한국지반공학회논문집 제40권 5호 2024년 10월 pp. 93 ~ 101
JOURNAL OF THE KOREAN GEOTECHNICAL SOCIETY
Vol.40, No.5, October 2024 pp. 93 ~ 101

ISSN 1229-2427 (Print)
ISSN 2288-646X (Online)
https://doi.org/10.7843/kgs.2024.40.5.93

# Oversampling 기법 및 이상치 제거 방법을 통한 데이터 구축 연구

# Data Construction through Oversampling Techniques and Outlier Removal Methods

장 병 수[1] Jang, Byeong-Su 고 규 현[2] Go, Gyu-Hyun
김 영 석[3] Kim, YoungSeok 김 세 원[4] Kim, Sewon
최 현 준[5] Choi, Hyun-Jun 윤 형 구[6] Yoon, Hyung-Koo

## Abstract

Numerical analysis methods are widely used to assess the safety of hydrogen storage facilities; however, obtaining data under various conditions poses significant challenges. This study aims to expand the dataset using oversampling algorithms and utilize these enhanced datasets as diverse input parameters for numerical analysis. The oversampling techniques applied include SMOTE, Borderline-SMOTE, ADASYN, and CTGAN, with data amplified by factors of 2, 5, and 100 relative to the original dataset. This approach increases data volume based on the characteristics of the existing data, which may consequently introduce outliers. To address this, statistical methods such as the 3-sigma rule and the confidence level method are employed to identify and remove outliers beyond the normal distribution range. The reliability of the conditions generated through data amplification and outlier analysis is evaluated by comparing them with trends observed in the original dataset. Additionally, the SHAP algorithm is utilized to analyze changes in the importance values of each parameter. The SHAP values derived from the original dataset and those processed through AI techniques and outlier analysis exhibit similar trends, validating the proposed methodologies. The methods proposed in this paper are applicable not only to hydrogen storage facilities but also to the systematic construction of data for assessing the stability of various geotechnical structures.

## 요 지

수소 저장 시설의 안정성을 평가하기 위해서 주로 수치해석 방법이 활용되나, 여러 조건의 데이터 확보에는 어려움이 따른다. 해당 연구의 목적은 oversampling 알고리즘을 활용하여 데이터 그룹의 양을 확대하고 수치해석 시 다양한 입력 인자로 이용되도록 하는 것이다. Oversampling 알고리즘은 AI 분야에서 데이터 불균형 문제를 해소하고자 제안된

1 정회원, 대전대학교 재난안전공학과 박사과정 (Member, Ph.D. Student, Dept. of Disaster Safety Engineering, Daejeon Univ.)
2 정회원, 금오공과대학교 토목공학과 부교수 (Member, Associate Prof., Dept. of civil Engineering, Kumoh National Institute of Tech.)
3 정회원, 한국건설기술연구원 북방인프라특화팀 선임연구위원 (Member, Senior Research Fellow, Northern Infrastructure Specialized Team, Korea Institute of Civil Engineering and Building Technology)
4 비회원, 한국건설기술연구원 지반연구본부 연구원 (Researcher, Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology)
5 정회원, 한국건설기술연구원 북방인프라특화팀 수석연구원 (Member, Senior Researcher, Northern Infrastructure Specialized Team, Korea Institute of Civil Engineering and Building Technology)
6 정회원, 대전대학교 재난안전공학과 정교수 (Member, Prof., Dept. of Disaster Safety Engineering, Daejeon Univ., +Tel: +82-42-280-2570, +Fax: +82-42-280-2576, hyungkoo@dju.ac.kr, Corresponding author, 교신저자)

* 본 논문에 대한 토의를 원하는 회원은 2025년 4월 30일까지 그 내용을 학회로 보내주시기 바랍니다. 저자의 검토 내용과 함께 논문집에 게재하여 드립니다.

SMOTE, Borderline-SMOTE, ADASYN 그리고 CTGAN 기법을 적용하였으며, 기존 데이터 대비 2배, 5배 그리고 100배로 증폭하였다. 해당 방법은 기존 데이터 특성을 기반으로 양을 증폭하는 방식으로 최종 데이터 그룹은 이상치가 포함될 가능성이 있다. 이를 해소하고자 통계기법인 3 sigma rule과 confidence level 방법으로 데이터의 정규분포 특성의 일정한 범위 외에 있는 값들은 이상치로 판단하여 제거하였다. 데이터 증폭과 이상치 분석을 통해 구축된 다양한 조건의 값의 신뢰성은 기존 데이터의 경향과 비교하여 판단하고자 하였으며, SHAP 알고리즘을 통해 각 물성치들의 중요도 값의 변화를 살펴보았다. 기존 데이터와 AI 기법 및 이상치 분석을 수행한 데이터의 SHAP 값은 모두 유사하게 나타나 해당 논문에서 제안한 방법이 타당함을 입증하였다. 해당 논문에서 제안한 방법은 수소 저장 시설뿐 아니라 다양한 지반 구조물의 안정성 평가 시 합리적인 데이터 구축에 활용 가능할 것으로 판단된다.

**Keywords :** Hydrogen storage facilities, Outlier analysis, Oversampling algorithm, SHAP, Stability

## 1. 서 론

화석 연료 등의 사용량이 증가함에 따라 지구 온난화 등의 문제가 발생되고 있으며, 이를 해결하기 위한 신재생 에너지에 대한 관심이 증가하고 있다. 대표적으로 내연기관으로 인한 온실가스 배출을 감소하기 위해 수소 에너지에 대한 관심이 증가하고 있다(Lee et al., 2021). 이처럼 전세계적으로 수소에 대한 관심이 증가함에 따라 수소 저장시설에 대한 안전성에 대한 관심이 함께 동반되고 있다. 지중저장은 대량의 수소 저장이 가능하다는 장점이 있어 합리적인 저장 방식으로 인식되어(Taylor et al., 1986) 지반공학 측면의 지하 저장시설에 대한 연구가 진행되고 있다(Panfilov, 2016; Ning et al., 2021; Zivar et al., 2021). 국내에서는 Go et al.(2022) 연구자가 국내 지형조건을 고려하여 수소 저장 시설이 얕은 깊이에 매설될 경우 안정성 분석에 대한 연구를 수행한 사례가 있다. 또한 수소저장 시설의 안정성과 관련된 문제를 해결하고자 저장 탱크의 폭발에 관련된 연구도 수행되고 있다. Choi et al.(2022)은 수소저장탱크의 폭발에 따른 인접 구조물의 영향성을 평가 하기위한 수치모델을 제안하였고, Shin(2023)은 수소저장시설의 주요설계변수에 따른 폭발 피해를 평가하고 인접 시설물의 손상도를 분석하여 시설물의 안전한 이격 거리를 제안하였다.

이처럼 안정성을 평가하기 위해서는 모든 지반 환경에서 직접적인 실험을 진행하기에 시간적 경제적으로 어려움이 발생하여 주로 시뮬레이션의 방법론이 활용되고 있다. 하지만 시뮬레이션 방법은 저장시설의 지반공학적 표본 데이터를 설정하고 다양한 위치에서 DB가 구축되어야 한다. 즉 시뮬레이션을 통한 수소저장시설의 안정성 평가 연구에서 신뢰성 있는 결과를 얻기 위해서는 다양한 지반 환경과 설계 변수를 고려한 충분한 데이터가 요구된다. 그러나 실험 데이터는 지리적 및 경제적 등과 같은 문제로 인해 충분한 개수를 확보하기에는 한계가 있다. 이를 해결하고자 해당 연구에서는 최소의 데이터 개수를 통해 수치해석이 가능한 수준의 데이터 개수가 확보될 수 있도록 기계학습 알고리즘 중 oversampling 기법을 적용하고자 하였다.

해당 논문은 oversampling(SMOTE, Borderline-SMOTE, ADASYN, CTGAN) 기법의 배경이론을 소개하였고, 원 데이터와 증폭된 데이터의 특성차이를 분석하였다. 증폭된 데이터의 신뢰성은 증폭 방법에 따른 출력 값의 범위, 이상치 제거에 따른 표준편차 그리고 box plot을 통해 평가 후 결과를 도시하였다. 최종적으로, SHAP 값을 사용하여 각 인자들의 중요도를 분석하고, 증폭된 데이터 마다 중요인자의 차별화가 발생하는지 관찰 후 oversampling 기법의 신뢰성을 정리하였다.

## 2. 배경이론

### 2.1 Oversampling 알고리즘

분류 문제는 입력 데이터 클래스의 분포가 균일할 때 신뢰성 있는 결과가 동반되나, 분포가 비대칭일 경우 편향된 결과를 제공하는 한계를 보인다. 이를 해결하기 위해 oversampling 알고리즘이 제안되었으며, 기계학습의 신뢰성 향상을 위해 다양한 연구에 활용되고 있다(Cordón et al., 2018; Kim et al., 2024). Synthetic Minority Oversampling Technique(SMOTE)는 K-nearest neighbors(K-NN) 알고리즘의 거리계수를 활용하여 데이터의 불균형 문제를 해결하는 방식으로 Fig. 1(a)의 SMOTE와 같이 소수

데이터의 영향 범위를 결정하여 새로운 데이터를 생성한다(Chawla et al., 2002). K-NN 알고리즘에서 빨간색 원은 소수 데이터 포인트와 근접한 데이터 포인트의 범위를 의미하며, SMOTE 알고리즘은 근접한 데이터 간의 거리 관계를 활용하여 새로운 데이터를 생성한다. 원의 크기는 K-NN 알고리즘에서 설정한 k 값에 따라 달라지며, k 값이 클수록 더 넓은 범위의 데이터가 포함된다. 새로운 데이터는 기본 데이터에 가중치를 부여하는 형태로 생성되며, 해당 가중치는 0~1 사이의 범위에서 무작위로 부여된 거리 계수를 기반으로 한다. Borderline-SMOTE는 SMOTE와 유사하지만 다수의 데이터와 소수 데이터의 경계에 위치한 데이터에 초점을 맞춰 oversampling이 진행된다(Han et al., 2005). Fig. 1(a)의 Borderline-SMOTE과 같이 다수 및 소수 데이터의 경계면을 식별하고 새로운 데이터가 생성된다. 경계면에서 생성된 데이터는 분류 경계에서의 정확도를 높이고, 소수 데이터 생성시 오차가 감소하는 특징을 보인다. Adaptive Synthetic Sampling Approach for Imbalanced Learning(ADASYN)는 새로운 데이터 생성시 소수 데이터의 중요도를 고려하여 가중치가 부가되는 방식이 이용된다(He et al., 2008). Fig. 1(a)의 ADASYN와 같이 K-NN 알고리즘을 통해 영향 범위가 설정되어 소수 데이터와 근접된 위치를 찾고, 다수 데이터의 밀집 정도를 통해 새로운 데이터 생성 방향이 결정된다. Conditional Tabular GAN(CTGAN)은 Fig. 1(b)와 같이 Generative Adversarial Networks(GAN) 알고리즘을 기반으로 생성자와 판별자 두 부분으로 구성되며 판별자가 생성한 데이터를 지속적으로 검토해서 원래 데이터와 유사한 특성을 보이는 데이터가 산출될 수 있도록 한다. 해당 알고리즘은 샘플링 된 데이터에서 임의의 노이즈를 추가하는 방식이 사용되며 train과 test 과정을 통해 데이터의 신뢰성이 향상된다.

### 2.2 이상치(outlier) 제거

이상치 제거는 oversampling을 적용한 후 데이터의 품질을 높이고 모델의 성능을 향상시키기 위해 필수적인 과정이다. 이상치(outlier)는 데이터셋에서 다른 데이터와 현저히 다른 값을 가지는 데이터 포인트로, 모델 학습 시 노이즈로 작용할 수 있다. 이러한 이상치는 모델의 일반화 능력을 저하시켜 예측 성능을 떨어뜨리고 과적합(overfitting)을 유발할 수 있다. 이상치 제거는 주로 통계적 방법을 통해 이루어지며, 데이터를 정규분포로 가정

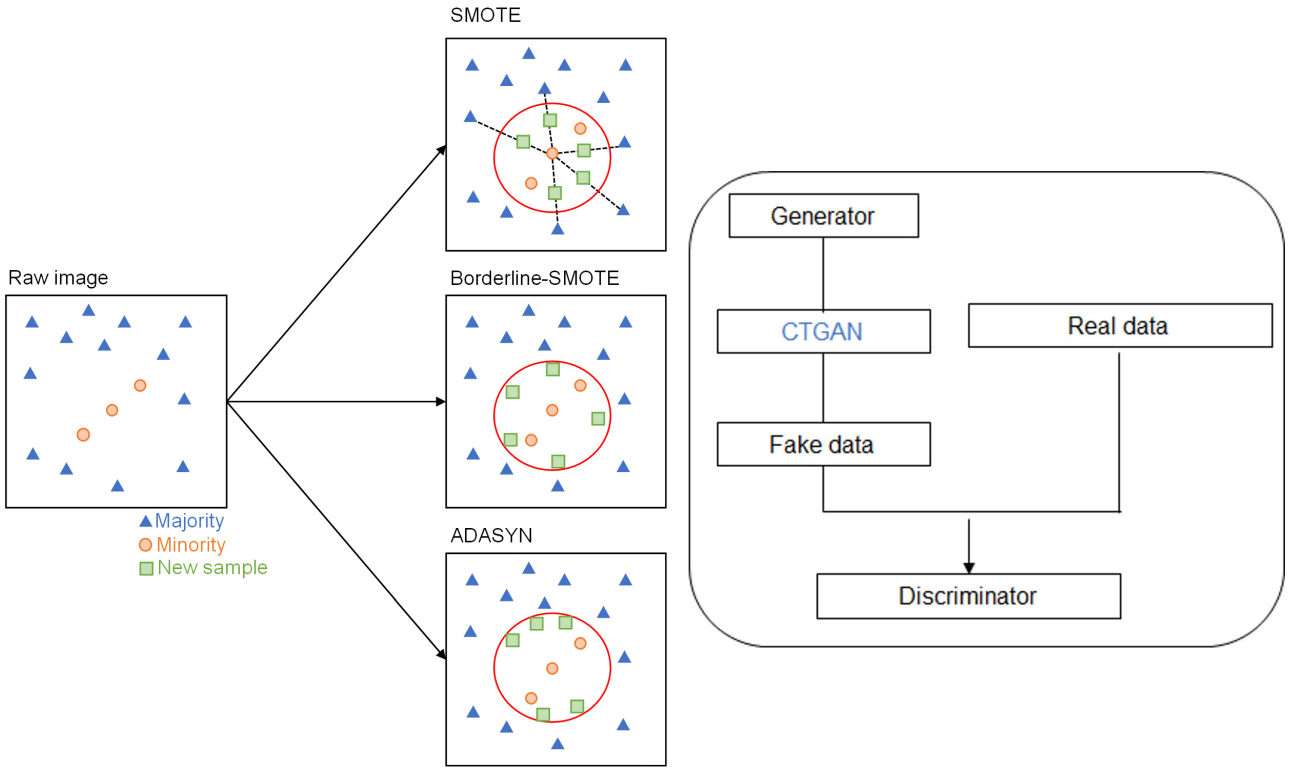



Fig. 1. Applied oversampling techniques: (a) Over-sampling using SMOTE, Borderline-SMOTE, and ADASYN methods; (b) CTGAN method. The red circle is determined by the K-nearest neighbors (K-NN) algorithm, and its size is determined by the value of k

할 때, 평균으로부터 ±표준편차의 x 시그마 범위를 벗어난 데이터 포인트를 이상치(outlier)로 가정한다. 3 sigma rule은 평균 값에서 ± 표준편차(σ)를 벗어나는 데이터를 이상치로 간주하며, ± 1σ은 68%의 데이터를 포함하며 32%를 이상치로 간주한다. ± 2σ 및 ± 3σ는 각각 95%, 99%를 포함하고 있으며, 5% 및 1%를 이상치로 간주한다. Confidence level은 사용자의 선택에 따라 포함된 데이터의 비율을 결정한다. 해당 논문에서 3 sigma rule는 ± 1~3σ, confidence level의 비율은 90%, 95% 그리고 99%를 적용하였다.

## 3. Oversampling을 통한 DB 구축

수치해석시 매개 변수의 선택은 지반에 저장된 수소 시설물의 안정성을 평가하기 위해 매우 중요하다. 수소 저장 시설물이 지중에서 폭발하면 에너지가 발생하며, 이는 주기, 최대진폭 그리고 속고 값을 갖는 파형으로 간주할 수 있다. 따라서 출력 값은 최대속도로 설정하였으며, 입력 물성치는 수소 이온 농도, 흙의 종류, 단위 중량, 점착력, 내부 마찰각, 동적 탄성 계수 그리고 동적 포아송 비로 설정하였다. 원 데이터는 10가지 종류의 입력 물성치에 따라 구축하였으며, 최대 속도는 거리에 따른 영향을 고려하여서 11m까지 1m 간격으로 최대 속도를 계산하였다. 구축된 데이터는 다양한 특성을 포함하기 위해 Fig. 2와 같이 원 데이터 개수 기반으로 2배, 5배 그리고 100배 개수를 증폭하였으며, SMOTE, Borderline-SMOTE, ADASYN 그리고 CTGAN 알고리즘을 적용하였다. 증폭된 데이터를 기반으로 원데이터와의 신뢰성을 분석하기 위해 이상치 제거와 이상치 제거에 따른 인자들의 특성 변화를 분석하기 위해 SHAP value를 통한 신뢰성을 확보하고자하였다.

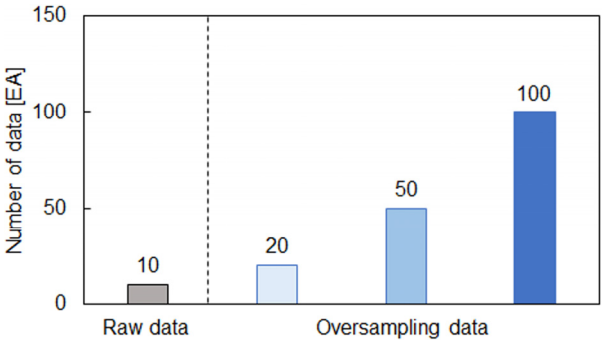



Fig. 2. Number of data points changed by applying the oversampling algorithms

## 4. Oversampling 방법에 따른 변화 분석

### 4.1 Oversampling 방법과 개수에 따른 변화 분석

Oversampling알고리즘으로 확대된 데이터 개수의 분포는 Fig. 3에 box plot을 통해 도시하였으며, Fig. 3(a)과 (b)는 출력 값의 거리와 최대속도를 의미한다. Fig. 3에서 보라색은 원 데이터이며, 노란색, 회색, 파란색 그리고 초록색은 각각 SMOTE, Borderline-SMOTE, ADASYN 그리고 CTGAN을 의미한다. 원 데이터의 최대 값, 중간 값 그리고 최소 값은 각각 1m, 6m 그리고 11m로 나타났으며, SMOTE, Borderline-SMOTE 그리고 ADASYN 기법들은 원 데이터의 최대 값, 중간 값 그리고 최소값의 범위와 유사하게 나타났다. 그러나 CTGAN 기법으로 5배 증폭된 결과에서 최대 값은 약 2.5m 차이가 발생하였고, 중간 값은 약 0.8m 그리고 최소 값은 약 0.4m의 차이가 발생되었다. Fig. 3(b)의 원 데이터 최대 값, 중간 값 그리고 최소 값은 각각 1.13m/s, 0.2m/s 그리고 0.04m/s로 나타났다. CTGAN을 제외한 기법들은 원 데이터와 차이가 0.1~0.2m/s 정도의 오차를 보였지만, CTGAN의 오차는 최대 값 0.18m/s, 중간 값 0.04m/s 그리고 최소 값은 0.13m/s으로 관측되었다.

SMOTE, Borderline-SMOTE, ADASYN 기법으로 생성된 데이터는 원 데이터의 최대 값, 중간 값 그리고 최소 값의 특성을 반영하여 원 데이터 범위내에서 데이터의 개수를 효과적으로 증폭하였다. CTGAN은 새로운 데이터를 생성하는 과정에서 보다 다양한 데이터를 생성하는 경향이 관측되었다. 이는 GAN의 특성을 기반으로 구동되는 CTGAN이 다른 기법들에 비해 데이터의 다양성을 높이는 장점이 있지만, CTGAN 기법 사용시 데이터 분포의 변동성을 충분히 고려해야 함을 의미한다.

### 4.2 이상치 제거 특성 분석

Rekha and Reddy(2018)은 oversampling 과정에서 발생하는 이상치가 모델의 성능에 부정적인 영향을 줄 수 있어, 이상치는 필수적으로 제거되어야 함을 명시하였다. Fig. 4는 Fig. 3의 증폭된 데이터 개수를 기반으로 신뢰성을 높이기 위해 3 sigma rule 및 confidence level로 이상치를 제거한 결과이다. Fig. 4에서 빨간색 점선은 원 데이터이며, 노란색, 주황색, 파란색 그리고 회색은 각각 SMOTE, Borderline-SMOTE, ADASYN 그리고

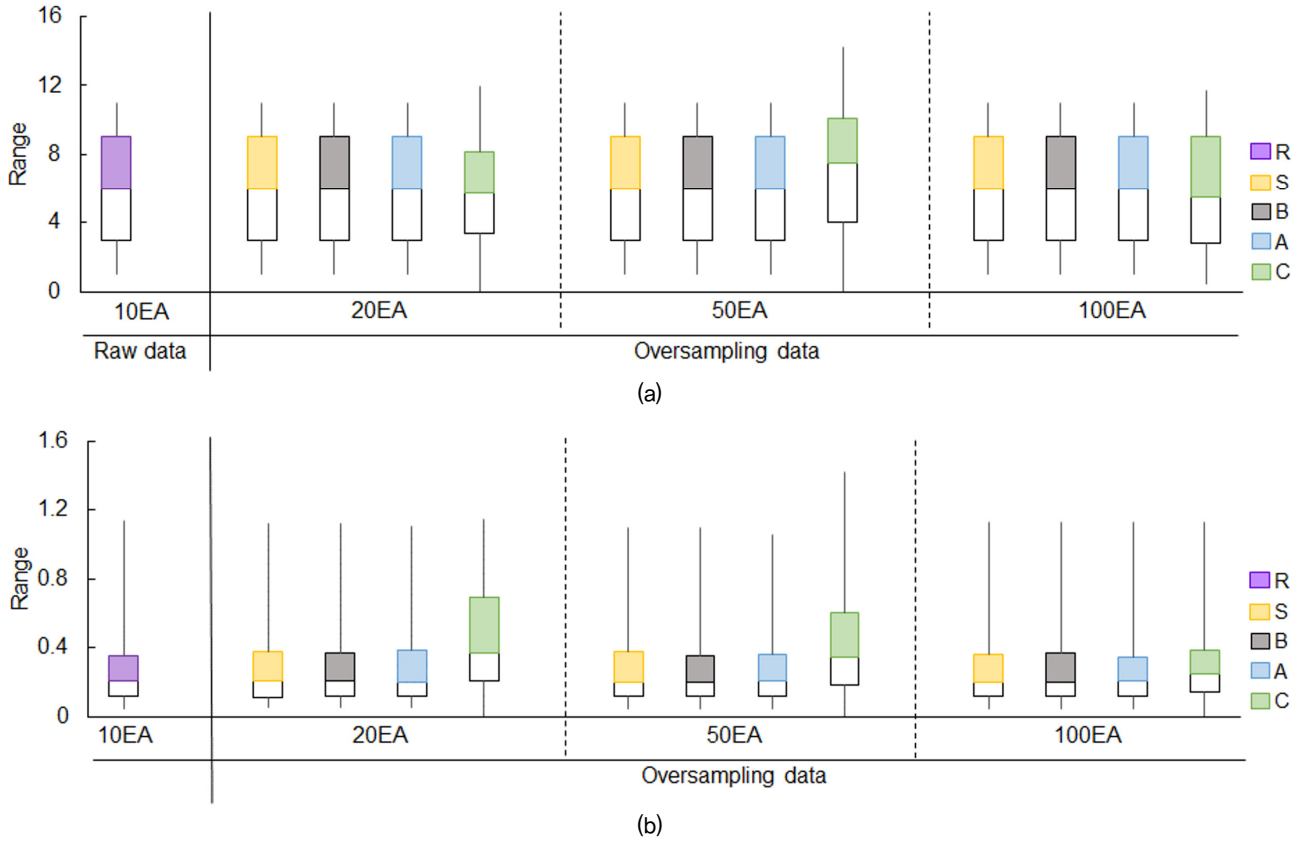


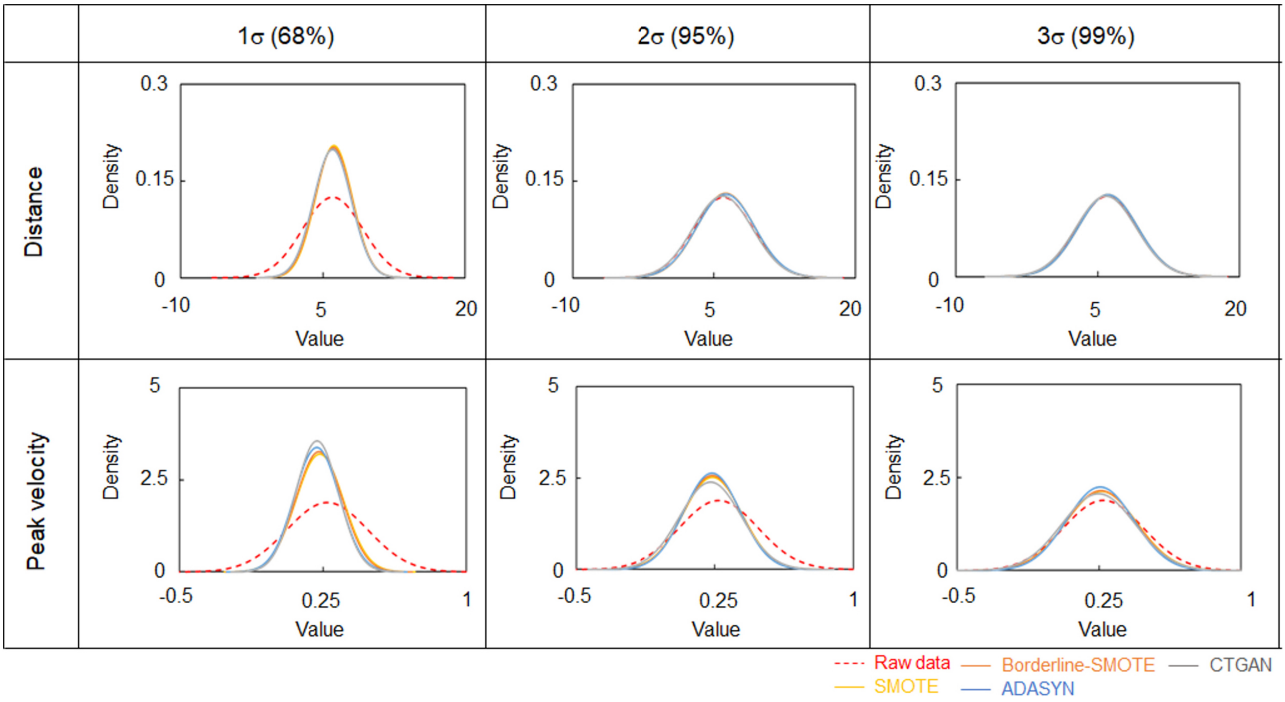

Fig. 3. Comparison of boxplots after applying oversampling algorithm in terms of: (a) distance; (b) peak velocity. R represents Raw data, S stands for SMOTE, B denotes Borderline-SMOTE, A refers to ADASYN, and C indicates CTGAN

CTGAN 알고리즘에 해당하는 결과이다. Fig. 4(a)와 (b)는 각각 3 sigma rule 및 confidence level을 통해 이상치가 제거된 데이터의 정규분포 특성을 보여준다.

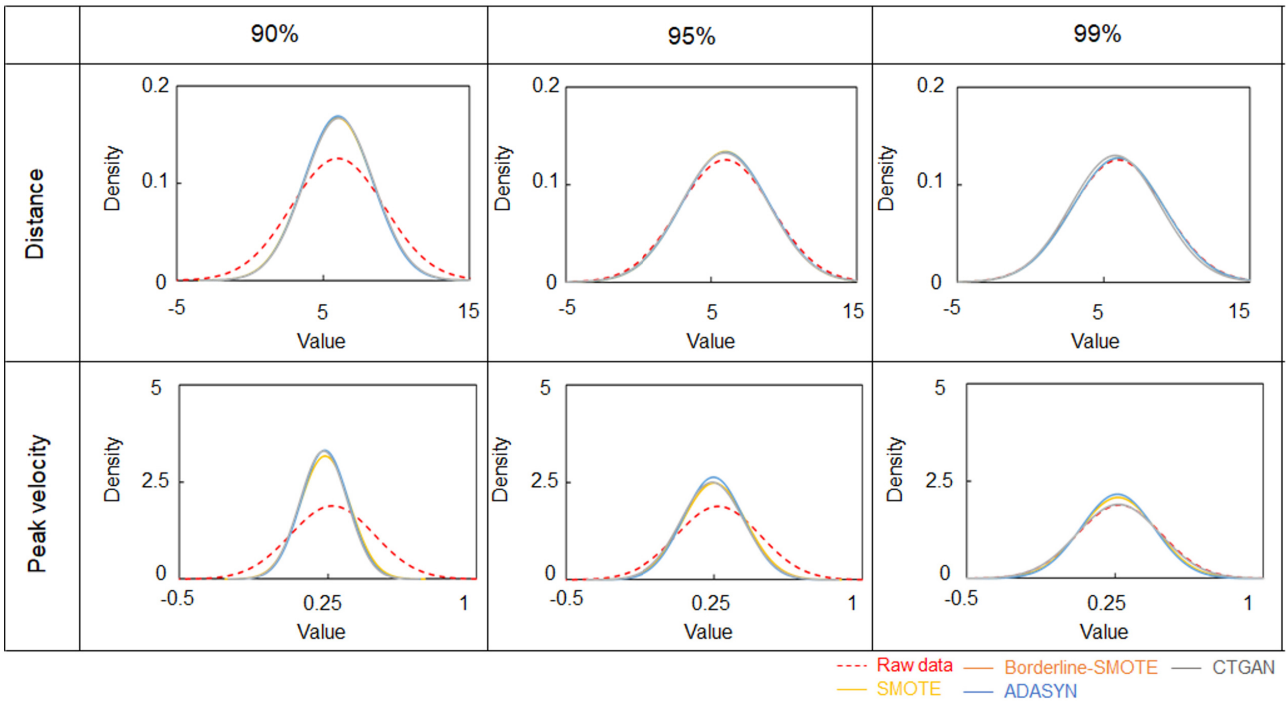
원 데이터를 기준으로 거리 값에 대한 특성은 최대 0.12, 평균 값은 5.9로 나타났으며, 최대 속도는 1.87과 0.26으로 나타났다. 3 sigma rule 방법으로 이상치를 제거한 결과 중 거리 값은 평균 6으로 일정하게 나타났고, 밀도의 최대값은 1σ, 2σ, 3σ에서 모두 0.12로 나타났다. 최대 속도의 평균 값은 1σ 및 2σ에서 0.22, 3σ는 0.25로 다수 큰 값이 나타났고, 최대 값은 1σ, 2σ, 그리고 3σ에서 각각 3.35, 2.52 그리고 2.14로 나타났다. Confidence level 결과인 Fig. 4(b)에서 원 데이터의 밀도와 평균은 각각 0.12와 5.93으로 나타났고, 밀도 값의 변화는 신뢰구간이 90%, 95% 그리고 99%에서 각각 5.98, 5.93 그리고 5.9로 관찰되었다. 최대 속도의 최대 밀도는 0.26, 평균 값은 1.88로 나타났으며, 신뢰구간이 90%, 95% 그리고 99% 일 때 밀도는 각각 0.23, 0.24 그리고 0.26으로 나타났다. 평균은 위와 같이 동일한 신뢰구간에서 각각 3.26, 2.53 그리고 2.06을 보였다. 원 데이터와 이상치를 제거한 데이터를 비교해보면, 3 sigma rule 방법과 confidence level 방법 모두 유사한 정규분포 값을 보인다. 이와 같은 결과는 이상치 제거가 데이터의 전체 거동과 불일치하는 데이터만을 제거하고 전체 특징의 변화는 없는 것을 보여준다.

## 5. 토 의

수소 저장 시설 폭파 시 지반의 안정성을 평가하기 위해 oversampling(SMOTE, Borderline-SMOTE, ADASYN, CTGAN) 기법으로 데이터 개수를 확장하였으며, 이상치 존재 유무도 파악하여 이를 제거하였다. 이상치 제거가 데이터 분포에 어떻게 영향을 미치는지 확인하기 위하여 출력인자 예측 시 입력인자의 중요성을 분석하였으며, 알고리즘은 SHapley Additive exPlanations(SHAP) 기법을 적용하였다. 중요도 결과는 Fig. 5에 각각 두배, 5배 그리고 100배로 증폭된 결과를 도시하였고 도시하였고, Fig. 5(a), (b) 그리고 (c)는 증폭된 데이터에서 이상치 제거 전과 후의 결과이다. Fig. 5에서 A, B, C, D, E, F, G 그리고 H는 각각 수소 이온 농도, 흙의 종류, 단위 중량, 점착력, 내부 마찰각, 동적 탄성 계수, 동적 포




(a)

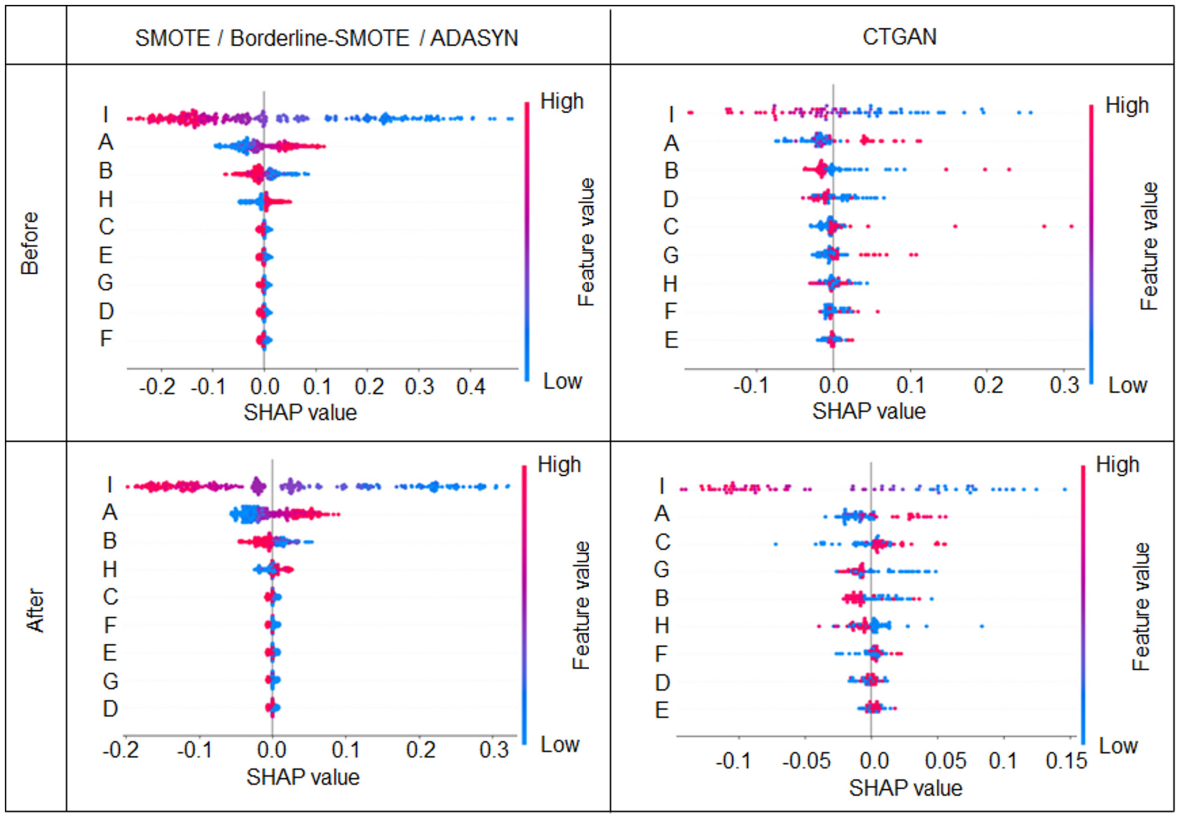



(b)

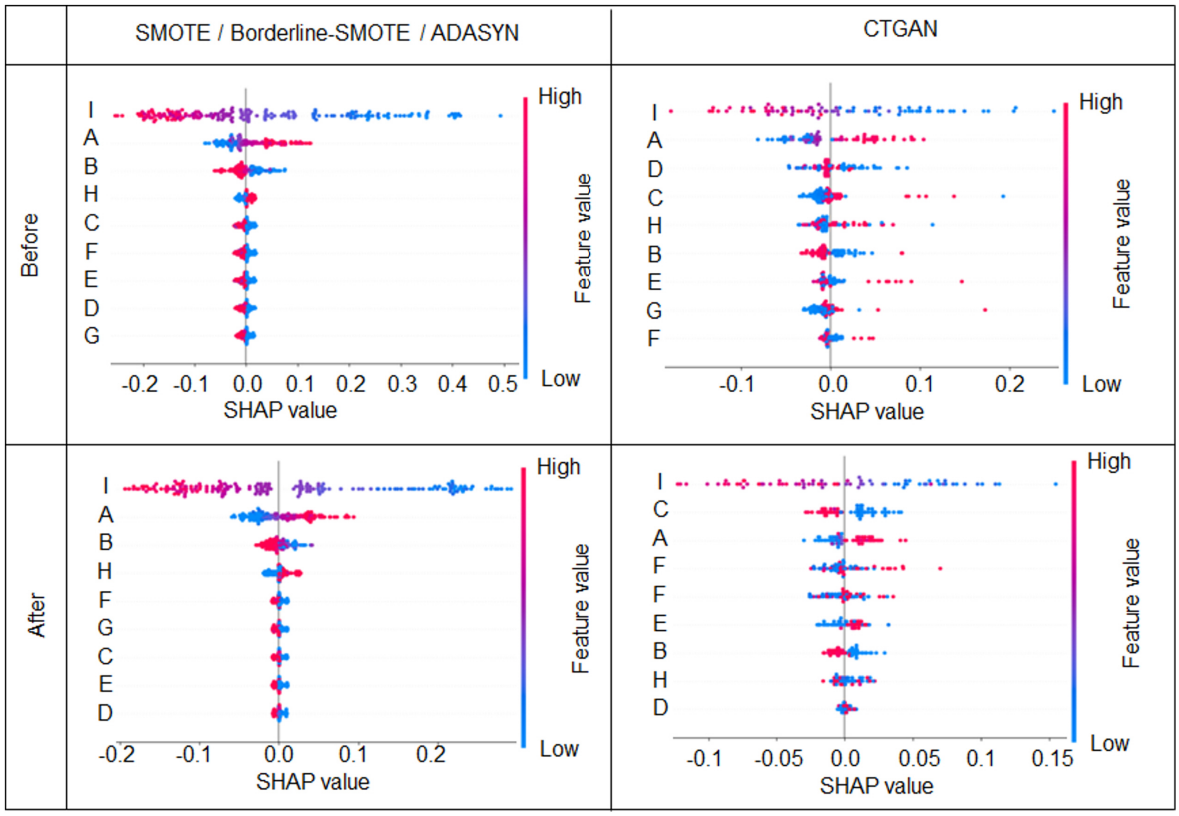
Fig. 4. Distributions of normal distribution curve according to the outlier removal method: (a) 3 sigma rule; (b) confidence level

아송 그리고 거리를 의미하며, I는 최고 속도를 보여준다.

SMOTE, Borderline-SMOTE 그리고 ADASYN 기법으로 2배, 5배 그리고 100배로 증가된 데이터의 중요도 순서는 I, A, B 그리고 H로 모두 동일하게 나타났다. I의 SHAP 수치는 -0.25~0.4 범위를 가지며 (-) 범위에 수치 값이 집중되어 있어 최고 속도와 반비례 관계를 나타내는 것을 의미한다. A와 B는 각각 -0.11~0.1, -0.1~0.1의 범위를 가지며 대부분의 SHAP 값이 음수에 분포하여 I와 동일하게 반비례 상관성을 보인다. H는 -0.5~0.2의 범위를 가지며 (+) 범위에 결과가 집중되어 최고 속도와

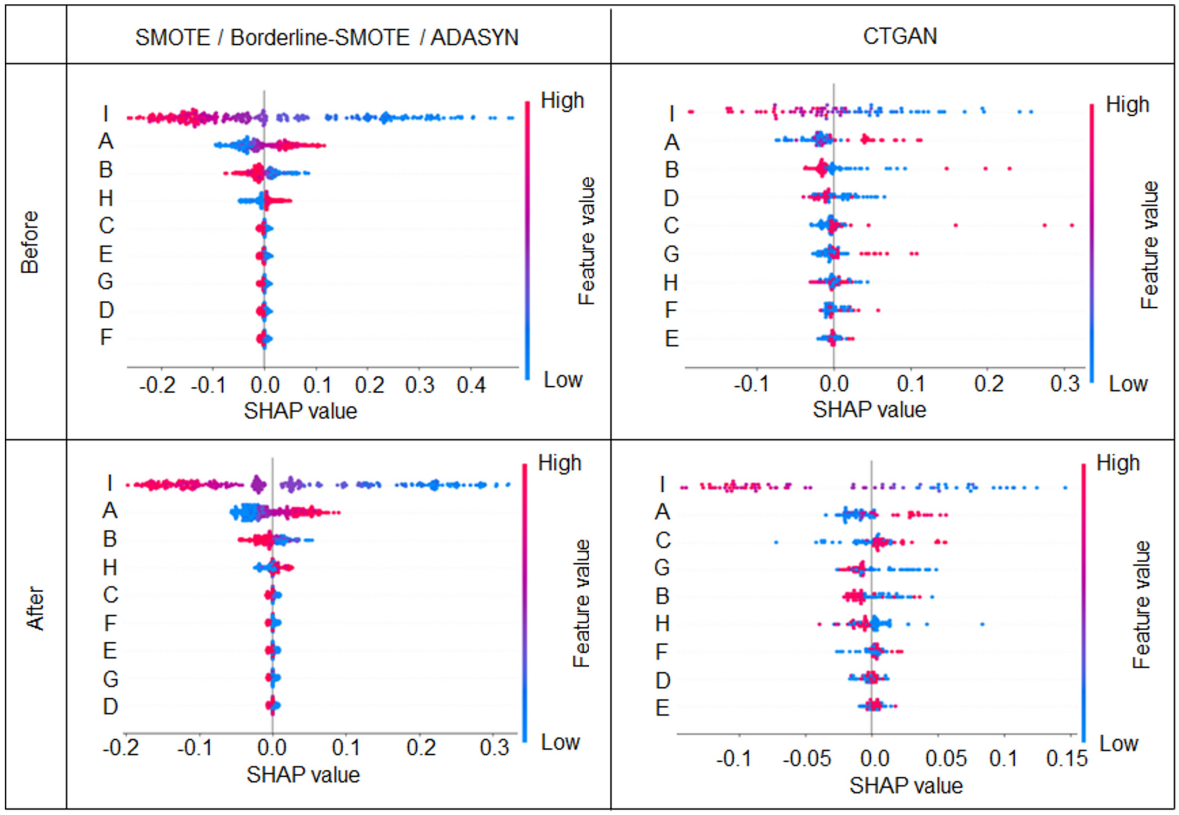



(a)




(b)

Fig. 5. Distributions of SHAP values: (a) Doubled data; (b) Quintupled data; (c) 100-fold amplified data. SHAP values represent the order of influence on the model's output. A, B, C, D, E, F, G, H, and I denote the hydrogen ion concentration, soil type, unit weight, cohesion, internal friction angle, dynamic elastic modulus, dynamic Poisson's ratio, distance, and peak speed, respectively




(c)

Fig. 5. Distributions of SHAP values: (a) Doubled data; (b) Quintupled data; (c) 100-fold amplified data. SHAP values represent the order of influence on the model's output. A, B, C, D, E, F, G, H, and I denote the hydrogen ion concentration, soil type, unit weight, cohesion, internal friction angle, dynamic elastic modulus, dynamic Poisson's ratio, distance, and peak speed, respectively (Continued)

비례 관계임을 확인하였으며, 이후 각 입력인자의 중요도 순서는 oversampling 방법에 따라 상이하게 나타났다. CTGAN 기법으로 증폭된 데이터의 중요도의 순서는 I, A, G 그리고 B로 나타났으며, I와 A는 타 증폭기법과 동일하게 큰 영향을 미치는 것으로 나타났지만, 나머지 인자는 상이한 순서를 보였다. Fig. 5(b)는 이상치 제거 후 결과로서 이상치 제거 전 결과와 동일하게 SMOTE, Borderline-SMOTE 그리고 ADASYN에서는 I, A, B 그리고 H 인자의 순서대로 중요도가 높은 것으로 나타났고, CTGAN는 I의 인자가 가장 중요한 결과로 나타났지만 그 외에 인자들의 중요도 순서는 증폭된 데이터의 개수에 따라 다른 결과가 나타났다.

SHAP 기법을 적용 후 SMOTE, Borderline-SMOTE 그리고 ADASYN 기법으로 증폭된 인자들의 중요도 순서는 모두 동일하게 나타났지만, CTGAN 기법으로 증폭된 데이터는 다소 상이한 결과를 보였다. 이는 증폭된 데이터가 사용된 알고리즘 특성에 따라 상이하게 구축된 것을 의미하며, 특히 CTGAN 알고리즘의 경우 앞서 살펴보았듯이 다양성이 강조되어 중요인자가 다르게 나타난 것으로 사료된다. 물론 이상치 제거 전과 후의 SHAP 결과는 각 알고리즘마다 모두 동일하게 나타나 이상치 제거의 신뢰성은 확보된 것으로 보인다. 해당 연구는 논문에서 사용한 이상치 제거 방법이 타당함을 보여주며, 입력인자의 중요도는 데이터 증폭 시 활용한 알고리즘을 고려해야 신뢰성이 있음을 시사한다. 본 연구에서는 구조물의 안전성을 평가하기 위한 시뮬레이션 실행 시 데이터 부족 문제를 해결하기 위한 방안을 모색하였다. 이러한 시뮬레이션 과정에서 데이터의 불충분함은 평가의 정확도를 저하시킬 수 있으므로, 본 연구에서 제시한 방법은 해당 문제를 효과적으로 완화하는 데 기여할 수 있을 것으로 사료된다.

## 6. 결 론

해당 연구에서는 지중 수소 저장 시설의 안정성 평가 시 활용할 수 있는 데이터 베이스를 구축하기 위해 oversampling(SMOTE, Borderline-SMOTE, ADASYN, CTGAN) 알고리즘을 적용하였다. Oversampling을 통해 증폭된

데이터의 신뢰성을 분석하기 위해 이상치 제거(3 sigma rule, confidence level)을 통해 증폭된 데이터의 변화를 분석하였고, SHAP value 기법을 활용해 인자들의 변화를 분석하였다. 연구에서 도출된 결론은 다음과 같이 요약된다.

(1) SMOTE, Borderline-SMOTE 그리고 ADASYN 기법은 원 데이터의 특성을 비교적 일관되게 유지하며 데이터의 개수를 증폭시킬 수 있었으나, CTGAN 기법은 다양성이 향상되는 결과를 확인하였다. 이는 사용자 목적에 따라 적합한 증폭 알고리즘을 고려해야 됨을 보여준다.

(2) SMOTE, Borderline-SMOTE 그리고 ADASYN 기법으로 생성된 데이터는 SHAP 기법 적용 후 주요 인자들의 중요도 순서가 일관되게 나타났다. CTGAN 기법에서는 데이터 증폭 수준에 따라 인자들의 중요도 순서가 달라지는 경향을 보였다. 이는 CTGAN 기법이 데이터의 다양성 확보에 기여할 수 있지만, 증폭된 데이터의 일관성은 저하될 수 있음을 시사한다.

(3) 이상치가 제거된 데이터에 SHAP 알고리즘을 적용하여 중요도 분석을 수행하였으며, 이상치가 포함된 데이터의 결과와 동일하게 나타났다. 이와 같은 결과는 해당 연구에서 제시한 이상치 제거 방법이 타당함을 보여준다.

## 감사의 글

본 연구는 과학기술정보통신부의 한국연구재단(NRF-2020R1A2C2012113)과 과학기술정보통신부 한국건설기술연구원 '수소도시 기반시설의 안전 및 수용성 확보 기술 개발(No.20240176-001)' 사업의 지원으로 수행되었으며 이에 감사드립니다.

## 참고문헌 (References)

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, Vol.16, pp.321-357.
2. Choi, H. J., Kim, S. W., and Kim, Y. S. (2022), "A Basic Study on Effect Analysis of Adjacent Structures due to Explosion of Underground Hydrogen Infrastructure", *Journal of Korean Geosynthetics Society*, pp.21-27.
3. Cordón, I., García, S., Fernández, A., and Herrera, F. (2018), "Imbalance: Oversampling algorithms for Imbalanced Classification in R", *Knowledge-Based Systems*, Vol.161, pp.329-341.
4. Go, G. H., Jeon, J. S., Kim, Y. S., Kim, H. W., and Choi, H. J. (2022), "Prediction of Hydrodynamic Behavior of Unsaturated Ground Due to Hydrogen Gas Leakage in a Low-depth Underground Hydrogen Storage Facility", *Journal of the Korean Geotechnical Society*, Vol.38, No.11, pp.107-118.
5. Han, H., Wang, W. Y., and Mao, B. H. (2005, August), "Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning", *In International Conference on Intelligent Computing* (pp.878-887), Berlin, Heidelberg: Springer Berlin Heidelberg.
6. He, H., Bai, Y., Garcia, E. A., and Li, S. (2008, June), "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", *In 2008 IEEE International Joint Conference on Neural Networks (IEEE world congress on computational intelligence)* (pp.1322-1328), Ieee.
7. Kim, S. Y., Lee, D., Yu, J. D., and Yoon, H. K. (2024), "A Study on the Characteristics of Applying Oversampling Algorithms to Fosberg Fire-Weather Index (FFWI) data", *Smart Structures and Systems*, Vol.34, No.1, p.9.
8. Lee, W. S., Kim, Y., Shinn, Y., Wang, J., Moon, B., Park, H., ... and Kwon, O. (2021), "Role of Blue Hydrogen for Developing National Hydrogen Supply Infrastructure", *Journal of the Korean Society of Mineral and Energy Resources Engineers*, Vol.58, No.5, pp.503-520.
9. Ning, Z. X., Su, M. X., Xue, Y. G., Qiu, D. H., Li, Z. Q., and Fu, K. (2021), "Reevaluation of the Design and Excavation of Underground Oil Storage Cavern Groups Using Numerical and Monitoring Approaches", *Geomech Eng*, Vol.27, No.3, pp.291-307.
10. Panfilov, M. (2016), Underground and Pipeline Hydrogen Storage, In Compendium of Hydrogen Energy (pp.91-115), Woodhead Publishing.
11. Rekha, G. and Reddy, V. K. (2018), "A Novel Approach for Handling Outliers in Imbalance Data", *International Journal of Engineering & Technology*, Vol.7, No.3.1, pp.1-5.
12. Shin, J. W. (2023), "Damage Evaluation of Adjacent Structures for Detonation of Hydrogen Storage Facilities", *Korean Society of Disaster & Security*, Vol.16, No.1, pp.61-70.
13. Taylor, J. B., Alderson, J. E. A., Kalyanam, K. M., Lyle, A. B., and Phillips, L. A. (1986), "Technical and Economic Assessment of Methods for the Storage of Large Quantities of Hydrogen", *International Journal of Hydrogen Energy*, Vol.11, No.1, pp.5-22.
14. Zivar, D., Kumar, S., and Foroozesh, J. (2021), "Underground Hydrogen Storage: A Comprehensive Review", *International Journal of Hydrogen Energy*, Vol.46, No.45, pp.23436-23462.

Received : September 30th, 2024
Revised : October 15th, 2024
Accepted : October 18th, 2024