The Journal of Engineering Geology
pISSN : 1226-5268 eISSN : 2287-7169

Vol. 33, No. 4, December, 2023, pp. 673-687
https://doi.org/10.9720/kseg.2023.4.673

RESEARCH ARTICLE

데이터 탐색을 활용한 딥러닝 기반 제천 지역 산사태 취약성 분석

안상아[1] · 이정현[2] · 박혁진[3*]

[1]세종대학교 에너지자원공학과 석사과정, [2]세종대학교 지구자원시스템공학과 박사후 연구원,
[3]세종대학교 지구자원시스템공학과 교수

Assessment of Landslide Susceptibility in Jecheon Using Deep Learning Based on Exploratory Data Analysis

Sang-A Ahn[1] · Jung-Hyun Lee[2] · Hyuck-Jin Park[3*]

[1]Master Candidate, Department of Energy and Mineral Resources Engineering, Sejong University
[2]Post-Doctoral Fellow, Department of Energy Resources and Geosystem Engineering, Sejong University
[3]Professor, Department of Energy Resources and Geosystem Engineering, Sejong University

Abstract

Exploratory data analysis is the process of observing and understanding data collected from various sources to identify their distributions and correlations through their structures and characterization. This process can be used to identify correlations among conditioning factors and select the most effective factors for analysis. This can help the assessment of landslide susceptibility, because landslides are usually triggered by multiple factors, and the impacts of these factors vary by region. This study compared two stages of exploratory data analysis to examine the impact of the data exploration procedure on the landslide prediction model's performance with respect to factor selection. Deep-learning-based landslide susceptibility analysis used either a combinations of selected factors or all 23 factors. During the data exploration phase, we used a Pearson correlation coefficient heat map and a histogram of random forest feature importance. We then assessed the accuracy of our deep-learning-based analysis of landslide susceptibility using a confusion matrix. Finally, a landslide susceptibility map was generated using the landslide susceptibility index derived from the proposed analysis. The analysis revealed that using all 23 factors resulted in low accuracy (55.90%), but using the 13 factors selected in one step of exploration improved the accuracy to 81.25%. This was further improved to 92.80% using only the nine conditioning factors selected during both steps of the data exploration. Therefore, exploratory data analysis selected the conditioning factors most suitable for landslide susceptibility analysis and thereby improving the performance of the analysis.

Keywords: landslide susceptibility, exploratory data analysis, correlation coefficient, feature importance, multi-layer perceptron

초 록

데이터 탐색은 수집한 데이터를 다양한 각도에서 관찰 및 이해하는 과정으로 데이터 구조 및 특성 분석을 통해 데이터의 분포와 상관관계를 파악하는 과정이다. 일반적으로 산사태는 다양한 인자들에 의해

OPEN ACCESS

***Corresponding author:** Hyuck-Jin Park
E-mail: hjpark@sejong.ac.kr

Received: 12 December, 2023
Revised: 26 December, 2023
Accepted: 26 December, 2023

Ⓒ 2023 The Korean Society of Engineering Geology

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

유발되고 발생 지역에 따라 유발 인자들이 미치는 영향이 상이하기 때문에 산사태 취약성 분석 이전에 데이터 탐색을 통해 유발 인자 사이의 상관관계를 파악하고 특징적인 유발 인자를 선별한다면 효과적인 분석을 수행할 수 있다. 따라서 본 연구는 데이터 탐색이 예측 모델의 성능에 미치는 결과를 확인하기 위해 두 단계에 걸친 데이터 탐색을 수행하여 인자를 선별하고, 선별된 유발 인자들 사이의 조합과 23개의 전체 유발 인자 조합을 활용하여 딥러닝 기반의 산사태 취약성 분석을 진행하였다. 데이터 탐색 과정에서는 Pearson 상관계수 heat map과 random forest의 인자 중요도 histogram을 활용하였으며, 딥러닝 기반 산사태 취약성 분석 결과의 정확도는 분석을 통해 획득한 산사태 취약 지수 값을 이용해 제작한 산사태 취약성 지도를 confusion matrix 기반의 정확도 검증 방법을 통해 분석하였다. 분석 결과, 전체 23개의 인자를 사용한 산사태 취약성 해석 결과는 55.90%의 낮은 정확도를 보였지만 한 단계의 탐색을 거쳐 선별한 13개 인자를 활용한 취약성 해석 결과는 81.25%의 분석 정확도를 보였고, 두 단계 데이터 탐색을 모두 수행하여 선별된 9개의 유발 인자를 활용한 산사태 취약성 분석 결과는 92.80%로 가장 높은 정확도를 보였다. 따라서 데이터 탐색을 통해 특징적인 유발 인자를 선별하고 분석에 활용하는 것이 산사태 취약성 분석에서 더 좋은 분석 성능을 기대할 수 있음을 확인하였다.

주요어: 산사태 취약성, 데이터 탐색, 상관계수, 인자 중요도, 다층 퍼셉트론

서 론

산사태 취약성 분석은 산사태를 유발하는 다양한 인자들의 공간 분석을 통해 산사태가 발생할 가능성이 있는 지역을 예측하는 방법이다. 다양한 산사태 취약성 분석 기법 중에서 가장 많은 연구가 진행되고 있는 데이터 기반 분석 기법(data-driven methods)은 과거의 산사태 발생 위치와 유발 인자 사이의 상관성을 분석하여 미래에 산사태가 발생할 가능성을 예측하는 기법이다(Park and Lee, 2022). 과거에는 주로 전통적인 통계 및 확률 기법을 사용하여 통계적 분석 기법(statistical methods)으로 분류되었지만, 최근 컴퓨팅 기술의 발전과 좀 더 빠르고 복잡한 알고리즘이 개발되면서 비선형적 정보들을 대상으로 분석할 수 있는 다양한 분석 기법들이 제안되고 있다. 특히 최근 들어 인공지능(Artificial Intelligence, AI) 기법이 산사태 취약성 분석에 활발히 적용되면서 예측 기법의 성능(performance) 향상에 상당한 발전이 이루어지고 있다.

데이터 기반 산사태 취약성 분석에서 사용되는 다양한 인공지능 기법 중에서 인공신경망(Artificial Neural Network, ANN)을 기반으로 하는 딥러닝(Deep Learning) 기법이 활발히 적용되고 있다. 딥러닝은 분류 또는 회귀 문제에 가장 널리 사용되는 신경망 구조인 MLP(Multi-Layer Perceptron) 모델을 기반으로 다수의 은닉층(hidden layers)을 포함하는 신경망 모델이다. 딥러닝 기법은 신경망의 깊이가 깊어질수록 복잡한 분석을 수행하는 데 유용하므로 다양한 산사태 유발 인자의 비선형적 상관관계를 분석하기에 적합한 기법이다. 이러한 특징 때문에 대부분의 딥러닝 기반 산사태 취약성 연구에서는 획득할 수 있는 산사태 유발 인자 자료를 모두 적용하여 분석을 수행한다. 그러나 유발 인자 사이의 상관성이 매우 높거나 산사태 발생과 관련성이 현저히 낮은 인자를 분석에 적용할 경우, 예측 기법의 성능 향상을 기대하기 어려울 수 있다. 따라서 산사태 취약성 분석과 같이 다양한 유발 인자를 사용하여 분석을 수행하는 경우, 필수적으로 데이터 탐색(Exploratory Data Analysis, EDA)을 수행하여야 한다.

데이터 탐색은 수집한 데이터를 다양한 각도에서 관찰하고 이해하는 과정으로 데이터 활용 방법에 따라 모델링이 결정되고 분석 기법의 예측 성능에 결정적인 영향을 미치게 된다. 따라서 탐색을 통해 산사태 발생 위치와 유발 인자 사이의 상관관계를 파악하여 중요도가 낮은 인자를 제외하고, 특징적인 유발 인자를 선별하여 효과적인 분석을 수행할 수 있다. 결과적으로 데이터 탐색을 통해 예측 모델의 성능이 향상될 뿐만 아니라 분석에 요구되는 자료처리 시간도 단축할 수 있다.

산사태 유발 인자의 데이터 탐색

데이터 탐색은 수집한 데이터를 다양한 각도에서 관찰 및 이해하는 과정으로 데이터 구조와 특성을 통해 데이터의 분포와 상관관계를 파악하는 과정이다. 히스토그램과 같은 그래프를 이용한 시각화를 통해 데이터의 분포와 상관관계를 파악하거나 데이터의 구조적 특성을 알아내기 위한 통계적 분석 기법들을 사용한다(Huh and Jung, 1990; Brownley, 2016). 이러한 탐색을 통해 산사태의 다양한 패턴과 산사태와 유발 인자 사이의 관계를 파악하고 특징적인 유발 인자를 선정할 수 있다. 또한 불필요한 산사태 유발 인자를 제외하여 산사태 취약성 분석을 위한 딥러닝 모델의 성능을 향상시킬 수 있다는 장점이 있다.

상관계수 분석

상관계수 분석은 데이터 탐색에서 널리 사용되는 방법이다. 이 방법은 크거나 고차원인 데이터셋에서 한 변수와 다른 하나 이상의 변수 간의 상관 정도를 파악하는 데이터 탐색 기술이다. 분석 결과는 다변량 데이터셋을 탐색하고 단순화하는데 사용될 수 있다. 가장 잘 알려진 상관계수에는 Pearson 상관계수, Spearman 상관계수, Kendall 순위 상관계수가 있다(Xiao et al., 2016). 그중 본 연구에서 사용한 Pearson 상관계수는 1895년에 Karl Pearson이 제안한 방법으로 두 변수 간 선형 관계의 강도를 측정한다(Hauke and Kossowski, 2011). 두 변수 X, Y 사이의 선형상관관계(r_{XY})를 측정하는 식은 다음 식 (1)과 같이 정의할 수 있다.

$$r_{XY} = \frac{\frac{1}{n}\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\left(\frac{1}{n}\sum(x_i - \overline{X})^2\right)\left(\frac{1}{n}\sum(y_i - \overline{Y})^2\right)}} \tag{1}$$

여기서, n은 자료 쌍(paired)의 수 이다. 결과값은 -1에서 1 사이의 값을 가지며 1은 완전한 양의 상관을 나타내고 -1은 완전한 음의 상관을 0은 완전히 상관이 없음을 나타낸다(Lee and Kim, 2006). 본 연구에서는 Pearson 상관계수를 산사태 유발 인자 간의 상관관계를 파악하여 상관성이 높은 인자들을 골라내는 데 활용하였다.

인자 중요도

인자 중요도(feature importance)는 random forest 패키지에서 제공하는 추가적인 정보로, random forest의 트리 생성 과정에서 각 산사태 유발 인자들이 얼마나 중요하게 작용했는지 평가하는 값이다. Random forest 알고리즘은 각 인자가 split될 때 불순도(imputity) 감소의 평균을 모든 트리에 걸쳐 계산한 값인 Mean Decrease Impurity(MDI)을 통해 인자 중요도를 평가한다(Breiman, 2001; Liaw and Wiener, 2002; Biau and Scornet, 2016; Muller and Guido, 2016; Kang and Park, 2019). M개의 트리를 집계하여 얻은 랜덤 포레스트의 경우, 변수($X^{(j)}$)의 MDI는 다음 식 (2)와 같이 정의된다.

$$\widehat{MDI}(X^{(j)}) = \frac{1}{M}\sum_{\ell=1}^{M}\sum_{\substack{t\in T_\ell \\ j^*_{n,t}=j}} p_{n,t} L_{reg,n}(j^*_{n,t}, z^*_{n,t}) \tag{2}$$

여기서, $p_{n,t}$는 노드 t의 관측치 하락 비율이며, $T_{\ell_{1 \le \ell \le M}}$은 random forest의 트리 집합이다. 그리고 $(j^*_{n,t}, z^*_{n,t})$은 노드 t에서 경험적 기준을 최대화하는 분할이다(Biau and Scornet, 2016). 데이터의 결과값은 0에서 1 사이의 양수 값을 가지며 값이 클수록 중요도가 높다. Random forest의 인자 중요도는 특히 특성을 선택해야 할 때 어떤 특성이 중요한지 빠르게 확인할 수 있어 편리하다(Géron, 2022). 본 연구에서는 상관계수 분석을 통해 얻어낸 서로 상관성이 높은 데이터들 중 어떤 인자가 가장 중요한지 파악하는 데 활용하였다.

딥러닝

딥러닝은 인간의 두뇌에서 영감을 얻은 인공신경망 기반의 머신러닝 기법이다. 딥러닝은 입력층, 은닉층, 출력층의 3단계 구조로 되어 있으며, 은닉층의 뉴런(노드)이 여러 개로 층층이 쌓여 있다. 인공신경망의 뉴런들은 서로 연결되어 있으며 입력 데이터와 응답 간의 복잡한 관계가 있는 문제를 모델링하는데 적합하다. 그렇기 때문에 산사태와 같은 복잡한 현상의 모델링에 유용하게 사용될 수 있다(Tien Bui et al., 2016; Lee et al., 2021). 다양한 딥러닝 기법 중 본 연구에서 활용한 MLP 모델은 분류나 회귀 문제에서 가장 널리 사용되는 신경망 구조로 은닉층이 여러 개인 모델을 말한다. MLP의 은닉층은 비선형 활성화 함수(non-linear activation function)를 통해 예측을 수행하고, 순전파(feedforward)와 역전파(backpropagation)를 반복하며 모델을 훈련한다. 순전파는 신경망 네트워크에 입력 데이터의 패턴을 제시하고, 역전파는 출력층에서 시작하여 직전 계층을 통과하면서 가중치 조절 및 오류를 최소화한다(Gomez and Kavzoglu, 2005; Lee et al., 2021). 본 연구에서는 3개의 은닉층을 사용했으며, 신경망 모델에서 가장 많이 사용되는 활성화 함수인 ReLU(Rectified Linear Unit)을 적용하여 가중치를 계산하였다. 또한 epoch 반복 횟수는 1,000, learning rate는 0.0005로 설정하여 분석을 진행하였다.

연구 지역 및 데이터베이스 구축

연구 지역

본 연구에서는 2020년 7~8월에 걸친 집중호우로 인해 304개소의 산사태가 발생한 충청북도 제천시 일대를 연구 지역으로 선정하였다(Figs. 1a and 1b). 연구 지역은 지리좌표 상으로 경도 128° 2' 27.59"~128° 18' 29.37", 위도 37° 1' 19.29"~37° 14' 11.33"에 위치하며, 해발고도 113 m에서 982 m 사이의 평균 337 m에 해당하는 구릉 지형의 형태를 보인다. 지질학적으로는 연구 지역 중앙에 중생대 백악기 반상 화강암과 쥬라기 흑운모 화강암, 북서쪽에는 선캄브리아기 편마암이 주로 분포해 있고 제4기 충적층이 일부 존재한다.

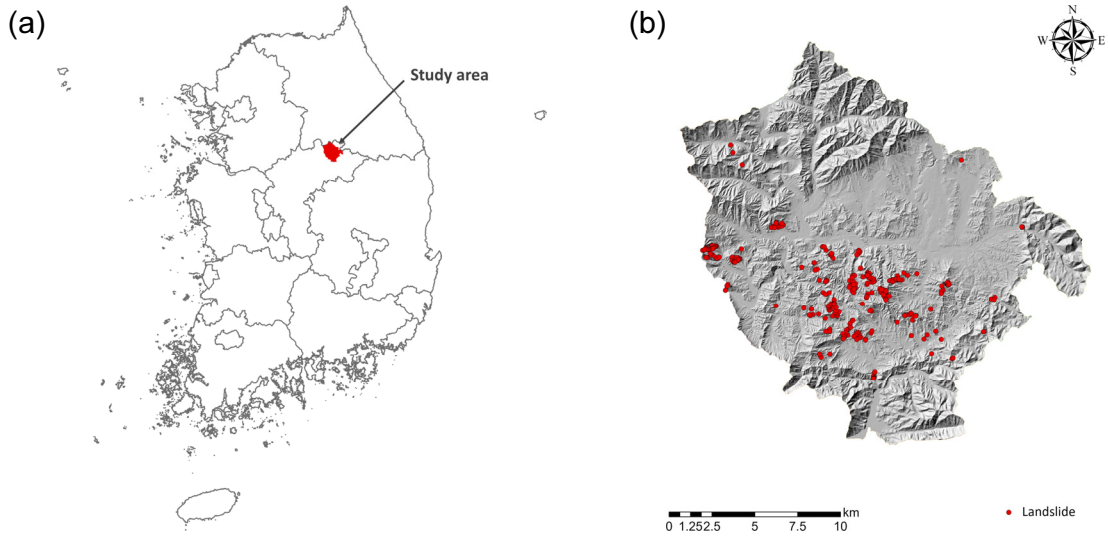

Fig. 1. Study area: (a) location of Jecheon area, (b) landslide locations in Jecheon.

공간 데이터베이스 구축

산사태 취약성 분석에 있어 산사태 발생 위치 정보는 필수적이다. 산사태 발생 위치 정보를 획득하기 위해서 연구 지역의 산사태 발생 시점 전후의 항공사진과 위성사진을 비교하여 위치를 결정하는 항공사진의 시각적 해석은 산사태 위치 지도화의 대표적인 접근법 중 하나이다(Kääb, 2002; Martha et al., 2010; Lee and Oh, 2019). 본 연구에서도 산사태 발생 전후의 사진을 비교하여 산사태 발생 위치를 결정하였으며, 1:5,000 축척의 수치지형도를 이용하여 GIS 기반 벡터 데이터 형태의 산사태 발생 위치도로 작성하였다.

한편 산사태는 다양한 요인의 상호작용으로 발생하기 때문에 연구 지역으로부터 요인들을 획득하는 것이 중요하다. 본 연구는 이용 가능한 공간데이터베이스를 통해 지형, 임상, 토양, 지질, 토지 이용 등과 관련된 인자를 획득하였다. 지형 인자는 국토지리정보원(NGII)에서 2019년 제작된 1:5,000 축척의 수치지형도를 통해 디지털 고도 모델(Digital Elevation Model, DEM)을 획득하였다. 그리고 DEM을 사용하여 고도(elevation, Fig. 2a), 사면 경사(slope angle, Fig. 2b), 사면 방향(aspect, Fig. 2c), 표준 곡률(standard curvature, Fig. 2d), 횡단 곡률(planform curvature, Fig. 2e), 종단 곡률(profile curvature, Fig. 2f), SCA(Specific Catchment Area, Fig. 2g), TWI(Topographic Wetness Index, Fig. 2h), SPI(Stream Power Index, Fig. 2i), STI(Sediment Transport Index, Fig. 2j)를 계산하여 획득하였다. 임상 인자는 산림청에서 제공하는 1:5,000 축척의 2019년 임상도에서 임상 종류(forest type, Fig. 2k), 영급(timber age, Fig. 2l), 경급(timber diameter, Fig. 2m), 임분고(forest height, Fig. 2n), 임상 밀도(forest density, Fig. 2o)를 추출하였다. 한편, 토양 인자는 농촌진흥청에서 제공하는 1:25,000 축척의 토양도에서 유효 토심(soil depth, Fig. 2p), 표토 토성(soil texture, Fig. 2q), 심토 토성(soil sub-texture, Fig. 2r), 배수 등급(soil drainage, Fig. 2s)을 획득하였으며, 암종(lithology, Fig. 2t) 인자는 한국지질자원연구원에서 제공하는 1:50,000 축척의 수치지질도에서 추출하였다. 토지 이용(land use, Fig. 2u)은 환경부에서 제공하는 2019년 제작된 1:5,000 축척의 토지피복지도에서 추출한 토지 이용 주제도에 제천시의 산사태 발생 특성을 고려하기 위해 임상도에서 추출한 벌채지(미립목지) 데이터를 추가하여 제작하였다. 기타 인자로는 산림청에서 제공하는 1:5,000 임도망도으로부터 임도를, 그리고 지질도로부터 단층을 추출하여 각각 임도와의 거리(distance from forest road, Fig. 2v), 단층과의 거리(distance from fault, Fig. 2w)를 GIS를 통해 보간하여 획득하였다. 획득한 인자들은 총 23개로 축척을 고려하여 10 m 래스터 형식으로 공간 데이터베이스를 구축하였다.

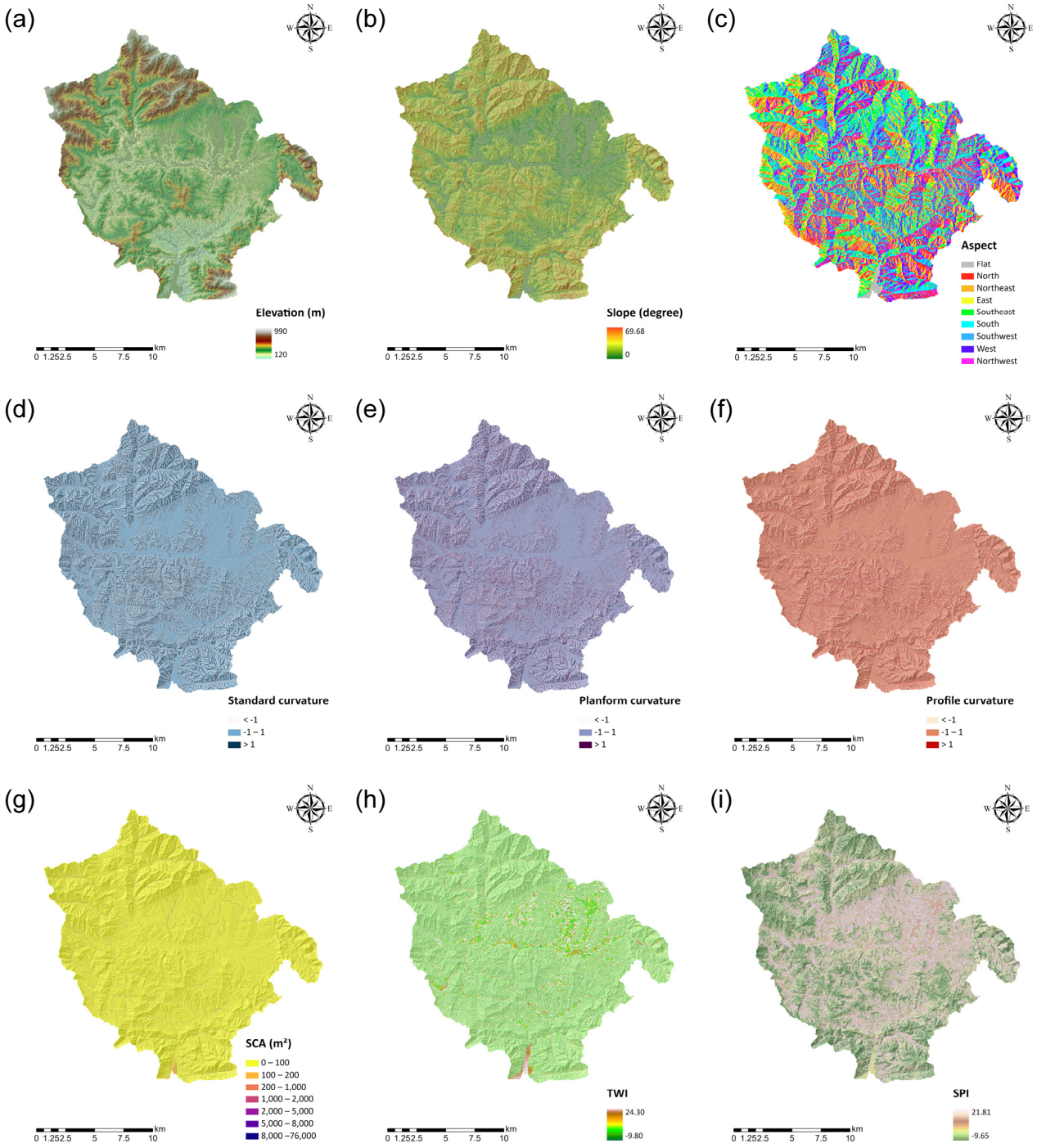

Fig. 2. Thematic maps of landslide conditioning factor: (a) elevation, (b) slope angle, (c) aspect, (d) standard curvature, (e) planform curvature, (f) profile curvature, (g) SCA, (h) TWI, (i) SPI, (j) STI, (k) forest type, (l) timber age, (m) timber diameter, (n) forest height, (o) forest density, (p) soil depth, (q) soil texture, (r) soil sub-texture, (s) soil drainage, (t) lithology, (u) land use, (v) distance from forest road, (w) distance from fault.

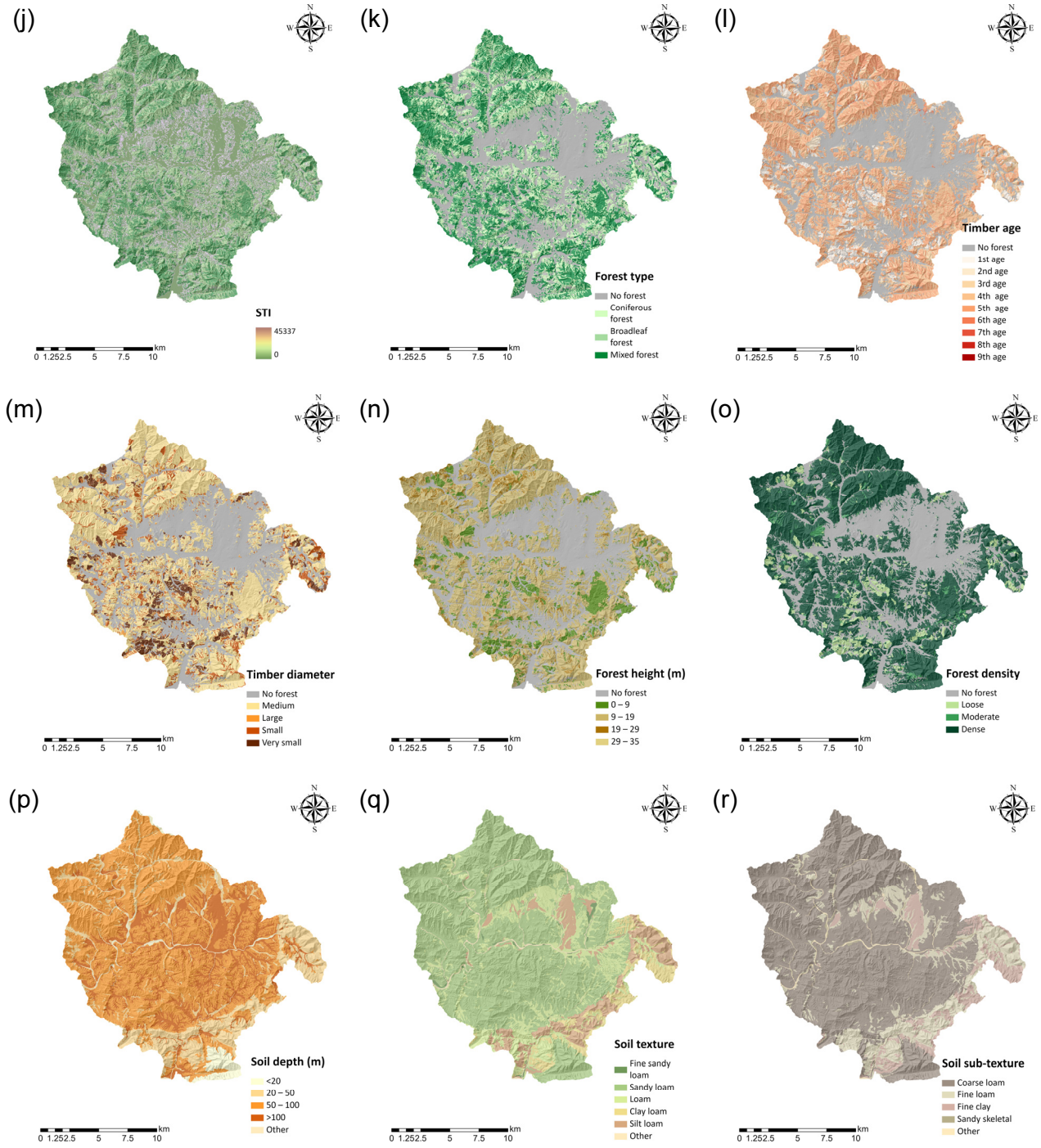

Fig. 2. Continued.

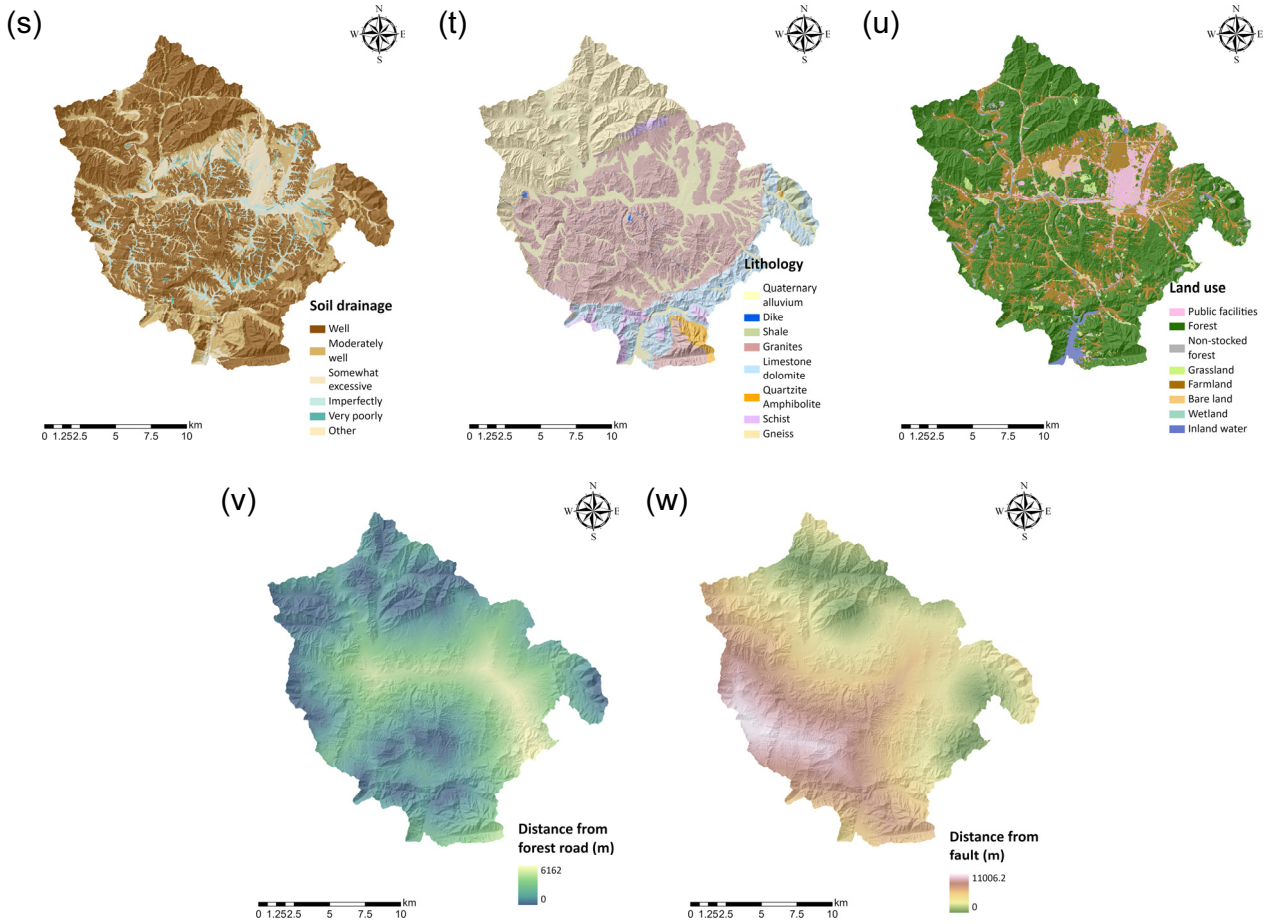

Fig. 2. Continued.

분석 데이터셋 구축

산사태 취약성 분석을 위한 딥러닝 모델링을 위해 획득한 산사태 위치 및 유발 인자 데이터는 훈련 데이터(train data)와 검증 데이터(test data)로 분할이 필요하다.

산사태 데이터는 산사태 발생부(332개)의 개수에 비해 미발생부(3,080,103개)의 개수가 많은 불균형 데이터이기 때문에 훈련 데이터셋 내의 산사태 위치 수와 샘플링 전략은 분석의 품질에 큰 영향을 미친다(Kalantar et al., 2018). 연구지역에 따라 산사태의 발생 개수가 다르기 때문에 훈련에 사용될 자료의 갯수가 다르고, 훈련 데이터와 검증 데이터의 비율을 나누는 구성 방식 역시 논문에 따라 상이하다. 그러나 대부분의 관련 연구에서는 훈련 데이터의 산사태 발생부와 미발생부의 비율이 1:1이 되도록 랜덤하게 선택하는 방식을 사용하고 있으며, 따라서 본 연구에서도 산사태 발생부와 미발생부를 각 259개 씩 1:1 비율로 샘플링하였다(Bui et al., 2011; Tien Bui et al., 2016; Kalantar et al., 2018; Kang and Park, 2019). 딥러닝 분석을 위한 훈련 데이터와 검증 데이터 분할 비율은 보편적으로 사용하는 비율인 70:30을 사용하였다.

데이터 탐색을 고려한 산사태 유발 인자 선정

본 연구에서는 산사태 취약성 분석에 앞서 데이터 탐색 기법을 적용하여 산사태 발생 위치와 다양한 산사태 유발 인자의 상관성을 분석하였고, 산사태 발생에 영향을 미치는 중요도가 높은 인자순으로 분석에 사용되는 유발 인자를 선정하

였다. 선정된 유발 인자 조합은 딥러닝 기법인 MLP 모델에 적용하여 산사태 취약성 분석을 수행하였다. 제안된 방법을 통해 구축된 예측 모델의 성능을 검증하기 위해 전체 유발 인자 조합을 활용한 취약성 분석 결과와 각각의 인자 조합에 따른 산사태 취약성 분석 결과의 정확도를 비교 및 검증하였다.

상관관계 분석

먼저 데이터 탐색 기법을 적용하기 위해서 연구 지역의 산사태 유발 인자 총 23개의 주제도를 분석에 적용가능한 데이터베이스 형태로 구축하였다. 구축된 데이터셋은 상관계수 분석에 대입하여 산사태 유발 인자의 Pearson 상관계수를 계산하였고, 이를 Fig. 3과 같이 heat map으로 구축하였다.

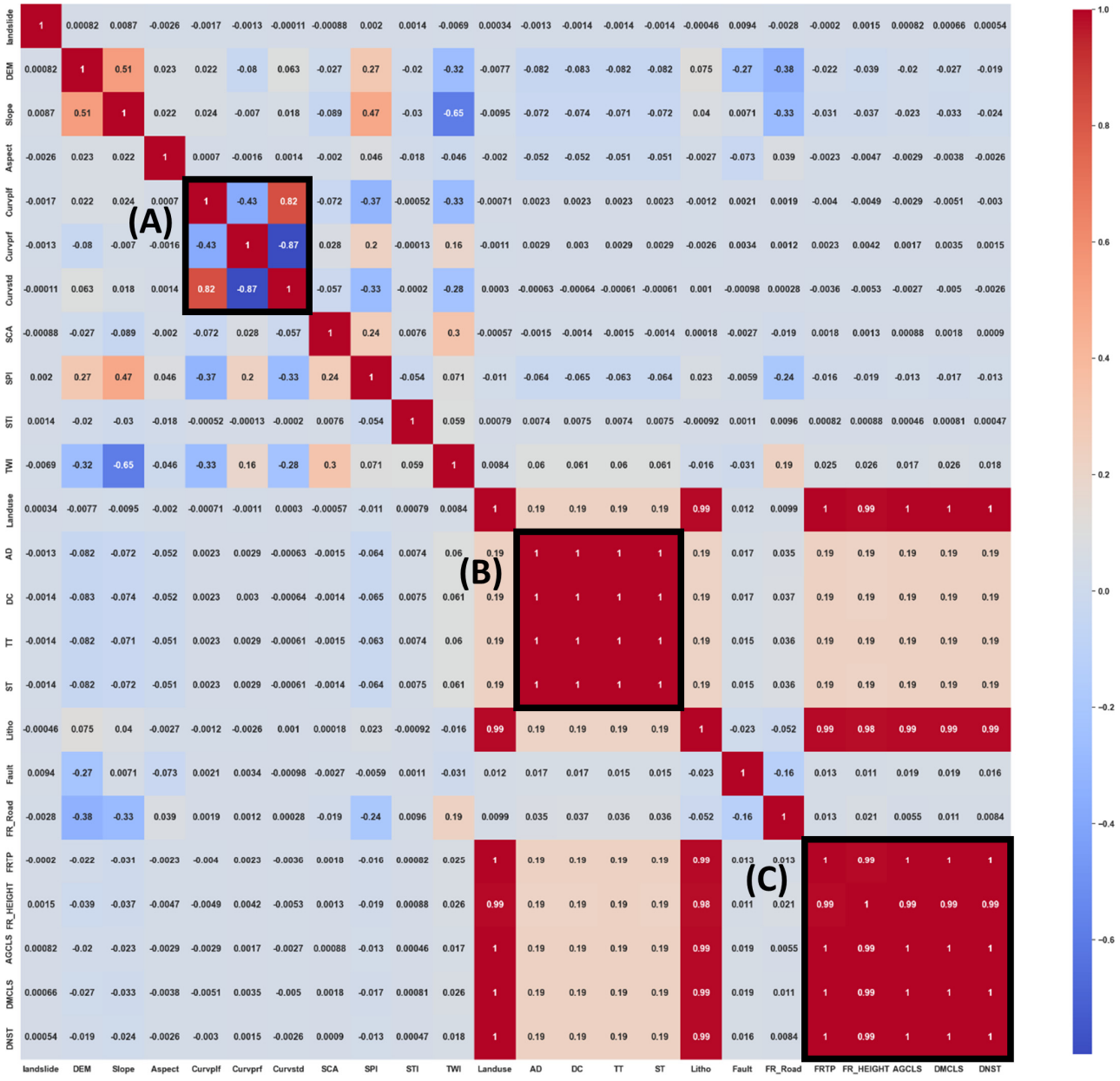

Fig. 3. Pearson correlation coefficient heat map.

Pearson 상관계수를 분석하여 0.8 이상의 매우 높은 상관성을 보이는 인자를 파악하였다. Fig. 3의 (A) 그룹은 지형 인자 중 곡률 인자(표준 곡률, 횡단 곡률, 종단 곡률), (B) 그룹은 토양 인자(유효 토심, 표토 토성, 심토 토성, 배수 등급), (C) 그룹은 임상 인자(임상 종류, 영급, 경급, 임분고, 임상 밀도)에 해당한다. 특히 토양 인자와 임상 인자 그룹은 각 그룹 인자 사이의 상관계수 값이 1로 계산되었는데, 이는 토양 인자의 경우 수치토양도에서 주제도를 구축하는 과정에서 토양통을 기준으로 나머지 인자 특성이 나열되기 때문에 유사한 상관성을 보이는 것으로 판단된다. 임상 인자의 경우도 수치임상도에서 임상 종류와 나머지 특성 자료의 관련성이 매우 밀접하기 때문에 높은 상관계수를 보인다. 따라서 Pearson 상관계수의 heat map을 통해 확인된 (A), (B), (C) 각 그룹의 인자는 매우 높은 수준의 상관관계에 해당하므로 동일한 인자를 반복해서 사용하는 것과 같다. 따라서 (A), (B), (C) 그룹에서 상관성이 높은 인자를 제거하기 위해 각 그룹의 인자 중요도를 분석하였다. 먼저 총 23개 유발 인자를 대상으로 인자 중요도 분석을 수행하였다. Fig. 4의 인자 중요도 결과를 기반으로 각 그룹의 유발 인자에 해당하는 인자 중요도 값을 확인하였다(Table 1). 인자 중요도가 높은 순으로 나열한 결과를 살펴보면 지형 인자 중에서 종단 곡률이 0.028, 토양 인자 중 표토 토성이 0.007, 임상 인자 중 영급 0.166, 임상 밀도 0.137로 각 그룹에서 가장 높은 인자 중요도 값을 보이는 인자로 확인되었다. 따라서 각 그룹 인자 중에서 인자 중요도가 낮은 표준 곡률, 횡단 곡률, 유효 토심, 심토 토성, 배수 등급, 임상 종류, 경급, 임분고 인자는 분석에서 제외하였다.

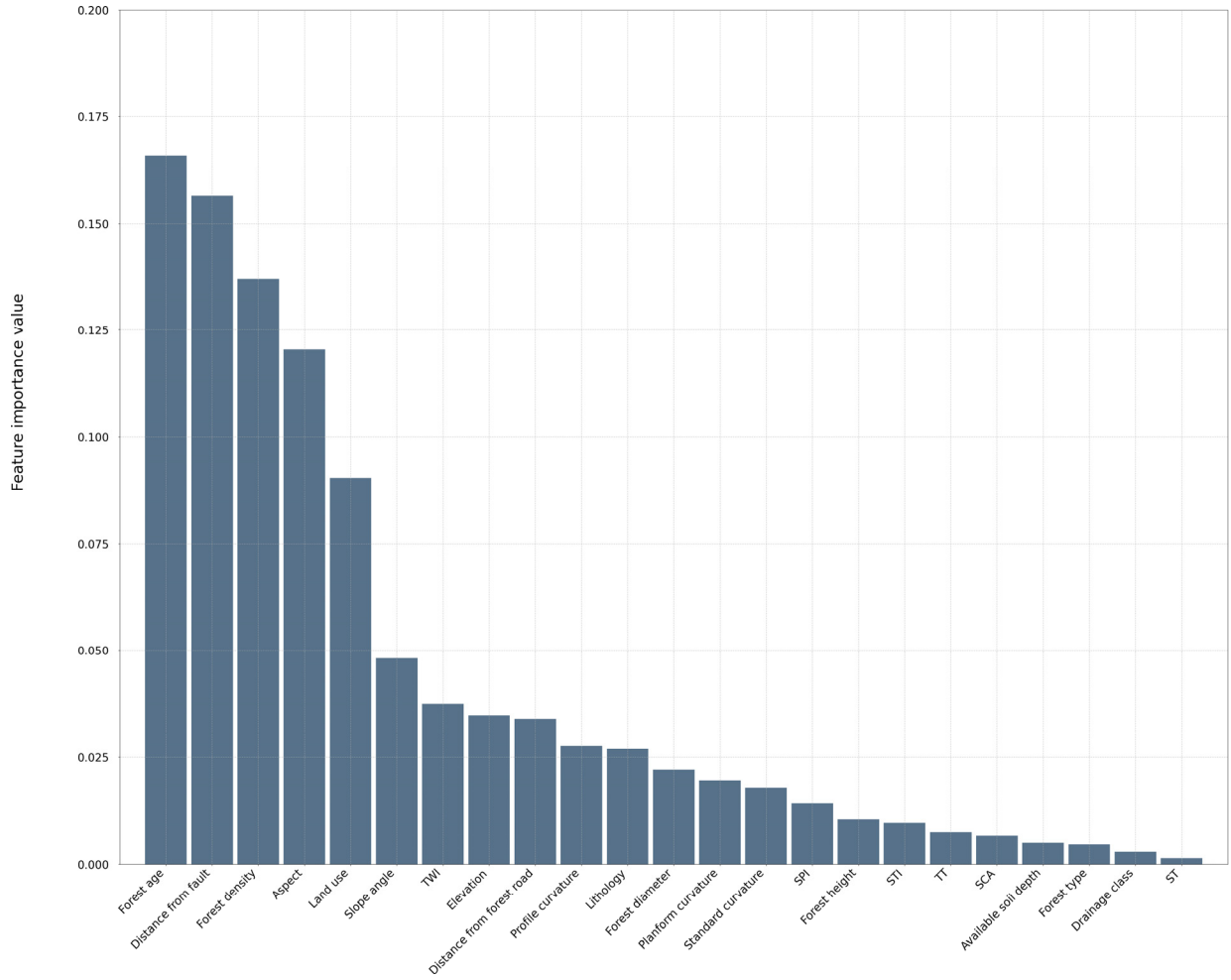

Fig. 4. Random forest feature importance for 23 factors.

Table 1. Rankings of groups of factors by importance (A : curvature, B : soil, C : forest)

Ranking	Curvature (A group)	Soil (B group)	Forest (C group)
1	Profile (0.027643)	Texture (0.007335)	Age (0.165890)
2	Planform (0.019568)	Thickness (0.004899)	Density (0.136919)
3	Standard (0.017866)	Drainage (0.002786)	Diameter (0.022035)
4	-	Sub-texture (0.001420)	Height (0.010336)
5	-	-	Type (0.004495)

또한 본 연구에서 사용한 인자 중 단층 그리고 임도와 관련된 인자의 경우 산사태 발생에 영향력이 높은 인자이지만 그 분포가 매우 한정된 지역에만 밀집되어 있어 연구 지역 전반에 걸쳐 분포한 산사태 발생에 대해 분석하는 것은 어렵다고 판단하여, 인자 선정에서 단층과 임도 인자는 제외하였다.

인자 중요도 분석

앞서 상관관계 분석을 통해 연구 지역의 총 23개 인자 중에서 10개의 인자를 제외한 13개의 인자로 인자 중요도 분석을 실시하였다. 인자 중요도는 산사태 유발 인자가 산사태 발생 위치를 예측하는데 어느 정도 영향을 주었는지 판단하는 지표이다. 본 연구에서는 상관관계 분석을 통해 선정된 13개 유발 인자를 random forest 모델에 적용하여 인자 중요도를 계산하였고, Fig. 5와 같이 인자 중요도를 확인하였다.

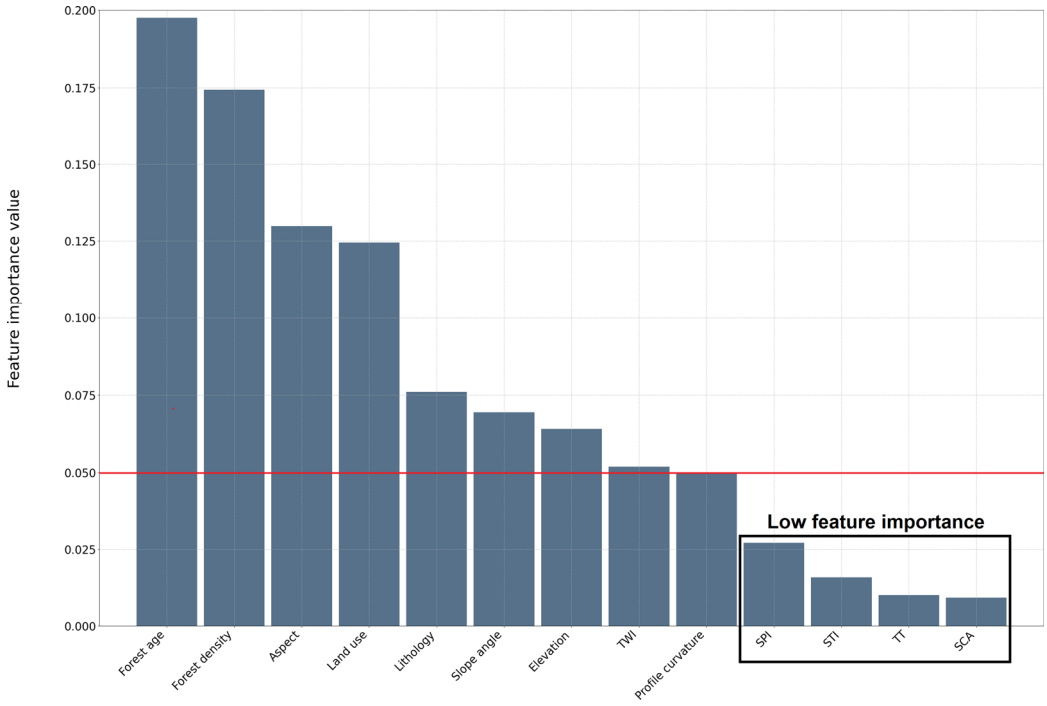

Fig. 5. Random forest feature importance for 13 factors.

그 결과, 임상 인자인 영급과 임상 밀도가 각각 0.197과 0.174로 가장 높은 인자 중요도 값을 나타냈다. 이는 연구 지역의 산사태가 벌채의 영향으로 인해 발생하였으며, 연구 지역의 벌채지(임상도 상의 미립목지)를 본 연구에서 사용된 4차

임상도에 수치임상도 상에 반영하여 임상 인자를 추출하였기 때문에 산사태 발생 예측에 큰 영향을 미친 것으로 판단된다. 또한 인자 중요도 값이 0.05 미만인 유발 인자는 지형 인자인 SPI(0.027), STI(0.016), SCA(0.010), 토양 인자인 표토토성(0.010)이 매우 낮은 값을 보였다. 따라서 앞서 상관관계 분석을 통해 선정된 13개 인자 중 낮은 인자 중요도 값을 보이는 4개 인자를 제외하고, Table 2와 같이 총 9개 인자를 최종 선정하였다.

Table 2. Ranking of features by importance

Ranking	Feature name	Importance value
1	Forest age	0.20
2	Forest density	0.17
3	Aspect	0.13
4	Land use	0.13
5	Lithology	0.08
6	Slope angle	0.07
7	Elevation	0.06
8	TWI	0.05
9	Profile curvature	0.05

딥러닝 기반 산사태 취약성 분석 및 결과

본 연구에서는 두 차례의 데이터 탐색 분석 과정을 통해 유발 인자의 선별 과정이 산사태 취약성 분석 결과에 미치는 영향에 대해 분석하였다. 비교를 위해 전체 23개의 유발 인자와 두 가지 데이터 탐색을 순서대로 적용하여 상관관계 분석을 통해 선별된 13개 인자, 그리고 인자 중요도 값이 낮은 인자를 제외한 9개 인자를 활용하여 취약성 분석에 적용하였다. 딥러닝 기반 산사태 취약성 분석에 3가지 인자 조합을 적용하여 0에서 1사이의 취약 지수(Susceptibility Index, SI) 값을 획득하였으며, 그 결과를 0.2 간격으로 나누어 Very high(0.8~1.0), High(0.6~0.8), Moderate(0.4~0.6), Low(0.2~0.4), Very low(0~0.2)로 구분된 취약성 지도로 구축하였다. 취약성 분석 결과의 모델 성능을 비교하기 위해서 confusion matrix 기반의 정확도 검증을 수행하였으며, ROC(Receiver Operating Characteristics) 그래프 상의 AUC(Area Under Curve) 값을 획득하여 데이터 탐색 분석을 통한 모델의 예측 성능을 확인하였다.

먼저 연구 지역의 전체 유발 인자 23개를 사용한 딥러닝 기반의 취약성 분석을 수행한 결과, AUC는 55.90%로 매우 낮은 성능을 보였다(Figs. 6a and 7). 이는 Table 3에서 보이는 바와 같이 Very high와 Very low 등급이 각각 50.08%와 48.84%의 면적 비율을 차지하였는데, 연구 지역의 산사태 유발과 상관성이 낮은 인자들이 적용되면서 예측 성능이 현저히 낮아진 것으로 판단된다.

첫 번째 상관관계 분석으로 선정된 13개 인자를 취약성 분석에 적용한 결과는 81.25%의 AUC를 획득하였으며, 앞서 전체 유발 인자를 사용한 결과와 비교하여 모델 성능이 향상된 것으로 분석되었다(Figs. 6b and 7). 특히 취약 지수가 Very high인 경우를 살펴보면 전체 유발 인자를 적용한 분석에서는 50.08%로 나타났지만, 1차 선별된 유발 인자를 사용한 분석에서는 8.01%로 위험 지역 면적 비율이 매우 낮아진 것을 확인할 수 있었다. 이러한 결과는 유발 인자를 선정하는 과정이 취약성 분석 과정에서 모델 성능 향상에 확실한 영향을 미쳤다고 판단된다.

마지막으로 인자 중요도 분석까지 모든 데이터 탐색 분석을 수행한 결과는 AUC가 92.80%이며 본 연구에서 진행한 취약성 분석 모델 중 가장 높은 정확도를 획득하였다(Figs. 6c and 7). 또한, 취약 지수가 Very high인 경우 앞선 분석에서

8.01% 비율이 2.74%로 낮아지면서 모델의 성능이 향상될 뿐만 아니라 안전한 지역을 위험하다고 판단하는 비율이 줄어들면서 예측 모델의 신뢰성 향상에도 기여한 것으로 보인다.

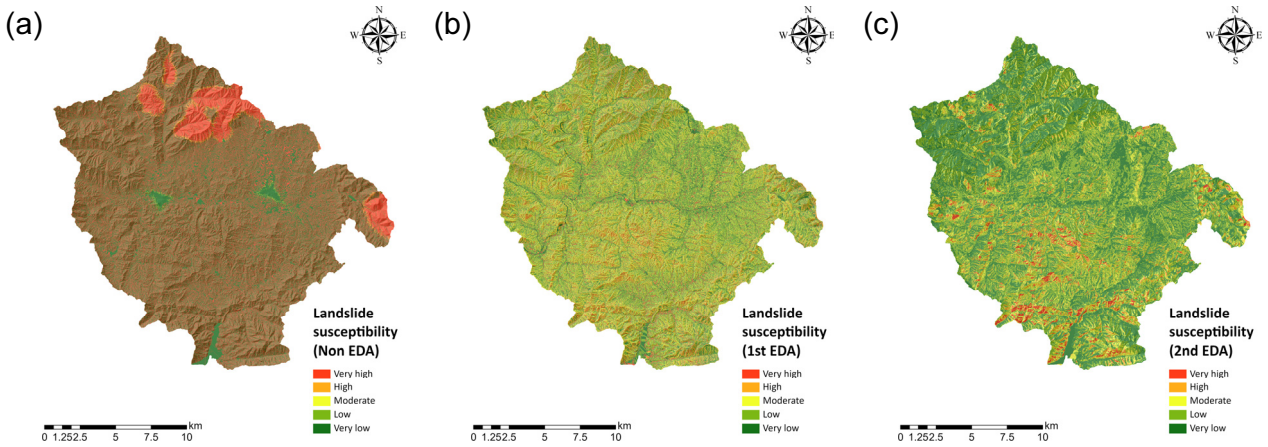

Fig. 6. Landslide susceptibility maps: (a) non EDA, (b) first EDA, (c) second EDA.

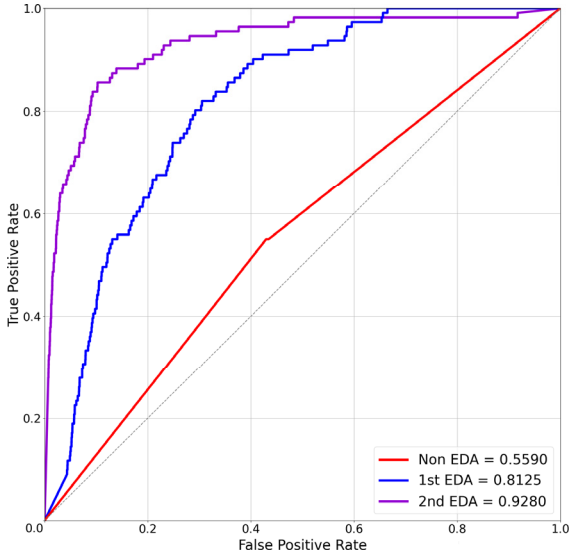

Fig. 7. Receiver operating characteristics graph.

Table 3. Comparison of model performance

	Non EDA		1st EDA		2nd EDA	
	Number of cells	Percentage	Number of cells	Percentage	Number of cells	Percentage
Very low	1,504,394	48.84	784,711	25.47	1,130,399	36.70
Low	11,083	0.36	802,090	26.04	1,205,605	39.14
Moderate	11,071	0.36	931,940	30.25	488,531	15.86
High	11,131	0.36	314,990	10.23	171,398	5.56
Very high	1,542,727	50.08	246,675	8.01	84,473	2.74
AUC	55.90%		81.25%		92.80%	

결 론

본 연구에서는 데이터 탐색 과정이 산사태 취약성 분석의 성능에 미치는 영향을 알아보고자 Pearson 상관계수와 random forest 인자 중요도를 이용하여 2단계에 걸친 데이터 탐색을 진행하였다. 첫 번째 탐색에서 Pearson 상관계수 분석을 진행한 결과, 지형 인자 중 곡률 인자(표준 곡률, 횡단 곡률, 종단 곡률), 토양 인자(유효 토심, 표토 토성, 심토 토성, 배수 등급), 임상 인자(임상 종류, 영급, 경급, 임분고, 임상 밀도)가 높은 상관성을 보였다. 서로 높은 상관성을 보인 인자들끼리 random forest의 인자 중요도를 비교해 표준 곡률, 횡단 곡률, 유효 토심, 심토 토성, 배수 등급, 임상 종류, 경급, 임분고 인자를 우선적으로 제외하였다. 추가적으로 단층 및 임도와 관련된 인자는 연구 지역의 매우 한정된 지역에만 존재하여 산사태 요인으로 고려하기 어렵다고 판단하여 첫 번째 데이터 탐색에서 단층과 임도를 포함하여 총 10개의 인자를 제외하였다. 두 번째 데이터 탐색은 남은 13개의 인자로 random forest 인자 중요도 분석을 실시하였고 인자 중요도 값이 0.05 미만인 SPI, STI, SCA, 표토 토성 인자를 제외하였다. 데이터 탐색이 산사태 취약성 분석 성능에 얼마나 영향을 미치는지 판단하기 위해 각 단계별로 선정된 인자를 이용해 학습한 모델의 검증 결과와 전체 유발 인자를 모두 사용하여 학습한 모델의 검증 결과를 비교하였다. 검증 결과, 전체 유발 인자를 모두 사용한 모델의 AUC는 55.90%, 첫 번째 상관관계 분석으로 선정된 인자를 분석에 적용한 모델은 81.25%의 AUC, 인자 중요도까지 모두 활용하여 선정한 인자를 분석에 활용한 모델은 92.80%의 AUC를 보였다. 또한 계산된 취약 지수가 Very high인 부분의 비율을 비교해 봤을 때 모든 인자를 사용했을 때 비율은 50.08%, 첫 번째 탐색 후 분석 결과는 8.01%, 최종 선정 인자를 활용한 분석은 2.74%의 비율을 보였다.

본 연구에서는 제천 지역에 데이터 탐색 기법을 적용하여 산사태 발생에 영향을 미친 중요도가 높은 인자들로 구분하였으며, 이를 딥러닝 기법에 적용하여 산사태 취약성 분석 및 검증을 수행하였다. 데이터 탐색 과정에서 제천 지역 특성이 반영되어 산사태 발생지 주요 특성 중 하나인 벌채지와 관련된 임상 인자들이 높은 중요도를 보여 본 연구에서 학습된 모델을 타지역에 적용한다면 좋은 성능을 기대하기 어렵다. 하지만 본 연구에서 제안된 분석 기법인 유발 인자 상관성과 중요도를 확인하는 데이터 탐색 기법을 타지역에 적용하면 취약성 분석의 정확도를 높일 수 있는 매우 유용한 방법이라고 판단된다. 따라서 딥러닝을 활용한 산사태 취약성 분석에서 데이터 탐색 과정을 통해 연구 지역 데이터의 특성을 고려한 분석을 수행한다면 모델 성능과 취약성도의 신뢰성 향상에 기여할 수 있다.

사 사

본 연구는 산림청(한국임업진흥원) 산림과학기술 연구개발사업(FTIS 2022424A00-2324-0801)의 지원을 받아 수행된 연구입니다.

References

Biau, G., Scornet, E., 2016, A random forest guided tour, Test, 25, 197-227.

Breiman, L., 2001, Random forests, Machine Learning, 45, 5-32.

Brownley, C.W., 2016, Foundations for analytics with Python: From non-programmer to hacker, O'Reilly Media, 352p.

Bui, D.T., Lofman, O., Revhaug, I., Dick, O., 2011, Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression, Natural hazards, 59, 1413-1444.

Géron, A., 2022, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, Inc., 864p.

Gomez, H., Kavzoglu, T., 2005, Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela, Engineering Geology, 78(1-2), 11-27.

Hauke, J., Kossowski, T., 2011, Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data, Quaestiones Geographicae, 30(2), 87-93.

Huh, M.H., Jung, J.J., 1990, Software review of statistical package programs on EDA aspects, The Korean Journal of Applied Statistics, 3(2), 17-25 (in Korean with English abstract).

Kääb, A., 2002, Monitoring high-mountain terrain deformation from repeated air- and spaceborne optical data: Examples using digital aerial imagery and ASTER data, ISPRS Journal of Photogrammetry and Remote Sensing, 57(1-2), 39-52.

Kalantar, B., Pradhan, B., Naghibi, S.A., Motevalli, A., Mansor, S., 2018, Assessment of the effects of training data selection on the landslide susceptibility mapping: A comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN), Geomatics, Natural Hazards and Risk, 9(1), 49-69.

Kang, K.H., Park, H.J., 2019, Study on the effect of training data sampling strategy on the accuracy of the landslide susceptibility analysis using random forest method, Economic and Environmental Geology, 52(2), 199-212 (in Korean with English abstract).

Lee, J.H., Kim, H., Park, H.J., Heo, J.H., 2021, Temporal prediction modeling for rainfall-induced shallow landslide hazards using extreme value distribution, Landslides, 18, 321-338.

Lee, J.N., Kim, T.S., 2006, Statistics (with R), Freeacademy Inc., 456p.

Lee, S., Oh, H.J., 2019, Landslide susceptibility prediction using evidential belief function, weight of evidence and artificial neural network models, Korean Journal of Remote Sensing, 35(2), 299-316 (in Korean with English abstract).

Liaw, A., Wiener, M., 2002, Classification and regression by randomForest, R News, 2, 18-22.

Martha, T.R., N., Kerle, V., Jetten, C.J., Van Westen, K., Vinod Kumar, 2010, Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods, Geomorphology, 116 (1-2), 24-36.

Muller, A.C., Guido, S., 2016, Introduction to machine learning with Python: A guide for data scientists, O'Reilly Media, Inc., 386p.

Park, H.J., Lee, J.H., 2022, A review of quantitative landslide susceptibility analysis methods using physically based modelling, The Journal of Engineering Geology, 32(1), 27-40 (in Korean with English abstract).

Tien Bui, D., Tuan, T.A., Klempe, H., Pradhan, B., Revhaug, I., 2016, Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree, Landslides, 13, 361-378.

Xiao, C., Ye, J., Esteves, R.M., Rong, C., 2016, Using Spearman's correlation coefficients for exploratory data analysis on big dataset, Concurrency and Computation: Practice and Experience, 28(14), 3866-3878.