A Statistical Model for Evaluating Landslide Susceptibility in Northwest Colorado

Justin H. Cai

December 10, 2022

Abstract Landslides in northwestern Colorado represent a significant risk to both people and infrastructure, especially in light of Colorado's rapidly growing population and development in tourism, urban sprawl, and infrastructure. However, there does not yet exist a robust and accurate model that maps landslide susceptibility across northwestern Colorado. This study aims to construct such a model by using twelve different landslide conditioning factors commonly used in landslide susceptibility literature as highly influential in predicting landslide occurrence, using a landslide inventory prepared by the Colorado Geological Survey, and testing several modern classification models before applying the best performing model to a full-scale spatial classification of landslide susceptibility in northern Colorado. Model performance was evaluated using cross-validation and confusion matrices as well as visual observation of the final spatial classification products. The results demonstrate that random forest is the most effective model for classifying and predicting landslide susceptibility in northern Colorado. The maps produced by the spatial random forest classification model can be used for urban planners and natural resource managers in informing sustainable development and infrastructure.

1 Introduction

Landslides include several forms of mass movement of rocks and soil down a slope under the influence of gravity, including rockfalls, mudflows, and debris flows. Landslides are the result of downward driving force exceeding resistance forces due to soil and rock destabilization, and can occur on a slope of any gradient and in any kind of environment, can move very quickly, and transport several thousands of cubic feet of material. Often, landslides are due to certain events such as rainfall, earthquakes, or human activity; landslides can also be due to a non-identifiable source or seemingly at random.

Understanding how and when landslides occur is of paramount importance due to their unpredictable nature and significant capacity to cause infrastructural damage. In the U.S. state of Colorado, known for its high mountains, steep slopes, rugged wilderness, and inclement weather, landslides present an especially high risk to people and infrastructure. In 2010, nine deaths and four injuries, as well as over $9 million in

damages, have been documented across Colorado (8), with unreported casualties and indirect losses making the true figures likely much larger. With Colorado's rapidly expanding population, tallying 5,773,714 residents in 2020, up from 5,029,196 in 2010 (**?**), expansion of urban developments into the undeveloped slopes of the Rocky Mountains, large year-round tourism industry ($21.9 billion in 2021 with some 84.2 million trips) (**?**), and likewise expanding road and rail infrastructure, landslides continue to pose an increasingly large risk to people and infrastructure.

Yet, despite these risks, there does not yet exist a detailed model for mapping and accurately predicting landslide susceptibility across Colorado. While the Colorado Geological Survey has mapped landslide locations and past debris flow activity to create a detailed inventory from all available U.S. Geological Survey maps (**?**), there currently is no rigorous way to assess the landslide susceptibility across unmapped areas of Colorado within the unique environmental, geomorphological, and spatial setting of the state. This paper seeks to address the question of whether it is possible to construct a model that accurately predicts and maps the susceptibility of regions in Northwestern Colorado to landslides, and if so, what would constitute such a model. A robust landslide susceptibility model of northwestern Colorado would significantly enhance disaster risk assessments and inform landslide prevention efforts, sustainable land use practices, and proper infrastructure management.

Landslide susceptibility is defined as the likelihood of a landslide occurring based on local terrain conditions (3) and attempts to predict where landslides are most likely to occur (7). Landslide susceptibility modelling has been an active field of research since the mid-1970s and is primarily divided into two categories: deterministic modelling and statistical modelling. Deterministic modelling employs purely mathematical relationships based on the physical laws governing the interactions between driving and resistance forces during debris flows. However, these models require extremely detailed data on the geometry of landslide-prone slopes and geologic foundations, and thus are suitable only for small, highly-studied regions (1). Statistical and machine learning models, in contrast, can be applied to much wider regions as long as data on landslide controlling factors are available. Random forest, logistic regression, neural networks, weighted linear combination, and regression trees have all been used to assess landslide susceptibility in regions around the world (2), (14), (9), (12), (11), (5), (16). This paper studies a variety of statistical models for landslide susceptibility in northwestern Colorado.

Choosing the factors used to predict landslide susceptibility is critical to building an effective model. This study uses geomorphological features (elevation, slope, slope angle, profile curvature, plan curvature, rock lithology, soil type, distance to faults), environmental features (NDVI, distance to water) and anthropogenic features (distance to roads, land cover). These conditioning factors are generally regarded in the literature as highly important for determining landslide susceptibility and have been frequently used in past landslide susceptibility models (16), (10), (15), (13). The definitions of these features and their sources and calculation are detailed in Data.

The study area is displayed in figure 1. This region constitutes most of northwestern Colorado, including most of the Front Range of the Rocky Mountains, and encompasses most of the rugged terrain and wilder-

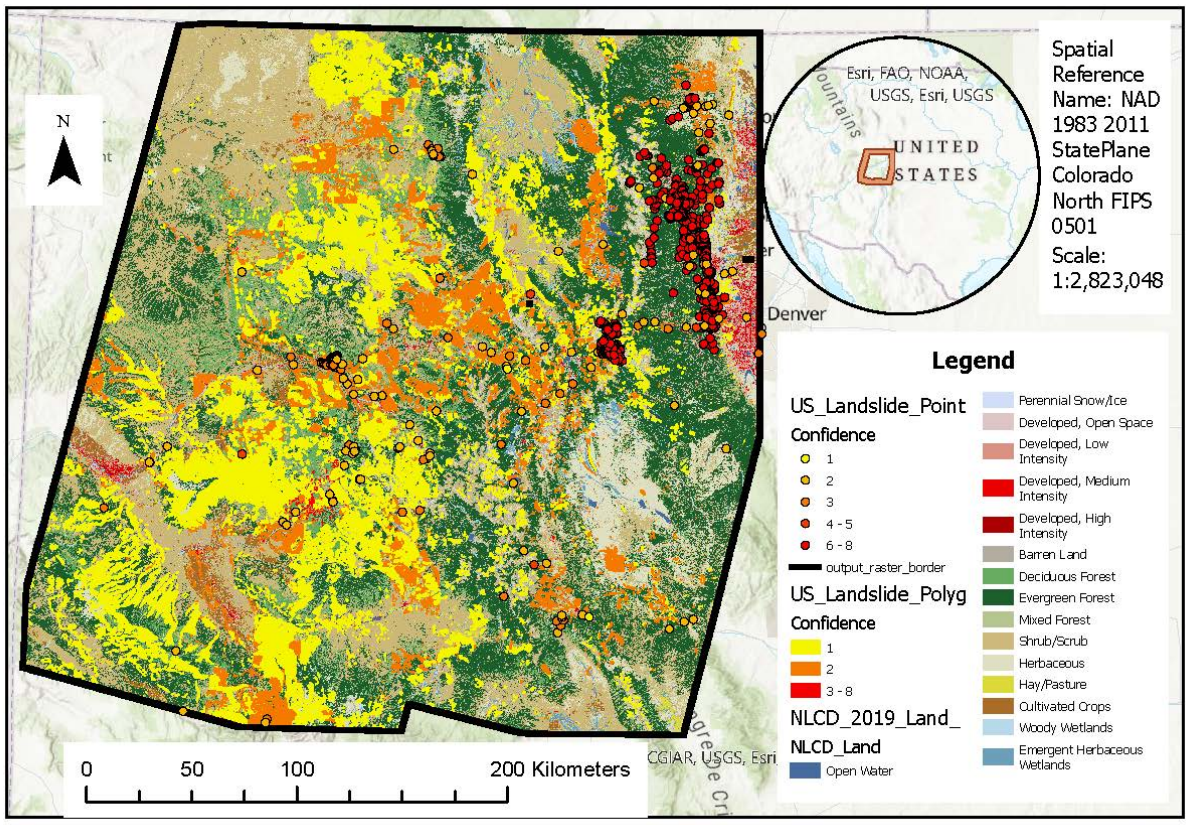

Figure 1: The study area and landslide inventory polygons and points. land cover is displayed as the baselayer, and landslide polygons and points are laid above it, with all three layers symbolized as described in the legend. An inset provides the location of the study area within the greater United States, and the spatial reference, map scale, and a scale bar are shown.

ness between Grand Junction, CO and Denver, CO. The region is approximately 200 by 200 miles wide. This study area was chosen due to its diverse and rugged terrain, concentration of scattered towns, roads, resorts, and parks, and growing tourism and residential development.

2 Data

This study uses a suite of remotely sensed and survey-based data sources to map factors that are likely to strongly affect landslide susceptibility in northwestern Colorado. A brief technical specification of each dataset and their transformation into usable data products is given below, along with maps of the data products produced from these sources after preprocessing. The preprocessing of these sources into the products used for modelling are detailed in Methods, and the motivation for the use of each data product is detailed in Introduction.

2.1 Landsat 8

Landsat 8 is an American Earth observation satellite launched in 2013 as part of the NASA/USGS Landsat program that provides medium-resolution satellite imagery in 12 spectral bands. It uses a sun-synchronous orbit at an altitude of 705 km, with an orbital period of 99 minutes, and acquires about 740 images a day on the Worldwide Reference System-2 system. This study uses bands 4 and 5 of Landsat 8's Operational Land Imager, comprising red visible light at $0.63 - 0.67$ μm wavelength and near-infrared light at $0.85 - 0.88$ μm wavelength, both at 30m resolution acquired on 10/19/2022.

The area of study was covered by four scenes (images), which were mosaicked (combined) together using ArcGIS Pro's Mosaic to New Raster tool. The Normalized Difference Vegetation Index (NDVI) was calculated using bands 4 and 5. NDVI is given by

$$\frac{NIR - Red}{NIR - Red}$$

where NIR represents near-infrared pixel values and Red represents red pixel values. NDVI was calculated at each pixel using ArcGIS Pro's Raster Calculator tool. The result is displayed in subfigure (g) of figure 3. NDVI measures the absorbance of visible light and reflectance of near-infrared light, which serves as the primary proxy for measuring photosynthetic potential, and thus the health and density of vegetation cover. Larger positive values of NDVI indicate dense, healthier vegetation growth.

2.2 USGS TNM 1/3 Arc-second DEM

Digital Elevation Models (DEM) at 1/3 arc-second resolution were derived from the U.S. Geological Survey 3D Elevation Program (3DEP) for The National Map (TNM). The data are derived from Lidar point cloud data and interferometric synthetic aperture radar (IfSAR) data, are seamless, and are referenced to the North American Vertical Datum of 1988 (NAVD 1988).

The area of study was covered by twelve scenes, which were mosaicked together using the Mosaic to New Raster tool. The output is displayed in subfigure (i) of figure 3. From the DEM, slope angle (subfigure (h) of figure 3), slope aspect (subfigure (d) of figure 2), profile curvature (subfigure (e) of figure 2), and plan curvature (subfigure (f) of figure 2) were calculated using the Slope, Aspect, and Curvature tools.

Slope angle represents the angle of slope from the horizontal plane of each pixel and is given in degrees ranging from 0° (flat) to 90° (vertical). Slope aspect is defined as the angle with respect to north at which each slope face is facing in degrees, ranging from 0° (true north) to 360° (true north again). Profile curvature is the curvature parallel to the slope and indicates the direction of maximum slope, which affects the acceleration and deceleration of flow across the surface. Planform (plan) curvature is the curvature perpendicular to the slope and dictates the convergence and divergence of flow on a slope.

2.3 SSURGO

The Soil Survey Geographic Database (SSURGO) is a collection of soil data as collected by the National Cooperative Soil Survey. The data were collected by walking over the land and observing the soil as well as laboratory soil sample analysis. The data were collected at scales ranging from 1:12000 to 1:63360.

This study uses data collected in the river basins corresponding to the study area. The soil datasets from the individual river basins were combined into a new vector dataset, which was then rasterized. The result is displayed in subfigure (c) of figure 2. This dataset is the only dataset that did not contain data for the entire region.

2.4 TIGER/Line Shapefiles

Topologically Integrated Geographic Encoding and Referencing (TIGER/Line) shapefiles are produced by the U.S. Cenus Bureau to describe physical features such as roads, buildings, and rivers.

This study uses linear water (rivers/streams) and road shapefiles. The data were obtained via county Federal Information Processing System (FIPS) codes and joined together into linear water and road system layers covering the study area. The Euclidean Distance tool was used to produce a raster that constitutes the straight-line distance of each pixel to the nearest linear water or road feature. The results are displayed in subfigure (j) and subfigure (k) of figure 3.

2.5 MRLC NLCD 2019

The National Land Cover Database (NLCD) is a set of 34 vector-based different land cover products produced by the Multi-Resolution Land Characteristics Consortium (MRLC), a collaboration of U.S. federal agencies who coordinate and generate land cover information on a national scale. NLCD 2019 represents the most recent set of land cover products and includes land cover and land cover change from $2001 - 2019$ and represents the most comprehensive land cover database ever produced by the USGS. This dataset is displayed in subfigure (a) of 2.

2.6 USGS SGMC

The State Geologic Map Compilation (SGMC) geodatabase represents a seamless, vector-based spatial database of 48 state geological maps at scales ranging from 1:50000 to 1:1000000 scale. The SGMC is a compilation of individual U.S. Geological Survey releases of the Preliminary Integrated Geologic Map Databases for the United States. This study uses only SGMC data in Colorado, which eliminates data issues such as differing units across state boundaries.

The SGMC dataset contains data on bedrock geology, which is displayed in subfigure (b) of figure 2. The dataset also maps geological fault locations as feature lines. The Euclidean Distance tool was used to

produce a raster that constitutes the straight-line distance of each pixel to the nearest fault line. The result is displayed in subfigure (l) of figure 3.

2.7 U.S. Landslide Inventory

The U.S. Landslide Inventory contains landslide polygons digitized from a variety of USGS maps and points at observed landslide locations from multiple sources. This study uses landslide polygons digitized by the Colorado Geological Survey based on all landslides mapped by geological hazard maps in Colorado and includes digitized landslides from maps at 1:24000, 1:48000 to 1:100000, and 1:250000 scales, as well as point landslides identified by Coe et al. (2014) and Godt and Coe (1999). Landslide polygons represent regions of observed landsliding activity, while landslide points identify specific locations where landslides were observed.

Both the points and polygons contain a value named "Confidence" which represents the degree to which a debris flow occurred, ranging from 1 (possible landslide in this area) to 8 (high confidence in nature or extent of landslide). A larger value of "Confidence" indicates a larger more well-defined observed debris flow, and represents a proxy for the severity of a landslide. This field is the target variable upon which the models (see Methods) attempt to predict based on the previous data sources. The points and polygons are displayed in figure 4. These data were not rasterized.

2.8 EDA

The distribution of each landslide controlling variable and the Confidence field are displayed in figure 5. (OBJECTID refers to the unique identifier of each point and can be ignored.) Immediately, the target variable Confidence is unbalanced, having many more landslides classified in high confidence than those in low confidence. This imbalanced dataset leads to several methodological and model validation modifications described in Methodology and Results. In addition, while the graph is unable to display the labelled names for soil, geology, and land cover, it is clear that most landslides occur under a few certain types of these three categorical features. Both profile and plan curvature are centered around 0, implying that the degree of slope curvature is normally distributed, and high curvature (both positive and negative) is rare. Distance to water, faults, and roads possess a right-skewed distribution, indicating that landslides far from these three features are much rarer. Slope tends to be centered around $30°$, which is relatively steep but still far from vertical. In addition, most landslides seem to occur at (relatively) low elevations of around 2000 metres and at low positive NDVI values, indicating low amounts of vegetation.

The binary relationship between each landslide controlling factor and the Confidence field are displayed in figure 6. Landslides classified with high confidence and severity tend to have lower NDVI values and distance to faults and higher elevation and slope. Lack of vegetation and high slope degrees have been shown to be associated with higher landslide risk (13), agreeing with this exploratory data analysis. However, the relationship between the other controlling factors and the Confidence field is inconclusive.

Observing the mapped landslide polygons and points on a map (figures 1, 4), we observe a heavy concentration of point landslides classified with high confidence and severity near the Front Range of the Rocky Mountains near Boulder and Denver, CO. High confidence landslides tend to be observed and recorded in clusters due to most publications employing field surveying within a given region and time period after a disturbance (e.g. torrential rainfall) and less attention drawn to regions more remote and across longer time periods, exposing some landslide inventory bias and limitations of data. However, the extensive inventory of landslide polygons in the more remote parts of northwestern Colorado (away from Denver) somewhat remedies this problem; however, the amount of polygons classified as highly confident and severe are much lower, leading to an unbalanced dataset. Similarly to the landslide points, this imbalance is experimented with in Methods and Results.

3 Methods

This study's methodology aims to find a robust, highly accurate model for classifying landslide susceptibility by incorporating all landslide controlling factors and testing for the most important factors, testing various different classification models and spatial vs nonspatial classification paradigms, and tests compensating for sparse variables due to the imbalanced nature of the landslide inventory. For a flowchart summarizing the methodology, see figure 7.

3.1 Data Preprocessing

Because the raw data originate from a variety of sources, the data types, resolution, spatial reference systems, and collection methods were highly inconsistent. Resolution is defined as the land size of each pixel in a dataset; for instance, a 30-metre resolution satellite imagery product such as Landsat 8 means that each pixel in the image is of size 30 by 30 metres on the ground. Spatial reference refers to the coordinate projection that each data source is based in; because the Earth is a sphere, it is impossible to perfectly represent it on a 2D rectangular space like a map without deforming either shape or size. Map projections define a transformation geographical coordinates onto a plane depending on what spatial qualities are desired; for instance, the Mercator projection preserves the shape of landforms while deforming sizes, especially for regions far from the equator. Inconsistent map projections lead to improperly aligned map features, which leads to inaccurate spatial analysis.

For this study, each data product was downloaded from its source and mosaicked or joined as detailed in the Data section such that each natural factor is contained in one layer, and each vector layer was rasterized such that all data were in a consistent file type. Each data product was then masked out using the study area boundaries, as any data outside the study area is not needed.

The data were then all projected into one coordinate system: NAD 1983 (2011) StatePlane Colorado North FIPS 0501 (Meters). This is a Lambert conformal conic projection designed for use in northern Colorado, where deformation in both shape and size are minimized.

The data were converted to 32-metre resolution using the Resample tool, where categorical non-continuous datasets like land cover and bedrock geology were interpolated using the nearest neighbor algorithm option in the Resample tool, and continuous features were interpolated using the bilinear algorithm option (excluding profile and plan curvature, which were interpolated with the cubic algorithm option). Slope angle, aspect, NDVI, elevation, and Euclidean distances are linear functions and thus better represented by a bilinear interpolation algorithm, while curvature tends to be nonlinear and better represented by a cubic algorithm.

3.2 Model Construction

The construction of models was separated into two categories: one corresponding to the landslide points, and one corresponding to the landslide polygons. Each feature dataset lent itself to different kinds of model building, application, and analysis, and are detailed in the following subsections. The results from the nonspatial classification of landslide points were used to determine which model was to be used in the spatial classification of landslide susceptibility across all of the study region.

3.2.1 Landslide Points

The specific landslide factor values (slope, elevation, land cover, etc.) at each of the 2223 points representing an observed/mapped landslide were extracted, creating a table listing the values of each of these factors at each observed landslide point. This table formed the basis upon which a classification model can be trained to predict the Confidence field of the landslide points based on each of the values of the landslide factors. This transforms the spatial problem into a nonspatial standard multiclass classification problem, upon which most classification models could be applied and tested.

From this table, points that did not have soil type data were masked out, to ensure that all points had complete data on landslide controlling factors. The table was separated into columns for the 12 landslide controlling factors (Geology, Soil, Aspect, NDVI, Road Distance, Fault Distance, Water Distance, Land Cover, Slope, Aspect, Elevation, Plan Curvature, Profile Curvature) and the variable to be predicted (Confidence). The dataset was split into 80% training data and 20% testing data. Then, the following models were tested with the given parameters.

Random Forest Estimators: 500, maximum depth: 10, minimum samples per leaf: 5.

Gaussian Naive Bayes All default parameters.

Support Vector Machine Kernel: Radial basis function.

Logistic regression Penalty: l2, solver: lbfgs.

Multi-layer perceptron Activation: ReLU, solver: lbfgs, learning rate: constant.

Based on the results of this nonspatial classification (see Results), random forest was determined to be the most accurate and robust model for mapping landslide susceptibility. Because classifying landslide susceptibility across the entire region of northwest Colorado is computationally expensive and requires classifying each individual pixel, only one model could be applied towards spatial classification.

The application of random forest towards a spatial classification was first performed using all 12 landslide controlling factors. Then, based on the calculated feature importance from this classification (see figure 11), another spatial classification with random forest was performed using only the 6 features with highest importance. Each classification was performed with 500 trees, a maximum depth of 10, and 5 minimum samples per leaf to provide a high capacity for learning ability and avoiding overfitting while minimizing the computational resources needed to perform the classification.

3.2.2 Landslide Polygons

Unlike landslide points, landslide polygons inherently have a spatial component and cannot be dissolved down into a nonspatial classification problem. Based on the results of the nonspatial classification with points, the random forest model was chosen again to perform spatial classifiation of landslide susceptibility with landslide polygons. First, the classification was performed using all 12 controlling features, once with compensating for sparse datasets, and once without. Due to the imbalanced nature of the landslide inventory and the Confidence target variable, correcting for sparseness by forcing each feature and unique value of target variable in each tree was tested as a possible remediation to the issue. Based on the calculated feature importance (see figure 10), another spatial classification was performed using only the 6 features with high importance. The same parameters used for spatial classification based on landslide points were used for the classifications here.

4 Results

4.1 Landslide points

The nonspatial classification component of landslide points have cross-validation accuracies and standard deviations as displayed below, along with confusion matrices in figure 9.

4.1.1 Random forest

Cross-validation accuracy: 0.845
Cross-validation standard deviation: 0.047

4.1.2 Gaussian Naive Bayes

Cross-validation accuracy: 0.682
Cross-validation standard deviation: 0.18

4.1.3 Support vector machine

Cross-validation accuracy: 0.804
Cross-validation standard deviation: 0.002

4.1.4 Logistic regression

Cross-validation accuracy: 0.801
Cross-validation standard deviation: 0.057

4.1.5 Multi-layer perceptron

Cross-validation accuracy: 0.802
Cross-validation standard deviation: 0.001

The nonspatial point-based classification using the five models produced mostly accurate results, with random forest producing the consistently highest cross-validation accuracy and minimal number of misclassified features as seen in its confusion matrix in figure 9. However, support vector machines, logistic regression, and multi-layer perceptrons all had extremely similar cross-validation accuracies not far behind that of random forest, making them reasonably suitable algorithms for classification as well. These three algorithms were more likely to misclassify points with high landslide confidence. Gaussian Naive Bayes performed the worst, with a cross-validation accuracy lower than the other four models.

The result of the spatial component of landslide point classification is shown in subfigures (d) and (e) of figure 8. Each point is coloured blue if it was classified correctly and coloured orange if it was not. The landslide susceptibility classification using only important features is much more grainy in texture. Additionally, the classification seems to be influenced by the heavy concentration of high confidence landslides on the Front Range of the Rocky Mountains.

4.2 Landslide polygons

The result of random forest classification using landslide polygons is displayed in subfigures (a), (b), and (c) of figure 8. Each polygon is coloured blue if it was classified correctly and coloured orange if it was not. The classification that corrected for sparseness of the target Confidence variable classified classified

much more land at the highest confidence value than did the classification without correcting for sparseness. The polygon landslide classification with only important features, like the point landslide classification, is highly grainy, and classified even more points with the highest confidence level than the classification that corrected for sparseness.

5 Discussion

These results indicate promising new models to model landslide susceptibility in Colorado. The reasonably high accuracy of random forest-based classification of severity of landslides, along with its small cross-validation score, indicate both an accurate and robust model for predicting landslide susceptibility across northwestern Colorado. In addition, while a bit less accurate, the support vector machine, logistic regression, and multi-layer perceptron models are not far behind in accuracy.

However, the accuracy of these models was likely inhibited by the limitations of the landslide inventory. Because observed landslides are highly concentrated in places where past studies have mapped landslides in small regions due to identifiable events, the landslide inventory tends to be highly imbalanced, which can make a landslide susceptibility classification across the entire region of northwest Colorado inaccurate.

In addition, a researcher with more time and computational resources could attempt applying more machine learning and deep learning models towards spatial classification, as this study was only able to do so for random forest. Of particular interest could be support vector machines, logistic regressions, and deep learning methods. The variation of parameters in each model, and the optimization of such parameters to construct an accurate and robust model, is also an avenue for further research.

While this study successfully prepared a reasonably accurate and robust model for performing landslide susceptibility classification, there is still yet much work to do in terms of understanding which landslide conditioning factors are most important from both a physics-based perspective and a model feature importance perspective. This study does not examine the physical dynamics behind the causes of landslides or what caused certain features to be more important than others; both could be pathways for further research.

In addition, the variation between the spatial classifications of landslide susceptibility are objects of further study. Point landslide-based classification classified much more land as high confidence and severity landslides than did polygon-based classification models. While this is likely due to the much higher concentration of high confidence and severe landslides in the point dataset, it remains a question if so, and how, the imbalance can be remedied, and why correction for sparse data categories did not fully account for this discrepancy.

While the spatial classification maps produced by the differing variations of the random forest model have significant differences, these maps serve as a starting point for urban planners and natural resource managers to identify which regions of northwestern Colorado are most prone to landslide activity, and adjust sustainable design and infrastructure planning accordingly.

6 Acknowledgments

I would like to thank Jonathan Hanke of CSML for being the instructor for SML312, teaching me the tools and practices used in data science, and offering extensive guidance on my project. I would like to thank Daniel Melese for enhancing my learning of data science concepts and tools and also offering guidance on my project. I would like to thank Wangyal Shawa of the Princeton University Library for providing geospatial analysis-specific guidance and technical guidance on model implementation and debugging.

References

[1] Ayalew L, Yamagishi H (2005) The Application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda–Yahiko Mountains, central Japan. Geomorphology 65:15–31. doi:10.1016/j.geomorph.2004.06.010

[2] Ayalew L, Yamagishi H, Ugawa N (2004) Landslide susceptibility mapping using GISbased weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. Landslides 1(1):73–81. doi:10.1007/s10346-003-0006-9

[3] Brabb, E.E., 1984. Innovative approaches to landslide hazard mapping. In: Proc. 4th Int. Symp. Landslides, Toronto. 1. pp. 307–324.

[4] Coe, Jeffrey, Kean, Jason, Godt, Jonathan, Baum, Rex, Jones, Eric, Gochis, David, Anderson, Gregory. (2014). New insights into debris-flow hazards from an extraordinary event in the Colorado Front Range. GSA Today. 24. 4-10. 10.1130/GSATG214A.1.

[5] Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhital MR, Althuwaynee OF (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling-Narayanghat road section in Nepal Himalaya. Nat Hazards 65(1):135–165. doi:10.1007/s11069-012-0347-6

[6] Jonathan W. Godt, Jeffrey A. Coe, Alpine debris flows triggered by a 28 July 1999 thunderstorm in the central Front Range, Colorado, Geomorphology, Volume 84, Issues 1–2, 2007, Pages 80-97, ISSN 0169-555X, https://doi.org/10.1016/j.geomorph.2006.07.009.

[7] Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., Ardizzone, F., 2005. Probabilistic landslide hazard assessment at the basin scale. Geomorphology 72, 272–299. http://dx.doi.org/10.1016/j.geomorph.2005.06.002.

[8] Highland, L.M., 2012, Landslides in Colorado, USA—Impacts and loss estimation for 2010: U.S. Geological Survey Open-File Report 2012–1204, 49 p.

[9] Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides 4(1):33–41. doi:10.1007/ s10346-006-0047-y

[10] Meena, S. R., Puliero, S., Bhuyan, K., Floris, M., and Catani, F.: Assessing the importance of conditioning factor selection in landslide susceptibility for the province of Belluno (region of Veneto, northeastern Italy), Nat. Hazards Earth Syst. Sci., 22, 1395–1417, https://doi.org/10.5194/nhess-22-1395-2022, 2022

[11] Pradhan B, Lee S (2010) Regional landslide susceptibility analysis using back-propagation neural networks model at Cameron Highland, Malaysia. Landslides 7(1):13–30. doi:10.1007/s10346-009-0183-2

[12] Mathew J, Jha VK, Rawat GS (2009) Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. Landslides 6(1):17–26. doi:10.1007/s10346-008-0138-z

[13] Paola Reichenbach, Mauro Rossi, Bruce D. Malamud, Monika Mihir, Fausto Guzzetti, A review of statistically-based landslide susceptibility models, Earth-Science Reviews, Volume 180, 2018, Pages 60-91, ISSN 0012-8252, https://doi.org/10.1016/j.earscirev.2018.03.001.

[14] Remondo J, Bonachea J, Cendrero A (2005) A statistical approach to landslide risk modelling at basin scale: from landslide susceptibility to quantitative risk assessment. Landslides 2(4):321–328. doi:10.1007/s10346-005-0016-x Ridgeway G (2006) Generalized boosted regression models.

[15] Wubalem, A. Landslide susceptibility mapping using statistical methods in Uatzau catchment area, northwestern Ethiopia. Geoenviron Disasters 8, 1 (2021). https://doi.org/10.1186/s40677-020-00170-y

[16] Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S. et al. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13, 839–856 (2016). https://doi.org/10.1007/s10346-015-0614-1

7 Honor Statement

This paper represents my own work in accordance with University regulations.
Justin Cai

8 Appendix

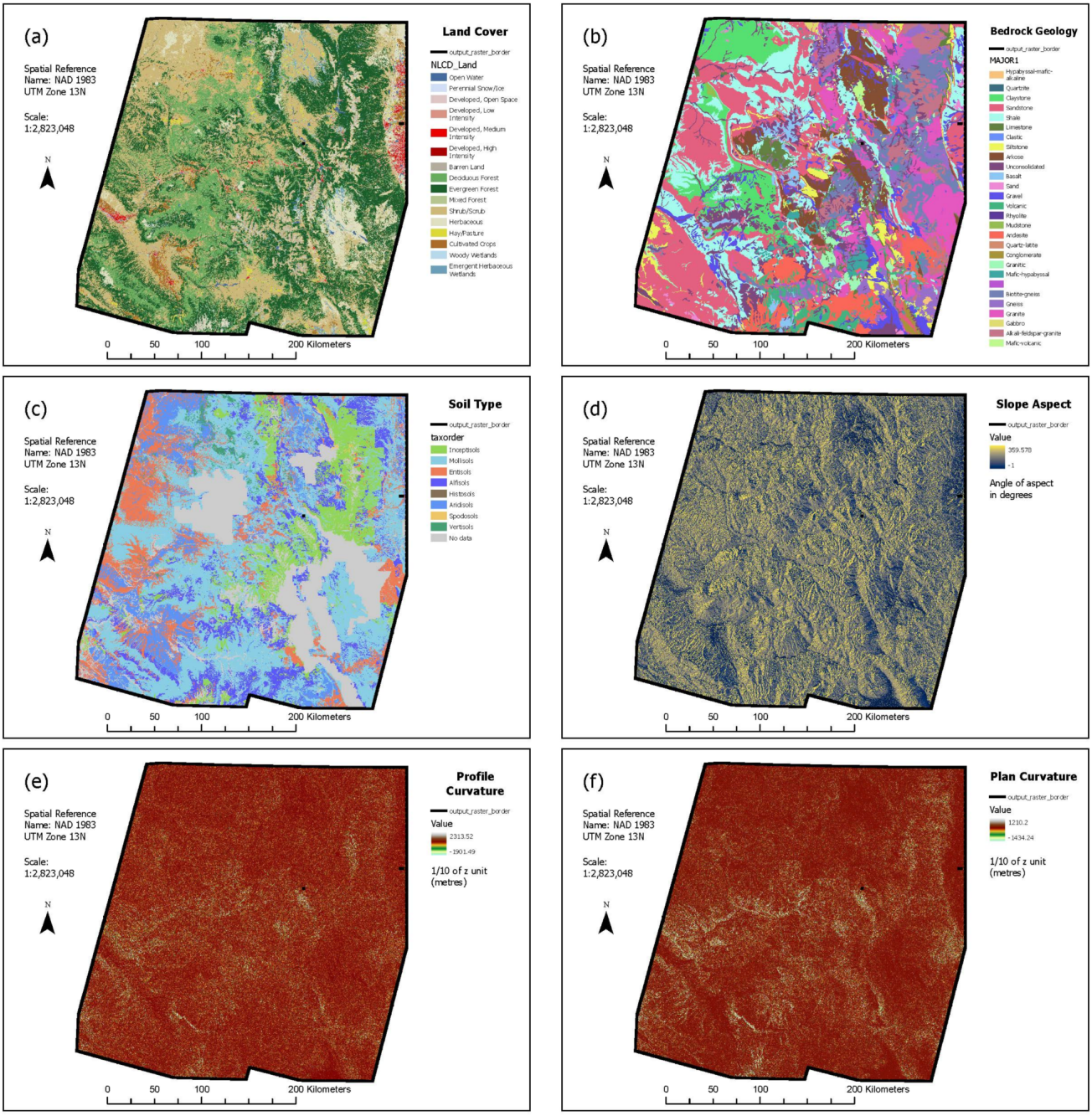

Figure 2: Maps of landslide controlling factors. (a): land cover; (b): bedrock geology; (c): soil type; (d): slope aspect; (e): profile curvature; (f): plan curvature. All maps were constructed in the same spatial reference system and map scale.

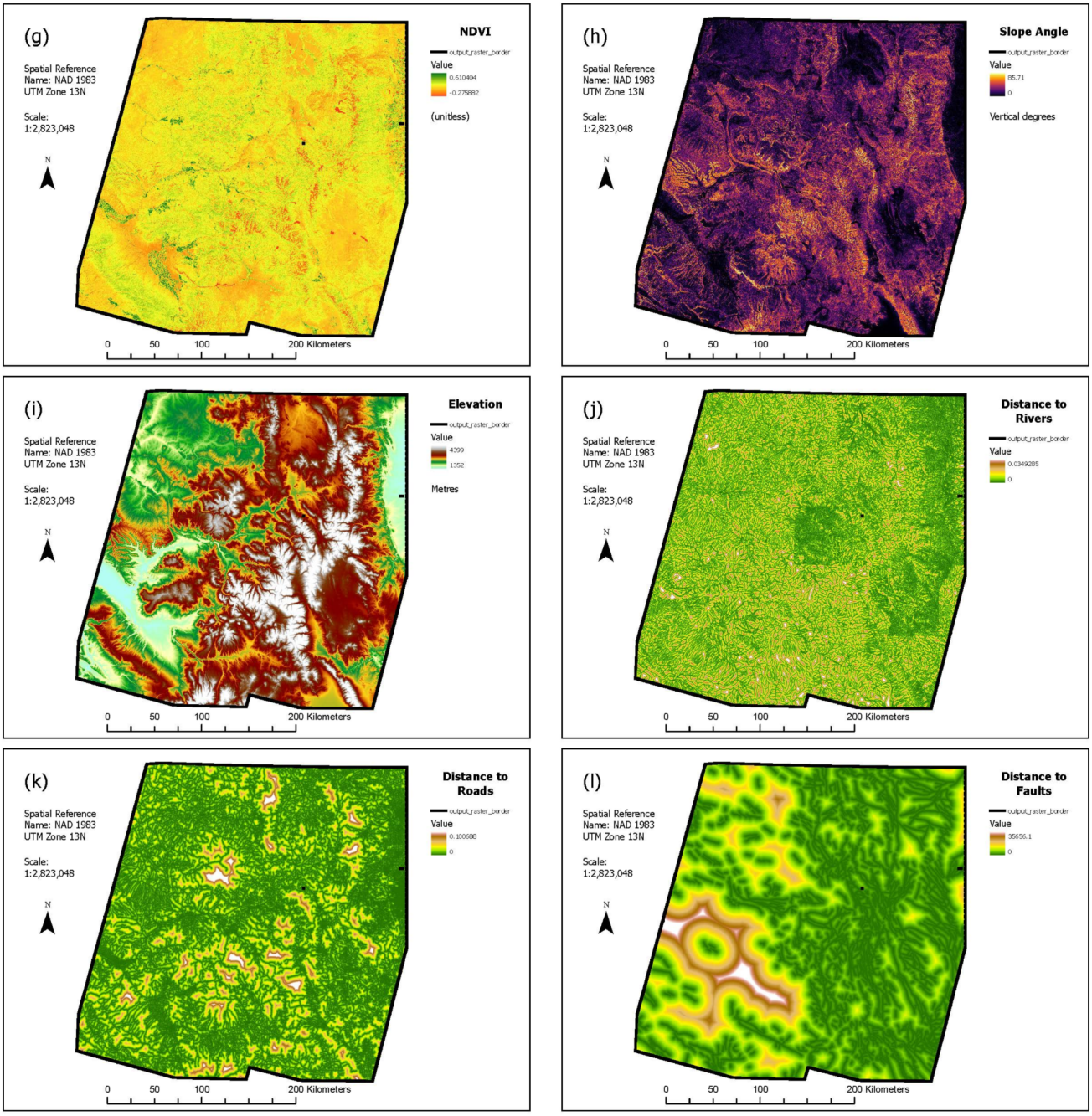

Figure 3: Maps of landslide controlling factors. (g): NDVI; (h): slope angle; (i): elevation; (j): distance to rivers; (k): distance to roads; (l): distance to faults. All maps were constructed in the same spatial reference system and map scale.

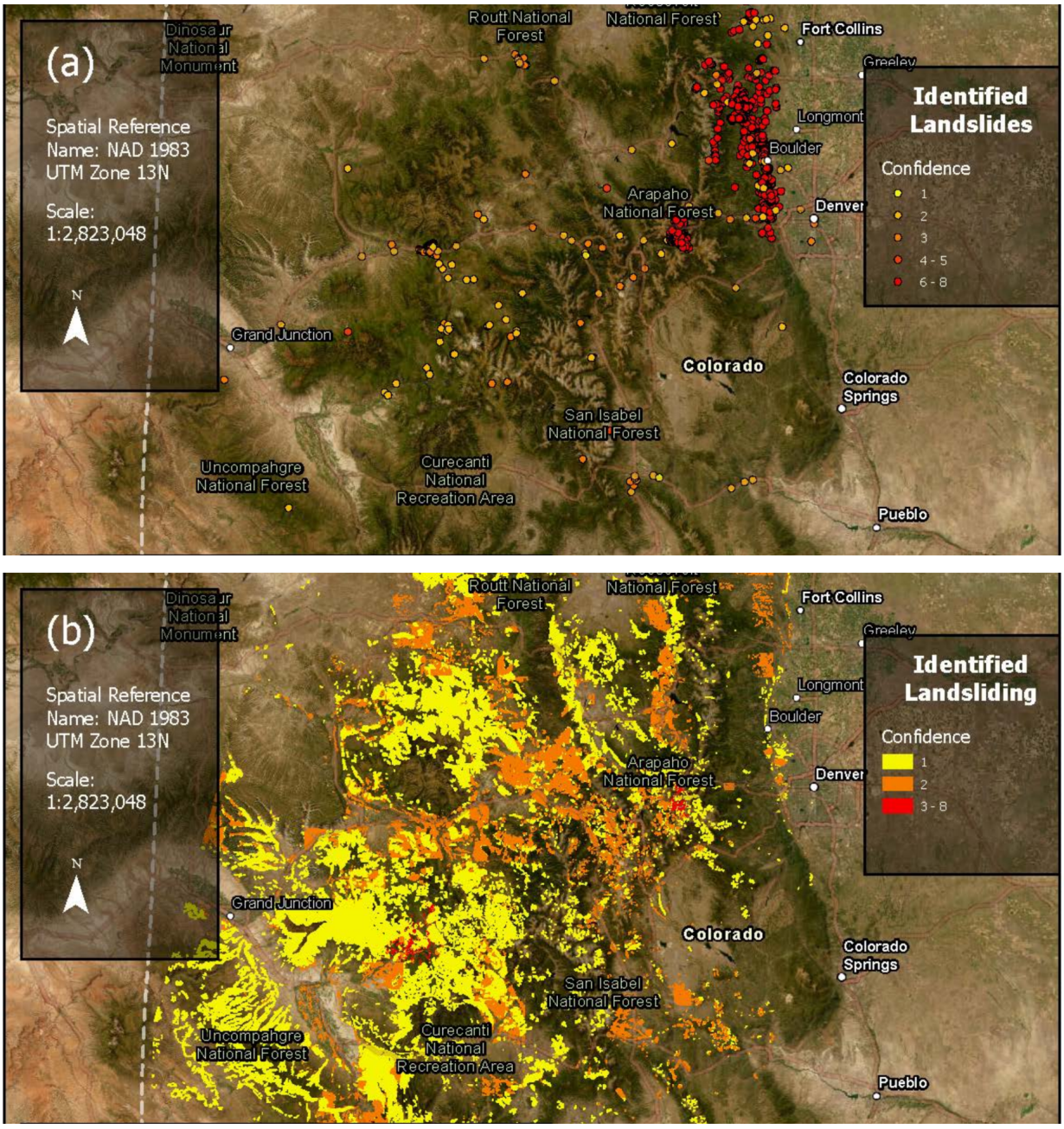

Figure 4: Map of landslide points and polygons on a true-color satellite image of Northern Colorado. (a): landslide points; (b): landslide polygons. The points and polygons are symbolized by their confidence values.

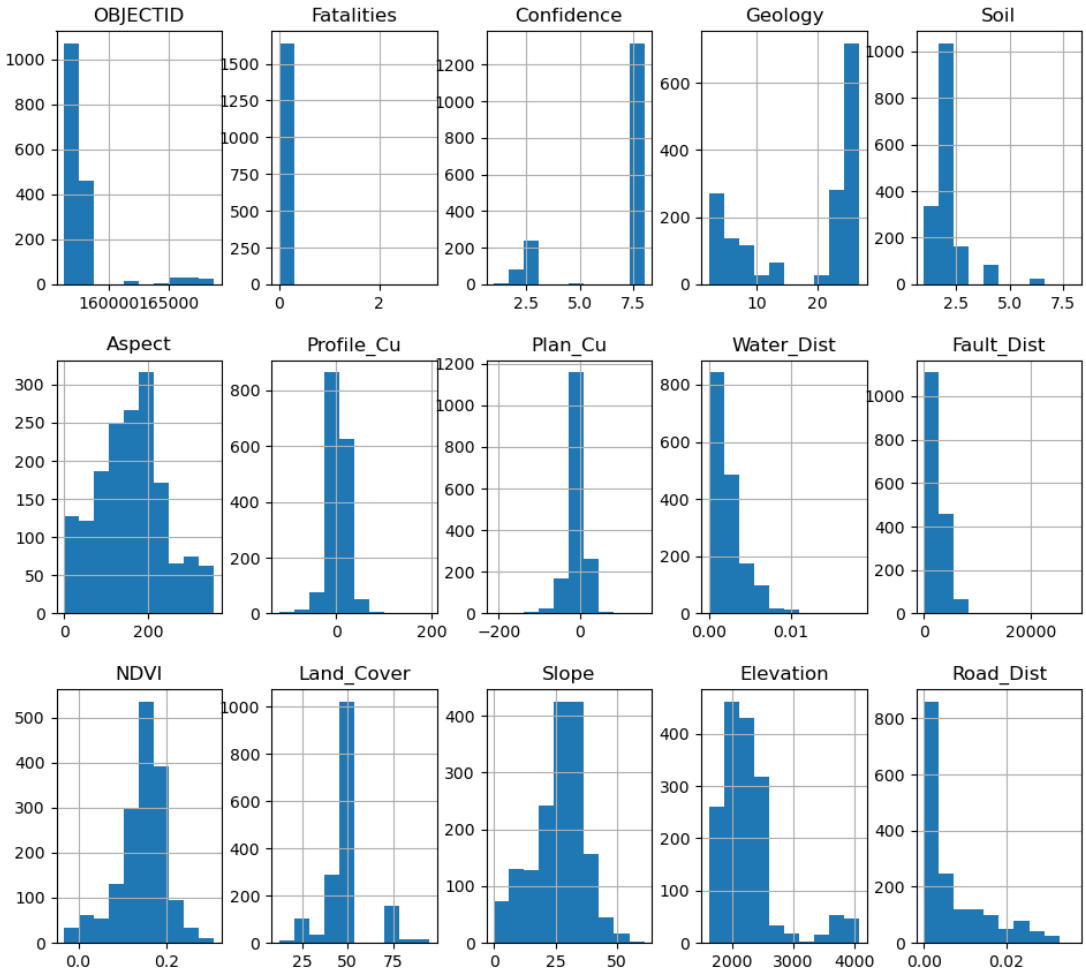

Figure 5: Histogram depicting frequency of the target variable Confidence and each of the landslide controlling factors. OBJECTID represents the unique identifier of each point and can be ignored.

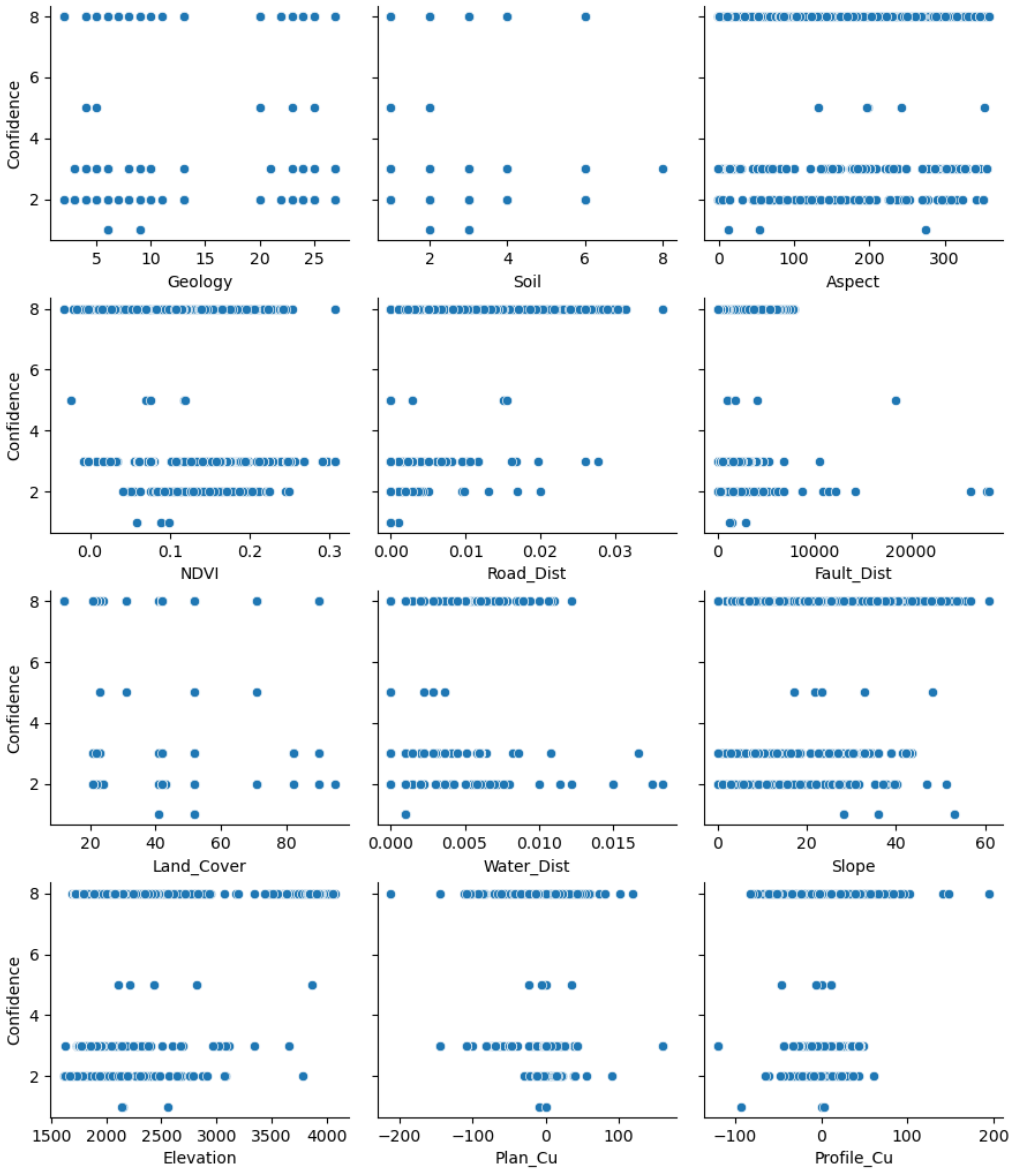

Figure 6: Binary plots of each landslide controlling factor against the Confidence field.

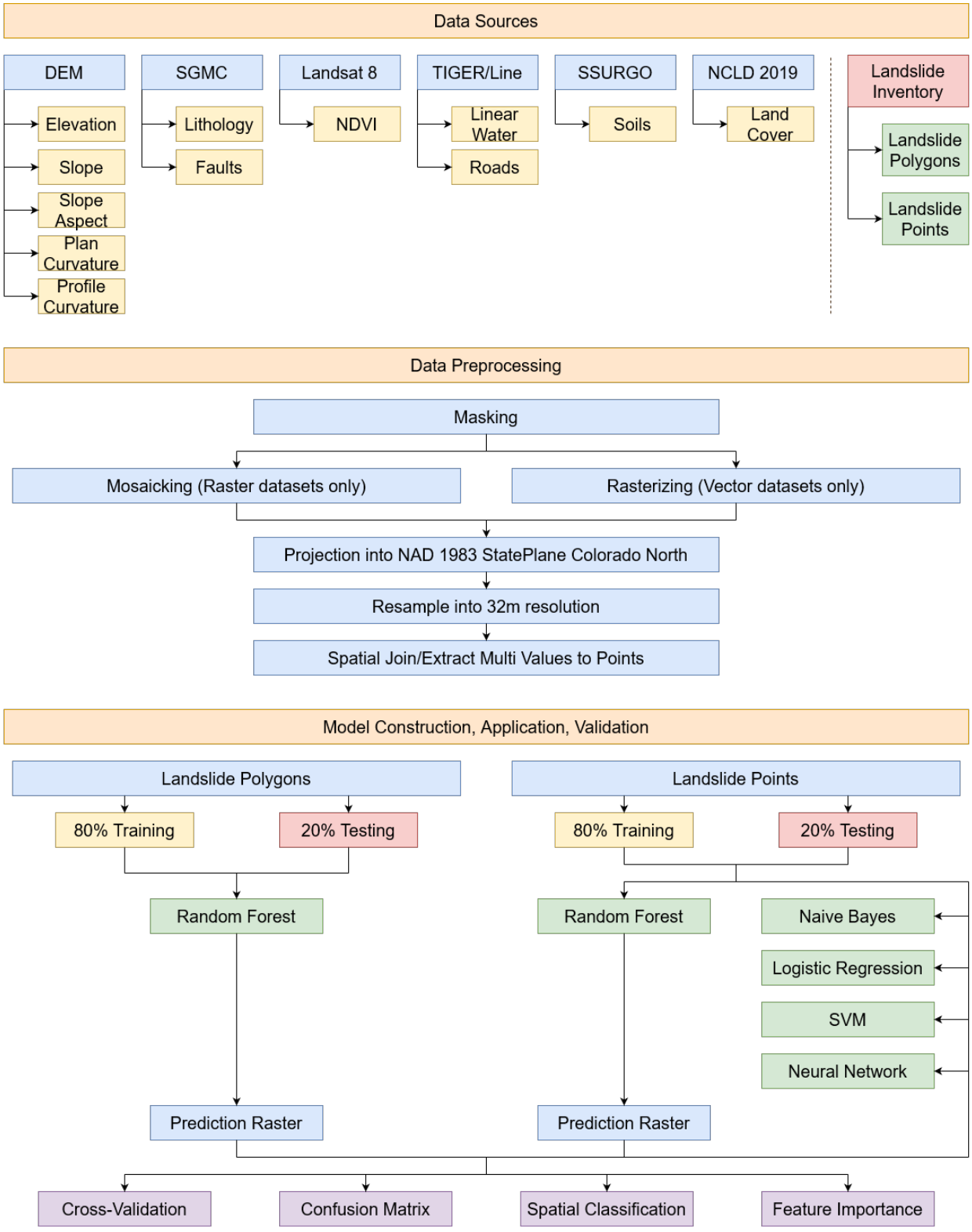

Figure 7: Flowchart showing general methodology.

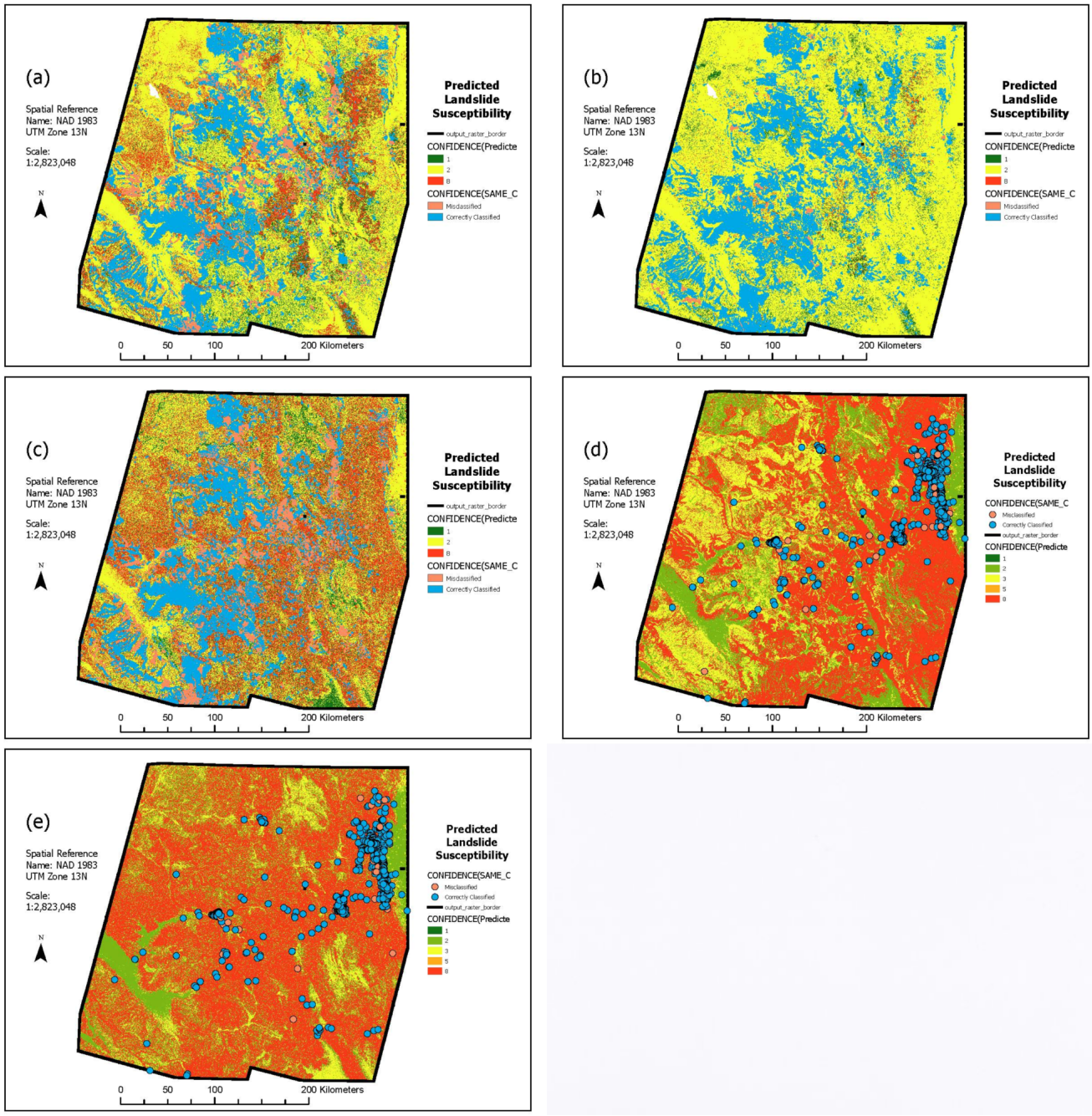

Figure 8: Results of applying several forest-based classification models to the study region to predict landslide susceptibility. Each point/polygon is coloured blue if it was classified correctly and coloured orange if it was not. (a): Random forest classification of landslide susceptibility based on landslide polygons with compensation for sparse categories. (b): Random forest classification of landslide susceptibility without compensation for sparse categories. (c): Random forest classification of landslide susceptibility based on landslide polygons with compensation for sparse categories and using only features with high importance. (d): Random forest classification of landslide susceptibility based on landslide points. (e): Random forest classification of landslide susceptibility based on landslide points using only important features.

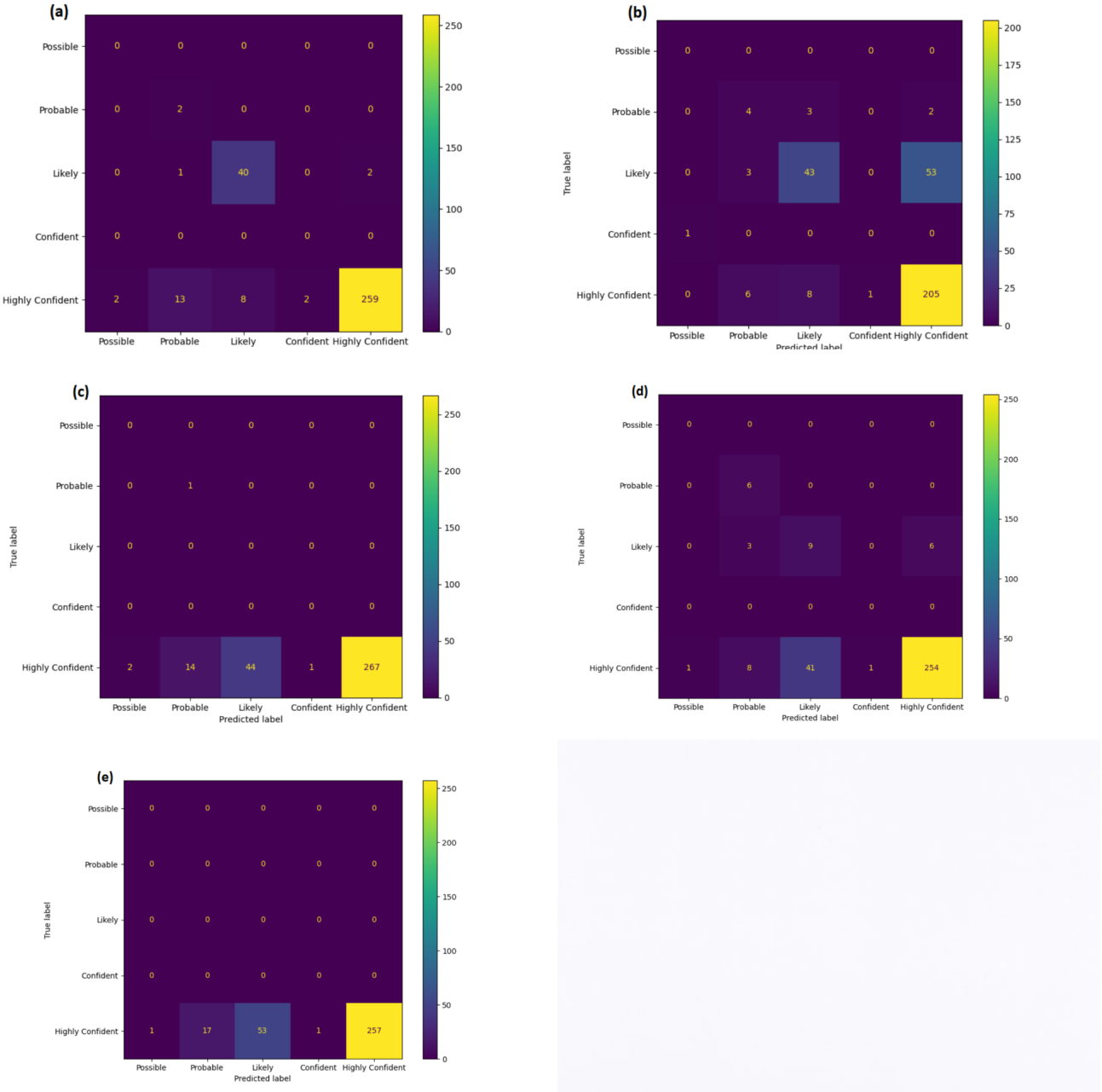

Figure 9: Confusion matrices for each point-based landslide classification model. (a): random forest; (b): Gaussian Naive Bayes; (c): support vector machine; (d): logistic regression; (e): multi-layer perceptron.

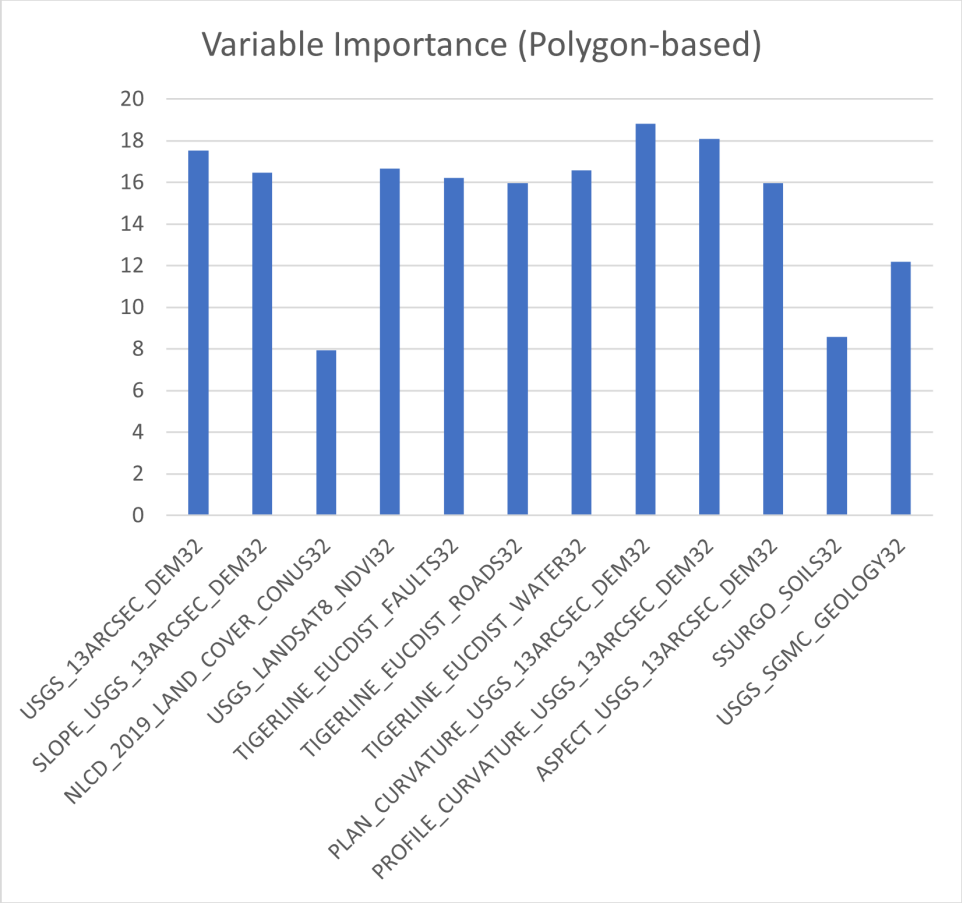

Figure 10: Feature importance based on mean decrease in impurity for the spatial random forest classification model based on landslide polygons.

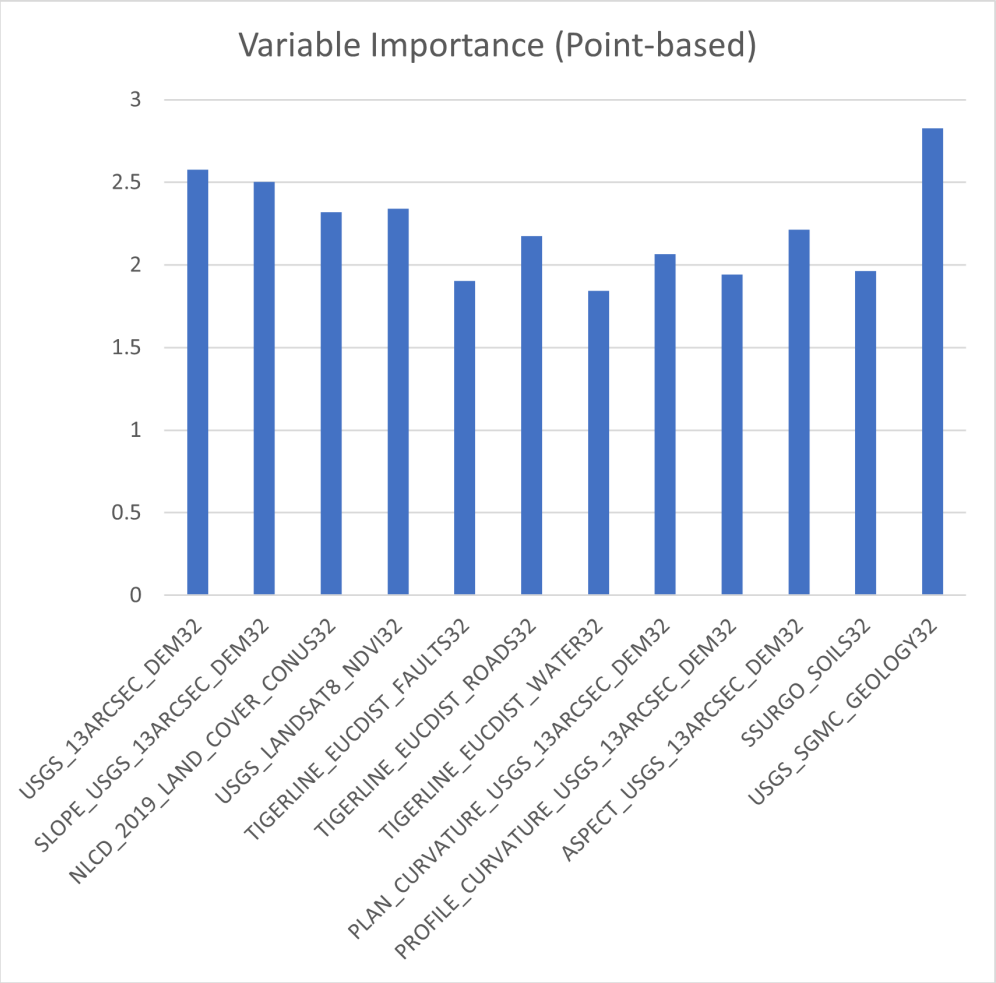

Figure 11: Feature importance based on mean decrease in impurity for the spatial random forest classification model based on landslide points.