

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

DOI: 10.1109/ACCESS.2017.Doi Number

Knowledge Discovery of Global Landslides Using Automated Machine Learning Algorithms

Fahim K. Sufi,¹ Musleh Alsulami²

¹ School of Computer Science and Information Technology, RMIT University, Melbourne, Vic. 3000, Australia

² Information Systems Department, Umm-Al-Qura University, Makkah, Saudi Arabia

Corresponding author: Fahim Sufi (e-mail: research@fahimsufi.com), Musleh Alsulami (e-mail: mhsulami@uqu.edu.sa).

This work received no external funding.

ABSTRACT: Understanding the complex dynamics of global landslides is essential for disaster planners to make timely and effective decisions that save lives and reduce the economic impacts on society. Using NASA's inventory of global landslide data, we developed a new machine learning (ML)-based system for town planners, disaster recovery strategists, and landslide researchers. Our system revealed hidden knowledge about a range of complex scenarios created from five landslide feature attributes. Users of our system can select from a list of 1.295×10^{64} possible global landslide scenarios to discover valuable knowledge and predictions about the selected scenario in an interactive manner. Three ML algorithms—anomaly detection, decomposition analysis, and automated regression analysis—are used to elicit detailed knowledge about 25 scenarios selected from 14,532 global landslide records covering 12,220 injuries and 63,573 fatalities across 157 countries. Anomaly detection, logistic regression, and decomposition analysis performed well for all scenarios under study, with the area under the curve averaging 0.951, 0.911, and 0.896, respectively. Moreover, the prediction accuracy of linear regression had a mean absolute percentage error of 0.255. To the best of our knowledge, our scenario-based ML knowledge discovery system is the first of its kind to provide a comprehensive understanding of global landslide data.

Keywords: Strategic decision support tool for landslides, machine learning, anomaly detection, regression analysis, decomposition analysis, knowledge discovery.

I. INTRODUCTION

Landslides are natural events that have adverse effects on human life, infrastructure, the economy, and society [1]. To reduce the negative effects of landslides and increase the level of disaster preparedness, in-depth research on global landslides is essential [2].

In the past, strategic decision-makers needed a data scientist to prepare the data, develop machine learning (ML) models, and summarize the results. Depending on the complexity of the problem, the data scientist may need an information technology administrator to run high-performance computing on infrastructure capable of handling the ML load [3], enabling the data scientist to execute the ML model and manually summarize the results for the strategic decision-maker. This task delegation process can undergo several iterations until the required model satisfies the needs of the strategic decision-maker (see Fig. 1). The delays caused by this task delegation may become critical if additional roles (e.g., business analysts, data engineers, artificial intelligence (AI) engineers, statisticians, database administrators, etc.) are introduced

As shown in Fig. 1, the proposed system eliminates the delays associated with task delegation. Users of the proposed

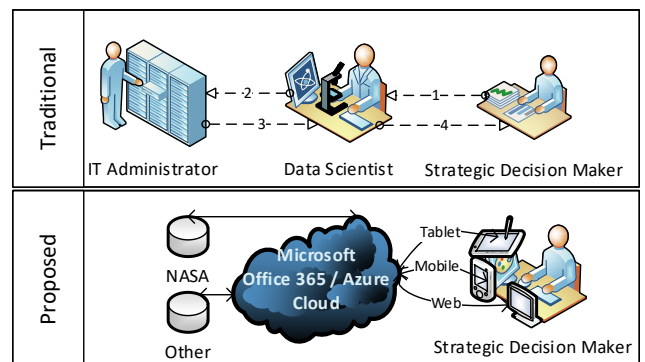


Figure 1. Evidence-based decision-making allows town planners and strategic decision-makers to implement effective landslide policies that can save lives, protect infrastructure, and reduce economic impacts on society.

system can access secure cloud-based solutions for specific scenarios using a range of mobile, tablet and other web-enabled devices [5]. Because the system uses recently developed natural language processing algorithms, the user can select a landslide scenario and obtain instant insights in plain english texts [6]. The fully automated summarized insights produced by the proposed system can support

evidence-based decision-making to save lives, protect infrastructure, and reduce economic impacts on society.

In its current version, the system is connected to the National Aeronautics and Space Administration (NASA) landslide database, which contains 14,532 records of global landslides covering 12,220 injuries and 63,573 fatalities in 157 countries. From these data, 1.295×10^{64} possible scenarios can be constructed (see section III). From this vast number of scenarios, 25 were randomly selected to demonstrate the applicability of the system using laptops or web-enabled mobile devices, with a clustering accuracy (area under the curve (AUC)) of up to 0.951 and a prediction accuracy (mean absolute percentage error (MAPE)) of up to 0.10.

Based on the existing literature [7]–[15], the proposed system is the first to utilize clustering algorithms such as automated anomaly detection (AUC up to 0.951) and decomposition tree analysis (AUC up to 0.896) in the landslide domain.

II. BACKGROUND

ML has been used in landslide research for landslide detection [7], characterization [8], susceptibility assessment [9]–[11], prediction [12], and early warning systems [13]. The most common ML algorithms used are support vector machines [7], [9]–[11], followed by logistic regression [7], [10], [11], random forests [7], [9], artificial neural networks [9], [11], and Bayesian networks [9], [10]. The performance of ML algorithms in landslide research is mostly evaluated using AUC [7], [9]–[11] or MAPE [12]. AUC is more suitable for measuring the performance of clustering algorithms, while MAPE is preferred for measuring the accuracy of prediction algorithms [7], [9]–[12]. Therefore, we evaluated ML algorithms using AUC and MAPE.

Existing landslide research using ML is based on the use of R, Python and other statistical programming languages that do not support the use of ML programs in the mobile environment. However, to make the proposed solution available to strategic decision-makers via mobile, tablet and other web-enabled devices, the proposed system was coded using Microsoft .NET and Azure. Microsoft documentation in [16], provides the details of supported ML algorithms on Microsoft ecosystem.

Gaps in the literature on the use of ML algorithms in landslide research include the following:

1. Researchers require a complex understanding of ML models [14].
2. ML is not being utilized to facilitate strategic decision-making about disaster preparedness or risk management [15].
3. The results obtained from existing ML-based landslide research were not expanded in a natural language [14].
4. The potential for ML models to analyze the root cause of landslides is not being harnessed [17].

5. ML models and data must be manually handled, making knowledge discovery a time-consuming and labor-intensive task [7]–[14].
6. Existing implementation of ML algorithms are not suitable for use in integrated cloud-based applications or web-enabled mobile devices.

Table I shows the algorithms used in the literature and whether they may be implemented in a .NET-based cloud environment.

TABLE I: ML ALGORITHMS USED IN EXISTING LANDSLIDE RESEARCH

ML algorithm	Accuracy	Support in .NET
Support vector machine	AUC: Up to 0.750 [7], 0.950 [10], 0.930 [9], 0.912 [11]	Supported; resource intensive
Logistic regression	AUC: Up to 0.823 [7], 0.922 [10], 0.748 [11]	Supported; lightweight
Random forest	AUC: Up to 0.991 [7], 0.951 [9]	Supported; resource intensive
Convolutional neural networks	Up to 0.917 [7]	Supported; lightweight
Bayesian network	AUC: 0.915 [10], 0.916 [9]	Supported; resource intensive
Support vector regression	MAPE: 0.125 and 0.777% [12]	Supported; lightweight
Artificial neural network	AUC: 0.934 [9], 0.852 [11]	Not supported
Naïve Bayes	AUC: 0.910 [10]	Supported; resource intensive
Backpropagation neural network	MAPE: 0.147 and 0.899% [12]	Not supported
Fisher's linear discriminant analysis	AUC: 0.921 [10]	Not supported
Multiple factor particle swarm optimization-kernel extreme learning machine	MAPE: 0.083 and 0.494% [12]	Not supported

NOTE: ML: MACHINE LEARNING; AUC: AREA UNDER CURVE; MAPE: MEAN ABSOLUTE PERCENTAGE ERROR.

We have previously reported on the use of ML to solve problems ranging from abnormality detection [18]–[20] to person identification [21]. Here, ML is used in knowledge discovery and analysis of the root cause of global landslides.

III. METHODOLOGY

The primary motivation for designing a new ML-based knowledge discovery solution arose from the deficiencies in existing ML-based landslide research. The proposed solution has the following features:

7. System users (i.e., strategic decision-makers) do not need a deep understanding of ML models. Using natural language processing [6], the information is translated into a language that the strategic decision-maker can understand.
8. Multiple interactive interfaces facilitate strategic decision-making on disaster preparedness and risk management using multiple ML algorithms.

9. The insights obtained from one ML algorithm (e.g., regression) may be expanded to other algorithms (e.g., anomaly detection, decomposition tree analysis).
10. Decompression tree analysis is used to analyze the root causes of landslides. This is the first time that decomposition tree analysis has been used in landslide research.
11. The solution is fully automated, and the appropriate ML algorithm is automatically executed on the correct set of data without the need for manual intervention from data scientists, data engineers, statisticians or database programmers, minimizing delays.
12. The solution is programmed using .NET, facilitating its use in Microsoft Office 365 and Azure [5], [6], [16], [22]. This will allow strategic decision-makers to access the solution via laptops and mobile devices.



Figure 3. Categorization of global landslide feature attributes.

TABLE II: DATA DISTRIBUTION OF GLOBAL LANDSLIDE ATTRIBUTES

Attribute name	Attribute type	Attribute distribution and statistics
1. Object identifier	Text	D: 14,532 U: 14,532 V: 100% E: 0%
2. Event title	Text	D: 12,162 U: 11,794 V: 95% E: 5%
3. Event description	Text	D: 11,061 U: 10,534 V: 82% E: 18%
4. Event date	Date	D: 7,512 U: 5,442 V: 90% E: 10%
5. Country	Text	D: 157 U: 24 V: 99% E: < 1%
5. Latitude	Decimal	D: 13,980 U: 13,612 V: 100% E: 0%
6. Longitude	Decimal	D: 14,071 U: 13,780 V: 100% E: 0%
7. Location description	Text	D: 12,237 U: 11,779 V: 90% E: 0%
8. Location accuracy	Text	D: 10 U: 0 V: 99% E: < 1%

To develop the proposed solution, data were obtained from NASA’s global landslide inventory [23] before being cleaned and transformed prior to modeling. Data modeling was then performed using best practice [24]. Finally, the data were visualized and analyzed using ML algorithms (see section III D). Fig. 2 shows the step-by-step process used to generate AI insights into global landslide data.

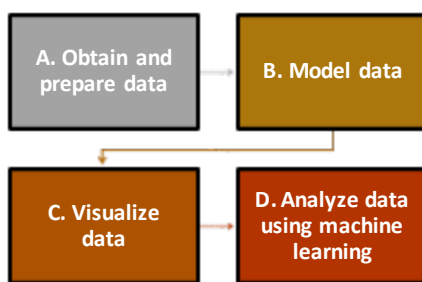


Figure 2. Methodology for knowledge discovery of global landslides using automated machine learning.

A. OBTAIN AND PREPARE DATA

Data may be accessed from a range of sources, including online databases, websites, Excel files, flat files, web-based application programming interfaces, and even pdf files. After identifying the data source, data integration tools (e.g. SQL Server Integration Services or Power BI Query Editor) may be used to facilitate the export, transformation, and loading of data from the source into a data warehouse. Data transformation and cleaning, also known as data preparation, transform the data into the correct format for modeling or or analysis using ML. For this research, we obtained data from an online source in a comma-separated values format [23]. The data were then transformed into a suitable format, allowing a more rapid analysis and better understanding of the feature attributes of global landslides.

Fig. 3 shows the categorization of the data fields, while Table II presents the detailed statistics of the landslide data. Understanding the statistics of landslide feature attributes is crucial before proceeding to the next step, namely data modeling, visualization, and analysis using ML.

Attribute name	Attribute type	Attribute distribution and statistics
9. Trigger (cause of the landslide)	Text	D: 19 U: 1 V: 99% E: < 1%
10. Category (e.g., topple, mudslide, rock fall etc.)	Text	D: 15 U: 0 V: 100% E: 0%
11. Setting (e.g., above river, above road, mine etc.)	Text	D: 15 U: 1 V: 99% E: < 1%
12. Size (e.g., small, medium, large)	Text	D: 6 U: 0 V: 100% E: 0%
13. Injury count	Integer	D: 57 U: 27 V: 75% E: 25%
14. Fatality count	Integer	D: 116 U: 57 V: 75% E: 25%

Note: D: distinct; U: unique; V: valid; E: empty.

B. MODEL DATA

Data modeling is the most important stage in the process of knowledge discovery using ML. With effective data modeling, ML-driven solutions can produce powerful insights with minimal delays. During this phase, the relationships among different sets of data with the correct cardinality are drawn.

Fig. 4 shows that the data obtained in this study were arranged in a star schema [24], with the main factual data (fatality and injury counts) in the center, surrounded by the following dimensions: category, size, setting, trigger, and country. This arrangement enabled the analysis of the main facts by category, size, setting, trigger, and country using one-way filtering. The benefits of the star schema over other data modeling techniques (e.g., flattened tables, snowflakes) is that it provides faster and more accurate results [24].

In our system, we created the scenario (S) using five dimensional features: category (G), size (I), setting (E), trigger (T), and country (C). Therefore:

$$S = \{x, y, z, m, n, p, q \mid x \subseteq G, y \subseteq I, z \subseteq E, m \subseteq T, n \subseteq C\} \quad (1)$$

$$G = \{\text{landslide, mudslide, rock_fall, debris_flow, complex, rotational_slide, translational_slide, riverbank_collapse, creep, snow_avalanche, lahar, earth_flow, topple, other, unknown}\} \quad (2)$$

$$I = \{\text{small, medium, large, very_large, catastrophic, unknown}\} \quad (3)$$

$$E = \{\emptyset, \text{above_road, natural_slope, urban, burned_area, below_road, above_river, mine, deforested_slope, retaining_wall, engineered_slope, bluff, above_coast, other, unknown}\} \quad (4)$$

$$T = \{\emptyset, \text{downpour, rain, continuous_rain, tropical_cyclone, monsoon, snowfall_snowmelt, construction, mining, earthquake, flooding, freeze_thaw, dam_embankment_collapse, leaking_pipe, volcano, vibration, other, no_apparent_trigger, unknown}\} \quad (5)$$

$$C = \{\text{United States, India, Myanmar, Philippines, Nepal, China, Colombia, Indonesia, United Kingdom, Canada, Macedonia, Malaysia, Brazil, Pakistan, Czech Republic, New Zealand, Vietnam, Australia, Japan, Mexico, Uganda, Thailand, Bangladesh, Trinidad and Tobago, Sri Lanka, Guatemala, Italy, Peru, Costa Rica, Congo, Kenya, Kyrgyzstan, Switzerland, Fiji, Jamaica, Germany, Panama, Georgia, Honduras, Rwanda, Ecuador, Austria, Bulgaria, Nicaragua, Papua New Guinea, Ireland, Tajikistan, Azerbaijan, Norway, etc.}\} \quad (6)$$

Equations (4) and (5) contains null values represented by \emptyset .

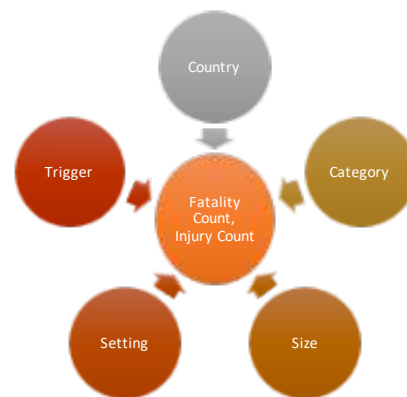


Figure 4. The main facts (fatality and injury count) are filtered by dimension (i.e., category, size, setting, trigger and country). The arrows represent filter direction.

To calculate the number of possible scenarios, we first needed to calculate the possible filter options for each dimension. For example, if $A = \{\text{slide, fall, topple}\}$, the following filter combinations are possible:

1. $\{\text{slide}\}$
2. $\{\text{fall}\}$
3. $\{\text{topple}\}$
4. $\{\text{slide, fall}\}$
5. $\{\text{slide, topple}\}$
6. $\{\text{fall, topple}\}$
7. $\{\text{slide, fall, topple}\}$

Therefore, for A, there are seven possible filter settings, represented by $(2^{|A|} - 1)$, which is the formula used to calculate the power set of the attribute minus 1 (i.e., $P(A) - 1$), where $|A|$ is the cardinality of A. One is deduced because the power set also includes an empty set, and the selection of an empty set is not supported by the proposed system.

Hence, the total number of possible scenarios for our global landslide can be calculated as:

$$|S| = (2^{|G|} - 1) \times (2^{|I|} - 1) \times (2^{|E|} - 1) \times (2^{|T|} - 1) \times (2^{|C|} - 1) = 1.296 \times 10^{64} \quad (7)$$

The purpose of this study is not to produce an exhaustive list of global landslide data covering all 1.296×10^{64} scenarios. However, we demonstrate the ability to dynamically discover knowledge based on automated ML algorithms for any possible scenario out of 1.296×10^{64} scenarios.

C. VISUALIZE DATA

Once the data modeling was complete, we used category, size, setting, trigger, and country to filter the factual data to drive the ML-based knowledge discovery. A wide range of visualizations, including slicers, Bing Maps, key influencers, decomposition analysis, and anomaly detection on a line chart were used in the dashboards. Changing the value of each filter (e.g., landslide size to *small*, *medium*, or *large*) filtered the fact table containing the number of injuries and fatalities, in turn changing the key influencers, anomaly detection, or decomposition analysis. Table III shows how a change in a filter such as landslide size affects the number of injuries and fatalities.

TABLE III: FILTER OPTIONS FOR GLOBAL LANDSLIDE SIZE

Landslide size	No. records	No. injuries	No. fatalities
No filter	14,532	12,220	63,573
Catastrophic	18	2,794	24,212
Very large	144	2,768	19,547
Large	995	3,622	8,937
Medium	7,144	2,187	9,854
Small	3,484	450	618
Unknown	2,747	399	405

Table III shows that the total number of fatalities caused by medium-sized landslides was higher than that caused by large landslides. Table III was generated using the exploration dashboard of our ML-based knowledge discovery system (see Fig. 6). Fig. 6 shows that from 1915 to 2021, approximately 14,532 global landslides caused 12,220 injuries and 63,573 fatalities.

Fig. 5 shows how different types of algorithms enable different dashboards of the proposed system to function. For this research, we created the following four dashboards:

1. General analysis of global landslides (see Fig. 6)
2. Linear regression (i.e., a key influencer) to determine the influence of factors on the number of fatalities (see Fig. 7)
3. Time-series anomaly detection (i.e., a key influencer) to identify anomalies in the number of total casualties, fatalities, injuries and countries by year (see Fig. 8).
4. Decomposition analysis to identify root causes and explore data (see Fig. 9).

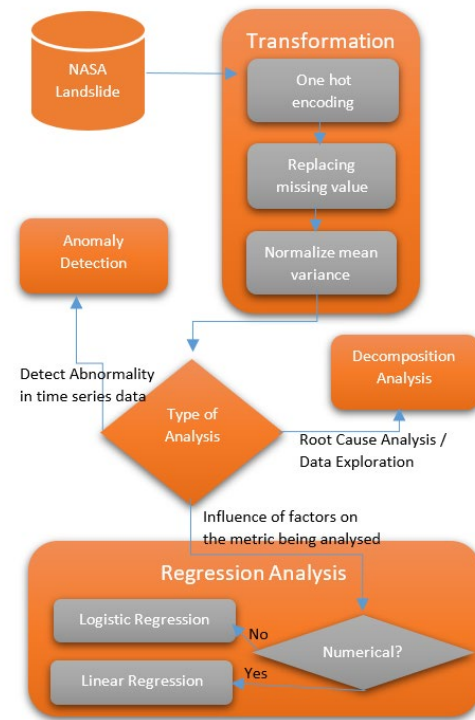


Figure 5. Knowledge discovery using three different types of machine learning algorithms: regression analysis, decomposition analysis, and anomaly detection.

The dashboards are publicly available and hosted in Microsoft Cloud [5] (for data exploration purposes without ML features only). To utilize the fully functional version of ML-based knowledge discovery, a user can download our solution from GitHub [22], which provides regression analysis (Fig. 7), anomaly detection (Fig. 8), and decomposition analysis (Fig. 9).

D. ANALYZE DATA USING ML

For this study, we conducted an extensive analysis of NASA’s global landslide data, comprising 14,532 records [23]. We used the ML algorithms in ML.NET [16], including regression analysis [25]–[27], anomaly detection [28], [29], and decomposition analysis [30], to analyze the number of total casualties, fatalities, and injuries according to the following feature attributes: landslide category, size, setting, trigger, and country. As shown in Fig. 5, before executing any ML algorithm, the data must be prepared and cleaned for faster ML operations. To achieve this, a series of data transformations is undertaken [31]. Once the data transformation is complete, depending on the type of analysis and intent, the proposed solution will perform regression analysis [16], anomaly detection [28], [29], or decomposition tree analysis [30]. For regression analysis, if the data are numerical, then linear regression [25] is performed, and if the data are categorical, then logistic regression [26], [27] is performed.

Analysis of Global Landslide

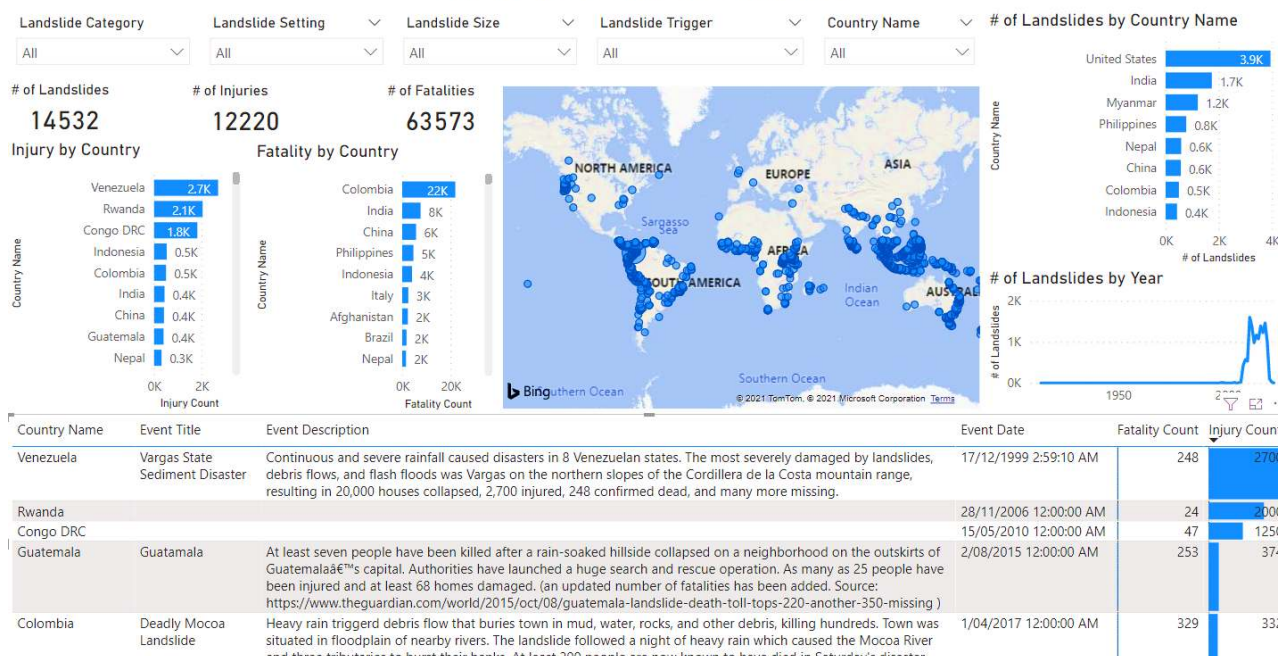


Figure 6. General analysis of global landslides. This dashboard demonstrates that from 1915 to 2021, there have been 14,532 global landslides reported, causing 12,220 injuries and 63,573 fatalities.

Key Influencers Analysis on Global Landslide

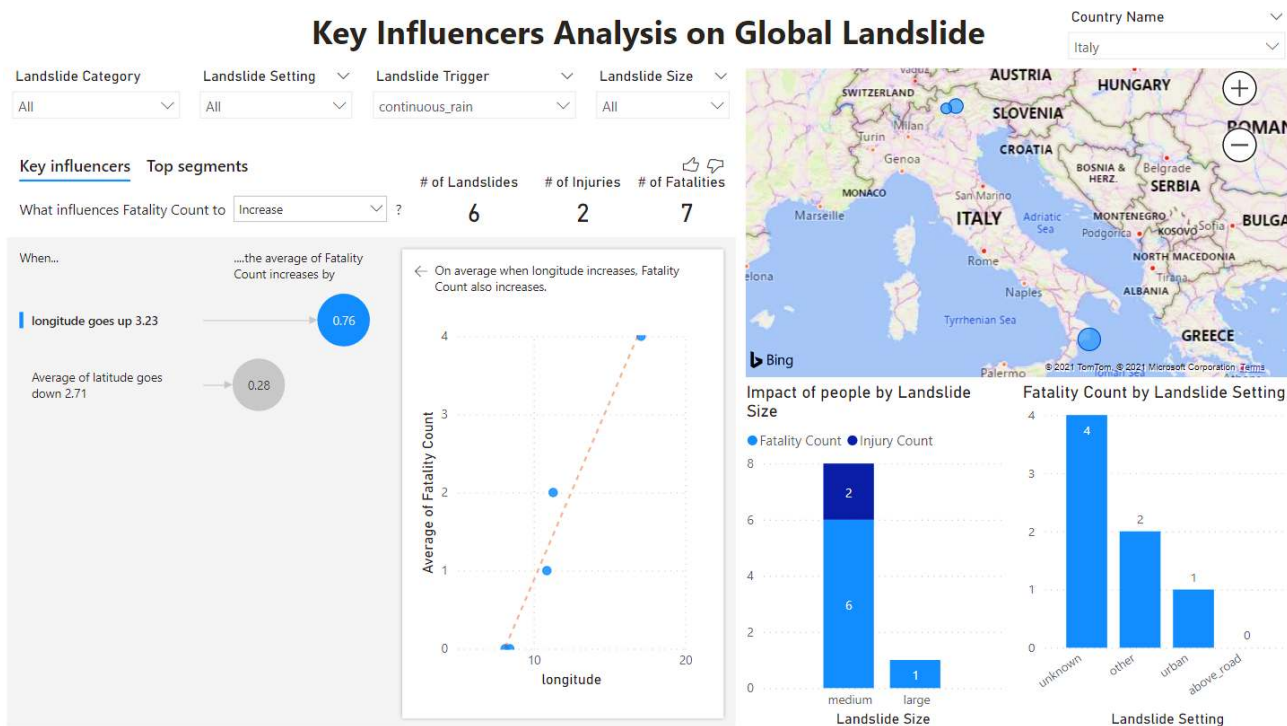


Figure 7. Linear regression solution (i.e., key Influencer) for the influence of factors on the number of fatalities (when the country is Italy and landslide trigger is continuous rain).

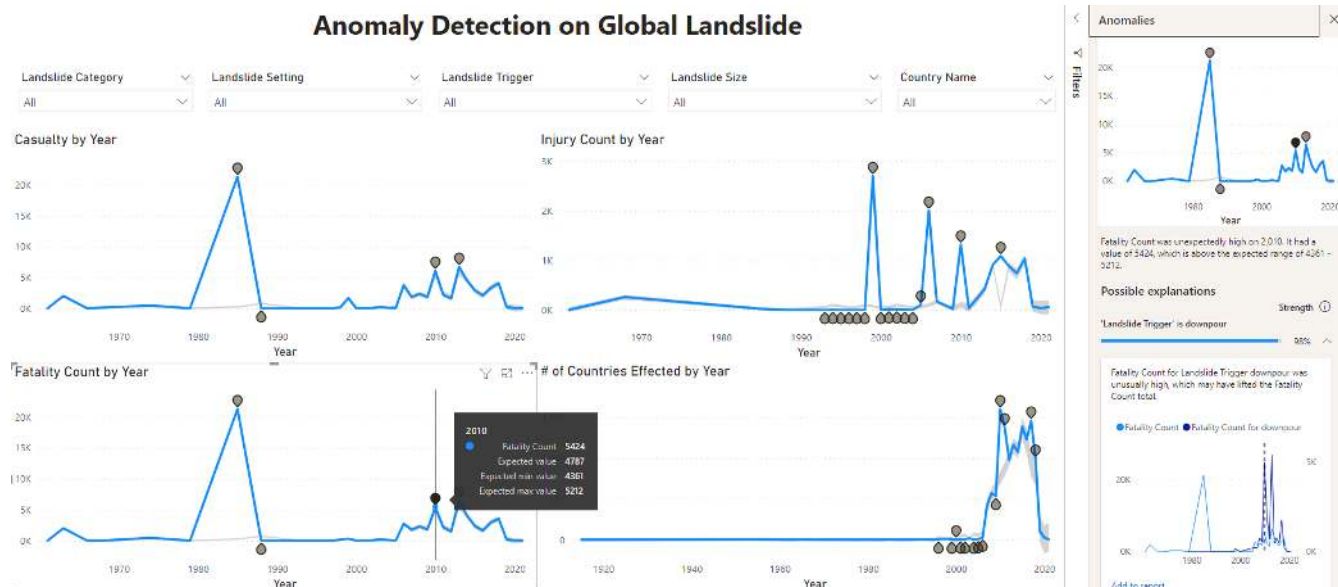


Figure 8. Anomaly detection on time-series data (i.e., key Influencer) to identify anomalies in the number of casualties, fatalities, injuries and countries per year.

Decomposition Analysis on Global Landslide

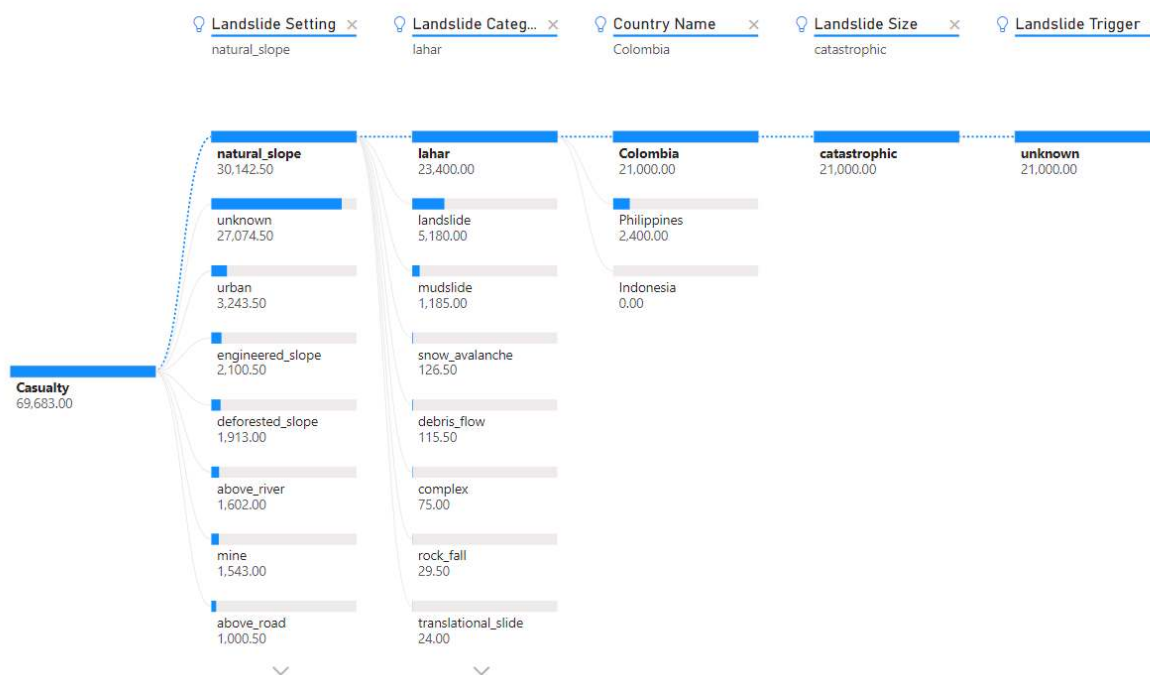


Figure 9. Decomposition analysis for root cause analysis and data exploration.

1. Transformation

Transformation was undertaken to prepare the global landslide data for regression analysis, anomaly detection, and decomposition analysis. During the transformation, the following three algorithms in Microsoft.ML.Transform were executed:

13. The *OneHotEncoding* function converts categorical data into numerical values for efficient and effective processing of ML algorithms [32].
14. The *ReplaceMissingValues* function replaces missing values with default, minimum, maximum, mean, or most frequent values [33].
15. The *NormalizeMeanVariance* function adjusts values measured on different scales to a notionally common scale with computed mean and variance of the data [34].

2. Regression analysis

In this paper, regression analysis was used to identify the most important landslide feature attributes associated with landslide-related fatalities. Regression analysis automatically ranks factors by their relative importance and displays them as key influencers of both categorical and numerical metrics. Two types of regression were used (see Fig. 5). For numerical features, linear regression was performed using ML.Net’s stochastic dual coordinate ascent function [25]. Linear regression is one of the simplest ML algorithms in supervised learning techniques and is used to solve regression problems and predict continuous dependent variables with the help of independent variables. The goal of linear regression is to identify the best-fit line that can accurately predict the output of the continuous dependent variable. By finding the best-fit line, the algorithm establishes a linear relationship between the dependent and independent variables in the form $y = b_0 + b_1x_1 + \varepsilon$.

In contrast, for categorical features, logistic regression was performed using ML.Net’s L-BFGS logistic regression [26], [27]. Logistic regression is one of the most popular ML algorithms for supervised learning techniques. It can also be used for classification and regression problems. Logistic regression was used to predict the categorical dependent variable with the help of independent variables using $\text{Log}[y/y - 1] = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$. The output of the logistic regression problem can only be between 0 and 1; therefore, logistic regression may be used when the probabilities of two classes are required, such as whether it will rain or not, 0 or 1, true or false, etc.

MAPE has been used in previous ML-based landslide research [12] to evaluate the performance of prediction algorithms. Therefore, in this study, MAPE was used to evaluate the accuracy of linear regression. To measure the accuracy of logistic regression, AUC was used because it has been used in previous research to measure the performance of clustering algorithms.

3. Anomaly detection

Anomaly detection enhances line charts by automatically detecting anomalies within time-series data. It also provides explanations of anomalies to help with root cause analysis. Before delving into the details of anomaly detection, we consider the problem definition.

Problem 1: Given a sequence of real values (i.e., $x = x_1, x_2, x_3, \dots, x_n$), the task of time-series anomaly detection is to produce an output sequence ($y = y_1, y_2, y_3, \dots, y_n$), where $y_i \in \{0, 1\}$ denotes whether x_i is an anomaly point.

The implemented solution was informed by the spectral residual (SR) approach used in the visual saliency detection domain. Then, a convolutional neural network (CNN) was applied to the results produced by the SR model [28].

The SR algorithm consists of three major steps:

1. Fourier transform to obtain the log amplitude spectrum
2. Calculation of SR
3. Inverse Fourier transform to transform the sequence back to the spatial domain:

$$A(f) = \text{Amplitude}(f(x)) \quad (8)$$

$$P(f) = \text{Phase}(f(x)) \quad (9)$$

$$L(f) = \log(A(f)) \quad (10)$$

$$AL(f) = h_q(f) \cdot L(f) \quad (11)$$

$$R(f) = L(f) - AL(f) \quad (12)$$

$$S(x) = \left| |f^{-1}(\exp(R(f) + iP(f)))| \right|, \quad (13)$$

where f and f^{-1} denote Fourier transform and inverse Fourier transform, respectively; x is the input sequence with shape $n \times 1$; $A(f)$ is the amplitude spectrum of sequence x ; $P(f)$ is the corresponding phase spectrum of sequence x ; $L(f)$ is the log representation of $A(f)$, and $AL(f)$ is the averaged spectrum of $L(f)$, which can be approximated by convoluting the input sequence by $h_q(f)$, where $h_q(f)$ is a $q \times q$ matrix defined as:

$$h_q(f) = \frac{1}{q^2} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} \quad (14)$$

$R(f)$ is the SR; that is, the log spectrum $L(f)$ minus the averaged log spectrum $AL(f)$. The SR serves as a compressed representation of the sequence, whereas the innovation part of the original sequence becomes more significant. Finally, the sequence was transferred back to the spatial domain using inverse Fourier transform. The resultant sequence $S(x)$ is referred to as the saliency map [29]. The values of the anomaly points are calculated as follows:

$$x = (\bar{x} + \text{mean})(1 + \text{var}).r + x, \quad (15)$$

where \bar{x} is the local average of the preceding points, mean and var are the mean and variance, respectively, of all points in the current sliding window, and $r \sim N(0, 1)$ is randomly sampled. In this process, CNN instead of raw input is applied

to the saliency map, making the overall process of anomaly detection more efficient [28], [29].

Given that this paper is the first to report the use of anomaly detection in landslide research, AUC was used to measure its performance against that of clustering algorithms in previous studies [7], [9]–[11].

4. Decomposition tree analysis

Decomposition tree visualization is a valuable tool for ad hoc exploration and root cause analysis when visualizing data across multiple filter attributes or dimensions [30].

Our implementation of decomposition analysis enables the visualization of landslide casualty data over a range of landslide feature attributes, namely, trigger, category, setting, size, and country. As shown in Fig. 9, interactive root cause analysis and data exploration are supported by the aggregation of data and drilling down into the dimensions.

For filter attributes $T = \{T^1, T^2, T^3, \dots, T^N\}$, where N is the number of total filter attributes within a dataset (i.e., the cardinality of T , $|T| = N$), each filter attribute can form one or many filtered conditions, as follows:

$$T^1 = \{T_1^1, T_2^1, T_3^1, \dots, T_p^1\}, \text{ such that } |T^1| = P \quad (16)$$

$$T^2 = \{T_1^2, T_2^2, T_3^2, \dots, T_q^2\}, \text{ such that } |T^2| = Q \quad (17)$$

$$T^3 = \{T_1^3, T_2^3, T_3^3, \dots, T_u^3\}, \text{ such that } |T^3| = U \quad (18)$$

$$T^N = \{T_1^N, T_2^N, T_3^N, \dots, T_N^N\}, \text{ such that } |T^N| = V \quad (19)$$

Each filter condition can filter r number of rows ($r \in \{1, 2, 3, \dots, R\}$) from the dataset. For example, when $Country_name = \text{Ecuador}$ was selected, the filter condition T_{58}^1 selected 39 records (i.e., $r = 39$) from the global landslide dataset (where T^1 is the $country_name$ filter attribute, and Ecuador is the 58th item in that attribute).

Continuing on, we defined landslide casualties as:

$$C_i^n = \sum_{r=0}^r (fatality_{count} + 0.5 * injury_{count}),$$

Where, r is the rows effected by filter attribute condition T_i^n (20)

Our decomposition tree visualization (supported by AI) enables the user to find the next filter attribute condition in which to drill down based on either high or low values [30]:

1. High value: This mode considers all available filter attribute conditions and determines that into which to drill down to obtain the highest value of the measure being analyzed. Therefore, the high-value AI split mode finds the most influential filter attribute condition T_i^n for which the highest level of casualty occurs, as represented by:

$$\exists T_i^n \subseteq T | C_i^n > C_j^m, \forall n, m \subseteq \{1, 2, 3, \dots, N\} \wedge \forall i, j \subseteq \{1, 2, 3, \dots\} \quad (21)$$

2. Low value: This mode considers all available filter attribute conditions and determines that into which to drill down to obtain the lowest value of the measure being analyzed. Therefore, the low-value AI split mode finds the most influential filter attribute condition T_i^n

for which the lowest level of casualty occurs, as represented by:

$$\exists T_i^n \subseteq T | C_i^n < C_j^m, \forall n, m \subseteq \{1, 2, 3, \dots, N\} \wedge \forall i, j \subseteq \{1, 2, 3, \dots\} \quad (22)$$

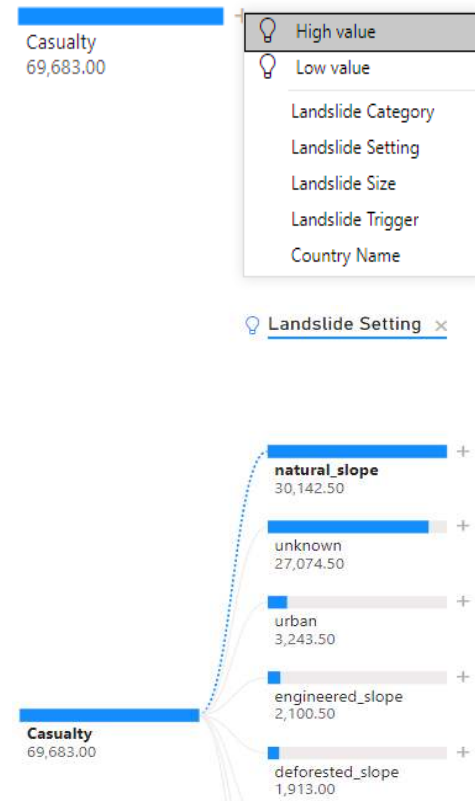


Figure 10. Selecting “High value” for “Casualty” reveals that the highest number of casualties (30,142.50) occurred when the landslide setting was a natural slope.

As shown by Fig. 10, selecting “High value” for the measure “Casualty” reveals that the highest number of casualties (30,142.50) occurred when the landslide setting was a natural slope (i.e. $C_i^n = 30142.50$, where T_i^n represents filter attribute condition “Landslide_Setting = natural slope”). In this way, the AI split allows the user to delve into the root cause. AUC was used to measure the performance of the decomposition tree analysis algorithm.

IV. RESULTS

We executed the ML algorithms described in the previous section (i.e., decomposition analysis, regression analysis, and anomaly detection) on global landslide data containing 14,532 records of landslide events worldwide. By selecting the filter settings for one or more landslide feature attributes, we created particular scenarios from the set of 1.296×10^{64} scenarios, as shown in Equation (7). A change in an attribute filter causes the fact table containing injuries and fatalities to change, as shown in Fig. 4 and Table II.

We used 25 scenarios of 1.296×10^{64} possible scenarios to demonstrate the applicability and usability of the proposed

ML-based knowledge discovery solution. As shown in Table IV, Scenarios 1–3 were based on decomposition analysis, Scenarios 4–22 were based on automated regression analysis, and Scenarios 23–25 were based on anomaly detection.

TABLE IV: RESULTS OF KNOWLEDGE DISCOVERY ON 25 SCENARIOS

Scenario	ML algorithm utilized	Table/figure reference
1–3	Decomposition analysis	Figs. 9, 11 & 12
4–22	Regression analysis	Table IV
23–25	Anomaly detection	Figs. 13, 14, 15, 16, 17, 18, 19

A. DECOMPOSITION ANALYSIS

Using decomposition tree analysis in Scenarios 1–3, our system answered the following strategic questions:

- 16. What causes the highest number of casualties?
- 17. What causes the highest number of casualties when the landslide setting is urban?
- 18. What causes the highest number of casualties when the landslide trigger is a tropical cyclone?

Fig. 9 shows that we delved into the fifth level of detail to identify the causes of the highest number of casualties.

1. Scenario 1 (Fig. 9): What causes the highest number of casualties?

- 19. Level 1 (landslide setting): natural slope
- 20. Level 2 (landslide category): lahar
- 21. Level 3 (country): Colombia
- 22. Level 4 (landslide size): catastrophic
- 23. Level 5 (landslide trigger): unknown.

2. Scenario 2 (Fig. 11): What causes the highest number of casualties when the landslide setting is urban?

- 24. Level 1 (landslide category): landslide
- 25. Level 2 (landslide size): very large
- 26. Level 3 (country): Indonesia
- 27. Level 4 (landslide trigger): earthquake.



Figure 11. Scenario 2 decomposition analysis (i.e., what causes the highest number of casualties when the landslide setting is urban?).

3. Scenario 3 (Fig. 12): What causes the highest number of casualties when the landslide trigger is a tropical cyclone?

- 28. Level 1 (country): Philippines
- 29. Level 2 (landslide setting): natural slope
- 30. Level 3 (landslide size): very large
- 31. Level 4 (landslide category): lahar.

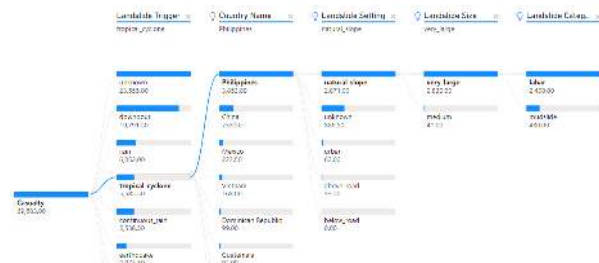


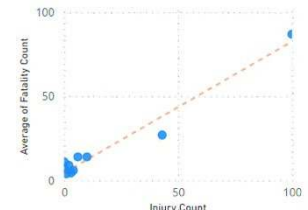
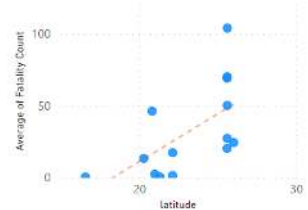
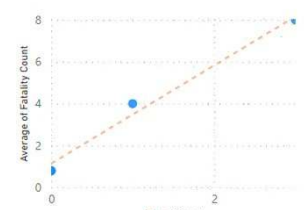
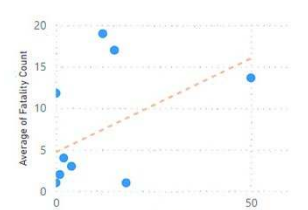
Figure 12. Scenario 3 decomposition analysis (i.e., what causes the highest number of casualties when the landslide trigger is a tropical cyclone?).

B. REGRESSION ANALYSIS

As shown in Fig. 7, the dataset may be filtered by any combination of values from feature attributes such as country, landslide setting, landslide trigger, landslide category, and landslide size. When country name was set to “Italy” and trigger was set to “continuous rain”, the dataset was filtered to only six landslides (containing six injuries and seven fatalities). Therefore, regression analysis was only applied to these six landslide events, finding a positive correlation between fatality count and longitude and a negative correlation between fatality count and latitude (Scenario 10 of Table V). When longitude increased by 3.23, the average fatality count increased by 0.76. When average latitude decreased by 2.71, the average fatality count increased by 0.28.

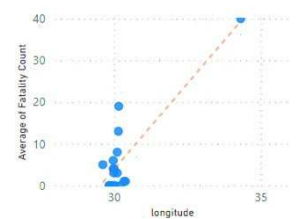
Using regression analysis in Scenarios 4 to 22, our system demonstrated that latitude and longitude exhibit a wide range of behaviors depending on the selected country and landslide features.

TABLE V: SCENARIOS 4–22: AUTOMATED REGRESSION ANALYSIS

Scenario	Scenario condition	Filtered data	Variable correlated with fatality	Correlation direction	Results	Linear graph
4	Country: Myanmar	L = 1,213 I = 278 F = 614	Number of injuries	+	When the number of injuries increases by 26.67, the average fatality count increases by 15.07	<p>← On average when Injury Count increases, Fatality Count also increases.</p> 
5	Country: Myanmar Size: Large	L = 15 I = 243 F = 443	Latitude	+	When latitude increases by 2.82, average fatality count increases by 11.08	<p>← On average when Average of latitude increases, Fatality Count also increases.</p> 
6	Country: Nigeria	L = 24 I = 7 F = 35	Number of injuries	+	When the number of injuries increases by 0.82, the average fatality count increases by 2.11	<p>← On average when Injury Count increases, Fatality Count also increases.</p> 
7	Country: Uganda	L = 97 I = 204 F = 773	Number of injuries	+	When the number of injuries increases by 17.81, the average fatality count increases by 7.48	<p>← On average when Injury Count increases, Fatality Count also increases.</p> 

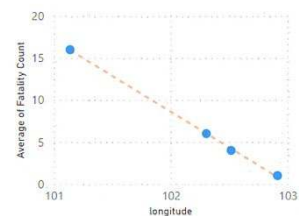
8 Country: Uganda
Trigger: Rain
Size: Medium
L = 26
I = 114
F = 120
Longitude +
When longitude increases by 0.99, the average fatality count increases by 5.5

← On average when longitude increases, Fatality Count also increases.



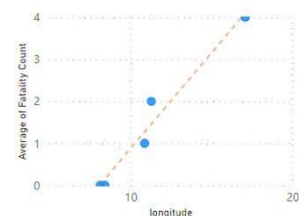
9 Country: Laos
L = 6
I = 0
F = 27
Longitude -
When longitude decreases by 0.66, the average fatality count increases by 4.4

← On average when longitude decreases, Fatality Count increases.



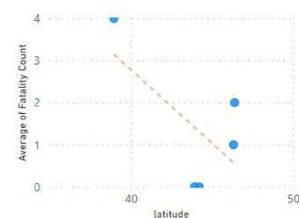
10 Country: Italy
Trigger: continuous rain
L = 6
I = 3
F = 7
Longitude +
When longitude increases by 3.23, the average fatality count increases by 0.76

← On average when longitude increases, Fatality Count also increases.



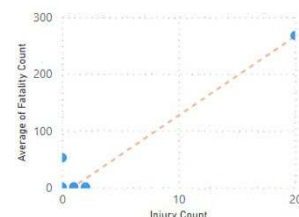
10 Country: Italy
Trigger: continuous rain
L = 6
I = 3
F = 7
Latitude -
When average latitude decreases by 2.71, the average fatality count increases by 0.28

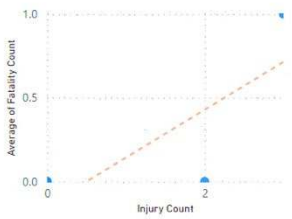
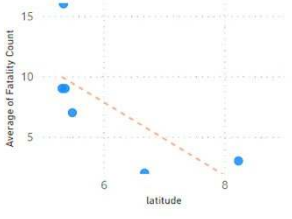
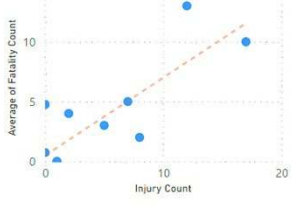
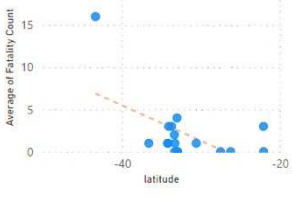
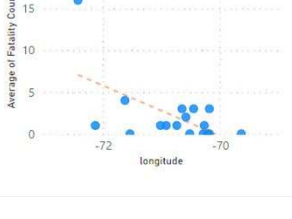
← On average when Average of latitude decreases, Fatality Count increases.



11 Country: Italy
L = 84
I = 33
F = 2534
Number of injuries +
When the number of injuries decreases by 3.65, the average fatality count increases by 87.91

← On average when Injury Count decreases, Fatality Count increases.



12	Country: Lebanon	L = 9 I = 5 F = 1	Number of injuries	+	When the number of injuries increases by 1.25, the average fatality count increases by 0.36	<p>On average when Injury Count increases, Fatality Count also increases.</p> 
13	Country: Cote d'Ivoire	L = 7 I = 12 F = 55	Latitude	-	When latitude decreases by 1.03, the average fatality count increases by 0.82	<p>On average when Average of latitude decreases, Fatality Count increases.</p> 
14	Country: Ecuador	L = 39 I = 52 F = 105	Number of injuries	+	When the number of injuries increases by 4.54, the average fatality count increases by 1.36	<p>On average when Injury Count increases, Fatality Count also increases.</p> 
15	Country: Chile	L = 20 I = 2 F = 36	Latitude	-	When latitude decreases by 5.02, the average fatality count increases by 0.3	<p>On average when Average of latitude decreases, Fatality Count increases.</p> 
16	Country: Chile	L = 20 I = 2 F = 36	Longitude	-	When longitude decreases by 0.73, the average fatality count increases by 0.15	<p>On average when longitude decreases, Fatality Count increases.</p> 

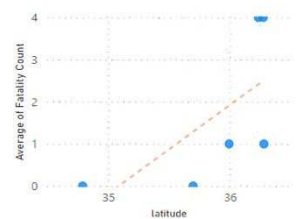
17 Country: Iran
Category: landslide
L = 9
I = 1
F = 10

Latitude

+

When latitude increases by 0.53, the average fatality count increases by 0.59

← On average when Average of latitude increases, Fatality Count also increases.



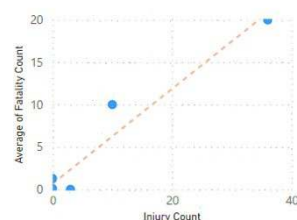
18 Country: Georgia
L = 46
I = 49
F = 54

Number of injuries

+

When the number of injuries increases by 7.52, the average fatality count increases by 4.37

← On average when Injury Count increases, Fatality Count also increases.



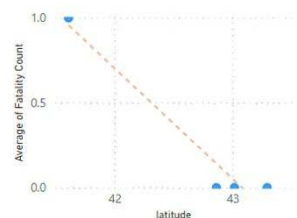
19 Country: Spain
Category: landslide
L = 8
I = 50
F = 1

Latitude

-

When latitude decreases by 0.54, the average fatality count increases by 0.17

← On average when Average of latitude decreases, Fatality Count increases.



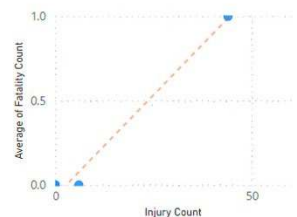
19 Country: Spain
Category: landslide
L = 8
I = 50
F = 1

Latitude

+

When latitude decreases by 19.48, the average fatality count increases by 0.16

← On average when Injury Count increases, Fatality Count also increases.



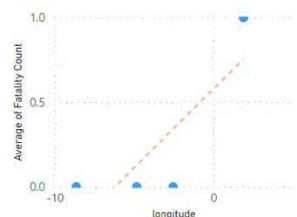
19 Country: Spain
Category: landslide
L = 8
I = 50
F = 1

Longitude

+

When longitude increases by 3.81, the average fatality count increases by 0.09

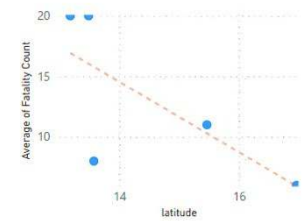
← On average when longitude increases, Fatality Count also increases.



20 Country: Yemen
L = 7
I = 44
F = 65
Latitude

–
When latitude decreases by 1.46, the average fatality count increases by 0.84

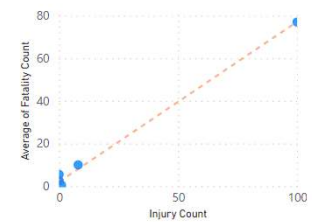
← On average when Average of latitude decreases, Fatality Count increases.



21 Country: Burundi
L = 14
I = 118
F = 125
Number of injuries

+
When the number of injuries increases by 34.10, the average fatality count increases by 23.79

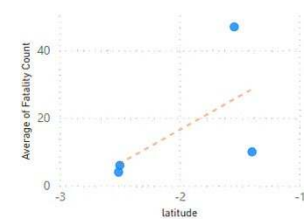
← On average when Injury Count increases, Fatality Count also increases.



22 Country: Democratic Republic of Congo
L = 4
I = 2
F = 67
Latitude

+
When latitude increases by 0.52, the average of fatality count increases by 8.57

← On average when Average of latitude increases, Fatality Count also increases.



Note: L: number of landslides; I: number of injuries; F: number of fatalities.

C. ANOMALY DETECTION

Anomaly detection automatically identifies anomalies in time-series data, along with supporting explanations and the strength of each explanation. As shown in Fig. 13, an anomaly was detected in 2010, when the fatality count was abnormally high (5,424). This value was substantially higher than the expected value of 4,787 and fell outside of the expected range of 4,361–5,212.

Fig. 13 attempts to explain this particular anomaly in Scenario 23, with possible explanations as follows:

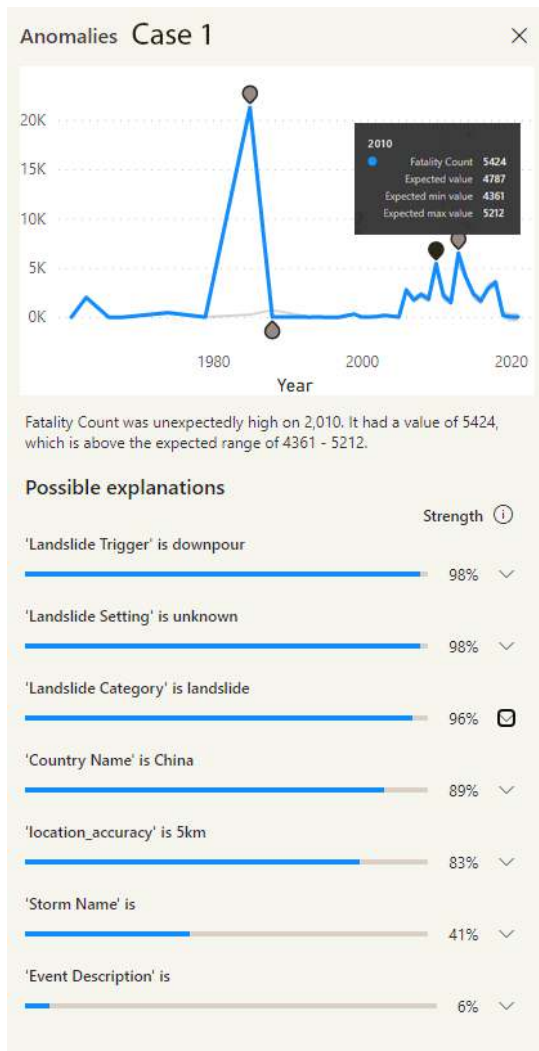
- 32. Landslide trigger: downpour (strength = 98%)
- 33. Landslide setting: unknown (strength = 98%)
- 34. Landslide category: landslide (strength = 96%)
- 35. Country: China (strength = 89%)
- 36. Location accuracy: 5 km (strength = 83%)
- 37. Storm name: empty (strength = 41%)
- 38. Event description: empty (strength = 6%).

Fig. 14 provides possible explanations for Scenario 23 (Anomaly Case 1) in detail.

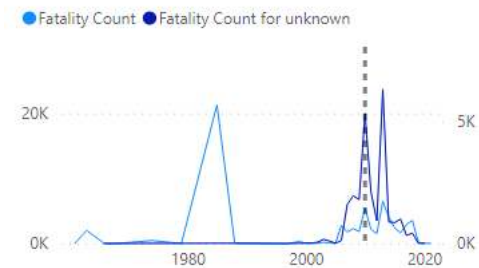
In Scenario 24 (Anomaly Case 2), the total number of landslide events in 2018 was 992, which was higher than the expected range of 664–980 (see Fig. 15). The anomaly detection algorithm in Fig. 15 attempts to explain this anomaly with the following three possible explanations (see Fig. 16):

- 39. Storm name: empty (strength = 43%)
- 40. Landslide trigger: monsoon (strength = 37%)
- 41. Landslide size: large (strength = 28%).

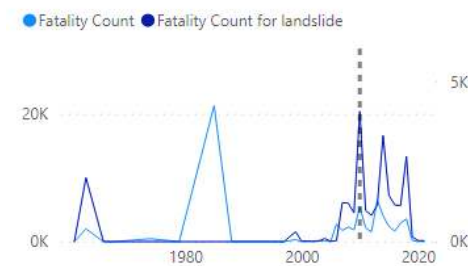
Finally, Scenario 25 (Anomaly Case 3) is depicted in Fig. 17. Here, the injury count in 2010 was exceptionally high at 1,317 (substantially higher than the range 41–159). Fig. 17 also shows a possible explanation for Scenario 25 with the corresponding strengths. One of these explanations is that in 2010, Congo observed a substantially higher number of injuries compared with the usual range, increasing the global number of injuries from landslides in 2010. Our automated knowledge discovery solution assigned a strength of 48% for this explanation (see Fig. 18).



Fatality Count for Landslide Setting unknown was unusually high, which may have lifted the Fatality Count total.



Fatality Count for Landslide Category landslide was unusually high, which may have lifted the Fatality Count total.



Fatality Count for Country Name China was unusually high, which may have lifted the Fatality Count total.

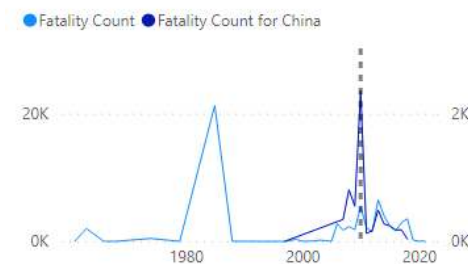
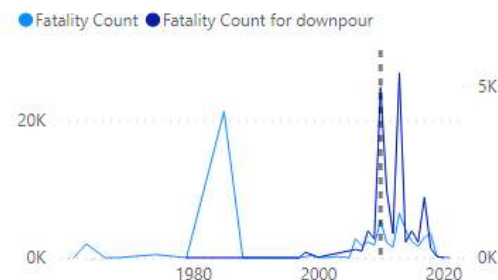


Figure 13. Scenario 23 (Anomaly Case 1): anomaly detection for fatality count of 2010.

Fatality Count for Landslide Trigger downpour was unusually high, which may have lifted the Fatality Count total.



Fatality Count for Storm Name was unusually high, which may have lifted the Fatality Count total.



Fatality Count for Event Description was unusually high, which may have lifted the Fatality Count total.



Count of Country Name for Storm Name was unusually high, which may have lifted the Count of Country Name total.

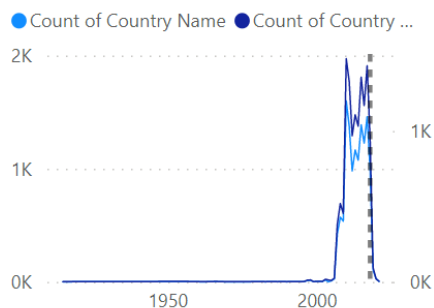
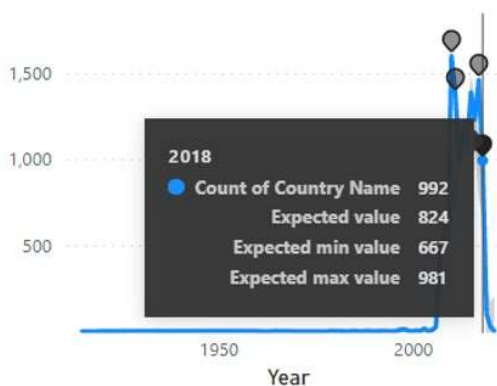


Figure 14. Possible explanations for Anomaly Case 1.

Anomalies Case 2 >> X



Count of Country Name was unexpectedly high on 2,018. It had a value of 992, which is above the expected range of 667 - 981.

Possible explanations

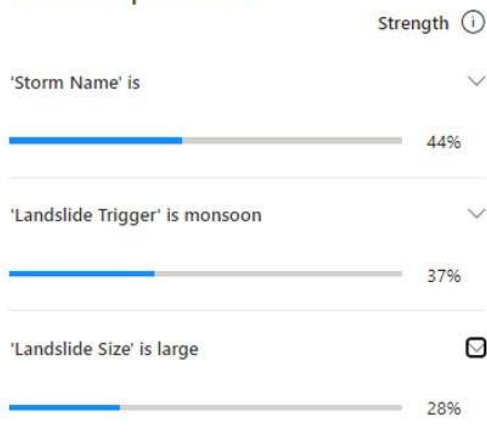
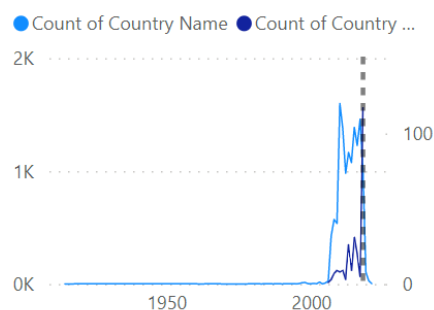


Figure 15. Scenario 24 (Anomaly Case 2): anomaly detection in the number of countries affected in 2018.

Count of Country Name for Landslide Trigger monsoon was unusually high, which may have lifted the Count of Country Name total.



Count of Country Name for Landslide Size large was unusually high, which may have lifted the Count of Country Name total.

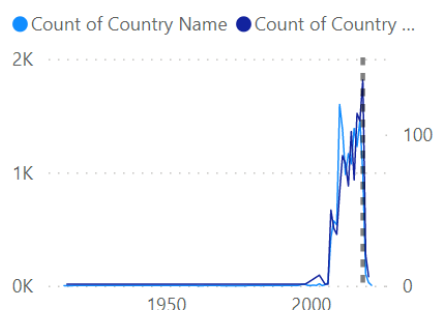


Figure 16. Possible explanations for Anomaly Case 2.



Figure 17. Scenario 25 (Anomaly Case 3): anomaly detection for the number of injuries in 2010.

Therefore, in Scenarios 23 to 25, our system detected and automatically provided explanations for the following three anomalies:

42. In 2010, the fatality count was abnormally high (5,424), which was substantially higher than the expected value of 4,787 and fell outside of the expected range of 4,361–5,212.
43. In 2018, the number of landslide events was 992, which was higher than the expected range of 664–980.
44. In 2010, the number of injuries was exceptionally high at 1,317 (substantially higher than the range of 41–159).

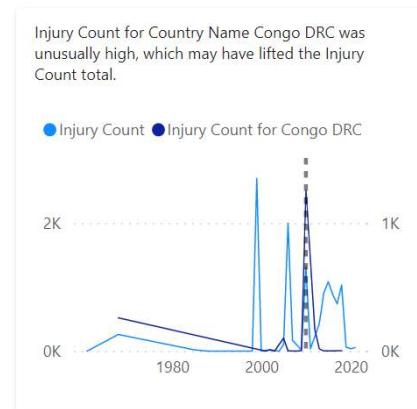


Figure 18. Possible explanations for Anomaly Case 3.

D. VALIDATION OF ML ALGORITHMS

In previous research, ML models have been evaluated using data splitting or cross-validation in which some of the data are used to estimate the model coefficients, and the remainder are used to measure a range of evaluation metrics [7], [9]–[14]. This process of model evaluation is suitable for studies based on a single static dataset. Unlike previous studies, this study reflects a comprehensive dataset of 14,532 global landslide records that dynamically updates to a smaller set of datasets based on the user's selected scenario. Hence, for the present study, there are 1.295×10^{64} smaller sets of dynamic data on which multiple ML algorithms are executed concurrently. Given the feasibility limitations of evaluating models using extremely large dynamic datasets, our ML models were only evaluated using a selected set of scenarios associated with a selected set of data to demonstrate the feasibility of the system.

Sensitivity, specificity, the receiver operating characteristic curve and AUC were used during model evaluation for anomaly detection, logistic regression and decomposition analysis (see Fig. 19). Anomaly detection, logistic regression, and decomposition analysis performed well for all scenarios under experimentation, with AUC averaging 0.951, 0.911, and 0.896, respectively.

In contrast, when evaluating performance of linear regression algorithms, MAPE was used as follows:

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|, \quad (23)$$

where M is MAPE, n is the number of summation iterations, A_t is actual value and F_t is the forecast or predicted value. MAPE has been used for model evaluation by other researchers along with root-mean-square error (RMSE) [12].

In the best-case scenarios, $M < 0.10$ was obtained from Scenarios 4, 6, 9, and 10 (see Table IV). However, for the other scenarios, M was found to be 0.255 on average.

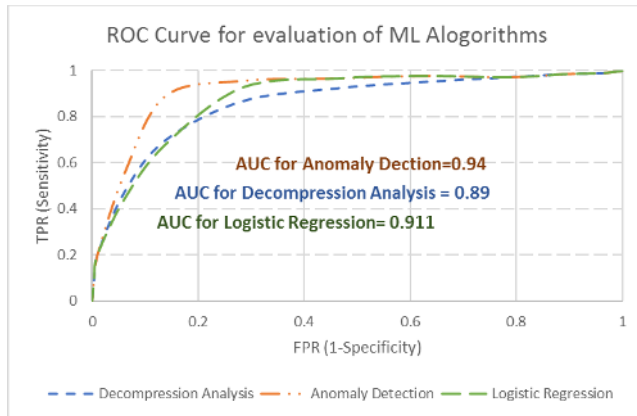


Figure 19. Evaluation of classification algorithms using receiver operating characteristic (ROC) curve and area under the curve (AUC).

V. DISCUSSION

Fig. 20 shows a user receiving an ML-based insight on her mobile device immediately after selecting a particular scenario. The ML-based insight states, “When longitude decreases by 0.66, the average fatality count increases by 4.4”. The prediction accuracy of this insight was MAPE = 0.105, which is higher than that of other algorithms used in landslide research [12]. Therefore, using this instant ML-based insight, the user can decide to increase landslide preparedness in cities at lower longitudes.

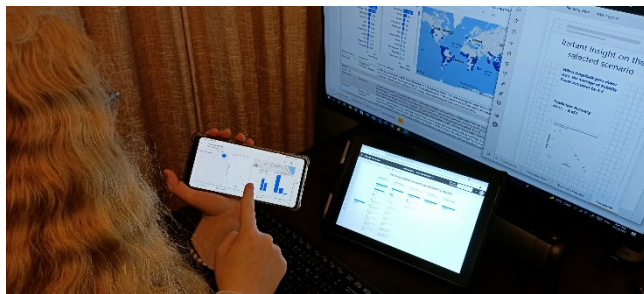


Figure 20. ML-based insight in natural language for a selected scenario using a mobile phone, tablet or personal computer. The insight states, “When longitude decreases by 0.66, the average fatality count increases by 4.4”, with a prediction accuracy of MAPE = 0.105.

In all previous studies, an expert data scientist was in charge of manually preparing and modeling the data and manually training and testing the ML model [7], [9]–[14]. Given that strategic decision-makers are often required to make quick decisions, delegating data science tasks to experts is often not feasible. The system presented in this paper completely automates the task of data preparation and ML modeling. Our decision support system is hosted in Microsoft Cloud, which can be accessed by users on their laptops, tablets or mobile phones. The moment a user selects a scenario from our 1.295×10^{64} possible scenarios, our system automatically prepares the data for that scenario and executes the appropriate ML algorithm to provide hidden insights about the selected scenario to the

decision-maker. A fully automated ML-based decision support system has not been previously reported.

The benefits of the proposed system include the following:

45. It can be executed by users with no prior knowledge of data science and ML algorithms.
46. It instantly prepares the data and executes ML modeling without delay, supporting quick decisions.
47. It is completely scalable, supporting multiple data sources with unlimited concurrent users.
48. It uses two ML algorithms (anomaly detection and decision tree analysis), which have not been used in previous landslide research, with a clustering accuracy of up to AUC = 0.941.
49. It elaborates and explains the result in plain language to the strategic decision maker

Given these benefits, the proposed solution is being trialed by three different town planning agencies.

Being a fully automated decision support system, it lacks the manual rigor of ML modeling with multiple algorithms, as demonstrated in previous research. Therefore, some studies have reported a higher accuracy in classification and prediction using a different set of ML algorithms (e.g., AUC of 0.951 in [9], AUC of up to 0.991 in [7], MAPE of 0.125 in [12]). In our future work, we will include modified versions of random forests and CNNs (as reported in [7] and [9]) as well as support vector regression (as reported in [12]) into our fully automated decision support system, which may improve the accuracy of the current version.

Moreover, the current version only analyzes textual information. In future versions, we plan to include multidimensional data, including imagery, light detection and ranging data, synthetic aperture radar (SAR) and interferometric SAR data [14]. We believe that adding multiple sources of data will increase our system’s capability in terms of more insightful knowledge discovery with higher accuracy.

VI. USER NOTES

The ML-based knowledge discovery solution proposed in this study was implemented using Microsoft Power BI, which is freely available for download from <https://app.powerbi.com/>. The user can download the complete source files (.pbix) along with global landslide data (.csv) files from the authors’ GitHub source control site at <https://github.com/DrSufi/GlobalLandslide> [17]. After downloading and opening the entire solution using MS Power BI Desktop, the user can host the solution either in Microsoft Cloud or a local network to make it available to other researchers or strategic planners.

Typical users of this system are town planners, policymakers, and disaster recovery strategists concerned about landslides in any region or country (the system is capable of generating insights for 157 countries). The system will allow users to understand the characteristics of

landslides in a particular area and provide useful guidance for policy implementation to mitigate the risks associated with landslides in that area.

VII. CONCLUSION

Traditionally, town planners and strategic decision-makers have relied on traditional statistical analyses of regional landslide databases for strategic planning and policy implementation [2], [15], [17], [35], [36]. Previous research into ML-based algorithms has also relied on regional and local datasets, where the data were manually prepared and ML models were manually created by expert data scientists, researchers, and engineers.

In this study, we used advances in ML and AI on NASA's robust database of global landslides to create an automated ML-based solution that provides interactive knowledge discovery in user-defined scenarios with a higher degree of accuracy compared with other ML algorithms in the literature.

We found that anomaly detection had a higher level of accuracy (AUC = 0.941) than decompression tree analysis (AUC = 0.896) and regression (AUC = 0.911, MAPE = 0.255). It should be noted that this study is the first to report on the use of anomaly detection and decompression tree analysis algorithms for analyzing landslide data. Moreover, to the best of our knowledge, our system is the first of its kind to directly provide ML-based hidden trends and insights into a vast set global landslide data containing 1.296×10^{64} scenarios to strategic decision-makers.

In the future version of our solution, we plan to enhance system capacity and capability by studying the potential for:

50. additional landslide data sources
51. additional ML algorithms (e.g. random forest, CNN, support vector regression)
52. additional types of data (e.g. imagery, light detection and ranging data, SAR and interferometric SAR data).

REFERENCES

- [1] Y. W. Rabby and Y. Li, "Landslide inventory (2001–2017) of Chittagong Hilly Areas, Bangladesh," *Data*, vol. 5, no. 1, Art. no. 4, Dec. 2019, doi: 10.3390/data5010004.
- [2] E. Alam, "Landslide hazard knowledge, risk perception and preparedness in southeast Bangladesh," *Sustainability*, vol. 12, no. 16, Art. no. 6305, Aug. 2020, doi: 10.3390/su12166305.
- [3] T. Ben-Nun, M. Besta, S. Huber, A. N. Ziogas, D. Peter, and T. Hoefler, "A modular benchmarking infrastructure for high-performance and reproducible deep learning," in *2019 IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, 2019, pp. 66–77, doi: 10.1109/IPDPS.2019.00018.
- [4] A. Ho, A. Nguyen, J. L. Pafford, and R. Slater, "A data science approach to defining a data scientist," *SMU Data Sci. Rev.*, vol. 2, no. 3, Art. no. 4, 2019. Available: <https://scholar.smu.edu/datasciencereview/vol2/iss3/4>
- [5] F. Sufi, "Analysis of global landslide." Microsoft Power BI. <https://app.powerbi.com/view?r=eyJrIjojYmIzZDU3MmMtNjgxMy00NjgyLWJlZWMtMjcyM2JhMWMwNjY0Yy05NjAwLTJhNzUzZGFjYmEwNSJ9&pageName=ReportSection66d73d81f99936b75aee> (accessed Jun. 2, 2021).
- [6] "Choosing a natural language processing technology in Azure." Microsoft. <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/natural-language-processing> (accessed Jun. 6, 2021).
- [7] H. Wang, L. Zhang, K. Yin, H. Luo, and J. Li, "Landslide identification using machine learning," *Geosci. Front.*, vol. 12, no. 1, pp. 351–364, Jan. 2021, doi: 10.1016/j.gsf.2020.02.012.
- [8] C. Zhou *et al.*, "Landslide characterization applying Sentinel-1 images and InSAR technique: The Muyubao landslide in the Three Gorges Reservoir area, China," *Remote Sens.*, vol. 12, no. 12, Art. no. 3385, Oct. 2020, doi: 10.3390/rs12203385.
- [9] A. M. Youssef and H. R. Pourghasemi, "Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia," *Geosci. Front.*, vol. 12, no. 2, pp. 639–655, Mar. 2021, doi: 10.1016/j.gsf.2020.05.010.
- [10] B. T. Pham, B. Pradhan, D. T. Bui, I. Prakash, and M. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environ. Model. Softw.*, vol. 84, pp. 240–250, Oct. 2016, doi: 10.1016/j.envsoft.2016.07.005.
- [11] C. Zhou *et al.*, "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China," *Comput. Geosci.*, vol. 112, pp. 23–37, Mar. 2018, doi: 10.1016/j.cageo.2017.11.019.
- [12] C. Zhou, K. Yin, Y. Cao, E. Intrieri, B. Ahmed, and F. Catani, "Displacement prediction of step-like landslide by applying a novel kernel extreme learning machine method," *Landslides*, vol. 15, no. 11, pp. 2211–2225, Nov. 2018, doi: 10.1007/s10346-018-1022-0.
- [13] H. Thirugnanam, M. V. Ramesh, and V. P. Rangan, "Enhancing the reliability of landslide early warning systems by machine learning," *Landslides*, vol. 17,

- no. 9, pp. 2231–2246, Sep. 2020, doi: 10.1007/s10346-020-01453-z.
- [14] Z. Ma, G. Mei, and F. Piccialli, “Machine learning for landslides prevention: A survey,” *Neural Comput. Appl.*, vol. 33, no. 17, pp. 10881–10907, Sep. 2020, doi: 10.1007/s00521-020-05529-8.
- [15] E. Alam and N. S. Ray-Bennett, “Disaster risk governance for district-level landslide risk management,” *Int. J. Disaster Risk Reduct.*, vol. 59, Art. no. 102220, Jun. 2021, doi: 10.1016/j.ijdr.2021.102220.
- [16] “How to choose an ML.NET algorithm.” Microsoft. <https://docs.microsoft.com/en-gb/dotnet/machine-learning/how-to-choose-an-ml-net-algorithm> (accessed Jun. 6, 2021).
- [17] B. Ahmed, “The root causes of landslide vulnerability in Bangladesh,” *Landslides*, vol. 18, no. 5, pp. 1707–1720, May 2020, doi: 10.1007/s10346-020-01606-0.
- [18] F. Sufi, Q. Fang, I. Khalil, and S. S. Mahmoud, “Novel methods of faster cardiovascular diagnosis in wireless telecardiology,” *IEEE J. Sel. Areas Commun.*, vol. 27, no. 4, pp. 537–552, May 2009, doi: 10.1109/jsac.2009.090515.
- [19] F. Sufi and I. Khalil, “Diagnosis of cardiovascular abnormalities from compressed ECG: A data mining-based approach,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 33–39, Jan. 2011, doi: 10.1109/titb.2010.2094197.
- [20] F. Sufi and I. Khalil, “A clustering based system for instant detection of cardiac abnormalities from compressed ECG,” *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4705–4713, May 2011, doi: 10.1016/j.eswa.2010.08.149.
- [21] F. Sufi and I. Khalil, “Faster person identification using compressed ECG in time critical wireless telecardiology applications,” *J. Netw. Comput. Appl.*, vol. 34, no. 1, pp. 282–293, Jan. 2011, doi: 10.1016/j.jnca.2010.07.004.
- [22] T. Stanley, “The NASA Cooperative Open Online Landslide Repository (COOLR) points, downloadable as a .csv file.” Aug. 31, 2021. Distributed by NASA Global Landslide Catalogue Points (CSR). <https://maps.nccs.nasa.gov/arcgis/home/item.html?id=eec7aee8d2e040c7b8d3ee5fd0e0d7b9>.
- [23] A. Ferrari, “The importance of star schemas in Power BI.” Sqlbi. <https://www.sqlbi.com/articles/the-importance-of-star-schemas-in-power-bi/> (accessed Jun. 2, 2021).
- [24] F. Sufi, “Machine learning based knowledge discovery solution source files.” Jul. 12, 2021. Distributed by GitHub. <https://github.com/DrSufi/GlobalLandslide> (accessed Jun. 6, 2021).
- [25] H.-F. Yu, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin, “Large linear classification when data cannot fit in memory,” in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '10)*, 2010, pp. 833–842, doi: 10.1145/1835804.1835910.
- [26] H. Matthies and G. Strang, “The solution of nonlinear finite element equations,” *Int. J. Numer. Method. Eng.*, vol. 14, no. 11, p. 1613–1626, 1979, doi: 10.1002/nme.1620141104.
- [27] J. Nocedal, “Updating quasi-Newton matrices with limited storage,” *Math. Comput.*, vol. 35, no. 151, pp. 773–782, Sep. 1980, doi: 10.1090/s0025-5718-1980-0572855-7.
- [28] H. Ren *et al.*, “Time-series anomaly detection service at Microsoft,” in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '19)*, (New York), 2019, pp. 3009–3017, doi: 10.1145/3292500.3330680.
- [29] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multi-context deep learning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2015, pp. 1265–1274, doi: 10.1109/cvpr.2015.7298731.
- [30] “Create and view decomposition tree visuals in Power BI.” Microsoft. <https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-decomposition-tree> (accessed Jun. 10, 2021).
- [31] “Data transformations.” Microsoft. <https://docs.microsoft.com/en-gb/dotnet/machine-learning/resources/transforms> (accessed Jun. 6, 2021).
- [32] “CategoricalCatalog.OneHotEncoding method.” Microsoft. <https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.categoricalcatalog.onehotencoding?view=ml-dotnet> (accessed Jun. 6, 2021).
- [33] “ExtensionsCatalog.ReplaceMissingValues method.” Microsoft. <https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.extensionscatalog.replacemissingvalues?view=ml-dotnet> (accessed Jun. 6, 2021).
- [34] “NormalizationCatalog.NormalizeMeanVariance method.” Microsoft. <https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.normalizationcatalog.normalizemeanvariance?view=ml-dotnet> (accessed Jun. 6, 2021).
- [35] M. Rahman, B. Ahmed, F. Huq, S. Rahman, and T. Al-Hussaini, “Landslide inventory in an urban setting in the context of Chittagong Metropolitan Area, Bangladesh,” in *Proc. 3rd Int. Conf. Adv. Civil Eng.*, (Chittagong, Bangladesh), 2016, pp. 170–178. Available: <http://103.99.128.19:8080/jspui/bitstream/123456789/134/1/36.pdf>.
- [36] B. Ahmed, M. S. Rahman, S. Rahman, F. F. Huq, and S. Ara, “Landslide inventory report of Chittagong Metropolitan Area, Bangladesh,” Bangladesh University of Engineering and Technology (BUET)–Japan Institute of Disaster

Prevention & Urban Safety, Dhaka, Bangladesh, Aug. 2014. Accessed: May 29, 2021. [Online] Available: <http://www.landslidebd.com/wp-content/uploads/2014/09/Landslide-Inventory-Report-BAYES.pdf>.



Dr. Fahim Sufi is a senior artificial intelligence solution architect with the federal government. He has held lead solution architect roles in several federal and state government agencies, including the Australian Department of Defence, the Australian Institute of Family Studies, the Victorian Department of Health, and the Victorian Department of Human Services. He obtained his PhD in computer science and information technology as well as a Master of Engineering in Computer Systems from RMIT University, Australia. His research interests include artificial intelligence, machine learning, software development, big data analysis, cyber security, and encryption.

Dr. Sufi currently possesses several highly revered industry certifications from Microsoft, The Open Group, and others. His certifications include TOGAF 9 (Level 1 and Level 2), ArchiMate 3 (Level 1 and Level 2), Microsoft-certified data analyst associate, PRINCE2 (foundation and practitioner), ITIL v3, and several other Microsoft certifications in cloud technologies. He has published many articles in top-ranking journals, including the *IEEE Journal on Selected Areas in Communication* and *IEEE Transactions on Information Technology in Biomedicine*.

Dr. Sufi's PhD thesis received the best PhD thesis award from the School of Computer Science and Information Technology at RMIT University in 2011. He has won several other awards and scholarships, including an Australian Postgraduate Award, the Victorian Government ICT Postgraduate Scholarship, the Commonwealth Government Commercialization Scheme, RMIT School of Electrical and Computer Engineering (SECE) Scholarship, IEEE and SECE Best Literature Review Award, and a Victorian Government Certificate of Appreciation.



Dr. Musleh Alsulami is an assistant professor of information systems at Umm Al-Qura University. He completed a BSc in software engineering at Imam University, KSA, in 2004, an MSc in information technology at Monash University, Australia, in 2010, and a PhD in information systems at Monash University, Australia, in 2017.

His current research interests include enterprise resources planning (ERP), including ERP life cycles, implementation conflicts, stakeholders, and cloud-based ERP, as well as digital transformation in government organizations, software quality, and human-computer interactions.