
Non-elevation-based features	Intensity raster grid	No	No
	Return raster grid	No	No
	Slope raster grids (from surface, elevation, height and filtered height models)	Yes	Yes (from DEM and DSM)
Textural features	Roughness raster grids (from surface, elevation, height and filtered height models)	Yes	Yes (from fDHM and DSM)
	Laplacian-filtered raster grids (from surface, elevation, height and filtered height models)	Yes	Yes (from DSM)
	Haralick textures (generated for digital height model only)	No	No

A total of 26 products were generated

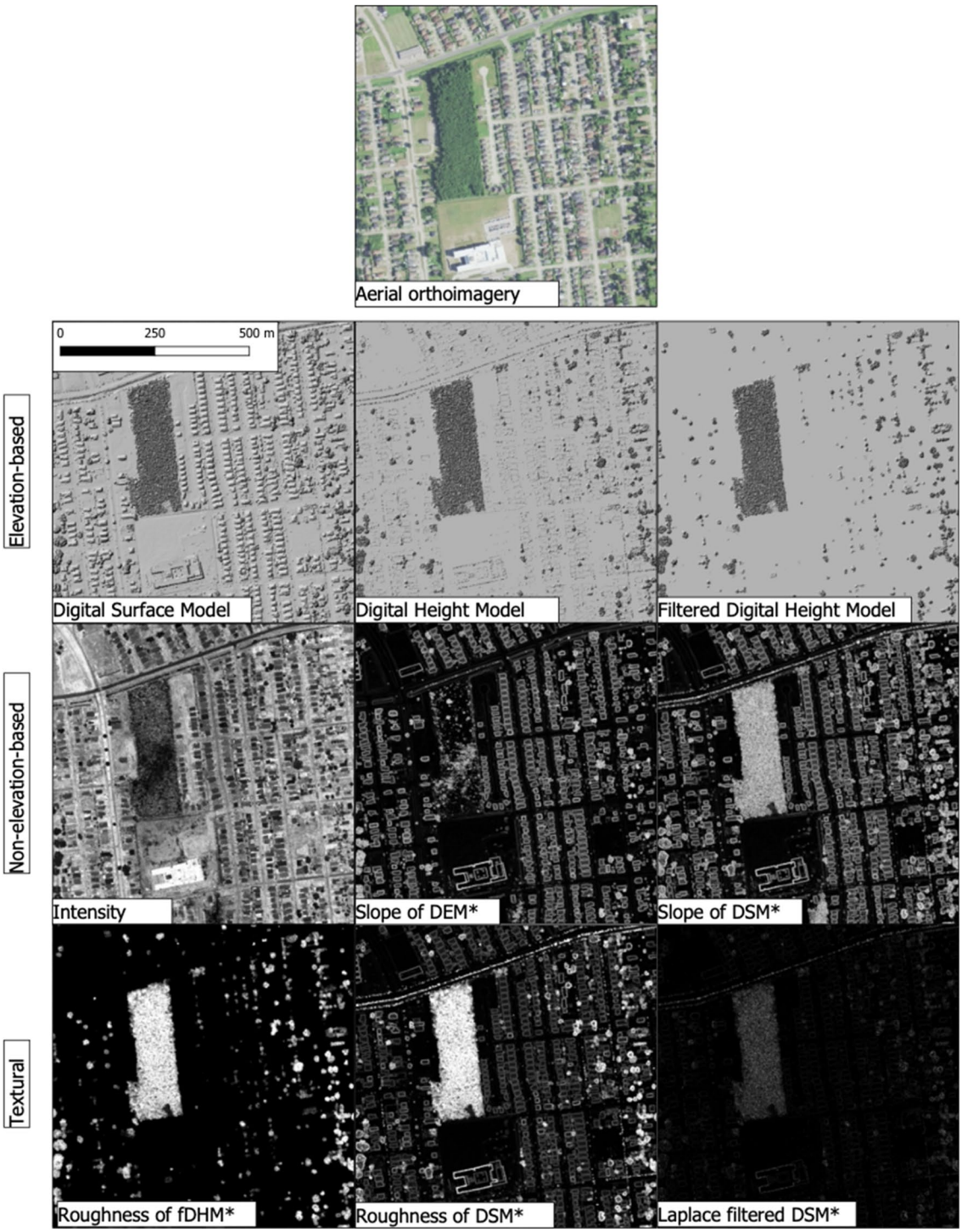

Fig. 2 Representation of selected raster grids derived from LiDAR for a portion of the study area in watershed 080902030201 (New Orleans, LA) as an example of the generated inputs. Note that the elevation-based raster grids are displayed as hillshade raster grids for improved visualization. *Indicates features that are included in the final model

excluding some aboveground structures and tree canopies, often referred to as "bare ground". Some versions of the DEM also require removal of built structures, but because this study did not utilize vendor-supplied classification codes, buildings could not be removed from the DEM directly and so our generated DEM products may have included some built structures.

A digital height model (DHM), sometimes referred to as a normalized digital surface model (nDSM) or height-aboveground model, is a raster grid representing the result of the subtraction of the DEM from the DSM. The DHM represents the distance between the ground and any overhead structures, such as tree canopies or powerlines. Because building edges often produce multiple returns, they frequently appear in DHMs. The DHM is often used as a rough proxy for canopy height because its signature is dominated by trees, while building edges, powerlines, and other anthropogenic structures are minor contributors.

Though the DHM often serves as a proxy for canopy height, anthropogenic structures such as building edges and powerlines contribute significantly to the DHM in urban areas. Because of this, investigators have proposed methods to distinguish the signature of trees from built structures through various methods, such as filtering the linear signatures representing powerlines or building edges (Maxwell et al. 2018). While many methods exist to filter built structures from a DHM, we elected to use a simple density filter due to its ease of implementation, quick runtime, and high selectivity for canopy signals. The filter works such that, for every target pixel in the DHM, a 3 by 3 window encompassing its eight neighbors is used. All pixels within the window are evaluated and if over 30% of the pixels (3 or more) within the window are nonzero, it causes the target pixel's value to be retained. Otherwise, the target pixel is set to 0. This filter tends to remove linear features while retaining more rounded features such as tree canopies. The result is a filtered DHM referred to as fDHM.

Non-elevation-based features

An intensity raster grid representing the average backscatter return intensity for all LiDAR points at each pixel was generated through standard GIS interpolation. Intensity values are not normalized and are a function of the sensor type, outgoing pulse strength, and aircraft flight altitude in addition to the inherent reflectance of the material being surveyed. Intensity raster grids were used qualitatively in the process of manual land cover classification.

For each LiDAR pulse emitted by the sensor, it is possible to receive multiple backscatters (returns) which, in turn, would yield multiple points. Modern LiDAR systems record the total number of returns and the respective return number for each backscatter. For example, the same pulse could generate two points where both points would have a value of 2 as the number of returns attribute. However, they would have the different return number attribute values, with one point assigned a value of 1 (indicating first return) and the other a value of 2 (indicating last return). A return raster grid representing the average number of returns in a pixel was created by interpolating the return number of the last return for each LiDAR pulse. Similar to the intensity raster grid, the number of returns is a function of variables including information beyond the land cover type and varies with sensor characteristics and post-processing parameters.

For each elevation-based raster grid (DSM, DEM, DHM, fDHM), a slope raster grid was created representing the maximum slope between a given pixel and its eight neighbors. The slope was generated in degrees.

Textural features

Texture is often defined as the frequency of elevation/tonal changes in a raster grid. Based on this definition, a texture "roughness" raster grid was created for each of the elevation-based raster grids (DSM, DEM, DHM, fDHM) by applying a filter that returns the maximum difference in values between a target pixel and its eight neighbors, often referred to as range.

Similarly, Laplacian filters are also based on moving windows and calculate the value of the target pixels as the weighted average of all the pixels within the moving window.

Laplacian filters are commonly used to emphasize areas of rapid change of tone or discontinuities in imagery (Kimmel and Bruckstein 2003). This emphasis on rapid image textural change has led to widespread use of the Laplace filter for edge detection, but a Laplace-filtered image can be thought of more broadly as a textural derivative of the parent image. Laplacian filters are often referred to as convolution operations and the used convolution matrix representing the weights (Eq. 1). A Laplacian-filtered raster grid was created for each elevation-based model (DSM, DEM, DHM, fDHM).

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}. \quad (1)$$

Haralick textures, or Haralick features, refers to a set of image/grid texture operators derived from a matrix describing frequency of occurrence of specific value pairs within subsets of an image (Haralick et al. 1973). This frequency matrix is called the gray-level co-occurrence matrix (GLCM) and the derived features are referred to as textures, because they tend to quantify the relationships between a given pixel and its neighbors. From the DHM, we generated eight Haralick textures: energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence and Haralick correlation.

Raster grid exclusions

Although a total of 26 raster grids were generated and used as inputs into a machine learning algorithm, not all the input raster grids were used in the development of the general model. Raster grids DEM and DSM were excluded because meaningful comparison of raw elevation values between watersheds is not possible, and thus their inclusion provides little additional predictive power. Similarly, the DHM was also excluded because the fDHM is a better approximation of canopy coverage as the DHM contains significant building edge artifacts. The intensity and return raster grids were both excluded, because intensity and return values are in large part of a function of sensor and flight characteristics rather than landform properties, making quantitative comparison between missions impossible.

All remaining raster grids were used as input in the development of the proposed general canopy classification model using machine learning algorithm. The objective was to derive relationships between datasets, generate classification rules and, more importantly, select key datasets based on their predictive power.

Canopy classification model design and evaluation

The selected machine learning classification algorithm in this study was decision trees. A decision tree scheme, represented by a branching series of conditional statements, was chosen over other machine learning algorithms because its candidate solutions are user readable, easily interpretable, and can be translated into a set of if–then-else statements, thus simplifying their incorporation into GIS software packages. In addition, decision trees and decision tree-like machine learning algorithms have been reported to perform highly in canopy discrimination using LiDAR data (Michałowska and Rapiński 2021) and other remote sensing datasets (Hakam et al. 2022; Kumar et al. 2021). The decision tree was created using a variation of the CART regression technique for decision tree optimization implemented in the scikit-learn v0.23.1 Python library. Input data was weighted such that each study area contributed equally to the model.

Out of the ten considered watersheds, seven were selected for training and testing and three for testing only. In watersheds selected for training and testing, the reference datasets were randomly divided into 80 and 20%, for training and testing, respectively. Three canopy modeling strategies were developed: general, specific, and singular.

The canopy classification model, referred to as general, was developed using 80% of the reference data from all seven watersheds. This model was then evaluated against the remaining 20% of the data from the seven watersheds, as well as 100% of the data from the three testing watersheds.

Canopy classification models specific to individual watersheds, referred to as "specific models", were also constructed for comparison against the general model. The general and specific models used similar input raster grid types. The only difference was the inclusion of the intensity and the return number raster grids which were included in specific models to better simulate conditions of traditional classification models that use all collected LiDAR-derived datasets. An additional model was generated using only data from HUC 10500021301 to test the generalizability of models trained in a single region (referred to as the "singular" model).

Accuracy scores using the overall accuracy (% of pixels correctly classified) and the Kappa coefficient of agreement (Cohen's κ) statistics were calculated individually for each watershed in this study. The Cohen's κ represents an accuracy statistic designed to account for the random chance of agreement or bias in the dataset (Cohen 1960). Kappa coefficient of agreement values are interpreted as agreement compared to the random chance of agreement. Values greater than 0.75 are considered excellent agreement beyond the random chance of agreement (Landis and Kock 1977). The resulting overall accuracy and Kappa scores for each model (general, specific, and singular) as applied to each watershed were calculated. The model design and implementation steps are summarized in Fig. 3.

Results

The final general canopy classification model is presented in Fig. 4. As a result of the iterative development of the canopy classification model by the machine learning algorithm, five raster grids were present in the final solution: filtered digital height model roughness (FIDHR), digital surface model roughness (DSMRO), Laplace filter of digital surface model

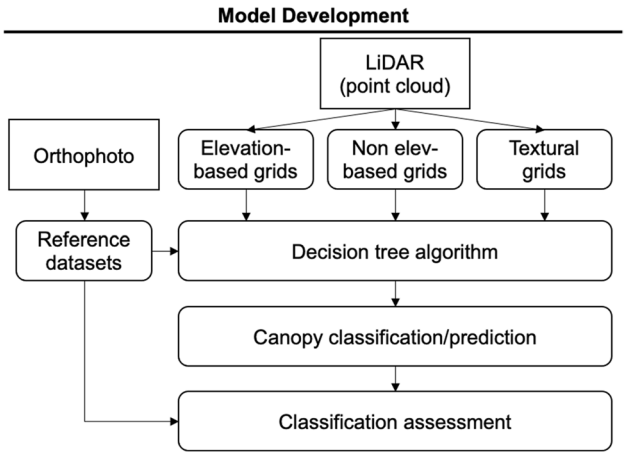

Fig. 3 General overview of the design and implementation of the classification model development and evaluation

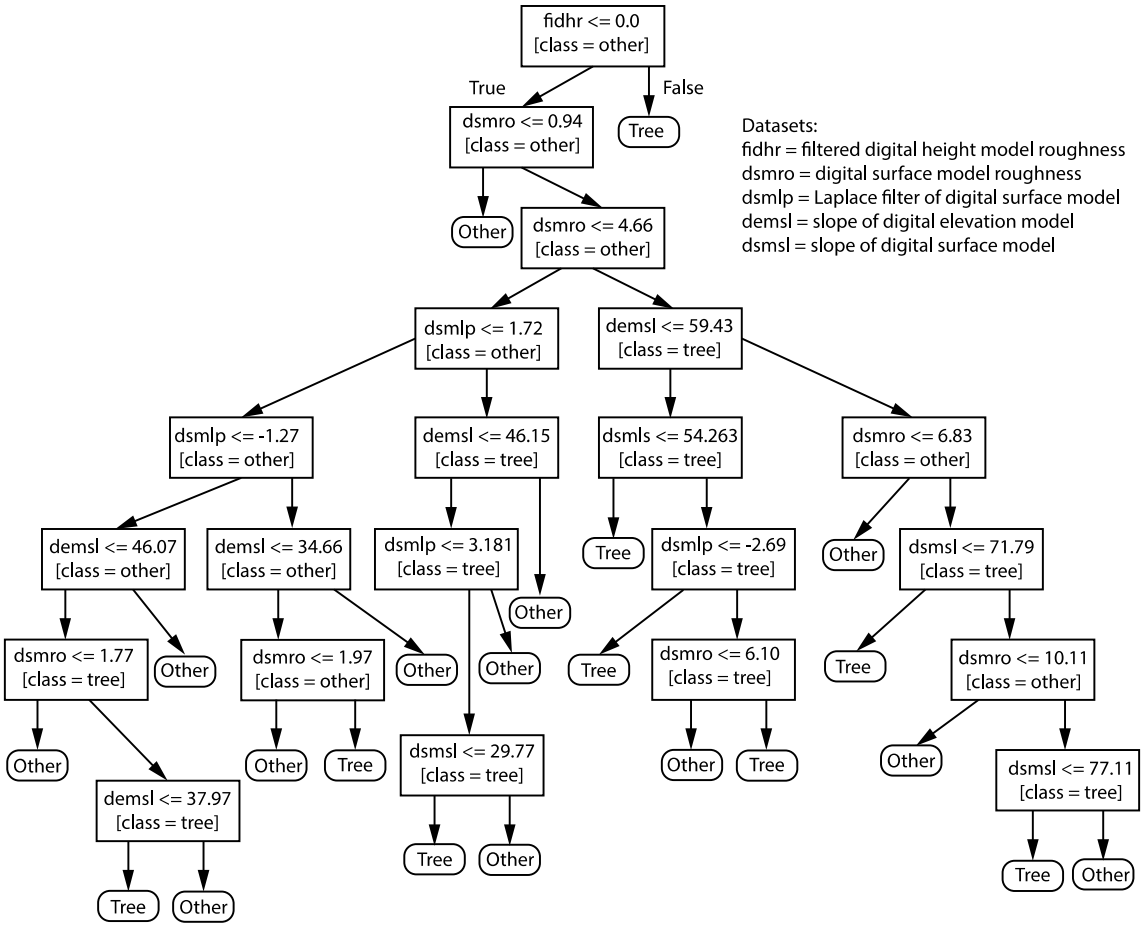

Fig. 4 Final decision tree representing the general model for canopy detection using LiDAR-derived raster grids

(DSMLP), slope of digital elevation model (DEMSL), and slope of digital surface model (DSMSL). The final solution generated by the decision tree algorithm constitutes a series of if–then statements in which threshold values are agnostic to location or dataset-specific characteristics (Fig. 4).

The constructed decision tree from these five raster grids yielded an average general model overall accuracy of 96.3% and Cohen's κ of 0.901 based on testing data from the seven watersheds used to train the model (Table 3). The average general model accuracy for the three testing watersheds is slightly higher at 98.5% overall accuracy and 0.958 Cohen's κ. Most watersheds show, at best, marginal improvement when classification is performed using a specific model rather than the general model (Table 3), thus demonstrating the generalization capability of the general model.

Overall model performance is excellent both on datasets from which subsets were used to train the model and on entirely new datasets, though the study areas that comprise the new testing data are dissimilar in terms of physiography, land use and tree characteristics. Contrasting the final classified dataset with three-dimensional transects and aerial photography demonstrates the classification results (Fig. 5). For instance, the transect from New Orleans illustrates a flat topography at a predominantly urban environment and a LiDAR dataset with nominal sampling intensity of 0.7 m. In contrast, the transect from Missouri illustrates a sloping topography at a predominantly agricultural environment and a LiDAR dataset with nominal sampling intensity of 1.33 m. Despite the differences in data and landscape characteristics, visual inspection of these cross sections reveals a good agreement of the classified canopy with aerial imagery. Graphical results are shown in Fig. 6. It is notable that of the two testing watersheds, New Orleans shows the greatest density of false positives as a result of the high number of building edges and powerlines that filtering of the DHM did not remove. Additionally, tree canopy was properly identified, despite the different land use features (i.e., urban landscape, agricultural fields, and coastal environment) and local height information.

Table 3 Accuracy metrics for the general model, specific models and singular model

HUC12	Location	Cohen's κ (general/specific/singular)	Overall accuracy (general/specific/singular)
Watersheds used to train and test the general model			
010500021301	Penobscot, ME	0.94/0.98/0.99	0.97/0.99/0.98
030902040303	Naples, FL	0.86/0.97/0.80	0.95/0.99/0.93
070801050901	Ames, IA	0.98/0.95/0.96	1.00/0.99/0.99
080102040304	Trenton, TN	0.90/0.96/0.90	0.96/0.98/0.96
130202090102	Datil, NM	0.82/0.89/0.63	0.95/0.97/0.87
140801040103	Telluride, CO	0.87/0.88/0.60	0.94/0.95/0.83
180500020905	San Francisco, CA	0.95/0.97/0.85	0.97/0.99/0.92
Average		0.90/0.94/0.82	0.96/0.98/0.93
Watersheds used for testing only			
080902030201	New Orleans, LA	0.95/0.96/0.93	0.98/0.98/0.97
100301011309	Helena, MT	0.93/0.94/0.87	0.98/0.98/0.96
102901110304	Freeburg, MO	0.99/1.00/0.78	1.00/1.00/0.90
Average		0.96/0.96/0.86	0.98/0.99/0.94

The general model was trained using data from seven watersheds. Specific models were trained using only data from the corresponding watershed and allowed access to intensity data. The singular model was trained using only data from HUC 010500021301

Discussion

LiDAR attribute selection

The contribution of an input to a decision tree's classification quality can be quantified by calculating feature importance score, which roughly describes how much the impurity of nodes in the tree is reduced when the node is split using a decision based on a given input. Normalizing feature importance allows the relative contributions of each input to be compared. The fDHM roughness raster grid is by far the most significant input to the model, with a normalized feature importance of 0.9468. The DSM roughness, Laplace filter of DSM, slope of DEM and slope of DSM raster grids have a normalized feature importance of 0.0421, 0.0047, 0.0046 and 0.0018, respectively. This suggests that nearly all of the model's classification power is derived from the fDHM roughness and DSM roughness, while the Laplace filter of DSM, slope of DEM and slope of DSM provide only minor contributions. It is interesting to note the fDHM itself does not appear in the final decision tree, suggesting that though the DHM (or a filtered DHM) is often used as a proxy for canopy cover (Dong et al. 2020; Dupuy et al. 2012; Oh et al. 2022; Ponce et al. 2022), the DHM signal itself is far less useful for canopy detection than raster grids derived from elevation products, particularly the fDHM roughness.

Canopy classification model performance

The general model yielded high accuracy results in all selected watersheds. These results are comparable with most of the misclassified false positive pixels (i.e., classification of non-tree pixels as tree pixels) occurring at building edges and powerlines, which split LiDAR pulses in multiple backscatter returns, resulting in a canopy-like signal. This is a common challenge when classifying trees in urban environments given the physical complexity resulting from constructed features (Wang et al. 2019; You et al. 2021).

Although the DHM filtering removes many of these false signals, it does not remove all of them. The use of more sophisticated filtering algorithms on the DHM raster grid or a post-classification removal of isolated false positive pixels may enhance classification quality in urban areas. We elected not to include this step to minimize the complexity in preparing input datasets by a nonexpert user. Filtering also removes some true canopy signals, resulting in false negatives (i.e., classification of tree pixels as non-tree pixels). This is a smaller contributor to model misclassification than false positives. Model misclassification is summarized as a confusion matrix in Table 4.

Despite the reduced number of used LiDAR attributes and the misclassification reported in Table 4, the proposed general model performance is in the high end of the reported site-specific LiDAR-based models/methods ranging between 69 and 99% (Guan et al. 2015; Hanssen et al. 2021; Rahman and Rashed 2015).

Model generalization

Two of the three LiDAR datasets for the testing watersheds have a point sampling density that is significantly lower than the point sampling density in the training watersheds.

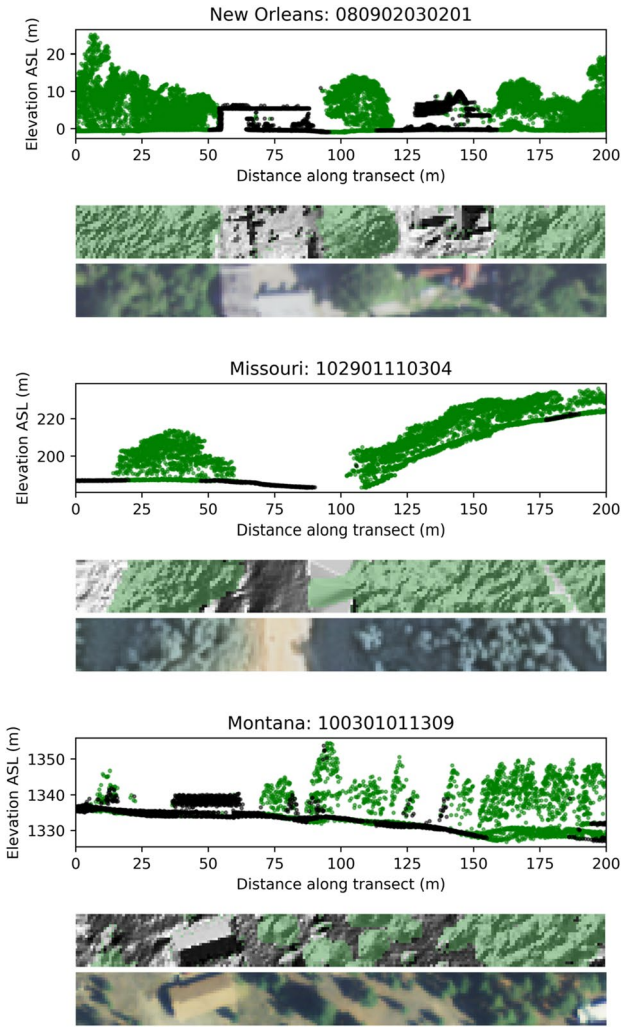

Fig. 5 Transects illustrating classification results using the general model. Transects were 200-m long by 20-m wide

Accuracy results suggest that the canopy signal in the selected sites is consistent regardless of the LiDAR sampling density.

Additionally, the negligible improvement when using a watershed-specific classification model that also includes intensity and return number raster grids shows that the performance of the general model is comparable to the custom models (Table 3), despite that the general model is more restricted in the diversity of the used raster grid datasets. It is recognized that not every component of the canopy signal is universal. For example, in a study designed to differentiate tree species, findings have documented that an accurate species classification requires a combination of a complex set of LiDAR information and that the use of height alone cannot describe the unique biomass of all tree species (Gao and Zhang 2021). However, findings in this study suggest that the canopy cover signal in the different study sites from LiDAR-derived datasets has at least some components that are common and widespread (Table 3). Furthermore, when reviewing the singular model accuracy, the classification results are acceptable for some watersheds, indicating that the canopy signatures are largely similar in those watersheds (Table 3). This implies that the most useful signals for canopy classification in some datasets are confounding in others. It seems that the general model was constructed with datasets capturing the general canopy properties while ignoring signals useful for canopy classification in specific sites. Conversely, the performance of the singular model was significantly smaller at Telluride, CO and Datil, NM watersheds, denoting that some specific training watershed characteristic was included but not found in these two watersheds.

Pretrained models, such as the one presented here, are useful for investigators seeking to efficiently quantify land cover. Canopy coverage quantification, as an example, is critical to the monitoring of forest changes or riparian buffer health. By generalizing canopy detection, a pretrained model lowers the level of user technical ability and data quality needed to produce accurate canopy cover models. Generalization, however, imposes limitations on the types of data that can be used in the model. Many previously published site-specific models that make use of LiDAR datasets use return intensity values as key input (MacFaden et al. 2012; Morsy et al. 2017; Sasaki et al. 2012). In these studies, a significant portion of land cover variance is explained by the inclusion of backscatter intensity raster grids and as a result some of these models can accurately differentiate multiple land cover classes or even tree species. However, it is important to note that LiDAR intensities are not standardized and vary depending on the type of material that is struck by a pulse in addition to variations in sensor and flight characteristics, such as original pulse signal intensity, receiver sensitivity, flight altitude and other mission-specific characteristics as well. Therefore, return intensities and number cannot be included in a pretrained general model as there is no general relation between landcover and signal intensity. LiDAR pulses with larger footprints will result in more "pulse splitting", in which a multi-level feature such as a tree, creates multiple returns since the different levels reflect different parts of the pulse footprint. By restricting model input, the generalizability of the model is increased at the expense of potential model granularity. However, comparison of the general model to the generated site-specific models using intensity and number of returns indicates that the inclusion of these datasets provide a modest advantage for canopy detection. Still, developments in LiDAR data standards, such as intensity normalization, may allow the use of additional LiDAR data fields in a general model and thus allow finer land cover classification, broadening the usefulness of general models.

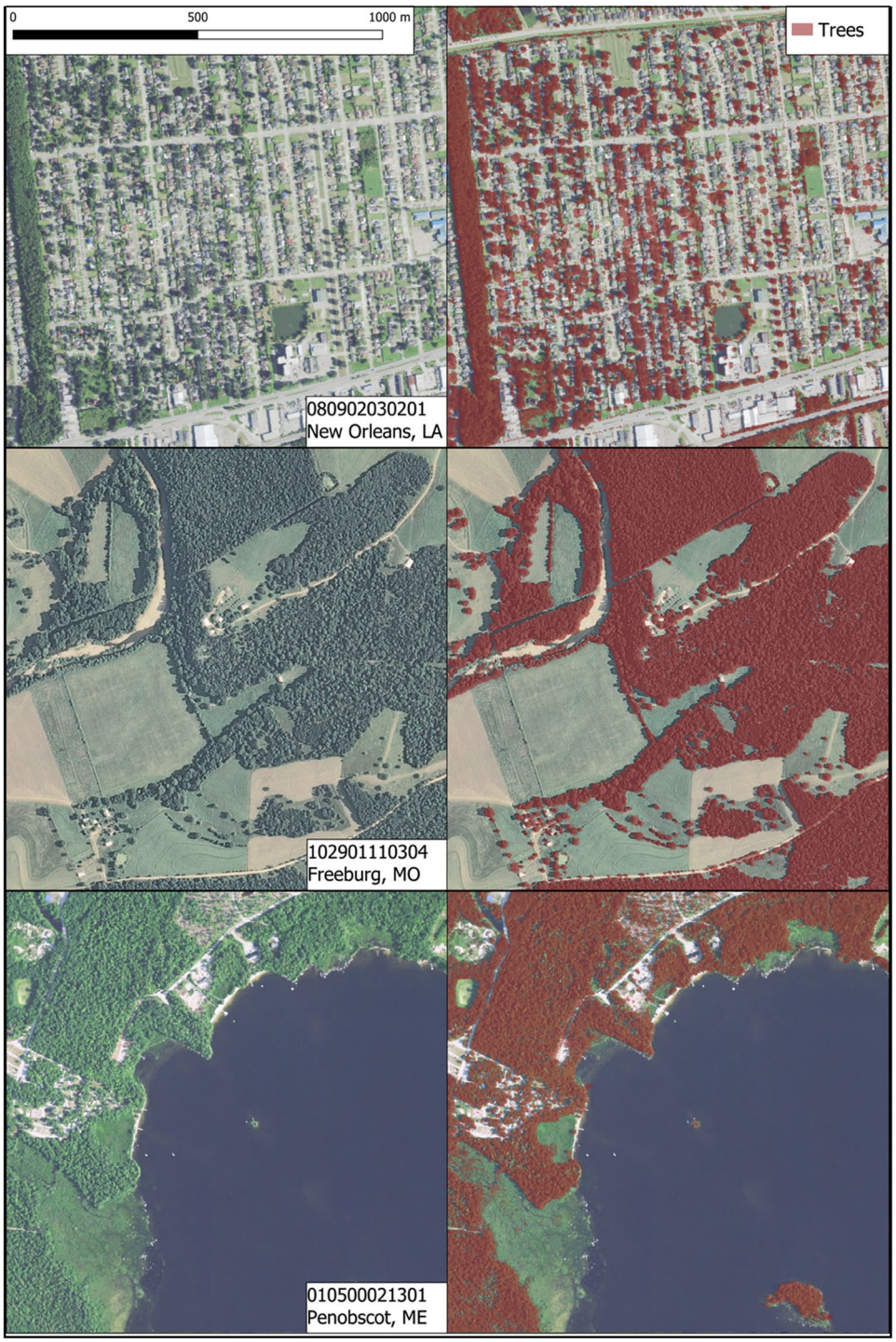

Fig. 6 Sample classifications for two of the testing-only study areas (080902030201, 102901110304) and one training study area (010500021301). Note that the imagery for the New Orleans watershed is not perfectly contemporaneous with the LiDAR input, and that some trees in the study area were removed between LiDAR and image data collection

Table 4 Confusion matrix for the general model as applied to both the testing datasets of the training watersheds and the testing watersheds, providing the number of pixels (1 m^2 each) and percentage for each category

	Reference data			
	Training		Testing	
	Trees	Other	Trees	Other
Predicted data				
Trees	6,270,072 (97.7%)	322,156 (5.9%)	9,057,777 (99.5%)	129,929 (3.1%)
Other	149,893 (2.3%)	5,167,202 (94.1%)	43,050 (0.5%)	4,039,993 (96.9%)

It is also important to recognize that the physiographic diversity of the entire contiguous US cannot be entirely captured from sampling just ten watersheds. However, the results obtained for different physiographical areas suggests that canopy signals are relatively consistent regardless of site. Thus, results indicate that the model proposed in this study can identify canopy coverage in the study sites from LiDAR alone and without any user input regardless of the study area and LiDAR sampling density and characteristics.

Conclusions

Findings from this study suggested the existence of a consistent tree canopy signal that can be captured by LiDAR datasets across the contiguous US and used to create a generalizable canopy classification model that is functional regardless of the study area or LiDAR data characteristics. This outcome was informed by: (a) the inclusion of key LiDAR-derived datasets in the final solution, and (b) the testing accuracy results. In the general classification model, only five raster grids were present in the final solution, with the roughness of the filtered DHM yielding the highest normalized feature importance of 0.9. Additionally, no vendor-specific datasets (point classification, backscatter intensity, number of results, etc.) or site-specific datasets (DEM, DSM, etc.) were included in the final solution. The general model performed with high accuracy on both training/testing and testing datasets and its classification performance was comparable to that of the singular and the specific models. Furthermore, the proposed generalizable model can be used without any input parameter adjustment nor the need for input of additional training data, allowing investigators with limited knowledge of machine learning or land cover classification to perform high-quality canopy delineations. Moreover, the robust characteristic of the general classification model suggests it could be included in remote sensing software packages, providing an alternative to existing preset tools that rely on vendor-supplied classification codes or height-based heuristics. Though the general model is currently limited in its classification ability to simple binary canopy detection, future normalization and standardization of LiDAR parameters or data fusion with other sources of remotely sensed information may enable the design of general models that can perform more detailed land cover classification. Such models would be useful for allowing nonexpert investigators to efficiently perform classification tasks that have previously been time consuming and expertise dependent.

Author contributions Conceptualization: RSJ, RE, and HGM; methodology: RSJ, RE, and HGM; software: RSJ; validation: RSJ; formal analysis: RSJ, RE, and HGM; writing: RSJ, RE, and HGM; review and editing: RE, and HGM. All authors have read and agreed to the published version of the manuscript.

Funding This work was partially supported by the Natural Resources Conservation Service (agreement number NR194741XXXXC005) and the US Department of Agriculture (agreement number 58-6060-7-014).

Data availability A package including the model Python script, the associated ESRI Toolbox file, as well as supporting documentation (i.e., tutorial and sample datasets) can be shared upon request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104

de Sosa LL, Glanville HC, Marshall MR, Abood SA, Williams A, Jones DL (2018) Delineating and mapping riparian areas for ecosystem service assessment. Ecohydrology 11(2):e1928. https://doi.org/10.1002/eco.1928

Derwin JM, Thomas VA, Wynne RH, Coulston JW, Liknes GC, Bender S, Blinn CE, Brooks EB, Ruefanacht B, Benton B, Finco MV, Megown K (2020) Estimating tree canopy cover using harmonic regression coefficients derived from multitemporal Landsat data. Int J Appl Earth Obs Geoinf 86:101985. https://doi.org/10.1016/j.jag.2019.101985

Dong X, Zhang Z, Yu R, Tian Q, Zhu X (2020) Extraction of information about individual trees from high-spatial-resolution UAV-acquired images of an orchard. Remote Sens 12:1. https://doi.org/10.3390/rs12010133

Dubayah RO, Drake JB (2000) LiDAR remote sensing for forestry. J for 98(6):44–46. https://doi.org/10.1093/jof/98.6.44

Dupuy S, Laine G, Tormos T (2012) OBIA for combining LiDAR and multispectral data to characterize forested areas and land cover in a tropical region. In: Proceedings of the 4th GEOBIA, May 7–9, Rio de Janeiro, Brazil

Gao L, Zhang X (2021) Above-ground biomass estimation of plantation with complex forest stand structure using multiple features from airborne laser scanning point cloud data. Forests 12:12. https://doi.org/10.3390/f12121713

García-Quijano MJ, Jensen JR, Hodgson ME, Hadley BC, Gladden JB, Lapine LA (2008) Significance of altitude and posting density on LiDAR-derived elevation accuracy on hazardous waste sites. Photogramm Eng Remote Sens 74(9):1137–1146. https://doi.org/10.14358/PERS.74.9.1137

Ghamisi P, Benediktsson JA, Phinn S (2015) Land-cover classification using both hyperspectral and LiDAR data. Int J Image Data Fusion 6(3):189–215. https://doi.org/10.1080/19479832.2015.1055833

Giles MF (2002) Status of land cover classification accuracy assessment. Remote Sens Environ 80(1):185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Goodwin NR, Coops NC, Tooke TR, Christen A, Voogt JA (2009) Characterizing urban surface cover and structure with airborne LiDAR technology. Can J Remote Sens 35(3):297–309. https://doi.org/10.5589/m09-015

Guan H, Yu Y, Ji Z, Li J, Zhang Q (2015) Deep learning-based tree classification using mobile LiDAR data. Remote Sens Lett 6(11):864–873. https://doi.org/10.1080/2150704X.2015.1088668

Guiotte F, Pham MT, Dambreville R, Corpetti T, Lefèvre S (2020) Semantic segmentation of LiDAR points clouds: rasterization beyond digital elevation models. IEEE Remote Sens Lett 17(11):2016–2019. https://doi.org/10.1109/LGRS.2019.2958858

Hakam O, Baali A, Belhaj Ali A (2022) Modeling drought-related yield losses using new geospatial technologies and machine learning approaches: case of the Gharb plain, North-West Morocco. Model Earth Syst Environ 2:3. https://doi.org/10.1007/s40808-022-01523-2

Hanssen F, Barton DN, Venter ZS, Nowell MS, Cimburova Z (2021) Utilizing LiDAR data to map tree canopy for urban ecosystem extent and condition accounts in Oslo. Ecol Indic 130:108007. https://doi.org/10.1016/j.ecolind.2021.108007

Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst SMC 3(6):610–621. https://doi.org/10.1109/TSMC.1973.4309314

Heinzel J, Koch B (2011) Exploring full-waveform LiDAR parameters for tree species classification. Int J Appl Earth Obs Geoinf 13(1):152–160. https://doi.org/10.1016/j.jag.2010.09.010

Herzfeld UC, McDonald BW, Wallin BF, Neumann TA, Markus T, Brenner A, Field C (2014) Algorithm for detection of ground and canopy cover in micropulse photon-counting LiDAR altimeter data in preparation for the ICESat-2 mission. IEEE Trans Geosci Remote Sens 52(4):2109–2125. https://doi.org/10.1109/TGRS.2013.2258350

Holmgren J, Lindberg E (2019) Tree crown segmentation based on a tree crown density model derived from airborne laser scanning. Remote Sens Lett 10(12):1143–1152. https://doi.org/10.1080/2150704X.2019.1658237

Hopkinson C (2007) The influence of flying altitude, beam divergence, and pulse repetition frequency on laser pulse return intensity and canopy frequency distribution. Can J Remote Sens 33(4):312–324. https://doi.org/10.5589/m07-029

Jamali A (2020) Land use land cover modeling using optimized machine learning classifiers: a case study of Shiraz, Iran. Model Earth Syst Environ 7:1539–1550. https://doi.org/10.1007/s40808-020-00859-x

Kimmel R, Bruckstein AM (2003) Regularized Laplacian zero crossings as optimal edge integrators. Int J Comput vis 53(3):225–243. https://doi.org/10.1023/A:1023030907417

Kumar N, Soni K, Agarwal R (2021) A comprehensive study of different feature selection methods and machine-learning techniques for SODAR structure classification. Model Earth Syst Environ 7:209–220. https://doi.org/10.1007/s40808-020-00872-0

Kuras A, Brell M, Rizzi J, Burud I (2021) Hyperspectral and LiDAR data applied to the urban land cover machine learning and neural-network-based classification: a review. Remote Sens 13:17. https://doi.org/10.3390/rs13173393

Kussul N, Lavreniuk M, Skakun S, Shelestov A (2017) Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci Remote Sens 14(5):778–782. https://doi.org/10.1109/LGRS.2017.2681128

Landis JR, Kock GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geosci Front 7(1):3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Lei M, Manchun L, Xiaoxue M, Liang C, Peijun D, Yongxue L (2017) A review of supervised object-based land-cover image classification. ISPRS J Photogramm Remote Sens 130:277–293. https://doi.org/10.1016/j.isprsjprs.2017.06.001

Luo S, Wang C, Xi X, Zeng H, Li D, Xia S, Wang P (2015) Fusion of airborne discrete-return LiDAR and hyperspectral data for land cover classification. Remote Sens 8:03. https://doi.org/10.3390/rs8010003

MacFaden S, O'Neil-Dunne J, Royar A, Lu J, Rundle A (2012) High-resolution tree canopy mapping for New York City using LIDAR and object-based image analysis. J Appl Remote Sens 6:3567. https://doi.org/10.1117/1.JRS.6.063567

Maxwell AE, Warner TA, Fang F (2018) Implementation of machine-learning classification in remote sensing: an applied review. Int J Remote Sens 39(9):2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Michałowska M, Rapiński J (2021) A review of tree species classification based on airborne LiDAR data and applied classifiers. Remote Sens 13:3. https://doi.org/10.3390/rs13030353

Morsy S, Shaker A, El-Rabbany A (2017) Multispectral LiDAR data for land cover classification of urban areas. Sensors 17:5. https://doi.org/10.3390/s17050958

Muftah H, Rowan TSL, Butler AP (2022) Towards open-source LOD2 modelling using convolutional neural networks. Model Earth Syst Environ 8:1693–1709. https://doi.org/10.1007/s40808-021-01159-8

Oh S, Jung J, Shao G, Shao G, Gallion J, Fei S (2022) High-resolution canopy height model generation and validation using USGS 3DEP LiDAR data in Indiana, USA. Remote Sens 14:4. https://doi.org/10.3390/rs14040935

O'Neil-Dunne JPM, MacFaden SW, Royar AR, Pelletier KC (2013) An object-based system for LiDAR data fusion and feature extraction. Geocarto Int 28(3):227–242. https://doi.org/10.1080/10106049.2012.689015

Ponce JM, Aquino A, Tejada D, Al-Hadithi BM, Andújar JM (2022) A methodology for the automated delineation of Crop Tree Crowns from UAV-based aerial imagery by means of morphological image analysis. Agronomy 12:1. https://doi.org/10.3390/agronomy12010043

Rahman MT, Rashed T (2015) Urban tree damage estimation using airborne laser scanner data and geographic information systems: an example from 2007 Oklahoma ice storm. Urban for Urban Green 14(3):562–572. https://doi.org/10.1016/j.ufug.2015.05.008

Sasaki T, Imanishi J, Ioki K, Morimoto Y, Kitada K (2012) Object-based classification of land cover and tree species by integrating airborne LiDAR and high spatial resolution imagery data. Landsc Ecol Eng 8(2):157–171. https://doi.org/10.1007/s11355-011-0158-z

Wang K, Wang T, Liu X (2019) A review: individual tree species classification using integrated airborne LiDAR and optical imagery with a focus on the urban environment. Forests 10:1. https://doi.org/10.3390/f10010001

You H, Li S, Xu Y, He Z, Wang D (2021) Tree extraction from airborne laser scanning data in urban areas. Remote Sens 13:17. https://doi.org/10.3390/rs13173428

Zhang J, Lin X (2017) Advances in fusion of optical imagery and LiDAR point cloud applied to photogrammetry and remote sensing. Int J Image Data Fusion 8(1):1–31. https://doi.org/10.1080/19479832.2016.1160960

Zhang L, Zhang L, Du B (2016) Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Trans Geosci 4(2):22–40. https://doi.org/10.1109/MGRS.2016.2540798

Zhen Z, Quackenbush LJ, Zhang L (2016) Trends in automatic individual Tree Crown detection and delineation—evolution of LiDAR data. Remote Sens 8:4. https://doi.org/10.3390/rs8040333