ARTICLE

머신러닝을 이용한 신호교차로 접근부 추돌사고 심각도 요인

양정훈[1] · 박정순[2] · 임희섭[3] · 김규혁[4] · 송태진[5*]
[1]도로교통공단 교통사고종합분석센터 차장, [2]도로교통공단 경북지부 안전조사운영부 부장,
[3]청주시정연구원 연구위원, [4]충북대학교 도시공학과 박사과정, [5]충북대학교 도시공학과 부교수

Identifying Factors Affecting the Severity of Rear-End Crashes at Signalized Intersection Approaches Using Machine Learning Technologies

YANG, Jeonghun[1] · PARK, Jeongsoon[2] · RIM, Heesub[3] · KIM, Kyuhyuk[4] · SONG, Tai-jin[5*]

[1]Deputy General Manager, Traffic Accident Analysis Center, Road Traffic Authority, Wonju 26466, Korea
[2]Department Head, Department of Safety Investigation and Operation, Road Traffic Authority Gyeongsangbuk-do branch, Gumi 39440, Korea
[3]Senior Research Fellow, Cheongju Research Institute, Cheongju 28394
[4]Ph.D Candidate, Department of Urban Engineering, Chungbuk National University, Cheongju 28644, Korea
[5]Associate Professor, Department of Urban Engineering, Chungbuk National University, Cheongju 28644, Korea

*Corresponding author: tj@chungbuk.ac.kr

J. Korean Soc. Transp.
Vol.42, No.2, pp.212-230, April 2024
https://doi.org/10.7470/jkst.2024.42.2.212

pISSN : 1229-1366
eISSN : 2234-4217

ARTICLE HISTORY
Received: 12 January 2024
Revised: 16 February 2024
Accepted: 14 March 2024

Abstract

Considering the high cost of road traffic accidents in Korea and the characteristics of accidents at signalized intersections, a study on the severity of accidents at intersections is necessary. Especially, if an analysis of accident factors at the intersection approach is conducted, it can prevent intersection accidents in advance. This study analyzed the severity factors of rear-end crashes occurring at the approach of signalized intersections. For this purpose, data on 171 rear-end crashes that occurred over the past three years (2020-2022) at the approach of 57 signalized intersections in Cheongju City, as well as traffic volume data, signal operation data, and road geometric structure data were collected. And two methods were applied (with/without the use of machine learning classification models). As a result of the study, the method using machine learning techniques showed higher performance with an Accuracy of 0.8338 and an F1 score of 0.9058 compared to the method without use. And a total of 8 factors (road surface, whether the offending driver was drunk driving, the vehicle type of the offending driver, the age group of the victim driver, vertical alignment of road, the number of straight lanes, U-turn area, skid-proof facility) were statistically significant ($p<0.05$). The results of this study can be utilized in establishing traffic safety measures to prevent severe rear-end crashes in advance. This is expected to contribute to the reduction of social costs due to traffic accidents. In the future, it is deemed necessary to conduct studies on the severity of accidents on collector roads or un-signalized intersections. Additionally, there is a need for comparing the characteristics between intersections with high and low rear-end collisions on the same hierarchical road, and analyzing driving behavior data.

Keywords: crash severity, F1 score, machine learning, rear-end crash, signalized intersection approach

초록

국내 도로교통 사고의 높은 인적피해비용과 신호교차로의 사고 특성을 감안시, 신호교차로에 대한 사고심각도 연구가 필요하다. 특히, 교차로 접근부에 대한 사고 요인 분석이 이루어진다면 교차로 사고를 미연에 방지할 수 있다. 본 연구는 신호교차로 접근부에서 발생하는 추돌사고의 심각도 요인을 분석하였다. 연구를 위해 충청북도 청주시의 57개 신호교차로 접근부에서 최근 3년간(2020-2022년) 발생되었던 171건의 추돌사고 자료와 교통량 자료, 신호운영 자료, 도로 기하구조 자료 등이 수집되었으며 두 가지 방법(머신러닝 분류모형의 사용/미사용)을 적용하였다. 연구 결과, 머신러닝 기법을 사용한 방법의 정확도(Accuracy)가 0.8338, F1 score가 0.9058로 미사용 방법 대비 높은 성능을 나타냈다. 그리고 총 8개(노면상태, 가해 운전자의 음주운전 여부, 가해 운전자의 차종, 피해 운전자의 연령대, 종단선형, 직진 차로수, 유턴구역, 미끄럼방지포장) 요인이 통계적으로 유의($p<0.05$)한 것으로 나타났다. 본 연구 결과는 심각 추돌사고를 미연에 방지하기 위한 교통안전대책 수립에 활용될 수 있을 것이며 이를 통해 교통사고로 인한 사회적 비용 감소에 기여할 수 있을 것으로 기대된다. 향후에는 집산도로나 비신호교차로 등에 대한 사고 심각도 연구, 동일 위계 도로에서의 추돌사고가 많은 교차로와 적은 교차로 간 특성 비교, 운전행태 데이터 분석 등이 필요할 것으로 판단된다.

주요어: 추돌사고, F1 score, 머신러닝, 사고심각도, 신호교차로 접근부

Introduction

Korea Road Traffic Authority(2023)에서 추계한 2022년 국내 도로교통 사고비용은 약 26조 2,833억원이며, 이는 2022년도 GDP의 약 1.2%, 국가 예산의 약 4.3% 수준에 해당한다. 그동안 정부 및 관계기관의 적극적인 교통안전대책으로 최근 교통사고 사망자수는 큰 폭으로 감소했지만, 도로교통 사고비용은 여전히 보합 또는 소폭 증가추세를 유지하고 있어 사고비용 감소를 위한 노력이 필요하다.

이러한 도로교통 사고비용 감소를 위해서는 전체 사고비용 중 가장 높은 비중을 차지하고 있는 인적 피해비용(약 12조 6,040억원, 48.0%)의 절감이 필수적이며, 이 중 중상자 비용은 약 4조 5,716억원으로 가장 높은 비중을 차지하고 있다. 또한, 사상자 피해종별 평균비용에서는 중상자 1명당 평균비용이 6,890만원으로 경상자 1명당 평균 비용(520만원) 대비 무려 13배 이상에 달하기 때문에 중상사고에 해당하는 심각사고에 신경을 써야만 사고비용 감소가 가능하다.

한편, 최근 5년(2018-2022년) 국내에서 발생된 총 997,238건[1]의 교통사고를 도로 형태별로 살펴보면 교차로 52.1%(519,524건), 단일로 47.9%(477,714건)으로 나타났으며, 전체 부상자수 중 중상자수 비율 또한 교차로 20.7%(159,332명/770,915명), 단일로 20.2%(140,444명/694,456명)로 나타났다. 이처럼 교차로는 사고건수 뿐만 아니라 심각도 또한 고속국도, 일반국도 등이 다수 포함된 단일로와 비교시 결코 낮지 않음을 알 수 있다. 따라서, 교차로를 대상으로 하는 사고심각도 연구 또한 충분히 의미가 있다. 최근 국외에서는 신호교차로를 대상으로 사고심각도 연구가 활발히 수행되고 있으나(Kidando et al., 2022; Sharafeldin et al., 2022a; Yuan et al., 2022 etc), 국내는 상대적으로 저조한 편이다.

이러한 배경하에 본 연구에서는 신호교차로에서 발생하는 추돌사고를 대상으로 사고심각도를 분석하였다. 국내 전체 사고 충돌유형 중 측면충돌 다음으로 높은 추돌사고는 다른 충돌유형과 달리 사고 발생시 당사자 과실 구분이 뚜렷한 특징[2]을 가지고 있어 사고원인 규명이 용이하다. 또한, 교차로에서의 추돌사고는 특히 교차로 부근에서 높은 비중을 나타낸다. Table 1은 최근 5년(2018-2022년) 교차로에서 발생된 차대차 사고 419,561건에 대한 충돌유형별 통계인데 교차로 부근에서의 추돌사고 비중이 34.2%나 됨을 알 수 있다.

1) 철길건널목 및 기타(불명) 사고 제외
2) 통상 추돌차량인 후행차량이 가해차량으로 분류됨

Table 1. Statistics on the occurrence of intersection traffic accidents by location and collision types (2018-2022)

Classification	Number of crashes					
	Sum	Head-on collision	Sideswipe collision	Rear-end collision	Backing collision	Etc.
Total	419,561 (100.0%)	20,640 (4.9%)	221,718 (52.8%)	63,402 (15.1%)	4,923 (1.2%)	108,878 (26.0%)
Inside the intersection	285,828 (100.0%)	16,283 (5.7%)	173,185 (60.6%)	17,654 (6.2%)	2,162 (0.7%)	76,544 (26.8%)
Near the intersection	133,733 (100.0%)	4,357 (3.2%)	48,533 (36.3%)	**45,748 (34.2%)**	2,761 (2.1%)	32,334 (24.2%)

Table 1을 감안할 때, 교차로에 대한 추돌사고 분석은 추돌사고가 빈번히 발생하는 교차로 부근을 대상으로 수행될 필요가 있다. 특히, 교차로 접근부에서 발생하는 추돌사고 요인이 분석된다면 교차로내 사고보다 더욱 효과적인 예방대책 수립도 가능하다. 그럼에도 불구하고, 그동안 신호교차로를 대상으로 수행된 국내외 대다수 연구들은 교차로 범위까지는 세분화하지 않고 진행되었다. 이렇게 불분명한 교차로 범위의 연구 결과는 정책결정권자 및 실무자들로 하여금 실질적 사고예방대책 수립에 큰 도움이 되기 힘들다.

그래도, 국외에서는 이미 교차로에서의 추돌사고 중요성을 감안하여 최근 심각도 분석 연구가 활발히 진행되고 있으며(Champahom et al., 2020; Zhang et al., 2022; Yuan et al., 2023 etc), 추돌사고 발생요인 분석 연구 또한 다수 수행되어졌다(Wang et al., 2003; Yan et al., 2005; Yan and Radwan, 2006 etc). 이에 반해, 국내는 발생요인 분석에 대한 연구만 소수 존재할 뿐이다(Park and Park, 2007; Park and In, 2009).

이에 따라, 본 연구에서는 충청북도 청주시의 주요 신호교차로 접근부에서 발생된 추돌사고 자료를 수집하고, 신호교차로에 대한 중상사고의 다양한 요인(발생환경적 요인, 인적 요인, 차량적 요인, 도로교통적 요인)이 될 만한 자료들을 수집하였다. 특히, 도로교통적 요인으로 도로 및 교통 조건, 신호운영 조건 뿐만 아니라 현장에 설치된 주요 안전시설까지를 포함하였다. 수집된 자료로 추돌사고 심각도의 영향 요인을 찾고자 하였으며 이를 통해 추돌사고의 사전예방과 교통사고로 인한 사회적 비용 감소를 도모하고자 하였다.

Literature Review

1. 신호교차로 사고심각도 연구 동향

최근에 수행된 신호교차로 대상의 사고심각도 주요 연구들이다. Mitra and Bhowmick(2020)은 인도 Kolkata 시에서 2011-2014년 발생된 8,324건의 전체사고 중 신호교차로 52개소에서 발생된 사고자료를 추출하여 사고발생과 사고심각도의 영향 요인을 분석하였다. 분석자료로 사고자료와 도로기하구조 및 시설현황이 수집되었으며 분석모형으로 상관분석과 t검정이 사용되었다. 연구 결과, 전체 교통량, 통행규제, 직진-회전차량 구성비, 보호 우회전, 교차로 규모, 트램 정류장 등이 사고발생 영향 요인으로 나타났으며 전적색 신호시간, 보호 우회전, 통행규제, 비동력 교통수단, 도로 표지 가시성이 사고심각도 영향 요인으로 나타났다.

Khattak et al.(2021)은 벨기에 Antwerp 시의 교차로 760개소(신호교차로 198개소, 비신호교차로 562개소)에서 2010-2015년 발생된 5,128건의 사고자료로 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료와 교통량, 도로 기하구조 자료가 수집되었으며 분석모형으로 음이항 모형과 포아송 모형이 사용되었다. 연구 결과, 신호교차로에서는 교통량, 소규모 접근로의 횡단보도 유무, 교차로 접속각이 사고심각도 영향 요인으로 나타났으며, 비신호교차로에서는 부도로 접근로의 횡단보도 유무, 부도로 접근로의 직진 차로수, 교차로 형태가 사고심각도 영향 요인으로 나타났다.

Kidando et al.(2022)은 미국 Florida 주의 Tallahassee 시에 위치한 22개의 신호교차로에서 2017-2019년 발생된 304건의 사고자료로 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료와 교통시설, 신호운영 자료가 수

집되었으며 분석모형으로 전통적 선형회귀 모형과 로지스틱 모형이 사용되었다. 연구 결과, 교차로 접근부의 지체, 차량군 비율, 충돌형태, 차내 탑승객 위치, 성별과 나이, 조명시설 유무 등이 사고심각도 영향 요인으로 나타났다.

Sharafeldin et al.(2022a)은 미국 Wyoming 주의 교차로 359개소에서 9년간(2007-2017년, 2010-2011년 제외) 발생된 9,108건의 사고자료로 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료, 도로경사, 노면상태 자료 등이 수집되었으며 분석모형으로 순서형 프로빗 모형이 사용되었다. 연구 결과, 노면 마찰력, 교차로 위치, 도로 기능 분류, 가드레일, 우측 길어깨폭, 조명시설이 사고심각도 영향 요인으로 나타났다.

Yuan et al.(2022)은 미국 전역에 위치한 주요 신호교차로에서 2012-2015년 발생된 1,843건의 사고자료로 교차로내 직각충돌 사고유형에 대한 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료, 도로기하구조, 교통시설 자료가 수집되었으며 분석모형으로 머신러닝 분류모형인 의사결정나무 모형이 사용되었다. 연구 결과, 차량 유효충돌 속도, 중앙분리시설, 기상 악천후가 사고심각도 영향 요인으로 나타났다.

2. 신호교차로 추돌사고 연구 동향

신호교차로 대상의 추돌사고 연구는 크게 사고심각도 연구와 사고발생 연구로 구분된다.

먼저 사고심각도 주요 연구들이다. Champahom et al.(2020)는 태국의 도시부와 지방부에서 2011-2015년 발생된 2,096건의 추돌사고 자료를 토대로 도로 형태별(고속도로와 신호교차로) 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료, 운전자 정보, 도로 특성 등이 수집되었으며 분석모형으로 계층적 로지스틱 모형이 사용되었다. 연구결과, 야간 시간대, 안전벨트 미착용, 차량 크기가 사고심각도 영향 요인으로 나타났다.

Sharafeldin et al.(2022b)은 미국 Wyoming 주의 신호교차로 240개소에서 9년간(2007-2017년, 2010-2011년 제외) 발생된 3,156건의 추돌사고 자료로 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료, 도로경사, 노면상태 자료 등이 수집되었으며 분석모형으로 순서형 프로빗 모형이 사용되었다. 연구 결과, 안전벨트 미착용, 운전자 나이 및 성별, 오토바이 차종, 노면 마찰력이 사고심각도 영향 요인으로 나타났다.

Zhang et al.(2022)은 중국 Harbin 시의 신호교차로를 대상으로 2015-2019년 발생된 1,236건의 추돌사고 자료를 통해 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료가 수집되었으며 분석모형으로 머신러닝 분류모형인 LightGBM과 랜덤 포레스트 모형이 사용되었다. 연구 결과, 온도, 날씨, 시간대, 차량 유형이 사고심각도의 주 영향 요인으로 나타났다.

Yuan et al.(2023)은 미국 전역에 위치한 주요 신호교차로를 대상으로 2017-2019년 발생된 2,062건의 추돌사고 자료를 통해 사고심각도 영향 요인을 분석하였다. 분석자료로 사고자료가 수집되었으며 분석모형으로 순서형 프로빗 모형이 사용되었다. 연구 결과, 차량 크기, 기상 악천후, 제한속도, 운전자 연령층이 사고심각도의 주요 영향 요인으로 나타났다.

다음은 사고발생 주요 연구들이다. Wang et al.(2003)은 일본 Tokyo 시의 4지 신호교차로 115개소에서 1992-1995년 발생된 589건의 추돌사고 자료로 사고발생 영향 요인을 분석하였다. 분석자료로 사고자료, 교통량, 신호운영, 도로기하구조 및 소음자료가 수집되었으며 분석모형으로 포아송 회귀분석이 사용되었다. 연구 결과, 제한속도, 종단경사, 교차로 접속각, 주야간 교통량 비율, 우회전 교통량 규모, 접근로 총 차로수가 사고발생 영향 요인으로 나타났다.

Yan et al.(2005)은 미국 Florida 내 신호교차로에 대해 2001년 사고자료와 도로기하구조 자료를 활용하여 사고발생 영향 요인을 분석하였다. 분석모형으로 준유발노출기법(Quasi-induced exposure technique)과 로지스틱 회귀분석이 사용되었다. 연구 결과, 운전자 연령층, 거주지, 차종, 차로수, 중앙분리시설, 사고시간, 노면상태, 토지이용 특성, 제한속도가 사고발생 영향 요인으로 나타났다.

Wang and Abdel-Aty(2006) 또한 Yan et al.(2005)의 연구 대상지와 동일한 미국 Florida 내 신호교차로를 대상으로 사고발생 영향 요인과 시공간적 관계를 분석하였다. 분석자료로 Florida 내 208개 4지 교차로를 실험군으로,

476개 교차로를 대조군으로 설정 후, 2000-2002년 발생된 사고자료, 도로기하구조 자료, 교통운영 자료 등이 수집되었으며 분석모형으로 음이항 회귀분석이 사용되었다. 연구 결과, 종단경사 및 토지이용 특성과 추돌사고 간 높은 상관관계가 나타났으며 주도로-부도로 교통량비, 주도로 회전차로수, 신호 현시수, 우회전 도류화, 전용 회전차로, 보호좌회전 신호운영이 사고발생 영향 요인으로 나타났다.

Yan and Radwan(2006) 또한 위와 동일한 미국 Florida 내 신호교차로의 2001년 사고자료와 도로기하구조 자료를 활용하여 사고발생 영향 요인을 분석하였다. 분석모형으로 머신러닝 분류모형인 의사결정나무가 사용되었다. 연구 결과, 운전자 연령층, 제한속도, 사고시간, 미끄러운 노면상태 등이 사고발생 영향 요인으로 나타났다.

Park and Park(2007)은 국내 청주시의 4지 신호교차로 106개소에서 2004년 발생한 308건의 사고자료, 도로기하구조 및 교통운영 자료를 통해 도로교통적 사고발생 영향 요인을 분석하였다. 분석모형으로 다중 회귀분석과 포아송 회귀분석이 사용되었다. 연구 결과, 평균 일 교통량(ADT; Average Daily Traffic), 종단경사, 좌회전 전용신호, 평균통행속도가 사고발생 영향 요인으로 나타났다.

Park and In(2009)은 국내 청원군의 일반국도에 위치한 43개 신호교차로에서 2005년 발생된 92건의 사고 자료와 도로기하구조 및 교통량 자료로 지방부 신호교차로의 도로교통 측면에 대한 사고발생 영향 요인을 분석하였다. 분석모형으로 포아송 회귀분석과 음이항 회귀분석이 사용되었다. 연구 결과, 주도로폭, 차량 유출입구수, 중차량 비율이 사고발생 영향 요인으로 나타났다.

이상 살펴본 추돌사고 연구 동향은 Table 2와 같이 정리된다.

Table 2. Summary of studies related to rear-end crashes

Classification	Year	Authors	Method (model types)	Key outcomes (affecting risk factors)
Severity study	2020	Champahom et al.	Hierarchical logistic	Incident time, seat belt use, vehicle size
	2022b	Sharafeldin et al.	Ordinal probit	Seat belt use, driver's age and gender, motorcycle, road pavement friction
	2022	Zhang et al.	LightGBM, Random forest	Temperature, weather, time, vehicle type
	2023	Yuan et al.	Ordinal probit	Vehicle size, adverse weather, speed limit, driver's age
Occurrence study	2003	Wang et al.	Poisson	Speed limit, longitudinal slope, angle of the approach, night-to-day traffic flow ratio, right-turn volume, total number of approach lanes
	2005	Yan et al.	Quasi-induced exposure, Logistic	Driver age and residence, vehicle type, number of lanes, median barrier, accident time, road surface condition, land use character, speed limit
	2006	Wang and Abdel-Aty	Negative binomial	Longitudinal slope, landuse character, heavy traffic on the major and minor roadway, right and left-turn lanes on the major roadway, number of phases per cycle, channelized right-turn, exclusive left-turn and right-turn lanes, protected left-turning signal
	2006	Yan and Radwan	Decision tree	Driver's age, speed limit, daytime, slippery road surface conditions
	2007	Park and Park	Multiple linear, Poisson	ADT, longitudinal slope, protected left-turning signal, average travel speed
	2009	Park and In	Poisson, Negative binomial	The width of major roadway, the number of exit/entry, the ratio of heavy vehicle

3. 시사점 및 연구 차별성

선행 연구들의 신호교차로 추돌사고 심각도 영향 요인은 크게 네 가지 측면(발생환경적 요인, 인적 요인, 차량적 요인, 도로교통적 요인)으로 구분된다. 발생환경적 요인으로는 사고시간, 온도, 기온, 인적 요인으로는 안전벨트 착

용여부, 운전자 나이 및 성별, 차량적 요인으로는 차종(차량 크기), 도로교통적 요인으로는 노면마찰력, 제한속도가 주요 요인으로 나타났다. 이러한 요인들은 모두 국외 연구결과이며 사고발생 영향 요인을 연구한 국내 연구에서는 도로교통적 요인으로 교통량, 중차량 비율, 종단경사, 보호좌회전 여부, 주도로 차로폭, 교차로 진출입구수가 주요 발생요인으로 나타났다.

본 연구의 차별성이다. 첫째, 신호교차로 추돌사고 심각도에 대한 국내 연구 부재에 따라 본 연구에서는 국내 신호교차로를 대상으로 사고심각도 영향 요인을 분석하였다. 둘째, 기존 연구들은 신호교차로에서의 추돌사고의 발생 위치까지는 고려하지 못했다. 본 연구는 신호교차로 유출부에서 발생된 추돌사고만을 엄격히 선별 · 분석함으로써 교차로 사고를 미연에 방지하기 위한 실질적 대책을 제언하였다. 셋째, 기존 연구들은 수집자료 한계로 추돌사고의 다양한 요인들까지는 도출하지 못했다. 본 연구에서는 선행 연구들에서 다루어졌던 변수 외 운전자의 음주운전 여부, 현장의 교통운영 여건과 주요 안전시설 등을 포함시킴으로써 보다 유의미한 요인들을 도출하고자 하였다.

Data and Methodology

1. 분석자료

본 연구의 분석절차는 Figure 1과 같다. 1단계는 자료수집 후 대상 교차로를 확정한다. 2단계는 수집된 자료의 전처리와 변수를 설정한다. 3단계는 정리된 자료의 기술통계량을 산출한다. 마지막 4단계는 사고심각도 영향 요인을 분석하는데, 두 가지 방법(머신러닝 기법의 사용/미사용)을 수행한다. 머신러닝 기법의 사용은 사고심각도를 분류하는 중요 변수들을 탐색하기 위해 수행되며, 두 가지 방법을 통해 로지스틱 회귀분석으로 변수의 영향력을 구체화하여 그 결과를 비교한다.

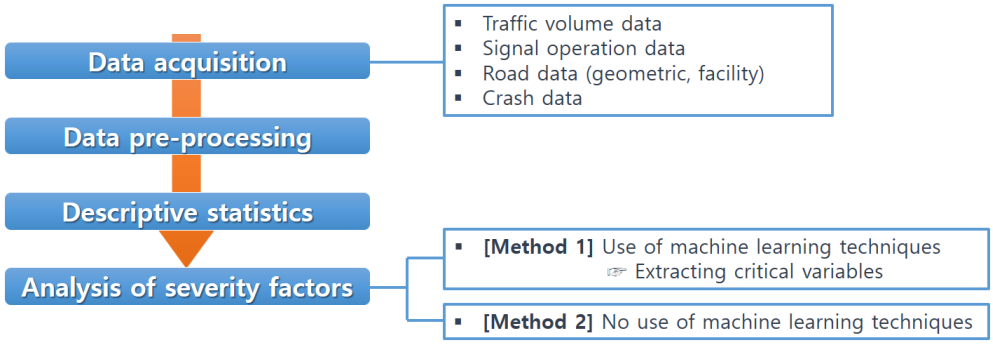

Figure 1. Overall research procedure

연구지역으로는 충청북도 청주시를 대상으로 하였다. 청주시에서 매년 수행하는 주요교차로 70개소의 교통량 자료를 가지고 개별 사고 발생시간대의 접근로 교통량 자료가 구축되었다. 신호운영 자료는 충북지방경찰청을 통해 수집되었다. 도로특성 자료는 청주시 교통지리정보시스템(T-Gis)과 로드뷰(road view)를 통해 차로수, 교차로 통과거리 등의 기하구조와 주요 교통시설(전방신호기, 단속카메라, 미끄럼방지포장) 자료가 수집되었다.

또한, 본 연구에서 가장 중요한 교통사고 자료는 충북지방경찰청 및 도로교통공단의 교통사고 잦은 곳 개선 DB에서 수집되었다. DB를 통해 최근 3년(2020-2022년) 경찰에서 조사된 개별 교통사고의 상황도와 그 밖의 사고정보를 분석하였으며 Figure 2와 같이 교차로 통과 중 추돌사고(type A)나 우회전 도류로 추돌사고(type B) 및 합류부 추돌사고 등을 제외하고 오로지 교차로 유출부에서의 신호대기나 정차 중 발생된 추돌사고(type C)만을 분석대상으로 하였다.

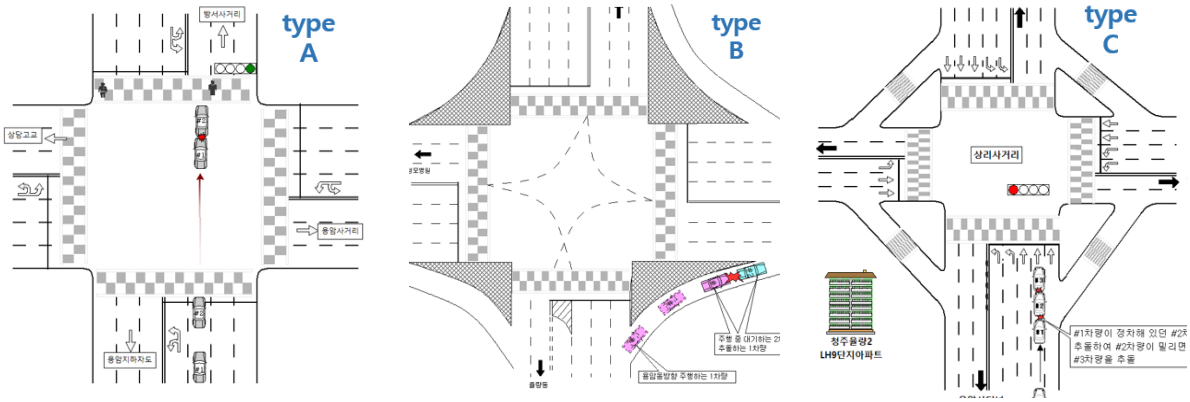

Figure 2. Types of rear-end crashes

한편, 본 연구의 교차로 범위는 경찰의 교통사고 통계작성 기준을 따랐다. 평면교차로 설계 지침(국토교통부)에서는 도시지역 교차로 영향권에 대한 범위가 명확히 제시되어 있지 않다. 따라서, 본 연구에서는 Figure 3과 같이 경찰 내부의 교통사고 통계업무 지침인 교차로 부근으로 분류되는 기준(횡단보도 측단에서 30m 이내)을 토대로 하여 신호교차로 유출부에서 발생한 사고를 교차로 접근부 사고로 정의하였다.

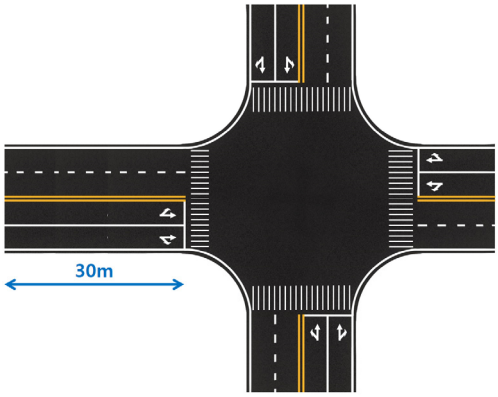

Figure 3. Spatial range of intersection approach (from intersection exit)

이에 따라 70개 교차로 중 교통사고 잦은 곳 개선 DB에서 관리되고 있는 교차로와 최근 3년(2020-2022년)간 1건 이상의 접근부 추돌사고가 발생된 교차로를 분석한 결과, 최종 57개 교차로가 이에 해당되는 것으로 파악되었다. 선정된 57개 교차로에서 3년간 발생된 차대차 교통사고는 총 1,128건(교차로당 19.8건)으로 나타났다. 이 중 전체 추돌사고는 총 220건이며 교차로 접근부에서만 발생한 추돌사고는 총 171건으로 확인되었다.

따라서, 본 연구에서는 171건의 추돌사고가 사고자료로 구축되었으며 데이터셋의 변수구성은 Table 3과 같다. 종속변수는 심각 사고로 사고의 경중에 따라 중상(Yes)/경상(No)으로 구분되었다. 독립변수는 4가지 측면(발생환경, 사람, 차량, 도로교통)에 해당하는 27개 변수로 구성되었으며 이 중 도로교통에는 기하구조, 교통량 구성, 신호운영, 안전시설에 대한 변수들이 포함되었다. 구축된 변수 대부분은 범주형(categorical)으로 구성되었으며 교차로 통과거리, 접근로 총 교통량, 접근로 차로당 교통량, 접근로 중차량 비율은 수치형(numerical)으로 구성되었다.

또한, 독립변수 중 가해자 및 피해자의 연령기준은 국내 청년기본법과 노인복지법에 따라 청년층은 19-34세, 중장년층은 35-64세, 노년층은 65세 이상으로 구분하였다. 자동차 유형분류는 세단에 전고 1,600mm 미만의 승용차가 포함되었으며 RV는 전고 1,600mm 이상의 승용차, 16인승 미만의 승합차와 2.5t 미만의 화물차가 포함되었으며

중차량은 16인승 이상의 승합차(버스)와 2.5t 이상의 화물차가 포함되었다. 직진과 좌회전의 신호운영 형태는 동시신호, 분리신호, 중첩신호로 구분하였다.

Table 3. The selected variables

Classification			Description
Dependent variable			Serious crash (Yes/No)
Independent variables	Environment		Time (Daytime/Nighttime), Weather (Sunny/Etc), Road surface (Dry/Etc)
	Human		Age of offender and victim (Youth/Middle/Senior), Gender of offender and victim (Male/Female), Drinking of offender (Yes/No)
	Car		Car type of offender and victim (Sedan/RV/Heavy vehicle)
	Road traffic	Geometric structure	Vertical alignment (Flat/Downhill), The distance through an intersection (Numeric), Total number of lanes (3 or less/4/5 or more), The number of TH lanes (2 or less/3 or more), LT lane (Yes/No), RT lane (Yes/No), U-turn (Yes/No), Speed limit (60 or less/70 or more)
		Traffic composition	Total volume (Numeric), Volume per lane (Numeric), Heavy vehicle composition ratio (Numeric)
		Signal operation	TH-LT signal operation type (Simultaneous/Separate/Overlap), Yellow time (3 second or less/4 second/5 second or more), All-red time (Yes/No)
		Safety facilities	Forward traffic light (Yes/No), Enforcement camera (Yes/No), Skid-proof facility (Yes/No)

2. 분석 방법론

본 연구와 같은 이분형 종속변수(중상 vs. 경상)를 전통적인 로짓 모형으로 분석할 경우, 다수의 예측변수 투입으로 인한 자유도 감소 문제가 발생한다(Lee and Kim, 2021). 본 연구에서는 다양한 머신러닝 분류 모형들(classification models) 중 분류성능이 우수하고 중요변수 추출이 가능하다는 장점에 따라 널리 활용되고 있는 대표적 세 가지 모형(랜덤 포레스트, 서포트벡터머신, XG부스트)을 적용하였다. 이들 모형으로 심각 사고에 미치는 중요 변수를 탐색한 후, 로지스틱 회귀분석으로 중요 변수들의 영향력을 구체적으로 분석하였다. 또한, 머신러닝 기능을 활용하지 않고 회귀모형의 자체적인 변수 선택방법(variable selection)도 적용하여 이 두 방법에 대한 비교를 수행하였다.

1) 랜덤 포레스트(Random Forest, RF)

Leo Breiman 및 Adele Cutler에 의해 제안된 랜덤 포레스트 기법은 머신러닝 분야의 대표적인 알고리즘이다. 랜덤 포레스트는 Figure 4와 같이 다수의 의사결정나무(Decision Tree)를 생성하되 개별 의사결정나무 내에서는 표본 및 변수 선택 과정에서의 무작위성을 최대한 부여함으로써 모형의 예측률을 높이는데 효과적인 앙상블 기법으로 평가받고 있다(Yoo, 2015).

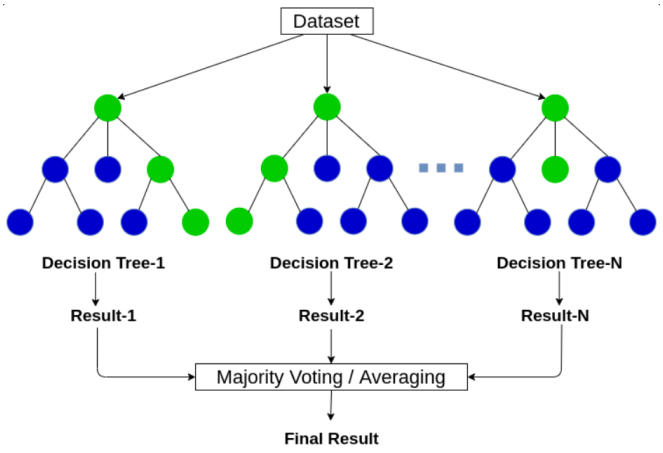

Figure 4. Conceptual diagram of random forest (Awasthi, 2020)

랜덤 포레스트의 기반 모델인 의사결정나무는 규칙 기반 모델로 최상위 노드와 중간 노드, 말단 노드로 구성된다. 최상위와 중간 노드는 특정 조건에 따른 데이터의 분기점을 학습하고 말단 노드에서는 해당 노드에 존재하는 데이터의 종속변수 그룹에 대한 비율 정보를 갖고 있다. 테스트 데이터가 최상위 노드로 들어오면 대상 데이터는 각 노드에서 학습한 조건에 의해 분기된다. 말단 노드에 도착하면 대상 노드에 존재하는 종속변수의 그룹 중 가장 비율이 높은 그룹으로 결과를 예측한다(Lee at al., 2019).

랜덤 포레스트는 범주형 변수뿐만 아니라 연속형 변수 예측에도 적용되며 높은 예측력을 보여주기 때문에 널리 쓰이는 장점으로 활용도가 높다(Lin et al., 2017). 특히, 무수히 많은 트리 구조를 갖고 있어도 무작위 중복표본 추출(bootstrapped)된 샘플이 서로 다른 특성 변수의 조합으로 구성되어 있어 과적합과 다차원성의 오류에 빠지지 않는다(Chung et al., 2021).

2) 서포트벡터머신(Support Vector Machine, SVM)

서포트벡터머신은 Vapnik이 제안한 머신러닝 기법으로 경험적 위험 최소화 원칙을 기반으로 하는 다른 통상적인 머신러닝 기법과는 달리 구조적 위험 최소화를 기반으로 하여 일반화 오류의 상한을 최소화하는 머신러닝 기법이다. 서포트벡터머신은 동일 클래스에 속한 데이터 포인트들이 모여있는 지역에 경계를 설정하는 분류기법으로 학습데이터의 마진을 최대로 하는 경계를 찾는 알고리즘이다(Cho et al., 2023).

서포트벡터머신에서는 퍼셉트론 기반의 모형에 Figure 5와 같이 가장 안정적인 판별 경계선을 찾기 위한 제한 조

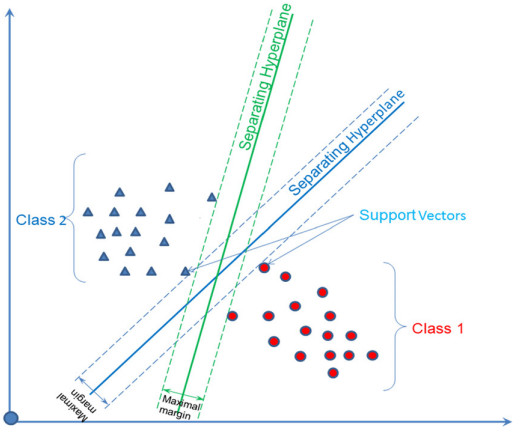

Figure 5. Conceptual diagram of support vector machine (Virnodkar et al., 2020)

건을 추가한 두 개의 클래스(모형 경계로 구분) 중 새로운 데이터가 어느 클래스에 속할지를 판단한다. 선형 분리시 각 클래스에 오분류된 데이터 포인트들에 벌점을 부과한 후 총 벌점이 최소가 되는 초평면을 선택한다. 이러한 서포트벡터머신은 선형 뿐만 아니라 비선형 분류에도 활용되는데 비선형 분리를 위한 데이터 집합의 경우는 커널 함수를 이용하여 분류를 수행한다(Park et al., 2021).

이러한 서포트벡터머신은 다른 알고리즘 대비 높은 분류성능 및 과적합(overfitting)에 대한 일반화 우수성, 다양한 커널 함수를 사용한 비선형 분류, 이상치(outlier)에 대한 강건성(robustness) 등의 장점이 있으며 본 연구와 같이 비교적 작은 데이터셋에서도 양호한 성능을 나타내는 장점이 있다.

3) XG부스트(Extreme Gradienting Boost, XGB)

XG부스트는 gradient boosting을 기반으로 알고리즘을 개선한 기법이다(Chen and Guestrin, 2016). 트리를 병렬로 작동하도록 구성하여 효율성을 개선하여 우수한 성능을 입증하였으며 랜덤 포레스트와 마찬가지로 분류 및 회귀문제를 해결하는 대표적인 모형이다(Son and Park, 2022).

랜덤 포레스트 모형이 여러 개의 의사결정나무를 사영해 결과를 평균 내는 방법이라면, XG부스트는 Figure 6과 같이 동일 원리를 적용 후 여러 개의 결과를 추천하여 오답에 가중치(ω)를 부여한다. 이렇게 가중치가 적용된 오답은 관심을 가지고 정답이 될 수 있도록 결과를 만들고 해당 결과에 대한 오답을 찾아 같은 작업을 반복한다(Yang et al., 2023).

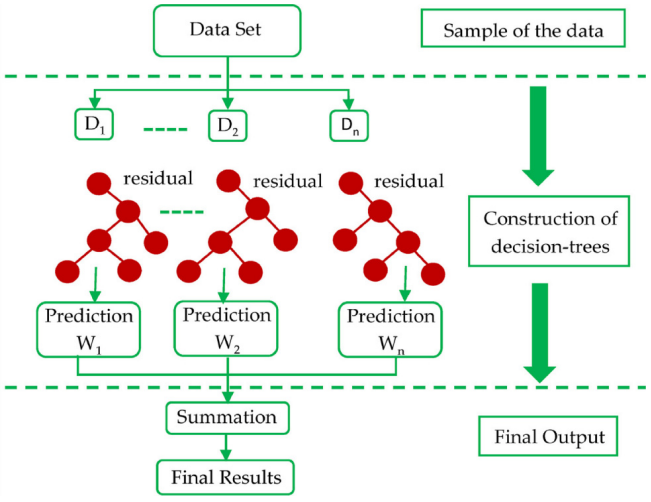

Figure 6. Conceptual diagram of extreme gradienting boost (Khan et al., 2022)

XG부스트는 분류, 회귀 및 순위 지정 작업에 모두 적용될 수 있는 확장성의 장점을 갖는다. 성능이 좋은 앙상블 또는 신경망 기법에 비해 속도가 빠르고 정확도가 높으며 단일 시스템에서 수십억 개로 확장된 예제 분석이 가능하기 때문에 빅데이터에 대한 컴퓨터 자원 활용률이 우수하다(Chen and Guestrin, 2016).

Result and Discussion

Table 4는 구축된 자료의 기술통계량이다. 총 171건의 추돌사고 중 종속변수인 중상사고와 경상사고는 각각 47건(27.5%), 124건(72.5%)으로 나타났으며 사망사고와 부상신고사고는 발생되지 않았다.

Table 4. Descriptive statistics of variables (number of samples = 171)

Classification		Descriptive statistics
Dependent variable	Serious crash	Yes=47 (27.5%), No=124 (72.5%)
Independent variables	Time	Daytime=100 (58.5%), Nighttime=71 (41.5%)
	Weather	Sunny=148 (86.5%), Etc=23 (13.5%)
	Road surface	Dry=145 (84.8%), Etc=26 (15.2%)
	Age of offender	Youth=53 (31.0%), Middle=102 (59.6%), Senior=16 (9.4%)
	Age of victim	Youth=61 (35.7%), Middle=91 (53.2%), Senior=19 (11.1%)
	Gender of offender	Male=139 (81.3%), Female=32 (18.7%)
	Gender of victim	Male=121 (70.8%), Female=50 (29.2%)
	Drinking of offender	Yes=43 (25.1%), No=128 (74.9%)
	Car type of offender	Sedan=87 (50.9%), RV=65 (30.8%), Heavy vehicle=19 (11.1%)
	Car type of victim	Sedan=94 (55.0%), RV=70 (40.9%), Heavy vehicle=7 (4.1%)
	Vertical alignment	Flat=146 (85.4%), Downhill=25 (14.6%)
	The distance through an intersection	Min=28.0m, Avg=51.6m, Max=85.0m, SD=10.9
	Total number of lanes	3 or less=38 (22.2%), 4=67 (39.2%), 5 or more=66 (38.6%)
	The number of TH lanes	2 or less=71 (41.5%), 3 or more=100 (58.5%)
	LT lane	Yes=154 (90.1%), No=17 (9.9%)
	RT lane	Yes=136 (79.5%), No=35 (20.5%)
	U-turn	Yes=117 (68.4%), No=54 (31.6%)
	Speed limit	60 or less=114 (66.7%), 70 or more=57 (33.3%)
	Total volume	Min=204.0vph, Avg=1,111.7vph, Max=3,046.0vph, SD=438.8
	Volume per lane	Min=92.7vph, Avg=266.4vph, Max=609.2vph, SD=84.2
	Heavy vehicle composition ratio	Min=0.3%, Avg=7.0%, Max=55.3%, SD=5.5
	TH-LT signal operation type	Simultaneous=80 (46.8%), Separate=24 (14.0%), Overlap=67 (39.2%)
	Yellow time	3 second or less=17 (9.9%), 4 second=113 (66.1%), 5 second or more=41 (66.1%)
	All-red time	Yes=60 (35.1%), No=111 (64.9%)
	Forward traffic light	Yes=118 (69.0%), No=53 (31.0%)
	Enforcement camera	Yes=45 (26.3%), No=126 (73.7%)
	Skid-proof facility	Yes=24 (14.0%), No=147 (86.0%)

머신러닝 모형들의 최적성능 발휘를 위한 하이퍼파라미터 튜닝은 여러 방법들이 존재하지만 본 연구에서는 보편적으로 사용하는 Grid Search 방법으로 최적 조합을 찾았으며, 데이터셋의 분할 또한 보편적 비율인 8(학습 데이터) : 2(테스트 데이터)로 분할하였다.

또한, 본 연구의 모형별 성능 평가는 머신러닝 분류 모형의 성능 평가에서 활용하는 정확도(Accuracy)에 더불어 F1 score까지 검토하였다. 통상적으로 분류 모형은 혼동행렬표(confusion matrix)를 활용하여 예측 결과를 평가하는데, 혼동행렬표는 실제 클래스와 예측된 클래스의 매칭을 이용하여 분류 모형을 평가하는 다양한 지표들을 산출하는 도구이다. 본 연구와 같은 이진 분류(binary classification) 문제에서 실제 클래스는 Positive/Negative로 나누어져 있고, 분류 모형은 샘플들을 Figure 7과 같이 Positive/Negative로 분류한다. 따라서 Figure 7처럼 TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)의 네 가지 경우가 발생할 수 있다.[3)]

3) 통계학에서는 FP를 Type I error(false alarm)이라 하고, FN은 Type II error(missed detection) 라고 정의하고 있음

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Figure 7. Confusion matrix of the machine learning model

혼동행렬표를 통해 분류 모형의 성능 측정에서 사용되는 정확도는 판별한 전체 샘플 중 TP와 TN의 비율로 분류 모형을 평가하는 가장 단순한 지표로 Equation 1과 같이 산출되며 가장 널리 활용되고 있지만 불균형한 클래스를 가진 데이터셋을 평가하기 어렵다는 단점이 있다.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

예를 들어, Positive와 Negative의 비율이 2:8로 불균형한 클래스를 가지는 데이터셋에서는 모든 예측을 Negative로 해버리는 엉터리 분류기의 정확도조차 80%로 측정될 수 있다. 이처럼 데이터셋의 클래스 불균형을 고려하기 위해 machinelearningmastery.com에서는 Figure 8과 같이 분석 목적과 데이터셋의 특성을 고려하여 다양한 성능 지표를 제시하고 있다(Brownlee, 2020).

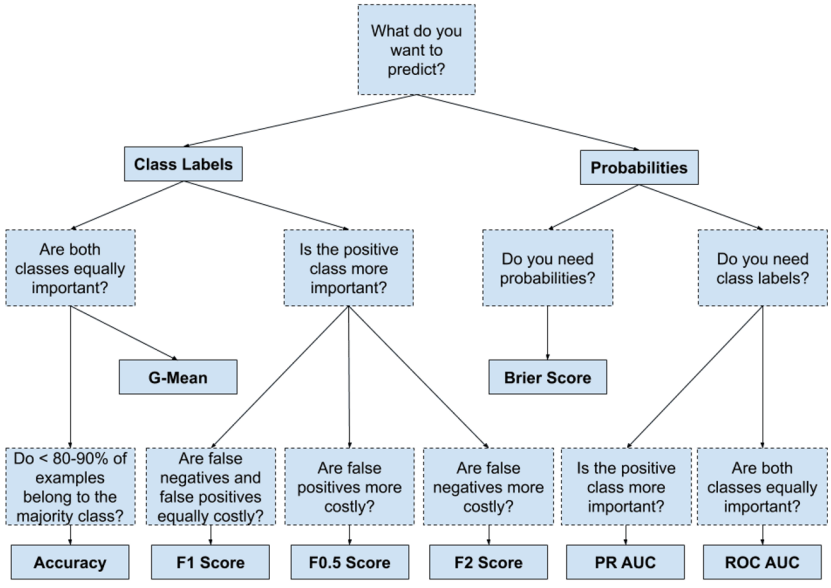

Figure 8. Performance metrics of binary classification for imbalanced data (Brownlee, 2020)

본 연구의 성능 지표는 Figure 8에서와 같이 다수 범주(경상사고) 비율이 80-90% 미만에 해당되기 때문에 정확도(Accuracy) 사용에 큰 무리는 없으나, 두 범주 간 어느 정도 불균형이 존재(경상사고 72.5%, 중상사고 27.5%) 하기 때문에 F1 score를 추가로 고려하였다. F1 score는 불균형 데이터에서의 정확도가 가지고 있는 약점을 보완하기

위해 사용되는 성능 지표로, 정밀도(Precision, 분류 모형이 Positive로 판정한 것 중 실제로 Positive인 샘플 비율)과 재현율(Recall, 실제 Positive 샘플 중 분류 모형이 Positive로 판정한 비율)의 성능을 동시에 고려하기 위해 이 둘을 조화평균한 지표이다. F1 score는 다른 지표들처럼 0과 1사이의 값을 가지며 1에 가까울수록 분류 성능이 좋음을 나타내는데, 통상 0.8 이상이면 상위 7.5% 수준, 0.7 이상이면 상위 15% 수준에 해당한다. Equation 2, 3, 4는 이들 정밀도, 재현율, F1 score의 산출식이다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\,score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \tag{4}$$

본 연구에서 구축된 자료를 통해 분석된 모형별 성능 평가 결과는 Table 5와 같이 랜덤 포레스트, 서포트벡터머신, XG부스트 순으로 나타났다. XG부스트의 정확도가 다소 낮은 수준(0.7118)을 나타냈지만 세 모형의 F1 score 값은 모두 0.8 이상으로 나타나 전반적으로 양호한 수준인 것으로 분석되었다.

Table 5. Performance evaluation results by each model

Classification	RF	SVM	XGB
Accuracy	0.8412	0.8088	0.7118
F1 score	0.9371	0.9200	0.8130

Figure 9는 중요 예측변수 파악을 위해 수행된 랜덤 포레스트 모형의 변수 추출 결과이다. 27개 독립변수 중 심각 사고의 영향 요인에 대한 예측력이 높은 상위 10개의 변수를 추출하였으며, 변수들은 랜덤 포레스트의 정확도개선지수(MDA; Mean Decrease Accuracy)에 대한 중요도 순서대로 도출되었다.[4)]

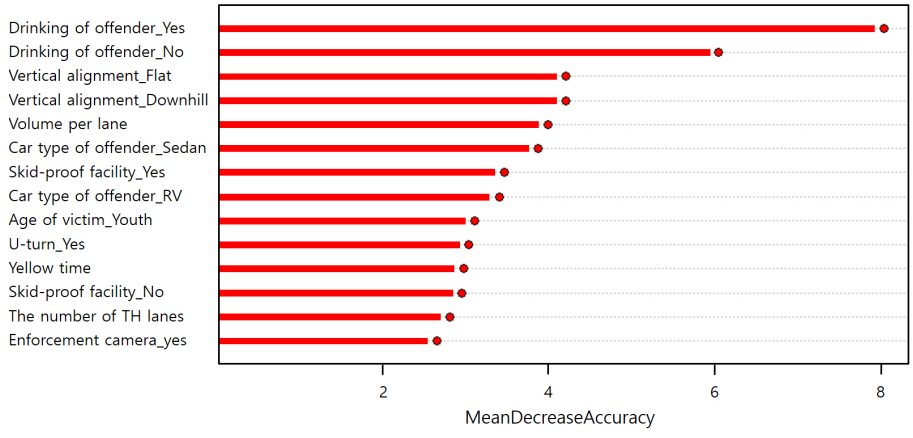

Figure 9. Selected top 10 variables in RF model (in case of MDA)

4) 랜덤 포레스트는 예측 변수의 예측력을 측정하기 위하여 정확도개선지수(MDA) 외 지니지수에 기반한 MDG(Mean Decrease Gini)도 함께 제공하나, MDG는 MDA에 비해 상대적으로 편의(bias)가 크며 추정 결과의 불안정성을 안고 있는 것으로 알려져 있음(Sandri and Zuccolotto, 2010)

같은 방법으로 서포트벡터머신과 XG부스트에 대해서도 각각 상위 10개의 변수를 추출하였으며 그 결과는 Table 6과 같다. 탐색된 중요 변수들은 모형별로 약간의 차이가 있으나 전반적으로는 유사한 수준을 나타냈다. 3개 모형에서 공통적으로 도출된 중요 변수는 3개(피해 운전자의 연령대, 가해 운전자의 음주운전 여부, 가해 운전자의 차종)에 해당되며 2개 모형에서 공통적으로 도출된 중요 변수는 8개(도로의 종단선형, 직진 차로수, 유턴구역, 총 교통량, 차로당 교통량, 황색신호시간, 단속카메라, 미끄럼방지포장)에 해당되며 1개 모형에서만 도출된 중요 변수는 5개(기상상태, 노면상태, 가해운전자의 연령대, 좌회전 전용차로, 직진-좌회전 분리신호)로 나타났다.

Table 6. Selected top 10 variables by each model

Ranking	RF	SVM	XGB
1st	Drinking of offender	Car type of offender	Total volume
2nd	Vertical alignment	Age of victim	Age of offender
3rd	Volume per lane	Drinking of offender	Drinking of offender
4th	Car type of offender	Total volume	Age of victim
5th	Skid-proof facility	Vertical alignment	Car type of offender
6th	Age of victim	Volume per lane	TH-LT signal operation type
7th	U-turn	U-turn	Enforcement camera
8th	Yellow time	Yellow time	The number of TH lanes
9th	The number of TH lanes	Skid-proof facility	Weather
10th	Enforcement camera	LT lane	Road surface

상기와 같이 세 모형에서 심각 사고의 영향 요인을 예측한 총 16개의 중요 변수를 가지고 로지스틱 회귀분석을 실시하였다. 로지스틱 회귀분석으로 다른 예측변수들을 통제한 상태에서 개별 변수들의 실질적인 영향력을 살펴보았으며 분석 결과는 Table 7과 같다. 심각 사고 요인에 통계적으로 유의(95% 신뢰수준)한 영향력을 미치는 것으로 나타난 변수는 총 8개(노면상태, 가해 운전자의 음주운전 여부, 가해 운전자의 차종, 피해 운전자의 연령대, 종단선형, 직진 차로수, 유턴구역, 미끄럼방지포장)인 것으로 나타났다.

Table 7. Logistic regression results for factors affected by severe rear-end crashes use of ML

Classification	Estimate	Std. error	Pr (>|z|)	Signif. level	Exp (B)
(Intercept)	2.7765	2.7969	0.3208		16.0628
Weather_Sunny	2.5728	1.4423	0.0745	.	13.1027
Road surface_Dry	-3.2207	1.3385	0.0161	*	0.0399
Age of offender_Youth	0.5012	0.6032	0.4061		1.6507
Drinking of offender_Yes	1.9907	0.6173	0.0013	**	7.3208
Car type of offender_Sedan	-1.6199	0.5613	0.0039	**	0.1979
Age of victim_Youth	-1.2280	0.6247	0.0493	*	0.2929
Vertical alignment_Flat	-2.0452	0.8004	0.0106	*	0.1294
The number of TH lanes	1.9543	0.7628	0.0104	*	7.0589
LT lane_Yes	-0.0025	1.0380	0.9981		0.9975
U-turn_Yes	-2.2232	0.7929	0.0051	**	0.1083
TH-LT signal operation type_Separate	-1.1635	0.8548	0.1735		0.3124
Yellow time	-0.0912	0.5029	0.8561		0.9128
Enforcement camera_Yes	-0.7171	0.6335	0.2577		0.4882
Skid-proof facility_Yes	-2.9375	1.3817	0.0335	*	0.0530
Volume per lane	-0.0055	0.0072	0.4382		0.9945
Total volume	-0.0006	0.0017	0.7371		0.9994

signif. levels: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
pchisq test: p-value=4.12071e-06 (q=161.78-107.17, df=136-120)

또한, 본 연구에서는 머신러닝 모형의 사용 없이 회귀모형의 변수선택(variable selection) 방법만으로도 분석을 수행하였으며 변수선택법은 세 가지 방법 중 가장 보편적으로 사용하는 단계선택법(stepwise selection)[5]을 사용하

였다. 분석 결과는 Table 8과 같이 나타났으며 머신러닝 모형을 사용한 Table 7과 비교시, 심각 사고 요인에 통계적으로 유의(95% 신뢰수준)한 영향력을 미치는 것으로 나타난 변수는 총 10개(노면상태, 가해 운전자의 음주운전 여부, 가해 운전자의 차종, 종단선형, 총 차로수, 직진 차로수, 유턴구역, 미끄럼방지포장, 총 교통량, 접근로 차로당 교통량)인 것으로 나타났다.

Table 8. Logistic regression results for factors affected by severe rear-end crashes no use of ML

Classification	Estimate	Std. error	Pr (>|z|)	Signif. level	Exp (B)
(Intercept)	−3.6133	2.5342	0.1539		0.0270
Weather_Sunny	2.5700	1.5199	0.0909	.	13.0663
Road surface_Dry	−3.2250	1.4690	0.0281	*	0.0398
Drinking of offender_Yes	2.7580	0.6475	0.0000	***	15.7677
Car type of offender_Sedan	−1.6245	0.5775	0.0049	**	0.1970
Vertical alignment_Flat	−1.9666	0.7785	0.0115	*	0.1399
Total number of lanes	2.1149	1.0603	0.0461	*	8.2891
The number of TH lanes	1.6534	0.7467	0.0268	*	5.2247
RT lane_Yes	1.6557	0.8977	0.0651	.	5.2366
U-turn_Yes	−2.1101	0.7954	0.0080	**	0.1212
Speed limit_60 or less	−1.3114	0.6721	0.0510	.	0.2694
TH-LT signal operation type_Separate	−1.4516	0.9498	0.1264		0.2342
Skid-proof facility_Yes	−3.7109	1.7812	0.0372	*	0.0245
Total volume	−0.0094	0.0041	0.0201	*	0.9906
Volume per lane	0.0342	0.0164	0.0367	*	1.0348

signif. levels: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
pchisq test: p-value=3.47457e-08 (q=161.78-98.79, df=136-122)

이와 같이, 두 가지 방법(머신러닝 기법의 사용/미사용)에서 도출된 모형의 최종 분석모형 선택을 위해 먼저 AIC (Akaike Information Criterion)를 측정하였다. AIC는 주어진 데이터셋에 대한 통계 모형의 상대적인 품질을 평가하는 방법 중 하나로, 최소의 정보 손실을 갖는 모형이 데이터와 가장 적합한 모형으로 선택되는 방법이며 Equation 5에 의해 구해진다.

$$AIC = -2\ln(L) + 2k \tag{5}$$

여기서 -2ln(L)은 모형의 적합도를 의미하는데 L은 likelihood function을 뜻한다. k는 모형의 추정된 파라미터의 개수이다. AIC 값이 낮다는 것은 즉 모형의 적합도가 높은 것을 의미한다. 본 연구에서 사용한 두 가지 방법에 의한 모형별 AIC는 머신러닝 기법 사용 모형이 141.17로 나타났으며 미사용 모형은 132.79로 나타났다. 이러한 결과는 회귀모형의 변수선택법이 최소 AIC를 산출하는 알고리즘에 의해 작동되기 때문에 두 모형 간 파라미터수 차이에 의한 것으로 판단된다. 그래도, 모든 변수가 전부 포함된 null model의 AIC가 164.01인 점을 감안할 때, 두 모형의 적합도는 많이 개선된 것으로 보인다.

위의 AIC 결과로는 최종 모형의 결정이 쉽지 않기 때문에 두 모형의 사고심각도에 대한 분류 성능평가를 위해 테스트 데이터를 활용하여 모형별 정확도(Accuracy)와 F1 score를 추가로 분석하였다. 분석 결과는 Table 9와 같이 머신러닝 기법을 사용한 방법의 정확도가 0.8338, F1 score가 0.9058로 나타나, 미사용 방법(정확도 0.7353, F1 score 0.8235) 대비 각각 1.14배와 1.1배 우수한 것으로 나타났다.

5) 모든 변수가 포함된 모델에서 출발하고 기준통계치에 가장 도움이 되지 않는 변수를 삭제하거나 모델에서 빠져있는 변수 중에서 기준 통계치를 가장 개선시키는 변수를 추가하는 방법으로, 이러한 변수의 추가 또는 제거를 반복적으로 수행하는 방법을 말함

Table 9. Performance evaluation results by each method

Classification	Method 1 (use of ML)	Method 2 (no use of ML)
Accuracy	0.8338	0.7353
F1 score	0.9058	0.8235

본 연구의 최종 모형 선택은 Table 9의 결과를 따랐다. 연구에서 적용한 두 가지 방법의 모형별 AIC가 크게 차이가 나지 않는 점과 테스트 데이터를 통한 모형의 성능평가 결과가 갖는 의미를 감안할 때, 머신러닝으로 중요 변수를 추출한 분석방법이 보다 적절한 것으로 판단된다(결과는 Table 7).

Table 7의 주요 결과를 토대로 한 신호교차로 접근부 추돌사고의 심각도 요인이다. 첫째, 발생환경적 요인으로 도로포장면이 건조(dry)한 경우는 젖음(wet)이나 결빙(iced) 상태인 경우 대비 심각 사고 발생 가능성이 약 94% 감소하는 것으로 나타났다. 미끄러운 노면이 추돌사고 발생 가능성을 높인다는 Yan and Radwan(2006)의 연구 결과와 함께 사고심각도까지 증가시킬 수 있음을 시사한다.

둘째, 인적 요인이다. 가해 운전자의 음주운전은 비음주 대비 심각 사고 발생 가능성이 무려 7.3배 이상인 것으로 나타났으며, 피해 운전자의 연령층이 청년층인 경우 중장년층, 노년층 대비 심각 사고 발생 가능성이 약 70% 감소하는 것으로 나타났다. 음주운전에 따른 운전자의 인지-반응 저하와 피해자의 신체조건이 사고심각도에 큰 영향을 미치는 것으로 해석되며 이는 기존의 추돌사고 연구들에서는 밝혀지지 않은 결과이다. 한편, 기존 Sharafeldin et al.(2022b), Yuan et al.(2023), Yan et al.(2005), Yan and Radwan(2006)의 연구 결과인 가해 운전자가 젊은 연령층인 경우는 본 연구의 중요 예측변수에는 포함되었으나, 로지스틱 회귀분석에서는 통계적 유의성이 없는 것으로 나타났다.

셋째, 차량적 요인이다. 가해 운전자의 차종이 승용차(sedan)인 경우는 RV나 중차량 대비 심각 사고 발생 가능성이 약 80% 감소하는 것으로 나타났다. 이 결과 또한 기존 Champahom et al.(2020), Zhang et al.(2022), Yuan et al.(2023)의 연구 결과와 일치하는 것으로 나타났다.

넷째, 도로교통적 요인이다. 종단선형이 평지인 경우에는 내리막 대비 심각 사고 발생 가능성이 약 87% 감소하는 것으로 나타났다. 이는 내리막 선형이 사고 발생 가능성을 높인다는 Park and Park(2007)의 연구 결과와 일치하며 사고심각도까지 증가시키는 요인이 될 수 있다. 직진 차로수가 3차로 이상인 도로에서는 2차로 이하 도로 대비 심각 사고 발생 가능성이 무려 7.3배 이상으로 나타났다. 이는 Yan et al.(2005)이나 Wang et al.(2003)의 연구 결과인 총 차로수와는 직접적 비교가 어려우나, 맥락은 유사하다고 볼 수 있다. 유턴구역과 미끄럼방지포장은 심각 사고의 발생 가능성을 각각 약 89% 및 약 95%까지 감소시키는 것으로 나타났다. 이는 유턴구역이 존재하는 교차로 접근부에서의 차량 주행속도 감소 경향과 미끄럼방지포장에 의한 노면의 마찰력 증진 효과가 사고심각도를 완화시키는데 기여한다고 볼 수 있다. 특히, 이러한 노면 마찰력의 효과는 기존 연구 중 Sharafeldin et al.(2022b)의 연구 결과와도 일치하는 것으로 분석되었다.

이상의 연구 결과를 토대로 신호교차로 접근부에서 발생하는 추돌사고의 심각도 감소를 위한 정책제언이다. 먼저 음주단속이 중요할 것으로 판단되며 특히, 비나 눈으로 노면이 미끄러울 수 있는 환경과 직진 차로수가 3차로 이상인 도로에서 실시한다면 보다 효과적일 수 있을 것이다. 또한, 도로시설의 공학적 개선으로는 추돌사고가 많은 곳에 대한 미끄럼방지포장 설치가 적극 권장된다. 미끄럼방지포장은 유지보수의 어려움으로 실제 현장에서 설치가 기피되고 있으나, 추돌사고의 발생 가능성이나 심각도 완화 측면에는 효과적일 것으로 보인다. 이러한 미끄럼방지포장은 특히 내리막 도로의 우선적 검토가 요구된다. 유턴구역의 경우는 추돌사고 심각도 감소 효과가 있을 것으로 예상되지만, 현장의 물리적 도로 여건과 유턴차량의 수요 등을 심도 있게 분석하여 설치 필요성을 판단해야 할 것으로 사료된다.

Conclusion

본 연구에서는 신호교차로 접근부에서 발생하는 추돌사고의 심각도 요인을 분석하고자 하였다. 연구를 위해 충청북도 청주시의 57개 신호교차로 접근부에서 최근 3년간(2020-2022년) 발생한 171건의 추돌사고 자료와 교통량 자료, 신호운영 자료, 도로 기하구조 자료 등을 활용하였으며, 이를 통해 머신러닝 분류모형의 사용 여부와 로지스틱 회귀분석으로 심각사고의 영향 요인을 도출하였다.

연구 결과, 머신러닝 기법을 사용한 방법의 정확도가 0.8338, F1 score가 0.9058로 미사용 방법 대비 높은 성능을 나타냈다. 그리고 심각도 영향 요인으로는 총 8개 요인이 심각사고에 통계적으로 유의($p<0.05$)한 것으로 나타났다. 이는 발생환경적 요인으로 사고 당시의 노면상태, 인적 요인으로 가해 운전자의 음주운전 여부와 피해 운전자의 연령대, 차량적 요인으로 가해운전자의 차종, 도로교통적 요인으로 종단선형, 직진 차로수, 유턴구역과 미끄럼방지포장인 것으로 나타났다.

본 연구의 주요 결과 중 가해운전자의 차종, 노면마찰력(미끄럼방지포장 효과로 간주)은 국외에서 수행된 기존의 신호교차로 추돌사고의 심각도 요인에도 해당되었으나, 나머지 요인들은 국내를 대상으로 수행한 본 연구에서 새롭게 밝혀진 심각도 요인으로 나타났다.

교차로 접근부를 대상으로 수행한 본 연구의 결과는 정책결정권자 및 실무자로 하여금 교차로 사고를 미연에 방지하기 위한 교통안전대책 수립에 활용될 수 있을 것이며 이를 통해 추돌사고의 예방과 교통사고로 인한 사회적 비용 감소에 기여할 수 있을 것으로 기대된다.

이러한 시사점에도 불구하고 본 연구는 다음과 같은 한계점들이 존재한다. 첫째, 본 연구의 분석대상은 CBD내 CI(Critical Intersection) 위주로 이루어진 관계로 도시부 가로망의 다양한 위계(hierarchy) 특성 까지는 다루지 못했다. 따라서, 향후에는 집산도로나 비신호교차로 등에 대한 유사 연구로 도로 특성별 비교가 필요할 것이다. 둘째, 같은 CI급 교차로에서도 빈번히 발생하는 충돌유형은 교차로마다 상이할 수 있다. 이 같은 점을 고려하여 유사한 위계의 교차로를 대상으로 전체 사고 대비 추돌사고 비중이 높은 교차로와 낮은 교차로 간 특성 비교 또한 필요할 것으로 보인다. 셋째, 연구대상 57개 교차로에 대한 접근부 추돌사고만을 선별한 관계로 표본수가 상당히 작은 수준이다. 따라서, 향후에는 분석기간의 확대나 연구대상 지역의 추가 등을 통해 보다 많은 표본이 확보된다면 분석 결과의 신뢰성이 더욱 높아질 것이다. 마지막으로 추돌사고가 빈번한 지역이나 지점에 대한 운전행태 데이터가 분석된다면 보다 의미 있는 결과들이 도출될 수 있을 것으로 전망된다.

알림

본 논문은 대한교통학회 제89회 학술발표회(2023.10.12)에서 발표된 내용을 수정 · 보완하여 작성된 것입니다.

ORCID

YANG Jeonghun http://orcid.org/0009-0001-2879-2260
PARK Jeongsoon http://orcid.org/0000-0002-2197-6796
RIM Heesub http://orcid.org/0000-0001-6059-5155
KIM Kyuhyuk http://orcid.org/0000-0002-1692-5054
SONG Tai-jin http://orcid.org/0000-0002-8700-4105

REFERENCES

Awasthi S. (2020), Random Forests in Machine Learning: A Detailed Explanation, datamahadev.com.

Brownlee J. (2020), Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-sensitive Learning, Machine Learning Mastery.

Champahom T., Jomnonkwao S., Watthanaklang D., Karoonsoontawong A., Chatpattananan V., Ratanavaraha V. (2020), Applying Hierarchical Logistic Models to Compare Urban and Rural Roadway Modeling of Severity of Rear-end Vehicular Crashes, Accident Analysis and Prevention, 141, 105537.

Chen T., Guestrin C. (2016), Xgboost: A Scalable Tree Boosting system, In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data mining, 785-794.

Cho E.S., Jo Y., Gu, Y. S., Oh C., Lee G. W. (2023), Detection of Risk Riding Events for Delivery Scooter : Methodology and Application, J. Korean Soc. Transp., 41(1), Korean Society of Transportation, 19-34.

Chung D. H., Yun J. S., Yang S. M. (2021), Machine Learning for Predicting Entrepreneurial Innovativeness, Asia-Pacific, Journal of Business Venturing and Entrepreneurship, 16(3), 73-86.

Khan K., Ahmad W., Amin M. N., Ahmad A., Nazar S., Alabdullah A. A. (2022), Compressive Strength Estimation of Steel-fiber-reinforced Concrete and Raw Material Interactions using Advanced Algorithms, Polymers, 14(15), 3065.

Khattak M. W., Pirdavani A., De Winne P., Brijs T., De Backer H. (2021), Estimation of Safety Performance Functions for Urban Intersections using Various Functional Forms of the Negative Binomial Regression Model and a Generalized Poisson Regression Model, Accident Analysis and Prevention, 151, 105964.

Kidando E., Kitali A. E., Kutela B., Karaer A., Ghorbanzadeh M., Koloushani M., Ozguven E. E. (2022), Use of Real-time Traffic and Signal Timing Data in Modeling Occupant Injury Severity at Signalized Intersections, Transportation Research Record, 2676(2), 825-839.

Korea Road Traffic Authority (2023), Estimation and Evaluation of Road Traffic Accident Costs.

Lee K. Y., Kim Y. S. (2021), Predictive Model for the Employment Retention of Persons with Disabilities, Journal of Special Education, 37(2), 73-95.

Lee M. W., Kim, Y. G., Jun Y. J., Shin Y. H. (2019), Random Forest based Prediction of Road Surface Condition Using Spatio-Temporal Features, J. Korean Soc. Transp., 37(4), Korean Society of Transportation, 338-349.

Lin W., Wu Z., Lin L., Wen A., Li J. (2017), An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, Ieee Access, 5, 16568-16575.

Mitra S., Bhowmick D. (2020), Status of Signalized Intersection Safety-a Case Study of Kolkata, Accident Analysis and Prevention, 141, 105525.

Park B. H., In B. C. (2009), Rear-end Accident Models of Rural Area Signalized Intersections in the Cases of Cheongju and Cheongwon, International Journal of Highway Engineering, 11(2), 151-158.

Park B. H., Park J. S. (2007), Analysis of Rear-End Accidents at 4-legged Signalized Intersections in Cheongju, J. Korean Soc. Transp., 25(5), Korean Society of Transportation, 57-66.

Park J. H., Lee G. W., Oh C., Kim J. H., Yun D. G. (2021), Development of Evaluation Methodology for Rear Collision Situation Using Vehicle Sensor Data, J. Korean Soc. Transp., 39(6), Korean Society of Transportation, 826-837.

Sandri M., Zoccolotto P. (2010), Analysis and Correction of Bias in Total Decrease in Node Impurity Measures for

Tree-based Algorithms, Statistics and Computing, 20(4), 393-407.

Sharafeldin M., Farid A., Ksaibati K. (2022a), Investigating the Impact of Roadway Characteristics on Intersection Crash Severity, Eng, 3(4), 412-423.

Sharafeldin M., Farid A., Ksaibati K. (2022b), Injury Severity Analysis of Rear-End Crashes at Signalized Intersections, Sustainability, 14(21), 13858.

Son S. O., Park J. Y. (2022), Assessment of Crash Prediction Models for Intersections with Severity Weight Parameters Using Data Science Approaches, J. Korean Soc. Transp., 40(2), Korean Society of Transportation, 190-204.

Virnodkar S. S., Pachghare V. K., Patil V. C., Jha S. K. (2020), Remote Sensing and Machine Learning for Crop Water Stress Determination in Various Crops: A Critical Review, Precision Agriculture, 21(5), 1121-1155.

Wang X., Abdel-Aty M. (2006), Temporal and Spatial Analyses of Rear-end Crashes at Signalized Intersections, Accident Analysis and Prevention, 38(6), 1137-1150.

Wang Y., Ieda H., Mannering F. (2003), Estimating Rear-end Accident Probabilities at Signalized Intersections: Occurrence-mechanism Approach, Journal of Transportation Engineering, 129(4), 377-384.

Yang J., Jiang P., Suhail S. A., Sufian M., Deifalla A. F. (2023), Experimental Investigation and AI Prediction Modelling of Ceramic Waste Powder Concrete: An Approach Towards Sustainable Construction, Journal of Materials Research and Technology, 23, 3676-3696.

Yan X., Radwan E. (2006), Analyses of Rear-end Crashes Based on Classification Tree Models, Traffic Injury Prevention, 7(3), 276-282.

Yan X., Radwan E., Abdel-Aty M. (2005), Characteristics of Rear-end Accidents at Signalized Intersections using Multiple Logistic Regression Model, Accident Analysis and Prevention, 37(6), 983-995.

Yoo J. E. (2015), Random Forests, An Alternative Data Mining Technique to Decision Tree, Educational Evaluation Study, 28(2), 427-448.

Yuan R., Gu X., Peng Z., Xiang, Q. (2023), Analysis of Factors Affecting Occupant Injury Severity in Rear-end Crashes by Different Struck Vehicle Groups: A Random Thresholds Random Parameters Hierarchical Ordered Probit Model, Journal of Transportation Safety and Security, 15(6), 636-657.

Yuan Y., Wang S., Liu Z., Cui G., Wang Y. (2022), Influencing Factors Analysis of Side Right-angle Collisions Severity at Intersections Based on Decision Tree, International Journal of Crashworthiness, 27(1), 59-69.

Zhang W., Liu T., Yi J. (2022), Exploring the Spatiotemporal Characteristics and Causes of Rear-end Collisions on Urban Roadways, Sustainability, 14(18), 11761.