기계학습을 활용한 고령운전자 교통사고 분석 및 교통사고 데이터 정책 제언

Elderly Driver-involved Crash Analysis and Crash Data Policy

김 승 훈*

* 주저자 및 교신저자 : 국토연구원 국토인프라연구본부 부연구위원

Seunghoon Kim*

* National Infrastructure Research Division in Korea Research Institute for Human Settlements

† Corresponding author : Seunghoon Kim, sh.kim@krihs.re.kr

Vol. 21 No.5(2022)
October, 2022
pp.90~102

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2022.21.5.90

Received 22 July 2022
Revised 9 August 2022
Accepted 20 September 2022

요 약

우리나라가 고령화시대에 진입하면서 고령운전자를 위한 교통 안전성 정책에 대한 관심이 높아지고 있다. 이를 위해서는 고령자 관련 교통사고의 영향요인을 분석하는 연구가 활성화될 필요가 있지만, 국내의 사고 데이터는 효과적인 사고분석 연구를 수행하기에는 한계가 있다. 이에 본 연구는 미국의 사고 데이터를 살펴보고 기계학습 알고리즘을 활용하여 고령운전자 사고심각도 예측 모형을 개발하고, 주요 사고 영향요인을 도출하여, 향후 국내 사고 데이터의 보완 방향을 제시하고자 한다. 분석 결과에 따르면, 주행속도, 제한속도, 사고 시 근접 주행 여부 등이 고령운전자 사고 심각도에 영향을 주는 요인으로 나타났는데, 한국의 사고 데이터에서 제공하지 않는 것으로 나타났다. 그러므로 이와 같은 정보들이 한국의 사고 데이터에서 제공된다면 고령운전자 교통안전성 제고에 기여할 수 있을 것이다.

핵심어 : 기계학습, 교통사고 데이터, 고령 운전자 교통사고

ABSTRACT

Currently, in our society with a substantial and increasing fraction of the elderly population, transport safety for elderly drivers is becoming the center of attention. However, deficient data on vehicle crashes in South Korea limits the growth of traffic accident research pertaining to the country. So, we complemented South Korean vehicle crash data by examining USA vehicle crash data, especially the data of Ohio State, and analyzing the influential factors of elderly driver-involved crashes of the State. Subsequently, we suggested a way of improving the South Korean dataset. Notably, our study showed that the influential factors were vehicle speed, posted speed, and following other vehicles too close and provided them in the South Korean dataset.

Key words : Machine learning, Crash data. Elderly driver-involved crash

Ⅰ. 서 론

1. 연구의 배경 및 목적

우리나라는 초고령 사회로 진입하고 있다. 통계청은 장래인구추계 분석을 통해 65세 이상 인구 비율이 2020년도 15.7%에서 2070년도 46.4%에 달할 것으로 예상하고 있다. 이에 고령인구 중심의 교통정책의 필요성이 강조되고 있고. 2005년 교통약자이동편의증진법 제정 이후, 고령 보행자의 이동권 및 안전을 보장하기 위한 노력이 지속되어 왔다(Yang, et al., 2021).

하지만, 고령자의 교통 안전성은 개선되지 않고 있다. 도로교통안전공단에 따르면 2010년 인구10만명당 교통사고 사망자 수가 약 11.1명으로 OECD 국가 중 칠레(12.1) 그리스(11.2)를 이어 3번째로 높았지만 2018년 기준 인구 10만명 당 사망자수 7.3명(6위)로 감소한 반면, 노인 교통사고 사망자 수는 2010년 1,752명(OECD 3위)에서 2018년 1,682명(OECD 3위)로 큰 변화가 없다. 또한, 2020년 전체 교통사고 사망자 3,081명 중 고령자가 1,342명으로 약 43%를 차지하고 있어 고령자 중심의 교통 안전성 제고가 시급하다고 하겠다.

고령 운전자는 비고령 운전자와 신체적 조건이나 인지 능력이 부족하므로 고령 운전자 교통사고는 일반적인 운전자의 교통사고와 다른 양상을 보인다(Yu and Choe, 2013; Lee and Lee, 2014). 하지만, 고령 운전자의 교통사고 요인을 규명하는 연구는 비고령자 혹은 전반적인 교통사고 원인을 규명하고자 하는 연구에 비해 미비한 실정이다.

고령 운전자 사고 관련 연구가 적은 이유에는 여러 가지가 있지만 사고 요인을 규명할 수 있는 데이터 부족이 가장 큰 이유라 하겠다. 교통사고분석시스템(Traffic Accident Analysis System, TAAS)에서 제공하고 있는 교통사고 데이터의 경우, 법규위반 사항과 같은 정보는 공개되어 있는 반면에, 사고 시 운전자의 행동이나 사고 상황과 같은 정보는 제공하지 않는다. 하지만, 국내에서 존재하지 않거나 공개하지 않는 데이터 중 어떤 데이터들이 왜 추가로 필요한 지에 대한 연구나 정책적 제안은 전무하다. 곧, 고령 운전자 사고 및 사고 심각도에 영향을 미칠 수 있는 주요 요인들을 규명하고 국내에서 제공하고 있지 않는 정보를 찾아내어 해당 데이터 조사, 수집 및 제공을 제안하는 연구가 필요하다.

한편, 미국(오하이오주)의 교통사고 데이터의 경우 사고 전 운전자 행동(Precrash action), 사고 상황(Contribution circumstance)와 같은 변수를 통해 사고 시 운전자의 행동과 사고 상황 등과 데이터가 상세하게 제공된다. 또한 미국도 한국과 유사하게 초고령 사회로 진입하고 있어, 2020년 기준 65세이상 인구비율이 17.03%(미국), 18.13%(오하이오주)로 대한민국보다 대략 2~3% 더 고령인구 비율이 높다. IIHS(Insurance Institute for Highway Safety)에 따르면 미국의 고령 운전자가 비고령 운전자에 비해 사고 시 사망할 확률이 약 5배 정도 높아 고령 운전자 교통사고를 줄이기 위한 노력이 지속되고 있다. 그러므로 본 연구는 한국과 유사한 상황의 미국 오하이오 주의 교통사고 데이터를 활용하여 고령 운전자의 사고심각도 요인을 분석하고 이를 기반으로 국내에서 추가적으로 필요한 교통사고 데이터를 규명하여 국내 고령자 교통 안전성 제고를 위한 정책적 시사점을 제시하고자 한다.

2. 연구의 방법 및 구성

본 연구의 핵심은 미국 오하이오주의 고령운전자 사고 심각도 주요 요인을 규명하여 국내에서 제공하는 사고 데이터와 비교하고자 하는 것이다. 이를 위해 첫째, 문헌 검토를 통해 고령 운전자 사고의 특성과 기존 연구 방법론을 고찰하고, 둘째, 연구의 착안점을 도출한다. 셋째, 미국 오하이오주 사고 데이터를 수집하여

고령 운전자 교통사고 심각도 모형을 개발한다. 넷째, 모형을 기반으로 오하이오주의 사고심각도 주요 영향요인을 분석하고 국내의 교통사고 데이터와 비교한다. 다섯째, 비교분석 결과를 기반으로 국내에 필요한 교통사고 데이터 항목을 제시하면 관련 정책적 시사점을 제안한다.

Ⅱ. 이론적 고찰 및 선행연구

1. 고령 운전자 특성

대부분의 연구는 고령 운전자 교통사고 심각도가 비고령 운전자 교통사고에 비해 높다고 결론 내리고 있다(Preusser et al., 1998; Zhang et al., 2000; Boufous, 2008). 고령 운전자의 교통사고 심각도가 높은 요인으로 가장 유력한 이론은 고령 운전자는 신체 능력이 저하되고 이에 심리적으로도 위축되어 주행 중 여러 상황에 대한 대응이 어렵고 늦어 사고의 규모를 키운다는 것이다(Oh et al., 2015; Lee and Gim, 2019).

신체 능력 저하 및 심리적인 위축 현상은 운전자의 인지 반응 속도에 결정적인 영향을 미친다. 예를 들어, 기존 연구자들이 복잡하고 혼잡한 교차로에서 고령운전자의 운전행태를 연구한 결과, 고령 운전자는 복잡한 교차로에서 빠른 반응시간과 복합적인 사고를 요구하는(좌회전) 운전 조작 미숙으로 교통사고를 일으킬 가능성이 높은 것으로 분석되었다(Preusser et al., 1998; Zhang et al., 2000; Boufous, 2008; Hakamies-Blomqvist and Henriksson, 1999; Lee, 2006). 또한, Lee et al.(2009)은 60대 운전자가 20대 운전자에 비해 이중과업 처리 능력이 떨어지고 주행속도를 낮추는 것을 보였다.

2. 고령운전자 교통사고 심각도 연구

고령운전자의 교통사고 심각도 요인을 분석한 연구는 비교적 적은 상황이다. Oh et al.(2015)는 2012~14년 사이의 65세 이상 고령운전자가 제1당사자인 교통사고 42,124건에 대한 분석을 수행하였고, 측면충돌사고에서 운전자 상해 정도가 가장 높고 보행자 추돌사고에서 가장 인명피해가 많이 난다는 사실을 확인하였다. 또한 교통사고 인적요인으로 운전자 연령이 높아질수록 향후 발생상황을 예측하지 못하여 발생하는 추정 미숙의 비율이 증가하는 것으로 나타났다.

Jang et al.(2017)는 이항로지스틱 모형을 사용하여 고령운전자 사고 심각도의 영향요인을 분석하였다. 그 결과, 음주운전일 경우, 일반도시일수록, 도로가 곡선일수록, 연령이 높을수록 사고심각도가 증가하는 것으로 나타났다.

해외의 경우, 비록 고령 운전자를 대상으로 하지는 않았지만, Zhang et al.(2000)이 여러 가지 기계학습 알고리즘을 활용하여 사고심각도 예측 모형을 개발하였다. 또한, Mafi et al.(2018)와 Al Mamlook et al.(2020)이 고령 운전자 교통사고 데이터를 대상으로 다양한 기계학습 알고리즘을 학습시켰으며 랜덤 포레스트가 최적 모형으로 분석되었다. 또한 주요 영향요인으로 나이, 차량 연식, 교통량 등을 꼽았다.

3. 연구의 착안점

기존의 국내 연구들은 고령 운전자의 인적요인(낮은 반응성) 및 환경적 요인(교차로 등 교통환경) 측면에서 교통사고 발생 빈도를 고려한 영향 요인을 규명하였다. 한편 해외에서는 기계학습 알고리즘을 이용한 다

양한 실증 분석들이 수행되었다. 또한 국내의 경우 데이터의 부족으로 고령운전자 사고 심각도 영향요인을 실증 분석하는 연구가 미진하였다. 그러므로 본 연구는 국내 데이터에 존재하지 않는 여러 교통사고 정보들이 포함된 미국 데이터를 이용하여 고령 운전자 사고심각도 모형을 개발하고자 한다.

한편, 개별 사고의 사고 심각도는 경상, 부상, 중상, 사망 등과 같이 순서형인데 반해 많은 국내 연구들은 특정한 근거나 정당화 과정 없이 순서형 혹은 비순서형(사망/비사망 등) 종속변수를 사용하였다. 이에 본 연구는 순서형 종속변수를 사용한 순서형 모형과 비순서형 모형의 성능을 비교하여 최적 모형을 선택하여 사용하고자 하는 분석의 틀을 제시하고자 한다.

종합해보면, 본 연구는 미국의 사고 데이터를 기반으로 순서형 종속변수와 비순서형 종속변수를 활용한 머신러닝 알고리즘을 활용하여 최적의 고령 운전자 사고심각도 모형을 개발하고 고령 운전자 사고심각도 영향 요인을 조사하여 고령 운전자의 교통안전성을 제고시킬 수 있는 시사점을 제공할 수 있다는 점에서 연구의 기여도가 있다.

Ⅲ. 연구 방법

1. 분석자료 구축 및 주요 변수

본 연구에 사용된 자료는 ODPS에서 제공하는(https://ohtrafficdata.dps.ohio.gov/CrashStatistics/Home에서 요청 후 다운로드 받을 수 있다) 2015-2019년도의 교통사고 데이터를 가공한 것이다. 오하이오 주의 데이터에는 교통사고의 특성, 차량 특성, 사고 상황, 운전자, 동승자 특성 등이 상세히 기록되어 있다. 본 연구에서는 원 데이터에서 65세 이상의 고령 운전자가 발생시킨 사고만을 추출하였다. 그 과정에서 1) 100세 이상 운전자, 2) 정확한 사고 원인을 파악하기 어려운 경우 3) 상업용 차량의 사고 등이 제외되었다. 그 결과 104,486개의 사고 데이터가 분석에 사용되었다.

대부분의 미국의 교통사고 심각도는 치명(Fatal), 중상(Incapacitating injury), 경상(Minor injury), 부상가능(Possible injury), 상해없음(No injury)로 구성되어 있는데, 분석의 용이성을 위해 사고심각도를 크게 3가지 Fatal(치명, 중상), Injury(경상, 부상가능), PDO(상해없음, Property Damage Only)으로 분류하였다.

분석에 사용된 변수는 다음 <Table 1>와 같다. 각 변수의 세부 내용은 참고자료 <Table A>에 있다. 미국 오하이오 주의 사고 데이터가 국내 사고 데이터와 구별되는 변수는 차량속도(Unit speed), 제한속도(Posted Speed), 사고 상황(Contributing Circumstance), 사고 전 운전자 행동(Precrash action)등이 있고, 해당 변수들의 고령 운전자 사고심각도에 미치는 영향을 분석하는 것이 본 연구의 중요한 목적 중의 하나이다.

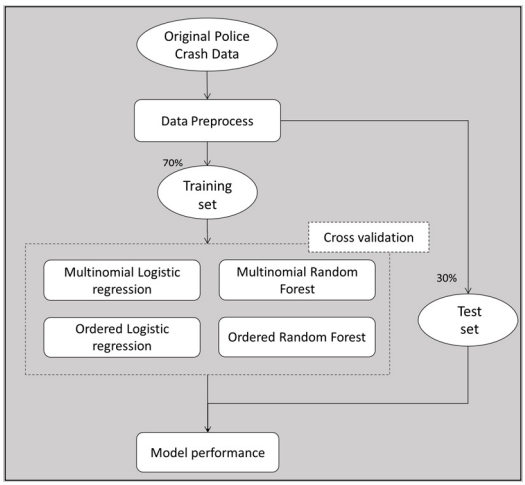
<Table 1> Variable availability of the Ohio crash dataset in the korea dataset

Availability	Variables
Available variable in the Korea dataset	Older driver's severity, Driver age, Number of units Weather, Light condition, Crash action, Manner of collision, Speed related, Semitruck related, Small truck related, Gender, Alcohol related, Drug related, Youth related, Teen related, DUI21 related, Commercial related, Location road type, Intersection or approach related, Within interchange area, Road condition, Road surface, Unit type, Number of occupants,
Unavailable variable in the Korea dataset	Unit Speed, Posted Speed, School zone related, Work zone related, Precrash action, Animal related, Motorcycle related, Contributing circumstance, Roadway divided, Road contour

2. 분석 방법

1) 분석의 틀

본 연구의 분석과정은 전통적인 기계학습(Machine Learning) 중 지도학습(Supervised Learning) 순서를 따른다.<Fig. 1> 먼저 원본 데이터를 전처리 과정을 통해 정리하고 비복원 랜덤 추출법을 이용하여 훈련데이터와 검증데이터로 분할한다.<Table 2> 4가지 기계학습 모형을 10-fold Cross-Validation법을 활용하여 훈련 데이터에 학습시키고 학습시킨 모형을 검증데이터에 적용하여 모형의 성능을 평가한다. 또한 최적 모형에서 선정된 상위 주요 영향요인을 국내 데이터와 비교하여 국내 도입이 필요한 데이터를 규명하여 국내 고령운전자를 위한 정책적 시사점을 도출하고자 한다.



<Fig. 1> Analytical Framework

<Table 2> Splitting Dataset

Fatality	Training set	Test set	Total dataset
Fatal	1,182	555	1,737 (1.7%)
Injury	10,505	4,406	14,911 (14.3%)
PDO	61,687	26,151	87,838 (84.1%)
Total	73,374	31,112	104,486 (100%)

2) 기계학습 모형

데이터 분석에 활용할 기계학습 모형은 크게 4가지이다. 첫 번째 알고리즘은 다항 로짓모형으로 전통적인 이산모형으로 종속변수가 둘 이상의 명목형 이산변수일 때, 독립변수와 종속변수의 영향관계를 분석하기 위해 사용된다. (McFadden, 1973; Ben-Akiva and Lerman, 1985) 두 번째 알고리즘은 순서형 로짓모형으로 종속변수를 순서형으로 해석하여 종속변수가 한단계 변할 때 독립변수가 미치는 영향을 분석한다(McCullagh, 1980). 순서형 로짓모형은 종속변수의 변화에 미치는 독립변수의 영향이 일정한 평행성 가정을 전제로 모형 내 독립변수의 계수가 한 개인 반면, 다항 로짓모형은 각 종속변수별 독립변수의 계수가 독립적으로 추정되기 때문에 각 사고심각도의 영향요인을 각각 분석할 수 있다는 장점이 있다.

세 번째 알고리즘은 랜덤 포레스트 모형으로 Breiman(2001)에 의해 개발·발전되었다. 랜덤 포레스트는 앙상블 기법 중의 하나로 Bootstrapping과 Bagging을 결합한 형태로 데이터를 반복적으로 랜덤 샘플링하여 여러 의사결정나무모형을 형성한 후, 다수의 의사결정나무모형의 예측값에 기반하여 관측치의 값을 예측하는 알고리즘이다. 다항로짓모형이나 순서형로짓모형과 같은 parametric 모형은 독립변수와 종속변수(종속변수의 오즈비) 간에 선형관계를 가정하지만, 랜덤 포레스트 모형은 이러한 가정에서 자유로운 특성이 있다. 랜덤 포레스트 모형에 대한 설명은 Breiman(2001)에 상세히 기술되어 있다. 네 번째 알고리즘은 순서형 랜덤포레스트이다. 순서형 랜덤포레스트는 순서형 로짓모형과 마찬가지로 종속변수가 순서형임을 전제로 데이터를

학습한다. 순서형 랜덤포레스트 모형의 구조는 순서형인 종속 변수를 다수의 이산형 변수로 변환하여 각 이산형 변수를 종속 변수로 설정하여 랜덤포레스트 모형을 개발하여 각 이산형 범주에 속할 누적(중상/사망보다 경미할) 확률을 계산한다. 최종적으로 각 누적 확률의 차이가 모형이 예측하는 사고 심각도에 속할 확률이 된다. 순서형 랜덤포레스트 알고리즘의 상세한 내용은 Lechner and Okasa(2022)을 참고하면 된다.

고령 운전자 사고 심각도의 주요 요인 분석을 위해서 다항 로짓모형과 순서형 로짓모형은 독립변수의 계수를 확인하여야 하는 반면, 랜덤 포레스트 계열의 모형은 다수의 의사결정나무에 기반하므로 애초에 계수가 존재하지 않으며, 종속변수 예측에 영향을 주는 정도(Mean Decrease Gini)로 변수의 영향력을 확인할 수 있다.

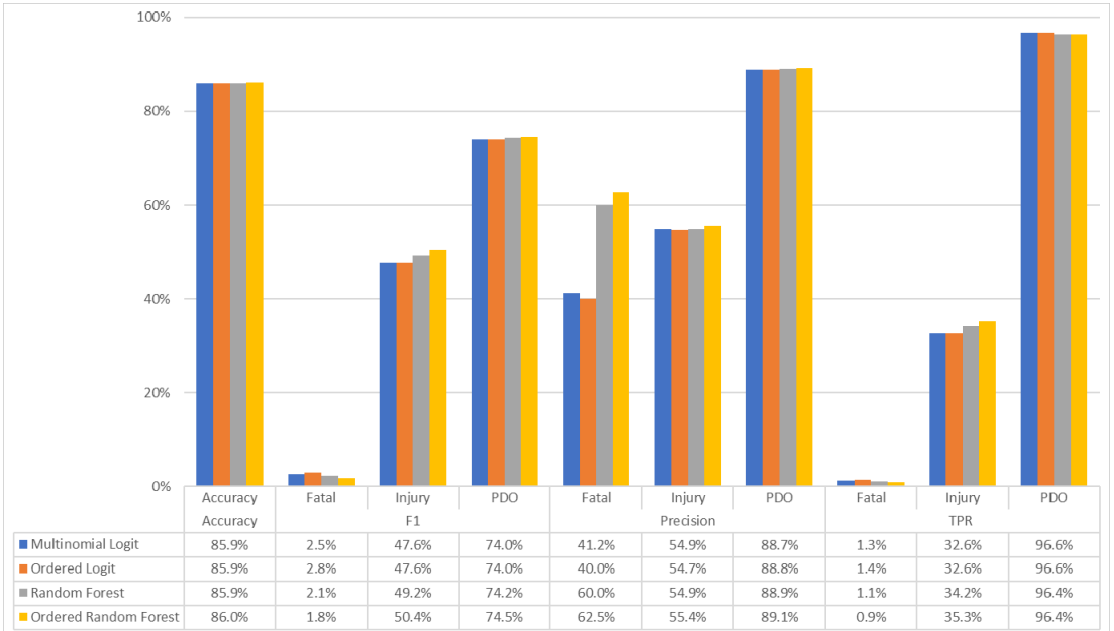
서두에서 언급했듯이 데이터 분석에 위 4가지 알고리즘을 쓰는 이유는 사고 심각도 데이터를 분석하는 데 있어, 비/순서형을 고려했을 때와 non-/parametric을 전제로 했을 때의 모형의 예측 성능 평가를 통해 어떠한 모형이 가장 효과적인지, 고령 운전자 사고심각도에 미치는 주요 영향요인이 다르게 인식되는지를 살펴보기 위해서이다.

Ⅳ. 실증 분석

1. 모형의 예측력 평가

훈련데이터를 기반으로 데이터를 학습시킨 모형을 검증데이터에 적용하여 모형의 성능을 평가하였다. 모형의 예측력을 평가하기 위하여 오분류표(confusion matrix)를 기반으로 한 예측 성능지표(Accuarcy, Precision, True Positive rate, F1)를 이용하였다.

Accuracy를 기반으로 보면 4개 모형의 전반적인 예측력은 85% 내외로 크게 차이가 없는 것으로 보인다. 각 사고 심각도 별 F1 및 TPR도 유사한 경향을 보인다.<Fig. 2> 그러나 각 사고심각도의 예측 정확도를 나타



	Accuracy	Fatal	Injury	PDO	Fatal	Injury	PDO	Fatal	Injury	PDO
	Accuracy	F1			Precision			TPR		
Multinomial Logit	85.9%	2.5%	47.6%	74.0%	41.2%	54.9%	88.7%	1.3%	32.6%	96.6%
Ordered Logit	85.9%	2.8%	47.6%	74.0%	40.0%	54.7%	88.8%	1.4%	32.6%	96.6%
Random Forest	85.9%	2.1%	49.2%	74.2%	60.0%	54.9%	88.9%	1.1%	34.2%	96.4%
Ordered Random Forest	86.0%	1.8%	50.4%	74.5%	62.5%	55.4%	89.1%	0.9%	35.3%	96.4%

<Fig. 2> Predictive performance

내는 Precision은 모형별로 차이를 보인다. 즉 로짓모형계열은 약 40%의 Fatal 사고심각도 예측 정확도를 보이는 반면, 랜덤 포레스트 계열은 대략 60% 정도의 예측 정확도를 보이고 있다. 이는 Fatal 사고 수가 다른 사고 수에 비해 현저히 적기 때문에 parametric 기반의 통계 모형이 Fatal 사고를 예측하는 데 있어 한계를 보이는 것으로 해석할 수 있다. 반면에 랜덤 포레스트 계열 모형의 경우, 통계이론에 기반을 두지 않아 이러한 제약에서 비교적 자유로워 비록 케이스 수가 적더라도 일정 수준의 예측력을 유지 한다고 볼 수 있다.

Fatal 사고의 예측 정확도는 최적모형을 선정함에 있어 중요하다. 고령자의 경우 중상 이상의 부상 정도에서 사망으로 이어지는 경우가 많기 때문이다. 그러므로 본 논문에서는 다른 예측 성능지표의 차이가 크지 않은 가운데, Fatal의 Precision 값이 가장 큰(62%) 순서형 랜덤 포레스트를 최적 모형으로 선정하고자 한다.

2. 영향요인 분석

고령운전자 사고심각도에 미치는 영향요인을 살펴보기 위해 각 모형에서 상위에 위치하는 영향요인을 추출하여 다음 <Table 3>과 같이 도시하였다.

먼저 4개의 모형에서 공통적으로 보여지는 영향요인으로 에어백 작동 여부, 날씨, 안전장비 유무, 차량 속도, 제한 속도, 사고 전 상황, 성별, 충돌 유형 등이 있었다. 모형 간 영향요인의 추세를 보면 다항로짓모형이 다른 3개의 모형에 비해 각 사고 심각도 수준별로 최상위 영향요인이 다르게 도출되는 것을 확인할 수 있었다. 예를 들면, 에어백 작동 여부는 다항로짓모형을 제외한 모든 모형에서 가장 큰 영향요인으로 분석된 반면, 다항로짓모형에서는 6순위에 위치하였다.

최적 모형인 순서형 랜덤 포레스트를 중심으로 고령운전자 사고 심각도 영향요인을 살펴보면 에어백 작동 여부, 주행 속도/제한속도, 사고 전 상황(앞차량에 과하게 근접하여 주행했는지 등), 성별, 정면/측면/진행방향 충돌, 도로의 유형, 사고 차량 수 등이 사고 심각도에 큰 영향을 미치는 것으로 나타났다. 이러한 주요 영향 요인들 중, 상당 수의 변수(에어백 작동 여부, 주행 속도/제한 속도, 근접 주행 여부)들이 국내 TAAS 데이터에서 제공되지 않고 있는 실정이다. 또한, 주행속도, 제한속도와 같은 변수들은 비고령 운전자 사고심각도 요인이기도 하다. 그러므로 만약 주행속도, 근접 주행 여부와 같은 사고 상황과 같은 데이터가 국내에서 제공이 될 수 있다면 국내 교통 사고심각도 모형개발에 기여하고 보다 효과적인 교통 안전 정책 수립에 기여 할 것이다.

<Table 3> Influential factors of elderly driver-involved crashes

Model		Common factor	Uncommon factor
Multinomial logit	Injury	Airbag usage, Unit Speed, Posted Speed, Weather, Manner of crash, Gender, Contributing Circumstance, Safety equipment usage(seat belt)	Road condition, Precrash action, Type of unit, Youth related, Distracted driving-related
	Fatal		Road condition
Ordered logit			Precrash action, Age
Random forest			Roadway divided, road type
Ordered Random forest			Number of units, road type

Ⅴ. 결 론

본 연구는 미국 오하이오 주의 사고 데이터를 수집하고 기계학습 알고리즘을 활용하여 고령운전자 교통

사고 심각도 예측 모형을 개발하였다. 또한 국내 TAAS에서 제공하는 사고 데이터와의 비교를 통해 국내 교통안전성 제고를 위한 정책적 시사점을 제안하고자 하였다.

첫째, 4가지 알고리즘을 이용하여 교통사고 데이터를 학습한 결과, 순서형 랜덤 포레스트 모형이 중상 이상의 심각한 사고를 예측하는 능력이 뛰어난 것으로 분석되었다. 이는 중상 이상의 사고 심각도인 교통 사고 수가 비교적 적기 때문에 전통적인 통계 모형인 로짓 모형이 해당 사고 심각도를 잘 예측할 수 없기 때문인 것으로 보인다. 반면에 랜덤 포레스트 계열의 알고리즘들은 앙상블 기법으로 샘플링 기반의 기계학습 알고리즘이기 때문에 비교적 해당 이슈에서 자유롭다고 볼 수 있다.

둘째, 사고 특성상 종속변수가 순서형임을 고려한 모형이 그렇지 않은 모형에 비해서 예측 성능이 그리 뛰어나지 않은 것으로 분석되었다. 이 것도 또한 데이터의 불균형 문제에서 비롯된 것으로 추정된다. 곧, Fatal 사고가 약 1.7% 밖에 되지 않기 때문에 다항로짓모형의 경우, Fatal을 예측함에 있어 한계가 있고, 순서형 로짓모형의 경우(독립 변수 당 계수가 1개 이므로) 큰 Injury 사고 비중(약 14.3%)에 맞춰 계수가 추정되는 것이다. 결국, 비/순서형 모형 모두 PDO와 Injury를 분류하기 위하여 데이터를 학습하게 되었다고 볼 수 있다. 이러한 데이터의 불균형 문제를 해결하기 위해서 Mafi et al.(2018)이 비용 민감도(Cost-sensitive) 기반의 학습 알고리즘을 제안하였는데 이는 향후 연구과제로 남겨둔다.

셋째, 최적 모형(순서형 랜덤 포레스트)을 기반으로 영향요인을 분석한 결과, 국내 TAAS 데이터에서 공개/제공하지 않는 일부 변수들이 주요 고령 운전자 사고 요인으로 나타났다(<Table 4> 참고). 특히, 에어백 작동 여부, 주행 속도/제한 속도, 사고 상황(근접 주행 여부)과 같은 정보들이 국내에서 제공되고 고령운전자 사고심각도에 미치는 영향들이 분석될 수 있다면, 교통안전성 제고에 기여할 수 있을 것이다. 사고상황, 곧 교통사고 시 주변 상황이나 요건에 대한 정보(Contributing Circumstance)를 좀 더 상세히 살펴보면(<Table A> 참고) 부적절한 회전(Improper Turn), 부적절한 차선 변경(Improper Lane Change) 등과 같이 조사자(경찰관)나 사고 당사자의 주관이 개입되는 경우가 많다. 이러한 한계점을 보완하고 해당 데이터를 수집하기 위해, 국내의 경우 먼저 해당 상황을 정의하는 기준을 마련하고 도로 상의 CCTV, 블랙박스 등을 활용하여 사고 상황을 파악할 수 있는 조사 방법이 정립되어야 할 것이다.

<Table 4> Influential factors of elderly driver-involved crashes NOT available in the TAAS data

Influential factors **unavailable** in the TAAS data	Influential factors **available** in the TAAS data
Airbag usage, Unit Speed, Posted Speed, Contributing Circumstance, Safety equipment usage(seat belt)	Weather, Manner of crash, Gender, Number of units, Road type

본 연구의 한계점은 다음과 같다. 먼저, 미국 오하이오 주의 도로 환경 및 법제도 등이 국내와 다르기 때문에 미국 고령자 사고심각도 주요 요인들을 일반화하여 국내에 적용할 수 있는지에 대한 논리적 근거가 미흡하다는 것이다. 이는 향후 미국의 다른 주 및 다양한 국가의 사고 요인 분석을 통해 어느정도 해결할 수 있을 것으로 보인다.

또한, 본 연구에서 사용된 사고 데이터의 불균형 문제로 인한 일부 사고심각도 예측 능력이 저하 현상이 있다는 것이다. 이는 대부분의 사고 데이터에서 나타나고 있는 문제로 향후 가중 회귀분석, 비용민감도 분석(Cost-sensitive analysis), Synthetic Majority Oversampling TEchnique(SMOTE) 기법 등을 이용하여 일부 극복할 수 있을 것으로 판단된다.

REFERENCES

Al Mamlook, R. E., Abdulhameed, T. Z., Hasan, R., Al-Shaikhli, H. I., Mohammed, I. and Tabatabai, S.(2020), "Utilizing Machine Learning Models to Predict the Car Crash Injury Severity among Elderly Drivers", *2020 IEEE International Conference on Electro Information Technology(EIT)*, pp.105-111.

Ben-Akiva, M. E. and Lerman, S. R.(1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA: MIT Press.

Boufous, S., Finch, C., Hayen, A. and Williamson, A.(2008), "The impact of environmental vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash", *Journal of Safety Research*, vol. 39, no. 1, pp.65-72.

Breiman, L.(2001), "Random forests", *Machine Learning*, vol. 45, pp.5-32.

Hakamies-Blomqvist, L. and Henriksson, P.(1999), "Cohort effect in older drivers' accident type distribution: Are older drivers as old they used to be?", *Transportation Research Part F*, vol. 2, no. 3, pp.131-138.

Jang, J., Choi, J. and Gim, T.(2017), "Analyzing Driving Environment Effects on Severity of Elderly Driver's Traffic Accidents", *Journal of Transport Research*, vol. 24, no. 1, pp.79-94.

Korea Road Traffic Authority Traffic Science Institute(2015), *A Study on the Major factor of High-risk Driver Groups' Accidents: Focusing on Elderly Drivers*, pp.1-85.

Lechner, M. and Okasa, G.(2019), *Random Forest Estimation of the Ordered Choice Model*, arXiv preprint (2022) Available online: https://arxiv.org/pdf/1907.02436.pdf, 2022.09.06.

Lee, J. and Gim, T.(2019), "Examining the Characteristics of Traffic Accidents Involving Elderly", *The Korea Spatial Planning Review*, vol. 102, pp.19-34.

Lee, M. J. and Lee, M. S.(2014), "Elderly Driver's Perceived Driving Ability and Driving Behavior Associated with Traffic Accident Risk", *Crisisonomy*, vol. 10, no. 12, pp.279-304.

Lee, S. C.(2006), "Psychological effects on elderly driver's traffic accidents", *Korean Journal of Psychological and Social Issues*, vol. 12, no. 5, pp.149-167.

Lee, Y. T., Kim, M. H. and Son, J. W.(2009), "The impact of cognitive workload on older driver's behavior", *The Korean Society of Automotive Engineers(KSAE)*, pp.982-987.

Mafi, S., AbdelRazig, Y. and Doczy, R.(2018), "Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups," *Transportation Research Record*, vol. 2672, no. 38, pp.171-183.

Mccullagh, P.(1980), "Regression Models for Ordinal Data", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, pp.109-127.

McFadden, D.(1973), *Conditional logit analysis of qualitative choice behavior*, University of California at Berkeley.

Oh, J. S., Lee, E. Y., Ryu, J. B. and Lee, W. Y.(2015), "An Analysis for Main Vulnerable Situations and Human Errors of Elderly Drivers' Traffic Accidents", *Journal of Transport Research*, vol. 22, no. 4, pp.57-75.

Ohio Department of Public Safety(ODPS) Crash Statistics System(2021), Available online: https://ohtrafficdata.dps.ohio.gov/CrashStatistics/Home, 2021.02.17.

Preusser, D. F., Williams, A. F., Ferguson, S. A., Ulmer, R. G. and Weinstein H. B.(1998), “Fatal crash risk for older drivers at intersections”, *Accident Analysis and Prevention*, vol. 30, no. 2, pp.151-159.

Yang, K. S., Kwon, S. M. and Youn, C. W.(2021), “An Analysis of the Determinants of Elderly Pedestrian Fatal Crash,” *Journal of the Korean Urban Management Association*, vol. 34, no. 1, pp.21-33.

Yu, J. H. and Choe, G. I.(2013), “A Comparative Analysis on Characteristics between Elderly Drivers and Younger Drivers by Accident Types: With Commercial Vehicles”, *Transportation Technology and Policy*, vol. 10, no. 5, pp.11-25.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G. and Mao, Y.(2000), “Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario”, *Accident Analysis and Prevention*, vol. 32, no. 1, pp.117-125.

<Appendix>

<Table A> A variable list of the Ohio crash dataset

Variables	Type of Factor	Label
Older driver's severity	Nominal (Ordinal)	Fatal
		Injury
		PDO
Driver age	Numeric	Driver age
Number of occupants	Numeric	Number of occupants
Number of units	Numeric	Number of vehicles involved in a crash
Weather	Nominal	Blowing Sand; Soil; Dirt; Snow
		Clear
		Cloudy
		Fog; Smog; Smoke
		Freezing Rain or Freezing Drizzle
		Other/Unknown
		Rain
		Severe Crosswinds
		Sleet; Hail
		Snow
Light condition	Nominal	Dark－Lighted Roadway
		Dark－Roadway Not Lighted
		Dark－Unknown Roadway Lighting
		Dawn/Dusk
		Daylight
		Other/Unknown
School zone related	Nominal	Active school zone related (1) or not (0)
Work zone related	Nominal	Work zone related (1) or not (0)
Crash action	Nominal	Both striking and struck
		Non-collision
		Non-contact
		Other/Unknown
		Striking
		Struck
Precrash action	Nominal	Backing
		Changing Lanes
		Driverless
		Entering Traffic Lane
		Leaving Traffic Lane
		Making Left Turn
		Making Right Turn
		Making U-Turn

Variables	Type of Factor	Label
		Negotiating a Curve
		Other/Unknown
		Overtaking/Passing
		Slowing or Stopped In Traffic
		Straight Ahead
Contributing circumstance	Nominal	Drove off Road
		Failure to Yield
		Following too Close / ACDA
		Improper Backing
		Improper Crossing
		Improper Lane Change
		Improper Passing
		Improper Start From a Parked Position
		Improper Turn
		Left of Center
		Load shifting/Falling/Spilling
		Lying in Roadway
		None
		Not Discernible
		Opening Door into Roadway
		Operating Defective Equipment
		Other Improper Action
		Ran Red Light
		Ran Stop Sign
		Stopped or Parked Illegally
		Swerving to Avoid
		Unsafe Speed
		Vision Obstruction
		Wrong Way
Manner of collision	Nominal	Angle
		Backing
		Head-on
		Not Collision Between Two Vehicles in Transport
		Other/Unknown
		Rear-end
		Rear-to-rear
		Sideswipe; opposite direction
		Sideswipe; same direction
Animal related	Nominal	Animal related (1) or not (0)
Motorcycle related	Nominal	Motorcycle related (1) or not (0)
Speed related	Nominal	Speed related (1) or not (0)

Variables	Type of Factor	Label
Semitruck related	Nominal	Semitruck related (1) or not (0)
Small truck related	Nominal	Small truck related (1) or not (0)
Gender	Nominal	Male (1) or female (0)
Alcohol related	Nominal	Alcohol related (1) or not (0)
Drug related	Nominal	Drug related (1) or not (0)
Youth related	Nominal	Youth related (1) or not (0)
Teen related	Nominal	Teen related (1) or not (0)
DUI21 related (If any driver's age is under 21)	Nominal	DUI21 related (1) or not (0)
Commercial related	Nominal	Commercial related (1) or not (0)
Location road type	Nominal	Highway
		Not highway
		No information
Intersection or approach related	Nominal	Intersection or Approach Related (1) or not (0)
Within interchange area	Nominal	Within Interchange Area (1) or not (0)
Roadway divided	Nominal	Roadway is Divided (1) or not (0)
Road contour	Nominal	Curve Grade
		Curve Level
		Other/Unknown
		Straight Grade
		Straight Level
Road condition	Nominal	Dry
		Ice
		Other/Unknown
		Sand; Mud; Dirt; Oil; Gravel
		Slush
		Snow
		Water (Standing; Moving)
		Wet
Road surface	Nominal	Blacktop; Bituminous; Asphalt
		Brick/Block
		Concrete
		Dirt
		Other/Unknown
		Slag; Gravel; Stone
		Unit Speed
		Posted Speed
Unit type	Nominal	Passenger Car
		Passenger Van (minivan)
		Pick up
		Sport Utility Vehicle