

공압기 소비전력에 대한 예측 모형의 비교연구

A Comparison Study on Forecasting Models for Air Compressor Power Consumption

김주헌¹, 장문수¹, 김예진¹, 허요섭², 정현상², 박소영^{1*}

Juhyeon Kim¹, Moonsoo Jang¹, Yejn Kim¹, Yoseob Heo²,
Hyunsang Chung², Soyoung Park^{1*}

〈Abstract〉

It's important to note that air compressors in the industrial sector are major energy consumers, accounting for a significant portion of total energy costs in manufacturing plants, ranging from 12% to 40%. To address this issue, researchers have compared forecasting models that can predict the power consumption of air compressors. The forecasting models were designed to incorporate variables such as flow rate, pressure, temperature, humidity, and dew point, utilizing statistical methods, machine learning, and deep learning techniques. The model performance was compared using measures such as RMSE, MAE and SMAPE. Out of the 21 models tested, the Elastic Net, a statistical method, proved to be the most effective in power consumption forecasting.

Keywords : Power, Regression, Machine Learning, Deep Learning, Time-series Forecasting

1 정회원, 부산대학교(Pusan National University), 통계학과 1 Department of Statistics, Pusan National University
2 정회원, 한국과학기술정보연구원(KISTI), 부산울산경남지원 2 Busan·Ulsan·Gyeongnam Branch, Korea Institute of Science and Technology Information(KISTI)
1* 교신저자, 부산대학교 통계학과, 조교수 1* Corresponding Author, Assistant Professor, Department of Statistics, Pusan National University
E-mail: soyoung@pusan.ac.kr E-mail: soyoung@pusan.ac.kr

1. 서론

인공지능과 머신러닝은 현재 산업 분야에서 빠르게 발전하고 있으며, 특히 에너지 소비 예측과 최적화에 있어 중요한 역할을 하고 있다. 이는 공기 압축기와 같은 에너지 집약적인 장치의 운영 비용과 효율성을 향상시키는 데 큰 도움이 된다. 그러나, 공기 압축기(이하 '공압기'라고 한다)는 산업에서 엄청난 양의 전력을 소비하는 반면(일반적으로 제조 공장 전체 에너지 비용의 12%에서 많게는 40%까지), 공압기의 전력 저감에 관한 연구는 상대적으로 부족한 상황이다. 본 논문은 이러한 연구 공백을 해소하기 위해 통계적 방법, 머신러닝, 딥러닝으로 공압기의 전력 사용량을 예측하는 방법을 제시한다.

공압기의 작동은 여러 환경변수에 의해 영향을 받는다. 이에 온도, 압력, 이슬점, 습도 등이 포함되며, 이런 변수들은 공압기의 전력 소비에 영향을 미친다. 그러나 이러한 변수들 사이의 관계는 매우 복잡하므로, 전통적인 방법만으로는 정확한 에너지 사용량을 예측하는 데 어려움이 있다. 이 문제를 해결하기 위해, 본 연구에서는 공압기 소비 전력과 환경변수들의 실증적인 자료를 바탕으로 회귀모형, 머신러닝 그리고 딥러닝을 사용하여 공압기의 전력 사용량을 예측한다.

또한 회귀모형, 머신러닝, 그리고 딥러닝 측면에서 공압기의 소비전력을 예측하는 모형을 비교하였다. 이를 통해 각 방법의 장단점을 파악하고, 가장 효과적인 공압기 소비전력 예측 방법을 제안하고자 한다. 이 연구의 결과는 공압기의 효율적인 운영을 위한 중요한 지침을 제공할 것으로 기대된다.

본 연구의 주요 목표는 공압기의 에너지 소비를 보다 정확하게 예측하고 이해하는 것으로, 이를 통해 공압기의 작동 효율성을 향상시키고 에너지 사용량을 감소시키는 방안을 모색하고자 한다.

이러한 노력은 전력 사용량에 미치는 변수의 영향을 파악함으로써 운영 비용을 절감하는 것을 목표로 하고 있다. 또한, 이 연구는 공압기에 대한 깊이 있는 연구를 통해 이 분야의 연구 공백을 메우는데 기여할 것으로 기대된다. 시계열 데이터의 예측 성능을 비교하는 연구로는 시계열 자료의 예측을 비교 분석한 (장문수 외, 2021)[1] 등이 있다.

2. 예측모형

본 연구에서는 회귀모형, 머신러닝, 딥러닝을 활용하여 다양한 예측모형을 학습하였다. 먼저 회귀모형은 단순 회귀모형부터 시작하여, 다중 회귀모형, 그리고 변수 규제화를 적용한 Ridge, Lasso, Elastic Net 등의 모형을 활용하였다. 머신러닝 모형에서는 의사 결정 트리 기반의 다양한 모형을 활용하였다. 배깅(bagging) 방식에 기반한 랜덤포레스트(randomforest)[2]와, 부스팅(boosting) 방식에 기반한 캣부스트(catboost)[3]를 주요 알고리즘으로 채택하였다. 또한, 신경망(Neural Net) 기반 모형과 머신러닝 기법들을 앙상블(ensemble)하여 보다 견고한 예측력을 확보하였다. 딥러닝 모형에서는 시계열 분석에 최적화된 NLinear[4], DLinear[4], SCINET[5]와 더불어, 일반적인 딥러닝 모형인 Long Short Term Memory (LSTM)[6]을 활용하였다. 다양한 모형과 기법을 활용함으로써, 복잡한 패턴을 가진 데이터에 대한 정확한 예측이 가능하였다.

2.1 회귀모형

통계적 기법은 데이터에 내재된 패턴을 이해하고, 예측모형을 구축하는 데 중요한 도구이다. 본 연구에서는 통계적 방법으로 선형회귀분석과

규제화 회귀분석을 사용하였다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi} + \epsilon_i, i = 1, 2, \dots, n \quad (2)$$

2.1.1 단순회귀분석

회귀분석은 하나 또는 그 이상의 독립 변수를 사용하여 종속변수의 변화를 예측하는 모형이다. 이 방법은 변수 간의 상관관계와 인과 관계를 수학적으로 표현하고 이해하는 데 필수적인 도구로 쓰인다.

단순회귀분석은 하나의 설명변수를 사용하여 목적변수의 변화를 예측하는 방법이다. 이를 이용하면 두 변수 사이의 선형관계를 명확히 파악할 수 있다. 단순회귀분석의 모형은 식 (1)에서 확인할 수 있다. 여기서 y 는 목적변수, x 는 설명변수, β_0 는 절편, β_1 은 회귀계수, ϵ 는 오차항이다. 회귀계수는 설명변수의 변화에 따른 목적변수의 변화량을 나타내며, 이 값이 클수록 설명변수가 목적변수에 큰 영향을 미친다는 것을 의미한다. 절편은 설명변수가 0일 때의 목적변수의 값이며, 오차항은 평균이 0이고 분산이 σ^2 인 정규분포를 따른다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, \dots, n \quad (1)$$

2.1.2 다중회귀분석

다중회귀분석은 두 개 이상의 독립 변수를 활용하여 종속변수의 변화를 예측하는 기법이다. 다중회귀분석의 모형은 식 (2)에서 확인할 수 있다. 여기서 y 는 목적변수, x_{mi} 는 m 번째 설명변수의 i 번째 값, β_0 는 절편, $\beta_1, \beta_2, \dots, \beta_n$ 은 각 설명변수의 회귀계수, ϵ 는 오차항이다. 다중회귀분석은 단순회귀분석과 달리 다중공선성을 주의해야된다. 다중공선성이란 설명변수들이 서로 높은 상관성을 가질 때 발생하는 문제로 계수 추정치의 불안정성을 초래하며, 해석을 어렵게 만든다. 이는 분산팽창계수(Variance Inflation Factor)를 통해 검정할 수 있고, 규제화 회귀 방법을 통해 완화할 수 있다.

2.1.3 규제화 회귀 방법: Ridge, Lasso Elastic Net

Ridge 회귀는 L2 규제화를 사용하여 회귀계수의 크기를 제한함으로써 과적합을 방지한다. 이로써 모형의 복잡성을 제어하고, 일반화 성능을 향상시킨다. 식 (3)에서 Ridge 회귀의 목적함수를 확인할 수 있다. 여기서 L 은 손실함수, y_i 는 샘플 i 의 목적변수, β_j 는 j 번째 설명변수의 회귀계수, x_{ij} 는 샘플 i 의 j 번째 종속변수의 값, λ 는 규제화 파라미터(regularization parameter)이다. 다음으로 식 (4)에서 Lasso 회귀의 목적함수를 확인할 수 있다. Lasso 회귀는 L1 정규화 기법을 사용하여 일부 회귀계수를 0으로 만들어 변수선택 기능을 수행하며, 이로써 모형의 해석이 용이해진다. Lasso 회귀에서는 가중치의 절댓값의 합을 추가한다. 마지막으로 식 (5)에서 Elastic Net의 목적함수를 확인할 수 있다. Elastic Net 회귀는 L1과 L2 정규화를 모두 사용하여 Ridge와 Lasso의 장점을 모두 결합한 방법으로, 다중공선성 문제를 효과적으로 해결하며, 필요 없는 변수를 제거함으로써 모형의 성능을 향상시킨다. 파라미터 α 는 Ridge와 Lasso의 비율을 결정하는 계수로, $0 \leq \alpha \leq 1$ 이다. 식 (3)-(5)에서, 첫 번째 항은 데이터에 대한 적합도를 나타내며, 두 번째 항은 규제화 항이다. λ 는 이 두 항 사이의 균형을 조절하는데 λ 가 클수록 모형은 더 단순해지고(즉, 가중치가 더 작아짐), λ 가 작을수록 모형은 데이터에 더 적합하게 된다.

$$L_{Ridge} = \sum_{i=1}^n (y_i - \sum_{j=1}^k \beta_j \times x_{ij})^2 + \lambda \sum_{j=1}^k \beta_j^2 \quad (3)$$

$$L_{Lasso} = \sum_{i=1}^n (y_i - \sum_{j=1}^k \beta_j \times x_{ij})^2 + \lambda \sum_{j=1}^k |\beta_j| \quad (4)$$

$$L_{Elastic} = \sum_{i=1}^n (y_i - \sum_{j=1}^k \beta_j x_{ij})^2 + \lambda \left\{ (1-\alpha) \sum_{j=1}^k \beta_j^2 + \alpha \sum_{j=1}^k |\beta_j| \right\} \quad (5)$$

2.2 머신러닝 기반 모형

다양한 머신러닝 기법들 중, 우수성이 인정받은 12개의 모형을 활용하였다. 그 중 대표적인 모형인 랜덤포레스트, 캣부스트, 앙상블 모형에 관하여 설명한다. 나머지 모형과 하이퍼 파라미터에 관해서는 (Klein & Aaron, 2020)[7]에서 확인할 수 있다.

2.2.1 랜덤 포레스트(Random forest)

랜덤포레스트는 앙상블 모형 기반의 기계학습 알고리즘으로, 종속변수의 분류 또는 예측에 사용된다. 주어진 설명변수와 변수를 무작위 표본 추출하여 각 시뮬레이션 상황별로 의사 결정 나무(decision tree)를 형성한다. 이를 통해 최대한 많은 비상관화된 의사결정 나무(uncorrelated decision tree)를 구성하고, 새로운 예측을 위한 설명변수가 입력되었을 때, 모든 의사결정 나무의 결정으로부터 다수의 결정을 최종 결정 값으로 선택한다. 기존 연구로부터, 랜덤 포레스트 알고리즘이 이미지 분류, 변수선택 등에 많이 사용되고 있고 시계열 데이터에서도 역시 사용이 가능하다고 알려져 있다.

2.2.2 캣부스트(CatBoost)

캣부스트는 기계학습에서 사용되는 앙상블 기반의 그래디언트 부스팅 알고리즘이다. 이 알고리즘은 “Category”와 “Boosting”의 합성어로, 범주형 데이터를 처리하는 능력이 뛰어나다는 의미를 내포하고 있다. 주어진 설명변수들에 대해 반복적으로 최적화를 진행하는 그래디언트 부스팅 알고리

즘의 일종이지만, 캣부스트는 기존 방법과는 다르게 순서 통계를 이용한 원-핫 인코딩 방법을 적용함으로써 범주형 데이터를 더욱 효과적으로 처리한다. 이러한 특징 덕분에 과적합(overfitting)을 피하면서 높은 예측 정확도를 달성할 수 있다. 캣부스트는 학습 과정에서 자동으로 범주형 변수를 인식하고 처리하는 기능도 갖추고 있어, 사전 데이터 전처리 작업이 상대적으로 간단하다는 장점이 있다. 또한, 이 알고리즘은 새로운 데이터에 대해 예측 시 모든 모형의 결과를 결합하여 최종 결정 값을 도출한다. 다양한 실용적인 분야에서 캣부스트는 그 효용성을 입증하고 있다. 이미지 분류, 텍스트 분류, 시계열 데이터 분석 등 뛰어난 성능을 보여주고 있으며, 이는 다양한 분야의 문제 해결에 적용 가능하다.

2.2.3 Neural Net

Neural Net은 Pytorch 라이브러리를 기반으로 한 신경망 모형이다. Neural Net은 다중 레이어 퍼셉트론(Multi-Layer Perceptron, MLP)을 기본 구조로 사용하며, 사용자가 지정한 하이퍼파라미터(hyperparameter)에 따라 모형의 구조와 학습 방식을 변경할 수 있다. 모형의 레이어(layer) 수, 뉴런 수, 활성화 함수, 최적화 기법, 손실 함수 등 다양한 요소를 사용자가 조절하여 특정 문제에 맞게 모형을 최적화할 수 있다. 특히, 이미지 분류, 텍스트 분류, 회귀분석 등 다양한 종류의 문제를 해결하는 데 사용할 수 있어 많은 연구자들에게 인기가 있다. 특히 이 모형은 특히 대용량의 데이터에 대해 높은 성능을 보이며, 특히 복잡한 비선형 패턴을 가진 데이터에서 그 능력을 발휘한다. 본 연구에서 구축한 Neural Net 모형을 (Erickson & Nick, 2020)[8]의 Fig. 1에서 확인할 수 있다. 해당 모형은 범주형 데이터와 수치형 데이터 모두 입력값으로 들어간다. 범주형 데이터는 임베딩

(embedding)을 거쳐 연속적인 표현으로 변환되며 수치형 데이터는 바로 신경망에 입력으로 사용된다. 두 종류의 데이터는 연결된 후 양방향 신경망 통과하게 되고 최종 예측값을 출력하게 된다.

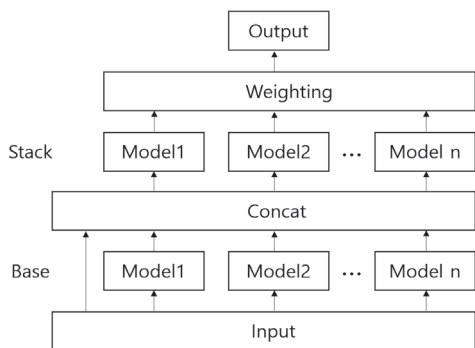


Fig. 1 Stack ensemble model

2.2.4 앙상블 모형(Ensemble Model)

앙상블 모형은 여러 가지의 학습 알고리즘을 결합하여 더 정확한 예측성능을 도출하는 기계학습 방법이다. 이러한 방법은 여러 모형의 예측을 결합함으로써 단일 모형으로는 도달하기 어려운 더 높은 정확도와 안정성을 얻을 수 있다. 앙상블 모형은 여러 가지 형태로 존재하는데, 주로 배깅, 부스팅, 스택킹(stackings)등의 방법이 있다.

배깅은 랜덤 포레스트와 같은 알고리즘에서 사용되며, 복원 추출을 통해 여러개의 서브셋(subset)을 생성하고 각각의 서브셋에 대해 독립적으로 모형을 학습하여 결과를 평균내는 방법이다. 부스팅은 약한 학습기를 순차적으로 학습시켜 가며 오차를 줄여나가는 방법으로, 그래디언트 부스팅이나 캣부스트 등이 이에 속한다. 스택킹은 다양한 모형의 예측결과를 새로운 메타 모형의 입력으로 사용하여 최종 예측을 하는 방법이다.

Autogluon의 경우, 앙상블 모델링에 매우 효과적인 방법으로 알려져 있다. 특히 사용자가 설정한 시간 및 리소스(resource) 한계 내에서 다양한 알고

리즘과 하이퍼파라미터를 사용하여 여러 모형을 학습하고, 이들을 앙상블 하는 기능을 가지고 있다. 따라서 복잡한 전처리 과정 없이도 다양한 데이터에 대해 뛰어난 성능을 보이며, 이러한 특성은 여러 분야에서 응용할 수 있는 유연성을 제공한다. 본 연구에서 구축한 앙상블 모형을 Fig. 1에서 확인할 수 있다. Fig. 1의 앙상블 모형은 앙상블에 활용한 모형들의 출력값과 입력값을 연결하여 입력값으로 만들어 준다. 그 후 다시 앙상블을 수행하여 최종 예측값을 출력하는 모형이다.

2.3 딥러닝 기반 모형

본 연구에서는 시계열 딥러닝에 기반을 둔 여러 모형들인, Nlinear, Dlinear, SCINET, LSTM 등을 활용하였다. 일반적인 시계열 딥러닝 모형은 Fig. 2와 같이 설명변수와 목적변수를 모두 예측하게 된다. 하지만 분석에 사용된 데이터는 시계열 데이터이지만, 추세 또는 계절성을 가지지 않는다는 점에서 그 특성이 일반적인 시계열 데이터와 다르다. 그래서 기존의 시계열 모형에서 사용하는 입력과 출력 형식을 그대로 적용할 수 없었고, 시계열 데이터의 특징과 표 형식(Tabular) 데이터의 특징을 함께 학습할 수 있도록 모형의 출력부분을 Fig. 3과 같이 수정하여 사용하였다. 이를 통해 본 연구에서는 데이터의 특성을 고려한 구조의 모형을 적용하였고, 더욱 정확한 예측 결과를 도출할 수 있었다. Fig. 2는 일반적인 시계열 모형의 입력값과 출력값의 형태를 도식화한 그림이다. Fig. 3은 본 연구에서 사용한 딥러닝 모형의 입력값과 출력값의 형태를 도식화한 그림이다. Fig. 2와 Fig. 3의 주된 차이로는 목적변수의 활용 방식에 있다. Fig. 2에서는 목적변수가 입력값 중 하나로 활용된다. 즉, 전 시점의 목적변수를 학습에 사용하여 현재 시점 이후의 목적변수를

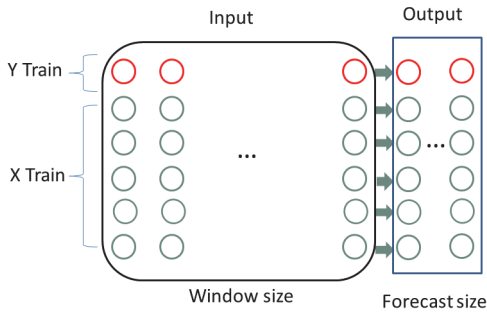


Fig. 2 General time series forecasting

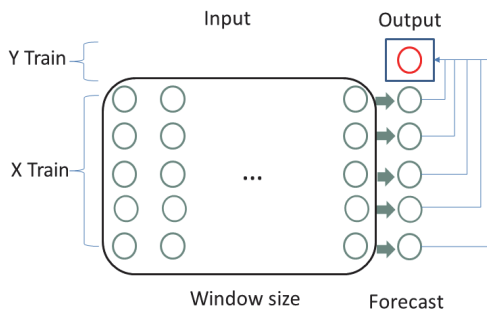


Fig. 3 Prediction method used in this paper

예측하게 된다. 반면, Fig. 3의 경우 목적변수는 입력값에 사용되지 않는다. 더불어, Fig. 3의 모형은 시계열 데이터로부터 얻어진 출력값을 다시 한번 선형 함수를 거쳐 최종 예측값으로 출력하게 된다. 이런 특성으로 인해 정형 데이터에 사용되는 모형들과 같은 조건에서 학습이 가능하고, 성능 비교가 가능하다.

2.3.1 NLinear, DLinear

NLinear 및 DLinear는 시계열 예측에서 뛰어난 성능을 보이는 모형으로, 특히 단변량 문제를 처리하는데 탁월하다. 시계열 모형 예측 성능 평가에 가장 많이 활용되는 Electricity Transformer Temperature(ETT) 데이터에서 가장 우수한 성능을 보여주는 것으로 알려져 있다. 두 모형은 Fig. 4의 Linear Model 기본 구조로 활용하였다. (Zeng &

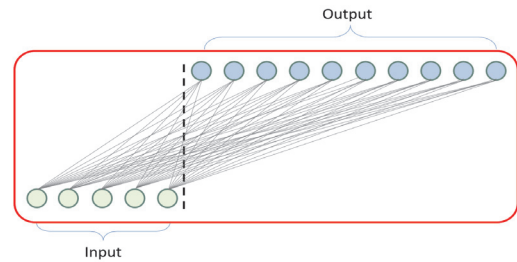


Fig. 4 Linear model

Ailing, 2022)[4]은 비교적 단순한 설명변수와 예측 변수의 선형구조모형을 통해 다른 딥러닝 방법보다 ETT 데이터에서 우수한 성능을 입증하였다.

NLinear는 데이터가 일정한 상승 또는 하락 추세를 보일 때 발생하는 문제를 해결하기 위해 설계되었다. 이 모형은 학습 데이터의 평균과 분산을 사용하여 데이터를 정규화하는데, 이 과정에서 평균 데이터의 분포가 이동할 수 있음을 인지하고 있다. 이러한 문제를 해결하기 위해, NLinear는 가장 최근의 값을 제외하고 모형을 학습시킨 후, 학습이 완료된 후에 이 값을 다시 추가함으로써 실제 값이 존재하는 분포로 이동시키는 방식을 사용한다.

한편, DLinear는 Autoformer[9]와 FEDformer [10]에서 사용되는 시계열 분해 방식을 단일 선형 레이어와 결합하여 사용하는 모형이다. 이 모형은 먼저 이동 평균을 계산하고, 이를 제거하여 시계열을 추세와 주기성 요소로 분해한다. 이후 각 요소에 대해 단일 선형 레이어를 적용하여 학습시킨 후, 이를 합산하여 최종 예측값을 도출한다. NLinear와 DLinear 모형은 각각의 독특한 방식으로 시계열 데이터의 복잡한 패턴을 학습하며, 이 과정에서 효과적인 성능을 보이는 것으로 알려져 있다.

2.3.2 SCINET

SCINET은 동적인 시간적 특성을 다양한 해상도에서 추출하기 위한 기법을 제공하는 모형이다. 이

모형은 두 개의 하위 시퀀스로 다운샘플링(down-sampling)을 수행하며, 이 과정에서도 시간적 연관성이 대부분 유지되는 특성을 활용한다. SCINET의 아키텍처는 재귀적인 다운샘플-컨볼루션-인터랙트의 구조로 이루어져 있어, 복잡한 시간적 동작을 효과적으로 모형링할 수 있다. 실험 결과, 이 모형은 다양한 시계열 예측 데이터셋에서 컨볼루션 모형과 transformer 기반 모형보다 우수한 성능을 보여주었다. 특정 작업에서는 NLinear 모형보다 더 나은 성능을 보여주기도 하며, 특히 예측 크기가 작은 경우 State of the Art(SOTA) 성능을 보여주었고 따라서 SCINET는 시간적 특성을 효과적으로 다루는 데에 유용한 도구로 평가될 수 있다.

2.3.3 LSTM

Hochreiter et al.[6]에 의해 개발된 Long Short Term Memory (LSTM) 모형은 순차적인 데이터를 학습시키는 딥러닝 모형 중 하나로, Recurrent Neural Network(RNN)의 단점을 보완한 모형이다. RNN은 순서가 있는 데이터를 학습시키고 처리하는데 적합한 모형이지만, 기울기의 소실(gradient vanishing) 문제나 기울기 폭발(gradient exploding)문제 때문에 장기적인 데이터는 기억하지 못하는 장기기억 의존성(long-term dependency)라는 치명적인 단점이 있다. LSTM은 RNN에 메모리 셀(memory cell)이라는 새로운 노드를 추가하여 기울기의 소실이나 기울기 폭발 문제를 해결해 장기기억 능력을 향상시킨 모형이다. 이러한 딥러닝 기반의 순차적인 데이터에 대한 모형들은 자연어 처리(NLP, Natural Language Processing)[11], 음성 인식(speech recognition)[12], 시계열 예측(time-series forecasting)[13] 등 광범위하게 사용되고 있다. RNN과 LSTM에 대한 자세한 내용은 Sherstinsky[14]에서 확인할 수 있다.

3. 공압기 전력 사용량 예측모형의 성능 비교

3.1 데이터 소개

본 연구에서 사용한 데이터는 2022년 8월 7일부터 2023년 1월 11일까지 (주)신한에어로의 공압기 제어 시스템, 인지컴퍼니(도장업체)에서 수집한 데이터이다. 공압기에 유량센서, 전력센서, 압력센서, 온도계, 습도계, 이슬점 측정기를 설치하여 관련 변수들을 10초 간격으로 매일 측정했다. 이 중, 센서와 기기의 결함으로 인해 잘못 측정된 날의 데이터는 사용하지 않고, 하루에 8640개의 관측값이 존재하는 날의 데이터만 사용했으며, 유량이 0인 데이터가 포함된 4일의 데이터 또한 본 분석에서는 제외하였다. 2022년 8월 7일부터 2022년 9월 25일까지 수집된 데이터는 (주)신한에어로에서 개발한 공압기 전력시스템을 적용하기 전이며, 2022년 9월 27일부터 2023년 1월 11일까지는 시스템을 적용한 후 수집된 데이터이다.

본 연구에서 사용한 데이터는 시스템 적용 후의 데이터 중 2022년 9월 27일부터 2023년 1월 11일까지 총 75일의 데이터를 사용했다. 그중 2022년 9월 27일부터 2022년 12월 22일까지 총 58일의 데이터가 Train 데이터로 사용되었고, 2022년 12월 23일부터 2023년 1월 11일 17일의 데이터가 Test 데이터로 사용되었다.

수집된 데이터의 변수는 총 29개이며 그중 8개 주요 변수의 시계열 그림과 기초 통계량을 각각 Fig. 5과 Table 1에 나타내었다. 빨간 점선은 (주)신한에어로의 공압에너지 절감 시스템의 적용 시점을 의미한다. 압력 센서 8개, 온도 센서 8개, 이슬점(dew point) 센서 3개, 온-습도계 센서 3개, 유량 측정기 1개로 데이터가 수집되었고, 제어 시스템의

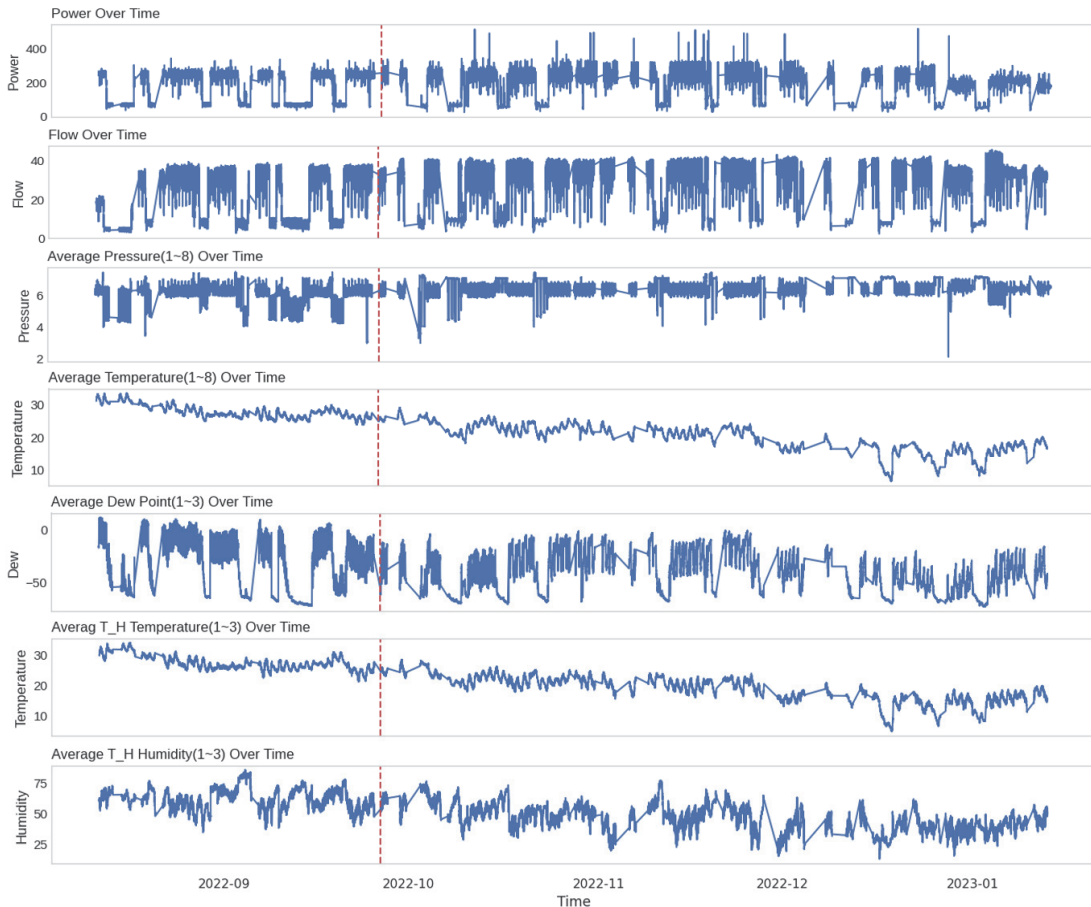


Fig. 5 Time-series plots for Variables(Power, Flow, Flow/Power, Pressure, Temperature, Dew point, Thermo-Hygrometer Temperature, Thermo-Hygrometer Humidity)

Table 1. Basic statistics for each variables

| Variable | Mean | Std. | KPSS (original) | KPSS (1-lag diff.) |
|--|--------|-------|-----------------|--------------------|
| Power | 188.24 | 74.19 | 1.86** | 0.006 |
| Flow | 28.10 | 12.38 | 1.10** | 0.003 |
| Power Efficiency | 0.14 | 0.02 | 15.47** | 0.001 |
| Average Pressure(1-8) | 6.37 | 0.48 | 6.04** | 0.005 |
| Average Temperature(1-8) | 19.54 | 4.14 | 93.13** | 0.048 |
| Average Dew Point(1-3) | -40.69 | 18.88 | 9.76** | 0.003 |
| Average thermo hygrometer Temperature(1-3) | 18.60 | 4.27 | 91.52** | 0.013 |
| Average thermo hygrometer Humidity(1-3) | 45.35 | 10.69 | 43.99** | 0.007 |

Std. : Standard deviation

Significance codes : ** for 1% and * for 5% levels;

사용 여부, 데이터 수집 시간, 전력 데이터를 확인할 수 있었다. 데이터 모형 학습에서는 수집시간을 제외한 27개의 변수가 설명변수로 사용되었고 전력 데이터가 목적변수로 사용되었다.

전력 데이터는 일반적으로 시간의 흐름에 따라 추세가 계절성이 있는 시계열 데이터의 특성을 가지고 있다. 그러나 이번 연구에서 분석한 공압기 전력 데이터는 1차 차분 후 KPSS(Kwiatkowski-Phillips-Schmidt-Shin) 검정결과 정상성은 만족하지만, 랜덤워크(random walk)의 특성을 가지는 것으로 나타났다. 데이터의 시각화를 통해 그 원인은 공압기의 사용량 즉, 유량이라는 목적변수의 무작위성에 있다는 것을 확인했다.

시계열 데이터가 랜덤워크의 성질을 가지면 통계적 모형은 물론 어떤 모형을 사용하더라도 종속변수를 잘 예측하지 못하게 된다. 실제로 이번 연구에서 ARIMA 모형과 단변수(univariate) 시계열 모형을 적용해 보았지만, 예측성능이 좋지 않다는 것을 확인했다. 따라서 시계열 분석 방법뿐 아니라 일반적인 정형 데이터에서 사용하는 예측모형 또한 사용하여 성능을 비교하기로 한다. 특히 시계열 분석에 사용되는 딥러닝 모형의 경우 시계열의 모형의 특성과 Tabular 모형의 특성을 모두 반영할 수 있도록 수정하여 사용하였고, Train set과 Test set의 경우 시계열 분석에 사용되는 시간 기반 분할 기법을 사용하였다.

3.2 모형 적합 방법

본 연구에서 사용한 모형 적합 모형으로는 설명변수들의 다중공선성을 고려하여 일반적인 다중선형회귀가 아닌 규제화(regularization) 회귀 모형인 Ridge, Lasso, Elastic Net을 사용하였다. 규제화 회귀방법으로 추정된 회귀계수를 통해 종속변수에 가장 큰 영향을 주는 변수를 확인할 수 있다.

Elastic Net으로 추정된 회귀계수 값들을 Fig. 6에 나타내었으며, 유량의 회귀계수가 가장 높은 것으로 나타났다.

비교적 다중공선성에 덜 민감한 랜덤포레스트 모형의 변수중요도를 Fig. 7에 나타내었다. Elastic Net과 랜덤포레스트 모두 유량이 종속변수에 가장 큰 영향을 주는 것으로 확인되었으나, Elastic Net과는 다르게 온도 변수의 중요도가 랜덤포레스트에서는 거의 영향이 없는 것으로 나타났다. 랜덤포레스트를 비롯한 다른 머신러닝 기법들은

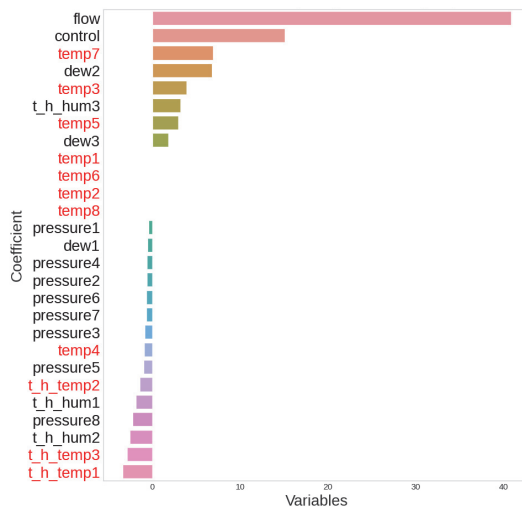


Fig. 6 Visualization of variable importance in elastic net model

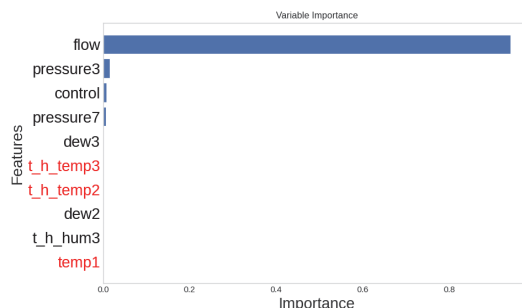


Fig. 7 Visualization of variable importance in random forest

시계열 모형이 아닌 일반적인 예측모형으로 구성하였으며, Bayesian Optimizer와 Hyperband 등의 알고리즘으로 최적화 한 뒤, 검증(Validation)을 통해 최적의 머신러닝 모형을 선택하였다.

딥러닝 모형은 시계열 모형을 현재 상황에 맞게 응용하였으며, 입력 데이터 크기를 결정하는 윈도우 사이즈(look-back window size)는 NLinear, DLinear, SCINet, LSTM 각각 64, 64, 96, 60을 사용하였고, 출력 데이터의 크기는 1로 설정하였다.

3.3 모형의 성능 비교

최종적으로 2022년 12월 23일부터 2023년 1월 11일까지 17일까지의 146,880개의 Test set으로 사용하여 21개의 모형에 적합하여 예측결과를 비교하였다. 예측모형의 성능 비교에 사용되는 평가 지표로는 RMSE (Root Mean Squared Error), MAE (Mean Absolute Error)와 전력예측 자료에 주로 사용되는 SMAPE (Symmetric Mean Absolute Percentage Error)를 사용하였다. 식 (6)-(8)에서 각 지표의 계산식을 확인할 수 있으며, 세 가지 지표 모두 값이 낮을수록 성능이 좋은 모형이다. 특히 SMAPE

의 경우 최솟값 0 최댓값 200을 가짐으로 스케일이 다른 변수 간 절대적인 비교가 가능하다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (7)$$

$$SMAPE = \frac{100}{n} \times \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (8)$$

Table 2에 각 모형에 대한 세 가지 지표들의 값을 나타내었다. 회귀분석, 머신러닝, 딥러닝 모형의 지표별 가장 낮게 나온 값을 굵게 표시하였으며, 소수점 셋째 자리까지 나타냈다. 모든 모형 중, 규제화 회귀분석 방법 중 하나인 Elastic Net이 RMSE는 21.36, MAE는 17.06, SMAPE는 10.14로, 세 지표 모두에서 값이 가장 낮아 우수한 성능을 보여주었다. Fig. 8에 Elastic Net으로 예측한 Test set의 전력을 그래프로 나타내었으며, 파란 실선이 실제 전력이며, 빨간 점선은 예측된

Table 2. Results of each model for power prediction

| Model type | Model | RMSE | MAE | SMAPE |
|--------------------|------------------------|--------------|--------------|--------------|
| Statistical Method | RFECV Regression | 26.67 | 21.32 | 13.28 |
| | Ridge Regression | 27.16 | 21.72 | 13.48 |
| | Lasso Regression | 23.15 | 18.26 | 10.54 |
| | Elastic Net Regression | 21.36 | 17.06 | 10.14 |
| Machine Learning | Neural Net Torch | 27.65 | 22.77 | 12.06 |
| | RandomForest | 30.06 | 25.31 | 13.08 |
| | CatBoost | 29.33 | 24.52 | 13.28 |
| | Weighted Ensemble | 29.86 | 25.43 | 13.85 |
| Deep Learning | NLinear | 26.78 | 21.39 | 12.95 |
| | DLinear | 27.10 | 22.04 | 13.35 |
| | SCINet | 29.27 | 24.70 | 14.83 |
| | LSTM | 34.24 | 29.02 | 21.42 |

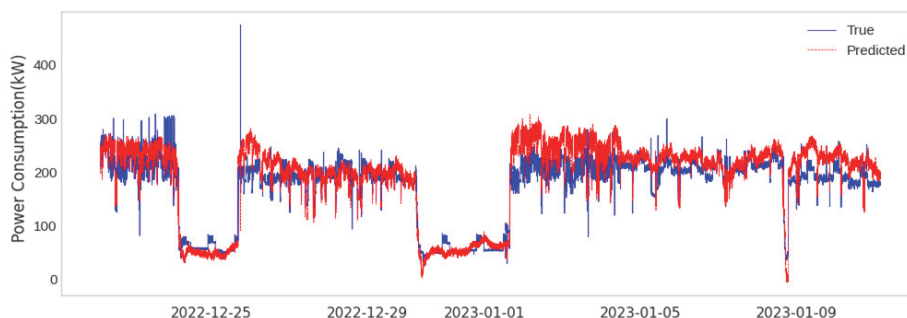


Fig. 8 Prediction results of the Elastic Net power forecasting model

전력이다. Elastic Net은 계절성에 따른 주기적인 변화를 반영하였으나, 일정 시점 이후부터 과대추정하는 경향을 보인다. 일반적으로 Tabular 데이터에 우수한 성능을 보이는 머신러닝 기반 모형들은 비교적 낮은 성능을 보여주고 있다. 이는 Test set이 Train set에 없는 특징을 가지고 있을 때 이러한 경향을 보이는데 Test set의 계절적 특징과 Train 셋의 계절적 특징이 서로 다르게 나타났기 때문인 것으로 추측된다. 또한 딥러닝 모형의 경우 긴 시간의 데이터를 통한 긴 시간의 예측에는 유리하지만 유량이 급변하는 추이를 바로 반영하지 못한다는 단점이 있다. 머신러닝 모형에서는 Neural Net이 RMSE, MAE, SMAPE가 각각 27.65, 22.77, 12.06으로 예측 성능이 가장 좋았고 딥러닝 모형 중에서는 Nlinear가 RMSE, MAE, SMAPE 각각 26.78, 21.39, 12.95로 가장 우수했다.

4. 결론

본 연구에서는 공압기의 소비전력 예측을 위해 통계적방법, 머신러닝 및 딥러닝 등 다양한 기법들을 사용해 모형을 개발하고 비교하였다. 그 결과, Elastic Net의 성능이 소비전력 예측에서 가장 뛰어났으며, 유량이 가장 큰 영향력을 가지는 것

으로 나타났다. 당연하게도 유량은 소비전력에 직접적인 영향을 주기 때문에 가장 큰 회귀계수를 보여주었으며, 이외에 온도와 이슬점 등이 소비전력에 큰 영향을 주는 것을 알 수 있었다. 또한 통계적 모형인 Elastic Net이 다른 복잡한 모형보다 좋은 성능을 보여주었고, 회귀계수를 통해 각 변수의 정확한 영향력을 계산할 수 있다.

이 연구 결과는 공압기의 전력 사용 최적화를 위한 첫걸음으로, 이를 바탕으로 추후 진행될 공압기의 전력 관리 시스템 설계와 개발에 중요한 참고 자료가 될 것으로 기대된다. 또한, Elastic Net 방법을 이용하여 효과적인 에너지 관리 전략을 수립하고, 이를 실제 현장에 적용함으로써 에너지 효율성을 극대화하는 방안을 모색할 수 있을 것이다. 이러한 접근법은 지속 가능한 공업 환경 구축에 기여하며, 에너지 소비를 줄이는 데 큰 도움이 될 것으로 예상된다.

사 사

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No. 2022-0-00761, 스마트제조 빅데이터 기반 인공지능 공압(Air) 통합제어 시스템 개발, 80%)과 2023

년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1C1C100711111, 통계적 이미지 매칭 알고리즘의 개발을 통한 범죄족적분석, 20%).

참고문헌

- [1] 장문수, 허요섭, 정현상 and 박소영: “유해가스 배출량에 대한 시계열 예측 모형의 비교연구” 한국산업융합학회논문집 24, no.3 (2021).
- [2] Breiman, Leo.: “Random forests.” *Machine learning* 45 (2001).
- [3] Prokhorenkova, Liudmila, et al.: “CatBoost: unbiased boosting with categorical features.” *Advances in neural information processing systems* 31 (2018).
- [4] Zeng, Ailing, et al.: “Are transformers effective for time series forecasting?.” *arXiv preprint arXiv:2205.13504* (2022).
- [5] Liu, Minhao, et al.: “Scinet: Time series modeling and forecasting with sample convolution and interaction.” *Advances in Neural Information Processing Systems* 35 (2022).
- [6] Hochreiter, Sepp, and Jürgen Schmidhuber.: “Long short-term memory.” *Neural computation* 9.8 (1997).
- [7] Klein, Aaron, et al.: “Model-based asynchronous hyperparameter and neural architecture search.” *arXiv preprint arXiv:2003.10865* (2020).
- [8] Erickson, Nick, et al.: “Autoglun-tabular: Robust and accurate automl for structured data.” *arXiv preprint arXiv:2003.06505* (2020).
- [9] Wu, Haixu, et al.: “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting.” *Advances in Neural Information Processing Systems* 34 (2021)
- [10] Zhou, Tian, et al.: “Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting.” *International Conference on Machine Learning*. PMLR, (2022).
- [11] Huang, Zhiheng, Wei Xu, and Kai Yu.: “Bidirectional LSTM-CRF models for sequence tagging.” *arXiv preprint arXiv:1508.01991* (2015).
- [12] Graves, Alex, Navdeep Jaitly, and Abdelrahman Mohamed.: “Hybrid speech recognition with deep bidirectional LSTM.” *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, (2013).
- [13] Sagheer, Alaa, and Mostafa Kotb.: “Time series forecasting of petroleum production using deep LSTM recurrent networks.” *Neurocomputing* 323 (2019).
- [14] Sherstinsky, Alex.: “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.” *Physica D: Nonlinear Phenomena* 404 (2020).

(접수: 2023.07.17. 수정: 2023.07.28. 게재확정: 2023.08.04.)