Geomechanics and Engineering, *Vol. 36, No. 5 (2024) 441-453*
https://doi.org/10.12989/gae.2024.36.5.441

Prediction of karst sinkhole collapse using a decision-tree (DT) classifier

Boo Hyun Nam[1a], Kyungwon Park[1b] and Yong Je Kim*[2]

[1]*Department of Civil Engineering, College of Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Republic of Korea*
[2]*Department of Civil and Environmental Engineering, Lamar University, 4400 MLK Blvd., Beaumont, TX 77710, USA*

(*Received October 18, 2023, Revised February 9, 2024, Accepted February 11, 2024*)

Abstract. Sinkhole subsidence and collapse is a common geohazard often formed in karst areas such as the state of Florida, United States of America. To predict the sinkhole occurrence, we need to understand the formation mechanism of sinkhole and its karst hydrogeology. For this purpose, investigating the factors affecting sinkholes is an essential and important step. The main objectives of the presenting study are (1) the development of a machine learning (ML)-based model, namely C5.0 decision tree (C5.0 DT), for the prediction of sinkhole susceptibility, which accounts for sinkhole/subsidence inventory and sinkhole contributing factors (e.g., geological/hydrogeological) and (2) the construction of a regional-scale sinkhole susceptibility map. The study area is east central Florida (ECF) where a cover-collapse type is commonly reported. The C5.0 DT algorithm was used to account for twelve (12) identified hydrogeological factors. In this study, a total of 1,113 sinkholes in ECF were identified and the dataset was then randomly divided into 70% and 30% subsets for training and testing, respectively. The performance of the sinkhole susceptibility model was evaluated using a receiver operating characteristic (ROC) curve, particularly the area under the curve (AUC). The C5.0 model showed a high prediction accuracy of 83.52%. It is concluded that a decision tree is a promising tool and classifier for spatial prediction of karst sinkholes and subsidence in the ECF area.

Keywords: C5.0 decision tree; karst sinkhole; sinkhole susceptibility prediction

1. Introduction

Karst topography refers to a terrain that results from the erosion and breakdown of soluble rock formations by groundwater through chemical or physical processes. The types of rock commonly involved in this process are limestone, dolomite, and gypsum, which are rich in carbonates. Sinkholes, springs, and caverns are characteristic features of karst topography. The United States Geological Survey (USGS) estimates that approximately 20% of the country's land surface is composed of karst topography, with significant examples found in Florida (Weary and Doctor 2014).

In karst terrain, sinkholes are prevalent and frequent, representing a common natural geohazard in such areas. (Zhou *et al.* 2015, Yuan *et al.* 2016, Genis *et al.* 2018, Xu *et al.* 2021). The failure mechanism of a sinkhole is similar to that of a tunnel, and sinkhole stability analysis generally involves an underground void, just like the stability analysis of tunnels (Kim *et al.* 2020, Soliman *et al.* 2019). It poses a threat not only to infrastructure and personal property but also to human safety. Each year in the United States, sinkholes cause damages amounting to a minimum of $300 million, although the actual extent of the damage surpasses this estimate, as stated by Weary in 2015. The Florida Office of Insurance Regulation (FOIR) provided data indicating that a total of 24,671 claims were reported for sinkhole-related damages, in only Florida, from 2006 to 2010, totaling $1.4 billion. This data also reveals a rising trend in the frequency and severity levels of sinkholes, leading to increased expenses for insurers (FOIR 2010).

Sinkholes that naturally occur in central Florida are caused by a combination of soluble/weathered bedrock, groundwater flow (or recharge), and soil conditions. Over geological time, fine-grained soils erode from the interface between the cover soil and bedrock, moving into the cavities within the bedrock. In regions where there exists a favorable hydraulic gradient, characterized by a higher water level in the surficial aquifer compared to the potentiometric surface of the Floridan aquifer, the hydraulic slopes play a significant role in facilitating the occurrence of internal erosion in concentrated zones. This erosion is driven by seepage forces and subsequent piping, as explained by Beck and Sinclair in 1986. Fig. 1 shows examples of sinkholes in central Florida and the conceptual progression and profile of cover collapse. The cover-collapse type is the most common type of sinkhole that occurs naturally in central Florida. This cover-collapse type is commonly observed under specific geological conditions where relatively thick cohesive strata are covered by sand. Internal erosion of soil within these “clay” layers enables the formation of cavities, but without a pre-sign on the surface, as finer particles of soil migrate into cavities in the limestone bedrock. The zone of this internal-erosion soil is

*Corresponding author, Professor
E-mail: ykim3@lamar.edu
[a]Professor
E-mail: boohyun.nam@khu.ac.kr
[b]Student
E-mail: kwpark@khu.ac.kr

ISSN: 2005-307X (Print), 2092-6219 (Online)

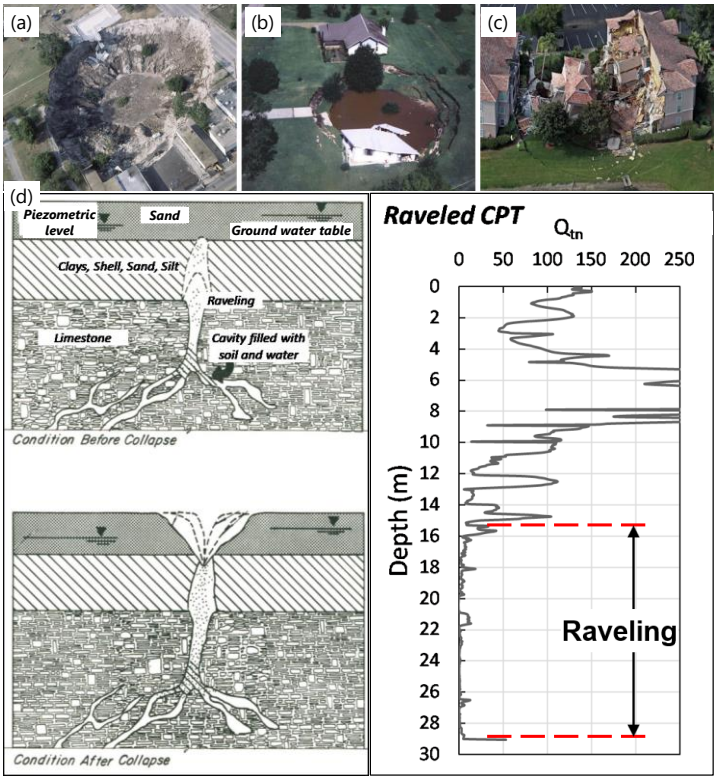

Fig. 1 Sinkhole mechanism: (a) Winter Park sinkhole (1981), (b) Crooked Lake sinkhole (1991), (c) Clermont sinkhole near Disney World (2013), and (d) raveling mechanism and CPT data

referred to as the zone of soil raveling (Shamet *et al.* 2018). Those cavities are often treated by grouting methods (Tacim *et al.* 2023).

Given the increasing economic and social impacts of sinkhole hazards, it is crucial to have reliable methods for evaluating these hazards. This assessment is critical in helping civil engineers not only design civil infrastructure but also establish mitigation strategies that effectively reduce the risks of casualties and infrastructure damage. Sinkhole susceptibility mapping emerges as a valuable approach for evaluating sinkhole hazards and ensuring their accurate assessment. Multiple methods have been developed for mapping the sinkhole susceptibility, such as heuristic (Taheri, Gutiérrez *et al.* 2015, Subedi *et al.* 2019), frequency ratio (Yilmaz 2007, Ozdemir 2015, Kim *et al.* 2020), logistic regression (Papadopoulou-Vrynioti *et al.* 2013, Ciotoli *et al.* 2016, Subedi *et al.* 2019, Naithani *et al.* 2022, Kim *et al.* 2022), artificial neural network (Yilmaz *et al.* 2013), and deterministic methods (Galve *et al.* 2009, Strzałkowski 2018). Some of these methods were adopted from landslide susceptibility analysis (Ding *et al.* 2019, Liu *et al.* 2021, Nanehkaran *et al.* 2021).

Sinkhole susceptibility mapping involves analyzing large and complex datasets. Data mining and machine learning techniques excel in handling such intricate data volumes, and among them, the C5.0 decision tree (C5.0 DT) algorithm has gained substantial popularity in susceptibility mapping applications. This rising preference for C5.0 DT can be attributed to its key characteristics. Its ability to generate transparent and interpretable decision rules resonates with the need for comprehensible models in sinkhole susceptibility studies, allowing stakeholders to readily understand the driving factors behind sinkhole occurrences. Furthermore, C5.0 DT's ease of use compared to other algorithms is advantageous, especially when dealing with limited resources or large datasets, often encountered in sinkhole assessments. Additionally, its competitive performance in terms of accuracy, demonstrated by studies such as Guo *et al*. (2021) achieving a high area under the curve (AUC) of 0.883, and its inherent versatility in handling diverse data types further establish its suitability for sinkhole susceptibility analysis. These combined features make C5.0 DT a well-suited choice for developing transparent and reliable sinkhole susceptibility models. This study presents the use of a C5.0 decision tree classifier in the assessment of sinkhole susceptibility by analyzing a number of sinkhole-affecting factors and the development of the sinkhole susceptibility model/map of the ECF area.

2. Study area and sinkhole contributing factors

2.1 Study area

Florida is considered to be very susceptible to sinkholes compared to other states in the United States. This vulnerability can be attributed to various factors, including the state's hydrogeology, geomorphology, climate, and human activities such as groundwater extraction for drinking water and irrigation purposes. The Florida Geological Survey (FGS) has documented a total of over

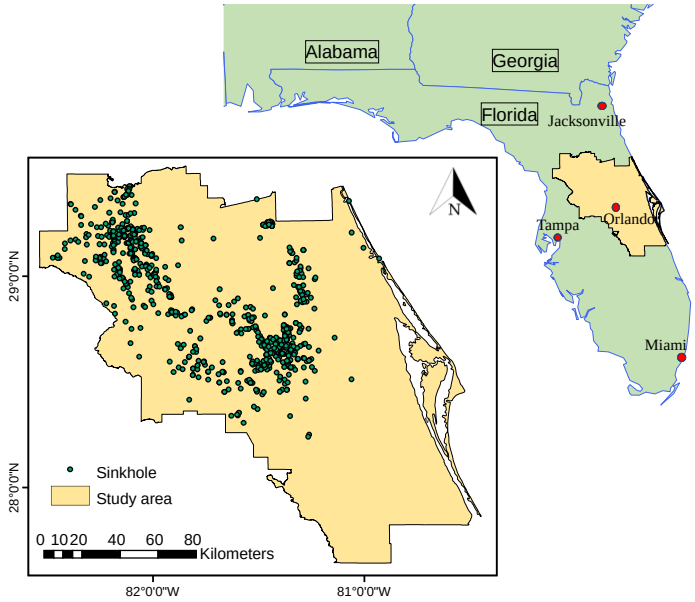

Fig. 2 Map showing the study area with the reported sinkholes (ECF area)

four thousand sinkholes that have been reported within the state of Florida since the 1950s (FDEP 2022). The region of Central Florida is widely recognized for its high incidence of sinkholes and the significant magnitude of resulting damages. The study area is delineated as the Central District, as designated by the Florida Department of Environmental Protection (FDEP). The area under consideration is situated within the ECF region, with its geographical coordinates ranging from 29º25' to 27º37' N in terms of latitude, and from 82º33' to 80º23' W in terms of longitude. The total land area covered by this region is approximately 22,010 square kilometers, as depicted in Fig. 2.

The geographical area exhibits predominantly low-lying and level terrain in close proximity to karst topographic features, including sinkholes, springs, sinking streams, and karst windows (depressions in the earth). The geological composition of Florida primarily consists of sedimentary rocks, with no presence of igneous or metamorphic rocks inside the state. The primary geological formation in the study area consists of limestone, with a layer of impermeable clay above the bedrock. The hydrostratigraphic units of ECF comprise a sequence of aquifer systems, namely the surficial aquifer system (SAS), intermediate aquifer system (IAS), and Floridan aquifer system (FAS), arranged in a vertical succession. The ECF region is subject to a variety of sinkhole dangers as a result of the interplay of hydrogeological, geological, and climatic factors. To present, a total of 1,113 sinkholes have been documented within the study area.

2.2 Sinkhole contributing factors

The authors investigated the key factors that contribute to sinkhole formation prior to the development of a sinkhole susceptibility model. Two significant hydrogeological factors that impact sinkhole development in karst terrains are the difference in hydraulic head and the rate of groundwater recharge. Additionally, the thickness of overlying materials plays a crucial role in determining the types of sinkholes that may occur. (Tihansky 1999). Additionally, various factors affecting the occurrence of sinkholes, such as soil permeability, the thickness of each aquifer unit, proximity to existing karst features, geology and geomorphology, precipitation, and land use and land cover patterns, were considered. In this study, for the C5.0 DT-based modeling and susceptibility mapping, a total of twelve factors associated with sinkhole formation were utilized. These factors include hydraulic head difference, groundwater recharge rate, soil permeability, proximity to or distance from karst features, overburden thickness, SAS thickness, IAS thickness, annual precipitation, geology, geomorphology, lithology, and land use land cover (LULC). All data of sinkhole contributing factors are represented in a raster format with a pixel size of 60 x 60 m, as shown in Fig. 3. In addition, a sinkhole inventory map was prepared by using Florida Subsidence Incident Reports (FSIR) published by FGS.

3. Sinkhole susceptibility model

3.1 Methodology framework

The present study employs a systematic five-part approach framework for assessing sinkhole susceptibility, as depicted in Fig. 4.

(1) The comprehensive acquisition of sinkhole inventory information and identification of twelve relevant elements involved an intricate process. A thorough literature review was conducted to gather existing knowledge, complemented by expert consultations, including engineering geologists, hydrogeologists and geotechnical engineers, for nuanced insights. Accessibility to relevant data sources was ensured

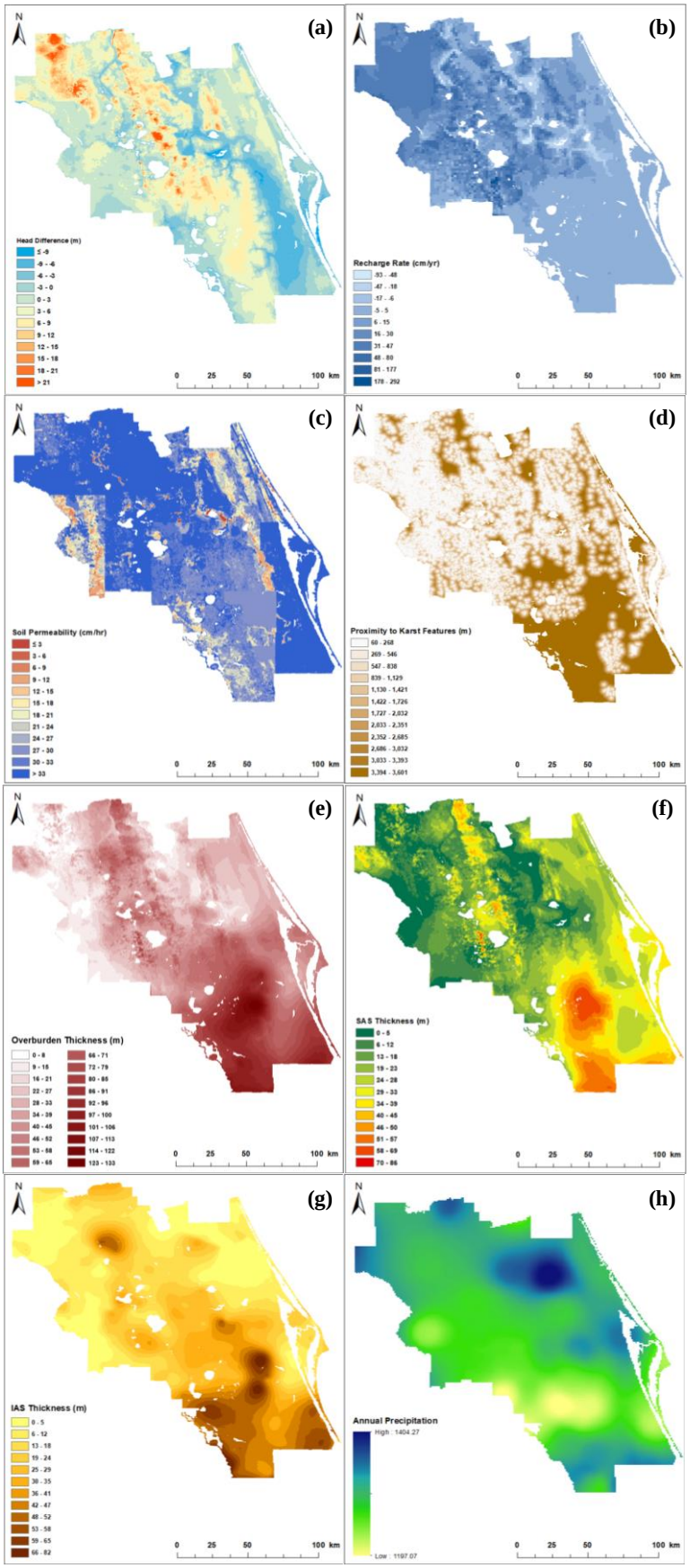

Fig. 3 Contributing factor maps used for sinkhole susceptibility modeling; (a) head difference, (b) recharge rate, (c) soil permeability, (d) proximity to karst features, (e) overburden thickness, (f) SAS thickness, (g) IAS thickness, (h) annual precipitation, (i) geology, (j) geomorphology, (k) lithology, and (l) land use land cover

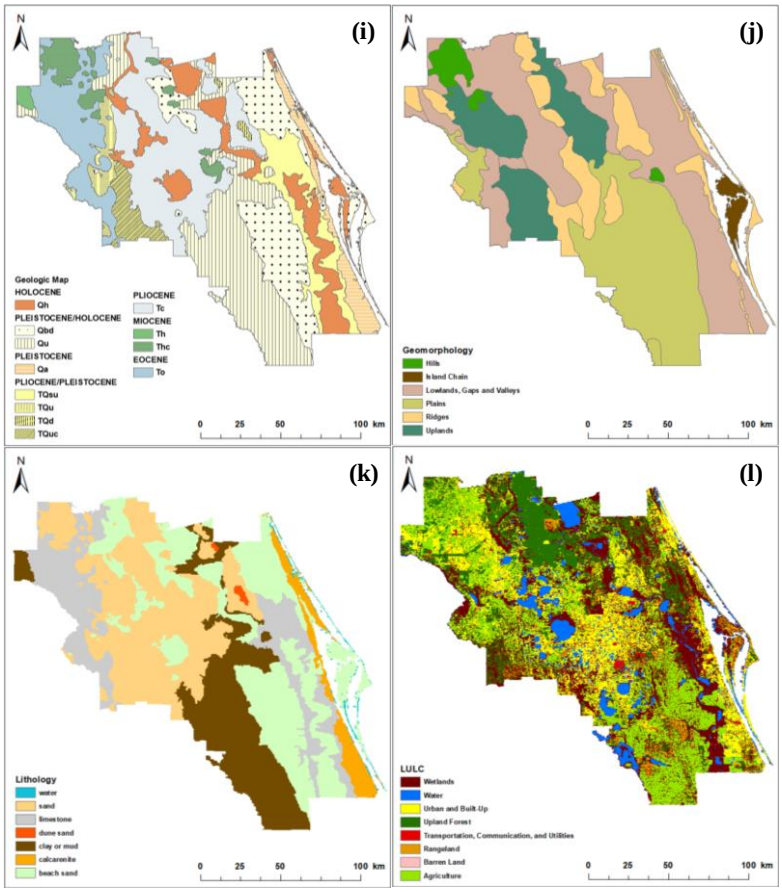

Fig. 3 Continued-

to capture diverse perspectives. Rigorous statistical analysis was employed, going beyond mere descriptive statistics, to delve into intricate relationships and patterns within the acquired data.(2) The utilization of the Frequency Ratio (FR) went beyond a mere correlation assessment; it involved a meticulous exploration of non-linear relationships between each contributing factor and the relative density of sinkholes. This process extended to recognizing and evaluating the significance of each contributing factor, providing a nuanced understanding of their impact on sinkhole occurrence.

(3) The C5.0 decision tree model utilized Feature Ratios (FR) of sinkhole contributing factors as input variables. These FR values underwent a detailed calculation considering the number of sinkholes associated with each factor and were subjected to advanced statistical analysis, extending beyond simple Pearson's correlation. The output variables represented sinkhole data, denoted by 0 for absence and 1 for presence. The establishment of mathematical correlation between these input (independent) and output (dependent) variables involved rigorous statistical methods, providing a nuanced understanding of their relationship.

(4) Input and output variables underwent strategic partitioning into modeling/training (70%) and validation/test (30%) datasets through a meticulous random allocation method with specified seed values. This process ensured transparency and reproducibility in subsequent analyses. Notably, the partitioning addressed potential imbalances by maintaining proportional representation of sinkhole presence and absence. For datasets with temporal information, a temporal splitting strategy preserved chronological order. Within the training dataset, cross-validation techniques, such as k-fold or leave-one-out, were employed for a comprehensive model assessment. While the training dataset played a central role in model construction, the deliberate exclusion of the test dataset preserved its integrity for unbiased evaluation. Evaluation on the test dataset included precision, recall, F1-score, and other metrics, offering a thorough understanding of the model's predictive capabilities.

(5) The evaluation of sinkhole susceptibility prediction involved the detailed utilization of the Receiver Operating Characteristic (ROC) curve. This comprehensive analysis included systematic threshold variation, consideration of optimal thresholds based on specific context, and subsequent calculation of the Area Under the Curve (AUC). The AUC provided a quantitative measure of the model's overall discriminatory power, considering values closer to 1 as indicative of superior predictive performance. Confidence intervals around the AUC were calculated to assess statistical significance and enhance the reliability of the evaluation.

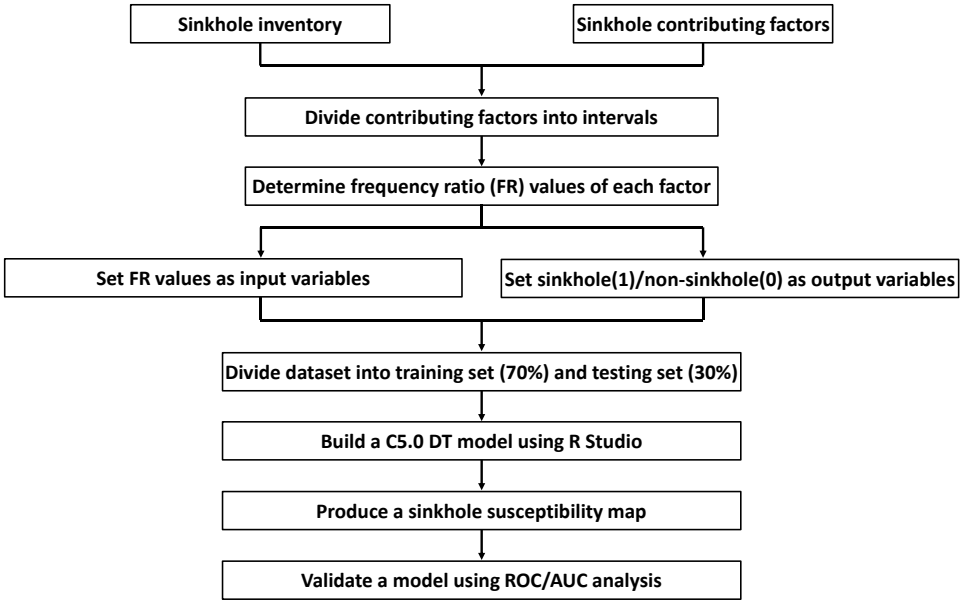

Fig. 4 Flowchart of the research methodology

3.2 Frequency ratio

The FR method enables the examination of the correlation between sinkholes and many contributing factors. Additionally, it facilitates the evaluation of the relative impact of different attribute intervals within these factors on the incidence of sinkholes. A FR value exceeding 1.0 signifies that the associated contributing factor actively facilitates the creation of sinkholes. On the other hand, a FR value nearing 1.0 indicates a limited correlation between the factor and the incidence of sinkholes within that specific range. Values of FR that are less than 1.0 imply a low probability of sinkhole occurrence within the specified attribute interval. The sinkhole contributing factors are categorized into distinct attribution interval levels by employing the FR approach in conjunction with the natural break method. The categorical factors are classified into intervals depending on their observed conditions. The formula used to calculate the FR is as follows

$$FR = \frac{A_i/A}{B_i/B} \tag{1}$$

The formula given denotes that A_i indicates the numerical value indicating the quantity of sinkhole grids present inside each interval for a particular factor category. The variable A represents the aggregate number of sinkhole grids within the specified region. The variable B_i denotes the cumulative number of grids falling inside the respective interval for each factor type, whilst B represents the total count of grids including the entire study region. FR represents the FR values associated with the factors under consideration.

3.3 C5.0 decision tree

According to Gkioulekas and Papageorgiou (2021), the decision tree (DT) approach is a very effective machine learning algorithm that is recognized for its adaptability and comprehensive characteristics. The application of this algorithm has been observed in numerous practical fields, such as radar signal classification, character identification, remote sensing, medical diagnostics, expert systems, speech recognition, and others (Safavian and Landgrebe 1991).

According to Sheng *et al*. (2022), the DT model demonstrates proficiency in analyzing and predicting the optimal label value for each pixel eigenvalue node. This capability facilitates the classification of the dataset by utilizing these labels. Significantly, the size of the dataset does not have an effect on the size of the decision tree model, rendering it appropriate for generating decision trees with substantial quantities of data. The objective of this study is to construct a classification model utilizing the C5.0 algorithm, known for its proficiency in categorizing and regulating extensive datasets while imparting significant insights. In contrast to other algorithms, C5.0 exhibits a notable aptitude for effectively managing large volumes of data. During the execution of the classification model, each classifier is assigned to a specific "leaf" (as depicted in Figure 5) and allocated to the class with the highest frequency. After the regression-tree processing step of the algorithm, an examination and rectification of missing data is conducted. Afterward, the classification tree proceeds to reclassify the data, resulting in the creation of an optimal binary tree. The determination of the ideal division or split value for each node is based on referencing the Gini coefficient of the economic category. The Gini coefficient is used to measure the inequality within this economic category, and the algorithm chooses splits that minimize the Gini impurity, leading to the creation of an optimal binary tree for classification within the context of economic data. The formula employed for this objective is

$$\omega(X) = \frac{Gini(X)}{\sum_1^n a_i} \tag{2}$$

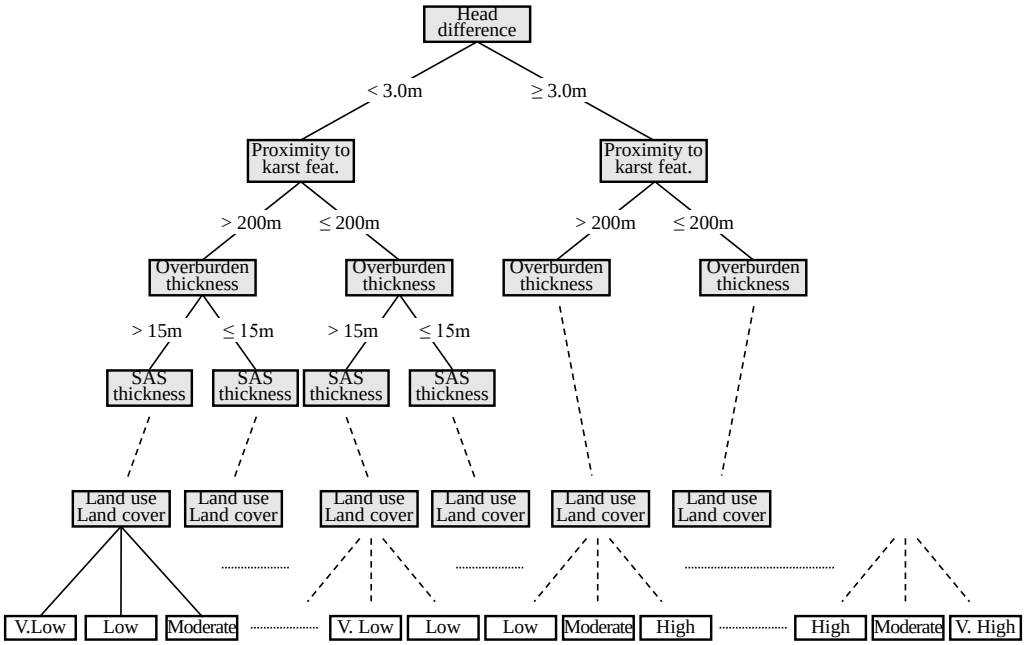

Fig. 5 Decision tree-based model

where n = number of branches, and a_i = number of leaves (for i branch).

4. Results

4.1 Frequency ratio analysis

The study employed the FR approach to evaluate the correlation between the occurrence of sinkholes and each factor that contributes to sinkhole formation in the study area. The results of this analysis are summarized in Table 1. Higher FR values signify stronger correlations with sinkhole occurrence, especially when surpassing 1.0, indicating a particularly robust association. Nevertheless, the attainable maximum FR is contingent upon the specific data and analysis methods utilized.The analysis of FR values reveals a clear correlation between the majority of parameters and the occurrence of sinkholes, hence highlighting their significance in the underlying process. While sinkhole occurrences are frequently documented in a particular area, the FR value typically exhibits a corresponding increase or reduction in response to changes in the associated factor's value.

4.2 Sinkhole susceptibility model

The area in which the sinkhole formed was designated as the center, with the surrounding grid cells being classified as sinkhole-affected areas. These areas were assigned a value of 1, indicating their status as unstable areas. Meanwhile, an equivalent number of grid cells that were not sinkholes were randomly chosen from the entire study area. A value of zero was assigned to the grid cells that were not classified as sinkholes. The grids that do not exhibit sinkholes are regarded as areas of stability. The sinkhole and non-sinkhole values that were assigned were regarded as the output variables of the model. The sinkhole and non-sinkhole grid cells were randomly partitioned by the *R* statistics program. 70% of the cells were used for training the C5.0 decision tree model, while the remaining 30% were used for testing the model.

Subsequently, the C5.0 decision tree model, which had been trained and evaluated, was utilized to provide predictions regarding the sinkhole susceptibility of the study area based on the identified contributing factors. The sinkhole susceptibility was displayed and divided into five distinct classes (Very Low, Low, Moderate, High, and Very High) using the natural-break method in ArcGIS 10.8 software.

4.3 Sinkhole Susceptibility Map

The sinkhole susceptibility map of the ECF is depicted in Fig. 6(a), generated using the C5.0 decision tree model. Additionally, for comparative analysis, Fig. 6(b) displays the susceptibility map generated by the frequency ratio model. Table 2 presents a summary of the susceptibility class and FR value utilized in the C5.0 decision tree model. The proportions of the projected area corresponding to the categories of Very Low, Low, Moderate, High, and Very High sinkhole-prone areas are 52.12%, 11.27%, 9.21%, 13.87%, and 13.52%, respectively. The C5.0 decision tree model predicts the FR values for the five classes as follows, in order: 0.07, 0.56, 0.89, 2.56, and 3.43. The portions of the study region classified as High and Very High sinkhole susceptibility classes encompass 27.39% of the total area under investigation. However, these classes contribute to 79.76% of the overall FR values.

Table 1 Frequency ratio (FR) values of sinkhole contributing factor

Factor	Type	Value	FR
Head difference (m)	Continuous	≤ 0.0	0.19
		0.0 ~ 3.0	1.11
		3.0 ~ 6.0	1.32
		6.0 ~ 9.0	2.08
		9.0 ~ 12.0	2.06
		12.0 ~ 15.0	1.59
		> 15.0	1.72
Recharge rate (cm/yr)	Continuous	≤ 0.0	0.09
		0.0 ~ 10.0	0.25
		10.0 ~ 20.0	1.17
		20.0 ~ 30.0	2.44
		30.0 ~ 40.0	1.89
		40.0 ~ 50.0	3.60
		> 50.0	3.27
Soil permeability (cm/hr)	Continuous	≤ 5.0	0.10
		5.0 ~ 10.0	0.28
		10.0 ~ 15.0	0.02
		15.0 ~ 20.0	0.13
		20.0 ~ 25.0	0.18
		25.0 ~ 30.0	0.23
		> 30.0	2.33
Proximity to karst features (m)	Continuous	≤ 200.0	4.99
		200.0 ~ 400	1.48
		400.0 ~ 600	0.91
		600.0 ~ 800	0.58
		800.0 ~ 1,000	0.75
		1,000.0 ~ 1,200	0.44
		1,200.0 ~ 1,400.0	0.42
		1,400.0 ~ 1,600.0	0.47
		> 1,600.0	0.16
Overburden thickness (m)	Continuous	≤ 5.0	2.59
		5.0 ~ 10.0	1.47
		10.0 ~ 15.0	1.48
		15.0 ~ 20.0	1.32
		20.0 ~ 25.0	1.03
		25.0 ~ 30.0	0.64
		30.0 ~ 35.0	1.63
		35.0 ~ 40.0	2.01
		40.0 ~ 45.0	2.59
		45.0 ~ 50.0	0.78
		50.0 ~ 55.0	0.73
		55.0 ~ 60.0	0.52
		> 60.0	0.03

Table 1 Continued-

SAS thickness (m)	Continuous	≤ 5.0	2.53
		5.0 ~ 10.0	1.62
		10.0 ~ 15.0	1.85
		15.0 ~ 20.0	1.32
		20.0 ~ 25.0	0.42
		25.0 ~ 30.0	0.19
		30.0 ~ 35.0	0.18
		35.0 ~ 40.0	0.15
		40.0 ~ 45.0	0.67
		45.0 ~ 50.0	0.00
		50.0 ~ 55.0	0.00
		> 55.0	0.00
IAS thickness (m)	Continuous	≤ 5.0	1.15
		5 .0 ~ 10.0	0.94
		10.0 ~ 15.0	1.56
		15.0 ~ 20.0	0.76
		20.0 ~ 25.0	0.79
		25.0 ~ 30.0	1.85
		30.0 ~ 35.0	0.63
		35.0 ~ 40.0	0.36
		40.0 ~ 45.0	0.14
		45.0 ~ 50.0	0.05
		> 50.0	0.00
Annual precipitation (mm/yr)	Continuous	≤ 1225.0	0.10
		1225.0 ~ 1250.0	0.04
		1250.0 ~ 1275.0	0.90
		1275.0 ~ 1300.0	1.87
		1300.0 ~ 1325.0	0.27
		1325.0 ~ 1350.0	0.31
		1350.0 ~ 1375.0	0.83
		> 1375.0	0.92
Geology	Discrete	Water [1]	0.00
		Qh* [2]	1.24
		Qbd* [3]	0.09
		Qu* [4]	0.57
		Qa* [5]	0.00
		TQsu* [6]	0.03
		TQu* [7]	1.94
		TQd* [8]	0.68
		TQuc* [9]	0.00
		Tc* [10]	1.51
		Th* [11]	0.63
		Thc* [12]	3.11
		To* [13]	2.16

Table 1 Continued-

Lithology	Discrete	Water [1]	0.00
		Sand [2]	2.86
		Limestone [3]	2.35
		dune sand [4]	1.15
		beach sand [5]	0.69
		clay or mud [6]	0.97
		Calcarenite [7]	0.00
Land use/Land cover	Discrete	trans., comm., & utilities [1]	3.60
		urban and built-up [2]	1.43
		rangeland [3]	0.47
		upland forest [4]	0.21
		Agriculture [5]	0.16
		Wetlands [6]	0.14
		Water [7]	0.14
		barren land [8]	0.00

*Geologic Units. ***Qh*** = Holocene sediments, ***Qbd*** = Beach ridge and dune, ***Qu*** = Undifferentiated sediments, ***Qa*** = Anastasia Formation, ***TQsu*** = Shelly sediments of Pilo-Pleistocene age, ***TQu*** = Undifferentiated sediments, ***TQd*** = Dunes, ***TQuc*** = Reworked Cypresshead sediments, ***Tc*** = Cypresshead Formation, ***Th*** = Hawthorn Group, ***Thc*** = Hawthorn Group – Coosawhatchie Formation, ***To*** = Ocala Limestone

Table 2 Susceptibility class and FR value used in C5.0 DT model

Susceptibility class	Sinkhole susceptibility index	% of sinkholes	% of areas	FR value
Very low	0 ~ 0.15	3.64	52.12	0.07
Low	0.15 ~ 0.34	6.36	11.27	0.56
Moderate	0.34 ~ 0.59	8.18	9.21	0.89
High	0.59 ~ 0.84	35.45	13.87	2.56
Very high	0.84 ~ 1	46.36	13.52	3.43

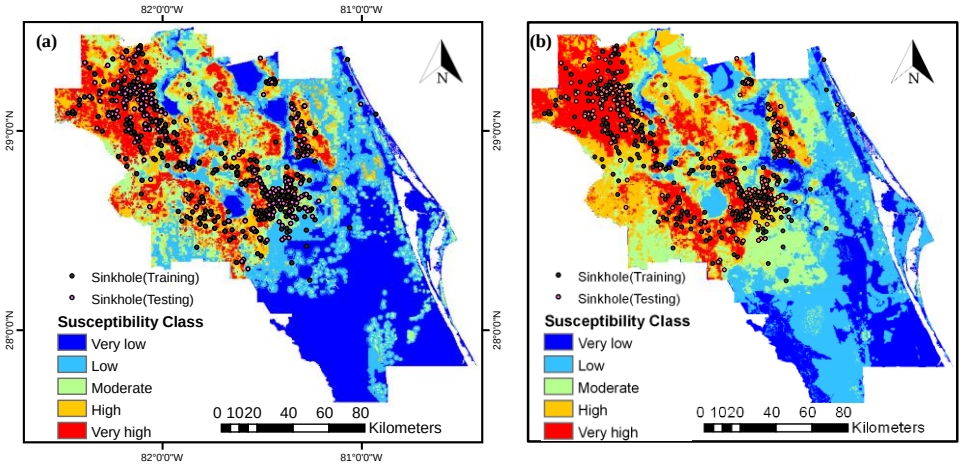

Fig. 6 Sinkhole susceptibility maps generated by (a) C5.0 decision tree, and (b) frequency ratio models

As shown in Figure 6 and Table 2, it can be observed that a significant proportion of the study area exhibits conditions that render the occurrence of sinkholes highly improbable. Furthermore, it can be observed that areas exhibiting elevated levels of sinkhole susceptibility, particularly those classified as high and very high, are predominantly found in the northwestern and central portions of the designated study area. Conversely, areas characterized by high and very high sinkhole susceptibility are infrequently encountered in the southern and eastern sectors of the aforementioned area.

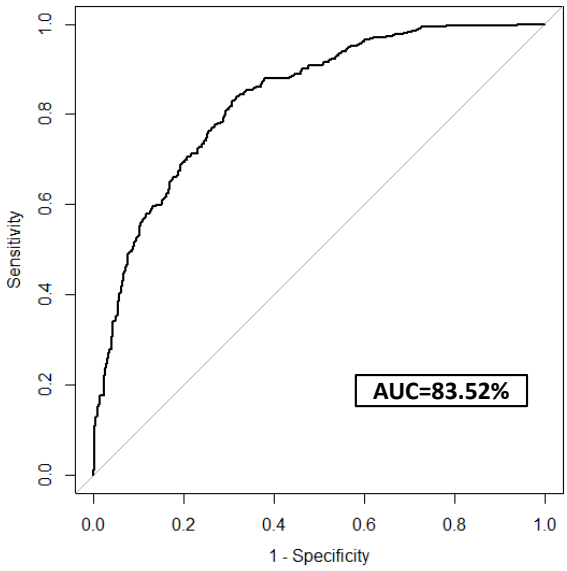

Fig. 7 ROC curve and AUC for a C5.0 DT model

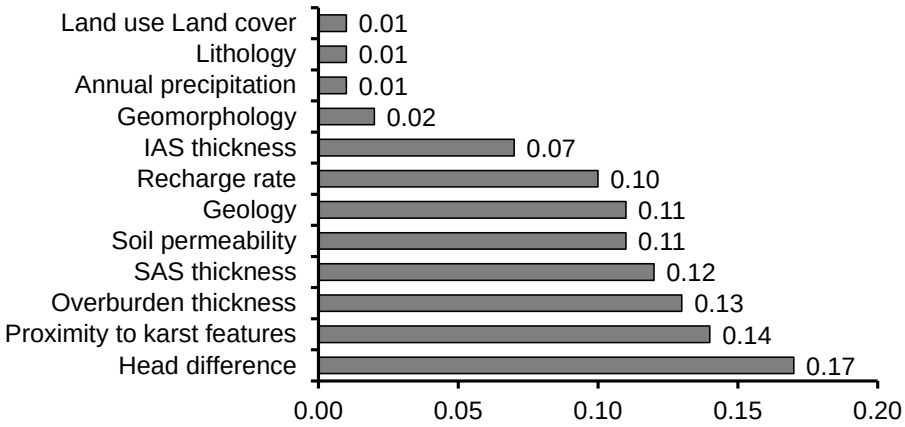

Fig. 8 Relative importance of sinkhole contributing factors

4.4 ROC curve analysis

The evaluation of the C5.0 decision tree model's performance in predicting sinkhole susceptibility was conducted by analyzing the receiver operating characteristic (ROC) curve for the test set, which accounted for 30% of the data. The construction of the Receiver Operating Characteristic (ROC) curve involves the charting of the "sensitivity," which represents the true-positive rate, versus the "1-specificity," which represents the false-positive rate, at various threshold levels. The AUC, which stands for area under the curve, is a measure that represents the accuracy or performance of a model. A bigger AUC indicates a superior performance of the model. For example, when the AUC value is 0.5, it indicates a 50% likelihood that the model would make correct predictions of future sinkhole occurrences. According to the data presented in Fig. 7, the suggested model has favorable performance, as indicated by an AUC value of 0.8352.

4.5 Analysis of sinkhole contributing factors

This study aimed to investigate 12 contributing factors that contribute to sinkhole formation using the C5.0 DT model implemented in the R statistical program. The primary objective was to determine the relative importance of each factor in the context of sinkhole occurrence. According to the findings presented in Fig. 8, the C5.0 decision tree (DT) model identifies several characteristics that exhibit significant importance. These factors include head difference, closeness to karst features, overburden thickness, SAS thickness, soil permeability, geology, and recharge rate, among others.

5. Discussion

Based on the FR analysis of the hydrogeological factors, we could see two interesting observations. First, the most

critical factor seems to be the hydraulic head difference between the surficial and the confined aquifers (higher head difference, higher potential of sinkhole collapse). The head difference is directly related to hydraulic gradient (i), thus related to the flow velocity (v) as well. Then, the proximity to existing karst features (or distance to existing sinkholes), thickness of overburden soil (including SAS and IAS), and soil permeability follow in order. It shows that other geological factors such as geomorphology, lithology, and land use are minor influences. It can be seen that the three triggering mechanisms are soil mechanics (stable vs. unstable), erosion-related hydraulics, and distance to the existing sinkholes. Second, another important point is that the east coastal area and south area of ECF are not susceptible to sinkhole collapse. It is known that those areas involve either low recharge or discharge of groundwater. Lastly, the authors did not look into the effect of the sinkhole factors on its size, which is out of the scope of the study, but the higher hydraulic head difference, along with a thicker overburden soil layer would cause a large size sinkhole, which will be valuable information for sinkhole georisk management.

While training and testing relied on gridded data, the underlying observations originated at a finer point-scale. To reconcile these scales, grid cells were populated by averaging all contained point-scale measurements. This approach facilitated efficient model training with spatially integrated data, preserving the dominant trends from the original data. However, the potential for information loss and uncertainty introduced by this aggregation is acknowledged. The number of point-scale observations per grid cell ranged from 1 to 16, potentially impacting the representativeness of averaged values in some cases.

Further investigations are planned to address these concerns. Sensitivity analyses employing various aggregation methods (e.g., median, maximum) will be conducted to assess their impact on model performance and potential biases introduced by averaging. Additionally, exploring advanced geospatial techniques like kriging interpolation is planned to refine the spatial representation within each grid cell and potentially reduce uncertainty. Such a deeper understanding of alternative approaches is expected to enhance data interpretation and strengthen the robustness of the model's conclusions.

6. Conclusions

A sinkhole is one of the major geohazards in carbonate karst areas (e.g., Florida), and its distribution is largely dependent on a combination of hydrogeological-geological-climatic conditions. An accurate sinkhole susceptibility model and spatial mapping are essential tools for engineers to effectively manage civil infrastructure and enhance public safety. This study aims to develop a sinkhole susceptibility model using data mining/machine learning and then construct the sinkhole susceptibility map for ECF.

The effects of twelve sinkhole contributing factors, including hydraulic head difference, groundwater recharge rate, soil permeability, proximity to or distance from karst features, overburden thickness, SAS thickness, IAS thickness, annual precipitation, geology, geomorphology, lithology, and land use land cover (LULC) on the sinkhole occurrence were investigated using the FR model. The influence of those factors on sinkhole occurrence was quantitatively evaluated. The randomly selected sinkhole dataset (70% of the entire data) was used to build the C5.0 DT model, the susceptibility map of ECF was constructed, and the map was validated by the AUC method, which is considered as an effective indicator of the predictive performance of the model. Using the remaining 30% of the dataset, the AUC was 0.8352, indicating that the model has robustness and good predictive ability for the study area.

The proposed susceptibility map predicts the relative risk of sinkhole occurrences; approximately 79.76% of reported sinkhole data falls into either the high or very high susceptibility of the map. These high-potential sinkhole areas in ECF are mainly located in the region where hydrogeological factors such as head difference, recharge rate, overburden thickness, etc., have a strong statistical correlation with sinkhole occurrence.

Acknowledgments

The authors thank the Florida Geological Survey (FGS) for sharing the data (sinkhole survey and geological data).

References

Beck, B.F. and Sinclair, W.C. (1986), Sinkholes in Florida: An Introduction, Florida Sinkhole Research Institute, Orlando, FL.

Ciotoli, G., Di Loreto, E., Finoia, M.G., Liperi, L., Meloni, F., Nisio, S. and Sericola, A. (2016), "Sinkhole susceptibility, Lazio Region, central Italy", *J. Maps.*, **12**(2), 287-294. https://doi.org/10.1080/17445647.2015.1014939.

Ding, H., Wu, Q., Zhao, D., Mu, W. and Yu, S. (2019), "Risk assessment of karst collapse using an integrated fuzzy analytic hierarchy process and grey relational analysis model", *Geomech. Eng.*, **18**(5), 515-525. https://doi.org/10.12989/gae.2019.18.5.515.

FDEP (2022), Florida Subsidence Incident Reports, FDEP, Florida Geological Survey.

Florida Office of Insurance Regulation (2010), Report on Review of the 2010 Sinkhole Data Call, Florida Office of Insurance Regulation (FOIR).

Galve, J.P., Gutiérrez, F., Remondo, J., Bonachea, J., Lucha, P. and Cendrero, A. (2009), "Evaluating and comparing methods of sinkhole susceptibility mapping in the Ebro Valley evaporite karst (NE Spain)", *Geomorphology*, **111**(3), 160-172. https://doi.org/10.1016/j.geomorph.2009.04.017.

Genis, M., Akcin, H., Aydan, O. and Bacak, G. (2018), "Investigation of possible causes of sinkhole incident at the Zonguldak Coal Basin, Turkey", *Geomech. Eng.*, **16**(2), 177-185. https://doi.org/10.12989/gae.2018.16.2.177.

Gkioulekas, I. and Papageorgiou, L.G. (2021), "Tree regression models using statistical testing and mixed integer programming", *Comput. Ind. Eng.*, **153**, 107059. https://doi.org/10.1016/j.cie.2020.107059.

Guo, Z., Shi, Y., Huang, F., Fan, X. and Huang, J. (2021), "Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management", *Geosci. Front.*, **12**(6), 101249. https://doi.org/10.1016/j.gsf.2021.101249.

Kim, J.M., Lee, S., Park, J.Y., Kihm, J.H. and Park, S. (2020), "A set of failure variables for analyzing stability of slopes and tunnels", *Geomech. Eng.*, **20**(3), 175-189. https://doi.org/10.12989/gae.2020.20.3.175.

Kim, Y.J., Nam, B.H., Jung, Y.H., Liu, X., Choi, S., Kim, D. and Kim, S. (2022), "Probabilistic spatial susceptibility modeling of carbonate karst sinkhole", *Eng. Geol*., **306**, 106728. https://doi.org/10.1016/j.enggeo.2022.106728.

Kim, Y.J., Nam, B.H., Shamet, R., Soliman, M. and Youn, H. (2020), "Development of sinkhole susceptibility map of east central Florida", *Nat. Hazard. Review*, **21**(4), 04020035. doi:10.1061/(ASCE)NH.1527-6996.0000404.

Li, C., Zou, J.F. and Sheng, Y.M. (2020), "Undrained solution for cavity expansion in strength degradation and tresca soils", *Geomech. Eng.*, **21**(6), 527-536. https://doi.org/10.12989/gae.2020.21.6.527.

Liu, L.L., Yang, C. and Wang, X.M. (2021), "Landslide susceptibility assessment using feature selection-based machine learning models", *Geomech. Eng.*, **25**(1), 1-16. https://doi.org/10.12989/gae.2019.18.5.515.

Nanehkaran, Y.A., Mao, Y., Azarafza, M., Kockar, M.K. and Zhu, H.H. (2021), "Fuzzy-based multiple decision method for landslide susceptibility and hazard assessment: A case study of Tabriz, Iran", *Geomech. Eng.*, **24**(5), 407-418. https://doi.org/10.12989/gae.2021.24.5.407.

Naithani, A.K., Jain, P., Singh, L.G. and Rawat, D.S. (2022), "Engineering geological characteristics of the underground surge pool cavern: a case study", *Int. J. Geo-Eng*., **13**(7). https://doi.org/10.1186/s40703-022-00172-9

Ozdemir, A. (2015), "Sinkhole Susceptibility Mapping Using a Frequency Ratio Method and GIS Technology Near Karapınar, Konya-Turkey", *Procedia Earth Planetary Sci.*, **15**, 502-506. https://doi.org/10.1016/j.proeps.2015.08.059.

Papadopoulou-Vrynioti, K., Bathrellos, G.D., Skilodimou, H.D., Kaviris, G. and Makropoulos, K. (2013), "Karst collapse susceptibility mapping considering peak ground acceleration in a rapidly growing urban area", *Eng. Geol*., **158**, 77-88. https://doi.org/10.1016/j.enggeo.2013.02.009.

Safavian, S.R. and Landgrebe, D. (1991), "A survey of decision tree classifier methodology", *IEEE T. Syst. Man Cy*., **21**(3), 660-674. https://doi.org/10.1109/21.97458.

Shamet, R., Nam, B.H. and Horhota, D. (2018), "Development of a sinkhole raveling chart based on Cone Penetration Test (CPT) data", *Proceedings of the 15th Multidisciplinary Conference on Sinkholes and the Engineering and Environmental Impacts of Karst and the 3rd Appalachian Karst Symposium*, Shepherdstown, WV, April 2-6.

Sheng, M., Zhou, J., Chen, X., Teng, Y., Hong, A. and Liu, G. (2022), "Landslide susceptibility prediction based on frequency ratio method and C5.0 decision tree model", *Front. Earth Sci*., **10**. https://doi.org/10.3389/feart.2022.918386.

Soliman, M.H., Shamet, R., Kim, Y.J., Youn, H. and Nam, B.H. (2019), "Numerical Investigation on the Mechanical Behavior of Karst Sinkholes", *Environ. Geotech.*, **8**(6), 367-381. https://doi.org/10.1680/jenge.18.00063.

Strzałkowski, P. (2018), "Sinkhole formation hazard assessment", *Environ. Earth Sci*., **78**(1), 9. https://doi.org/10.1007/s12665-018-8002-5.

Subedi, P., Subedi, K., Thapa, B. and Subedi, P. (2019), "Sinkhole susceptibility mapping in Marion County, Florida: Evaluation and comparison between analytical hierarchy process and logistic regression based approaches", *Scientific Reports*, **9**(1), 7140. https://doi.org/10.1038/s41598-019-43705-6.

Taheri, K., Gutiérrez, F., Mohseni, H., Raeisi, E. and Taheri, M. (2015), "Sinkhole susceptibility mapping using the analytical hierarchy process (AHP) and magnitude–frequency relationships: A case study in Hamadan province, Iran", *Geomorphology*, **234,** 64-79. https://doi.org/10.1016/j.geomorph.2015.01.005.

Tacim, G., Posluk, E. and Gokceoglu, C. (2023), "Importance of grouting for tunneling in karstic and complex environment (a case study from Türkiye)", *Int. J. Geo-Eng*., **14**(6). https://doi.org/10.1186/s40703-023-00183-0

Tihansky, A.B. (1999), "Sinkholes, west-central Florida", Land subsidence in the United States: US geological survey circular. 1182 121-140.

Weary, D.J. (2015), "The cost of Karst subsidence and sinkhole collapse in the united states compared with other natural hazards", *Proceedings of the 14th Multidisciplinary Conference on Sinkholes and the Engineering and Environmental Impacts of Karst*, Rochester, MN, October 5-9.

Weary, D.J. and Doctor, D.H. (2014), Karst in the United States: A digital map compilation and database, U.S. Geological Survey, Open-File Report 2014-1156.

Xu, Z., Xian, M., Li, X., Zhou, W., Wang, J., Wang, Y. and Chai, J. (2021), "Risk assessment of water inrush in karst shallow tunnel with stable surface water supply: Case study", *Geomech. Eng*., **25**(6), 495-508. https://doi.org/10.12989/gae.2021.25.6.495.

Xu, Z., Chengping, Z., Bo, M. and Youjun, X. (2020), "Experimental study on the mechanical response and failure behavior of double-arch tunnels with cavities behind the liner", *Geomech. Eng.*, **20**(5), 399-410. https://doi.org/10.12989/gae.2020.20.5.399

Yilmaz, I. (2007), "GIS based susceptibility mapping of karst depression in gypsum: A case study from Sivas basin (Turkey)", *Eng. Geol*., **90**(1), 89-103. https://doi.org/10.1016/j.enggeo.2006.12.004.

Yilmaz, I., Marschalko, M. and Bednarik, M. (2013), "An assessment on the use of bivariate, multivariate and soft computing techniques for collapse susceptibility in GIS environ", *J. Earth Syst. Sci*., **122**(2), 371-388. https://doi.org/10.1007/s12040-013-0281-3.

Yuan, Y.C., Li, S.C., Zhang, Q.Q., Li, L.P., Shi, S.S. and Zhou, Z.Q. (2016), "Risk assessment of water inrush in karst tunnels based on a modified grey evaluation model: Sample as Shangjiawan Tunnel", *Geomech. Eng.*, **11**(4), 493-513. https://doi.org/10.12989/gae.2016.11.4.493.

Zhou, Z.Q., Li, S.C., Li, L.P., Shi, S.S. and Xu, Z.H. (2015), "An optimal classification method for risk assessment of water inrush in karst tunnels based on the grey system", *Geomech. Eng*., **8**(5), 631-647. https://doi.org/10.12989/gae.2015.8.5.631.

IC