Journal of the Korean Society of Safety
Vol. 36, No. 3, pp. 31-39, June 2021
https://doi.org/10.14346/JKOSOS.2021.36.3.31

ISSN 1738-3803 (Print)
ISSN 2383-9953 (Online)
http://www.kosos.or.kr/jkosos

건설현장의 공사사전정보를 활용한 사망재해 예측 모델 개발

최승주* · 김진현** · 정기효***†

Development of Prediction Models for Fatal Accidents using Proactive Information in Construction Sites

Seung Ju Choi* · Jin Hyun Kim** · Kihyo Jung***†

†Corresponding Author
Kihyo Jung
Tel : +82-52-259-2709
E-mail : kjung@ulsan.ac.kr

Received : January 15, 2021
Revised : April 14, 2021
Accepted : April 19, 2021

Abstract : In Korea, more than half of work-related fatalities have occurred on construction sites. To reduce such occupational accidents, safety inspection by government agencies is essential in construction sites that present a high risk of serious accidents. To address this issue, this study developed risk prediction models of serious accidents in construction sites using five machine learning methods: support vector machine, random forest, XGBoost, LightGBM, and AutoML. To this end, 15 proactive information (e.g., number of stories and period of construction) that are usually available prior to construction were considered and two over-sampling techniques (SMOTE and ADASYN) were used to address the problem of class-imbalanced data. The results showed that all machine learning methods achieved 0.876~0.941 in the F1-score with the adoption of over-sampling techniques. LightGBM with ADASYN yielded the best prediction performance in both the F1-score (0.941) and the area under the ROC curve (0.941). The prediction models revealed four major features: number of stories, period of construction, excavation depth, and height. The prediction models developed in this study can be useful both for government agencies in prioritizing construction sites for safety inspection and for construction companies in establishing pre-construction preventive measures.

Key Words : occupational safety, construction, safety inspection, machine learning, imbalanced data

1. 서 론

건설현장은 다양한 중장비 혼재와 가설구조 등으로 인해 심각한 위험요인들이 근본적으로 잠재되어 있어 다른 산업에 비해 산업재해의 위험이 높다[1,2,3]. 영국의 경우 건설업의 치명적 부상비율은 전체 산업대비 3배 이상 높은 것으로 보고되고 있다[4]. 또한, 미국은 건설업의 위험성이 일반 제조업과 석유화학업 대비 2배와 6배 높은 것으로 알려지고 있다. 한국은 전체 산업재해 사고사망의 절반 이상이 건설업에서 발생하고 있으며, 건설업의 사고사망만인율은 전체 산업 대비 3.7배 높은 것으로 보고되고 있다[5].

기존 연구들은 건설현장의 산업재해 경감을 위해 사고인과모델 및 재해위험요인 등을 분석 및 제시하고 있다. Choi and Cho[4], Lim et al.[6], Zhang et al.[7], 그리고 Jeong and Jeong[8]은 건설현장의 사고원인 분석을 위한 시스템적 관점의 인과모델과 위험분류체계를 구축하였다. Yang and Paik[9], Park and Han[10], Jang and Go[11], Kim and Shin[12], 그리고 Kim et al.[13]은 재해통계 및 설문조사를 통해 건설기계, 공종, 추락재해 등에 대한 위험성평가 방안 및 개선 방안을 제안하였다. Lee et al.[14]와 Ha et al.[15]는 건설현장의 사망재해 다중요인과 건설공사의 위험도를 도출하였다. 마지막으로, Kim et al.[16]과 Chi et al.[17]는 건설업의 재해위험요인인 화재 및 감

*울산대학교 안전보건전문학과 박사과정 (Department of Safety Engineering, University of Ulsan)
**한국산업안전보건공단 산업안전보건연구원 산업안전연구실장 (Occupational Safety and Health Research Institute, Korea Occupational Safety and Health Agency)
***울산대학교 산업경영공학부 교수 (School of Industrial Engineering, University of Ulsan)

전사고 분석을 통해 사고 패턴과 예방 방법을 제안하였다.

건설현장의 위험요인을 체계적으로 통제하기 위해서는 상술한 재해모델 및 위험요인 분석과 같은 이론적인 연구에 더해 재해위험이 높은 건설현장에 대한 정부 차원의 안전점검이 필요하다. 산업안전보건 규제와 개입은 사업장의 사망과 부상 재해를 효과적으로 감소시키는데 공헌하는 것으로 알려지고 있다[18]. 예를 들면, 미국 캘리포니아의 산업안전보건청(Cal/OSHA)은 재해율이 높은 사업장 중에서 일부를 선정하여 안전점검을 실시하여 재해율을 감소시키는데 공헌하였다고 보고하였다[19]. 이러한 이유로 우리나라도 2019년부터 사고사망이 다발하고 있는 고위험 건설현장에 대해 집중적인 안전점검(특별기획점검 또는 패트롤 점검)을 실시하고 있다.

전체 건설현장을 대상으로 안전점검을 진행하는 것은 인력 및 자원의 한정으로 인해 현실적으로 어렵다. 따라서 효율적인 안전점검을 위해 재해발생의 위험도가 높은 건설현장을 선별하는 것이 필요하다. 현행 접근법은 중대재해가 발생하면 재해를 유발한 정보(예: 기인물)를 파악하여 동종의 재해를 예방할 수 있도록 위험 현장의 선정 및 점검을 진행하고 있다. 예를 들면, 타워크레인의 전복으로 인해 중대재해가 발생하면 타워크레인을 사용하는 공사현장에 대한 안전점검을 수행한다. 이러한 수동적(사후) 대응은 한계가 있으며 안전사고를 근본적으로 예방하기 위해서는 보다 적극적인(선제적인) 대응이 필요하다[20]. 한편, 공사 사전정보는 사고예방 방안을 선제적으로 찾는 데 도움을 주는 선행지표로 활용될 수 있는 것으로 알려지고 있다[21]. 공사 착공 전에 알 수 있는 공사사전정보를 활용하여 건설현장의 재해위험도를 예측할 수 있다면, 공사 착공 전 또는 직후에 해당 건설현장을 안전점검 대상으로 선별하여 효과적으로 점검할 수 있으며, 예측된 위험도를 고려하여 건설현장의 안전관리 체계 및 안전예산을 전향적으로 편성하는 근거로도 활용할 수 있다.

본 연구는 공사 착공 전에 확인할 수 있는 공사사전정보를 활용하여 건설현장의 사고사망 발생을 예측하는 기계학습 모델을 개발하였다. 본 연구는 기계학습을 위해 1,079개의 건설현장에서 사전적 정보(15종)와 사고사망 통계를 수집하였다. 또한, 본 연구는 수집된 자료의 표본불균형을 보정하기 위해 과표본화(Over-sampling)를 적용하였다. 마지막으로, 본 연구는 유관연구[22]와 예측 성능의 우수성에 대한 기존 연구를 참고하여 5종(Support vector machine (SVM), Random Forest (RF), Extreme gradient boosting (XGBoost), Light gradient boosting machine (LightGBM), Automated machine learning (AutoML))의 대표적 기계학습 방법을 적용하여 예측 모델을 개발하고 성능을 비교 평가하였다. SVM은 기계학습 분야에서 가장 기본이 되는 기법이며, RF, XGBoost, LightGBM은 의사결정나무를 기반으로 하는 앙상블(Ensemble) 학습을 통해 예측 정확성을 향상시킨 기법이다. 마지막으로, AutoML은 여러 가지 기계학습 방법 중에서 최적의 결과를 자동으로 도출하여 사용하는 가장 최근에 개발된 기법이다. 본 연구의 예측 모델은 한정된 자원을 활용하여 더욱 효과적으로 건설현장을 관리할 수 있도록 사고사망 발생 가능성이 높은 건설현장을 파악하여 안전점검 대상의 우선순위를 결정하는데 유용하게 활용될 수 있을 것으로 기대된다.

2. 연구 방법

본 연구는 Fig. 1과 같이 4단계 과정을 통해 건설현장의 사고사망 위험도를 예측하는 모델을 개발하였다. 첫째 단계에서는 기계학습을 통한 예측 모델을 만드는데 사용할 데이터를 수집하고 전처리를 수행하였다. 둘째 단계에서는 다양한 기계학습 방법(SVM, RF, XGBoost, LightGBM and AutoML)을 적용하여 건설현장의 사고사망을 예측하는 모델을 구축하였다. 셋째 단계에서는 도출된 예측 모델의 적합성을 검증하기 위해 성능 평가 척도를 선정하고 교차검증 방법을 적용하여 성능을 객관적으로 비교 분석하였다. 마지막 단계에서는 도출된 예측 모델을 분석하여 사고사망에 영향을 미치는 주요 인자를 도출하였다.

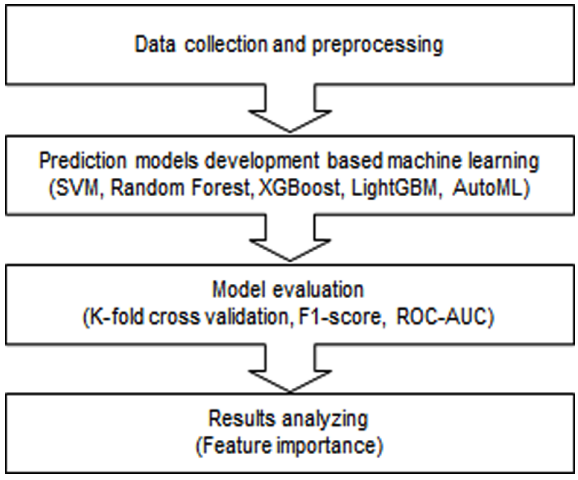

Fig. 1. Research process to develop prediction models for fatal accidents in construction sites.

2.1 데이터 수집 및 전처리

본 연구는 총 1,079개 건설현장의 공사 및 공법 정보(15종의 사전정보)와 산재 사고사망자 통계자료(각 건설현장의 사고 유무)를 활용하였다. 본 연구에서 고려된 15종의 사전정보는 건축계획서 및 건축허가 신청서 등에서 공통적으로 확인할 수 있는 정보로서 건설현장의 기본적인 특징을 나타내고 수집이 용이하다. 산재 사고사망자 통계자료는 지도학습을 위해 각 건설현장에서의 사고사망이 발생하였는지 여부로 조사되었다. 본 연구는 2014년부터 2020년까지의 건설현장 중에서 아파트, 근린생활시설, 기숙사 등과 같은 대규모 건설현장을 대상으로 하였다. 본 연구는 중대재해 위험이 높은 건설현장을 예측하기 위해 Table 1에 나타낸 공사 및 공법 정보(공사기간, 지하층수, 지상층수, 개소, 최고높이, 굴착깊이, 공사금액, 구조, 대지면적, 건축면적, 연면적, 외부비계, 외부마감, 흙막이 벽체공법, 지지공법)를 사용하였다. 그리고 산재 사고사망자는 연구 대상 건설현장 중에서 111개(전체 데이터의 10.3%)에서 발생한 것으로 나타났다. 공사사전정보(15종)에 따른 사고사망 발생 여부에 대한 자료는 Fig. 2와 같이 연속형 변수는 평균에 대한 95% 신뢰구간을 나타내는 구간 그림(Interval plot)으로 표시하였으며, 범주형 변수는 상위 5개 범주를 중심으로 구성 비율을 나타내었다.

Table 1. Definition of features and factors considered in this study

Feature name	Description	Type
Period of construction	Period in time during which the building was constructed	Continuous
Basement level	Number of floors below ground	Continuous
No. of stories	Number of floors including the ground level	Continuous
No. of buildings	Number of buildings constructed	Continuous
Height	The highest height of buildings	Continuous
Excavation depth	Depth of excavation for constructions	Continuous
Construction cost	Total cost of the construction project	Continuous
Structure	Frame materials of a building (reinforced concrete, steel frame/reinforce concrete, steel reinforced concrete, steel reinforced concrete/reinforced concrete, etc)	Categorical (8 categories)
Site area	Area of the land where a building is to be built	Continuous
Building area	Area occupied by building on the ground	Continuous
Total floor area	Sum of the floor area of a building	Continuous
Scaffolding	Type of scaffolding installed externally (steel pipe, system, steel pipe/system)	Categorical (3 categories)
External finishing	the finishing material such as ‘painting/stone’, ‘stone’, ‘painting’, ‘stone/panel’, ‘panel’, etc.	Categorical (25 categories)
Retaining wall method	Type of retaining wall constructed to withstand lateral pressure of soil (CIP, braced wall, braced wall/CIP SCW, braced wall/SCW, etc)	Categorical (19 categories)
Excavation support method	Type of excavation support system to prevent collapse of soil (STRUT, STRUT/RAKER, EA, STRUT/EA, STRUT/EA/RAKER, etc)	Categorical (37 categories)

본 연구에 사용된 공사 및 공법 정보는 데이터의 유형을 고려하여 전처리되었다. 연속형 정보(예: 공사기간, 공사금액)는 데이터 간의 규모(Scale) 차이가 크기 때문에 정규화(Normalization)되었다. 데이터 정규화는 기계학습 시 변수별 영향력의 차이를 줄이고, 학습 속도를 개선하며, 지역 최적화(Local optimum)의 가능성을 줄일 수 있다[23]. 한편, 범주형 정보(예: 구조, 외부비계)는 다양한 기계학습 방법에 범용적으로 적용될 수 있도록 원 핫 인코딩(One hot encoding)하였다. 마지막으로 결측치가 있는 경우 해당 데이터는 모두 제거(149개)되었다.

2.2 기계학습 모델 개발

2.2.1 예측 모델

본 연구는 건설현장의 사전 정보를 이용하여 사망재해가 발생할 가능성이 높은 현장을 선별하기 위한 기계학습 기반의 예측모델을 구축하였다. 이를 위해, 본 연구에서는 적용 데이터의 특성에 따라 우수한 기계학습 방법이 상이하다는 기존 연구를 참고하여[24], 5가지 기계학습 알고리즘(SVM, RF, XGBoost, LightGBM and AutoML)의 성능을 정량적으로 비교 평가하였다. 먼저 SVM은 분류 문제를 예측하는 기계학습 방법으로[25], 본 연구는 사고사망의 발생 유무를 분류하도록 개발되었다. RF는 다수의 의사결정나무를 학습한 후 다수결로 투표한 결과를 종합하여 사고사망의 발생 유무를 판별하도록 구현되었다[26]. XGBoost는 의사결정나무를 순차적으로 학습하는 부스팅(Boosting) 방법의 단점인 과적합 문제를 해결한 기법이다[27]. LightGBM은 XGBoost와 같이 부스팅을 활용하나 리프 중심(Leafwise) 방식을 채용한 차이가 있다[28]. AutoML은 데이터 전처리, 특징 분석, 분류 알고리즘 선택, 초매개변수의 최적화 등을 통계적 기법을 활용하여 자동화한 학습 방법이다[29,30]. 본 연구는 다양한 AutoML 중에서 Auto-sklearn 방법을 사용하였다.

본 연구의 예측 모델은 파이썬(v. 3.6.5)를 사용하여

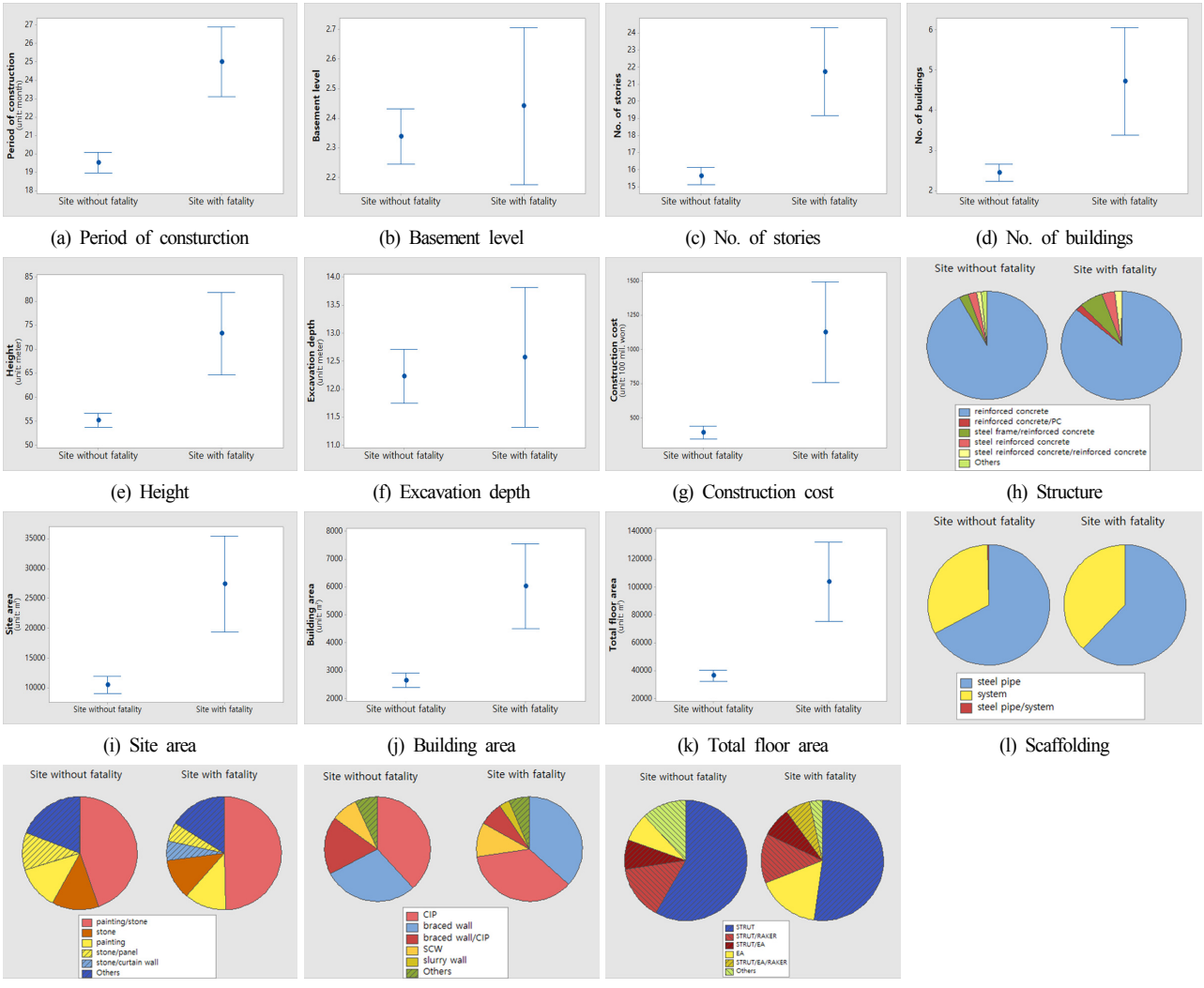

Fig. 2. Data distribution by fatality.

구현하였다. 데이터 처리 및 기계학습 및 성능평가를 위해 pandas, numpy, matplotlib, scikit-learn, imbalanced-learn, XGBoost, LightGBM, auto-sklearn, optuna 등의 라이브러리를 사용하였다.

2.2.2 초매개변수

본 연구는 기계학습 알고리즘의 초매개변수(Hyper-parameters)를 Optuna[32]를 활용하여 결정하였다. 본 연구에 고려된 기계학습 알고리즘은 데이터 학습을 위해 초매개변수 설정이 필요하다. 초매개변수는 예측 모델의 성능에 영향을 주기 때문에 적절하게 설정하는 것이 중요하다[31]. 따라서 초매개변수의 최적 조건을 찾기 위해 Optuna, 격자탐색(Grid search), Hyperopt 등의 다양한 방법이 사용되고 있다. 본 연구에서는 Optuna를 활용하여 예측 모델의 초매개변수를 최적화하였다.

2.2.2. 과표본화

본 연구는 사망재해 데이터의 불균형을 보완하기 위해 과표본화를 적용하였다. 일반적인 기계학습 알고리즘은 계급의 균형 분포(balanced class distribution)를 가정하고 전체 데이터의 오분류를 최소화 하도록 예측 모델을 구성한다[33]. 따라서 사망재해와 같이 발생빈도가 상대적으로 적은 불균형(imbalance) 데이터(본 연구의 사망재해는 전체 데이터의 약 10%)는 사망재해에 대한 예측 성능이 저하되는 한계점이 있다[34]. 불균형 데이터 문제를 해결하기 위해 다수 데이터를 줄이는 과소표본화(Under-sampling)나 소수 데이터를 늘리는 과표본화(Over-sampling)를 수행하게 된다[35]. 본 연구는 불균형 데이터 문제 해결을 위해 과표본화 기법인 Synthetic minority over-sampling technique (SMOTE)와 Adaptive synthetic sampling (ADASYN)을 적용하였다.

SMOTE는 k nearest neighbor (KNN) 알고리즘을 사용해 소수 계급의 데이터를 보간하여 새로운 데이터를 만들어 추가하는 방법이다[36]. ADASYN은 SMOTE를 적용한 보간 시 데이터의 밀도 분포를 가중치로 추가 고려하는 방법이다[37]. 본 연구는 소수 데이터(사고사망 발생)의 수를 다수 데이터(사고사망 미발생) 수와 동일하게 되도록 과표본화 비율을 1(사고사망 발생):1(사고사망 미발생)로 설정하였다.

2.3 예측 모델의 성능 평가

본 연구는 *k*겹 교차검증(*k*-fold cross validation)을 활용하여 예측 모델의 성능을 비교 평가하였다. 예측 모델의 일반화된 성능 향상을 위해서는 학습에 사용되는 데이터와 평가에 사용되는 데이터를 분리해야 한다. *k*겹 교차검증은 데이터를 무작위로 *k*개의 집단으로 나누고 *k*-1개는 학습에 사용하고 나머지 하나는 평가에 사용한다. 이러한 과정을 *k*번 반복하여 계산된 성능의 평균을 예측 모델의 최종 성능으로 간주한다. *k*겹 교차검증은 데이터가 적은 경우 정확도를 향상시킬 수 있으며, 학습 데이터와 평가 데이터로 한번 나눈 것보다 더 일반화된 결과를 얻을 수 있다[38]. 본 연구는 *k*가 5인 교차검증을 사용하였다. 한편, 교차검증 시 과표본화는 학습에 사용할 k-1개의 데이터 집단에 대해서만 적용하였고, 평가에 사용할 데이터에는 적용하지 않았다.

본 연구의 정량적 성능 평가는 F1-score와 AUC (Area under curve)를 적용하였다. 예측모델의 성능 평가 척도는 예측 성공 여부를 나타내는 Table 2의 혼동행렬(Confusion matrix)에 기반한다. 기계학습 분야에서 널리 사용되는 척도에는 정확도(Accuracy, 식 1), 정밀도(Precision, 식 2), 재현율(Recall, 식 2) 등이 있다.

Table 2. Example of confusion matrix

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

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

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN} \quad (2)$$

정확도는 전체 데이터 중에서 정확히 예측된 비율을 의미하므로, 본 연구와 같이 데이터가 불균형할 경우 다수 계급(예: 사고사망 미발생)만을 잘 예측해도 높은 정확도를 가지게 되므로 적합하지 않다. 반면, 재현율은 사망재해가 발생한 데이터에 대해 예측 모델이 정확히 분류한 정도(Sensitivity)를 나타내는 것으로 사망재해와 같이 가능한 많이 찾아내는 것이 중요한 경우에 사용되고 있다[39]. 정밀도는 사망재해가 발생할 것으로 예측된 경우 중에서 정확히 예측된 정도를 나타내는 지표이다. 일반적으로 정밀도와 재현율은 반비례 관계를 가지고 있어, 모델의 성능을 충분히 나타내기에는 한계가 있다. 따라서 본 연구에서는 정밀도와 재현율을 동시에 고려(조화평균)할 수 있는 F1-score(식 3)를 적용하였다.

$$F1-score = 2\times\frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\times\frac{Precision\times Recall}{Precision+Recall} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

ROC (Receiver operation characteristic) 곡선은 민감도(Sensitivity)와 특이도(Specificity, 식 4)의 관계를 곡선으로 나타낸 것으로 모델의 성능을 시각화한 것이다. 민감도는 재현율과 같으며, 특이도는 관심이 없는 계급에 대한 재현율(사망재해가 발생하지 않을 것으로 예측된 경우 중에서 정확히 예측된 정도)을 의미한다. ROC 곡선의 성능은 곡선 아래의 면적을 계산한 AUC를 사용해 정량화된다. AUC는 0~1사이의 값을 가지며 수치가 높을수록 성능이 우수함을 의미한다. 본 연구는 ROC 곡선이 불균형 데이터의 예측 정확도를 평가하는데 많이 사용되고 있는 추세[40]를 고려하여 F1-score와 함께 예측 모델의 성능 평가에 사용하였다.

3. 연구 결과

3.1 예측 모델의 성능 비교

과표본화를 적용한 모델의 F1-score는 Table 3에 나타낸 것과 같이 원시 데이터를 활용한 경우보다 현저히 높은 것으로 분석되었다. 과표본화 적용 모델의 평균 F1-score는 기계학습 방법에 따라 0.918~0.929의 범위를 보였다. 반면, 원시 데이터 활용 모델의 평균 F1-score는 0~0.181의 범위를 보여 현저히 낮은 것으로 분석되었다.

Table 3. F1-score for machine learning algorithms with/without over-sampling

Algorithm	Raw data	Over-sampling data		
		SMOTE (A)	ADASYN (B)	Difference (A-B)
SVM	0.181*	0.876	0.904	-0.028
RF	0.133	0.940*	0.939	0.001
XGBoost	0.150	0.921	0.936	-0.015
LightGBM	0.067	0.940*	0.941*	-0.001
AutoML	0.000	0.914	0.926	-0.012
Average	0.106	0.918	0.929	-0.011

* The best algorithm is highlighted for each data type.

과표본화 기법인 ADASYN은 SMOTE보다 F1-score가 다소 높은 것으로 나타났다. ADASYN의 F1-score는 기계학습 방법에 따라 0.926~0.941의 범위를 보이는 것으로 분석되었다. 반면, SMOTE의 F1-score는 0.876~0.940의 범위를 보여 ADASYN보다 평균적으로 0.011 낮은 것으로 파악되었다.

LightGBM은 두 가지 과표본화 기법에 대해 공히 가장 높은 F1-score를 보이는 것으로 파악되었다. RF와 LightGBM은 과표본화 기법인 SMOTE에 대해 가장 높은 F1-score(0.940)를 보였다. 한편, LightGBM은 ADASYN에 대해 가장 높은 F1-score(0.941)를 보이는 것으로 분석되었다.

가장 성능이 높았던 과표본화 기법인 ADASYN에 대한 ROC 곡선은 Fig. 3과 같다. ROC곡선 아래의 면적인 AUC는 기계학습 방법에 따라 0.904~0.941의 범위를 보이는 것으로 나타났으며, 이러한 결과는 F1-score와 매우 유사한 것으로 분석되었다.

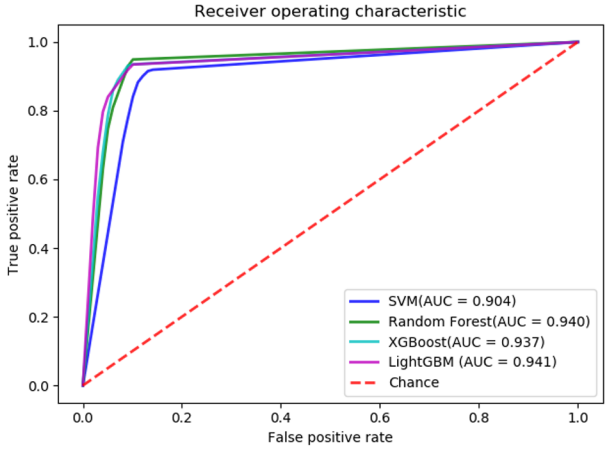

Fig. 3. ROC curve for machine learning algorithms with ADASYN (the ROC curve for AutoML is not visualized since the visualization function is not implemented in the AutoML library yet).

3.2 공사사전정보의 중요도 분석

본 연구는 과표본화 데이터에 대해 가장 성능이 우수한 RF와 LightGBM에 대해 공사사전정보의 중요도를 Table 4와 같이 도출하였다. 중요도는 예측 성능에 영향을 미치는 정도를 나타내는 것으로, 값이 클수록 영향력이 크다고 해석할 수 있다. 본 연구는 기계학습 기법에 따라 중요도의 표현방식이 달라 상대적 비교가 가능하도록 백분율로 환산하였다. 가장 중요도가 높은 사전적 정보는 두 기계학습 방법에서 동일하게 지상층수로 파악되었다. 또한, 공사기간, 굴착깊이, 최고높이는 공통적으로 두 기계학습 방법에서 상위 5위 내에 포함되었다. 반면, 지하층수는 RF에서 두 번째로 중요하였으나, LightGBM에서는 9번째로 중요한 요인으로 나타나 차이가 있는 것으로 분석되었다.

Table 4. Feature importances of two major algorithms with ADASYN

Rank	Random Forest		LightGBM	
	Feature	Importance	Feature	Importance
1	No. of stories	9.555	No. of stories	11.834
2	Basement level	8.035	Period of construction	10.918
3	Period of construction	7.857	Excavation depth	10.550
4	Height	7.485	Height	10.071
5	Excavation depth	7.473	Construction cost	9.146
6	Total floor area	6.852	Site area	8.990
7	Site area	6.537	Building Area	8.174
8	Construction cost	6.521	Total floor area	8.131
9	Building Area	5.805	Basement level	5.575
10	No. of buildings	4.568	No. of buildings	3.573

4. 결론 및 고찰

본 연구는 건설현장의 사망재해 감소를 목적으로 사망재해 발생 위험 현장을 예측하는 기계학습 모델을 개발하였다. 이를 위해 2014년부터 2020년까지의 건설현장의 사전적 정보(예: 공사기간, 층수, 건축면적)를 활용하였다. 이러한 사전적 정보의 활용은 공사 착공 전 또는 공사 초기에 건설현장의 사고 위험도를 예측할 수 있다는 장점이 있다. 또한, 본 연구는 다양한 기계학습 방법의 예측 성능을 비교 평가하여 본 연구의 목적에 가장 부합하는 기계학습 방법(예: LightGBM)을 식별하였다는 측면에서 의의가 있다.

본 연구는 사망재해와 같은 불균형 데이터에 대한 기계학습을 위해 과표본화 기법을 활용하였다. 건설현

장은 다른 산업에 비해 사고사망의 심각성이 상대적으로 높으나, 사망재해가 발생한 건설현장은 전체 현장 대비 소수에 불과하다(본 연구의 경우 약 10%). 이러한 데이터 불균형은 기계학습의 편이(bias)를 야기할 수 있다. 예를 들면, 본 연구의 원시 데이터에 대한 기계학습 모델은 F1-score가 0.2 이하로 매우 낮게 나타나 불균형 데이터를 그대로 사용할 경우 기계학습이 적합하게 이루어지지 못하는 것으로 나타났다. 반면, 과표본화 기법인 SMOTE와 ADASYN을 적용할 경우, 기계학습 모델의 F1-score가 0.88 이상으로 증가하여 데이터 불균형으로 인한 영향을 효과적으로 통제할 수 있는 것으로 파악되었다.

건설현장의 사고사망에는 지상층수, 공사기간, 굴착깊이, 최고높이가 상대적으로 중요한 요인인 것으로 나타났다. 본 연구는 RF와 LightGBM의 예측 모델로부터 공사사전정보 중에서 중요한 요인을 도출하였다. 일반적으로 지상층수, 공사기간, 굴착깊이, 최고높이가 증가할 경우 공사의 난이도가 높아져 사망재해의 위험이 높아지는 것으로 해석할 수 있다. 그러나 RF와 LightGBM은 앙상블 기법으로 예측 모델이 복잡하여 각 요인의 독립적인 영향을 정량적으로 계산하기에 어려움이 있다. 그로 인해, 본 연구에서 파악된 중요 요인과 사망재해 간의 연관성을 정량적으로 분석 및 해석할 수 없는 한계점이 있다. 따라서 두 앙상블 기법에서 공통적으로 파악된 중요 요인에 대해 사망재해와의 연관성을 과학적으로 분석하는 후속연구가 필요하다.

본 연구의 결과를 일반화하기 위해서는 보다 많은 데이터에 대한 후속 연구가 필요하다. 본 연구는 최근 7년 동안의 건설현장 사전 정보와 사망재해 통계자료를 활용하여 건설현장의 사고사망 위험도를 예측하는 모델을 개발하였다. 그러나 기계학습의 성능은 학습 데이터의 양과 질에 크게 영향을 받는다[41]. 또한, 입력변수 간에 상관관계가 높을 경우(예: 최고높이와 지상층수) 다중공선성 문제가 발생할 수 있다. 회귀분석에서는 이러한 다중공선성이 부정적 영향을 미치는 것으로 알려지고 있으나, 기계학습 분야에서는 다중공선성으로 인한 부정적 영향과 긍정적 영향이 공존하는 것으로 보고되고 있다. 예를 들면, 다중공선성을 줄이기 위해 상관관계가 높은 변수를 제외할 경우 의사결정나무의 예측성능이 낮아진다는 연구결과도 있다[42]. 따라서 다중공선성에 따른 예측 성능의 변화를 보다 심도있게 다루는 후속연구가 필요하다.

본 연구는 1,079개의 건설 현장에 대한 사전정보를 활용하여 예측 모델을 구축하였으나, 사고사망의 발생건수는 총 130건으로 제한적이었다. 따라서 사고사망과 공사사전정보 간에 연관성은 있으나 그 발생빈도가 낮아 상대적으로 중요도가 낮게 예측된 공사사전정보가 있을 수 있다는 한계점이 있다. 또한, 발생확률이 낮은 사고사망의 복합적인 원인은 데이터 부족으로 인해 예측 모형 개발에 반영되지 못했을 가능성도 있다. 향후 연구로 보다 많은 건설 현장과 재해 데이터를 확보하여 본 연구에서 개발된 예측모델을 검증과 보완하는 것이 필요하다. 또한, 본 연구에서 사용한 공사사전정보에 더해 기존 연구에서 산업재해와 연관이 있을 것으로 보고되고 있는 현장 안전관리 조직, 안전관리 상태, 안전교육 여부, 안전문화 등[43]의 데이터를 축적하여 예측 모델을 보완하는 후속연구가 필요하다.

본 연구에서 개발된 건설현장의 사망재해 예측 모델은 2가지 측면의 실용성이 있다. 첫째, 본 연구의 예측 모델은 안전 감독기관이 과학적인 근거에 의해 안전점검을 수행할 건설현장을 선별할 때 활용될 수 있다. 둘째, 본 연구의 예측 모델은 건설 사업자가 사망재해의 발생 가능성을 공사 전에 파악하고, 사망재해의 위험이 높을 경우 안전 대책을 사전에 마련(proactive)하는 근거로 활용될 수 있다.

Acknowledgement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; NRF-2019R1A2C4070310).

References

1) S. D. Choi and K. Carlson, “Occupational Safety Issues in Residential Construction Surveyed in Wisconsin, United States”, Industrial Health, Vol. 52, No. 6, pp. 541-547, 2014.
2) C. W. Liao and Y. H, Perng, “Data mining for Occupational Injuries in the Taiwan Construction Industry”, Safety Science, Vol. 46, No. 7, pp. 1091-1102, 2008.
3) Y. H. Chiang, F. K. W. Wong and S. Liang, “Fatal Construction Accidents in Hong Kong”, J. Constr. Eng. Manag., Vol. 144, No. 3, 04017121, 2018.
4) Y. G. Choi and K. T. Cho, “A Cause Analysis of the Construction Incident Using Causal Loop Diagram : Safety Culture Perspective”, J. Korean. Soc. Saf. Vol 35, No. 2, pp. 34-46, 2020.
5) Ministry of Employment and Labor (MOEL), “2019 Industrial Accident Statistics”
6) W. J. Lim, J. H. Kee, J. H. Seong and J. Y. Park, “Development of Accident Cause Analysis Model for

Construction Site", J. Korean Soc. Saf., Vol. 34, No. 1, pp. 45-52, 2019.
7) W. Zhang, S. Zhu, X. Zhang and T. Zhao, "Identification of Critical Cause of Construction Accidents in China using Model based on System Thinking and Cause Analysis", Saf. Sci., Vol. 121, pp. 606-618, 2019.
8) J. J. Jeong and J. J. Jeong, "Development of Framework for Integrated Work-Risk Breakdown Structure based on Fatal incident Cases in Construction Industry", Korean Journal of Construction Engineering and Management, Vol. 21, No. 3 pp. 11-19, 2020.
9) S. S. Yang and S. W. Paik, "Risk Management for Preventing Workers' Deaths in Construction Machinery Work", J. Korean Soc. Saf., Vol. 35, No. 3, pp. 16-23, 2020.
10) H. P. Park and J. G. Han, "Development of Risk Assesment Index for Construction Safety Using Statistical Data", J. Korea Inst. Build. Constr. Vol. 19, No. 4, pp. 361-371, 2019.
11) Y. R. Jang and S. S. Go, "A Risk Assessment Counterplan for Reducing the Accident Rates in Medium and Small sized Construction Sites", Korean Journal of Construction Engineering and Management, Vol. 19, No. 5, pp. 90-100, 2018.
12) D. S. Kim and Y. S. Shin, "A Study on the Risk Factors according to the Frequency of Falling Accidents in Construction Sites", J. Korea Inst. Build. Constr., Vol. 19, No. 2, pp. 185-192, 2019.
13) D. Y. Kim, S. M. Yun, J. M. Kim, S. Y. Lee and K. Y. Son, "Improvement of Fall Prevention Method in Construction Site through Comparison with Advanced Countries' Cases", J. Korea Inst. Build. Constr., Vol. 20, No. 5, pp. 471-480, 2020.
14) G. H. Lee, C. S. Lee, C. W. Koo and T. W. Kim, "Identification of Combinatorial Factors Affecting Fatal Accidents in Small Construction Sites: Association Rule Analysis", Korean Journal of Construction Engineering and Management, Vol. 21, No. 4, pp. 90-99, 2020.
15) S. G. Ha, T. H. Kim, K. Y. Son and J. M. Kim, "Quantification Model Development of Human Accidents based on the Insurance Claim Payout on Construction Site", J. Korea Inst. Build. Constr., Vol. 18, No. 2, pp. 151-159, 2018.
16) H. S. Kim, S. C. Jang and J. G. Joo, "Preventive Priority Methods Based on the Analysis of Fire Accident Causes in Construction Site", J. Korea Institute of Construction Safety, Vol. 2, No. 2, pp. 50-55, 2019.
17) C. F. Chi, C. C. Yang and Z. L. Chen, "In-depth Accident Analysis of Electrical Fatalities in the Construction Industry", Int. J. Ind. Ergo., Vol. 39, No. 4, pp. 635-644, 2009.
18) J. H. Andersen, P. Malmros, N. E. Ebbehoej and E. M. Flachs, E. Bengtsen and J. P. Bonde, "Systematic Literature Review on the Effects of Occupational Safety and Health (OSH) interventions at the workplace", Scand J. Work Environ. Health, Vol 45, No. 2, pp. 103-113, 2019.
19) D. I. Levine, M. W. Toffel and M. S. Johnson, "Randomized Government Safety Inspections Reduce Worker Injuries with No Detectable Job Loss", Science, Vol. 336, No. 6083, pp. 907-911, 2012.
20) S. Sarkar et al., "Predicting and Analyzing Injury Severity: A Machine Learning-based Approach using Class-imbalanced Proactive and Reactive Data", Saf. Sci., Vol. 125, 2020.
21) H. Li et al., "Proactive Behavior-based Safety Management for Construction Safety Improvement", Saf. Sci., Vol. 75, pp. 107-117, 2015.
22) R. Zhu et al., "Application of Machine Learning Techniques for Predicting the Consequences of Construction Accidents in China", Process Saf. Environ. Prot. Vol. 145, pp. 293-302, 2021.
23) M. Shanker, M. Y. Hu and M. S. Hung, "Effect of Data Standardization on Neural Network Training", Omega, Vol. 24, No. 4, pp. 385-397, 1996.
24) H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering, Vol. 2, No. 4, pp. 2231-2307, 2012.
25) V. N. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, 1999.
26) L. Breiman, "Random Forests", Machine Learning, Vol. 45, pp. 5-32, 2001.
27) T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", KDD '16: Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
28) G. Ke, Q. Meng, T. Finley and T. Wang, W. Chen, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems (NIPS 2017), Vol. 30, pp. 3146-3154, 2017.
29) Y. H. Moon, I. H. Shin, Y. J. Lee and O. G. Min, "Recent Research & Development Trends in Automated Machine Learning", Electronics and Telecommunications Trends, Vol. 34, No. 4, pp. 32-42, 2019.
30) M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M.

Blum and F. Hutter, “Efficient and Robust Automated Machine Learning”, Advances in Neural Information Processing Systems (NIPS 2015), Vol. 28, pp. 2962-2970, 2015.

31) M. Claesen and B. D. Moor, “Hyperparameter Search in Machine Learning”, arXiv preprint arXiv:1502.02127, 2015.
32) T. Akiba, S. Sano, T. Yanase T. Ohta and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework”, KDD '19: Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discoverty and Data Mining, pp. 2623-2631, 2019.
33) H. He and E. A. Garcia, “Learning from imbalanced data”, IEEE Transactions on Knowledge and Data Engineering, Vol 21, No. 9, pp. 1263-1284, 2009.
34) S. Ertekin, J. Huang, L. Bottou and L. Giles, “Learning on the border: Active Learning in Imbalanced Data Classification”, CIKM '07: Proc. of 16th ACM conference on Conference on Information and Knowledge Management, pp. 127-136, 2007.
35) M. J. Son, S. W. Jung and E. J. Hwang, “A Deep Learning Based Over-Sampling Scheme for Imbalanced Data Classification”, KIPS Trans. Softw. and Data Eng., Vol. 8, No. 7, pp. 311-316, 2019.
36) N. V. Chawla, L. O. Hall, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, 2002.
37) H. He, Y. Bai, E. A. Garcia and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning”, 2008 IEEE International Joint Conference on Neural Networks, pp. 1322-1328, 2008.
38) S. Arlot and A. Celisse, “A survey of Cross-validation Procedures for Model Selection”, Statistics Surveys, Vol. 4, pp. 40-79, 2010.
39) Y. Fang, Y. Zhang and C. Huang, “Credit Card Fraud Detection Based on Machine Learning”, Computers, Materials & Continua, Vol. 61, No. 1, pp. 185-195, 2019.
40) G. Haixiang, L. Yijing, J. Shang and G. Mingyun, H. Yuanyue and G. Bing, “Learning from Class-imbalanced Data: Review of Methods and Applications,” Expert Systems with Applications, Vol. 73, pp. 220-239, 2017.
41) J. H. Choi and H. G. Ryu, “Analysis of Occupational Injury and Feature Importance of Fall Accidents on the Construction Sites using Adaboost”, J. Architectural Institute of Korea Structure & Construction, Vol. 35, No. 11, pp. 155-162, 2019.
42) S. Piramuthu, “Input Data for Decision Trees”, Expert Systems with Applications, Vol. 34, pp. 1220-1226, 2008.
43) S. I. Choi and H. Kim, “A Study on the Safety Climate and Worker's Safe Work Behavior in Construction Site”, J. Korean Soc. Saf., Vol. 21, No. 5, pp. 60-71, 2006.