TUNNEL & UNDERGROUND SPACE Vol.33, No.6, 2023, pp.594-609
https://doi.org/10.7474/TUS.2023.33.6.594 ISSN: 1225-1275(Print) ISSN: 2287-1748(Online)

ORIGINAL ARTICLE

커터수명지수 예측을 위한 다중선형회귀분석과 트리 기반 머신러닝 기법 적용

홍주표[1], 고태영[2*]

[1]강원대학교 강원대학교 에너지·인프라 융합학과 석사과정, [2]강원대학교 에너지자원·산업공학부 조교수

Application of Multiple Linear Regression Analysis and Tree-Based Machine Learning Techniques for Cutter Life Index(CLI) Prediction

Ju-Pyo Hong[1] and Tae Young Ko[2*]

[1]Graduate Student, Department of Integrated Energy and Infra System, Kangwon National University
[2]Assistant Professor, Department of Energy and Resources Engineering, Kangwon National University

***Corresponding author:** tyko@kangwon.ac.kr

Received: December 11, 2023
Revised: December 12, 2023
Accepted: December 18, 2023

ABSTRACT

TBM (Tunnel Boring Machine) method is gaining popularity in urban and underwater tunneling projects due to its ability to ensure excavation face stability and minimize environmental impact. Among the prominent models for predicting disc cutter life, the NTNU model uses the Cutter Life Index(CLI) as a key parameter, but the complexity of testing procedures and rarity of equipment make measurement challenging. In this study, CLI was predicted using multiple linear regression analysis and tree-based machine learning techniques, utilizing rock properties. Through literature review, a database including rock uniaxial compressive strength, Brazilian tensile strength, equivalent quartz content, and Cerchar abrasivity index was built, and derived variables were added. The multiple linear regression analysis selected input variables based on statistical significance and multicollinearity, while the machine learning prediction model chose variables based on their importance. Dividing the data into 80% for training and 20% for testing, a comparative analysis of the predictive performance was conducted, and XGBoost was identified as the optimal model. The validity of the multiple linear regression and XGBoost models derived in this study was confirmed by comparing their predictive performance with prior research.

Keywords: Disc cutter wear, Cutter Life Index, TBM, Machine learning, Regression analysis

초록

TBM 공법은 굴착면 안정성 확보 및 주변환경에 비치는 영향을 최소화하기 때문에 도심지나 하·해저터널 등에서 적용 사례가 증가하는 추세이다. 디스크 커터의 수명을 예측하는 대표적인 모델 중 NTNU모델은 커터수명지수(Cutter Life Index, CLI)를 주요 매개 변수로 활용하지만 복잡한 시험절차와 시험장비의 희귀성으로 측정에 어려움이 있다. 본 연구에서는 다중선형회귀분석과 트리 기반의 머신러닝 기법으로 암석물성을 활용하여 CLI를 예측하였다. 문헌 조사를 통해 암석의 일축압축강도, 압열인장강도, 등가석영함량과 세르샤 마모지수 등을 포함한 데이터베이스를 구축하였고 파생변수를 계산하여 추가하였다. 다중선형회귀분석은 통계적 유의성과 다중공선성을 고려하여 입력 변수를 선정하였고 머신러닝 예

측 모델은 변수 중요도를 기반으로 입력 변수를 선정하였다. 학습용과 검증용 데이터를 8:2로 나누어 모델 간 예측 성능을 비교한 결과 XGBoost가 최적의 모델로 선정되었다. 본 연구에서 도출된 다중선형회귀모델과 XGBoost모델을 선행 연구와 예측 성능을 비교하여 타당성을 확인하였다.

핵심어: 디스크 커터 마모, 커터수명지수, TBM, 머신러닝, 회귀분석

1. 서 론

TBM (Tunnel Boring Machine)은 지반을 굴착하는 기계화 시공장비로 도심지나 하·해저터널 등에서 굴착면 안정성 확보 및 주변 환경에 끼치는 피해를 최소화하는 등의 이유로 전 세계적으로 널리 사용되고 있다. 특히 우리나라에서도 수도권 광역급행철도(GTX)나 김포-파주 한강터널 등 TBM을 사용한 터널 시공 사례가 점점 증가하는 추세이다. 디스크 커터는 TBM이 암반을 굴착할 때 사용되는 도구로 TBM의 커터헤드에 장착되어 TBM의 추력과 커터헤드의 회전력으로 굴착한다. 굴착 작업 중 굴착면과의 직접적인 마찰과 충격으로 인해 디스크 커터는 필연적으로 마모가 발생하며, 마모 한계에 이른 디스크 커터는 TBM의 굴착성능을 저하시키고 다른 디스크 커터의 마모를 가속시키기 때문에 적절한 시기에 교체해야 한다.

디스크 커터의 교체로 인한 문제는 공기 지연 및 공사비 증가로 이어질 수 있으며, 정확한 공사 기간과 공사비를 산정하기 위한 디스크 커터 마모의 신뢰성 있는 예측이 중요하다. 디스크 커터의 수명을 예측하는 모델은 대표적으로 CSM모델, Gehring모델, NTNU모델 등이 있다. 미국의 CSM (Colorado School of Mines)에서 개발한 CSM모델은 세르샤 마모지수(Cerchar Abrasivity Index, CAI)를 바탕으로 디스크 커터의 수명을 계산한다(Rostami and Ozdemir, 1993). Gehring모델은 디스크 커터의 마모 질량과 CAI의 상관관계를 이용하여 디스크 커터의 수명을 예측한다(Gehring, 1995). 노르웨이의 NTNU (Norwegian University of Science and Technology)에서 수십 년간 경암에서 굴착한 TBM시공사례를 바탕으로 개발한 NTNU모델은 커터수명지수(Cutter Life Index, CLI)를 주요 매개 변수로 활용하여 디스크 커터의 수명을 예측한다(Bruland, 2000).

CLI는 Sievers' J-Value (SJ) 미니 드릴 시험과 커터마모지수(Abrasion Value Steel Cutters, AVS) 시험으로 측정되나(Fig. 1), 그 시험절차가 복잡하고 시험장비를 보유한 곳이 많지 않아 측정에 어려움이 있다.

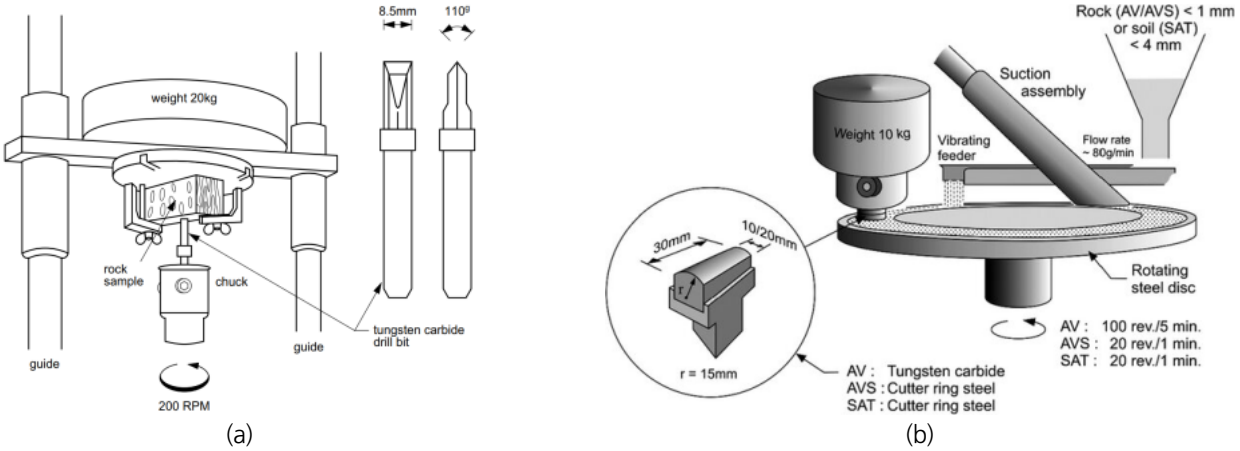


Fig. 1. Outline of Siever's J value (SJ) miniature drill test (a) and Abrasion Value Steel Cutters test (b) (Bruland, 2000)

암석의 물성을 활용하여 CLI를 예측하는 다양한 연구들이 진행되어 왔다. Massalov et al.(2022)는 CLI, 일축압축강도, 압열인장강도, 취성도(Brittleness)및 밀도를 포함한 80개의 암석물성자료를 바탕으로 CLI 예측 모델을 구현하였다. 암석물성자료는 퇴적암, 화성암과 변성암을 모두 포함하였다. 입력 변수의 조합으로 7 종류의 가지수를 활용하여 다중 선형 및 비선형 회귀분석, 인공신경망 및 퍼지 논리 예측 모델을 만들고 성능을 평가하였다. 다중 선형, 비선형 회귀분석, 인공신경망 및 퍼지 논리 모델의 결정계수 R^2는 각각 최대 0.71, 0.72, 0.71, 0.68로 나타났다. Aligholi et al.(2017)은 화성암에서 CLI를 측정하고 단일 및 다중 회귀 분석을 통해 암석학적 특성과 암석의 역학적 특성 간의 상관관계를 각각 조사하였다. CLI를 밀도, 공극률, P파속도, 슈미트 해머 반발 경도, 및 점하중 강도지수 등의 암석의 역학적 특성과 질감특성과 광물조성을 포함한 암석학적 특성으로 구분하여 다중 회귀 분석을 수행한 결과 결정계수 R^2 값이 각각 0.458과 0.709로 암석학적 특성이 더 높은 상관관계를 나타낸다고 밝혔다. Ko et al.(2016)은 문헌 조사를 통하여 CLI와 CAI를 포함한 암석 물성에 대한 데이터베이스를 구축하고 단일 및 다중회귀분석을 수행한 결과, CAI와 석영함량, 일축압축강도를 입력 변수로 활용하였을 때의 다중회귀분석이 단일회귀분석보다 더 높은 예측성능을 나타낸다고 보고하였다. Chang et al.(2011)은 국내 39개의 암석을 대상으로 NTNU모델의 주요 매개 변수인 CLI를 측정하기 위하여 NTNU시험을 수행하였고 측정된 CLI는 회귀분석을 통해 일축압축강도, 석영함량과의 상관관계를 조사하였다. 그 결과, CLI는 일축압축강도와는 유의한 관계를 보이지 않았으나 석영함량과 CLI는 선형적 반비례 관계임이 도출되었다.

NTNU 모델의 주요 매개변수인 CLI는 일반적인 암석물성 시험 장비가 아닌 별도의 시험 장비를 통해 측정되기 때문에 접근성과 활용성에 제약이 따른다. 일반적인 암석물성으로부터 CLI를 정확하게 추정할 수 있으면 NTNU 모델을 이용한 커터수명예측을 보다 쉽게 할 수 있을 것으로 판단된다. 따라서, 본 연구에서는 문헌연구를 통해 CLI, CAI, 일축압축강도, 압열인장강도, 등가석영함량 등을 포함한 데이터베이스를 구축하고, 다중선형회귀분석과 트리 기반 머신러닝 모델을 활용하여 CLI의 예측 모델을 구현하고 선행연구에서 제시된 예측 모델과 성능을 비교하여 타당성을 검증하고자 한다.

2. 문헌연구

2.1 CLI 예측에 대한 선행연구

Massalov et al.(2022)는 퇴적암, 화성암, 변성암의 암종을 가진 80개의 암석물성자료를 이용하여 CLI와 일축압축강도, 압열인장강도, 취성도(Brittleness)및 밀도와의 상관관계를 다중 선형 및 비선형 회귀분석을 통해 분석하였다. 그 결과, 일축압축강도와 압열인장강도를 이용한 관계식을 다음의 식 (1), (2)로 나타냈다.

$$CLI = -0.048\,UCS - 4.601\,BTS + 74.176 \quad (R^2 = 0.53) \tag{1}$$

$$CLI = -3.38\ln(UCS) - 40.63\ln(BTS) + 128.817 \quad (R^2 = 0.62) \tag{2}$$

여기서 UCS는 암석의 일축압축강도이고, BTS는 압열인장강도이다.

Ko et al.(2016)은 Bruland(2000)가 제시한 CLI와 CAI의 대략적 관계 그래프를 다음과 같은 식으로 나타냈다.

$$CLI = 2.87CAI^2 - 35.62CAI + 112.9 \quad (3)$$

또한 Ko et al.(2016)은 Dahl et al.(2012)이 제안한 CAI와 Sievers' J-Value (SJ)의 관계, 그리고, CAI와 AVS의 관계를 이용하여 CLI 예측식을 다음과 같이 제안하였다.

$$CLI = 115.24CAI^{-1.724} \quad (4)$$

Ko et al.(2016)은 문헌 조사를 통하여 수집된 화성암과 변성암에서의 38개 CLI, CAI, 일축압축강도, 석영함량(QC)의 자료를 이용하여 CLI 예측식을 다음과 같이 제안하였다.

$$CLI = -8.725\ln(CAI) + 18.898 \; (R^2 = 0.46) \quad (5)$$

$$CLI = 19.029 - 1.515CAI - 0.0796QC - 0.02654UCS \; (R^2 = 0.76) \quad (6)$$

$$CLI = 114.6QC^{-0.281} \times UCS^{-0.4} \; (R^2 = 0.72) \quad (7)$$

Macias(2016)은 Bruland(2000)의 그래프에 추가하여 새롭게 CAI와 CLI의 상관관계를 도출하고 다음의 식 (8)과 같이 제시하였다.

$$CAI = 7.50CLI^{-0.26} \quad (8)$$

2.2 트리 기반 머신러닝 모델 기법

본 연구에서는 의사결정나무(Decision tree)와 같은 트리 기반의 랜덤 포레스트(Random forest), 엑스트라 트리(Extra tree), 익스트림 그래디언트 부스팅(XGBoost), 그래디언트 부스팅 머신(GBM), 에이다 부스트(Adaboost) 알고리즘을 활용하여 예측 모델을 구현하였다. 트리 기반 모델들은 데이터를 분석하고 예측하기 위하여 계층적 구조를 가지고 데이터의 복잡한 구조와 비선형 관계를 모델링하는 능력이 뛰어나며, 각 변수의 중요도를 평가할 수 있는 기능을 가지고 있다.

2.2.1 의사결정나무(Decision Tree)

의사결정나무는 1개 이상의 독립변수에 대해 if-then 규칙으로 데이터를 나눠서 종속변수를 예측한다(Kuhn and Johnson, 2013). 먼저 나무의 가지를 따라 나무의 끝인 말단노드에 도달할 때까지 여러 if-then 규칙이 적용되며, 말단노드에 도달하면 해당 노드의 수식을 통해 최종 예측값이 생성되므로, 새로운 입력데이터에 대한 종속변수 예측이 가능하다.

2.2.2 랜덤 포레스트(Random Forest)

랜덤 포레스트는 여러 개의 모델을 생성하고 그 예측을 결합하여 단일모델보다 정확한 예측을 도출하는 기법인 앙상블 기법의 하나로 원본 데이터에서 무작위 추출을 통해 부분집합을 만들고, 만들어진 부분집합을 입력데이터로 하는 모델들을 만든다(Cutler et al., 2012). 각 모델의 노드 분할은 원래의 독립변수 조합에서 n개의 독립변수를 무작위로 선택하고 최적의 분할을 보이는 독립변수와 분할점을 찾아 해당 노드를 하위 두 개의 노드로 분할한다. 회귀분석의 경우 각 노드의 최종 예측의 평균이 최종 예측값이 된다.

2.2.3 엑스트라 트리(Extra Tree)

엑스트라 트리는 랜덤 포레스트와 매우 유사한 알고리즘이지만 두 가지 주요 차이점이 있다. 첫번째 차이점은 각 모델의 학습에는 무작위 추출을 하지 않은 전체 학습 데이터가 사용된다는 점이고, 두번째 차이점은 분할에 사용될 변수와 분할점 모두 무작위로 선택되어 노드분할이 진행된다는 점이다. 이처럼 무작위성이 높기 때문에 랜덤 포레스트에 비해 계산시간이 빠르다는 장점이 있다(Geurts et al., 2006).

2.2.4 익스트림 그래디언트 부스트(XGBoost)

익스트림 그래디언트 부스팅은 그래디언트 부스팅 모델의 개선된 형태로, 기본 원리는 그래디언트 부스팅 알고리즘과 같지만, 특정 변수의 모든 데이터를 크기순으로 정렬한 후, N개의 버킷으로 나누고, 각 버킷에 속하는 데이터에 대해 분할을 진행하며 최적의 분할점을 찾는 방법인 근사 전략(Approximation strategies)을 활용하여 처리 속도를 개선하였다. 근사 전략으로 전체 분할 탐색 수를 줄일 수 있을 뿐만 아니라, 각 버킷의 분할 탐색 과정을 병렬로 처리할 수 있다(Chen and Guestrin, 2016).

2.2.5 그래디언트 부스팅(Gradient Boosting)

그래디언트 부스팅은 이전 모델의 예측값과 실제값의 차이인 잔차를 다음 학습기의 목표 추정치로 설정한다. 학습은 이러한 잔차를 줄여나가는 방향으로 진행되며, 잔차를 감소시키는 과정은 손실함수를 미분하여 경사 하강법(Gradient descent)을 통해 이루어진다(Natekin and Knoll, 2013).

2.2.6 에이다부스트(Adaboost)

에이다부스트는 앙상블 기법으로 예측 성능이 낮은 데이터에 대해서 가중치를 크게 부여하여 다음 학습기의 입력 샘플로 선택될 확률을 증가시키는 방식으로 작동한다(Drucker, 1997). 즉, 이전 모델에서 예측 성능이 낮은 데이터에 대해서 집중적으로 학습을 진행하며 예측 성능을 개선 해나가는 알고리즘이다.

3. 데이터수집, 분석 및 변수선택

3.1 데이터수집

본 연구에 사용된 데이터는 문헌 조사를 통해 수집되었으며(Majeed et al., 2020, Majeed and Bakar, 2019, Eide, 2014,

Aligholi et al., 2017, Macias et al., 2016), 암석물성은 CLI, UCS, BTS, 등가석영함량(Equivalent Quartz Content, EQC), CAI를 포함한다. 또한, 예측 모델의 학습에 더 적합한 변수 형태 고려하여 기존 변수들의 함수로 표현되는 파생 변수를 데이터베이스에 추가하였다. 취성도는 암석의 파쇄성이나 충격에 대한 반응 정도를 측정하는 지표로서 일반적으로 인정되는 국제적인 표준이 없어 연구자마다 다양한 방식으로 정의된다(Meng et al., 2021). 본 연구에서 사용된 취성도(B_1~B_5)는 일축압축강도와 압열인장강도로 표현되며, 다음의 식 (9)~(13)과 같다(Hucka and Das, 1974, Özfirat et al., 2016, Altindag, 2003, Altindag, 2010).

$$B_1 = \frac{\sigma_c}{\sigma_t} \tag{9}$$

$$B_1 = \frac{\sigma_c - \sigma_t}{\sigma_c + \sigma_t} \tag{10}$$

$$B_3 = \frac{\sigma_c + \sigma_t}{2} \tag{11}$$

$$B_4 = \frac{\sigma_c \sigma_t}{2} \tag{12}$$

$$B_5 = \sqrt{\frac{\sigma_c \sigma_t}{2}} \tag{13}$$

여기서, σ_c는 일축압축강도, σ_t는 압열인장강도이다.

본 연구에서 수집된 자료를 학습하여 새롭게 취성도 BI를 정의하였다. 종속변수 CLI와 독립변수 일축압축강도와 압열인장강도의 관계를 비선형회귀분석으로 수행하여 CLI와의 결정계수가 가장 크게 될 때의 지수를 구하고, 다음과 같은 식으로 제안하였다.

$$BI = \sigma_c^{-0.3} \cdot \sigma_t^{-0.26} \tag{14}$$

점착력(Cohesion, C_o)은 물체의 입자 간에 서로 붙어있는 결합력을 나타내는 정도로 암반 공학에서 활용되는 주요 파괴 기준인 Mohr-Coulomb 파괴 기준을 이용한 Mohr-Coulomb 파괴 포락선 식으로부터 구할 수 있다. Piratheepan et al.(2012)이 제안하고, Moon and Yang(2020)이 정리한 일축압축강도와 압열인장강도로 점착력을 추정하는 식은 다음과 같다.

$$C_o = \frac{0.5\sigma_c\sigma_t}{\sqrt{\sigma_t(\sigma_c - 3\sigma_t)}} \tag{15}$$

Plinninger(2002)는 암석의 등가석영함량과 일축압축강도의 곱으로 표현되는 암석 마모 지수(Rock Abrasivity Index, RAI)를 제안하였으며, 식 (16)과 같다.

$$RAI = EQC \times UCS \tag{16}$$

수집된 암석물성들의 모든 변수 값이 다 있지 않기 때문에 결측치가 있는 자료는 제외하고 이상치를 제거하여 최종적으로 연구에 사용된 데이터베이스는 CLI, 일축압축강도(UCS), 압열인장강도(BTS), 등가석영함량(EQC), 취성도(B_1~B_5), BI, 점착력(C_o), 암석마모지수(RAI), CAI를 포함하는 87개의 데이터 세트로 구성된다. 화성암 64개, 변성암 13개, 퇴적암 10개로 변성암과 퇴적암의 수가 충분하지 않아 암종별로 나누어 분석하지 않았다.

3.2 탐색적 데이터 분석

수집된 데이터 세트의 기초통계분석, 시각화, 변수 간 관계 탐색을 위한 탐색적 데이터 분석(Exploratory data analysis)을 수행하여 본격적인 회귀분석과 머신러닝 모델 구현 전에 데이터를 다양한 관점에서 살펴보았다. Fig. 2에 수집된 데이터 세트 각 변수의 빈도 분포 히스토그램을 나타내었다. 등가석영함량과 점착력를 제외하면 대부분 비정규분포의 양상을 나타낸다.

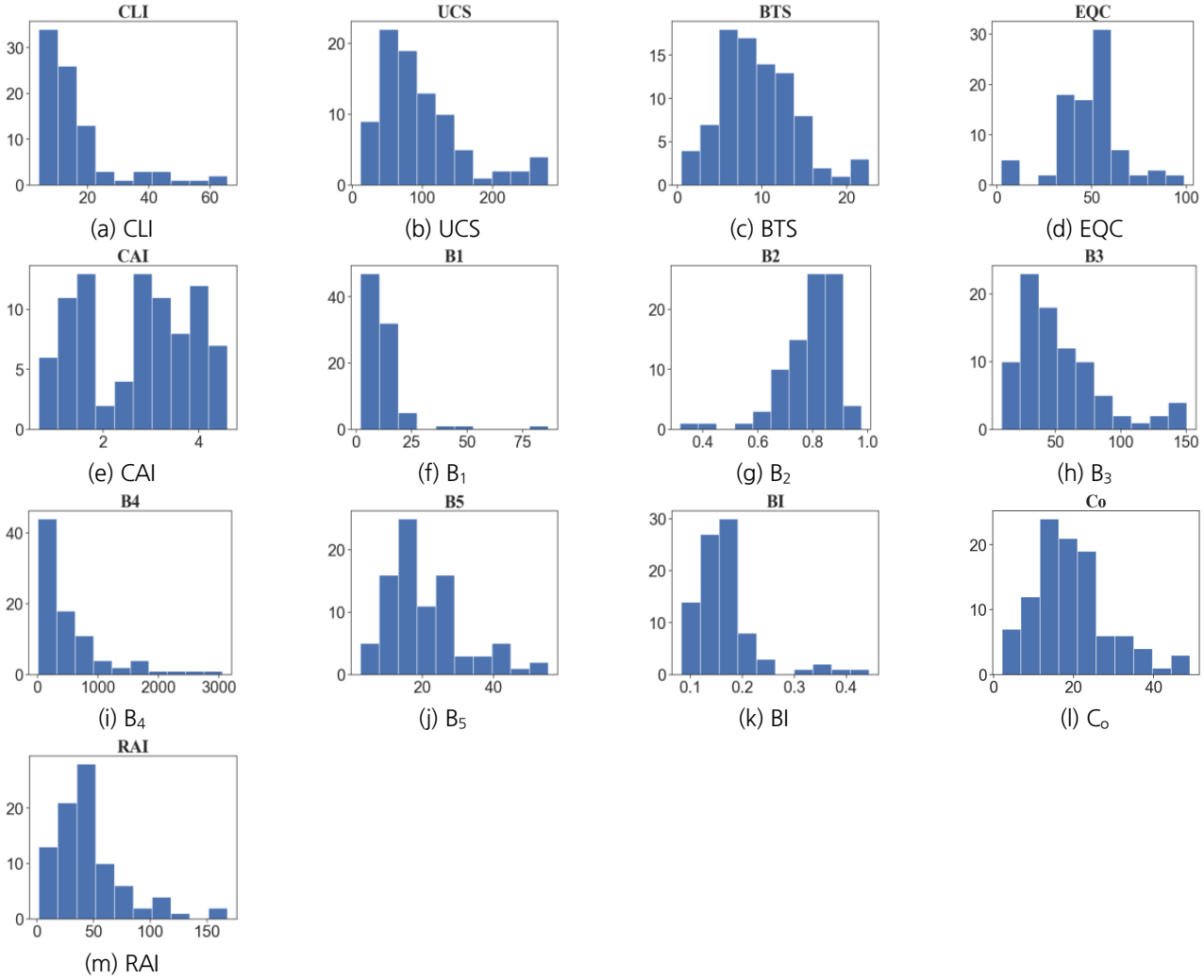


Fig. 2. Histogram of frequency distribution for each variable

변수들 사이의 상관관계는 Fig. 3의 히트맵을 통하여 시각적으로 표현하였다. 상관계수는 −1부터 +1까지의 범위를 가지며, +1에 근접하면 강한 양의 상관관계, -1에 근접하면 강한 음의 상관관계를 의미한다. 상관계수가 0에 근접하면 상관관계가 없음을 나타낸다. CLI는 CAI와 강한 음의 상관관계를 보여주며, RAI는 그 다음 강한 음의 상관관계를 나타낸다. B_5, C_o, BTS, EQC, B_3, UCS, B_4 변수들도 CLI와 음의 상관관계를 보인다. B_1, B_2 변수들을 CLI와 매우 약한 상관관계를 보인다. BI는 CLI와 가장 강한 양의 상관관계를 가지는 변수로 나타났다. 또한 취성도와 관련된 변수들과 UCS, BTS, C_o 변수들 사이에 높은 상관관계를 나타내므로 향후 변수들간의 다중공선성 문제를 확인해야 할 필요가 있는 것으로 판단되었다.

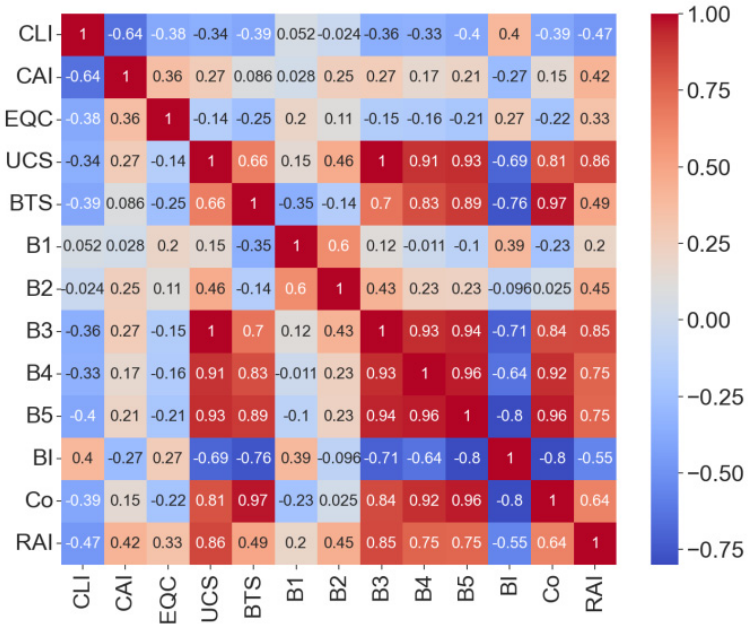
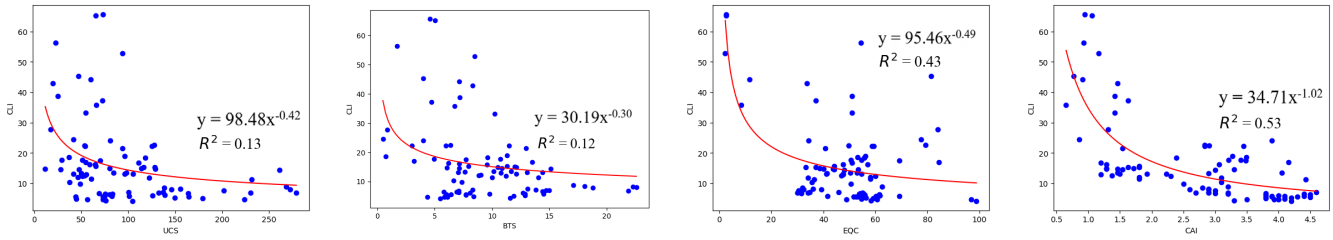


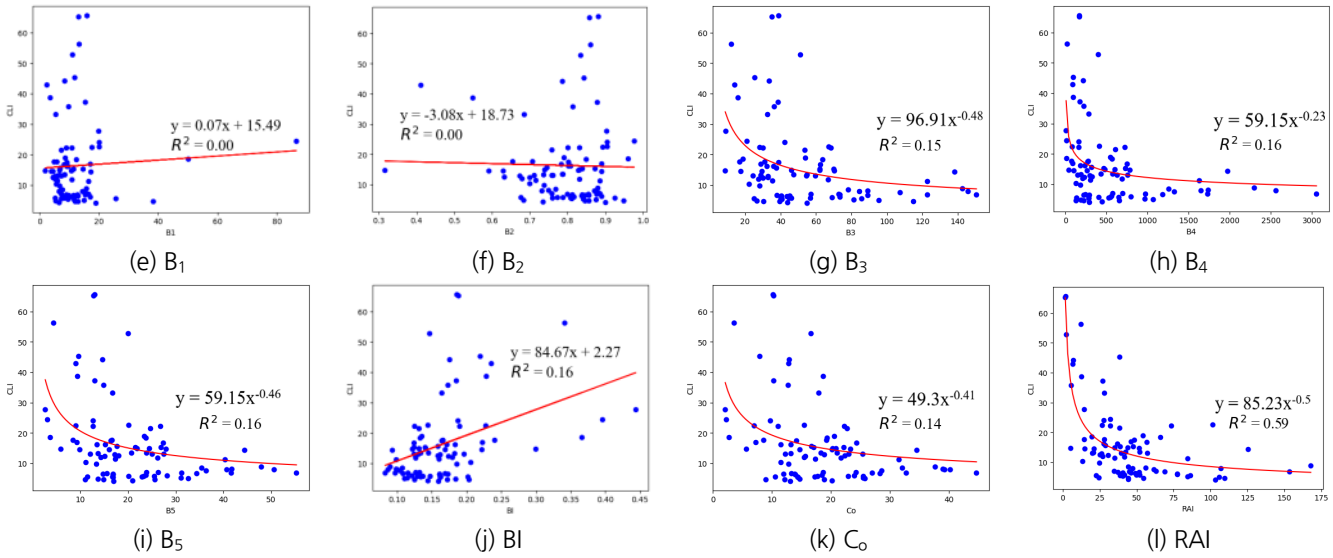
Fig. 3. Heatmap for representing correlation in CLI data

다음 Fig. 4는 CLI와 각 변수들 사이의 산포도로 CLI와 다른 변수간의 관계를 나타낸다. 데이터의 분포 형태로부터 변수들간의 선형 관계나 비선형 관계가 있음을 알 수 있다. CLI와 다른 변수들간의 관계는 대부분 비선형 관계를 보이며, 특히 EQC, CAI, RAI는 멱함수의 형태로 표현이 가능한 관계를 나타내며 결정계수 R^2도 0.40 이상으로 추가적인 파생변수를 도출할 수 있는 가능성을 보여주었다. Fig. 3의 히트맵의 결과와 동일하게 CLI는 UCS, BTS, EQC, CAI, B_3, B_4, B_5, C_o, RAI와 음의 상관관계가, BI와 양의 상관관계가 있으며, B_1과 B_2와는 뚜렷한 상관관계를 발견할 수 없었다.



(a) UCS (MPa) (b) BTS (MPa) (c) EQC (%) (d) CAI

Fig. 4. Relationships between CLI and other variables



(e) B_1 (f) B_2 (g) B_3 (h) B_4

(i) B_5 (j) BI (k) C_o (l) RAI

Fig. 4. Relationships between CLI and other variables (continued)

3.3 변수선택

3.3.1 다중선형회귀분석

다중선형회귀분석의 입력 변수 선택은 모델의 성능과 해석력에 영향을 미치는 단계로 변수의 유의성과 다중공선성을 고려해야 한다. 변수 간 높은 선형상관관계로 발생하는 다중공선성은 모델의 안정성과 해석력에 악영향을 미친다. 최적의 입력 변수 선정을 위하여 가능한 모든 변수 조합을 생성하여 각각에 대한 모델을 구현하고 평가하여 최적의 입력 변수를 선택하는 완전 탐색법(Exhaustive search method)은 변수의 수가 증가함에 따라 조합의 수가 기하급수적으로 증가하므로 계산시간과 비용이 많이 소요되는 단점이 있다.

이에 대한 대안으로, 본 연구에서는 계산 효율성을 고려하여 전진 선택법(Forward selection method), 후진 제거법(Backward elimination method), 단계적 선택법(Stepwise selection method)으로 변수 선택을 수행하였다. 전진 선택법은 출력 변수 상관관계 및 유의성이 가장 큰 입력 변수부터 입력하여 하나씩 입력 변수를 선택하는 방법이며, 후진 제거법은 전진 선택법과는 반대로 모든 변수를 선택한 후 설명력이 가장 낮은 입력 변수부터 한 개씩 순차적으로 제거하여 선정한다. 단계적 선택법은 전진 선택법과 후진 제거법의 단점을 보완하기 위해 전진 선택법의 각 단계에서 이미 선택된 변수들의 중요도를 다시 검사하여 더 이상 중요하지 않은 변수를 제거하는 방법이다. 변수 선택 과정에서 변수들의 유의성은 독립변수가 종속변수에 미치는 효과가 통계적으로 유의미한 정도를 나타내는 P-value를 고려하였으며 P-value가 0.05보다 크면 해당 변수의 회귀계수가 통계적으로 유의미하지 않다고 판단하였다. 선정된 변수의 다중공선성 문제를 고려하여 변수 간 상관성을 나타내는 통계적 지표인 분산 팽창 지수(VIF)를 확인하였고, 일반적으로 10 이상인 변수는 상관관계가 높다고 판단되기 때문에 10이하의 변수만을 선택하였다(Table 1). 다중회귀분석을 위해 선택된 변수는 전진 선택법과 단계적 선택법 모두에 의해서 선택된 BTS, EQC, CAI이다.

Table 1. Variables selected by the variable selection method along with their VIF and R^2

	Forward selection			Backward elimination			Stepwise selection		
Variables	BTS	EQC	CAI	EQC	B4	CAI	BTS	EQC	CAI
VIF	3.61	6.34	7.53	6.14	1.85	7.21	3.61	6.34	7.53
R^2		0.61			0.51			0.61	

또한 Fig. 4에서 EQC, CAI, RAI는 CLI와 멱함수의 관계로 표현되며 결정계수가 0.4이상으로 각 변수에 지수를 취한 것을 새로운 변수로 정의하고, P-value가 0.05 이하인 것만을 선택하여 다음의 Table 2에 나타내었다. $CAI^{-1.0}$과 $RAI^{-0.5}$는 분산 팽창 지수가 4.51로 다중공선성 문제는 발견되지 않는 것으로 판단된다.

Table 2. New variables obtained by applying exponent values to the variables

Variables	$CAI^{-1.0}$	$RAI^{-0.5}$
P-value	1.98×10^{-06}	2.93×10^{-13}
R^2	0.71	

3.3.2 머신러닝모델

본 연구에서 머신러닝 모델들은 트리 기반의 모델들로 변수의 중요도를 수치적으로 나타낼 수 있다. 변수의 중요도는 각 변수가 모델 예측에 기여한 정도를 평가하는 데 사용되는데, 트리의 각 분할에서 해당 변수가 기여하는 불순도(Impurity) 감소량을 기반으로 계산된다. 불순도는 각 노드에서 데이터의 동질성을 나타내는 지표로, 노드의 분할이 불순도를 줄이는 방향으로 이루어짐으로써 변수의 중요도를 결정한다. 트리 기반의 모델은 데이터의 복잡한 구조와 변수 간의 비선형적 상호작용을 고려할 수 있기 때문에, 선형 회귀 모델에서 문제가 되는 다중공선성의 영향을 상대적으로 덜 받는 것으로 여겨진다. 본 연구에서는 사용된 머신러닝 모델 각각의 변수 중요도를 확인하고 평균값을 구하여 다양한 모델에서 공통적으로 중요도가 높은 변수를 확인하였다(Fig. 5). 변수가 많아짐으로 발생하는 모델의 복잡성을 고려하여 상위 4개의 BTS, EQC, RAI, CAI를 머신러닝 모델의 입력 변수로 선정하였다.

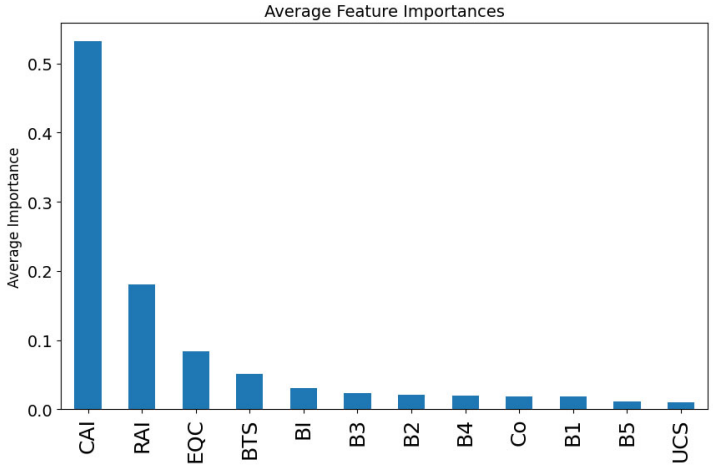


Fig. 5. Average variable importance ranking for each machine learning model

4. 모델 결과 및 성능 비교

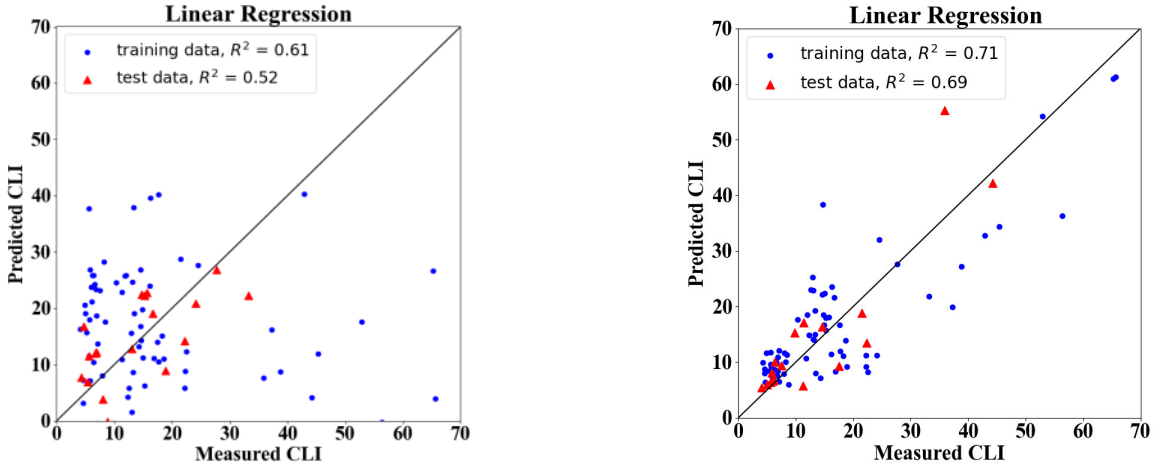
4.1 다중선형회귀분석

다중선형회귀분석에 사용된 입력변수는 BTS, EQC, CAI 및 $CAI^{-1.0}$, $RAI^{-0.5}$의 두 종류이다. 예측 모델의 성능을 확인하기 위하여 전체 데이터를 학습용과 검증용 8:2의 비율로 나누었다. 모델의 안정성과 데이터의 패턴이나 순서에 따른 편향을 최소화하기 위해 K-fold 교차 검증 방법을 활용하였다. K-fold 교차 검증은 트레이닝 데이터를 k개의 동일한 크기의 폴드(Fold)로 나누어 모델을 k번 학습시키고 평가하는 방법으로 각 반복에서 다른 폴드가 테스트 셋으로 사용되며 나머지 폴드는 학습에 사용된다. 본 연구에서는 5-fold 교차 검증을 수행하였고 5개의 값의 평균값을 최종 평가지표로 활용하였다. 모델 성능 평가를 위해 모델의 예측 수준을 파악할 수 있는 결정계수 R^2과 오차 추정을 위한 평균 제곱근 오차(Root-mean-square error, RMSE)를 평가지표로 선정했다. 결정계수 R^2는 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표로, 추정된 회귀 모델이 실제 관측값을 어느 정도 정확하게 예측하는지를 나타내는 데 사용된다. 일반적으로 결정계수는 0부터 1 사이의 값으로 표현되며, 값이 1에 가까울수록 모델의 예측 성능이 뛰어남을 나타낸다. 평균 제곱근 오차는 회귀모델의 예측값과 실제값의 차이를 기반으로 예측 성능을 판단하는 지표이며 일반적으로 수치가 낮을수록 성능이 우수하다.

다중선형회귀분석 결과는 다음의 Table 3에 정리하였으며, 예측된 CLI와 측정된 CLI사이의 산포도와 y=x 직선 기준의 결정계수를 Fig. 6에 나타내었다.

Table 3. Summary of multiple linear regression analysis results for CLI

Input variables	Train data		Test data	
	RMSE	R2	RMSE	R2
BTS, EQC, CAI	8.4	0.61	8.68	0.52
$CAI^{-1.0}$, $RAI^{-0.5}$	7.14	0.71	7.35	0.62



(a) Input variables : BTS, EQC, CAI (b) Input variables : $CAI^{-1.0}$, $RAI^{-0.5}$

Fig. 6. Scatter plot between predicted CLI and measured CLI in a multiple linear regression

본 연구를 통해 도출된 CLI의 예측식은 다음과 같다. 멱함수의 형태로 된 입력변수를 이용한 예측식의 성능이 더 뛰어났다.

$$CLI = -1.27BTS - 0.24EQC - 5.85CAI + 55.92 \ (R^2 = 0.61) \tag{17}$$

$$CLI = 20.59CAI^{-1.0} + 58.92RAI^{-0.5} - 4.78 \ (R^2 = 0.71) \tag{18}$$

4.2 머신러닝 회귀분석

머신러닝 회귀분석에 사용된 입력변수는 CAI, RAI, EQC 및 BTS의 네 종류이며, 분석에 사용된 머신러닝 모델은 의사결정나무, 랜덤 포레스트, 엑스트라 트리, 익스트림 그래디언트 부스팅, 그래디언트 부스팅 머신, 그리고 에이다 부스트의 6개이다. 최적의 머신러닝 모델을 구현하기 위하여 하이퍼파라미터 튜닝(Hyperparameter tuning)을 수행하였다. 하이퍼파라미터는 머신러닝 모델에서 자동으로 학습되지 않는 매개변수로, 사용자가 수동으로 설정하여 모델의 훈련 과정을 최적화하기 위해 조정하는 변수이다. 지정한 파라미터 안에서 가장 좋은 성능을 내는 하이퍼 파라미터 조합을 찾는 그리드서치(GridSearch) 튜닝 방법을 사용하여 예측 모델에 대한 최적의 파라미터를 결정하였다(Table 4).

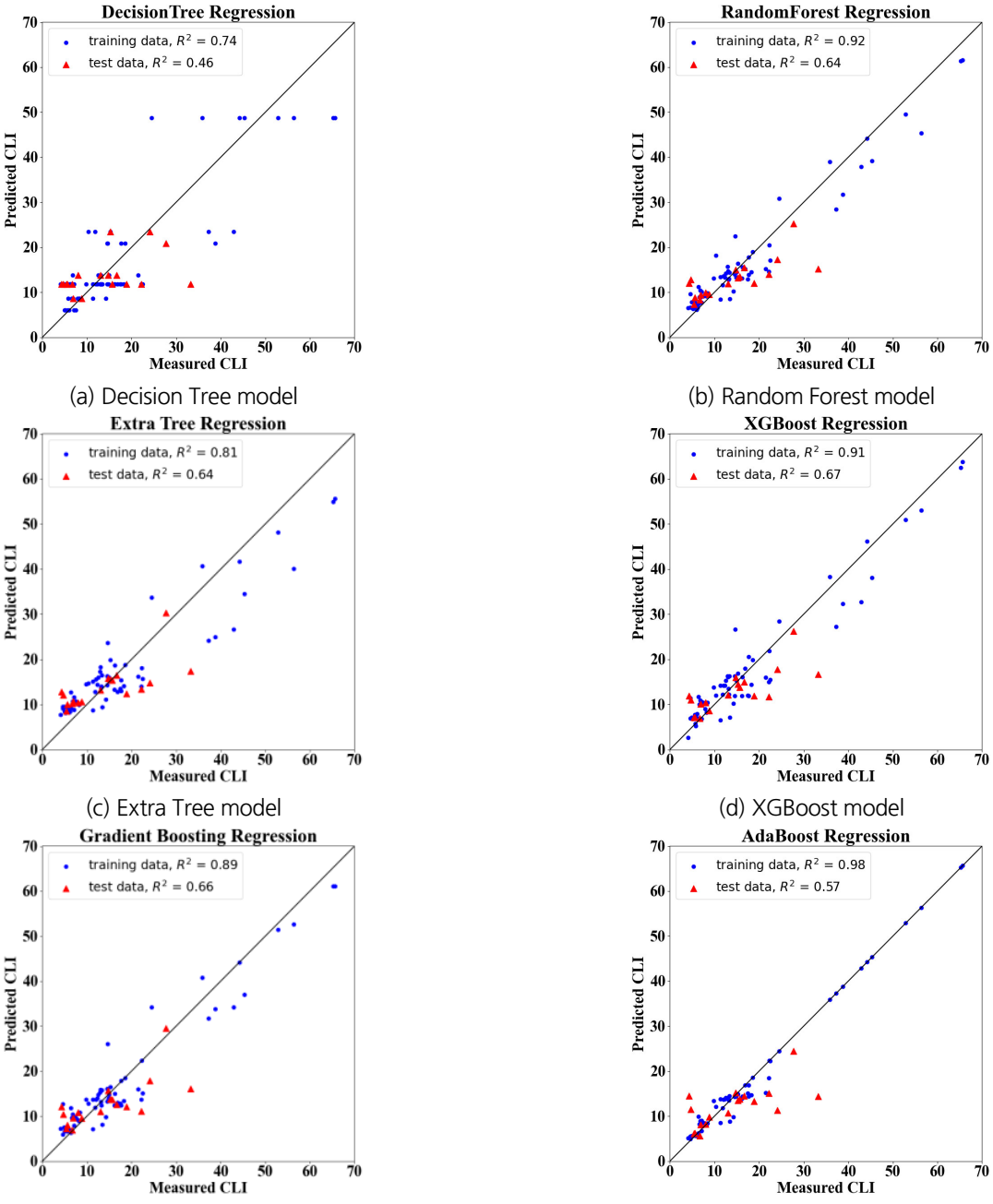
Table 4. Hyperparameters for CLI prediction models

Model	Hyperparameter
Decision Tree	max_depth: 4, max_features: sqrt, min_samples_leaf: 5, min_samples_split: 2
Random Forest	max_depth: 5, max_features: sqrt, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
Extra Tree	max_depth: 5, max_features: sqrt, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 70
XGBoost	learning_rate: 0.1, max_depth: 1, n_estimators: 300
GrdientBoost	max_depth: 2, max_features: sqrt, min_samples_leaf: 2, min_samples_split: 10, n_estimators: 30, learning rate : 0.1,
Ada Boost	estimator=DecisionTreeRegressor (max_depth=3), loss : linear, n_estimators: 30, learning_rate: 1

머신러닝 회귀분석의 결과는 다음의 Table 5와 같다. 전체적으로 학습용 데이터의 결정계수가 검증용 데이터의 결정계수보다 크며 그 차이는 최소 0.17부터 최대 0.41에 이른다. 학습용 데이터의 모델이 다소 과적합이 되었다고 볼 수 있다. 이는 수집된 데이터가 문헌연구를 통해 획득되어 각 암석이 채취된 지역이나 암종별 특이성으로 인한 데이터의 대표성이 부족할 수도 있고, 사용된 데이터 수가 87개로 머신러닝 학습에 다소 부족했을 수도 있다. 각 모델별 예측된 CLI와 실제 CLI값을 나타낸 산포도이며, y=x의 선을 기준으로 결정계수를 Fig. 7에 나타내었다. 각 모델들의 R^2값과 훈련용과 검증용 R^2값의 차이, 산포도의 형태 등을 종합적으로 고려하여 XGBoost 모델을 최적 모델로 선정하였다.

Table 5. Performance of machine learning models

Model	Train data		Test data	
	RMSE	R^2	RMSE	R^2
Decision Tree	6.83	0.74	9.08	0.46
Random Forest	3.72	0.92	7.17	0.64
Extra Tree	5.78	0.81	7.37	0.64
XGBoost	4.00	0.91	6.55	0.67
Gadient Boosting	4.32	0.89	7.08	0.66
AdaBoost	1.67	0.98	7.56	0.57



(a) Decision Tree model (b) Random Forest model

(c) Extra Tree model (d) XGBoost model

(e) Gradient Boosting model (f) AdaBoost model

Fig. 7. Scatter plot between predicted CLI and measured CLI in a machine learning model

4.3 선행 연구와 성능 비교

본 연구에서 도출된 선형회귀분석 모델과 최적의 모델로 선정된 XGBoost 모델을 선행연구들의 예측 모델과 성능을 비교하였다. 선행연구들의 예측모델들은 식 (1)~(5), (8)과 같으며, Fig. 8에 본 연구에서 사용된 87개의 실제 CLI와 예측된 CLI를 도식화하였다. 모델의 성능 지표는 RMSE를 사용하였으며, 그 결과는 Table 6에 정리하였다. 본 연구에서 도출된 선형회귀모델과 XGBoost 모델 모두 기존 선행연구의 예측모델보다 더 낮은 RMSE값을 나타내어 본 연구의 예측 모델의 타당성을 확인할 수 있다.

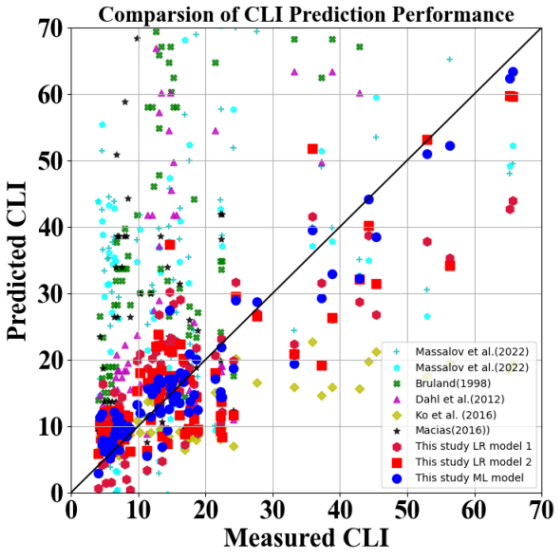


Fig. 8. Performance comparison of CLI prediction models

Table 6. Summary of CLI prediction models

No.	Reference	Model Equation	RMSE
1	Massalov et al.(2022)	$CLI = -0.048UCS - 4.601BTS + 74.176$	23.58
2	Massalov et al.(2022)	$CLI = -3.38\ln(UCS) - 40.63\ln(BTS) + 128.817$	28.94
3	Bruland(2000) (modified by Ko et al.(2016))	$CLI = 2.87CAI^2 - 35.62CAI + 112.9$	30.47
4	Dahl et al.(2012) (modified by Ko et al.(2016))	$CLI = 115.24CAI^{-1.724}$	40.47
5	Ko et al.(2016)	$CLI = -8.725\ln(CAI) + 18.898$	12.02
6	Macias(2016)	$CAI = 7.50CLI^{-0.26}$	19.53
7	This study LR model 1	$CLI = -1.27BTS - 0.24EQC - 5.85CAI + 55.92$	8.46
8	This study LR model 2	$CLI = 20.59CAI^{-1.0} + 58.92RAI^{-0.5} - 4.78$	7.18
9	This study ML model	XGBoost	4.20

5. 결 론

본 연구에서는 암석의 물성을 이용해 커터수명지수를 예측하기 위해 다중선형회귀분석과 여섯 가지 트리 기반 머신러닝 모델을 사용하였다. 사용된 머신러닝 모델은 의사결정나무, 랜덤 포레스트, 엑스트라 트리, 익스트림 그래디언트 부스팅, 그래디언트 부스팅 머신, 그리고 에이다 부스트로, 머신러닝 모델 간 성능 비교를 통해 최적 모델을 선정하였고, 본 연구에서 도출된 다중선형회귀분석 모델과 최적 머신러닝 모델을 선행 연구의 예측 모델과 비교하였다.

문헌 조사를 통하여 87개의 일축압축강도, 압열인장강도, 등가석영함량, 세르샤 마모지수가 포함된 데이터베이스를 구축하였고 이 데이터를 활용해 취성도, 점착력, 암석마모지수를 추가하여 분석하였다. 다중선형회귀분석의 입력 변수는 통계적 유의성과 다중공선성을 고려하여 등가석영함량, 세르샤 마모지수, 압열인장강도가 선정되었고, 멱함수의 형태로 된 입력변수 $CAI^{-1.0}$, $RAI^{-0.5}$도 선정되었다. 머신러닝 모델 입력 변수는 변수 중요도를 고려하여 등가석영 함량, 압열인장강도, 세르샤 마모지수, 암석마모지수

가 선정되었다. 머신러닝 모델의 학습 데이터와 테스트 데이터는 8:2 비율로 나눠서 적용하였고 평균 제곱근 오차와 결정계수로 성능을 비교하였다.

등가석영함량, 세르샤 마모지수, 압열인장강도를 입력변수로 하는 다중선형회귀분석의 결정계수는 0.61로 도출되었고, 멱함수의 형태로 된 입력변수를 이용한 다중선형회귀분석의 결정계수는 0.71로 더 높은 예측 성능을 나타내었다. 트리 기반 머신러닝 모델들의 R^2값과 훈련용과 검증용 R^2값의 차이, 산포도의 형태 등을 종합적으로 고려하여 XGBoost 모델을 최적 모델로 선정하였다.

도출한 선형회귀분석 모델과 XGBoost 모델을 선행 연구들의 예측 모델과 예측 성능을 비교하였고, 본 연구에서 도출된 예측 모델들의 성능이 더 높은 것으로 확인되었다.

감사의 글

이 논문은 2023년도 정부 (산업통상자원부)의 재원으로 해외자원개발협회의 지원(2021060003, 스마트 마이닝 전문 인력 양성)과 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2022R1F1A1063228)

REFERENCES

Aligholi, S., Lashkaripour, G.R., Ghafoori, M., and Azali, S.T., 2017, Evaluating the relationships between NTNU/SINTEF drillability indices with index properties and petrographic data of hard igneous rocks, Rock Mech. Rock Eng., 50, 2929-2953.

Altindag, R., 2003, Correlation of specific energy with rock brittleness concepts on rock cutting, J. South. Afr. Inst. Min. Metall., 103(3), 163-171.

Altindag, R., 2010, Assessment of some brittleness indexes in rock-drilling efficiency, Rock Mech. Rock Eng., 43, 361-370.

Bruland, A., 2000, Hard rock tunnel boring: vol. 8 drillability test methods, Ph. D. Thesis, Norwegian University of Science and Technology.

Chang, S.H., Choi, S.W., Lee, G.P., and Bae, G.J., 2011, Statistical analysis of NTNU test results to predict rock TBM performance, J. of Korean Tunn Undergr Sp. Assoc., 13(3), 243-260.

Chen, T. and Guestrin, C., 2016, Xgboost: A scalable tree boosting system, Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 785-794.

Cutler, A., Cutler, D.R., and Stevens, J. R., 2012, Random forests, Ensemble machine learning: Methods and applications, 157-175.

Dahl, F., Bruland, A., Jakobsen, P.D., Nilsen, B., and Grov E., 2012, Classifications of properties influencing the drillability of rocks, Tunn. Undergr. Space Technol., 28, 150-158.

Drucker, H., 1997, Improving regressors using boosting techniques, In icml, 97, 107-115.

Eide, L.N.R., 2014, TBM tunnelling at the Stillwater Mine, MS Thesis, NTNU.

Gehring, K., 1995, Leistungs-und Verschleißprognosen im maschinellen Tunnelbau, Felsbau, 13(6), 439-448.

Geurts, P., Ernst, D., and Wehenkel, L., 2006, Extremely randomized trees, Machine learning, 63, 3-42.

Hucka, V. and Das, B., 1974, Brittleness determination of rocks by different methods, Int. j. rock mech. min. sci. geomech. abstr., 11(10), 389-392.

Ko, T.Y., Kim, T.K., Son, Y., and Jeon, S., 2016, Effect of geomechanical properties on Cerchar Abrasivity Index (CAI) and its application to TBM tunnelling, Tunn. Undergr. Space Technol., 57, 99-111.

Kuhn, M. and Johnson, K., 2013, Applied predictive modeling, New York: Springer, 26, 13.

Macias, F.J. Dahl, F., and Bruland, A., 2016, New rock abrasivity test method for tool life assessments on hard rock tunnel boring: the rolling indentation abrasion test (RIAT), Rock Mech. Rock Eng., 49(5), 1679-1693.

Macias, F.J., 2016, Hard rock tunnel boring: performance predictions and cutter life assessments, Ph. D. Thesis, Norwegian University of Science and Technology.

Majeed, Y. and Abu Bakar, M.Z., 2019, Effects of variation in the particle size of the rock abrasion powder and standard rotational speed on the NTNU/SINTEF abrasion value steel test, Bull. Eng. Geol. Environ., 78, 1537-1554.

Majeed, Y., Abu Bakar, M.Z., and Butt, I.A., 2020, Abrasivity evaluation for wear prediction of button drill bits using geotechnical rock properties, Bull. Eng. Geol. Environ., 79, 767-787.

Massalov, T., Yagiz, S., and Adoko, A.C., 2022, Application of soft computing techniques to estimate cutter life index using mechanical properties of rocks, Appl. Sci., 12(3), 1446.

Meng, F., Wong, L.N.Y., and Zhou, H., 2021, Rock brittleness indices and their applications to different fields of rock engineering: A review, J. Rock Mech. Geotech. Eng., 13(1), 221-247.

Moon, K. and Yang, S.B., 2020, Cohesion and internal friction angle estimated from Brazilian tensile strength and unconfined compressive strength of volcanic rocks in Jeju Island, Journal of the Korean Geotechnical Society, 36(2), 17-28.

Natekin, A. and Knoll, A., 2013, Gradient boosting machines, a tutorial, Frontiers in neurorobotics, 7, 21.

Özfirat, M.K., Yenice, H., Şimşir, F., and Yaralı, O., 2016, A new approach to rock brittleness and its usability at prediction of drillability, J. Afr. Earth Sci., 119, 94-101.

Piratheepan, J., Gnanendran, C.T., and Arulrajah, A., 2012, Determination of c and φ from IDT and unconfined compression testing and numerical analysis, J. Mater. Civ. Eng., 24(9), 1153-1164.

Plinninger, R.J., 2002, Klassifizierung und Prognose von Werkzeugverschleiß bei konventionellen Gebirgslösungsverfahren im Festgestein.

Rostami, J. and Ozdemir, L., 1993, A new model for performance prediction of hard rock TBMs, Proc. Rapid Excav. Tunneling Conf., 793-809.