ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 52, No.2: 185-199, June 2024
https://dx.doi.org/10.7469/JKSQM.2024.52.2.185

CNC 가공 공정 불량 예측 및 변수 영향력 분석

홍지수* · 정영진*† · 강성우**

* 인하대학교 산업경영공학과

Defect Prediction and Variable Impact Analysis in CNC Machining Process

Hong, Ji Soo* · Jung, Young Jin*† · Kang, Sung Woo*

* Department of Industrial Engineering, Inha University

ABSTRACT

Purpose: The improvement of yield and quality in product manufacturing is crucial from the perspective of process management. Controlling key variables within the process is essential for enhancing the quality of the produced items. In this study, we aim to identify key variables influencing product defects and facilitate quality enhancement in CNC machining process using SHAP(SHapley Additive exPlanations)

Methods: Firstly, we conduct model training using boosting algorithm-based models such as AdaBoost, GBM, XGBoost, LightGBM, and CatBoost. The CNC machining process data is divided into training data and test data at a ratio 9:1 for model training and test experiments. Subsequently, we select a model with excellent Accuracy and F1-score performance and apply SHAP to extract variables influencing defects in the CNC machining process.

Results: By comparing the performances of different models, the selected CatBoost model demonstrated an Accuracy of 97% and an F1-score of 95%. Using Shapley Value, we extract key variables that positively of negatively impact the dependent variable(good/defective product). We identify variables with relatively low importance, suggesting variables that should be prioritized for management.

Conclusion: The extraction of key variables using SHAP provides explanatory power distinct from traditional machine learning techniques. This study holds significance in identifying key variables that should be prioritized for management in CNC machining process. It is expected to contribute to enhancing the production quality of the CNC machining process.

Key Words: CNC Machining Process, Defect Prediction, Variable Impact Analysis, Machine Learning, SHAP

● Received 2 February 2024, 1st revised 7 March 2024, accepted 19 March 2024
† Corresponding Author(YoungjinJung@inha.edu)

* 본 논문은 인하대학교의 지원에 의해 연구되었습니다.

1. 서 론

공정 관리는 제품을 생산함에 있어 높은 수준의 제품 품질과 수율을 달성하고 효율적인 생산을 위해 필수적이다. 공정 관리 중 공정 조건 등의 주요 변수를 적절히 관리하는 것은 제품 품질 특성치를 일정 수준으로 유지할 수 있는 효율적인 방법이다(Lee, 2020)(Kim, Iljung et al., 2022). CNC(Computerized Numerical Control) 가공 공정은 미리 프로그래밍 된 컴퓨터 소프트웨어를 이용하여 공장의 도구나 기계를 제어하여 비행기나 자동차 등 복잡한 형상을 가진 제품들을 효과적으로 생산하는 제조 공정 기법이다(Kang et al., 2016). 기존의 공정에서는 초중종품 샘플링 검사를 수행하여 제품별 검사 결과에 신뢰성이 저하되며, 샘플링 검사 결과를 바탕으로 전체 제품에 적용하여 불량이 발생한 경우 해당 검사 단위 전품목을 폐기해야 하는 단점이 존재한다. Lee et al.(2019)는 CNC 공구 마모도 예측에 관한 연구를 수행하였다. CNC 설비 가동에 영향을 미치는 요인을 파악하기 위해 SVM, XGB, RF 모형을 사용하였으며, 이들의 Accuracy와 Time 등의 성능 비교를 통해 가장 뛰어난 성능을 보인 RF 모형을 최종 모형으로 제안하였다. 해당 연구는 공구 마모도 예측 정확도가 99%로 매우 높은 성능을 보이나, 공정에서의 불량 예측이 아닌 공구 마모도에 초점을 맞춘 연구라는 한계가 있다.

CNC 공정에서 머신러닝 방법을 적용하여 공정 품질을 개선하기 위한 연구가 지속적으로 이어지고 있다. Han은 CNC 가공 공정 변수 데이터에 머신러닝 방법을 적용해 품질을 예측하는 연구를 수행하였다(Han, 2022). Kim은 CNC 공정 품질 예측 및 불량 원인 분석을 위해 비지도 학습 기법을 활용하여 연구를 수행하였다(Kim et al., 2022). 공정에 주요한 영향을 미치는 변수들의 정확한 영향력 파악을 통한 원인 규명과 이를 보완하여 공정 과정을 최적으로 관리하는 것은 효율적인 공정 관리 및 공정 품질 개선을 위해 필수적이다(Kim et al., 2022). 하지만, CNC 공정과 관련된 선행 연구들은 공정으로 생산된 제품의 불량을 예측하는 모델 개발 연구나 공정에 사용되는 공구 마모도 예측과 같이 제품 불량 생산에 있어 간접적인 영향을 예측하거나 분석하는 연구가 주로 이어져 왔으며, 불량품이 생산되는 직접적인 영향에 대한 세부적인 분석 연구가 부족한 실정이다. 본 연구는 불량 예측이나 생산 설비 손상 예방과 같이 간접적인 불량 개선이 아닌 직접적인 불량 개선을 위한 연구로, 불량품이 생산되는 원인에 대한 명확한 규명과 세부적인 분석을 위해 설명가능한 인공지능 기법을 활용하여 연구를 수행하고자 한다.

본 연구에서는 CNC 가공 공정 데이터로 학습한 머신러닝 모델과 설명가능한 인공지능 기법인 SHAP(SHapley Additive exPlanations)를 활용하여 불량에 영향을 미치는 변수를 파악하여 공정 품질 향상에 기여하고자 한다. SHAP 기법은 Y에 영향을 미치는 X 변수의 영향력을 보다 세부적으로 분석할 수 있고, 이를 본 연구에 적용하여 CNC 공정에서의 불량 여부에 영향을 미치는 원인을 보다 명확히 규명할 수 있으며, 이를 통해 CNC 공정에서의 불량률 개선에 도모하고자 한다. 본 논문은 총 5장으로 다음과 같이 구성되어 있다. 제2장은 설명가능한 인공지능과 CNC 가공 공정의 이론적 배경에 대해 논한다. 제3장에서는 CNC 가공 공정 데이터를 학습한 머신러닝 모델과 SHAP를 활용한 공정 변수 중요도를 파악하는 방법론을 제안한다. 제4장에서는 제안된 방법론을 사용해 머신러닝 모델 학습과 변수별 SHAP Value 해석 실험을 진행한다. 마지막으로 제5장에서는 결론 및 향후 연구에 대해 논하고자 한다.

2. 이론적 배경 및 선행연구

2.1 설명가능한 인공지능

XAI(eXplainable Artificial Intelligence, 설명가능한 인공지능)는 머신러닝이나 딥러닝 모델의 결과값에 대한 근거를 제공할 수 있는 기법이다(Arrieta et al., 2020). XAI는 AI의 작동 원리를 쉽게 이해할 수 있도록 개발되어, 이유를 알 수 없어 통제가 어려운 기존 AI의 한계를 극복할 수 있는 장점이 있다. 최근 여러 공정에서 생산량 증가, 불량률 감소 등의 기대효과를 얻기 위해 공정 개선 연구가 활발히 이루어지고 있다(Hong, et al., 2023). Nahm은 XAI를 활용하여 유입수 수질 센서 항목을 선택하고, 하수처리 활성오니공정 모델링 연구를 수행하였다(Nahm, 2023). 본 연구에서는 XAI 기법을 활용하여 CNC 가공 공정 과정에서 불량에 영향을 미치는 공정 변수를 제시하고자 한다. XAI에는 SHAP를 비롯한 LIME, PDP 등 여러 기법이 있다(Ahn and Cho, 2021). 본 연구에서는 SHAP 기법을 활용하여 CNC 가공 공정에 영향을 미치는 주요한 변수들을 추출하고 공정 관리에 도움이 되고자 한다.

2.1.1 SHAP(Shapley Additive exPlanations)

SHAP(SHapley Additive exPlanations)는 게임이론에 기반한 기법으로, 변수 간 독립성을 근거로 덧셈이 가능하여 다양한 상황에 활용될 수 있는 기법이다. SHAP에서 사용되는 Shapley Value는 각 독립변수가 종속변수에 미치는 영향력을 종합하여 수치로 표현한 값으로, 전체 모델에서의 각 독립변수의 기여도를 나타낸다. 이를 위해 SHAP는 영향력을 알아보고자 하는 변수의 채택 유무에 따른 종속변수의 변화를 비교하며 해당 변수의 기여도를 확인한다. 이를 모든 가능한 변수 조합을 샘플링하고 평균값을 계산하여 표현한다. 이를 통해, 기존 머신러닝 기법의 특징 중요도와는 달리, 일관된 결과를 통해 기법의 신뢰성을 높이며, 각 변수의 중요도뿐만 아니라 변수 간 의존성까지 고려하여 모델의 영향력을 계산한다는 장점을 가진다. Shapley Value는 양수 또는 음수가 될 수 있다. 양수이면 해당 독립변수가 종속변수의 예측값을 증가시켰음을 나타내고, 음수이면 해당 독립변수가 종속변수의 예측값을 감소시켰음을 의미한다. Shapley Value는 아래와 같은 수식으로 표현된다.

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S)) \cdots\cdots (1)$$

ϕ_i : 데이터 i에 대한 Shapley Value

n : 참여자 수

S : 총 그룹에서 데이터 i를 제외한 모든 집합

$v(S)$: 데이터 i를 제외하고 나머지 부분 집합의 결과에 공헌한 기여도

$v(S \cup i)$: 데이터 i를 포함한 전체 기여도

최근 여러 학술 연구에서 SHAP 기법을 활용한 연구가 지속되고 있다(Lee et al., 2022). Na 외 3명은 SHAP를 활용한 증권 금융 상품 거래 고객의 이탈 예측을 연구하였다(Na et al., 2019). Oh 외 2명은 SHAP 기법을 활용하여 산업재해 예측 모델링 및 분석 연구를 수행하였다(Oh et al., 2021). Kim 외 2명은 선박 벌크선 메인엔진 연료 소모량 예측을 위해 SHAP 기법을 활용하여 예측 결과에 대한 변수의 작용을 해석하였다(Kim et al., 2023). 최근, 교육학계에서도 SHAP 기법을 활용한 연구가 이어지고 있다. Han은 고등학생의 창의적 사고와 관련된 변수 예측을 위해 SHAP 지수를 활용하였다(Han, 2023). Seo는 사범대 과학교육 전공생의 전공만족도 및 학업만족도 영향 분석을 위

해 SHAP을 활용하였다(Seo et al., 2023). Seo는 SHAP의 특장점인 개별 변인의 영향력 뿐만 아니라 집단 전체에 대한 영향력 분석 또한 밝힐 수 있다는 특징을 활용하여 과학교육과 재학생의 전공 및 학업 만족도를 지원하기 위한 방법론을 제안하였다. 본 연구에서는 SHAP 기법을 CNC 가공 공정 데이터에 적용하여 공정 과정에서 제품 불량 여부에 영향을 미치는 공정 변수를 규명하여 공정 품질 향상에 기여하는 연구를 수행하고자 한다.

2.2 CNC 가공 공정 및 관련 선행 연구

CNC 가공 공정은 원재료를 절삭하여 형상을 가공하는 공정으로, 금속 절단, 연삭, 밀링과 같은 기존 기계 가공과 동일한 기능을 수행하지만, 컴퓨터를 통해 가공을 수행하는 공정이다. 컴퓨터를 통해 가공을 수행하므로, CNC 가공 공정을 통해 제작된 제품은 수작업으로 표현하기 힘든 작업을 구현하는데 용이하다(Lee, 2017). 따라서, CNC 가공 공정에서는 사전 프로그래밍 된 컴퓨터 소프트웨어를 이용하여 공장의 도구나 기계를 제어하며, 가공 언어(G-code)를 통해 기계의 좌표나 속도, 위치 등을 정밀하게 조절하여 제어한다.

이러한 특징을 바탕으로 높은 난이도의 형상을 정밀하게 구현할 수 있으며, 데이터 관리 및 공정 관리를 통해 다수의 복제 생산이 가능하다. 특히 다른 공정 기법과는 달리 제품의 재질이나 형상에 크게 구애받지 않아 금형 물품뿐만 아니라 목업 제품 제작에도 사용 가능하며, 비행기나 자동차 등 크고 복잡한 형상을 가진 제품들 역시 효율적으로 생산 가능하다는 장점을 지닌다. 많은 산업 분야의 생산 작업에 CNC 가공 공정을 사용한다는 점에서 높은 수준의 공정 관리는 매우 중요하다. 그러나 CNC 가공 공정 특성 상 절삭 공구의 마찰에 의한 마모나 파손으로 한계치에 도달할 경우 가공 정밀도가 급격히 떨어진다. 이러한 특성에 따라 CNC 가공 공정에서 제품 불량 여부에 영향을 미치는 원인을 규명하는 것은 공정 품질의 기본인 균일한 제품 생산 및 수율 향상을 위해 중요하다.

본 연구에서는 설명가능한 인공지능 기법인 SHAP로 공정에 미치는 주요 변수를 확인하고 제품 불량의 원인을 규명하기 위해 자동차 부품 J사 공장의 CNC 가공 공정 데이터를 활용하고자 한다. CNC 가공 공정은 헤드, 드릴, 슬롯커터, 탭, 페이스 밀링 커터 등과 같은 절삭 공구를 이용하여 제품 성형을 하며, 이러한 절삭 가공 과정에서 절삭 공구가 마모되거나 순간적인 절삭력 약화로 인한 불량 제품이 생겨날 수 있다. 반대로 절삭력이 기준치보다 증가할 경우 공작 기계의 수명이 단축될 수 있어 CNC 가공 공정에서의 제품 불량률 개선은 공정 품질 개선을 위해 유의한 연구라고 볼 수 있다.

선행 연구 분석을 통해 기존 CNC 가공 공정에서 공정 불량 개선 연구의 필요성, 인공지능 기법을 이용한 CNC 가공 공정 불량 개선 연구 가능성을 확인하였다. Kim et al.(2022)는 CNC 공정 품질 예측 및 불량 원인 분석을 위해 공정 조건(입력) 변수만을 이용하는 비지도 학습 기법 중 하나인 Isolation Forest 기법을 활용하였다. 하지만, 불량 제품을 판단하는 기준으로 각 사이클에서 데이터의 패턴이 유사한 경우 정상품, 상이한 경우를 불량품으로 구역화하여 인위 분리한 점, Isolation Forest 기법을 활용하여 불량 원인을 분석하였으나, 해당 변수가 어떠한 영향으로 인해 불량이 야기되었는지의 원인을 제공하기에 부족한 점이 있다. Han은 트리 기반 분류 알고리즘을 통해 CNC 가공의 불량 예측 모델 프레임워크를 제안하였으나, 인공지능 기법의 블랙박스 특성에 의해 불량의 원인을 규명하기에는 한계가 있다(Han, 2022). Ju et al.(2023)은 실험계획법과 머신러닝 기법을 활용한 CNC 절삭공정 개선과 품질예측 모델 개발 사례 연구를 수행하였다. 불량 분류 모델 개발을 위하여 로지스틱 회귀, 랜덤포레스트, 서포트벡터머신의 기법을 활용하여 모델 성능 평가를 통해 최적 품질 예측 모델을 개발하였으나, CNC 공정에서 영향을 미치는 개별 변수들이 불량품 제작에 어떤 영향을 미치는지의 원인을 규명하기에는 다소 한계가 있다.본 연구는 이를 보완하기 위해 설명가능한 인공지능을 활용하여 CNC 가공 공정에서의 제품 불량에 대한 변수 별 원인을 규명하여 공정 불량에 영향을 미치는 세부 요인들을 상세 분석하고자 하며, 본 연구 결과를 통해 공정 품질 개선에 기여하고자 한다.

3. 연구 방법

본 논문의 연구 방법은 총 4단계로, 제1단계는 데이터 수집 및 전처리, 제2단계는 모델 학습, 제3단계는 모델 성능 비교, 제4단계는 주요 변수 추출로 구성된다(Figure 1). 데이터 수집 및 전처리를 위해 인공지능 중소벤처 제조 플랫폼에서 제공하는 Open AI 데이터셋을 활용하고, 부스팅 기반 알고리즘을 활용하여 모델 학습을 수행한다. 이후, 성능 지표를 기준으로 학습을 진행한 모델들의 성능을 비교하여 최종 모델을 선택한다. 이후 SHAP을 활용하여 CNC 공정에 영향을 미치는 주요 변수를 추출하고, 세부 원인을 규명한다.

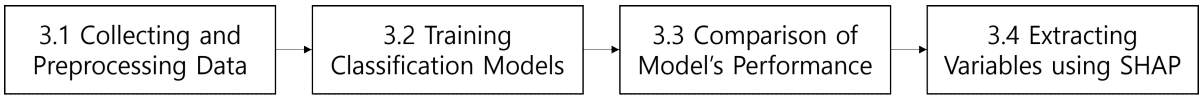

Figure 1. Methodology Process

3.1 데이터 수집 및 전처리

Table 1. Independent variables of CNC process data

Independent variable	Unit	Independent variable	Unit
X_ActualPosition	mm	Z_ActualVelocity	mm/s
X_ActualVelocity	mm/s	Z_ActualAcceleration	mm/s/s
X_ActualAcceleration	mm/s/s	Z_SetPosition	mm
X_SetPosition	mm	Z_SetVelocity	mm/s
X_SetVelocity	mm/s	Z_SetAcceleration	mm/s/s
X_SetAcceleration	mm/s/s	Z_CurrentFeedback	A
X_CurrentFeedback	A	Z_DCBusVoltage	V
X_DCBusVoltage	V	Z_OutputCurrent	A
X_OutputCurrent	A	Z_OutputVoltage	V
X_OutputVoltage	V	Z_OutputPower	kw
X_OutputPower	kw	S_ActualPosition	mm
Y_ActualPosition	mm	S_ActualVelocity	mm/s
Y_ActualVelocity	mm/s	S_ActualAcceleration	mm/s/s
Y_ActualAcceleration	mm/s/s	S_SetPosition	mm
Y_SetPosition	mm	S_SetVelocity	mm/s
Y_SetVelocity	mm/s	S_SetAcceleration	mm/s/s
Y_SetAcceleration	mm/s/s	S_CurrentFeedback	A
Y_CurrentFeedback	A	S_DCBusVoltage	V
Y_DCBusVoltage	V	S_OutputCurrent	A
Y_OutputCurrent	A	S_OutputVoltage	V
Y_OutputVoltage	V	S_OutputPower	kw
Y_OutputPower	kw	S_SystemInertia	kg*m^2
Z_ActualPosition	mm		

본 논문은 KAIST에서 제공하는 인공지능 중소벤처 제조 플랫폼(KAMP)의 CNC머신 AI 데이터셋을 사용한다(KAIST, 2020). 이 데이터는 2020년 10월 19일부터 2020년 10월 23일까지 자동차 부품의 CNC 가공 공정에서 수집되었다. 독립변수로 사용한 변수는 Table 1에서 확인할 수 있으며, 종속변수는 'PassOrFail'로, 양품은 0, 불량품은 1로 표시된다. 본 연구에서는 머신러닝 모델에 CNC 가공 공정 데이터를 학습시키고, SHAP 기법을 통해 CNC 가공 공정 과정에서 영향을 미치는 주요한 변수들을 도출하여 제품 불량의 원인을 제시함으로써 CNC 가공 공정 품질 향상 방법론을 제안하고자 한다.

3.2 모델 학습

본 연구에서는 부스팅(Boosting) 알고리즘 기반의 모델을 사용하여 모델 학습을 수행한다. 부스팅 알고리즘은 학습 모델의 성능 향상을 위해 분류기를 조정하는 개념에서 시작된다. 이 알고리즘은 데이터를 정제하여 여러 개의 분류 모델을 생성하는 방식으로 작동된다. 여러 개의 단순한 분류기를 조합하여 복잡한 분류기보다 우수한 성능을 달성하는 것이 목표이다. 각 단계에서 이전 단계의 학습 결과를 바탕으로 다음 단계의 분류 모델의 학습 데이터에 대한 가중치를 조정한다. 이 과정을 통해 이전 단계의 학습 결과가 다음 단계의 학습 결과에 영향을 미치게 된다. 따라서 학습이 진행될수록 분류 경계선 상의 데이터의 가중치가 증가하여 더욱 강력한 분별력을 갖출 수 있게 된다. 본 연구에서는 AdaBoost, GBM, XGBoost, LightGBM, CatBoost 총 5가지 부스팅 알고리즘 기반의 모델을 사용하여 학습을 진행한다.

3.2.1 AdaBoost

AdaBoost(Adaptive Boosting)는 부스팅 알고리즘 중에서도 가장 기본적인 기법으로, 약한 분류기들이 순차적으로 학습하고 이를 조합하여 강한 분류기를 형성하는 알고리즘이다(Freund et al., 1999). 약한 분류기들은 서로 상호보완적으로 동작하며 순차적으로 학습한다. 이 알고리즘은 초기에 학습된 분류기가 오분류한 샘플에 가중치를 부여하고, 이를 다음 분류기의 학습에 사용하여 이전 분류기의 오차를 보완한다. 이 과정을 반복하여 각 분류기가 더 중요하게 간주되는 데이터에 집중하여 학습한다. 최종적으로 이렇게 학습된 약한 분류기들을 조합하여 강한 분류기를 형성하며, 이는 전체 모델의 성능을 향상시키는 데 기여한다. 이 알고리즘의 주요 이점 중 하나는 이전 분류기가 잘못 분류한 샘플에 더욱 집중하여 학습하므로, 전반적인 성능 향상에 도움을 준다는 점이다.

3.2.2 GBM

그래디언트 부스팅 머신(Gradient Boosting Machine, GBM)은 부스팅 알고리즘의 한 종류로, 남은 잔차(Residual)를 개선하여 모델을 최적화하는 방법을 사용한다(Friedman, J. H., 2001). 부스팅은 초기에는 간단한 모델로부터 시작하여 반복적으로 모델을 개선해 나가는 방식으로 동작한다. 그래디언트 부스팅은 이러한 잔차를 줄여주는 최적의 파라미터를 찾기 위해 경사하강법(Gradient Descent)을 사용한다. 경사하강법은 손실 함수(Loss Function)의 기울기를 따라 가장 가파른 경사로 이동하여 손실을 최소화하는 방향으로 모델을 업데이트한다. GBM은 이를 활용하여 각 단계에서 잔차를 가장 효과적으로 줄이는 방향으로 모델을 개선한다. 이러한 방식으로 GBM은 강력한 예측 모델을 구축할 수 있으며, 특히 높은 성능을 보이는 모델 중 하나이다. 그러나 GBM은 과적합의 위험이 있기 때문에, 샘플링이나 정규화와 같은 기술을 활용하여 이를 완화하고 모델의 일반화 성능을 향상시킬 필요가 있다.

3.2.3 XGBoost

XGBoost(eXtreme Gradient Boosting) 알고리즘은 캐글(Kaggle) 등 다양한 데이터 분석 대회에서 뛰어난 성적을 거두는 알고리즘으로, 주로 수치 데이터 예측 모델로 활용되고 있다. XGBoost는 탐욕 알고리즘(Greedy algorithm)을 사용하여 내부에 생성된 다양한 모델들의 성능을 보완하는 가중치를 탐색한다(Chen and Guestrin, 2016.). 이 가중치는 CART(Classification And Regression Trees)라 불리는 앙상블 모델을 사용하며, 모든 최종 노드들이 최종 스코어를 계산하는 데 사용된다. XGBoost는 내부 하위 모델의 앙상블을 통해 과적합 문제를 효과적으로 해결할 수 있으며, 동시에 병렬 처리를 통해 그래디언트 부스팅 대비 학습 시간을 단축할 수 있는 장점이 있다. 이 알고리즘은 다양한 하이퍼파라미터 튜닝 옵션을 제공하여 모델의 성능을 최적화할 수 있는 유연성을 가지고 있다.

3.2.4 LightGBM

LightGBM은 XGBoost와 유사하게 GBM(Gradient Boosting Machine) 기반 알고리즘이다(Ke et al., 2017). 이 알고리즘의 주요 기술 중 하나는 Gradient-based One-Side Sampling(GOSS)로, 계산 시에 가중치가 작은 개체에 승수 상수를 적용하여 데이터를 효율적으로 증폭시키는 기술이다. LightGBM은 XGBoost보다 속도와 성능 면에서 우수하며, 일반적인 GBM 계열 트리의 특징인 level-wise 방식이 아닌 leaf-wise 방식을 통해 트리를 분할한다. Leaf-wise 방식은 가장 큰 손실을 가진 노드에 집중하여 분할하는 방식으로, level-wise 방식에 비해 시간적, 메모리적으로 효율적이다. Level-wise 방식은 트리를 균형적으로 만들기 위해 추가적인 연산이 필요하여 시간적으로 비효율적이다. 따라서 LightGBM은 대용량 데이터셋에서도 효율적으로 동작하면서 뛰어난 예측 성능을 제공하는 알고리즘 중 하나이다.

3.2.5 CatBoost

CatBoost는 XGBoost와 유사하게 level-wise 방식으로 트리를 부스팅하는 알고리즘이다. 이 알고리즘은 예측 속도가 빠르고, 불균형한 데이터에 대해서도 높은 예측 성능을 보이는 특징을 가지고 있다(Prokhorenkova et al. 2018). CatBoost의 특이한 점 중 하나는 Ordered Boosting이라는 기술을 사용한다는 것으로, 이는 모든 데이터를 대상으로 잔차 계산을 수행하는 기존 부스팅 알고리즘과는 다르게 일련의 순서를 가지고 데이터의 일부만을 선정하여 잔차를 계산한다. 이러한 방식은 순서 기준을 랜덤하게 섞어 과적합을 방지하는 데 도움이 되며, 효과적인 예측 모델을 구축한다. CatBoost는 특히 대용량 데이터셋에서도 뛰어난 성능을 보이면서, 사용자가 매개변수를 조정하기 쉽도록 설계되어 있는 장점이 있다.

3.3 모델 성능 비교

본 연구에서는 SHAP를 활용하여 CNC 가공 공정의 불량품에 영향을 미치는 주요 변수를 추출한다. 공정의 불량품에 영향을 미치는 원인을 잘 설명할 수 있는 변수를 추출할 수 있는 모델인지 평가하기 위해 모델 성능을 비교한다. 주요 변수 추출을 위해 모델 학습에 사용한 AdaBoost, GBM, XGBoost, LightGBM, CatBoost의 성능을 Accuracy(정확도), Precision(정밀도), Recall(재현율), F1-score 지표를 기준으로 비교한다. Accuracy는 전체 데이터 중 바르게 예측한 비율이다(Lee et al., 2021). 분류 모델을 평가하기에 가장 기본적인 지표이지만, 데이터 클래스가 불균형할 경우 성능 지표로 사용하기에 적절하지 않다. Precision은 True로 예측한 데이터 중 실제로 True

인 비율로, True로 예측한 결과가 얼마나 정확한지를 확인할 수 있다. Recall은 실제 True인 데이터를 True로 예측한 비율로, 실제 True를 얼마나 잘 예측하는지를 알 수 있다. F1-score는 Precision과 Recall을 동시에 고려하기 위한 방법으로, Precision과 Recall의 조화평균으로 정의된다. F1-score는 0과 1사이 값으로 1에 가까울수록 분류 성능이 좋다는 것을 의미한다. 본 논문에서는 Accuracy와 F1-score를 기준으로 모델을 선정한다(Table 2).

Table 2. Confusion Matrix and Evaluation Metrics

		Actual	
		True	False
Predicted	True	True Positive(TP)	False Positive(FP)
	False	False Negative(FN)	True Negative(TN)

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

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

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2\times\frac{\text{Precision}*Recall}{\text{Precision}+Recall}$$

3.4 주요 변수 추출

기존 부스팅 기반 머신러닝 모델들은 변수 중요도를 산출할 때, 기준에 따라 상이한 변수를 추출하여 주요 변수를 명확히 추출하기에 어려운 점이 존재한다. 예를 들어 XGBoost에서는 Weight, Cover, Gain을 기준으로 변수 중요도를 판단한다. Weight는 변수별 데이터를 분리하는 데 쓰인 횟수, Cover는 해당 변수로 분리된 데이터 수, Gain은 Feature를 사용했을 때 줄어드는 평균적인 Training Loss이다. 이와 같은 기준으로 변수 중요도를 산출할 경우 각 기준마다 중요한 변수들이 상이하게 선정된다. 따라서, 주요 변수를 명확히 추출하기 어렵다. 이를 보완하는 방법이 Shapley Value를 이용하는 것이다. Shapley Value는 여러 독립변수들의 조합을 기반으로 수행하여 얻은 변수 중요도의 평균이기에 기준에 따라 다른 변수가 추출되는 기존 기법들의 단점을 보완할 수 있으며, 긍/부정의 영향 역시 알 수 있다는 장점이 있다. 예를 들어, 어떤 독립변수의 Shapley Value가 음수라는 것은 종속변수의 값을 감소시켰다는 것을 의미하고, 양수라는 것은 종속변수의 값을 증가시켰다는 것을 의미한다. 본 논문에서 사용하는 CNC 가공 공정 데이터의 경우, 종속변수가 양품이면 0, 불량이면 1인 점을 고려할 때, 각 공정 변수들이 불량 여부에 양의 영향력을 미치는지 음의 영향력을 미치는지 알 수 있다는 것이다. 본 연구에서는 SHAP 기법을 사용하여 CNC 가공 공정에서 불량에 영향을 미치는 원인을 분석하고자 한다.

4. 연구 실험 및 결과

4.1 데이터 수집 및 전처리

본 연구에서 사용하는 자동차 부품 CNC 가공 공정 데이터 세트는 종속변수인 'PassOrFail'을 제외하고, 기계의 X축, Y축, Z축 및 스핀들 관련 변수들을 비롯하여 총 55개의 변수를 가진다. 그 중 'Press_Time'과 같이 공정에 영향을 미치지 않는 변수들, 이미 잘 통제되어 데이터 세트 전체에 걸쳐 같은 값을 가지는 변수들은 삭제하여 45개 변수만을 고려한다. 본 연구에서 사용하는 자동차 부품 CNC 가공 공정 데이터 세트는 총 32,048행을 가지는 데이터이며, 예시는 아래 Table 3과 같다.

Table 3. Example of CNC Machining Process Data & Used Feature

Index	Pass Or Fail(Y)	X_Actual_Position(X1)	X_Actual_Acceleration(X2)	· · ·	S_Output_Voltage(X45)
1	0	202	4	· · ·	6.96e-07
2	0	200	-13.8	· · ·	-5.27e-07
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
32,048	1	163	1.9	· · ·	0.158

4.2 모델 학습

본 연구의 목적은 공정에서의 제품 불량 여부에 영향을 미치는 원인을 분석하는 것으로, 종속변수가 제품의 양품 여부이며, 양품 여부를 분류하는 분류 모델을 사용한다. 모델 학습을 위해 부스팅 알고리즘 기반 분류 모델을 사용하였으며, 사용한 기법은 AdaBoost, GBM, XGBoost, LightGBM, CatBoost로 총 5개이다. 부스팅 알고리즘이란 앙상블 기법 중 하나로, 초기에 약한 학습기를 설정하여 오차를 수정해 나가면서 강건한 학습기를 만드는 알고리즘이다. 이전 모델이 오분류일 경우 더 높은 가중치를 부여하여 오분류에 집중하여 이를 더 잘 해결할 수 있는 모델이 되도록 수정해 나간다. 본 논문에서는 부스팅 알고리즘 기반의 모델들을 선정하여 학습에 사용한다. 모델 학습을 위해 학습, 검증 데이터 비율을 9:1로 나누어 실험을 진행한다. 모델 학습에 사용된 학습 데이터는 28,843개, 검증에 사용된 검증 데이터는 3,205개이다.

4.3 모델 성능 비교

Table 4. Comparison of model's performance

	Accuracy	Precision	Recall	F1-score
AdaBoost	0.93	0.88	0.89	0.89
GBM	0.95	0.91	0.91	0.91
XGBoost	0.97	0.93	0.95	0.94
LightGBM	0.97	0.93	0.95	0.94
CatBoost	**0.97**	**0.93**	**0.96**	**0.95**

AdaBoost, GBM. XGBoost, LightGBM, CatBoost 모델 성능 결과는 Table 4와 같다. Accuracy의 경우 CatBoost가 0.97로 가장 높은 성능을 보였고, Precision과 Recall의 조화 평균인 F1-score의 경우도 CatBoost가 0.95로 가장 높은 성능을 보였다. 성능 비교 결과에 따라 주요 변수 추출을 위한 SHAP 기법에 사용될 분류기로 CatBoost를 선정하여 실험을 진행한다.

4.4 주요 변수 추출

Figure 2는 Python 프로그래밍 CatBoost 모듈에 내장되어 있는 함수를 활용해 특징 중요도를 산출한 결과이다. 가장 높은 중요도를 갖는 변수인 '9. X_OutputCurrent'는 CNC 가공 기계의 X축 실제 출력 전류이다. 기계에 설정된 전류인 '18. X_CurrentFeedback'와 달리 실제 출력 전류는 중요도가 높다고 평가됨을 알 수 있다. 두 번째로 높은 중요도를 갖는 변수인 '37. S_SetPosition'은 CNC 가공 기계에 입력된 스핀들의 설정 위치이다. 스핀들의 실제 위치인 '34. S_ActualPosition'는 아홉 번째로 높은 중요도를 갖는 변수인 것을 보면 기계에 입력된 스핀들의 설정 위치와 실제 위치는 모두 중요하다고 볼 수 있다. 세 번째로 높은 중요도를 갖는 변수인 '41. S_DCBusVoltage'는 CNC 가공 기계에 입력된 스핀들의 설정 전압이다. 스핀들의 실제 전압인 '43. S_OutputVoltage'와 달리 기계에 입력된 스핀들의 설정 전압의 중요도가 높다고 평가됨을 알 수 있다.

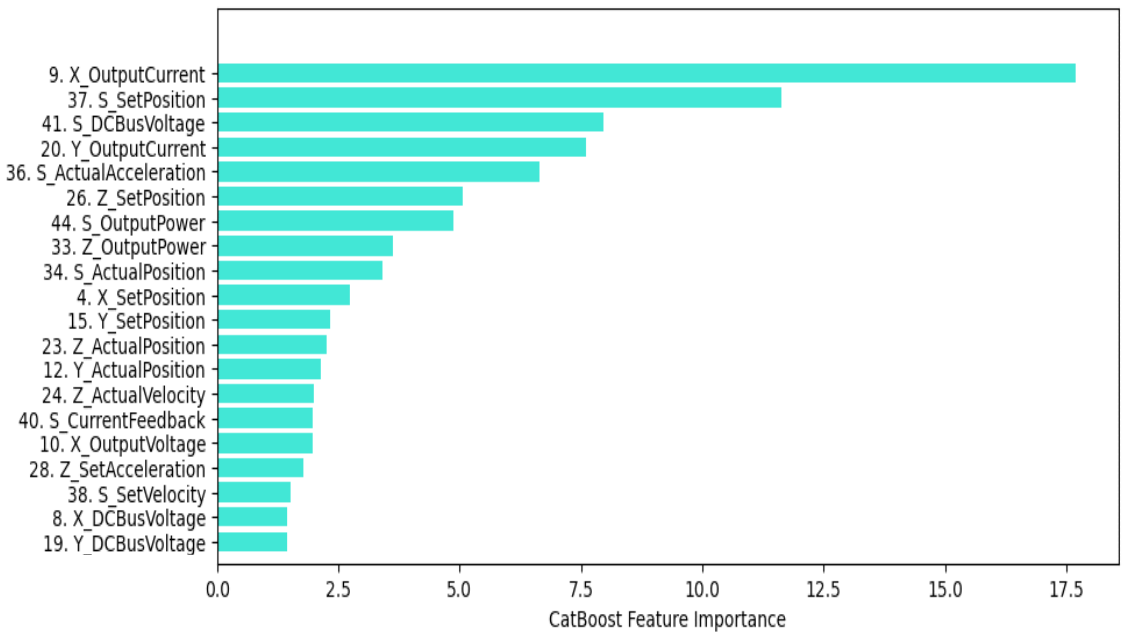

Figure 2. Feature Importance Analysis with CatBoost(Python Programming)

Figure 3은 데이터 샘플마다 독립변수의 Shapley Value를 산출하고, 그 절댓값을 평균내어 추출한 결과이다. 가장 높은 중요도를 갖는 변수는 '9. X_OutputCurrent'로, Figure 2와 동일하다. 기계에 설정된 전류인 '18. X_CurrentFeedback'는 Figure 3에서도 중요도가 낮게 평가된다. 두 번째로 높은 중요도를 갖는 변수 역시 '37. S_SetPosition'로, Figure 2와 동일하다. 주목할 점으로, 스핀들의 실제 위치인 '34. S_ActualPosition'이 세 번째로 높은 중요도를 갖는 것으로 평가된다. Figure 2에서 아홉 번째로 높은 중요도를 가진 것과 대조되는 점이다. 다음으로, Figure 2에서 세 번째로 높은 중요도를 갖는 변수인 '41. S_DCBusVoltage'는 Figure 3에서는 여덟 번째로 내려간 반면, '20. Y_OutputCurrent'는 Figure 2, 3 모두에서 네 번째로 높은 중요도를 갖는다.

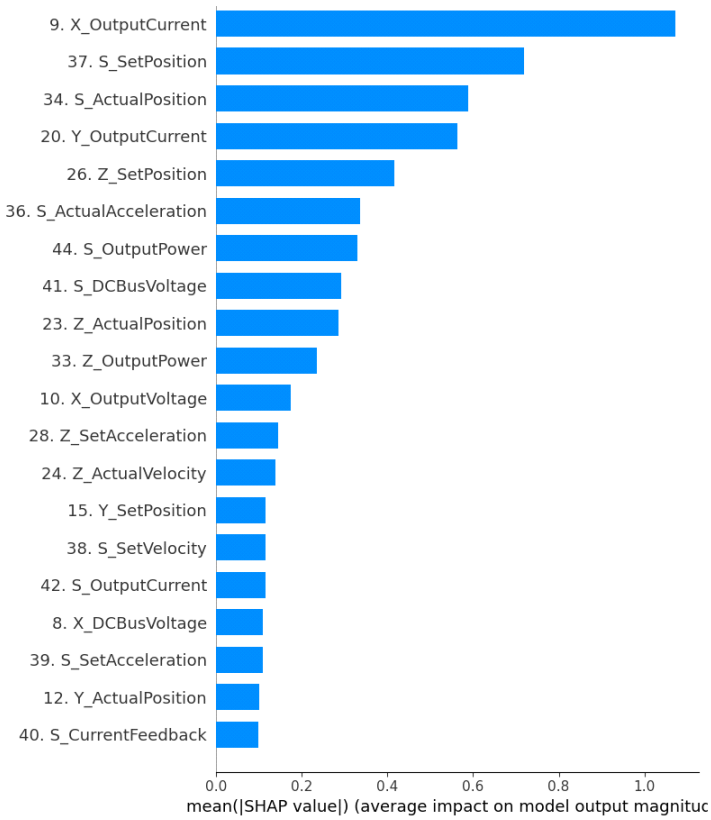

Figure 3. Shapley Value(Absolute Average Impact on Model Output Magnitude)

Figure 4는 데이터 샘플마다 독립변수의 Shapley Value를 산출하고, Shapley Value의 부호와 값을 통해 해당 변수가 종속변수인 불량에 주는 영향을 그래프로 나타낸 그림이다. 그래프의 X축은 Shapley Value의 값을. 그래프 Y축의 왼쪽 범례는 절댓값의 크기가 큰 20개의 독립변수를, 그래프 Y축의 오른쪽 범례는 해당 독립변수의 값이 높으면 적색으로, 낮으면 청색으로 표현한다는 것을 의미한다. 가장 높은 중요도를 갖는 '9. X_OutputCurrent' 변수의 경우, Shapley Value가 양수인 영역에서 청색으로, 해당 변수의 데이터가 낮은 것을 알 수 있다. 이는 해당 변수의 데이터 값이 낮아질수록, 종속변수인 불량의 값을 높인다는 의미로, 불량품을 만든다는 의미이다. 반면, Shapley Value가 음수인 영역에서 적색으로, 해당 변수의 데이터가 높은 것을 알 수 있다. 이는 해당 변수의 데이터 값이 높아질수록, 종속변수인 불량의 값을 낮춘다는 의미로, 양품을 만든다는 의미이다. 즉, CNC 가공 기계의 X축 실제 출력 전류가 높아질수록 양품을, 낮아질수록 불량품을 만든다는 것을 알 수 있다. 마찬가지로, '37_S_SetPosition' 변수, '34. S_ActualPosition' 변수, '20. Y_OutputCurrent' 변수도 데이터의 값이 높아질수록 양품을, 데이터의 값이 낮아질수록 불량품을 만든다.

'26. Z_SetPosition', '15. Y_SetPosition' 변수의 경우 앞서 나열한 변수들과는 반대로, 데이터의 값이 낮아질수록 양품을, 데이터의 값이 높아질수록 불량품을 만든다. '41. S_DCBusVoltage', '10. X_OutputVoltage' 변수의 경우 Shapley Value가 양수인 영역, 음수인 영역 모두에서 적색을 나타내어, 해당 독립변수의 데이터 변화가 종속변수인 불량품 여부에 일관된 영향을 보인다고 할 수 없다.

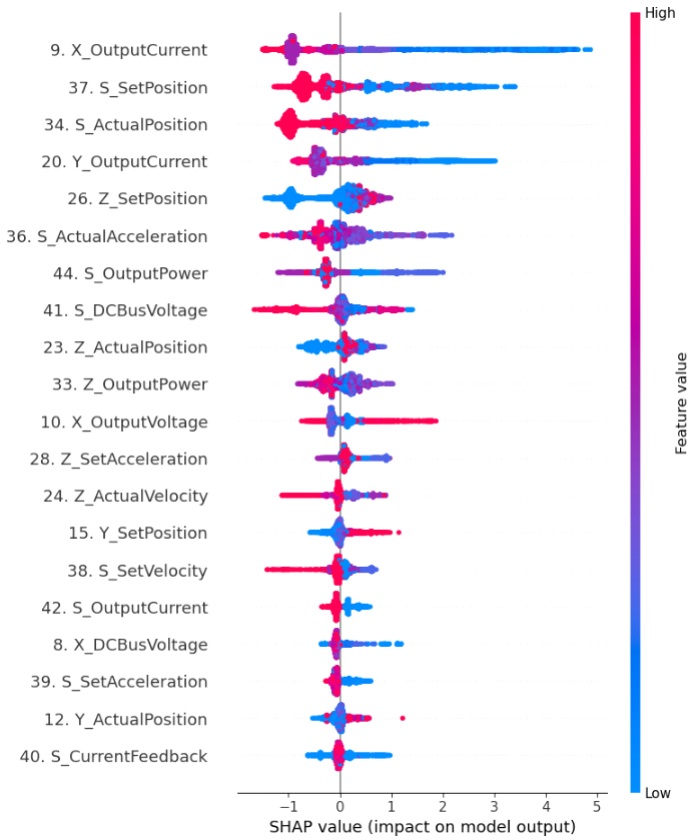

Figure 4. Shapley Value(Impact on Model Output)

5. 결 론

본 연구에서는 설명가능한 인공지능 기법 중 하나인 SHAP를 활용하여 CNC 가공 공정의 불량품에 영향을 미치는 주요 변수를 추출한다. 먼저, 공정의 불량에 영향을 미치는 원인을 잘 설명할 수 있는 변수를 추출할 수 있는 모델인지 평가하기 위해 모델 성능을 비교한다. 부스팅 알고리즘 기반 모델인 AdaBoost, GBM, XGBoost, LightGBM, CatBoost를 활용하여 모델 학습을 수행하였으며, 가장 성능이 높은 CatBoost 모델의 경우 Accuracy 97%, F1-score 95%의 성능으로 불량을 예측할 수 있다. 이 모델에 설명가능한 인공지능 기법인 SHAP를 활용하여 CNC 가공 공정에서 제품 불량에 영향을 미치는 변수를 추출하였다. Figure 4를 참고하면 변수의 데이터가 증가할수록 양품을 만드는 독립변수는 '9. X_OutputCurrent', '37_S_SetPosition', '34. S_ActualPosition', '20. Y_OutputCurrent' 등이 있다. 따라서 해당 공정에서는 이 변수들이 증가할 수 있도록 관리하여야 한다. 또한, 변수의 데이터가 감소할수록 양품을 만드는 독립변수는 '26. Z_SetPosition', '15. Y_SetPosition' 등이 있다. 이 변수들은 감소할 수 있도록 관리하여야 한다. '41. S_DCBusVoltage', '10. X_OutputVoltage' 변수와 같이 불량품 여부에 일관된 영향을 보이지 않는 변수들은 상대적으로 덜 주요한 변수로 보이므로, 앞서 일관된 영향을 보이는 변수들의 관리가 우선되어야 할 것이다.

기존의 CNC 관련 연구로는 제품 불량 예측을 위한 연구, 공정 설비 손상 예방을 위한 연구와 같이 공정 불량에 간접적인 영향을 미치는 연구들이 주로 선행되어 왔으며, 공정 불량에 직접적인 영향인 불량이 일어나는 원인에 대한 세부적인 분석을 위한 연구가 부족하였다. 본 연구에서는 Shapley Value를 활용하여 우선적으로 관리되어야 할 주요 변수들을 추출하고, 해당 변수들에 대한 세부적인 분석을 통해 CNC 공정 불량에 영향을 미치는 직접적인 요인을 연구 결과로 제시하였다. 본 연구 결과는 제품 불량의 원인을 공정 변수의 세부 분석을 통해 명확한 원인을 규명하여 공정 불량에 직접적인 영향을 미치는 결과를 도출하였다는 점에서, 그리고 특정 기준에 따라 상이한 변수들이 추출되는 기존 머신러닝 기법과 달리 주요 변수를 설명력 있게 추출할 수 있는 SHAP 기법을 활용하여 CNC 공정 불량에 미치는 변수 영향력을 일관되고 정확하게 설명할 수 있다는 점에서 연구적인 의의가 있다. CatBoost를 활용한 변수 중요도 산출 결과와 Shapley Value를 활용한 주요 변수 추출 결과를 비교하여 이를 확인하였으며, 주요 독립변수들의 데이터 변화에 따라 종속변수에 어떠한 영향을 미치는지 알 수 있었다. 본 연구 결과는 추후 CNC 가공 공정 데이터의 불량 개선을 위한 연구로 활용될 수 있다. 향후 연구로, CNC 가공 공정의 불량 여부에 영향을 미치는 주요 변수들을 제어하여 생산 과정에서의 불량품 생산을 조기에 예방할 수 있는 실시간 모니터링 시스템을 구축하고자 한다.

REFERENCES

Ahn, Yoonae, and Cho, Hanjin. 2021. A Study on XAI-based Clinical Decision Support System. The Journal of the Korea Contents Association 21(12):13-22.

Arrieta, A., B., Daz-Rodrguez, N., Ser, J., D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82-115.

Chen, Y., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016:785-794.

Freund, Y., Schapire, R., and Abe, N. 1999. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5):771-780.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29(5):1189-1232.

Han, Junga. 2023. Exploring Predictors Affecting Creative Thinking in High School Students Using Random Forest and SHAP. Korean Journal of Educational Research 61(4):101-131.

Han, Yonghee. 2022. Prediction Model of CNC Processing Defects using Machine Learning. Journal of the Korea Convergence Society 13(2):249-255.

Hong, Jisoo, Hong, Yongmin, Oh, Seungyong, Kang, Taeho, Lee, Hyeonjeong, and Kang, Sungwoo. 2023. Injection Process Yield Improvement Methodology Based on eXplainable Artificial Intelligence(XAI) Algotihm. Journal of Korean Society for Quality Management 51(1):55-65.

Ju, Hyejin, Seo, Hojin, Kim, Yeoungil, Kim, Sujin, Lee, Gunmyung, Kim, Sanghyeon, Jeong, Yoonhyeon, and Byun, Jaihyun. 2023. A Case Study of CNC Machining Process Improvement and Quality Prediction Model Development Using Design of Experiments and Machine Learning. Journal of the Korean Institute of Industrial Engineers 49(4):354-368.

KAIST. 2020. CNC Machine AI Dataset. Korea AI Manufacturing Platform(KAMP). 2020(December):01-58.

https://www.kamp-ai.kr/front/main/MAIN.01.01.jsp.

Kang, Seonghyeon, and Kim. Seoungbum. 2016. Multivariate Monitoring of the Metal Frame Process in Mobile Device Manufacturing. Journal of the Korean Insititute of Industrial Engineers 42(6):395-403.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu T. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30(2017).

Kim, Hyunju, Park, Mingyu, and Lee, Jihwan. 2023. A Study on the Prediction of Fuel Consumption of Bulk Ship Main Engine Using Explainable Artificial Intelligence. Journal of Navigation and Port Research 47(4):182-190.

Kim, Iljung, Kim, Woosoon, Kim, Joonyoung, Chae, Heesu, Woo, Jiyeong, Do Kyungmin, Lim, Sunghoon, Shin, Minsoo, Lee, Jieun, Kim, Heungnam. 2022. Discovering Essential AI-based Manufacturing Policy Issues for Competitive Reinforcement of Small and Medium Manufacturing Enterprises. Journal of Korean Society for Quality Management 50(4):647-664.

Kim, Kanghee, Kim, Hyunjung. 2022. A Study on the Build of a QbD Six Sigma System to Promote Quality Improvement(QbD) Based on Drug Design. Journal of Korean Society for Quality Management 50(3):373-386.

Kim, Namki, Jung, Minyoung, Park, Junpyo, Jin, Seungjong, and Wang, Jinam. 2022. Predict the Quality of CNC Processes and Analyze the Causes of Defects. Proceedings of Korean Institute of Industrial Engineers Spring Joint Conference. 2022.

Lee, Hyunggeun, Hong, Yongmin, and Kang, Sungwoo. 2021. Identifying Process Capability Index for Electricity Distribution System through Thermal Image Analysis. Journal of Korean Society for Quality Management 49(3):327-340.

Lee, Juyeon. 2020. Technologies for Collecting, Processing, Analyzing, and Utilizing Data for Intelligent Die-casting Processes. Journal of the Korean Society of Manufacturing Technology Engineers 29(6):441-448.

Lee, Kangbae, Park, Sungho, Sung, Sangha, and Park, Domyoung. 2019. A Study on the Predicition of CNC Tool Wear Using Machine Learning Technique. Journal of the Korea Convergence Society 10(11):15-21.

Lee, Seunghoon, Kim, Yongsoo. 2022. A Pre-processing Using TadGAN-based Time-series Anomaly Detection. Journal of Korean Society for Quality Management 50(3): 459-471.

Lee, Youngchoon. 2017. A Study on Design Method using CNC in Wooden Products. Journal of the Korea Furniture Society 28(4):371-379.

Na, Kwangtek, Lee, Jinyoung, Kim, Eunchan, and Lee, Hyochan. 2020. A Securities Company's Customer Churn Prediction Model and Causal Inference with SHAP Value. The Korea Journal of BigData 5(2):215-229

Nahm, Euiseok. 2023. A Study on Modeling of Activated Sludge Process in Wastewater Treatment System Utilizing XAI(eXplainable AI). the Transactions of the Korean Instititue of Electrical Engineers 72(2):263-269.

Oh, Hyungrok, Son, Aelin, and Lee, Zoonky. 2021. Occupational accident prediction modeling and analysis using SHAP. Journal of Digital Contents Society 22(7):1115-1123.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., V., and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31(2018).

Seo, Jibeom, and Kang, Namhwa. 2023. Exploration of Factors on Pre-service Science Teacher's Major Satisfaction and Academic Satisfaction using Machine Learning and Explainable AI SHAP. Journal of Science Education 47(1):37-51.

저자소개

홍지수 인하대학교 통계학과 학사를, 인하대학교 산업경영공학과 석사를 취득하고 현재 인하대학교 산업경영공학과 박사과정에 재학 중이다. 주요 관심 분야는 빅데이터 분석, 데이터 마이닝, 품질관리, 제조 최적화이다.

정영진 인하대학교 산업경영공학과 학사를 취득하고 현재 인하대학교 산업경영공학과 석박사통합과정에 재학 중이다. 주요 관심 분야는 데이터 사이언스이다.

강성우 인하대학교 산업경영공학과에서 학사를, 펜실베니아 주립대학교 산업제조공학과에서 석사와 박사를 취득하고 현재 인하대학교 산업경영공학과 부교수로 재직 중이다. 주요 관심 연구 분야는 빅데이터 프로세싱 기반 제품 설계, 공학 설계, 생산 장비 예측 진단 및 관리이다.