용역 결과 보고서

딥러닝 기반 다변량 스팀 사용 이상 감지 및 영향변수의 원인 분석 기능 제작

2023. 11. 28.

서울과학기술대학교 데이터사이언스학과

연구책임자 심재웅

목차

1. 용역 개요

1.1. 주요 내용

- 제지 공정 건조 설비에서 수집되는 다수의 시계열 센서 데이터를 활용하여 제품 생산을 반영한 스팀 사용량 이상 혹은 제품 생산 이상을 사전에 감지하고 대응하기 위한 딥러닝 기반 모델을 개발

1.2. 세부 내용

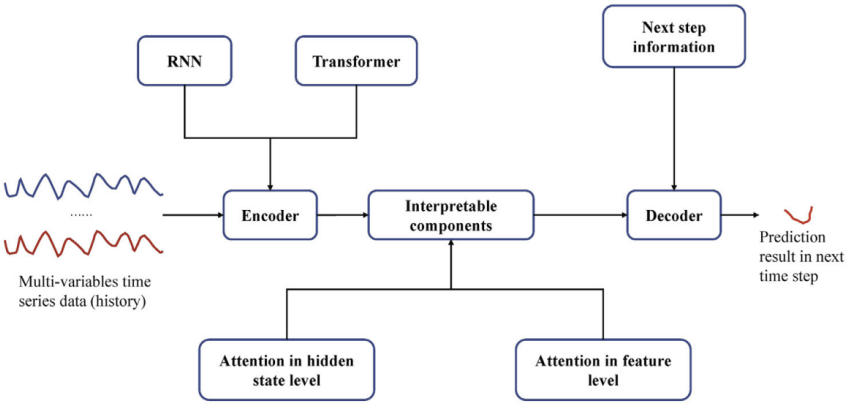
- 본 과제와 관련된 기존 연구에 대한 분석을 수행함. 응용 분야 측면에서 에너지 사용량 예측 관련 연구를 조사하였고, 모델링 방법론 측면에서 다변량 시계열 데이터에 대한 설명 가능한 예측 모델 관련 연구를 조사함.
- 제공된 데이터셋에 대해 스팀 에너지 사용량 예측 모델을 구축하고 그 성능을 비교 분석 수행함. 다음의 총 4종의 모델 (일반 머신러닝 모델 1종과 딥러닝 기반 모델 3종)을 구축하여 그 성능을 비교하였음.
    - Random Forest (일반 머신러닝 모델)
    - 1D Convolutional neural network 모델
    - Long short-term memory 모델
    - Dual attention 모델
- 각 구축된 모델에 대해 적절한 설명 가능 인공지능(XAI) 방법론을 적용하여, 모델의 예측에 주요한 영향을 미친 원인 인자를 탐색함. 다음의 3가지 방법을 각 모델에 적용하여 예측 원인 인자를 파악하여 비교하였음.
    - Feature importance (Random Forest)
    - SHAP (1D Convolutional neural network, Long short-term memory 모델)
    - Attention score (Dual attention 모델)

2. 관련 연구 소개

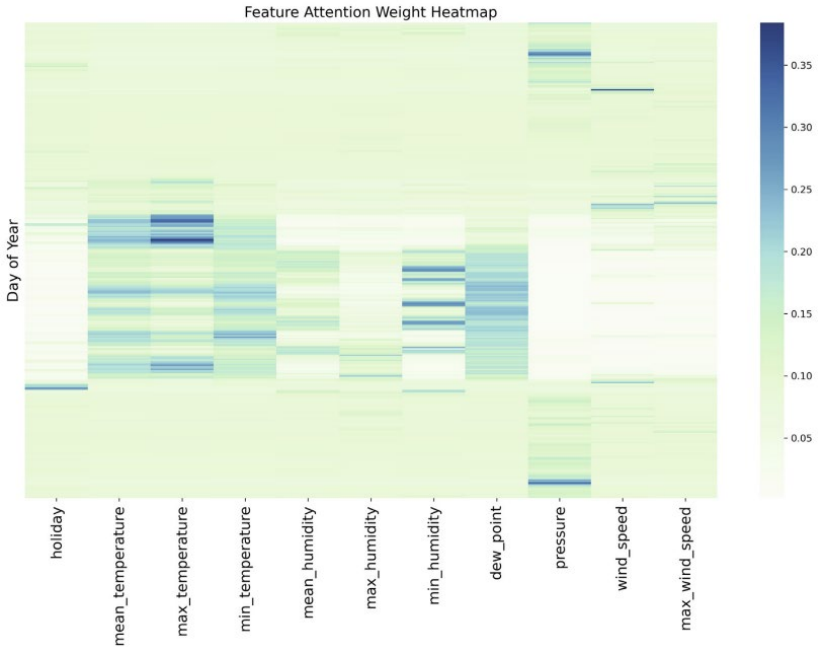
2.1. 에너지 사용량 예측 응용 관련

- **Interpretable deep learning model for building energy consumption prediction based on attention mechanism (Gao, Yuan, and Yingjun Ruan, Energy and Buildings, 2021)**

■ 사무용 빌딩의 에너지 사용량을 예측하는 모델을 구축하고, 예측 주요 변수를 찾아내는 연구. 방법론의 전체 흐름은 다음 그림과 같음.



■ LSTM과 Transformer 구조를 활용하였으며, 내부의 Attention 스코어를 통해 예측 원인 인자를 해석하고자 함. 도출된 Attention 스코어 예시는 다음과 같으며, Max temperature, dew point temperature 등이 주요 변수로 도출되었음.

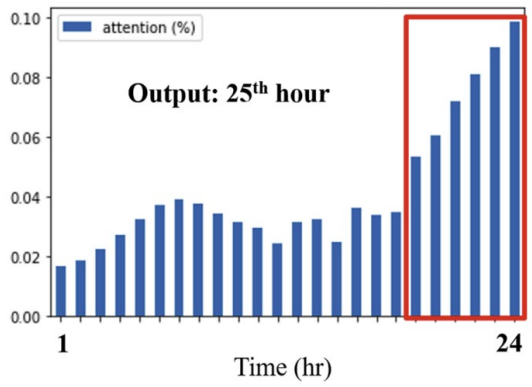


- **Attention-based interpretable neural network for building cooling load prediction (Li, Ao, et al., Applied Energy, 2021)**

■ 상업용 빌딩에서 다음 날의 냉방 부하를 사전에 예측하는 모델에 대한 연구.

■ 시간 정보, 외부 습도, 온도, 태양 복사량 등을 예측에 활용하였음.

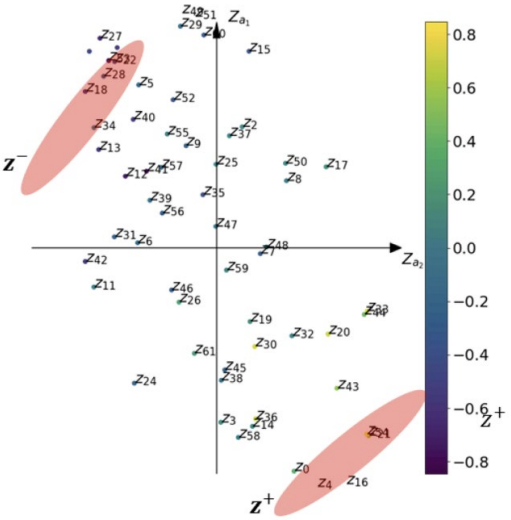
■ 시간 축으로의 attention score를 통해, 한시간 후의 cooling load 예측에 영향을 크게 미친 과거 cooling load를 시각화 하였음.



- **Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space (Kim, Jin-Young, and Sung-Bae Cho, Expert Systems with Applications, 2021)**

  - Deep autoencoder 구조를 기반으로 가정용 전기 에너지 수요를 예측하는 모델에 대한 연구.

  - Encoder를 통해 만들어진 latent space 상에서의 correlation을 통해, 에너지 수요 예측의 주요한 변수를 도출하고자 함. 학습된 latent space간의 correlation을 다음과 같이 시각화하여 해석하고자 하였음.



2.2. 설명가능한 시계열 예측 모델 관련

- **A dual-stage attention-based recurrent neural network for time series prediction (Qin, Yao, et al., Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017)**

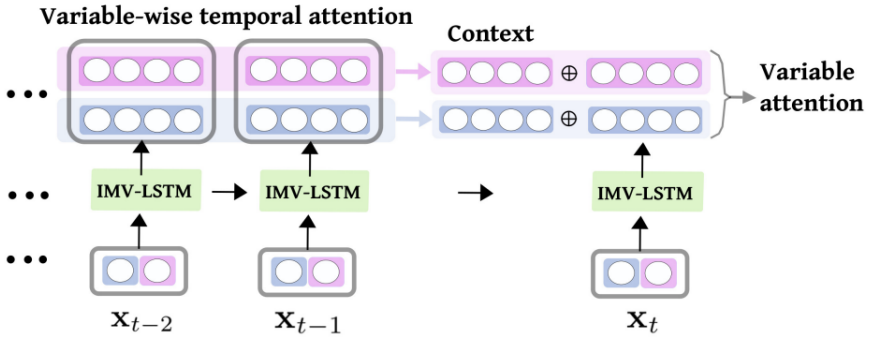
  - 본 과제의 스팀 에너지 사용량 예측에 활용된 모델 구조가 처음 제안된 연구.

  - LSTM과 함께 2단계의 attention mechanism을 사용함으로써, long-term temporal

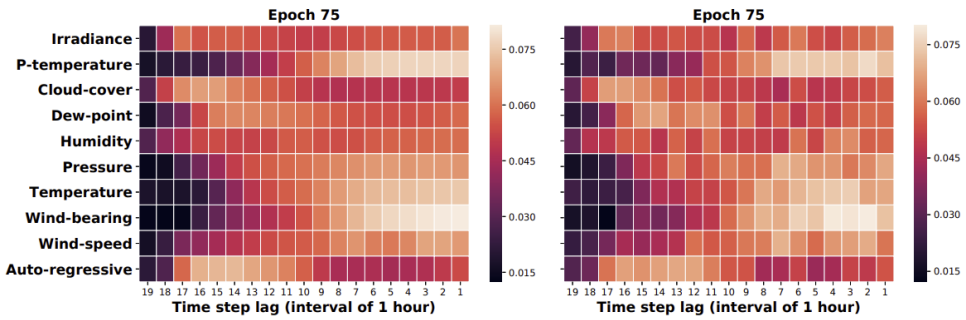
dependency 문제를 해결할 수 있게 하였고, 여러 변수 중 주요 관련 변수를 모델 자체적으로 선택하여 예측에 활용할 수 있도록 하였음.

- 1단계 attention: 주요 관련 변수를 선택적으로 예측에 활용할 수 있게 하고자 도입.
- 2단계 attention: 모든 시점의 정보를 충분히 고려할 수 있게 하고자 도입.
- 해당 연구에서는 NASDAQ 100 주식 데이터와 가정집 실내 온도 데이터셋을 활용하여 검증을 수행하였음.

- **Exploring interpretable LSTM neural networks over multi-variable data (Guo, Tian, Tao Lin, and Nino Antulov-Fantulin, International conference on machine learning, 2019)**
  - 변수 별 움직임을 개별적으로 반영할 수 있도록 변수 별로 hidden state를 학습하는 Interpretable multi-variable LSTM (IMV-LSTM) 구조를 제안함.
  - Mixture attention을 통해 모델의 해석이 가능하도록 하였음. 다음 그림과 같이 개별 변수 별 temporal attention과 이 후의 variable attention을 포함하고 있음.

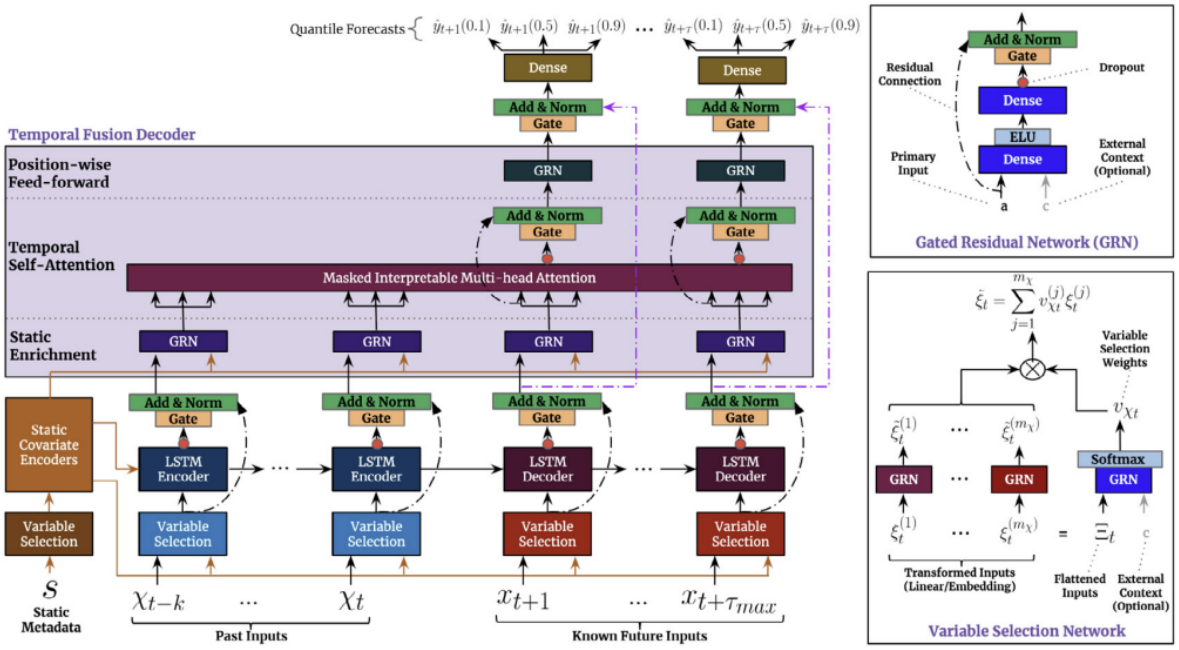


  - 변수 별 temporal attention을 활용하여 변수별/시점별 예측 기여도를 나타낼 수 있으며 예시는 다음과 같음.

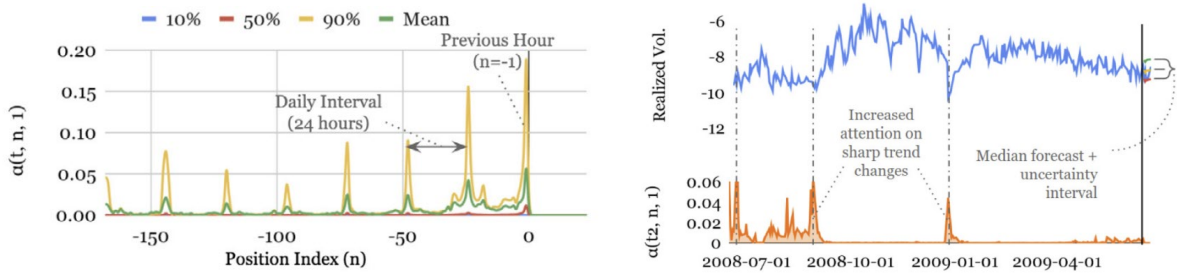


- **Temporal fusion transformers for interpretable multi-horizon time series forecasting (Lim, Bryan, et al., International Journal of Forecasting, 2021)**
  - Transformer를 기반으로 높은 성능의 예측 성능을 보이는 Temporal Fusion Transformer (TFT)를 제안한 연구.

- Recurrent layer와 self-attention layer를 활용하여 long-term dependency 문제를 극복하고 주요 시점에 대한 interpretability를 확보함.



- 모델 해석에 있어서 주요 use case로는, 변수 중요도 도출, seasonality와 같은 persistent 패턴 도출 (좌측 그래프), 금융위기와 같이 큰 변화를 일으키는 주요 사건 식별 (우측 그래프)을 예시로 제시함.



3. 활용 데이터

3.1. 데이터 소개

- 제조 공정: 제지 생산 공정 중 스팀을 이용한 열처리 공정
- 수집 기간: 2023년 4월 ~ 2022년 8월 (약 5개월)
- 수집 주기: 센서값 주기 사이클 타임 약 1분

date	tg04	tg05	tg02	tg06	tg10	tg09	tg12	tg11	tg14	tg03	tg13	tg17	tg33	tg26	tg34	tg35	tg36	tg22	tg37	tg38	tg23	tg24
2023-04-05 14:59:00	18.41245	4.443732	31.56803	4.710918	1092.585	62.62302	1208.394	72.93507	62.53452	1319.3	60.6363	578695.1	3.028413	75.1873	42.88396	31.84253	49.46822	605.2247	475.111	1000.628	1212.661	1148.518
2023-04-05 15:00:00	18.22477	4.443732	31.4176	4.710918	1092.585	62.62302	1208.394	73.43557	62.68712	1319.2	60.6363	578695.4	3.022546	75.79157	43.03349	31.7174	49.93668	604.8446	475.111	1000.25	1212.117	1148.518
2023-04-05 15:01:00	18.31815	4.390478	31.65214	4.692149	1092.585	62.56199	1208.394	73.56069	65.40017	1318.9	60.48371	578695.7	3.031201	73.98795	43.94598	31.46868	49.06233	605.5866	832.1172	1000.25	1210.428	1148.518
2023-04-05 15:02:00	21.33669	4.365606	31.52072	4.677958	1092.585	62.62302	1208.965	73.18532	67.20989	1318.8	60.71565	578696	3.030542	74.13443	44.99886	31.21843	48.78004	605.9666	834.3694	999.8889	1212.117	1148.518
2023-04-05 15:03:00	21.9556	4.328069	31.49057	4.677958	1092.585	62.68712	1208.394	73.18532	67.13969	1320.9	56.85817	578696.4	3.02337	73.83841	46.5095	31.21843	48.81208	604.4827	795.7336	1000.628	1212.661	1148.518
2023-04-05 15:04:00	21.13069	4.281224	31.50871	4.677958	1092.585	62.68712	1208.394	73.31045	62.2324	1320.8	50.60197	578696.8	3.022898	73.9086	45.60922	30.9987	48.99977	605.5866	774.7494	1000.25	1212.661	1148.518
2023-04-05 15:05:00	21.56191	4.312352	31.59648	4.677958	1092.585	62.56199	1208.394	73.18532	58.60685	1320.8	48.7129	578697.1	3.03128	73.38369	45.08125	30.9987	48.56184	604.4827	774.3649	999.5281	1212.661	1148.518
2023-04-05 15:06:00	21.03731	4.328069	31.47568	4.710918	1093.129	62.68712	1208.394	72.12329	56.79713	1320.9	49.24086	578697.4	3.028523	75.26666	44.16877	31.09331	48.78004	604.8446	774.3649	1000.25	1211.545	1148.518
2023-04-05 15:07:00	19.48089	4.324864	31.54022	4.729686	1092.585	62.62302	1208.394	72.5597	55.66491	1320.9	49.16152	578697.8	3.031923	75.04082	43.41192	30.7805	48.37415	605.5866	774.7494	1000.25	1212.117	1148.518
2023-04-05 15:08:00	19.72442	4.371862	31.50567	4.729686	1092.585	62.56199	1208.394	73.24941	54.91417	1320.9	49.24086	578698.1	3.035346	77.53414	43.64386	30.87358	48.34363	604.4827	775.1156	1000.25	1211.545	1148.518
2023-04-05 15:09:00	20.08057	4.437476	31.40915	4.734264	1092.585	62.62302	1208.394	72.99916	53.62936	1321	49.31716	578698.4	3.027886	77.07637	43.26543	30.46769	47.74853	605.2247	774.7494	1000.25	1212.661	1148.518
2023-04-05 15:10:00	19.36828	4.481117	31.53564	4.748226	1092.585	62.56199	1208.394	72.93507	52.19501	1321	49.4728	578698.8	3.035997	75.56878	42.81376	30.65538	47.87365	605.5866	774.7494	999.8889	1212.661	1147.973
2023-04-05 15:11:00	18.88121	4.468605	31.57417	4.729686	1093.129	62.62302	1208.394	73.12428	51.664	1323.4	49.61929	578699.1	3.041314	74.97063	42.66118	30.31205	47.15496	605.2247	774.7494	1000.25	1211.545	1148.518
2023-04-05 15:12:00	18.74937	4.434272	31.56984	4.692149	1092.585	62.4979	1208.394	72.87404	50.75761	1324	49.91531	578699.4	3.035542	75.64507	42.20951	30.40513	46.81163	605.5866	774.7494	999.8889	1211.545	1148.518
2023-04-05 15:13:00	18.7686	4.415503	31.59379	4.677958	1092.585	62.56199	1208.394	72.87404	50.08316	1324.1	50.14725	578699.7	3.036822	73.83841	41.30617	29.96872	46.31113	604.4827	774.7494	999.8889	1212.117	1148.518
2023-04-05 15:14:00	15.84344	4.396735	31.54505	4.677958	1092.585	62.56199	1208.394	73.0602	49.62539	1324	50.22355	578700	3.032195	75.79157	40.62257	29.65591	46.96727	605.5866	774.7494	1000.628	1209.883	1148.518
2023-04-05 15:15:00	17.83108	4.396735	31.47523	4.65919	1092.585	62.62302	1208.394	72.80995	50.15336	1324	49.69558	578700.3	3.020494	75.11711	40.39979	29.74899	46.31113	604.8446	774.7494	1000.25	1211	1148.518
2023-04-05 15:16:00	16.53742	4.409247	31.54215	4.710918	1092.585	62.62302	1208.394	72.74891	50.98955	1324	49.69558	578700.6	3.023317	73.3135	40.62257	29.24849	46.43626	605.2247	774.3649	1000.25	1212.661	1148.518
2023-04-05 15:17:00	16.63081	4.390478	31.54973	4.748226	1093.129	62.56199	1208.394	72.93507	52.72297	1324	49.16152	578700.8	3.028403	77.91257	41.08034	29.21798	46.43626	605.2247	774.7494	999.1494	1210.428	1148.518
2023-04-05 15:18:00	16.68666	4.378119	31.49482	4.799954	1092.585	62.62302	1208.394	73.18532	53.48287	1323.9	48.94484	578701.1	3.026158	78.05905	41.53506	28.74952	46.53086	605.2247	774.7494	999.5281	1212.661	1148.518
2023-04-05 15:19:00	17.28725	4.390478	31.49198	4.799954	1092.585	62.56199	1208.394	73.0602	53.93149	1323.9	48.56641	578701.4	3.014391	75.04082	42.13321	28.52979	46.15549	605.2247	774.7494	1000.25	1212.661	1148.518
2023-04-05 15:20:00	17.5491	4.362402	31.47872	4.818723	1092.585	62.62302	1208.394	73.0602	54.53574	1327.9	48.33448	578701.7	3.020903	78.66331	42.50553	28.28107	46.12497	605.9666	774.7494	1000.25	1212.661	1148.518
2023-04-05 15:21:00	17.71847	4.346838	31.44498	4.818723	1093.129	62.68712	1208.394	73.12428	54.61204	1328.4	48.49012	578702	3.014709	77.98276	42.81376	28.09339	45.71756	605.2247	775.1111	999.8889	1212.661	1148.518
2023-04-05 15:22:00	17.96201	4.456245	31.43833	4.832685	1092.585	62.56199	1208.394	73.12428	54.75853	1331	48.03235	578702.3	3.005552	73.53017	42.81376	28.34363	45.655	605.2247	475.111	1000.25	1212.661	1148.518

3.2. 데이터 유형/구조

- 데이터셋 구조 : 테이블 형식(Tabular)
- 원본 데이터셋
  - 파일명 : df_ext(2023-04-01~2023-08-31,51250385)_2023-10-17 10-58-30 - seoultec.xlsx
  - Column 수 : 46
  - Row 수 : 66106

- 데이터셋 주요 변수 정의
  - 데이터셋에는 일자, 시간, 배치번호 및 다양한 공정 변수 정보가 포함되며, 주요 변수는 다음과 같음

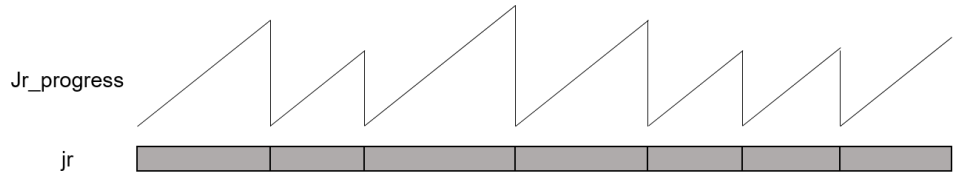
속성(column)	설명	데이터형
date	데이터가 수집된 시간	datetime
tg11	설비B2 온도	float
tg14	설비B2 절대습도	float
tg03	끝단 설비 속도	float
tg13	설비B1 절대습도	float
tg17	스팀 누적값	float
tg33	수분값	float
tg26	설비PE2 절대습도	float
tg34	설비AE 절대습도	float
tg35	설비S1 용액 높이	float
tg36	설비S2 용액 높이	float
tg22	설비 PE1 속도	float
tg37	설비 PE2 속도	float
tg38	설비 AE 속도	float
tg23	설비 BE1 속도	float
tg24	설비 BE2 속도	float
tg25	설비 PS1 속도	float
tg27	설비 PS2 속도	float

tg28	설비 AS 속도	float
tg29	설비 PS1 온도	float
tg30	설비 PS2 온도	float
tg31	설비 AS 온도	float
tg32	설비 S1 온도	float
tg39	설비 AS 절대습도	float
tg40	설비H 온도	float
tg41	스팀 온도	float
tg42	설비 PE1 온도	float
tg21	설비 PE1 절대습도	float
tg20	스팀 압력	float
tg43	설비 PE2 온도	float
tg44	설비 AE 온도	float
tg45	설비 S2 온도	float
stop	공정 분석값	str
output	재품 생산량 계산값	float
ei	원단위 계산값	float
jr	단위 공정값	str
shift	각각 작업팀 구분값	str
wclass	각 작업팀 구분값	str
sstable	원단위 상태 분석값	str

- 학습용 데이터

  ■ 목표(설비 센서의 패턴을 분석을 통해, 미래의 ei의 급격한 상승 요인 탐지)에 맞게 원본 데이터에서 학습에 필요한 부분만 추출 및 재가공

  ◆ jr_progress : 각 제품의 생산 시작 시점을 0으로 하여 새로운 제품이 생산되기 전까지 1분에 1씩 증가하는 각 제품의 생산 경과 시간 (분 단위)을 표현한 새로운 변수 도입



  ■ 독립변수 : 38개 sensor 데이터 + 공정진행도(jr_progress) + 과거 및 현재의 ei

  ■ 종속변수 : 특정 time step 이후의 ei

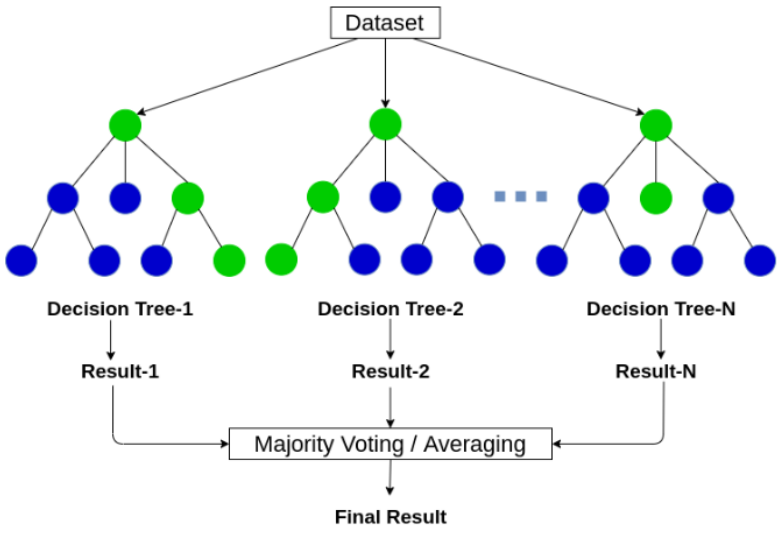
4. 스팀 사용량 추정 모델 구축 및 성능 비교 분석

4.1. Random Forest

4.1.1. 모델 설명

- 랜덤포레스트는 앙상블 학습 방법 중 하나로, 여러 결정 트리(Decision Trees)를 결합하여 구성된 강력하고 유연한 기계 학습 알고리즘
- 여러 개의 결정 트리 예측기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링 하여 개별적으로 학습한 후 최종적으로 모든 분류기가 보팅을 통해 예측을 결정하는 과정으로 결과를 도출
- 높은 정확도로 회귀 및 분류 작업을 모두 처리할 수 있으며, 각 트리가 서로 다른 데이터와 특성의 조합으로 학습되기 때문에 과적합 방지에 효과적인 머신러닝 기법
- 그러나 모델의 크기와 예측 시간이 비교적 크고 길다는 단점도 있음

4.1.2. 모델 아키텍처



- 랜덤포레스트는 다수의 결정 트리로 구성, 각 트리는 데이터의 다른 부분 집한을 사용하여 학습되며, 이들의 결합을 통해 최종 예측
- 각 결정 트리는 전체 데이터 세트에서 중복을 허용하는 무작위 추출방식(부트스트랩 샘플링)을 통해 생성된 서로 다른 훈련 데이터 세트에서 학습
- 각 트리에서 분할을 결정할 때 전체 특성 집합에서 부작위로 선택된 부분 집합의 특성을 사용
- 다양한 트리의 예측 결과의 평균으로 취합

4.1.3. 모델의 특징

- 여러 결정 트리를 결합함으로써 일반적인 단일 트리보다 높은 예측 정확도
- 개별 트리가 학습 데이터에 과적합되는 것을 방지하는 효과적인 매커니즘
- 각 특성이 예측 결과에 얼마나 중요한지를 평가할 수 있어 특성 선택과 데이터 이해에 유용

4.1.4. 모델 사용 이유

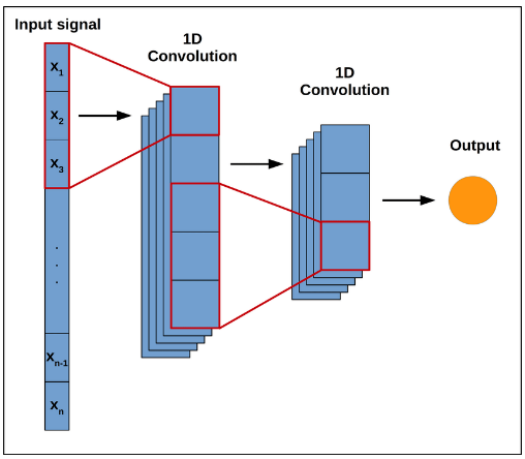
- 랜덤 포레스트는 복잡한 데이터 세트에서도 높은 정확도를 제공하는 모델로 다양한 유형의 데이터와 문제에 대해 효과적인 기본적인 예측 모델을 적용
- 주어진 데이터의 어떤 특성이 예측에 가장 중요한 역할을 파악하여 데이터를 깊게 이해하고 더 효과적인 특성 선택과 모델 개선에 기여하기 위함
- 일반적으로 과적합에 강하고 다양한 종류의 데이터에 대해 좋은 일반화 성능을 보이는 모델이기 때문에 사용

4.2. 1D Convolutional Neural Network model

4.2.1. 모델 설명

- 1D CNN(1D Convolutional Neural Network)은 시퀀스 데이터를 처리하기 위해 설계된 신경망의 한 유형으로 시계열 데이터, 오디오 신호, 텍스트 데이터 등에 적합
- 입력 데이터를 일련의 컨볼루션 레이어에 통과시켜 각 컨볼루션은 입력 데이터의 부분적인 연속 구간을 처리하여 특성을 추출
- 데이터의 불필요한 정보는 무시하고 중요한 정보를 캡처하여 최종적으로 예측을 위한 출력 생성

4.2.2. 모델 아키텍처



- 여러 샘플을 포함하는 1차원 시퀀스 데이터를 입력 데이터로 받음
- 입력 데이터에 대한 필터의 집합을 적용한 특성맵을 생성하는 레이어 통과
- 커널은 입력 신호를 슬라이딩 윈도우 방식으로 순회하면서 각 위치에서의 신호와 커널 간의 내적을 계산
- 추가적인 처리(평탄화, 완전 연결 레이어 등)을 거쳐 연속적인 수치 예측 값 출력

4.2.3. 모델의 특징

- 데이터의 순서를 유지하고, 데이터의 공간적 관계를 학습할 수 있는 특징 덕분에 시계열, 음성, 자연어 관련 데이터셋에 주로 사용
- 시간에 따른 연속적인 패턴이나 특성을 식별하는데 유용
- 다양한 길이의 시퀀스를 처리할 수 있으며, 각기 다른 윈도우 크기를 가질 수 있음

4.2.4. 모델 사용 이유

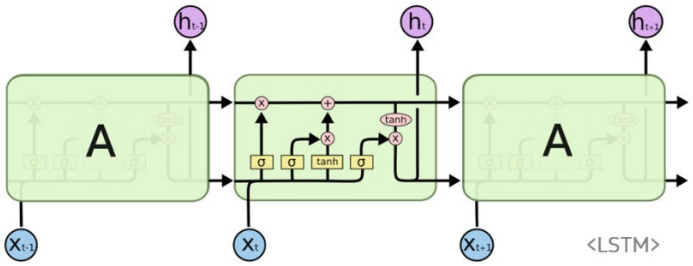
- 시퀀스 데이터의 중요한 정보를 효과적으로 추출할 수 있음
- SHAP과 같은 기법을 사용하여 각 입력 변수의 중요도 해석에 유용, 모델의 예측에 기여하는 주요 요인을 식별하여 모델의 해석 가능성을 높임

4.3. Long Short-Term Memory model

4.3.1. 모델 설명

- LSTM(Long Short-Term Memory) 모델은 시퀀스 데이터를 처리하기 위한 특별한 종류의 순환 신경망(RNN)
- 기존 RNN은 시퀀스가 길어질수록 이전 정보를 잃어버리는 경향이 있는데, LSTM은 '게이트'라는 메커니즘을 통해 이 문제를 해결
- 세 개의 게이트(입력, 삭제, 출력 게이트)를 사용하여 정보의 흐름을 조절하며 특히 셀 상태는 장기 정보를 저장하며, 게이트들을 통해 정보가 추가되거나 제거
- LSTM은 긴 시간에 걸친 의존성을 인식하고, 복잡한 시퀀스 데이터에서 패턴과 구조를 학습 가능

4.3.2. 모델 아키텍처



- 셀 상태(Cell State, C): LSTM 네트워크의 '메모리' 부분으로, 시간에 따라 정보를 전달
- 입력 게이트(Input Gate, σ): 셀 상태에 새로운 정보를 추가할지 결정
- 망각 게이트(Forget Gate, σ): 셀 상태에서 어떤 정보를 버릴지 결정
- 출력 게이트(Output Gate, σ): 셀 상태의 어느 부분을 출력으로 전달할지 결정
- 숨겨진 상태(Hidden State, h): 이전 타임 스텝의 출력으로, 다음 타임 스텝의 입력과 셀 상태에 영향을 줌

4.3.3. 모델의 특징

- 중요한 정보를 장기간 기억하고, 불필요한 정보는 시간이 지남에 따라 버릴 수 있음
- 데이터와 네트워크 상태에 따라 게이트를 동적으로 조정
- LSTM에 attention 매커니즘을 추가함으로써, 모델의 시퀀스의 어느 부분에 주목하는지 확인 가능

4.3.4. 모델 사용 이유

- 복잡한 시퀀스 데이터 패턴을 학습
- 시간적으로 연속적인 데이터에서 과거의 정보가 현재와 미래의 이벤트에 미치는 영향을 모델링
- 어텐션 메커니즘을 사용하여 모델이 시퀀스의 어느 부분을 중요하게 생각하는지 파악하고, 이를 통해 피처의 중요도 분석을 위함

4.4. Dual-Attention model

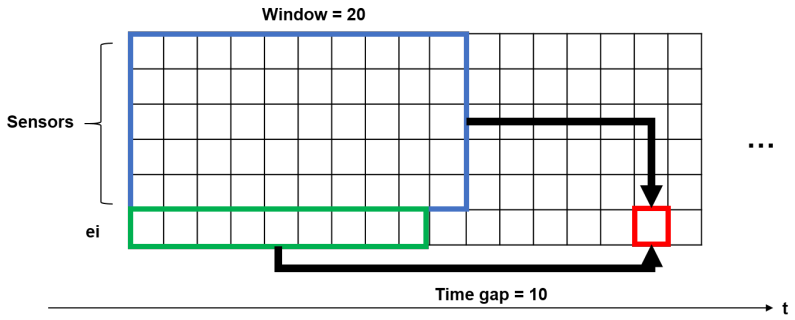
4.4.1. 모델 설명

- DARNN(dual-stage attention-based recurrent neural network)은 날씨, 주식 시장, 에너지 소비 등 시간에 따라 변하는 값을 예측하기 위한 시계열 회귀 모델

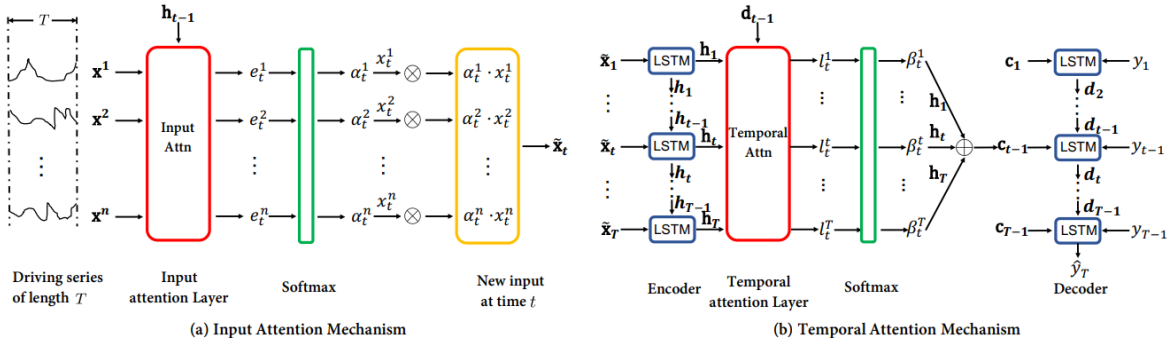
- DARNN은 2단계의 attention mechanism을 사용하여 feature 및 time stamp에 대한 중요한 정보를 모두 고려하여 시계열 예측 진행

- DARNN은 Input Attention + Encoder + Decoder 구조로 이루어져 있으며, Attention 메커니즘은 다양한 외부 변수들 사이에서 중요한 특징을 추출하기 위해 Encoder의 hidden state를 활용하며, 이렇게 추출된 데이터는 Encoder의 입력으로 사용되고, Encoder의 hidden state 업데이트 시에는 Decoder의 hidden state도 함께 사용

- 모델의 인풋 데이터는 특정 timestep T만큼의 센서 값과 예측하고자 하는 T-1개의 target series 'ei'이고 모델의 아웃풋 데이터는 T 시점에서의 target인 'ei'



4.4.2. 모델 아키텍처



- Input Attention Mechanism(그림 a)

  - 인코더는 이전 인코더의 hidden state를 참고해서 각 시점에서 예측과 가장 연관이 큰 input driving series를 파악하는 attention score를 계산

  - 즉, 인코더에서 relevant driving series를 선택하기 위해 이루어지는 attention 메커니즘에서 나온 feature weight를 이용해 새로운 input vector($\tilde{x}_t$)를 생성

- Temporal Attention Mechanism(그림 b)

  - 디코더는 이전 디코더의 은닉 상태를 참고하여, 전체 시점 중 target series와 연관이 가장 큰 인코더 hidden state가 무엇인지 파악하는 attention 스코어를 계산

  - 즉, 디코더는 relevant encoder hidden state를 선택하기 위해 이루어지는 attention

메커니즘

- Temporal Attention Mechanism을 적용한 context vector를 생성한 뒤, 주어진 T−1 개의 target series와 context vector를 concat하여 LSTM을 통해 최종적으로 T 시점에서의 target 예측

4.4.3. 모델 사용 이유

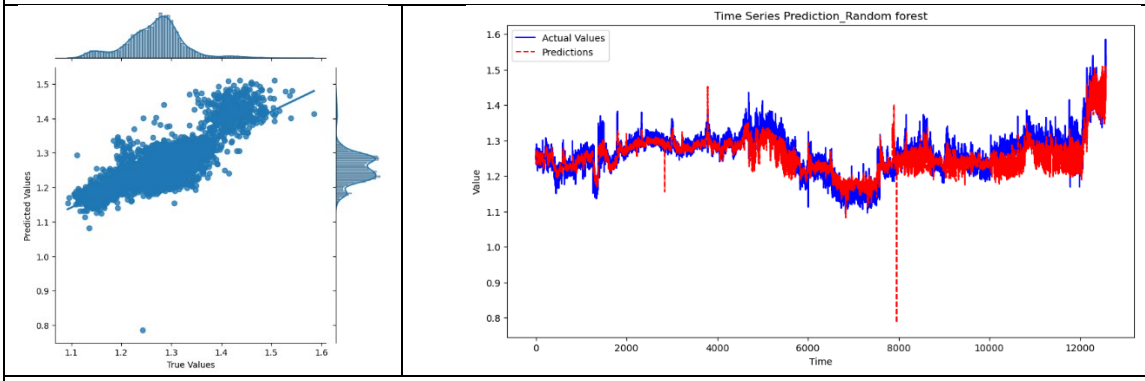
- 개별 변수 별로 모든 시점에 대해 attention score를 추출하여 해석할 수 있음.
- 2단계의 attention 및 LSTM을 사용해 다른 모델보다 더 복잡한 시퀀스 데이터 패턴을 학습 가능
- LSTM을 사용하기에 LSTM의 장점인 장기 의존성을 처리하는 능력이 뛰어남
- Attention mechanism을 사용하여 input driving series 및 time step에 대해 가중치를 부여하기에 보다 더 효과적으로 예측할 수 있음
- 사전 연구를 통해 금융, 기상학, 에너지 수요 예측 등 다양한 분야에서의 시계열 데이터 예측에 적용 가능한 사실이 증명됨

5. 성능 비교

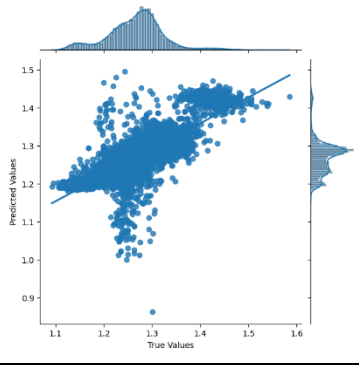
5.1. Random Forest

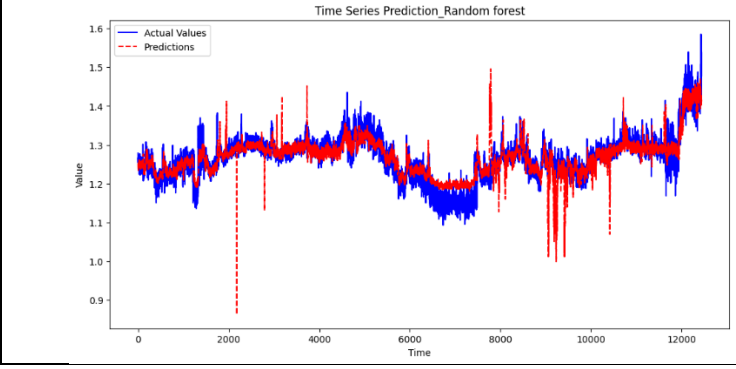
Time gap	R^2	MSE
5	0.6757	0.0011
10	0.6723	0.0011
15	0.3719	0.0021

Time gap 5

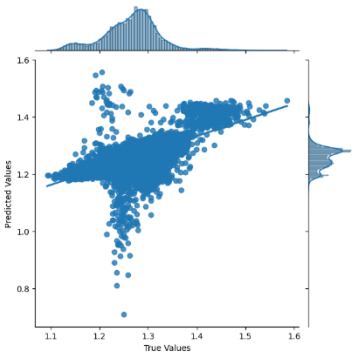


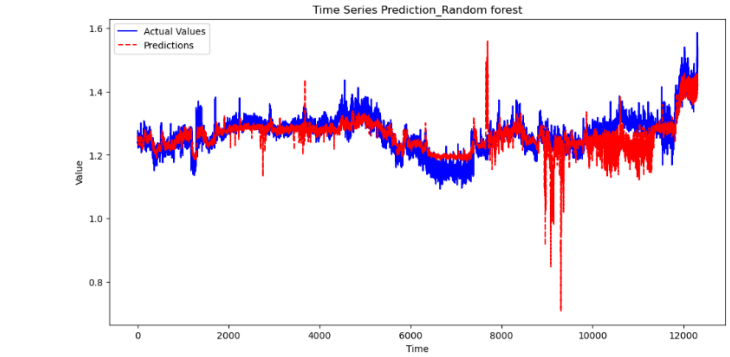
Time gap 10





Time gap 15

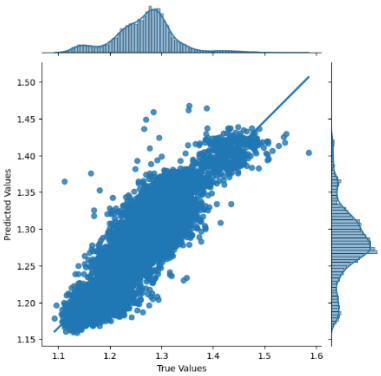


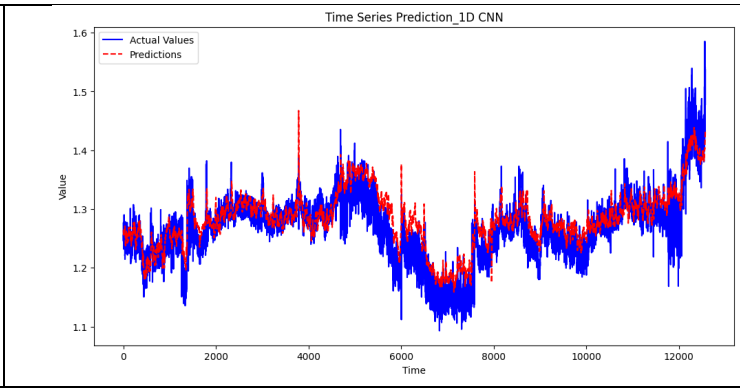


5.2. 1D CNN

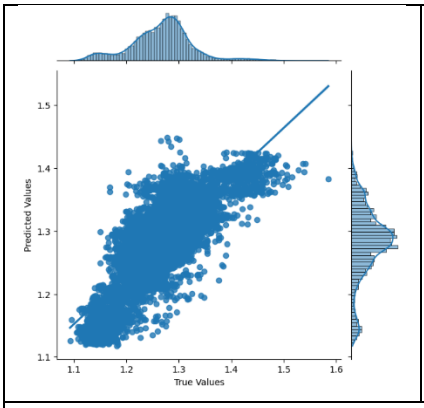
Time gap	R^2	MSE
5	0.6692	0.0011
10	0.5234	0.0016
15	-0.6541	0.0057

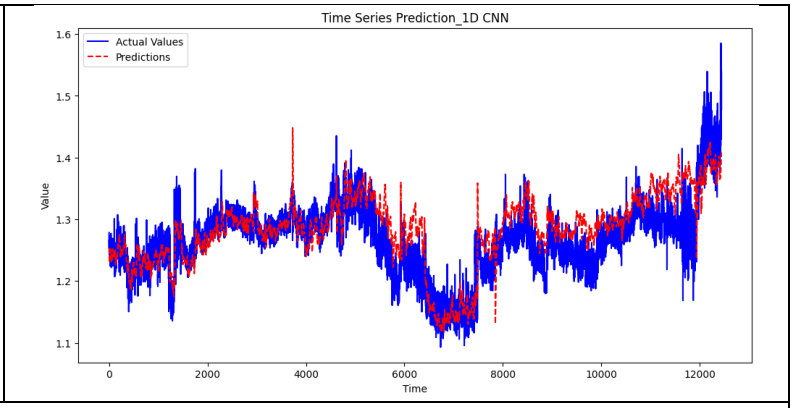
Time gap 5



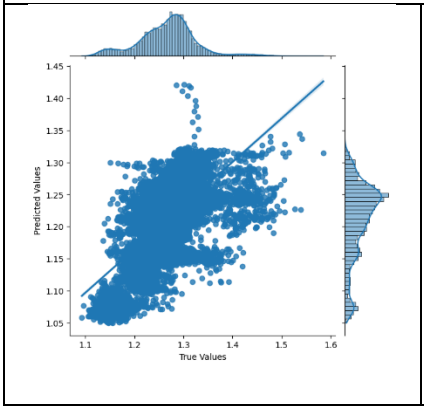


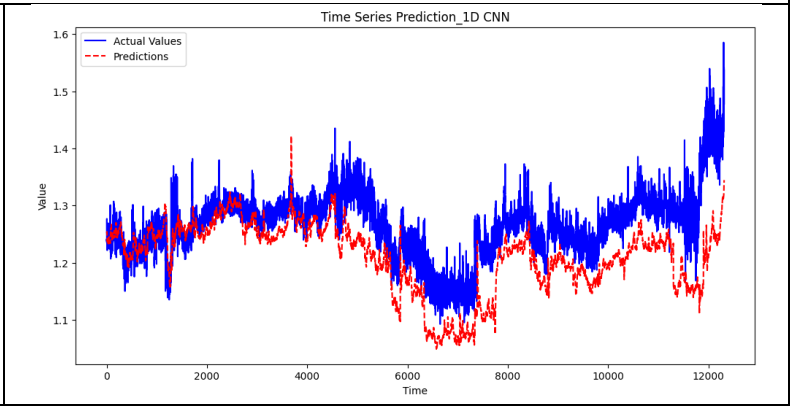
Time gap 10





Time gap 15

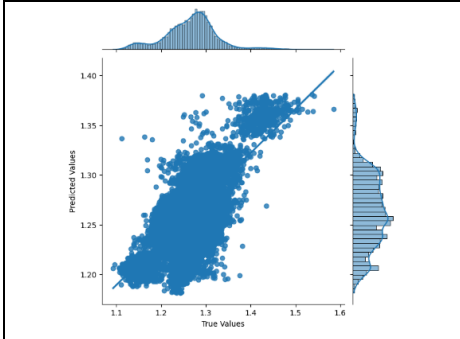


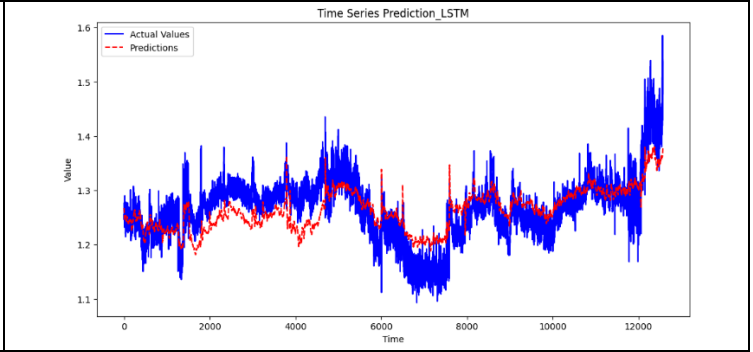


5.3. LSTM

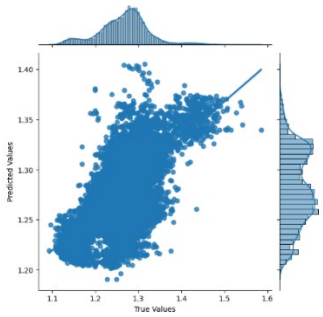
Time gap	R^2	MSE
5	0.4826	0.0018
10	0.2827	0.0025
15	0.0449	0.0033

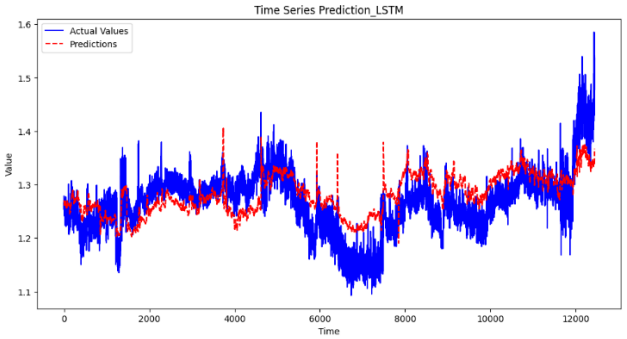
Time gap 5



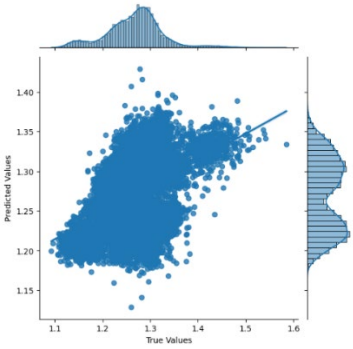


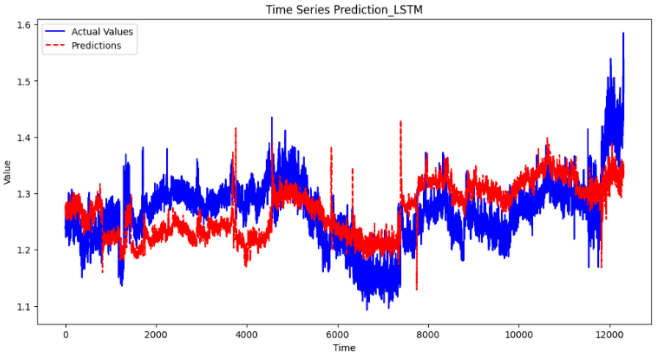
Time gap 10





Time gap 15

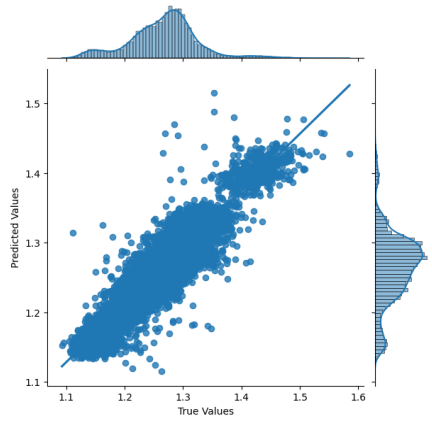


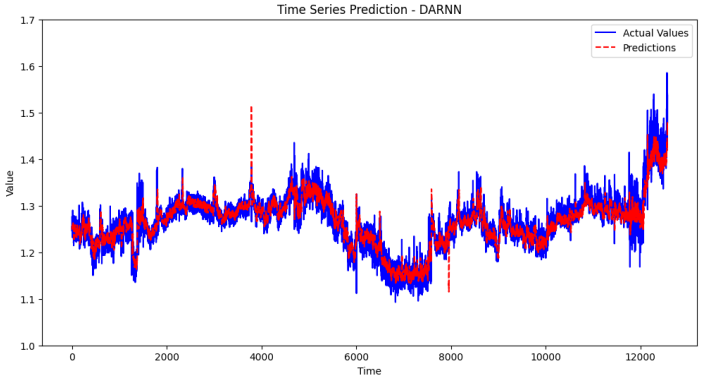


5.4. DARNN

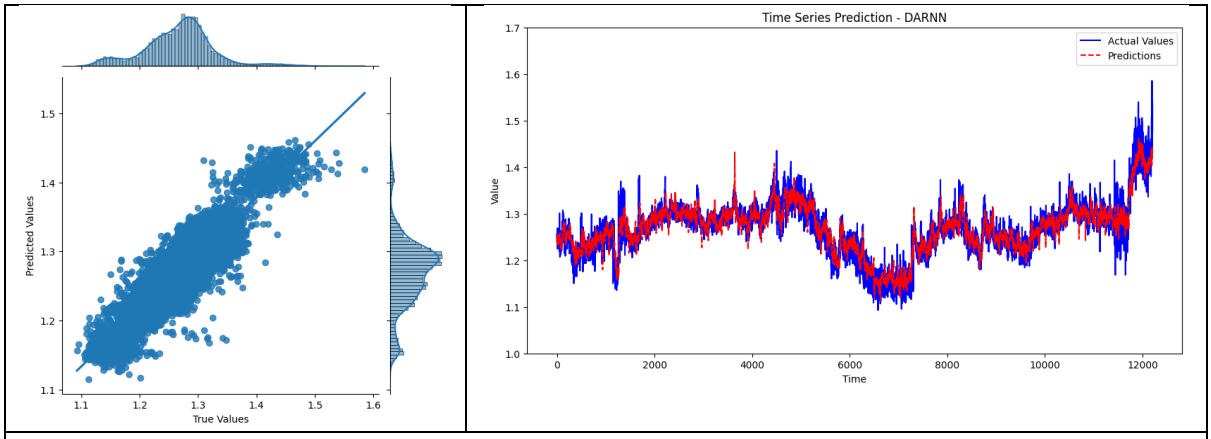
Time gap	R^2	MSE
5	0.8499	0.0005
10	0.8368	0.0006
15	0.8029	0.0008

Time gap 5

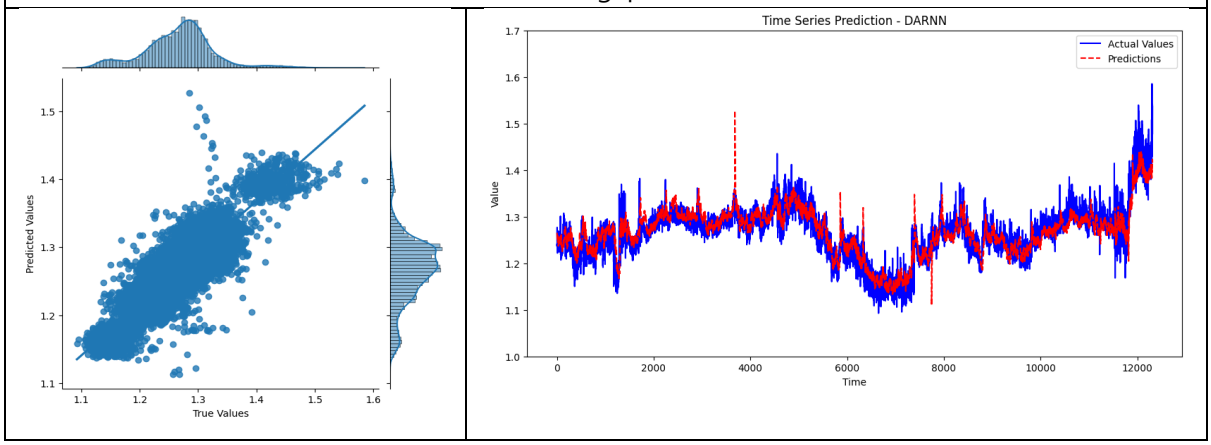




Time gap 10



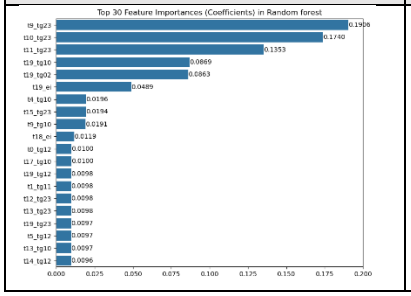
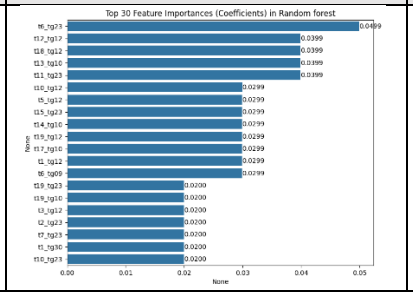
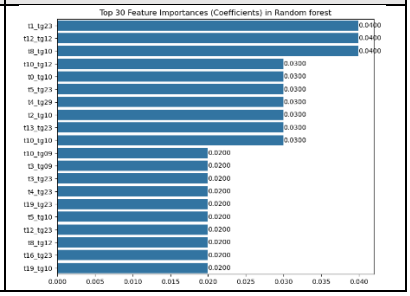
Time gap 15



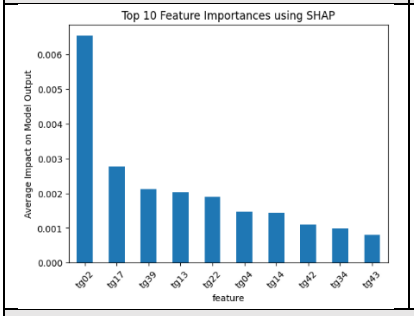
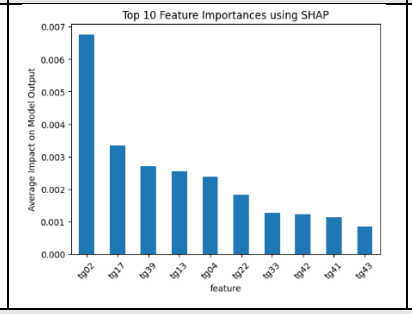
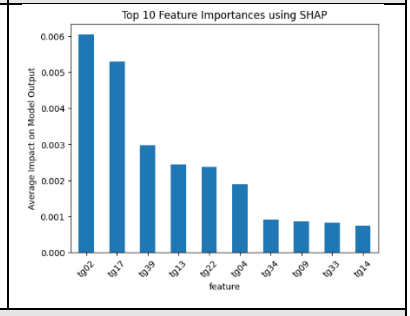
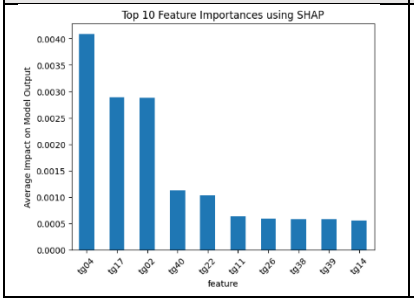
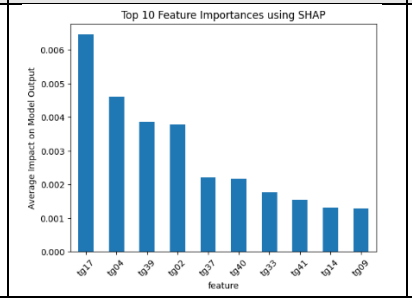
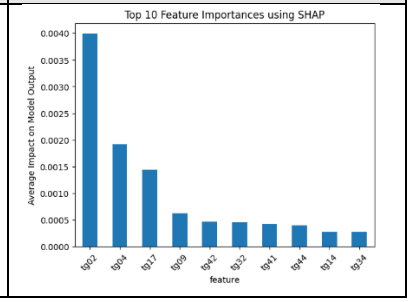
- 종합적으로 DARNN 에서 time gap 이 5 일 때의 결정계수(R^2)가 0.8499 로 가장 높은 성능을 보임
- 모든 모델에서 time gap 을 크게 할수록 모델의 성능이 크게 떨어지는 것을 확인

6. 현장에서의 해석 및 사전 조치를 위한 원인 인자 해석

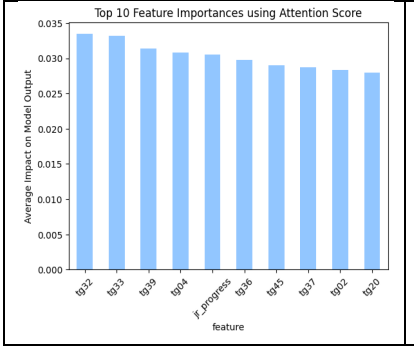
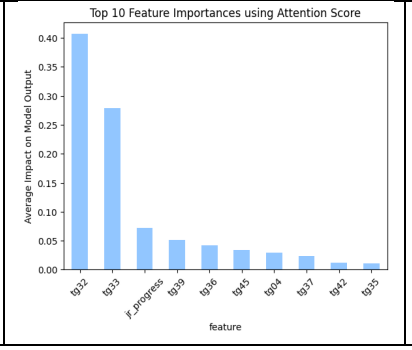
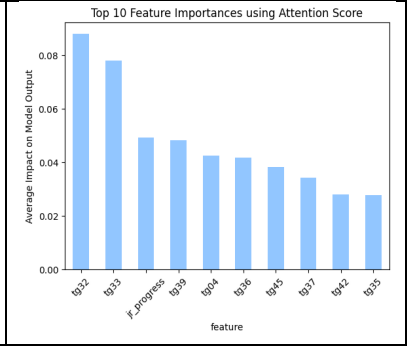
6.1. Feature-importance

Time gap		
5	10	15
		

6.2. SHAP

1D CNN Time gap		
5	10	15
		
LSTM Time gap		
5	10	15
		

6.3. Attention

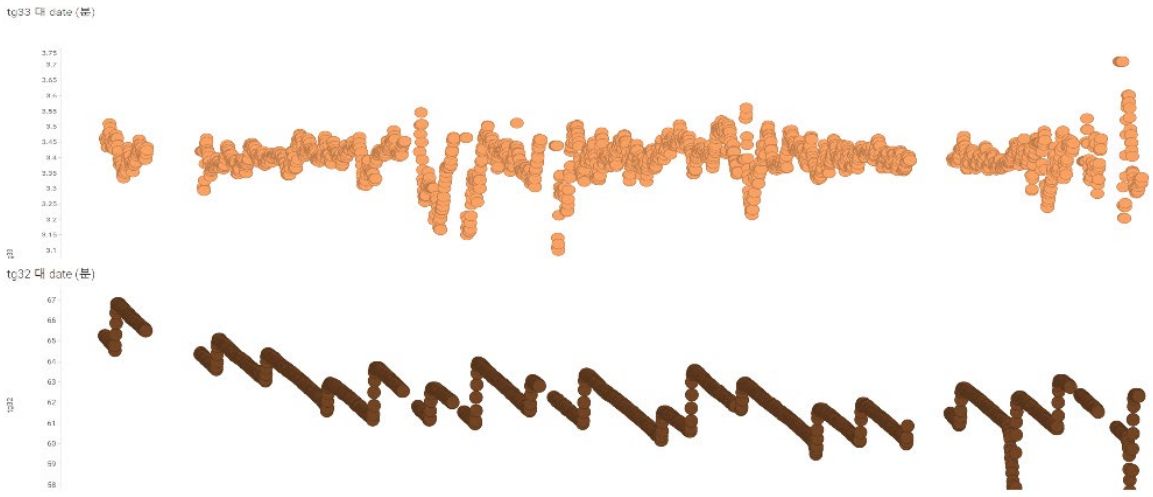
Time gap		
5	10	15
		

6.4. 비교

- 모델의 성능적 측면을 종합하자면, 모든 time gap 에서 DARNN 이 가장 좋은 성능을 보임
  - Timegap 이 5 일 때의 결정계수(R^2)는 DARNN(0.8499), Random forest(0.6757), 1D CNN(0.6692), LSTM(0.4826) 순

- Timegap 이 10 일 때의 결정계수(R^2)는 DARNN(0.8368), Random forest(0.6723), 1D CNN(0.5234), LSTM(0.2827) 순
- Timegap 이 15 일 때의 결정계수(R^2)는 DARNN(0.8029), Random forest(0.3719), LSTM(0.0449), 1D CNN(-0.6541) 순

- 모델의 예측에 대해 각 feature 에 대한 중요도를 계산한 결과 다음과 같은 결과를 보임
  - Random forest 의 경우 tg23(설비 BE1 속도), tg10(설비 B1 속도) 가 높은 중요도를 보임
  - LSTM 및 CNN 의 경우 tg02(종이별 측정 무게), tg04(스팀 순간값), tg17(스팀 누적값)이 높은 중요도를 보임
  - DARNN 의 경우 tg32, tg33 이 높은 중요도를 보임
- 종합하면 DARNN 모델이 센서값을 통해 ei 값을 예측하는데 있어 가장 좋은 모델로 평가되며, 해당 모델의 예측에 가장 영향력을 크게 발휘하는 tg32 (설비 S1 온도)와 tg33 (수분값)이 주요하다고 판단됨. 아래는 tg33, tg32 센서 값 예시



7. 개선 방안 및 고찰

- 주어진 제조 상황에서의 에너지 사용 이상에 대한 임계치를 명확히 정의하여 해당 상황에 대한 사전 탐지 여부를 검증하고 해석한다면 더 유의미한 모델로 발전될 수 있을 것임.
- 최근 시점의 스팀 에너지 사용량의 트렌드 대비 급격히 변동하는 시점을 '이상'으로 정의하고 나머지를 '정상'으로 정의한 후에, classification 모델을 활용하는 것도 고려해볼 수 있음.

- 해석 가능한 시계열 모델 구조로 IMV-LSTM 혹은 TFT 를 도입하여 변형하는 시도를 추가로 시도해 볼 수 있을 것임.
- 본 모델은 스팀 에너지 사용량의 급증을 사전에 탐지할 수 있으며 그 원인 인자를 도출하여 사전 조치를 취할 수 있게 함. 이에 따라 불필요한 스팀 에너지 사용을 줄이고 생산 제품의 불량을 감소시키는데 기여할 수 있을 것임.