Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 1, pp. 41-54, February 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.1.041

배양 공정에서 세포 생존율과 항체 생산성을 예측하기 위한 설명가능한 딥러닝 모델

김민지[1] · 양지영[2] · 이채현[2] · 윤지용[3] · 오대양[3] · 엄주명[1,2†]

[1]경희대학교 인공지능학과 / [2]경희대학교 산업경영공학과 / [2]프레스티지바이오로직스(주)

Explainable Deep Learning Model for Predicting Cell Viability and Antibody Productivity in Culture Processes

Minji Kim[1] · Jiyoung Yang[2] · Chaehyeon Lee[2] · Ji Yong Yoon[3] · Dae Yang Oh[3] · Jumyung Um[1,2]

[1]Department of Artificial Intelligence, Kyung Hee University
[2]Department of Industrial and Management Systems Engineering, Kyung Hee University

This paper proposes a machine learning framework for predicting key in antibody production in cell culture processes, focusing on antibody product concentration and cell viability. Three types of input data-process variables, output variables and their integration-were applied to Random Forest and Deep Neural Network models. Experimental results demonstrated that integrated datasets improved antibody product concentration prediction accuracy, while output variables were more effective for cell viability prediction. Shapley additive explanations analysis was employed to interpret variable contributions, revealing distinct drivers for each quality attribute. For antibody product concentration, effect of metabolic indicators such as glutamate and ammonia were dominant, while viability was mainly influenced by culture conditions and glutamine levels. These findings highlight that input composition strategies should be tailored to the target attribute and that combining prediction accuracy with interpretability provides actionable insights for process optimization. The proposed approach contributes to smart biomanufacturing and supports digital twin-based control strategies.

Keywords: Explainable AI, Biomanufacturing, Cell Culture Processes, Quality Prediction

1. 서 론

항체 기반 바이오의약품은 고난이도 치료 영역에서 핵심적인 역할을 수행하고 있으며, 전 세계적으로 그 수요가 지속적으로 증가하고 있다(Kothari *et al.*, 2024; Mekala *et al.*, 2024). 특히 면역질환, 암, 감염병 등의 치료에 효과적인 단클론 항체는 높은 생산성과 안정된 품질 특성을 요구하는 산업적 중요성을 지닌다(Ranbhor, 2025). 하지만 항체의 대량 생산 공정은 살아있는 세포를 기반으로 하기 때문에, 공정 조건, 배지 조성, 배양 기간 등 다양한 요소에 의해 생산성과 품질이 민감하게 변화한다(Birch *et al.*, 2006; Lee *et al.*, 2025; Zhang *et al.*, 2024). 이러한 생물학적 복잡성으로 인해 바이오의약품 제조 공정은 기존의 화학 기반 공정에 비해 제어와 예측이 어렵고, 생산성의 일관성을 확보하기 위해서는 보다 정밀한 공정 분석 및 예측 체계가 요구된다.

최근 Machine Learning(ML) 기반의 접근방식이 이러한 문제 해결의 대안으로 주목받고 있다. 특히 Deep Learning(DL)은 비선형적이며 고차원적인 공정 변수 간의 상관관계를 효과적으

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (RS-2025-02219563, Step by Step Autonomous Manufacturing: From Dataset Construction to Generative AI Utilization)
† 연락저자 : 엄주명 교수, 17104, 경기도 용인시 덕영대로 1732 경희대학교 국제캠퍼스, Tel : 031-201-5443, E-mail: jayum@khu.ac.kr
2025년 10월 28일 접수; 2025년 12월 18일 수정본 접수; 2026년 1월 15일 게재 확정.

로 모델링할 수 있는 장점이 있다(Baako *et al.*, 2024; Helleckes *et al.*, 2023). 더불어, Shapley Additive Explanations(SHAP)와 같은 설명 가능한 인공지능(Explainable AI, XAI) 기법은 예측 모델의 해석력을 확보함으로써 실제 공정 제어 인자 발굴 및 의사결정에 기여할 수 있다(Li *et al.*, 2024; Winter, 2002).

바이오의약품 제조에서는 일반적으로 공정 입력 변수(예: 배양 시간, 온도, pH, 용존 산소 등)와 함께 세포 배양 중에 측정 가능한 산출 지표(예: glucose, lactate, viability, osmolality 등) 데이터가 함께 축적된다(Reyes *et al.*, 2024). 하지만 기존 연구들은 주로 공정 입력 변수에만 기반한 품질 예측 모델 개발에 집중되어 있으며, 반응 산출 지표까지 통합한 예측 모델에 대한 체계적인 접근은 부족한 실정이다(Pham *et al.*, 2023; Mondal *et al.*, 2023). 또한, 예측 모델의 성능이 우수하더라도, 각 입력 변수가 품질 결과에 어떤 방식으로 기여했는지를 설명하기 어려운 '블랙박스' 문제로 인해 산업 현장에서의 활용에 제약이 존재한다(Lim *et al.*, 2023; Medl *et al.*, 2024).

이에 본 연구는 세포 배양 기반 항체 생산 공정에서 대표적인 생산성 지표인 항체 농도인 IgG와 세포 생존율인 Viability을 대상으로, 공정 입력 변수와 반응 산출 지표를 통합한 머신러닝 기반 생산성 예측 모델을 개발하고, SHAP 기법을 통해 각 변수의 영향력을 해석하고자 한다. 구체적으로는 (1) 공정 변수 기반, (2) 산출 지표 기반, (3) 두 요소를 통합한 세 가지 입력 조합에 대해 각각 Random Forest (RF)와 Deep Neural Network (DNN) 모델을 학습하고 성능을 비교하였다. 또한, SHAP 분석을 통해 변수별 영향도 및 방향성을 시각화함으로써, 생산성 특성에 대한 이해를 높이고 제어 변수 도출에 실질적인 인사이트를 제공하고자 하였다. 본 연구는 예측 정확도와 해석 가능성을 동시에 고려한 통합 모델을 제시함으로써, 향후 스마트 바이오제조와 디지털 트윈 기반 공정 최적화 체계 구축에 기여할 수 있을 것으로 기대된다.

본 논문은 제2장 관련연구, 제3장 방법론, 제4장 결과, 제5장 토론과 제6장 결론으로 마무리 된다.

2. 관련연구

2.1 바이오 의약품 공정 제어 및 품질 예측

바이오의약품 생산 공정은 세포 배양, 대사 작용, 스트레스 반응 등 복잡한 생물학적 상관관계가 존재하여, 품질 예측 및 공정 제어가 매우 어렵다. 최근 연구들은 비침습 측정 기술, 센서 데이터, 실시간 분석 기법을 공정 제어와 품질 예측에 접목하고 있다. Nik-Khorasani *et al.*(2025)은 항체 공정 전반에 머신러닝 적용 가능성을 제시하였으며, 특히 비선형 관계와 변수 간 상호작용을 다룰 수 있는 모델의 중요성을 강조하였다. Lai *et al.*(2022)은 항체 단백질의 응집 및 점도를 예측하기 위해 ML 모델을 도입하여 고농도 조건에서의 안정성 예측이 가능함을 보여주었다. Narayanan *et al.*(2021)은 바이오의약품 배합 설계 단계를 Bayesian 최적화 기법으로 가속화한 연구를 통해, 최적 배합 조건 탐색의 비용 절감 가능성을 실증하였다. Wossnig *et al.*(2024)은 항체 개발 과정에서의 머신러닝 최적 활용 전략을 다룬 리뷰를 통해, 전처리, 데이터 불균형, 모델 일반화 전략 등이 중요하다는 점을 강조하였다. Makowski *et al.*(2022)은 복잡한 항체 공학 설계에 있어 머신러닝을 활용한 모델링 접근을 제시하며, 설계 공간 탐색 속도를 크게 개선한 사례를 보고하였다. Joubbi *et al.*(2024)은 항체 서열 및 구조 수준에서의 딥러닝 기반 설계 접근을 정리한 리뷰를 통해, 단백질 디자인 측면에서도 ML의 가능성을 제시하였다.

이러한 연구들은 본 논문의 방향과 밀접하게 연결된다. 본 연구는 IgG와 Viability를 예측 대상으로 설정하고, 단순 입력 변수 중심 방식만이 아니라 생산성 지표를 통합하는 예측 구조를 제안함으로써 기존 연구의 한계를 보완하고자 한다.

2.2 머신러닝 기반 공정 예측 및 최적화

공정 데이터를 활용한 예측 및 최적화는 제조업 전반에서 활발히 연구되고 있으며, 바이오 분야에서도 그 응용이 증가하고 있다. Richter *et al.*(2025)은 CHO 세포 기반 배양 공정에 ANN 기반 딥러닝 모델을 적용해 생산량 예측 및 조건 최적화 연구를 수행했으며, 생산성을 최대 48% 향상시켰다. Pinto *et al.*(2023)은 CHO 세포 fed-batch 공정에 하이브리드 모델을 적용한 사례를 다루며, 전통적 모델과 딥러닝 기반 구성의 조합이 예측성능 향상 및 과적합 감소에 기여할 수 있음을 보여주었다. Duong-Trung *et al.*(2023)은 다양한 머신러닝 기법을 공정 데이터 분석 및 품질 예측에 활용하는 경향을 정리하고 있다. 특히 교차 검증, 과적합 제어, 해석 가능성 확보 전략의 중요성이 강조되고 있다. 일부 연구는 하이퍼파라미터 튜닝, 앙상블 모델, 교차 학습 방법을 적용하여 예측의 강건성을 향상시킨 사례를 보고했다(Khuat *et al.*, 2025; Yatipanthalawa *et al.*, 2024). 또한 Explainable AI(XAI)를 도입해 단순 예측을 넘어 변수 영향을 해석할 수 있는 예측 체계를 구축한 연구도 증가하고 있다(Di Bonito *et al.*, 2024). 본 연구에서도 RF 및 DNN 모델을 활용하며 SHAP 기반 해석을 도입하여, 예측 모델의 투명성과 중요 변수를 찾는 방법론을 채택했다.

2.3 설명 가능한 인공지능 기반 공정 해석과 디지털 트윈 응용

스마트 제조, Industry 4.0 흐름 속에서 디지털 트윈은 바이오 공정에서도 점차 핵심 기술로 부상하고 있다. Huang(2023) 은 디지털 트윈을 활용한 공정 제어 전략 개발과 자동화 구현 사례를 제시하며, 전통 제어 시스템(MPC, 피드백 제어 등)과의 융합 가능성을 탐색하였다. Shahaba *et al.*(2025) 은 디지털 트윈 내 인간-기계 협업 인텔리전스 개념을 강조하며, 디지털 트

윈이 단순 시뮬레이터를 넘어 운용자와 상호 작용하는 플랫폼으로 발전해야 한다고 주장했다. Isoko *et al.*(2024)은 Process analytical technology(PAT), 스마트 센서, 실시간 분석, 디지털 트윈 연계 기술 등이 바이오제조 업계의 미래 방향으로 제시되고 있다.

Rosen *et al.*(2015)은 디지털 트윈이 물리 시스템의 상태를 실시간으로 반영하고, 자율 제어 및 의사결정 과정에 활용되기 위해서는 모델의 신뢰성과 해석 가능성이 필수적임을 강조하였다. Tao *et al.*(2018)은 디지털 트윈과 사이버-물리 시스템의 결합을 통해 공정 상태의 해석, 피드백 제어, 품질 예측이 통합적으로 수행될 수 있음을 제시하였다. 디지털 트윈 기반 시스템이 공정 데이터를 단순히 처리하는 데 그치지 않고, 결과의 의미를 해석하고 이를 제어 및 의사결정으로 연결하는 역할을 수행해야 함을 강조하였다. 이러한 관점에서 모델 출력의 설명 가능성은 디지털 트윈 기반 피드백 구조의 신뢰성을 확보하는 핵심 요소로 작용한다.

한편, Samek *et al.*(2017)은 딥러닝 기반 모델의 의사결정 과정을 이해하고 시각화하기 위한 XAI 기법들을 체계적으로 정리하며, 복잡한 데이터 기반 모델이 실제 산업 환경에 적용되기 위해서는 해석 가능성이 반드시 수반되어야 함을 지적하였다. 이와 같은 XAI 접근법은 디지털 트윈 환경에서 활용되는 예측 및 제어 모델의 결과를 인간 운영자가 이해하고 검증할 수 있도록 지원함으로써, 디지털 트윈 기반 의사결정 시스템의 실용성과 신뢰성을 강화하는 데 기여할 수 있다.

디지털 트윈은 단순 예측이나 시뮬레이션을 넘어서, 실시간 공정 감시, 피드백 제어, 품질 예측, 비정상 탐지 등 복합 기능을 통합할 수 있는 방향으로 발전하고 있다. 본 연구의 예측과 해석 구조는 디지털 트윈의 핵심 모듈로 기능할 가능성을 가진다.

2.4 본 연구의 차별성 및 기여

기존의 많은 연구들이 공정 변수 또는 산출 지표와 같은 단일 입력 범주 기반 예측에 집중하거나, 단순한 제어 및 시뮬레이션 중심 접근에 머무르는 경향이 있다. 본 연구는 이러한 한계를 극복하기 위해 다음과 같은 차별점을 가진다: 공정 입력 (배양 조건 등) 및 산출 지표(대사 물질, 세포 상태 등)를 동시에 통합한 입력 구조를 구성하였다. 예측 모델의 해석 가능성 확보를 위해 SHAP 기법을 도입하여, 변수별 기여도와 방향성을 정량적·시각적으로 분석하였다. IgG 및 Viability 두 생산성 지표를 함께 다룸으로써, 지표별 특성 별로 입력 구성 전략이 달라질 수 있음을 제시하였다. SHAP 분석 결과를 바탕으로 공정 변수가 모델에 미치는 영향을 함께 제시하여, 단순 예측을 넘어 실질적으로 공정에 기여하는 변수를 찾았다. 이는 향후 스마트팩토리 시스템에 적용 가능한 예측 및 해석 구조를 제안했다는 점에서 학문적 그리고 산업적 응용 가능성이 크다.

3. 방법론

본 연구에서 사용된 실험 데이터는 관절염 치료용 단클론 항체를 생산하는 바이오팩토리 공정에서 수집되었다. 해당 항체는 Humira(Adalimumab)의 바이오시밀러 파이프라인인 PBP1502로, 실험실 규모 반응기에서 수행된 세포 배양 공정을 통해 생성되었다. 데이터는 3L 및 15L 규모의 Thermo Fisher Scientific사의 HyPerforma G3Lab Controller 기반 바이오리액터에서 획득되었으며, 서로 다른 공정 조건 하에서 수행된 다수의 배치 실험을 포함한다. 구체적으로 C2, C3, V1, V2, V3 실험은 3L 반응기에서, V4 실험은 15L 반응기에서 수행되었다. 본 연구는 해당 관절염 치료 항체 생산 공정을 대표적인 case study로 설정하여, 실제 바이오의약품 제조 환경에서의 생산성 예측 및 해석 가능성 검증을 목적으로 한다. <Figure 1>은 case study의 개요를 나타낸다.

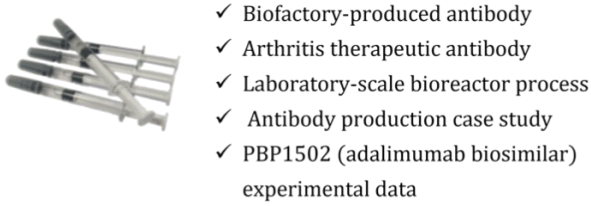


Figure 1. Case Study Overview of Monoclonal Antibody Production for Arthritis Treatment

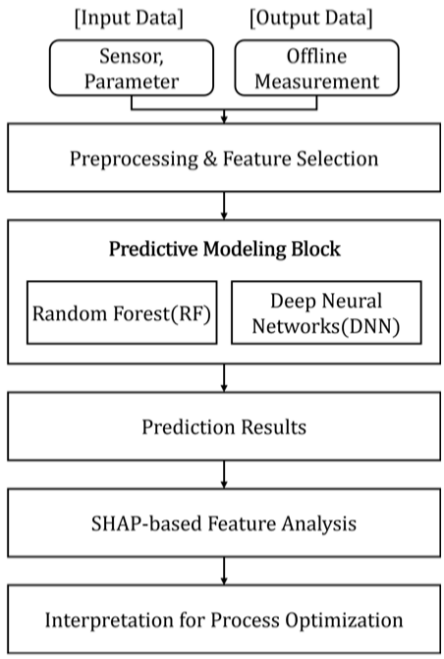


Figure 2. System Architecture

또한 본 연구에서 제안하는 시스템 아키텍처는 다음과 같다. <Figure 2>는 바이오공정 데이터를 활용한 예측 및 최적화

절차를 단계적으로 보여준다. 먼저 센서 및 공정 파라미터와 같은 입력 데이터와 오프라인 측정을 통해 얻은 출력 데이터가 수집되며, 이를 바탕으로 전처리와 특징 선택 과정을 거쳐 예측 모델링에 적합한 데이터셋이 구축된다. 이후 예측 모델링 블록에서는 RF와 DNN과 같은 다양한 ML 기법이 적용되어 공정 특성과 생산성을 예측한다. 모델의 결과는 단순히 예측 값 제공에 그치지 않고, XAI 기반의 특성 분석을 통해 각 입력 변수가 예측 결과에 미친 영향을 정량적으로 해석할 수 있도록 한다. 최종적으로 이러한 해석 과정은 공정 내 주요 인자를 식별하고, 이를 기반으로 최적화 전략을 도출하여 실질적인 공정 개선과 제어에 기여하는 의사결정 지원 체계로 활용된다. 이에 대한 상세한 방법론은 본 장에서 다룰 예정이다.

3.1 데이터 전처리 및 입력 변수 구성

본 연구에서는 바이오 의약품 배양 공정에서 수집된 데이터를 기반으로 생산성 지표 예측을 위한 학습 데이터를 구성하고, 분석 목적에 맞게 전처리 과정을 수행하였다. 사용된 데이터셋은 서로 다른 공정 조건에서 수행된 총 40개의 항체 생산 배치 실험으로부터 수집되었다. 각 배치 실험은 배양 시간에 따라 공정 변수 및 산출 지표가 측정되었으며, 원시 데이터 기준으로 총 384개의 row로 구성되어 있다. 그러나 배치별 샘플링 주기 및 측정 시점이 상이하고 일부 결측 구간이 존재하여, 기계학습 모델 학습을 위한 일관된 시간 축 데이터 구성이 필요하였다.

이에 따라 배양 시간 축을 기준으로 보간을 적용하여 데이터를 정렬하였으며, 해당 전처리 과정을 통해 최종적으로 11,375개의 데이터 포인트로 구성된 학습 데이터셋을 구축하였다. 본 연구에서의 보간은 데이터 증강을 목적으로 하지 않으며, 불균일한 시간 간격과 결측치를 보완하기 위한 전처리 단계이다. 또한 보간은 입력 변수에 대해서만 적용하였고, IgG와 Viability과 같은 예측 대상 변수는 실측값을 그대로 사용하여 정보 누수를 방지하였다.

데이터는 크게 공정 변수(process variables)와 산출 지표(output variables)으로 구분된다. 공정 변수는 배양 시간(Day), 초기 세포 밀도(Initial VCD), 설정 온도(target temperature), pH, DO, 교반 속도, Feeding 정보 등 생산 조건을 나타내며, 산출 지표는 실험 종료 후 측정된 IgG와 Viability로 구성된다. 예측 성능 비교를 위해 다음과 같이 세 가지 유형의 데이터셋을 구성하였다. 첫째, 공정 변수 기반 데이터셋은 공정 조건만을 포함하며, 실제 제조 시점에서의 생산성 예측 가능성을 평가하기 위해 사용되었다. 둘째, 산출 지표 기반 데이터셋은 실험 도중 또는 종료 직전에 획득 가능한 결과 데이터를 입력으로 활용한 형태로, 공정의 상태 평가 및 최종 생산성 추정을 목적으로 한다. 셋째, 통합 데이터셋은 공정 변수와 산출 지표를 모두 포함하여, 입력 정보의 복합성을 높인 형태로 구성하였다. 이를 통해 다양한 정보 조합이 예측 정확도와 해석 가능성에 미치는 영향을 분석하고자 하였다. 전처리 과정에서는 데이터 누락 항목에 대한 보간, 범주형 변수에 대한 원핫 인코딩, 그리고 정규화를 수행하였다. 특히 정규화는 DNN 기반 예측 모델에서만 적용되었으며, 예측 결과에 대해서는 역정규화 과정을 통해 해석 가능성을 확보하였다. 최종적으로 csv를 기반으로 실험에 사용될 세 가지 데이터셋을 구축하였다.

3.2 Random Forest 및 Deep Neural Network를 이용한 예측 모델링

본 연구에서는 바이오 의약품 공정 데이터를 기반으로 생산성 지표 예측 모델을 구축하기 위해, 머신러닝 기반의 RF 회귀 모델과 DNN을 활용하였다. 두 모델 모두 IgG와 Viability을 예측 대상으로 설정하였으며, 학습에는 전처리된 세 가지 유형의 데이터셋(공정 변수 기반, 산출 지표 기반, 통합형) 을 각각 독립적으로 적용하였다.

RF는 결정 트리 기반의 앙상블 학습 기법으로, 과적합에 대한 내성이 강하고 변수 간 비선형 관계에 대한 해석이 용이하다는 장점을 지닌다. 본 연구에서는 학습 성능 향상을 위해 n_estimators와 max depth를 중심으로 그리드 탐색을 수행하였으며, 예측 정확도와 일반화 성능을 균형 있게 고려하였다. 모델 학습 시 정규화를 적용하지 않았으며, 변수 중요도 기반 해석을 위한 SHAP 분석이 가능하도록 모델 구조를 유지하였다. n_estimators는 {50, 100, 200}, max_depth는 {None, 10, 20}의 범위로 설정하였으며, 각 조합에 대해 학습 데이터를 이용해 모델을 학습한 후 검증 데이터셋에서의 평균제곱오차(Mean squared error, MSE)를 기준으로 최적의 모델을 선택하였다. RF 모델의 특성을 고려하여 입력 변수에 대한 정규화는 적용하지 않았다. <Table 1>은 RF의 하이퍼파라미터를 나타내는 표이다.

Table 1. Hyperparameter Search Space for Random Forest

Parameter	Search range
n_estimators	50, 100, 200
max_depth	None, 10, 20
Normalization	Not applied

DNN은 Multilayer Perceptron 기반으로 구성하였으며, 각 은닉층은 64~128개의 노드를 포함하고 Rectified Linear Unit(ReLU) 활성화 함수를 적용하였다. 과적합 방지를 위해 드롭아웃(Dropout)을 각 은닉층에 도입하였으며, 옵티마이저는 Adam을 사용하였다. 학습률은 0.001로 설정하였으며, 검증 성능이 일정 epoch 이상 개선되지 않을 경우 학습을 조기 종료하는 Early stopping 전략을 병행하였다. 모든 입력 변수는 min-max scaling를 거쳤으며, 예측값에 대해서는 역정규화를

적용하여 원 단위 해석이 가능하도록 처리하였다. <Table 2>는 DNN의 하이퍼파라미터를 나타내는 표이다.

Table 2. Hyperparameter for DNN

Parameter	Search range
Layer	128-64-32
Activation Fuction	ReLU
Optimizer	Adam
Learning rate	0.001
Normalization	Min-max scaling

모델의 성능 평가는 다음의 세 가지 지표를 기반으로 수행하였다. 평균절대오차(Mean absolute error, MAE), 평균제곱오차(Mean squared error, MSE), 평균 제곱근 오차(Root mean squared error, RMSE) 예측값과 실제값 간의 절대적 오차 크기를 측정하는 지표이다. 데이터셋은 학습, 검증Validation, 테스트Test 세 집합으로 분할하였다. 각 집합은 6:2:2의 비율로 구성하였으며, 가장 최신에 수행된 실험 데이터를 테스트셋에 포함하여 예측 모델의 실효성을 평가하고자 하였다.

3.3 변수 중요도 분석을 통한 모델 해석

ML 기반 예측 모델은 높은 예측 정확도를 확보할 수 있다는 장점이 있으나, 대부분의 비선형 모델은 내부 작동 원리가 불투명하여 결과 해석에 제약이 따른다. 특히 바이오 의약품 제조 공정과 같이 변수 간 상호작용이 복잡한 시스템에서는, 단순한 예측값 제공을 넘어 각 변수의 영향력을 정량적으로 파악하는 것이 공정 최적화 및 제어 전략 수립에 중요한 단서를 제공할 수 있다.

이에 따라 본 연구에서는 예측 모델의 해석 가능성을 확보하고, 주요 변수들이 생산성 지표에 미치는 영향을 분석하기 위해 SHAP(SHapley Additive exPlanations) 기법을 활용하였다. SHAP은 게임 이론 기반의 해석 방법으로, 각 변수(feature)가 모델의 예측 결과에 기여한 정도를 샘플 단위로 산출한다. 특히 SHAP 값은 모델의 구조와 관계없이 일관된 기준으로 변수 기여도를 비교할 수 있다는 점에서, 랜덤 포레스트와 DNN 모델 모두에 적용 가능하다.

본 연구에서는 예측 모델의 유형에 따라 SHAP 값 계산 방식을 구분하여 적용하였다. RF에 대해서는 TreeExplainer를 사용하여 SHAP 값을 계산하였으며, DNN에 대해서는 DeepExplainer를 적용하였다. 이는 각 모델 구조의 특성을 반영하여 보다 정확한 변수 기여도 해석을 수행하기 위한 것으로, SHAP 공식 가이드라인 및 기존 연구에서 권장되는 접근 방식이다. SHAP 분석 결과는 각 모델 내부에서의 변수 중요도 및 영향 방향성을 해석하는 용도로 활용되었으며, 서로 다른 모델 간 SHAP 값의 절대적 비교는 수행하지 않았다.

SHAP 값 계산에는 학습 데이터 분포를 반영하기 위해 학습 데이터셋을 기반으로 explainer를 구성하였으며, test set 중 일부 샘플을 대상으로 SHAP 값을 산출하고, beeswarm plot을 통해 변수 중요도와 영향 방향성을 시각화하였다.

분석에는 beeswarm plot을 활용하였으며, 이는 SHAP 값의 분포를 통해 변수의 상대적 중요도, 영향 방향성, 샘플 간 분산을 시각적으로 표현할 수 있는 방식이다. Y축에는 모델에 입력된 변수들이 중요도 순으로 정렬되어 있으며, X축은 각 샘플의 SHAP 값을 나타낸다. 점의 색상은 해당 변수의 실제 값 크기를 의미하므로, 변수 값의 크기에 따른 영향력 차이도 함께 확인할 수 있다.

SHAP 분석은 예측 대상에 따라 구분하여 수행되었으며, IgG와 Viability 각각에 대해 공정 변수 기반, 산출 지표 기반, 통합형 데이터셋을 입력으로 사용한 모델에 대해 독립적으로 적용하였다. 이를 통해 각 데이터 유형별로 중요하게 작용하는 변수들을 식별하고, 생산성 지표에 영향을 미치는 핵심 인자를 도출하였다. SHAP 값은 각 모델 내부에서의 변수 기여도를 해석하기 위한 용도로 사용되었으며, 서로 다른 모델 간 SHAP 값의 절대적인 크기는 비교하지 않았다.

분석 결과는 후속 장에서 구체적으로 제시되며, 해당 결과를 통해 다음과 같은 해석이 가능하다. 첫째, IgG 예측에서는 산출 지표 기반 변수들의 영향력이 상대적으로 높게 나타났으며, 이는 항체 생성량이 실시간 공정 조건보다는 반응 산물에 더 직접적으로 연동된다는 점을 시사한다. 둘째, Viability 예측에서는 공정 변수만으로도 높은 설명력이 확보되었으며, 이는 초기 배양 조건과 환경 조절 요소들이 세포 생존에 결정적인 역할을 한다는 것을 의미한다. 셋째, Feeding 관련 변수들은 두 생산성 지표 모두에 대해 중간 수준의 영향력을 보이며, 생산성과 생존율 간의 균형을 조절할 수 있는 제어 인자로 활용 가능성이 확인되었다.

이와 같은 해석 기반 접근은 예측 모델의 단순 성능 비교를 넘어, 실제 바이오공정에서 적용 가능한 제어 전략 수립, 디지털 트윈 시뮬레이션 연계, 고위험 조건 사전 탐지 등 다양한 공정 혁신 방안으로 확장될 수 있다.

4. 결과

4.1 IgG 예측결과 분석

IgG는 바이오 의약품의 주요 생산성 지표 중 하나로, 세포의 성장 및 대사에 따른 누적적 반응 결과로 나타난다. 본 연구에서는 공정 변수 기반, 산출 지표 기반, 통합형 세 가지 데이터셋을 각각 입력으로 활용하여 IgG를 예측하고, 실제 측정값과 비교함으로써 각 데이터 구성 방식의 적합성을 검토하였다.

<Table 3>의 결과는 RF와 DNN 모델을 활용해 IgG를 예측한 성능을 비교한 것이다. 전반적으로 DNN이 RF보다 낮은 MAE,

Table 3. IgG Prediction Results of RF and DNN Models by Input Data

Model	Input data	MAE	MSE	RMSE
RF	Process Variable Data	0.3778	0.2617	0.5115
DNN	**Process Variable Data**	**0.3104**	**0.1338**	**0.3659**
RF	Output Variable Data	0.3720	0.2529	0.5029
DNN	**Output Variable Data**	**0.3160**	**0.1433**	**0.3785**
RF	Process Variable and Output Variable Data	0.3994	0.2631	0.5129
DNN	**Process Variable and Output Variable Data**	**0.2174**	**0.0798**	**0.2824**

MSE, RMSE 값을 보여 더 우수한 예측 성능을 나타냈으며, 특히 공정변수와 산출 지표를 통합한 경우 DNN 의 성능이 가장 뛰어나 MAE 0.2174, MSE 0.0798, RMSE 0.2824로 관측되었다. 이는 공정변수 또는 산출 지표와 같은 단일 입력 데이터보다 통합 데이터를 활용했을 때 예측력이 향상됨을 의미하며, 복잡한 데이터 패턴을 학습할 수 있는 DNN의 특성이 공정 최적화와 생산성 예측에 효과적으로 기여할 수 있음을 보여준다.

<Figure 3>은 공정 변수만을 입력으로 활용한 경우의 예측 결과이며, 전반적인 추세는 일정 수준 유지되었으나, 실제 데이터의 곡선 형태를 따라가지 못하거나 피크 구간에서의 오차가 크게 나타나는 구간이 관찰되었다. 이는 배양 조건만으로는 항체 생성의 복잡한 동역학을 충분히 설명하기 어렵다는 한계를 시사한다.

<Figure 4>는 산출 지표만을 입력 변수로 사용한 예측 결과로, 예측 선이 실제 관측치와 유사한 흐름을 따랐다. 특히, 각 배양 주기의 급격한 변화 구간에서도 예측값이 빠르게 반응하며, 모델이 산출 기반 변수들(예: TCD, Glutamate 등)의 내재적 상관관계를 효과적으로 학습했음을 확인할 수 있었다. 그러나 일부 구간에서는 과도하게 진폭이 확대되어 실제 값을 초과하는 경향도 나타났는데, 이는 결과 기반 변수 간 중복 정보 또는 과적합 가능성으로 해석될 수 있다.

<Figure 5>는 공정 변수와 산출 지표를 통합한 형태의 데이터셋을 기반으로 예측한 결과이며, 가장 실제 측정치에 근접한 예측선을 나타냈다. 모델은 각 실험군에서의 증감 패턴을 정밀하게 추적하였으며, 특히 반복되는 배양 주기의 형태, 상승 곡선의 기울기, 피크 시점에서의 예측 안정성이 향상되었음을 확인할 수 있다. 이는 통합된 입력 구성이 항체 생성 과정의 입력 조건과 출력 반응 간 관계를 보다 정교하게 포착할 수 있도록 함을 의미한다.

세 가지 데이터셋 구성 방식 간 비교를 통해, IgG 예측에는 산출 지표의 반영이 성능 향상에 핵심적인 역할을 함을 확인할 수 있었으며, 공정 변수와의 통합적 활용은 예측 정확도와 해석 가능성 모두를 확보하는 데 효과적인 접근임을 시사한다. 또한 IgG 예측 성능이 통합 데이터 구성에서 향상된 결과는 단순히 입력 변수의 개수가 증가했기 때문만으로 설명되기 어렵다. SHAP 분석 결과에 따르면, IgG 예측에는 Day, 대사 관

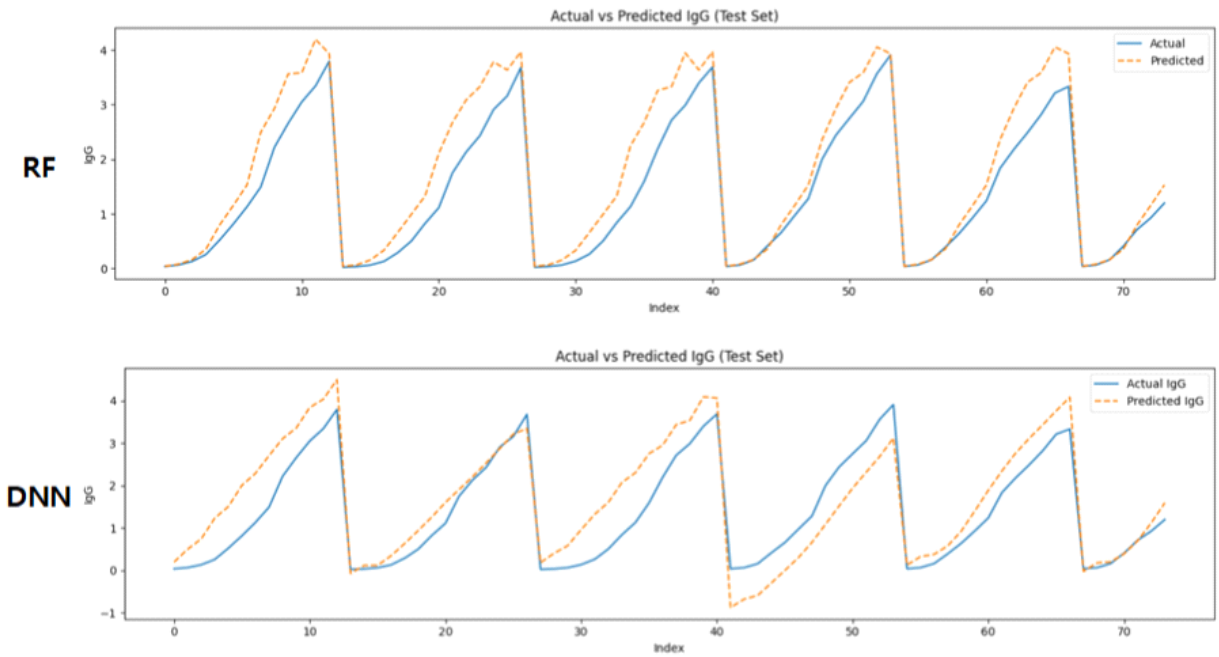


Figure 3. Visualization Results of IgG Prediction Using RF and DNN Models Trained on Process Variable Data

련 변수, 그리고 일부 공정 조건 변수들이 복합적으로 기여하고 있으며, 특정 변수 하나에 의존하기보다는 여러 변수들의 상호작용을 통해 예측 성능이 향상되는 경향을 보였다. 이는 IgG가 항체 생산 공정 전반의 누적적인 공정 상태를 반영하는 지표임을 시사하며, 다양한 입력 변수의 통합이 예측 정확도 향상에 실질적으로 기여하고 있음을 의미한다.

본 연구에서 보고된 IgG의 MAE, RMSE 결과는 공정 운용 중 항체 생산성의 절대 값을 정확히 대체하기 위한 수준을 목표로 하지는 않는다. 바이오 배양 공정에서는 배치 간 생산성 변동성이 크기 때문에, 공정 의사결정 관점에서는 정확한 수치 예측보다는 생산성 변화 추세 및 상대적 비교 정보가 더욱 중요하다. 이러한 관점에서 본 연구의 예측 성능은 배치 간 IgG 생산성 차이의 조기 인지, 공정 조건 변경 여부 판단, 생산성 저하 가능성 탐지를 위한 의사결정 보조 도구로 활용 가능

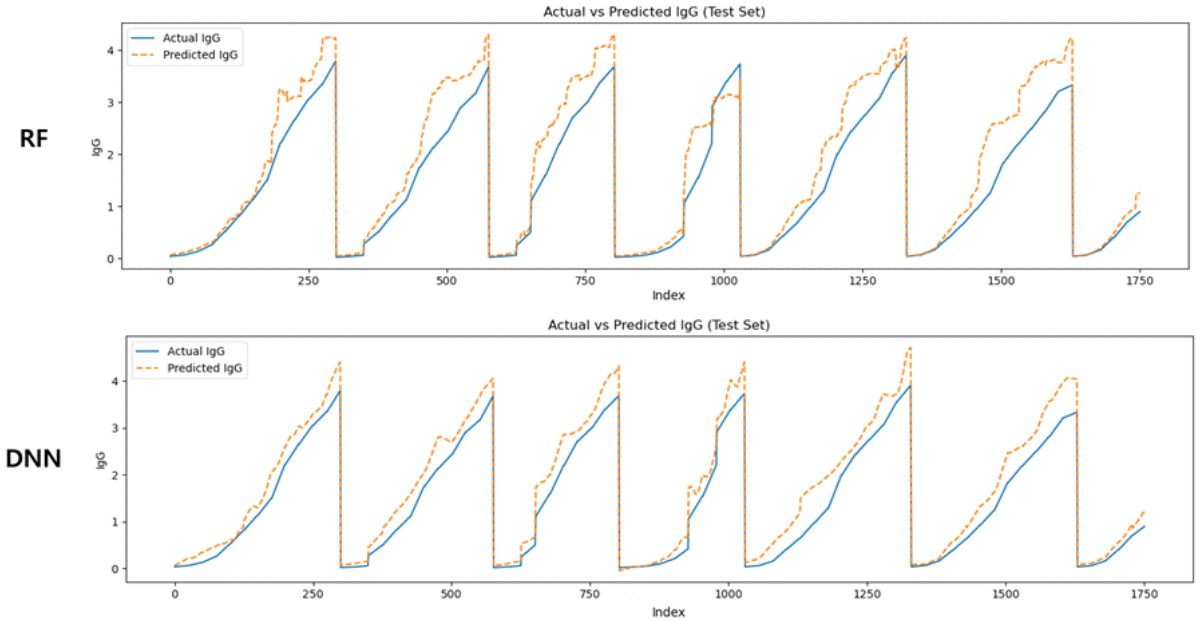


Figure 4. Visualization Results of IgG Prediction Using RF and DNN Models Trained on Output Variable Data

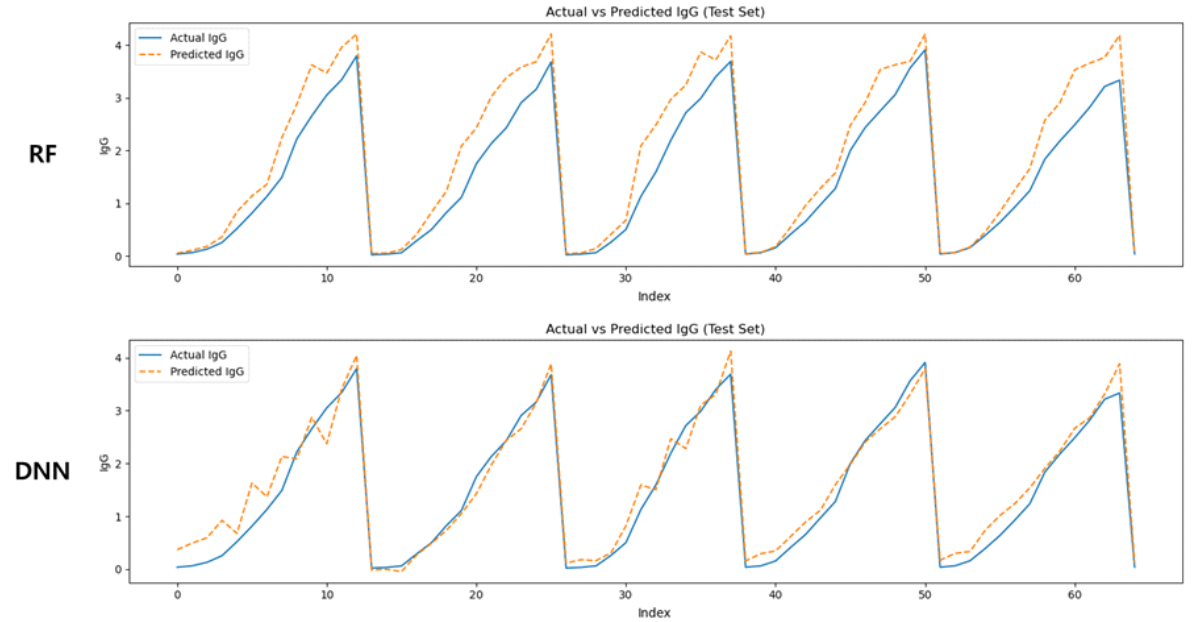


Figure 5. Visualization Results of IgG Prediction Using RF and DNN Models Trained on Integrated Process Variable and Output Variable Data

Table 4. Viability Prediction Results of RF and DNN Models by Input Data

Model	Input data	MAE	MSE	RMSE
RF	Process Variable Data	1.9620	4.8646	2.2056
DNN	**Process Variable Data**	**1.4841**	**2.6482**	**1.6273**
RF	Output Variable Data	1.7125	3.8372	1.9580
DNN	**Output Variable Data**	**1.4633**	**3.2010**	**1.7880**
RF	Process Variable and Output Variable Data	2.2062	6.3154	2.5129
DNN	**Process Variable and Output Variable Data**	**1.8214**	**4.0976**	**2.0243**

한 수준의 정보를 제공한다.

4.2 Viability 예측 결과 분석

Viability는 세포의 생존율을 나타내는 지표로, 배양 조건과 환경 요인에 민감하게 반응하는 생산성 지표이다. 본 연구에서는 다양한 입력 정보 구성을 바탕으로 Viability를 예측하고, 실제 측정치와의 비교를 통해 모델의 표현력 및 데이터셋 구성의 타당성을 검토하였다.

<Table 4>는 RF와 DNN 모델이 서로 다른 입력 데이터에 대해 Viability를 예측 성능을 비교한 결과이다. 전반적으로 모든 입력 조건에서 DNN이 RF보다 낮은 MAE, MSE, RMSE 값을 기록하여 더 우수한 성능을 나타냈으며, 특히 공정변수를 입력으로 활용했을 때 DNN의 예측력이 가장 두드러졌다. 반면 공정변수와 산출 지표를 통합한 경우에는 두 모델 모두 성능이 저하되는 경향을 보였는데, 이는 데이터 통합이 항상 예측 성능을 향상시키지 않음을 시사한다. 따라서 DNN은 복잡한 공정 데이터를 효과적으로 학습할 수 있는 장점을 지니고 있으나, 입력 변수의 조합에 따라 모델 성능이 달라질 수 있음을 확인할 수 있다.

<Figure 6>은 공정 변수 기반 데이터셋을 활용한 예측 결과를 나타낸다. 전반적으로 예측 곡선은 실제 값의 하강 경향과 반복 주기를 따라가고 있으나, 일부 구간에서는 급격한 변화에 대한 민감도가 떨어져 예측 오차가 다소 크게 나타났다. 특히 Viability가 급락하는 구간에서 모델의 반응 속도가 실제보다 늦게 나타나는 경향이 있었으며, 이는 공정 변수만으로는 세포 생존에 영향을 주는 미세한 생리학적 반응을 완전히 설명하기 어렵기 때문으로 판단된다.

<Figure 7>은 는 산출 지표 기반 데이터셋을 활용한 결과로, 예측 곡선은 실제 값과 유사한 추세를 보이나, 일부 실험 반복 구간에서는 예측값이 과도하게 낙폭을 키우는 양상이 관찰되었다. 이는 산출 지표 간의 상관 관계는 비교적 잘 반영되었으나, 세포 스트레스 누적이나 시간 경과에 따른 점진적 변화까지는 정교하게 반영되지 못한 것으로 해석된다. 또한 산출 변수 간 중복성이 존재할 경우, 학습 과정에서 일부 정보가 과대 반영될 가능성도 배제할 수 없다.

<Figure 8 >은 공정 변수와 산출 지표를 모두 포함한 통합 데이터셋 기반 예측 결과이며, 전체적인 곡선의 형태와 정합성

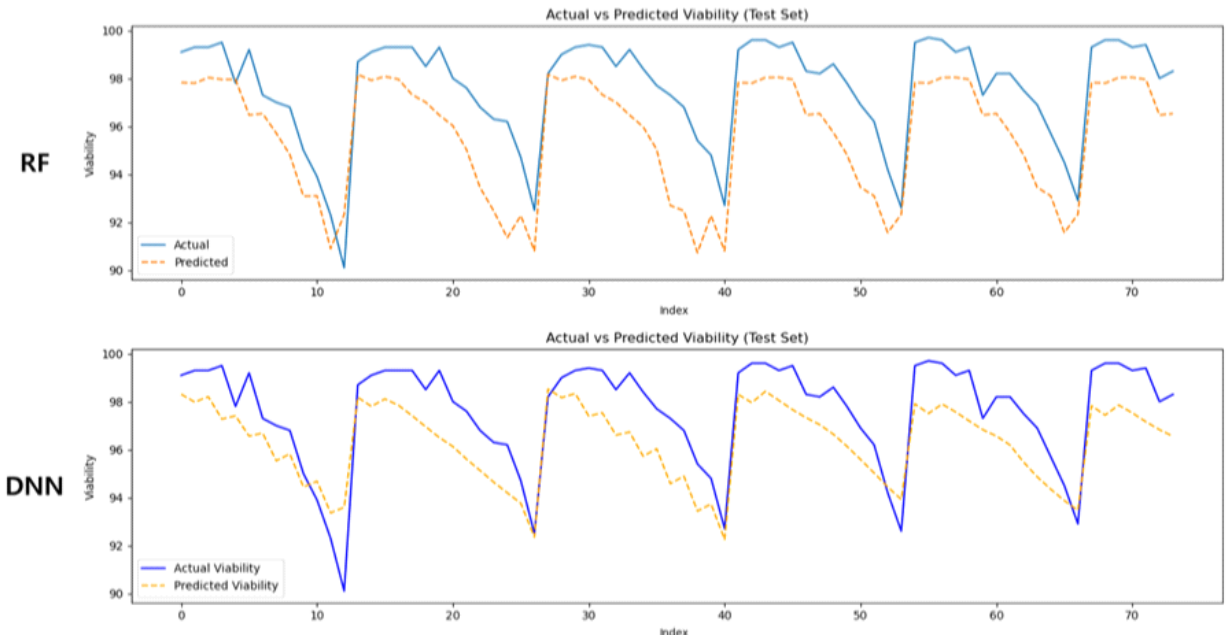


Figure 6. Visualization Results of Viability Prediction Using RF and DNN Models Trained on Process Variable Data

이 가장 떨어졌다. 특히 Viability의 점진적 하락 구간에서 가장 큰 차이를 보였으며, 회복 구간에서만 예측값이 실측치에 근접한 양상을 보였다. 이러한 결과는 공정 조건과 실시간 반응 데이터를 통합적으로 활용할 경우, 세포 생존에 영향을 미치는 다양한 내·외부 요인을 정밀하게 반영할 수는 없음을 의미한다.

Viability 예측 결과에서는 통합 데이터를 사용할 경우 예측 성능이 오히려 급격히 저하되는 현상이 관찰되었다. SHAP 기반 변수 중요도 분석 결과, Viability 예측은 배양 일자(Day)에 의해 지배적으로 설명되며, 그 외 공정 변수들은 Viability와의 상관성이 매우 낮거나 음(-)의 기여도를 보이는 경향을 나타냈다. 이로 인해 Viability와 직접적인 관련성이 낮은 변수들이 통합 데이터 구성에 포함될 경우, 예측 모델에 노이즈로 작용하여 성능 저하를 유발한 것으로 해석된다. 이러한 결과는

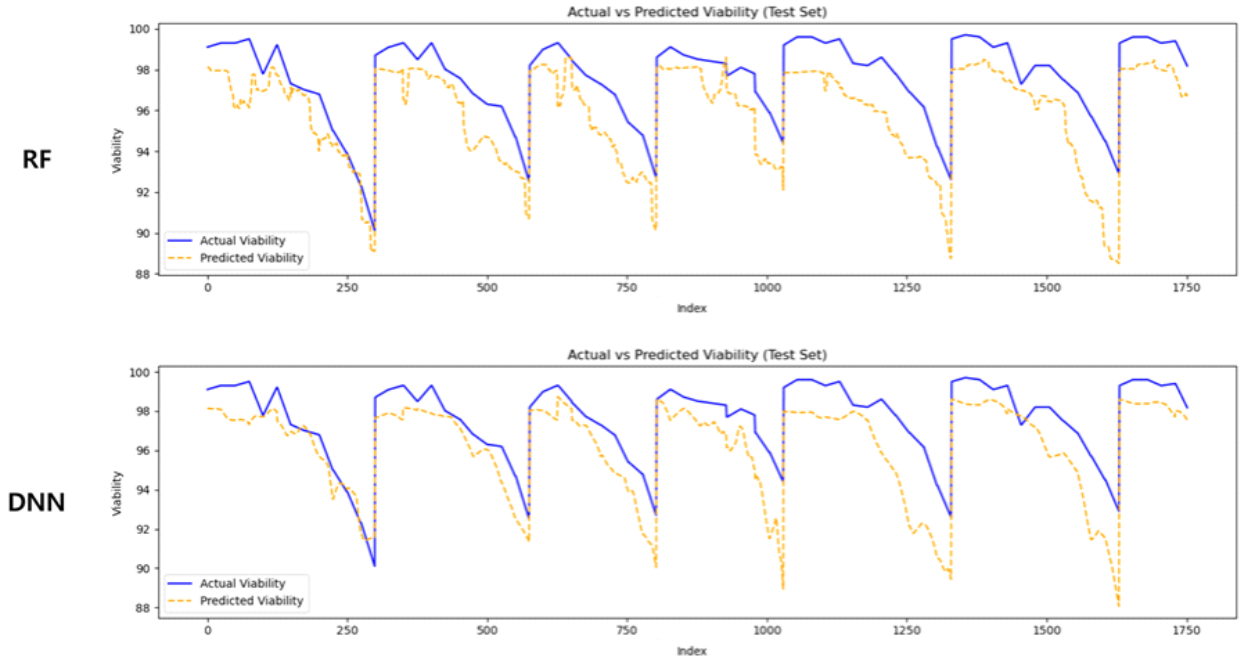


Figure 7. Visualization Results of Viability Prediction Using RF and DNN Models Trained on Output Variable Data

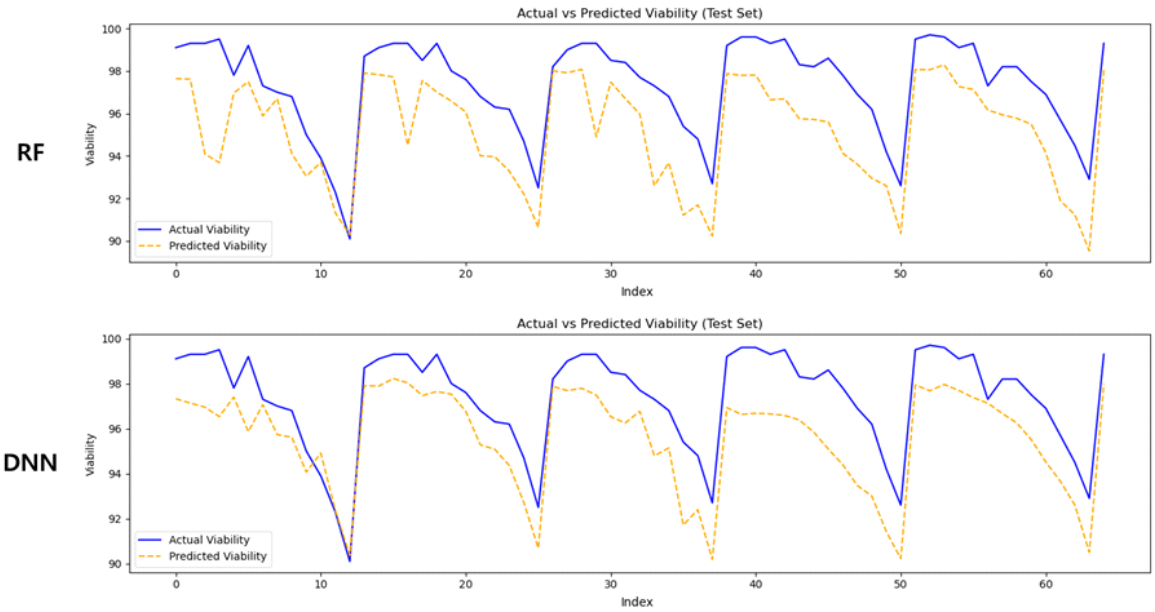


Figure 8. Visualization Results of Viability Prediction Using RF and DNN Models Trained on Integrated Process Variable and Output Variable Data

Viability가 특정 공정 시점의 상태를 반영하는 국소적 지표임을 시사하며, IgG와 달리 다양한 공정 변수의 누적 효과를 반영하지 않는 특성을 가진다는 점에서 두 지표 간 예측 특성의 차이를 보여준다.

Viability 예측 결과에서 제시된 오차 수준은 세포 생존율의 미세한 수치를 정확히 예측하기보다는, 배양 과정에서의 생존율 저하 경향과 위험 구간을 사전에 인지하기 위한 목적에 부합한다. 따라서 본 연구의 예측 결과는 공정 중단 여부 판단이나 공정 조건 조정 시점을 검토하는 데 참고 지표로 활용될 수 있다.

4.3 변수 중요도 기반 SHAP 해석 결과

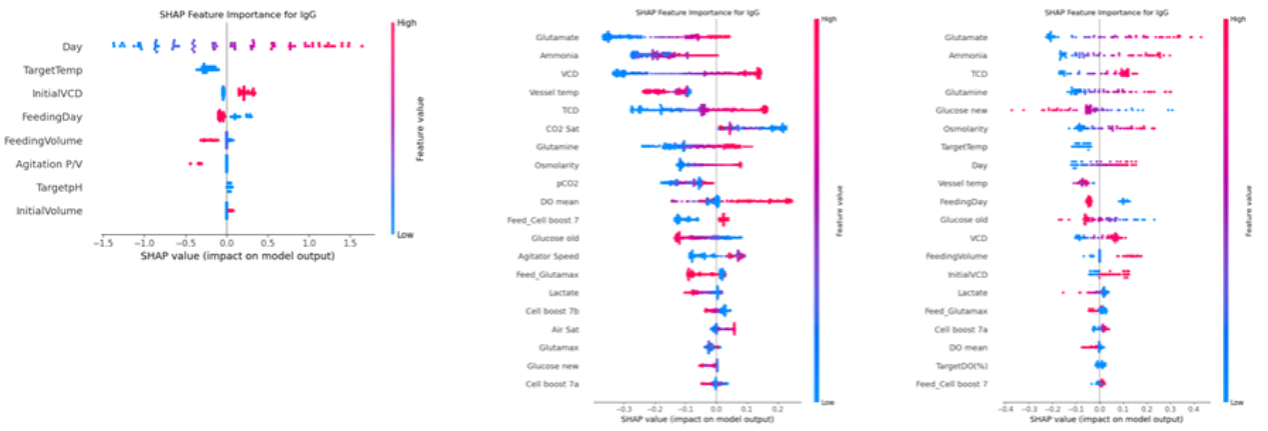
SHAP 분석을 통해 예측 모델의 결과에 기여하는 주요 변수들을 도출하고, 각 변수의 영향 방향성과 상대적 중요도를 정량적으로 평가하였다. SHAP 분석은 예측 모델의 해석 가능성을 확보함과 동시에, 공정 제어에 활용 가능한 주요 인자를 발굴하는 데 목적이 있다. 본 절에서는 예측 대상인 IgG와 Viability 각각에 대해 세 가지 입력 구성(공정 변수 기반, 산출 지표 기반, 통합형)에서 도출된 SHAP 결과를 종합적으로 분석하였다.

(1) IgG 예측에 대한 변수 영향 분석

IgG 예측에서는 데이터 구성 방식에 따라 중요 변수의 유형과 영향 방향이 달라짐을 보였다. <Figure 9> (a) 에서는 공정 변수 기반 모델의 SHAP 해석을 나타낸다. 이때 'Day'가 가장 우세한 영향력을 보였다. SHAP 분포에 따르면 배양 시간이 길어질수록 IgG 예측값은 일관되게 증가하는 방향으로 작용하였고, 점들의 분산 역시 명확한 패턴을 보였다. 'InitialVCD'와 'FeedingDay'는 각각 IgG 생성에 양의 기여와 복합적인 영향을 주는 변수로 확인되었으며, FeedingDay는 빠를수록 항체 생성에 기여하는 경향이 나타났다. 'TargetTemp'는 비교적 일관된 양의 영향을 보였고, 'Agitation P/V'나 'FeedingVolume' 등은 영향도가 낮고 분산도 작아 제한적인 역할을 하는 것으로 해석된다. 전반적으로 공정 변수 기반 모델은 IgG의 누적 생산 특성은 반영하나, 반응산물이나 세포 상태와 같은 생리학적 정보를 반영하지 못하여 예측 정확도에는 한계가 있었다.

<Figure 9> (b)는 산출 지표 기반 모델에서 대사 부산물 및 세포 밀도와 관련된 변수들이 지배적인 영향력을 보였다. 'Glutamate'는 높은 값일수록 예측값을 상승시키는 강한 양(+)의 영향을 나타냈고, 'Ammonia'는 일반적으로 음(-)의 방향으로 작용하였다. 'VCD', 'TCD', 'Vessel temp', 'Glutamine'도 상위권 변수로 나타났으며, 전체적으로 IgG 생성이 반응 산물의 축적 및 세포 상태에 의해 지배된다는 점을 잘 보여주었다. 특히 산출 지표 기반 모델은 피크 시점과 곡선의 변화에 민감하게 반응하는 예측 구조를 가지며, 반응 기반 예측의 가능성을 시사한다.

<Figure 9> (c)는 공정 - 산출 통합 모델에서 공정 변수와 산출 지표의 정보가 함께 반영되면서, 각 변수의 중요도가 복합적인 형태로 나타남을 보여준다. 'Glutamate', 'Ammonia', 'TCD', 'Glutamine', 'Glucose new' 등 산출 중심 변수들이 여전히 상위권에 위치하였으며, 'Day', 'FeedingDay', 'TargetTemp' 등 공정 변수들도 중간 이상의 영향력을 가지며 예측 정밀도를 보완하였다. 특히 Glutamate는 모든 구성에서 예측값에 일관된 양의 기여를 한 변수로, IgG 생성 예측의 핵심 지표로 기능할 수 있음을 보여준다. 통합 모델은 SHAP 값의 분포가 양

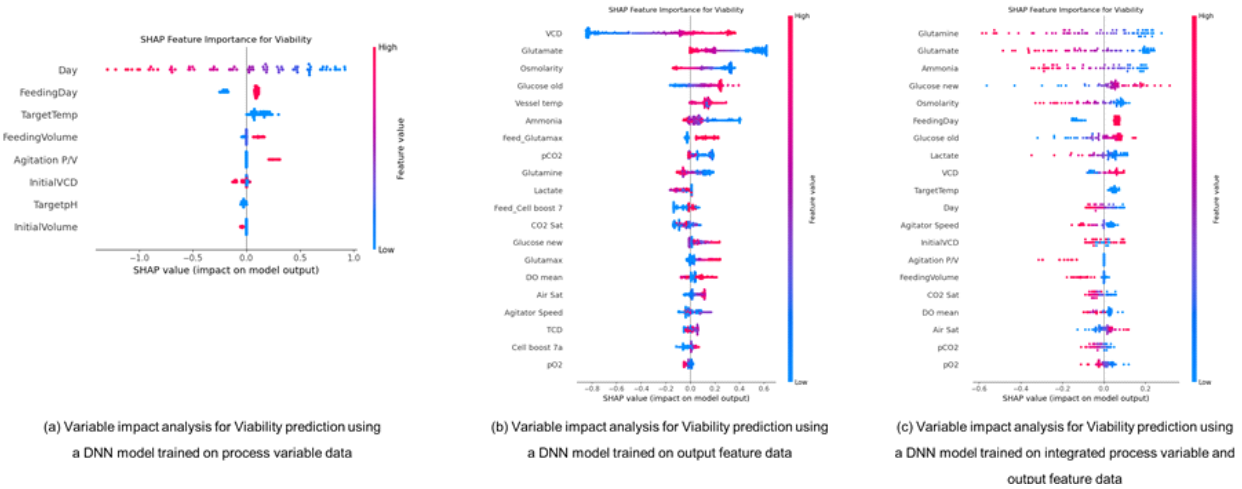


(a) Variable impact analysis for IgG prediction using a DNN model trained on process variable data

(b) Variable impact analysis for IgG prediction using a DNN model trained on output feature data

(c) Variable impact analysis for IgG prediction using a DNN model trained on integrated process variable and output feature data

Figure 9. SHAP-based Feature Importance Results for IgG Prediction under Different Input Configurations. (a) Process variable-based model, (b) output variable-based model, and (c) integrated model combining process and output variables. The plots show the distribution of SHAP values for each input variable across the evaluated samples.



(a) Variable impact analysis for Viability prediction using a DNN model trained on process variable data

(b) Variable impact analysis for Viability prediction using a DNN model trained on output feature data

(c) Variable impact analysis for Viability prediction using a DNN model trained on integrated process variable and output feature data

Figure 10. SHAP-based feature importance results for Viability prediction under different input configurations. (a) Process variable-based model, (b) output variable-based model, and (c) integrated model combining process and output variables. The plots show the distribution of SHAP values for each input variable across the evaluated samples.

방향적으로 균형 잡혀 있으며, 예측 정확도뿐 아니라 해석력에서도 가장 안정적인 결과를 보였다.

(2) Viability 예측에 대한 변수 영향 분석

본 절에서는 Viability 예측 모델에 대해 SHAP 분석을 수행하여, 공정 변수와 산출 지표가 Viability 예측에 미치는 상대적 영향과 기여 방향을 분석한다. <Figure 10> (a)는 공정 변수 기반 DNN 모델에 대한 Viability 예측 결과를 SHAP 값으로 분석한 변수 영향도를 나타낸 것이다. <Figure 10> (a)에서 나타내는 공정 변수 기반 모델 에서는 'Day'가 가장 강한 음(-)의 영향을 나타냈으며, 시간 경과에 따라 Viability가 지속적으로 감소하는 경향이 예측 결과에 그대로 반영되었다. 'FeedingDay'는 조기 급여일수록 Viability에 양(+)의 기여를 하였고, 'TargetTemp'는 높은 온도에서 음의 영향이 나타났다. 'FeedingVolume'은 전체적으로 낮은 중요도를 보였으며, 'InitialVCD'는 지나치게 높을 경우 음의 방향으로 작용하는 경향이 관찰되었다. 이러한 분석은 공정 초반 조건 설정이 Viability에 큰 영향을 미친다는 점을 시사한다.

<Figure 10> (b) 에서 나타내는 산출 지표 기반 모델에서는 'VCD', 'Glutamine', 'Glutamate', 'Ammonia', Osmolarity' 등이 주요 변수로 나타났으며, Glutamine은 Viability을 높이는 양(+)의 영향, Glutamate와 Ammonia는 축적 시 Viability를 저하시키는 음(-)의 영향을 주는 것으로 해석되었다. 'Glucose old', 'pCO_2', 'Vessel temp' 등도 일정 수준의 영향력을 보였으며, 전반적으로 산출 변수는 세포 내 대사 환경의 상태를 반영하여 예측 정밀도를 높였다. 특히 Glutamate의 경우 Viability와 IgG에서 상반된 영향을 보이며, 생산성 지표 간 trade-off의 실질적 분석이 가능함을 보여준다.

<Figure 10> (c)에서 나타내는 공정-산출 통합 모델에서는 산출 지표와 공정 변수가 상호 보완적으로 작용하였다. Glutamine, Glutamate, Ammonia 등은 여전히 중심 변수로 나타났고, FeedingDay, Glucose new, Lactate, Osmolarity 등이 예측에 의미 있는 기여를 하였다. FeedingDay는 일정 수준 이하에서는 Viability를 향상시키는 반면, FeedingVolume의 과도한 증가나 높은 Osmolarity는 음(-)의 영향을 보여, 조정 가능한 변수로서의 제어 전략 수립에 중요한 근거를 제공한다.

본 연구에서의 예측 문제는 시계열 예측이 아니라, 특정 배양 시점에서의 공정 상태를 기반으로 생산성 지표를 추정하는 상태 기반 예측으로 정의된다. 'Day' 변수는 시간 순서에 따른 과거 정보의 누적을 반영하기 위한 시계열 인덱스가 아니라, 배양 공정 내에서의 현재 공정 단계(stage)를 나타내는 상태 변수로 사용되었다. 따라서 SHAP 분석에서 'Day'의 높은 영향도는 모델이 시간적 의존성을 학습했음을 의미하기보다는, 배양 진행 단계 자체가 Viability를 설명하는 중요한 상태 정보임을 시사한다.

5. 토 론

본 연구에서는 항체 기반 바이오의약품 생산 공정에서의 생산성 예측을 위해, 공정 입력 변수와 산출 지표를 통합한 머신러닝 기반 예측 모델을 제안하고, SHAP 기반 해석 기법을 통해 변수 기여도를 정량적으로 분석하였다. 트리 기반 모델과 신경망 기반 모델을 대표하는 RF와 DNN 을 선택하여 비교 분석을 수행하였는데, 이는 모델 구조의 다양성을 확보하면서도, 입력 데이터 구성 전략에 따른 예측 특성과 해석 가능성을 명확히 비교하기 위함이다. 예측 모델의 성능 비교 결과, IgG 예측의 경우 공정 변수와 산출 지표를 통합한 입력 구성에서 가장 우수한 결과를 기록하며, 단일 입력 변수 기반 모델에 비해 높은 예측 정확도를 보였다. 이는 Glucose, Lactate, Viability 등 세포의 대사 및 생리 상태를 반영하는 산출 지표가 항체 생산

량과 유의미한 상관관계를 가지며, 공정 입력 변수와 함께 사용할 경우 예측 성능이 개선될 수 있음을 시사한다.

반면, Viability 예측의 경우 산출 지표 기반 모델의 예측력이 가장 우수하였으며, 통합 입력 구성에서도 별도의 성능 향상은 제한적이었다. 이러한 결과는 Viability가 공정 말기 시점에서 직접 측정된 반응 변수로서, 해당 시점의 세포 상태를 가장 잘 반영하기 때문으로 해석할 수 있다. 즉, 예측하려는 생산성 지표에 따라 입력 변수 구성 전략이 달라져야 하며, 특정 생산성 지표에 대해 불필요한 변수 결합은 오히려 예측 성능을 저해할 수 있음을 시사한다.

SHAP 해석을 통해 각 변수의 영향력을 시각화한 결과, IgG 예측에서는 Day, Viable Cell Density(VCD), Glucose, Lactate 등의 변수들이 높은 영향도를 가지는 것으로 나타났으며, 특히 배양 일수가 길수록 IgG가 증가하는 경향이 뚜렷하였다. 반면 Viability 예측에서는 배양 일수가 가장 큰 음의 영향력을 가지는 변수로 나타났으며, Lactate 및 Osmolality 등의 변수도 Viability 저하에 기여하는 부정적 요소로 분석되었다. 이 결과는 동일한 변수라도 예측 대상에 따라 모델 내에서의 기여 방향과 중요도가 달라질 수 있음을 보여준다. 또한 IgG 예측에서는 산출 지표를 포함한 입력 구성이 예측 성능에 영향을 미칠 수 있음을 실험적으로 확인하였다.

또한 본 연구는 기존 문헌에서 상대적으로 간과되었던 입력 변수 통합 효과와 모델 해석 가능성을 함께 고려하였다는 점에서 차별성을 가진다. 기존 연구들은 대부분 공정 변수 기반 모델링에 집중하거나, 예측 성능만을 중심으로 결과를 해석하는 경우가 많았다. 반면 본 연구는 공정 입력과 반응 산출 지표의 상호 작용을 통합적으로 분석하고, SHAP 기반 해석을 통해 변수별 영향도를 정량화함으로써 실제 공정 제어 인자 도출에도 활용 가능한 구조를 제시하였다.

그럼에도 불구하고 본 연구는 몇 가지 한계를 가진다. 첫째, 실험에 사용된 데이터는 특정 조건 하에서 수집된 제한된 규모의 배치 공정 데이터로, 시간 간격이 불균형하거나 누락된 값이 존재할 수 있다. 둘째, 본 연구에서 사용한 예측 모델은 고정된 입력 구조를 기반으로 학습되었으며, RNN이나 LSTM과 같이 시계열 의존성을 명시적으로 모델링하는 구조는 적용하지 않았다. 이는 본 연구에서 사용한 데이터셋이 단일 공정의 연속적인 시간 흐름을 나타내는 시계열 데이터가 아니라, 서로 다른 배양 조건에서 수행된 다수의 배치 실험으로부터 수집된 관측치를 통합하여 구성되었기 때문이다. 셋째, 예측 모델의 실시간 적용 가능성 및 일반화 가능성은 아직 검증되지 않았기 때문에, 향후 다양한 배양 조건과 공정 환경을 포함한 다기관 데이터셋에 대한 확장이 필요하다.

본 연구에서는 배치별로 상이한 측정 시점을 정렬하기 위해 보간을 적용하였다. 그러나 보간된 값은 원시 관측 데이터의 변동성을 일부 평활화할 수 있으며, 이에 따라 예측 모델이 실제 배치 데이터의 불확실성을 충분히 반영하지 못했을 가능성이 있다. 따라서 본 연구의 예측 성능은 보간된 배치 데이터 구조를 전제로 해석되어야 하며, 향후 연구에서는 원시 배치 데이터 기반 분석을 통해 해당 전처리의 영향을 추가로 검증할 필요가 있다.

향후 연구에서는 공정 시간에 따른 생산성 지표의 변화 양상을 정량화하고, 이를 기반으로 최적 배양 시점 판단을 지원하는 실시간 품질 예측 및 공정 제어 전략으로의 확장이 필요하다. 또한 Gradient Boosting 계열 모델과의 추가 비교는 향후 연구 과제로 남긴다. 본 연구에서 개발한 예측 및 해석 구조는 디지털 트윈 기반 바이오 공정 모듈에 통합하여, 운용자 의사결정을 보조하고 공정 최적화를 지원하는 핵심 구성요소로 활용될 수 있을 것으로 기대된다.

6. 결 론

본 연구는 바이오의약품 제조 공정에서 대표적인 생산성 지표인 IgG와 Viability을 예측하기 위해, 공정 입력 변수와 반응 산출 지표를 통합한 머신러닝 기반 예측 모델을 제안하였다. 특히, RF와 DNN 모델을 기반으로 세 가지 입력 조합(공정 변수 기반, 산출 지표 기반, 통합 입력 기반)에 대해 예측 성능을 비교하고, SHAP 해석 기법을 활용하여 각 변수의 기여도와 영향 방향성을 정량적으로 분석하였다.

실험 결과, IgG 예측에서는 통합 입력 구성의 예측 성능이 가장 우수하였으며, 반응 산출 지표가 예측력 향상에 크게 기여하는 것으로 나타났다. 반면 Viability 예측에서는 공정 변수 기반 모델이 가장 안정적인 예측 성능을 보였으며, 이는 예측 대상에 따라 최적의 입력 구성 전략이 달라져야 함을 시사한다. 또한 SHAP 분석 결과, 동일한 변수라 하더라도 생산성 지표에 따라 기여 방향과 영향도가 상이하게 나타나는 것으로 확인되었으며, 이는 공정 종료 시점 결정 등 실제 운전 전략 수립 시 품질 간 균형 고려가 필요함을 의미한다.

본 연구의 기여는 다음과 같다. 첫째, 공정 입력 변수와 산출 지표를 통합한 예측 모델을 구성함으로써 생산성 예측 정확도를 향상시켰고, 둘째, SHAP 기반 해석 기법을 통해 예측 모델의 설명 가능성을 확보하였다. 셋째, IgG와 Viability의 특성별 입력 구성 전략을 실증적으로 제시함으로써, 향후 공정 설계 및 제어 변수 선택 시 근거 기반 의사결정이 가능하도록 하였다.

향후에는 시계열 정보를 반영한 예측 모델 구조의 고도화, 다기관 공정 데이터를 활용한 일반화 성능 검증, 디지털 트윈 시스템 내 예측 모듈로의 통합 적용 등의 후속 연구가 필요하다. 특히 본 연구에서 제안한 예측 및 해석 구조는 스마트 바이오제조 시스템의 생산성 관리 및 실시간 공정 제어에 활용 가능한 핵심 기반 기술로서의 확장 가능성을 가진다.

참고문헌

Kothari, M., Wanjari, A., Acharya, S., Karwa, V., Chavhan, R., Kumar,

S., ... and Patil, R. (2024), A comprehensive review of monoclonal antibodies in modern medicine: Tracing the evolution of a revolutionary therapeutic approach, *Cureus*, **16**(6).

Mekala, J. R., Nalluri, H. P., Reddy, P. N., Sb, S., NS, S. K., Gvsd, S. K., ... and Dirisala, V. R. (2024), Emerging trends and therapeutic applications of monoclonal antibodies, *Gene*, **925**, 148607.

Ranbhor, R. (2025), Advancing Monoclonal Antibody Manufacturing: Process Optimization, Cost Reduction Strategies, and Emerging Technologies, *Biologics: Targets and Therapy*, 177-187.

Birch, J. R. and Racher, A. J. (2006), Antibody production, *Advanced drug Delivery Reviews*, **58**(5-6), 671-685.

Lee, J., Ortega-Rodriguez, U., Madhavarao, C. N., Ju, T., O'Connor, T., Ashraf, M., and Yoon, S. (2025), Effect of different cell culture media on the production and glycosylation of a monoclonal antibody from a CHO cell line, *Cytotechnology*, **77**(3), 1-18.

Zhang, S., Chen, H., Wan, Y., Wang, H., and Qu, H. (2024), A Data-Driven Approach for Leveraging Inline and Offline Data to Determine the Causes of Monoclonal Antibody Productivity Reduction in the Commercial-Scale Cell Culture Process, *Pharmaceutics*, **16**(8), 1082.

Reyes, S. J., Durocher, Y., Pham, P. L., and Henry, O. (2022), Modern sensor tools and techniques for monitoring, controlling, and improving cell culture processes, *Processes*, **10**(2), 189.

Pham, T. D., Manapragada, C., Sun, Y., Bassett, R., and Aickelin, U. (2023). A scoping review of supervised learning modelling and data-driven optimisation in monoclonal antibody process development, *Digital Chemical Engineering*, **7**, 100080.

Mondal, P. P., Galodha, A., Verma, V. K., Singh, V., Show, P. L., Awasthi, M. K., ... and Jain, R. (2023), Review on machine learning-based bioprocess optimization, monitoring, and control systems, *Bioresource Technology*, **370**, 128523.

Lim, S. J., Son, M., Ki, S. J., Suh, S. I., and Chung, J. (2023), Opportunities and challenges of machine learning in bioprocesses: categorization from different perspectives and future direction, *Bioresource Technology*, **370**, 128518.

Medl, M., Leisch, F., Dürauer, A., and Scharl, T. (2024), Explainable deep learning enhances robust and reliable real-time monitoring of a chromatographic protein A capture step, *Biotechnology Journal*, **19**(2), 2300554.

Baako, T. M. D., Kulkarni, S. K., McClendon, J. L., Harcum, S. W., and Gilmore, J. (2024), Machine learning and deep learning strategies for Chinese hamster ovary cell bioprocess optimization, *Fermentation*, **10**(5), 234.

Helleckes, L. M., Hemmerich, J., Wiechert, W., von Lieres, E., and Grünberger, A. (2023), Machine learning in bioprocess development: from promise to practice, *Trends in Biotechnology*, 41(6), 817-835.

Li, M., Sun, H., Huang, Y., and Chen, H. (2024), Shapley value: from cooperative game to explainable artificial intelligence, *Autonomous Intelligent Systems*, **4**(1), 2.

Winter, E. (2002), The shapley value, *Handbook of Game Theory with Economic Applications*, **3**, 2025-2054.

Nik-Khorasani, A., Khuat, T. T., and Gabrys, B. (2025), Hyperbox Mixture Regression for process performance prediction in antibody production, *Digital Chemical Engineering*, **14**, 100221.

Lai, P. K., Gallegos, A., Mody, N., Sathish, H. A., and Trout, B. L. (2022, December), Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics, In *MAbs* (Vol. 14, No. 1, p. 2026208), Taylor & Francis.

Narayanan, H., Dingfelder, F., Condado Morales, I., Patel, B., Heding, K. E., Bjelke, J. R., ... and Arosio, P. (2021), Design of biopharmaceutical formulations accelerated by machine learning, *Molecular Pharmaceutics*, **18**(10), 3843-3853.

Wossnig, L., Furtmann, N., Buchanan, A., Kumar, S., and Greiff, V. (2024), Best practices for machine learning in antibody discovery and development, *Drug Discovery Today*, **29**(7), 104025.

Makowski, E. K., Kinnunen, P. C., Huang, J., Wu, L., Smith, M. D., Wang, T., ... and Tessier, P. M. (2022), Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space, *Nature Communications*, **13**(1), 3788.

Joubbi, S., Micheli, A., Milazzo, P., Maccari, G., Ciano, G., Cardamone, D., and Medini, D. (2024), Antibody design using deep learning: from sequence and structure design to affinity maturation, *Briefings in Bioinformatics*, **25**(4), bbae307.

Richter, J., Wang, Q., Lange, F., Thiel, P., Yilmaz, N., Solle, D., ... and Beutel, S. (2025), Machine learning-powered optimization of a CHO cell cultivation process, *Biotechnology and Bioengineering*, **122**(5), 1153-1164.

Pinto, J., Ramos, J. R., Costa, R. S., Rossell, S., Dumas, P., and Oliveira, R. (2023), Hybrid deep modeling of a CHO-K1 fed-batch process: combining first-principles with deep neural networks, *Frontiers in Bioengineering and Biotechnology*, **11**, 1237963.

Duong-Trung, N., Born, S., Kim, J. W., Schermeyer, M. T., Paulick, K., Borisyak, M., ... and Martinez, E. (2023), When bioprocess engineering meets machine learning: A survey from the perspective of automated bioprocess development, *Biochemical Engineering Journal*, **190**, 108764.

Khuat, T. T., Bassett, R., Otte, E., and Gabrys, B. (2025), Uncertainty quantification using ensemble learning and Monte Carlo sampling for performance prediction and monitoring in cell culture processes, *Journal of Raman Spectroscopy*.

Yatipanthalawa, B. S., Fitzsimons, S. E. W., Horning, T., Lee, Y. Y., and Gras, S. L. (2024), Development and validation of a hybrid model for prediction of viable cell density, titer and cumulative glucose consumption in a mammalian cell culture system, *Computers & Chemical Engineering*, **184**, 108648.

Di Bonito, L. P., Campanile, L., Di Natale, F., Mastroianni, M., and Iacono, M. (2024), Explainable artificial intelligence in process engineering: Promises, facts, and current limitations, *Applied System Innovation*, **7**(6), 121.

Huang, Y. S. (2023), Digital Twin Development and Advanced Process Control for Continuous Pharmaceutical Manufacturing (Doctoral dissertation, Purdue University Graduate School).

Shahab, M. A., Destro, F., and Braatz, R. D. (2025), Digital Twins in Biopharmaceutical Manufacturing: Review and Perspective on Human-Machine Collaborative Intelligence, *arXiv preprint* arXiv:2504.00286.

Isoko, K., Cordiner, J. L., Kis, Z., and Moghadam, P. Z. (2024), Bioprocessing 4.0: A pragmatic review and future perspectives, *Digital Discovery*, **3**(9), 1662-1681.

Rosen, R., Von Wichert, G., Lo, G., and Bettenhausen, K. D. (2015), About the importance of autonomy and digital twins for the future of manufacturing, *Ifac-papersonline*, **48**(3), 567-572.

Tao, F., Qi, Q., Wang, L., and Nee, A. Y. C. (2019), Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison, *Engineering*, **5**(4), 653-661.

Samek, W., Wiegand, T., Müller, K. (2017), Explainable artificial

intelligence: Understanding, visualizing and interpreting deep learning models, *arXiv preprint* arXiv:1708.08296.

저자소개

김민지 : 경희대학교 산업경영공학과에서 2024년 학사학위를 취득하고, 2026년 동 대학교 인공지능학과 석사학위를 취득하였다. 연구분야는 Smart Factory, Physical AI, Digital Product Passport이다.

양지영 : 2022년부터 경희대학교 산업경영공학과에 재학 중이며, 주요 연구분야는 스마트 제조, 공정 최적화이다.

이채현 : 2022년부터 경희대학교 산업경영공학과에 재학 중이며, 주요 연구 관심 분야는 공정 최적화, 시뮬레이션 기반 디지털 트윈, 스마트 제조 시스템이다.

윤지용 : 경희대학교 화학공학과에서 학사 및 석사 학위를 취득하였으며, 현재 프레스티지바이오로직스에서 공정개발 및 MSAT 팀장을 맡고 있다. 항체의약품 및 단백질의약품 제조 공정 분야에서 20년 이상의 풍부한 경력을 보유하고 있으며, 현재까지도 활발한 연구와 실무를 이어가고 있다.

오대양 : 2010년 한양대학교 토목공학과 학사를 취득하였고, 2012년 동 대학교 건설환경공학과(환경) 석사 학위를 취득하였다. 현재는 프레스티지바이오로직스 엔지니어링실장으로 재직 중으로 재직 중이며, 프레스티지바이오로직스 1, 2, 3, 4 공장 구축을 담당하였다.

엄주명 : 성균관대학교 기계공학부에서 2003년 학사, POSTECH에서 산업공학 석/박사학위를 취득하였다. 영국 캠브리지대학교 Research associate와 독일 인공지능연구소 senior researcher를 역임하고 2018년부터 경희대학교 산업경영공학과 교수로 재직하고 있다. 연구분야는 Smart Factory, Augmented reality, CAD/CAM이다.