AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA002845

Exploring the Influence of Input Feature Space on CNN-Based Geomorphic Feature Extraction From Digital Terrain Data

Aaron E. Maxwell[1], William E. Odom[2], Charles M. Shobe[1], Daniel H. Doctor[2], Michelle S. Bester[1], and Tobi Ore[1]

[1]Department of Geology and Geography, West Virginia University, Morgantown, WV, USA, [2]Florence Bascom Geoscience Center, U.S. Geological Survey, Reston, VA, USA

Key Points:

- We test the sensitivity of convolutional neural network geomorphic feature extraction approaches to the input feature space
- A three-layer composite incorporating slope and topographic position index outperformed common inputs like hillshades and slopeshades
- Results held true across four use cases including both natural and anthropogenic landforms

Correspondence to:

A. E. Maxwell,
Aaron.Maxwell@mail.wvu.edu

Citation:

Maxwell, A. E., Odom, W. E., Shobe, C. M., Doctor, D. H., Bester, M. S., & Ore, T. (2023). Exploring the influence of input feature space on CNN-based geomorphic feature extraction from digital terrain data. *Earth and Space Science*, *10*, e2023EA002845. https://doi.org/10.1029/2023EA002845

Received 17 JAN 2023
Accepted 17 APR 2023

Author Contributions:

Conceptualization: Aaron E. Maxwell, William E. Odom, Charles M. Shobe, Daniel H. Doctor
Data curation: Aaron E. Maxwell
Formal analysis: Aaron E. Maxwell, Michelle S. Bester, Tobi Ore
Funding acquisition: Aaron E. Maxwell
Investigation: Aaron E. Maxwell, William E. Odom, Michelle S. Bester, Tobi Ore
Methodology: Aaron E. Maxwell, William E. Odom, Charles M. Shobe, Daniel H. Doctor
Validation: Aaron E. Maxwell
Writing – original draft: Aaron E. Maxwell

Abstract Many studies of Earth surface processes and landscape evolution rely on having accurate and extensive data sets of surficial geologic units and landforms. Automated extraction of geomorphic features using deep learning provides an objective way to consistently map landforms over large spatial extents. However, there is no consensus on the optimal input feature space for such analyses. We explore the impact of input feature space for extracting geomorphic features from land surface parameters (LSPs) derived from digital terrain models (DTMs) using convolutional neural network (CNN)-based semantic segmentation deep learning. We compare four input feature space configurations: (a) a three-layer composite consisting of a topographic position index (TPI) calculated using a 50 m radius circular window, square root of topographic slope, and TPI calculated using an annulus with a 2 m inner radius and 10 m outer radius, (b) a single illuminating position hillshade, (c) a multidirectional hillshade, and (d) a slopeshade. We test each feature space input using three deep learning algorithms and four use cases: two with natural features and two with anthropogenic features. The three-layer composite generally provided lower overall losses for the training samples, a higher F1-score for the withheld validation data, and better performance for generalizing to withheld testing data from a new geographic extent. Results suggest that CNN-based deep learning for mapping geomorphic features or landforms from LSPs is sensitive to input feature space. Given the large number of LSPs that can be derived from DTM data and the variety of geomorphic mapping tasks that can be undertaken using CNN-based methods, we argue that additional research focused on feature space considerations is needed and suggest future research directions. We also suggest that the three-layer composite implemented here can offer better performance in comparison to using hillshades or other common terrain visualization surfaces and is, thus, worth considering for different mapping and feature extraction tasks.

Plain Language Summary Characteristics of the land surface, such as steepness, relative slope position (e.g., ridge vs. valley), and roughness, can be digitally represented using a variety of methods. These digital representations, along with human-annotated labels, can be provided to artificial intelligence algorithms to generate maps of landforms, such as those associated with river systems, glaciers, or human-induced changes to the landscape. Given the large number of terrain derivatives that can be generated from digital elevation data, it is unclear how the chosen inputs impact the utility of the resulting maps. This study suggests that artificial intelligence mapping algorithms are sensitive to representations provided to them since different inputs resulted in varying map accuracies. We suggest a combination of features, which collectively describe relative slope position, steepness, and local terrain texture, as a means to represent the landscape and to serve as input to artificial intelligence algorithms. Our results can make it easier to train artificial intelligence algorithms to consistently and objectively find surficial features of interest across large swaths of Earth's surface.

1. Introduction

Maps and digital geospatial data that differentiate and characterize landforms, surficial geologic units, and anthropogenic landscape alterations are central to investigating a variety of Earth surface processes and landscape evolution research questions (Baker, 1986; Bishop et al., 2012; Dramis et al., 2011; Jacek, 1997; Minár & Evans, 2008; Pavlopoulos et al., 2009; Tucker & Hancock, 2010; Verstappen, 2011). Such data also have applications in other fields such as soil science (Lagacherie, 2008; Y. Ma et al., 2019; Minasny & McBratney, 2016) and archeology (Albrecht et al., 2019; Fernandez-Diaz et al., 2014; Guyot et al., 2021). They are also used in applied workflows associated with geohazard risk assessment, site selection, civil engineering, and environmental conservation (Bishop et al., 2012; Brunsden et al., 1975; Jacek, 1997; Pavlopoulos et al., 2009). Surficial

Writing – review & editing: Aaron E. Maxwell, William E. Odom, Charles M. Shobe, Daniel H. Doctor, Michelle S. Bester, Tobi Ore

mapping tasks have a rich history that has been impacted by changes in our understanding of Earth surface processes. For example, physiographic and surficial geologic mapping developed prior to plate tectonic theory (Bishop et al., 2012; Verstappen, 2011).

Surficial mapping tasks have long proven difficult for several reasons. Important mapping units may vary with scale and purpose, and the appropriate level of detail and/or mapping scale can be unclear (Dramis et al., 2011; Evans, 2012; McMaster & Sheppard, 2004; Minár & Evans, 2008; Quattrochi & Goodchild, 1997; Sheppard & McMaster, 2004; Smith et al., 2011). Even defining what constitutes a landform unit can be difficult (Evans, 2012), and landform features often exist as a hierarchy and can have gradational or fuzzy boundaries (Dramis et al., 2011; Gustavsson et al., 2006), making it difficult to differentiate discrete, non-overlapping objects and assess mapping products in an accurate and not overly harsh manner (Burrough, 2020; Foody, 2008). There may also be a need to map processes and features that change with time (Flageollet, 1996; Tucker & Hancock, 2010).

Despite these difficulties, there is an increasing abundance of remotely sensed and digital terrain data, such as those made available by light detection and ranging (lidar), that can serve as input to both manual, semi-automated, and automated methods. One example data source is the publicly available products provided by the U.S. Geological Survey (USGS) 3D Elevation Program (3DEP) (Sugarbaker et al., 2014). Simultaneously, advances in machine learning (ML) and deep learning (DL) methods are improving our ability to extract landform and geomorphic information from these data sources (Maxwell & Shobe, 2022; Sofia et al., 2016; Tarolli, 2014; L. Zhang et al., 2016; Zhu et al., 2017). Thus, there is a need to investigate such methods for extracting actionable geomorphic information and generating map data to support research and applications that require landform or surficial geology data.

Convolutional neural network (CNN)-based DL methods in particular have yielded advances in the extraction of information from a wide variety of data sources that can be represented as structured, multidimensional arrays of measurements. CNNs can improve upon traditional feature recognition and image classification methods, such as shallow ML, by incorporating characterizations of patterns within the domain(s) of interest (e.g., two- or three-dimensional space, time, depth below ground or height above ground, or spectral reflectance patterns across a defined range of electromagnetic wavelengths) (L. Ma et al., 2019; L. Zhang et al., 2016; Zhu et al., 2017). CNN-based methods have enabled progress in computer vision (Hassaballah & Awad, 2020; Voulodimos et al., 2018), autonomous vehicle technologies (Fayyad et al., 2020; Miglani & Kumar, 2019), medical imaging and interpretation (Greenspan et al., 2016; Mainak et al., 2019; Sahiner et al., 2019), geophysical subsurface analysis (Yu & Ma, 2021), and processing of three-dimensional point clouds (Lu & Shi, 2020; Pierdicca et al., 2020; J. Zhang et al., 2019), such as those generated using lidar.

Within the fields of geospatial modeling and remote sensing specifically, many studies have explored CNNs for extracting information from true color and multispectral imagery (L. Ma et al., 2019; L. Zhang et al., 2016; Zhu et al., 2017). However, research on using CNNs to support geomorphic, landform, or surficial geologic mapping through the recognition of natural or anthropogenic landforms from raster-based land surface parameters (LSPs) (i.e., digital terrain variables) derived from digital terrain models (DTMs) is more limited.

This study explores the impact of the feature space, or the LSPs provided to the algorithm, on model performance. Due to the large number of LSPs that can be calculated from a DTM, selecting a subset of features to provide as input to the DL algorithm can be daunting. This contrasts with working with true color or multispectral images, where inputs are generally the original image bands or a finite set of derivatives from these bands (e.g., principal components or band ratios) commonly used in the discipline. Given the infinite number of LSP combinations that can be calculated and assessed, comparisons made in this study are informed by the authors' experiences manually interpreting and mapping surficial geology and landform features from digital terrain data. We explore a three-layer composite that we have found to be especially effective for manual interpretation of landscape characteristics. It consists of the square root of topographic slope and the topographic position index (TPI) calculated using window sizes selected to capture patterns at the local- and hillslope-scales. This composite is compared to LSPs commonly used to visualize DTMs and manually interpret landscape characteristics and landforms: hillshades (HS) and a slopeshade. We are specifically interested in whether this composite, when used as the input feature space to train CNN-based semantic segmentation algorithms, allows for improved automated mapping in comparison to the more commonly used terrain visualization surfaces. The four use cases that we explore specifically rely on binary semantic segmentation, which consists of labeling each pixel as representing the class or

feature of interest or the background (Hoeser et al., 2020; Hoeser & Kuenzer, 2020; L. Ma et al., 2019; Maxwell et al., 2021a, 2021b).

2. Background

2.1. Convolutional Neural Networks for Modeling Spatial Context

In an attempt to complement or replace manual landform or surficial geologic mapping methods, such as manual digitizing via on-screen interpretation of digital terrain data, a variety of techniques have been investigated for extracting geomorphic information and mapping landform features from remotely sensed data and LSPs. Methods rely on supervised classification, clustering or unsupervised classification, or expert-defined workflows that attempt to procedurally detect features from digital terrain data based on unique landscape characteristics that differentiate them from other features on the landscape (Pavlopoulos et al., 2009; Smith et al., 2011; Wilson & Gallant, 2000). One example of an expert-defined approach is Yang et al. (2019), which explored the extraction of bank gullies in the Loess Plateau of China (Yang et al., 2019). Here we specifically focus on the application of supervised learning techniques in which labeled data are used to train a learning algorithm to recognize or classify features from remotely sensed data (Lillesand et al., 2015).

The integration of spatial context information into the supervised learning process has been explored in remote sensing research for several decades (e.g., Warner, 2011). Traditionally, pixel-based classification was performed using only the image band digital number (DN) values as predictor variables within a supervised classification workflow (i.e., learning from labeled data) and using parametric or, later, ML algorithms. To characterize local, two-dimensional spatial patterns in pixel DN values, hand-crafted weights associated with moving windows or kernels are often applied (M. Li et al., 2014; Warner, 2011). To allow for greater flexibility in defining local patterns, measures of texture derived from the gray-level co-occurrence matrix (GLCM) after Haralick et al. (1973) have been used. Such methods allow for the direction and distance between pixels used in the creation of the co-occurrence matrix and associated metrics to be defined by the user, resulting in more control over how spatial patterns are represented (Hall-Beyer, 2017; Haralick et al., 1973; Haralick & Shanmugam, 1974; Warner, 2011). Expanding upon these earlier techniques, geographic object-based image analysis (GEOBIA) methods first require that the image be segmented into polygons or areas based on similarity between adjacent pixel values. These objects, as opposed to each individual pixel, then become the unit for subsequent analysis and classification (Blaschke, 2010; Blaschke et al., 2014; G. Chen et al., 2018; Hay & Castilla, 2008). Classification of these objects can be conducted using rulesets or ML, both of which require generating object-specific measures of spectral band central tendency, variability, and/or texture (Blaschke, 2010; Maxwell et al., 2018). Thus, all of these existing techniques require the user to select and calculate measures for inclusion in the feature space. Given the infinite number of measures that can be calculated, and the common lack of a priori information as to which features will be useful for the task of interest, developing an optimal feature space for a specific task can be challenging.

When applying CNN-based algorithms and architectures the user does not need to engineer or apply hand-crafted kernels or convolutional operations to create predictor variables as input to the modeling process. Instead, the CNN learns weights associated with kernels (i.e., moving window or convolutional filters) via a supervised learning process, which are then applied to the image data to generate feature maps (i.e., spatial abstractions of the input data). By reducing the size of the resulting array by aggregating pixel values (e.g., applying max pooling operations) and continuing to learn kernel weights at these reduced resolutions, patterns over multiple scales can be characterized (Hoeser et al., 2020; Hoeser & Kuenzer, 2020; L. Ma et al., 2019; L. Zhang et al., 2016; Zhu et al., 2017). Such methods allow for the CNN algorithm to learn abstractions of spatial patterns of values for differentiating features or separating classes, as opposed to the analyst providing such representations via a feature engineering process in which it is often necessary to generate and assess a large number of derivatives in order to select a subset of features that adequately characterize patterns in the data.

We argue that techniques that can characterize spatial context or textural patterns in data are valuable when the features of interest are characterized or differentiated by such patterns as opposed to individual pixel DN values. Methods that can learn spatial context information are especially valuable when a large number of hand-crafted derivatives can be generated and the optimal set of features is difficult to determine. Both criteria are true for extracting geomorphic features from gridded LSPs. Visual inspection of digital terrain surfaces, such as

hillshades, suggests that the information content of the data is associated with local textures and patterns. Further, a large number of LSPs that characterize different aspects of the terrain surface (e.g., steepness, curvature, local topographic position, roughness, and incision) can be manually generated from raster-based DTMs using varying moving window sizes, shapes, and weightings of pixel values within the window, resulting in an infinite set of derivatives that can be considered (Franklin, 2020; Maxwell & Shobe, 2022). The architecture of CNNs allows for applying kernels with trainable weights to create spatial data abstractions. Being that these architectures incorporate the learning of a large number of kernels at varying spatial scales by integrating a series of convolution and pooling or downsampling operations, we argue that such techniques build upon and overcome limitations of prior pixel-, object-, and knowledge-based methods, such as the difficulty of characterizing textural information in pixel-based classification or segmenting the landscape into meaningful units in object-based classification. Thus, there is practical utility in implementing CNN-based methods to potentially reduce the number of inputs that must be generated and assessed and allow for learning local patterns that are predictive for the specific use case.

2.2. CNNs for Geomorphic Mapping and Feature Extraction

The characterization of spatial context information has been investigated for the mapping and extraction of landforms, geomorphic and archeological features, and anthropogenic terrain alterations (Maxwell & Shobe, 2022; Sofia et al., 2014, 2016; Tarolli, 2014; Verhagen & Drăguţ, 2012). Prior to the availability of CNN-based DL methods, GEOBIA methods were applied to imagery (e.g., d'Oleire-Oltmanns et al., 2013; Stumpf & Kerle, 2011), DTM-derived LSPs (e.g., Drăguţ & Blaschke, 2006; Feizizadeh et al., 2021; Janowski et al., 2022; K. Saha et al., 2011; Verhagen & Drăguţ, 2012), or a combination of spectral and terrain data (e.g., Diaz-Varela et al., 2014; Dornik et al., 2018; Kazemi Garajeh et al., 2022). More generally, authors have investigated how best to characterize spatial patterns in digital terrain data at varying scales using different methods; for example, Jordan and Schott (2005) applied wavelet analysis via Fourier transforms to DTMs to study spatial patterns of geologic lineaments. Behrens et al. (2018) proposed a Gaussian pyramid method that allows for the generalization of DTMs at varying scales using downscaling and subsequent upscaling.

More recently, DL methods have been explored for slope failure mapping or susceptibility modelling (e.g., Gholami et al., 2021; Huang et al., 2020; S. Li et al., 2020; Prakash et al., 2020; S. Saha et al., 2021; Schönfeldt et al., 2022; Thi Ngo et al., 2021), landform and geomorphic feature extraction (e.g., Bickel et al., 2021; Du et al., 2019; S. Li et al., 2020; Moseley et al., 2021; Robson et al., 2020; van der Meij et al., 2022; Xie et al., 2020; Xu et al., 2021; W. Zhang et al., 2018, 2020), mapping of anthropogenic landscape alterations or archaeological features (e.g., Guyot et al., 2018; Maxwell et al., 2020; Suh et al., 2021; Trier et al., 2015, 2019), and digital soil unit mapping (e.g., Behrens et al., 2018; Padarian et al., 2019; Wadoux, 2019). Generally, these studies highlight the value of modelling spatial patterns in LSPs and/or other spatial data to improve the use of such data for geomorphic mapping and feature extraction.

However, few prior studies have explicitly explored the impact of feature space on the resulting model performance. One exception is Suh et al. (2021), who mapped charcoal hearths in New England, USA using LSPs and UNet-based semantic segmentation DL. They explored the following input feature spaces: slope, VAT (Visualization for Archaeological Topography tool), slope and an HS, and all available terrain derivatives. They documented variability in model performance and differences in the best set of features depending on the landscape being predicted (Raab et al., 2022; Suh et al., 2021). Given this observed variability, we argue that there is a need to further investigate the impact of feature space and selected LSPs on resulting model performance for mapping of geomorphic features. This work expands upon the work of Suh et al. (2021) by exploring different LSPs, with a specific focus on a three-layer combination that we have found to be especially useful for undertaking manual mapping of landform and surficial geology features, generalizing findings to multiple mapping problems (both natural and anthropogenic), and assessing the impact of input feature space on multiple DL semantic segmentation architectures. Given that many LSPs (e.g., slope, surface curvatures, local topographic position, and topographic roughness) are calculated using convolutional operations and local moving windows, it is important to understand whether the original feature space is of great importance or if the CNN can model or generate predictive spatial abstractions using a small set of predictor variables. We investigate this question by exploring the extraction of four geomorphic features: valley fill faces resulting from mountaintop removal coal mining, agricultural terraces used for erosion control, surficial alluvial deposits, and thick glacial till. In order to generalize the results, we also implement three different semantic segmentation deep learning architectures: DeepLabv3+, UNet, and UNet++.

Table 1
Summary of LSPs Used in Prior GEOBIA or CNN-Based DL Studies Focused on Landform, Surficial Geology, or Archeological Mapping

Study	Method	Task	Variables
K. Saha et al. (2011)	GEOBIA	Drumlins	Elevation, slope, aspect
Verhagen & Drăguţ (2012)	GEOBIA	Archeological	Elevation, slope, and curvature
d'Oleire-Oltmanns et al. (2013)	GEOBIA	Drumlins	Normalized relative elevation
Pedersen (2016)	GEOBIA	Glaciovolcanic landforms	Slope, profile curvature
Trier et al. (2019)	DL	Archeological	Topographic position index
Maxwell et al. (2020)	DL	Coal mining valley fill faces	Slopeshade
Guyot et al. (2021)	DL	Archeological	Multiscale topographic position index, slope, topographic openness
Na et al. (2021)	GEOBIA	General landforms	Slope, slope gradient, terrain relief, surface roughness, elevation, elevation coefficient variation, hillshade, accumulative curvature
Suh et al. (2021)	DL	Relict charcoal hearths	Slope, hillshade, Visualization for Archeological Topography (VAT)
van der Meij et al. (2022)	DL	General landforms	Smoothed DEM, relief map
Salas & Argialas (2022)	DL	Seafloor landforms	Elevation, slope, topographic position index
Janowski et al. (2022)	GEOBIA	Glacial landforms	Slope, aspect, surface curvature
Kazemi Garajeh et al. (2022)	GEOBIA	Desert landforms	Elevation, slope, aspect, hillshade

Different LSPs have been explored as input to DL- and GEOBIA-based workflows for landform, surficial geologic, or archeological mapping tasks, as summarized in Table 1. This table specifically highlights studies that used only LSPs as opposed to combining them with spectral data. As Table 1 demonstrates, a wide variety of LSPs have been used, and there is not currently a standard set of LSPs that are considered. Some commonly used features include slope or other representations of topographic steepness, curvature, measure of local topographic position (e.g., the TPI), hillshades, and measures of topographic roughness or local variabiliy. This highlights the need to further investigate the impact of feature space on model performance. Given the large number of features that can be calculated and the variable settings available, a full comparison of a large number of LSP combinations is not possible, especially considering the computational demand of training DL algorithms. Given these limitations, this study contributes to a better understanding of LSP feature space impacts on model performance by comparing an LSP composite that we have found to be useful for manual interpretation to some commonly used terrain visualization LSPs and by exploring multiple algorithms and mapping problems.

3. Methods

3.1. Use Cases and Input Data

In order to generalize the results of the study and offer a more robust assessment of the impact of feature space on model performance, multiple studies were explored using binary semantic segmentation. The landforms of interest had varying characteristics, levels of difficulty in regards to differentiating them from other features within the landscape, and abundance of reference data to train and assess models. The study area extents are shown in Figure 1, including the geographic stratification of training, validation, and testing samples, while Figure 2 provides examples of the features of interest. Data from within the training extents were used to train models while data from the validation areas were used to assess models at the end of each training epoch. Data from within testing areas were withheld to assess and compare final models.

Valley fills occur predominantly in southern West Virginia, eastern Kentucky, and southwestern Virginia in the eastern United States and result from mountaintop removal coal mining, a prolific anthropogenic landscape alteration in this region. This coal mining practice consists of removing mountaintops to expose coal seams. After the coal seams are extracted, it is generally not possible to reclaim the landscape's original topography

10.1029/2023EA002845

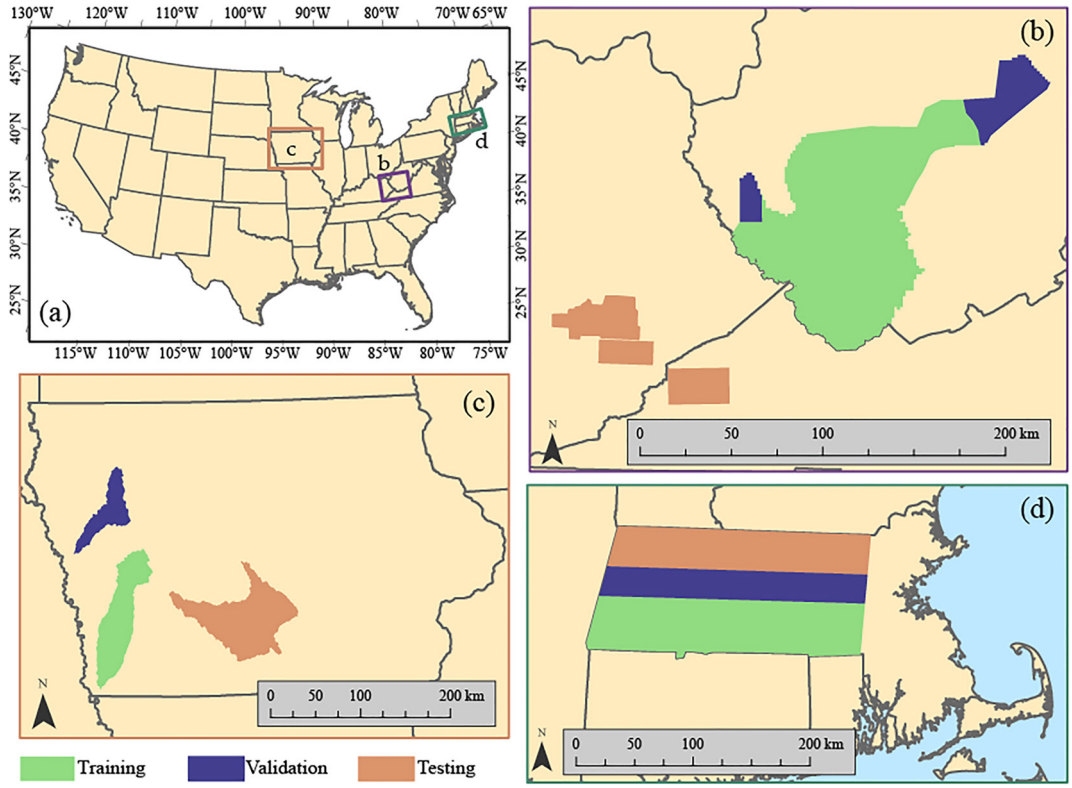

Figure 1. Study areas associated with each use case explored with training, validation, and testing areas differentiated. (a) Extent of study areas in United States. (b) Valley fill faces study area in West Virginia, Kentucky, and Virginia. (c) Agricultural terraces study area in Iowa. (d) Alluvium and thick glacial till study areas in Massachusetts.

while also maintaining stable slopes. As a result, displaced overburden material is placed in adjacent valleys. The faces of these filled valleys have a characteristic triangular shape, a steep, terraced surface, and drainage ditches (Fritz et al., 2010). These data were generated by the researchers for use in a prior study associated with applying mask region-based CNN DL (mask R-CNN), an instance segmentation method, for geomorphic feature extraction (Maxwell et al., 2020). Valley fill faces in the training, validation, and testing extents were manually mapped by the researchers based on interpretation of LSPs and other ancillary data (e.g., aerial imagery and mine permit boundaries). LSPs were generated from a DTM created from ground classified lidar point clouds. The West Virginia data were obtained from the West Virginia GIS Technical Center (WVGISTC) while the data from Kentucky and Virginia, which served as testing data in the study, were obtained from the USGS 3DEP (Sugarbaker et al., 2014). DTMs were created at a 2 m spatial resolution using the LAS Dataset to Raster Tool (LAS Dataset To Raster (Conversion)—ArcGIS Pro | Documentation, 2023) in ArcGIS Pro (2D 3D & 4D GIS Mapping Software | ArcGIS Pro, 2023). The average ground return elevation was calculated within each cell, and linear interpolation was used to fill data gaps.

Agricultural terrace data were provided by the Iowa Best Management Practices (BMP) Mapping Project at Iowa State University (Iowa BMP Mapping Project—Geographic Information Systems, 2023) as geospatial vector line data. These data were generated using manual interpretation of HSs and aerial imagery. Terraces are anthropogenic landforms designed to reduce soil loss by hindering sheet and rill erosion and discouraging gully development. Two general types of terracing practices are used in Iowa: narrow base and broadbase. Narrow base terraces have sloped surfaces in both the upslope and downslope directions and are vegetated with perennial grasses while broadbase terraces are generally wider and flatter. Types of terraces were not differentiated in the provided data set (Iowa BMP Mapping Project—Geographic Information Systems, 2023). For this study specifically, we defined training, validation, and testing partitions using hydrologic unit code (HUC) 8-digit watershed boundaries. Features mapped in the West Nishnabotna watershed were used to train models while those in the Maple watershed were used to test the model performance at the end of each training epoch. Features in the Lake

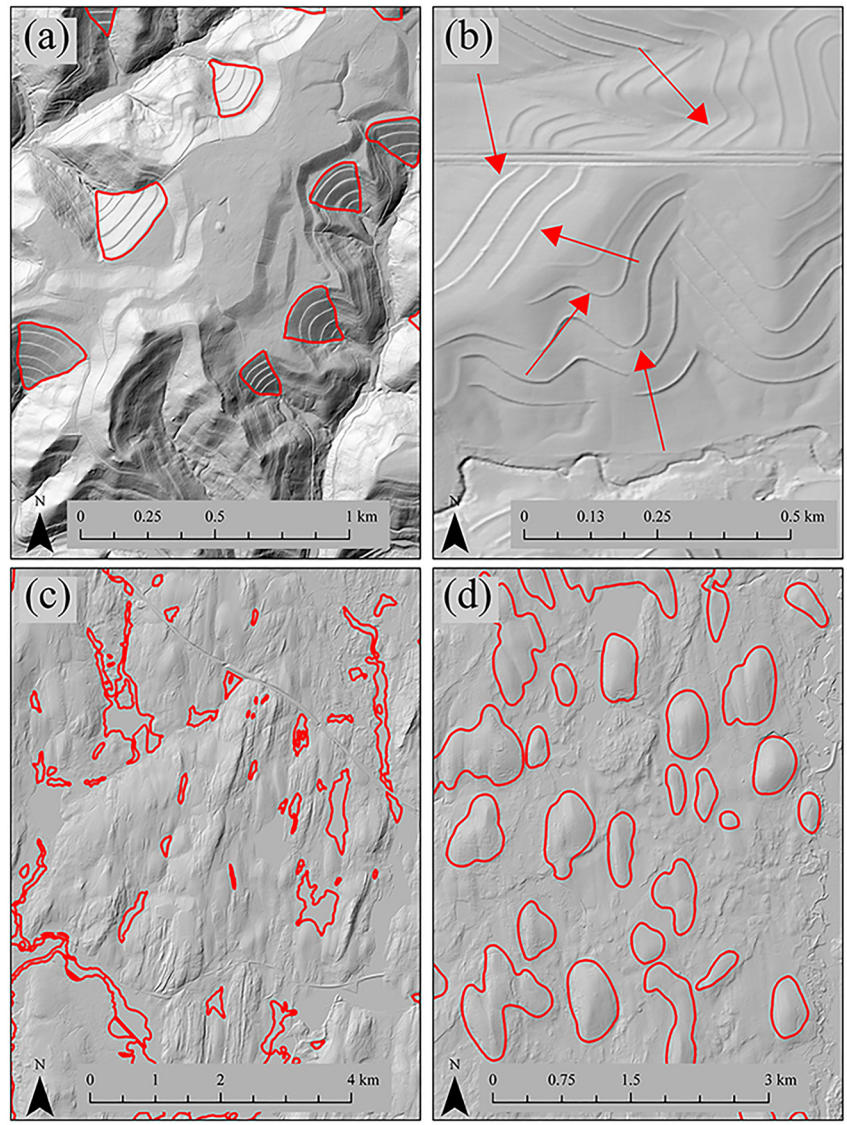

Figure 2. Examples of the topographic features investigated. (a) Valley fill faces resulting from mountaintop removal coal mining; (b) Agricultural terraces; (c) Alluvium; (d) Thick glacial till.

Red Rock watershed were withheld to test the final models. The line features were buffered using a 4 m distance to generate areal features then subsequently rasterized. The lidar-derived DTMs used during the manual interpretation process were provided by the Iowa BMP Mapping Project and resampled to a 2 m cell size using cubic convolution.

Two natural surficial geologic landform features were explored in the western portion of Massachusetts, USA: alluvial deposits and thick glacial till. These features were extracted from a complete surficial geologic mapping of the state conducted for each 7.5-min topographic quadrangle occurring within the state at a scale of 1:24,000. This mapping differentiates nonlithified earth materials and the boundaries between exposed bedrock, glacial till, glacial stratified deposits, and post-glacial deposits. Digital geospatial data were generated from the original hardcopy maps by the USGS ("Surficial Materials of Massachusetts—A 1," 2018). From the larger set of mapped surficial deposits, we chose to map post-glacial alluvial deposits and other related classes (e.g., alluvial-fan deposits, flood-plain alluvium, swamp deposits, and valley-floor fluvial deposits) grouped as a single class. Such deposits are commonly found at lower relative slope positions, such as valleys. Second, we investigated thick-till glacial deposits, defined as till deposits greater than 15 feet (4.6 m) thick and characterized by drumlin landforms. We did not use data for the eastern, coastal region of Massachusetts (Figure 1) due to the differences in landforms

Table 2
Summary of Derived LSPs and Tested Feature Spaces Explored in This Study

Feature space	Bands	Abbreviation
3 Band Stack	Band 1 = TPI (50 m Radius circular window) Band 2 = Square root of slope Band 3 = TPI (2 m Inner, 10 m Outer radius annulus window)	Stack
Hillshade		HS
Multidirectional hillshade		MHS
Slopeshade		SlpS

and physiography in comparison to the western part of the state. The DTM data for this study was provided by MassGIS and consists of a mosaic of best available data. The provided DEM was resampled to a 2 m spatial resolution using cubic convolution.

3.2. Land Surface Parameters

Topographic slope in degrees units was produced using the Slope Tool (Slope (Spatial Analyst)—ArcGIS Pro Documentation, 2023) from the Spatial Analyst Extension of ArcGIS Pro (2D 3D & 4D GIS Mapping Software | ArcGIS Pro, 2023). From this variable, we produced two surfaces: the square root of slope and a slopeshade (SlpS) (Maxwell & Shobe, 2022). In the square root of slope output, values smaller than 0 were recoded to 0, values larger than 10 were recoded to 10, and all values were then linearly rescaled from 0 to 1. The SlpS was calculated by dividing slope by 90 then subtracting the result from 1 (Equation 1). This resulted in values scaled from 0 to 1 in which low values represent steep slopes and high values represent flatter terrain. Four HSs were produced using the Hillshade Tool (Hillshade (Spatial Analyst)—ArcGIS Pro | Documentation, 2023) in ArcGIS Pro (2D 3D & 4D GIS Mapping Software | ArcGIS Pro, 2023) using illuminating positions in the north, northwest, west, and southeast. The HS generated with a northwest illuminating position was included in the final feature space, and a multidirectional hillshade (MHS) was calculated by averaging all HSs and double weighting the HS created using a northwest illuminating position (Equation 2). All HSs were then rescaled from 0 to 1 by dividing by 255. The TPI (Wilson & Gallant, 2000) was calculated using a circular moving window with a radius of 50 m and also with an annulus with a 2 m inner radius and a 10 m outer radius. The mean was calculated within the focal windows then subtracted from the center cell value to obtain an index in which larger, positive values indicate topographic high points and negative values indicate topographic low points (Equation 3). The TPI produced using a circular window captured patterns at a hillslope spatial scale while the annulus-based TPI captured more local patterns. One complexity of calculating LSPs that rely on moving window-based computation, such as the TPI, is the difficulty in determining an optimal window size, shape, and/or cell weightings. The size of the larger, circular window was selected based on common ridge-to-valley distances and in order to capture patterns at a hillslope scale. The smaller, annulus window was parameterized to capture more local topographic patterns and variability. Both configurations were guided by experimentation with varying window configurations and professional judgment. Although methods have been proposed to select window sizes (see Maxwell & Shobe, 2022 for a review), these methods have not been applied or validated in the context of DL methods. For both TPIs, values smaller than −10 were recoded to −10 while values larger than 10 were recoded to 10. The data were then rescaled from 0 to 1 by subtracting −10 then dividing by 20.

$$\mathrm{SlpS} = 1 - \frac{\text{Slope in Degrees}}{90} \tag{1}$$

$$\mathrm{MHS} = (2 * \mathrm{HS}_{\mathrm{Northwest}} + \mathrm{HS}_{\mathrm{North}} + \mathrm{HS}_{\mathrm{West}} + \mathrm{HS}_{\mathrm{Southeast}})/5 \tag{2}$$

$$\mathrm{TPI} = \mathrm{DTM} - \mathrm{DTM}_{\mathrm{Mean}} \tag{3}$$

The TPI calculated using a circular moving window, square root of slope, TPI calculated using an annulus moving window, HS with a northwest illuminating position, MHS, and SlpS were stacked into a multiband raster grid with all layers represented as floating point data with values scaled from 0 to 1. From these 6 inputs, 4 feature spaces were compared (Table 2 and Figure 3). The TPIs and square root of slope were provided as input to the DL

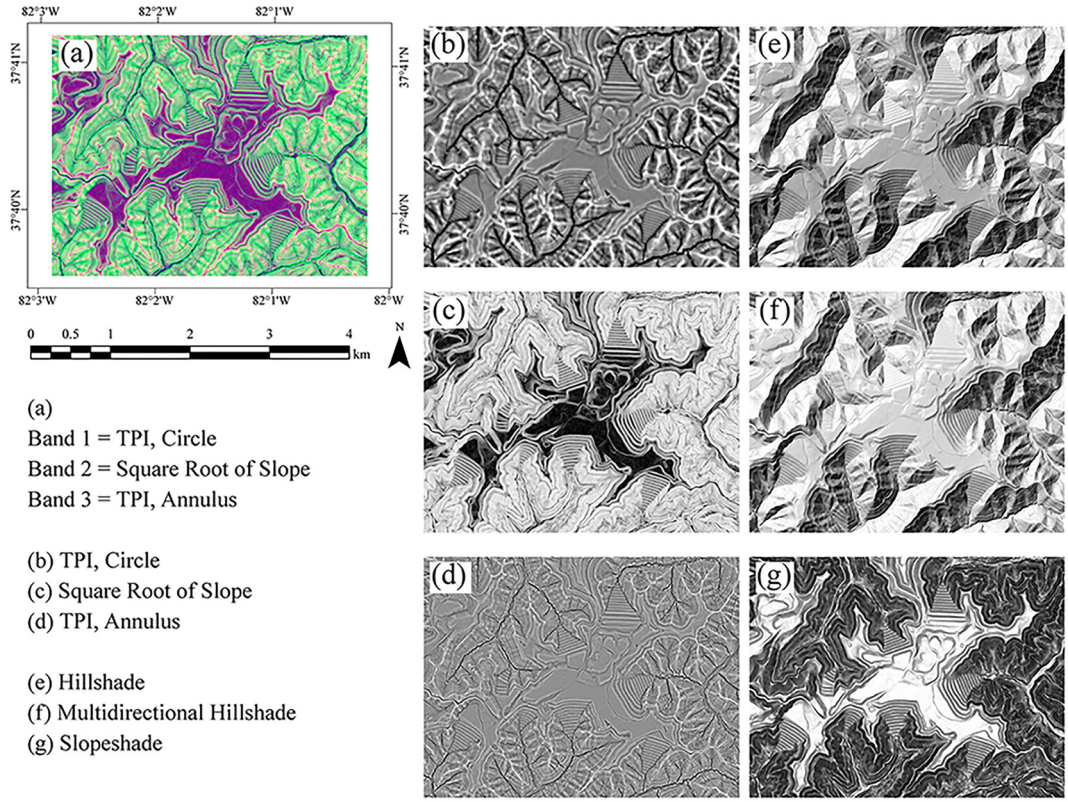

Figure 3. Example of the six land surface parameters used in this study. (a) Three-layer composite. (b) Topographic position index (TPI) with 50 m circular radius. (c) Square root of slope. (d) TPI with 2 m inner and 10 m outer radius annulus (e) Hillshade (northwest illuminating position). (f) Multidirectional hillshade. (g) Slopeshade.

models as a three-layer composite (Stack) while the HS, MDS, and SlpS were treated as separate, one-band layers in the analysis. As noted above, this specific three-layer composite was chosen based on some of the authors' experiences visually interpreting digital terrain data. We have found that this input combination is effective for visualizing local topographic characteristics in the context of surficial geologic mapping as it highlights local terrain textures, steepness, and topographic position. Given that the potential feature space is infinite, basing our test input on expert experience is a good starting point. Due to the complexity and computational demands of training and evaluating multiple algorithms using different feature spaces or LSP combinations, it was not possible to undertake a study to compare a large number of features spaces. Instead, we focused on comparing the three-layer composite that we have found useful for manual interpretation tasks to some commonly used LSPs. As highlighted in the Background section, slope and hillshades are commonly used to represent digital terrain characteristics in feature extraction and mapping tasks.

Semantic segmentation CNN-based deep learning requires that training, validation, and testing data be provided as small image chips with associated pixel-level masks (Hoeser et al., 2020; Hoeser & Kuenzer, 2020; Maxwell et al., 2021a, 2021b). All vector-based features were converted to binary grids at a 2 m spatial resolution to match that of the digital terrain data. Chips were then generated from the raster masks and LSPs data using a 256-by-256 cell chip size (i.e., 512-by-512 m) and a custom R (R Core Team, 2022) script making use of the terra package (Hijmans, 2022). A stride of 256 pixels was applied so that there was no overlap between chips.

Table 3 summarizes the number of training, validation, and testing features; chips containing some cells mapped to the positive class; and chips containing only background cells for each use case. To mitigate issues of data imbalance, only chips that included positive-class pixels were used to train the models. Due to the geographic stratification used in this study, there was no overlap between chips in the training, validation, and testing partition so as not to bias the assessment and to allow for quantification of model generalization to new geographic extents.

Table 3
Summarization of Number of Features and Positive and Background-Only Chips in Each Data Partition for Each Use Case

Use case	Set	Number of features	Positive chips	Background-only chips
Valley fill faces	Training	1,105	2,324	31,388
	Validation	304	424	4,202
	Testing	874	1,290	5,595
Agricultural terraces	Training	67,744	13,548	4,666
	Validation	15,606	4,341	4,478
	Testing	26,933	10,059	18,351
Alluvium	Training	5,084	8,925	30,489
	Validation	2,810	4,407	16,610
	Testing	4,141	4,280	23,875
Thick Till	Training	1,453	14,535	24,879
	Validation	797	7,449	13,568
	Testing	833	9,908	18,247

3.3. Models and Implementation

The model training and validation processes are conceptualized in Figure 4. Once available chips were partitioned into separate and non-overlapping training, validation, and testing sets, three separate semantic segmentation architectures were trained for all use cases: DeepLabv3+, UNet, and UNet++. Multiple architectures were tested to further generalize the findings of the study. All models were implemented using Python 3.8 (Welcome to Python.org, 2023), PyTorch 1.12 (PyTorch, 2023), and the Segmentation Models (Iakubovskii, 2022) library. Model training was conducted on a custom Linux-based workstation with the Ubuntu 20.04 operating system, an AMD Ryzen Threadripper Pro 3955WX 16-core CPU, 128 GB of RAM, and three NVIDIA RTX A5000 graphics cards with a combined 72 GB of VRAM. The CUDA 11.1 toolkit (CUDA Toolkit—Free Tools and Training | NVIDIA Developer, 2023) and cuDNN 8.5 library (CUDA Deep Neural Network, 2014) were used to implement GPU computation.

A UNet consists of an encoder component followed by a decoder component. In the original implementation of UNet, the encoder consists of multiple 3-by-3 cell double convolution blocks that learn weights associated with multiple kernels. The learned kernels and a rectified linear unit (ReLU) activation function are then applied to the convolution block inputs to generate feature maps. The sizes of the feature maps are then decreased using max pooling before being provided as input to the next convolution block, allowing for the learning of patterns at multiple spatial scales. Since the goal of semantic segmentation is to make predictions at each pixel location, the decoder must then increase the size of the data arrays in the spatial dimensions to recover the original array size. This process consists of upsampling the array using transposed convolution. In order to make use of the learned patterns at each stage in the encoder, the feature maps from the same level or array size in the encoder are concatenated via skip connections with the associated upsampled arrays as additional input to the 3-by-3 cell double convolution blocks in the decoder. Lastly, 1-by-1 convolution is used to obtain the positive class logit at each cell location, which can be converted to a class probability using a sigmoid activation function (Ronneberger et al., 2015).

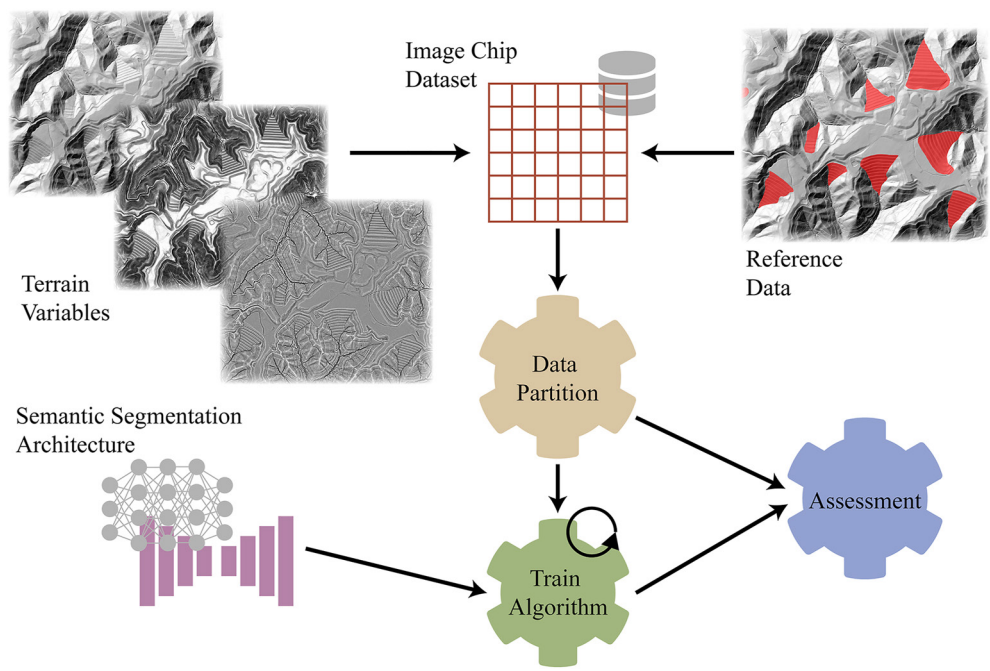

Figure 4. Conceptualization of model training and assessment processes.

UNet++ expands upon traditional UNet by incorporating densely connected convolution blocks between the encoder and decoder at each level with a goal of bridging the semantic gap between the feature maps of the encoder and the associated decoder block prior to concatenation. In other words, simple concatenation is replaced with additional convolution blocks (Zhou et al., 2019).

DeepLabv3+ is similar to UNet except with some additional components. First, it incorporates atrous or dilated convolution, in which 0s are added into the kernels, which allow for the modeling of spatial patterns between cells that are not direct neighbors, effectively increasing the field-of-view. It also incorporates atrous spatial pyramid pooling (ASPP). On top of the feature maps learned using traditional 3-by-3 and atrous convolution, an additional four parallel atrous convolutions with different atrous rates (e.g., number of inserted 0s) are applied. Further, image-level features are incorporated using global average pooling. After applying all the operations in parallel, the results are concatenated and 1-by-1 convolution is applied to obtain the positive class logit (L.-C. Chen et al., 2014, 2018; Chen, Papandreou, Kokkinos, et al., 2017; Chen, Papandreou, Schroff, & Adam, 2017).

UNet, UNet++, and DeepLabv3+ can all accommodate different CNN-architectures within the encoder blocks (Hoeser et al., 2020; Hoeser & Kuenzer, 2020). In this study, the encoder components of the semantic segmentation models were built using a ResNet-34 (He et al., 2015) architecture with a total of 5 convolutional blocks. Batch normalization was applied in all stages of the encoder and decoder. For UNet and UNet++, spatial and channel squeeze and excitation attention was applied in the decoder. In order to further increase the variability in the training data and to potentially reduce overfitting, random changes to brightness and/or contrast and blurring were applied using the Albumentations Python library (Buslaev et al., 2020). We did not apply random flips or rotations since this would artificially change the illuminating characteristics of the HSs. All models were trained for 50 epochs using all available training chips that contained some cells mapped to the positive class. The AdamW optimizer (Loshchilov & Hutter, 2017) was used with an initial learning rate of 0.001, which was reduced to 0.00001 after 25 epochs. Due to issues of class imbalance, the focal Tversky loss (Abraham & Khan, 2018) was used:

$$1 - \left(\frac{\mathrm{TP}}{\mathrm{TP} + \alpha \mathrm{FN} + \beta \mathrm{FP}}\right)^{\gamma} \tag{4}$$

where $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.75$. The training epoch that yielded the best performance, as measured using the F1-score for predicting to the validation data, was chosen as the final model to apply to the testing data set and calculate final assessment metrics.

3.4. Accuracy Assessment and Comparison

The epoch that provided the highest F1-score for predicting to the withheld validation data was saved as the final model for each use case, feature space, and algorithm combination for a total of 48 models. The 12 models for each use case were then applied to the withheld testing data. The predictions and reference labels were used to obtain counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) pixels within the testing extents. We performed model assessments using (a) just chips containing some pixels mapped to the positive case in the reference labels and (b) all chips within the validation extents. In the second case and because all chips were used, the relative proportions of the positive and background classes within the landscape were maintained, resulting in a population confusion matrix (Stehman, 2013, 2014).

Once a confusion matrix was obtained, we calculated the following summary statistics: overall accuracy (OA) (Equation 5), F1-score (Equation 6), precision (Equation 7), and recall (Equation 8). OA accuracy is the proportion of pixels that were correctly classified relative to the total number of pixels. Precision and recall are measures of 1-commission and 1-omission error relative to the positive class, respectively. F1-score is the harmonic mean of precision and recall (Tharwat, 2020).

$$\text{Overall Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{F1} - \mathrm{Score} = \frac{2\,\mathrm{X}\,\mathrm{TP}}{2\,\mathrm{X}\,\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

4. Results and Discussion

4.1. Training Losses and Prediction to Validation Data

The focal Tversky losses calculated for the training samples for each use case, algorithm, and feature space combination included in the study across 50 training epochs are summarized in Figure 5. Training loss stabilized before the final training epoch, suggesting that 50 epochs was adequate for these feature recognition tasks. The rate of stabilization varied between the use cases; for example, the losses for the agricultural terrace models tended to stabilize faster, or with fewer training epochs, than the valley fill faces experiments. The valley fill faces and agricultural terraces models generally stabilized at a lower final loss than those for the alluvium and thick till models, suggesting the valley fill faces and terraces were more easily differentiated from the background landscape or other landforms using the input feature spaces and DL-based semantic segmentation. This result was expected, as these two anthropogenic features tended to have a more distinguishable presentation and were generally easier to visually differentiate from the landscape background. For example, valley fill faces have distinctive slope patterns. In contrast, the alluvium and thick till tended to be more difficult to visually distinguish from other landforms.

For all algorithms and use cases, the model trained using the three-layer stack stabilized at a lower training loss than the HS, MHS, and SlpS models, suggesting that the three-layer stack allowed for better differentiation of the features of interest from the landscape background. The differences between feature spaces varied between models. Specifically, the three-layer models tended to perform better than the other three feature spaces for the valley fill faces and thick till use cases. Less difference was noted for the agricultural terraces. The worst performing feature space, based on training loss, was generally the SlpS models; however, all single band models performed similarly for the valley fill faces experiments.

The losses for the training data generally stabilized to a similar loss for all three implemented algorithms for all four use cases, suggesting minimal differences in performance between the three algorithms for the explored feature extraction problems. The feature spaces had a larger impact on model performance than the algorithm used, though algorithm comparison was not an objective of this study. Multiple algorithms were tested to generalize the findings associated with feature space. A study designed to compare algorithms would require more experimentation with algorithm hyperparameters and tuning processes. The performance of the tested algorithms could potentially be improved with further augmentations of architectures, training processes, and/or hyperparameter tuning. Our experiments were designed to provide a consistent training implementation to compare feature spaces in an unbiased manner as opposed to optimizing algorithm performance. It would be interesting to explore the impact of using different atrous rates within the DeepLabv3+ architecture. We expected better performance from the DeepLabv3+ algorithm in comparison to the UNet-based methods due to the use of dilated or atrous convolution and the associated increase in the size of the field-of-view (Chen et al., 2014, 2018; Chen, Papandreou, Kokkinos, et al., 2017; Chen, Papandreou, Schroff, & Adam, 2017). Similar to the impact of moving window size on LSPs and their value within modeling tasks, it would be expected that varying the sizes of the learned kernels within the DL semantic segmentation architecture would affect model performance. Additional experimentation with kernel sizes, dilated convolution, and model architectures that focus on the impact of varying field-of-view, while not the primary focus of this study, would be valuable.

The F1-score calculated for the withheld validation data at the end of the training epochs is summarized in Figure 6. Note that the first 10 epochs were excluded from the graphs due to a high degree of noise. In comparison to the training results, the validation losses show a high degree of noise before stabilizing. However, 50 epochs were found to be adequate to stabilize both the training and validation loss. Similar to the training results, the three algorithms tended to show similar performance for predicting the validation samples. Generally, the three-layer feature space provided better performance for predicting to the validation data, with the differences between feature spaces generally being more pronounced for the valley fill faces, alluvium, and thick till use cases in comparison to the agricultural terraces.

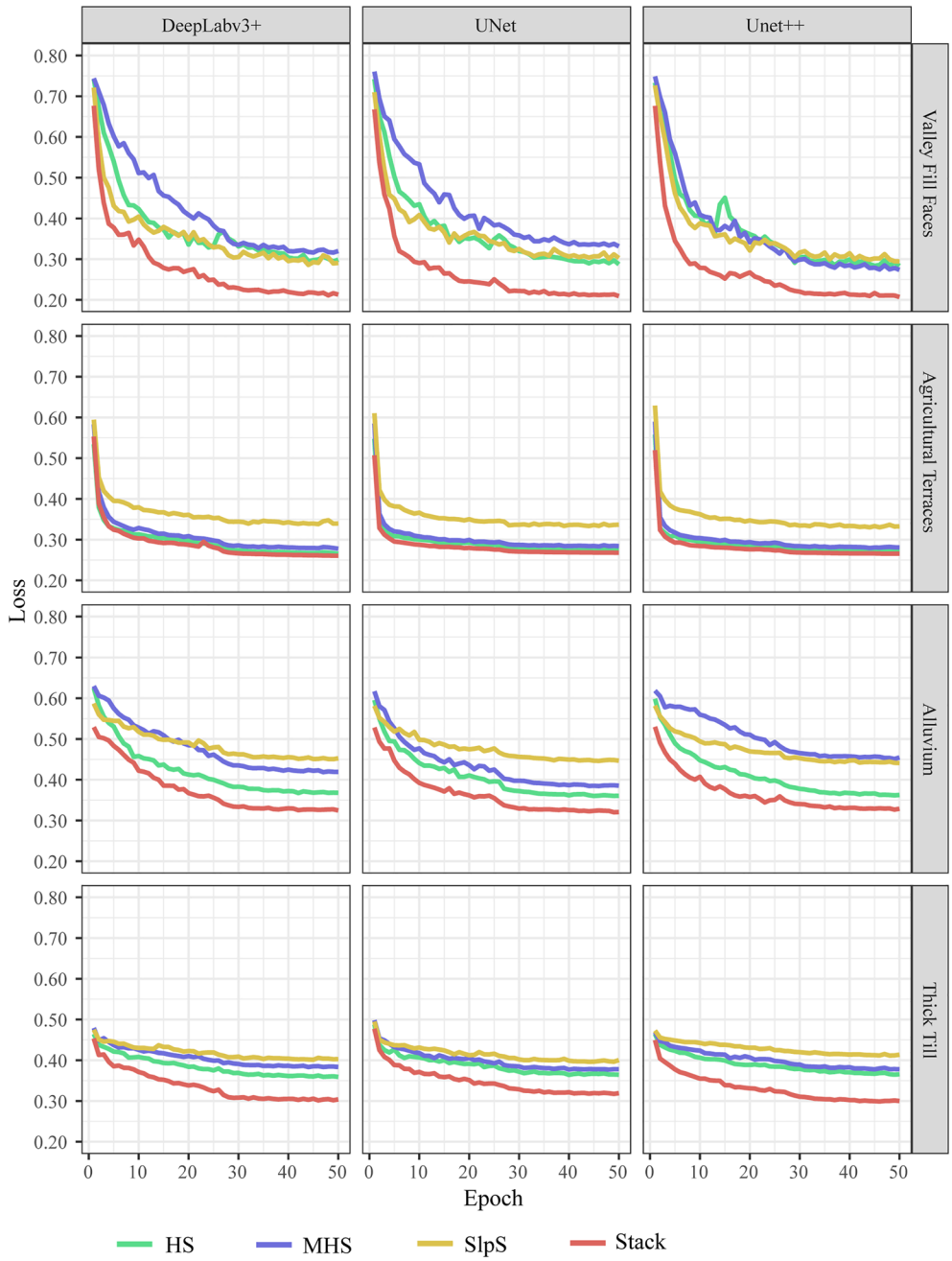

Figure 5. Change in training loss for each feature set, algorithm, and use case combination.

In summary, the losses calculated for the training data and the F1-scores calculated for the withheld validation data at the end of each training epoch suggest that the three-layer feature space outperformed the single-band feature spaces for predicting to both the training and validation data for multiple use cases, regardless of the algorithm used. In other words, all tested semantic segmentation DL algorithms were sensitive to the input feature space, meaning that the way the terrain surface was represented as LSPs impacted model performance.

4.2. Performance for Predicting to Testing Data

As noted above, the model weights associated with the epoch that yielded the highest F1-score for predicting to the validation data were selected as the final model used to predict to the testing data, as opposed to using

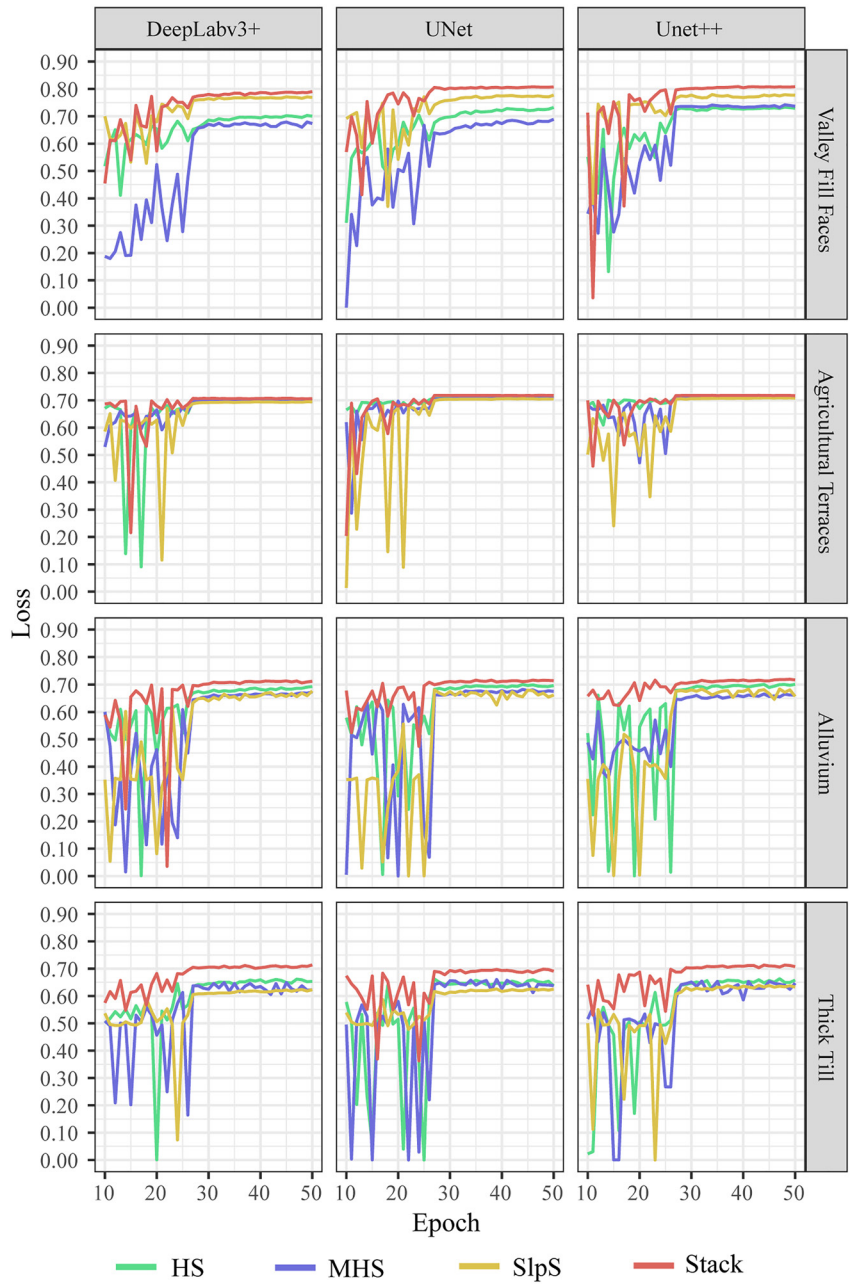

Figure 6. Change in F1-score for the validation data by epoch for each feature set and use case.

the results from the 50th epoch. Figure 7 provides the model assessment results for both the validation and testing data for the selected model for each algorithm, use case, and feature space combination. Figures 8 and 9 show example classification results for all four use cases and feature spaces using the DeepLabv3+ results. We calculated the F1-score (Figure 7a), precision (Figure 7b), and recall (Figure 7c) using just chips that had some pixels labeled to the positive case in the reference labels (P) and all chips in the validation extents ($P + B$). As noted above, the $P + B$ results maintain the relative proportions of the positive and background classes within the landscape, allowing for the estimation of a population confusion matrix and associated summary metrics (Stehman, 2013, 2014). Including the background-only chips in the assessment resulted in notable decreases in estimated precision. This is expected because precision takes into account FPs. Since the features of interest, especially for the valley fill faces and terraces use cases, were not abundant on the landscape, a relatively small number of pixels were labeled to the positive class in comparison to the background class. Due to the larger

10.1029/2023EA002845

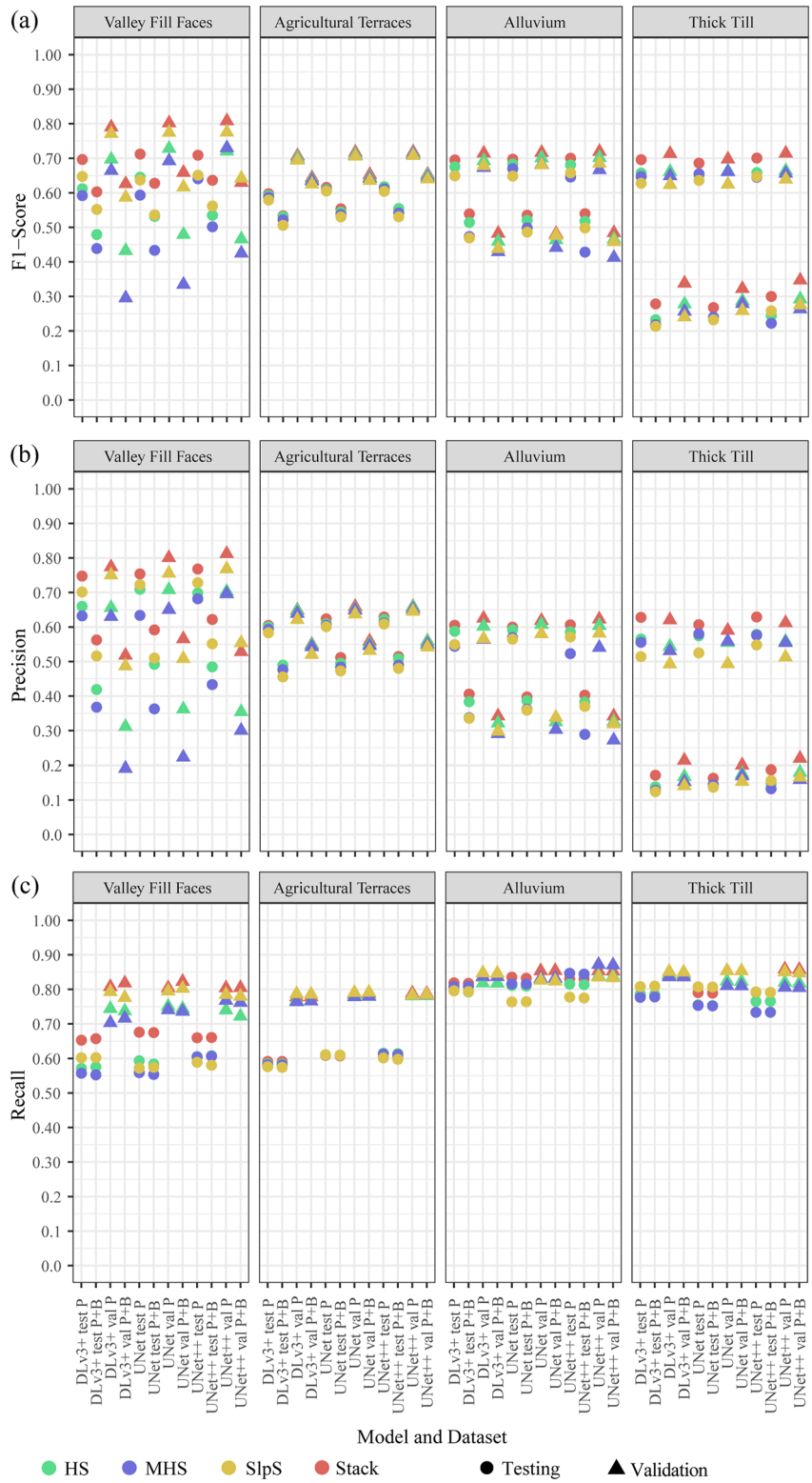

Figure 7. Assessment results using validation and testing datasets for each use case and algorithm. (a) F1-score. (b) Precision. (c) Recall. *P* indicates results for assessment using just chips containing pixels labeled to the positive class while *P* + *B* indicates all image chips in the testing and training extents.

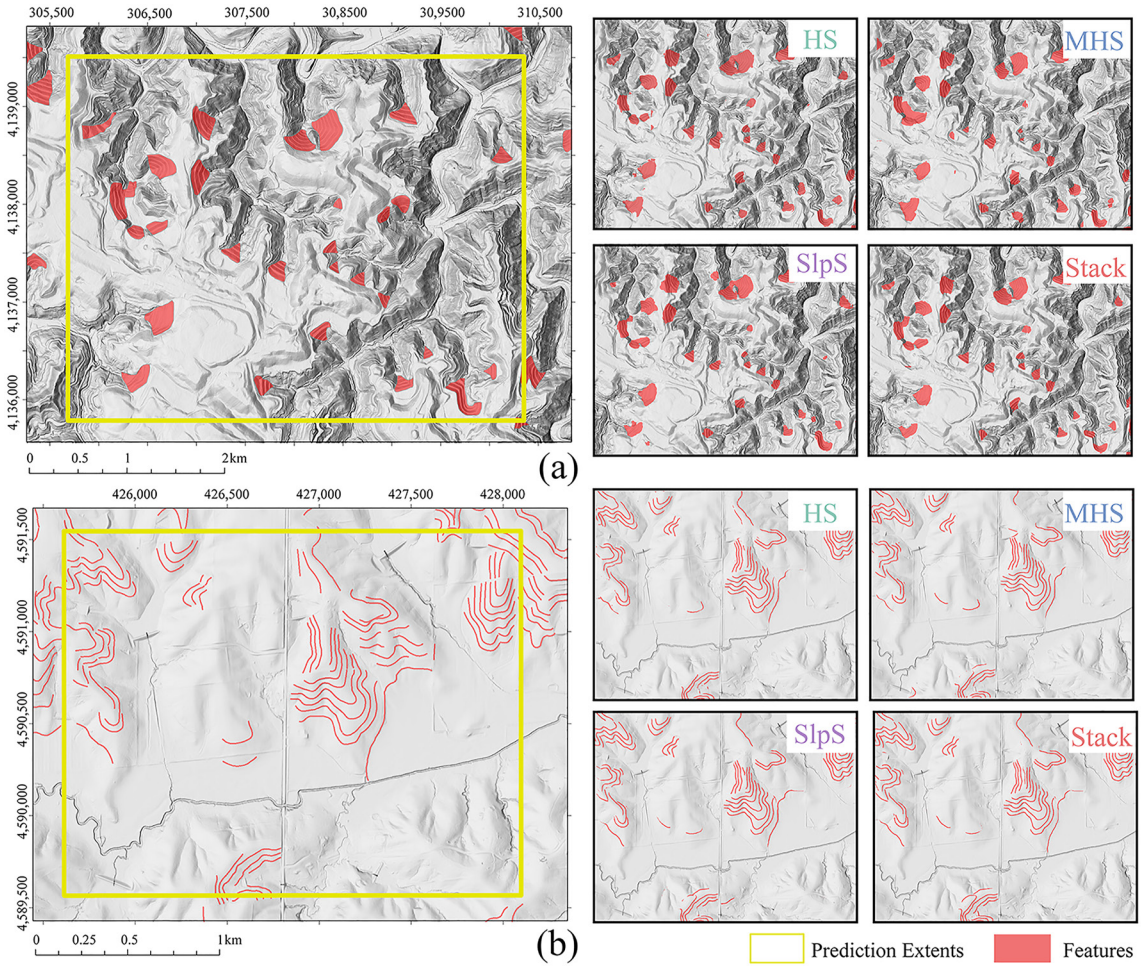

Figure 8. Example predictions using for different features spaces and the DeepLabv3+ algorithm. (a) Valley fill faces (NAD83 UTM Zone 17N). (b) Agricultural terraces (NAD83 UTM Zone 15N). Associated assessment metrics are provided in Table 4.

number of background pixels relative to positive-class pixels, there were many chances for FP outcomes in comparison to the total number of reference positive pixels, resulting in lower reported precision with an increase in the number of background pixels included in the assessment. Although alluvium and thick till generally made up a larger proportion of the landscapes predicted, reductions in precision were also noted with the inclusion of the background-only samples. We attribute this to an overprediction of alluvium and thick till (i.e., FPs), as evident in the example results presented in Figure 9. A comparable level of FP occurrence is not noted for the valley fill faces and terraces, as visualized in Figure 8. Precision was generally lower than recall for all use cases. This suggests that FPs were generally more of an issue that FNs. In other words, the algorithms tended to overpredict the extent of the features of interest, especially for the alluvium and thick till use cases, as opposed to missing samples. It should be noted that FPs and FNs may not be equally undesirable in applied mapping tasks. For example, deleting FPs from the resulting datasets would generally require less manual labor than digitizing missed features or FNs.

Model performance was generally higher for the validation data in comparison to the testing data. Since the final models were selected based on the highest F1-score for the validation data, it makes sense that model performance would be higher for the validation data set in comparison to the testing data, which were not used to select the final model from the set of weights associated with specific training epochs.

Similar to the training losses (Figure 5) and validation F1-scores (Figure 6), the assessment results using the testing data suggest that the three-band feature spaces provided better performance than the one-band feature spaces based on reported OA, F1-score, precision, and recall. Table 4 summarizes the testing assessment results

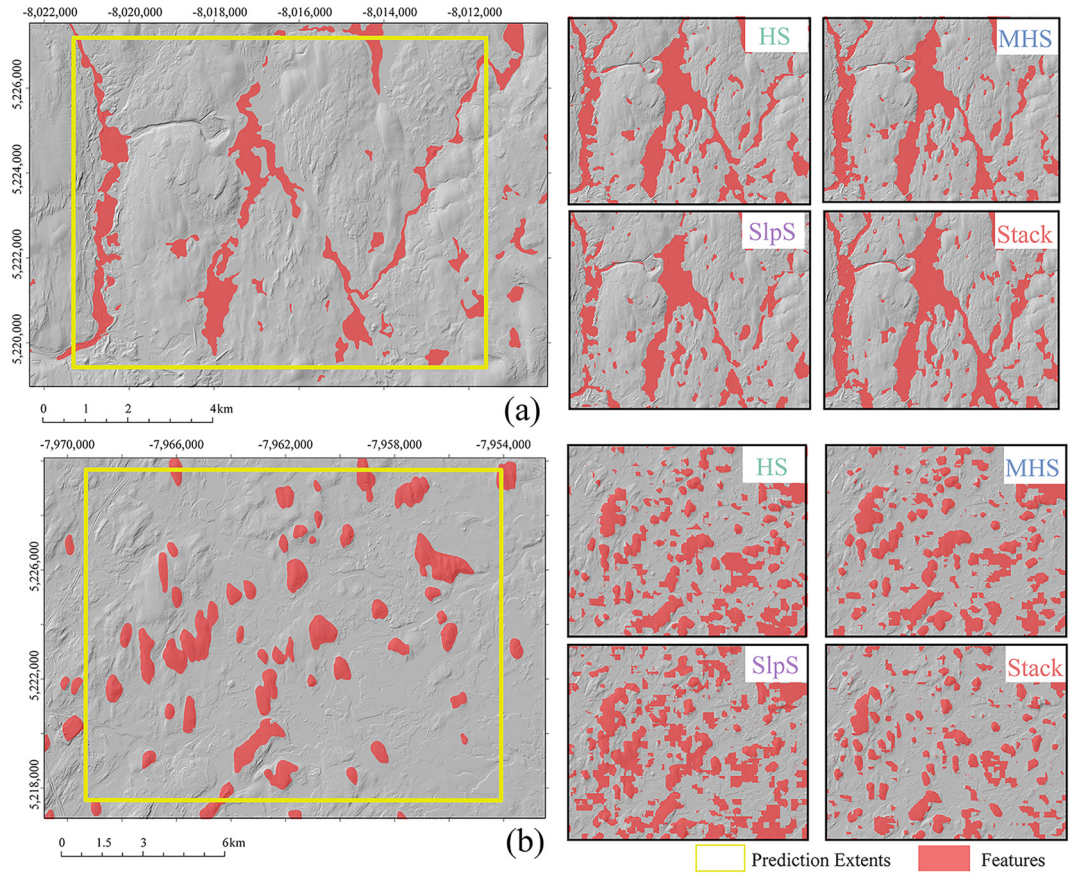

Figure 9. Example predictions using for different features spaces and the DeepLabv3+ algorithm. (a) Alluvium (WGS84 Web Mercator). (b) Glacial Till (WGS84 Web Mercator). Associated assessment metrics are provided in Table 4.

using the positive and background-only chips and just the positive chips. The three-band feature space provided the best performance in regards to F1-score, precision, and recall. Note that OA for all models tended to be higher when including the background only chips. This can be attributed to the large number of background pixels and the associated low number of FNs. As noted above, precision was generally lower when incorporating the background-only chips due to the large number of background pixels and the associated increase in FPs relative to the low number of FNs. In this study, the assessment using withheld testing data can be interpreted as an assessment of model generalization to new geographic extents but not to new input data since the same DTMs were used in the study but the training, validation, and testing partitions were defined based on geographic extents (Maxwell et al., 2021b).

Comparing the four investigated use cases, the assessment results for the testing data generally suggest that the valley fill faces and agricultural terraces were easier to predict than the alluvium and thick till. Again, this was anticipated based on the relative difficulty in manually interpreting these features in the LSPs in comparison to the valley fill faces and agricultural terraces. This highlights the value of exploring the impact of feature space using multiple use cases with varying levels of difficulty and different proportions of positive and background classes (i.e., differing levels of class imbalance). Another factor to consider is that LSPs not explored here may be valuable for improving the mapping of alluvium and thick till; the features explored here may have been more predictive for the valley fill faces and agricultural terraces features as opposed to the alluvium and thick till features.

5. Recommendations and Conclusions

Deep learning has great potential as a surficial mapping tool, but requires feature spaces that enable accurate mapping across a range of geomorphic features. Our results suggest that the input feature space influences model

Table 4
Assessment Results for Testing Set Predictions for all Use Cases, Algorithms, and Feature Spaces

			Positive + background-only				Positive			
Use case	Algorithm	Feature space	OA	*F*	*R*	*P*	OA	*F*	*R*	*P*
Valley fill faces	DeepLabv3+	HS	0.980	0.479	0.575	0.419	0.937	0.612	0.571	0.660
		MDHS	0.978	0.439	0.552	0.368	0.933	0.592	0.557	0.632
		SlpSd	0.985	0.552	0.602	0.516	0.943	0.647	0.602	0.701
		Stack	0.986	0.603	0.657	0.562	0.951	0.696	0.653	0.747
	UNet	HS	0.984	0.532	0.584	0.492	0.943	0.645	0.593	0.709
		MDHS	0.977	0.433	0.554	0.363	0.933	0.593	0.559	0.634
		SlpSd	0.984	0.536	0.574	0.510	0.943	0.637	0.572	0.722
		Stack	0.987	0.627	0.675	0.592	0.952	0.712	0.676	0.754
	UNet++	HS	0.983	0.535	0.608	0.485	0.943	0.647	0.604	0.698
		MDHS	0.981	0.502	0.607	0.433	0.941	0.641	0.605	0.682
		SlpSd	0.986	0.562	0.580	0.551	0.945	0.651	0.589	0.728
		Stack	0.988	0.636	0.661	0.622	0.953	0.709	0.660	0.768
Agricultural terraces	DeepLabv3+	HS	0.993	0.531	0.584	0.490	0.984	0.592	0.585	0.601
		MDHS	0.993	0.522	0.582	0.476	0.984	0.587	0.582	0.593
		SlpSd	0.992	0.506	0.574	0.455	0.983	0.579	0.576	0.583
		Stack	0.993	0.533	0.591	0.489	0.984	0.597	0.591	0.605
	UNet	HS	0.993	0.544	0.610	0.495	0.985	0.610	0.611	0.609
		MDHS	0.993	0.537	0.609	0.484	0.984	0.607	0.610	0.605
		SlpSd	0.993	0.530	0.609	0.473	0.984	0.605	0.610	0.601
		Stack	0.993	0.554	0.608	0.512	0.985	0.616	0.609	0.624
	UNet++	HS	0.993	0.554	0.614	0.508	0.985	0.618	0.615	0.622
		MDHS	0.993	0.541	0.611	0.490	0.985	0.612	0.613	0.612
		SlpSd	0.993	0.531	0.597	0.480	0.984	0.604	0.601	0.609
		Stack	0.993	0.553	0.603	0.515	0.985	0.615	0.603	0.629
Alluvium	DeepLabv3+	HS	0.746	0.232	0.785	0.138	0.855	0.675	0.796	0.588
		MDHS	0.727	0.218	0.778	0.128	0.836	0.650	0.808	0.544
		SlpSd	0.709	0.213	0.809	0.124	0.838	0.649	0.796	0.549
		Stack	0.804	0.278	0.781	0.172	0.865	0.695	0.819	0.605
	UNet	HS	0.759	0.234	0.753	0.140	0.859	0.683	0.811	0.592
		MDHS	0.767	0.240	0.752	0.145	0.848	0.670	0.816	0.569
		SlpSd	0.739	0.232	0.807	0.137	0.844	0.648	0.764	0.565
		Stack	0.789	0.268	0.789	0.163	0.863	0.697	0.835	0.599
	UNet++	HS	0.768	0.244	0.765	0.147	0.856	0.681	0.815	0.586
		MDHS	0.750	0.222	0.734	0.132	0.825	0.646	0.846	0.523
		SlpSd	0.778	0.258	0.791	0.156	0.847	0.658	0.777	0.571
		Stack	0.821	0.300	0.791	0.187	0.866	0.700	0.830	0.606
Thick till	DeepLabv3+	HS	0.902	0.514	0.792	0.383	0.733	0.657	0.786	0.565
		MDHS	0.882	0.473	0.808	0.337	0.724	0.647	0.777	0.556
		SlpSd	0.882	0.469	0.794	0.336	0.689	0.628	0.808	0.514
		Stack	0.908	0.539	0.817	0.406	0.778	0.696	0.783	0.628

Table 4
Continued

Use case	Algorithm	Feature space	Positive + background-only				Positive			
			OA	*F*	*R*	*P*	OA	*F*	*R*	*P*
	UNet	HS	0.902	0.520	0.810	0.387	0.738	0.652	0.755	0.575
		MDHS	0.892	0.498	0.815	0.362	0.742	0.655	0.754	0.580
		SlpSd	0.893	0.486	0.764	0.359	0.700	0.636	0.807	0.525
		Stack	0.906	0.535	0.832	0.398	0.765	0.686	0.791	0.607
	UNet++	HS	0.900	0.518	0.814	0.383	0.742	0.658	0.765	0.579
		MDHS	0.852	0.428	0.844	0.289	0.736	0.645	0.734	0.576
		SlpSd	0.898	0.498	0.775	0.370	0.719	0.647	0.793	0.548
		Stack	0.907	0.540	0.830	0.403	0.779	0.701	0.792	0.629

Note. OA = Overall Accuracy, *F* = F1-Score, *R* = Recall, *P* = Precision. Gray shades indicate the best performing model or models based on that metric for each use case.

performance, as the three-layer composite tended to provide better predictions than the single-band HS, MHS, and SlpS. This result was consistent across three algorithms (DeepLabv3+, UNet, and UNet++) and four separate use cases (valley fill faces, agricultural terraces, alluvium, and thick till). The three-band feature space generally resulted in lower losses for the training data across training epochs; higher validation F1-scores across training epochs; and higher OA, F1-score, precision, and recall for predicting or generalizing to withheld testing data using the model epoch that yielded the highest F1-score for predicting to the validation data. Figure 10 highlights the landscapes explored in this study as presented in the proposed three-layer stack. We argue that this combination is effective because it characterizes hillslope position, steepness, and local surface roughness or textures. HSs and SlpS do not offer the same level of contextual detail.

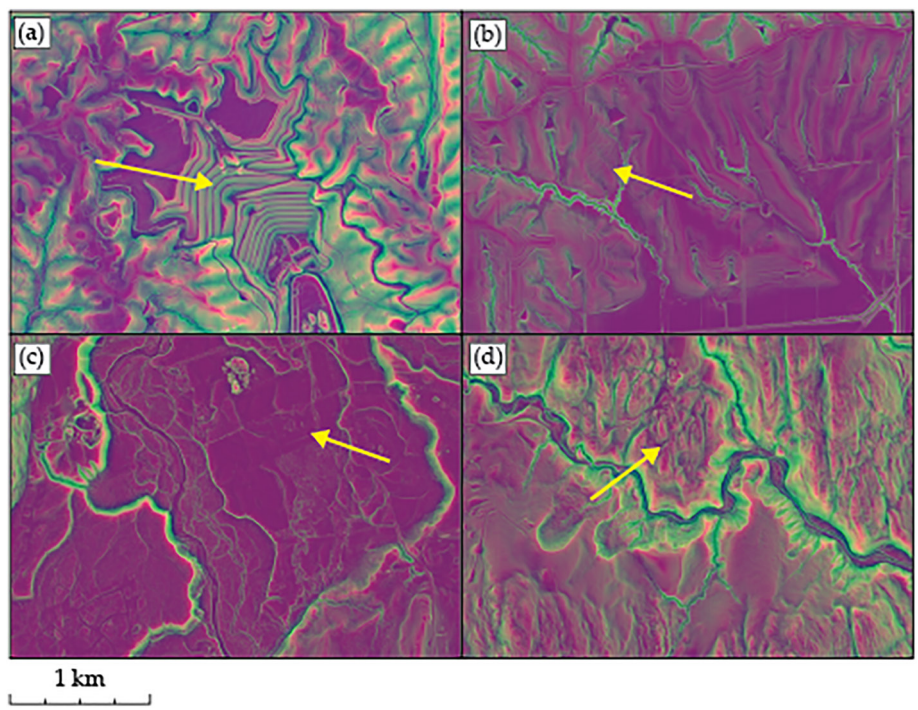

Figure 10. Three-layer stack imagery of the four mapped feature types. All images are at the same spatial and color scales. (a) Valley fill faces along Caney Creek, Virginia. (b) Agricultural terraces on hillslopes of subcatchments of Clanton Creek, Iowa. (c) Alluvium north of Greenfield Town, Massachusetts. (d) Thick till west of Greenfield Town, Massachusetts. Note slope failure of till material. Arrows indicate features of interest.

There are some notable limitations and complexities in this study. First, we only ran a single model for each use case, algorithm, and feature space combination. Running multiple models initialized using different random weights could allow for assessing the variability in model results. However, this was not feasible in this study due to computational time and cost. Even when executing model training using three GPUs and a combined 72 GB of VRAM, running all 48 models for 50 epochs required over three days of computational time. Further, it was not possible to test a wide variety of algorithm architectural manipulations, training methodologies, or hyperparameters due to the large number of models that had to be run and the computational and time costs of undertaking these experiments. Instead, we opted to use consistent settings and algorithm architectures to foster unbiased comparisons across feature spaces, use cases, and algorithms. Given the infinite set of LSPs that can be derived from DTMs (Franklin, 2020; Maxwell & Shobe, 2022), testing a larger set of LSP combinations could have been informative. However, this was not possible due to computational costs. There is a need to explore additional feature spaces and other mapping problems to further quantify the impact of selected input LSPs on model performance. There are a variety of issues associated with mapping and extracting landform features that will need to be explored in the context of CNN-based DL. For example, mapping hierarchical features at varying scales and/or features with inherently uncertain or gradational boundaries continue to be persistent challenges. Multiclass, scale-appropriate, and application-specific landform classification schemes will need to be defined as research on applying DL to such mapping tasks progresses.

This study highlights the difficulty in training and assessing models when classes are heavily imbalanced, as discussed above. The focal Tversky loss helped alleviate this issue to some degree; however, class imbalance still presented a challenge. The reference data used in this study were imperfect due to errors in manually digitizing, inconsistencies in interpretations, mapping difficulties, fuzzy or gradational boundaries, and/or potential landscape changes resulting in misalignment between the labels and DTMs. However, we argue that these datasets were of adequate quality to address the proposed research question. One notable issue is that the features mapped may have fuzzy or gradational boundaries with the background class or other landforms. This can result in overly harsh assessment results that assume "hard" boundaries between classes (Foody, 2008; Maxwell & Warner, 2020).

Future studies should investigate multi-class feature extraction or landform mapping problems, instance segmentation methods, such as mask R-CNN, and the potential benefits of using semi-supervised learning methods. As noted above, further investigation of architectural changes that impact the field-of-view, such as kernel sizes and dilation rates when using atrous convolution, would be valuable. The DeepLabv3+ architecture is of specific interest in such explorations due to its reliance on atrous convolution. However, other base architectures can be augmented to incorporate dilated or atrous convolution. It would also be possible to explore data augmentations, such as applying smoothing operations, and the associated impact on model performance. Behrens et al. (2018) proposed a Gaussian pyramid method that allows for the generalization of DTMs at varying scales using downscaling and subsequent upscaling. It would be interesting to explore this method as a means to augment input features during different stages in the semantic segmentation architecture.

Our study generally complements the findings of Suh et al. (2021): CNN-based semantic segmentation DL for geomorphic mapping and feature extraction is sensitive to input feature space. This study expands upon the prior study by exploring different LSPs, use cases, and DL semantic segmentation architectures. Given the variety of LSPs that can be produced; the complexity of sematic segmentation architectures; and options for changing window sizes, shapes, and cell weightings, additional research focused on feature space considerations could improve model performance. Given the results of this and prior studies, we argue that researchers and analysts exploring such mapping problems should carefully consider the input features provided to the algorithm. Commonly used terrain visualization surfaces, such as HSs and MHSs, may not serve as an optimal feature space. We have qualitatively found the three-layer stack used here, which consists of TPI calculated with a 50 m circular radius, the square root of topographic slope, and TPI calculated with an annulus with a 2 m inner radius and 10 m outer radius, to be especially informative and applicable for manual interpretation for surficial geologic mapping and landform identification in comparison to HSs. This study quantifies this in the context of automated mapping using semantic segmentation DL. The DL models were able to extract both natural and anthropogenic landforms more accurately from LSPs that we have found to be appropriate for visual interpretation. Other researchers may find it beneficial to implement this specific LSP combination for other geomorphic mapping problems. Further

refinement of the optimal feature space for common surficial mapping problems could allow more efficient generation of accurate geomorphic datasets from increasingly available high-resolution topography.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The valley fill faces data are available at Maxwell (2023c; https://doi.org/10.6084/m9.figshare.22318522.v2). The agricultural terraces data are available at Maxwell (2023b; https://doi.org/10.6084/m9.figshare.22320373.v2). The thick glacial till and alluvium surficial features data are available at Maxwell (2023a; https://doi.org/10.6084/m9.figshare.22320481.v1). Each data set includes ArcGIS Pro ModelBuilder tools, Python notebooks, and R scripts for creating terrain derivatives from DTMs, generating image chips and associated masks, generating lists of chips in a directory, training deep learning semantic segmentation models, making inferences to new data, and assessing model performance with withheld testing data. Data have been partitioned into separate training, validation, and testing sets. Study area extents and features are stored as vector geospatial data in shapefile format. DTMs are provided in TIFF format.

Acknowledgments
We would like to thank Phillip Goodling, Robert Stamm, Josh Woda, and two anonymous reviewers whose comments strengthened the work, as well as Robin McNeely and Josh Obrecht, the West Virginia GIS Technical Center, and Paul Nutting for providing access to data. Funding was provided by the National Science Foundation (Federal Award ID No. 2046059: "CAREER: Mapping Anthropocene Geomorphology with Deep Learning, Big Data Spatial Analytics, and lidar"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Funding was provided by AmericaView, which is supported by the U.S. Geological Survey under Grant/Cooperative Agreement No. G18AP00077. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

2D, 3D & 4D GIS Mapping Software | ArcGIS Pro. (2023). Retrieved from https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Abraham, N., & Khan, N. M. (2018). A novel focal Tversky loss function with improved attention U-net for lesion segmentation. ArXiv Preprint ArXiv:1810.07842.

Albrecht, C. M., Fisher, C., Freitag, M., Hamann, H. F., Pankanti, S., Pezzutti, F., & Rossi, F. (2019). Learning and recognizing archeological features from LiDAR data. *2019 IEEE International Conference on Big Data (Big Data)*, 5630–5636. https://doi.org/10.1109/BigData47090.2019.9005548

Baker, V. (1986). Introduction: Regional landforms analysis. In *Geomorphology from space: A global overview of regional landforms* (p. 717). NASA. (NASA SP-486).

Behrens, T., Schmidt, K., MacMillan, R. A., & Viscarra Rossel, R. A. (2018). Multi-scale digital soil mapping with deep learning. *Scientific Reports*, *8*(1), 15244. Article 1. https://doi.org/10.1038/s41598-018-33516-6

Bickel, V. T., Moseley, B., Lopez-Francos, I., & Shirley, M. (2021). Peering into lunar permanently shadowed regions with deep learning. *Nature Communications*, *12*(1), 5607. Article 1. https://doi.org/10.1038/s41467-021-25882-z

Bishop, M. P., James, L. A., Shroder, J. F., & Walsh, S. J. (2012). Geospatial technologies and digital geomorphological mapping: Concepts, issues and research. *Geomorphology*, *137*(1), 5–26. https://doi.org/10.1016/j.geomorph.2011.06.027

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(1), 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., et al. (2014). Geographic object-based image analysis–towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, *87*, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014

Brunsden, D., Doornkamp, J. C., Fookes, P. G., Jones, D. K. C., & Kelly, J. M. H. (1975). Large scale geomorphological mapping and highway engineering design. *The Quarterly Journal of Engineering Geology*, *8*(4), 227–253. https://doi.org/10.1144/gsl.qjeg.1975.008.04.01

Burrough, P. A. (2020). *Natural objects with indeterminate boundaries*. CRC Press.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Chen, G., Weng, Q., Hay, G. J., & He, Y. (2018). Geographic object-based image analysis (GEOBIA): Emerging trends and future opportunities. *GIScience and Remote Sensing*, *55*(2), 159–182. https://doi.org/10.1080/15481603.2018.1426092

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected CRFs. ArXiv Preprint ArXiv:1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/tpami.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. ArXiv Preprint ArXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

CUDA Deep Neural Network. (2014). NVIDIA developer. Retrieved from https://developer.nvidia.com/cudnn

CUDA Toolkit—Free Tools and Training | NVIDIA Developer. (2023). Retrieved from https://developer.nvidia.com/cuda-toolkit

Diaz-Varela, R. A., Zarco-Tejada, P. J., Angileri, V., & Loudjani, P. (2014). Automatic identification of agricultural terraces through object-oriented analysis of very high resolution DSMs and multispectral imagery obtained from an unmanned aerial vehicle. *Journal of Environmental Management*, *134*, 117–126. https://doi.org/10.1016/j.jenvman.2014.01.006

d'Oleire-Oltmanns, S., Eisank, C., Drăgut, L., & Blaschke, T. (2013). An object-based workflow to extract landforms at multiple scales from two distinct data types. *IEEE Geoscience and Remote Sensing Letters*, *10*(4), 947–951. https://doi.org/10.1109/LGRS.2013.2254465

Dornik, A., Drăguţ, L., & Urdea, P. (2018). Classification of soil types using geographic object-based image analysis and random forests. *Pedosphere*, *28*(6), 913–925. https://doi.org/10.1016/S1002-0160(17)60377-1

Drăguţ, L., & Blaschke, T. (2006). Automated classification of landform elements using object-based image analysis. *Geomorphology*, *81*(3), 330–344. https://doi.org/10.1016/j.geomorph.2006.04.013

Dramis, F., Guida, D., & Cestari, A. (2011). Chapter three—nature and aims of geomorphological mapping. In M. J. Smith, P. Paron, & J. S. Griffiths (Eds.) *Developments in Earth surface processes* (Vol. 15, pp. 39–73). Elsevier. https://doi.org/10.1016/B978-0-444-53446-0.00003-3

Du, L., You, X., Li, K., Meng, L., Cheng, G., Xiong, L., & Wang, G. (2019). Multi-modal deep learning for landform recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 63–75. https://doi.org/10.1016/j.isprsjprs.2019.09.018

Evans, I. S. (2012). Geomorphometry and landform mapping: What is a landform? *Geomorphology*, *137*(1), 94–106. https://doi.org/10.1016/j.geomorph.2010.09.029

Fayyad, J., Jaradat, M. A., Gruyer, D., & Najjaran, H. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, *20*(15), 4220. https://doi.org/10.3390/s20154220

Feizizadeh, B., Kazemi Garajeh, M., Blaschke, T., & Lakes, T. (2021). An object based image analysis applied for volcanic and glacial landforms mapping in Sahand Mountain, Iran. *CATENA*, *198*, 105073. https://doi.org/10.1016/j.catena.2020.105073

Fernandez-Diaz, J. C., Carter, W. E., Shrestha, R. L., Leisz, S. J., Fisher, C. T., González, A. M., et al. (2014). Archaeological prospection of north Eastern Honduras with airborne mapping LiDAR. *2014 IEEE Geoscience and Remote Sensing Symposium*, 902–905. https://doi.org/10.1109/IGARSS.2014.6946571

Flageollet, J.-C. (1996). The time dimension in the study of mass movements. *Geomorphology*, *15*(3–4), 185–190. https://doi.org/10.1016/0169-555x(95)00069-h

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, *29*(11), 3137–3158. https://doi.org/10.1080/01431160701442120

Franklin, S. E. (2020). Interpretation and use of geomorphometry in remote sensing: A guide and review of integrated applications. *International Journal of Remote Sensing*, *41*(19), 7700–7733. https://doi.org/10.1080/01431161.2020.1792577

Fritz, K. M., Fulton, S., Johnson, B. R., Barton, C. D., Jack, J. D., Word, D. A., & Burke, R. A. (2010). Structural and functional characteristics of natural and constructed channels draining a reclaimed mountaintop removal and valley fill coal mine. *Journal of the North American Benthological Society*, *29*(2), 673–689. https://doi.org/10.1899/09-060.1

Gholami, H., Mohammadifar, A., Golzari, S., Kaskaoutis, D. G., & Collins, A. L. (2021). Using the Boruta algorithm and deep learning models for mapping land susceptibility to atmospheric dust emissions in Iran. *Aeolian Research*, *50*, 100682. https://doi.org/10.1016/j.aeolia.2021.100682

Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159. https://doi.org/10.1109/tmi.2016.2553401

Gustavsson, M., Kolstrup, E., & Seijmonsbergen, A. C. (2006). A new symbol-and-GIS based detailed geomorphological mapping system: Renewal of a scientific discipline for understanding landscape development. *Geomorphology*, *77*(1), 90–111. https://doi.org/10.1016/j.geomorph.2006.01.026

Guyot, A., Hubert-Moy, L., & Lorho, T. (2018). Detecting neolithic burial mounds from LiDAR-derived elevation data using a multi-scale approach and machine learning techniques. *Remote Sensing*, *10*(2), 225. Article 2. https://doi.org/10.3390/rs10020225

Guyot, A., Lennon, M., Lorho, T., & Hubert-Moy, L. (2021). Combined detection and segmentation of archeological structures from LiDAR data using a deep learning approach. *Journal of Computer Applications in Archaeology*, *4*(1), 1. https://doi.org/10.5334/jcaa.64

Hall-Beyer, M. (2017). Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. *International Journal of Remote Sensing*, *38*(5), 1312–1338. https://doi.org/10.1080/01431161.2016.1278314

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*, 610–621. https://doi.org/10.1109/tsmc.1973.4309314

Haralick, R. M., & Shanmugam, K. S. (1974). Combined spectral and spatial processing of ERTS imagery data. *Remote Sensing of Environment*, *3*(1), 3–13. https://doi.org/10.1016/0034-4257(74)90033-9

Hassaballah, M., & Awad, A. I. (2020). *Deep learning in computer vision: Principles and applications*. CRC Press.

Hay, G. J., & Castilla, G. (2008). Geographic object-based image analysis (GEOBIA): A new name for a new discipline. In *Object-based image analysis* (pp. 75–89). Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. https://doi.org/10.48550/arXiv.1512.03385

Hijmans, R. J. (2022). terra: Spatial data analysis. Retrieved from https://CRAN.R-project.org/package=terra

Hillshade (Spatial Analyst)—ArcGIS Pro | Documentation. (2023). Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/hillshade.htm

Hoeser, T., Bachofer, F., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. *Remote Sensing*, *12*(18), 3053. Article 18. https://doi.org/10.3390/rs12183053

Hoeser, T., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review-Part I: Evolution and recent trends. *Remote Sensing*, *12*(10), 1667. Article 10. https://doi.org/10.3390/rs12101667

Huang, F., Zhang, J., Zhou, C., Wang, Y., Huang, J., & Zhu, L. (2020). A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides*, *17*(1), 217–229. https://doi.org/10.1007/s10346-019-01274-9

Iakubovskii, P. (2022). Qubvel/segmentation_models.pytorch [Python]. (Original work published 2019). Retrieved from https://github.com/qubvel/segmentation_models.pytorch

Iowa BMP Mapping Project—Geographic Information Systems. (2023). Retrieved from https://www.gis.iastate.edu/BMPs

Jacek, S. (1997). Landform characterization with geographic information systems. *Photogrammetric Engineering & Remote Sensing*, *63*(2), 183–191.

Janowski, L., Tylmann, K., Trzcinska, K., Rudowski, S., & Tegowski, J. (2022). Exploration of glacial landforms by object-based image analysis and spectral parameters of digital elevation model. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–17. https://doi.org/10.1109/TGRS.2021.3091771

Jordan, G., & Schott, B. (2005). Application of wavelet analysis to the study of spatial pattern of morphotectonic lineaments in digital terrain models. A case study. *Remote Sensing of Environment*, *94*(1), 31–38. https://doi.org/10.1016/j.rse.2004.08.013

Kazemi Garajeh, M., Feizizadeh, B., Weng, Q., Rezaei Moghaddam, M. H., & Kazemi Garajeh, A. (2022). Desert landform detection and mapping using a semi-automated object-based image analysis approach. *Journal of Arid Environments*, *199*, 104721. https://doi.org/10.1016/j.jaridenv.2022.104721

Lagacherie, P. (2008). Digital soil mapping: A state of the art. In A. E. Hartemink, A. McBratney, & M. L. deMendonça-Santos (Eds.) *Digital soil mapping with limited data* (pp. 3–14). Springer. https://doi.org/10.1007/978-1-4020-8592-5_1

LAS Dataset To Raster (Conversion)—ArcGIS Pro | Documentation. (2023). Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/conversion/las-dataset-to-raster.htm

Li, M., Zang, S., Zhang, B., Li, S., & Wu, C. (2014). A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing*, *47*(1), 389–411. https://doi.org/10.5721/eujrs20144723
Li, S., Xiong, L., Tang, G., & Strobl, J. (2020). Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery. *Geomorphology*, *354*, 107045. https://doi.org/10.1016/j.geomorph.2020.107045
Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.
Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.1711.05101
Lu, H., & Shi, H. (2020). Deep learning for 3D point cloud understanding: A survey. ArXiv Preprint ArXiv:2009.08920.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015
Ma, Y., Minasny, B., Malone, B. P., & Mcbratney, A. B. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science*, *70*(2), 216–235. https://doi.org/10.1111/ejss.12790
Mainak, B., Venkatanareshbabu, K., Luca, S., Damodar, R. E., Elisa, C.-G., Tato, M. R., et al. (2019). State-of-the-art review on deep learning in medical imaging. *Frontiers in Bioscience-Landmark*, *24*(3), 380–406.
Maxwell, A. (2023a). surficialDL: A geomorphology deep learning dataset of alluvium and thick glacial till derived form 1:24,000 scale surficial geology data for the Western portion of Massachusetts, USA (Version 1). [Dataset]. figshare. https://doi.org/10.6084/m9.figshare.22320481.v1
Maxwell, A. (2023b). terraceDL: A geomorphology deep learning dataset of agricultural terraces in Iowa, USA (Version 2). [Dataset]. figshare. https://doi.org/10.6084/m9.figshare.22320373.v2
Maxwell, A. (2023c). vfillDL: A geomorphology deep learning dataset of valley fill faces resulting from mountaintop removal coal mining (southern West Virginia, eastern Kentucky, and southwestern Virginia, USA) (Version 2). [Dataset]. figshare. https://doi.org/10.6084/m9.figshare.22318522.v2
Maxwell, A. E., Pourmohammadi, P., & Poyner, J. D. (2020). Mapping the topographic features of mining-related valley fills using mask R-CNN deep learning and digital elevation data. *Remote Sensing*, *12*(3), 547. Article 3. https://doi.org/10.3390/rs12030547
Maxwell, A. E., & Shobe, C. M. (2022). Land-surface parameters for spatial predictive mapping and modeling. *Earth-Science Reviews*, *226*, 103944. https://doi.org/10.1016/j.earscirev.2022.103944
Maxwell, A. E., & Warner, T. A. (2020). Thematic classification accuracy assessment with inherently uncertain boundaries: An argument for center-weighted accuracy assessment metrics. *Remote Sensing*, *12*(12), 1905. Article 12. https://doi.org/10.3390/rs12121905
Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343
Maxwell, A. E., Warner, T. A., & Guillén, L. A. (2021a). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sensing*, *13*(13), 2450. Article 13. https://doi.org/10.3390/rs13132450
Maxwell, A. E., Warner, T. A., & Guillén, L. A. (2021b). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 2: Recommendations and best practices. *Remote Sensing*, *13*(13), 2591. Article 13. https://doi.org/10.3390/rs13132591
McMaster, R. B., & Sheppard, E. (2004). Introduction: Scale and geographic inquiry. *Scale and Geographic Inquiry: Nature, Society, and Method*, 1–22.
Miglani, A., & Kumar, N. (2019). Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. *Vehicular Communications*, *20*, 100184. https://doi.org/10.1016/j.vehcom.2019.100184
Minár, J., & Evans, I. S. (2008). Elementary forms for land surface segmentation: The theoretical basis of terrain analysis and geomorphological mapping. *Geomorphology*, *95*(3), 236–259. https://doi.org/10.1016/j.geomorph.2007.06.003
Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017
Moseley, B., Bickel, V., López-Francos, I. G., & Rana, L. (2021). Extreme low-light environment-driven image denoising over permanently shadowed lunar regions with a physical noise model. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6313–6323. https://doi.org/10.1109/CVPR46437.2021.00625
Na, J., Ding, H., Zhao, W., Liu, K., Tang, G., & Pfeifer, N. (2021). Object-based large-scale terrain classification combined with segmentation optimization and terrain features: A case study in China. *Transactions in GIS*, *25*(6), 2939–2962. https://doi.org/10.1111/tgis.12795
Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. *Soil*, *5*(1), 79–89. https://doi.org/10.5194/soil-5-79-2019
Pavlopoulos, K., Evelpidou, N., & Vassilopoulos, A. (2009). *Mapping geomorphological environments*. Springer Science & Business Media.
Pedersen, G. (2016). Semi-automatic classification of glaciovolcanic landforms: An object-based mapping approach based on geomorphometry. *Journal of Volcanology and Geothermal Research*, *311*, 29–40. https://doi.org/10.1016/j.jvolgeores.2015.12.015
Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., et al. (2020). Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, *12*(6), 1005. https://doi.org/10.3390/rs12061005
Prakash, N., Manconi, A., & Loew, S. (2020). Mapping landslides on EO data: Performance of deep learning models vs. Traditional machine learning models. *Remote Sensing*, *12*(3), 346. Article 3. https://doi.org/10.3390/rs12030346
PyTorch. (2023). Retrieved from https://www.pytorch.org
Quattrochi, D. A., & Goodchild, M. F. (1997). *Scale in remote sensing and GIS*. CRC press.
Raab, T., Raab, A., Bonhage, A., Schneider, A., Hirsch, F., Birkhofer, K., et al. (2022). Do small landforms have large effects? A review on the legacies of pre-industrial charcoal burning. *Geomorphology*, *413*, 108332. https://doi.org/10.1016/j.geomorph.2022.108332
R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Robson, B. A., Bolch, T., MacDonell, S., Hölbling, D., Rastner, P., & Schaffer, N. (2020). Automated detection of rock glaciers using deep learning and object-based image analysis. *Remote Sensing of Environment*, *250*, 112033. https://doi.org/10.1016/j.rse.2020.112033
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.
Saha, K., Wells, N. A., & Munro-Stasiuk, M. (2011). An object-oriented approach to automated landform mapping: A case study of drumlins. *Computers & Geosciences*, *37*(9), 1324–1336. https://doi.org/10.1016/j.cageo.2011.04.001
Saha, S., Sarkar, R., Thapa, G., & Roy, J. (2021). Modeling gully erosion susceptibility in Phuentsholing, Bhutan using deep learning and basic machine learning algorithms. *Environmental Earth Sciences*, *80*(8), 295. https://doi.org/10.1007/s12665-021-09599-2
Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., et al. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, *46*(1), e1–e36. https://doi.org/10.1002/mp.13264

Salas, E., & Argialas, D. (2022). Automatic identification of marine geomorphologic features using convolutional neural networks in seafloor digital elevation models: Segmentation of DEM for marine geomorphologic feature mapping with deep learning algorithms. *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, *1–8*.

Schönfeldt, E., Winocur, D., Pánek, T., & Korup, O. (2022). Deep learning reveals one of Earth's largest landslide terrain in Patagonia. *Earth and Planetary Science Letters*, *593*, 117642. https://doi.org/10.1016/j.epsl.2022.117642

Sheppard, E., & McMaster, R. B. (2004). Introduction: Scale and geographic inquiry. In E. Sheppard & R. M. McMaster (Eds.) *Scale and geographic inquiry*. (pp. 1–22). Blackwell.

Slope (Spatial Analyst)—ArcGIS Pro | Documentation. (2023). Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/slope.htm

Smith, M. J., Paron, P., & Griffiths, J. S. (2011). *Geomorphological mapping: Methods and applications* (Vol. 15). Elsevier.

Sofia, G., Fontana, G. D., & Tarolli, P. (2014). High-resolution topography and anthropogenic feature extraction: Testing geomorphometric parameters in floodplains. *Hydrological Processes*, *28*(4), 2046–2061. https://doi.org/10.1002/hyp.9727

Sofia, G., Hillier, J. K., & Conway, S. J. (2016). Frontiers in geomorphometry and Earth surface dynamics: Possibilities, limitations and perspectives. *Earth Surface Dynamics*, *4*(3), 721–725. https://doi.org/10.5194/esurf-4-721-2016

Stehman, S. V. (2013). Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, *132*, 202–211. https://doi.org/10.1016/j.rse.2013.01.016

Stehman, S. V. (2014). Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing*, *35*(13), 4923–4939. https://doi.org/10.1080/01431161.2014.930207

Stumpf, A., & Kerle, N. (2011). Object-oriented mapping of landslides using random forests. *Remote Sensing of Environment*, *115*(10), 2564–2577. https://doi.org/10.1016/j.rse.2011.05.013

Sugarbaker, L., Constance, E. W., Heidemann, H. K., Jason, A. L., Lucas, V., Saghy, D., & Stoker, J. M. (2014). *The 3D elevation program initiative: A call for action*. US Geological Survey Reston.

Suh, J. W., Anderson, E., Ouimet, W., Johnson, K. M., & Witharana, C. (2021). Mapping relict charcoal hearths in New England using deep convolutional neural networks and LiDAR data. *Remote Sensing*, *13*(22), 4630. Article 22. https://doi.org/10.3390/rs13224630

Surficial materials of Massachusetts—A 1:24,000-scale geologic map database. (2018). *Scientific investigations map (No. 3402)*. U.S. Geological Survey. https://doi.org/10.3133/sim3402

Tarolli, P. (2014). High-resolution topography for understanding Earth surface processes: Opportunities and challenges. *Geomorphology*, *216*, 295–312. https://doi.org/10.1016/j.geomorph.2014.03.008

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. ahead-of-print(ahead-of-print). https://doi.org/10.1016/j.aci.2018.08.003

Thi Ngo, P. T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., & Lee, S. (2021). Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geoscience Frontiers*, *12*(2), 505–519. https://doi.org/10.1016/j.gsf.2020.06.013

Trier, Ø. D., Cowley, D. C., & Waldeland, A. U. (2019). Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeological Prospection*, *26*(2), 165–175. https://doi.org/10.1002/arp.1731

Trier, Ø. D., Zortea, M., & Tonning, C. (2015). Automatic detection of mound structures in airborne laser scanning data. *Journal of Archaeological Science: Report*, *2*, 69–79. https://doi.org/10.1016/j.jasrep.2015.01.005

Tucker, G. E., & Hancock, G. R. (2010). Modelling landscape evolution. *Earth Surface Processes and Landforms*, *35*(1), 28–50. https://doi.org/10.1002/esp.1952

van der Meij, W. M., Meijles, E. W., Marcos, D., Harkema, T. T., Candel, J. H., & Maas, G. J. (2022). Comparing geomorphological maps made manually and by deep learning. *Earth Surface Processes and Landforms*, *47*(4), 1089–1107. https://doi.org/10.1002/esp.5305

Verhagen, P., & Drăguţ, L. (2012). Object-based landform delineation and classification from DEMs for archaeological predictive mapping. *Journal of Archaeological Science*, *39*(3), 698–703. https://doi.org/10.1016/j.jas.2011.11.001

Verstappen, H. T. (2011). Chapter two—Old and new trends in geomorphological and landform mapping. In M. J. Smith, P. Paron, & J. S. Griffiths (Eds.) *Developments in Earth surface processes* (Vol. 15, pp. 13–38). Elsevier. https://doi.org/10.1016/B978-0-444-53446-0.00002-1

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). *Deep learning for computer vision: A brief review*. Computational Intelligence and Neuroscience, 2018.

Wadoux, A. M. J.-C. (2019). Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma*, *351*, 59–70. https://doi.org/10.1016/j.geoderma.2019.05.012

Warner, T. (2011). Kernel-based texture in remote sensing image classification. *Geography Compass*, *5*(10), 781–798. https://doi.org/10.1111/j.1749-8198.2011.00451.x

Welcome to Python.org. (2023). Python.Org. Retrieved from https://www.python.org/

Wilson, J. P., & Gallant, J. C. (2000). *Terrain analysis: Principles and applications*. John Wiley & Sons.

Xie, Z., Haritashya, U. K., Asari, V. K., Young, B. W., Bishop, M. P., & Kargel, J. S. (2020). GlacierNet: A deep-learning approach for debris-covered glacier mapping. *IEEE Access*, *8*, 83495–83510. https://doi.org/10.1109/ACCESS.2020.2991187

Xu, Y., Zhu, H., Hu, C., Liu, H., & Cheng, Y. (2021). Deep learning of DEM image texture for landform classification in the Shandong area, China. *Frontiers of Earth Science*, *16*(2), 352–367. https://doi.org/10.1007/s11707-021-0884-y

Yang, X., Na, J., Tang, G., Wang, T., & Zhu, A. (2019). Bank gully extraction from DEMs utilizing the geomorphologic features of a loess hilly area in China. *Frontiers of Earth Science*, *13*(1), 151–168. https://doi.org/10.1007/s11707-018-0700-5

Yu, S., & Ma, J. (2021). Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, *59*(3), e2021RG000742. https://doi.org/10.1029/2021rg000742

Zhang, J., Zhao, X., Chen, Z., & Lu, Z. (2019). A review of deep learning-based semantic segmentation for point cloud. *IEEE Access*, *7*, 179118–179133. https://doi.org/10.1109/access.2019.2958671

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 22–40. https://doi.org/10.1109/MGRS.2016.2540798

Zhang, W., Liljedahl, A. K., Kanevskiy, M., Epstein, H. E., Jones, B. M., Jorgenson, M. T., & Kent, K. (2020). Transferability of the deep learning mask R-CNN model for automated mapping of ice-wedge polygons in high-resolution satellite and UAV images. *Remote Sensing*, *12*(7), 1085. Article 7. https://doi.org/10.3390/rs12071085

Zhang, W., Witharana, C., Liljedahl, A. K., & Kanevskiy, M. (2018). Deep convolutional neural networks for automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery. *Remote Sensing*, *10*(9), 1487. Article 9. https://doi.org/10.3390/rs10091487

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867. https://doi.org/10.1109/tmi.2019.2959609

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307