◎ 논 문
◎ Original Paper

DOI: https://doi.org/10.5293/kfma.2026.29.1.006
ISSN (Print): 2287-9706

데이터 증강 기법을 활용한 가스터빈 최대출력 예측 모델의 성능 개선 연구

김재동*† · 서정석* · 박준수* · 이가영* · 장성용*

A Study on Improving the Performance of a Gas Turbine Maximum Output Prediction Model Using Data Augmentation Techniques

Jaedong Kim*†, Jung-Seok Seo*, Jun-Su Park*, Ga-Young Lee*, Sung-Yong Chang*

Key Words : Gas Turbine(가스터빈), Maximum Power(최대출력), Data Augmentation(데이터 증강), Gaussian Noise(가우시안 노이즈), XGBoost, LightGBM.

ABSTRACT

This study investigates the performance improvement of a model to predict gas turbine maximum power output through the application of data augmentation techniques. The accurate prediction of a gas turbine maximum output is crucial for stable power system operation and economic dispatch, as it is complexly determined by environmental conditions (such as outside temperature, humidity, and pressure) and equipment status. A prediction model was initially developed by integrating gas turbine operating data and corresponding meteorological data from a combined cycle power plant. A critical challenge in this process was the limited volume and variability of the real-world dataset, which constrains the model's accuracy and generalizability. To overcome this limitation, we introduced and applied data augmentation techniques to generate a more robust and diverse training set. The prediction model, utilizing the XGBoost and LightGBM algorithms, showed significant performance enhancement after applying data augmentation. Specifically, the application of data augmentation successfully reduced the RMSE to approximately 2 MW, achieving a maximum error reduction of 18% compared to the model trained on the original data alone. These results empirically demonstrate the high effectiveness of data augmentation techniques in mitigating data scarcity issues and substantially improving the prediction accuracy and reliability of gas turbine maximum output models.

약어 목록

AT: Ambient Temperature
AH: Absolute Humidity
RH: Relative Humidity
AP: Ambient Pressure
CCPP: Combined Cycle Power Plant
RMSE: Root Mean Squared Error

1. 서 론

1.1 연구배경

전 세계적으로 탄소중립 목표와 기후변화 대응이 강화되면서, 석탄발전을 줄이고 LNG 및 수소를 사용하는 가스터빈 발전 비중이 증가하고 있다. 2015년 파리협정 이후 전 세계적으로 온실가스 감축을 위해 노력하고 있고, 가스터빈은 저탄소 에너지 전환의 핵심 기술로 자리잡고 있다. 이러한 글로벌 트렌드와 더불어 국내 전력 수요의 지속적인 증가와 에

* 한국전력공사 전력연구원(KEPCO Research institute)
† 교신저자, E-mail : jaedong.kim@kepco.co.kr

너지 시장의 복잡성이 심화되면서, 발전설비의 정확한 출력 예측은 안정적인 전력공급과 경제적 운영을 위해 중요한 요소이다. 특히 가스터빈은 높은 효율성과 빠른 기동특성으로 인해 기저부하와 첨두부하 운영에서 중요한 역할을 담당하고 있다. 이 설비의 최대출력을 정확히 예측하는 것은 발전소 수익과 직결될 뿐만 아니라 전력시스템 전체의 안정성과 신뢰성 유지에 매우 중요하다. 가스터빈의 최대출력은 외기온도, 습도와 압력 등의 환경조건, 설비의 운전이력과 정비상태 등 다양한 요인들의 상호작용에 의해 결정된다. 이러한 복잡성으로 인해 기존의 단순한 설계용량 기반 예측방법에는 한계가 있으며, 최근 설비의 상태를 반영한 데이터기반 예측모델이 필요하다.

1.2 연구목적

발전설비가 노후화됨에 따라 성능 보정곡선의 정확도는 떨어진다. 이는 물리적 성능 저하, 작동 특성 변화, 환경 및 유지보수 방식 등과 관련된 복잡한 문제이다. 이러한 요인들로 인해 가스터빈의 효율과 출력이 서서히 감소하고 결과적으로 성능 예측의 신뢰성에 영향을 미칠 수 있다. 따라서, 과거 다양한 기상조건에 따른 최대출력값을 학습하여 데이터기반 예측모델을 개발하고자 한다. 이전 연구에서 복합화력발전소의 출력을 데이터로 학습하여 예측하는 여러 연구가 있다[1-6]. 하지만, 대부분의 연구는 UCI CCPP Dataset[7]을 활용하여 모델을 만들고 성능을 비교하는 방식이다. UCI CCPP Dataset은 데이터 포인트가 9568개로서 충분한 데이터가 존재하며, 가스터빈출력이 아닌 복합발전소 전체출력만 제공되어 가스터빈의 예측모델 구축 및 실무 활용에 어려움이 있다. 실제 발전소 데이터의 경우 최대출력 데이터가 많지 않으며 데이터의 부족은 모델 학습에 큰 제약이 되며, 이는 예측의 신뢰성을 저하시킬 수 있다. 본 연구의 주요 목적은 데이터 부족의 한계를 극복하고 가스터빈 최대출력 예측 모델의 성능을 개선하는 것이다. 이를 위해 실제 발전소 데이터에 적합한 데이터 증강 기법을 적용하여, 모델 학습에 필요한 충분한 다양성과 양을 갖춘 데이터를 확보한다. 궁극적으로, 데이터 증강 기법이 가스터빈 최대출력 예측 모델의 예측 정확도 및 일반화 성능에 미치는 긍정적인 영향을 정량적으로 검증하고, 이를 통해 실제 발전소 운영 환경에서 예측 모델의 신뢰성을 향상시키는데 기여하고자 한다.

2. 데이터 취득 및 분석

2.1 데이터 취득

2.1.1 가스터빈 운전 데이터

복합화력발전소에서 현재 운전 중인 F-Class 가스터빈 운전 데이터를 취득하였다. 데이터를 학습하여 예측모델을 만들기 때문에 너무 오래전 데이터를 학습하게 되면 현재 설비의 상태를 학습할 수 없다. 따라서 최근 4년간의 가스터빈 출력 데이터를 취득하였다.

2.1.2 기상데이터

기상조건은 가스터빈의 출력에 큰 영향을 미치기 때문에 데이터의 정확성이 중요하다. 기상데이터는 기상청에서 제공하는 기상자료개방포털(http://data.kma.go.kr)에서 취득하였다. 기상자료개방포털은 기상청이 제공하는 공식 데이터 플랫폼으로 기상관측 및 예보 자료를 다운로드할 수 있도록 구축된 사이트이다. 여기서 종관기상관측(ASOS) 데이터 중 온도, 상대습도 및 기압의 시간별 데이터를 취득하였다.

2.2 최대출력 데이터

재생에너지 확대, 가스터빈의 유연운전 요구, 경제성 변화 등 복합적인 요인에 의해 부분부하 운전이 빈번해지고 있다. 이로 인해 전 부하(Full load)운전 구간이 많지 않고, 데이터만 보고 최대출력구간을 정확히 특정하기는 어렵다. 본 연구의 목적을 위해서는 정확한 최대 출력 구간의 데이터가 필요하기 때문에 성능시험 데이터를 활용하기로 하였다. 성능시험이란 복합화력발전소의 실제 운전 조건에서 보증된 성능(출력, 효율 등)이 충족되는지 확인하고, 설비의 효율적 운영과 유지관리를 위해 정기적으로 수행하는 시험을 말한다. 이때 당시 기상조건에서 최대 출력을 측정하기 때문에 이 구간의 데이터를 사용하여 학습데이터셋을 구축하였다.

2.3 학습특성(Train features)

기상 데이터에는 다양한 기상요소가 있지만 가스터빈의 출력에 영향을 미치는 대표적인 요소는 기온, 습도 및 기압이다. 이 중 기상청에서 제공하는 습도는 상대습도이다. 상대습도는 현재 공기 중에 포함된 수증기량이 같은 온도에서 포화 상태일 때 포함될 수 있는 최대 수증기량에 대해 몇 퍼센트나 포함되어 있는지를 나타낸다. 반면 절대습도는 단위 부피 안에 포함된 수증기의 질량을 나타낸다. 절대습도가 실제 수증기의 양을 나타내기 때문에 가스터빈 출력에 더 직접적인 영향을 미치는 요소는 절대습도이다. 따라서 상대습도를 절대습도로 변환하여 학습 특성에 포함하였다. 이렇게 얻어진 총 4가지 기상변수(기온, 상대습도, 절대습도, 기압)와 가스터빈 출력과의 상관관계 분석을 수행하였다.

Table 1에서 보는 바와 같이, 가스터빈 출력(GT Power)은 대기온도(AT)와의 상관성이 가장 크고, 상대습도(RH)보다는 절대습도(AH)와의 관련성이 더 크다. 따라서 상관성이 낮은 상대습도를 제외한 나머지 3가지 특성(대기온도, 절대

Table 1 Pearson correlation coefficient of weather features and gas turbine power outputs

Features	AT	AH	RH	AP	GT Power
AT	1.00	0.92	0.74	-0.61	-0.98
AH	0.92	1.00	0.87	-0.64	-0.92
RH	0.76	0.87	1.00	-0.51	-0.75
AP	-0.61	-0.64	-0.51	1.00	0.72
GT Power	-0.98	-0.92	-0.75	0.72	1.00

Table 2 Train and target features

Train features			Target features
AT	AH	AP	GT Power output

습도, 대기압력)을 학습 특성으로 사용하였고, Table 2와 같이 3가지 특성으로 가스터빈 출력을 예측하는 모델을 구축하였다.

3. 데이터 증강(Data Augmentation)

3.1 데이터 증강의 필요성

머신러닝 및 딥러닝 기반 모델의 성능은 주어진 학습 데이터의 양과 질에 크게 의존한다. 하지만 복합화력발전소의 성능시험은 일반적으로 월 3~4회 밖에 수행하지 않기 때문에 확보할 수 있는 데이터가 많지 않다. 이러한 데이터 부족 문제는 모델의 일반화 성능 저하, 과적합 문제, 그리고 실제 적용에서의 신뢰도 하락으로 이어질 수 있다. 이런 문제를 해결하기 위해 데이터 증강이 필요하다. 데이터 증강은 기존의 학습 데이터를 인위적으로 변형하거나 확장하여 새로운 데이터를 생성하는 기법이다. 데이터 부족 상황에서 데이터 증강을 이용하여 모델의 성능을 향상시킬 수 있다.

3.2 가우시안 노이즈를 이용한 증강

수치형 데이터를 증강하는 방법은 여러가지가 있지만, 본 연구에서는 Fig. 1과 같이 원본데이터에 가우시안 노이즈(Gaussian noise)를 추가하는 방식으로 증강을 수행하였다. 가우시안 노이즈 추가는 데이터 증강에서 자주 사용되는 방법[8-9]으로, 데이터에 수학적으로 정의된 가우시안 분포를 따르는 무작위 잡음을 인위적으로 더하는 기법이다. 이는 산업현장에서 흔히 나타나는 잡음패턴을 모사하며, 모델이 다양한 변동 상황에 적응하도록 학습시키는데 효과적이다. 가우시안 노이즈 추가는 특정한 패턴이 아닌, 자연스러운 형태의 무작위 변동을 데이터에 부여하여, 학습하는 AI모델이 현실적인 다양한 상황에 유연하게 반응할 수 있도록 하는 검증된 데이터 증강 방법이다. Onishi 등[10]의 연구에서는 정형

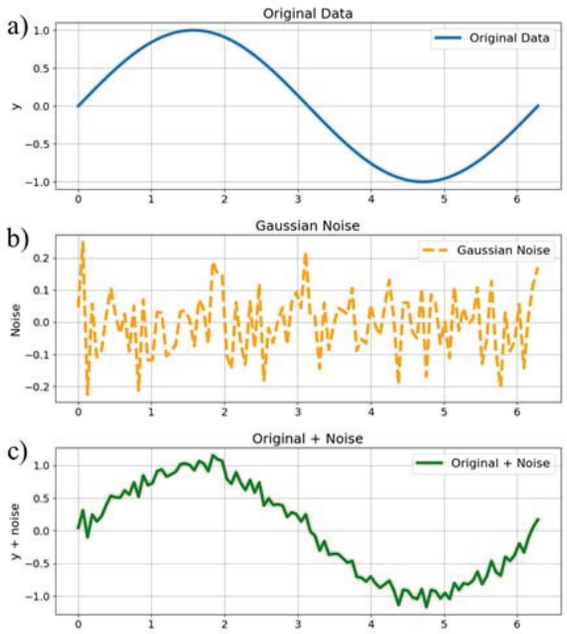

Fig. 1 Gaussian noise augmentation example: a) Original data; b) Gaussian noise data; c) Augmented data

데이터에 대한 다양한 증강 기법을 비교 분석한 결과, 가우시안 노이즈 기반 증강이 데이터의 구조적 특성을 유지하면서도 효과적인 정규화를 제공함을 확인하였다. 이는 가스터빈 출력예측과 같이 연속형 수치 변수가 주를 이루는 회귀 문제에 특히 적합하다. 또한, Li 등[11]의 비교 연구에 따르면, SMOTE, VAE, GAN 등 다양한 tabular data 증강 기법 중에서 가우시안 노이즈 증강은 계산 효율성과 구현 용이성 측면에서 우수하며, 회귀 문제에서는 복잡한 생성 모델과 비슷하거나 더 나은 성능을 보이는 경우가 많다.

4. 예측모델 학습

4.1 사용 알고리즘

본 연구에서는 예측모델을 만들기 위해 정형 데이터 회귀 모델 중 가장 안정적이고 성능이 검증된 아래 2가지 알고리즘을 사용하였다.

4.1.1 XGBoost(eXtreme Gradient Boosting)[12]

XGBoost는 여러 개의 약한 학습기를 연속적으로 결합해 예측력을 높이는 앙상블 기법이다. 일반적인 Gradient Boosting 기법을 빠르고 유연하며 효과적으로 확장한 버전이라고 할 수 있다. 높은 예측 성능과 빠른 처리 속도로 정형 데이터 회귀(tabular data regression) 분야에서 널리 사용되고 있는 알고리즘이다.

4.1.2 LightGBM(Light Gradient Boosting Machine)[13]

LightGBM은 고성능과 효율성을 강조하는 머신러닝 알고리즘으로, 대규모 데이터와 고차원 특성에서도 빠른 학습과 예측이 가능하도록 설계된 알고리즘이다. XGBoost와 근본적 작동원리는 비슷하지만, 속도와 효율성 등을 보완하기 위해 설계된 알고리즘이다. 하지만 모든 상황에서 예측 성능이 일방적으로 우수하다고 볼 수는 없기 때문에 모델의 비교 분석이 필요하다.

4.2 모델학습 및 검증

4.2.1 데이터 셋 (Dataset)

앞에서 설명한 내용과 같이 4년간 발전소 성능시험시 출력데이터와 동일 기간의 기상데이터를 결합하여 Table 3과 같이 학습 및 검증 데이터 셋을 구축하였다. 전체 데이터 개수는 69개 밖에 되지 않기 때문에 모델 학습을 위해서는 데이터 증강이 필수적이다. 결정트리(Decision Tree) 알고리즘을 사용하였기 때문에 데이터 스케일링은 진행하지 않았다.

4.2.2 K-fold 교차검증을 이용한 데이터 분할

전통적인 학습/검증/테스트 분할 방식은 데이터가 충분할 때는 효과적이고 일반적인 방법이다. 하지만 데이터가 부족한 문제에서 검증 세트를 따로 분리해 놓을 경우 학습에 사용할 수 있는 데이터가 더 줄어들어 모델의 성능이 저하될 수 있다. K-fold 교차 검증은 데이터를 K개의 폴드(fold)로 나누어, 각각을 한 번씩 검증데이터로 사용하고 나머지를 학습 세트로 사용하는 방법이다. 이를 통해 모든 데이터가 학습과 검증에 사용되어 데이터 활용도를 극대화할 수 있다. 본 연구에서는 K=4로 설정하여 4개의 학습데이터셋을 만들었다.

Table 3 Example of Dataset (n_data=69)

Date	AT(℃)	AH(g/m^3)	AP(hPa)	Power (MW)
2019-01-03	7.95	1.31	1023.1	193.12
2019-01-08	7.50	2.32	1014.1	190.96
2019-01-15	9.15	3.20	1010.8	187.91
...	...	...	...	...

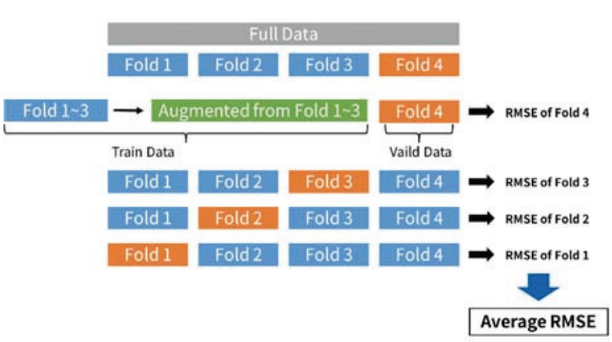

Fig. 2 Model training and validation method using K-Fold cross validation

4.2.3 데이터 증강을 적용한 모델 학습 및 검증

4개로 나누어진 학습/검증 데이터 중 학습데이터에만 데이터 증강을 적용한다. 검증 데이터에 증강을 적용하면 모델의 실제 성능을 과대평가하게 되어 객관적인 평가가 불가능하다. 학습 시에는 다양한 증강 데이터로 강건성을 학습하되, 검증 시에는 원본 데이터로 실제 성능을 측정해야 모델의 실제 성능을 객관적으로 평가할 수 있다. 따라서 Fig. 2과 같이 4개의 폴드로 나누어 3개 폴드에 증강을 적용하고 이 데이터로 모델을 학습한 후 나머지 1개 폴드로 검증을 하고 이를 각 폴드에 순차적으로 적용하였다. 이렇게 하면 모든 데이터에 대해 학습과 검증을 진행할 수 있고 검증데이터 누출 없이 데이터를 최대한 활용하여 모델의 일반화 성능을 검증할 수 있다. 각 폴드 데이터에 대한 RMSE를 계산한 후, 4개 폴드의 RMSE를 평균하여 모델의 오차를 계산하였다.

Table 4 Effect of data augmentation on average RMSE for XGBoost model

	Augmentation multiplier			
Noise intensity	No Aug. (Baseline)	×2	×4	×6
0	2.30	-	-	-
0.01	-	2.30	2.13	2.16
0.05	-	2.20	2.16	**2.02 (-11.96%)**
0.10	-	2.07	2.11	2.20
0.20	-	2.23	2.17	2.11

* Number of data points
No Aug: 69, ×2: 138, ×4: 276, ×6: 414

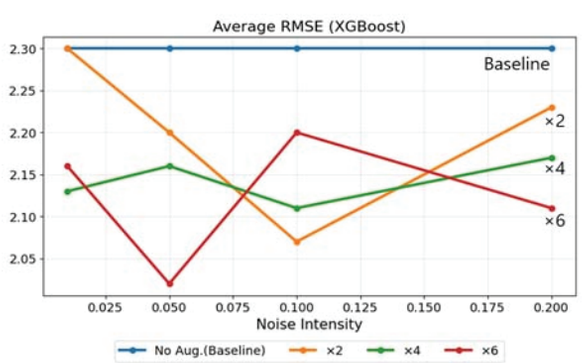

Fig. 3 Effect of data augmentation on average RMSE for XGBoost model

Table 5 Effect of data augmentation on average RMSE for LightGBM model

	Augmentation multiplier			
Noise intensity	No Aug. (Baseline)	×2	×4	×6
0	2.35	-	-	-
0.01	-	2.16	2.21	2.01
0.05	-	2.09	2.01	2.09
0.10	-	1.99	**1.92 (-18.26%)**	2.08
0.20	-	2.15	2.06	2.12

* Number of data points
No Aug: 69, ×2: 138, ×4: 276, ×6: 414

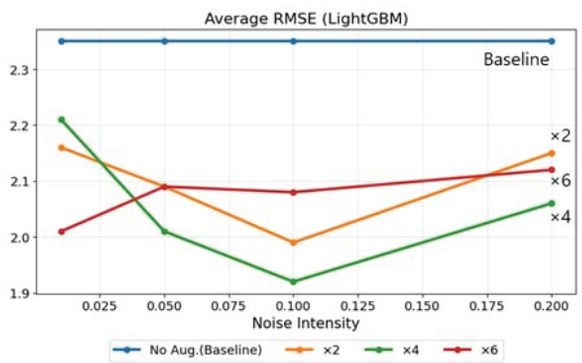

Fig. 4 Effect of data augmentation on average RMSE for LightGBM model

5. 결과 분석

5.1 데이터 증강 효과 사례 연구(case study)

데이터 증강에 따른 모델 성능 변화를 확인하기 위해 데이터 증강배수와 노이즈 강도를 변화해 가며 RMSE 값의 변화를 확인하였다. Table 4는 데이터 증강에 따른 XGBoost 모델의 검증오차 변화를 보여준다. 증강배수와 노이즈 강도(noise intensity) 변화에 따른 RMSE변화를 확인한 결과 거의 모든 경우에서 데이터 증강했을 경우의 RMSE가 낮았다. 증강을 하지 않았을 경우를 기준으로 가장 RMSE가 낮은 경우는 노이즈 강도 0.05, 증강배수 6배 일 때이고 오차가 11.96% 감소하였다.

Table 6 RMSE and R2 of LightGBM prediction(Aug. multiplier: ×6, Noise intensity: 0.10)

LightGBM	RMSE	R^2
Fold 1	2.478	0.940
Fold 2	1.722	0.968
Fold 3	2.186	0.943
Fold 4	1.307	0.979
Average	**1.923**	**0.958**

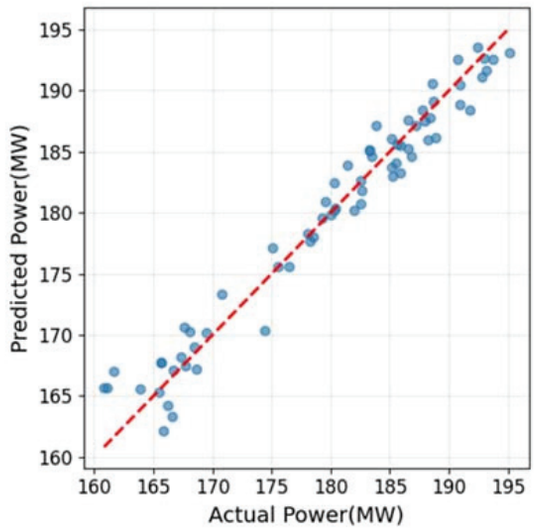

Fig. 5 Result plot for predicted power output

반면 LightGBM모델에서는 Table 5에서와 같이 노이즈 강도 0.1, 증강배수 4배 일 때 오차가 18.26% 감소하여 가장 낮은 RMSE를 보였다.

5.2 예측 결과

위 사례 연구(case study)에서 RMSE가 가장 낮게 나타난 모델의 각 폴드별 RMSE와 R2를 Table 6에 나타내었다. 폴드에 따라 예측오차의 편차가 존재하지만 Fig. 5에서 확인할 수 있듯이 과도한 예측오차를 보이는 데이터는 관찰되지 않았다. R^2값이 모두 94%이상으로 데이터 증강이 원본데이터의 경향성에서 크게 벗어나지 않았다고 판단된다.

6. 결 론

본 연구에서는 데이터 증강 기법을 활용하여 기상 데이터 기반의 가스터빈 최대출력 예측 모델의 성능 개선에 대한 연구를 진행하였다. 연구 결과, 적은 데이터 수에도 불구하고 데이터 증강을 통해 모델의 예측 정확도가 향상되었음을 확인할 수 있었다. 특히, LightGBM 알고리즘을 사용한 예측 모델에서 데이터 증강을 적용한 경우, 평균적으로 2MW 내외의 오차를 보이며 최대 18%의 오차 감소 효과를 보였다. 다만, 모델과 데이터에 따라 최적의 증강 파라미터가 달라지기 때문에 실제 적용 시에는 충분한 사례 연구(case study)가 필요할 것으로 보인다.

향후 연구에서는 다양한 데이터 증강 기법을 적용하여 예측 모델의 성능을 더욱 향상시키고, 가스터빈 출력뿐만 아니라 증기터빈 출력도 예측하여 복합화력발전소 전체의 출력을 예측하는 것이 필요하다.

후 기

본 연구는 2025년도 정부(산업통상자원부)의 재원으로 한국에너지기술평가원의 지원을 받아 수행된 연구임(과제명: 표준가스복합발전 플랜트 운영 최적화 기술개발, 과제번호: 20217010100020)

References

(1) C. Ahamed Saleel, 2021, “Forecasting the energy output from a combined cycle thermal power plant using deep learning models”, Case Studies in Thermal Engineering, Vol. 28, 101693.

(2) Afzal, A., Alshahrani, S., Alrobaian, A., Buradi, A., & Khan, S. A., 2021, “Power plant energy predictions based on thermal factors using ridge and support vector regressor algorithms” Energies, Vol. 14, No. 21, 7254.

(3) Siddiqui, R., Anwar, H., Ullah, F., Ullah, R., Rehman, M. A., Jan, N., & Zaman, F., 2021, “Power prediction of combined cycle power plant (CCPP) using machine learning algorithm-based paradigm.”, Wireless Communications and Mobile Computing, Vol. 2021, No. 1, 9966395.

(4) Aslan, Asiye, and Ali Osman Büyükköse. 2025, “Comparative Performance Analysis of Machine Learning-Based Annual and Seasonal Approaches for Power Output Prediction in Combined Cycle Power Plants” Energies, Vol. 18, No. 19, 5110.

(5) Song, Yujeong, Jisu Park, Myoung-Seok Suh, and Chansoo Kim. 2024, “Prediction of Full-Load Electrical Power Output of Combined Cycle Power Plant Using a Super Learner Ensemble” Applied Sciences, Vol. 14, No. 24, 11638.

(6) Saeed MA, El-Kenawy E-SM, Ibrahim A, Abdelhamid AA, Eid MM, Karim FK, Khafaga DS and Abualigah L, 2023, Electrical power output prediction of combined cycle power plants using a recurrent neural network optimized by waterwheel plant algorithm. Front. Energy Res., Vol. 11, 1234624.

(7) P. Tfekci and H. Kaya., 2014, “Combined Cycle Power Plant,” UCI Machine Learning Repository.

(8) H. Dong, J. Wang, X. Wu, M. Zhou and J. Lü., 2023, “Gaussian noise data augmentation-based delay prediction for high-speed railways.” IEEE Intelligent Transportation Systems Magazine, Vol. 15, No. 6, pp. 8～18.

(9) Arora, A., Shoeibi, N., Sati, V., González-Briones, A., Chamoso, P., Corchado, E., 2020, “Data augmentation using gaussian mixture model on csv files.” International Symposium on Distributed Computing and Artificial Intelligence. pp. 258～265.

(10) Onishi, S., & Meguro, S., 2023, Rethinking data augm-entation for tabular data in deep learning. arXiv preprint arXiv:2305.10308.

(11) Cui, L., Li, H., Chen, K., Shou, L., & Chen, G., 2024, Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai. arXiv preprint arXiv:2407.21523.

(12) Chen, T., & Guestrin, C., 2016, “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785～794.

(13) Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y., 2017, “Lightgbm: A highly efficient gradient boosting decision tree.” Advances in neural information processing systems 30.