Natural Hazards
https://doi.org/10.1007/s11069-021-04780-z

ORIGINAL PAPER

An evolutionary approach for spatial prediction of landslide susceptibility using LiDAR and symbolic classification with genetic programming

Pece V. Gorsevski[1]

Received: 8 July 2020 / Accepted: 3 May 2021

Abstract

This research examines the potential of spatial prediction of landslide susceptibility by implementing an evolutionary approach using symbolic classification with genetic programming (GP). Specifically, the light detection and ranging (LiDAR)-based digital elevation model was used to generate topographic prediction attributes and to digitize the location of shallow landslides by derivatives such as hillshade maps and contours. The presented approach tested a total of 72 runs with different parameter configurations for producing a good outcome among a number of possible solutions by varying population size, tournament group size and mutation probability. The final solution depicted a total of three important variables including slope, wetness index and solar insulation that were used in the prediction. The GP methodology used symbolic expression trees for the development of the predictive models that were tested and validated in the northern portion of the Cuyahoga Valley National Park located in northeast Ohio. The selected solution from the implemented approach showed that the area under the curve from the receiver operating characteristic curves had a high discrimination power in separating the areas with high susceptibility. The presented model yielded an accuracy of 85.0 % classifying a total of 13.4 % as high susceptibility area with an overall quantitative index of accuracy corresponding to 0.9082. Based on obtained results, the potential of the presented GP approach for mapping landslide susceptibility is promising and further exploration of its capabilities is suggested for finding new avenues of possible landslide research and practical implementations.

Keywords LiDAR · Genetic programming · Symbolic classification · Machine learning · Landslides

✉ Pece V. Gorsevski
peterg@bgsu.edu

[1] School of Earth, Environment and Society, Bowling Green State University, Bowling Green, OH 43403, USA

Published online: 27 May 2021

1 Introduction

Recent research for mapping landslide susceptibility shows different spatial and temporal approaches that evaluate "where" or "when" landslides are most likely to occur (Gorsevski 2002; Gorsevski et al. 2000, 2003, 2005, 2006a, 2006b, 2006c, 2010, 2016; Gorsevski and Jankowski 2008, 2010; Bathurst et al. 2010; Feizizadeh and Blaschke 2013; Saro et al. 2016; Tien Bui et al. 2016). Within those approaches, the statistical models which link spatial correlations of landslides and environmental attributes are the most widely used methods (Carrara et al. 1995; Chung et al. 1995; Westen et al. 1997; Gorsevski 2002; Gorsevski et al. 2003, 2005, 2006b; Ayalew and Yamagishi 2005; Gokceoglu and Sezer 2009; Atkinson and Massari 2011; Reichenbach et al. 2018). According to a recent review of statistically based landslide susceptibility approaches by Reichenbach et al. (2018), the logistic regression model is the most frequently applied method (Dai and Lee 2003; Ayalew and Yamagishi 2005; Gorsevski et al. 2006b; Bai et al. 2010; Ercanoglu and Temiz 2011; Saro et al. 2016; Zêzere et al. 2017). The literature also shows that in logistic regression studies the most commonly used covariates are slope, aspect and geology/lithology, but there is variation based on different trigger mechanisms (Budimir et al. 2015).

Other approaches for mapping landslide susceptibility that are also widely used include the map overlay methods (Reichenbach et al. 2018). The overlay methods represent the first generation of spatially based attempts which superimpose multiple maps of susceptibility factors. For instance, Boolean overlay implements a combination of maps (criteria) by logical operators such as intersection (AND) or union (OR), while the weighted linear combination (WLC) implements numeric standardization of evaluation criteria aggregated by weighted average (Gorsevski and Jankowski 2010). Moreover, through time such methods have evolved and they are often coupled with heuristic and statistical components (Gupta and Joshi 1990; Carrara et al. 1995; Westen et al. 1997; Dai and Lee 2002; Ayalew et al. 2004; Gorsevski et al. 2006c). For example, assessments that extend multicriteria evaluation (MCE) and analytical hierarchy process (AHP) often integrate additional techniques that account for uncertainties such as the fuzzy sets, Bayesian and Dempster-Shafer theories (Gorsevski et al. 2003; Ayalew et al. 2004; Gorsevski et al. 2005, 2006c; Gorsevski and Jankowski 2010; Gorsevski et al. 2010; Park 2011; Ercanoglu and Temiz 2011; Feizizadeh and Blaschke 2013).

At present, the most recent approaches are focused on data mining techniques that perform predictive (i.e., what may happen in the future) or descriptive (i.e., what happened in the past) analysis for discovering patterns in large datasets. There are several major data mining techniques for extracting patterns such as groups of data records (clustering, classification), dependences (associations, sequential patterns), relationships (predictions) and conditions (decision trees) (Chawla et al. 2001; Yao and Hamilton 2008; Tien Bui et al. 2014, 2016; Taalab et al. 2018). Those techniques have been mainly developed because of the significant expansion of "big data" and proliferation of cost-effective technologies for collecting the data. The rapid growth of big data in the field of landslide susceptibility is mostly driven by implementation of different types of distributed sensors for monitoring small-scale processes and/or the integration of heterogeneous datasets, high-resolution imaging, measurements and observations by technologies such as light detection and ranging (LiDAR) and global navigation satellite system (GNSS) (Buffat et al. 2017).

Specifically, data mining is the process of analyzing large datasets ("big data") from different perspectives and uncovering correlations and patterns through methods such as artificial intelligence, statistics, data science, database theory and machine learning. Data mining techniques reported in the landslide susceptibility literature include approaches such as rough sets, classification and regression trees, artificial neural network (ANN), random forest models and support vector machines (Gorsevski and Jankowski 2008; Stumpf and Kerle 2011; Marjanović et al. 2011; Song et al. 2012; Catani et al. 2013; Tsai et al. 2013; Micheletti et al. 2014; Korup and Stolle 2014; Tien Bui et al. 2016, 2017; Gorsevski et al. 2016; Krušić et al. 2017). Such approaches also rely on a strong geospatial component that often has multidimensional, heterogeneous and spatially autocorrelated characters. Thus, analyzing geospatial datasets that are large and complex require new robust approaches such as artificial intelligence (human-like intelligence coded by software) or machine learning (predictive algorithms that learn from data).

Artificial intelligence example of landslide susceptibility model was developed and tested by Gorsevski et al. (2016) using LiDAR and ANN with backpropagation method in the Cuyahoga Valley National Park (CVNP), Ohio. The study used LiDAR-based digital elevation model (DEM) derivatives for the extraction and mapping of the shallow landslides as well as for the development of the landslide susceptibility model. The landslides were identified and acquired through high-resolution ground models, including different sets of hillshades with varying sun azimuths and sun angles, combined with slope maps and draped topographic contours. The landslide susceptibility mapping was based on the relationship between landslides and predictor attributes including slope, profile and plan curvatures, upslope drainage area, annual solar radiation and wetness index. The validation from the predicted landslide susceptibility showed that within the very high susceptibility class (i.e., degree of membership between 0.8 and 1.0), a total of 42.6% of known landslides were correctly predicted. The very high susceptibility class associated with the landslides included a total of 1.6% from the entire study area. In contrast, the very low susceptibility class (i.e., 0–0.2) that represented 82.7% of the total area was associated with 1.2% of known landslides. In the same study, by using a wider range of 0.4–1.0 degree of membership a total of 87.8% of the known landslides were correctly classified within a total of 9.95% from the entire study area. The results from the ANN simulations yielded an RMS error of 0.292 with a total accuracy rate of 87.9% for the final model.

A machine learning approach that has gained a significant amount of popularity is the genetic algorithm (GA) (Holland 1975; Goldberg 1989). GA represents a heuristic search and optimization technique inspired by natural evolution. The approach has been successfully applied to a wide range of complex real-world problems in science, medical research, engineering and industry (Yang 2007; Winkler et al. 2009; Tong et al. 2010; Yang et al. 2014). Unlike ANNs which are used for modeling complex relationships between inputs and outputs (i.e., landslides and susceptibility factors) or for finding patterns in data, the GA is used for computation of exact or approximate solutions for optimization and search-based problems. In particular, the GA is well suited for finding optimal combinations of parameters, since they make no assumption about the problem being solved. For instance, Kavzoglu et al. (2015) used GA for finding best suited predictor variables (i.e., investigated the performance of the best 4–15 combinations of factors) that were used as inputs in logistic regression model for mapping landslide susceptibility.

This paper expands upon the previous machine learning efforts by implementing a symbolic classification with genetic programming (GP) (Koza 1992; Winkler et al. 2007; Affenzeller et al. 2009; Kommenda et al. 2012). The GP applies similar evolutionary ideas as the GA for generating an optimal solution (i.e., maximization or minimization) for a fitness function. However, unlike other machine learning approaches such as decision trees, evolutionary algorithms do not directly construct a solution to a problem but rather search in a space of possible solutions. The search space is often large and the population of solution candidates evolves through many generations by implementation of evolutionary operators applied to a selection, crossover and mutation schemes. The main difference between GA and GP resides in the representation of the final solution, where GA creates a string of numbers that represent the solution, and the GP creates computer programs or mathematical expressions to generate the solution. For instance, the best-so-far solution that appeared in any generation is often used as the final result from the GP (Koza 1992).

The proposed approach is demonstrated using a case study of the CVNP. This paper builds on earlier work by Gorsevski et al. (2016), which used LiDAR-derived high-resolution ground models for the acquisition of the landslides and subsequent implementation of the ANN backpropagation method that was used for the mapping of the landslide susceptibility. The present paper is organized as follows. The method development for GA and GP used in the study is highlighted in Sect. 2. The study area, datasets and methodology that concentrate on the implementation of the approach are presented in Sect. 3. The results and discussion produced by the GP and validation of the approach are presented in Sect. 4, and finally, conclusions are discussed in Sect. 5.

2 Method development

2.1 Genetic algorithm

GA is an optimization technique that is based on an adaptive heuristic search methodology and evolutionary ideas inspired by the principles of genetics and natural selection (Holland 1975, 1992; Goldberg 1989; Koza 1992; Wagner et al. 2014). GA implements targeted intelligent searches to solve an optimization problem based on the Darwin's theory of evolution where "survival of the fittest" (i.e., producing offspring of the next generation) is propagated through the process of reproduction, crossover and mutation. The GA simulates the survival of the fittest (or most well adapted) from a set of individuals referred as population over sequential generations where each individual has a set of rules encoded as genes and connected together into chromosomes. In practice, the search space for the GA is represented by the input data that contains all feasible solutions, while individual records represent one feasible solution. Each solution (a chromosome) is represented by variables (genes), while the fitness score assigned to the solution represents the measure of the quality (Winkler et al. 2007).

The GA process starts with a selection of set of individuals where each individual is evaluated based on "fitness" or "goodness" (e.g., a high level of adaptation to a given environmental circumstance). The fitness function evaluates the fitness of each chromosome x in the population where ability of an individual to survive in its environment is compared with the rest of individuals using a fitness score. The fitness score is used to select individuals for the reproduction, usually by ranking the effectiveness of potential solutions (hypothesis) (Mitchell 1997). The purpose of this step is to ensure that the best individuals within the population survive and proceed to the next phase. The selection of appropriate solution candidates is performed by different selection techniques such as roulette wheel, ranking, tournament, steady-state GA and elitism (Blickle and Thiele 1996). For example, the tournament selection is a method that runs several "tournaments" among a few individuals that are randomly chosen from the population. The technique is inspired by nature, as it is similar to the competition for food or mating in populations in natural environments. The winner of each tournament is the individual with the highest fitness score (or best adaptation) that is selected to generate new offspring. The selection pressure is a probabilistic measure of a chromosome's likelihood of participation in the tournament which ensures selection of better individuals in the mating pool. The convergence rate of the GA is mostly governed by the selection pressure where low selection pressure generates slow convergence rate (i.e., longer computation for finding an optimal solution), while high selection pressure can generate premature suboptimal convergence of the solution. Increased selection pressure is controlled by tournament size where the winner from a large tournament has a higher fitness than the winner from a small tournament (Miller and Goldberg 1995). However, a large tournament size can also increase a loss in diversity and lower the selection variance. The probability of reproduction and expected number of copies for most of these methods relies on the computation of the fitness score. Thus, the probability of reproduction is strongly controlled by different selection operators.

The next step is the crossover, which is the most significant phase in GA. The mechanism of crossover is similar to sexual reproduction in ecosystems in the natural world. Here, the new individuals are created by combining (mating) two parent hypotheses to produce offspring. In GA, crossover represents a genetic operator that is used to create genetic diversity by varying chromosomes from one generation to the next. The crossover in GA also uses different techniques such as standard and alternative crossover for binary encoding and recombination for non-binary encoding (Chipperfield et al. 1994). For instance, the GA as proposed by Holland (1975) implements encoding on the individuals as binary strings. The encoding and decoding function has the following form:

$$h : M \rightarrow \{0,1\}^l, h' : \{0,1\}^l \rightarrow M \tag{1}$$

which maps solution candidates $\vec{x} \in M$ to binary strings $h(\vec{x}) \in \{0{,}1\}^l$ and vice versa, where the individuals are represented as binary vectors of fixed length l and the search space of the optimization problem is denoted by M. The complexity of the mapping h and h' depends on the differences in the parameters in terms of types and ranges. The binary representation of individuals is mainly for the purpose of implementing the schema theory. The term schema refers to a string over the alphabet $\{0, 1, *\}$ that can be viewed as a template for binary strings. A template is made up of characters 0, 1 which are fixed, and the symbol * which represents either 0 or 1. For example, the template 00* will match two binary strings 000 and 001. The usefulness of the schema theory is that it can provide a better understanding of the evolution process and population dynamics that are aimed at increased performance and predicting the behavior of GA (i.e., convergence in early stage and optimal configuration such as combined use of operators, parameter settings and selection of a fitness function) (Poli 2001; Zojaji and Ebadzadeh 2016).

The most common crossover techniques with binary encoding include single-point crossover, two-point crossover and uniform crossover. The main purpose of the crossover operators is to mix and match attributes between the parent chromosomes through random processes. In a single-point crossover a random value of an arbitrary gene position within the chromosomes (referred to as the crossover point) is selected. To produce the offspring, all data beyond the crossover point are swapped between the two chromosomes. The two-point crossover calls for two random points to be selected on the parent chromosomes, and data between the crossover points are exchanged for the production of the offspring. The uniform crossover uses a fixed mixing ratio between two parents that enables a crossover at gene level instead of at segment level. A mixing ratio of 0.5 generates 50 % genes from each parent with randomly chosen crossover points.

Following crossover, the mutation operator is applied occasionally to maintain genetic diversity between sequential generations. Mutation alters one or more gene values in a chromosome from its initial state which can change an outcome of a solution entirely, compared to previous solution. Every gene within a chromosome has an equal chance for mutation that is assigned by mutation probability. The probability is usually kept at a low value to avoid losing a large number of good chromosomes and to prevent a random search. Thus, a mutation represents a small random change to a value that occurs with low probability that is essential to the convergence of the GA. Some of the most common mutation operators include a bit flip mutation, random resetting, swap mutation, scramble and inversion mutation (Ronco and Benini 2014). In bit flip mutation one or more bits (genes) are flipped, whereas in swap mutation two positions on the chromosome are selected at random and values are interchanged.

Following mutation after the child population is created, all children are evaluated and fitness values are generated. The fitness values are used as an evaluation for ranking and assigning probabilities that are used for both selection and replacement that follows. Replacement is the final generational step that merges the new population of children with the parents. Replacement represents a strategy that controls population diversity by implementing extended exploration of the search space regions where global optimum is not contained (i.e., preventing premature convergence) (Lozano et al. 2008; Hamblin 2012). In generational replacement, all parents die at the end of each generation, while selection persists until the entire population has been replaced by new child population. In the case of overlapping also called steady-state systems, parents and offspring compete for survival and some percentage of the population is deleted (i.e., replacement rate) for each generation and replaced with children. Often, during the selection process elitism is applied to some percentage of the fittest individuals that guarantee inclusion in the next generation.

Finally, the termination condition determines the end of the GA run. Although, at the beginning stages of the run, better solutions emerge frequently, but at the later stages the improvements in the solutions stagnate and are very small. For that reason, a termination condition is applied to end the GA run while producing a good solution that is close to optimal. The common termination conditions are applied (1) when there is no improvement in the solution in a specified number of iterations (i.e., convergence); (2) when an absolute number of generations are reached; and (3) when the objective function is optimized based on a pre-defined value.

2.2 Genetic programming

GP (Koza 1992; Poli et al. 2008) is a machine learning method that builds on the fundamental principles of GA but which is more powerful. The main difference is that the output of the GA represents a quantity, while the output of the GP represents a computer program that is used to produce the solution. The derivation of an automated GP computer program is achieved through evolutionary computation using the principles of GA's population breeding without user's interaction. Typically this is done by the creation of a population of individuals which go through the process of simulated evolution implemented as selection, crossover and mutation operators. The GP uses the following four steps for solving problems (1) initiates a population of random compositions of functions and terminals or computer programs; (2) executes each program in the population and assigns a fitness value; (3) creates a new population of computer programs by coping best existing programs and creates new computer programs by crossover and mutation; and (4) the best computer program produced in any generation (i.e., the best-so-far solution) represents the result of the GP (Koza 1992).

Although GP can use diverse representations to evolve programs, the most common representation is the syntax tree. Within the tree structure the functions and the terminals are the building blocks of the computer programs. For example, the leaves of the tree are the terminals (terminal nodes), while the internal nodes are the functions such as arithmetic operations addition, subtraction, division, multiplication or other more complex mathematical expressions. Also, the terminals are divided in constants and arguments where the constants remain the same for the entire evolution, while the arguments are the program inputs that represent the variables from the dataset. The genetic crossover operation on the tree structure involves the selection of two parents (i.e., solutions) by fitness function that are combined to form two new solutions or offspring. The tree structure is particularly useful for representing computer programs, and the most common recombination operator is based on a subtree crossover. Subtree crossover creates new programs from two parental programs (i.e., existing programs) that are selected from the population based on fitness where the first is called the *receiving* parent and the second is called the *contributing* parent. This type of crossover selects a crossover point in each parent tree at a random position such as a node that divides the parent trees. Given two parents, the parts are exchanged to produce the offspring. Thus, the subtree rooted at the crossover point of the first parent is deleted and replaced by the subtree from the second parent.

Often, the selection of parent solutions, which serve as an input to the crossover operation, is selected by a different method such as probability based on fitness of the solution, tournament selection and rank selection. For example, using the probability based on fitness of the solution (Koza 1992), where $f\left(S_i(t)\right)$ is the fitness of the solution S_i and

$$\sum_{j=1}^{M} f\left(S_j|(t)\right) \tag{2}$$

the total sum of all members of the population M is represented by Eq. 2, than the probability that the solution S_i will be copied to the next generation is:

$$\frac{f\left(S_i(t)\right)}{\sum_{j=1}^{M} f\left(S_j|(t)\right)} \tag{3}$$

The tournament selection chooses random solutions (i.e., tournament group size) where a higher fitness score wins. The approach simulates biological mating patterns by a tournament where members of the same sex compete to mate with another member of a different sex. On the other hand, the rank selection is based on the rank of the fitness values of the solutions of the population (Koza 1992). In a tree structure representation, where each individual of the population is represented as one tree structure the crossover and mutation processes are applied to randomly chosen branches (i.e., subtree) within the trees. In GP the tree structures are predefined by minimum and maximum tree sizes, while the branch represents a tree subset that has been selected by a random value (i.e., lying below the random value in the tree). During the crossing process branches in each parent tree are randomly selected and swapped to produce offspring. Finally, the mutation process randomly modifies a solution candidate by choosing a node and altering its function or branches that are replaced by other branches (Winkler et al. 2007).

3 Materials and methods

3.1 Study area

The study area (Fig. 1) is within the CVNP, in northeast Ohio. The area (133.6 km^2) is located along the lower Cuyahoga River, which is one of the most landslide-prone regions of the watershed. The elevational range is between 305 and 381 m above sea level. The

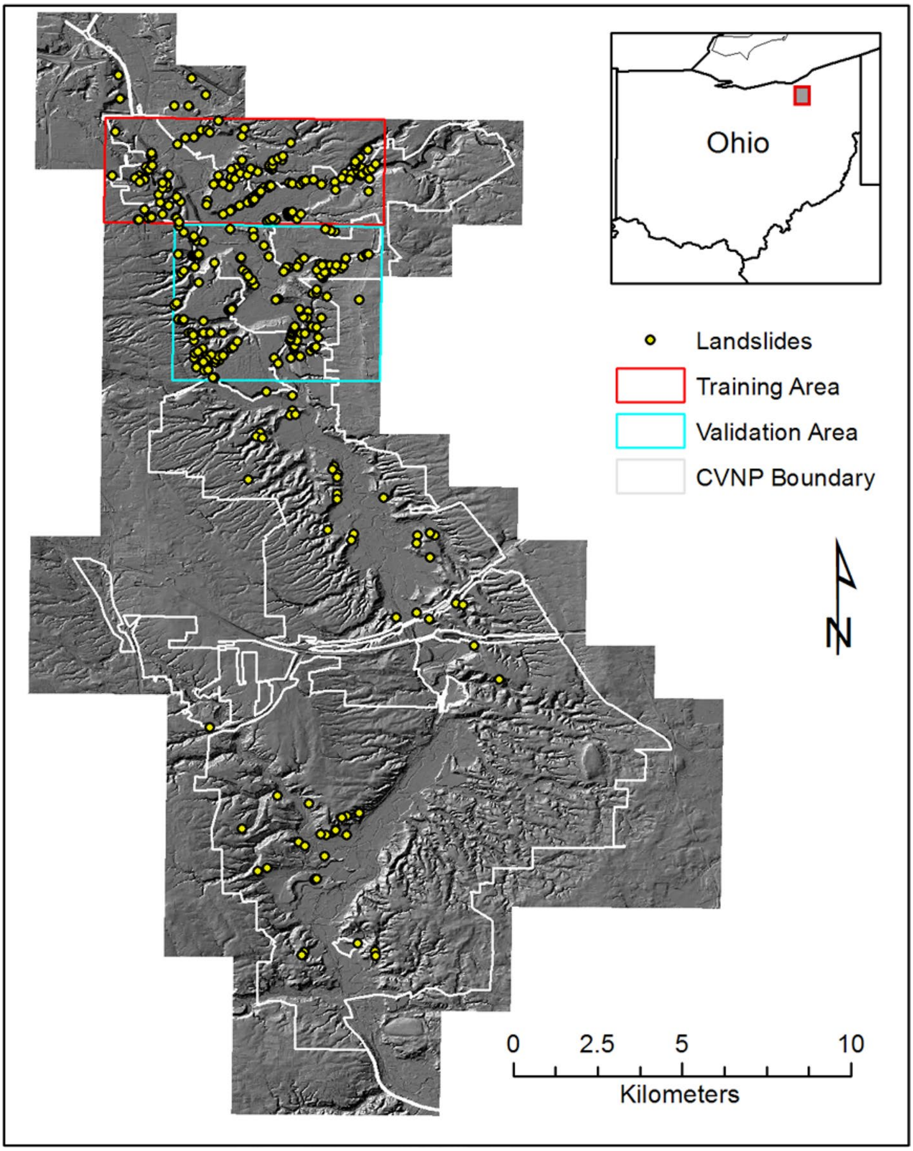


Fig. 1 Distribution of landslides organized by training and validation areas in the CVNP

annual precipitation is mostly driven by localized ‘lake effect’ precipitation from Lake Erie and ranges between 927 and 1027 mm. The length of the river through the National Park is approximately 35.4 km with well-developed drainage network which contributes to a significant surface runoff (Nandi and Shakoor 2010; Gorsevski et al. 2016).

The Cuyahoga River Valley is a part of the glaciated Allegheny Plateau (Szabo 1987). The valley of the river is well known for rotational slumps in clays and silts deposits that are from lacustrine origin formed during the Pleistocene glaciation when various segments of the valley were blocked (Hansen 1995). The modern environment is represented by deep incisions through these deposits and steep valley walls of unstable sediments with high landslide potential.

The hillshade LiDAR-derived map in Fig. 1 shows the training area (red outline) used for the model development (18.6 km^2) and the testing area (turquoise outline) used for the validation of landslides (20.9 km^2). The areas were selected based on the similarity of the lithology derived from geological maps. The training area consists of 31.15% Berea Sandstone and Bedford Shale, 68.49% Ohio Shale and 0.36% Maxville Limestone, while the validation area consists of 30.23% Berea Sandstone and Bedford Shale, 69.2% Ohio Shale and 0.57% Maxville limestone. Because the lithology of the study area was derived from bedrock geology at a scale of 1:500,000, it lacks detail; therefore, the lithology in this study is represented by similarity in the study units and it was not used in the GP analysis (Gorsevski et al. 2016). However, there is a possibility that differences in the lithological units existed within the training and the validation dataset that could contribute to class imbalance that could influence reported validation results. Although unbalanced data were not considered here, the literature reports different techniques for improving classification accuracy especially for the minority classes such as a borderline-stroke technique and evolutionary parameter tuning strategy with combined gradient boosting and differential evolution (Deng et al. 2017; Saporetti et al. 2019).

3.2 LiDAR datasets and landslides

The DEMs were constructed using publicly available pre-processed LiDAR datasets from the Ohio Statewide Imagery Program (OSIP) (OGRIP 2018). The OSIP datasets include high-resolution imagery and elevation data from LiDAR for the State of Ohio, which were acquired in 2006 and 2007. The LiDAR dataset was produced with an average laser pulse spacing of 7 ft (2.1336 m) and with an accuracy of 1 foot (0.3048 m). The DEMs are available in ArcGIS and ASCII grid format which are organized in tiles. The coverage of each tile is approximately an area of 2.3 km^2. In the study area, the multiple tile covers were mosaiced to create training and validation DEM grid layers. The training study area used a total of eight DEMs, whereas the validation study area used a total of nine DEMs.

A total of six topographic attributes derived from LiDAR OSIP datasets were used in the model development (Wilson and Gallant 2000; Hengl and Reuter 2009). These topographic attributes included: slope, solar radiation, profile curvature, plan curvature, upslope drainage area and wetness index. The same set of topographic attributes were also used for the prediction and testing of landslide susceptibility in the same study area by (Gorsevski et al. 2016) when ANN with backpropagation method was implemented. Similar topographic attribute-based approaches were also used for the prediction of landslide susceptibility in earlier work by Gorsevski et al. (2003, 2005, 2006b, 2006c). Such attributes are often used to quantitatively describe water movement, hydrological process, morphometry, catchment position and soil-landscape processes (Wilson and Gallant 2000; Hengl and Reuter 2009).

Landslides were acquired from the training and the testing areas using LiDAR-based hillshades combined with slope maps and draped topographic contours as described by Gorsevski et al. (2016). The delineation of landslides included two separate inventory trails using varying sun azimuths and sun angles that were intended to enhance the landform shadowing and to assist with the recognition of the landslides. Only the landslides that appeared in both inventory trails were used for the analysis. The final coverage had a total of 190 landslides where 83 and 107 were recorded within the training and the testing study areas, respectively. In addition, the field validation suggested an accuracy of 90.7 and 78.9 % for the training and the testing study areas, respectively (Gorsevski et al. 2016). Furthermore, the landslides generated from the hillshades were classified into three different classes that were used to depict the main geomorphological elements of a landslide: the scarp, the body and toe of the landslide. The six predictor attributes were used as inputs for the classification, while the centroids derived from the classification (mean cluster values) were used for the interpretation and characterization of the scarp. For instance, the scarp was represented based on the cluster center values using criteria such as highest slope, concave and wetness characteristics. The grid cells associated with the scarp represented the presence of a landslide, while non-scarp grid cells represented the absence of a landslide.

3.3 Methods

The dataset was constructed by implementing random sampling design using binary classification for presence and absence of landslides. The presence or absence of a landslide was represented by grid cells with values of 1 for presence and 0 for absence. The initiation area of each landslide (i.e., the area where the main scarp of the landslide occurred) was used as the point representing the presence of a landslide. Areas outside the landslide polygons were used for sampling the absence. A total of 1000 grid cells represented presence (landslides) and a total of 1000 represented absence (non-landslides) that were sampled from the training area. The training set used a total of 75 % of the records for the development of the GP model, while the rest of the records (i.e., 25 %) were used for testing the fit of the model. Additional validation area was used as an independent testing site to test the GP model, which was developed from the training area (Fig. 1). The distributions of landslides (presence) and non-landslides (absence) from the training area are shown in Fig. 2. The plots associated with the individual predictor attributes show the probability density distributions of landslide grid cells (turquoise color) and non-landslide grid cells (red color). For instance, the slope attribute plot shows that most of the grid cells affected by the landslides have higher slope values, while grid cells not affected by the landslides have lower slope values. The distinction in the solar radiation and the wetness index plots is more pronounced compared to the rest of the attributes.

In this research the symbolic classification algorithm based on GP was implemented using HeuristicLab framework which is evolutionary algorithm used for prototyping and analyzing optimization techniques. The software is an open source developed by the Heuristic and Evolutionary Algorithms Laboratory (HEAL) (HeuristicLab 2018). The main goal of the symbolic classification is to produce a model that predicts discrete target class (i.e., high susceptibility) using predictor variables that are both discrete or continuous (i.e., topographic attributes). The symbolic classification is characterized by different manipulation operators and operands to represent the discriminating function or the classification model. The classification model developed by the GP symbolic classifier was subsequently used in an open-source GIS environment (GRASS 2018; SAGA 2018) to produce

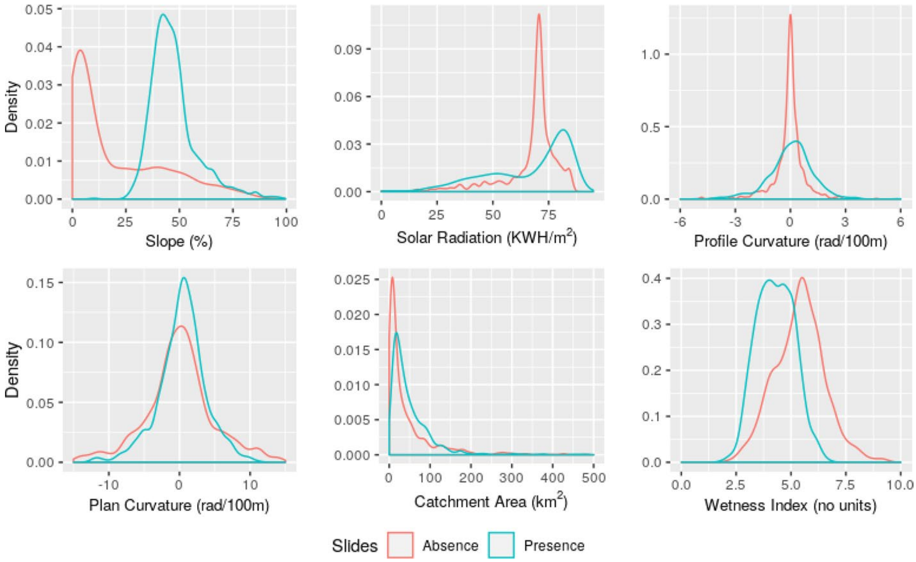


Fig. 2 Probability density distributions of landslide's presence and absence associated with individual topographic attributes used in modeling

a dichotomous maps of landslide susceptibility outputs (i.e., landslide and non-landslide areas). A number of attempts were explored for generating the final model and the final set of parameters used for implementing the symbolic classification algorithm that are shown in Table 1. The table shows that the parameter configuration used in the GP classifier implemented a range of different values for the population size, group size and mutation probability, while other parameter values such as crossover, maximum generations and maximum symbolic expression tree depth and length were fixed. Other settings included the use of default parameters suggested by HeuristicLab such as the symbolic expression tree grammar where settings included all available library functions in the software.

4 Results and discussion

A total of 72 runs were generated using the parameter configuration from Table 1. The final solution presented in this research used a population size of 1000, a mutation rate of 5% with multisymbolic expression tree manipulator with the following inner operators: change node type manipulation, full tree shaker, one point shaker, remove branch manipulation, replace branch manipulation, a sub-tree swapping method for the crossover and a tournament selection with group size of 7. The termination criteria for all runs used a maximum number of 1000 generations, while the number of elite solutions for each generation was kept to one (i.e., a single best solution was produced for each run). Although there are many considerations for finding good (i.e., candidate) solutions for complex optimization problems, in this research the best quality measure was used for the selection of the solution presented here. Figure 3 illustrates the metaheuristic approach for the comparison of runs and selection of good solution outcome among large number of possible solutions. Such an approach does not guarantee finding an optimal solution, but the aim is to find a good solution in large search spaces and choices of different parameter values. Although it may be possible to achieve an even better solution by increasing population size beyond 1000 and testing additional runs, the main goal of this research was to demonstrate the methodological steps for the selection for a good solution which is described below.

4.1 Selection and analysis of parameter values

The assessment for comparison of the quality of the runs is shown in Fig. 3a, b and c where different parameter values were explored. The *x*-axis in the plots groups the solutions from the runs by different parameter values, while the *y*-axis shows the best quality generated by fitness values of the best found solutions that have the least mean squared error (MSE). The best solution quality in the figure is represented by lower *y*-axis values. For instance, Fig. 3a shows better mean quality for the population size parameter that is achieved by the increase in the population size which is 1000 (mean = 0.1052; SD = 0.0029). According to Gotshall and Rylander (2002) smaller population-sized batches are most likely to find a suboptimal and premature solution during the initial state, while a larger population size yields a high probability that the solution would not converge in a reasonable number of generations. Although the authors propose a method for optimal population size based on point of inflection where quick convergence can be balanced by increased inaccuracy, the overall benefit of larger population size is that increases the overall accuracy of the solution.

Figure 3b shows that the best quality for the selection parameter that is generated by group size of 7 (mean = 0.1059; SD = 0.0039). The tournament-based selection involves number of individuals that are chosen randomly where the fittest one is selected as a parent for crossover. The tournament size controls the selection pressure, which is a probabilistic

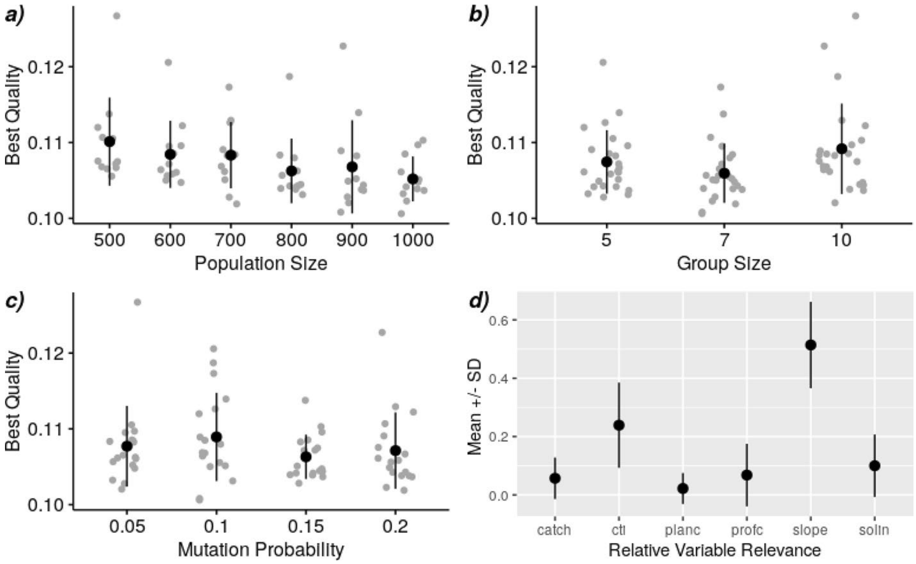


Fig. 3 Comparison of the runs based on best quality generated by varying **a** population size, **b** tournament group size and **c** mutation probability, while **d** represents the importance of the variables used for the GP predictions

measure of the likelihood for individuals to participate in the tournament. For example, when the tournament size is small the selection preserves the diversity (i.e., repeated rounds of selection) and allows a chance to all individuals to be selected in the tournament, but it may degrade the convergence speed. On the other hand, a larger tournament size could have a loss of diversity (i.e., prevents weak individuals to be selected) that will result in higher fitness individuals winning the tournament or being selected more often. Figure 3c shows the mutation probability that also contributes to the retention diversity objective. The mutation probability yields relatively similar quality across the groups. The best quality for the mean is generated by the 15% mutation probability (mean = 0.1063; SD = 0.0029). The 5% mutation probability has the lowest dispersion associated with the quality using the interquartile range (0.0033). In this research, the 5% mutation probability was used with a mean = 0.1077 and SD = 0.0053.

The relevance of variables associated with the selection of a minimal subset necessary to describe relationships in the dataset is shown in Fig. 3d. The attributes in the figure include upslope drainage area or catchment (catch), wetness index or compound topographic index (cti), plan curvatures (planc), profile curvatures (profc), slope (slope) and annual solar radiation (solin). The figure shows that the slope variable is the most important (mean = 0.514; SD = 0.148) followed by wetness index (mean = 0.239; SD = 0.146), solar radiation (mean = 0.100; SD = 0.107), profile curvature (mean = 0.068; SD = 0.107), catchment (mean = 0.057; SD = 0.071) and plan curvature (mean = 0.022; SD = 0.053). The relative variable relevance is derived from frequency-based approach from all runs using count of individual variables being referenced. From the figure it is clear that the slope variable is with the highest importance followed by wetness index and solar radiation. The selected solution (from a single run) illustrated in this research is based on the top three important variables which are shown in Fig. 4a.

Lastly, symbolic expression grammar frequency used for the classification is shown in Fig. 4b. The plot illustrates the changes of the symbolic grammar expression through the classification and the relative frequency of symbols used to construct the final symbolic expression. For instance, at the initial state around generation 250 there is higher variation of symbols used, but after generation 500 the variation diminishes and the grammar frequency is relatively stable. The order of symbols at generation 1000 is the following: variable, constant, addition, multiplication, start symbol, division and greater than.

4.2 Symbolic classification model

The predictive models for landslide susceptibility are shown in Fig. 5a, b and c represented as symbolic expression trees with their corresponding mathematical representations. The original symbolic classification solution has a depth of seven levels and a length of 24 nodes (Fig. 5a). On the other hand, the automatically simplified solution shows a semantically equal model as the original which resulted from mathematical simplification where the depth is reduced to six levels, while the length is reduced to 15 nodes (Fig. 5b). Finally, Fig. 5c shows an oversimplified version of the model generated by manual pruning of branches where the depth is reduced to two levels, while the length is reduced to 3 nodes. The coloring scheme in the automatically simplified solution represents the visual relevance of the branches that are used for analysis and further simplification through manual pruning. For instance, the strong coloring is associated with high relevance, while lighter coloring is associated with branches that have weak impact on the model.

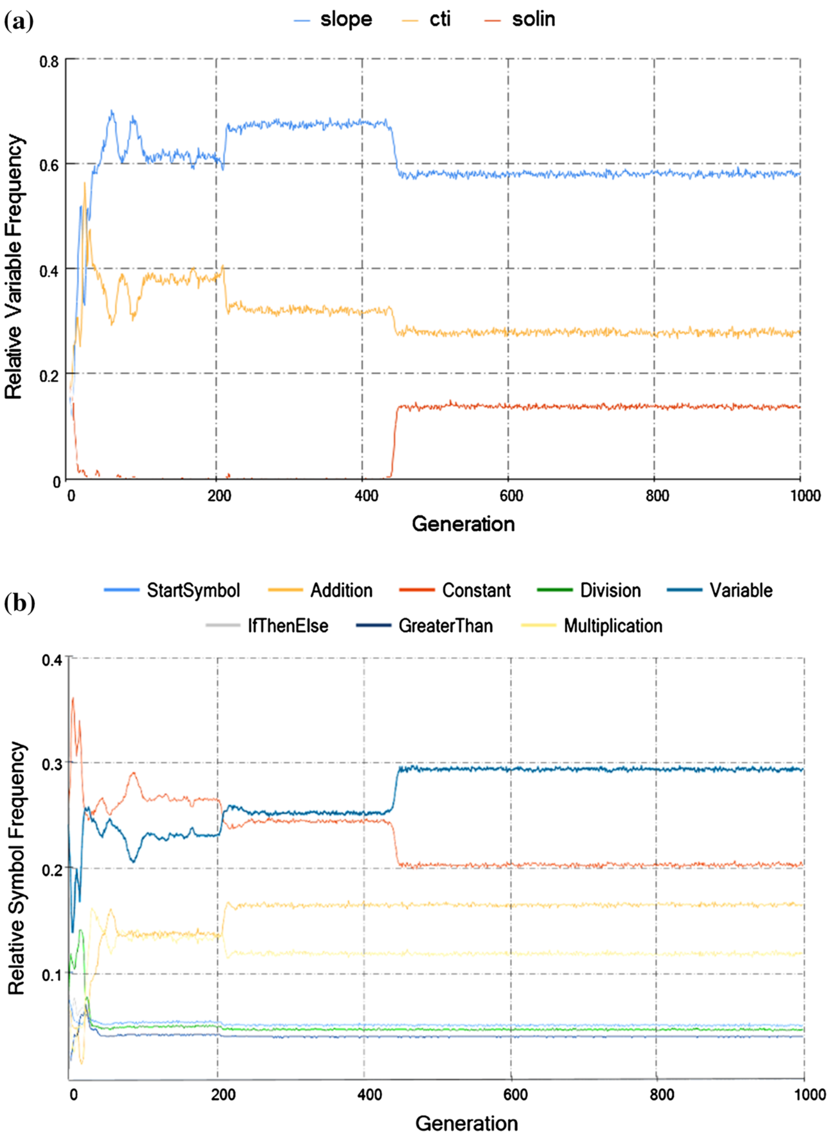
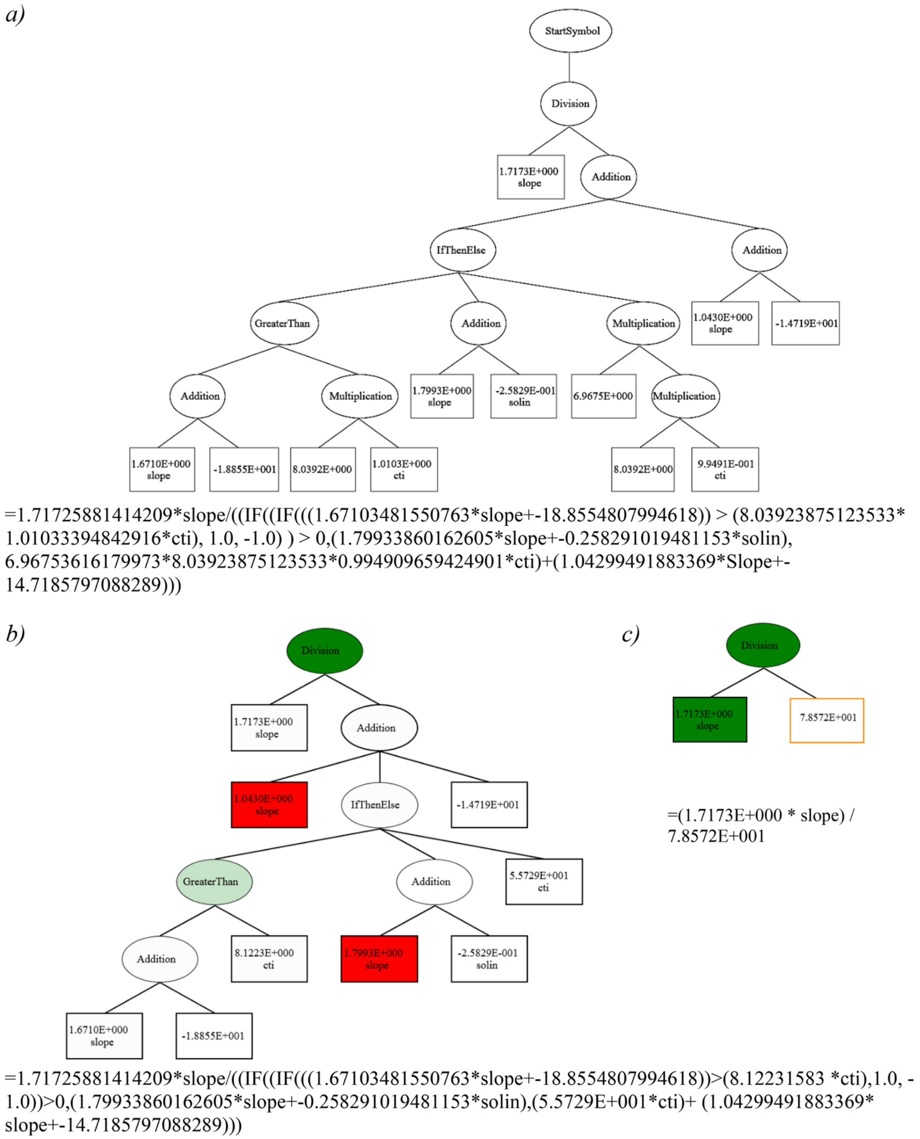


Fig. 4 Parameter analysis showing the relative frequency of **a** important variables and **b** symbolic expression grammar for the classification problem which is used as the final solution.

The symbolic tree structure produces a model that integrates functions and terminal primitives to represent nonlinear mathematical expressions. For example, the graphical representation of the symbolic tree in Fig. 5a, b and c shows that the terminals (i.e., tree leaves) correspond to coefficients or to constants (i.e., random coefficients) that are multiplied by an explanatory variable. The original symbolic classification and the automatically simplified solutions in Fig. 5a and b use all important variables from Fig. 4a including slope, wetness index and solar radiation, while the model after manual pruning of branches



=1.71725881414209*slope/((IF((IF(((1.67103481550763*slope+-18.8554807994618)) > (8.03923875123533* 1.01033394842916*cti), 1.0, -1.0)) > 0,(1.79933860162605*slope+-0.258291019481153*solin), 6.96753616179973*8.03923875123533*0.994909659424901*cti)+(1.04299491883369*Slope+- 14.7185797088289)))

=(1.7173E+000 * slope) / 7.8572E+001

=1.71725881414209*slope/((IF((IF(((1.67103481550763*slope+-18.8554807994618))>(8.12231583 *cti),1.0, - 1.0))>0,(1.79933860162605*slope+-0.258291019481153*solin),(5.5729E+001*cti)+ (1.04299491883369* slope+-14.7185797088289)))

Fig. 5 Symbolic expression trees and mathematical representation models for predicting landslide susceptibility where **a** is the original representation, **b** is the automatically simplified solution and **c** is the model after manual pruning of branches with low impacts

in Fig. 5c is based on a single variable which is the slope. The non-terminals are not associated with any additional parameters but represent mathematical expressions such as combined formulas of arithmetic and logical functions for receiving the parameters as input branches for the creation of a complex mathematical expression.

The classification results are shown in Tables 2 and 3 where confusion matrices are generated from the training area. The training area dataset was partitioned into training and

Table 1 The parameter configuration used in the GP classifier

Parameter	Values
Population size	500, 600, 700, 800, 900, 1000
Selection	Tournament selector with group size: 5, 7, 10
Crossover	Sub-tree swapping
Mutation probability	Multi symbolic expression tree manipulator: Change Node Type Manipulation, Full Tree Shaker, One Point Shaker, Remove Branch Manipulation, Replace Branch Manipulation with mutation rates: 5 %, 10 %, 15 %, 20 %
Maximum generations	1000
Maximum symbolic expression tree depth	8
Maximum symbolic expression tree length	25
Symbolic expression tree grammar	Arithmetic functions, Conditional symbols, Terminals and constants

test sets with a split of 1500 and 500 grid cells, respectively. While the training partition was used by the algorithm for model learning, the test partition was used for estimating the generalization error which is a measure of prediction accuracy. The results in Table 2 represent the original or the automatically simplified solution from Fig. 5a or b which are equivalent. The visual difference in the symbolic tree structure is due to removal of constants in the simplified solution which improves the clarity by reducing depth and length of the final solution. The confusion matrix in Table 2 shows that a total of 574 absence (non-landslide) and 724 presence (landslide) cells were correctly classified in the training partition, while a total of 183 non-landslide and 242 landslide cells were correctly classified in the test partition. The accuracy for the training is 86.5 % and for the test is 85.0 %, while the F1 score for the training is 85.0 % and for the test is 82.9 %. The classification accuracy from the confusion matrix depicts the performance in terms of proportion of true results among the total number of items (i.e., importance of true positives and true negatives), while the F1 score, which calculates a mean of precision and recall by emphasizing the lowest value, depicts the importance of the false negatives and false positives (Jayawardhana and Gorsevski 2019). In the confusion matrix, a true positive is a prediction of a landslide for a location where a landslide occurred, while a false positive is a prediction of a landslide for a location where a landslide did not occur. The confusion matrix produced after manual pruning from Fig. 5c is shown in Table 3. The table shows that a total of 561 non-landslide and 722 landslide cells were correctly classified in the training, while a total of 170 non-landslide and 243 landslide cells were correctly classified in the test. The accuracy for the training is 85.5 % and for the test is 82.6 %, while the F1 score for the training is 83.7 % and for the test is 79.6 %. The quality of this model is a slightly lower, but the model is compact and comprehensive.

The classified model predictions were compared using the area under the curve (AUC) derived from receiver operating characteristic (ROC) curves (Fig. 6). The ROC curves are graphical representations of the relationship between the true positive (correctly predicted landslides) and the false positive (falsely predicted landslides), while the AUC is used as a measure of overall model fit and comparison of different predictive outcomes. An ideal model would have an area equal to 1, because then P (true positive) = 1 and P (false positive) = 0 regardless of the cutoff point (Gorsevski et al. 2000, 2006b; Gorsevski 2013). The

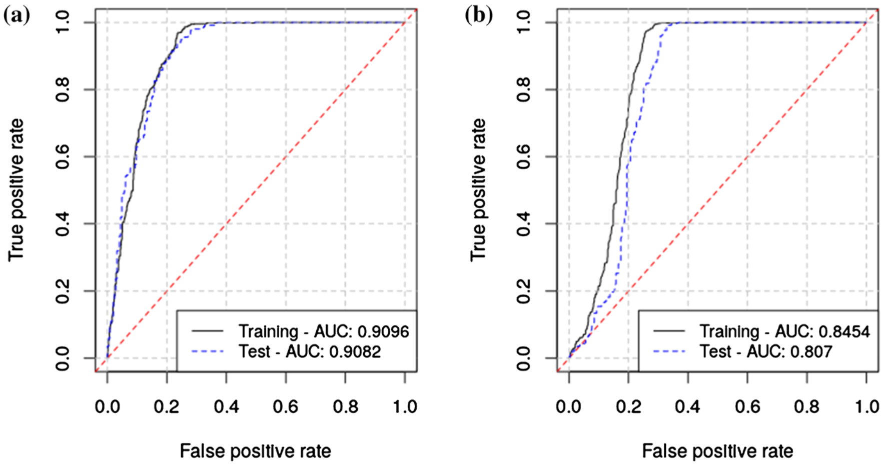


Fig. 6 ROC curves from **a** the automatically simplified solution and **b** the model after manual pruning of branches

AUC under the red diagonal line in the figures represents an area of 0.5 which indicates a random classification model. The true-positive rate on the *y*-axis represents the ratio of true positive out of condition positive (i.e., true positive/true positive + false negative), while the false-positive rate on the *x*-axis represents the ratio of false positive out of condition negative (i.e., false positive/false positive + true negative). The positive and negative conditions correspond to the columns of the confusion matrices (Tables 2 and 3). Figure 6a and b shows the differences produced from the automatically simplified solution and the model after manual pruning of branches. The figure shows that the AUC for the training is 0.9096 and 0.8454 and the AUC for the test is 0.9082 and 0.8070 for the automatically simplified solution and the model after manual pruning, respectively. The AUC differences between both models for the training are 0.0642 (*standard error of the difference* = 0.0134) and for the test are 0.1012 (*standard error of the difference* = 0.0253). Both AUC differences showed statistical significance with *p* values = < 0.0001 and *z* statistics for the training $z = 4.781$ and tor the testing $z = 4.0033$. As a result, the impact of the removal of the two important variables (wetness index and solar radiation) in the model after manual pruning shows a reduction in the overall fit of the data, which contributes to a different model outcome.

4.3 Spatial prediction of landslide susceptibility

The predicted landslide susceptibility is shown in Fig. 7a and b for both the training and the validation study areas. The landslide susceptibility shown in the figure is solely based on the landslide initiation zones or the scarps. The predicted landslide susceptibility was derived by implementing the automatically simplified solution model from Fig. 5b. The visual representation of this dichotomous classification shows areas of low and high susceptibility where the blue color represents areas of low susceptibility, while the red color represents areas of high susceptibility. In the validation study area (Fig. 7b), the susceptibility classification implemented by the automatically simplified solution model generated

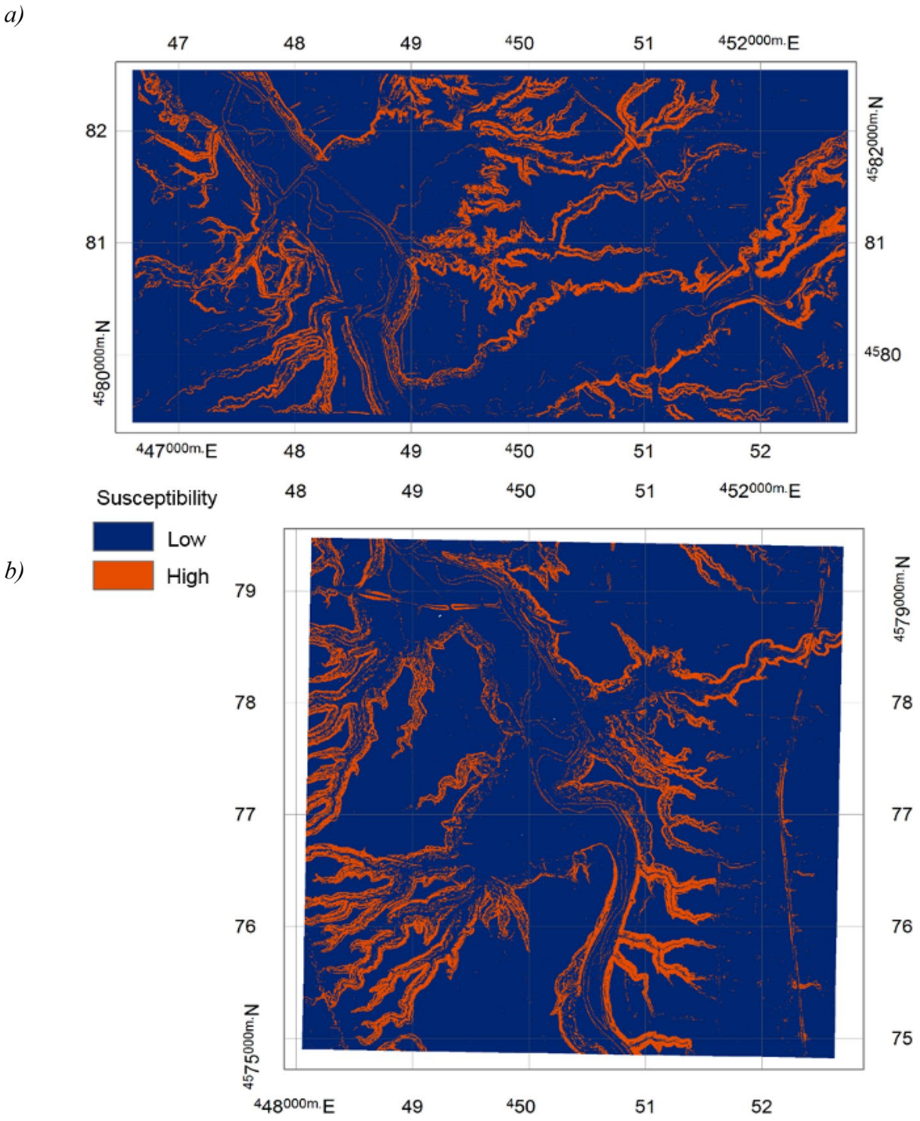


Fig. 7 Predicted landslide susceptibility derived by the automatically simplified solution where **a** is the training area used for the development of the GP model and **b** is the validation area used for testing the GP model developed from the training area

a total of 86.6 % low susceptibility pixels and a total of 13.4 % high susceptibility pixels (Table 4). Table 4 also shows the results from the implementation of the manual pruning solution model which generated a total of 77.5 % low susceptibility pixels and a total of 22.5 % high susceptibility pixels in the validation area. Interestingly from the total landslide scarp pixels only 0.94 % of the pixels were misclassified by the automatically simplified solution model, while 0.22 % of the pixels were misclassified by the manual pruning

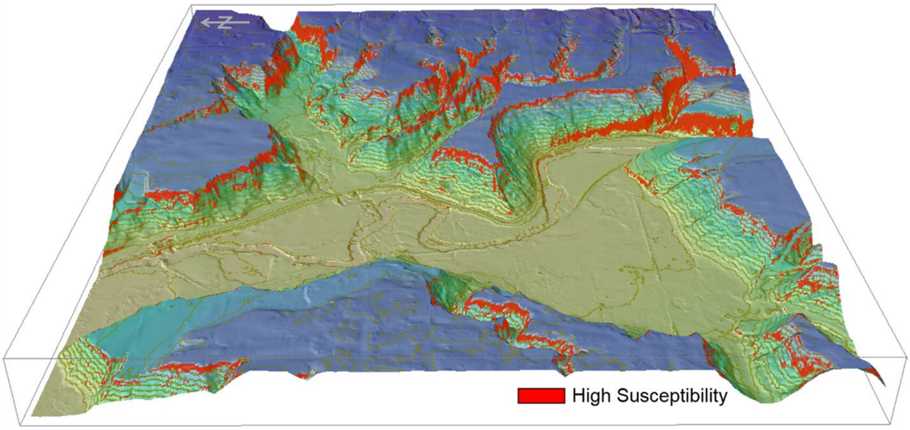


Fig. 8 A 3D visualization of landslide susceptibility associated with the validation area

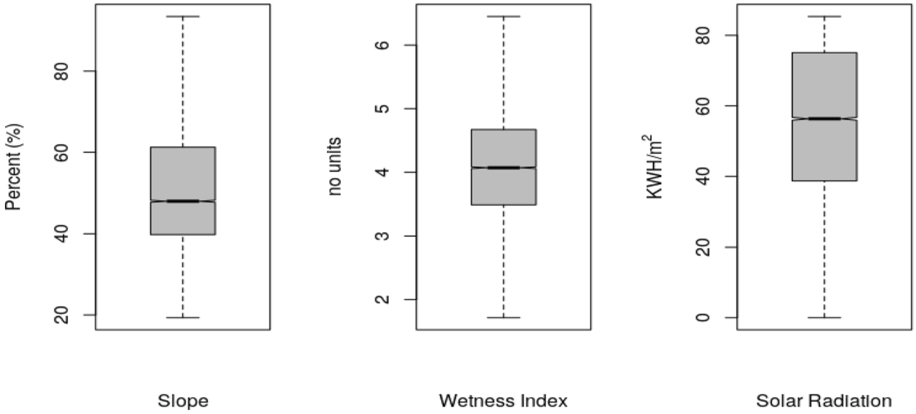


Fig. 9 Box plots of modeled high susceptibility areas vs. predictor attributes for the validation area

Table 2 Confusion matrix from the classification results produced by the original or automatically simplified solution from Fig. 5a or b

		Actual			
		Training		Test	
		Absence	Presence	Absence	Presence
Predicted	Absence	574	23	183	11
	Presence	179	724	64	242

Table 3 Confusion matrix from the classification results produced after manual pruning from Fig. 5c

		Actual			
		Training		Test	
		Absence	Presence	Absence	Presence
Predicted	Absence	561	25	170	10
	Presence	192	722	77	243

Table 4 Predicted areas associated with the susceptibility of landslides and non-landslides from the validation areas. The values are percentages from total pixels in the validation area and correspond to Fig. 7b

	Validation area	
	Automatically simplified solution	Manual pruning solution
Low susceptibility area	86.6 %	77.5 %
High susceptibility area	13.4 %	22.5 %

solution model. However, in comparison the automatically simplified solution model provided a classified proportion that is smaller by a total of 9.1 % in the high susceptibility area.

A close-up of three-dimensional (3-D) landslide susceptibility is shown in Fig. 8 where the areas of high susceptibility in red color are draped over an elevational terrain model. The north arrow on the map is pointing to the left showing susceptibility of valley walls that have exposed slopes mostly toward northern, northwestern, western, southwestern and southern directions. The 3-D visualization also depicts that the landslide susceptibility varies along the walls of the main valley and the smaller tributaries. For instance, the figure shows that most of the location of high susceptibility areas occur along the base and the rim of the valleys, while smaller portion of high susceptibility areas appear at different parts of the hillslopes. Additional analysis in the validation study area demonstrated that the proportion of susceptible cells with aspect direction pointing to northern, northwestern and western orientation was the highest, while the proportion of susceptible cells with aspect pointing to northeastern and eastern orientation was the lowest.

The high susceptibility areas associated with individual predictor attributes are illustrated as box plots in Fig. 9. Each box plot shows the distribution of predictor values that relate to high susceptibility. For instance, the plot for the slope attribute suggests that the high susceptibility class corresponds to slopes which range between 39.8 and 61.3 % based on interquartile range (IQR) (i.e., upper quartile–lower quartile) with mean value of 52.9 %. The distribution of IQR values for the wetness index ranges between 3.5 and 4.7 with mean value of 4.1, while the distribution of IQR values for the solar radiation ranges between 38.7 and 75.1 with mean value of 55.0 KWH/m^2. From the figure, the high susceptibility is associated with steeper slopes, relatively drier areas in terms of wetness as the rim of the valleys shown in Fig. 8, and lower annual solar radiation. The dominant orientations of the river valley within the validation area are in north and northwest directions (43.3 % of the valley walls), while a smaller portion of the valley walls are oriented in south and southwest directions (14.7 %). Such distributions influence amounts of sunlight received in the steep valley walls and subsequent soil moisture processes that influence slope susceptibility. For example, additional analysis of aspect directions associated with the solar radiation box plot shows that the lower quartile is representative of the northwest direction (mean = 36.8 KWH/m^2), the median is representative of the north direction (mean = 56.5 KWH/m^2), and the upper quartile is representative of the southeast direction (mean = 73.2 KWH/m^2). Since the validation area is dominated by north and northwest aspect directions, the lower values of the solar radiation govern the box plot.

5 Conclusions

The goal of this work was to test the capabilities of symbolic classification with GP for spatial prediction of landslide susceptibility. The input data employed in this research were LiDAR-based DEM that was used for the derivation of the topographic attributes (i.e., predictor variables) and to identify landslides (i.e., feature extraction) through high-resolution ground models, such as hillshades, combined with slope maps and draped topographic contours. The methodology was implemented in HeuristicLab, which is an evolution-inspired optimization framework that has a large set of algorithms for combinatorial optimization, including symbolic classification with GP.

The metaheuristic optimization approach tested a total of 72 runs generated by different parameter configurations where best quality solutions obtained by varying population size, tournament group size and mutation probability were evaluated. Additional criteria for the selection of the good solution included the relative variable frequency generated from multiple independent GP runs which depicted a total of three important variables including slope, wetness index and solar insulation. The spatial prediction of landslide susceptibility used mathematical representations of symbolic expression trees generated by the GP. The automatically simplified solution and the model after manual pruning solutions were used for the validation and assessment of the generalization error. The accuracy for the testing area was 85.0 % for the automatically simplified solution and 82.6 % for the manual pruning solutions, while the AUC from the ROC curves showed significant difference in the fit of the models. The AUC for the testing area was 0.9082 for the automatically simplified solution and 0.807 for the manual pruning solutions. The results also showed differences in classified solution associated with high susceptibility areas. The automatically simplified solution yielded a total of 13.4 % of the area, while the manual pruning solutions yielded a total of 22.5 % of the area. Evaluation of the areas with high susceptibility suggested that the potential for landslide occurrence was associated with steep slopes, relatively drier areas in terms of wetness as the rim of the valleys and lower annual solar radiation such as areas with north and northwestern expositions.

In summary, the implementation of the symbolic classification with GP in the validation area further demonstrated that the goodness-of-fit is relatively high. Almost all existing landslide scarps were predicted within a small portion of the study area which represented the high susceptibility class. Some of the highlighted advantages of the illustrated model include the ability to discern important subsets of predictor attributes among different runs and the ability to automatically generate appropriate nonlinear mathematical functions for predicting complex relationships for mapping landslide susceptibility. Besides the potential and the usefulness of this approach to support decision-makers in identifying landslide susceptibility, the interpretability of this approach is straightforward and intuitive. In addition, a large tree-structured GP models can be easily controlled by limiting the trees to some specified depth and size or by manual pruning while interactively assessing the impact of different sub-trees on the simplification of the model.

Possible directions for future work would be to implement multi-objective optimization, formulation of a fitness functions that can integrate both the spatial and the temporal distribution patterns, and additional testing of population size, the choice of important parameters, crossover and mutation methods. Other directions can be focused on big data challenges for processing and analyzing large-scale data using techniques such as parallelism or distributed computing.

Acknowledgements The author would like to thank the two anonymous reviewers, who provided helpful suggestions and excellent additions to the manuscript.

Funding This research received no external funding.

Data availability All datasets are derived from LiDAR-based DEMs.

Code availability Not applicable, "Free and open-source software" (FOSS) software used and referenced in the manuscript.

Declarations

Conflict of interest The author declares no conflicts of interest

References

Affenzeller M, Wagner S, Winkler S, Beham A (2009) Genetic algorithms and genetic programming: modernconcepts and practical applications. CRC Press, Boca Raton

Atkinson PM, Massari R (2011) Autologistic modelling of susceptibility to landsliding in the central apennines. Italy Geomorphol 130:55–64. https://doi.org/10.1016/j.geomorph.2011.02.001

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. https://doi.org/10.1016/j.geomorph.2004.06.010

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

Bai S-B, Wang J, Lü G-N et al (2010) GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology 115:23–31. https://doi.org/10.1016/j.geomorph.2009.09.025

Bathurst JC, Bovolo CI, Cisneros F (2010) Modelling the effect of forest cover on shallow landslides at the river basin scale. Ecol Eng 36:317–327. https://doi.org/10.1016/j.ecoleng.2009.05.001

Blickle T, Thiele L (1996) A comparison of selection schemes used in evolutionary algorithms. Evol Comput 4:361–394. https://doi.org/10.1162/evco.1996.4.4.361

Budimir MEA, Atkinson PM, Lewis HG (2015) A systematic review of landslide probability mapping using logistic regression. Landslides 12:419–436. https://doi.org/10.1007/s10346-014-0550-5

Buffat R, Froemelt A, Heeren N et al (2017) Big data GIS analysis for novel approaches in building stock modelling. Appl Energy 208:277–290. https://doi.org/10.1016/j.apenergy.2017.10.041

Carrara A, Cardinali M, Guzzetti F, Reichenbach P (1995) Gis technology in mapping landslide hazard. Geographical information systems in assessing natural hazards. Springer, Dordrecht, pp 135–175

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831. https://doi.org/10.5194/nhess-13-2815-2013

Chawla S, Shekhar S, Wu W, Ozesmi U (2001) Modeling Spatial Dependencies for Mining Geospatial Data. In: Proceedings of the 2001 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp 1–17

Chipperfield A, Fleming P, Pohlheim H (1994a) A genetic algorithm toolbox for MATLAB. Proc Int Conf Syst Eng 200–207

Chung C-JF, Fabbri AG, Westen CJV (1995) Multivariate regression analysis for landslide hazard zonation. Geographical information systems in assessing natural hazards. Springer, Dordrecht, pp 107–133

Dai FC, Lee CF (2003) A spatiotemporal probabilistic modelling of storm-induced shallow landsliding using aerial photographs and logistic regression. Earth Surf Process Landf 28:527–545. https://doi.org/10.1002/esp.456

Dai FC, Lee C-F (2002) Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology 42:213–228.

Deng C, Pan H, Fang S et al (2017) Support vector machine as an alternative method for lithology classification of crystalline rocks. J Geophys Eng 14:341–349. https://doi.org/10.1088/1742-2140/aa5b5b

Ercanoglu M, Temiz FA (2011) Application of logistic regression and fuzzy operators to landslide susceptibility assessment in Azdavay (Kastamonu, Turkey). Environ Earth Sci 64:949–964. https://doi.org/10.1007/s12665-011-0912-4

Feizizadeh B, Blaschke T (2013) GIS-multicriteria decision analysis for landslide susceptibility mapping: comparing three methods for the Urmia lake basin, Iran. Nat Hazards 65:2105–2128. https://doi.org/10.1007/s11069-012-0463-3

Gokceoglu C, Sezer E (2009) A statistical assessment on international landslide literature (1945–2008). Landslides 6:345–351. https://doi.org/10.1007/s10346-009-0166-3

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading

Gorsevski P (2002) Landslide Hazard Modeling Using GIS. Ph.D. dissertation. University of Idaho, Moscow

Gorsevski P, Gessler P, Jankowski P (2010) A Fuzzy k Means Classification and a Bayesian Approach for Spatial Prediction of Landslide Hazard. https://doi.org/10.1007/978-3-642-03647-7_31

Gorsevski P, Gessler PE, Foltz RB (2000) Spatial prediction of landslides hazard using logistic regression and GIS. 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4), Problems, Prospects and Research Needs, Banff, Alberta, Canada. September 2–8

Gorsevski PV (2013) Using Bayesian inference to account for uncertainty in parameter estimates in modelled invasive flowering rush. Remote Sens Lett 4:279–287. https://doi.org/10.1080/2150704X.2012.724539

Gorsevski PV, Brown MK, Panter K et al (2016) Landslide detection and susceptibility mapping using LiDAR and an artificial neural network approach: a case study in the Cuyahoga Valley National Park, Ohio. Landslides 13:467–484. https://doi.org/10.1007/s10346-015-0587-0

Gorsevski PV, Gessler PE, Boll J et al (2006a) Spatially and temporally distributed modeling of landslide susceptibility. Geomorphology 80:178–198. https://doi.org/10.1016/j.geomorph.2006.02.011

Gorsevski PV, Gessler PE, Foltz RB, Elliot WJ (2006b) Spatial prediction of landslide hazard usinglogistic regression and ROC analysis. Trans GIS 10:395–415. https://doi.org/10.1111/j.1467-9671.2006.01004.x

Gorsevski PV, Gessler PE, Jankowski P (2003) Integrating a fuzzy k -means classification and a Bayesian approach for spatial prediction of landslide hazard. J Geogr Syst 5:223–251. https://doi.org/10.1007/s10109-003-0113-0

Gorsevski PV, Jankowski P (2008) Discerning landslide susceptibility using rough sets. Comput Environ Urban Syst 32:53–65. https://doi.org/10.1016/j.compenvurbsys.2007.04.001

Gorsevski PV, Jankowski P (2010) An optimized solution of multi-criteria evaluation analysis of landslide susceptibility using fuzzy sets and Kalman filter. Comput Geosci 36:1005–1020. https://doi.org/10.1016/j.cageo.2010.03.001

Gorsevski PV, Jankowski P, Gessler PE (2005) Spatial Prediction of Landslide Hazard Using Fuzzy k-means and Dempster-Shafer Theory. Trans GIS 9:455–474. https://doi.org/10.1111/j.1467-9671.2005.00229.x

Gorsevski PV, Jankowski P, Gessler PE (2006c) An heuristic approach for mapping landslide hazard by integrating fuzzy logic with analytic hierarchy process. Control Cybern 35:121–146

Gotshall S, Rylander B (2002) Optimal Population Size and the Genetic Algorithm. In: WSEAS 2002. Interlaken, Switzerland, p 5

GRASS (2018) Geographic Resources Analysis Support System (GRASS) GIS. https://grass.osgeo.org. Accessed 4 Jun 2018

Gupta RP, Joshi BC (1990) Landslide hazard zoning using the GIS approach—a case study from the Ramganga catchment. Himalayas Eng Geol 28:119–131. https://doi.org/10.1016/0013-7952(90)90037-2

Hamblin S (2012) On the practical usage of genetic algorithms in ecology and evolution. Methods Ecol Evol 11:598

Hansen M (1995) Landslides in Ohio. https://www.dnr.state.oh.us/Portals/10/pdf/GeoFacts/geof08.pdf. Accessed 29 May 2018

Hengl T, Reuter HI (2009) Geomorphometry: concepts, software, applications. development in soil science 33. Elsevier, Amsterdam, p 772. https://www.sciencedirect.com/bookseries/developments-in-soil-science

HeuristicLab (2018) Heuristic and Evolutionary Algorithms Laboratory (HEAL). https://dev.heuristiclab.com/trac.fcgi/. Accessed 1 Jun 2018

Holland JH (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, Michigan

Holland JH (1992) Adaptation in natural and artificial systems: anintroductory analysis with applications to biology, control, and artificialintelligence. MIT Press, Cambridge

Jayawardhana UK, Gorsevski PV (2019) An ontology-based framework for extracting spatio-temporal influenza data using Twitter. Int J Digit Earth 12:2–24. https://doi.org/10.1080/17538947.2017.1411535

Kavzoglu T, Kutlug Sahin E, Colkesen I (2015) Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Eng Geol 192:101–112. https://doi.org/10.1016/j.enggeo.2015.04.004

Kommenda M, Kronberger G, Wagner S et al (2012) On the architecture and implementation of tree-based genetic programming in HeuristicLab. ACM Press, New York

Korup O, Stolle A (2014) Landslide prediction from machine learning. Geol Today 30:26–33. https://doi.org/10.1111/gto.12034

Koza JR (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection, 1 edition. A Bradford Book, Cambridge, Mass

Krušić J, Marjanović M, Samardžić-Petrović M et al (2017) Comparison of expert, deterministic and Machine Learning approach for landslide susceptibility assessment in Ljubovija Municipality, Serbia. Geofizika 34:251–273. https://doi.org/10.15233/gfz.2017.34.15

Lozano M, Herrera F, Cano J (2008) Replacement strategies to preserve useful diversity in steady-state genetic algorithms. Inf Sci 178:4421–4433. https://doi.org/10.1016/j.ins.2008.07.031

Marjanović M, Kovačević M, Bajat B, Voženílek V (2011) Landslide susceptibility assessment using SVM machine learning algorithm. Eng Geol 123:225–234. https://doi.org/10.1016/j.enggeo.2011.09.006

Micheletti N, Foresti L, Robert S et al (2014) Machine learning feature selection methods forlandslide susceptibility mapping. Math Geosci 46:33–57. https://doi.org/10.1007/s11004-013-9511-0

Miller BL, Goldberg DE (1995) Genetic Algorithms, Tournament Selection, and the Effects of Noise. 20

Mitchell TM (1997) Machine learning. McGraw-Hill, New York

Nandi A, Shakoor A (2010) A GIS-based landslide susceptibility evaluation using bivariate and multivariate statistical analyses. Eng Geol 110:11–20. https://doi.org/10.1016/j.enggeo.2009.10.001

OGRIP (2018) Ohio Geographically Referenced Information Program. http://ogrip.oit.ohio.gov/. Accessed 31 May 2018

Park N-W (2011) Application of Dempster-Shafer theory of evidence to GIS-based landslide susceptibility analysis. Environ Earth Sci 62:367–376. https://doi.org/10.1007/s12665-010-0531-5

Poli R (2001) Exact Schema Theory for Genetic Programming and Variable-Length Genetic Algorithms with One-Point Crossover. 41

Poli R, Langdon WB, McPhee NF, Koza JR (2008) A field guide to genetic programming. Lulu Press, Morrisville

Reichenbach P, Rossi M, Malamud BD et al (2018) A review of statistically-based landslide susceptibility models. Earth-Sci Rev 180:60–91. https://doi.org/10.1016/j.earscirev.2018.03.001

Ronco CCD, Benini E (2014) A simplex-crossover-based multi-objective evolutionary algorithm. IAENG transactions on engineering technologies. Springer, Dordrecht, pp 583–598

SAGA (2018) System for Automated Geoscientific Analyses (SAGA) GIS. http://www.saga-gis.org/en/index.html. Accessed 4 Jun 2018

Saporetti CM, da Fonseca LG, Pereira E (2019) A lithology identification approach based onmachine learning with evolutionary parameter tuning. IEEE Geosci Remote Sens Lett 16:1819–1823. https://doi.org/10.1109/LGRS.2019.2911473

Saro L, Woo JS, Kwan-Young O, Moung-Jin L (2016) The spatial prediction of landslide susceptibility applying artificial neural network and logistic regression models: A case study of inje, Korea. Open Geosci. https://doi.org/10.1515/geo-2016-0010

Song K-Y, Oh H-J, Choi J et al (2012) Prediction of landslides using ASTER imagery and data mining models. Adv Space Res 49:978–993. https://doi.org/10.1016/j.asr.2011.11.035

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

Szabo JP (1987) Wisconsinan stratigraphy of the Cuyahoga Valley in the Erie Basin, northeastern Ohio. Can J Earth Sci 24:279–290. https://doi.org/10.1139/e87-029

Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using Random Forest. Big Earth Data 2:159–178. https://doi.org/10.1080/20964471.2018.1472392

Tien Bui D, Ho TC, Revhaug I et al (2014) Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based J48 decision tree classifier and its ensembles. In: Buchroithner M, Prechtel N, Burghardt D et al (eds) Cartography from pole to pole. Springer Berlin Heidelberg, Berlin, pp 303–317

Tien Bui D, Tuan TA, Hoang N-D et al (2017) Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. Landslides 14:447–458. https://doi.org/10.1007/s10346-016-0711-9

Tien Bui D, Tuan TA, Klempe H et al (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13:361–378. https://doi.org/10.1007/s10346-015-0557-6

Tong X, Zhang X, Liu M (2010) Detection of urban sprawl using a genetic algorithm-evolved artificial neural network classification in remote sensing: a case study in Jiading and Putuo districts of Shanghai, China. Int J Remote Sens 31:1485–1504. https://doi.org/10.1080/01431160903475290

Tsai F, Lai J-S, Chen WW, Lin T-H (2013) Analysis of topographic and vegetative factors with data mining for landslide verification. Ecol Eng 61:669–677. https://doi.org/10.1016/j.ecoleng.2013.07.070

Wagner S, Kronberger G, Beham A et al (2014) Architecture and design of the HeuristicLab optimization environment. In: Klempous R, Nikodem J, Jacak W, Chaczko Z et al (eds) Advanced methods and applications in computational intelligence. Springer International Publishing, Heidelberg, pp 197–261

Westen CJ van, Rengers N, Terlien MTJ, Soeters R (1997) Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. Geol Rundsch 86:404–414. https://doi.org/10.1007/s005310050149

Wilson J, Gallant J (2000) Terrain Analysis: Principles and Applications

Winkler S, Affenzeller M, Wagner S (2007) Advanced genetic programming based machine learning. J Math Model Algorithms 6:455–480. https://doi.org/10.1007/s10852-007-9065-6

Winkler SM, Affenzeller M, Wagner S (2009) Using enhanced genetic programming techniques for evolving classifiers in the context of medical diagnosis. Genet Program Evolvable Mach 10:111–140. https://doi.org/10.1007/s10710-008-9076-8

Yang M-D (2007) A genetic algorithm (GA) based automated classifier for remote sensing imagery. Can J Remote Sens 33:203–213. https://doi.org/10.5589/m07-020

Yang M-D, Yang Y-F, Su T-C, Huang K-S (2014) An efficient fitness function in genetic algorithm classifier for Landuse recognition on satellite images. Sci World J. https://doi.org/10.1155/2014/264512

Yao H, Hamilton HJ (2008) Mining functional dependencies from data. Data Min Knowl Discov 16:197–219. https://doi.org/10.1007/s10618-007-0083-9

Zêzere JL, Pereira S, Melo R et al (2017) Mapping landslide susceptibility using data-driven methods. Sci Total Environ 589:250–267. https://doi.org/10.1016/j.scitotenv.2017.02.188

Zojaji Z, Ebadzadeh MM (2016) Semantic schema theory for genetic programming. Appl Intell 44:67–87. https://doi.org/10.1007/s10489-015-0696-4