JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 1, pp. 235-248, Jan. 2026

Check for updates

시그널 특성함수를 이용한 발전설비 감시 AI 플랫폼에 관한 연구

김 종 호*
경성대학교 경영학과 부교수

An AI Platform to Monitor Power Generation Equipment Using Signal Feature Functions

Jong-Ho Kim*

Associate Professor, Department of Business Administration, Kyungsung University, Busan 48434, Korea

[요 약]

본 연구는 발전설비의 비 계획적 기동정지로 인한 효율 저하와 유지비용 증가 문제를 해결하기 위해, 시그널 특성함수(Signal Feature Functions)를 활용한 인공지능 기반 감시 플랫폼을 개발하였다. 석탄발전소 1개 호기의 1초 단위 운전데이터를 수집하여 탐색적 데이터 분석을 수행하고, 이상탐지 모델과 가혹도 예측 회귀모델을 구축하였다. 분석 결과, 발전량과 주증기 유량 간의 상관계수 저하가 불시정지의 주요 예측 인자로 확인되었으며, AI 모델은 고장 발생 24~48시간 전에 위험 패턴을 사전에 탐지할 수 있었다. 또한, 실시간 데이터 적재·변환·이상경보·시각화를 통합적으로 지원하는 감시 시스템을 구현하였다. 개발된 플랫폼은 설비 운전데이터의 시그널 특성함수를 실시간으로 분석함으로써 이상징후를 조기에 감지하고, 가혹도 수준을 정량적으로 예측하여 예방정비를 가능하게 한다. 이를 통해 발전설비의 안정성, 신뢰성, 운영 효율성을 향상하며, 공공데이터 기반의 AI 감시 기술의 산업적 확장 가능성을 제시한다.

[Abstract]

In this study, an AI-based platform is proposed to monitor power generation equipment using signal features in order to mitigate increased maintenance costs and reduced efficiency caused by unplanned shutdowns at power generation facilities. Operational data were collected at a single coal-fired power unit at one-second intervals for an exploratory analysis. An Isolation Forest model and an Autoencoder model were trained to detect anomalies, and a conventional regression model was trained to predict their severity. The results of the analysis revealed that a decrease in the correlation coefficient between power generation and primary steam flow served as a key indicator for predicting unplanned outages. Moreover, the AI models successfully detected risk patterns 24–48 hours prior to failure. A monitoring system was also implemented to integrate real-time data collection, signal transformation, anomaly detection, and visualization. By analyzing the features extracted from input signals comprising operational data in real time, the proposed platform enables early detection of abnormal conditions and quantitative prediction of their severity to facilitate predictive maintenance. Consequently, the results demonstrate that the proposed approach can enhance the stability, reliability, and operational efficiency of power generation facilities, and they also confirm the industrial scalability of AI-based monitoring technologies utilizing public data.

색인어 : 플랫폼, 발전설비, 시그널 특성함수, 인공지능, 이상탐지

Keyword : Platform, Power Plant, Signal Feature Functions, Artificial Intelligence, Anomaly Detection

http://dx.doi.org/10.9728/dcs.2026.27.1.235

Received 30 October 2025; **Revised** 08 December 2025
Accepted 07 January 2026

***Corresponding Author; Jong-Ho Kim**

Tel: ※ 개인정보 표시제한

E-mail: jonghokim@ks.ac.kr

Ⅰ. 서 론

1-1 연구배경 및 목적

최근 인공지능(AI)과 빅데이터 기술의 발전은 산업 전반의 운영 효율성과 지능화 수준을 획기적으로 향상하고 있다[1],[2]. 특히 발전산업 분야에서는 설비의 안정적인 운전이 국가 에너지 공급망의 핵심 요소로 인식되며, 비 계획적 기동정지로 인한 손실이 심각한 문제로 대두되고 있다[3].

예를 들어, 대형 화력발전소의 보일러 튜브 또는 주증기 계통의 고장으로 인해 발전기가 갑작스럽게 정지하면, 단일 사건으로도 수억 원 규모의 직접 손실과 수일간의 가동 중단이 발생한다. 이러한 현상은 단순한 설비 결함뿐 아니라 운전 조건의 변화, 열적 스트레스, 유량 불균형 등 복합적 요인에 의해 유발되는 다변량 이상 패턴의 결과로 해석된다.

기존의 발전설비 관리체계는 주기적 점검과 예방정비 위주의 사후적 유지보수 체계가 주를 이루었으나, 이는 설비의 실시간 상태를 반영하지 못하고 데이터의 미활용이라는 구조적 한계를 가진다. 또한, 발전소에는 수천 개의 센서가 초당 단위로 데이터를 생성하기 때문에, 단순 모니터링만으로는 복잡한 이상 패턴을 조기에 식별하기 어렵다.

이러한 한계를 극복하기 위해 최근 AI 기반 예지 보전(Predictive Maintenance) 기술이 주목받고 있으며, 특히 시계열 신호 기반 이상탐지 모델이 높은 신뢰성을 보인다[4]. 예지 보전은 정상 상태의 운전데이터를 기반으로 패턴을 학습하고, 이상징후를 조기에 탐지함으로써 고장을 예방하는 기술로 정의된다[5].

따라서, 설비 신호의 시간적·주파수적 특성을 정량화하는 시그널 특성함수를 정의하고, 이를 이용한 AI 기반 감시 플랫폼을 개발함으로써, 발전설비의 안정적 운영과 효율성 향상을 달성하는 것이 필요하다. 본 연구는 이러한 관점에서, 기존의 수동형 설비감시 체계를 데이터 기반 지능형 감시 체계로 전환하기 위한 실증적 기술모델을 제시하고자 한다. 이를 위해 A사 석탄발전소 1개 호기의 실제 운전데이터를 활용하여 이상징후를 탐지하고, 가혹도 수준을 정량화하는 AI 기반 감시 플랫폼을 구축하는 것을 목표로 한다.

기존 예지 보전연구는 주로 단일 변수의 추세 변화나 FFT 기반의 진동 신호 특성에 초점을 맞추어 왔으며, 변수 간 상관 구조의 동적 붕괴(correlation breakdown)를 고장예측 인자로 활용한 사례는 거의 보고되지 않았다. 그러나 실제 발전설비 고장 현상은 단일 변수의 이상값보다 변수 간 물리적 관계의 약화에서 더 명확한 전조가 나타나는 경우가 많다. 본 연구는 이러한 점에 착안하여 발전량, 주증기 유량, 급수 유량, 압력 등 주요 운전변수 간 상관계수의 시간적 하락 현상이 고장 2~3일 전에 반복적으로 나타난다는 실증적 패턴을 규명하고, 이를 고장예측 인자로 활용하는 새로운 접근을 시도하였다.

아울러 기존 예지 보전연구는 이상탐지와 고장 강도 예측을 별도로 수행하는 경우가 많았으나, 본 연구는 비지도 이상탐지와 이상 지수 기반 가혹도 회귀모델을 결합한 두 단계 통합 구조를 제안하였다. 이를 통해 단순 이상탐지를 넘어, 이상 원인의 정량적 해석과 리스크 수준 평가가 가능한 통합 프레임워크를 구축하였다.

기존 연구는 평균, 표준편차 등 일부 기초 통계량 또는 특정 주파수 성분을 이용하는 방식에 의존하는 경향이 있었으며, 설비 신호의 복잡도, 비선형성, 변동성, 상관성 등 다양한 특성을 포괄적으로 반영한 분석체계가 부족하였다. 본 연구는 특징기반 시계열 분석 연구에서 제시된 관점을 발전설비 도메인에 적용하여, 시그널 특성함수(Signal Feature Functions)를 기반으로 한 고차원적 신호 표현 방식을 설계하였다. 이는 기존의 단순 수치 중심 분석을 넘어 발전설비의 물리적 상태 변화를 정량화할 수 있는 새로운 신호 기반 모델이라는 점에서 의미가 있다.

1-2 연구방법

본 연구는 발전설비의 실시간 운전데이터를 활용하여 이상탐지 및 성능 예측모델을 개발하기 위한 데이터 기반 실증 연구로 수행되었다. 연구 절차는 다음과 같이 구성된다.

첫째, 데이터 수집 단계에서는 1개 석탄 발전기의 주요 설비(발전기, 보일러, 주증기, 급수기, 복수기 등)로부터 1초 단위의 시계열 데이터를 수집하였다. 수집된 데이터는 54개 변수, 약 3,100만 건으로 구성되었으며, 총 13개월(2023년 10월~2024년 10월)의 운전기록을 확보하였다.

둘째, 탐색적 데이터 분석(EDA) 단계에서는 일평균 발전량을 기준으로 정상운전, 이상운전, 기동·정지운전, 불시정지 등 8개 운전 패턴을 분류하고, 변수 간 상관관계를 분석하여 고장 예측 인자를 도출하였다. 특히 발전량과 주증기 유량 간 상관계수의 저하가 불시정지의 주요 징후로 확인되었다.

셋째, 이상탐지 모델링 단계에서는 시계열 데이터를 시간 구간(Time Window)과 슬라이딩 윈도우(Sliding Window) 방식으로 구간화하여 학습용 데이터셋을 구축하였다. 비지도 학습 기반의 Isolation Forest와 AutoEncoder 모델을 적용하여 이상운전 구간을 탐지하였으며, 이상탐지 결과와 기술분석 결과를 상호 검증하였다.

넷째, 가혹도 예측 모델을 구축하기 위해 Isolation Forest로 산출된 Anomaly Score를 회귀모델의 종속변수로 활용하였다. Random Forest, LightGBM, CatBoost 등의 앙상블 회귀모델을 적용하여 주요 영향변수(발전량, 급수유량, 주증기압력, 표준편차 등)를 도출하였고, R^2 0.97~0.99 수준의 높은 예측정확도를 확보하였다.

마지막으로, AI 감시 플랫폼 구축 단계에서는 PostgreSQL, DuckDB 하이브리드 DB 구조와 Apache Airflow, FastAPI, Nginx Unit 등 오픈소스 기반의 시스템 아키텍처를 구현하였다. 이를 통해 실시간 데이터 적재·변환·시각화 및 이상경보 자동화를 실현하였다.

II. 이론적 배경

2-1 이상탐지 모델

본 연구에서는 발전설비의 이상운전 상태를 조기에 탐지하고 예측하기 위한 비지도 학습 기반의 이상탐지 알고리즘으로서 Isolation Forest와 AutoEncoder 기법을 적용하였다. 이 두 모델은 각각 서로 다른 수학적 원리를 기반으로 이상치를 판별하며, 시계열 기반의 고차원 설비 데이터에 대한 이상탐지 문제를 효과적으로 해결할 수 있는 대표적인 기법이다. 두 기법 모두 발전설비와 같이 복잡하고 대규모의 센서 기반 시계열 데이터에 적합하며, 학습된 모델은 비정상적인 운영 패턴이나 장비 이상을 조기에 탐지하는 데 효과적으로 활용될 수 있다. 특히, 본 연구에서는 기술 분석을 통해 정의된 특성함수 기반 지표들을 입력 변수로 사용하여 이상탐지 모델을 학습하였으며, 각각의 기법이 탐지한 이상운전 구간의 적중률을 비교함으로써 모델의 실효성을 평가하였다. 또한, 탐지 결과를 탐색적 분석결과와 교차 분석하여 모델이 포착한 이상신호가 실제 고장 전조와 어떤 상관성을 가지는지 정량적으로 검증하였다.

1) Isolation Forest

Isolation Forest는 Liu 등[6]에 의해 제안된 이상탐지 알고리즘으로, 데이터 내 이상치를 다른 점들과 얼마나 쉽게 고립(isolate)시킬 수 있는지를 기반으로 한다(그림 1 참조).

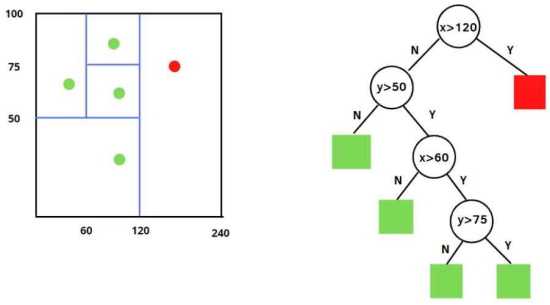


그림 1. Isolation Forest의 분할구조와 분류경계
Fig. 1. Partition structure and decision boundaries of a Isolation Forest

본 기법은 데이터를 무작위로 선택된 특성과 값을 기준으로 반복적으로 이진 분할하는 트리 구조를 사용하여, 각 데이터 포인트를 고립시키는 데 필요한 분할 수(또는 깊이)를 측정한다. 일반적으로 이상치는 정상 데이터보다 훨씬 적은 수의 분할만으로 고립되기 때문에, 평균 분할 깊이를 기반으로 이상 점수(anomaly score)를 산출하여 이상 여부를 판별한다. Isolation Forest는 거리나 밀도 기반이 아닌, 분할 기반의 이상 탐지 알고리즘이라는 점에서 계산 복잡도가 낮고, 고차원 데이터셋에도 효율적으로 적용할 수 있는 장점이 있다[6],[7].

2) AutoEncoder

AutoEncoder는 입력 데이터를 저차원(latent space)으로 압축하였다가 다시 원래의 차원으로 복원하는 과정을 학습하는 인공신경망 기반의 비지도 학습 모델이다(그림 2 참조). 이 모델은 일반적으로 다층 퍼셉트론(multi-layer perceptron, MLP) 구조로 구성되며, 인코딩 단계에서 입력의 중요한 특징만을 추출하고, 디코딩 단계에서 이를 바탕으로 원래 입력과 유사한 출력을 생성한다. 이상탐지에 AutoEncoder를 활용하는 경우, 정상 데이터만을 학습시킨 후 테스트 시 입력과 출력 간의 복원오차(reconstruction error)를 계산하여 이상 여부를 판단한다. 이상치는 정상 데이터 분포에서 벗어난 형태이므로, 복원오차가 상대적으로 크게 나타나는 특성이 있다. 이러한 방식은 비선형적인 데이터 구조를 효과적으로 모델링할 수 있으며, 시계열, 이미지, 센서 데이터 등 다양한 도메인에서 사용되고 있다[8]-[10].

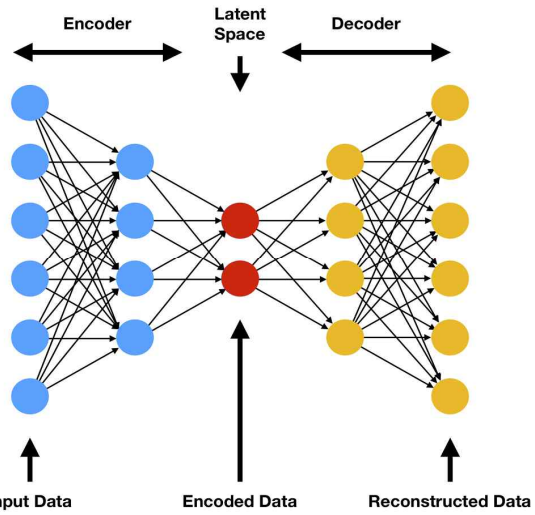


그림 2. 오토인코더(Autoencoder)의 구조 개념도
Fig. 2. Conceptual structure of an Autoencoder

2-2 가혹도 모델

본 연구에서는 발전설비의 이상징후 강도를 정량화하기 위한 "가혹도 모델"의 회귀 예측을 위해, 트리 기반의 앙상블 모델인 Random Forest, LightGBM, CatBoost를 적용하였다. 이들 알고리즘은 각각 고유한 방식의 트리 결합 전략을 가지고 있으며, 예측 정확도 향상과 과적합 방지, 범주형 데이터 처리 최적화 등의 장점이 있다.

1) Random Forest

Random Forest는 Breiman[11]이 제안한 대표적인 앙상블 학습 기법으로, 여러 개의 결정 트리(Decision Trees)를 생성하고, 그 결과를 평균 또는 다수결로 종합하는 방식이다. 각 트리는 서로 다른 부트스트랩 샘플(bootstrapped samples)을 기반으로 학습되며, 노드 분할 시 일부 특성만 사용함으로써 트리 간의 상관성을 줄이고 예측 성능을 향상

한다.

무작위성과 평균화를 통해 개별 트리의 과적합 문제를 해결하며, 회귀 및 분류 문제 모두에서 안정적인 성능을 보인다. 특히 노이즈가 포함된 고차원 데이터셋에서도 강건한 예측이 가능하다는 장점이 있다.

2) LightGBM (Light Gradient Boosting Machine)

LightGBM은 Microsoft Research에서 개발한 그래디언트 부스팅 프레임워크로, 대규모 데이터셋을 빠르고 효율적으로 처리할 수 있도록 최적화된 알고리즘이다. 기존의 Gradient Boosting Decision Tree(GBDT) 방식과 달리, LightGBM은 리프 중심 트리 성장 방식(leaf-wise growth)을 채택하여, 정보 이득이 가장 큰 리프를 우선적으로 분할함으로써 더 낮은 손실을 달성한다[12].

또한, LightGBM은 히스토그램 기반의 학습 방식을 사용하여 연산 속도를 개선하고 메모리 사용량을 줄인다. 이러한 구조는 고차원/대규모 데이터에 적합하며, 병렬 학습이 용이하여 실시간 예측 환경에서도 활용 가능하다.

3) CatBoost

CatBoost는 Yandex에서 개발한 범주형 데이터 처리에 최적화된 Gradient Boosting 알고리즘이다. 기존 GBDT와 달리, CatBoost는 범주형 변수에 대한 전처리 없이도 자동 인코딩을 수행할 수 있는 Target Statistics 방식을 채택하고 있으며, 대칭 트리 구조(Symmetric Tree Structure)를 사용하여 추론 속도를 크게 향상시킨다[13].

2-3 시그널 특성함수

시계열 기반의 센서 데이터에서 의미 있는 정보를 추출하기 위해서는 원시 데이터를 그대로 사용하는 대신, 특정 통계적·물리적 특성을 정량화한 지표들을 생성하는 것이 일반적이다. 이러한 지표를 시그널 특성함수(Signal Feature Functions)라고 하며, 이는 원시 데이터의 복잡한 시간적 구조를 요약하고, 기계적·물리적 상태의 변화를 효과적으로 감지하는 데 활용된다[14],[15].

시그널 특성함수는 통상적으로 데이터의 평균, 표준편차, 최대값 및 최소값과 같은 기초 통계량에서부터, 변화율, 첨도(kurtosis), 왜도(skewness) 등과 같은 시간 도메인 기반의 동적 특성, 그리고 푸리에 변환을 통해 얻는 주파수 도메인 기반 특성까지 다양한 방식으로 구성된다. 또한, 여러 센서 간의 상관관계 분석(correlation analysis)을 통해 다변량 신호 간의 동적 상호작용을 정량화할 수도 있다[16].

설비 내 주요 변수(예: 온도, 압력, 유량) 간의 상관계수가 갑작스럽게 저하될 경우, 이는 시스템의 불균형 상태나 이상 동작 가능성을 시사할 수 있다. 이처럼 시그널 특성함수는 단순히 수치 요약을 넘어, 이상 탐지, 고장예측, 수명 추정 등 다양한 산업 응용에 활용 가능한 기반 지표로 기능한다[17]. 특히 본 연구에서 가장 핵심적인 시그널 특성함수는 변수 간 동적 상관계수(correlation coefficient)의 변화율과 붕괴 패턴이다. 이는 기존 연구에서 주로 사용된 단일 변수의 평균이나 분산 기반 특성과 달리, 발전설비 내 다변량 시스템의 관계 구조가 약화되는 현상을 직접적으로 포착한다는 점에서 본질적인 차별성을 가진다.

표 1. 시그널 특성함수
Table 1. Signal feature functions

Signal Feature Function	Description
length	length of time series
absolute energy	sum over the squared values
absolute_sum_of_changes	sum over the absolute value of consecutive changes in the series x
cid_ce	an estimate for a time series complexity (A more complex time series has more peaks, valleys etc)
count_above	percentage of values in x that are higher than t (threshold)
count_below	the percentage of values in x that are lower than t
number of peaks	Calculates the number of peaks in the time series
ratio_beyond_2_sigma	Ratio of values that are more than 2 * std(x) (so r times sigma) away from the mean of x
approximate entropy	a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data
correlation_coefficient	Pearson correlation coefficient between key operational variables used to quantify dynamic relationship strength and detect correlation breakdown preceding failures
root_mean_square	RMS
standard_deviation	Standard Deviation
skewness	Skewness
kurtosis	Kurtosis
maximum	the highest value of the time series
minimum	the lowest value of the time series
absolute maximum	the highest absolute value
crest factor	Peak(Absolute Maximum)/RMS
peak to peak	maximum - minimum

Ⅲ. 데이터 수집과 탐색적 데이터 분석

3-1 데이터 수집

1개 석탄발전소의 주요 발전기, 보일러, 과열기, 복수기, 급수기, 주증기 계통에서 1초 단위로 데이터를 수집하였다. 발전기에서 발전량(MW 단위)은 1초 단위로 측정하고 과열기 총 9개 부위에 대한 온도(섭씨)는 1초 단위로 측정하였다. 배관 중 복수기 계통, 급수 계통, 주증기 계통 총 3곳의 유량(시간당 톤)을 1초 단위로 측정하고 주증기 계통은 유량과 함께 압력도 측정하였다. 또한 보일러 내부에서는 노내 압력(mmH_2O)과 Tube Leak를 탐지를 위한 총 38개의 음압센서로부터 1초 단위로 음압데이터를 수집하였다.

2023년 10월부터 2024년 10월까지 총 13개월의 데이터

를 입수하였다. 이 중 2023년 11월부터 2024년 10월까지 일 년 치 366일간 1초 단위로 수집된 데이터를 활용하여 총 54개의 변수와 약 3,100만개(366일×86,400초)의 레코드로 구성된 데이터셋을 생성하였다.

3-2 탐색적 데이터 분석

1) 일평균 발전량 추이와 일별 패턴추출

일평균 발전량을 기준으로 정상출력구간, 기동정지 구간, 운전전환 구간으로 구분하고 각 구간을 더 세분하여 8개의 일별 패턴을 도출하였다(그림 3, 표 2 참조). 본 연구에서 정의한 운전패턴 8종은 발전설비의 실제 운전 로직과 현장 운전 절차를 반영하여 구분한 것이다. 운전패턴 분류는 단순한 통계적 군집이 아니라, 출력 수준, 부하 변화율, 기동·정지 상태, 제어 신호 조합을 기준으로 정의하였다.

일평균 발전량이 400MW 수준을 유지하는 정상출력 구간과 발전량이 0인 기동정지 구간, 정상출력구간과 기동정지 구간이 상호전환되는 운전전환구간으로 크게 구분하였다. 정상출력 구간은 발전량과 과열기, 복수기, 주증기, 급수기 간의 상관성 강도에 따로 정상운전 구간과 이상운전 구간으로 다시 구분하였다. 이상 운전 구간은 발전량, 주증기 유량, 주증기 압력, 급수 유량 간 상관계수가 0.8 미만으로 떨어지는 구간으로 정의하였다. 본 연구에서는 탐색적 분석 결과를 바탕으로 0.8을 기준값으로 설정하였으나, 이는 절대적인 고정값이라기보다는 분석 대상 설비와 운전 조건에 따라 조정 가능한 경험적 기준이다. 실제로 0.75~0.85 구간에 대한 민감도 분석을 수행한 결과, 0.8 부근에서 정상·비정상 구간의 구분력이 가장 높게 나타났다. 기동정지 구간은 고장 예방을 위한 정지 목적으로 계획적으로 운전을 정지하는 계획 예방정비구간과 고장, 장애 등으로 인하여 기동정지하는 불시 정지구간으로 다시 구분하였다. 운전전환구간은 기동정지상태에서 정상출력 구간으로 상승하는 기동운전 구간과 정상출력에서 기동정지로 떨어지는 정지운전구간, 그리고 정지운전 과정에서 발전량을 줄이는 출력 감발구간으로 구분하였다. 그밖에, 센서나 네트워크 이상으로 데이터의 수집이 이루어지지 않는 통신장애구간도 존재한다.

2) 운전패턴별 비중과 MTBF, MTTR

분석기간에서 불시정지가 발생한 횟수는 총 15회이고 발생한 일수는 총 27일로서 전체의 7.4%를 차지한다.

MTBF는 450.8h(18.8일), MTTR은 79.2h(3.3일). 즉 운전시작 후 평균적으로 18.8일 후 기동정지가 발생하고 평균 수리기간은 3.3일이다.

3) 고장예측인자의 발견

정상운전구간에서는 발전량과 주증기 유량, 발전량과 주증기 압력, 발전량과 급수 유량 간 상관계수는 1에 근접하지만, 불시정지가 발생하기 2~3일 전 발전량과 주증기 유량, 발전량과 주증기 압력, 발전량과 급수 유량 간 상관계수가 0.8 미만으로 떨어지는 현상이 발생한다(그림 4 참조). 따라서 정상적인 발전출력(380MW 이상)을 생성하면서도 주요변수 간 상관계수가 낮은 구간을 이상 운전 구간으로 따로 분류하여 중점적으로 분석하는 것이 필요하다.

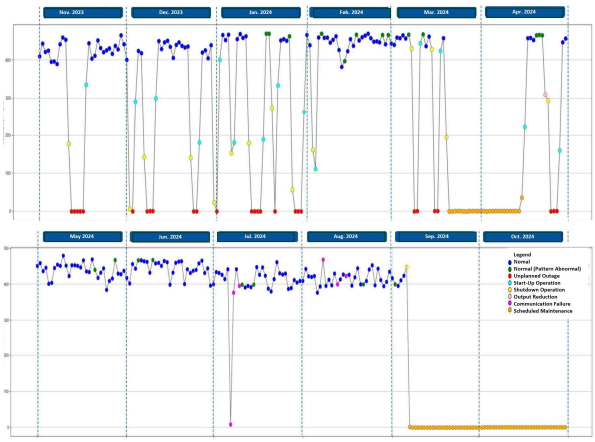


그림 3. 일별 운전패턴
Fig. 3. Daily operating pattern

표 2. 일별 운전패턴 비중
Table 2. Daily operating pattern distribution

Category	Days	Ratio
Normal Operation	201	54.9%
Abnormal Operation	22	5.7%
Shutdown Operation	15	4.1%
Start-up Operation	14	3.8%
Unplanned Outage	27	7.4%
Scheduled Maintenance	81	22.1%
Output Reduction	1	0.3%
Communication Failure	6	1.7%

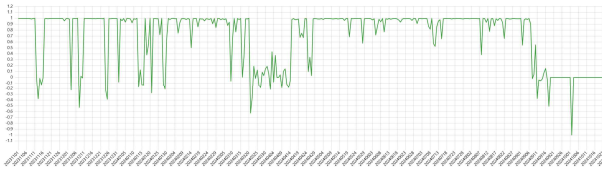


그림 4. 발전량과 주증기 유량 간 상관계수 연간 추이
Fig. 4. Annual trend of the correlation coefficient between power generation and main steam flow

2024년 1월 21일 불시정지 발생 2~3일 전 발전량과 주증기 유량 간 상관관계가 0.55와 0.38까지 하락하는 현상이 발생하였다(그림 5 참조). 발전기 배관 파열이 일어난 4월 22일 이전 3일 동안 발전량과 주증기 유량 간 상관계수가 0.67과 0.74 사이로 떨어지는 현상이 발생하였다(그림 6 참조).

이상 운전이 발생한다고 해서 꼭 불시정지가 일어나는 것은 아니지만 불시정지가 일어났다면 이상 운전이 있었을 확률은 54%이다. 특히 이상 운전이 지속한 기간이 길어질수록 불시정지의 발생 가능성은 매우 커진다.

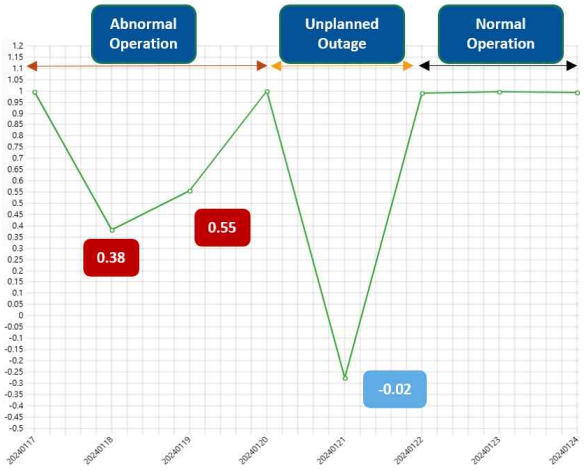


그림 5. 발전량과 주증기 유량 간 상관계수 추이 (2024. 1.17 ~ 2024. 1. 24)

Fig. 5. Trend of the correlation coefficient between power generation and main steam flow (Jan. 17, 2024 - Jan. 24, 2024)

그림 6. 발전량과 주증기 유량 간 상관계수 추이 (2024. 4. 19 ~ 2024. 4. 28)

Fig. 6. Trend of the correlation coefficient between power generation and main steam flow (Apr. 19, 2024 - Apr. 28, 2024)

그림 5 및 그림 6에서 확인할 수 있듯이, 불시정지가 발생한 사례에서는 고장 발생 시점 기준 약 2~3일 이전부터 발전량과 주증기유량 간 상관계수가 급격히 하락하는 현상이 반복적으로 관측되었다. 본 연구에서 적용한 이상탐지 모델은 이러한 상관구조 변화가 시작되는 시점에서 이상 점수를 상승시키며, 실제 고장 발생 시점 대비 약 24~48시간 이전에 위험 상태를 탐지하였다.

정상구간과 이상 구간에서 각 변수의 분포를 비교한 결과 발전량, 급수 유량, 주증기 압력, 주증기 유량의 차이가 두드러진다. 정상운전 구간인 2024년 2월 16일과 이상 운전 구간인 2024년 4월 20일을 비교 분석하였을 때, 발전량, 급수 유량, 주증기 압력, 주증기 유량은 큰 분포의 차이가 발생하고 있다. 비교분석을 통해, 발전량, 급수 유량, 주증기 유량, 주증기 압력의 분포와 이 변수 간의 상관계수가 고장예측의 주요 인자로 파악되었다(그림 7 참조).

그림 7. 정상구간 (2024/02/16)과 이상 구간 (2024/01/18)간 비교분석

Fig. 7. Comparative analysis between normal operation period (Feb. 16, 2024) and abnormal operation period (Jan. 18, 2024)

Ⅳ. 이상탐지 모델 개발

4-1 학습 데이터셋 생성

본 연구에서는 시간 기반(Time Window) 구간화 기법을 적용하여 원본 데이터를 시계열 특성에 따라 구간화하고, 슬라이딩(Sliding) 방식으로 연속적인 분석 단위를 생성하였다. 이를 통해 기술분석, 이상치 탐색, 및 가혹도 모델 개발을 위한 데이터셋을 체계적으로 구축하였다. 분석 대상 데이터는 2023년 11월부터 2024년 10월까지의 1년간 수집된 시계열 데이터를 활용하였다.

우선, 기술 분석(Technical Analysis) 단계에서는 데이터의 전반적 경향과 특성을 파악하기 위해 윈도우 크기(window size)를 1일(1 day), 이동 크기(shift size)를 1일(1 day)로 설정하였다. 이 설정을 통해 총 366개의 구간 데이터를 생성하였으며, 각 구간은 하루 단위의 데이터 변동을 반영하도록 구성하였다.

다음으로, 이상치 탐색(Anomaly Detection) 및 가혹도 모델(Harshness Model) 개발 단계에서는 더욱 세밀한 시간 단위의 패턴 변화를 탐지하기 위하여 윈도우 크기를 86,400초(1일, window size: 86,400s), 이동 크기를 3,600초(1시간, shift size: 3,600s)로 설정하였다. 이와 같은 세밀한 슬

라이싱 기법을 적용함으로써 총 8,784개의 구간 데이터를 생성하였다(그림 8 참조).

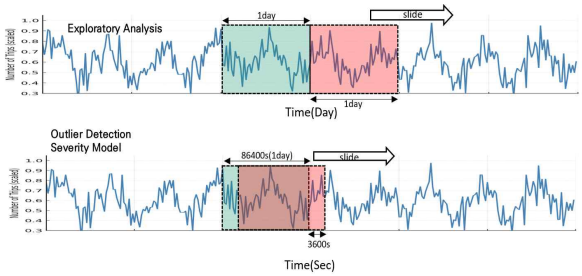


그림 8. 기술분석, 이상치탐색, 가혹도모델 학습을 위한 데이터셋 생성

Fig. 8. Data set construction for exploratory analysis, outlier detection, and severity-model training

전체 데이터 중 70%는 학습데이터로 30%는 검증 데이터로 분리하여, 모형 학습을 수행하였다. 단일 분할에 따른 결과의 우연성을 줄이기 위하여 5겹 교차검증을 하였다. 각 반복 분석에서 회귀계수 및 모형 적합도의 일관성을 확인함으로써 분석결과의 안정성을 검증하였다. 데이터 누수방지를 위하여 모형 추정과 변수 선택 과정은 학습데이터에서만 수행하였고 검증 데이터에는 학습 단계에서 확정된 구조만을 적용하였으며, 검증 데이터 정보가 사전 분석에 사용되지 않도록 절차를 엄격히 통제하였다.

물리적으로 불가능한 값은 사전 규칙 기반으로 제거하였다. 센서 통신 오류 또는 기록 누락으로 발생한 결측치는 결측 구간 길이에 따라 차등 처리하였는데 단기 결측(연속 3초 이내)은 선형 보간법을 적용하였으며 장기 결측 구간은 분석 대상에서 제외하였다.

4-2 Isolation Forest 기반 이상탐지 모델

기술분석에서 중요하게 파악된 주요변수들의 특징(상관관계, 변동성, 수준 등)을 추출하여 Isolation forest 모델에 기반하여 이상탐지 모델을 학습하였다. 모델의 하이퍼파라미터로서 앙상블 트리갯수, 각 트리 학습에 사용되는 샘플의 수, 전체 데이터 중 이상치 비율, 각 분기(Split)에서 사용하는

표 3. 기술분석과 Isolation 이상탐지 결과 비교

Table 3. Comparison between exploratory analysis and isolation-based anomaly detection results

Abnormal Operation Dates Identified by Exploratory Analysis	Isolation-Based Anomaly Detection Model (Anomaly Ratio: 5%)	Isolation-Based Anomaly Detection Model (Anomaly Ratio: 10%)
2024-01-18	Abnormal	Abnormal
2024-01-19	Abnormal	Abnormal
2024-02-06	Abnormal	Abnormal
2024-02-14	Abnormal	Abnormal
2024-02-18	Normal	Normal
2024-02-27	Abnormal	Abnormal
2024-02-29	Abnormal	Abnormal
2024-03-10	Abnormal	Abnormal
2024-04-19	Normal	Abnormal
2024-04-20	Normal	Abnormal
2024-04-21	Normal	Abnormal
2024-05-20	Normal	Abnormal
2024-05-28	Abnormal	Abnormal
2024-06-10	Normal	Abnormal
2024-06-30	Abnormal	Abnormal
Summary	Matching Ratio of Abnormal Cases: 9/15 = 60%	Matching Ratio of Abnormal Cases: 14/15 = 93.3%

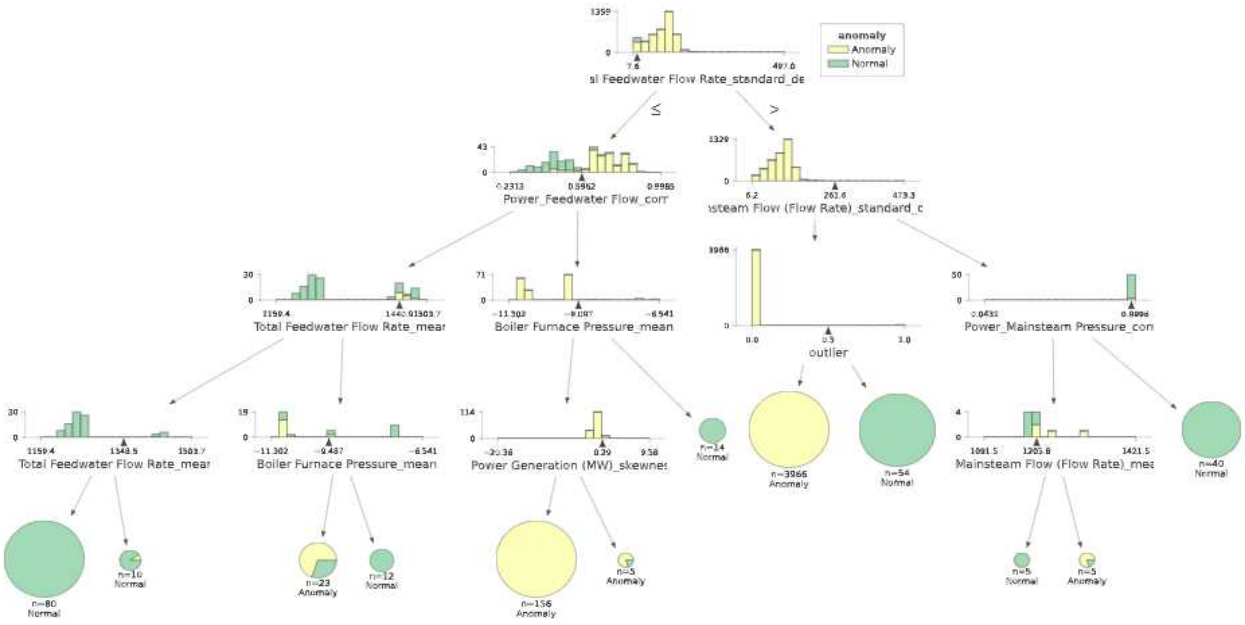


그림 9. Isolation Forest 적용결과의 Decision Tree

Fig. 9. Decision Tree of Isolation Forest application results

feature의 수 조합을 변경하면서 최적의 모델을 탐색하였다.

기술분석의 결과와 이상탐지 결과 간 비교를 하면 표 3과 같다. 분석 대상인 불시정지 15건 중 14건에서 이상탐지 모델의 Anomaly Score가 고장 발생 시점 기준 최소 24시간 이전에 임계값을 초과하였으며, 이 중 8건에서는 48시간 이전부터 위험 신호가 지속적으로 관측되었다.

Isolation Forest 모델은 그림 9와 같이 결정 나무(Decision Tree) 형태로 표현되어 Anomaly 즉 이상 운전이 발생할 수 있는 조건들과 그 확률을 구할 수 있다. 예를 들어, 급수 유량의 표준편차가 7.6 이하이고 전력량과 급수 유량 간 상관계수가 0.39 보다 크고, 보일러 노내 압력이 -8 이하이면서 발전량의 왜도가 0.29 이하이면 100% Anomaly로 분류된다. 또한 급수 유량의 표준편차가 7.6을 초과하고 주증기 유량의 표준편차가 261.6보다 작고 outlier 계수가 0.3 보다 작다면 100% Anomaly로 분류된다.

4-3 Auto Encoder 기반 이상탐지 모델

본 연구에서는 AutoEncoder 모델을 활용하여 데이터의 이상 여부를 판단하기 위한 이상치 탐지 기법을 적용하였다. AutoEncoder는 입력 데이터를 저차원(latent space)으로 압축한 후, 이를 다시 원래의 차원으로 복원하는 비지도 학습 기반의 신경망 구조이다. 이 과정에서 입력 데이터와 복원된 데이터 간의 차이를 오차(Error)로 계산하며, 일반적으로 평균제곱오차를 이상치 판단 지표로 활용한다. 정상 데이터의 경우 AutoEncoder가 학습 과정에서 그 특성을 충분히 학습하므로 복원 시 오차값이 낮게 나타난다. 반면, 이상 데이터(Outlier)는 정상 데이터의 패턴과 상이한 특성을 가지기 때

표 4. 기술분석과 AutoEncoder 이상탐지 결과 비교
Table 4. Comparison between exploratory analysis and AutoEncoder-based anomaly detection results

Abnormal Operation Dates Identified by Exploratory Analysis	AutoEncoder Anomaly Detection Model (Anomaly Ratio: 5%)	AutoEncoder Anomaly Detection Model (Anomaly Ratio: 10%)
2024-01-18	Normal	Abnormal
2024-01-19	Abnormal	Abnormal
2024-02-06	Abnormal	Abnormal
2024-02-14	Abnormal	Abnormal
2024-02-18	Normal	Abnormal
2024-02-27	Abnormal	Abnormal
2024-02-29	Abnormal	Abnormal
2024-03-10	Abnormal	Abnormal
2024-04-19	Normal	Abnormal
2024-04-20	Normal	Normal
2024-04-21	Normal	Abnormal
2024-05-20	Normal	Abnormal
2024-05-28	Abnormal	Abnormal
2024-06-10	Abnormal	Abnormal
2024-06-30	Abnormal	Abnormal
Summary	Matching Ratio of Abnormal Cases: 10/15 = 67%	Matching Ratio of Abnormal Cases: 14/15 = 93.3%

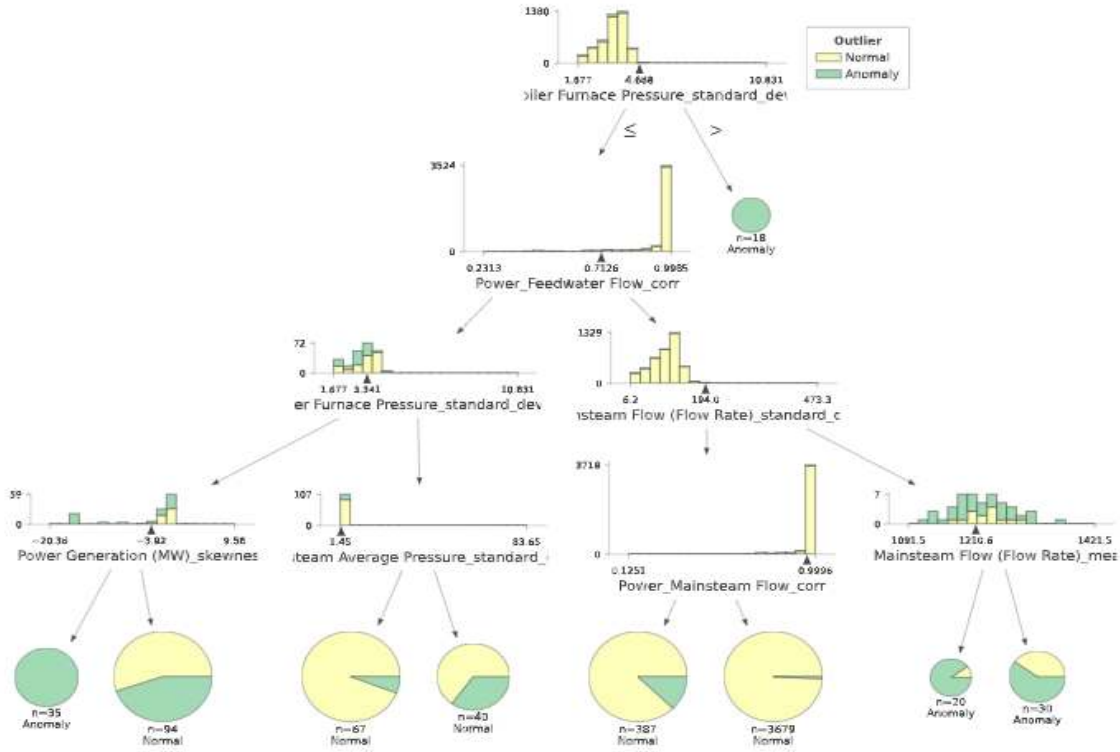


그림 10. AutoEncoder 적용결과의 Decision Tree
Fig. 10. Decision Tree of AutoEncoder application results

문에 복원 시 오차값이 상대적으로 높게 나타나며, 이는 이상치로 판단할 수 있는 중요한 근거가 된다. 따라서 복원 오차(MSE)가 일정 임계값 이상으로 증가한 경우, 해당 데이터를 이상치로 분류할 수 있다. 모델의 하이퍼파라미터로서 은닉층 수, 노드 수, 학습률, 최적화 알고리즘, 손실함수 등의 조합을 변경하면서 최적의 모델을 탐색하였다.

이와 같은 방식으로 도출된 이상 탐지 결과를 기술적 분석 결과와 비교한 결과, AutoEncoder 기반 이상 탐지 모델은 데이터의 비정상적 패턴을 효과적으로 식별할 수 있으며, 기술적 분석에서 제시된 이상 구간과 높은 일치도를 보였다. 이는 AutoEncoder가 데이터의 내재된 비선형적 구조를 학습하여 잠재적 이상을 정량적으로 검출하는 데 유용함을 시사한다. 기술분석의 결과와 이상 탐지 결과 간 비교를 하면 표 4와 같다.

AutoEncoder 모델 역시 아래와 같이 결정나무 (Decision Tree) 형태로 표현되어 Anomaly 즉 이상 운전이 발생할 수 있는 조건들과 그 확률을 구할 수 있다. 그림 10의 결정트리는 AutoEncoder 모델 그 자체의 구조를 나타내는 것이 아니라, AutoEncoder가 판단한 이상 결과를 어떤 시그널 특성함수 조합이 설명하는지를 해석하기 위한 설명용 모델이다.결정트리로의 표현을 통해, 특정 조건 하에서 이상 탐지 확률이 높아지는 규칙을 명시적으로 도출할 수 있다.

한 예로서 노내압력의 표준편차가 4.04 이하이고 발전량과 급수유량의 상관계수가 0.71 이상, 주증기 유량의 표준편차가 194.0 이상이면서 주증기 유량 평균이 1210.6 이하이면 이상치 발생확률은 거의 100%에 이른다는 것을 알 수 있다(그림 10 참조).

본 연구에서는 모델 성능의 신뢰성과 재현성을 확보하기 위하여 성능 지표 선택, 검증 설계, 데이터 누수 방지, 조기 탐지 성과의 평가 기준을 일관된 기준하에 설계하였다. 이상탐지 모델의 성능 평가는 개별 시점의 분류 정확도보다는 실제 고장 이벤트를 사전에 탐지할 수 있는지를 판단하는 것이 핵심 목적이므로, 고장 발생 이전 일정 시간 내에 이상 점수가 임계값을 초과한 경우를 성공으로 정의하는 이벤트 기반 일치율(event-based matching ratio)을 주요 성능 지표로 사용하였다. 데이터 분할은 시계열 특성을 고려하여 시간 순서를 유지한 상태에서 학습 데이터 70%, 검증 데이터 30%로 구성하였으며, 단일 분할에 따른 우연성을 최소화하기 위해 5겹 교차검증을 수행하여 모델 성능의 안정성과 일반화 가능성을 검증하였다. 데이터 누수를 방지하기 위해 특성함수 생성, 변수 선택, 임계값 설정 및 모델 추정 과정은 모두 학습 데이터에 한해 수행하였고, 검증 데이터는 최종 성능 평가에만 사용하였다. 특히 시그널 특성함수는 각 시간 윈도우 내부 데이터만을 이용하여 산출함으로써 미래 시점 정보가 학습 과정에 유입되는 가능성을 차단하였다. 조기 탐지 성과는 이상 점수가 사전 정의된 임계값을 최초로 초과한 시점과 실제 불시정지 발생 시점 간의 시간 차이를 기준으로 산정하였다.

V. 가혹도 예측 모델 개발

본 연구에서는 이상탐지 모델을 통해 일별(Time Window 단위) 구간마다 산출된 Anomaly Score를 설명할 수 있는 주요 변수들을 식별하고, 그 중요도를 분석하기 위하여 가혹도(Severity) 모델을 개발하였다. 가혹도는 시스템이 정상적인 출력 상태를 유지하고 있음에도 불구하고, 발전량과 관련된 유량 및 압력 간의 상관관계가 약화되는 현상이 발생하는 경우를 의미하며, 이는 설비가 정상 범위 내에서 작동하더라도 내부적으로는 발전기를 과도하게 가동(이상운전) 하고 있을 가능성이 있다는 가설에서 출발하였다. 따라서 본 연구의 가혹도 모델은 단순한 이상 탐지 결과의 재현에 그치지 않고, 이상운전의 원인을 변수 간 관계 약화의 관점에서 해석하고 정량화하기 위한 목적으로 설계되었다. 즉, Anomaly Score의 변동을 가장 잘 설명하는 인자들을 도출함으로써, 시스템 내 잠재적 비정상 운전 패턴의 근본 원인 규명과 설비 상태 예측의 정밀화를 가능하게 하였다. 이와 같은 접근은 단순한 이상치 탐지(Detection)를 넘어, 이상현상의 원인 진단(Causality Analysis)과 운전 리스크 수준의 정량적 평가(Harshness Index 산출)를 결합한 고도화된 분석 체계를 제시한다는 점에서 의의가 있다.

5-1 학습 데이터셋 생성

초단위 일별 데이터셋에 대하여 Isolation Forest 이상탐지 알고리즘 적용을 통해 얻은 Anomaly Score를 라벨로 활용하여 일별 단위 총 366개의 레코드로 구성된 학습용 데이터셋을 개발하였다(그림 11 참조). 이상탐지 모델개발과 동일하게 전체 데이터의 70%는 학습 데이터로 30%는 검증 데이터로 분리하여, 모형 학습을 수행하였다. 5겹 교차검증을 수행하였고 누수방지를 위해 검증 데이터는 모형 추정이나 변수 선택 과정에서 사용하지 않았다.

5-2 모델 학습

본 연구에서는 가혹도 예측 모델의 학습을 위해 Random Forest, LightGBM, 및 CatBoost 세 가지 머신러닝 회귀 모델을 적용하였다. 각 모델은 시계열 기반의 윈도우 데이터셋을 입력으로 하여, 센서 데이터의 변동 패턴과 이상 신호의 상관관계를 학습하도록 설계되었다. 모델 학습 결과, 세 가지 모델 모두 전반적으로 우수한 예측 성능을 보였다. 성능 평가 지표로 사용된 평균제곱오차(Mean Squared Error, MSE)와 평균절대오차(Mean Absolute Error, MAE) 값은 모두 매우 낮게 나타났으며, 결정계수(R^2 Score)는 0.97에서 0.99 사이의 높은 값을 기록하였다(표 5 참조). 이러한 결과는 각 모델이 입력 변수 간의 비선형적 관계를 효과적으로 학습하여, 가혹도 수준을 정밀하게 예측할 수 있음을 의미한다. 특히, CatBoost 모델이 세 모델 중 가장 높은 예측 정확도를 보였다. 이는 CatBoost가 범주형 변수 처리와 학습 속도 최적

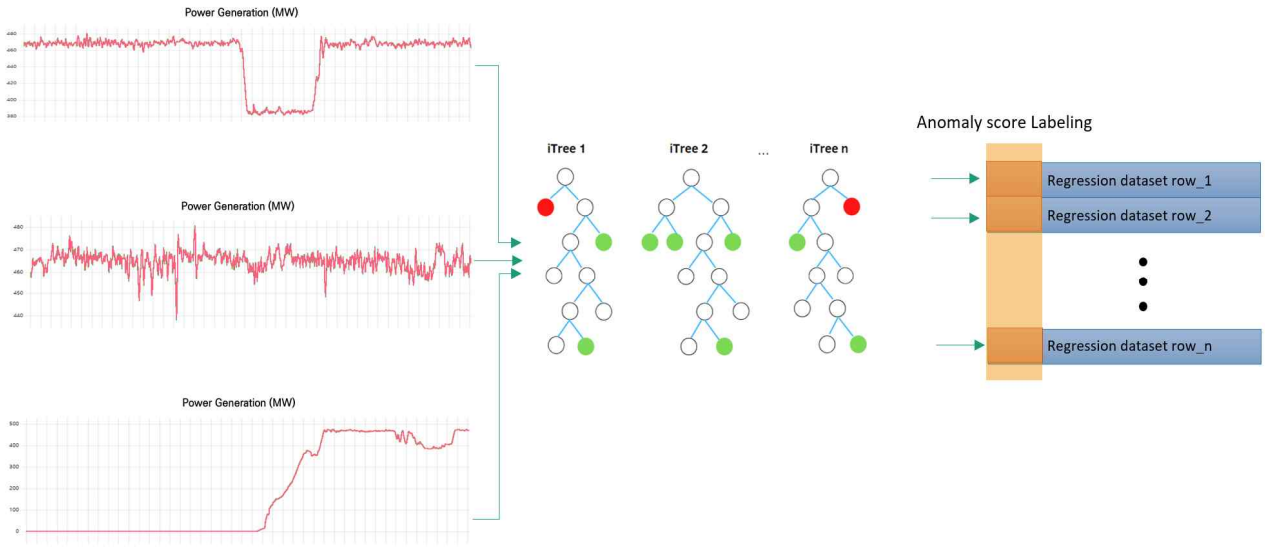


그림 11. 가혹도 모델 개발을 위한 학습 데이터셋 생성
Fig. 11. Training dataset generation for severity model development

화에 강점을 가지며, 과적합(overfitting)을 방지하는 내장 정규화 기법을 포함하고 있기 때문이다. 따라서 본 연구에서는 CatBoost 모델을 가혹도 예측의 최적 모델로 선정하였으며, 향후 가혹도 지수 산출 및 이상 탐지 시스템의 핵심 예측 엔진으로 적용할 수 있음을 확인하였다.

3가지 모델 모두에서 급수량의 표준편차, 발전량의 peak to peak 차이, 주증기량 표준편차를 공통적으로 중요 영향인자로 지목하였다. 즉 발전량, 급수량, 주증기량의기복이 심할수록 발전의 가혹도가 높아지고, 주요변수 간 상관강도가 감소하여 고장으로 이어질 가능성이 높다고 할 수 있다.

표 5. 각 모델의 성능지표
Table 5. Performance metrics of each model

	Random Forest	Light GBM	Cat Boost
MSE	0.0001	0.0001	0.0001
MAE	0.0066	0.0046	0.0041
R^2	0.9746	0.9865	0.9894

Ⅵ. 발전설비 감시 AI 플랫폼의 개발

6-1 아키텍처 설계

본 연구의 시스템 아키텍처는 전 구성요소를 오픈 소스 소프트웨어로만 구성하여 라이선스 비용을 제로화한 것이 핵심 설계 원칙이다(그림 12 참조).

데이터 적재 모듈과 변환 모듈은 모두 Python으로 구현되어 있으며, 작업 간 의존성과 실행 순서를 명시적으로 표현하기 위해 DAG(Directed Acyclic Graph) 구조로 모델링되어 있다. 서비스 계층의 REST API는 Python 기반으로 구현되어 있으며, 경량형 WAS인 Nginx Unit에 배포되어 있다.

Apache Airflow, FastAPI, Nginx Unit을 중심으로 시스템 아키텍처를 구성한 이유는 대규모 시계열 데이터의 안정적 운영, 실시간 분석 확장성, 그리고 장기 운영 환경에서의 유지보수 효율성을 동시에 확보하기 위함이다. 단순한 프로토타입 구현이 아니라, 실제 발전설비 감시 환경에서 지속적으로 운영 가능한 구조를 설계하는 것이 본 연구의 목표이다.

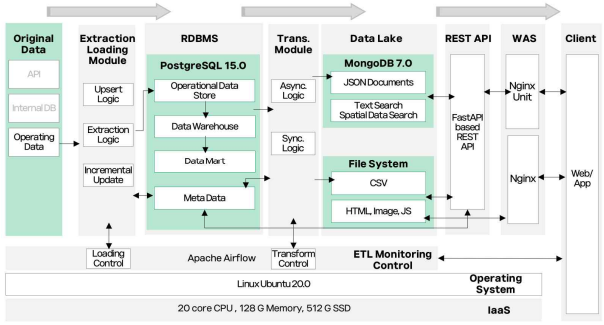


그림 12. 플랫폼 아키텍처
Fig. 12. Platform architecture

6-2 플랫폼의 주요 기능

1) Apache Airflow를 이용한 ETL 모니터링과 제어

Apache Airflow는 초단위 시계열 데이터의 적재·전처리·특성함수 생성·모델 실행 과정을 워크플로우 단위로 명시적으로 관리할 수 있어, 데이터 파이프라인의 재현성과 운영 안정성을 확보하는 데 적합하다. 특히 작업 간 의존성과 실패 지점을 명확히 추적할 수 있어, 발전설비와 같이 데이터 신뢰성이 중요한 환경에서 운영상 장점이 있다(그림 13 참조).

2) FastAPI와 Nginx Unit을 이용한 REST API 개발

FastAPI는 비동기 기반의 경량 API 프레임워크로서, 이상탐지 결과와 가혹도 지수를 실시간으로 외부 시스템 및 시각

화 모듈에 제공하는 데 적합하다. 입력 데이터 검증과 자동 문서화를 기본적으로 지원하여, 운영 중 발생할 수 있는 인터페이스 오류를 최소화할 수 있다. Nginx Unit은 애플리케이션 서버와 웹 서버 기능을 통합한 경량 런타임 환경으로, FastAPI와 결합할 경우 높은 처리 성능과 빠른 배포가 가능하다. 설정 변경 시 서비스 중단 없이 애플리케이션을 재로딩할 수 있어, 무중단 운영과 신속한 모델 업데이트가 필요한 발전설비 감시 환경에 적합하다.

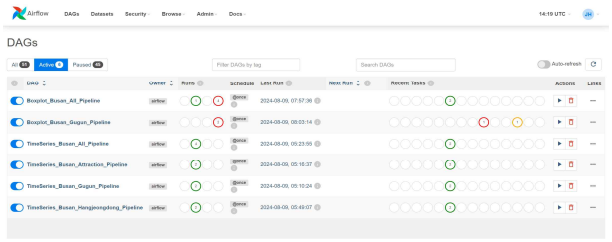


그림 13. Apache Airflow 내 DAGs
Fig. 13. DAGs in Apache Airflow

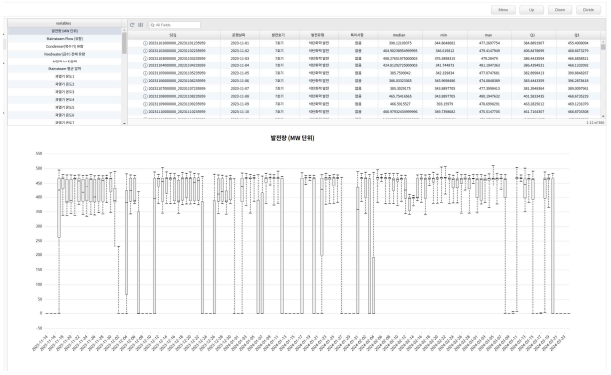
3) 개별구간 분석 기능의 개발

본 연구에서는 2023년 11월부터 2024년 10월까지의 데이터를 하루 단위로 분할하여 구간 리스트를 구성하였다. 각 구간은 색상으로 정상·비정상 패턴을 구분하였으며, 선택된 구간에서 1,000개의 샘플 데이터를 추출하여 시각화하였다. Y-Profiling을 통해 목표 변수의 분포, 이상치, 결측치 등을 분석하고 그래프로 제시하였다. 하루 데이터의 19개 Feature에 대해 시계열 그래프와 박스플롯으로 변수별 변화 추이를 분석하였다. 변수별 평균, 중앙값, 표준편차 등 통계량을 grid 형태로 제공하여 데이터의 중심 경향성을 파악하였다. 단위 차이가 큰 변수는 로그 스케일(그림 14 참조)로 표현하였다.

*The above figure is a computer screenshot and therefore contains Korean text
그림 14. 로그스케일 그래프
Fig. 14. Log-scale graph

4) 추이 분석 기능의 개발

본 연구는 장기 발전소 운영 데이터를 기반으로 변수별, Feature별 추이 분석을 수행하여 성능 저하와 손상을 조기 탐지하는 기법을 제안하였다. 다변량 시계열 데이터를 분석하여 운전 변수 간의 상관성과 변동 추세를 시각화하였다(그림 15 참조). 성능 저하 및 손상 심도를 정량화하기 위해 진단 함수를 설계하고 이상 구간을 탐지하였다. 운전 조건의 차이를 보정하기 위해 가혹도 지수를 산출하여 데이터 패턴을 표준화하였다. 이를 통해 발전설비의 상태를 일관성 있게 평가하고 예지보전 체계의 구축이 가능하다.



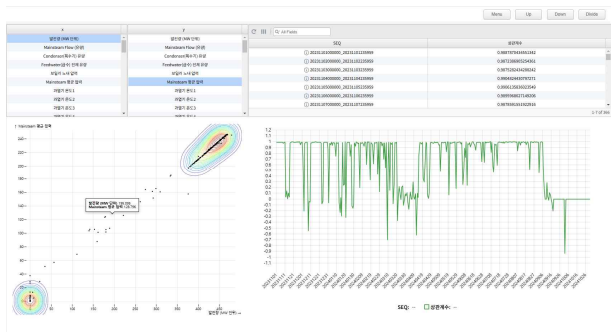
*The above figure is a computer screenshot and therefore contains Korean text
그림 15. Box Plot을 이용한 추이 분석
Fig. 15. Trend analysis using Box Plot

5) 상관 분석 기능의 개발

본 연구에서는 성능 저하나 손상 발생이 확인된 이후, 그 원인을 규명하기 위해 상관계수의 시계열 분석을 수행하였다(그림 16 참조). 시간에 따른 변수 간 상관 강도의 변화를 분석함으로써 고장에 영향을 미치는 주요 요인을 식별하였다. Heatmap 시각화를 통해 다수의 변수 간 상관관계를 직관적으로 파악할 수 있도록 하였으며, Pair Plot 분석을 통해 변수 간 상호작용 및 이상 패턴을 정밀하게 탐색하였다. 이를 통해 엔지니어는 고장 원인 진단 및 문제 해결의 실마리를 효과적으로 도출할 수 있다.

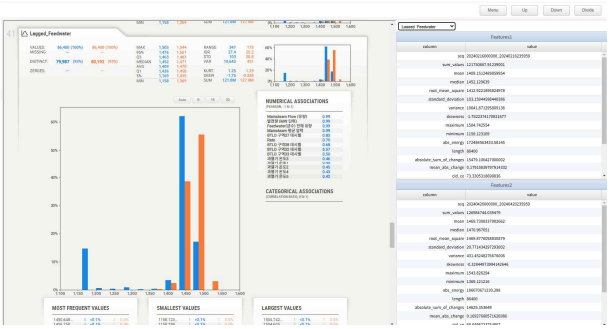
6) 비교 분석 기능의 개발

본 연구에서는 기동 간, 기동 내 구간 간, 또는 특정 조건을 만족하는 운전 요소를 샘플링하여 상호 비교함으로써 성능



*The above figure is a computer screenshot and therefore contains Korean text
그림 16. 상관계수 시계열 분석 기능
Fig. 16. Time-series analysis function of correlation coefficient

저하의 원인과 개선 가능성을 분석하였다(그림 17 참조). 하드웨어적·소프트웨어적 운전 성능 개선 후, 개선 전후의 주행 시험 데이터를 비교하였다. 이를 통해 개선 조치의 효과를 정량적으로 평가하고, 성능 향상 메커니즘을 검증하였다. 본 분석은 발전설비의 기능 최적화와 예지정비 전략 수립에 유용한 근거를 제공한다.



*The above figure is a computer screenshot and therefore contains Korean text

그림 17. 비교분석 기능의 개발

Fig. 17. Comparative analysis function

Ⅶ. 결 론

7-1 연구수행 요약

본 연구는 석탄발전소 1개 호기의 발전 데이터를 활용하여 이상탐지 및 성능진단을 위한 AI 기반 분석체계를 구축하였다. 2023년 11월부터 2024년 10월까지 1초 단위로 수집된 3,162만여 건의 데이터(54개 변수)를 분석하여 8가지 일별 패턴을 도출하였다. 정상운전 후 약 18.8일 만에 불시정지가 발생하고, 평균 3.3일의 수리 기간 후 복귀하는 경향이 나타났다. 발전량과 주증기 유량·압력·급수유량 간 상관계수가 1에 근접하나, 고장 2~3일 전 0.8 미만으로 하락하여 상관계수 변화가 주요 고장예측 인자로 확인되었다. 이상탐지 모델로 Isolation Forest와 AutoEncoder를 적용한 결과, 이상치 일치율 93.3%로 높은 정확도를 보였다. 가혹도 모델은 Anomaly Score를 라벨로 부여하고 Random Forest, CatBoost, LightGBM 회귀모델을 적용하여 R^2 0.97~0.99의 우수한 성능을 보였다. 급수량 표준편차, 발전량, peak to peak 차이, 주증기량 표준편차가 공통 중요 변수로 도출되어 변동성 증가 시 고장 발생과의 인과관계가 규명되었다. 플랫폼은 Apache Airflow와 Nginx Unit 기반 오픈소스로 설계되어 라이선스 비용을 제로화하였다. CSV 데이터를 전처리하여 샘플링, 프로파일링, 시계열 분석, 상관계수 산출 등을 수행하였다. 구간별 추이, 통계량, 로그·선형 스케일, 상관분석 및 Pairplot 기능을 통해 데이터 패턴과 이상을 시각적으로 탐색할 수 있다. Heatmap 기반 상관추이 분석을 통해 고장 영향 요인을 규명하고, Pairplot으로 문제 해결 단서를 시각화하였다. 기동 간 비교분석을 통해 성능 저하 원인과 개선 효과를 검증하였다. 결과적으로 본 연구는 발전설비의 상태를 실시간 모니터링하고, 이상징후를 조기 탐지하여 예지보전 체계를 구현할 수 있는 AI 기반 플랫폼을 제시하였다.

7-2 연구결과의 우수성

표 6은 AutoEncoder, FFT 기반 특성추출, TadGAN을 활용한 기존 이상탐지 연구들과 본 연구의 성능을 정량적으로 비교한 것이다. 기존 연구들은 단변량 기반 재구성 오차나 주파수 특성에 의존하는 방식이 많아 고장 전조 신호를 장기간 이전에 포착하는 데 한계가 있었으며, 실제 고장 발생 0~12시간 이전의 짧은 탐지 구간에서만 이상을 확인하는 경우가 대부분이었다. 정확도 역시 75~87% 수준에 머물러, 조기경보체계로서의 활용성에는 제약이 있었다. 반면, 본 연구는 다변량 운전변수들 간 상관구조의 시간적 붕괴를 핵심 지표로 활용함으로써 고장 발생 24~48시간 이전의 전조 패턴을 탐지할 수 있었으며, 이는 기존 연구 대비 약 4~8배 향상된 조기 탐지 성능이다. 또한 이상탐지 정확도는 93.3%로 기존 대비 약 10~13%p 높은 성능을 보였다. 종합적으로 본 연구가 기존 이상탐지 연구에 비해 조기 탐지 능력, 탐지 정확도, 전조 패턴 재현성 등 주요 지표에서 모두 우수한 성능을 보임을 명확하게 보여주고 있다.

7-3 향후 연구과제

본 연구는 발전설비의 이상탐지 및 예지보전을 위한 AI 기반 감시 플랫폼의 설계와 구현을 통해 성능 저하 및 고장을 조기 예측할 수 있는 기술적 가능성을 입증하였다. 그러나 실시간 예측 정밀도 향상과 사용자 중심의 운영 환경 고도화를 위해 다음과 같은 후속 과제가 필요하다.

1) 예측모델 성능 향상

향후에는 데이터 품질과 모델의 예측 정밀도를 동시에 개선하기 위한 통합적 접근이 필요하다. 우선, 데이터 전처리 및 품질 향상 측면에서 노이즈 제거, 이상치 보정, 결측값 보완 등의 정제 과정을 강화하고, 데이터 증강(Augmentation)을 통해 정상·비정상 구간 간 데이터 불균형 문제를 해소할 것이다. 또한 시계열 데이터의 연속성과 상관 구조를 반영하기 위한 정교한 전처리 기법을 적용할 예정이다. 다음으로, 고도화된 특성함수(Feature Functions)를 설계하여 데이터의 내재된 패턴을 세밀하게 반영하고자 한다. 평균, 표준편차뿐 아니라 왜도, 첨도, 주파수 도메인, 동적 상관계수 등의 고급 통계함수를 추가하고, 발전설비의 물리적 특성을 반영한 도메인 지식 기반 지표를 확장할 필요가 있다.

AI 모델 고도화를 위해서는 LightGBM, XGBoost, Random Forest 등 앙상블 기법에 대한 체계적인 하이퍼파라미터 최적화(Grid Search, Random Search)를 수행하고, CNN, LSTM, Transformer 등 시계열 분석에 특화된 딥러닝 구조를 적용하여 1~2일 전 고장 예측 모델을 개발할 계획

이다. 또한 전이학습(Transfer Learning)을 통해 학습 효율성을 높이고, 유사 설비 간 모델 이식성을 강화하는 것이 필요하다. 운영 측면에서는 실시간 데이터 수집 및 피드백 루프를 구축하여 모델 예측 결과를 지속적으로 보정하고, Canary Deployment 방식으로 무중단 모델 업데이트 체계를 확립할 예정이다. 모델 성능 평가는 정확도, 정밀도, 재현율, F1-score, AUC 등 다중 지표를 활용하며, 교차 검증을 통해 모델의 일반화 성능을 정량적으로 검증할 예정이다.

표 6. 선행연구와 본 연구의 성능 비교
Table 6. Comparative performance between previous studies and the present study

Research	Anomaly Detection Performance	Accuracy	Characteristics
Sakurada & Yairi (2014), AutoEncode [8]	Detects anomalies 6–12 hours before failure	80–85%	Limited ability to capture multivariate relationship changes; primarily single-variable patterns
Tzanetakis & Cook (2002), FFT based Feature Engineering [18]	Detects anomalies 0–6 hours before failure	75–82%	Effective for vibration patterns but weak in identifying early precursor signals
Lee & Kim (2022), TadGAN based Anomaly Detection [3]	Detects anomalies at approximately 12 hours before failure	83–87%	Detects anomaly spikes but exhibits limited capability in capturing long-term precursor patterns
Present Study	**Detects precursor patterns 24–48 hours before failure (4–8× improvement over prior studies)**	**93.3%(≈ +10–13%p improvement compared to previous work)**	**Captures dynamic breakdown of multivariate correlation structures; early-warning capability**

2) 플랫폼 기능 개선

플랫폼의 실효성과 사용성을 높이기 위해 대시보드의 시각화 및 사용자 경험(UX) 개선이 필요하다. 이상 상태, 기동정지 위험도 등을 게이지, 히트맵, 트리맵, 선형 그래프 등 다양한 시각 요소로 표현하고, 위험 수준별 색상 코드(녹색–정상, 주황–경고, 빨강–위험)를 적용하여 직관적 모니터링이 가능하도록 할 예정이다. 또한 사용자가 필요에 따라 위젯을 자유롭게 구성할 수 있는 대시보드 커스터마이징 기능을 제공하는 것이 필요하다. 사용자 경험 향상을 위해 실시간 알림 및 피드백 시스템을 강화하고, 인터랙티브 튜토리얼·도움말·툴팁 등을 통해 접근성을 개선할 계획이다. 실시간 데이터 갱신과 드릴다운(Drill-down) 분석 기능을 결합하여, 이상 발생 시 원인 변수 및 상관 관계를 즉시 탐색할 수 있는 인과 기반 분석환경을 구축해야 한다. 또한 사용자 역할에 따라 UI를 차별화하고, 다국어 지원과 개인화된 알림 설정을 통해 글로벌·현장 맞춤형 운영을 지원한다. RESTful API를 통한 시스템 통합 및 확장성 확보, 클라우드 기반 UI 배포를 통한 유연한 업데이트도 병행 추진해야 한다. 보안성 강화를 위해 다중 인증(MFA), 접근 권한 제어, 사용자 행동 로그 및 감사 기록 관리 기능을 확립하고, 실시간 피드백 루프를 통해 사용자 의견을 반영한 지속적 UI/UX 개선을 수행해야 한다. 종합적으로, 향후 연구는 AI 예측모델의 정밀화와 실시간 운영 플랫폼의 지능화를 병행하여, 발전설비의 자율 진단·예지보전 체계를 완성하는 방향으로 확장되어야 한다.

참고문헌

[1] M. Kim and K. Jin, “Development of a Deep Learning Algorithm for Anomaly Detection of Manufacturing Facility,” *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 26, No. 2, pp. 199-206, 2022. https://doi.org/10.6109/jkiice.2022.26.2.199

[2] Y. Jeong and Y. S. Kim, “Analysis of Domestic Research Trends on Artificial Intelligence-Based Prognostics and Health Management,” *Journal of Korean Society for Quality Management*, Vol. 51, No. 2, pp. 223-245, 2023. https://doi.org/10.7469/JKSQM.2023.51.2.223

[3] S. Lee and Y. Kim, “A Pre-Processing Process Using TadGAN-Based Time-Series Anomaly Detection,” *Journal of Korean Society for Quality Management*, Vol. 50, No. 3, pp. 459-471, 2022. https://doi.org/10.7469/JKSQM.2022.50.3.459

[4] B. D. Fulcher and N. S. Jones, “Highly Comparative Feature-Based Time-Series Classification,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 12, pp. 3026-3037, 2014. https://doi.org/10.1109/TKDE.2014.2316504

[5] S. Lee, H. Shin, P. Tadic, and Z. Durovic, “Applications of Predictive Maintenance Techniques in Industrial Systems,” *Serbian Journal of Electrical Engineering*, Vol. 8, No. 3, pp. 263-279, 2011. https://doi.org/10.2298/SJEE1103263M

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in *Proceedings of the 8th IEEE International. Conference on Data Mining (ICDM)*, Pisa: Italy, pp. 413-422, 2008. https://doi.org/10.1109/ICDM.2008.17

[7] S. Hariri, M. C. Kind, and R. J. Brunner, “Extended Isolation Forest,” arXiv:1811.02141, 2019. https://doi.org/10.48550/arXiv.1811.02141

[8] M. Sakurada and T. Yairi, “Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction,” *in Proceedings of the MLSDA 2014 2nd Workshop on*

Machine Learning for Sensory Data Analysis, Gold Coast: Australia, pp. 4-11, 2014. https://doi.org/10.1145/2689746.2689747

[9] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier Detection with Autoencoder Ensembles," in *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, Houston: TX, pp. 90-98, 2017. https://doi.org/10.1137/1.9781611974973.11

[10] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection," in *Proceedings of the Internation Conference on Learning Representations (ICLR)*, Vancouver: Canada, 2018.

[11] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach: CA, pp. 3146-3154, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal: Canada, pp. 6639-6649, 2018. https://doi.org/10.48550/arXiv.1706.09516

[14] P. Schäfer and U. Leser, "Multivariate Time Series Classification with WEASEL+MUSE," arXiv:1704.08055, 2017. https://doi.org/10.48550/arXiv.1711.11343

[15] B. D. Fulcher and N. S. Jones, "Highly Comparative Feature-Based Time-Series Classification," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 12, pp. 3026-3037, 2014. https://doi.org/10.48550/arXiv.1401.3531

[16] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293-302, 2002. https://doi.org/10.1109/TSA.2002.800560

[17] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ: Wiley, 2015.

[18] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293-302, 2002. https://doi.org/10.1109/TSA.2002.800560

김종호(Jong-Ho Kim)

1994년 : KAIST 경영정책학과 (공학사)
1996년 : KAIST 경영정보공학과 (공학석사)
2003년 : KAIST 경영공학과 (공학박사-빅데이터)
1996년～2003년: 비트컴퓨터
2003년～2006년: 삼성SDS
2006년～2008년: 가톨릭대학교 연구조교수
2008년～2011년: 현대경제연구원
2011년～현 재: 경성대학교 경영학과 부교수, 주식회사 하이퍼로직 대표
※관심분야 : 빅데이터, 인공지능, 시스템 분석 및 설계