Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.3.123

Vol. 46, No. 3 : 123-138, September 2023
ISSN : 2005-0461 / EISSN : 2287-7975

A Study on the Health Index Based on Degradation Patterns in Time Series Data Using ProphetNet Model

Sun-Ju Won* · Yong Soo Kim**†

*Department of Industrial and Systems Engineering, Kyonggi University Graduate School
**Department of Industrial and Systems Engineering, Kyonggi University

ProphetNet 모델을 활용한 시계열 데이터의 열화 패턴 기반 Health Index 연구

원선주* · 김용수**†

*경기대학교 일반대학원 산업시스템공학과
**경기대학교 산업시스템공학과

The Fourth Industrial Revolution and sensor technology have led to increased utilization of sensor data. In our modern society, data complexity is rising, and the extraction of valuable information has become crucial with the rapid changes in information technology (IT). Recurrent neural networks (RNN) and long short-term memory (LSTM) models have shown remarkable performance in natural language processing (NLP) and time series prediction. Consequently, there is a strong expectation that models excelling in NLP will also excel in time series prediction. However, current research on Transformer models for time series prediction remains limited. Traditional RNN and LSTM models have demonstrated superior performance compared to Transformers in big data analysis. Nevertheless, with continuous advancements in Transformer models, such as GPT-2 (Generative Pre-trained Transformer 2) and ProphetNet, they have gained attention in the field of time series prediction. This study aims to evaluate the classification performance and interval prediction of remaining useful life (RUL) using an advanced Transformer model. The performance of each model will be utilized to establish a health index (HI) for cutting blades, enabling real-time monitoring of machine health. The results are expected to provide valuable insights for machine monitoring, evaluation, and management, confirming the effectiveness of advanced Transformer models in time series analysis when applied in industrial settings.

Keywords : Health Index, Degradation Patterns, Time Series Data, ProphetNet Model

1. 서 론

1.1 연구 배경 및 필요성

4차 산업혁명과 센서 기술의 발전으로 인해 센서 데이터의 활용이 증가하고 있다. 현대 사회에서는 데이터의 복잡성이 점점 더 증가하고 있으며, IT(Information Technology) 기술의 빠른 변화에 따라 데이터에서 얻을 수 있는 정보의 중요성도 커지고 있다[13].

자연어 처리에서 널리 사용되었던 RNN(Recurrent Neural Network)과 LSTM(Long Short-Term Memory) 모델은 시계열 예측에서도 우수한 성능을 보여왔고, 현재 자연어 처리 분야에서 뛰어난 성과를 보이는 트랜스포머(Transformer)

Received 21 July 2023; Finally Revised 5 September 2023; Accepted 5 September 2023
† Corresponding Author : kimys@kgu.ac.kr

도 시계열 예측에 관한 많은 연구가 진행되고 있다[11]. 자연어 처리 분야에서 뛰어난 성능을 보이는 모델이 시계열 예측 분야에서도 우수한 성능을 발휘할 수 있으므로, 트랜스포머 모델의 시계열 예측 성능을 확인하고자 한다.

시계열 데이터는 어려운 불규칙성을 가지고 있으며, 외부 요인이나 예기치 않은 이벤트로 인해 예측 결과에 영향을 주는 경우가 많다. 이러한 도전과제를 해결하기 위해 기계 학습과 인공지능 기술이 시계열 데이터 처리에 활용되고 있고, GPT-2(Generative Pre-trained Transformer 2)와 ProphetNet은 그중에서도 주목받는 모델이다[19]. GPT-2는 시퀀스 데이터에 대한 생성 모델로, 문맥을 이해하고 다음 단어를 예측하는 능력을 갖추고 있어 시계열 데이터에서도 패턴 인식과 예측 작업에 활용할 수 있다.

ProphetNet은 Facebook에서 개발한 시퀀스 예측 모델로, 시계열 데이터의 특징인 주기성, 계절성(Seasonality) 및 추세성(Trend)을 모델링하고 예측 결과를 생성하는 능력을 갖추고 있어 긴 시퀀스를 처리할 수 있는 능력에 초점을 맞추어 개발되었다. 따라서, 대표적인 모델을 활용하여 잔여수명 범위를 예측할 때, 각 모델의 분류 정확도를 확인하고자 한다.

본 연구는 지속적으로 센서 데이터를 수집하여 신호 처리를 수행하고 특징을 추출한 뒤, 모델을 활용하여 시스템의 잔여수명 범위에 대한 확률값을 계산하여 건전성 상태를 평가하고자 한다. 건전성 지표(Health Index, HI)는 다양한 산업 분야에서 시스템, 장비, 또는 구성 요소의 건강 상태를 정량화하고 모니터링하기 위해 개발되었다.

열 수축 필름 포장기의 절단 블레이드의 성공적인 유지보수는 기계의 수명을 늘릴 수 있으며, 기계 신뢰성을 향상시킬 수 있는 대안으로 절단 블레이드의 열화 상태 관리가 필수적이다[28]. 절단 블레이드는 절단기, 포장기, 회전기의 위치 오차, 속도 등 다양한 센서를 통해 블레이드의 상태를 파악할 수 있다.

따라서, 본 연구는 센서 데이터를 활용하여 기계의 HI를 기반으로 체계적인 상태 진단을 수행하는 프로세스를 제안하고자 한다. 연구의 최종 목표는 실제 응용에 적합한 상태 진단을 기반으로 적절한 관리 정책을 수립하여 다양한 산업에서의 운용손실을 최소화하고 제조 산업에서 효율성을 향상시켜 제품 품질을 향상시키는 데에 도모하는 것이다.

1.2 연구절차 및 구성

본 연구는 열 수축 필름 포장기의 절단 블레이드의 열화 패턴을 분석하고, 잔여수명 범위의 확률을 확인하여 HI를 수립하기 위한 프로세스를 제안하기 위해 연구절차 및 구성은 다음과 같다.

먼저, 필드 데이터를 수집하고 변동성이 없는 변수를 제거한 후 사이클을 설정한다. 그 이후, 분석 모델을 선정하고 데이터를 사용하여 모델 학습을 진행한다. 하이퍼파라미터 조정과 교차 검증을 통해 학습된 모델을 평가하고, 잔여수명 범위 예측 성능을 확인하여 가장 우수한 성능을 보이는 모델을 식별하고자 하였다. 이와 함께, 본 연구절차를 수행하는 과정에서 체계적인 HI 수립방안을 도출할 수 있다.

너무 많은 센서 데이터는 데이터의 양과 복잡성으로 인해 전처리 과정이 필요하며, 중요한 공정 변수를 식별하는 과정이 필요하다. 데이터셋은 노이즈를 포함하고 있으며 불균형한 특성이 있어 변수 선택(feature selection)은 어려운 문제 중 하나이다. 이를 해결하기 위해 Stepwise 변수 선택법을 활용하였고, 변동성이 없는 변수를 제거하였다.

또한, 성능이 가장 좋은 모델을 식별하고, 절단 블레이드의 열화 패턴을 가장 잘 분류하는 모델을 활용하여 잔여수명 범위 예측 확률을 도출하였다. 이를 통해 효율적인 의사결정을 지원하고 리스크를 최소화하며, 예방적 유지관리가 가능해졌다.

2. 관련 문헌 연구

본 연구에서는 문헌연구를 시계열 데이터 분석, 패턴 분류, 그리고 HI 세 가지 부문으로 분류하여 수행하였다. 시계열 데이터 분석 부문에서는 다양한 알고리즘을 활용하여 예측 성능을 비교하는 연구를 중점적으로 조사하였다. 패턴 분류 부문에서는 모델의 예측 정확도를 향상시키는 분류 방법을 제시하는 연구를 대상으로 실시하였다. 또한, HI 방법의 시장 동향과 수행된 연구 현황에 대한 조사를 진행하였다.

시계열 데이터를 사용하여 모델의 예측 성능을 비교하는 연구가 중요한 주제로 다루어지고 있다. 최근 몇 년간 시계열 데이터 분석에서 다양한 알고리즘을 활용하여 예측 정확도를 향상시키는 연구가 많이 이루어졌다. Lim et al.[14]은 네트워크 이상 탐지와 관련된 작업에서 LSTM, GRU(Gate Recurrent Unit), BERT(Bi-directional Encoder Representations from Transformers)와 같은 자연어 처리에서 뛰어난 성능을 보여주었던 모델들의 성능을 비교 분석하였다. Oh and Kim[18]은 미세한 움직임을 분류하기 위해 SVM(Support Vector Machine), LSTM, RNN, ConvLSTM(Convolutional LSTM) 등 4가지 인공지능 모델의 성능을 비교하고 분석하였다. Jang et al.[10]은 LSTM과 트랜스포머 모델의 성능을 비교하고, 고속도로의 교통흐름 예측에 대한 실험을 수행하였다.

최근 몇 년간 패턴 분류 분야에서는 다양한 알고리즘과

이미지 형태의 데이터를 활용하여 패턴 분류 모델의 예측 정확도를 향상시키는 연구가 활발하게 이루어지고 있다. 이러한 연구들은 주로 분류 모델을 사용하여 새로운 분류 방법을 제안하고 모델의 예측 정확도를 개선하는 데 초점을 맞추고 있다. Yi et al.[26]은 강건 회귀와 시계열 클러스터링 K-Shape 알고리즘을 활용하여 건물 에너지 패턴의 분류 방법을 제안하였고, Yoo and Park[27]은 디지털 음원의 순위 변화의 패턴을 세분화하여 음원 연구에 새로운 접근을 시도하였다. Qi et al.[19]은 ProphetNet 모델을 활용하여 n-gram 예측을 수행하였고, 대규모 데이터셋을 학습하여 동일한 규모를 가진 모델과 비교하여 최고의 성능을 보였다.

Wafer map과 같이 2차원 공간 데이터를 가진 이미지 형태를 활용하고 이미지 분류에 강점을 가진 CNN (Convolutional Neural Network)을 사용하여 불량 패턴 분류를 시도하였고, Saqlain et al.[22]은 로지스틱 회귀, 랜덤 포레스트(Random Forest, RF), 그래디언트 부스팅 머신 및 인공 신경망 등을 기계 학습 분류기로 활용하고, 추출된 특징을 사용하여 학습을 수행하였다. Sandoval et al.[21]은 IoT 모트에서 제공되는 수신 신호 강도 지시자를 활용하여 일련의 수학적 보정 및 측정 절차를 거쳐 패턴을 분류하였다. 이미지 분류와 패턴 분류에 적용 가능한 다양한 기법과 알고리즘을 제시하고 있으며, 실제 응용에서의 성능과 유용성을 탐구하는 데에 중점을 두고 있다.

마지막으로 HI 분야에서는 주로 기계 및 환경 자원 데이터의 유지보수를 위해 HI 방법을 제안하는 연구가 중심적으로 이루어지고 있다. 다양한 알고리즘을 활용하여 HI의 성능을 향상시키거나 효과적으로 HI 점수를 도출하는 새로운 연구가 많이 이루어졌다. Rediansyah et al.[20]은 전력 변압기의 상태를 빠르고 효율적으로 평가하는 도구로 HI를 사용하였으며, KNN(K-Nearest Neighbors), SVM, AdaBoost, RF, NB(Naive Bayes), ANN(Artificial Neural Network), Decision Tree의 방법을 비교하였다. Bohatyrewicz and Mrozik[4]은 주기적인 오일 진단을 포함하는 HI 방법을 제안하였고, 4개의 하위 집단으로 분류하여 보정 유지관리가 HI 점수의 평균값에 미치는 영향을 보여주었다. Murugan and Ramasamy[16]는 고장 데이터를 통계 분석하여 전기적, 열적, 기계적 및 절연 저하를 포함하는 네 가지 주요 고장 수준으로 그룹화하여 유지보수를 위한 HI 방식을 제안하였다.

시계열 데이터를 패턴 분류하여 분류된 데이터로 HI를 수립하는 연구는 현재 매우 부족한 실정이다. 고장 패턴이 분명하게 나타나는 실제 시계열 필드 데이터를 기반으로 하는 분석 및 HI 수립 방법론의 개발이 시급하다. 본 연구는 최근 연구에서 사용된 시계열 데이터 패턴 분류 방법론과 HI 방법론을 기반으로 필드의 특성을 반영한 시계열 데이터의 패턴 분류 및 HI 수립 프로세스를 개발 및 제시하고자 한다. 이 과정에서, 변압기의 건전성 상태를 분류하기 위해 입력 데이터 특성에 기반한 훈련 데이터와 패턴 인식 도구를 사용한 Alqudsi and El-Hag[3]의 연구를 참고하였다. 또한, 현재 트랜스포머 모델은 자연어 처리 분야에서 주로 활용되고 있으므로, 이 모델들의 시계열 분야에서의 성능을 확인하고 시계열 데이터 분석을 위해 모델을 사용할 수 있는 환경을 조성하는 데 활용하고자 한다.

3. 이론적 배경

3.1 시계열 데이터 분석

본 논문은 시계열 데이터의 패턴 분석을 통해 패턴을 분류하고 고장 시점 범주를 예측하여 HI를 도출하는 문제를 다룬다. 시계열 데이터 분석은 시간에 따라 기록되는 데이터를 분석하는 기술이다. 시계열 데이터는 일련의 측정값으로 시간에 따라 변화하는 어떤 현상을 나타내며, 주요 목표는 데이터의 패턴을 파악하고 예측하는 것이다. 시계열 데이터 분석에는 시계열 모델링 기술이 활용된다. 시계열 모델링은 시계열 데이터의 특성을 이해하고 예측 모델을 구축하는 것을 목표로 한다. 시계열 데이터 분석은 다양한 산업 분야에서 활용되며, 공정 제어, 장비 예측 등에 활용된다. 이를 위해 RNN, LSTM, GPT-2, ProphetNet과 같은 시퀀스 모델링 기법이 적합하다. RNN과 LSTM은 과거 시점의 정보를 현재 시점의 예측에 활용하여 시계열 데이터 분석에 우수한 성능을 보인다.

3.1.1 RNN

RNN은 연속적으로 변화하는 시계열 데이터를 다루기 위한 인공 신경망 중 하나로, 시퀀스 데이터를 처리하는 데에 주로 사용된다. 각 시퀀스의 요소를 차례로 입력으로 받아들이며 내부 상태를 갱신하고 출력을 생성한다. 이 내부 상태는 이전 시간의 입력과 현재 입력을 고려하여 업데이트된다[7]. RNN은 순환 구조로 되어 있어 이전 시간 단계의 정보를 현재 시간 단계로 전달할 수 있다.

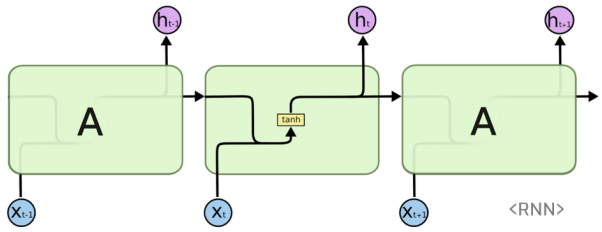
RNN은 <Figure 1>과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성된다. 입력층은 시퀀스의 각 요소를 받아들이는 역할을 담당한다. 은닉층은 내부 상태를 갱신하고 다음 단계로 전달하는 임무를 수행하며, 출력층은 RNN의 최종 출력을 생성한다. 은닉층과 출력층은 일반적으로 활성화 함수를 통해 비선형 변환을 수행한다[25]. 역전파(backpropagation) 알고리즘을 사용하여 학습이 이루어지며, 역전파는 출력과 정답 간의 오차를 최소화

하기 위해 가중치를 조정한다. 이렇게 시간의 흐름에 따라 그래디언트를 전파하면서 RNN은 학습된다[6].

그러나 RNN은 긴 시퀀스 데이터에서 기울기 소실(gradient vanishing) 문제가 발생한다. 역전파 시 그래디언트가 지수적으로 감소하거나 증가하여 학습이 어려워지는 문제이며, 이를 해결하기 위해 LSTM처럼 변형된 RNN 구조가 개발되었다[23].

$$y = f(\sum_{i=1}^{n} w_i x_i - h) \tag{1}$$

처음으로 제안된 RNN의 수식은 식 (1)과 같은 다중 선형 모형이다. RNN의 활동은 누적된 자극의 크기($\sum w_i x_i$)가 역치 h보다 작으면 0을 유지하고, 커지면 1을 출력하는 함수 f로 해석 가능하며 초기에는 뉴런들을 조합하여 AND와 OR 같은 논리 회로를 구현하는 목적의 연구가 진행되었다.

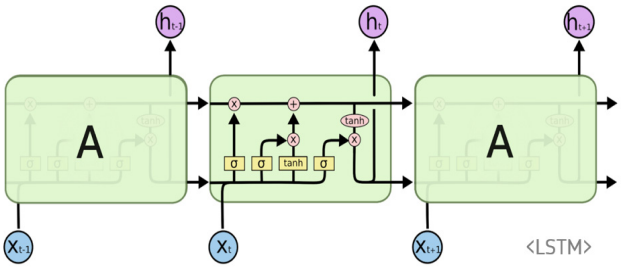


<Figure 1> RNN Structure

3.1.2 LSTM

LSTM은 RNN의 변형된 구조로, 주로 시퀀스 데이터를 처리하고 장기 의존성(Long-term Dependency)을 모델링하는 데에 사용된다. RNN의 기울기 소실 문제를 해결하기 위해 개발되었으며, RNN과 유사하게 시퀀스 데이터의 각 요소를 차례로 입력으로 받아들이면서 내부 상태를 갱신하고 출력을 생성한다. 하지만 LSTM은 메모리 셀(memory cell)이라고 불리는 구조적인 유닛을 도입하여 장기 의존성을 기억하고 관리한다[5].

메모리 셀은 <Figure 2>와 같이 입력 게이트(input gate), 망각 게이트(forget gate), 출력 게이트(output gate)로 구성된다. 입력 게이트(x_t)는 시그모이드 함수를 활용하여 현재 입력을 얼마나 기억할지 업데이트 여부를 결정하고, 후보 값($\widetilde{h_t}$)을 만든 후 이전 셀 상태(h_{t-1})를 업데이트한다. 망각 게이트는 이전 셀 상태에서 어떤 정보를 삭제할지를 결정한다. 출력 게이트(y_t)는 메모리 셀의 상태를 얼마나 출력할지를 결정한다. LSTM의 주요 장점은 각 게이트에 의해 제어되는 메모리 셀의 상태를 통해 장기 의존성을 유지할 수 있다는 것이다.



<Figure 2> The Structure of LSTM and Memory Cell

3.1.3 GPT-2

GPT-2는 OpenAI에서 개발한 자연어 처리 모델로, 대규모 텍스트 데이터를 사전 학습한 후 다양한 자연어 처리 작업에 활용할 수 있는 강력한 언어 모델이다. GPT-2는 GPT-1(Generative Pre-trained Transformer 1) 모델의 개선된 버전으로, 더 많은 매개변수와 더 큰 모델 크기를 가지고 있어 더욱 풍부하고 의미 있는 텍스트를 생성할 수 있다[29].

최근에는 GPT-2가 시계열 데이터 분석에서도 활용되고 있다. 시계열 데이터를 처리하기 위해서는 시계열 데이터를 일련의 텍스트로 변환한 후 GPT-2에 입력으로 제공한다. 이를 위해 시계열 데이터의 값을 텍스트 형식으로 변환한 후, 입력된 텍스트를 기반으로 다음 값을 예측하거나 특정 시점에서의 데이터를 생성하는 것이 가능하다[15].

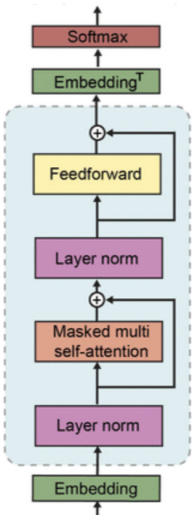
GPT-2는 L개의 토큰으로 구성된 시퀀스 X를 생성하는 모델로, 이때 발생 가능한 총 토큰 개수를 K라고 가정한다면, x_L은 $0 \le x_L \le (K-1)$을 만족하는 정수가 된다.

$$X = [x_0 \cdots x_{(L-1)}]^\top \tag{2}$$

입력된 토큰을 기반으로 다음에 등장할 토큰을 예측하므로, 입력 시퀀스 X에 대한 출력 시퀀스인 식 (3)의 확률을 예측한다고 할 수 있다.

$$\widetilde{X} = [x_1 \cdots x_L]^\top \tag{3}$$

이러한 구성에서 토큰 x_L은 E차원 임베딩 벡터로 변환된 후, 데이터 내의 해당 위치에 따라 E차원 위치 임베딩(positional embedding) 벡터가 더해져 y_L로 변환된다. 이때, 위치 임베딩 벡터는 데이터의 길이(L)만큼 서로 다른 벡터들로 구성된 벡터 집합으로부터 해당 데이터의 위치에 따라 해당 벡터를 선택하여 사용된다. 결과적으로 시퀀스 X는 식 (4)로 변환된다. Y를 구성하는 L개의 벡터 $\{y_t\}$는 H개의 헤드(head)들로 구성된 Masked multi self-attention 계층에 동시에 입력된다.



<Figure 3> GPT-2 Structure

$$Y = [y_0 \cdots y_{(L-1)}] \tag{4}$$

3.1.4 ProphetNet

ProphetNet은 Facebook에서 개발된 시퀀스 예측 모델로, 주기성, 계절성, 추세성과 같은 시계열 데이터의 특성을 모델링하고 예측 결과를 생성하는 능력을 갖추고 있다. 이 모델은 트랜스포머 아키텍처를 기반으로 하며, 세그먼트 어텐션(segment attention) 및 프로젝션(projection) 메커니즘과 같은 기술을 활용하여 입력 시퀀스를 처리하는 데 뛰어난 성능을 발휘한다[9].

주어진 시퀀스 쌍 (x, y)에서 $x = (x_1, \cdots, x_M)$은 M개의 토큰으로 구성된 입력 시퀀스이고, $y = (y_1, \cdots, y_T)$는 T개의 토큰으로 구성된 대상 시퀀스이다. Seq2Seq 모델은 조건부 확률 $p(y|x)$를 모델링하는 것을 목표로, 식(5)로 분해될 수 있다.

$$p(y|x) = \prod_{t=1}^{T} p(y_t|y_{<t}, x) \tag{5}$$

식 (5)에서 $y_{<t}$는 위치 t 이전의 토큰을 나타낸다. 일반적으로 Seq2Seq 모델은 입력 시퀀스 표현을 인코딩하는 인코더와 이전 대상 토큰을 입력으로 사용하여 조건부 확률을 모델링하는 디코더로 구성된다. 입력 시퀀스는 인코더에 주입되어 데이터를 임베딩하고, 인코더의 셀프 어텐션(self-attention) 메커니즘을 통해 입력 시퀀스의 특징을 추출한다. 디코더는 이전 데이터들을 기반으로 다음 데이터를 예측하며, 셀프 어텐션과 다중 헤드 어텐션(multi-head attention)을 사용하여 데이터의 흐름을 파악하고 패턴을 이해한다[24]. 모델은 각 시간 단계에서 이전의 정답 문맥 토큰 $y_{<t}$와 x가 주어졌을 때 다음 대상 토큰 y_t를 예측하도록 최적화된다.

트랜스포머와는 다른 새로운 접근으로 자기 지도 학습(Self-Supervised Learning) 기법 중 하나인 미래 n-gram 예측을 도입하였다. ProphetNet은 원래 Seq2Seq 모델의 최적화 방법인 다음 단일 토큰 예측을 식 (6)과 같이 수행하는 대신, 시간 단계마다 변환되는 방식으로 변경하여 식 (7)과 같이 구현하였다. 식 (7)에서 $y_{t:t+n-1}$은 다음 연속된 n개의 미래 토큰을 나타내며, 동시에 예측한다.

$$p(y_t|y_{<t}, x) \tag{6}$$

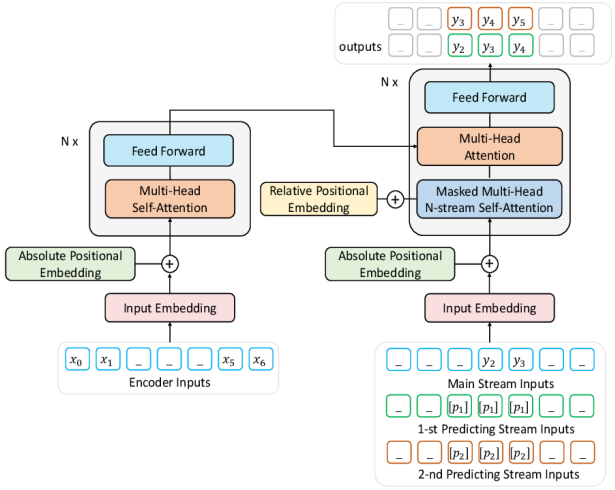
$$p(y_{t:t+n-1}|y_{<t}, x) \tag{7}$$

ProphetNet은 다중 헤드 셀프 어텐션 메커니즘을 가진 다층 트랜스포머 인코더와 제안된 다중 헤드 n-stream 셀프 어텐션 메커니즘을 가진 다층 트랜스포머 디코더로 구성된다. 입력 시퀀스가 주어지면, ProphetNet은 식 (8)에서 나타내는 방식으로 입력 시퀀스를 인코딩하여 시퀀스 표현을 생성한다. 이는 원래의 트랜스포머 인코더와 동일하다.

$$H_{enc} = \text{Encoder}(x_1, \cdots, x_M) \tag{8}$$

ProphetNet 디코더는 각 시간 단계에서 다음 토큰 하나만 예측하는 것이 아니라, 앞에서 언급한 것처럼 n개의 미래 토큰을 동시에 예측한다.

$$\begin{aligned} &p(y_y|y_{<t}, x), \cdots, p(y_{t+n-1}|y_{<t}, x) \\ &= \text{Decoder}(y_{<t}, H_{enc}) \end{aligned} \tag{9}$$



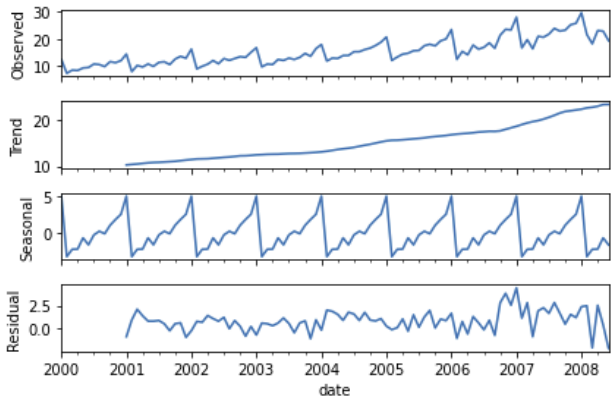
<Figure 4> ProphetNet Structure

3.2 패턴 분류

시계열 데이터의 패턴 분류는 시계열 데이터에서 특정한 패턴을 식별하고, 해당 패턴에 기반하여 데이터를 분류하는 과정을 의미한다. 시계열 데이터는 다양한 패턴을 포함하고 있고, 이러한 패턴을 분석하고 분류하는 것은 데이터의 이해와 예측에 매우 유용하다. 패턴 분류는 다양한 방법과 기법을 활용할 수 있다. 일반적으로 사용되는 패턴 분류 기법으로는 시계열 분해, 패턴 인식 기법, 시계열 군집 분석 등이 있다. 본 논문에서는 이 세 가지 기법을 모두 적절히 조합하여 시계열 데이터의 패턴을 분류하는 데 사용하였다.

3.2.1 시계열 분해

시계열 분해는 시계열 데이터를 추세성, 계절성, 순환성(Cyclic), 불규칙성(Irregularity)의 구성 요소로 분해하는 기법이다. 이를 통해 데이터의 각 요소에 대한 패턴을 식별할 수 있으며, 본 연구에서는 추세성과 순환성을 중점적으로 분석하였다. 추세성은 시계열 데이터의 장기적인 변동 패턴을 나타내며, 데이터가 점진적으로 증가하거나 감소하는 경향을 가지는 경우를 의미한다. 추세성을 분석하면 시계열 데이터의 장기적인 변동을 파악할 수 있으며, 예측 모델링에 유용한 기법이다. 순환성은 시계열 데이터에서 나타나는 주기적인 패턴을 의미한다. 계절성과 유사한 개념으로 이해할 수 있지만, 주기의 길이가 정해져 있지 않을 수 있다. 시계열 데이터에서 순환성을 분석하고 모델링하면, 데이터의 특정 주기를 이해하고 주기적인 변동을 예측하는 데 도움이 된다[12].



<Figure 5> Example Graph of Time Series Decomposition

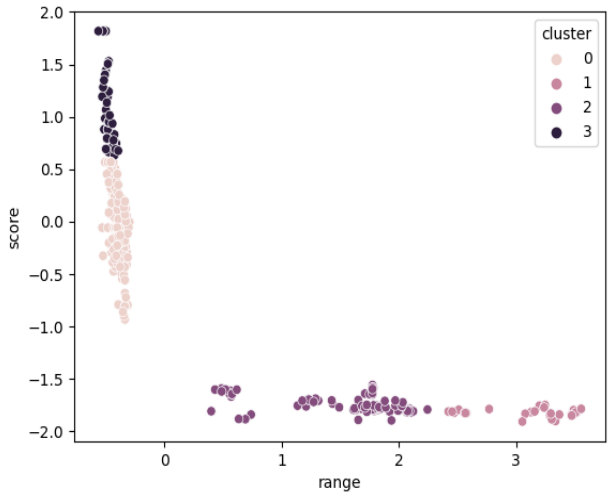
3.2.2 패턴 인식 기법

패턴 인식 기법은 시계열 데이터의 패턴을 자동으로 학습하고 분류하는 기법이다. 이를 위해 머신러닝 및 딥러닝 기법을 활용하여 모델을 구축하고 데이터를 분류한다. 입력된 시계열 데이터를 특징적인 패턴이나 구조로 변환하고, 이를 기반으로 데이터의 클래스 또는 카테고리를 분류한다. 주요한 단계로는 데이터 전처리, 특징 추출, 분류 모델 구축이 포함된다. 데이터 전처리 단계에서는 시계열 데이터를 정규화하거나 표준화하는 등의 전처리 작업을 수행한다. 다음으로, 특징 추출 단계에서는 시계열 데이터에서 유용한 특징을 추출하기 위해 다양한 방법을 활용한다. 이를 통해 데이터의 중요한 패턴이나 구조를 표현하는 특징 벡터를 생성한다. 마지막으로, 분류 모델 구축 단계에서는 추출된 특징 벡터를 입력으로 사용하여 학습 알고리즘을 적용하여 모델을 구축한다.

시계열 데이터의 패턴을 학습하고, 새로운 입력 데이터를 분류하는 데 사용되는 알고리즘으로는 딥러닝 모델인 RNN, LSTM, GPT-2, ProphetNet 등이 활용될 수 있다. 패턴 인식 기법은 데이터의 패턴을 자동으로 학습하므로 사전에 정의된 규칙이나 패턴을 사람이 수동으로 지정하지 않아도 된다. 따라서 다양한 시계열 데이터에 적용 가능하며, 예측, 분류, 이상 탐지 등 다양한 응용 분야에서 활용될 수 있다[1].

3.2.3 시계열 군집 분석

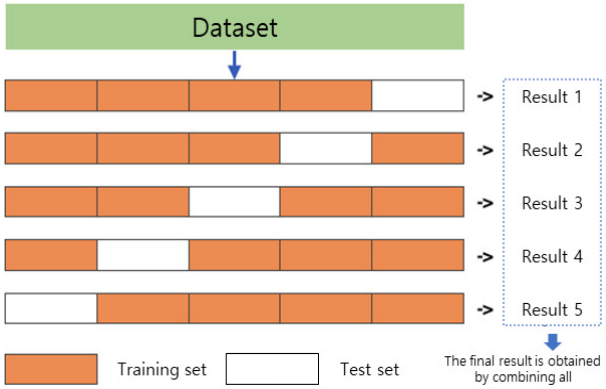
시계열 군집 분석은 비슷한 패턴을 보인 시계열 데이터를 군집화하는 기법이다. 서로 다른 그룹으로 분류할 수 있으며, 각 그룹은 특정한 패턴을 공유한다. 데이터 간의 유사성을 측정하고 클러스터링 알고리즘을 적용하여 유사한 패턴을 보인 데이터를 동일한 클러스터에 할당한다. 일반적으로는 거리 기반 측정 방법이 주로 활용되고, 클러스터링 알고리즘으로는 계층적 군집화, K-means, DBSCAN 등이 일반적으로 활용된다[2].



<Figure 6> Example of K-means Clustering Analysis Graph

3.2.4 5-fold 교차 검증 방법

5-fold 교차 검증은 머신러닝 모델의 성능을 평가하는데 널리 사용되는 교차 검증 방법 중 하나이다. 이 방법은 데이터셋을 5개의 서로 다른 부분 집합으로 분할하여 모델의 성능을 평가하는 과정을 반복한다. 먼저, 원본 데이터셋을 무작위로 섞은 후 데이터셋을 5개의 동일한 크기로 분할한다. 이때, 각 부분 집합을 '폴드'라고 한다. 다음으로, 첫 번째 폴드를 테스트 데이터셋으로 선택하고, 나머지 4개의 폴드를 훈련 데이터셋으로 사용하여 모델을 학습시킨다. 학습된 모델을 첫 번째 폴드의 테스트 데이터셋에 적용하여 예측 결과를 얻는다. 이후, 두 번째부터 다섯 번째까지의 폴드를 차례로 테스트 데이터셋으로 선택하고, 나머지 폴드를 훈련 데이터셋으로 사용하여 모델을 학습시키고 예측을 수행한다.



<Figure 7> An Example of 5-fold Cross-validation

예측을 수행한 후, 정확도, 정밀도, 재현율, F1-score 등의 평가 지표를 사용하여 5개의 폴드에 대한 예측 결과를 평가하고 모델의 성능을 측정한다. 위 과정을 5번 반복하여 각 폴드에서 얻은 성능 평가 지표를 평균 내어 최종적인 성능 평가를 얻는다. 5-fold 교차 검증은 데이터를 효과적으로 활용하여 모델의 일반화 성능을 평가할 수 있도록 한다. 또한, 과적합을 방지할 수 있다는 장점이 있으며, 이를 통해 모델의 성능을 신뢰할 수 있게 평가할 수 있다.

3.3 Health Index(HI)

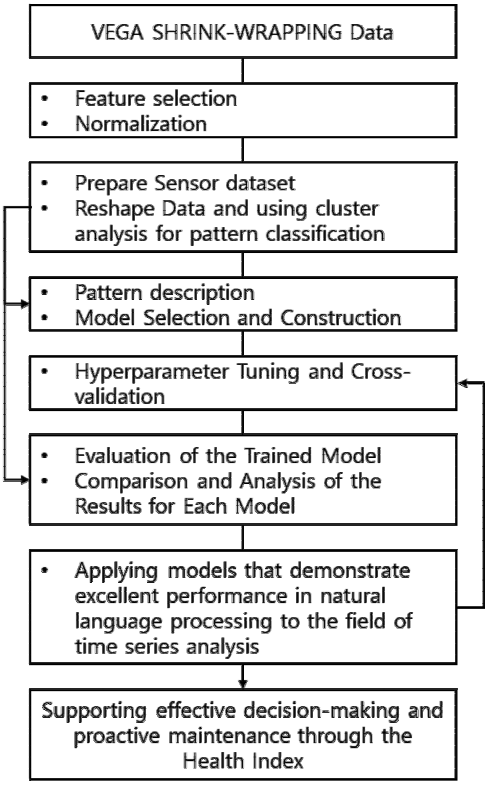
기계의 HI는 기계의 상태와 성능을 평가하고 모니터링하기 위해 사용되는 지표이다. 기계의 작동 상태, 신뢰성, 안전성 등을 종합적으로 평가하여 기계의 건강 상태를 파악하고 개선할 수 있는 정보를 제공한다. 일반적으로 센서 데이터, 작동 기록, 오류 등 다양한 정보를 수집하여 분석한다.

이를 통해 기계의 작동 상태, 성능 변화, 부품의 고장 가능성 등을 예측 및 확인하는 것이 가능하다. HI를 활용하여 예방 정비, 고장 진단, 유지보수 계획 등을 수립하는데 도움을 줄 수 있으며, 기계의 안정성과 생산성을 향상시킬 수 있다[3].

4. 패턴 기반 HI 수립 프로세스 개발

4.1 패턴 기반 HI 수립 프로세스 제안

본 연구에서는 시계열 데이터의 특징에 따라 분류된 패턴을 기반으로 HI를 수립하기 위해 다음 <Figure 8>과 같은 연구 프로세스를 제안한다.



<Figure 8> The Proposed Process for Establishing HI

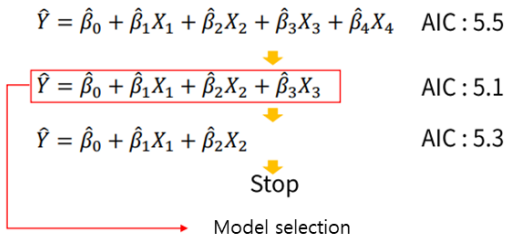
연구 프로세스는 데이터 전처리, 패턴 분류, HI 수립 세 단계로 구성된다. 이를 위해 Stepwise, Min-Max Normalization, 군집 분석과 같은 다양한 분석방법을 사용하여 시계열 데이터를 정확하게 분석하고, 기계의 현재 상태를 진단하는 HI를 수립하고자 한다.

4.2 변수 선택

Stepwise 변수 선택법은 종속 변수(y)가 없는 분석에서

변수 간 상호 관계나 변수의 중요성을 평가하는 데에 활용될 수 있으며, 군집화(clustering)에서 변수 간 유사성을 측정하고 가장 유의미한 변수를 선택하는 데에도 적용될 수 있다. 변수의 중요도를 판단하기 위해 통계적 기준을 사용하여 변수를 추가하거나 제거하는 방식으로 수행할 수 있다. 따라서, Stepwise 변수 선택법을 활용한 변수 선택은 가능하며, 선택된 변수로 분석이 가능하게 된다. Stepwise를 통해 변수 선택을 수행하고, 이에 따라 선택된 변수로 시계열 패턴 분석을 수행하고자 한다.

제거된 변수를 제외한 나머지 변수로 모델을 재적합하고, 제거된 변수의 유의성을 평가한다. 유의성이 충분히 낮으면 해당 변수를 영구적으로 제거하는 과정을 반복하면서 변수의 수를 줄여 모델의 해석력을 향상시켰다. <Figure 9>는 Stepwise 후진 제거법의 예시로, 변수 X_4를 제거하면 성능이 향상되고 변수 X_3을 제거하면 성능이 저하되므로 변수 X_3을 유지하는 모델을 선택하였다.



<Figure 9> Example of Stepwise Backward Elimination Method

4.3 군집 분석 기반 패턴 분류

군집 분석(Cluster Analysis)은 비슷한 특성을 가진 개체들을 그룹으로 나누는 기법으로, 데이터 내의 유사성이 높은 개체들을 동일한 군집으로 묶을 수 있다. 이는 비지도 학습(Unsupervised Learning)의 한 종류로, 사전에 레이블이 없는 데이터에서 패턴을 찾아내는 데 사용된다. 본 연구에서 사용하는 데이터는 레이블이 없는 데이터로, 군집 분석을 통해 패턴을 분류하여 모델별 분류 성능을 평가하고자 하였다.

모델의 분류 성능 평가에는 F1-score의 정밀도(Precision)와 재현율(Recall)을 활용하였다. F1-score는 주로 이진 분류(Binary Classification)에서 사용되며, 클래스의 불균형이 있거나 정확도(accuracy)만으로 평가하기에 부족한 경우 유용하게 활용된다. 정밀도는 모델이 참으로 예측한 샘플 중 실제로 참인 비율을 나타내며, 식 (10)과 같이 계산된다. 여기서 TP는 True Positive로 실제 값과 예측값이 모두 참인 샘플의 수이고, FP는 False Positive로 실제 값은 거짓이지만 예측값은 참인 샘플의 수이다.

$$\text{정밀도} = \frac{TP}{TP+FP} \tag{10}$$

재현율은 실제 참인 샘플 중 모델이 참으로 예측한 비율을 나타내며, 식 (11)과 같이 계산된다. 여기서 FN은 False Negative로 실제 값은 참이지만 예측값은 거짓인 샘플의 수이다.

$$\text{재현율} = \frac{TP}{TP+FN} \tag{11}$$

패턴 분류 후, 가장 성능이 우수한 모델을 활용하여 잔여수명 범위의 구간 예측을 수행하였다. 잔여수명 범위의 구간 예측은 기계의 수명이 어느 범위 내에서 얼마나 남았는지를 예측하는 작업을 의미한다. 기계의 수명은 여러 요인에 의해 영향받으며, 정확하게 언제 고장이 발생할지 예측하기는 어렵다. 그러나 기계의 상태를 모니터링하고 수집한 데이터를 기반으로 잔여수명 범위를 구간 추정하는 것은 가능하다.

4.4 현재 상태 기반 Health Index 개발

본 연구의 주요 목표는 데이터셋의 각 사이클을 기반으로 기계 학습을 수행하여 각 패턴으로 분류한 후, 잔여수명 범위의 구간 예측을 통해 절단 블레이드의 상태를 추정하는 HI를 수립하는 것이다. Jahromi et al.[8]과 Naderian et al.[17]은 HI를 계산하기 위한 기본 방법을 제시하였다. 이 방법에서는 모든 입력 데이터 특성에 대해 사전에 정의된 척도에 따라 점수가 할당되며, 점수화된 데이터 변수는 사전에 정의된 가중치 요인으로 곱해진다. 이 방법의 수식을 참고하여 본 연구에 맞는 HI를 식 (12)와 같이 수립하였다.

$$HI = \frac{(\alpha_1 \times \beta_1) + (\alpha_2 \times \beta_2) + (\alpha_3 \times \beta_3)}{(\alpha_1 \times \gamma_1) + (\alpha_2 \times \gamma_2) + (\alpha_3 \times \gamma_3)} - \beta_4 \tag{12}$$

α_i: i 패턴의 가중치 값(상수)
β_i: 현재 데이터 내 i 패턴의 비율(변수)
γ_i: 현재 데이터 내 i 패턴의 존재 여부

식 (12)에서 표시된 각 패턴의 가중치 값(α)은 잔여수명 범위의 구간 예측에서 가장 성능이 우수한 모델의 패턴별 재현율과 잔여수명 예측 범위를 기반으로 도출된 척도 값을 곱한 값이다. 현재 시점에서의 사이클 경과에 따라 각 패턴의 비율과 존재 여부를 고려하여 현재 절단 블레이드의 상태를 판단할 수 있는 HI 점수를 얻을 수 있으며, 이를 통해 관리자는 적절한 조치를 취할 수 있다. 점수의 감소 추세에 따라 장비의 검사 및 확인을 진행하거나, 장비의

교체 여부 등을 파악할 수 있다.

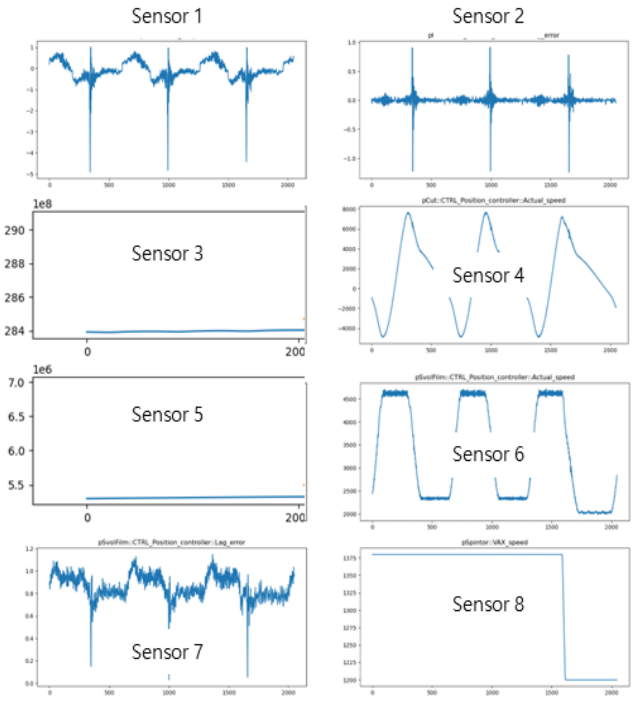
5. 분석 및 평가

5.1 데이터 전처리

본 연구에서 사용된 데이터는 열 수축 필름 포장기의 절단 블레이드의 센서 데이터이다. 총 8개의 데이터셋이 있으며, 각 데이터셋은 서로 다른 포장기에 해당되어 모드로 구별한다. 각 블레이드 데이터에는 약 1년 동안의 마모 정보가 포함되어 있다. 본 연구에서는 가장 많은 데이터를 포함하고 있으며, 두 번째로 긴 측정 기간을 가진 mode 1 데이터셋을 선택하여 분석에 활용하였다.

<Table 1> Dataset Information by Mode

Dataset No.	Samples	Cycles	Period (days)
Mode 1	1,349	100	356
Mode 2	121	100	357
Mode 3	70	100	210
Mode 4	12	100	13
Mode 5	60	100	342
Mode 6	18	100	134
Mode 7	3	100	1
Mode 8	6	100	1



<Figure 10> Trend Graph by Sensor

해당 데이터셋은 8개의 센서 측정 변수가 초 단위로 구성되어 있다. 이 8개의 센서 측정 변수는 블레이드와 필름 공급 기계의 위치, 속도, 지연 오차 등 각 구성 요소의 정보를 파악하기 위해 수집되었다. 일부 센서 정보는 블레이드의 직접적인 추세를 나타내지만, 다른 센서는 성능 저하에 대한 정보를 거의 포함하지 않는다. <Table 2>는 해당 변수들의 물리적인 의미를 나타내며, <Figure 10>은 초기 데이터를 시각화한 그래프이다.

<Table 2> Dataset sensor list

No.	Variable name	Variable meaning
1	pCut: Motor Torque	Blade motor torque
2	pCut: Lag error	Delay error of the blade position controller
3	pCut: Actual position	The actual position of the blade position controller
4	pCut: Actual speed	The actual velocity of the blade position controller
5	pSvolFilm: Actual position	The actual position of the film supply machine
6	pSvolFilm: Actual speed	The actual speed of the film supply machine
7	pSvolFilm: Lag error	The delay error of the film supply machine
8	pSpintor: VAX speed	The rotational speed of the turntable

변동성이 없는 변수를 제거하기 위해 Stepwise 후진 제거법을 사용하였다. 총 3단계로 진행되었으며, 각 단계에서 P-값이 가장 큰 변수를 차례로 제거하는 방식을 채택하여 Sensor 3, 5, 8을 제거하였다. 4단계에서 Sensor 6을 제거하면 모델의 성능이 저하되는 것을 확인하였으므로 3단계에서 Stepwise 후진 제거법을 중지하였다. 제거된 변수를 제외한 5개의 센서를 입력 데이터로 사용하여 분석에 활용하였고, 이 변수들을 결합하여 절댓값 변환한 후 이를 기반으로 사이클을 생성하였다.

<Table 3> Stepwise Backward Elimination (Step 1)

Sensor No.	Coefficient	P-value
1	0.0009	0.001
2	-0.0000	0.188
3	0.0000	0.986
4	0.0003	0.203
5	-24	0.898
6	-0.724	0.420
7	-0.0011	0.193
8	1.81	0.424

<Table 4> Stepwise Backward Elimination (Step 2)

Sensor No.	Coefficient	P-value
1	0.0009	0.001
2	-0.0000	0.153
3	-	-
4	0.0003	0.197
5	-24	0.896
6	-0.726	0.410
7	-0.0011	0.180
8	1.81	0.419

<Table 5> Stepwise Backward Elimination (Step 3)

Sensor No.	Coefficient	P-value
1	0.0009	0.001
2	-0.0000	0.151
3	-	-
4	0.0003	0.183
5	-	-
6	-0.737	0.398
7	-0.0011	0.177
8	1.79	0.421

입력 데이터의 정규화를 위해 Min-Max 정규화를 사용하여 데이터의 범위를 [0, 1]로 조정하였다. 정규화는 식 (13)과 같이 수행된다.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{13}$$

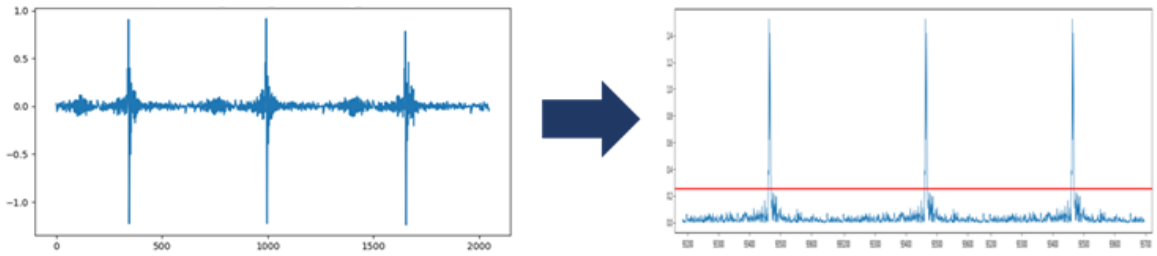
정규화된 5개의 센서 데이터를 결합하여 하나의 변수로 생성하였고, 데이터셋을 구성하기 위해 512개의 행을 한 개의 사이클로 설정하였다. 총 100개의 사이클이 데이터에 포함되어 있다. 이상 분석을 통해 임계값 0.25를 도출하였으며, 이를 사용하여 사이클 내 이상값을 판별하였다. <Figure 12>와 같이 0.25 이상의 값이 포함되어 있지 않은 사이클은 패턴 0으로 설정하였으며, 이는 패턴 분류나 잔여수명 범위 예측에서 제외되었다.

5.2 분석 결과

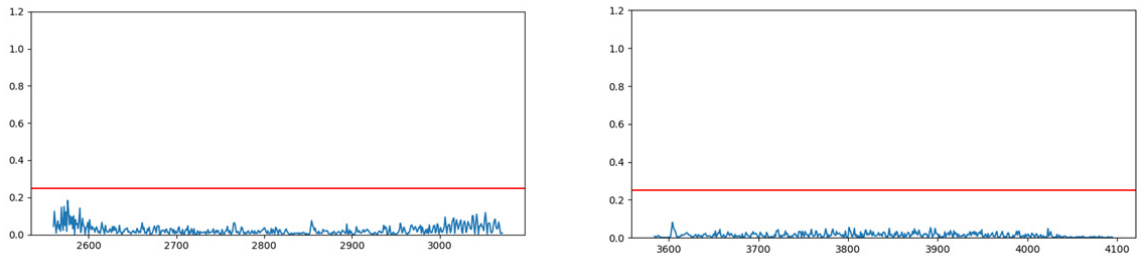
데이터셋에 포함된 사이클을 활용하여 군집 분석을 수행하였다. x축과 y축은 각각 사이클 내에서 이상값이 발생하는 구간의 크기(range)와 이상값이 발생하는 구간의 크기 대 개수의 비율(ratio)을 나타낸다. 비율을 A로, 구간의 크기를 B로 설정했을 때, <Figure 13>과 같이 패턴 1, 2, 3, 4를 설정할 수 있다. 이 중 패턴 4는 고장을 나타낸다.

$$A = \frac{\text{이상값의 개수}}{\text{이상값이 발생하는 구간의 크기}(B)} \tag{14}$$

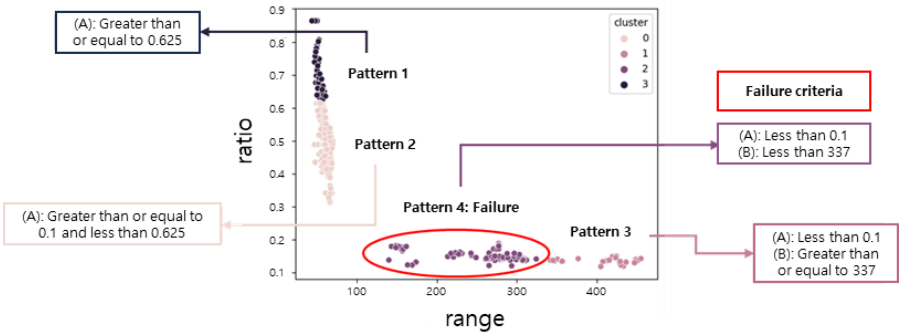
패턴 1은 A 값이 0.625 이상인 경우로, 이는 이상값이



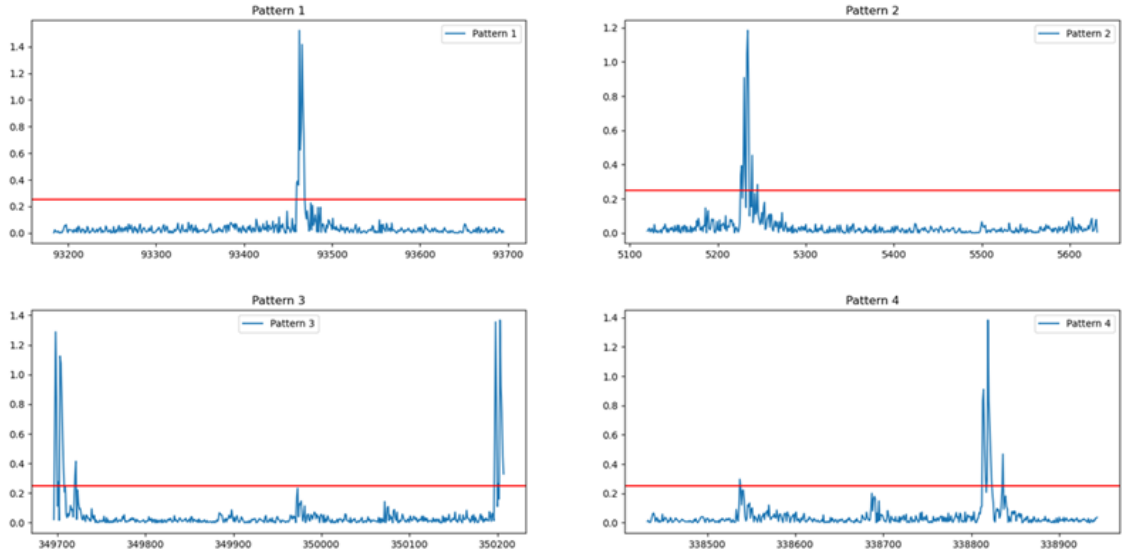
<Figure 11> The Combined Graph of the Original Data (left) and the Graph after Applying Absolute Value Transformation (right)



<Figure 12> The graph of Pattern 0



<Figure 13> Pattern Classification Based on Cluster Analysis



<Figure 14> Graphs by Pattern

발생하는 구간의 크기와 개수의 차이가 크지 않으므로 해당 구간에서만 이상값의 분포가 집중되어 있다는 것을 나타낸다. 패턴 2는 A 값이 0.1 이상 0.625 미만인 경우로, 패턴 1보다 더 큰 구간에서 이상값이 발생한다는 것을 의미하므로 이상값의 분포가 더 넓게 퍼져있다는 것을 나타낸다. 패턴 3은 A 값이 0.1 미만이고 B 값이 337 이상인 경우로, B 값이 기준값을 초과하므로 이상값이 나타나는 분포가 넓게 퍼져있다기보다는 블레이드의 작동 주기가 짧아져 서로 다른 이상값 분포 간의 간격이 좁아졌으므로 한 사이클 내에서 두 개의 이상값 분포를 확인할 수 있다고 판단하는 것이 더 적절하다. 패턴 4는 A 값이 0.1 미만이고 B 값이 337 미만인 경우로, 패턴 3에서 나타나는 두 개의 이상값 분포 간의 간격이 점차 좁아져 이상값이 나타나는 주기가 짧아지는 형태이므로 이를 블레이드의 고장으로 판단하였다.

하나의 데이터셋에는 100개의 사이클이 포함되어 있으며, 패턴 1, 2, 3, 4가 차례대로 나타나고, 패턴 0은 무작위로 나타난다. 각 데이터셋에 패턴 1은 약 40 사이클, 패턴 2는 약 45 사이클, 패턴 3은 약 5 사이클, 패턴 4는 약 10 사이클, 패턴 0은 약 5 사이클 미만이 포함되어 있다.

<Table 6> Pattern classification performance by model

Model	Pattern			
	1	2	3	4
RNN	0.7442	0.8122	0.8648	0.9016
LSTM	0.7373	0.7726	0.8441	0.8624
GPT-2	0.8345	0.8472	0.8718	0.9275
ProphetNet	0.8462	0.9118	0.9348	0.9957

패턴을 분류할 때 사용되는 RNN, LSTM, GPT-2, ProphetNet 모델의 성능을 비교하기 위해 F1-score의 정밀도를 사용하였다. 클래스 분포가 불균형하므로 정확도 대신 정밀도를 사용하였으며, 5-fold 교차 검증 방법을 채택하여 활용하였다. 데이터셋의 사이클을 5등분으로 분류하

여 5번의 실험을 통해 각 정밀도 값을 계산하고, 이들의 평균을 산출하여 모델의 성능을 비교하였다.

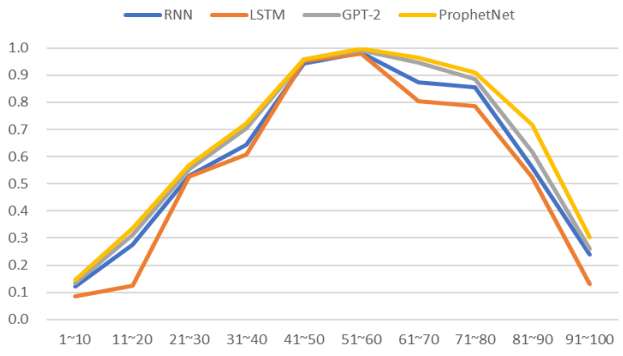
정밀도를 사용하여 모델들의 패턴 분류 성능을 비교한 결과, ProphetNet 모델이 가장 우수한 성능을 보였다. 패턴 1과 2는 데이터 분포가 매우 유사하여 패턴 3과 4보다 비교적 구분이 어렵다. 패턴 1부터 4까지를 가장 잘 분류하는 모델로써 ProphetNet 모델을 선택하는 것이 가장 적절하다고 판단되었다.

F1-score의 재현율을 활용하여 잔여수명 범위를 구간 예측하였다. 재현율은 실제로 참인 샘플 중에서 모델이 정확히 참으로 예측한 샘플의 비율을 나타낸다. 패턴별 잔여수명 범위의 구간 예측에서 재현율을 사용하는 이유는 제품이 실제로 고장 났을 때 모델이 정확히 참으로 예측하는 것이 중요하기 때문이다. 따라서, 재현율이 높은 모델을 선택하고자 하였다.

패턴별 잔여수명 범위의 구간 예측 재현율을 확인하여 표와 그래프로 정리하였다. 패턴 1의 잔여수명 범위별 예측 재현율이 가장 높은 구간은 51~60 사이클 구간이다. 또한, 성능이 가장 좋은 모델은 ProphetNet 모델이다.

<Table 7> Prediction recall of residual life range for Pattern 1

Cycles	RNN	LSTM	GPT-2	ProphetNet
1~10	0.1203	0.0851	0.1332	0.1468
11~20	0.2757	0.1261	0.3119	0.3351
21~30	0.5285	0.5264	0.5537	0.5673
31~40	0.6449	0.6087	0.7044	0.7219
41~50	0.9427	0.9514	0.9581	0.9587
51~60	0.9828	0.9799	0.9921	0.9969
61~70	0.8739	0.8049	0.9447	0.9637
71~80	0.8542	0.7847	0.8867	0.9093
81~90	0.5588	0.5229	0.6177	0.7178
91~100	0.2405	0.1293	0.2604	0.3026

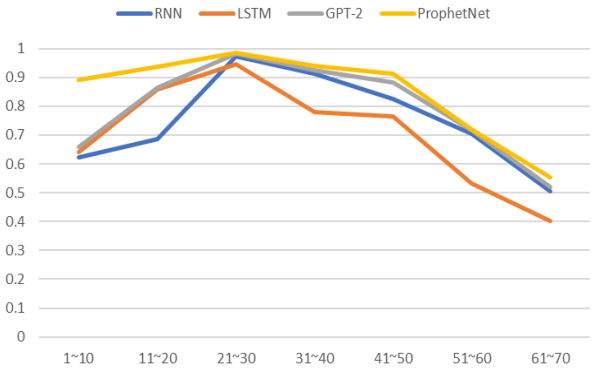


<Figure 15> Graph of prediction recall for residual life range of Pattern 1

데이터 내에서 패턴 2가 나타나기 시작할 때 잔여수명 범위 예측의 최대 사이클은 약 70 사이클을 넘지 않는다. 따라서, 잔여수명 범위가 70 사이클 이상으로 예측될 가능성은 전혀 없다. 예측 재현율이 가장 높은 구간은 21~30 사이클 구간이며, 성능이 좋은 모델 순서는 ProphetNet, GPT-2, RNN, LSTM이다.

<Table 8> Prediction Recall of Residual Life Range for Pattern 2

Cycles	RNN	LSTM	GPT-2	ProphetNet
1~10	0.6234	0.6409	0.6576	0.8916
11~20	0.6848	0.8571	0.8651	0.9366
21~30	0.9741	0.9457	0.9827	0.9853
31~40	0.9133	0.7785	0.9253	0.9395
41~50	0.8245	0.7657	0.8838	0.9123
51~60	0.7031	0.5315	0.7177	0.7203
61~70	0.5041	0.4025	0.5192	0.5537

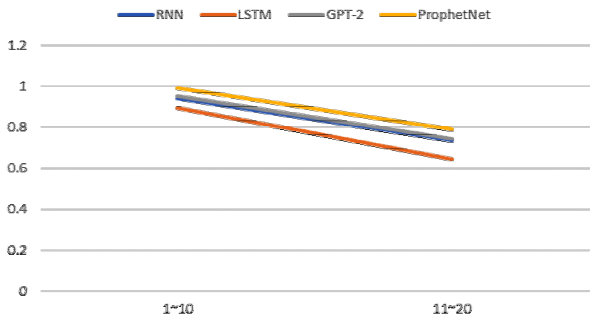


<Figure 16> Graph of Prediction Recall for Residual Life Range of Pattern 2

데이터에서 패턴 3이 나타날 때 잔여수명 범위 예측의 최대 사이클은 약 20 사이클을 넘지 않는다. 따라서, 잔여수명 범위가 20 사이클 이상으로 예측될 가능성은 전혀 없다. 예측 재현율이 가장 높은 구간은 1~10 사이클 구간이며, 성능이 좋은 모델 순서는 ProphetNet, GPT-2, RNN, LSTM이다. 패턴 3에서는 고장까지의 잔여수명 사이클이 비교적 짧으므로 이에 따른 적절한 조치를 빠르게 취해야 할 필요가 있다고 판단하였다.

<Table 9> Prediction Recall of Residual Life Range for Pattern 3

Cycles	RNN	LSTM	GPT-2	ProphetNet
1~10	0.9402	0.8958	0.9517	0.9914
11~20	0.7328	0.6436	0.7414	0.7899



<Figure 17> Graph of Prediction Recall for Residual Life Range of Pattern 3

패턴별 잔여수명 범위의 예측 재현율에 따라 절단 블레이드의 상태를 파악할 수 있는 HI를 수립하고자 하였다. HI를 활용하여 절단 블레이드의 교체 시기나 교체 전 취해야 할 조치 사항을 결정할 수 있다. 4가지의 모델 중 가장 성능이 우수한 ProphetNet 모델의 패턴별 재현율과 잔여수명 범위의 구간 예측을 통해 얻은 척도 값을 곱하여 패턴별 가중치(α)를 도출하였다. 척도 값을 곱하는 이유는 잔여수명이 많이 남아있을수록 HI 점수를 높게 주기 위함이다.

<Table 10> The weights for each pattern derived by combining the scale and recall of the ProphetNet model(α)

Scale	Cycles	Pattern		
		1	2	3
1	1~10	0.1468	0.8916	0.9914
2	11~20	0.3351	0.9366	0.7899
3	21~30	0.5673	0.9853	-
4	31~40	0.7219	0.9395	
5	41~50	0.9587	0.9123	
6	51~60	0.9969	0.7203	
7	61~70	0.9637	0.5537	
8	71~80	0.9093	-	
9	81~90	0.7178		
10	91~100	0.3026		
α		39.6879	22.2379	2.5712

$$HI=\left\{\frac{\sum_{i=1}^{3}(\alpha_i\times\beta_i)}{\sum_{i=1}^{3}(\alpha_i\times\gamma_i)}-\beta_4\right\}\times 100 \qquad (15)$$

이를 활용하여 본 연구에서는 시간의 흐름에 따라 데이터에 포함된 패턴 1, 2, 3, 4의 비율에 따라 β 값이 계산된다. 네 가지 비율의 합은 1이며, 초반에는 데이터셋에 패턴 1만 존재하기에 β_1은 1이다. 사이클이 흐를수록 β_1의 값은 작아지고 β_2와 β_3의 값은 커지므로 HI의 값은 작아진다. 식 (15)에서 β_4 값만 따로 빼는 이유는 고장이 난 시점부터 β_4 값이 발생하므로 HI 점수를 급격히 감소시키기 위함이다.

$$\beta_i=\frac{k_i}{n} \qquad (16)$$

n: 현재 지난 사이클의 총 개수
k_i: 현재 지난 사이클 중 i 패턴의 개수

예를 들어, 데이터셋에 패턴 1이 38 사이클, 패턴 2가 45 사이클, 패턴 3이 6 사이클, 패턴 4가 11 사이클이 포함되어 있다고 가정한다. 38 사이클이 진행되는 동안은 n과 k_1 값이 같으므로 β_1은 1이다. 39 사이클이 되었을 때, β_1은 38/39이 되고, β_2는 1/39이 된다. 사이클의 경과에 따라 각 패턴의 비율을 <Table 11>에 표현하였다.

<Table 11> The Ratio of Each Pattern as the Cycle Progresses (β)

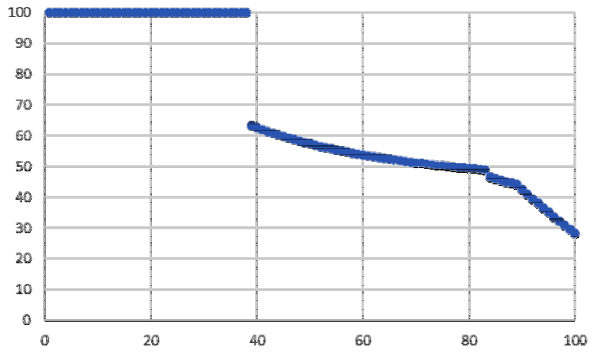
Cycles	β_1	β_2	β_3	β_4
1	1/1	-	-	-
2	2/2			
3	3/3			
⋮	⋮			
38	38/38			
39	38/39	1/39		
40	38/40	2/40		
41	38/41	3/41		
⋮	⋮	⋮		
83	38/83	45/83		
84	38/84	45/84	1/84	
85	38/85	45/85	2/85	
86	38/86	45/86	3/86	
⋮	⋮	⋮	⋮	
89	38/89	45/89	6/89	
90	38/90	45/90	6/90	1/90
91	38/91	45/91	6/91	2/91
92	38/92	45/92	6/92	3/92
⋮	⋮	⋮	⋮	⋮
100	38/100	45/100	6/100	11/100

식 (15)를 활용하여 사이클 경과에 따른 HI 점수를 아래 <Table 12>와 같이 작성하였고, <Figure 18>과 같이 HI 점수 그래프를 시각화하였다. 패턴 1만 존재했을 때는 HI 점수가 계속 100점을 유지한다. 그러다가 패턴 2가 나타나면서 점수가 많이 감소하고, 패턴 3이 나타나고 패턴 4가 발

생하기 전까지 천천히 점수가 감소한다. 절단 블레이드가 고장이 났다고 판단되는 패턴 4가 발생할 때부터 급격하게 HI 점수가 감소하는 것을 확인할 수 있다. 현재 데이터셋에는 100 사이클까지밖에 없으므로 마지막 순간에 HI 점수가 28.1378점이지만, 더 많은 사이클이 생긴다면 0점에 가까운 점수가 나타날 것으로 판단된다. 따라서, 점수가 많이 감소할 때에는 관리자가 블레이드의 현재 상태를 확인하며 검사할 필요가 있고, 점수가 40점 이하로 떨어지면서 급격히 감소할 때에는 블레이드를 교체해야 할 것으로 판단하였다.

<Table 12> HI Score as the Cycle Progresses

Cycles	HI score
1	100
2	100
3	100
⋮	⋮
38	100
39	63.3669
40	62.6805
41	62.0276
⋮	⋮
83	48.8117
84	46.3554
85	45.8569
86	45.3700
⋮	⋮
89	43.9751
90	42.3754
91	40.8108
92	39.2803
⋮	⋮
100	28.1378



<Figure 18> HI Score Graph as the Cycle Progresses

6. 결론 및 추후 연구

시계열 예측 분야와 자연어 처리 분야에서 기존의 RNN과 LSTM 모델이 널리 사용되며 우수한 성능을 보여왔다. 현재 자연어 처리 분야에서는 트랜스포머 모델이 뛰어난 성능을 보인다. 그러나 시계열 예측에 트랜스포머 모델을 적용한 연구는 아직 많이 이루어지지 않았으며, 트랜스포머 모델은 빅데이터 분석 분야에서의 성능이 크게 강조되지 않았다. 따라서, 트랜스포머 모델을 기반으로 한 시계열 예측 연구는 계속해서 진행되고 있으며, 발전된 트랜스포머 모델인 GPT-2와 ProphetNet 모델이 시계열 예측 분야에서 주목을 받는 추세이다.

본 연구에서는 발전된 트랜스포머 모델을 활용하여 시계열 데이터를 패턴별로 분류하는 분류 성능과 잔여수명 범위의 구간 예측 성능을 확인하고자 하였다. 절단 블레이드의 HI를 수립하는 데 성능이 가장 좋았던 ProphetNet 모델의 재현율을 활용하였으며, 이를 통해 실시간으로 절단 블레이드의 건강 상태를 확인할 수 있는 프로세스를 제안하였다. 제안한 프로세스가 시계열 데이터의 열화 패턴을 분석하는 데 있어 유용하다는 것을 확인할 수 있었으며, 이를 활용하여 절단 블레이드의 건강 상태를 예측하는 HI를 구성할 수 있었다. 이러한 프로세스는 기계의 시계열 데이터를 수집한 후, 필요한 데이터의 전처리 과정을 수행하고 패턴을 분류하여 기계의 상태를 실시간으로 모니터링하고 관리함으로써 유지보수를 포함한 기계의 관리를 위한 의사결정을 내리는 데 많은 산업 분야에서 도움이 될 것으로 기대된다.

본 연구는 HI를 활용하여 실제 건강 상태를 판단하는 것은 가능하지만, 잔여수명을 정확하게 예측하는 데 있어 어려움이 있다. 따라서, 추후 연구에서는 잔여수명 범위를 구간 예측하는 것이 아닌 잔여수명 자체를 정확히 예측할 수 있는 모델을 구성하기 위해 추가적인 특징을 고려하거나 앙상블 기법 등의 활용 가능성을 확인할 필요가 있다. 이를 기반으로 맞춤형 시스템 관리를 수행할 수 있는 실용적인 연구가 가능할 것으로 사료된다.

Acknowledgement

This work was supported by the GRRC program of Gyeonggi province. [GRRC KGU 2023-B01, Research on Intelligent Industrial Data Analytics]

References

[1] Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Umar, A.M., Linus, O.U., Arshad, H., Kazaure, A.A.,

Gana, U., and Kiru, M.U., Comprehensive review of artificial neural network applications to pattern recognition, *IEEE Access*, 2019, Vol. 7, pp. 158820-158846.

[2] Ahn, H.J., Kim, H.B., Jung, D.W., Kim, D.J., and Lee, D.K., Development of Artificial Intelligence-Based Anomaly Detection Method for Aerospace Systems, Korea Aerospace Research Institute, 2019.

[3] Alqudsi, A. and El-Hag, A., Application of machine learning in transformer health index prediction, *Energies*, 2019, Vol. 12, No. 14, p. 2694.

[4] Bohatyrewicz, P. and Mrozik, A., The analysis of power transformer population working in different operating conditions with the use of health index, *Energies*, 2021, Vol. 14, No. 16, p. 5213.

[5] Chandar, S., Sankar, C., Vorontsov, E., Kahou, S.E., and Bengio, Y., Towards non-saturating recurrent units for modelling long-term dependencies, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, Vol. 33, No. 01, pp. 3280-3287.

[6] Chen, G., A gentle tutorial of recurrent neural network with error backpropagation, arXiv preprint arXiv, 2016, 1610.02583.

[7] Fang, W., Chen, Y., and Xue, Q., Survey on research of RNN-based spatio-temporal sequence prediction algorithms, *Journal on Big Data*, 2021, Vol. 3, No. 3, p. 97.

[8] Jahromi, A., Piercy, R., Cress, S., Service, J., and Fan, W., An approach to power transformer asset management using health index, *IEEE Electrical Insulation Magazine*, 2009, Vol. 25, No. 2, pp. 20-34.

[9] Jana, R.K., Ghosh, I., and Wallin, M.W., Taming energy and electronic waste generation in bitcoin mining: Insights from Facebook prophet and deep neural network, *Technological Forecasting and Social Change*, 2022, Vol. 178, p. 121584.

[10] Jang, S.M., Moon, J-H., and Sohn, K-A., Comparison of Transformer and LSTM for threat detection and traffic prediction on long time-series data, *Journal of Korean Institute of Next Generation Computing*, 2021, pp. 18-21.

[11] Jin, Y-H., Ji, S-H., and Han, K-H., Time Series Data Analysis and Prediction System Using PCA, *Journal of the Korea Convergence Society*, 2020, Vol. 12, No. 11, pp. 99-107.

[12] Jung, I.K., Park, D.K., and Jun, D.B., Performance of Pairs Trading Algorithm with the Implementation of Structural Changes Detection Procedure, *Journal of The Korean Operations Research and Management Science Society*, 2017, Vol. 42, No. 3, pp. 13-24.

[13] Kim, Y.S. and Park, K.S., The Multivariate Sensor Data Classification using Time Series Imaging, *Journal of Korean Institute of Information Scientists and Engineers*, 2022, Vol. 49, No. 8, pp. 593-600.

[14] Lim, J., Kim, I.K., Lee, M.H., Ha, J.M., and Lee, J.K., Performance comparison of network anomaly detection using BERT, LSTM and GRU, *Journal of Korean Institute of Communications and Information Sciences*, 2022, pp. 1268-1269.

[15] Mizuno, T., Fujimoto, S., and Ishikawa, A., Generation of individual daily trajectories by GPT-2, *Frontiers in Physics*, 2022, p. 1118.

[16] Murugan, R. and Ramasamy, R., Understanding the power transformer component failures for health index-based maintenance planning in electric utilities, *Engineering Failure Analysis*, 2019, Vol. 96, pp. 274-288.

[17] Naderian, A., Cress, S., Piercy, R., Wang, F., and Service, J., An approach to determine the health index of power transformers, *Conference Record of the 2008 IEEE International Symposium on Electrical Insulation*, 2008, pp. 192-196.

[18] Oh, H-W. and Kim, W-S., A Study on the Small Motion Classification Model based on Time Serial Data, *Journal of Korean Institute of Communications and Information Sciences*, 2022, pp. 949-950.

[19] Qi, W., Yan, Y., Gonkg, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M., Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training, arXiv preprint arXiv, 2020, 2001.04063.

[20] Rediansyah, D., Prasojo, R.A., and Abu-Siada, A., Artificial intelligence-based power transformer health index for handling data uncertainty, *IEEE Access*, 2021, Vol. 9, pp. 150637-150648.

[21] Sandoval, R.M., Garcia-Sanchez, A.J., Garcia-Haro, J., and Chen, T.M., Optimal policy derivation for transmission duty-cycle constrained LPWAN, *IEEE Internet of Things Journal*, 2018, Vol. 5, No. 4, pp. 3114-3125.

[22] Saqlain, M., Jargalsaikhan, B., and Lee, J.Y., A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing*, 2019, Vol. 32, No. 2, pp. 171-182.

[23] Sherstinsky, A., Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Physica D: Nonlinear Phenomena*, 2020, Vol. 404, p. 132306.

[24] Tan, C., Deep Reinforcement Learning with Copy-oriented Context Awareness and Weighted Rewards for Abstractive Summarizatio, *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*, 2023, pp. 84-89.

[25] Xiao, J. and Zhou, Z., Research progress of RNN language model, *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 1285-1288.

[26] Yi, D-H., Yu, Y.S., Ko, Y.D., Jo, H.K., and Park, C.S., Classification of Building Energy Pattern based on Robust Regression and Time series clustering, *Journal of the Architectural Institute of Korea*, 2020, Vol. 40, No. 2, pp. 392-393.

[27] Yoo, I-J. and Park, D-H., Derivation of Digital Music's Ranking Change Through Time Series Clustering, *Journal of Intelligence and Information Systems*, 2020, Vol. 26, No. 3, pp. 171-191.

[28] Yook, H.N., Kim, Y-J., Choi, Y.S., Oh, J.W., Lee, K., and Ha, S.L., Comprehensive Studies on Reliability of Cutting Blade According to the Material of Knife Mill Used in Biomass Pretreatment Process, *Transactions of the Korean Society of Mechanical Engineers*, 2022, p. 171.

[29] Zhang, M. and Li, J., A commentary of GPT-3 in MIT Technology Review 2021, *Fundamental Research*, 2021, Vol. 1, No. 6, pp. 831-833.

ORCID

Sun-Ju Won | https://orcid.org/0009-0002-2688-2016
Yong Soo Kim | https://orcid.org/0000-0003-3362-4496