ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 52, No.4: 603-620, December 2024
https://dx.doi.org/10.7469/JKSQM.2024.52.4.603

AI 기술 기반의 단계적 예측 실험계획법 개발

박경진* · 정제한* · 장준혁** · 신상문*†

* 동아대학교 산업경영공학과
** 선박해양플랜트연구소

Development of Stepwise Forecasting Experimental Design Methods Based on AI Technologies

Park, Kyungjin* · Jeong, Jehan* · Jang, Junhyuk** · Shin, Sangmun*†

* Department of Industrial & Management Systems Engineering, Dong-A University, Republic of Korea
** Korea Research Institute of Ships & Ocean Engineering (KRISO)

ABSTRACT

Purpose: The objective of this paper is to develop forecasting experiment procedures increasing the efficiency and effectiveness of experiments by combining DoE (Design of Experiments) and AI (Artificial Intelligence) algorithms to reduce unnecessary cost and period in phase of animal experiments in the field of new drug development.

Methods: A methodology utilizing AI algorithms like k-NN and XGBoost for interpolating outliers and missing values of DoE results and for predicting results at remaining experimental points of FD (Factorial Design) based on FFD (Fractional Factorial Design) results is proposed in a stepwise experimental design methods.

Results: In this case study, a proposed methodology utilizing AI algorithms for predicting results at remaining experimental points show performance of XGBoost is better than k-NN and the predicting results are significant. Especially, when predicting results at remaining experimental points of FD (Factorial Design) based on FFD (Fractional Factorial Design) results, predicting results are sensitive from whether or not data of center points. This proposed methodology can reduce the cost and period for retesting by utilizing an appropriate AI algorithm in a stepwise experimental design methods.

Conclusion: Combining DoE based on traditional statistical methods with AI algorithms for predicting experimental results is shown that a stepwise experimental design methods can become more efficient and effective.

Key Words: Stepwise Experimental Design, Design of Experiment, k-NN, XGBoost, Design Space

● Received 29 August 2024, 1st revised 4 September 2024, accepted 19 September 2024
† Corresponding Author(sshin@dau.ac.kr)

* This research was supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (RS-2021-KS211529). This work was supported by the National Research Foundation of Korea (NRF) (NRF-2022R1F1A107386411).

1. 서 론

의약품의 품질을 고도화하기 위한 목적으로 의약품 개발단계에서 사전에 리스크를 예측하여 개발 단계별로 의약품의 체계적인 품질관리를 위한 QbD(Quality by design) 방법론이 2008년 도입되었다. QbD 적용시 장점으로 의약품 전주기적 관리 용이, 의약품 품질 향상, 의약품 불량 최소화로 비용 절감, 규제에 대한 탄력적인 대처, 의약품 시장 출시 기간 단축, 재작업 최소화 등이 있지만, 특히 QbD 허가 품목은 제품의 제조공정의 일정 범위 내에서 제약사가 변경을 직접 할 수 있게 된다. 즉, 제품의 일정 범위내 변경에 있어 허가와 규제사항에 과학적 근거를 제시할 수 있는 디자인 스페이스를 확보하게 되어 유연성 있게 관리 및 대처할 수 있게 된다. 검증된 디자인 스페이스 내에서 제약사의 자율 관리 아래 공정 변경이 가능해진다는 것이다. 이 디자인 스페이스(Design Space)를 설정하기 위해 실험계획법을 통해 최소의 시험으로 주요 공정의 관리 범위를 결정할 수 있다(Kim, 2016).

신약 개발은 개발후보물질 선정부터 시판까지 많은 실험이 필요하다. 크게 원천기술연구를 시작으로 개발 후보물질 선정 이후 전임상 시험 및 임상 시험, 신약 허가 및 시판의 5개의 절차로 구성되어있다. 이 절차 중 전임상 시험은 사람에게 신약후보물질을 투여하기 전, 독성 및 활성을 확인하기 위해 동물에게 투여하여 시험하는 단계이다. 따라서 사람에게 사용하는 임상 시험 전 동물실험은 반드시 거쳐야 하는 과정인데, 신약 개발의 발전과 함께 동물실험도 많아지고 있고 이로 인해 동물에 대한 윤리적 문제 또한 지속적으로 제기되고 있다. 따라서 최소한의 동물실험과 통계적 분석을 통해 최대의 정보를 도출하여야만 약리적 효능을 활용할 수 있으므로 다양한 통계방법론 중 실험계획법이 널리 활용되고 있다. 실험계획법은 품질특성에 대한 인자의 영향, 인자들 간의 교호작용 등을 파악할 수 있고 이를 바탕으로 회귀식을 도출하여 품질특성을 최적화하는 인자들의 조합 및 공정 조건을 찾아낼 수 있다(Jung, 2021).

전임상 단계 개발후보물질의 주요변수들에 대한 디자인 스페이스 설정을 위해 동물을 대상으로한 실험계획법 수행 중, 일부 실험에서 실험이 중단되거나 동물에 이상이 발생하는 등의 이상치가 발생한 경우, 재실험이 필요하며 추가적인 시간과 비용이 발생하게 된다. 그리고 여러 변수를 비교분석하기 위해서는 재실험 결과룰 기다려야하는 비효율적인 상황이 발생될 수 있다. 이런 경우, 재실험을 하지 않고 이상치를 예측 혹은 보정 할 수 있다면 비용과 시간을 줄일 수 있는 효율적인 실험이 될 수 있을 것이다.

따라서 본 연구에서는 신약개발과정에 QbD 적용을 위한 실험계획시 이상치나 결측치가 발생할 경우에 인공지능(Artificial Intelligence, AI) 기반의 예측 방법과 통계적 추정 방법을 결합하여 결과를 분석하는 방법을 제시하고자 한다. 첫째, 본 연구에서는 실험계획법 기반의 실험에서 이상치 및 결측치가 발생한 경우, 이를 효과적으로 보간할 수 있는 AI 알고리즘(데이터 증강 및 학습 알고리즘)을 제시하였다. 둘째, 소수의 실험점에서 이상치 및 결측치가 발생할 경우 뿐만 아니라, 1/2 부분요인 실험을 기반으로 완전요인 실험의 나머지 실험점에 대한 결과를 예측할 수 있는 방법을 제시하였다. 셋째, AI 알고리즘 중 k-NN (k-Nearest Neighbor)과 XGBoost (Extreme Gradient Boosting)의 예측 결과를 사례연구를 통해 비교분석 하고, ANOVA와 디자인스페이스를 통해 결과에 대한 통계적 검증을 진행하였다.

2. 이론적 배경 및 선행연구

2.1 AI 기술 활용 현황

AI는 4차 산업혁명의 핵심 기술 중의 하나로 산업 및 사회 부문의 패러다임 변화를 가져오고 있다. AI를 활용한

연구로 산업 부문에서 Lee (2023)는 전기자동차 헤어핀 권선 모터의 레지저용접 불량 실시간 검출에 CNN을 활용하여 공정효율을 향상하였고, Kim (2023)은 경전철 타이어 정비 및 유지 관리 비용을 최소화하기 위해 CNN과 이미지 증강 기법을 통해 타이어 이상 징후 및 수명 예측에 활용하였다. 또한 Jung (2024)은 사출 성형 공정에 Decision Tree, Random Forest 및 XGBoost를 활용한 예측모델 개발을 통해 최적 공정조건을 도출하였다. 사회 부문에서는 Kim (2020)은 서울시의 초미세먼지량을 예측과 시간별 일사량을 예측을 위해 XGBoost, k-NN과 LSTM 알고리즘을 각각 적용하여 성능을 비교평가 하였고, Cheon (2021)은 교통 흐름을 단계별로 나누어 정교한 예측을 위해 Catboost 알고리즘을 이용하여 최적 파라미터를 도출하였다.

신약개발 분야에서도 효율성의 한계와 R&D 비용 증가를 해결하기 위해 빅데이터와 AI를 활용하여 효율성 개선과 효과성 향상을 위한 기술개발이 이루어지는 중이다(K. Mak and M. Pichika, 2019). 신약개발은 화합물의 구조와 효능에 대한 임상데이터, 의료데이터 관련 빅데이터와 연구논문, 특허 자료 등의 방대한 자료 분석을 통해 화합물 활성 및 효능 최대화부터 독성 및 부적용 최소화까지 여러 품질특성을 동시에 최적화해야 한다. 이러한 과정에서 무작위성과 오류를 줄여 약물 개발의 효율성을 향상시키고 빅데이터의 지능적 탐색과 패턴인식을 가능하기 하기 위해 특정 부분을 자동화할 수 있다. 차세대 기술로 주목받고 있는 AI 기술은 신약개발의 모든 단계에 활용될 수 있는데 데이터에 대한 단순한 분석을 넘어 데이터를 통한 학습으로 쉽게 보이지 않은 내재된 패턴에 대한 통찰까지 줄 수 있다(G. Hessler and K. Baringhaus, 2018).

신약개발에 사용되는 AI 기술은 머신러닝(Machine Learning), 세부적으로는 딥러닝(Deep Learing) 알고리즘이 주로 사용된다(J. Vamathevan et. al., 2019). 머신러닝의 대표적인 적용 예로는 바이오마커 확인, 약물 효능 발견, 분자의 생물학적 활성 최적화, 약물-단백질 상호작용 예측, 약물의 새로운 용도 발견 등이 있다(L. Patel et al., 2020). 임상 시험을 위한 단계에서도 유효 물질 도출 및 도출된 화합물의 효능에 대한 평가를 통해 최종후보물질을 도출하기 위해 AI 기술이 활용된다. 또한 생체 적용 환경 뿐만 아니라 약물 화합물 성분의 구성도 고려해야 하는 약물전달체계의 성공적인 구축을 위한 과정에서 AI는 전임상 시험 설계 및 실험 결과 분석의 정확성을 높이고 독성 예측에도 사용된다(M. Ragoza et al., 2017). 임상시험은 최종 약물 후보를 상용 의약품으로 승인받기 위한 단계이다. 임상시험의 성공을 위해 가장 중요한 부분은 임상 대상자를 적절히 선별하는 일이며 AI 기술을 활용해 이러한 임상 시험 설계를 최적화하고 임상시험의 신뢰성과 효율성을 향상할 수 있다. 또한, 신약이 상용화된 후에도 AI를 사용하여 정보 분석이나 약물 부작용 등 약물 감시 모니터링의 효율성을 높일 수 있다(S. Woo, 2018).

AI 기술은 기존 자료를 통한 학습에 의존하기 때문에 자료의 질이 중요하다. 특히 신약개발에서는 계산적으로 생물의 복잡성을 100% 구현할 수 없으므로 AI 기술을 더 효율적으로 사용하기 위해 소량의 데이터로 학습하는 방법 등 학습자료에 대한 많은 연구가 진행되고 있다. 예를 들어 제한된 데이터를 사용하는 모델들로 원샷 학습(one-shot learning)을 변형한 매칭 네트워크와 같은 방법은 소수 데이터 학습에 대한 유의미한 결과를 보여주었고, 생물학적 자료와 같이 양적 자료 확보의 어려움이 있는 분야에 대한 대안으로 연구되고 있다(R. C. Mohs and N. H. Greig, 2017).

따라서 본 연구에서는 신약개발에서 체계적으로 실험 횟수를 최소화하기 위해 단계적인 실험계획(스크리닝, 특성화, 최적화 단계) 적용시 소수, 양질의 생물학적 실험데이터 학습을 통해 추가적인 실험점의 데이터를 예측하는 절차 및 방법론을 제시하였다. AI 알고리즘 중 k-NN 및 XGBoost를 활용하여 알고리즘 성능 비교 및 데이터에 따른 적합한 알고리즘 선정을 진행하였다. 해당 알고리즘 선정 사유은 쉽게 구현이 가능하며, k-NN의 빠른 분석 속도 특성과 XGBoost의 과적합 규제 기능 특성을 고려하였다.

2.2 실험계획법과 AI 적용 현황

실험계획법(Design Of Experiments, DoE)은 실험 횟수를 최소화하기 위해 통계적 근거를 기반으로 실험을 설계하고 영향인자, 품질특성 간의 관계를 통계적 분석을 통해 함수로 나타냄으로써 실험과 비용을 최소화하고 최대의 정보를 얻기 위한 방법이다. 실험계획법(DoE)은 고품질에 대한 지속적인 고객 요구사항을 만족하기 위해 제품 개발에 적용되고 있으며 설계, 제조, 전기, 공정 능력 개선, 의약품 개발 등 다양한 분야에 적용되고 있다. 반응변수인 결과물에 대한 규격이 정해져 있고 입력변수인 공정 변수의 수준이 서로 각각 낮은 수준과 높은 수준을 원하는 값으로 설정할 수 있을 경우 주로 요인설계(Factorial Design, FD) 방법을 활용한다(Han, Y. et al., 2021). 이 밖에도 실험 목적에 따라 부분요인설계(Fractional Factorial Design, FFD), 반응표면법(Response Surface Method, RSM) 등이 활용되고 있다.

품질 특성을 만족시키기 위해 실험계획을 통한 영향인자들의 최적 조합을 도출하는데 RSM이 활용되고 있다. RSM은 2차의 다항 회귀를 사용하며, 중심점과 축점을 포함하는 곡률효과를 통해 정밀한 최적해를 도출할 수 있다. RSM의 회귀식은 아래 식 (1)과 같다.

$$\hat{y}(x) = \hat{\beta_0} + X^T\hat{b} + X^T\hat{B}X \;\; where, X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \hat{b} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \text{and } \hat{B} = \begin{bmatrix} \hat{\beta}_{11} & \hat{\beta}_{12}/2 & \cdots & \hat{\beta}_{1p}/2 \\ & & \cdots & \\ & & \ddots & \\ sym & & & \hat{\beta}_{pp} \end{bmatrix} \quad (1)$$

RSM 방법으로 박스-벤킨법(Box-Behnken Design, BBD)과 중심합성법(Central Composite Design, CCD)이 있다. CCD의 경우 실험점이 꼭지점, 축점, 중심점으로 구성되기 때문에 완전요인설계(꼭지점 및 일부 중심점)를 먼저 수행하여 입력변수의 유의성 및 실험 범위의 적절성을 판단하고, 범위가 적절하다고 판단될 경우 축점과 나머지 중심점을 추가하여 실험을 진행함으로써 단계적 실험이 가능하다는 장점이 있다(Lee, H. et al., 2022).

도출된 RSM의 회귀식을 기반으로 품질특성의 규격을 만족시키는 영향인자의 설계 가능 범위인 디자인스페이스(Design Space, DS)를 도출하고 이 디자인스페이스 내에서 실험의 분산을 고려한 95% 신뢰구간이 적용된 안전가용영역(Operating Space)을 설정해 최적화를 진행할 수 있다. 디자인 스페이스는 품질 특성을 만족시키는 신뢰성 높은 통계적 설계 가능 영역을 의미하며, 디자인 스페이스 안에서의 변동은 목표 품질의 저하를 초래하지 않음을 의미한다(ICH guideline, 2009).

실험계획법에도 AI 기술이 다양하게 적용되어지고 있다. 우선 실험계획의 최적화 단계에서 머신러닝(Machine Learning), 딥러닝(Deep Learing) 알고리즘은 입력변수와 출력변수간의 함수관계를 통계적 분석인 LSM(Lesat Square Method)의 기본 error에 대한 가정없이 효과적으로 도출할 수 있다. Chang and Chen(2011)은 유전 알고리즘에 딥러닝 예측 방법을 통합함으로써 최적공정조건을 도출하였다. Le et al.(2021)은 딥러닝을 활용한 최적해 도출시 신경망 노드의 출력을 계산하는 활성화 함수 최적화 연구를 통해 AI 알고리즘 성능을 최대화 하였으며, Le et al.(2022)은 통계적 분석인 LSM과 딥러닝, 머신러닝을 통한 최적해 비교분석을 통해 AI 알고리즘 성능의 우수성을 보여주었다. 그리고 실험계획의 설계부문에서 Viana FA et al.(2010)은 컴퓨터 시뮬레이션에 사용되는 특수한 실험 설계인 Latin Hypercube 설계에 머신러닝을 활용하여 실험에서 유의미한 결과를 얻을 수 있는 최적 실험점을 도출하였다. 또한 실험계획법은 최소한의 실험데이터로 최대의 정보를 얻는 분석 방법으로 빅데이터 바탕의 AI 분석을 통한 예측이 어렵다. 그래서 Lou et al.(2019)은 실험계획을 통해 얻은 적은 데이터로 유의한 예측값을 도출하기 위해 여러 AI 알고리즘 성능을 비교 평가하였고, Le and Shin(2021)은 실험계획을 통해 얻은 데이터로부터 AI 학습

을 위한 데이터 증강시 분산이 크고 작음에 따른 AI 알고리즘 성능의 차이를 비교 분석하였다.

이와 같이 AI 기술이 실험계획법에 적용된 사례는 입력변수와 출력변수 간의 관계를 통한 최적해 도출과 최적의 실험점 도출 부문이었으며, J. Freiesleben et al.(2020)은 현재 실험계획 절차에 따른 인자선정 및 실험계획 설계에 대한 의사결정이 사람에 의해 수행되지만 이러한 부분 또한 AI 알고리즘 활용에 대한 연구의 필요성을 제시하였다.

2.3 k-NN 알고리즘

k-최근접 이웃 알고리즘(k-Nearest Neighbors, k-NN)의 설명에 앞서 NN(Nearest Neighbors)은 최근접한 이웃을 사용한 분류이며, k-NN은 1968년에 Hart에 의해 제안된 알고리즘으로 k개의 가장 근접한 이웃을 이용하여 훈련시킨 데이터 집합에 속해있는 표본들 간의 유사도에 따라서 라벨이 없는 표본들을 직관적인 방법으로 분류하는 방법이다(Dasarathy, 1991). 즉, 라벨이 없는 데이터 집합에서 가장 가까운 k개의 표본을 묶었을 때 나타나는 집합에서 빈도수가 많은 그룹에 k개의 표본을 할당시키는 방법이다. k-NN 유사도를 측정하는 다양한 방법들 중 대표적으로 유클리드 거리(Euclidean distance)가 있으며 아래 식 (2)와 같다.

$$Euclidean\,distance\;\; D(\underline{x},\underline{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

여기서 $\underline{x} = (x_1, \cdots, x_p)$, $\underline{y} = (y_1, \cdots, y_p)$는 공간상의 두 점을 뜻하고, p는 차원을 뜻한다. 최적의 k값을 결정하는 것은 데이터에 의존적이므로, k 값을 크게 설정하면 데이터 구조 분석이 어렵고 기존의 데이터의 분류결과에 따라 편향될 수 있으며, k 값을 작게 설정하면 이상치의 영향을 많이 받을 수 있다(Won, C., 2018).

2.4 XGBoost 알고리즘

XGBoost 학습 모델은 Chen, Guestrin에 의해 제안된 방법론으로 트리모델(tree model)에 boosting 방법론을 적용하였다. Boosting 방법은 오차가 최소화되는 방향으로 약한 학습기(weak learner)들을 결합하여 강한 예측기(strong learner)를 만드는 앙상블 알고리즘으로 새로운 예측기를 만들고 이를 학습하는 과정에 발생하는 오류를 최소화하는 방향으로 가중치를 조정하여 다음 예측기를 만들고 학습하는 과정을 반복한다. 기본적인 트리의 boosting 모델 결과값 예측은 식 (3)을 이용할 수 있다(Chen and Guestrin, 2016).

$$\hat{y_i} = \varnothing(x_i) = \sum_{k=1}^{K} f_k(x_i),\; f_k \in F \tag{3}$$

여기서 K는 트리의 개수, F는 모든 CART(Classification And Regression Tree) 집합을 의미하며, f_k는 트리와 가중치에 대응되며, K개의 트리 가중치를 모두 더하여 결과를 예측한다. gradient boosting은 boosting의 대표적인 모델로 경사 하강법으로 오차를 최소화하는 방법이며, 경사하강법은 손실 함수(loss function)를 정의한 후 손실의 크기를 최소화하는 방향으로 모형을 변화시키는 것이다. Gradient boosting은 손실 함수의 손실을 최소화시키는 최적 함수를 찾기 위해 모든 경우의 수를 탐색하므로 과적합의 단점이 있고, 고려할 변수가 많을 경우 연산의 효율성이 떨어져 분석시간이 느려진다. 하지만 XGBoost는 트리의 복잡도 증가에 따라 손실함수에 페널티(penalty)를 부여하는 방식으로 복잡도를 제한하여 과적합을 방지한다. XGBoost의 손실함수는 식 (4)와 같다.

$$\ell(\varnothing) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k), \ \ where, \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad (4)$$

여기서 l은 손실함수로 예측값 $\hat{y_i}$과 실제값 y_i사이 차이를 의미하며, Ω는 트리의 정규화 항(regularization term)으로 γT는 트리 잎의 갯수, $1/2\lambda \parallel w \parallel^2$는 트리 잎의 점수를 의미한다. 즉 트리의 잎의 갯수와 점수가 작아지는 방향으로 학습되며, 이것은 모델의 과적합을 방지하는 역할을 한다. 또한 XGBoost는 분산 및 병렬 처리를 통한 학습이 가능하여 빠른 분석시간을 가진다. 기존 gradient boosting의 단점을 이러한 방법으로 보완하여 연산 효율성을 높이고 안정적인 예측이 가능하다.

3. 연구 방법

3.1 AI 기술 기반의 단계적 예측 실험계획법

위의 선행 연구를 통해 AI 알고리즘 및 실험계획법을 활용한 방법론에 대한 연구들을 알아보았다. 이를 바탕으로 AI 알고리즘과 실험계획법을 결합하는 방법론을 제안하고자 한다. 아래의 Figure 1은 본 연구가 제안하는 AI 알고리즘을 결합한 예측 실험계획법 분석 절차를 도식화한 것이다. 실험 분석 진행 과정은 연구에 대한 목표, 기대 효과를 설정하고 실험 횟수, 비용 및 상황을 고려하여 실험을 설계한 후, 실험설계표에 따라 실험을 수행한다. 실험이 완료되었으면 분석 절차에 맞추어 결과 분석을 수행한다. 분석 결과를 바탕으로 최적값이 도출되고 재현성 실험을 통해 검증 실험을 수행하고 검증이 완료되었으면 그 분석 결과를 바탕으로 관리범위를 수립하는 절차로 구성된다.

실험계획법(DOE)은 영향인자가 품질특성에 미치는 영향 및 관계를 체계적으로 접근하기 위해 실험을 설계하고 통계적으로 분석하는 방법론이다. 단계적 실험계획법의 절차는 Figure 1과 같이 실험 목적에 따라 3단계로 나누어 구성된다. 첫 번째 단계는 다수의 입력변수가 있는 경우 실험 횟수가 증가하는 것을 방지하기 위한 변수선별 단계(Screening step)이며 주로 플라켓 버만 설계(Plackett-Burman)과 부분요인설계법(FFD)를 활용하여 가장 유의한 최소의 영향인자를 선별한다. 두 번째는 특성화 단계(Characterization step)로 유의하게 선별된 영향인자들을 기반으로 실험 범위의 적절성을 분석하는 단계로 완전요인설계법(FD)이 주로 활용된다. 마직막인 최적화 단계(Optimization step)에서는 더욱 정밀한 실험을 위해 중심점과 축점 등의 실험점을 추가하여 도출된 회귀식을 통한 최적화 단계로 반응표면법(RSM) 등이 활용된다. 먼저 실험에 필요한 영향인자와 품질특성을 식별한 뒤, 영향인자들 중 품질특성에 영향을 크게 미치는 주요 영향인자를 실험을 통해 선별하고 완전요인설계법(FD)을 기반으로 실험 범위의 적절성 및 통계적 유의성을 판단한다. 그리고 최적화 단계에서는 반응표면법(RSM)의 중심합성계획법(CCD)을 통해 정밀한 실험으로 영향인자와 품질특성 간의 함수관계를 규명할 수 있다. 마지막으로 도출된 회귀식을 기반으로 품질특성의 규격을 만족시키는 디자인스페이스(DS) 및 최적조건 도출을 진행할 수 있다.

본 연구에서는 신약개발과정의 단계적 실험계획시 이상치가 발생한 경우나 결측치가 발생할 때 AI 기반의 예측 방법과 통계적 추정 방법을 결합하여 실험결과를 분석하는 방법론을 제시하고자 한다. 첫째, 실험계획법 기반의 실험에서 이상치 및 결측치가 발생한 경우, 이를 적절한 AI 알고리즘의 활용을 통해 효과적으로 보간하는 방법을 제시하였다. 둘째, 실험점에서 소수의 이상치 및 결측치가 발생했을 경우뿐 아니라, 1/2 부분요인 실험을 기반으로 완전요인실험의 나머지 실험점의 결과를 예측하는 방법을 제시하였다. 셋째, 사례연구를 통하여 AI 알고리즘 중 k-NN(k-Nearest Neighbor)과 XGBoost(eXtreme Gradient Boosting)를 활용하여 예측 결과를 비교하고 통계적으로 검증하였다.

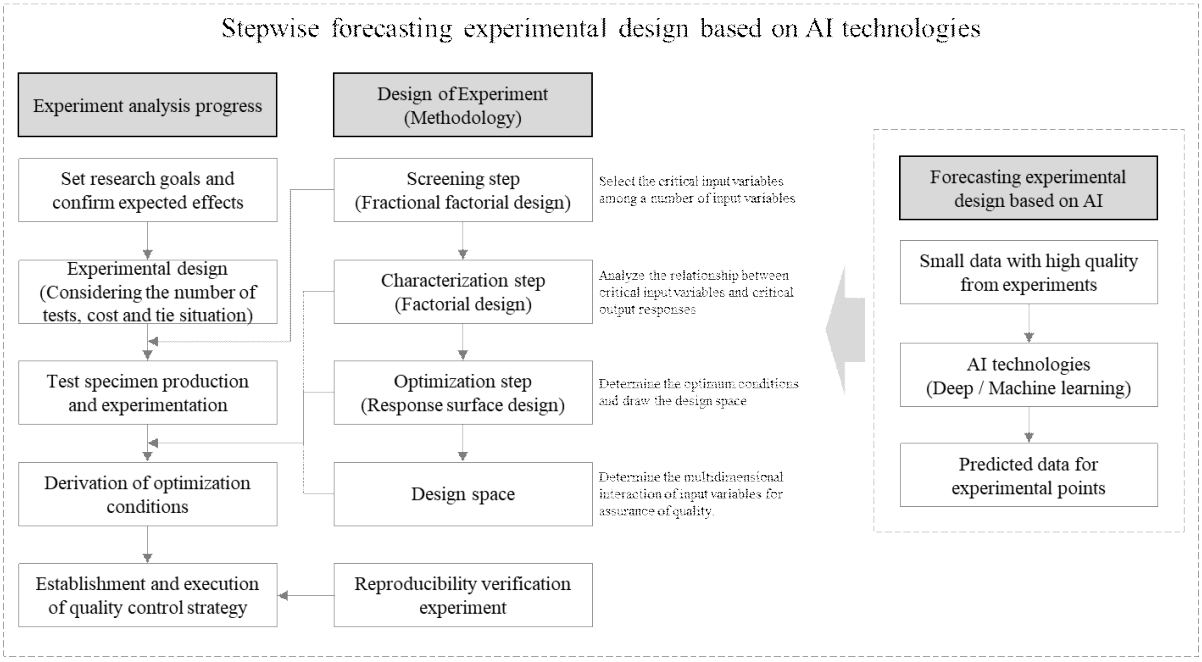

Figure 1. Overview of stepwise forecasting experimental design methods based on AI technologies

3.2 실험점에서 결측 및 이상치에 대한 예측

SO	EO	Input factors			Output responses	
		x1	x2	x3	y1	y2
1	7	-1	-1	-1		52.14
2	2	1	-1	-1	93.71	42.45
3	1	-1	1	-1	77.88	45.14
4	5	1	1	-1	80.52	43.42
5	3	-1	-1	1	84.31	52.03
6	4	1	-1	1	80.71	44.43
7	9	-1	1	1	83.95	43.9
8	6	1	1	1	77.96	43.78
9	8	0	0	0	85.31	47.1

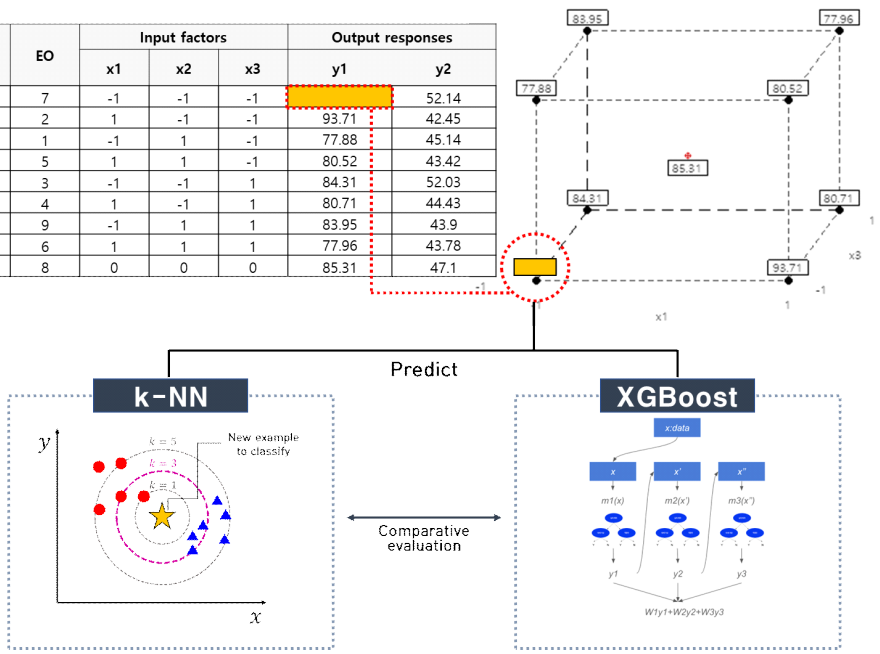

Figure 2. Prediction of results at a experimental point

실험계획법의 실험점을 대상으로 실험결과의 결측 발생시 AI 알고리즘을 활용한 결과 예측을 수행하고자 한다. 실험계획법을 통해 확보된 실험결과 데이터를 바탕으로 Figure 2에서와 같이 임의의 실험점 1개에 대해 결측을 가정하고 AI 알고리즘을 활용하여 예측값을 도출하였고 예측값과 실제값의 비교분석은 유클리드 거리와 R^2 를 통해 AI 알고리즘의 성능을 평가하였다. AI 학습을 위한 데이터는 각 실험점의 결과값을 기준으로 0.3의 표준편차를 부여해 각 실험점마다 100개의 랜덤 데이터를 생성하였다. 통계 상용 소프트웨어 프로그램 Minitab를 통해 실험설계 및 분석을 수행하였으며, 클라우드 기반 개발환경 구글 코랩을 통해 AI 알고리즘은 분석을 수행하였다.

3.3 부분요인실험점을 통한 완전요인실험점 결과 예측

부분요인실험 결과 데이터에 대한 AI 학습을 통해 완전요인실험의 나머지 실험점에 대한 결과를 예측한 후 그 결과에 대한 ANOVA 및 디자인 스페이스를 통해 결과의 유의성을 검증하였다. 분석 과정은 3.2와 동일하게 데이터 생성 후 아래 Figure 3.과 같이 입방체도에서 부분요인실험점을 제외한 나머지 한 실험점의 결과를 예측하고 그 예측값을 기준으로 0.3의 표준편차를 적용하여 AI 학습데이터 100개를 생성하였고 나머지 실험점들 또한 같은 방식으로 결과 예측 및 학습데이터 생성을 반복하였다. 실험 데이터는 부분요인실험점의 결과를 바탕으로 예측 및 학습데이터 생성을 수행하였고, 예측값과 실제 값과 비교하여 알고리즘 성능을 평가하였다.

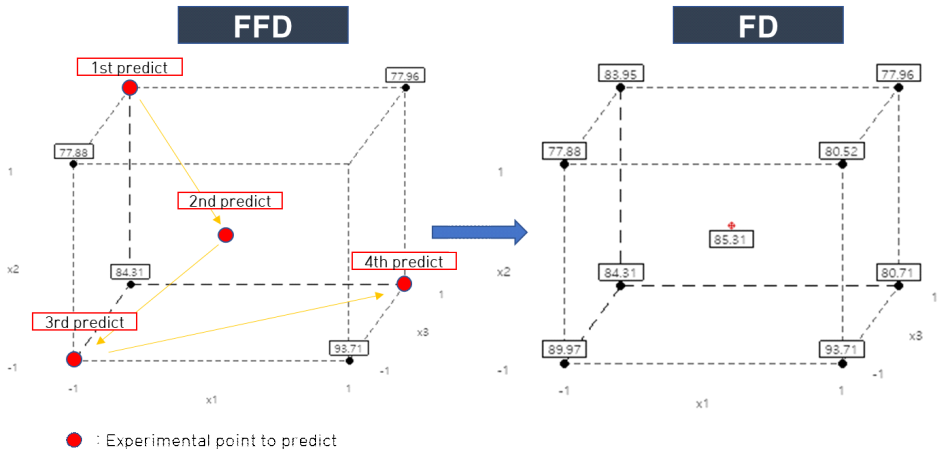

Figure 3. Prediction of results at remaining experimental points of FD based on FFD results

4. 연구 결과

4.1 실험점 결측 발생에 따른 예측

본 실험에서는 전임상 단계 동물실험을 위하여 Table 1과 같이 2개의 종속변수를 가진 2수준 3인자의 완전요인실험에 중심점을 추가한 실험계획과 1/2 부분요인실험에 따라 실제 실험을 수행한 결과를 대상으로 임의로 하나의 실험점에 결측을 부여하고 제안한 AI 알고리즘을 활용하여 예측한 후 임의로 결측한 값과 실제값을 비교 분석하였다. 실험점 9개 중 1개씩의 실험점 종속변수에 결측을 주어 총 각각의 9개 실험점에 대한 예측 성능을 분석하였다.

Table 1. Experimental data for FD and FFD (3factors, 2levels, 1center point with 2responses)

SO	factorial design					1/2 fractional factorial design				
	Input factors			Output responses		Input factors			Output responses	
	x_1	x_2	x_3	y_1	y_2	x_1	x_2	x_3	y_1	y_2
1	-1	-1	-1	89.97	52.14					
2	1	-1	-1	93.71	42.45	1	-1	-1	93.71	42.45
3	-1	1	-1	77.88	45.14	-1	1	-1	77.88	45.14
4	1	1	-1	80.52	43.42					
5	-1	-1	1	84.31	52.03	-1	-1	1	84.31	52.03
6	1	-1	1	80.71	44.43					
7	-1	1	1	83.95	43.90					
8	1	1	1	77.96	43.78	1	1	1	77.96	43.78
9	0	0	0	85.31	47.10					

k-NN 알고리즘을 이용한 예측 결과, Table 2와 Figure 4와 같이 y_1과 y_2의 실제값과 예측값의 R^2 가 각각 6.27%, 68.3%로 낮게 나타나 본 사례연구 데이터에 대한 해당 알고리즘 성능이 좋지 않음을 의미한다. XGBoost 알고리즘을 사용하여 실험점 예측 결과, Table 3와 Figure 5와 같이 y_1과 y_2의 실제값과 예측값의 R^2 가 각각 98.4%, 92.02%로 높게 나타나 XGBoost가 k-NN 알고리즘보다 오차값이 작고 예측 성능이 좋은 것으로 나타났다.

Table 2. Prediction of results at each experimental points using k-NN

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	52.14	81.90	51.94				
2	93.71	42.45	83.57	46.55				
3	77.88	45.14	88.02	43.40				
4	80.52	43.42	84.6	43.58				
5	84.31	52.03	88.66	50.09	19.79	6.20	6.27%	68.3%
6	80.71	44.43	87.97	44.81				
7	83.95	43.90	79.3	46.94				
8	77.96	43.78	80.9	44.00				
9	85.31	47.10	83.81	49.42				

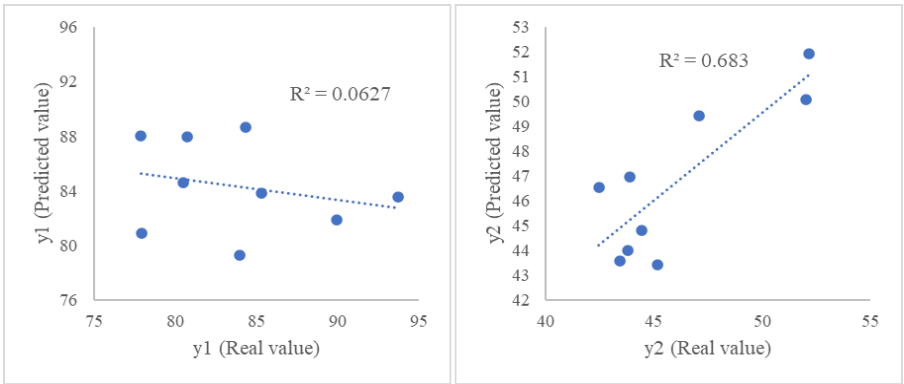

Figure 4. Comparison between real value and predicted value from k-NN

Table 3. Prediction of results at each experimental points using XGBoost

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	52.14	88.29	50.04				
2	93.71	42.45	92.01	44.57				
3	77.88	45.14	79.25	45.10				
4	80.52	43.42	80.61	43.36				
5	84.31	52.03	85.10	51.27	2.91	3.31	98.4%	92.02%
6	80.71	44.43	80.87	45.36				
7	83.95	43.90	83.73	43.68				
8	77.96	43.78	78.16	43.13				
9	85.31	47.10	85.67	47.48				

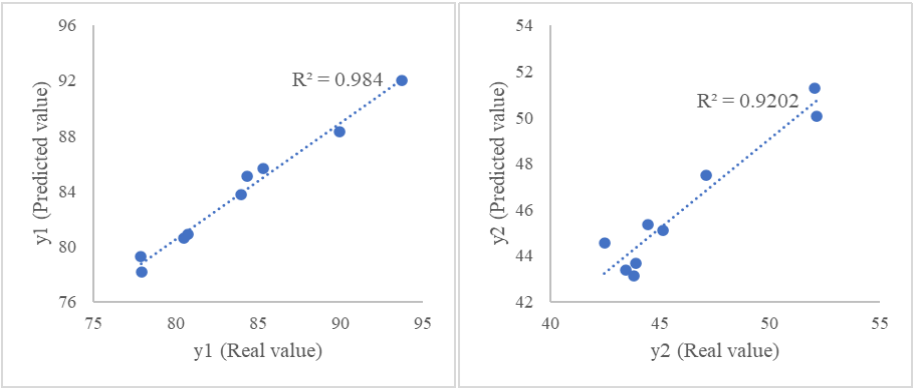

Figure 5. Comparison between real value and predicted value from XGBoost

4.2 부분요인실험점에 대한 완전요인실험점 예측

본 절에서는 부분요인실험 결과를 바탕으로 중심점값의 유무에 따른 완전요인실험의 나머지 실험점에 대한 AI 알고리즘 예측 성능을 비교 분석하였다. 우선, 중심점값을 반영한 k-NN 알고리즘의 예측 성능을 분석하였고, 결과는 Table 4 및 Figure 6와 같다. 중심점값을 반영한 부분요인실험점을 통한 나머지 실험점에 대한 k-NN 분석 결과, y_1과 y_2의 실제값과 예측값의 R^2이 각각 68.94%, 11.21%로 낮게 나타나며, 예측값이 모두 중심점 결과값과 비슷하게 도출되어 예측이 불가하였다. k-NN 알고리즘은 k의 값 및 거리 기반의 예측 알고리즘이며, 중심점의 값이 훈련 데이터에 존재할 때 다른 점들의 값보다 중심점의 값들이 훈련에 포함되는 횟수가 많기 때문으로 판단된다. 따라서 중심점 값이 포함되어 있을 때 k-NN 알고리즘을 이용한 실험점 결과 예측은 적절하지 않은 것을 확인할 수 있다. XGBoost 알고리즘을 이용한 예측 결과는 Table 5와 Figure 7과 같이 y_1과 y_2의 실제값과 예측값의 R^2가 각각 91.78%, 13.73%로 나타나 중심점 값을 포함한 실험점 결과 예측에 적절하지 않은 것으로 판단되었다.

중심점을 반영하지 않고 부분요인실험점을 통한 k-NN 알고리즘을 이용하여 완전요인실험의 나머지 실험점에 대한 예측 결과는 Table 6와 Figure 8과 같이 y_1과 y_2의 실제값과 예측값의 R^2가 각각 0.12%, 14.67%로 k-NN 알고리즘을 이용한 실험점 결과 예측은 적절하지 않은 것을 확인할 수 있다. XGBoost 알고리즘을 사용한 경우의 결과는 Table 7, Figure 9과 같으며 y_1과 y_2의 실제값과 예측값의 R^2가 각각 96.04%, 72.32%로 높게 나타나며 k-NN 알고리즘보다 예측 성능이 우수하였다. 종합적으로 부분요인실험 결과를 바탕으로 완전요인실험의 나머지 실험점에 대한 AI 알고리즘 예측시 중심점 값을 반영하지 않은 XGBoost 알고리즘이 k-NN 알고리즘보다 예측 성능이 좋은 것

으로 판단되었다.

Table 4. Prediction of results at remaining experimental points of FD based on FFD results with a center point using k-NN

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	52.14	85.25	47.10	8.21	7.48	68.94%	11.21%
4	80.52	43.42	85.27	47.08				
6	80.71	44.43	85.27	47.06				
7	83.95	43.90	85.25	47.11				

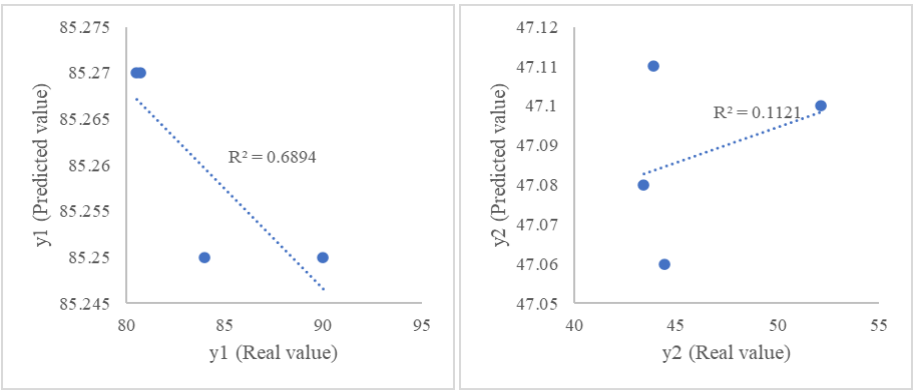

Figure 6. Comparison between real value and predicted value from k-NN

Table 5. Prediction of results at remaining experimental points of FD based on FFD results with a center point using XGBoost

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	44.43	88.84	43.44	2.60	10.79	91.78%	13.73%
4	80.52	43.42	81.38	42.23				
6	80.71	43.90	82.32	47.42				
7	83.95	52.14	82.49	49.42				

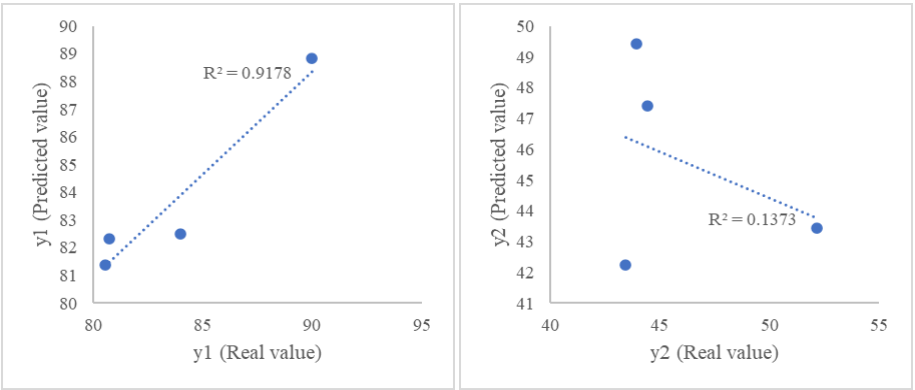

Figure 7. Comparison between real value and predicted value from XGBoost

Table 6. Prediction of results at remaining experimental points of FD with a center point based on FFD results without a center point using k-NN

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	52.14	82.52	44.91				
4	80.52	43.42	77.89	51.89				
6	80.71	44.43	87.20	43.13	10.38	11.77	0.12%	14.67%
7	83.95	43.90	83.71	47.27				
9	85.31	47.10	83.53	48.33				

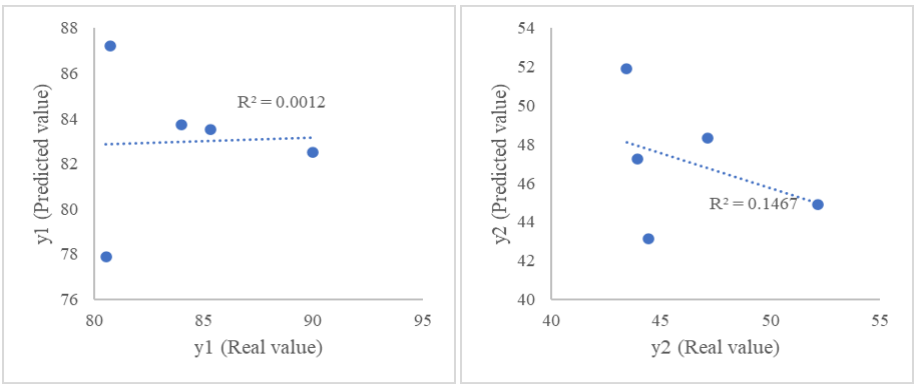

Figure 8. Comparison between real value and predicted value from k-NN

Table 7. Prediction of results at remaining experimental points of FD with a center point based on FFD results without a center point using XGBoost

SO	Real value		Predicted value		Euclidean distance		R^2	
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2
1	89.97	52.14	88.71	47.46				
4	80.52	43.42	81.40	42.86				
6	80.71	44.43	80.63	45.40	1.79	4.82	96.04%	72.32%
7	83.95	43.90	83.45	44.05				
9	85.31	47.10	86.08	46.99				

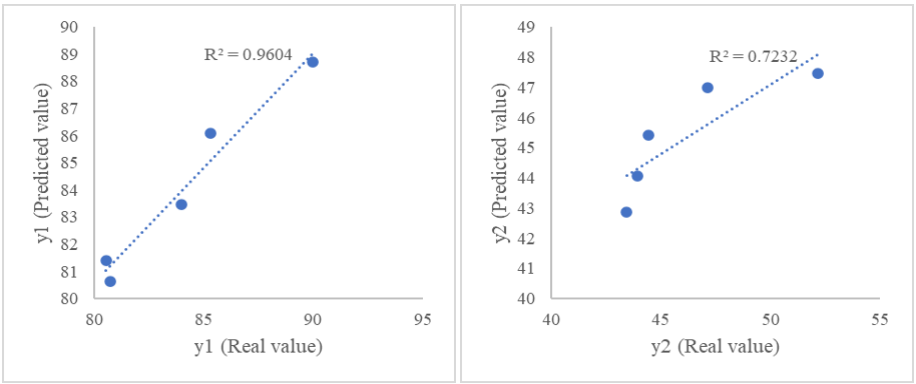

Figure 9. Comparison between real value and predicted value from XGBoost

4.3 RSM(반응표면법) 및 디자인 스페이스 분석을 통한 예측결과 검증

위 예측 결과를 바탕으로 기존 부분요인실험 실측값과 완전요인실험점의 나머지 실험점과 중심점에 대한 실측값과 k-NN, XGBoost 예측값을 통한 RSM 분석을 수행하였다. 실측값과 예측값 모두 모형의 p-값 0.05 이하로 유의하게 나타났으며, 각각에 대한 추정 반응함수와 ANOVA는 아래 식 (5)~(10)과 Table 8~10과 같다. 실측값과 XGBoost 예측값의 y_1에 대한 주요 영향인자는 x_2, x_3, 교호작용은 x_1*x_3, x_2*x_3로 동일하게 p값 0.05 이하로 나타났고, y_2에 대한 주요 영향인자 또한 x_1, x_2, 교호작용은 x_1*x_2가 p값 0.05 이하로 동일하게 나타났으나, k-NN 예측값의 주요 영향인자, 교호작용 모두 실측값 결과와 다르게 나타났다.

$$y_1\_Real = 85.310-0.401*x_1-3.549*x_2-1.894*x_3-1.684*x_1^2-1.996*x_1*x_3+2.771*x_2*x_3 \quad (5)$$

$$y_2\_Real = 47.100-2.391*x_1-1.851*x_2+0.123*x_3-1.189*x_1^2+1.931*x_2*x_2+0.461*x_1*x_3-0.343*x_2*x_3 \quad (6)$$

$$y_1\_KNN = 83.190+1.043*x_1-3.788*x_2+0.148*x_3-2.477*x_1*x_2-1.758*x_1*x_3+1.327*x_2*x_3 \quad (7)$$

$$y_2\_KNN = 46.548-1.012*x_1+0.695*x_2+1.827*x_1*x_2-2.085*x_1*x_3-1.723*x_2*x_3 \quad (8)$$

$$y_1\_XGB = 86.080-0.081*x_1-3.333*x_2-1.918*x_3-2.574*x1^2-0.411*x_1*x_2-2.211*x_1*x_3+2.451*x_2*x_3 \quad (9)$$

$$y_2\_XGB = 45.573-1.774*x_1-1.439*x_2+0.919*x_3+1.136*x_1*x_2-0.961*x_2*x_3 \quad (10)$$

위 RSM 분석 결과를 바탕으로 실측값과 k-NN, XGBoost 예측값에 대한 디자인 스페이스 및 최적값을 비교하였다. 반응변수에 대한 예측값의 평균이 허용된 범위 내의 영역을 효과적으로 식별하기 위해서 중첩등고선도를 활용하였으며, 흰색 영역은 가용 영역을 나타내며 다른 변수들의 값을 고정한 상태로 계량형 변수에 따라 형성되는 영역을 의미한다. 각 반응에 대한 가용 영역의 95% 신뢰구간을 적용한 디자인 스페이스는 Figure 10.과 같다. 실측값과 XGBoost 예측값의 디자인 스페이스가 유사하게 나타나며, 또한 실측값에 대한 최적조건 x_1 0.057, x_2 0.551, x_3 -1.000, XGBoost 예측값에 대한 최적조건 x_1 -0.854, x_2 0.090, x_3 -1.000 으로 모두 디자인 스페이스 내에서 도출되었다, 하지만 실측값과 k-NN 예측값에 대한 디자인 스페이스는 서로 다르게 나타난 것을 볼 수 있다.

Table 8. ANOVA table for real value

구분	y_1_real					y_2_real				
	DF	Adj SS	Adj MS	F값	P값	DF	Adj SS	Adj MS	F값	P값
모형	6	226.56	37.71	43.64	0.023	7	107.02	15.28	509.43	0.034
선형	3	130.77	43.56	50.36	0.02	3	73.28	24.42	813.93	0.026
x_1	1	1.28	1.28	1.49	0.347	1	45.74	45.74	1524.19	**0.016**
x_2	1	100.79	100.79	116.44	**0.008**	1	27.41	27.41	913.52	**0.021**
x_3	1	28.69	28.69	33.16	**0.029**	1	0.12	0.12	4.08	0.293
제곱	1	2.52	2.52	2.91	0.23	1	1.25	1.25	41.85	0.098
x_1*x_1	1	2.52	2.52	2.91	0.23	1	1.25	1.25	41.85	0.098
2차 교호작용	2	93.39	46.69	53.93	0.018	3	32.48	10.82	360.8	0.039
x_1*x_2						1	29.83	29.83	994.18	**0.02**
x_1*x_3	1	31.88	31.88	36.84	**0.026**	1	1.70	1.70	56.71	0.084
x_2*x_3	1	61.49	61.49	71.01	**0.014**	1	0.94	0.94	31.5	0.112
오차	2	1.71	0.85			1	0.03	0.03		
총계	8	228.27		R^2 = 99.24%		8	107.05		R^2 = 99.97%	

Table 9. ANOVA table for predicted value at remaining experimental points of FD with a center point based on FFD results without a center point using k-NN

구분	y_1_KNN					y_2_KNN				
	DF	Adj SS	Adj MS	F값	P값	DF	Adj SS	Adj MS	F값	P값
모형	6	211.54	35.25	75.3	0.013	5	97.29	19.45	10.08	0.043
선형	3	123.63	41.21	88.01	0.011	2	12.06	6.03	3.12	0.185
x_1	1	8.69	8.69	18.57	0.05	1	8.20	8.20	4.25	0.131
x_2	1	114.76	114.76	245.09	0.004	1	3.86	3.86	2	0.252
x_3	1	0.17	0.17	0.37	0.604					
2차 교호작용	3	87.91	29.30	62.58	0.016					
x_1*x_2	1	49.10	49.10	104.87	0.009	1	26.71	26.71	13.84	0.034
x_1*x_3	1	24.71	24.71	52.77	0.018	1	34.77	34.77	18.01	0.024
x_2*x_3	1	14.09	14.09	30.11	0.032	1	23.73	23.73	12.29	0.039
오차	2	0.93	0.46			3	5.79	1.93		
총계	8	212.47		R^2 = 99.56%		8	103.09		R^2 = 94.38%	

Table 10. ANOVA table for predicted value at remaining experimental points of FD with a center point based on FFD results without a center point using XGBoost

구분	y_1_XGB					y_2_XGB				
	DF	Adj SS	Adj MS	F값	P값	DF	Adj SS	Adj MS	F값	P값
모형	7	212.84	30.40	2233.7	0.016	5	66.20	13.24	10.12	0.043
선형	3	118.41	39.47	2899.71	0.014	3	48.48	16.16	12.36	0.034
x_1	1	0.05	0.052	3.88	0.299	1	25.17	25.17	19.24	**0.022**
x_2	1	88.91	88.91	6531.58	**0.008**	1	16.56	16.56	12.66	**0.038**
x_3	1	29.45	29.45	2163.66	**0.014**	1	6.75	6.75	5.16	0.108
제곱	1	5.88	5.88	432.56	0.031					
x_1*x_1	1	5.88	5.88	432.56	0.031					
2차 교호작용	3	88.53	29.51	2168.08	0.016	2	17.72	8.86	6.77	0.077
x_1*x_2	1	1.35	1.35	99.39	0.064	1	10.32	10.32	7.9	**0.047**
x_1*x_3	1	39.11	39.11	2873.61	**0.012**					
x_2*x_3	1	48.06	48.06	3531.24	**0.011**	1	7.39	7.39	5.65	0.098
오차	1	0.014	0.013			3	3.92	1.30		
총계	8	212.85		R^2 = 99.99%		8	70.12		R^2 = 94.40%	

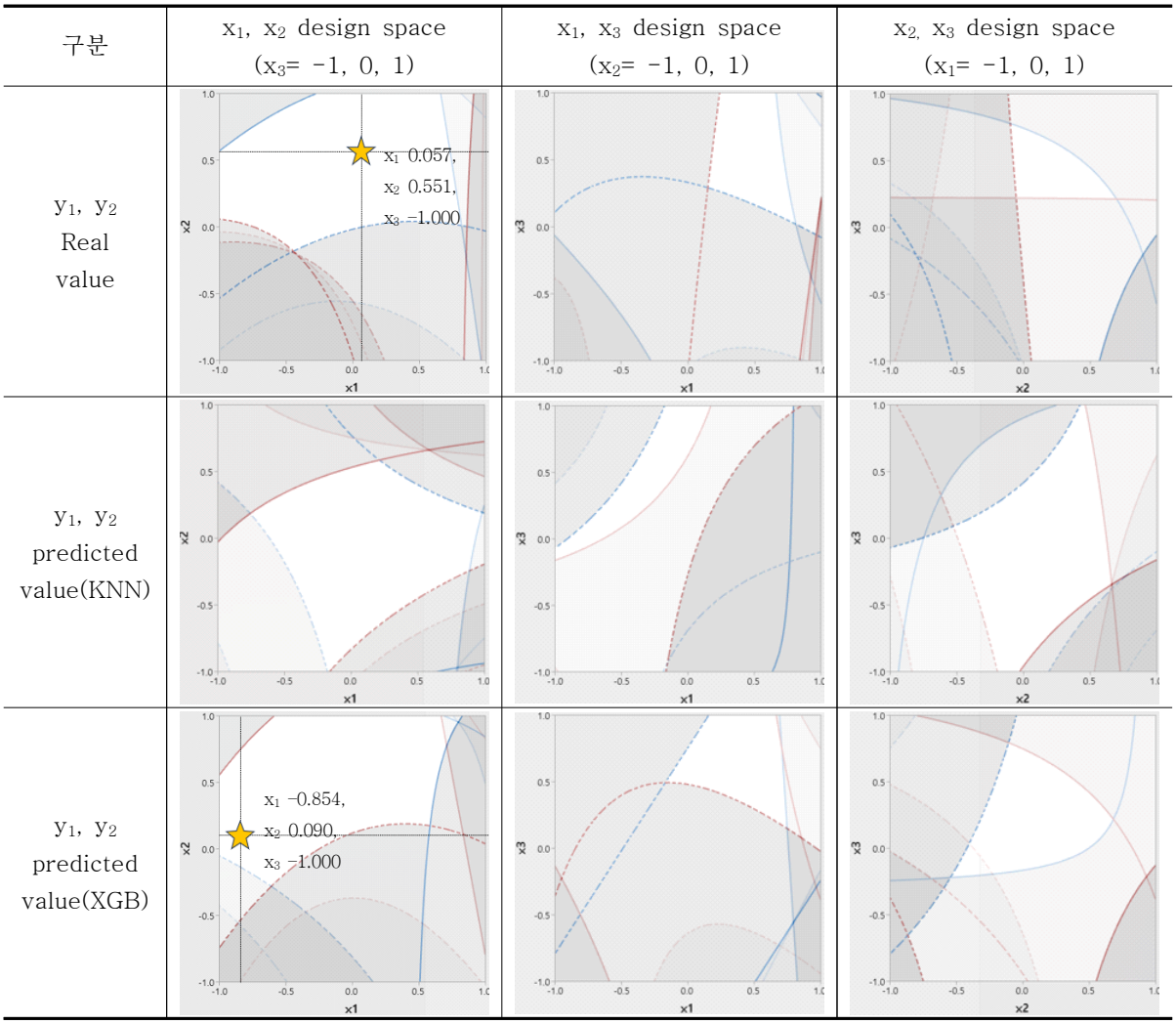

Figure 10. Comparison with design space between real value and predected value from k-NN and XGBoost

5. 결 론

본 연구는 실험계획에 따른 실험 결과값을 바탕으로 나머지 실험점에 대한 결과를 예측 및 분석하는 절차를 연구하였다. 실험계획법을 이용한 연구 주요 목표는 최소의 실험으로 최대의 정보를 확보하는 것이며 신뢰성 있는 최적값을 도출하는 것이다. 만약 실험 결과에 이상치 및 결측치가 발생할 때 실험을 재수행해야 하며 실험 횟수와 비용이 더 발생하기 때문에 소수의 결측치 및 이상치 결과에 대해서 통계적인 근거와 데이터에 알맞은 AI 알고리즘을 활용하여 데이터 보간을 수행하는 것이 연구개발 측면에서 유리하다고 볼 수 있다. 본 연구에서 제안하는 AI 알고리즘을 이용한 소수 실험점 결측치 예측 결과, k-NN보다는 XGBoost 알고리즘의 예측 성능이 우수함을 확인하였다. 따라서 재실험을 수행할 때 발생되는 비용 및 기간 측면에서 본 연구에서 제안한 방법론은 적절한 AI 알고리즘을 선택하여 결측치를 예측함으로써 비용과 기간의 부담을 감소시킬 수 있을 것이다. 또한, 실험계획법 바탕의 데이터에서 부분요인실험 결과만을 활용하여 나머지 절반의 실험점에 대한 결과를 예측하기 위해 데이터의 형태에 따라 AI 알고리즘

을 활용할 수 있음을 보여주었다. 기존 부분요인실험값을 바탕으로 완전요인실험점의 나머지 실험점과 중심점에 대한 XGBoost 알고리즘 예측값과 실측값을 비교분석한 결과 통계적으로 유의함을 확인하였다. 실측값과 XGBoost 예측값의 ANOVA 분석 결과 y_1에 대한 주요 영향인자는 x_2, x_3, 교호작용은 x_1*x_3, x_2*x_3로 동일하게 p값 0.05 이하로 나타났고, y_2에 대한 주요 영향인자 또한 x_1, x_2, 교호작용은 x_1*x_2가 p값 0.05 이하로 동일하게 유의하게 나타났고, y_1, y_2에 대한 디자인 스페이스 또한 유사하게 나타났다. 최적조건 또한 실측값, XGBoost 예측값 모두 디자인 스페이스 내에서 도출되었다, 기존의 통계적 절차를 기반으로 한 실험계획법에 AI 알고리즘을 결합하여 더욱 효율적인 실험계획이 될 수 있음을 보여주었다. 결론적으로 AI 알고리즘을 결합한 예측 실험계획 방법론 연구를 통해 향후 실험계획법을 활용하는 연구의 효율성을 향상시킬 수 있을 것으로 기대한다.

향후 연구 방향에서는 실험을 통해 얻은 최소한의 데이터를 기반으로 다양한 AI 알고리즘에 적합한 적정양의 학습 데이터를 체계적으로 증강시키는 연구가 필요할 것이다. 또한 단계적 실험계획시 다양한 AI 알고리즘을 통한 변수선정 및 검증 등의 의사결정을 수행하는 연구가 필요하며, 다양한 데이터 속성 및 구조에 따른 적절한 예측을 위한 AI 알고리즘 선정 연구가 진행되어 최적의 실험계획법 설계 및 분석 플랫폼을 개발할 필요가 있다.

REFERENCES

Chang, H. H. and Chen, Y. K. 2011. Neuro-genetic approach to optimize parameter design of dynamic multiresponse experiments. Applied Soft Computing 11(1):436-442.

Cheon, M., Choi, H., Park, J., Choi, H., Lee, D., and Lee, O. 2021. A Study on the traffic flow prediction through Catboost algorithm. Journal of the Korea Academia-Industrial cooperation Society 22(3):58-64.

Dasarathy, B. V. 1991. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Tutorial.

G. Hessler and K. Baringhaus. 2018. Artificial Intelligence in Drug Design. Molecules 23(10):2520.

Han, Y., Jeong J., Lee, S. and Shin, S. 2021. Study on the Parameter Optimization of the Heat Treatment Process Using the Design of Experiment (DoE). Korean Journal of Computational Design and Engineering 26(4): 273-284.

ICH Harmonised Tripartite Guideline. 2009. Pharmaceutical Development Q8 (R2). International conference of harmonisation.

J. Freiesleben, J. Keim and M. Grutsch. 2020. Machine learning and Design of Experiments: Alternative approaches or complementary methodologies for quality improvement?. Qual Reliab Engng Int. 36:1837-1848.

J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao. 2019. Applications of machine learning in drug discovery and development. Nature Review of Drug Discovery 18(6):463-477.

J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao. 2019. Applications of machine learning in drug discovery and development. Nature Review of Drug Discovery 18(6):463-477.

Jung, M. and Kwon, W. 2021. Present Status and Future of AI-based Drug Discovery. Journal of the Korea Institute of Information and Communication Engineering 25(12):1797-1808.

Jung, Y., Kang, T., Park, J., Cho, J., Hong, J., and Kang, S. 2024. Methodology for Variable Optimization in Injection Molding Process. J Korean Soc Qual Manag. 52(1):43-56

K. Mak and M. Pichika. 2019. Artificial intelligence in drug development: present status and future prospects. Drug Discovery Today 24(3):773-780.

Kim, H. 2020. The Prediction of PM2. 5 in Seoul through XGBoost ensemble. Journal of the Korean Data Analysis Society 22(4):1661-1671.

Kim, K., Kim, S. and Kim, Y. 2023. A Study on Optimization of Classification Performance through Fourier Transform and Image Augmentation. J Korean Soc Qual Manag. 51(1):119-129

Kim, S. 2016. The analysis of current performance and perception on QbD(Quality by Design) of pharmaceutical companies in south Korea. MS diss. Sungkyunkwan University.

L. Patel, T. Shukla, X. Huang, D. Ussery, and S. Wang. 2020. Machine Learning Methods in Drug Discovery. Molecules 25:5277.

Lee, H., Kim, Y. and Shin, S. 2020. Optimal Parameter Design for a Cryogenic Submerged Arc Welding(SAW) Process by Utilizing Stepwise Experimental Design and Multi-dimensional Design Space Analysis. J Korean Soc Qual Manag. 48(1):1-18.

Lee, S., Sim, J. and Choi, J. 2023. A Case Study on Quality Improvement of Electric Vehicle Hairpin Winding Motor Using Deep Learning AI Solution. J Korean Soc Qual Manag. 5(2):283-296

Lou, H., Chung, J. I., Kiang, Y. H., Xiao, L. Y., and Hageman, M. J. 2019. The application of machine learning algorithms in understanding the effect of core/shell technique on improving powder compactability. International Journal of Pharmaceutics 555:368-379.

M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. Koes. 2017. Protein-ligand scoring with convolutional neural etworks. Journal of Chemical Information and Modeling 57:942-957.

R. C. Mohs and N. H. Greig. 2017. Drug discovery and development: role of basic biological research. Alzheimer's & Dementia 3(4):651-657.

S. Woo. 2018. Drug Discovery Enhanced by Artificial Intelligence. Biomedical Journal of Scientific & Technical Research 12(1).

T. Chen, C and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System, Association for Computing Machinery. KDD'16, August:13-17.

T. H. Le and S. Shin. 2021. Structured neural network models to improve robust design solutions. Computers & Industrial Engineering 156:107231.

T. H. Le, H. Jang and S. Shin. 2021. Determination of the Optimal Neural Network Transfer Function for Response Surface Methodology and Robust Design. Appl. Sci. 11:6768.

T. H. Le. L. Dai, H. Jang and S. Shin. 2022. Robust Process Parameter Design Methodology: A New Estimation Approach by Using Feed-Forward Neural Network Structures and Machine Learning Algorithms. Appl. Sci. 12:2904.

Viana FA, Venter G and Balabanov V. 2010. An algorithm for fast optimal Latin hypercube design of experiments. Int J Numer Methods Eng. 82(2):135-156.

Won, C. 2018. Reflections on the importance of variables in k-NN. MS diss. Korea University.

저자소개

박경진 인제대학교에서 시스템경영공학과를 졸업하고 석사를 취득하였으며, 현재 동아대학교 산업경영공학과 박사과정 재학 중이다. 주요 관심분야는 실험계획법, 강건설계 및 최적화이다.

정제한 경남과학기술대학교 자동차공학과를 졸업하고, 동아대학교 산업경영공학과 석사를 취득하였다. 주요 관심분야는 실험계획법 및 머신러닝이다.

장준혁 현재 선박해양플랜트연구소 재직중이며, 주요 관심 분야는 실험계획법, 강건설계 및 최적화이다.

신상문 동아대학교 산업공학과를 졸업하고, 미국 Clemson University에서 산업공학 석사와 박사를 취득하였으며, 현재 동아대학교 산업경영공학과에서 교수로 재직중이다. 주요관심분야는 의약품 개발을 위한 QbD, 임상-전임상 PK/PD 분석, 실험계획법, 강건설계 등이다. 국제저널인 International Journal of Quality Engineering and Technology의 편집위원장으로도 활동하고 있다.