지질학회지 제62권 제1호, p. 123-133 (2026년 3월)
J. Geol. Soc. Korea, v. 62, no. 1, p. 123-133 (March 2026)
DOI http://dx.doi.org/10.14770/jgsk.2026.003

ISSN 0435-4036 (Print)
ISSN 2288-7377 (Online)

Article

제주분지 근원암 평가를 위한 물리검층 및 측벽코아 자료 활용 기계학습기반 TOC 추정모델 개발

홍요셉[1], 박주영[1], 김영석[1,2], 홍성경[3], 최지영[4], 백세익[5], 전재호[5], 임동현[5], 이경북[1,2,‡]

[1]국립공주대학교 지질환경과학과
[2]국립공주대학교 지구시스템연구소
[3]강원대학교 지질학과
[4]한국지질자원연구원 자원탐사개발연구본부 석유·미래에너지연구센터
[5]한국석유공사 서남해탐사팀

Development of machine learning-based TOC estimation model using well logging and sidewall core data for source rock evaluation in the Jeju Basin, South Sea of Korea

Yosep Hong[1], Ju Young Park[1], Youngseuk Keehm[1,2], Sungkyung Hong[3], Jiyoung Choi[4], Seik Paik[5], Jae-Ho Jeon[5], Donghyun Lim[5], Kyungbook Lee[1,2,‡]

[1]Department of Geoenvironmental Sciences, Kongju National University, Gongju 32588, Republic of Korea
[2]Institute of Earth system Research, Kongju National University, Gongju 32588, Republic of Korea
[3]Department of Geology, Kangwon National University, Chuncheon 24341, Republic of Korea
[4]Petroleum and Future Energy Research Center, esource Exploration and Development Research Division, Korea Institute of Geoscience and Mineral Resources, Daejeon 34132, Republic of Korea
[5]West & South Sea Exploration Team, Korea National Oil Corporation, Ulsan 44538, Republic of Korea

Received: December 14, 2025 / Revised: January 22, 2026 / Accepted: January 22, 2026
‡Corresponding author: +82-41-850-8511 / E-mail: kblee@kongju.ac.kr

요약: 본 연구에서는 효율적인 근원암 평가를 위해 물리검층자료를 활용한 기계학습기반 총유기탄소량(total organic carbon, TOC) 추정모델을 개발하였다. 해당 모델은 코아시료가 없는 구간에서도 TOC를 연속적으로 추정할 수 있는 장점이 있다. 가용자료는 제주분지 O, J3, J5 시추공의 물리검층자료와 측벽코아 대상 TOC 분석결과이다. 먼저 물리검층-TOC간 자료쌍 매칭을 수행하였으며, 가용자료 수를 고려해 입력인자(밀도, 음파, 감마선 검층)를 선정하였다. 감마선검층은 시추공단위 표준정규화를 수행하였으며, 113개 자료를 훈련, 5개를 테스트 자료로 사용하였다. 랜덤포레스트와 극한경사부스팅의 성능을 비교한 결과, 극한경사부스팅은 테스트자료 기준 결정계수(R^2) 0.84, 평균절대오차 0.09 wt.%로, 랜덤포레스트의 R^2(0.11)와 평균절대오차(0.19 wt.%) 대비 준수한 성능을 보였다. 개발된 모델을 O 시추공 내 코아가 부재한 구간에 적용한 결과, 1 wt.% 이상의 TOC 구간에서는 과소평가되었다. 향후 자료증강 또는 암편시료를 활용해 높은 TOC 구간에 대한 성능개선연구가 필요하다.

주요어: 제주분지, 총유기탄소량, 기계학습, 측벽코아, 물리검층

ABSTRACT: In this study, a machine learning-based model was developed to estimate total organic carbon (TOC) from well logging data for enhanced source rock evaluation. The model enables estimation of a continuous TOC curve along well logs, even in core-limited intervals. The dataset consisted of well logging data and TOC analyses of sidewall core samples obtained from the O, J3, and J5 wells in the Jeju Basin. After pairing well logging data with corresponding TOC values, input features were selected considering the number of available data pairs. Subsequently, a well-to-well standard normalization was performed to account for inter-well variability of the gamma-ray log responses. A total of 118 datasets were divided into 113 for train and 5 for test data. A performance comparison between random forest (RF) and extreme gradient boosting (XGBoost) models revealed that XGBoost demonstrated superior performance. Specifically, on the test dataset, XGBoost achieved a coefficient of determination (R^2) of 0.84 and a mean absolute error (MAE) of 0.09 wt.%, significantly outperforming RF (R^2 = 0.11, MAE = 0.19 wt.%). Application of the developed model to intervals within the O well where core data were unavailable revealed an underestimation in sections with TOC exceeding 1 wt.%. Future improvements within high-TOC intervals can be achieved through data augmentation or training with TOC experimental data using cutting samples.

Key words: Jeju basin, total organic carbon, machine learning, sidewall core, well logging

1. 서 론

유기물이 풍부한 근원암의 존재는 석유시스템이 성립되기 위한 가장 기본적인 전제조건이다. 유효한 근원암이 없다면 탄화수소의 생성이 불가능하므로, 퇴적분지 내 근원암의 탄화수소 생성가능성을 평가하는 것은 석유탐사의 핵심적인 과제이다(Magoon and Dow, 1994). 근원암 평가의 가장 기본적인 지표는 총유기탄소량(total organic carbon, TOC)으로, 이는 유기물의 함량과 탄화수소 생성가능성을 직접적으로 지시한다(Peters and Cassa, 1994).

TOC는 직접분석법 또는 간접추정법으로 평가할 수 있다. 직접분석법의 대표적 방법인 Rock-Eval 열분석법은 분석 정확도가 높지만, 특정 심도에서만 불연속적인 정보를 제공하는 한계가 존재한다. 이는 코아시료 확보에 어려움이 있을 뿐만 아니라, 실험비용이 크기 때문이다. 간접추정법으로는 물리검층자료를 활용한 경험식이 있다. Schmoker (1979, 1981)는 밀도(density)검층의 역수와 TOC의 선형관계를 제안하였고, Passey *et al*. (1990)은 전기비저항(resistivity)과 음파(sonic) 검층을 이용해 TOC를 추정하는 $\Delta \log R$ 기법을 제안하였다. 해당 방법들은 지역적 특성에 따라 신뢰도가 상이하며, 경험식에 분석자가 결정해야하는 인자가 포함되어 분석자의 주관이 개입될 수 있다.

기존 TOC 평가기법의 한계를 보완하기 위해, 최근 기계학습(machine learning)을 활용해 물리검층자료로 TOC를 추정하는 연구가 활발히 진행되고 있다(Lai *et al*., 2024). 이는 물리검층의 연속성과 코아실험의 높은 신뢰도라는 장점을 결합한 방법이다. 구체적으로, Sun *et al*. (2023)은 중국 세일 분지(Ordos, Bohai Bay, Sichuan, Southern North China Basin)에서 랜덤포레스트(random forest, RF), 서포트벡터머신(support vector machine) 및 극한경사부스팅(extreme gradient boosting, XGBoost)모델로 TOC를 추정한 결과, RF가 가장 우수한 TOC 추정성능을 보임을 확인하였다. Salaim *et al*. (2024)은 수단 Muglad 분지에서 인공신경망(artificial neural network)모델을 적용하여 Passey 추정법보다 우수한 TOC 추정결과를 보고하였다.

이처럼 기계학습기법의 유용성은 여러 퇴적분지에서 입증되었으나, 각 모델의 성능은 연구대상 지역의 지질학적 특성과 자료의 질에 의존하므로, 특정 연구지역에 적합한 TOC 추정모델을 개발하는 것이 필요하다(Han *et al*., 2022).

우리나라 남해대륙붕에 위치한 제주분지의 경우, 유·가스를 생산하고 있는 동중국해 시후(Xihu)분지와 지질학적

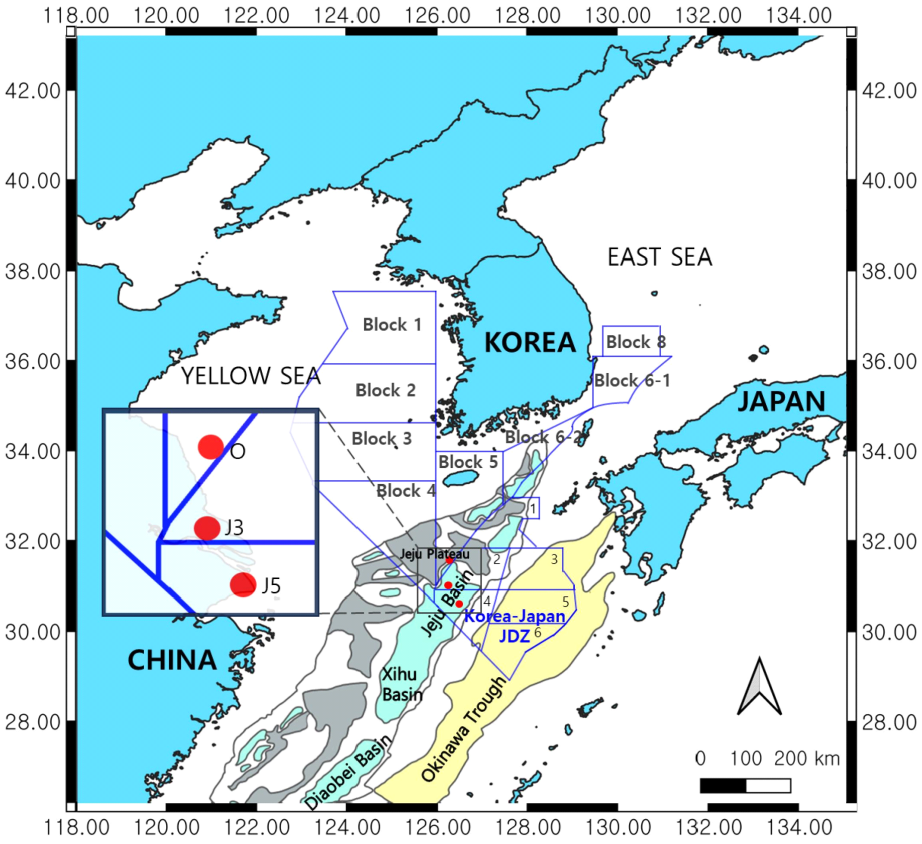

Fig. 1. Exploration blocks and well locations in the Jeju basin, Republic of Korea.

Table 1. Summary of logging and TOC from sidewall core for the three wells (O, J3, and J5 well).

Data	Parameter	O well	J3 well	J5 well
Logging	Depth	474.8-2900.2 m	450.04-3179.9 m	456.2-4181.7 m
	Type	Density, DTC, Gamma-ray, Neutron, RDEEP		
Sidewall core	Depth	1403.2-2822.6 m	1376.8-3062.6 m	805.5-1368 m
	Number of data	44	44	30

Table 2. Statistical summary of logging data and TOC for each well.

Well	Statistics	Density (g/cm^3)	DTC (us/ft)	GR (gAPI)	Neutron (m^3/m^3)	RDEEP (ohm.m)	TOC (wt.%)
O	Mean	2.24	119.17	61.34	0.34	14.24	0.88
	STD	0.14	24.73	16.61	0.08	143.67	1.18
	Range	1.28-2.82	33.06-198.95	22.34-122.93	0.02-0.94	0.55-2000.00	0.19-6.31
J3	Mean	2.29	109.13	70.54	0.26	2.55	0.44
	STD	0.18	25.41	20.77	0.08	2.22	0.34
	Range	1.45-2.86	53.37-191.07	12.96-136.32	0.07-0.59	0.77-19.09	0.08-2.02
J5	Mean	2.36	96.94	62.46	0.22	5.83	0.32
	STD	0.31	28.47	23.37	0.10	3.90	0.25
	Range	1.43-2.96	58.73-185.43	13.43-154.87	0.03-0.57	0.44-37.71	0.04-1.05

으로 연결성이 존재해 유·가스 부존가능성이 제기되어왔다 (Lee *et al*., 2019; Qu *et al*., 2023). 이에 제주분지 근원암 평가의 일환으로, 시추암편시료 대상 Rock-Eval 열분석법을 통해 근원암 특성을 평가하는 연구가 보고되었다(Lee *et al*., 1998). 기존 연구에서는 시추공별로 100개 내외 지점을 선별하여 분석을 수행하였지만, 직접분석법 특성상 TOC의 연속적인 변화 파악이 불가능하다, 또한 시추코아의 부재로 인해 시추암편시료를 사용함으로써 3-10 m 내외의 심도 불확실성이 존재한다. 즉, 상·하부 지층의 암편 혼입 문제로 인해 전반적인 근원암 특성 경향성만 파악이 가능하다.

본 연구에서는 제주분지를 대상으로 기계학습기법을 활용해 물리검층 해상도의 TOC 추정모델을 개발하고자 한다. 특히, 기존 연구들은 주로 단일 시추공자료 또는 동일 지층을 대상으로 TOC 추정모델을 개발한 것에 비해, 본 연구에서는 다수의 시추공과 시추공별로 다양한 심도에 산재된 TOC 분석결과를 통합하여 제주분지에 적합한 TOC 추정모델을 개발하는 것이 특징이다. 이때 지질학적 특성을 고려한 자료 전처리 과정을 수행한 후, 기존 연구에서 성능이 입증된 RF와 XGBoost 모델의 성능을 비교하였다.

2. 연구방법

2.1. 자료개요

연구지역인 제주분지는 수심이 70-150 m로 비교적 얕은 대륙붕 지역이며, 남동 방향으로 갈수록 점차 깊어지는 특징을 지닌다(Kwon, 1996). 지리적으로 동중국해대륙붕의 북동쪽 경계부에 위치하며 국내대륙붕 5광구와 한일공동개발구역(Korea-Japan Joint Development Zone, JDZ)을 포함한다(Kim and Son, 2013). 본 연구에서는 제주분지 5광구에서 획득된 두 개 시추공(O, J3)과 JDZ에서 획득된 한 개 시추공(J5)에 대한 물리검층자료와 측벽코아(sidewall core)를 통해 획득된 TOC 분석결과를 활용하였다(그림 1).

기계학습기반 TOC 추정모델의 입력자료는 세 개 시추공에서 공통적으로 취득된 밀도, 음파, 감마선, 중성자 및 심부전기비저항 검층자료이며, 출력자료는 측벽코아의 TOC 분석결과이다. 표 1은 각 시추공의 물리검층 및 코아시료 취득구간을 나타낸 것으로, 각 시추공별 물리검층이 취득된 구간이 상이한 것을 확인할 수 있다. O 시추공의 경우 약 2900 m까지 검층이 취득되었으나, J5의 경우 약 4200 m까지 취득되었다. 또한, 측벽코아 취득심도의 범위는 O와 J3 시추공의 경우 약 1400 m 이상으로 넓지만, J5의 경우 약 500 m 정도의 비교적 좁은 구간에서 측벽코아가 취득되었고, 그 개수도 30개로 타 시추공 대비 적다.

표 2는 세 시추공에 대한 물리검층과 TOC의 통계분석 결과로, 뚜렷한 분포차이를 보인다. 구체적으로, O 시추공은 심부전기비저항검층 최댓값이 2000 ohm.m인 반면, 다른 두 시추공은 약 38 ohm.m 미만의 낮은 값을 갖는다. 또한 평균 음파검층 값은 J5 시추공(96.94 us/ft)이 O 시추공(119.17 us/ft)보다 현저히 낮게 나타나는 등 자료의 이질성

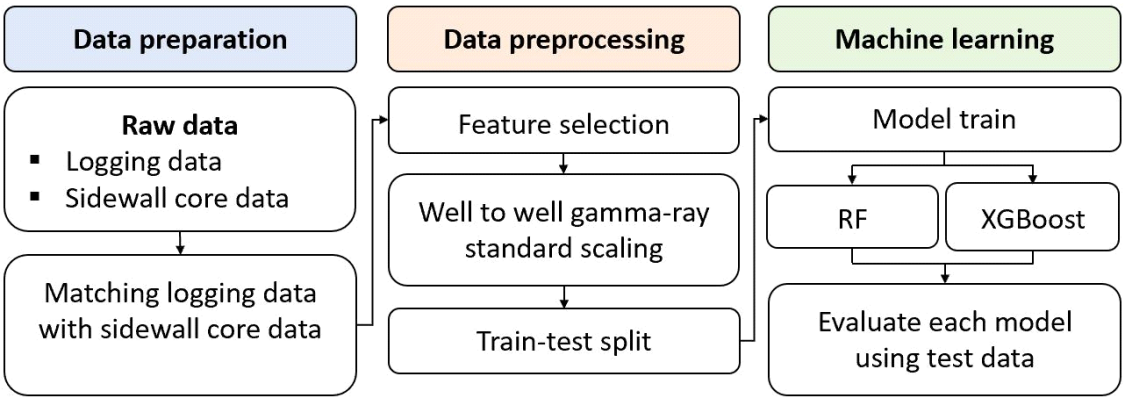

Fig. 2. Workflow of TOC estimation using well logging data based on machine learning.

이 존재한다. TOC 분포 또한 O 시추공은 평균 0.88 wt.%, 최대 6.31 wt.%로 타 시추공에 비해 높은 수치를 보인다.

2.2. 가용자료 전처리

본 연구의 기계학습모델 개발과정은 그림 2와 같으며, 다음의 네 단계를 통해 자료쌍(입력자료-출력자료) 매칭과 전처리가 수행되었다. 첫 번째 단계는 측벽코아의 심도를 기준으로 물리검층과 자료쌍을 매칭하는 것이다(그림 2 좌측 두 번째 과정). 구체적으로, 측벽코아 심도에 대해 가장 근접한 물리검층 측정심도를 대응시켜 입·출력 자료를 매칭하였다. 타당성을 검증하기 위해 매칭한 두 자료간의 심도 차이를 분석한 결과, 모든 시추공에서 최대 심도오차가 8 cm 이내로 나타나, 물리검층 자료간격(15 cm)을 고려했을 때 자료매칭이 합리적으로 수행되었음을 확인하였다(그림 3).

두 번째 단계는 자료전처리 중 특징추출(feature selection) 과정이다(그림 2 중간 첫 번째 과정). 그림 4를 보면 J3 시추공은 3000 m 하부구간에서 심부전기비저항검층이, J5 시추공은 1300 m 상부구간에서 중성자 및 심부전기비저항

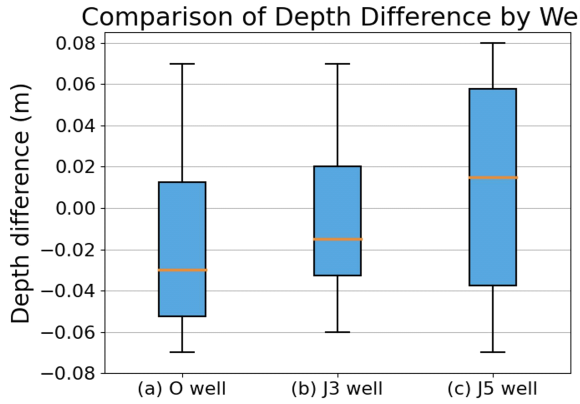

Fig. 3. Boxplot for depth difference between TOC data and corresponding logging data for each well: (a) O, (b) J3, and (c) J5 well.

검층이 부재하여 TOC 분석결과와 매칭되지 않음을 알 수 있다. 이로 인해 심부전기비저항검층을 입력변수로 고려할 경우, J3 및 J5 시추공에서 각각 23개, 1개의 TOC 자료만을 활용할 수 있어 가용자료의 수가 118개에서 68개로 대폭 감소한다. 따라서 최대한 많은 자료를 확보하기 위해 밀도, 음파 및 감마선 검층을 최종 입력변수로 선정하였다. 해당 검층자료를 입력변수로 사용할 경우, 측벽코아시료에 대한 모든 TOC 분석결과를 가용할 수 있다.

세 번째 단계는 검층자료의 특성을 고려하여 전처리하는 것이다(그림 2 중간 두 번째 과정). 감마선검층은 검층장비의 방사선 민감도차이, 시추공의 크기 및 이수로 발생하는 감쇠 효과로 인해 시추공별 분포가 달라질 수 있다(Shier, 2004). 이러한 시추공간 편차를 보정하기 위해 개별 시추공 단위(well-to-well) 표준정규화(standard normalization)를 수행하였다(식 1).

$$GR(SS)_{i,n} = \frac{GR_{i,n} - \mu_i}{\sigma_i} \quad (1)$$

여기서, $GR_{i,n}$는 i 시추공의 n번째 감마선 검층값, $GR(SS)_{i,n}$은 개별시추공단위로 표준화된 감마선 검층값을 나타내며, μ_i과 σ_i는 각각 i 시추공 감마선검층의 평균과 표준편차를 의미한다.

마지막 네 번째 단계에서는 정규화된 118개 자료를 훈련 및 테스트 자료로 분할하였다(그림 2 중간 세 번째 과정). TOC는 퇴적환경의 영향을 받아 특정 층서구간에 높은 값이 편중되어 분포하는 경향이 있어 편향을 방지하기 위한 적절한 자료분할방법이 요구된다. 자료분할 시 일반적으로 사용되는 무작위 분할 또는 k-fold 교차검증은 검증자료 또는 테스트자료에 특정 층서의 자료가 편향될 가능성이 있다. 특히 k-fold 교차검증의 경우, 훈련자료의 일부를 검증자료로 활용하므로 가용자료 수가 감소해 추정성능이 저하

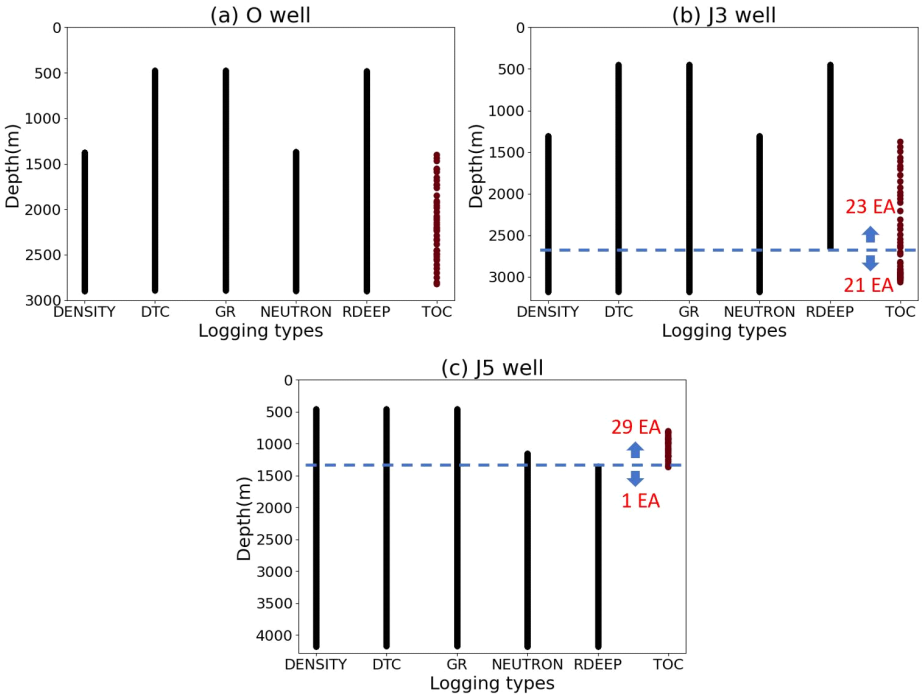

Fig. 4. Data availability of well logs and TOC for each well: (a) O, (b) J3, and (c) J5 well.

될 수 있다. 본 연구에서는 이를 고려하여 층서기반 자료분할을 적용하였다. 구체적으로, O 시추공의 각 층서마다 테스트자료를 1개씩 선별하였다. 해당 방식을 적용할 경우, O 시추공을 중심으로 추정성능이 검증되므로 J3 및 J5 시추공에 대한 추정성능과 불확실성은 충분히 평가되지 못하는 한계가 존재한다. 하지만 테스트자료의 편향을 방지하고 훈련자료를 확보할 수 있다는 점에서 합리적인 자료분할 방법이라고 판단된다. 결과적으로, 전체 118개 중 113개를 훈련자료로, 5개를 테스트자료로 분할하였다(표 3).

2.3. TOC 추정모델 개발

앞서 전처리된 113개 훈련자료는 기계학습모델 개발에 활용된다(그림 2 우측 첫 번째 과정). 기계학습 알고리즘 선정에 앞서 전처리된 자료의 상관성을 분석하였다(그림 5). 출력인자인 TOC는 밀도검층과 0.1로 약한 양의 상관관계를 보였으나, 감마선검층과 0.06, 음파검층과는 -0.03의 상관관계를 보여 뚜렷한 상관성이 나타나지 않았다. 이는 시추공별 상이한 퇴적환경과 개별시추공의 다양한 층서로 인한 자료의 이질성으로 해석된다. 이에 따라, 변수간 복잡한 비선형 및 상호작용 관계를 효과적으로 반영할 수 있고, 낮은 선형상관성에도 비교적 우수한 성능을 보이는 트리기반 앙상블모델을 기계학습 알고리즘으로 선정하였다.

본 연구에서는 대표적인 트리기반 앙상블모델인 RF와 XGBoost를 활용하여 TOC 추정모델을 구축하였다(그림 2

Table 3. Train and test data split based on stratigraphic units.

Stratigraphy	O well		J3 well	J5 well
	Train	Test	Train	Train
Late Pliocene	-	-	-	1
Early Pliocene	-	-	-	13
Middle Miocene	2	1	11	16
Early Miocene	8	1	5	-
Late Oligocene	10	1	12	-
Early Oligocene	12	1	19	-
Late Eocene	7	1	-	-

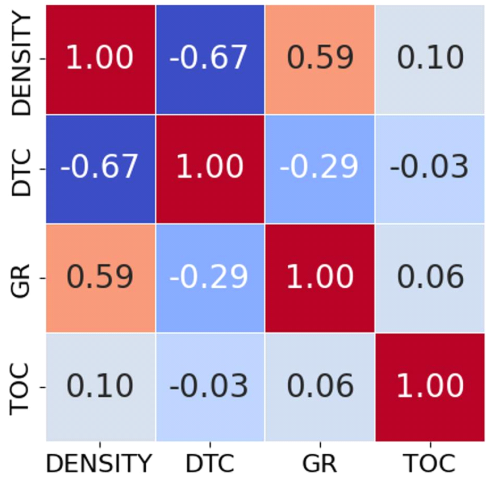

Fig. 5. Pearson correlation matrix among well logs and TOC.

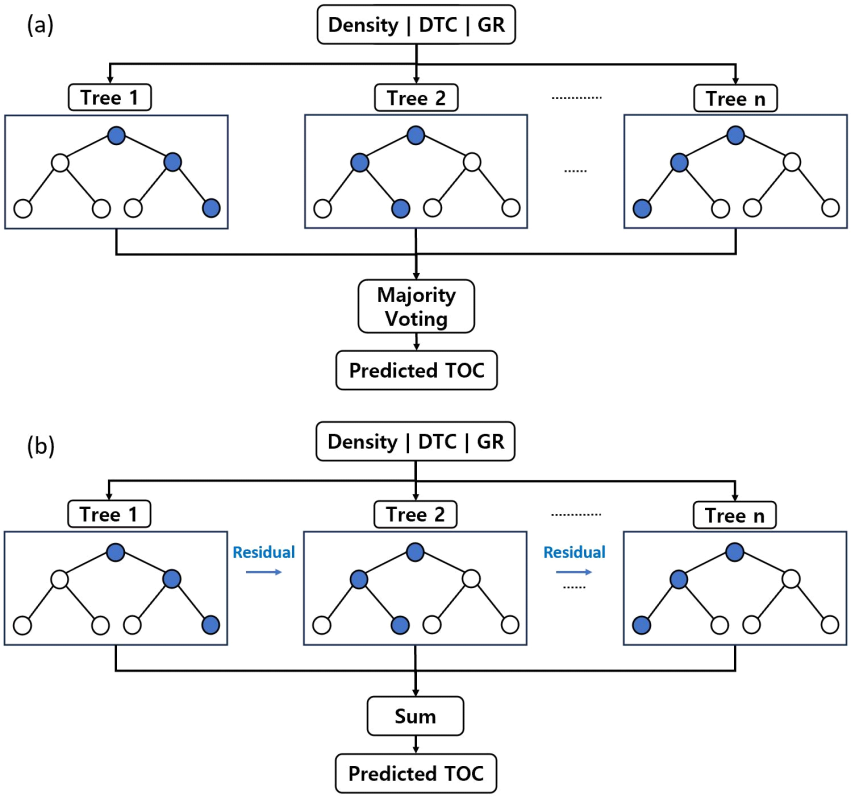

Fig. 6. Structure of developed TOC estimation model (modified from Yeom *et al*., 2023): (a) RF and (b) XGBoost.

우측 두 번째 과정). 두 알고리즘은 각각 병렬적 학습을 통한 안정성과 순차적 학습을 통한 정확도 향상이라는 특징을 가진다.

RF는 Breiman (2001)에 의해 제안되었으며, 여러 개의 의사결정트리(decision tree)를 독립적으로 학습시킨 후 각 트리의 추정결과에 평균을 취하거나 다수결로 종합하는 앙상블모델이다(그림 6a). RF는 자료와 변수를 무작위로 샘플링하여 개별 트리의 상관관계를 줄임으로써 과적합에 강하고 안정적인 성능을 보이는 장점이 있다.

XGBoost는 Chen and Guestrin (2016)에 의해 제안되었으며, 이전 트리의 추정오차(residual)를 다음 트리가 순차적으로 보완하며 학습을 진행하는 모델이다(그림 6b). 잔여오차를 효과적으로 줄여나가는 방식으로 학습하여 높은 정확도를 보이지만, RF에 비해 과적합 위험이 비교적 높다는 한계가 있다.

RF와 XGBoost 모델의 추정성능을 극대화하기 위해 격자탐색(grid search)기법을 이용하여 주요 하이퍼파라미터(hyper-parameter)들을 최적화하였다. 격자탐색은 사전에 정의된 하이퍼파라미터들의 모든 조합을 탐색하여 가장 우수한 성능을 보인 조합을 최종파라미터로 선정하는 방식이다. 본 연구에서는 RF모델에 대해 트리개수(n_estimators), 최대깊이(max_depth), 최소분할샘플수(min_samples_split), 리프노드최소샘플수(min_samples_leaf) 및 최대변수개수(max_features)를 최적화하였으며, XGBoost모델에 대해서는 트리개수, 최대깊이, 학습률(learning_rate), 부스팅단계별 샘플링비율(subsample) 및 트리별변수샘플링비율(colsample_bytree)을 최적화하였다(표 4).

Table 4. Optimized hyper-parameter using grid search.

RF		XGBoost	
n_estimator	150	n_estimator	5
max_depth	20	max_depth	10
min_samples_split	9	learning_rate	0.3
min_samples_leaf	9	subsample	0.8
max_features	1.0	colsample_bytree	1.0

3. 연구결과

3.1. 알고리즘에 따른 TOC 추정성능 비교

최적화된 두 모델의 성능을 정량적으로 비교하기 위해 훈련자료와 테스트자료에 대한 결정계수(coefficient of determination, R^2), 평균제곱오차(mean squared error, MSE), 평균절대오차(mean absolute error, MAE)를 계산하였다(식 2-4; 표 5).

Table 5. Performance metrics of RF and XGBoost models on train and test datasets.

Model	Data type	R^2	MSE	MAE
RF	Train	0.549	0.285	0.261
	Test	0.107	0.058	0.186
XGBoost	Train	0.883	0.078	0.108
	Test	0.835	0.011	0.088

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (2)$$

여기서, y_i와 $\hat{y}_i$는 각각 실제값과 추정값을 나타내며, $\bar{y}_i$는 실제값의 평균을 의미한다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

훈련자료에 대하여 XGBoost모델의 R^2는 0.883으로, RF 모델(0.549) 대비 높은 설명력을 보였다. MSE와 MAE 또한 XGBoost모델이 RF모델에 비해 낮은 값을 보여 전반적인 추정오차가 작음을 확인하였다. 테스트자료에서는 두 모델간 성능차이가 더욱 뚜렷하게 나타났다. XGBoost모델은 R^2 0.835, MSE 0.011로 매우 우수한 일반화 성능을 보이며, 과적합이 발생하지 않았음을 확인하였다. 반면, RF 모델은 R^2 0.107, MSE 0.058로 낮은 추정성능을 보였다. 이는 학습에 사용되지 않은 테스트자료에 대해 XGBoost 모델이 훨씬 더 안정적이고 신뢰도 높은 추정을 수행하였음을 의미한다. 이러한 성능차이는 각 모델의 학습방식 차이에서 기인한다.

본 연구의 훈련자료는 세 개 시추공(O, J3 및 J5)으로 구성되어 있어 시추공별 이질성이 존재한다. 이때 RF모델의 트리들은 각 시추공의 자료를 기반으로 서로 다른 규칙을 학습하게 되며, 일부 트리는 O 시추공보다 타 시추공의 특성에 더 적합한 형태로 형성된다. 따라서 O 시추공으로만 구성된 테스트자료에서는 RF모델의 일부 트리만 정상적으로 작동해, 다수결 원리로 최종 추정값을 도출하는 RF의 성능이 저하된 것으로 사료된다.

반면 XGBoost는 트리를 순차적으로 학습하기 때문에 각 시추공의 개별특성보다 전체 자료에서 공통적으로 나타나는 패턴을 학습한다. 즉, 시추공의 차이를 그대로 반영하기보다 세 시추공에 적용가능한 일반적인 관계를 우선적으로 포착해 테스트자료에서도 안정적이고 일관된 추정을 수행한 것으로 해석된다.

그림 7과 8은 두 모델의 추정 경향성을 상세히 분석하기 위해 실제 TOC와 모델 추정값을 산점도로 나타낸 것이다. 대부분의 훈련자료가 분포하는 1.5 wt.% 미만 구간에서는 XGBoost모델의 추정값이 y=x 직선 주변에 밀집하여 분포하는 준수한 추정성능을 보였다(그림 8a). 반면, RF모델의 추정값은 상대적으로 분산과 오차가 크게 나타났다(그림 7a). 이러한 경향은 테스트자료에서 더욱 명확해져, XGBoost 모델이 RF모델에 비해 TOC 범위 전반에 걸쳐 일관되고 신뢰도 높은 추정을 수행하였음을 확인할 수 있다(그림 7b, 8b).

3.2. 경험식을 활용한 TOC 추정성능 검증

기계학습모델의 성능을 검증하기 위해 Schmoker 경험

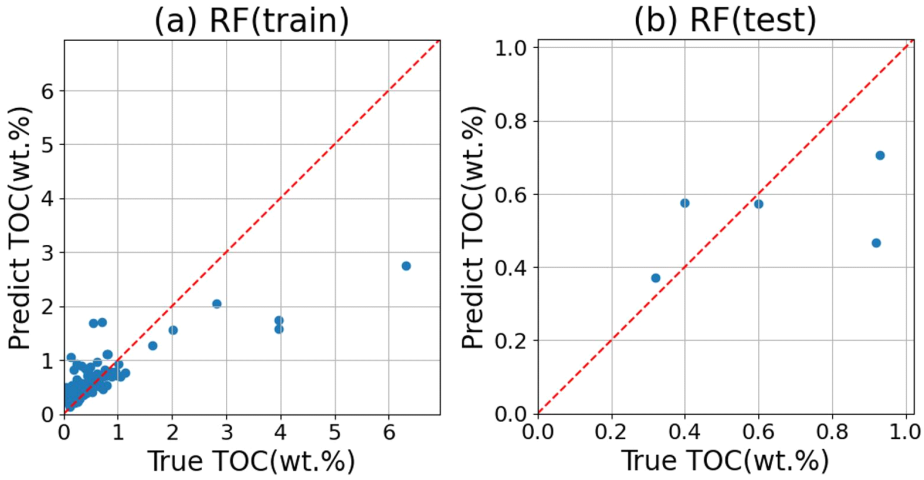

Fig. 7. Scatter plot between estimated TOC using RF model and true TOC: (a) Train and (b) Test datasets.

식(식 5)과 추정결과를 비교하였다.

$$TOC[wt.\%] = (A \times \frac{1}{\rho}) - B \quad (5)$$

여기서, ρ는 밀도검층을 나타내며, A와 B는 밀도검층의 역수와 TOC간 선형회귀를 통해 계산되는 상수이다.

본 연구에서는 훈련자료를 활용하여 상수 A(-1.9)와 B(1.45)를 도출하였다(그림 9a). 경험식을 테스트자료에 적용한 결과, R^2는 -0.145, MSE는 0.074로 기계학습모델 대비 저조한 성능을 보였다(그림 9b). 이는 연구지역(신생대 제일층)의 지질·지화학적 특성이 경험식이 개발된 지역(데본기 제일층)의 특성과 상이하여 나타난 결과로 해석된다.

구체적으로, 연구지역은 총 황 함량이 최대 10,000 ppm 이며(Kim and Lee, 1999), 이는 황철석을 포함한 다량의 황화광물이 존재했을 가능성을 시사한다. 고밀도(약 5 g/cm³) 특성을 보이는 황철석은 환원환경에서 유기물과 함께 형성되므로(Jiang *et al*., 2018; Zhou *et al*., 2024), TOC와 밀도검층값이 양의 상관관계를 보인다(그림 9a). 그러나, Schmoker 경험식은 데본기 제일층 특성에 따라 TOC와 밀도검층간 음의 상관관계를 가정하므로, 본 연구지역의 지질학적 특성과 경험식의 기본 가정이 부합하지 않음을 알 수 있다. 결론적으로, 본 연구지역은 경험식 보다는 기계학습 기반으로 TOC를 추정하는 것이 합리적이며, 밀도, 음파 및 감마선 검층을 종합적으로 고려하여 TOC를 추정하는 것이 적합하다고 판단된다.

3.3. 각 모델별 입력변수 영향력 분석

TOC 추정모델 개발과정에서 각 입력변수의 상대적 영향력을 평가하기 위해 트리기반모델의 변수중요도(feature

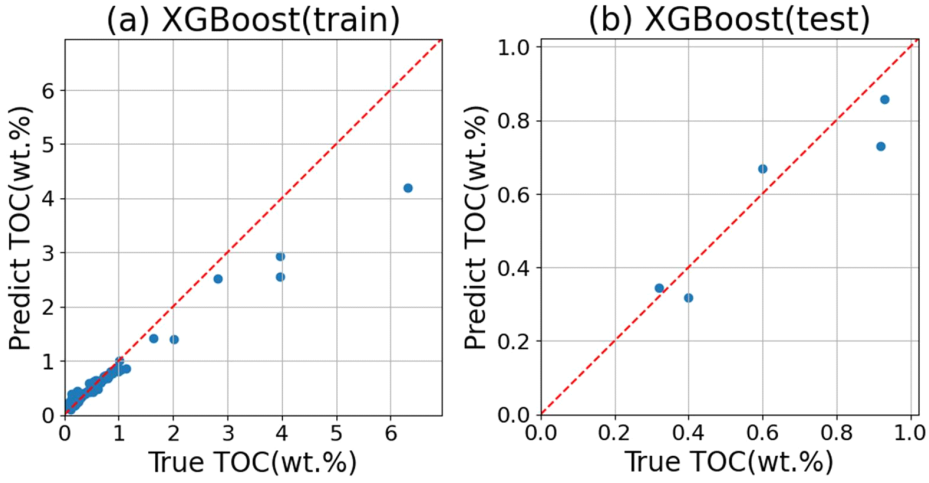

Fig. 8. Scatter plot between estimated TOC using XGBoost model and true TOC: (a) Train and (b) Test datasets.

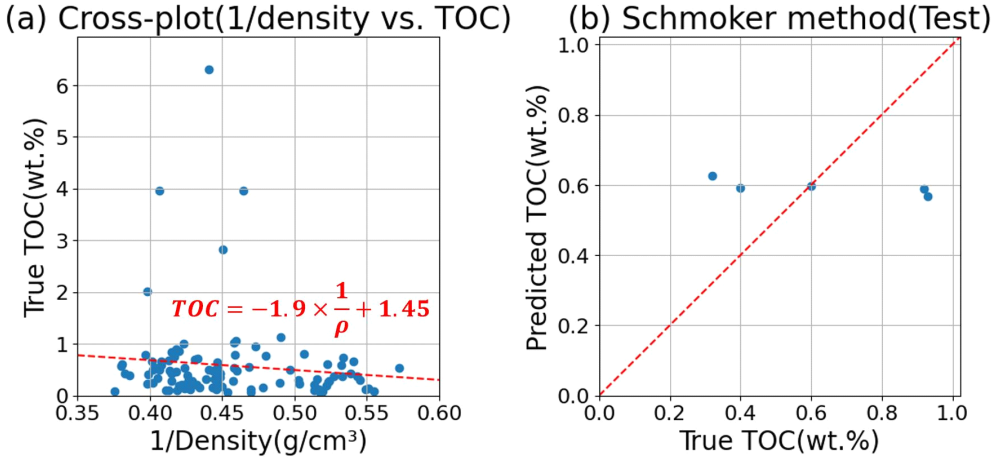

Fig. 9. TOC estimation results using Schmoker method: (a) Cross-plot of 1/density vs. true TOC and (b) estimated vs. true TOC for Test dataset.

importance)를 분석하였다(그림 10). RF모델은 감마선검층의 중요도가 0.5로 가장 높게 나타났다. 이는 연구지역의 근원암이 주로 유기물이 풍부한 셰일층에 해당하며, 이러한 셰일층은 방사성 원소인 우라늄을 흡착하는 경향이 있어 높은 감마선 값을 보이기 때문으로 해석된다. 또한 감마선검층은 개별 시추공단위 표준화가 수행되어, 세 시추공에서 평균 0, 표준편차 1에 근사하도록 자료분포를 변형시켰다. 즉, 감마선검층은 세 시추공에서 유사한 범위를 가져

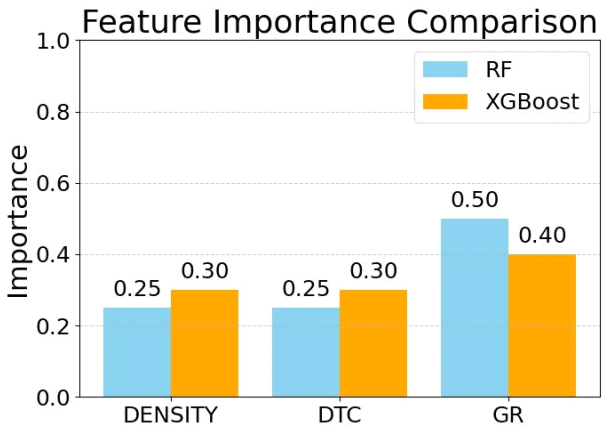

Fig. 10. Comparison of feature importance between RF and XGBoost models.

학습 시 큰 이질성이 나타나지 않았으며, 따라서 RF모델은 감마선검층을 중심으로 학습하였다고 볼 수 있다. 결론적으로, RF모델은 감마선검층에 지나치게 의존성이 높아 테스트자료와 같은 신규자료에 대해서는 추정성능이 낮은 것으로 판단된다.

반면, XGBoost모델은 밀도와 음파 검층의 중요도가 RF모델 대비 증가하였다. 이는 감마선검층을 우선적으로 활용하여 층서적·지질학적 특성을 파악하고, 감마선만으로 설명되지 않는 TOC 변화양상을 밀도와 음파 검층이 보완적으로 설명하도록 학습되었기 때문이다. 따라서 세 검층 모두가 TOC 추정에 균형있게 기여함으로써, 테스트자료에도 과적합없이 우수한 성능을 보인 것으로 판단된다.

3.4. 시추암편시료 활용 TOC 추정성능 분석

개발된 모델 중 우수한 성능을 보인 XGBoost모델의 실용성을 검증하기 위해 O 시추공 1600-1700 m와 2600-2700 m 구간의 물리검층자료에 적용한 후 시추암편시료 대상 TOC 분석결과와 비교하였다(그림 11). 그 결과, 1600-1700 m 구간에서는 시추암편시료로 추정된 TOC와 모델의 TOC 추정 경향성이 유사하게 나타나는 것을 확인하였다(그림 11a). 2600-2700 m 구간에서는 시추암편시료 TOC 대비 모델의 추정 TOC가 전반적으로 과소평가되는 것을 확인하였다

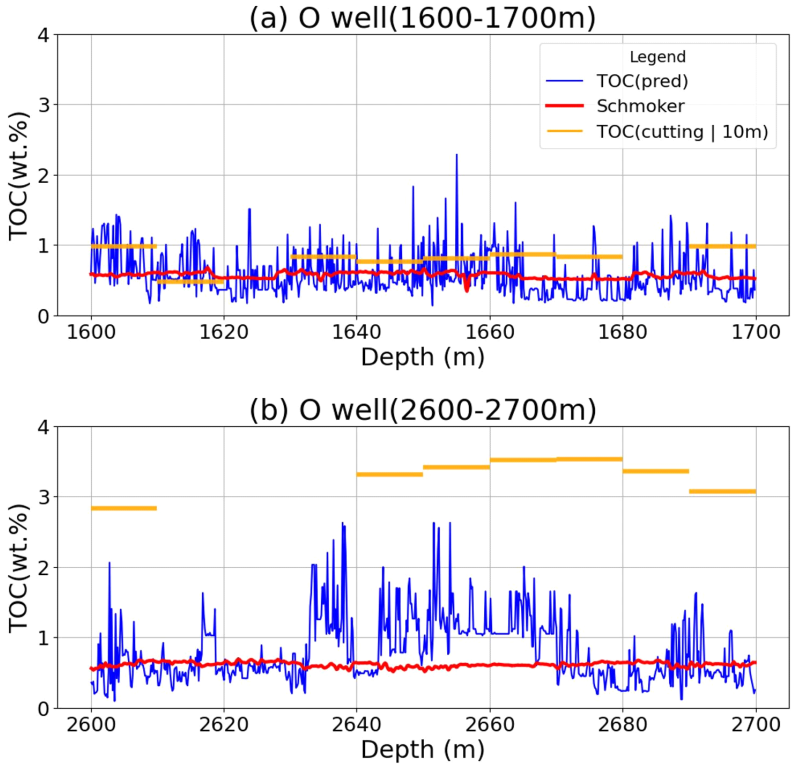

Fig. 11. Estimated TOC curve from XGBoost (blue) compared with TOC from Schmoker method (red) and drilling cutting (orange): (a) 1600-1700 m and (b) 2600-2700 m for O well.

(그림 11b). 2600-2700 m 구간에서의 저조한 성능은 외삽에 취약한 트리기반모델의 특성에 기인한 것으로 해석된다. 본 연구의 훈련자료는 대부분 1 wt.% 이하인 TOC 값에 집중되어 있어 1 wt.% 이상 구간의 자료 수가 제한적이다. 이러한 편중으로 인해 모델이 높은 TOC 구간에서의 물리검층 패턴과 TOC간 비선형적 관계를 충분히 학습하지 못하였으며, 이에 학습범위를 벗어난 구간에서는 과소추정하는 경향을 보인다. 따라서 그림 11a의 1 wt.% 미만 구간에서는 모델이 안정적으로 TOC를 추정하지만, 그림 11b의 3 wt.%에 해당하는 구간에서는 추정값이 과소평가되어 상대적으로 낮은 추정성능을 보이는 것으로 판단된다.

나아가 지질학적 관점에서, 1600-1700 m 구간과 같이 1 wt.% 미만으로 나타나는 대부분의 시료는 유기물 기원이 Type III에 해당하며, 2600-2700 m 구간과 같이 1 wt.% 이상인 일부 시료는 Type II로 분류된다. Type III에 해당하는 시료는 훈련자료의 대부분을 구성하므로 모델이 충분히 학습되어 높은 정확도를 보이는 반면, Type II는 훈련자료가 제한적이어서 상대적으로 낮은 성능을 보인 것으로 판단된다.

Schmoker 경험식은 두 구간 모두에서 약 0.6 wt.%로 일정하게 TOC를 추정하여 암편시료의 TOC 변화양상을 파악하는데 한계가 있음을 볼 수 있다(그림 11). 이는 3.2에서 언급한 바와 같이, 연구지역의 지질·지화학적 특성이 경험식의 기본 가정과 부합하지 않기 때문이다. 결론적으로, XGBoost모델은 TOC 변화양상을 경험식보다 합리적으로 반영하는 것으로 해석된다.

이와 더불어, 그림 11에서 제시된 TOC 실제값은 시추암편시료 내 셰일만을 선별하여 분석한 결과로, 모델학습에 사용된 측벽코아 TOC 분석결과와 시료 채취 방식, 실험 절차 및 장비 측면에서 차이가 존재한다. 이러한 분석환경의 상이성은 두 자료의 대비에 있어 일정수준의 불확실성을 야기할 가능성이 있다.

4. 결 론

본 연구에서는 제주분지 내 세 개 시추공의 물리검층 및 측벽코아시료를 활용하여 TOC를 추정하는 기계학습모델을 구축하고 그 성능을 비교하였다. 이를 통해 다음과 같은 결과를 얻을 수 있었다.

본 연구에서는 측벽코아와 물리검층자료를 가장 근접한 측정심도로 매칭하여 신뢰도 높은 자료쌍을 구축하였다. 그 결과, 모든 시추공에서 최대 심도오차가 8 cm 이내로 나타나 자료매칭의 높은 신뢰도를 확인하였다. 입력변수는 모든 시추공에서 확보가능한 밀도, 음파, 감마선 검층을 선정하여 자료 가용성을 극대화하였다.

테스트자료에 대해 XGBoost모델이 RF모델 대비 R^2가 681% 상승하였고, MSE 81%, MAE 53% 감소하였다. 이는 이질적인 다중 시추공 자료를 통합하여 학습하는 과정에서 XGBoost가 변수간 비선형 관계와 시추공간 편차를 효과적으로 학습했기 때문이다. 특히 XGBoost는 감마선, 밀도 및 음파 검층의 특성을 균형있게 반영하여 감마선검층에 의존적인 RF모델 대비 높은 TOC 추정성능을 보인다.

테스트자료에 Schmoker 경험식을 적용한 결과, R^2가 -0.145, MSE 0.074로 저조한 성능을 보임을 확인하였다. 이는 경험식에서 가정하는 TOC와 밀도검층간 상관관계가 연구지역에서 확인되지 않기 때문이다. 이에 따라 연구지역인 제주분지는 경험식보다는 기계학습을 활용하여 TOC를 추정하는 것이 합리적인 것으로 사료되며, 개발된 모델을 통해 추후 TOC 분석결과가 부재한 구간에 대해서도 경험식 대비 준수한 추정결과를 제시할 수 있을 것으로 기대된다.

XGBoost모델의 실용성을 검증하기 위해 O 시추공 1600-1700 m 및 2600-2700 m 구간의 물리검층자료에 적용한 후 시추암편시료 TOC 분석결과와 비교하였다. 시추암편시료로 분석 시 낮은 TOC를 보였던 1600-1700 m에서는 모델의 추정성능과 시추암편시료의 분석결과가 유사한 경향성을 보이는 것을 확인하였다. 하지만 약 3 wt.%의 높은 실험 TOC를 보인 2600-2700 m에서는 기계학습모델의 TOC가 전반적으로 과소추정된 것을 확인하였다. 향후 변분오토인코더(variational auto encoder) 또는 합성검층을 통한 자료증강, 암편시료를 활용한 TOC 추정모델 개발과 같은 방법을 통해 높은 TOC에 대한 추정성능을 개선하는 연구가 필요하다.

본 연구에서는 가용한 전체자료 수가 118개로 한정되어, 5개의 테스트자료만을 활용하여 모델의 성능을 평가하였다. 이로 인해 모델의 일반화 성능을 충분히 검증하는데 한계가 있으며, 실제 현장자료에 적용할 경우 추정결과에 불확실성이 존재할 수 있다. 따라서 향후, 시추암편시료의 TOC 분석결과와 교차검증을 통한 추가적인 성능평가가 필요할 것으로 판단된다.

감사의 글

본 연구는 한국석유공사 “남해대륙붕 종합기술평가” 사업과 산업통상부 유전개발사업출자의 지원을 받아 수행되었습니다. 또한, 2025년도 정부(기후에너지환경부)의 재원으로 한국에너지기술평가원의 지원(RS-2025-14383289, 동해대륙붕 셰일층 CO_2 기밀성 및 저장성 나노기술기반 평가모델 개발)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. 2021R1C1C1004460)이며, 이에 감사드립니다.

REFERENCES

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

Chen, T. and Guestrin, C., 2016, XGBoost: A scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C., Shen, D. and Rastogi, R., (eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, 785-794.

Han, X., Hou, D., Cheng, X., Li, Y., Niu, C. and Chen, S., 2022, Prediction of TOC in Lishui-Jiaojiang Sag Using Geochemical Analysis, Well Logs, and Machine Learning. Energies, 15, 9480.

Jiang, S., Mokhtari, M., Borrok, D. and Lee, J., 2018, Improving the Total Organic Carbon Estimation of the Eagle Ford Shale with Density Logs by Considering the Effect of Pyrite. Minerals, 8, 154.

Kim, J.H. and Lee, Y.I., 1999, Stratigraphy and Depositional Environment of the Oligocene Series in the Jeju Basin. In: Proceedings of the 6th Annual Conference of The Korean Society of Petroleum Geology. The Korean Society of Petroleum Geology, 29-37 (in Korean with English abstract).

Kim, S.H. and Son, B.K., 2013, Petroleum system modeling of the Jeju basin, offshore southern Korea. Journal of the Geological Society of Korea, 49, 493-491 (in Korean with English abstract).

Kwon, Y.I., 1996, Sequences Stratigraphy and Rift-related Sediment Characteristics in the Cheju Basin. Ph.D. thesis, Yonsei University, Seoul, 302 p (in Korean with English abstract).

Lai, J., Zhao, F., Xia, Z., Su, Y., Zhang, C., Tian, Y., Wang, G. and Qin, Z., 2024, Well log prediction of total organic carbon: A comprehensive review. Earth-Science Reviews, 258, 104913.

Lee, C.Y., Shinn, Y.J., Han, H.C. and Ryu, I.C., 2019, Structural evolution of two-stage rifting in the northern East China Sea Shelf Basin. Geological Journal, 54, 2229-2240.

Lee, Y.J., Yun, H.S., Cheong, T.J., Kwak, Y.H. and Oh, J.H., 1998, Geochemical characteristics of organic matter in the Tertiary sediments from the JDZ Blocks, offshore Korea. The Korean Journal of Petroleum Geology, 6, 25-36 (in Korean with English abstract).

Magoon, L.B. and Dow, W.G., 1994, The petroleum system. In: Magoon, L.B. and Dow, W.G. (eds.), The petroleum system-from source to trap. American Association of Petroleum Geologists, Tulsa, 3-24.

Passey, Q.R., Creaney, S., Kulla, J.B., Moretti, F.J. and Stroud, J.D., 1990, A Practical Model for Organic Richness from Porosity and Resistivity Logs. AAPG Bulletin, 74, 1777-1794.

Peters, K.E. and Cassa, M.R., 1994, Applied source rock geochemistry. In: Magoon, L.B. and Dow, W.G. (eds.), The Petroleum System-From Source to Trap. American Association of Petroleum Geologists, Tulsa, 93-120.

Qu, T., Huang, Z., Li, T., Yang, Y., Wang, B. and Wang, R., 2023, Sedimentology and geochemistry of Eocene source rocks in the east China sea: Controls on hydrocarbon generation and source rock preservation. Journal of Asian Earth Sciences, 255, 105770.

Salaim, M.E., Ma, H., Hu, X. and Quer, H., 2024, An Artificial Neural Network Approach for Predicting TOC and Comprehensive Pyrolysis Parameters from Well Logs and Applications to Source Rock Evaluation. Natural Resources Research, 33, 2063-2087.

Schmoker, J.W., 1979, Determination of Organic Content of Appalachian Devonian Shales from Formation-Density Logs : GEOLOGIC NOTES. AAPG Bulletin, 63, 1504-1509.

Schmoker, J.W., 1981, Determination of Organic-Matter Content of Appalachian Devonian Shales from Gamma-Ray Logs. AAPG Bulletin, 65, 1285-1298.

Shier, D.E., 2004, Well Log Normalization: Methods and Guidelines. Petrophysics-The SPWLA Journal of Formation Evaluation and Reservoir Description, 45, SPWLA-2004-v45n3a4.

Sun, J., Dang, W., Wang, F., Nie, H., Wei, X., Li, P., Zhang, S., Feng, Y. and Li, F., 2023, Prediction of TOC Content in Organic-Rich Shale Using Machine Learning Algorithms: Comparative Study of Random Forest, Support Vector Machine, and XGBoost. Energies, 16, 4159.

Yeom, J.Y., Kim, H.Y., Lee, K.B., Chang, C.D. and Jo, Y.U., 2023, Detection of Borehole Breakout Depth in Image Logs Using Machine Learning Algorithms. Journal of The Korean Society of Mineral and Energy Resources Engineers, 60, 223-230 (in Korean with English abstract).

Zhou, Q., Liu, J., Ma, B., Li, C., Xiao, Y., Chen, G. and Lyu, C., 2024, Pyrite characteristics in Lacustrine Shale and Implications for Organic Matter Enrichment and Shale Oil: A Case Study from the Triassic Yanchang Formation in the Ordos Basin, NW China. ACS Omega, 9, 16519-16535.