

# 인공지능 모델 평가 방식에 따른 도금공정 운영 최적화

이성주<sup>1\*</sup>, 이강호<sup>2</sup>, 이수연<sup>1</sup>, 신요섭<sup>1</sup>  
<sup>1</sup>한동대학교 기계제어공학부, <sup>2</sup>한동대학교 ICT융합전공

## Electroplating Process Optimization through AI Model Evaluation Methods

Sung-Joo Lee<sup>1\*</sup>, Gang-Ho Lee<sup>2</sup>, Soo-Yeon Lee<sup>1</sup>, Yo-Sup Shin<sup>1</sup>  
<sup>1</sup>Division of Mechanical & Control Engineering, Handong University  
<sup>2</sup>Division of Global Entrepreneurship and ICT, Handong University

**요약** 이노징크(Innozinc) 전기 도금 처리 과정 중, 산제 전처리 과정을 통해 공정들의 밀착성을 향상하는 효과를 기대할 수 있으나, 용액 농도 및 수소취성에 따른 품질 저하의 문제점이 발생한다. 문제를 해결하기 위한 기존 연구에서는 결정 트리 알고리즘을 통해 공정 운영 최적화 모델을 생성하였으나, 해당 모델의 낮은 정확도로 학습이 올바르게 진행되지 않았다고 평가하였다. 본 연구에서는 기존 연구에서 사용된 분류 목적 모델이 아닌, 도금의 정량적 예측을 위한 회귀 모델의 사용을 제안하였다. 또한, 데이터 전처리 과정을 통해 시계열 데이터로 재구성하여 RNN과 LSTM을 통해 처리하였으며, 회귀 모델의 성능 측정 지표를 통해 연구의 타당성을 검증하였다. 이를 통해 프로세스가 중요한 도금 과정의 품질 예측에는 시계열을 무시한 모델 방법보다 시계열을 고려한 시퀀스 모델이 적절함을 입증하였다.

**Abstract** During the Innozinc electroplating process, an effect of improving the adhesion of the processes can be expected through a powder pretreatment process. However, there is a problem of quality deterioration due to solution concentration and hydrogen embrittlement. In previous studies to solve the problem, a process operation optimization model was created through the decision tree algorithm, but the learning did not proceed correctly due to the low accuracy of the model. In this study, we propose a method of using a regression model for prediction purposes instead of the classification objective model used in previous studies. In addition, the sequence was processed through RNN and LSTM by reconstructing data into time series data through preprocessing. The validity of the study was verified through the performance measurement index of the regression model. As a result, we proved that the sequence model that considers the time series is more appropriate than a regression model method that ignores the time series in predicting the quality of the plating process, where the sequence is important.

**Keywords** : Artificial Intelligence, LSTM, Machine Learning, Plating Treatment Process, RNN

본 논문은 한동대학교 AI 활용 프로그래밍 연구과제로 수행되었음.

\*Corresponding Author : Sung-Joo Lee(Handong Univ.)

email: sungjoo.lee153@gmail.com

Received October 13, 2022

Revised December 6, 2022

Accepted December 7, 2022

Published December 31, 2022

## 1. 서론

### 1.1 연구배경

산제 전처리리는 이노징크, 세라믹 아연도금 처리 시 전처리 공정의 첫 번째 단계로 제품 표면의 산도 및 피막을 제거하는 작업이다. 제품에 존재하는 산화층을 제거하여 표면의 활성화율을 유도하고, 공정들의 밀착성을 향상하는 전처리 방법[1]이다.

산제 전처리 과정 중 용액 농도 및 수소가스 침투에 의한 품질 저하가 발생하는 문제[2]가 존재한다. 따라서 KAMP(Korea AI Manufacturing Platform, 이하 KAMP) 제조 플랫폼에서는 도금 공정에서 발생하는 데이터에 대한 분석을 통해 산제 전처리 공정 운영 최적화 모델을 만드는 것을 제안[3]하였다.

KAMP는 중소벤처기업이 갖추기 어려운 데이터 저장/분석 인프라, 인공지능 개발/활용 도구, 인공지능 데이터셋, 전문가 컨설팅/교육 서비스 등을 모아 중소 제조업의 인공지능화를 지원하는 종합 플랫폼이다. KAMP는 이노징크에서의 품질 이상 탐지, 진단 문제를 해결하기 위해 ‘공정운영 최적화 AI 데이터셋’의 제목으로 산제 전처리 과정의 제조 AI 데이터를 분석하였다.

KAMP에서 진행되었던 기존 연구에서는 분류 모델인 결정 트리(Decision Tree) 알고리즘을 사용한 결과, 평균 제곱근 오차(RMSE: Root Mean Square Error, 이하 RMSE)=3.16의 오차율을 나타내는 성능을 확인하였다. 또한, 기존 연구에서 시계열 데이터의 특성을 고려하지 않은 모델을 적용하였다는 한계가 존재하였다. 본 연구에서는 예측을 목적으로 하는 모델을 선정하는데 분류 모델보다 회귀 모델, 그리고 시계열을 고려한 시퀀스 모델이 더욱 적절함을 제시한다.

## 2. 이론적 배경

### 2.1 이노징크

철의 부식 방지를 위한 아연도금은 철 대신 산화를 하여 철의 내구성과 수명을 연장하는 효과가 있다. 아연도금은 이온화 방법에 따른 전기 아연도금과 용융 방법에 따른 용융아연도금으로 분류할 수 있다. 이노징크는 ㈜켄프에서 개발한 세라믹 아연도금 공정으로, 기존 아연도금 방식보다 내식성이 뛰어나며, 강력한 물성의 장점이 있으며, 저렴하고 친환경 기술의 장점이 존재[4]한다.

산제 전처리 공정의 경우, 공정에서 사용하는 용액의 농도가 일정 기준 이상일 경우, 전체 도금 품질 저하가 발생하여 도금 불량으로 취급되는 경우가 발생할 수 있다. 또한, 산제 반응 과정에서 발생하는 수소가스가 제품의 금속 내부로 침투하면서 수소취성이 발생할 수 있다. 수소취성으로 인해 제품의 연성 및 인성이 저하되며, 소성변형 없이도 파괴되는 경향이 증대되는 문제점[3]이 있다.

이러한 문제점을 해결하기 위한 최적의 독립변수를 알아내는 것이 본 연구의 목적이며, 이를 수행하기 위해 최근 다양한 산업에서 활용되고 있는 인공지능을 통해 연구를 수행하고자 한다.

본 연구를 수행하기 위해 제공된 데이터셋은 Table 1과 같다. 데이터셋은 제품 생산 시 산제 전처리 설비 내에 설치된 PLC Data, HDH Data 센서를 통해 수집되었다[1]. 5초의 주기로 공정 산도 및 온도 측정값, 그리고 총 360초의 과정 이후의 정량 지표화된 처리율의 데이터로 구성되어 있다.

Table 1. Specifications of Innozinc dataset

Items	Specification
Data collection time(sec)	360
Data collection interval(sec)	5
Process acidity readings	50,094 data
Process temperature readings	50,094 data
Process rate	726 data

데이터의 개수, 평균 및 표준편차는 Table 2와 같다. 독립변수로 산도와 온도 데이터가, 종속변수로 처리율 데이터가 된다.

Table 2. Mean & Standard deviation of variables

	pH	Temperature	Process Rate
Count	50,094	50,094	50,094
Mean	2.01	49.9	96.1
Standard Deviation	0.552	1.35	3.21

기존 연구에서는 결정 트리 알고리즘[5]을 사용하여 종속변수를 예측하였다. 결정 트리 알고리즘은 머신러닝 알고리즘으로, 데이터의 규칙을 자동으로 찾아내어 분류, 예측, 회귀 등을 진행한다. 결정 트리는 데이터 전처리 작업이 거의 수행되지 않으며, 데이터 구조를 직관적으로 살펴볼 수 있는 큰 장점을 갖고 있다. 해당 알고리즘

을 바탕으로 생성된 모델은 RMSE=3.16의 오차를 보여 주었다. 그러나 해당 모델은 0.37의 낮은 정확도를 가지며, 모델이 정확한 값을 예측하지 않는다고 기존 연구에서 판단하였다.

본 연구는 기존 연구에서 성능 측정 지표를 정확도로 오용하였다고 판단하여 진행되었다. 측정 지표로서의 정확도는 분류 모델링을 목적으로 하는 모델에 대한 통계적 분석이다. 그러나 산재 전처리 데이터는 어떠한 기준에 따라 분류하는 것이 아닌, 독립변수에 따른 종속변수의 값을 예측하는 회귀 모델을 사용하였다.

Fig. 1의 무작위로 선택한 데이터셋의 시간별 산도 및 온도의 변화를 나타낸 그래프이다. 산도와 온도가 시간의 흐름에 대해 우상향 곡선을 그리는 일정한 패턴을 띄고 있으며, 시퀀스를 고려한 예측 모델을 통해 데이터를 분석하는 것이 적절하다고 예상하였다.

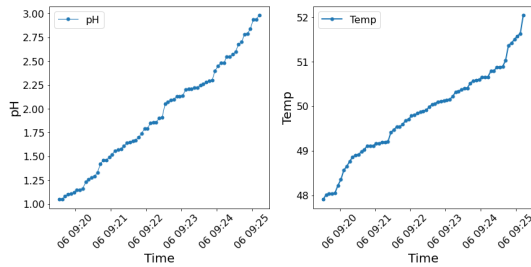


Fig. 1. Data distribution of independent variables over time

기존 연구에서와 동일하게 학습용 데이터와 검증용 데이터를 8:2의 비율로 나누어 진행했으며, 추가로 진행한 모델로는 선형 회귀(Linear Regression)[6], 랜덤 포레스트(Random Forest)[5]를 사용하였다. 또한, 시계열 데이터 분석을 위해 순환 신경망(RNN: Recurrent Neural Network, 이하 RNN)[7,8] 및 장단기 메모리(LSTM: Long Short Term Memory, 이하 LSTM)[7,8] 모델을 사용하였다.

## 2.2 회귀 모델 성능 평가 지표

MAE, RMSE, MAPE는 회귀 모델의 대표적인 성능 평가 지표[9]이다. 회귀 모델은 다수의 특징값을 입력으로 하나의 특징값을 산출하는 회귀 분석 시에 사용된다. 데이터셋에서 공정 운영 최적화의 정도를 예측하는 것이 목표이기에 회귀 모델을 사용하였으며, 다음 세 가지의 성능 평가 지표를 통해 모델의 성능을 평가하였다.

### 2.2.1 MAE

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

where,  $N$  denotes total count,  $y_i$  denotes actual value,  $\hat{y}_i$  denotes predicted value

MAE는 실제값과 예측값의 차이를 절댓값으로 변환해 평균화한 성능지수이다. MAE는 직관적으로 차이를 알 수 있으며, 오차값이 데이터 이상치의 영향을 상대적으로 적게 받는다는 장점이 있지만, 오차에 덜 민감하다는 단점이 존재한다.

### 2.2.2 RMSE

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

where,  $n$  denotes total count,  $y_i$  denotes actual value,  $\hat{y}_i$  denotes predicted value

RMSE의 수식적 전개는 위와 같으며, 지표 자체가 직관적이고 단순하다는 장점이 있다. 그러나 RMSE를 구하는 과정에서 오차를 제곱하므로 데이터 이상치 값으로 인해 수치가 급격하게 증가하며 이상치에 민감하다는 단점이 있다. 오차에 대해 제곱을 하므로 오차에 따라 가중치가 반영되며, 오차에 따른 손실이 기하급수적으로 증가하는 상황에서 쓰기에 적합하다.

### 2.2.3 MAPE

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{f}(x_i)}{y_i} \right| \quad (3)$$

where,  $n$  denotes total count,  $y_i$  denotes actual value,  $\hat{f}(x_i)$  denotes predicted value

MAPE는 MAE를 상대적 척도로 변환한 값이다. 따라서 MAE가 가지는 장단점을 동일하게 가진다. 덧붙여, 백분율의 상대적 확률값을 가지기 때문에 결과 해석에 유리하며, 데이터의 성능 비교에 쉬운 장점이 있다.

## 2.3 머신러닝 및 딥러닝 기반 회귀 분석

기존 연구에서는 결정 트리 모델을 통해 데이터 분석을 진행하였다. 다음은 본 연구에서 결정 트리보다 적은

오차율을 보여 더 좋을 것으로 예상하는 머신러닝(ML: Machine Learning, 이하 ML) 및 딥러닝(DL: Deep Learning, 이하 DL) 모델의 장단점이다.

### 2.3.1 선형 회귀(Linear Regression)

선형 회귀는 고려된 변수 간 관계를 정의하고 수량화하기 위해 데이터셋에 적용되는 통계 테스트이다. 선형 회귀는 종속 변수와 독립 변수 간의 연관성 정도를 분석하기에 적합한 방법이며, 이에 대한 수식적 전개는 다음과 같은 선형 방정식 형태로 표현할 수 있다.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n. \quad (4)$$

where,  $\hat{y}$  denotes predicted value,  $n$  denotes number of characteristics,  $x_i$  denotes  $i$ th characteristic value,  $\theta_j$  denotes  $j$ th model parameter

선형 회귀에서 가장 널리 사용되는 성능 측정 지표는 RMSE이다. RMSE를 최소화하는 모델 파라미터의 집합을 찾는 과정이 선형 회귀이며, 학습 및 예측 속도가 빠르다는 장점이 있지만, 데이터셋의 상관관계가 깊은 경우에는 모델 파라미터를 분석하기 어렵다는 단점을 가지고 있다.

### 2.3.2 랜덤 포레스트(Random Forest)

랜덤 포레스트 모델은 결정 트리 모델의 집합체이며, 단일 결정 트리 모델이 가지고 있는 한계를 보완한다. 트리의 노드를 분할하는 과정에서 전체 특성 중 최선의 특성을 찾는 방법 대신, 특성 후보를 무작위로 선택하며, 선택된 후보 중에서 최적의 특성을 찾는 방식을 통해 무작위성을 더한다. 이는 트리를 더욱 다양하게 만들며, 편향을 손해 본다는 단점이 있음에도 분산을 낮추어 전체적으로 더 훌륭한 모델을 생성한다.

## 2.4 순서(Sequence) 고려

기존의 방법들은 종속변수를 독립변수의 길이에 맞춰 보간하여 데이터셋을 구축하였다. 이 방법은 아연도금 처리가 시퀀스에 대해 무관한 공정이라는 것을 가정한 상황에서 머신러닝 및 딥러닝 모델들을 적용한 것이다. 본 연구에서는 데이터 전처리 작업을 통해 시계열 데이터를 형성하고, 순환 신경망 및 장단기 메모리셀을 통한 학습을 진행하였다. 각 모델에 대해 이전과 같이 MAE, MAPE, RMSE의 성능 평가 지표를 산출하였다.

### 2.4.1 순환 신경망(RNN)

RNN은 과거 학습 결과를 현재 학습에 사용하는 딥러닝 네트워크로서, 시계열 데이터를 효과적으로 처리하는 장점이 있다. RNN은 하나의 tanh 혹은 ReLU 활성화 함수를 가진 구조로서, 체인이 길어지면 과거의 학습 결과가 사라지는 장기 의존성 문제가 있다. 병렬 체인 구조로 순환 신경망을 통해 시계열 데이터를 분석하여 어떤 출력이 나올지를 예상할 수 있다. 일반적으로 순환 신경망은 고정 길이가 아닌, 임의의 길이를 가진 시퀀스를 다룰 수 있는 장점이 있다.

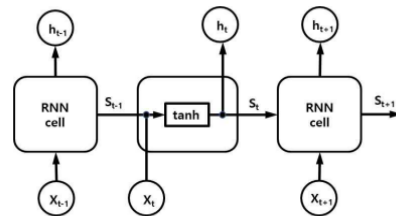


Fig. 2. Structure of RNN

### 2.4.2 장단기 메모리(LSTM)

순환 신경망을 통해 데이터가 가공되는 과정에서 데이터가 변환되기 때문에 어느 정도의 과정을 거치면 초기에 있는 데이터는 승화하게 된다. 장단기 메모리셀은 이러한 문제를 해결한다. LSTM 셀은 중요한 입력을 인식, 장기 상태에 저장, 필요기간 동안 보존, 그리고 필요할 때마다 이를 추출하기 위해 학습을 진행한다. LSTM은 셀 스테이트를 통해 과거 학습 결과를 큰 변함없이 전달하는 구조로 장기 의존성 문제를 해결한다. 먼저 Sigmoid 함수를 사용해 삭제할 정보를 결정한다. Sigmoid 함수, tanh 함수를 이용해 새로운 정보가 셀 스테이트에 저장될지 결정한다. 셀 스테이트를 업데이트하고 마지막 Sigmoid 함수와 셀 스테이트에서 나온 출력을 통과시킨 마지막 tanh 함수로 어떤 출력값을 출력할지 결정한다.

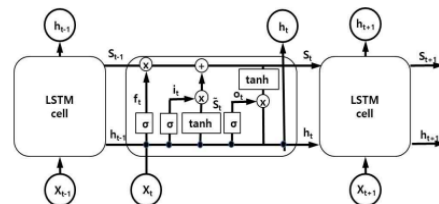


Fig. 3. Structure of LSTM

본 논문의 연구 절차는 Fig. 4와 같다.

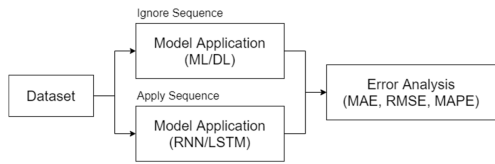


Fig. 4. Research procedure

### 3. 머신러닝 모델 예측 결과

기존 연구에서 적용한 결정 트리 및 다른 모델들 간의 오차율을 통한 성능 지표는 Table 3과 같다. 시계열 데이터를 무시한 조건 하 다양한 모델을 적용한 결과, 랜덤 포레스트 모델에서 기존 연구 대비 MAE 7.69%, RMSE 18.8% 향상된 성능 지표를 확인하였다. 시계열에 대한 데이터 전처리 작업을 진행한 경우, LSTM 모델에서 기존 연구 대비 MAE 53.8%, RMSE 75.0%, MAPE 50.0% 향상으로 기존 연구에 대해 월등히 적은 오차율을 보여 줌으로써 더욱 향상된 예측률을 보여주었다.

Table 3. Performance measurement through estimated error rate

Model	MAE	RMSE	MAPE
Decision Tree	1.3	3.2	1.4
Linear Regression	1.4	3.1	1.6
Random Forest	1.2	2.6	1.3
RNN	0.7	0.9	0.7
LSTM	0.6	0.8	0.7

RNN의 장기 의존성 문제를 고려하여 LSTM 모델을 이용해 예측하였으나, 두 모델 간 평가 지표에서 큰 차이가 없음을 확인하였다. 이는 주어진 데이터가 5초 단위로 약 6분간, 69개의 단위로 시계열을 구성하고 있으며, 장기 의존성 문제를 일으키기엔 데이터 길이가 짧다고 판단하였다. 데이터가 1초 단위로 시계열의 길이가 길어 진다면 LSTM 모델이 효과적인 것이라 예상한다.

산도 및 온도 변수는 0.47의 상관관계를 가지고 있다. 산도 및 온도 변수를 모두 사용하지 않고 처리율을 예측할 수 있는지 확인하기 위해 하나의 독립변수만을 설정하여 각각 학습을 진행하였다. 이에 따른 결과는 Table 4와 같다.

Table 4의 결과를 통해 하나의 독립변수를 사용하였을 때, 기존의 학습 방법보다 낮은 성능을 확인할 수 있

다. 이를 통해 산도와 온도가 각각 학습과는 독립적으로 연관성이 있어, 두 변수를 동시에 학습에 사용했을 때 좋은 성능을 나타냄을 확인할 수 있었다.

Table 4. Performance measurement through different methods using independent variables

Model	Variable	MAE	RMSE	MAPE
RNN	pH	2.0	2.6	2.1
	Temp	0.8	1.2	0.8
	pH, Temp	0.7	0.9	0.7
LSTM	pH	3.3	4.7	3.5
	Temp	3.5	3.6	3.6
	pH, Temp	0.6	0.8	0.7

Table 3의 결정 트리와 선형 회귀 모델의 RMSE 값이 종속변수의 처리율의 표준편차(Table 2 참고)와 비슷한 수준임을 통해, 이 두 모델은 본 공정의 예측에 적합하지 않다는 본 연구의 판단이 적절했음을 확인할 수 있다.

### 4. 결론

본 연구에서는 산제 전처리 공정 운영 최적화 모델에 있어 기존 연구보다 처리율을 정밀하게 예측하였다. 기존 연구에서 이노징크 데이터셋에 대해 분류 목적인 결정 트리 모델을 사용하여 처리율을 예측하였다. 하지만 이노징크 공정 특성상 예측이 목적이며, 이에 따라 회귀 모델을 적용함이 더욱 적절하다고 판단하였다. 가설을 검증하기 위해 MAE, RMSE, MAPE의 성능 측정 지표를 기준으로 삼았다. 연구 결과는 다음과 같다.

1. 시계열 데이터를 고려하지 않은 경우, 랜덤 포레스트 모델이 기존 연구 대비 MAE 7.69%, RMSE 18.8% 향상된 우수한 모델임을 입증하였다.
2. 6분의 처리 과정을 하나의 시퀀스로 할당하는 전처리 과정을 적용한 경우, LSTM 모델이 기존 연구 대비 MAE 53.8%, RMSE 75.0%, MAPE 50.0% 향상의 우수한 성적을 확인하였다.

따라서 산제 전처리 공정은 시퀀스에 민감한 공정이기 에 시계열을 고려하는 RNN, LSTM 모델에서의 더욱 나은 예측 성능을 확인하였다.

본 연구는 일정 시간을 소요하여 결과가 도출되는 도금, 화학, 반도체, 농업 등 다양한 공정에서의 활용이 가능하다. 적절한 모델을 선정함으로써 실제값에 근접한 정

밀한 예측이 가능하며, 이는 최적의 공정을 찾는 시간 및 비용의 절감, 그리고 생산성 증대의 효과를 기대한다.

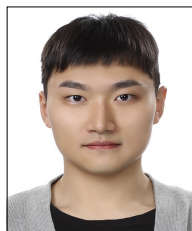
## References

- [1] KAIST(ABH, Impix), AI Dataset for Process Operation Optimization, KAMP(Korea AI Manufacturing Platform), Korea, 2022., Available From: <https://www.kamp-ai.kr/> (accessed Aug. 20, 2022)
- [2] Ivanov, V., Lebedev, V., Davydova, I., & Atoyan, T., "Analysis of the zinc covering quality formed by the vibration chemical-mechanical synthesis", MATEC Web of Conferences, pp.132, 01003. 2017. DOI: <https://doi.org/10.1051/mateconf/201713201003>
- [3] Bolobov, V. I., Latipov, I. U., Popov, G. G., Buslaev, G. V., & Martynenko, Y. V. "Estimation of the influence of compressed hydrogen on the mechanical properties of pipeline steels". Energies, 14(19), 6085. 2021. DOI: <https://doi.org/10.3390/en14196085>
- [4] KEMP Corporation, Products, Available From: <http://kempkorea.com/kr/main/main.php> (accessed Aug. 20, 2022)
- [5] A. Geİron, Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.), p.29-639, O'Reilly.
- [6] Kumari, K., & Yadav, S., "Linear Regression Analysis Study", Journal of the Practice of Cardiovascular Sciences, 4(1), pp.33, 2018. DOI: <https://doi.org/10.4103/jpcs.jpacs.8.18>
- [7] Sherstinsky, A., "Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network", Physica D: Nonlinear Phenomena, pp.404, 132306. 2020. DOI: <https://doi.org/10.1016/i.physd.2019.132306>
- [8] D. H. Shin, "Deep Learning Model for Prediction Rate Improvement of Stock Price Using RNN and LSTM", Korean Institute of Information Technology, v.15 no.10, pp.9-16, 2017. DOI: <https://doi.org/10.14801/ikiit.2017.15.10.9>
- [9] Chai, T., Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development, pp.1247-1250. June 2014. DOI: <https://doi.org/10.5194/gmd-7-1247-2014>

---

Sung-Joo Lee

[Associate member]



- Mar. 2016 ~ current : Handong Univ., Control Engineering, BS
- Mar. 2021 ~ current : Handong Univ., Electronic Engineering, BS

<Research Interests>

Artificial Intelligence, Image Processing, Signal Processing

---

Gang-Ho Lee

[Associate member]



- Mar. 2020 ~ current : Handong Univ., Information Communications Technology Convergence, BS
- Mar. 2020 ~ current : Handong Univ., Mathematics and Statistics, BS

<Research Interests>

Artificial Intelligence, Data Analysis

---

Soo-Yeon Lee

[Associate member]



- Mar. 2018 ~ current : Handong Univ., Mechanical & Control Engineering, BS

<Research Interests>

Artificial Intelligence

Yo-Sup Shin

[Associate member]



- Mar. 2017 ~ current :  
Handong Univ., Mechanical &  
Control Engineering, BS

〈Research Interests〉  
Artificial Intelligence