Korean Chem. Eng. Res., 62(3), 214-224 (2024)
https://doi.org/10.9713/kcer.2024.62.3.214
PISSN 0304-128X, EISSN 2233-9558

머신 러닝을 이용한 밸브 사이즈 및 종류 예측 모델 개발

김찬호* · 최민식** · 주종효*,** · 이아름*** · 윤 건*** · 조성호*** · 김정환*,†

*연세대학교 화공생명공학과
03722 서울특별시 서대문구 연세로 50
**한국생산기술연구원 친환경재료공정연구그룹
44413 울산광역시 중구 종가로 55
***삼성 E&A
05288 서울특별시 강동구 상일동 500
(2023년 12월 7일 접수, 2024년 4월 30일 수정본 접수, 2024년 5월 22일 채택)

Data-driven Modeling for Valve Size and Type Prediction Using Machine Learning

Chanho Kim*, Minshick Choi, Chonghyo Joo*,**, A-Reum Lee***, Yun Gun***, Sungho Cho*** and Junghwan Kim*,†**

**Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul, 03722, Korea*
***Green Materials and Processes R&D Group, Korea Institute of Industrial Technology, 55, Jongga-ro, Jung-gu, Ulsan, 44413, Korea*
****Samsung E&A Co., Ltd., 26, Sangil-ro 6-gil, Gangdong-gu, Seoul, 05288, Korea*
(Received 7 December 2023; Received in revised from 30 April 2024; Accepted 22 May 2024)

요 약

밸브는 유량과 압력 조절 등의 중요한 역할을 수행하며, 적절한 밸브 사이즈와 유형 선택이 필요하다. Engineering Procurement Construction (EPC) 산업에선 밸브 사이즈 계수(C_v)의 수식적 계산을 바탕으로 사이즈와 유형을 선정해 왔으나 이러한 방식은 전문가의 많은 시간과 비용이 요구되어 비효율적이다. 본 연구는 이를 해결하기위해 머신 러닝 기법을 이용한 밸브 사이즈 및 유형 예측 모델을 개발하였다. Artificial neural network (ANN), Random Forest, XGBoost, Catboost의 알고리즘을 적용하여 모델들을 개발하였으며, 평가 지표로는 사이즈 예측에는 Normalized root mean squared error (NRMSE)와 R^2를, 종류 예측에는 F1 score를 적용하였다. 또한, 유체 상에 따른 영향을 확인하고자 유체 전체, 액체, 기체, 스팀의 4개의 데이터 세트로 사례 연구를 진행하였다. 연구 결과, 사이즈의 경우 전체, 액체, 기체에선 Catboost(R^2 기준, 전체: 0.99216, 액체: 0.98602, 기체: 0.99300. NRMSE 기준, 전체: 0.04072, 액체: 0.04886, 기체: 0.03619)가, 스팀에선 Random Forest가(R^2: 0.99028, NRMSE: 0.03493) 가장 뛰어난 모델임을 확인하였다. 종류의 경우 Catboost가 모든 데이터에서 가장 높은 성과를 제시하였다(F1 score 기준, 전체: 0.95766, 액체: 0.96264, 기체: 0.95770, 스팀: 1.0000). 본 연구에서 제안한 모델들을 적용할 경우, 주어진 조건에 따른 밸브 선택을 도와 의사결정 속도를 높여줄 것으로 기대된다.

Abstract – Valves play an essential role in a chemical plant such as regulating fluid flow and pressure. Therefore, optimal selection of the valve size and type is essential task. Valve size and type have been selected based on theoretical formulas about calculating valve sizing coefficient (C_v). However, this approach has limitations such as requiring expert knowledge and consuming substantial time and costs. Herein, this study developed a model for predicting valve sizes and types using machine learning. We developed models using four algorithms: ANN, Random Forest, XGBoost, and Catboost and model performances were evaluated using NRMSE & R^2 score for size prediction and F1 score for type prediction. Additionally, a case study was conducted to explore the impact of phases on valve selection, using four datasets: total fluids, liquids, gases, and steam. As a result of the study, for valve size prediction, total fluid, liquid, and gas dataset demonstrated the best performance with Catboost (Based on R^2, total: 0.99216, liquid: 0.98602, gas: 0.99300. Based on NRMSE, total: 0.04072, liquid: 0.04886, gas: 0.03619) and steam dataset showed the best performance with

† To whom correspondence should be addressed.
E-mail: kjh24@yonsei.ac.kr

RandomForest (R^2: 0.99028, NRMSE: 0.03493). For valve type prediction, Catboost outperformed all datasets with the highest F1 scores (total: 0.95766, liquids: 0.96264, gases: 0.95770, steam: 1.0000). In Engineering Procurement Construction industry, the proposed fluid-specific machine learning-based model is expected to guide the selection of suitable valves based on given process conditions and facilitate faster decision-making.

Key words: Engineering valve, Valve size, Valve type, Data-driven modeling, Machine learning

1. 서 론

엔지니어링 밸브는 화학 플랜트에서 압력, 유량, 온도 등을 조건에 맞게 조절하고, 원하는 방향으로 유체를 흐르게 하는 중요한 역할을 한다. 넓은 범위의 온도와 압력 조건에서 대량의 유체가 흐르는 화학 플랜트 산업의 특성상 부적절한 밸브 선택은 막대한 금전적 손실과 인명 피해를 야기할 수 있다. 따라서, 운전 범위에 적합한 밸브 선택은 필수적이다.

부적절한 밸브 사이즈와 유형을 공정에 적용하면 제어 효율과 반응성의 하락, 에너지 손실, 파이프라인 손상 등의 문제가 발생할 수 있다. 너무 작은 사이즈의 밸브는 유체의 원활한 흐름을 방해하여 플랜트의 공정 효율을 낮출 수 있다. 반대로, 너무 큰 사이즈의 밸브는 과도한 유속과 유량으로 인해 밸브의 부식, Cavitation, Flashing 등의 문제가 발생할 수 있다. 뿐만 아니라, 사이즈가 같더라도 Globe, Ball, Gate, 그리고 Butterfly 등 다양한 유형[1]이 존재하며, 유형에 따라 적합한 공정 조건과 역할이 다르다. 따라서, 플랜트 설계 시, 엔지니어링 밸브의 사이즈와 유형을 찾는 것은 매우 중요하다.

밸브 사이즈는 일반적으로 밸브 사이즈 계수(Valve sizing coefficient, C_v)를 계산하고, C_v에 따른 사이즈 표를 통해 선택된다. C_v는 "완전히 열린 밸브 사이에 1psi의 압력 강하가 존재할 때 통과하는 60°F 물의 분당 갤런"라는 Driskell의 정의에 기반한다[2]. 반복 실험을 통해 C_v를 측정할 수 있으나 시간과 비용의 소모가 크기 때문에 EPC 업계에서는 수식 계산을 통해 구한다. C_v는 다음의 수식 (1)로 표현된다.

$$C_v = Q\sqrt{(S/\Delta P)} \tag{1}$$

Q는 gal/min 단위의 유량, S는 60°F 물에 대한 유체의 비중, 그리고 ΔP는 psi 단위의 압력 강하를 의미한다.

수식 (1)을 통한 C_v 계산은 액체, 기체/스팀의 유체 상에 따른 차이와 온도, 압력 등의 공정 파라미터 영향을 충분히 반영하지 못해 오늘날 EPC 업계는 IEC 60534-2-1[3], ISA75.01[4]와 같은 산업 기준에 따라 C_v를 계산하고 있다. 그러나, 이러한 수학적 방법은 상황에 따라 정확성이 떨어지거나 특정 공정 조건에는 적용하기 어려울 수 있는 등의 문제가 발생할 수 있으며, 부적절한 밸브 사이즈 선정 문제로 이어진다.

이를 해결하기 위해 몇몇 연구들이 진행되었다. Grace는 최소자승법(Least Square Method)로 Choke Valve의 velocity of approach (α)에 대한 C_v의 함수를 도출하는 방법론을 제시하였다[5]. Long은 angle-seat valve에 대하여 IEC의 몰 농도 유량 계산 수식에서부터 non-choked와 choked valve의 C_v를 구하는 식을 유도하여 시뮬레이션을 설계하고 최소자승법을 통해 곡선을 적합 시키는 과정을 진행했다[6]. Zhou는 전산유체역학(CFD)을 이용해 실험을 통한 측정보다 적은 시간과 비용을 들여 3% 이하의 상대 오차로 C_v를 예측하였다[7]. Lisowski 역시 CFD를 이용한 연구를 진행하였다[8]. 그는 CFD 결과를 기반으로 C_v 데이터를 확보한 뒤 MATLAB을 이용해 Linear curve, Linear polynomial surface, 그리고 Quadratic polynomial surface 중 가장 C_v에 잘 적합할 수 있는 함수를 선정하였다. Valdés는 CFD를 이용해 압력 강화와 유량을 brake master cylinder와 ABS 밸브에 대해 시뮬레이션하였고 그 결과를 토대로 최소자승법을 이용해 C_v 함수의 파라미터를 추정하였다[9]. Nguyen는 파이프의 단면적의 일부만 유체가 흐르는 Partially Filled Pipe Flow(PFPF)의 반복 실험에서의 거품 생성에 대한 관측을 토대로 Butterfly 밸브에서의 C_v를 구하는 과정을 진행하였다[10].

그러나, 이러한 C_v 예측 방법에는 여러 한계점들이 있다. C_v 계산에는 공정을 운전해보기 전에는 모르는 여러 불확실성을 고려한 전문가의 가정이 요구되며, 계산 결과의 적용 가능 여부에 대한 검토와 반복 계산이 필요하여 전문가 의존도가 높다. 또한, 공정 조건이나 파라미터 변경 시, 이를 반영하기 위한 계산 과정이 반복되고, 밸브 구조가 달라지는 경우, effective valve opening percentage, coefficient of diffusion, pressure drop 등의 여러 파라미터가 달라 수식 (1)의 Q와 ΔP에 영향을 주게 된다[11-13]. 이러한 문제들로 인해 C_v와 밸브 유형 사이의 직접적인 연관 관계가 불분명하여 유형 선정 시, 전문가의 노하우와 경험에 의존도가 높다. CFD를 활용한 연구들은 이러한 문제를 어느 정도 보완할 수 있으나 계산 비용이 크다는 근본적인 한계가 있다. Mahalleh는 이를 해결하기 위해, 객체 인식 딥러닝 모델인 YOLO 모델을 이용해 유체의 흐름을 센서로 읽고 이를 통해 밸브 종류를 인식하는 연구를 진행한 바 있으나[14] 여전히 공정 설계 단계에서는 활용하기 어렵다는 문제점이 있다. Hlubek은 Solenoid 밸브 내부의 Diaphragm 유형을 여러 머신러닝 모델을 이용해 예측하는 연구를 진행하였으며[15] Convolutional Neural Network가 0.993의 정확도로 가장 높은 성능을 제시함을 보여주었다. 그러나, 그의 연구는 여러 밸브 유형은 다루지 않고 있으며 Solenoid 밸브에만 한정되어 있다는 점과 시계열 데이터가 요구되어 실제 산업에 적용되기에는 난관이 있다는 한계가 존재한다.

본 연구에서는 이러한 문제를 효과적으로 해결하기 위해 실제 엔지니어링 밸브 데이터와 머신러닝을 이용한 밸브 사이즈 및 유형 예측 모델을 개발하였다. 머신러닝은 화학공학 분야의 다양한 문제 해결을 위한 방법론으로 다방면의 복잡한 상황과 조건에 적극적으로 적용되고 있는 기법이다[16-19]. 2,206개의 실제 엔지니어링 밸브 데이터를 액체(1,303개), 기체(787개), 스팀(116개)의 유체 별 데이터 세트로 분리하고(Fig. 1a)) 전체 데이터 세트와 함께 총 네 종류의 데이터 세트로 정리하였다. 밸브의 유형에 대한 라벨링은 연구에 참여한 기업측 전문가의 지식을 통해 선택 및 적용이 이루어졌다. 기업에서 제시한 보안 규정으로 인해 밸브 데이터의 출처인 세부 공정 관련 정보, 밸브 라벨링 기준, 그리고 데이터 원본 등의 중요한 영업 기밀은 본 연구에서 공개할 수 없음을 미리 밝히고자 한다. 생산량 및 수율 예측[20,21], 물성 예측[22-24], 온도 예측[25,26] 등등 여러 분야에서 활발하게 사용되고 있는 ANN, Random Forest, XGBoost, 그리고 CatBoost 네 가지 머신러닝 알고리즘을 활용한 모델들을 개발하고(Fig. 1b)) 각각의 데이터 세트에 대한 성능 평가

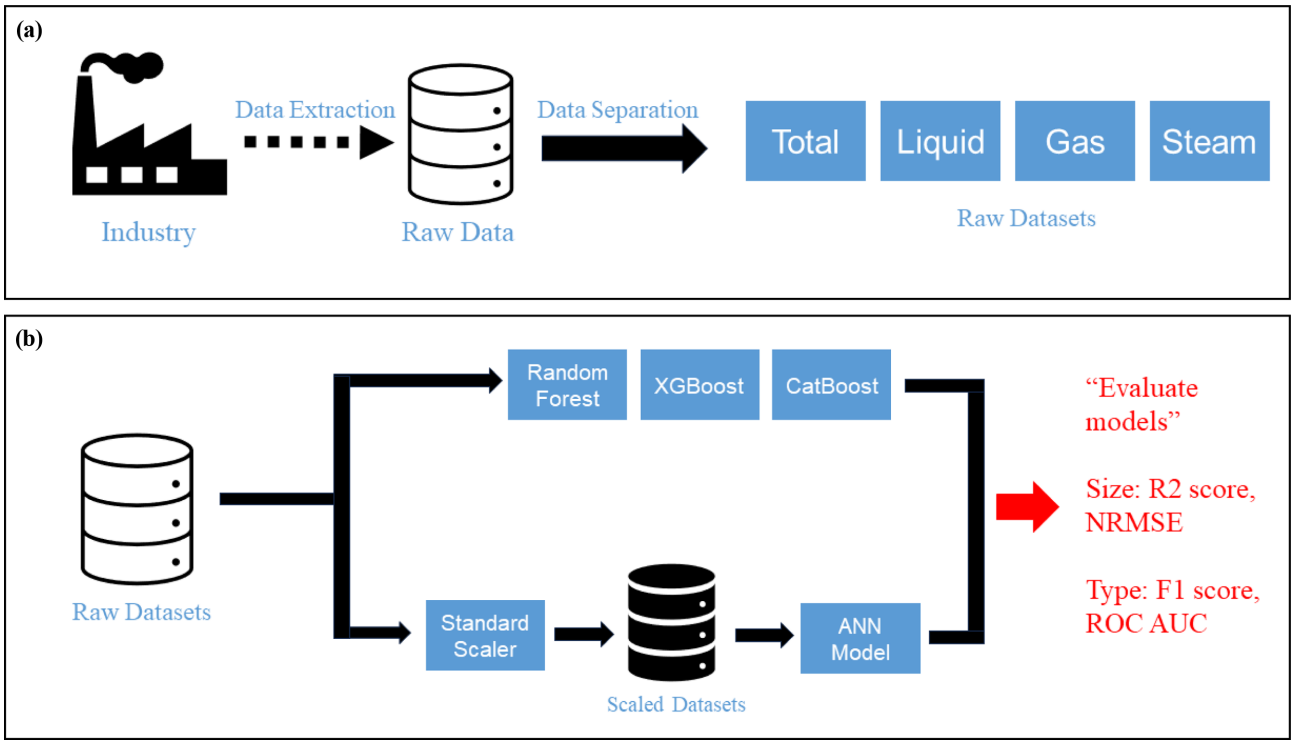

Fig. 1. Overview of the research.

를 진행했다.

본 논문 구성은 다음과 같다. 논문 2장에서는 이론적 배경과 모델 설명을 다룬다. 데이터 전처리 과정과 사용된 기법에 관한 이론적 내용을 서술한 뒤 본 연구에서 활용한 네 가지 알고리즘 ANN, Random Forest, XGBoost, 그리고 CatBoost의 작동 원리를 설명한다. 3장에서는 연구 결과에 대한 분석을 진행한다. 모델 평가 지표에 대한 이론적 설명을 바탕으로 각 모델이 밸브 사이즈와 유형에 대하여 어느 정도의 예측 성능을 나타내는지 비교 분석했다. 추가적으로 각 유체 상 별 모델이 전체 데이터에 대한 모델 대비 어떠한 성능 차이가 있는지 확인하였다. 4장에서는 지금까지의 내용을 정리하여 결론을 내리고 본 연구의 의의와 가치를 다룬다.

2. 예측모델 개발

2-1. 데이터 전처리

총 2,206개의 밸브 데이터를 담고 있는 원본 데이터 세트는 과 같이 15개의 입력 변수와 3개의 출력 변수로 구성되어 있다.

예측 대상이 밸브 사이즈와 유형이기에 Sizing Coefficient, 즉, C_v는 예측 대상으로 하지 않았다. 밸브 사이즈는 최소 0.5인치에서부터 최대 36인치까지로 이루어져 있다. 밸브 유형은 Globe, Angle, Ball, Butterfly의 네 가지이며 순서대로 1부터 4의 라벨을 부여했다. Globe 유형이 1,868개로 전체 데이터의 84.7%를 차지하며 Angle이 2.8%, Ball이 5%, Butterfly가 7.5%로 나머지를 구성한다.

유체 상 별 훈련 모델의 성능에 대한 사례 연구를 진행하기 위해 해당 데이터 세트를 상 별로 나누는 과정이 진행되었다. 전체 데이터를 구성하는 Process Fluid는 액체 데이터 1,303개, 기체 데이터 787개, 스팀 데이터 116개로 구성되어 있다.

추후 서술할 네 개의 알고리즘 중 Tree 기반 모델인 Random Forest, XGBoost, 그리고 Catboost와는 달리, 인공 신경망 (ANN) 모델의 경우, 알고리즘 내부에 존재하는 경사하강법이 크기가 큰 변수에 영향을 많이 받고 크기가 작은 변수는 거의 무시하므로 데이터 정규화 과정이 필요하다.

본 연구에서는 Standard Scaling을 이용하여 데이터를 정규화 하였으며, 정규화 과정은 수식 (2)와 같다.

$$x_{new} = \frac{x - x_{mean}}{\sigma} \quad (2)$$

Table 1. Lists of input variables and output variables

Input Variable	
Volumetric Flow Rate	Nm^3/h
Mass Flow Rate	tonne/h
Inlet Pressure	kg/cm^2g
Pressure Change	kg/cm^2
Inlet Temperature	°C
Inlet fluid density	kg/m^3
Dynamic Viscosity	cP
Vapor Pressure	kg/cm^2a
M / Gg M	
Specific heats ratio	
Design Temperature	°C
Design Pressure	kg/cm^2g
Line in	Inch
Line Out	Inch
Process Fluid	Liquid, Gas, Steam
Output Variable	
Sizing Coefficient	
Size	Inch
Body Style	Globe, Angle, Ball, Butterfly

x_{mean}은 x값의 평균 σ는 x값의 표준편차, 그리고 x_{new}는 정규화 이후의 x값을 의미한다.

Standard Scaling이 적용된 입력 변수는 비슷한 크기로 정규화되기 때문에 ANN의 경사하강법이 모든 입력 변수에 균등하게 영향을 받게 된다.

2-2. 모델 알고리즘

2-2-1. 인공 신경망(ANN)

ANN은 Artificial Neural Network의 약어로 Multi-Layer Perceptron (MLP)라고도 한다. 하나 이상의 노드로 이루어진 레이어를 층층이 쌓아서 입력 변수와 출력 변수 사이의 패턴을 학습시키는 이 모델은 생명체의 신경망을 모방한 구조를 지니고 있다. 레이어는 입력층, 은닉층, 출력층으로 나뉜다(Fig. 2). 은닉층은 실질적인 학습을 담당하는 레이어들로 외부에 직접적으로 드러나지 않는다. 노드는 정보의 수신 및 출력 역할을 하며, 이 과정에서 가중치 업데이트와 활성화 함수를 통해 정보들의 복잡한 상관관계를 수학적으로 반영한다. 또한, 각 노드들은 인접한 레이어 안의 모든 노드와 연결되어 있으며 초반 레이어가 간단한 패턴을 학습하면 후반 레이어가 복잡한 패턴을 학습하는 방식이다.

노드는 전달되는 입력 벡터인 χ와 가중치 벡터 w 및 편향 b를 이용해 데이터의 관계를 수학적으로 반영하며(Fig. 3), 이는 수식 (3)과 같다.

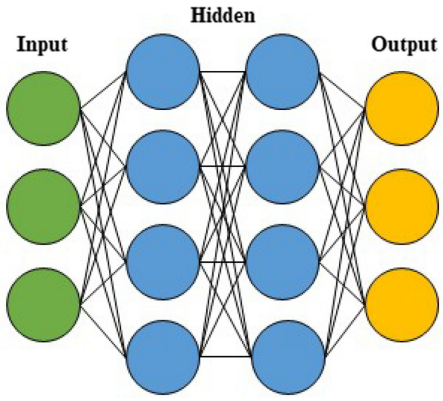

Fig. 2. Schematic figure of ANN model.

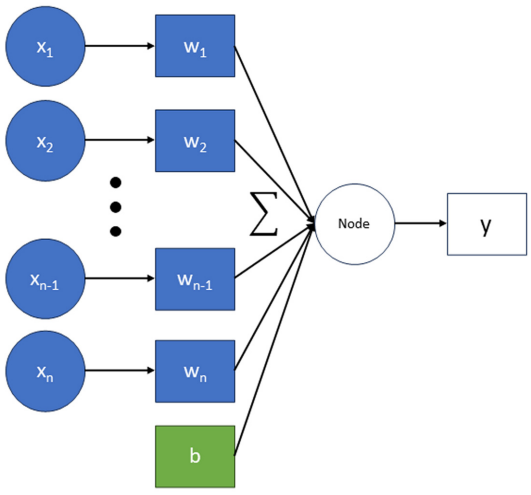

Fig. 3. Calculation principle of node output.

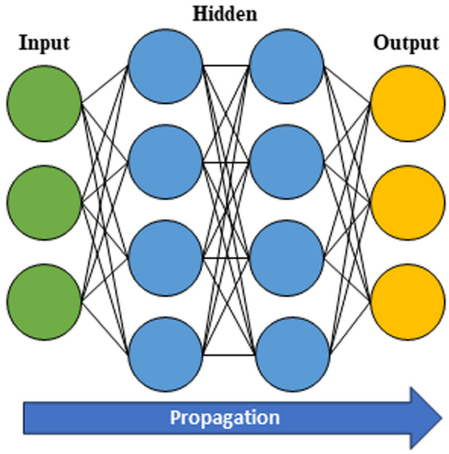

Fig. 4. Forward propagation of ANN model.

$$y = w^T x + b \tag{3}$$

w는 노드의 가중치로 이루어진 벡터이며 χ는 노드에 입력되는 변수들의 벡터다. b는 노드의 편향이며 두 벡터의 곱과 b를 합한 스칼라 값이 노드의 출력 y이다.

출력 y는 같은 레이어의 다른 노드들의 출력과 함께 다음 레이어의 입력 벡터 χ가 된다. 이러한 과정이 입력층에서부터 시작해 출력층까지 순서대로 전파된다(Fig. 4).

이후 전파가 끝나면 출력층에서 예측 결과를 내놓고 이를 실제 값과 비교하여 생긴 오차를 역전파하여 노드의 가중치를 수정한다(Fig. 5). 이 때 수정하는 정도는 오차에 대한 경사하강법을 통해 이루어진다. 경사하강법을 통해 최적화가 중단되는 지점에 도달하면 모델의 훈련이 끝난다.

2-2-2. Random Forest (RF)

Random Forest 알고리즘은 Decision Tree라는 간단한 알고리즘을 기반으로 한다. Decision Tree는 각각의 노드에서 조건을 제시하고 그에 따라 결과를 분리하여 다음 노드로 넘기는 기법이다(Fig. 6). 더 이상 분리할 수 없는 단계의 노드를 '잎(leaf)'라고 부르며 잎에 도달하면 그 값을 결과로 제시하게 된다.

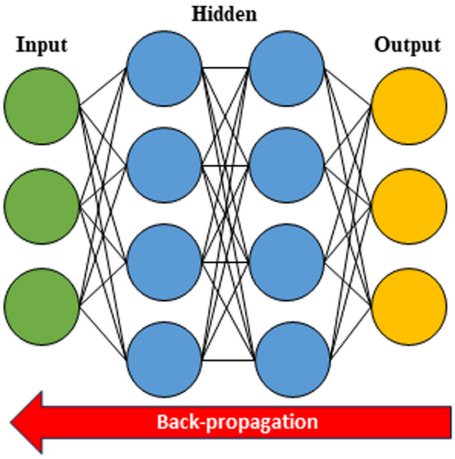

Fig. 5. Backward Propagation of ANN model.

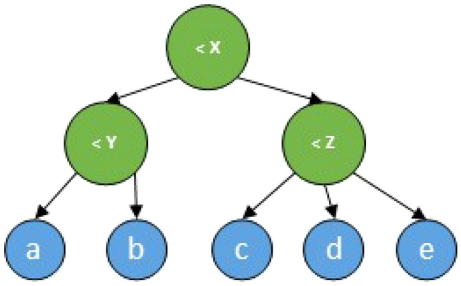

Fig. 6. Schematic illustration of Decision Tree model.

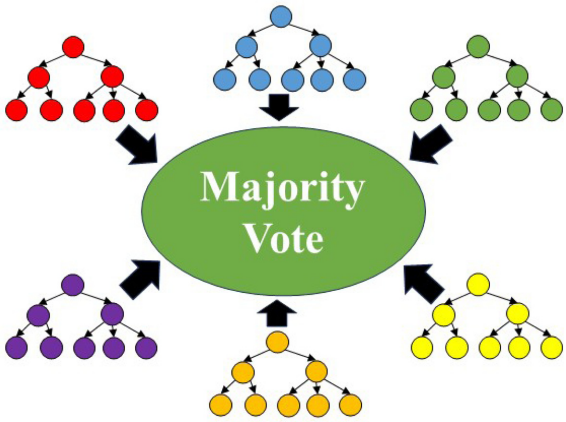

Fig. 7. Schematic illustration of Random Forest model.

Decision Tree는 주어진 데이터를 가장 잘 분리하도록 분류 기준을 설정하게 훈련된다. 이 알고리즘만으로 충분한 경우도 있으나 구조상 높은 분산을 지니는 특성으로 인해 종종 낮은 예측 성능이 나오는 문제가 있다. 이를 보완하는 두 가지 알고리즘이 Random Forest와 Boosting이다.

Random Forest는 Decision Tree를 여러 개 생성하여 그 결과로 다수결 투표를 하거나(분류) 평균을 사용하여(회귀) 분산의 영향을 줄이고 예측 성능을 높이는 모델이다(Fig. 7). 주어진 데이터를 특성이나 행을 기준으로 샘플링하고 이에 대해 Decision Tree들을 훈련시킨다. 이 때의 Decision Tree 훈련 데이터는 Bootstrap 알고리즘을 통해 무작위적으로 선정되기 때문에 Random Forest를 구성하는 Decision Tree들 사이의 다양성을 유지할 수 있다.

2-3. XGBoost

XGBoost는 Boosting 알고리즘의 일종으로 Decision Tree가 연속적으로 이어지는 방식으로 이루어져 있다. Decision Tree를 훈련시키고 예측 결과에 존재하는 오차 정보를 다음 Decision Tree에 전달한다(Fig. 8). 이 오차는 훈련 데이터에 대한 가중치로 작용하거나(AdaBoost 계열) 잔여 오차 자체를 새로운 훈련 라벨로 사용하여(Gradient Boost 계열) 오차를 줄이는 방향으로 훈련된다. 이 과정을 반복하면 높은 예측 성능을 얻어낼 수 있다.

XGBoost는 Gradient Boosting 알고리즘 중에서도 뛰어난 성능을 보여주는 모델 중 하나다. 일반적인 Gradient Boosting에 비해 좋은 성능을 지니는 이유는 세부적인 내부 알고리즘 차이에서 기인한다.

XGBoost [27]는 eXtreme Gradient Boosting의 준말로 정규화 기법을 적용하여 기존의 Gradient Boosting보다 예측 성능이 높고 속

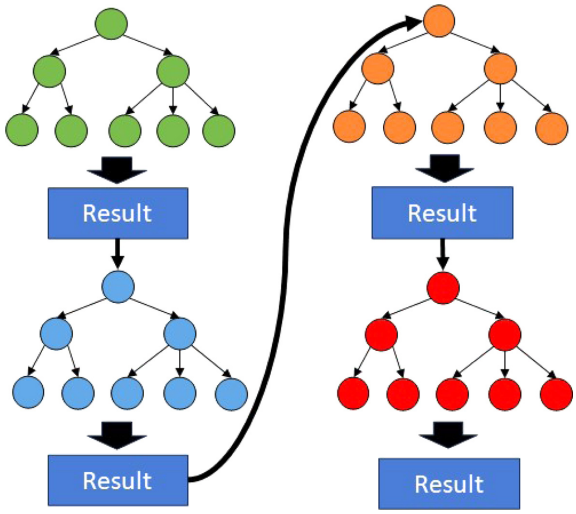

Fig. 8. Schematic illustration of Boosting Algorithm.

도가 빠르다. XGBoost의 목적함수는 다음의 수식 (4)로 표현된다.

$$\text{obj}^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \omega(f_k) \tag{4}$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

예측 대상의 개수는 n개이며, t는 몇 번째 Boosting 과정인지를 의미한다. $l(y_i, \hat{y}_i^{(t)})$ 는 오차 함수이며 $\omega(f_k)$는 정규화 함수이다. y_i는 실제값이며 $\hat{y}_i^{(t)}$ 는 t 번째 Boosting의 예측값이다. $\hat{y}_i^{(t)}$ 는 수식 (5)와 같이 계산되며 x_i는 예측 대상을, $f_k(x_i)$는 예측 대상에 대한 점수를 의미한다.

2-4. CatBoost

XGBoost와 마찬가지로 CatBoost는 일반적인 Gradient Boosting 알고리즘보다 뛰어난 성능을 보여주는 기법이다. CatBoost 알고리즘[28]은 자체적으로 카테고리 변수를 연속 변수로 변환하고 결측치를 처리할 수 있기에 여러 변수가 혼재된 상황에서도 잘 작동한다. 다른 Boosting 기법과의 핵심적인 차별점은 특유의 Decision Tree 생성 방식에 있다.

CatBoost는 가능한 객체 범위에 따라 주어진 데이터를 양자화하고 작은 데이터 세트로 분할한다. 양자화란 실수형(float) 변수를 정수형(int) 변수로 만드는 것을 의미하며 알고리즘 처리 속도에 기여한다. 이 양자화 분할 데이터는 추후 Decision Tree 구조를 결정하는데 사용된다. 이 때 사용된 분할 기준을 Decision Tree 노드의 분할 기준으로 활용하고 분할된 데이터 세트를 Tree의 말단인 leaf로 적용한다. 이러한 Decision Tree를 Bootstrap 알고리즘으로 무작위 데이터 세트를 선정하여 다수 만들고 비용 함수(cost function)를 적용해 가장 작은 비용 값을 가지는 Tree를 사용한다. 이 때의 비용 함수는 L2 정규화 함수에 기반한다.

$$L2 = -\sum_{i} w_i (a_i - g_i)^2 \tag{6}$$

w_i는 가중치, g_i는 오차 함수의 기울기(Gradient), 그리고 a_i는 결과값을 의미한다.

Tree의 선정에 사용되는 L2 정규화 함수는 다음과 같다.

$$S(a,g) = -\sum_i w_i(a_i-g_i)^2 = -\left[\sum_{i:x_{i,j}\le c} w_i(a_L-g_i)^2 + \sum_{i:x_{i,j}>c} w_i(a_R-g_i)^2\right] \tag{7}$$

leaf의 분할 기준이 c이며 i는 대상 데이터의 인덱스 j는 특성 변수의 인덱스에 해당한다. $x_{i,j}$는 i 인덱스의 대상 데이터의 j 특성 변수 값이다. 기준 c보다 작은 왼쪽 leaf의 결과값을 a_L, 큰 오른쪽 leaf의 결과값을 a_R라고 한다.

최적의 Decision Tree는 a_L과 a_R의 가중 평균을 통해 얻어진다. 두 값의 가중 평균(Weighted average)을 구하면 다음과 같다.

$$\bar{a}_L = \frac{\sum_{i:x_{i,j}\le c} w_i g_i}{\sum_{i:x_{i,j}\le c} w_i} \tag{8}$$

$$\bar{a}_R = \frac{\sum_{i:x_{i,j}>c} w_i g_i}{\sum_{i:x_{i,j}>c} w_i} \tag{9}$$

식 (7)에 (-)가 곱해져 있으므로 (7)의 최소값은 대괄호 안의 식이 최대일 때 구해진다. 식 (8), (9)를 이용해 해당 조건을 만족하는 j, c를 다음 식 (10)와 같이 도출할 수 있다.

$$j^*, c^* = argmax_{j,c}\left(\bar{a}_L^2 * \sum_{i:x_{i,j}\le c} w_i + \bar{a}_R^2 * \sum_{i:x_{i,j}>c} w_i\right) \tag{10}$$

식 (6)~(10)는 Decision Tree의 깊이가 1을 기준으로 하며, 깊이가 증가하면 (11)~(13)을 적용하게 된다.

$$S(a,g) = \sum_{lf\in leaf} S(a_{lf}, g_{lf}) \tag{11}$$

$$j^*, c^* = argmax_{j,c}(S(lf_L) + S(lf_R) - S(lf_{before\ split})) \tag{12}$$

$$S(a,g) = \sum_{lf\in leaf} S(lf) \tag{13}$$

식 (11)은 모든 leaf에 대한 비용 함수의 합이 최종 비용 함수가 됨을 의미한다. (12)은 leaf가 분리되는 기준이 tree 깊이의 증가에 따라 변경되는 것을 나타낸다. CatBoost에 의해 생성되는 Tree는 대칭적이기 때문에 (11)은 (13)로 간소화할 수 있다.

2-5. 모델 개발 환경

본 연구에선 ANN, Random Forest, XGBoost, CatBoost를 적용하여 밸브의 사이즈와 유형 예측 모델을 개발하는 과정을 진행했다. 상세 하드웨어 환경은 Table 2와 같다.

Table 2. Hardware and software environment of model development

Hardware Environment	
CPU	12th Gen Intel(R) Core(TM) i7-12700, 2100 Mhz, 12 Cores, 20 Logical processors
OS	Microsoft Windows 11 Home, 10.0.22621 Build 22621
RAM	32GB
GPU	NVIDIA GeForce RTX 3060
Software Environment	
Python	3.9.16v
Tensorflow	2.10.1v
cudatoolkit	11.2.2v
cudnn	8.1.0.77v
scikit-learn	1.2.2v
Xgboost	1.7.5v
Catboost	1.2v

2-6. 모델 하이퍼파라미터 튜닝

하이퍼파라미터는 학습의 진행에 따라 조정되는 일반적인 파라미터와 달리 사용자가 사전에 설정하기 때문에 학습에 영향을 받지 않는 파라미터를 말한다. 일반적으로 하이퍼파라미터는 예측 성능과 과대 적합에 대한 민감도 등 모델의 성능에 큰 영향을 미치기 때문에 적절한 하이퍼파라미터의 선정이 중요하다.

본 연구에서는 그리드 탐색(Grid Search) 기법을 적용하여 하이퍼파라미터를 튜닝하였다. 이 때, ANN은 레이어 개수와 레이어 별 노드 개수를 튜닝 대상으로 삼았고 Tree 계열 모델들은 Tree의 개수와 최대 깊이를 중심으로 튜닝하였다. 탐색하는 그리드의 간격은 휴리스틱 방법을 통해 정해졌으며 전체 데이터 세트에 대해 훈련한 모델의 주요 성능 평가 지표(회귀: R^2, 분류: F1)를 기준으로 최적의 하이퍼파라미터를 탐색하였다.

탐색 결과, ANN은 3개의 Hidden Layer에 각각 200, 100, 100개의 노드가 있는 모델이 선정되었으며 Tree 계열 모델은 최대 깊이는 3인 상태에서 CatBoost와 XGBoost는 300개의 tree, Random Forest는 350개의 tree를 가지는 모델이 최고의 성능을 제시하였다. 따라서, 본 연구에서는 해당 하이퍼파라미터를 적용한 모델들을 기준으로 최종적인 모델 성능 평가를 진행하였다.

3. 결과 및 고찰

3-1. 모델 평가

전체, 액체, 기체, 스팀의 4개 데이터 세트에 대한 모델 성능 평가는 Scikit-learn의 RepeatedKFold를 이용한 100번의 검증 결과의 평균과 표준 편차를 통해 진행되었다. 이는 데이터를 K개의 Fold로 나누고 그 중 한 개를 검증 데이터 나머지를 훈련 데이터로 사용하여 모델의 성능을 교차 검증하는 기법으로 일반적으로 과대적합을 고려한 성능 평가에 자주 활용된다. 과대 적합이 발생할 경우 교차 검증 결과 예측 성능의 평균 값이 하락하고 표준 편차가 증가하는 것을 통해 과대 적합의 발생 여부를 판단 가능하다. 본 연구에서는 5개의 Fold로 나누어 20%의 검증 데이터와 80% 훈련 데이터로 평가하였으며, 이를 20번 반복하여 총 100개 결과를 분석하여 모델을 비교하였다.

3-1-1. Valve size prediction

사이즈 예측 모델은 회귀 모델 평가에 빈번히 사용되는 Normalized Root Mean Squared Error (NRMSE)와 결정 계수(Coefficient of determination, R^2)를 적용했다.

NRMSE는 Root Mean Squared (RMSE)를 데이터의 평균값으로 정규화한 값으로 각각 다음의 식을 통해 표현된다.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_{true,i}-y_{predict,i})^2}{N}} \tag{14}$$

$$\mathrm{NRMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_{true,i}-y_{predict,i})^2}{\bar{y}_{true}}} \tag{15}$$

N 개의 데이터에 대해 $y_{predict,i}$는 데이터 i의 예측값, $y_{true,i}$는 i번째 데이터 i의 실제값, 그리고 $\bar{y}_{true}$ 는 실제값의 평균을 의미한다.

Table 3. K-fold Cross validation results for the size prediction models

	Total Fluid		Liquid		Gas		Steam	
	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE
ANN	0.9751 ± 0.0067	0.1474 ± 0.0247	0.9622 ± 0.0266	0.1696 ± 0.0478	0.9806 ± 0.0135	0.1372 ± 0.0388	0.9578 ± 0.0472	0.0670 ± 0.0327
Random Forest	0.9887 ± 0.0090	0.0555 ± 0.0432	0.9766 ± 0.0128	0.0852 ± 0.0550	0.9906 ± 0.0053	0.0492 ± 0.0277	0.9721 ± 0.0256	0.0316 ± 0.0326
XGBoost	0.9908 ± 0.0067	0.0409 ± 0.0310	0.9741 ± 0.0359	0.0957 ± 0.1557	0.9922 ± 0.0065	0.0407 ± 0.0332	0.9701 ± 0.0298	0.0332 ± 0.0352
CatBoost	0.9922 ± 0.0065	0.0407 ± 0.0332	0.9860 ± 0.0045	0.0489 ± 0.0144	0.9930 ± 0.0040	0.0362 ± 0.0193	0.9691 ± 0.0383	0.0333 ± 0.0404

RMSE는 데이터의 크기에 영향을 받기 때문에 유체 상 별 데이터의 성능 평가에는 적합하지 않다. NRMSE는 이러한 데이터 크기의 영향을 줄여 비교적 공정한 성능 비교가 가능하다는 장점이 있다.

결정 계수는 선형 모형이 데이터를 얼마나 표현하는지를 나타내는 지표로 회귀 모델에서는 주로 x 축을 예측값, y 축을 실제값으로 하여 기울기가 1인 직선에 대해서 구해지는 값을 사용한다. 1에 가까울수록 모델 예측 성능이 뛰어나며 0에 가까울수록 성능이 저조함을 알 수 있다. 결정 계수는 다음의 수식과 같이 구해진다.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{true,i} - y_{predict,i})^2}{\sum_{i=1}^{N}(y_{true,i} - \bar{y}_{true})^2} \quad (16)$$

NRMSE와 R^2를 평가 지표로 하여 네 개의 데이터 세트에 대해 훈련한 ANN, Random Forest, XGBoost, CatBoost 모델의 사이즈 예측 성능은 다음의 Table 3과 같다.

전체, 액체, 기체 데이터 세트에 대해선 CatBoost가 순서대로 0.99216±0.00654, 0.98602±0.00457, 0.99300±0.00397의 결정 계수로 가장 뛰어난 성능을 보여주었다. 스팀의 경우엔 Random Forest가 0.97207±0.02563의 결정 계수로 가장 좋은 모델임을 확인했다. 한편, 높은 평균 R^2 값과 낮은 표준편차를 통해 과대적합이 발생하지 않았다고 해석될 수 있다.

유체 상 별 데이터에 대한 성능의 경우 기체 데이터 모델이 전체 데이터 모델보다 성능이 좋았으나 액체와 스팀 데이터의 경우에는 오히려 성능이 저하되는 모습을 보여주었다.

Fig. 9-12는 각각의 데이터 세트에 따른 예측 모델에 대한 Parity plot이다. x축은 실제값, y축은 예측값이며 기울기 1인 직선에 가깝게 분산도가 분포할수록 성능이 뛰어남을 의미한다.

Parity plot에서 6 인치에서의 큰 분산으로 보아 예측 모델은 6 인치에 대한 예측 성능이 가장 불안정한 것으로 추정된다. 그 다음으로는 8인치, 3인치, 4인치 순서대로 예측 성능이 낮은 것으로 보인다. 6인치가 8인치로 예측되는 경우와 그 반대가 자주 발견되는 것으로 보아 6인치와 8인치 밸브의 공정 조건이 유사한 것이 원인으로 보인다. 3인치와 4인치의 관계도 그러한 관점에서 해석 가능할

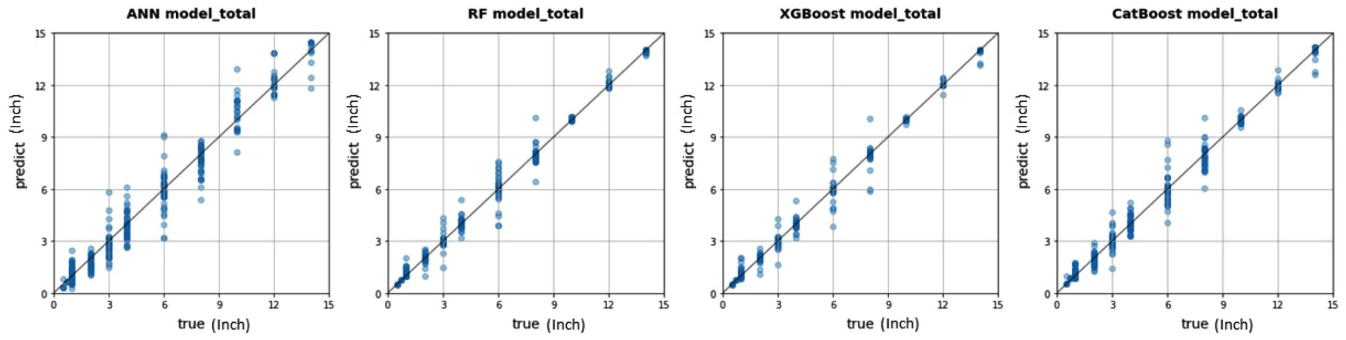

Fig. 9. Size prediction result for Total dataset.

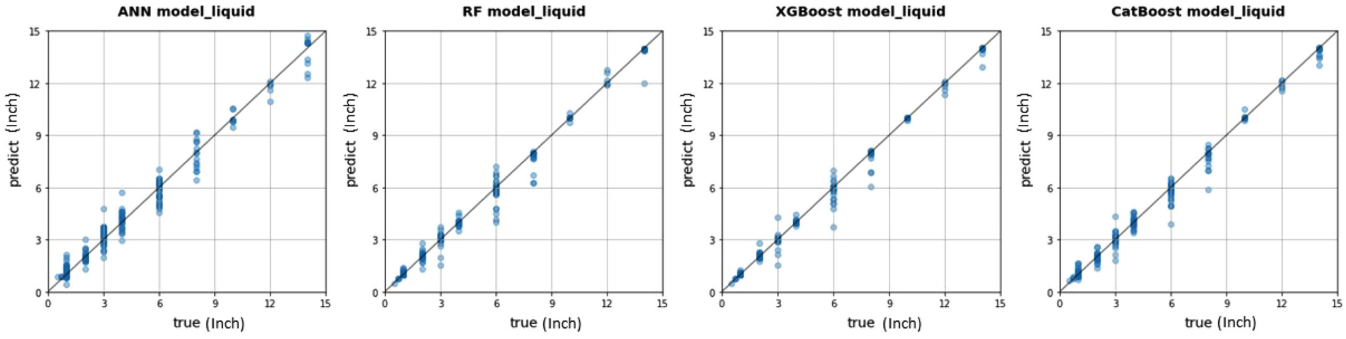

Fig. 10. Size prediction result for Liquid dataset.

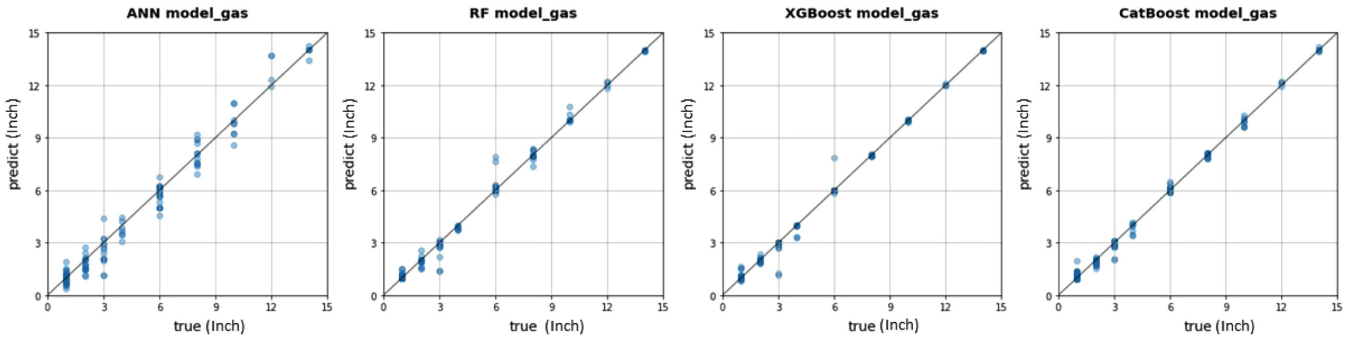

Fig. 11. Size prediction result for Gas dataset.

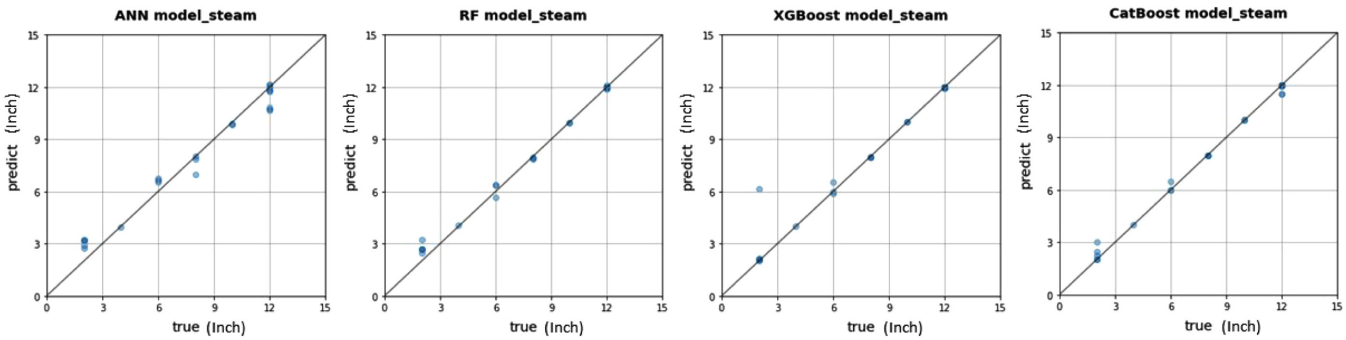

Fig. 12. Size prediction result for Steam dataset.

Table 4. Performance evaluation on test datasets for size prediction model

	Total Fluid		Liquid		Gas		Steam	
	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE
ANN	0.9856	0.1136	0.9758	0.1331	0.9903	0.1045	0.9615	0.0986
Random Forest	0.9936	0.0757	0.9861	0.1009	0.9974	0.0536	0.9916	0.0461
XGBoost	0.9888	0.1002	0.9918	0.0775	0.9981	0.0457	0.9583	0.1027
CatBoost	0.9942	0.0720	0.9893	0.0882	0.9988	0.0372	0.9953	0.0346

것으로 판단된다.

한편 Table 4는 액체와 스팀에서의 교차 검증 결과와 다른 결과를 보여준다. 액체에서는 XGBoost가 가장 좋고, CatBoost 가 그 다음으로 높은 성능을 보여주었다. 이는 XGBoost의 교차 검증 표준 편차가 상대적으로 큰 것을 고려하면 해당 시험 데이터 세트의 분포가 XGBoost에 유리하게 분할되었기 때문으로 해석된다. 같은 이유로, 스팀에서도 Random Forest가 아닌 CatBoost가 가장 좋은 성능을 제시한 것으로 판단된다.

3-1-2. Valve Type

밸브 유형 예측 모델은 수치를 예측하는 사이즈 예측과는 달리, 4개의 유형으로 분류하는 모델이다. 따라서, 밸브 유형 예측 모델은 분류 모델을 활용하였으며, 모델의 예측 결과는 과 같이 네 가지로 분류 가능하다. 분류된 TP, TN, FP, 그리고 FN은 수식 (22)-(26)와 같이 다양한 모델 성능 평가 지표 계산에 사용된다.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{18}$$

	Real: Positive	Real: Negative
Prediction: Positive	True Positive(TP)	False Positive(FP)
Prediction: Negative	False Negative(FN)	True Negative(TN)

Fig. 13. Four kinds of predictions explanation with binary confusion matrix.

$$\text{Recall} = \frac{TP}{TP+FN} = \text{TPR} \tag{19}$$

분류 모델에서 가장 흔히 사용되는 평가 지표는 정확도(Accuracy) (17)다. 그러나, 정확도는 데이터가 불균형한 경우 가장 많은 데이터로 무조건적 예측을 하게 훈련될 가능성이 크다. 따라서, 본 연구에서는 데이터 불균형의 영향을 줄이기 위해 F1 점수를 밸브 유형 예측에 대한 평가 지표로 사용하였다. 이 지표를 사용하면 단순히 맞은 예측의 개수를 늘리도록 훈련되는 것이 아닌 예측과 실제 사이의 일치율을 높이도록 모델을 훈련할 수 있다.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (20)$$

F1 점수는 정밀도(Precision)와 재현율(Recall)의 조화 평균이며 수식 (20)로 표현된다. 정밀도와 재현율 사이엔 트레이드 오프가 존재하며 조화 평균을 적용하였기 때문에 정밀도와 재현율 중 더 작은 값에 영향을 많이 받게 된다. 이는 양성 예측 중 실제 양성 비율인 정밀도와 실제 양성 중 양성 예측 비율인 재현율 중 하나에 편향되지 않게 막는 효과가 있다.

F1 점수를 평가 지표로 했을 때 데이터 세트에 따른 분류 모델 별 예측 성능은 다음의 Table 5와 같다.

F1 점수를 기준으로 했을 때 CatBoost가 모든 데이터 세트에 대해서 가장 뛰어난 예측 성능을 보여주었다. 유체 상 별 성능의 관점에서는 밸브 사이즈 예측과는 반대의 결과가 나왔다. 사이즈 예측 모델에서는 기체 데이터에 대한 모델만 전체 데이터에 대한 모델보다 성능이 좋았던 데 반해 이번에는 반대로 기체 데이터에 대한 모델의 성능이 전체 데이터에 대한 모델보다 낮고 액체와 스팀 데이터에 대한 모델의 성능이 더 높은 것을 확인할 수 있었다. 한편, 모델 과대적합의 관점에서 보았을 때 스팀 데이터 세트에 대한 ANN 모델의 낮은 성능을 보아 ANN은 과대적합이 진행되었다고 해석할 수 있다. 그 외의 데이터의 경우 높은 평균 F1 점수와 낮은 표준 편차를 통해 과대적합이 발생하지 않았음을 확인할 수 있다.

Fig. 14-17은 시험 데이터 세트에 따른 네 가지 모델의 예측에 대한 혼동 행렬(confusion matrix)다. 실제값인 True label와 예측값인 Predicted label을 각각 행과 열로 사용하고 있으며 대각선에 해당하는 것이 올바르게 예측된 값의 비율이다. Fig. 13에서는 참과 거짓으로만 표현되었으나 Fig. 14-16 밸브의 유형 네 가지에 대하여 4×4의 크기로 제시되었다. 예외적으로 스팀 데이터 세트에 해당하는 Fig. 17은 3번 유형의 밸브가 포함되지 않아 3×3의 크기의 행렬이다. 행렬 내부의 값은(Predicted label의 개수/True label의 개수)로 정규화되었다.

Table 5. K-fold Cross validation results for the type prediction models

F1 Score	Total Fluid	Liquid	Gas	Steam
ANN	0.9307 ± 0.0226	0.9464 ± 0.0193	0.9382 ± 0.0397	0.6589 ± 0.0000
Random Forest	0.9538 ± 0.0209	0.9563 ± 0.0278	0.9566 ± 0.0594	1.0000 ± 0.0000
XGBoost	0.9513 ± 0.0217	0.9601 ± 0.0292	0.9545 ± 0.0588	0.9956 ± 0.0259
CatBoost	0.9577 ± 0.0196	0.9626 ± 0.0268	0.9576 ± 0.0586	1.0000 ± 0.0000

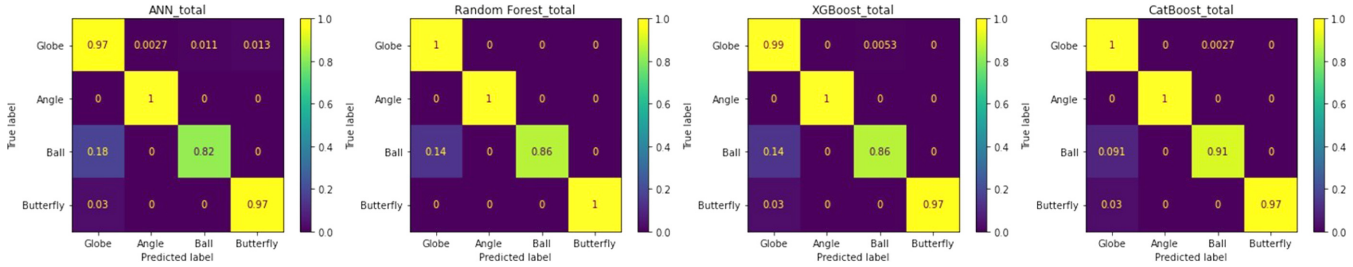

Fig. 14. Confusion matrix results for Total dataset type prediction.

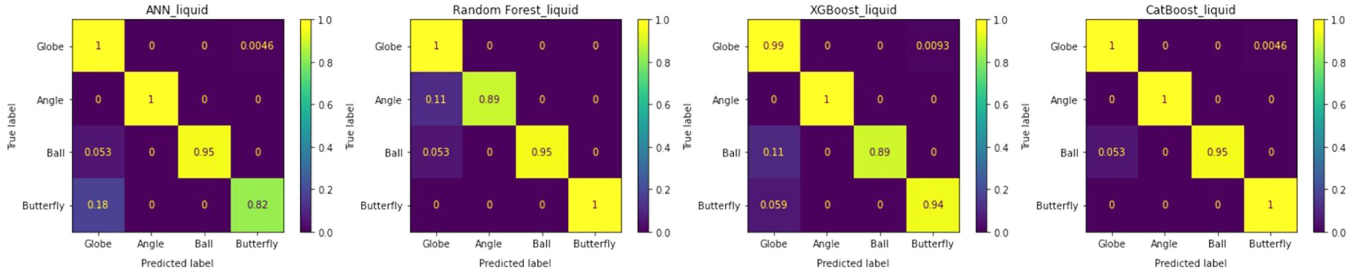

Fig. 15. Confusion matrix results for Liquid dataset type prediction.

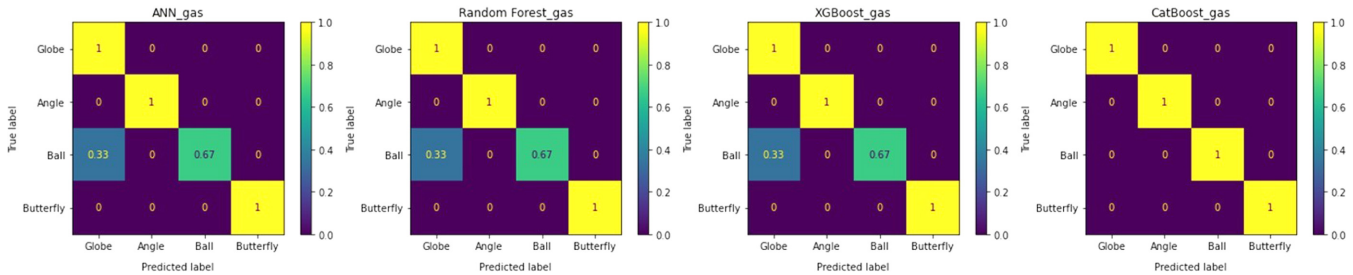

Fig. 16. Confusion matrix results for Gas dataset type prediction.

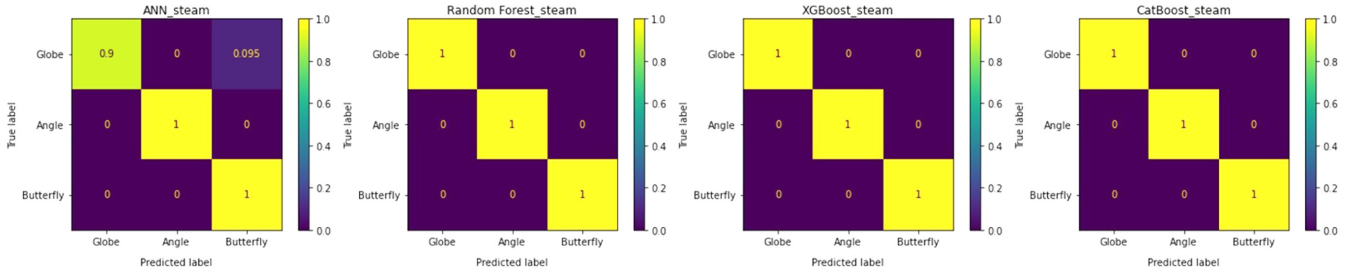

Fig. 17. Confusion matrix results for Steam dataset type prediction.

Table 6. Performance evaluation on test datasets for type prediction model

F1 Score	Total Fluid	Liquid	Gas	Steam
ANN	0.9292	0.9759	0.9490	0.8166
Random Forest	0.9581	0.9774	0.9491	1.0000
XGBoost	0.9651	0.9618	0.9491	1.0000
CatBoost	0.9774	0.9849	1.0000	1.0000

혼동 행렬을 분석해보면 전체적으로 F1 점수에 의거한 해석과 일치하는 것을 확인할 수 있다. 유체 상 별 모델에서 CatBoost가 전부 정확한 예측에 성공함을 보여주었고, 전체 데이터에 대해서도 가장 높은 예측 성공률을 확인하였다.

3번 유형에 해당하는 Ball 유형 밸브에 대한 예측 성능이 떨어지는 모습이 여러 혼동 행렬에서 관측되었다. 주로 1번 유형인 Globe로 예측되었음을 고려하면 85% 가량을 차지하는 Globe 데이터의 불균형 문제를 중심으로 여러 문제가 같이 작용한 것으로 보인다. Globe 유형 밸브와 Ball 유형 밸브 사이의 공정 조건 유사성이 겹치면서 생긴 문제일 가능성이 존재한다고 추정된다.

또한, 2번 유형에 해당하는 Angle 유형 밸브에 대한 예측 성능이 전반적으로 높게 나왔는데 이는 Angle 밸브가 사용되는 공정 조건이 나머지 밸브들과 유의미한 차이가 있다는 뜻으로 생각된다.

Table 6의 결과는 Table 5와 거의 유사함을 보여주는데, 이는 유체 상 별 최적 모델로 제시된 모델들이 과대 적합 없이 적절하게 훈련됨을 뒷받침하는 근거로 해석된다.

4. 결 론

본 연구에서는 주어진 공정 조건에 적합한 밸브의 사이즈와 유형의 선정에 소요되는 시간과 비용을 줄이기 위해 머신러닝 기반 예측 모델을 개발하였다. 이를 위해 ANN, Random Forest, XGBoost, 그리고 CatBoost를 적용한 예측 모델을 만들고. 이를 전체, 액체, 기체, 스팀 데이터 세트에 대해 훈련 및 평가하는 과정을 진행하였다. 사이즈 예측의 경우 결정 계수를 기준으로 하였을 때 전체 데이터에서는 CatBoost가 0.99216±0.00654로 최고의 성능을 보여주었다. 액체에서도 CatBoost가 0.98602±0.00457로 가장 뛰어났으나 전체 모델에 비해 성능이 낮았다. 기체에서도 마찬가지로 CatBoost가 0.99300±0.00397의 결정 계수로 가장 뛰어났으며 전체 모델에 비해서 높은 예측 성능을 제시했다. 스팀에선 Random Forest가 0.97207±0.02563의 결정 계수로 높은 성능을 얻을 수 있었다. 유형 예측의 경우 CatBoost가 모든 데이터에서 최고의 예측 능력을 보여주었다. 전체 데이터 모델은 0.95766±0.01959의 예측 성능을 보여주었고, 액체 데이터에선 0.96264±0.02677, 기체 데이터에선 0.95760±0.05863, 그리고 스팀 데이터에선 1.00000±0.00000의 F1 점수를 확인할 수 있었다. 사이즈 예측과는 반대로 액체에서의 예측 성능이 전체에서의 예측 성능보다 높았고, 기체에서의 예측 성능은 전체보다 낮게 측정되었다.

데이터의 불균형에도 불구하고 전반적으로 높은 예측 성능을 확보할 수 있었으나, 예측 모델이 잘못된 예측을 내리는 경향성이 일부 남아 있음을 Parity plot과 Confusion matrix를 통해 확인하였다. 이는 데이터 불균형이 영향을 미친 것으로 보이며 추가적인 데이터의 확보로 개선할 수 있으리라고 생각된다. 하지만 추가 데이터 확보가 불가능한 경우, 여러 Upsampling 기법을 이용하여 데이터 불균형 문제를 어느정도 해결 가능하다. 예를 들어 Random Upsampling이나 SMOTE Upsampling 등 상대적으로 비율이 적은 데이터를 증폭시키는 기법을 적용하여 데이터 불균형 문제 해결을 시도해볼 수 있다.

본 연구에서 제시한 머신러닝 모델을 활용하면 주어진 조건에 따른 적절한 밸브의 사이즈와 유형을 제시하여 전문가의 수고와 비용을 줄이고 효율적인 의사 결정을 도울 수 있을 것으로 기대된다. 또한, 공정 조건이나 파라미터에 변경 사항이 존재하면 기존처럼 C_v 계산에 많은 시간을 소모하는 대신 변경 내용을 모델에 입력하여 빠른 대응이 가능할 것이라 생각된다.

감 사

본 논문은 “AI 지능화기반 엔지니어링 예측 모델 개발(2023-11-0458)”의 지원으로 수행한 연구입니다.

References

1. Park, G., “How to Select Control Valve,” *HWAHAK KONGHAK*, **10**(3), 141-152(1972).
2. Driskell, Les. Control valve selection and sizing. 1st ed. North Carolina: Creative Services Inc; 1983.
3. IEC 60534-2-1 Mod: flow equations for sizing control valves, Switzerland, International Electro-technical Commission.
4. ISA75.01, Control Valve Sizing Equations, International Society of Automation.
5. Grace, A. and Frawley, P., “Experimental Parametric Equation for the Prediction of Valve Coefficient (C_v) for Choke Valve Trims,” *International Journal of Pressure Vessels and Piping*, **88**(2–3), 109-118(2011).
6. Long, C. and Guan, J., “A Method for Determining Valve Coefficient and Resistance Coefficient for Predicting Gas Flowrate,”

Experimental Thermal and Fluid Science, **35**(6), 1162-1168(2011).
7. Zhou, X.-M., Wang, Z.-K., Zhang, Y.-F., "A Simple Method for High-precision Evaluation of Valve Flow Coefficient by Computational Fluid Dynamics Simulation," *Advances in Mechanical Engineering*, **9**(7), 1-7(2017).
8. Lisowski, E. and Filo, G., "Analysis of a Proportional Control Valve Flow Coefficient with the Usage of a CFD Method," **53**, Part B, 269-278(2017).
9. Valdés, J. R., Rodríguez, J. M., Saumell, J, Pütz, T., "A Methodology for the Parametric Modelling of the Flow Coefficients and Flow Rate in Hydraulic Valves," *Energy Conversion and Management*, **88**, 598-611(2014).
10. Nguyen, Q. K. and Jung, K. H., "Experimental Study on Pressure Characteristics and Flow Coefficient of Butterfly Valve," *International Journal of Naval Architecture and Ocean Engineering*, **15**, (2023).
11. Al-Zaidi, B. M. and Ismaeel, A. J., "Effect of Hydraulic Characteristics on Fluid Transients Analysis under Different Types of Control Valves," *Journal of Ecological Engineering*, **23**(12), 111-123(2022).
12. Fu, W.-S. and Ger, J.-S., "A Concise Method for Determining a Valve Flow Coefficient of a Valve Under Compressible Gas Flow," *Experimental Thermal and Fluid Science*, **18**, 307-313(1999).
13. Boccardi, G., Bubbico, R. and Celata, G. P., "Geometry Influence on Safety Valves Sizing in Two-phase Flow," *Journal of Loss Prevention in the Process Industries*, **21**(1), 66-73(2008).
14. Mahalleh, VBS, YOLO-Based Valve Type Recognition and Localization, 2019 IEEE 6th International Conference On Industrial Engineering and Applications (Iciea), 37-40(2019).
15. Hlubek, N. and Baumann, M., Sebastian Heinze Florian Ostermaier, Using Machine Learning for Diaphragm Prediction in Solenoid Valves, IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA).
16. Roh, J., Park, H., Kwon, H., Joo, C., Moon, I., Cho, H., Ro, I. and Kim, J., "Interpretable Machine Learning Framework for Catalyst Performance Prediction and Validation with Dry Reforming of Methane," *Applied Catalysis B: Environmental*, **343**, 123454(2024).
17. Roh, J., Oh, S., Lee, D., Joo, C., Park, J., Moon, I., Ro, I., Kim, J., "Hybrid Quantum Neural Network Model with Catalyst Experimental Validation: Application for the Dry Reforming of Methane," *ACS Sustainable Chemistry & Engineering*, **12**(10), 4121-4131(2024).
18. Kwon, H., Oh, K. C., Choi, Y., Chung, Y. G. and Kim, J., "Development and Application of Machine Learning-based Prediction Model for Distillation Column," *International Journal of Intelligent Systems*, **36**(5), 1970-1997(2021).
19. Jeong, S., Joo, C., Lim, J., Cho, H., Lim, S. and Kim, J., "A Novel Graph-based Missing Values Imputation Method for Industrial Lubricant Data," *Computers in Industry*, **150**, 103937(2023).
20. Lee, J., Hong, S., Kim, J. and Moon, I., "Machine Learning-based Energy Optimization for on-site SMR Hydrogen Production," *Energy Conversion and Management*, **244**(15), 114438(2021).
21. Lim, J., Jeong, S. and Kim, J., "Deep Neural Network-based Optimal Selection and Blending Ratio of Waste Seashells as an Alternative to High-grade Limestone Depletion for SOX Capture and Utilization," *Chemical Engineering Journal*, **431**, Part 3, 133244(2022).
22. Joo, C., Park, H. and Kim, J., "Development of Physical Property Prediction Models for Polypropylene Composites with Optimizing Random Forest Hyperparameters," *International Journal of Intelligent Systems*, **37**(6), 3189-3771(2022).
23. Joo, C., Park, H., Kwon, H. and Kim, J., "Machine Learning Approach to Predict Physical Properties of Polypropylene Composites: Application of MLR, DNN, and Random Forest to Industrial Data," *Polymers*, **14**(17), 3500(2022).
24. Joo, C., Park, H. and Kim, J., "Data-driven Modeling for Physical Property Prediction of Polypropylene Composites Using Artificial Neural Network and Principal Component Analysis," *Computer Aided Chemical Engineering*, **51**, 1369-1374(2022).
25. Lee, Y., Choi, Y., Cho, H. and Kim, J., "Prediction of Distillation Column Temperature Using Machine Learning and Data Preprocessing," *Korean Chem. Eng. Res.*, **59**(2), 191-199(2021).
26. Joo, C., Park, H., Lim, J., Cho, H. and Kim, J., "Machine Learning-based Heat Deflection Temperature Prediction and Effect Analysis in Polypropylene Composites Using Catboost and Shapley Additive Explanations," *Engineering Applications of Artificial Intelligence*, **126**, Part A, 1801-1806(2022).
27. Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 785-794(2016).
28. Dorogush, A. V., Ershov, V. and Gulin, A., "CatBoost: Gradient Boosting with Categorical Features Support," Workshop on ML Systems at NIPS 2017.

Authors

Chanho Kim: Undergraduate researcher, Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yonsei-ro, Seodae-mun-gu, Seoul, Republic of Korea; kygon5801@yonsei.ac.kr
Minshick Choi: Undergraduate researcher, Energy Resources Upcycling Research Laboratory, Korea Institute of Energy Research, 152, Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea; als980@kitech.re.kr
Chonghyo Joo: PhD Student, Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea; hyo156@yonsei.ac.kr
A-Reum Lee: Researcher, Samsung E&A Co., Ltd., 26, Sangil-ro 6-gil, Gangdong-gu, Seoul, Republic of Korea; a_reum.lee@samsung.com
Yun Gun: Researcher, Samsung E&A Co., Ltd., 26, Sangil-ro 6-gil, Gangdong-gu, Seoul, Republic of Korea; gabriel.yun@samsung.com
Sungho Cho: Vice President, Samsung E&A Co., Ltd., 26, Sangil-ro 6-gil, Gangdong-gu, Seoul, Republic of Korea; s.h.cho@samsung.com
Junghwan Kim: Associate professor, Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yonsei-ro, Seodae-mun-gu, Seoul, Republic of Korea; kjh24@yonsei.ac.kr