Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2021.44.4.012

Vol. 44, No. 4 : 12-22, December 2021
ISSN : 2005-0461 / EISSN : 2287-7975

Quality Prediction Model for Manufacturing Process of Free-Machining 303-series Stainless Steel Small Rolling Wire Rods

Seokjun Seo · Heungseob Kim†

Department of Smart Manufacturing Engineering, Changwon National University

쾌삭 303계 스테인리스강 소형 압연 선재 제조 공정의 생산품질 예측 모형

서석준 · 김흥섭†

창원대학교 스마트제조융합협동과정

This article suggests the machine learning model, i.e., classifier, for predicting the production quality of free-machining 303-series stainless steel(STS303) small rolling wire rods according to the operating condition of the manufacturing process. For the development of the classifier, manufacturing data for 37 operating variables were collected from the manufacturing execution system(MES) of Company S, and the 12 types of derived variables were generated based on literature review and interviews with field experts. This research was performed with data preprocessing, exploratory data analysis, feature selection, machine learning modeling, and the evaluation of alternative models. In the preprocessing stage, missing values and outliers are removed, and oversampling using SMOTE(Synthetic oversampling technique) to resolve data imbalance. Features are selected by variable importance of LASSO(Least absolute shrinkage and selection operator) regression, extreme gradient boosting(XGBoost), and random forest models. Finally, logistic regression, support vector machine(SVM), random forest, and XGBoost are developed as a classifier to predict the adequate or defective products with new operating conditions. The optimal hyper-parameters for each model are investigated by the grid search and random search methods based on k-fold cross-validation. As a result of the experiment, XGBoost showed relatively high predictive performance compared to other models with an accuracy of 0.9929, specificity of 0.9372, F_1-score of 0.9963, and logarithmic loss of 0.0209. The classifier developed in this study is expected to improve productivity by enabling effective management of the manufacturing process for the STS303 small rolling wire rods.

Keywords : Smart Manufacturing, Machine Learning, Manufacturing Execution System(MES), Free-machining 303-series Stainless Steel(STS303), Extreme Gradient Boosting(XGBoost)

1. 서 론

국내 · 외 제조업에서는 정보통신기술(ICT)을 이용하여 광범위하게 수집/축적되고 있는 양질의 데이터를 바탕으로 생산성 향상과 비용 절감을 위한 공정관리 개선, 생산설비 예지 정비, 불량률 감소, 생산품질 예측 등과 같은 다양한 노력을 시도하고 있으며, 그러한 노력의 중심에서 기계학습(ML: Machine learning) 기술이 활발하게 적용되고 있다. 또한, 기계학습(ML)은 스마트공장(Smart factory)으로 진화하기 위한 핵심 기반기술로 인식되고 있다. 철강산업에서는 연속적인 제철 공정의 특성상 제조설비의 자동화를

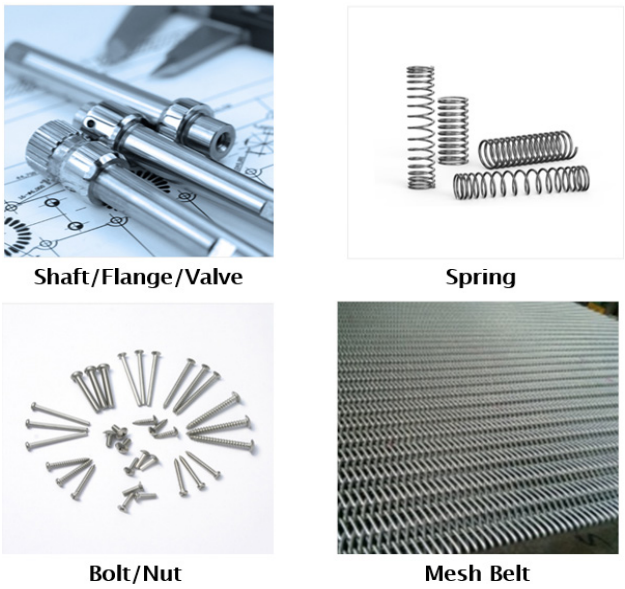
Received 6 October 2021; Finally Revised 27 November 2021; Accepted 6 December 2021
† Corresponding Author : heungseob79@changwon.ac.kr

중요하게 인식해왔으며, 세계적인 철강・제강기업들은 공급과잉, 환경, 인건비 등 산업의 구조적 문제 해결과 질적 성장을 위한 신성장 방안으로 스마트공장을 추진하고 있는 것으로 알려지고 있다. 이에 따라 철강・제강 공정의 스마트화와 관련한 연구가 활성화되었을 것으로 추정되나, 기업의 보안 문제로 인해 노출되지 않고 있어 철강・제강 제품을 생산하는 대규모 제조공정 데이터와 연구결과를 확인하기 어려운 실정이다.

국내・외 철강산업 분야에서 확인되는 대표적인 기계학습 응용 사례로는, 국내 리딩기업인 P사는 고로(용광로)의 스마트화를 통해 용선 생산량과 품질을 향상시키면서 연료량(비용)은 감소시키는 효과를 거두었으며, 용융아연도금 공정에 도금량 제어 자동화를 위한 인공지능(AI: Artificial Intelligence) 적용을 통해 자동차용 도금강판의 편차를 획기적으로 저감하였다. H사는 철분말 공장에 빅데이터(Big data)와 인공지능 인프라 구축을 통해 품질/공정 모니터링, 품질/설비상태 예측 등의 공정 스마트화를 추진하였다. 또한, 국외 사례로는, 중국의 최대 국영 철강기업인 B사는 빅데이터 기반 수요 맞춤형 자동 생산 계획 시스템, 설비상태 진단, 인공지능(AI)을 활용한 품질 검사 시스템 등 다양한 솔루션을 개발 적용하였으며, 일본은 과거 사고사례와 원인, 복구 작업 데이터를 기반으로 노후된 공장에서의 각종 사고의 위험을 예측하고 적절하게 대처할 수 있도록 하는데 인공지능(AI) 기술을 적용하고 있다[21].

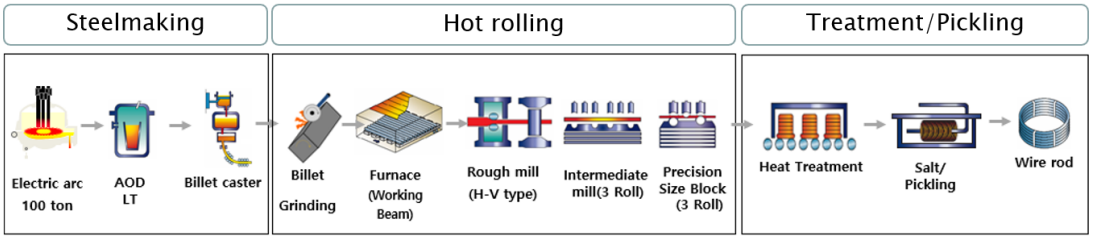
최근 철강산업을 포함한 제조업 분야에서 중요하게 다뤄지고 있는 연구주제 중 하나는 생산공정에서 수집되는 빅데이터를 활용하여 제품의 생산품질을 예측하고, 품질 영향 인자를 식별하기 위한 연구이다. Jeong and Kang[7]은 국내 D기업의 자동차 범퍼 제조공정 데이터를 대상으로 연관규칙(Association rules) 분석을 통해 공정 불량요인을 식별하고, 불량항목과 공정 간의 변화 패턴 관계를 분석하였으며, Kim and Baek[10]은 반도체 제조 공정의 셀 레벨에서 밀도 기반 클러스터링 기법인 OPTICS(Ordering Points To Identify the Clustering Structure)를 통해 품질 특성치를 파악하여 칩의 품질을 예측하는 방법론을 제시하였다. Lee et al.[23]은 비모수 ANOVA 검정을 통해 양품 집단과 불량품 집단 간의 차이를 보이는 변수와 패턴`을 파악하고, 서포트벡터머신(SVM: Support vector machine), 앙상블모형(Ensemble model)을 이용하여 고무 생산 공정의 품질 예측 모형을 개발하였다. 철강산업 분야에서는, Nkonyana et al.[18]은 랜덤포레스트(Random forest), 인공신경망(ANN: Artificial neural network) 등의 기계학습 모형을 이용하여 철강 제조공정의 결함 진단을 위한 기법을 제안하였다. Kang et al.[9]은 후판(Steelplate) 제품을 대상으로 품질(인장강도, 항복강도)에 영향을 미치는 조업조건(인자)을 탐색하고, 이를 바탕으로 인공신경망(ANN)을 이용해 품질을 예측하기 위한 시스템을 제안하였다. Park et al.[21]은 랜덤포레스트 모형을 이용하여 특수강 소형 압연 공정에서 탈탄(Decarburization) 두께와 같은 품질 인자를 예측하였으며, Ruiz et al.[22]은 기계학습을 사용하여 철강 선재(Wire rod)의 길이에 영향을 미치는 인자들을 식별하고, 부분의존도(PDP: Partial dependence plots)를 통해 구리(Cu), 탄소(C), 망간(Mn) 등과 같은 제강 성분의 변화에 따른 선재 길이의 변화량을 확인하였다.

본 연구는 쾌삭 303계 오스테나이트계 스테인리스강(STS303: Free-Machining 303-series Austenitic Stainless Steels) 소재의 압연 선재(Wire rod) 제조공정을 대상으로, Kang et al.[9], Park et al.[21]에서와 같이 제품의 품질에 영향을 미치는 조업조건을 탐색하고, 조업조건에 따른 생산품질을 예측하기 위한 기계학습 모형을 제안한다. STS303은 일반적인 대기 부식(Atmospheric corrosion), 식품, 대부분의 유기화학 물질과 일부 무기화학 물질에 상당한 내성을 가짐에 따라 <Figure 1>과 같은 복잡한 형상의 기계 부품부터 건설, 가전, 항공, 방위산업 분야 등에 폭넓게 사용되고 있다.



<Figure 1> Use of STS303 Wire Rods

쾌삭 스테인리스강은 황(Sulfur) 또는 셀레늄(Selenium)을 추가하여 기계 절삭성(Machinability)을 향상시킨 스테인리스 강이다[8]. 이러한 원소들은 0.15~0.35% 정도 첨가되며, 기계 절삭성 향상을 통해 가공 속도 증가, 가공 툴(Tool)의 수명 연장, 그리고 표면 마감 품질 향상을 도모할 수 있게 된다. 하지만 STS303의 기계 절삭성을 향상시키기 위해 형성시킨 황화망간(MnS)으로 인해 열간 성형성(Hot workability)이 열악해지는 문제점이 대두되고 있으며, 이로 인해 압연 선재 제조공정에서는 일반 스테인리스강



<Figure 2> Manufacturing process for STS303 wire rod

대비 개구(Alligatoring), 스캡(Scab) 등과 같은 불량이 증가하여 생산성이 저하되고 있다[15, 16]. 재료공학 분야에서, 스테인리스강의 가공, 마모, 표면 결함 등의 영향 인자 등에 대한 연구들은 일부 수행되어 왔으나[2, 11, 13, 14], STS303 소재의 열간 성형성에 대한 영향 인자에 대한 연구는 매우 드물어 제조현장의 높은 불량률 문제를 해결하는데 어려움을 겪고 있다[8]. 더욱이 철강 제품의 품질 영향 인자에 대한 연구들은 대부분 단일 인자의 품질 영향성 분석을 목적으로 하고 있어, 다단계 제조공정에서 각 세부 공정의 조업조건 변화에 의한 생산품질 변동을 분석/예측하는데 적용하기에 제한이 따르고 있다.

따라서 본 연구는 S사의 STS303 압연 선재 제조공정의 생산관리시스템(MES: Manufacturing execution system) 데이터를 기반으로 양품과 불량품의 조업조건 패턴을 식별하고, 조업조건에 따른 생산품질을 예측할 수 있는 기계학습 모형을 제안한다. STS303 소재의 압연 선재는 국내에서 S사가 유일하게 생산하고 있는 제품으로, 직경(Diameter) 5㎜부터 34㎜까지 41종의 선재를 생산하고 있다. 예비적 연구(Preliminary study)에서 직경 5㎜부터 15.5㎜까지의 소형 선재와 직경 16㎜부터 34㎜까지의 중형 선재의 불량률 추이, 생산품질의 주요 영향 인자 등이 상이함이 확인되었다. 이에 따라 본 연구는 직경 5㎜부터 15.5㎜까지의 소형 선재를 대상으로, 제조 빅데이터에 대한 탐색적 분석(EDA: Exploratory data analysis), 차원 축소(Dimensionality reduction), 그리고 기계학습 모델링 및 성능평가 결과를 제시한다. 기계학습 모델링 단계에서는 다양한 지도학습(Supervised learning) 모형들이 고려되었으나, 그중 유의미한 수준의 성능을 보인 로지스틱 회귀 모형(Logistic regression model), 서포트벡터머신(SVM: Support vector machine) 모형, 랜덤포레스트(Random forest) 모형, 그리고 XGBoost(Extreme gradient boosting) 모형의 연구결과를 중심으로 제시한다. 위의 연구절차는 R의 통합개발환경(IDE: Integrated Development Environment)인 R-Studio에서 수행되었으며, 주요하게 활용된 패키지(Package)는 부록으로 제시되었다.

본 연구의 제2장에서는 S사의 STS303 압연 선재 제조공정과 수집된 데이터 현황 소개와 연구 절차를 제시하고, 제3장에서는 데이터 전처리(Preprocessing) 및 차원 축소 결과를 제시한다. 제4장에서는 조업조건에 따른 생산품질 예측 모형, 즉, 기계학습 모델링과 모형 성능평가 결과를 제시하며, 제5장에서는 결론 및 향후 연구방향을 제시한다.

2. 문제정의 및 연구절차

2.1 문제정의

S사의 STS303 소형 선재 제조공정은 크게 <Figure 2>와 같이 제강(Steelmaking), 열간압연(Hot rolling), 열처리/산세(Treatment/Pickling) 공정으로 구성된다. 제강공정에서는 빌렛(Billet)을 제품 용도에 따라 저온에서 충분히 예열하고, 선재의 표면 품질을 좌우하는 탈탄 현상을 억제하는 것이 관건이다. 이를 위해 가열로 내 온도 및 재로시간(Residence time), 연료 및 공기비 등을 관리하고, AOD-LT (Argon Oxygen Decarburization-Ladle Treatment) 공정에서 아르곤 가스와 산소 가스를 함께 취입하여 용강 내 존재하는 탄소를 제거한다. 열간압연 공정에서는 고객이 요구하는 선재의 규격과 재질 특성을 확보하기 위해 압연온도, 압하량, 변형속도 등 압연조건을 조정하여야 한다. 열처리 공정에서는 열간압연으로 경화된 조직을 연화시키고, 산세 공정에서 스케일이 효과적으로 제거될 수 있도록 스케일/모재 계면에 크롬(Cr)-고갈층을 형성시키기 위해 일정 온도에서 소둔 열처리를 실시한다. 마지막으로, 산세 공정에서는 제품 표면의 오염 물질과 스케일을 제거하여 표면을 청정하게 한다. 이와 같이 각 세부 공정은 적정한 품질의 제품을 생산하기 위한 개별적인 목적을 가지며, 각각의 목적을 달성하기 위해 세밀한 조업조건을 유지하여야 한다. S사에서는 각 세부 공정의 목적과 관련된 조업조건들에 대한 데이터를 측정하여 저장·관리하고 있다.

따라서 본 연구는 S사의 생산관리시스템(MES)으로부터 STS303 소형 압연 선재 7,550 로트(Lot)에 대한 제조공정 데이터를 기반으로 하며, 이중 양품(Good) 데이터는 7,311 로트(96.8%), 불량품(Defective) 데이터는 239 로트(3.2%)로 상당한 수준의 불균형 데이터(Imbalanced data)이다. 생산관

리시스템(MES)으로부터 수집된 제조공정 데이터는 황(Sulfur) 함유량 등 <Table 1>에 제시된 총 37종의 조업변수를 포함하고 있다. 또한, <Table 2>와 같은 스테인리스강 제품의 품질 영향 인자에 대한 문헌조사를 통해 Mn/S 비(Mn/S ratio) 등 3종, 현장 전문가들에 대한 인터뷰를 통해 9종의 파생변수(Derived variable)들을 추가하였다. 결과적으로, 본 연구에서는 생산관리시스템(MES)의 37종과 문헌조사 및 현장 전문가 인터뷰를 통해 추가된 12종의 파생변수를 포함하여 총 49종의 조업변수를 고려한다.

<Table 1> MES Data

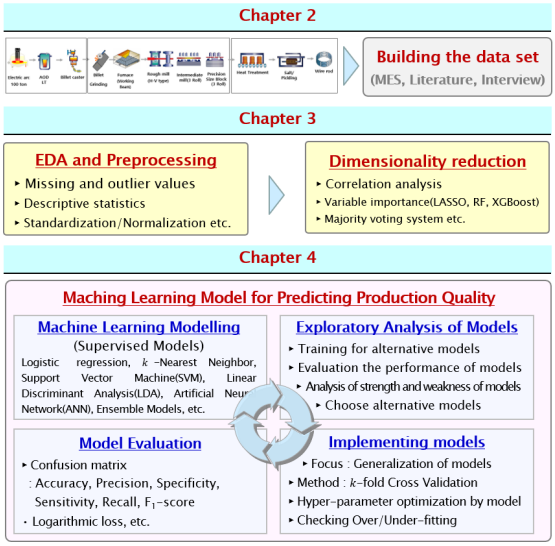
Steel making	Hot rolling	Treatment/Pickling
Refine time etc.(31)	Heating time etc.(6)	Good / Defective

<Table 2> Literature Review on Quality Influencing Factors

Factors	Source	Literature
Delta-Ferrite	MES	Ahn et al.[1], Jung[8], Kim[11], Lee et al.[14]
Mn/S ratio	**Added**	Bae and Nam[3], Brinkman and Garvin[4], Jung[8], Nazábal et al.[17], Toledo et al.[24]
Homogenizing Annealing	**Added**	Cheruvathur et al.[5], Jung[8]
Temperature and Strain rate in forming process	**Added**	Kim et al.[12], Nazábal et al.[17], Osakada[20]
Content of S	MES	Lee et al.[13], Lee et al.[14]
Content of Mn	MES	Lee et al.[13], Oliver et al.[19]
Content of Cu, Si, N	MES	Oliver et al.[19]

2.2 연구절차

본 연구의 흐름은 <Figure 3>과 같으며, 첫 번째 단계인 제조공정 빅데이터 정립은 2.1절에서 소개한 바와 같다. 두 번째 단계에서는 데이터 전처리와 차원 축소를 수행하였다. 데이터 전처리의 주요 내용으로는, 결측치와 이상치 제거, 표준화(Standardization), 정합성 확보를 위한 데이터 정제, 그리고 데이터 불균형성 해소를 위한 조치들을 수행하였다. 또한, 차원 축소(특성치 추출 및 선택, Feature extraction and selection) 검토를 통해 기계학습 모형의 다중공선성(Multicollinearity) 문제와 과적합(Overfitting) 방지, 시간 복잡도와 공간 복잡도(Time and space complexity) 감소를 통한 모형의 효율성 향상을 도모하고자 하였다. 마지막으로, 조업조건에 따른 선재의 생산품질을 예측하기 위한 후보 모형들의 학습과 성능 평가를 수행하였다. 기계학습 모델링에 있어서는 모형별 하이퍼 파라미터들(Hyperparameters)을 최적화하고 일반화(Generalization) 성능을 확보하기 위해 k-겹 교차검증(k-fold cross validation) 기법을 적용하였다. 최종적으로, 생산품질 예측 모형을 선정하기 위한 성능 평가를 수행하며, 성능 평가에는 정분류율(Accuracy), 민감도(Sensitivity), 특이도(Specificity)와 같은 오분류표(Confusion matrix) 기반의 지표들과 로그 손실(Logarithmic loss) 함수를 활용하였다.



<Figure 3> Study process

3. 데이터 전처리 및 차원 축소

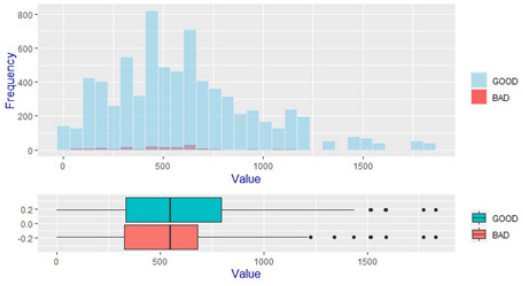
3.1 데이터 전처리(Preprocessing)

데이터 전처리 단계는 획득된 데이터를 대상으로 결측치와 이상치를 제거하고 탐색적 데이터 분석을 통해 데이터를 다양한 각도에서 관찰하고 이해하는 과정이다. 이는 그래프나 통계적인 기법들을 이용하여 데이터를 직관적으로 파악하여 데이터 품질에 대한 잠재적인 문제점을 식별하고 인사이트(Insight)를 발견할 수 있다.

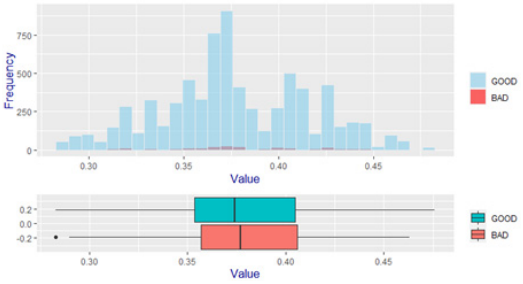
생산관리시스템(MES)으로부터 수집된 7,550개의 데이터를 대상으로 결측치를 제거하고, 히스토그램(Histogram)과 박스 플롯(Box plot)을 사용하여 데이터 탐색을 진행하였다. 데이터 탐색 예시는 <Figure 4>와 같으며, 몇몇 공정변수에서 극저빈도의 이상치가 확인된다. 이에 따라 IQR(Interquartile range) 규칙(1.5×IQR)에 의한 잠재적 이상치들 중 전체 데이터의 1% 이하로 나타난 데이터를 최종 이상치로 판정하고 제거하였다.

또한, <Figure 4>에서도 나타난 바와 같이 다수의 공정변수에서 양품과 불량품의 조업조건이 혼재된 것이 확인되었다. 이에 따라 데이터를 상세히 관찰한 결과, 모든 공정의 조업조건이 매우 유사함에도 생산품이 양품인 경우와 불량

품인 경우가 혼재되어 있음이 확인되었다. 이에 대한 현장 전문가들과의 토의 결과는 다음과 같았다. 선재의 제조 특성상 <Figure 5>와 같이 하나의 빌렛(Billet)이 투입(Input)되어 수 Km의 선재 제품으로 생산되고, 최종 양품과 불량품의 분류는 부분 결함의 정도, 즉, 결점 수에 따른다. 즉, <Figure 5>와 같이 수 Km의 선재에서 특정 지점의 공정 데이터가 대표값으로 생산관리시스템(MES)으로 보고될 때, 최종 생산품질과 공정 데이터상에 불일치(Mismatch)가 발생할 수 있다는 것이다. 따라서 본 연구에서는 데이터의 정합성 확보를 위해 데이터 정제를 수행하였다.

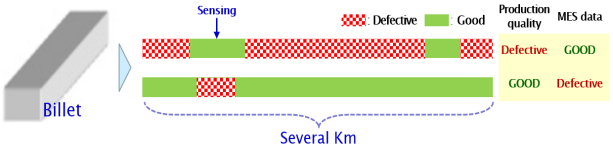


(a) Variable 1



(b) Variable 2

<Figure 4> Examples of data exploration



<Figure 5> Example of data mismatch

양품과 불량품의 혼재 구간에 대한 공정 데이터를 정제하기 위해 군집분석(Clustering)을 활용하였으며, 군집분석 전에 조업변수별 표준화를 수행하였다. 또한, 다수 공정의 조업조건 조합에 따라 양품 또는 불량품이 생산될 수 있기 때문에 코사인 유사도(Cosine similarity)를 적용하였다. 즉, 군집분석에서 <Figure 6>과 같이 임의의 군집(Cluster)에 양품과 불량품이 혼재될 수 있으며, 이때 불량품의 비율이 특정 수준을 넘어서는 군집을 불량품 군집으로 판정하고, 해당 군집에 속한 양품 데이터를 제거하는 언더샘플링(Undersampling)을 진행하였다. 본 연구에서는 불량품의 비율이 전체 데이터에서의 불량률 3.2%를 초과하는 군집에 적용하였다. 이와 같이 데이터를 정제한 결과는 <Table 3>과 같다.

※ ①, ②, and ③ are determined as clusters for defective products → Eliminating data on good products.

# of Cluster	1	2	3	4	5	6	7	8	···
GOOD	100	200	30	158	216	19	2	5	···
Defective (rate)	20 16.7%	1 0.5%	25 45.5%	0 0.0%	6 2.7%	0 0.0%	33 94.3%	0 0.0%	···

<Figure 6> Data cleansing using clustering

<Table 3> Results of Data Cleaning

Cleaning	Good	Defective	Total
Before	7,311(96.8%)	239(3.2%)	7,311(100%)
After	3,888(95.3%)	191(4.7%)	4,079(100%)

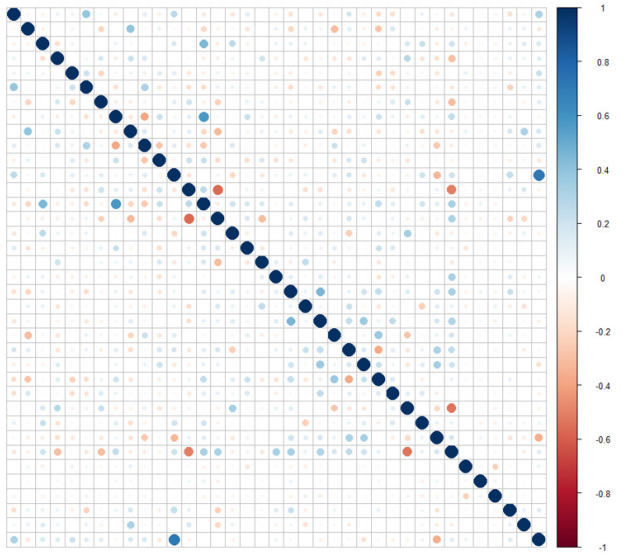
<Table 3>에서 확인되는 바와 같이 데이터 정제 후 데이터의 불균형성이 약간 완화되기는 하였으나, 데이터 정제 후에도 양품과 불량품 데이터의 비율이 약 95:5로 불균형이 심함을 확인할 수 있다. 따라서 데이터 불균형을 해소하기 위해 SMOTE(Synthetic oversampling technique)를 사용하여 오버샘플링을 수행하였다. 마지막으로, 각 공정변수의 표준화(Standardization)를 통해 데이터 스케일링(Data scaling)을 수행하고, 층화 표본추출(Stratified sampling)을 통해 7:3의 비율로 훈련 데이터(Training set)와 시험 데이터(Test set)을 구성하였다.

3.2 차원축소(Dimensionality Reduction)

기계학습 모형에서 학습 데이터의 변수(특성치, Feature) 간의 상관관계가 높은 경우, 다중공선성 문제로 인해 모형의 분산(Variance)과 편이(Bias)가 증가하여 성능이 저하되는 문제가 발생한다. 또한, 양적으로 제한된 학습 데이터에서 변수(차원)의 수가 증가할수록 모형의 과적합(Overfitting) 가능성도 증가하게 된다. 따라서 본 연구에서는 기계학습 모형의 상관관계 분석을 통해 다중공선성 문제를 검토하고, 라쏘(LASSO: Least Absolute Shrinkage and Selection Operator) 회귀분석, 랜덤포레스트와 XGBoost 모형에서의 변수 중요도(Variable importance)를 바탕으로 생산품질에 대한 예측력이 강한 변수들을 선정한다.

본 연구에서 다루는 데이터에는 생산관리시스템(MES)로부터 수집된 37종의 변수, 문헌조사와 현장 전문가 인터뷰를 통해 생성한 12종의 파생변수가 포함된다. 파생변수는 생산관리시스템(MES)로부터 수집된 변수들을 특정 함수를 통해 생성되었기 때문에 변수들 간의 상관관계 분석에서 파생변수는 제외하였다. 또한, 상관관계 분석 전 시

각화를 통해 확인한 결과, 변수 간 유의미한 비선형 관계 등은 확인되지 않았다. 결과적으로, 생산관리시스템(MES)로부터 수집된 37종의 변수에 대한 상관관계 분석 결과는 <Figure 7>과 같았으며, 상관계수의 절대값이 0.6을 초과하는 변수 조합은 식별되지 않았다.



<Figure 7> Correlation Matrix Plot

기계학습 모형에 반영할 변수를 선택하기 위해 데이터 전처리 과정을 거친 4,079개의 공정 데이터를 바탕으로 변수 중요도를 검토한 결과는 <Figure 8>과 같다. <Figure 8>에서 각 모형에 대한 그래프는 변수들을 변수 중요도를 기준으로 내림차순하였으며, 선택되는 변수의 범위는 변수 중요도의 변곡점을 기준으로 선정하였다. 각 모형에서 추천된 변수는 모두 30종으로 결정되었으나, 모형별 변수 중요도 순위가 상이함에 따라 추천된 변수들이 동일하지는 않다. 따라서 차원 축소의 신뢰성 향상을 위해 Voting system 개념을 도입하여 2가지 이상의 모형에서 추천된 변수를 선택하였다. 차원 축소(변수 선택) 결과는 <Table 4>에 제시된 바와 같이 총 49종의 변수 중 38종이 선택되었으며, 생산관리시스템(MES)에 수집된 총 37종의 변수 중 27종, 연구문헌을 통해 추천된 3종의 파생변수, 그리고 현장 전문가들이 추천한 9종의 파생변수 중 8종이 선택되었다.

<Table 4> Result of Feature Selection by Voting System

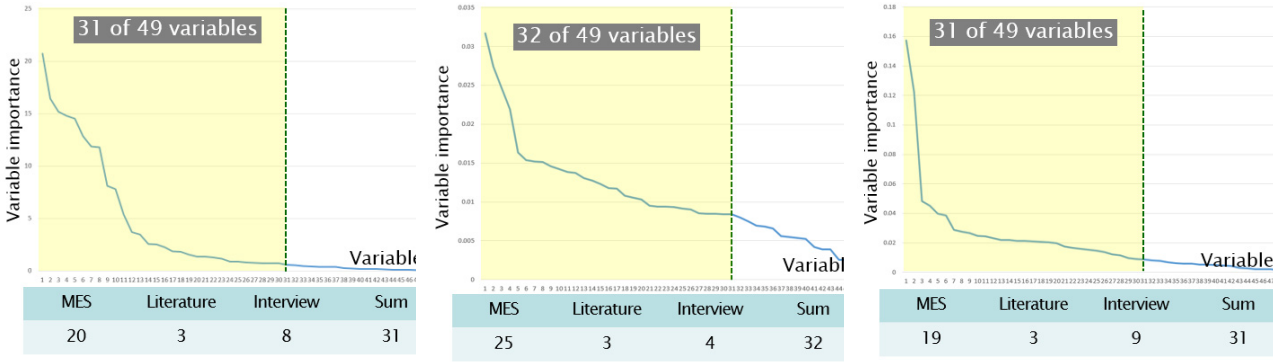
Source	MES	Literature	Interview
Feature selection	27 of 37 variables	3 of 3 variables	8 of 9 variables
Total	38 of 49 variables		

4. 생산품질 예측 모형

4.1 후보 모형 선정 및 모델링

본 연구에서는 선형판별분석(LDA: Linear Discriminant Analysis), 이차판별분석(QDA: Quadratic Discriminant Analysis), 의사결정나무(Decision Tree) 모형, 인공신경망(ANN: Artificial Neural Network), 심층신경망(Deep Neural Network) 등 다양한 지도학습 모형들을 대상으로 예비적인 연구를 수행하고, 개략적 분류 성능, 계산비용(Computational cost)과 과적합 가능성 등을 고려하여 로지스틱 회귀분석 모형, 서포트벡터머신(SVM) 모형, 랜덤포레스트 모형, 그리고 XGBoost 모형을 후보 모형으로 선정하였다.

로지스틱 회귀분석 모형은 새로운 독립(설명)변수가 주어질 때, 종속(반응)변수의 각 범주(Class)에 속할 확률을 추정하여 추정 확률의 기준치에 따라 종속변수를 분류하는 목적으로 활용되는 통계기법이다. 서포트벡터머신(SVM)은 최대 마진(Margin)을 갖는 선형판별에 기초하며, 고차원 공간에서 탐색된 초평면(Hyperplane)의 집합을 통해 분류하는 방법이다. 랜덤포레스트는 의사결정 나무와



(a) LASSO model (b) Random forest model (c) XGBoost model

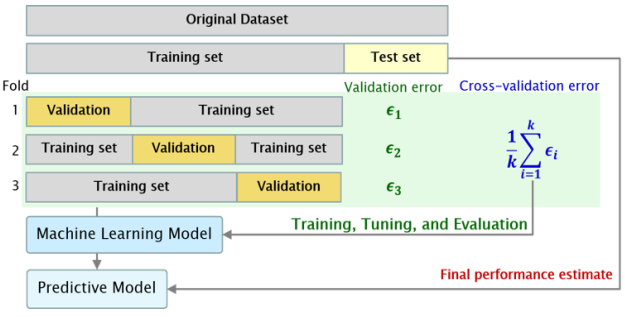
<Figure 8> Variable importance

<Table 5> Hyperparameter Tuning in Each Model

Model	Hyper parameters	Range for the random search
Logistic Regression	· Decision boundary(DB) = **0.54** (by Matthews Correlation Coefficient)	$DB \in (0, 1)$
SVM	· Kernel function : $K(x_i, x_j) = \gamma(x_i^T x_j + c)^{\mathbf{d}} = \mathbf{0.25}(x_i^T x_j + \mathbf{2})^{\mathbf{2}}$	$\gamma \in [0, 0.4]$, $d \in [1, 3]$, $c \in [0, 2]$
XGBoost	· Maximum number of iterations(N) = **500**, · Step size of each boosting step(Learning rate, η) = **0.2**, · Gamma(γ) : **0**, · Maximum depth of the tree($D_{\max}$) = **100**, · Minimum sum of weights of all observations required in a child($O_{\min}$) : **2**	$N = 500$ $\eta \in [0.1, 0.3]$ $\gamma \in [0, 2]$ $D_{\max} \in [50, 150]$ $O_{\min} \in [1, 5]$
Random Forest	· Number of trees to grow(G) = 100, · Number of variables randomly sampled as candidates at each split = $\sqrt{38}$ · Minimum size of terminal nodes($N_{\min}$) = 2, · Maximum number of terminal nodes trees($N_{\max}$) = 200	$G \in [75, 150]$ Díaz-Uriarte and De Andres[6] $N_{\min} \in [2, 5]$ $N_{\max} \in [100, 500]$

배깅(Bagging)을 혼합한 형태이며 여러 개의 의사결정나무를 형성한 후에 각 의사결정나무의 분류 결과를 종합하여 다수결 또는 평균에 따라 분류하는 방법이다. XGBoost는 예측력이 약한 모델들의 학습 에러(Training error)에 가중치(Weight)를 두고, 순차적으로 다음 학습 모델에 반영하여 강한 예측 모델을 만드는 앙상블 기법이다.

또한, 기계학습 모델링에서 각 모델의 하이퍼 파라미터(Hyper-parameter)는 모델의 성능을 결정짓는 매우 중요한 요소이다. 따라서 본 연구는 후보 모형들의 최적 하이퍼 파라미터를 탐색하기 위해 <Figure 9>와 같은 k-겹 교차검증(k-fold cross validation)을 기반으로 격자탐색법(Grid search method)와 무작위탐색법(Random search method)을 병용하였다. k-겹 교차 검증은 학습 데이터셋을 균등하게 k개의 그룹(Fold)으로 나누고 1개의 검증 데이터셋 그룹과 $(k-1)$개의 학습 데이터셋 그룹을 지정하여 학습 및 성능 평가를 실시하는 방법이다. 학습·검증 데이터셋을 k번 반복하여 변경하는 과정에 따라 얻어진 결과를 통해 성능의 평균을 다른 하이퍼 파라미터 조합과 비교하여 가장 높았던 성능의 하이퍼 파라미터를 최종적으로 결정한다.



<Figure 9> k-fold Cross Validation Method

또한, 각 모형의 최적 하이퍼 파라미터 조합을 효율적으로 탐색하기 위해 격자탐색법과 무작위탐색법을 상호보완적으로 활용하는 2단계 탐색전략을 적용하였다. 즉, 격자탐색법은 안정적인 탐색 방법이지만 격자를 세분화하면 계산시간이 많이 소요되고, 계산시간 감소를 위해 격자를 크게 적용하면 세밀한 탐색이 불가능하다. 무작위탐색법은 계산시간 측면에서 격자탐색법보다 유리할 수 있으나, 넓은 탐색범위에 적용할수록 탐색 결과에 대한 불확실성이 높다. 따라서 본 연구에서는 격자탐색법을 통해 (1단계) 넓은 탐색범위에서 최적의 하이퍼 파라미터 조합이 존재할 수 있는 격자(Grid)를 탐색하고, 이후 (2단계) 해당 격자(Grid) 내에서 무작위탐색법을 이용하여 최적 조합을 탐색하였다. 단, 로지스틱 회귀분석 모형의 결정경계(Decision boundary)는 매튜의 상관계수(Matthews Correlation Coefficient)를 기준으로 격자탐색법만으로 탐색되었으며, 랜덤포레스트 모형의 '부트스트랩 표본에 포함되는 변수 개수'는 Díaz-Uriarte and De Andres[6]에서 추천된 $[\text{조업변수 개수}(38)]^{1/2}$로 적용하였다.

5-fold 교차 검증 기반으로 탐색된 각 후보 모형의 하이퍼 파라미터를 최적화한 결과는 <Table 5>와 같다. <Table 5>에 제시된 탐색범위는 격자탐색법을 통해 탐색된 최적 하이퍼 파라미터 조합의 격자(Grid)이며, 무작위탐색법에 적용된 탐색범위이다.

4.2 후보 모형 성능 평가

본 연구와 같은 이진 분류 시스템(Binary Classifier System)의 성능평가에서 가장 일반적으로 <Figure 10>과 같은 오분류표(Confusion matrix)에 기반한 성능지표들이 사용된다. 오분류표에 기반한 지표들은 정분류율(Accuracy), 특이도(Specificity), F_1 지표(F_1-score) 등이 있으며, 식 (1)-(5)와 같이 정의된다. 하지만, 오분류표 기반의 지표들은 실제 범주(Class)에 대한 적중 여부로만 판단하고, 상호보완적 의미를 갖기 때문에 최종 모형 선정 시에 모호성을 보일 수 있다. 따라서 본 연구에서는 추가적으로 손실 함수(Loss function), 즉, Log loss(또는, Binary Cross-Entropy) 함수를

고려하였다. Log loss 함수는 식 (6)과 같이 정의되며, 모형의 분류 확신도, 즉, 해당 클래스로의 분류 확률을 이용하여 모형의 성능을 평가할 수 있으며, 모형의 분류성능이 높을수록 Log loss 함수의 값이 낮게 나타난다.

<표기(Notation)>

- $\mathbf{M}$: 클래스(Class) 집합; $j \in \mathbf{M}$
- $\mathbf{N}$: 샘플(Sample)의 수; $i \in \mathbf{N}$
- $y_{ij} = \begin{cases} 1, & \text{샘플 } i\text{가 클래스 } j\text{에 속하면,} \\ 0, & \text{Otherwise} \end{cases}$
- p_{ij} : 샘플 i가 클래스 j로 분류될 확률

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

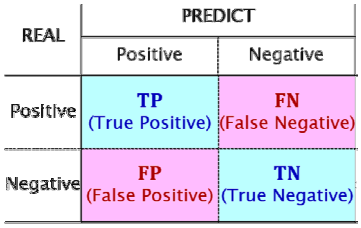
$$\text{Specificity} = \frac{TN}{FP + TN} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Log. Loss (Class)} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \quad (6)$$



<Figure 10> Confusion Matrix

각 후보 모형의 분류 예측 결과는 <Table 6>과 같이 나타났으며, 각 성능 지표(6종)에 대한 평가결과는 <Table 7>과 같다. 정분류율 지표는 로지스틱 회귀분석이 77.1%, 서포트벡터머신(SVM)이 98.46%, XGBoost가 99.29%, 랜덤포레스트가 99.56%로 나타났으며, 특이도 지표는 로지스틱 회귀분석이 74.87%, 서포트벡터머신(SVM)이 78.01%, XGBoost가 93.71%, 랜덤포레스트가 92.15%로 나타났다. F_1-지표는 로지스틱 회귀분석이 86.53%, 서포트벡터머신(SVM)이 99.19%, XGBoost가 99.63%, 랜덤포레스트가 99.77%로 나타났다. 결론적으로, 대부분의 지표에서 XGBoost와 랜덤포레스트 모형이 다른 모형들에 비해 우수한 것으로 나타났으며, 정분류율 등 오분류표 기반 성능지표에서 두 모형은 성능은 유사하게 평가되었다. 따라서 두 모형 중 최종 모형은 Log. loss 함수값을 기준으로 선정하였다. XGBoost 모형의 Log. loss 함수값은 0.0299, 랜덤포레스트 모형의 Log. loss 함수값은 0.0809로 측정되어 최종 모형은 XGBoost 모형으로 결정되었다.

<Table 6> Confusion Matrix of the Models

Logistic Regression		Predict		Total
		Positive	Negative	
Real	Positive	3,002	886	3,888
	Negative	48	143	191
Total		3,050	1,029	
SVM		Predict		Total
		Positive	Negative	
Real	Positive	3,867	21	3,888
	Negative	42	149	191
		3,909	170	
XGBoost		Predict		Total
		Positive	Negative	
Real	Positive	3,885	3	3,888
	Negative	15	176	191
Total		3,900	179	
Random Forest		Predict		Total
		Positive	Negative	
Real	Positive	3,870	17	3,888
	Negative	12	179	191
Total		3,882	196	

<Table 7> Performance Metrics of the Alternative Models

Model	Accuracy	Specificity	Recall	Precision	F_1-score	Log. loss
Logistic Regression	0.7710	0.7487	0.7721	0.9843	0.8653	0.4754
SVM	0.9846	0.7801	0.9946	0.9893	0.9919	0.0899
XGBoost	**0.9929**	**0.9372**	**0.9956**	**0.9969**	**0.9963**	**0.0299**
Random Forest	0.9956	0.9215	0.9992	0.9962	0.9977	0.0809

5. 결론 및 향후 연구방향

본 연구는 국내 S사의 STS303 소형 압연 선재 제조 공정의 생산관리시스템(MES) 데이터를 기반으로 문헌조사와 현장 전문가 인터뷰를 통해 파생변수를 생성하고, 탐색적 분석 및 데이터 전처리, 차원축소, 기계학습 모델링을 통해 공정 조건에 따른 선재의 생산품질 예측 모형을 개발하였다. 데이터 전처리 단계에서는 공정 데이터와 실제 생산품질 간의 불일치 발생이 의심되는 데이터를 제거하여 데이터 셋의 정합성을 향상시키고자 하였으며, 이때 코사인 유사도 기반의 군집분석이 활용되었다. 차원축소 단계에서는 라쏘(LASSO) 회귀 모형, 랜덤포레스트 모형, 그리고 XGBoost 모형의 변수 중요도를 기준으로 Voting system 개념을 도입하여 변수 선택의 신뢰성을 향상시키고자 하였다. 또한, 기계학습 모델링 단계에서는 다수의 후보 모형들을 검토하여, 그 중 로지스틱 회귀분석, 서포트벡터머신(SVM), XGBoost와 랜덤포레스트 모형의 성능 분석 결과를 제시하였다. 후보 모형들 중 XGBoost 모형이 최선의 성능을 보였으며, 각 성능지표 측정 결과는 정분류율(Accuracy) 0.9929, 특이도(Specificity) 0.9372, F_1-지표 0.9963, 그리고 Log. loss 함수값은 0.0299로 나타났다. 즉, S사의 STS303 소형 압연 선재의 공정 조건에 따른 생산품질 예측 모형으로 개발된 XGBoost 모형의 성능은 우수한 것으로 판단된다.

향후 연구 방향으로는, 첫째, 본 연구에서 제안한 연구 절차와 방법론을 다른 제품군에 적용하는 것이다. 다수 제품군에서 공정조건과 생산품질 간의 상관성 분석은 경험적/실험적 지식에 의존도가 높은 철강/제강 산업에 새로운 통찰력(Insight)를 제공할 수 있을 것이다. 둘째, 본 연구에서 개발된 생산품질 예측 모형을 기반으로 불량률 최소화, 즉, 정상품 생산 확률을 최대화하기 위한 공정 조건을 탐색하는 것이다. 즉, 생산품질 예측 모형에서 제공하는 정상품으로의 분류 확률을 최대화하기 위한 각 공정의 조업조건을 탐색하는 비선형 최적화(Nonlinear optimization) 문제로 고려될 수 있으며, 이에 대한 해법으로는 메타 휴리스틱(Meta-heuristic) 등 다양한 비선형 최적화 알고리즘이 고려될 수 있다. 마지막으로, 압연 선재와 같이 다수의 공정을 거쳐 생산되는 제품을 고려할 때, 이미 수행된 선수 공정의 조업조건을 바탕으로 후공정의 조업조건을 최적화하는 연구도 필요하다. 즉, 실시간으로 공정을 제어함으로써 선수 공정에서의 부적합한 조업을 후공정에서 만회할 수 있는 기회를 기대할 수 있을 것으로 사료된다. 이를 통해 공정 간의 상관관계를 파악할 수 있을 것이며, 공정 스마트화를 통해 생산성을 향상시킬 수 있을 것으로 기대된다.

Acknowledgements

This results was supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (MOE) (2021RIS-003)

References

[1] Ahn, Y. S., Lee, Y. H. and Lee, Y. D., Effect of Ti, B, Element and δ-ferrite on the Hot Ductility of STS 304, *Korean Journal of Metals and Materials*, 1992, Vol. 30, No. 7, pp. 799-806.

[2] Ahn, Y., Lee. Y., and Lee, Y, Effect of Ti, B.Element and δ-ferrite on the Hot Ductility of STS304, *Journal of the Korean Inst. of Met. & Mater.*, 1992, Vol. 30, No. 7, pp. 799-806.

[3] Bae, C. M. and Nam, W. J., Effects of Sulfur and Boron Content on Hot Ductility and Machinability of STS303, *Korean Journal of Metals and Materials*, 1994, Vol. 32, No. 6, pp. 706-713.

[4] Brinkman, C. and Gavin, H., *Properties of austenitic stainless steels and their weld metals (influence of slight chemistry variations)*, West Conshohocken, PA: ASTM International, 1979.

[5] Cheruvathur, S., Lass, E. A., and Campbell, C. E., Additive manufacturing of 17-4 PH stainless steel: post-processing heat treatment to achieve uniform reproducible microstructure, *JOM*, 2016, Vol. 68, No. 3, pp. 930-942.

[6] Díaz-Uriarte, R. and De Andres, S. A., Gene selection

and classification of microarray data using random forest, *BMC Bioinformatics*, 2006, Vol. 7, No. 1, pp. 1-13.

[7] Jeong, Y. and Kang, C., Defect Type Prediction in Manufacturing Process using Association Rules, *KSIE Spring Conference*, 2004, pp. 310-313.

[8] Jung, J., Effects of Homogenization Treatment Mn/S ratio and δ-ferrite on the Hot Workability of Free-Machining 303-series Austenitic Stainless Steels, *Korean Journal of Metal and Materials*, 2017, Vol. 55, No. 12, pp. 880-887.

[9] Kang, D., Oh, S., Park, J., and Seo, M., Predicting the quality of steelplate products using artificial neural networks, *KSQM Autumn Conference*, 2019, p. 146.

[10] Kim, K. H. and Baek, J. G., A Prediction of Chip Quality using OPTICS(Ordering Points to Identify the Clustering Structure)-based Feature Extraction at the Cell Level, *Journal of the Korean Institute of Industrial Engineers*, 2014, Vol. 40, No .3, pp. 257-266.

[11] Kim, S. W., Effect of δ-Ferrite on the Hot Workability and Surface Defect of STS 304 Billets Containing 3 wt.% Cu, *Korean Journal of Materials Research*, 2004, Vol. 14, No. 6, pp. 379-388.

[12] Kim, Y. S., Park, J. G., Ahn, D. C., and Kim, Y. H., Review of Formability and Forming Property for Stainless Steel, *Transactions of Materials Processing*, 2011, Vol. 20, No. 3, pp.193-205.

[13] Lee, S., Kim, Y., and Lee, Y, Effect of S and Mn on the Hot Workability of STS 316L and 309S steels, *Journal of the Korean Inst. of Met. & Mater*, 1998, Vol. 36, No. 10, pp. 1590-1598.

[14] Lee, Y., Sim, S., and Kim. D., Effect of castring structure and S contents on the how workability of 18-8 stainless steel, *Korean Journal of Metal and Materials*, 1995, Vol. 1995, No. 1, p. 204.

[15] Liu, H., Hu, D., and Fu, J., Analysis of MnS inclusions formation in resulphurised steel via modeling and experiments, *Materials*, 2019, Vol. 12, No. 2, p. 2028.

[16] Maciejewski, J., The effects of sulfide inclusions on mechanical properties and failures of steel components, *Journal of Failure Analysis and Prevention*, 2015, Vol. 15, No. 2, pp. 169-178.

[17] Nazábal, J. L., Urcola, J. J. and Fuentes, M., High-temperature deformation characteristics of free-machining steels, *Metals Technology*, 1982, Vol. 9, No. 1, pp. 323-326.

[18] Nkonyana, T., Sun, Y., Twala, B., and Dogo, E., Performance evaluation of data mining techniques in steel manufacturing industry, *Procedia Manufacturing*, Vol. 35, pp. 623-628.

[19] Oliver, J., Jonsson, J. Y. and Talonen, J., Method for manufacturing and utilizing ferritic-austenitic stainless steel with high formability, *U.S. Patent Application*, 2013.

[20] Osakada, K., Effects of strain rate and temperature in forming processes of metals, *Le Journal de Physique IV*, 1997, Vol. 7, No. C3, p. C3-XXXVII.

[21] Park, J., Lee, J., and Seo, Kyung., Reliability Evaluation of Small Rolling Process Temperature Prediction Using Machine Learning, *Journal of Institute of Control, Robotics and Systems*, 2020, Vol. 26, No. 6, pp. 412-422.

[22] Ruiz, E., Ferreño, D., Cuartas, M., López, A., Arroyo, V., and Gutiérrez-Solana, F., Machine learning algorithms for the prediction of the strength of steel rods: an example of data-driven manufacturing in steelmaking, *International Journal of Computer Integrated Manufacturing*, 2020, Vol. 33, No. 9, pp. 880-894.

[23] Sa, G., Lee, S., Jang, Y. and Park, C., Analysis of Data from a Rubber Manufacturing Process Based on Hadoop Ecosystem and Machine Learning for Smart Factor, *Journal of KIISE*, 2020, Vol. 26, No. 12, pp. 519-527.

[24] Toledo, G. A., Campo, O., and Lainez, E., Influence of sulfur and Mn/S ratio on the hot ductility of steels during continuous casting, *Steel Research*, 1993, Vol. 64, No. 6, pp. 292-299.

ORCID

Seokjun Seo | https://orcid.org/0000-0002-9479-4811

Heungseob Kim | https://orcid.org/0000-0003-0090-5670

〈Appendix〉

Chapter	Research process	Used packages
Ch. 2	· Building dataset	{data.table}, {dplyr}
Ch. 3	· Exploratory Data Analysis	{ggplot2}, {hrbrthemes}, {tidyr}, {viridis}, {gridExtra}
	· Data Preprocessing	{data.table}, {dplyr}, {proxy}, {NbClust}
	· Dimensionality reduction	{corrplot}. {caret}. {glmnet}, {randomForest}, {xgboost}
Ch. 4	· Machine learning	{smotefamily}, {e1071}, {glmnet}, {randomForest}, {xgboost}
	· Model evaluation	{caret}, {MLmetrics}