KSCE Journal of Civil and Environmental Engineering Research
Vol. 42, No. 6: 843-852/ December, 2022
DOI: https://doi.org/10.12652/Ksce.2022.42.6.0843

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

Construction Management **시공관리**

빅데이터를 활용한 AI 기반 우선점검 대상현장 선정 모델

황윤호* · 지석호** · 이현승*** · 정현준****
Hwang, Yun-Ho*, Chi, Seokho**, Lee, Hyeon-Seung***, Jung, Hyunjun****

AI-based Construction Site Prioritization for Safety Inspection Using Big Data

ABSTRACT

Despite continuous safety management, the death rate of construction workers is not decreasing every year. Accordingly, various studies are in progress to prevent construction site accidents. In this paper, we developed an AI-based priority inspection target selection model that preferentially selects sites are expected to cause construction accidents among construction sites with construction costs of less than 5 billion won (KRW). In particular, Random Forest (90.48 % of accident prediction AUC-ROC) showed the best performance among applied AI algorithms (Classification analysis). The main factors causing construction accidents were construction costs, total number of construction days and the number of construction performance evaluations. In this study an ROI (return of investment) of about 917.7 % can be predicted over 8 years as a result of better efficiency of manual inspections human resource and a preemptive response to construction accidents.

Key words : Construction site inspection, Random forest, AI, Classification analysis, Prediction

초 록

지속적인 안전관리에도 불구하고 매년 건설업 근로자 사망율은 줄어들지 않는 추세다. 이에 따라 건설현장 사고를 예방하기 위한 다양한 연구가 진행 중이다. 본 논문에서는 건설공사 비용 50억원 미만의 건설현장 중 건설사고가 발생할 것으로 예상되는 현장을 우선적으로 선별하는 AI기반 우선점검대상 선정 모델을 개발하였다. 특히, 적용한 AI 알고리즘 중 분류분석에서 가장 뛰어난 성능(사고발생예측 AUC-ROC 90.48 %)을 보인 랜덤 포레스트를 모델 개발에 활용하였으며, 건설사고를 유발하는 주요한 요인으로는 공사비, 총공사일수, 공사실적평가액이 확인되었다. 본 연구를 통해 점검인력 효율화와 건설사고에 대한 선제적 대응의 결과로 8년간 약 917.7 % ROI(투자수익률)를 기대할 수 있다.

검색어 : 건설현장점검, 랜덤 포레스트, 인공지능, 분류, 예측

1. 서 론

최근 10년간 전체 산업재해 사망자는 감소하는 추세이다. 하지만 지속적인 안전관리 대책에도 불구하고 매년 추락, 낙하, 붕괴, 화재 등으로 인한 건설업 근로자 사망자 수는 줄어들지 않고 있다(MOEL, 2021). 특히 건설 붕괴사고는 작업자뿐만 아니라 무고한 국민의 생명에도 위협이 되며 최근 광주 철거현장 사고, 광주 아파트 붕괴사고 등 연이은 건설사고 발생으로 국민의 불안감이 증가하고 있다.

* 국토안전관리원 디지털혁신추진단 빅데이터전략팀, 부산대학교 통계학과 석사과정 (KALIS · Pusan National University · unohwang@kalis.or.kr)
** 종신회원 · 서울대학교 건설환경공학부 교수 (Seoul National University · shchi@snu.ac.kr)
*** 부산대학교 통계학과 학사 (Pusan National University · shine96@kakao.com)
**** 종신회원 · 교신저자 · 국토안전관리원 디지털혁신추진단 빅데이터전략팀 팀장, 공학박사 (**Corresponding Author** · KALIS · hyunjun.jung@kalis.or.kr)

Received May 6, 2022/ **revised** August 18, 2022/ **accepted** October 12, 2022

특히, 중대재해처벌법(시행일 2022. 1. 27)(Korean Law Information Center, 2022)이 시행됨에 따라 사망자가 1명 이상 발생하거나 동일한 사고로 6개월 이상 치료가 필요한 부상자가 2명 이상 발생, 또는 동일한 유해요인으로 급성중독 등 대통령령으로 정하는 직업성 질병자가 1년 이내에 3명 이상 발생했을 시, 중대재해처벌 등에 관한 법률 제6조에 따라 사업주 또는 경영책임자 등이 1년 이상의 징역 또는 10억원 이하의 벌금을 받게 되었다. 따라서, 기업의 입장에서도 건설사고를 선제적으로 예방하는 것이 기업 리스크 관리에 있어서도 과거보다 더욱 중요한 사안이 되었다.

국토안전관리원에서는 제한된 인적·물적 리소스로 인해 전국의 모든 건설현장을 점검할 수 있는 인원이 부족한 실정이다. Table 1과 같이 연 평균 50억 미만의 소규모 민간 건설공사는 52,467건이지만 2021년도 기준 점검 가능한 건설현장은 12,804건으로 전체의 24.4 % 수준이다. 특히, 건축·토목공사의 경우 모든 인원이 현장점검에 투입되더라도 전체 현장 27,790건 중 43 %만이 점검을 수행할 수 있다. 따라서 건설현장의 특성과 건설업체의 특성을 고려한 데이터를 분석하여 사고 고위험도 현장을 선정하고 효율적으로 현장점검 인력을 배치할 필요가 있다. 하지만, 국토안전관리원은 현장 우선점검대상을 선정하기 위해 건설공사대장을 공사금액, 민간공사 여부, 공종 등과 같은 큰 범위의 조건에 따라 분류하고 있어 과학적인 우선점검 대상선정 프로세스를 갖추고 있다고 보기 어렵다.

최근 디지털 사회로의 변화가 가속화되고 있고 인공지능 및 빅데이터 기술이 발전함에 따라 건설현장 위험성 예측을 통한 선제적 안전관리가 가능해졌다. 점검대상 현장 선정에 있어서도 데이터를 기반으로 한 과학적인 의사결정이 필요하다. 국토안전관리원 내부 실무 담당자와의 인터뷰 결과 현재 건설사고 원인에 대한 세밀한 분석이 미흡하며 기존 사례에 대한 자료 이외의 부분을 사전에 분석해주는 서비스가 필요한 것이 확인되었다. 따라서, 빅데이터 및 AI기반의 사고발생 예측모델을 개발하여 우선점검대상 건설현장을 선정할 수 있다면 인력 효율화로 건설현장 점검을 통해 최종적으로 건설사고 발생 건수 감소 및 건설 사망자 수 저감효과를 기대할 수 있다.

선행연구로 Choi et al.(2021)은 건설현장의 공사 사전정보를 활용한 사망재해 예측 모델 개발 연구를 진행하였으며, Yoon et al.(2020)은 의사결정나무와 랜덤 포레스트를 이용한 건설재해 예측모델 개발과 건설업 재해보고서 데이터를 이용한 사고예측모델 개발을 진행하였다. Cho et al.(2017)은 의사결정나무를 기반으로 건설사고 데이터를 분석해 건설사고유형을 분류하였다. 해외에서는 머신러닝을 적용한 건설사고 유형 예측 연구가 진행되었다 (Tixier et al, 2016).

그러나 기존 연구의 경우 주로 건설사고의 유형을 예측하는 연구가 이루어져 건설사고의 발생을 예측하는 데는 어려움이 있었고, 건설사고 자체의 발생여부를 예측하는 기존의 연구의 경우에도 건설사고의 방지를 위한 실무적인 해결책을 제시하지 못해 현장점검에 있어서 실질적인 건설사고의 예방에는 한계가 있다. 그리고 시공능력평가액과 같은 건설업체의 특성을 고려하지 않았기 때문에 건설사고 예측에 있어서 주요인으로 예상되는 변수를 모두 고려하였다고는 볼 수 없다.

따라서, 본 연구는 특정 현장의 위험요소를 제시할 시 해당 공사에서 실제로 사고가 발생할지를 예측하는 모델을 개발하였다. 이후 개발한 모델을 바탕으로 건설사고의 주요인을 예측한 위험요소 프로파일을 구성하고, 현장점검대상의 우선순위를 선정해 한정된 자원 내에서 선제적으로 건설사고를 예방하는데 중점을 두었다. 또한, 데이터에 기반한 과학적인 우선점검 대상 선정을 위해 보다 정량화된 객관적인 지표(공사비, 공사일수, 계절, 시공능력평가액 등)에 근거한 예측모델을 개발하여 우선점검대상 현장선정 업무의 효율성을 향상시키고자 하였다. 수집한 데이터는 국토안전관리원 건설공사 안전관리 종합정보망(CSI) 건설사고 데이터, 건설산업지식정보시스템(KISCON) 건설현장 데이터, 대한건설협회의 2021년도 시공능력평가액 데이터를 포함하였다. 3가지 데이터를 융합해 건설사고에 영향을 줄 것으로 예상되는 현장적 특성과 건설업체특성을 반영한 모델을 개발하였고 AUC-ROC값을 지표로 모델의 성능을 평가하였다. 구축한 모델을 바탕으로 건설사고 위험요소를 도출한 뒤 최종적으로 건설현장 점검대상을 선정하였다. 모델에 사용된 알고리즘은 이중분류모델 중 다변량 데이터셋 분석에 효과적인 트리기반 알고리즘인 랜덤 포레스트(Random Forest), XGBoost (eXtreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine), 의사결정나무(Decision Tree) 등 총 4개 알고리즘과 이중분류모델의 대표적인 알고리즘인 SVM (Support Vector Machine), Logistic Regression, K-NN (K-Nearest Neighbor) 등 총 3개의 알고리즘을 적용하였다.

Table 1. Private Construction Sites between '18 and '20 with a Construction Cost Under 5 Billion KRW (KALIS, 2021a)

	Total Number of on-Site (EA)	0~0.1 billion (EA)	0.1~1 billion (EA)	1~5 billion (EA)
Overall average ('18~'20)	**52,467**	Unreported subject	41,545	10,922
Average of building and infrastructure ('18~'20)	**27,790**		21,364	6,426
Inspection plan ('21)	**12,804**	-	4,377	8,427

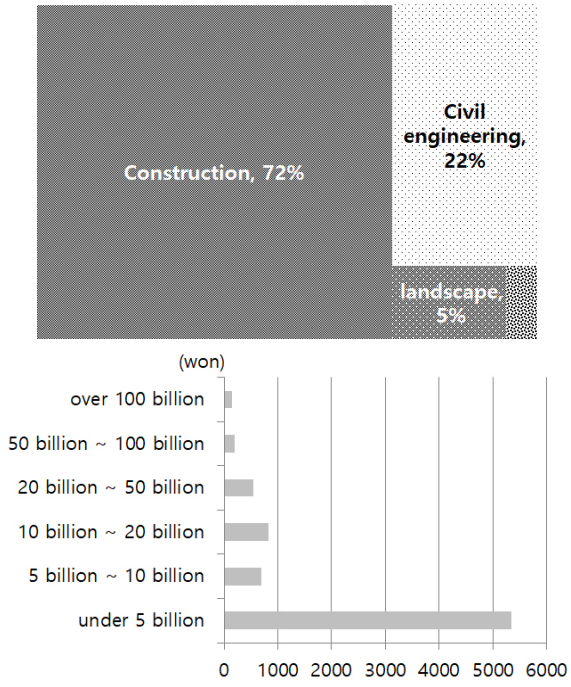

Fig. 1. Ratio of Construction Accidents in Building and Infrastructure Sectors (left) and Accidents in Construction Sites Under 5 Billion KRW (right) (KALIS, 2021b)

본 연구는 건축·토목 분야의 50억 미만의 소규모 건설공사 현장 데이터를 대상으로 분석을 진행하였다. Fig. 1과 같이 전체 시설물 건설사고 중 건축·토목 분야가 전체의 94 %(건축 72 %, 토목 22 %)를 차지하며 이 중에서도 50억 미만의 소규모 건설공사 현장에서 발생하는 사망사고가 50억 이상의 건설공사에 비해 발생 현황이 2.5배로 나타났다. 소규모 건설현장에서 사망 안전사고가 빈번하게 발생하는 이유는 여러 가지가 있겠지만, 그 중에서도 안전교육 미흡, 안전관리 감독 미흡, 부적절한 공사 운영 및 공사계획, 작업자의 과실 등이 주요 원인으로 나타났다(Lim et al., 2021). 본 연구의 범위를 건축·토목 분야의 50억 미만의 소규모 건설공사 현장 데이터로 지정함에 따라 데이터를 기반으로 우선점검 대상현장을 지정하여 인력 효율화를 통하여 건설현장을 점검한다면 건설사고를 저감할 수 있을 것이다.

2. AI 기반 건설현장 점검대상 분류 모델

본 연구는 건설공사대장, 위험요소 프로파일 데이터 분석을 통해 사고 발생과 유의미한 상관관계를 가지고 있는 요소를 발굴하고, 다양한 위험도 지수를 기반으로 현장점검 대상을 선정할 수 있는 모델을 실무에 적용 가능하도록 개발하였다. 이를 위해 데이터 수집(Data Collection), 데이터 전처리(Data Preprocessing), 분류 모델 구축(Classification Model Development), 위험요소 도출(Hazard Factor Identification) 및 우선점검대상 현장선정(Inspection Site Selection) 순으로 4단계 연구를 진행하였다(Fig. 2).

데이터 수집단계에서는 내부에서는 건설공사 안전관리 종합정보망의 사고사례정보를, 외부에서는 건설산업지식정보시스템의 건설공사대장과 대한건설협회의 종합건설사업자 시공능력평가액 정보를 수집하였다. 데이터 전처리 단계에서는 수집된 3가지 데이터를 공사명을 기준으로 매칭 후, 결측치가 많이 발생하는 변수와 이상치를 제거하는 전처리를 실시하였다. 분류 모델 구축단계에서는 건설공사별 사고발생률을 예측하여 건설공사 우선점검대상 현장을 선정할 수 있는 다양한 알고리즘의 예측성능을 비교하여 최적의 모델을 선정하였다. 위험요소 도출 및 현장점검 대상선정 단계에서는 모델을 새로운 현장데이터로 검증한 후 상위 17개의 우선점검대상을 선정하여 대상선정에 필요한 기준을 정립하였다.

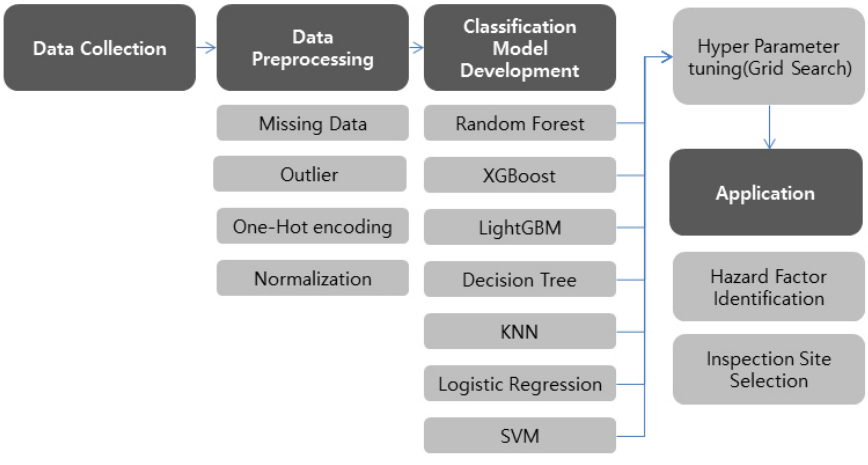

Fig. 2. Research Process for Model Development

2.1 데이터 수집 및 전처리

본 연구에서는 건설사고를 예측하기 위해 건설사고가 발생한 데이터와 모든 건설현장 데이터를 수집하였다. 먼저 건설사고 발생 데이터의 수집을 위해 2020년 7월부터 2021년 6월까지 발생한 1년간의 건설공사 안전관리 종합정보망 건설사고 데이터인 '사고사례정보' 9,247건을 수집하였다. 수집된 데이터는 사고명, 일자, 공사명 등 23개의 변수로 구성되어 있다. 다음으로 모든 건설현장 데이터를 수집하기 위해 '2019년도 건설산업지식정보시스템 건설현장 데이터' 123,841건과 '2020년도 건설산업지식정보시스템 건설현장 데이터'113,051건 총 236,892건을 수집하였다. 수집된 데이터는 공사명, 건설업체명, 일자, 공사비 등 건설공사의 정보를 파악할 수 있는 24개의 변수로 구성되어 있다. 마지막으로 건설업체의 특성을 파악할 수 있는 대한건설협회의 '2021년 종합건설사업자 시공능력평가 현황 데이터'72건을 수집하였다. 수집된 데이터는 건설업체명, 건설업체소재지, 시공능력평가액, 건설공사실적 등 16개의 변수로 구성되어 있다. 수집된 데이터를 먼저 건설공사 안전관리 종합정보망 데이터인 사고사례정보의 '사고명'과 건설산업지식정보시스템 데이터인 건설현장 데이터의 '공사명'을 기준으로 결합하였다. 결합한 데이터 중 사고사례정보를 가지고 있는 데이터는 1(사고발생), 사고사례정보가 없는 데이터는 0(사고미발생)의 값을 가지는 사고여부(0,1) 파생변수를 생성하였다. 이후 대한건설협회의 '2021년도 종합건설사업자 시공능력평가 현황 데이터'를 건설업체명을 기준으로 결합함으로써 모델구축에 활용할 수 있는 최종 데이터 연계를 완료하였다. 본 연구에서 최종적으로 고려한 변수는 23개의 입력변수와 1개의 출력변수로 구성되어 있고 구체적으로 Table 2와 같다.

수집 데이터를 분석 가능한 형태로 변환하기 위하여 우선 결측치 기준(빈값, 총공사일수 또는 공사비가 '0'인 값)을 정한 뒤 결측치가 존재하는 행을 모두 제거하였다. 결측치 처리 시, 변수의 결측치 비율이 60 % 이상인 경우, 변수 자체를 제거하는 것이 데이터 손실을 줄이는 일반적인 방법이지만, 결측치가 60 % 이상인 변수가 절반가량이 되며 애초에 분석에 활용한 최종 데이터셋은 약 23만개의 대용량 데이터이기 때문에 결측치를 가지는 데이터 행을 기준으로 제거하였다. 이상치 또한 모델의 과적합이나 성능 저하를 초래하기 때문에 표준화점수(Z-score) 기준 3 이상이거나 -3 이하인 데이터를 제거하였다. 데이터 결합 시 시스템마다 표준이 달라

Table 2. Features Considered in This Study

Features		type	Feature description
Output	**Accidence**	**int**	**Whether an accident occurred (0,1)**
input	Construction site	chr	Construction site
	Construction company	chr	Construction company
	Date	int	Construction start date
	Month	int	Derived feature of 'Date'
	Day of week	chr	Derived feature of 'Date'
	Season	chr	Derived feature of 'Date'[Spring, Summer, Fall, Winter]
	Construction type	chr	Building, Infrastructure
	Owner	chr	Public, Private
	Construction company location	chr	Construction company location
	Successful bid rate	int	Successful bid rate
	Number of construction days	int	Construction end date − Construction start date + 1
	Construction cost	int	Construction cost
	Construction capability evaluation (building and infrastructure)	int	Construction capability evaluation data
	Construction performance evaluation	float	Construction capability evaluation data
	Management evaluation	float	Construction capability evaluation data
	Technical competency evaluation	float	Construction capability evaluation data
	Credit evaluation	float	Construction capability evaluation data
	Construction capability evaluation (Infrastructure)	float	Construction capability evaluation data
	Construction capability evaluation (Building)	float	Construction capability evaluation data
	Construction work performance (building and infrastructure)	float	Construction capability evaluation data
	Construction work performance (infrastructure)	float	Construction capability evaluation data
	Construction work performance (building)	float	Construction capability evaluation data
	Number of technicians	int	Construction capability evaluation data

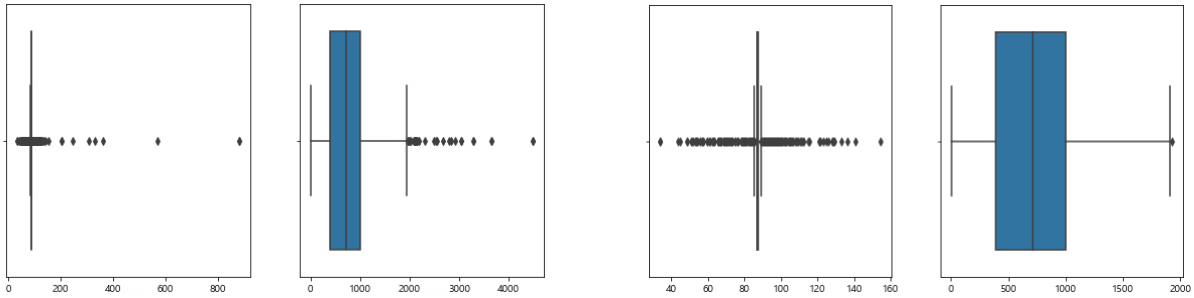

Fig. 3. Distribution of Some Features Before (left) and After (right) Outlier Removal

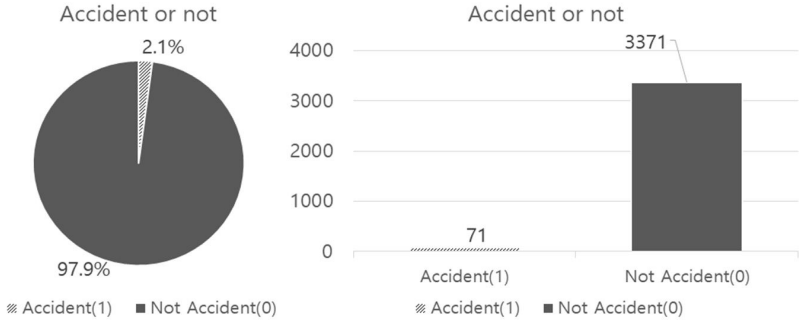

Fig. 4. Ratio of Accident Data

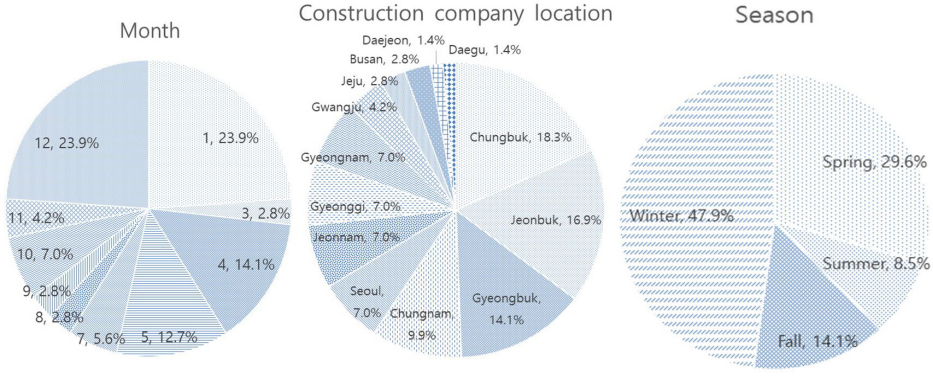

Fig. 5. Categorical Features in Accident Data

키(Primary Key) 매핑 과정에서 결측치 발생으로 다수의 데이터가 제거되었다. 결측치와 이상치 제거로 최종 분석 데이터로는 3,442건을 사용하였다. 이상치 제거 전후의 데이터를 비교한 결과는 Fig. 3과 같다. 이상치 제거결과 높은 수치의 값들이 제거되었고 변수의 분포가 정규분포와 유사한 경향을 확인하였다.

명목변수의 경우, 데이터의 수치적 의미를 나타내는 것이 아니다. 따라서, 명목변수를 모델링에 포함시키기 위해 원-핫 인코딩(One-Hot Encoding)을 통해 '0'과 '1'로 변환하였다. 또한, 데이터 셋의 변수들의 단위가 다르기 때문에 그대로 모델링을 하게 된다면 단위가 큰 변수의 영향을 많이 받을 수 있어 모든 변수의 단위를 통일시키기 위한 min-max 정규화를 실시하였다.

Fig. 4와 같이 최종적으로 구성된 데이터는 총 3,442건이며 종속변수인 사고여부가 1(사고발생 71건, 2.1 %), 0(사고미발생 3,371건, 97.9 %)으로 구성된 불균형데이터이다. Fig. 5와 같이 겨울철, 특히 1월과 12월에 건설사고가 많이 발생하며 지역별로는 충북에서 사고가 많이 발생하는 경향을 보인다.

Fig. 6과 같이 사고여부에 따라 50억 미만 소규모 현장 내에서 '공사비'가 많은 경우 사고가 많이 발생하는 경향을 보인다. 또한,

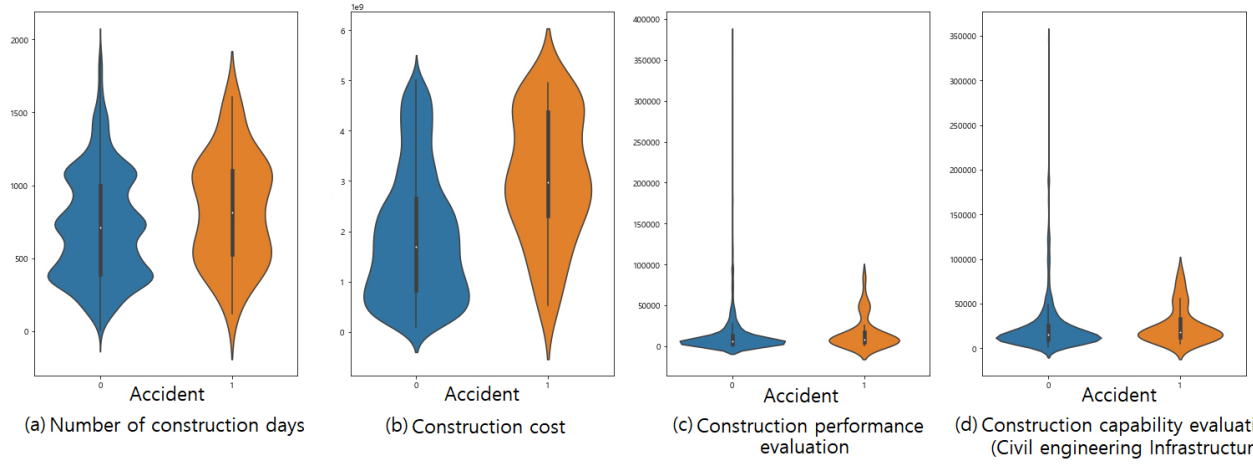

Fig. 6. Distribution of Major Features according to Accident (1) or not (0)

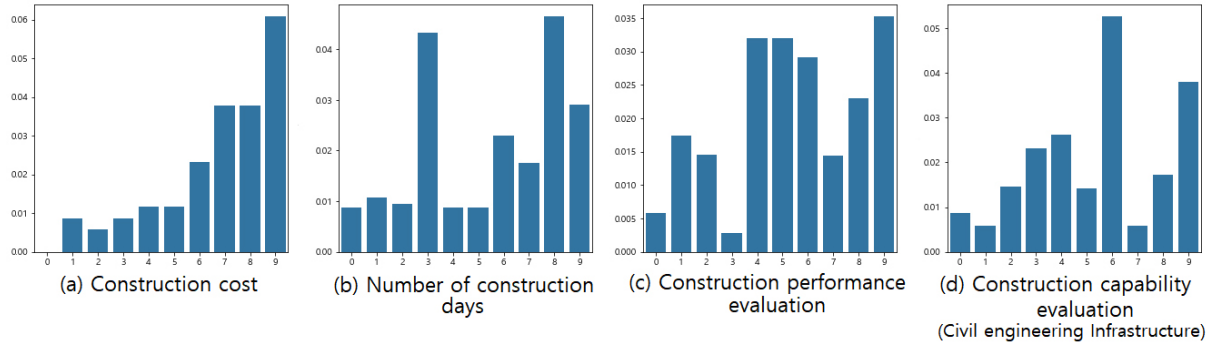

Fig. 7. Accident Ratio of Major Features by Percentile Interval

'총공사일수', '공사실적평가액', '시공능력평가액(토목)'의 값이 클수록 사고발생이 늘었다.

Fig. 7은 주요 변수들의 백분위 구간별 사고비율을 그래프로 나타낸 것이다. '공사비'가 커질수록 사고율이 증가하는 것을 명확히 확인할 수 있고 '총공사일수', '공사실적평가액', '시공능력평가액(토목)'변수 또한 값이 커질수록 사고율이 증가하는 경향을 보인다.

2.2 AI 기반 건설현장 점검대상 분류 모델

점검대상 선정모델 개발을 위해 이중분류모델 중 다변량 데이터셋 분석에 효과적인 트리기반 알고리즘인 랜덤 포레스트, XGBoost, LightGBM, 의사결정나무 등 총 4개 알고리즘과 이중분류모델의 대표적인 알고리즘인 SVM, Logistic Regression, K-NN 등 총 3개의 알고리즘을 적용하였다. 이 중 모델 성능 평가를 통해 랜덤 포레스트, XGBoost, LightGBM을 성능이 우수한 모델로 선정 후 모델 최적화 결과를 비교하였다.

랜덤 포레스트는 트리 구조의 분류기법으로 사용되는 앙상블(Ensemble) 기법 중 하나로 예측 정확도를 높이기 위해 다수의 의사결정나무를 사용하여 각 트리가 분류한 결과를 가지고 투표해 가장 많이 득표한 결과를 최종 분류 결과로 선택하는 방식이며(Fig. 8), 대수의 법칙에 의해 과적합을 방지할 수 있다는 특징이 있다

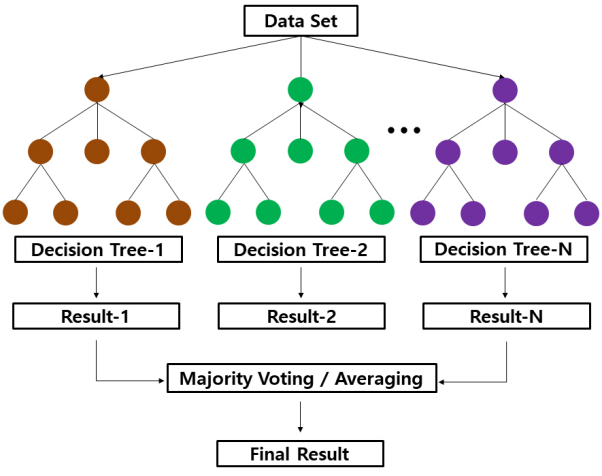

Fig. 8. Random Forest Algorithm

(Breiman, 2001). 또한, 고차원데이터와 다중공선성을 효과적으로 처리할 수 있다(Belgiu and Drăguţ, 2016). Adaboost와 비교를 해보면 거의 유사한 정확도를 얻을 수 있지만 학습시간이 더 빠르고 더 안정적이다(Chan and Paelinckx, 2008). XGBoost도 랜덤 포레스트와 유사하게 다수의 의사결정나무를 조합하여 사용하는 앙상블 기법 중 하나인 gradient boosting의 단점을 보완한 알고리즘으로 2015년 캐글 대회에서 우수한 성적을 거둔 대부분의 팀이 사용하여 널리 알려지게 되었으며, 소량의 데이터를 잘 표현하는데 장점이 있다고 알려진 부스팅 기법 중 하나이다(Chen and Guestrin, 2016). LightGBM은 트리 기반의 앙상블 기법 중 하나로 고차원 변수의 데이터가 많은 경우 효율성과 확장성이 떨어지는 XGBoost의 단점을 보완하기 위해 개발된 알고리즘으로 기존 XGBoost에 비해 계산 속도와 메모리 소비면에서 20배의 성능을 보여준다(Ke et al., 2017).

본 연구에서는 각 알고리즘의 최적의 하이퍼파라미터를 찾기 위해 그리드 서치(Grid-search) 방식을 사용하였다. 하이퍼파라미터를 찾는 기법으로는 랜덤 서치(Randomize-search) 방식과 그리드 서치(Grid-search) 방식이 존재한다. 다수의 랜덤조합을 시험하는 랜덤 서치 방식은 속도는 빠르지만 성능은 떨어진다(Feurer and Hutter, 2019). 따라서, 본 연구에서는 모든 조합을 적용하여보고 최적의 결과를 도출하는 조합을 선정하는 그리드 서치 방식을 사용하여 하이퍼파라미터를 최적화하였다.

Table 3과 같이 혼동행렬(Confusion Matrix)로 예측모델의 성능을 평가하여 재현율(Recall, Eq. (1)), AUC-ROC, F1-score (Eq. (2))를 검증기준으로 가장 우수한 성능을 보이는 모델을 선정하였다. 이때 재현율은 실제 사고발생 데이터를 사고가 발생한 것으로 올바르게 예측한 확률을 의미한다. AUC-ROC는 다양한 임계값에서 분류 모델의 성능을 측정할 수 있는 지표이며, ROC는 확률곡선을 의미하고 AUC는 분류성의 척도를 나타낸다. 이는 모델이 분류를 얼마나 잘하는지 설명하며 AUC값이 높을수록 모델의 성능이 훌륭하다는 것을 의미한다(Narkhede, 2018). 본 연구에서 사용한 데이터는 사고미발생 데이터에 비해 사고발생 데이터의 양이 적은 불균형 데이터이다. 따라서, 일반적으로 분류 알고리즘 성능 판단에 사용되는 정확도(정분류율)를 사용하게 된다면 알고리즘이 사고발생 데이터가 아닌 사고미발생 데이터만 올바르게 분류하게 되더라도 높은 성능을 나타낼 수 있다. 그렇기 때문에 이를 보완할 수 있도록 실제 사고미발생 데이터를 올바르게 예측하였는지 확인할 수 있는 지표인 재현율을 사용했고, 재현율과 정밀도(Precision, Eq. (3))의 조화평균값인 F1-score를 사용해 사고발생 데이터를 정확하게 예측할 수 있는 성능 확인을 하였다. 마지막으로 중요한 변수를 선정하기 위해 가장 우수한 성능을 보이는 알고리즘에 대한 중요도 분석을 수행하였고, 우선점검 대상현장 10개소의 95 % 신뢰구간을 기준으로 대상 구간을 선정하였다.

Table 3. Confusion Matrics

	Predict Accident (1)	Predict No Accident (0)	
Real Accident (1)	True Positive (TP)	False Negative (FN)	Recall (=Sensitivity)
Real No Accident (0)	False Positive (FP)	True Negative (TN)	Specificity
	Precision		

$$Recall = \frac{TP(True\ Positive)}{TP(True\ Positive) + FN(False\ Nagative)} \quad (1)$$

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

$$\mathrm{Precision} = \frac{TP(True\ Positive)}{TP(True\ Positive) + FP(False\ Positive)} \quad (3)$$

3. AI 기반 건설현장 우선점검대상 선정 모델

3.1 건설사고 예측 데이터 구축

데이터는 훈련용 데이터(training set)와 테스트 데이터(test set) 2가지 세트로 구성되며, 일반적으로 7:3 혹은 8:2 비율로 구축한다. 본 연구에서는 구축한 모델을 검증하기 위해 7:3의 비율로 전체 데이터 3,442개를 훈련용 데이터 2,409개와 테스트 데이터 1,033개로 분리하였다. 훈련용 데이터로 모델을 훈련시키고 훈련시킨 모델을 테스트 데이터에 적용해 모델을 검증하였다. 일반적으로 훈련용 데이터 결과와 검정 데이터 결과에 대한 과적합 상태를 별도 검토하는데 본 논문에서 적용한 기법인 K-분할 교차검증의 특징인 훈련용 데이터의 과적합 상태를 검증하므로 이에 대한 별도 검토가 필요하지 않고 수행하였다. 사용된 데이터의 종속변수 3,442건 중 사고발생 데이터는 71건으로 전체 데이터 중 2.1 %에 불과한 불균형 데이터이다. 따라서 훈련용 데이터와 테스트 데이터의 사고여부 비율을 전체 데이터와 동일한 비율로 분리하였다.

3.2 선정 모델의 성능 분석

K-분할 교차검증(k=5)을 통해 데이터를 분할하여 모델을 학습하고 과적합을 방지하였다. 도출되는 확률값 중 AUC를 최대로 하는 모델을 선정하였다. 총 7개의 알고리즘(LightGBM, XGBoost, 랜덤 포레스트, 의사결정나무, SVM, Logistic Regression, K-NN) 중 AUC를 기준으로 가장 높은 분류성능을 보이는 모델인 LightGBM, XGBoost, 랜덤 포레스트를 선택(Table 4)하였다.

예측 모델의 성능 분석 결과에 따라 선정된 LightGBM 알고리즘과 XGBoost 알고리즘 그리고 랜덤 포레스트 알고리즘을 하이퍼파

라미터 수정을 통해 성능을 증대시켰다. LightGBM에서 하이퍼파라미터 최적화 결과 알고리즘의 성능은 88.68 %로 기존 성능 85.09 % 대비 3.59 %p 증가하였다. XGBoost에서는 하이퍼파라미터 최적화 결과 알고리즘의 성능은 85.74 %로 기존 성능 83.79 % 대비 1.95 %p 증가하였으며, 랜덤 포레스트에서는 하이퍼파라미터 최적화 결과 알고리즘의 성능은 88.32 %로 기존 성능 83.19 % 대비 5.13 %p 증가하였다.

3.3 서비스 모델의 성능 평가

Table 5과 같이 테스트 데이터를 활용하여 세 가지 알고리즘의 분류성능을 평가하였다. 재현율과 AUC-ROC, 그리고 F1-score 값을 비교한 결과, 랜덤 포레스트가 각각 80.95 %, 90.48 %, 89.47 %로 모든 지표에서 다른 알고리즘에 비해 우수한 분류성능을 보였다. 알고리즘별 혼동행렬(Confusion matrix)은 Tables 6~8과 같다. 따라서, 본 연구에서는 랜덤 포레스트 알고리즘을 최종 모델로 선정하였다.

Table 4. Comparison of Algorithm Performance (AUC-ROC)

Algorithm	AUC-ROC	Posterior AUC-ROC (Optimization)
LightGBM	0.8509	0.8868
XGBoost	0.8379	0.8574
Random Forest	0.8319	0.8832

Table 5. Performances of Algorithm (%)

Algorithm	Recall	AUC-ROC	F1-score
LightGBM	47.6190	74.8095	64.5161
XGBoost	76.1905	88.0952	86.4865
Random Forest	**80.9524***	**90.4762***	**89.4737***

*The highest predictive performance of each Algorithm

Table 6. Confusion Matrix of LightGBM

	Predict (0)	Predict (1)
Real (0)	1,012	0
Real (1)	11	10

Table 7. Confusion Matrix of XGBoost

	Predict (0)	Predict (1)
Real (0)	1,012	0
Real (1)	5	16

Table 8. Confusion Matrix of Random Forest

	Predict (0)	Predict (1)
Real (0)	1,012	0
Real (1)	4	17

Table 9. Feature Importances of Random Forest and LightGBM

Features	Rank of Importances	
	Random Forest	LightGBM
Construction cost	**1**	**1**
Number of construction days	**2**	**3**
Construction performance evaluation	**3**	**5**
Construction work performance (building and infrastructure)	**4**	10
Technical competency evaluation	**5**	7
Credit evaluation	6	**4**
Management evaluation	7	–
Construction capability evaluation (building and infrastructure)	8	–
Construction work performance (building)	9	7
Construction capability evaluation (Infrastructure)	10	**2**
Construction capability evaluation (Building)	–	6
Construction work performance (infrastructure)		9

3.4 건설사고 주요인 도출

선정된 최종 모델인 랜덤 포레스트를 활용하여 건설사고를 유발하는 중요한 요인을 도출하기 위해 알고리즘 내 'feature importance'를 사용하였다. Table 9과 같이 중요도(importance)가 큰 것을 기준으로 상위 10개의 주요인을 도출하였다. 그 결과 건설사고에 가장 큰 영향을 주는 변수는 공사비였으며, 뒤이어 총공사일수, 공사실적평가액, 건설공사실적 등이 중요한 요인으로 도출되었다.

랜덤 포레스트보다는 낮은 성능을 가진 LightGBM과 랜덤 포레스트의 건설사고 주요인을 비교해보면 마찬가지로 가장 중요한 요소는 '공사비'였다. 또한, 상위 10개의 주요인 중 두 알고리즘에 공통적으로 중요하다고 나온 변수는 '총공사일수', '공사실적평가액', '건설공사실적(토건)', '기술능력평가액', '신인도평가액', '건설공사실적(건축)', '시공능력평가액(토목)'이었다. '총공사일수' 변수의 경우 두 알고리즘에서 각각 2순위, 3순위로 높은 중요도를 보였다. '시공능력평가액(토목)'의 경우 LightGBM 알고리즘에서는 2순위로 높은 중요도를 보였지만, 랜덤 포레스트 알고리즘에서는 10순위로 낮은 중요도를 보였다. 두 알고리즘에 공통적으로 상위 5순위 내의 중요도를 보인 변수는 '공사비','총공사일수','공사실적평가액'이었다. 두 알고리즘 모두 상위 3개 변수를 제외한 나머지 변수들은 중요도의 차이가 크지 않았다.

3.5 우선점검 대상현장 선정

테스트 데이터셋 1,033건에 적용했을 때 가장 우수한 분류 성능을 보인 랜덤 포레스트 알고리즘으로 사고확률이 높은 건설현장

Table 10. Construction Site Accident Prediction Result

Construction site	Construction cost (million)	Number of construction days (day)	Construction performance evaluation (million)	Real	Prediction
1	2,978	527	2,675	accident	accident
2	4,497	540	7,200	accident	accident
3	2,917	799	2,675	accident	accident
4	2,741	1057	50,103	accident	accident
5	4,598	547	17,026	accident	accident
6	4,148	1079	10,216	accident	accident
7	1,507	352	2,850	accident	accident
8	2,084	1154	9,445	accident	accident
9	1,335	772	4,942	accident	accident
10	3,165	547	17,538	accident	accident
11	4,372	1444	12,166	accident	accident
12	2,431	1111	4,993	accident	accident
13	756	730	52,613	accident	accident
14	1,453	1096	4,732	accident	accident
15	2,612	481	84,224	accident	accident
16	2,491	527	13,622	accident	accident
17	3,329	1201	5,146	accident	accident
18	2,268	891	43,131	accident	Non-accident
19	3,241	607	6,142	accident	Non-accident
20	4,929	539	4,649	accident	Non-accident
21	1,730	1173	7,888	accident	Non-accident

21개를 Table 10과 같이 정렬하였을 때, 실제 사고가 발생한 21개 건설현장 중 17개 현장을 예측하여 본 모델을 활용한 사고발생 예측율이 81 %로 확인되었다. 향후에 데이터 변수 추가 및 보완에 따른 모델 업데이팅이 이루어지고 최적화하여 90 %이상의 사고발생 예측을 하게 되면 개발된 모델을 통해 데이터에 기반하여 현장 점검 인력을 효율적으로 투입하여 건설사고를 저감할 수 있을 것이다. 본 개발 모델에서 실제로 사고가 발생한 17개 건설현장의 주요인을 활용한다면 실무에서 우선적으로 현장점검 대상 선정기준으로 즉시 활용 가능하여 의사결정자가 선정하는데 기초자료로 사용될 수 있을 것으로 사료된다.

4. 결론

본 논문에서는 건설현장의 현장 데이터와 건설업체 데이터를 활용하여 AI 기반 50억원 미만의 소규모 건설공사 현장의 우선점검 대상현장 선정 모델을 개발하였다. 이를 위해 내외부의 약 23만건의 데이터를 수집하여 활용하였다. 하지만 데이터를 결합하는 과정에서 결측치 발생으로 많은 데이터 손실이 발생하였다. 그 결과 총 3,442건의 데이터를 활용하여 가장 우수한 분류 성능을 보인 랜덤 포레스트 알고리즘을 적용하여 모델을 개발하였다. 테스트 데이터에 알고리즘을 적용한 결과 AUC가 90.48 %로 높은 예측 성능을 보였다.

본 논문에서 구축한 모델에 건설현장 데이터를 입력할 시, 사고예측확률을 도출할 수 있으며, 모델을 사용하지 않더라도 제시된 대상선정기준에 따라 사고예상현장을 선제적으로 관리해 최종적으로 건설사고 사고 발생률을 줄일 수 있을 것이다. 점검인력 효율화 및 건설사고에 대한 선제적 대응에 따른 피해저감 효과로 8년간 약 917.7 %의 ROI를 기대할 수 있다. 먼저, 모델 적용에 따른 점검 인력의 효율화로 인해 국토안전관리원 현업인원, 연간 업무시간, 업무절감 효과 그리고 시간당 인건비를 고려해 보았을 때 8년간 약 191.7 %를, 선제적 대응에 따른 피해저감 효과로는 연간 사망자 발생 건설사고 건수, 사망사고 1건당 손실대가 그리고 사고예방 효과를 산정인자로 두고 계산한 결과 8년간 약 726 %의 ROI를 기대할 수 있다.

추후 데이터를 추가로 수집해 모델을 고도화할 수 있을 것이다. 국토안전관리원의 기타 내부 데이터, 국토교통부, 기상청 그리고

대한전문건설협회 등의 데이터를 추가 확보하여 개발한 모델을 고도화할 수 있을 것이다. 단기적으로는 축적되는 건설현장 데이터와 사고발생 데이터로 모델을 보완하여 실무 매뉴얼으로 활용할 수 있으며, 장기적으로는 원클릭시스템 구축을 통한 대상현장 점검 서비스를 제공할 수 있을 것이다. 또한, 국토안전관리원 내 5개 지사(수도권지사, 강원지사, 중부지사, 호남지사, 영남지사)의 권역별 건설현장 목록과 건설현장 관련 자료를 수집해 권역별 예측모델을 구축하여 지역적 특성을 반영한 우선점검 대상현장 선정을 통해 예측모델을 더욱 정교화 할 수 있을 것이다.

References

Belgiu, M. and Drăguţ, L. (2016). “Random forest in remote sensing: A review of applications and future directions.” *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 114, pp. 24-31.

Breiman, L. (2001). “Random forest.” *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

Chan, J. C. W. and Paelinckx, D. (2008). “Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery.” *Remote Sensing of Environment*, Vol. 112, No. 6, pp. 2999-3011.

Chen, T. 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*, California, USA, pp. 785-794.

Cho, Y. R., Kim, Y. C. and Shin, Y. S. (2017). “Prediction model of construction safety accidents using decision tree technique.” *Journal of the Korea Institute of Building Construction*, Vol. 17, No. 3, pp. 295-303 (in Korean).

Choi, S. J., Kim, J. H. and Jung, K. H. (2021). “Development of prediction models for fatal accidents using proactive information in construction sites.” *Journal of the Korean Society of Safety*, Vol. 36, No. 3, pp. 31-39 (in Korean).

Feurer, M. and Hutter, F. (2019). “Hyperparameter optimization.” *Automated machine learning*, Springer, Cham, pp. 3-33.

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.” *In Advances in Neural Information Processing Systems*, pp. 3149-3157.

Korea Authority of Land & Infrastructure Safety (KALIS) (2021a). *Submission of documents for participation in public big data analysis contest by MOIS in 2021*, pp. 8 (in Korean).

Korea Authority of Land & Infrastructure Safety (KALIS) (2021b). *Submission of documents for participation in public big data analysis contest by MOIS in 2021*, pp. 6 (in Korean).

Korean Law Information Center (2022). *Act on punishment of serious accidents, etc., Article 6,* Available at: https://www.law.go.kr (September 23, 2022).

Lim, J. R., Park, C. Y. and Yun, S. M. (2021). “A study on safety management measures for small and medium-sized construction sites-Focused on reinforced concrete construction.” *Journal of the Architectural Institute of Korea,* Vol. 23, No. 6, pp. 197-204 (in Korean).

Ministry of Employment and Labor (MOEL) (2021). *2020 Industrial accident and death statistics announced*, pp. 13 (in Korean).

Narkhede, S. (2018). “Understanding auc-roc curve.” *Towards Data Science*, Vol. 26, No. 1, pp. 220-227.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B. and Bowman, D. (2016). “Application of machine learning to construction injury prediction.” *Article of Automation in Construction*, Vol. 69, pp. 102-114.

Yoon, Y. G., Lee, J. Y. and Oh, T. G. (2020). “Development of accident prediction model with construction accident report data.” *Korea Institute for Structural Maintenance and Inspection*, Vol. 24, No. 2, pp. 6 (in Korean).