KIPS Trans. Softw. and Data Eng.
Vol.10, No.8 pp.301~310
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.8.301

Proposed TATI Model for Predicting the Traffic Accident Severity

Min-Ji Choo† · So-Hyun Park†† · Young-Ho Park†††

ABSTRACT

The TATI model is a Traffic Accident Text to RGB Image model, which is a methodology proposed in this paper for predicting the severity of traffic accidents. Traffic fatalities are decreasing every year, but they are among the low in the OECD members. Many studies have been conducted to reduce the death rate of traffic accidents, and among them, studies have been steadily conducted to reduce the incidence and mortality rate by predicting the severity of traffic accidents. In this regard, research has recently been active to predict the severity of traffic accidents by utilizing statistical models and deep learning models. In this paper, traffic accident dataset is converted to color images to predict the severity of traffic accidents, and this is done via CNN models. For performance comparison, we experiment that train the same data and compare the prediction results with the proposed model and other models. Through 10 experiments, we compare the accuracy and error range of four deep learning models. Experimental results show that the accuracy of the proposed model was the highest at 0.85, and the second lowest error range at 0.03 was shown to confirm the superiority of the performance.

Keywords : TATI, Color Representation, Severity Prediction, Traffic Accident

교통사고 심각 정도 예측을 위한 TATI 모델 제안

추 민 지† · 박 소 현†† · 박 영 호†††

요 약

TATI 모델이란 Traffic Accident Text to RGB Image 모델로, 교통사고 심각 정도 예측을 위한 본 논문에서 제안하는 방법론이다. 교통사고 치사율은 매년 감소하는 추세이나 OECD 회원국 중 하위권에 속해있다. 교통사고 치사율 감소를 위해 많은 연구들이 진행되었고, 그 중에서 교통사고 심각 정도를 예측하여 발생 및 치사율을 줄이기 위한 연구가 꾸준하게 진행되고 있다. 이와 관련하여 최근에는 통계 모델과 딥러닝 모델을 활용하여 교통사고 심각 정도 예측을 하는 연구가 활발하다. 본 논문에서는 교통사고 심각 정도를 예측하기 위해서 교통사고 데이터를 컬러 이미지로 변환하고, CNN 모델을 통해 이를 수행한다. 성능 비교를 위해 제안하는 모델과 다른 모델들을 같은 데이터로 학습시키고, 예측결과를 비교하는 실험을 진행했다. 10번의 실험을 통해 4개의 딥러닝 모델의 정확도와 오차 범위를 비교하였다. 실험 결과에 따르면 제안하는 TATI 모델의 정확도가 0.85로 가장 높은 정확도를 보였고, 0.03으로 두 번째로 낮은 오차 범위를 보여 성능의 우수성을 확인하였다.

키워드 : TATI, 컬러 표현, 심각 정도 예측, 교통사고

1. 서 론

교통안전은 지속 가능한 교통개발에서 항상 중요한 이슈이다[1]. 최근 첨단 운전자 보조 시스템(Advanced Driver Assistance System)과 같은 첨단 교통기술을 대부분의 자동차 제조업체에서 제공하고[2], 스마트폰 및 내비게이션 이용률이 증가하면서 지능형 교통 시스템에 대한 관심은 더욱 높아졌다. 첨단 교통기술들은 생활의 편리함 등 다양한 이점을 주었지만, 아직까지 교통사고 발생률을 감소시키는 데에는 큰 효과를 얻지 못하고 있다. 국내 교통사고 통계에 따르면, 교통사고 발생률 및 치사율은 꾸준하게 감소하고 있는 추세이다[3]. 그러나 통계청에 의하면 아직까지도 2017년 기준 교통사고로 인한 사망자 수는 OECD 회원국 35개 나라 중 32위로 평균적으로 높은 수치를 보여주고 있다. 이는 발전하는 기술에 비해 뚜렷한 효과를 얻지 못했으며, 교통사고 치사율 감소를 위한 연구의 필요성을 보여준다.

교통사고의 심각 정도를 예측한다는 것은 기존의 도로와 환경 조건에서 사고를 미연에 방지하고, 사고 규모를 미리 파악하여 더 심각한 피해를 예방할 수 있다는 점에서 큰 의미를 가진다. 그렇기 때문에 최근까지도 교통사고의 원인을 분석하고 교통사고 심각 정도를 예측하여 교통사고의 발생률을

※ 이 논문은 2021년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2016-0-00406. (기반 SW-창소씨앗 2단계)SIAT형 CCTV 클라우드 플랫폼 기술 개발).
† 준 회 원 : 숙명여자대학교 IT공학과 석사과정
†† 준 회 원 : 숙명여자대학교 빅데이터활용 연구센터 책임연구원
††† 종신회원 : 숙명여자대학교 IT공학과 교수
Manuscript Received : March 26, 2021
First Revision : May 31, 2021
Accepted : June 26, 2021
* Corresponding Author : Young-Ho Park(yhpark@sm.ac.kr)

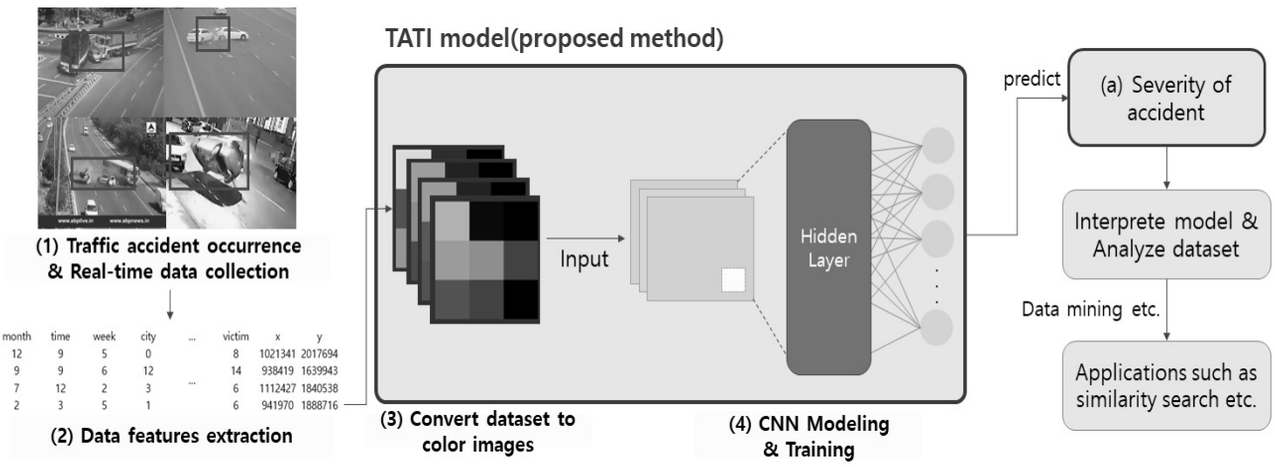

Fig. 1. Overall Diagram of Severity Estimation and its Application

줄이기 위한 연구들이 꾸준하게 이루어지고 있다[4-11].

교통사고 심각 정도 예측 연구의 초반에는 로짓 또는 프로빗 모델과 같은 통계 모델이 주로 사용되었다. 교통사고 데이터를 통계 이론을 기반으로 분석하여 독립변수들 간의 종속변수에 끼치는 영향정도를 추론하고, 교통사고 심각 정도를 예측하는 방법이다. 통계 모델은 예측할 수 있는 데이터의 종류가 한정되어 있고, 상관관계 추론의 결과에 따라 예측 결과의 차이가 크다는 단점이 있다[12]. 이후 컴퓨터 과학의 발전으로 빅데이터의 시대가 도래하면서 통계 모델의 단점을 보안하기 위한 딥러닝 모델을 적용한 연구들이 제안되었다. 이 연구들은 학습 데이터를 통해 딥러닝 모델을 학습시켜서 원하는 데이터의 결과값을 예측하는 방법이다. 딥러닝을 이용한 예측 방법은 통계 모델에 비해 다양한 형태의 데이터를 사용할 수 있는 장점을 가지고 있다. 그러나 이러한 선행 연구들은 대부분 심각 정도를 레벨별로 나누어 예측하는데, 이때 레벨을 나누는 기준이 명확하지 않거나 예측 가능한 정도가 포괄적이라는 문제가 있다. 교통사고 심각 정도와 같이 교통안전과 관련된 연구들은 생명으로 이어지는 분야이기 때문에 좀 더 정확하고 세분화된 예측이 필요할 것으로 보인다.

단순히 정제된 데이터셋을 사용하여 교통사고의 심각 정도를 예측하는 연구뿐만 아니라 실제로 사고가 발생했을 때 활용이 가능하도록 실시간 예측을 하는 연구들도 활발하게 이루어지고 있다[13-15]. 이에 본 연구는 교통사고 이미지나 영상을 이용하여 교통사고 심각 정도를 예측하고, 모델 해석 및 데이터셋 분석을 통해 유사도 검색 등의 다양한 활용이 가능한 하나의 교통안전 시스템을 연구하고자 한다. 본 논문은 교통안전 시스템을 위한 선행 연구로써, 실시간이 아닌 정제된 교통사고 데이터셋을 활용하여 교통사고의 심각 정도를 예측하는 CNN(Convolution Neural Networks) 기반의 딥러닝 모델을 제안한다. Fig. 1은 본 연구의 교통사고 심각 정도 예측 시스템 및 활용방안 및 과정을 보여주는 전체 다이어그램이다. Fig. 1의 TATI model(proposed method) 부분과 (a)는 본 논문에서 제안하는 교통사고 심각 정도 예측 모델 부분이며, (2)번 단계의 데이터 특징 추출 대신 이미 정제된 교통사고 데이터셋을 이용한다.

본 논문은 관련 연구로 조사한 통계 모델의 단점을 보완하고, 더 정확하고 세분화된 교통사고 심각 정도를 예측한다. 데이터셋을 RGB 색 기반 컬러 이미지로 변환하고, 이를 활용하여 CNN 모델 기반의 교통사고 심각 정도를 예측하는 TATI(Traffic Accident text data to RGB color Image) 모델을 제안하며 공헌하는 바는 다음과 같다.

- 교통사고 심각 정도의 판단 지표로써 전체 부상자 중 사망자와 중상자의 합의 비율을 제안한다. 구체적인 심각 정도를 예측함으로써 실제 환경에 적용할 시, 보다 효율적인 의료자원의 배치나 통제를 가능하게 하는 의미가 있다.
- TATI 모델과 기존의 딥러닝 모델의 성능 비교를 통해 제안된 방법의 우수성을 입증하였다. 또한 기존 교통사고 심각 정도 예측 연구보다 더 구체적인 심각 정도를 예측함에도 비슷한 분류 정확도를 달성하였다.

본 논문의 구성은 다음과 같다. 2장에서 통계 모델과 딥러닝 모델을 기반으로 한 교통사고 심각 정도 예측과 관련된 선행 연구를 설명한다. 3장에서는 본 논문에서 제안하는 텍스트 데이터셋을 컬러 이미지로 변환하는 과정에 대해 설명한다. 4장에서 TATI 모델의 실험 과정 및 결과를 설명하고, 다른 모델들과 비교한 결과를 보여준다. 마지막으로 5장에서 결론과 추후 연구에 대해 설명하고 본 논문을 마친다.

2. 관련 연구

교통사고 심각 정도를 예측하는 대부분의 선행 연구들은 독립변수들의 상관관계를 분석한 후, 그 결과를 바탕으로 심각 정도를 예측한다. 교통사고 데이터는 사고 시간 및 주변 환경, 운전자의 특징, 운전 방식, 차량 종류, 센서 데이터 등의 많은 요소가 포함될 수 있다. 다양한 독립변수가 결합하여

하나의 상황을 만들지만, 그 독립변수들이 모두 상황의 결과에 동등하게 기여하는 것은 아니다. 그렇기 때문에 각 독립변수들 간의 종속변수 결과에 더 중요한 영향을 끼친 변수를 파악하는 일은 정확한 예측을 위해 중요한 과정 중 하나이며[16], 이후 본 논문에서는 이 과정을 독립변수들의 상관관계 분석이라고 부르겠다. 상관관계 분석 방법에는 피어슨 상관관계, 변수 중요도 측정, p-value 등이 있고, 상관관계를 분석하여 교통사고 심각 정도를 예측하는 다양한 연구 사례가 있다. 본 장에서는 독립변수들의 상관관계를 분석하여 교통사고 심각 정도를 예측한 관련 연구들을 2.1절에서 통계 모델을 이용한 방법과 2.2절에서 딥러닝 모델을 이용한 방법으로 나누어 설명한다.

2.1 통계 모델을 이용한 방법

로짓 모델이나 순서형 프로빗 모델과 같은 회귀 모델은 독립변수의 상관관계 분석 및 예측을 위한 방법으로 널리 채택되어 왔다[4]. 회귀 모델을 이용한 선행 연구에는 Grigorios et al.은 기존의 순서형 프로빗 모델의 정확도를 개선한 zero-inflated hierarchical ordered probit 모델을 제안하였고[17], 특정 요소의 환경에 대한 차량 충돌 빈도를 예상하기 위해 Karim et al. 이 제안한 multivariate Poisson-lognormal 회귀 모델 등의 연구가 있다[5]. 위와 같은 회귀 모델들은 요소들 사이에 미리 정의된 상관관계가 존재하며, 이를 바탕으로 결과를 예측한다. 이 때문에 상관관계의 결과에 따라 예측 결과가 크게 바뀔 수 있고, 독립변수들 간에 너무 강한 상관관계 가지는 다중공성선의 단점이 있다.

이러한 회귀 모델의 상관관계 분석 결과에 대한 의존성을 낮추고, 다중공성선 문제를 해결하기 위해 트리 모델을 활용한 다양한 연구가 나오게 되었다. 심각 정도 예측을 위해 자주 사용되는 트리 모델인 CART(Classification And Regression Tree)는 미리 정의된 상관관계를 요구하지 않고, 이상치나 누락된 데이터, 다중공성선 문제를 처리할 수 있다는 장점이 있다. 또한 트리를 통해 교통사고 심각 정도 예측과 동시에 변수 중요도 측정을 통해 독립변수들의 상관관계를 분석한다[6,7]. 그러나 트리 모델만을 이용한 방법은 변수의 수가 많아질수록 처리 방식이 복잡해지고, 변수의 개수만큼 변수 중요도 수치가 분배되기 때문에 예측 결과에 큰 영향을 끼치는 변수가 숨겨질 가능성이 있다는 문제가 있다[18].

지금까지 연구돼 온 통계 모델을 기반으로 교통사고 심각 정도를 예측하는 선행 연구들은 대부분 독립요소 간의 상관관계나 수에 따라 예측 결과가 크게 좌우된다. 또한 통계 모델의 특성상 사용할 수 있는 데이터의 형식에 제한이 있다는 한계점이 존재한다. 최근에는 위의 문제들을 해결하기 위해 통계 모델과 딥러닝 모델을 결합하여 다양한 데이터셋을 활용한 연구들이 활발하게 이루어지는 추세이다.

2.2 딥러닝 모델을 이용한 방법

딥러닝 모델을 기반으로 제안하는 논문들은 MLP 네트워크(Multi-Layer Perceptron Networks), RNN(Recurrent Neural Networks), CNN 등 다양한 신경 네트워크를 기반으로 한 딥러닝 모델을 학습시켜 교통사고 심각 정도를 예측한다. Hassan et al.가 제안한 MLP 네트워크와 적응 공명 이론 신경망을 사용하여 교차로를 중심으로 운전자의 성별, 차량 종류, 속도, 벨트의 유무와 교통사고 심각 정도와의 관계를 분석하였다[8]. 심각 정도는 부상 없음과 부상 가능성을 포함한 부상 있음 두 개로 분류하여 예측하였다. Mehme et al.가 제안한 알고리즘은 유전 알고리즘과 패턴 검색 방법을 결합하여 유전 알고리즘 인공신경망을 제안하였다[9]. 고속도로 교통사고를 중심으로 심각 정도를 부상 없음, 부상, 사망 3단계로 분류하였고, 12개의 변수를 사용하였다. 또 다른 연구로는 Maher et al.가 제안하였는데, 9개의 요소를 사용하여 고속도로에서의 교통사고 심각 정도를 예측하는 RNN 모델을 제안하였다[10]. 이외에도 교통사고 데이터셋을 이미지로 변환하여 교통사고 심각 정도를 예측하는 연구도 제안되었다. Zheng et al.이 제안하는 CNN 기반의 TASP-CNN은 12개의 요소를 회색 이미지로 표현하여 교통사고 심각 정도를 미약, 심각, 사망 3단계로 예측하였다[11]. 딥러닝을 기반으로 하는 교통사고 심각 정도 예측 연구들은 대부분 60%~ 75%로 통계 모델과 비교하였을 때 낮은 정확도를 보였다. 또한 심각 정도를 2단계 혹은 3단계로 나누는 방법은 단계를 나눈 기준이 정확하지 않거나 예측할 수 있는 정도가 포괄적이라는 한계점이 있다.

이에 본 논문에서는 독립요소의 상관관계를 결합한 데이터셋을 컬러 이미지로 표현하고, 그 결과를 CNN 모델에 적용하여 심각 정도를 세부적으로 예측하는 TATI 모델을 제안한다. 분석한 상관관계 결과를 직접 모델에 적용하지 않고 이미지 데이터셋에 결합함으로써 상관관계에 대한 의존도를 낮추는 효과를 기대할 수 있고, 심각 정도를 사망자와 중상자의 비율을 예측하여 심각 정도를 세분화한다.

3. Traffic Accident Text Data to RGB Color Image Model

본 장에서는 TATI 모델의 과정을 설명한다. 텍스트 데이터를 컬러 이미지로 변환하기 위한 과정을 설명한다. Fig. 2는 TATI 모델이 작동하는 과정을 단계별로 나타낸 전체 흐름도이다. 3.1절에서는 본 논문에서 사용하는 교통사고 데이터셋에 대한 설명을 한다. 3.2절에서는 독립요소들의 상관관계를 추론하여 가중치를 계산하는 방법에 대해 설명하고, 3.3절에서는 데이터 전처리 과정을 포함한 교통사고 데이터셋을 RGB 기반의 컬러 이미지로 변환하는 알고리즘에 대해 설명한다.

3.1 데이터셋 설명

본 연구에서 사용한 데이터셋은 도로교통공단(KoROAD)에서 제공하는 TAAS(Traffic Accident Information Distribution System)의 사망교통사고정보 데이터셋이다. 본 논문에서는 데이터 표현 방법에 대한 교통사고 심각 정도 예측 모델의 성

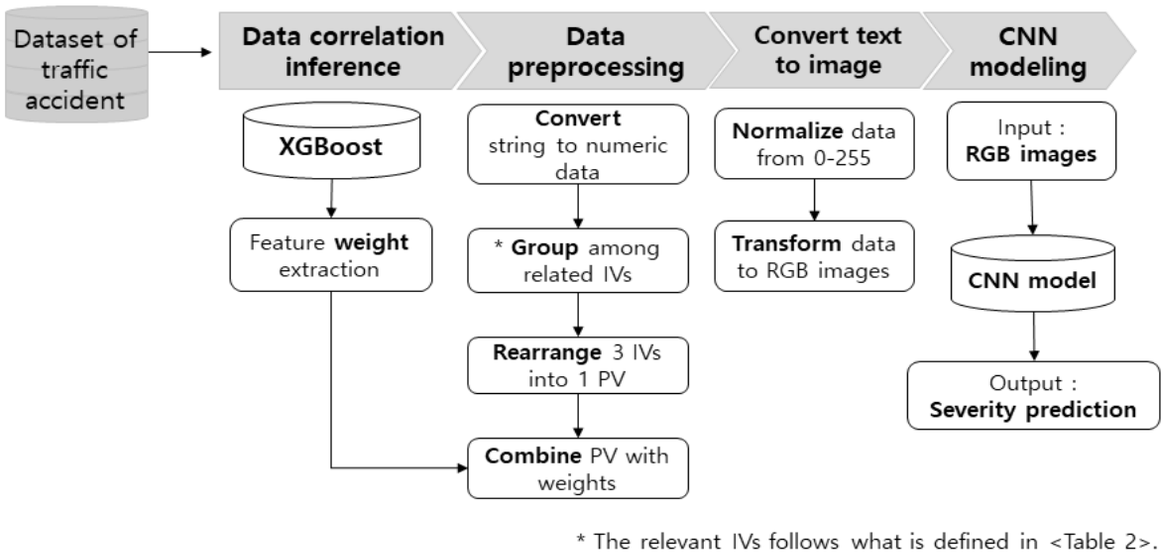

Fig. 2. Flow Chart of TATI Model

능을 중점적으로 관찰하기 위하여 공식적으로 제공되는 데이터셋을 사용하였다.

사망교통사고정보 데이터셋은 사고 발생 시간, 위치 등의 정보와 사망자, 부상자의 수 등이 포함되어 있다. 사망 기준은 교통사고 일시부터 30일 이내에 사망한 경우를 의미한다. 국내 전국에서 발생한 2012.01.01.-2019.12.29. 까지의 총 8년간의 교통사고 정보를 제공한다. 총 34,146건의 데이터가 존재하는데, 이는 선행 연구들[8-11]이 사용한 데이터셋의 개수가 30,000~35,000 건 사이로 사용하기 적합한 데이터셋이다. 23개의 항목 중 본 논문에서는 15개의 항목만을 사용한다. 교통사고 발생 월 및 요일, 발생 지역의 시 및 도 단위의 지명과 군 및 구 단위의 지명, 사고의 유형, 사고 가해자가 위반한 법률, 교통사고 발생 구역의 도로 유형, 가해자와 피해자의 차량 종류가 속해있다. 사용하지 않는 항목은 사고 연도, 주야간, 경도 및 위도 좌표이다. 주야간 항목은 시간 항목에 이미 내포되어 있기 때문에 데이터의 중복을 피하기 위해 삭제하였다. 경도와 위도 좌표는 한 위치를 특정화하는 데이터로 딥러닝 모델의 과적합을 발생시킬 가능성이 있기 때문에 사용 항목에서 제외하였고, 사고 연도와 일자 또한 어떤 시대를 특정하는 지표가 될 수 있기 때문에 제외하였다.

텍스트 형식의 데이터인 요일, 지역, 자치구, 사고 종류, 위반 법률, 도로 유형, 가해 차량, 피해 차량 요소들은 컬러 이미지로 표현하기 위해 우선 숫자 데이터로 변환하였다. 요일 데이터는 월-일 순서로 인덱스를 지정하였다. 그 외의 사고 위치, 사고 유형, 차량 유형 등의 항목들은 사전식 오름차순 배열 후, 순서대로 인덱스를 지정하였다. 시간 항목은 한국교통연구원에서 조사한 교통수단 이용 실태조사 연구(2014)에 따른 시간대별 분포 자료를 이용하여 8개의 시간대로 나누어 시간의 흐름 순서로 인덱스를 지정하였다. 데이터셋에 대한 자세한 설명 및 인덱스 생성 규칙은 Table 1에서 설명한다.

3.2 독립요소와 상관관계 분석

본 논문에서는 경사 부스팅 알고리즘을 기반으로 한 XGBoost(Extreme Gradient Boosting)의 변수 중요도 측정 방법을 이용하여 독립변수들의 상관관계를 분석하였다. 변수 중요도 측정 방법은 독립변수들이 종속변수에 얼마나 영향을 미치는지, 즉 독립 변수가 종속변수의 결과에 얼마나 중요한지를 계산하는 방법이다. XGBoost는 일반적인 경사 부스팅 알고리즘과 다르게 병렬 학습을 사용하여 학습과 분류가 빠르고, 과적합을 방지할 수 있는 장점을 가지고 있다[19,20]. 이 때문에 상관관계 분석에 널리 사용되고 있는 방법 중 하나다.

XGBoost가 변수 중요도를 측정하기 위해 우선 하나의 의사결정 트리에서 각 독립변수들로 이루어진 분할 지점마다 트리의 성능이 얼마나 올라가는지 계산한다. 이때 얼마나 많은 데이터가 해당 분할 지점에 의해서 영향을 받았는지에 따라 트리 별 가중치를 계산하여 변수 중요도를 측정한다. 하나의 의사결정 트리에서 계산된 변수 중요도를 모든 트리에 대해 계산한 후, 평균을 내어 최종적인 상관관계를 분석한다[21,22].

3.3 데이터 전처리 및 변환

3.2장에서 분석한 상관관계와 교통사고 데이터셋을 결합하여 한 건의 교통사고 데이터를 하나의 RGB 색 기반의 컬러 이미지로 변환한다. 먼저 데이터 독립변수 간의 관계성의 정의는 Definition 1과 같다.

Definition 1 (*CFV*: Correlation Feature Variables): *CFV*는 *PV*의 집합으로, 하나의 *CFV*는 한 건의 교통사고 데이터를 나타낸다. 이때 c는 *CFV*의 총 개수를 표현할 때 사용하며, 전체 데이터셋의 개수이다. 본 논문에서 *CFV*의 총 개수 c

Table 1. Data Variable Description and Code Generation Rules

Variable	Description
month	Month of traffic accident. January - December =1-12
week	The day of the week when a traffic accident occurred. monday=1, tuesday=2, wednesday=3, thursday=4, friday=5, saturday=6, sunday=7
time_zone	The day of the week when a traffic accident occurred. 05-07=1, 07-09=2, 09-12=3, 12-15=4, 15-18=5, 18-20=6, 20-22=7, 22-05=8
city	An area where traffic accidents. Sort in the ascending order, code from 1 to 17. Ex) Seoul, Jeju, Busan
borough	The region of the county and district where traffic accidents occurred. Sort in ascending order, then code from 1 to 208. Ex) GangNam-Gu, Jung-Gu, YongSan-Gu
ac_large/ ac_medium/ ac_small	Traffic accident type large, medium, small category. After sorting in ascending order, code the large category up to 1-4, the mediun category up to 1-19, and the small category up to 1-22. Ex) ac_large : vehicle vs vehicle, vehicle vs person, vehicle alone / ac_medium : collision, passing, crossing / ac_small : workpiece collision, passing through edge of road, overturning
law	A traffic-crash law. Sort in ascending order, then code from 1 to 8. Ex) Signal violation, no safe distance, midline breach
road _large/road_ small	The type of road in the traffic accident zone. After sorting in ascending order, code the large category up to 1-8, and the small category up to 1-15. Ex) road_large : general national highway, high-speed national highway / road_small : unitary, intersection, unknown
perpetrator	The type of vehicle that caused a traffic accident. Sort in ascending order, then code from 1 to 13. Ex) taxi, people, bus
victim	Type of vehicle affected by traffic accident. Sort in ascending order, then code from 1 to 16. Ex) taxi, car, bus
x	The x-coordinate of the location of the traffic accident.
y	The y-coordinate of the location of the traffic accident.

는 34,146개이다. *CFV*와 *PV*의 관계는 아래 Equation (1)과 같다.

$$CFV_{(i=1\cdots c)} = \{PV_{[1]}, \cdots, PV_{[m]}\}_i \quad (1)$$

Equation (1)에서 나타난 *PV*는 데이터셋의 독립변수인 *IV*와 그에 해당하는 가중치 *cw*를 결합한 후, 관련성이 높은 주제를 가지는 것들끼리 묶은 집합을 나타낸다. *PV*와 *IV*의 관계는 아래 Equation (2)과 같다. Equation (2)의 *m*은 *PV* 집합의 총 개수, *j*는 *IV*의 총 개수인 *k*보다 작고 0보다 큰 임의의 수를 의미한다.

$$PV_{i(i=1\cdots m)}=\{IV_{[j-2]}\times cw_{[j-2]}, \cdots, IV_{[j]}\times cw_{[j]}\} \quad (2)$$
(j : k보다 작고 0보다 큰 숫자)

1) *PV*(Parent Variables) : 하나의 *PV*는 컬러 이미지로 변환 시 하나의 픽셀로 표현되며, 본 논문은 RGB 컬러를 사용하기 때문에 하나의 *PV*에 들어갈 수 있는 최대 개수는 3개이다. 본 논문에서 *PV*는 시간, 지역, 사고 유형, 위법 종류, 도로 종류, 차량 종류, 좌표까지 총 7개로 분류하였다.

2) *IV*(Independent Variables) : 데이터의 모든 독립변수를 나타내며 $\{IV_1,\cdots,IV_k\}$로 표현한다. *IV*의 총 개수는 *k*로 표현하며, 본 논문에서 *IV*의 개수 k는 15개이다.

3) *cw*(Correlation Weight) : 각 IV의 가중치를 의미한다. $cw = \{cw_1, \cdots, cw_k\}$로 표현하며, 본 논문에서 *cw*의 개수는 *IV*와 동일한 *k*로 표기한다.

위의 Definition 1을 이용하여 본 논문에서 사용하는 교통사고 데이터셋의 *PV*, *IV*, *cw*를 정의하면 아래 (a), (b), (c)와 같다.

(a) *PV = {Time, Region, Accident Type, Law, Road Type, Car Type, Coordinate}*
(b) *IV = {month, week, time-zone, city, borough, ac_large, ac_medium, ac_small, law, road_large, road_small, perpetrator, victim, x, y}*
(c) *cw* = Weight of each *IV*

Table 2는 위에 정의된 Definition 1의 1), 2)를 이용하여 PV와 IV의 관계성을 나타낸다.

시간과 관련된 *month, week*, *time-zone* 항목은 *Time*, 사고 지역과 관련된 *city*, *borough* 항목은 *Region*에 속한다. *Accident type*은 *ac_large*, *ac_medium*, *ac_small*를 포함한다. *Road type*은 *road_large*, *road_small*을 포함하며, *Car Type*은 *perpetrator*, *victim*을 포함한다. 좌표 *x*, *y*는 *Coordinate* 항목에 함께 포함되고, *Law*은 단일 변수 law 하나만을 포함한다.

아래 알고리즘은 Definition 1과 Table 2의 내용을 바탕으로 교통사고 데이터셋을 재정렬하고, *CFV*을 생성하여 컬러 이미지로 변환하는 알고리즘이다.

Table 2. Relativity of 7 PV(Parent Variables) and 15 IV(Independent Variables)

PV	IV
Time	month
	week
	time-zone
Region	city
	borough
Accident type	ac_large
	ac_medium
	ac_small
Law	law
Road type	road_large
	road_small
Car Type	perpetrator
	victim
Coordinate	x
	y

Text to Color images based on RGB algorithm

Input: *IV*, *cw* of *k*-size

1: **Initialize** an zero array: *CFV*[*c*][*m*], *PV*[*m*][3]

2: **for** *i*=1 to *c* **do**
3: **for** *j*=1 to *m* **do**
4: $IV[j] = IV[j] \times cw[j]$ about *i*-th dataset
5: **end**
6: **end**
7: **Grouping** related topics among *IV* according to 〈Table 2〉
8: **if** the number of *IV*s grouped 〈 3
9: **Initialize** empty digits to zero
10: **Rearrange** to grouped *IV*s in *PV*

11: **for** *i*=1 to *c* **do**
12: **for** *j*=1 to *m* **do**
13: *CFV*[*i*]=Normalize range 0 to 255 of *PV*[*j*] about *i*-th dataset.
14: **end**
15: **end**
16: Convert *CFV* to RGB color images

Output: RGB images for all dataset

먼저 Definition 1의 2), 3)에 따라 전체 데이터셋에 대한 독립변수 *IV*와 가중치 *cw*를 초기값으로 선언한다. [Line 1]에서 컬러 이미지로 변환하기 전, *CFV*와 *PV* 변수를 0으로 초기화한다. 이때 이중배열 *CFV*의 크기는 Equation (1)에서 정의한 [전체 데이터셋의 개수 *c*] × [Equation (2)에서 정의한 *PV*의 개수 *k*로 정의한다. 이중배열 *PV*의 크기는 [*m*] × [3]으로 정의한다. 3으로 정의하는 이유는 이미지로 변환할 때 각 각 R, G, B 값이 필요하기 때문이다. 위함이다. [Line 2-6]에서 *IV*와 가중치 *cw*를 결합한 후, [Line 7-10]은 Table 2에서 정의한 *PV*와 *IV*의 관계성에 따라 *IV* 데이터셋을 그룹화하여 *PV* 변수에 재정렬한다. 이때 그룹화한 *IV*의 개수는 3개가 포함되는 것이 원칙이다. 하지만 예외적으로 2개 이하일 경우, 우선적으로 *IV* 값을 채운 후 빈자리를 0 값으로 대체하여 저장한다. [Line 11-15]에서 *CFV* 값을 *RGB* 색 범위인 0-255 사이의 값으로 정규화 하여 *CFV* 변수에 업데이트한다. [Line 16]에서 최종 *CFV* 값을 *RGB* 컬러 이미지로 변환한 후, 리턴하여 알고리즘을 마친다.

4. 실 험

본 장에서는 TATI 모델의 성능을 위한 실험 및 결과를 설명한다. 4.1장에서는 실험 환경 및 파라미터 세팅과 실험에 사용된 데이터셋에 대해 설명한다. 4.2장에서는 본 논문에서 진행된 실험의 결과를 보여주고 분석한다.

4.1 실험 환경 및 실험 데이터

실험은 Ubuntu 20.04.1 LTS OS, Intel Core i7-10700 CPU 2.90GHz × 16, NVIDIA TU102 [GeForce RTX 2080 Ti] GPU를 사용해서 수행하였다. 제안된 모델은 구글이 개발한 오픈소스 TensorFlow 딥러닝 프레임워크[23]를 사용하여 파이썬으로 구현되었다. 실험에 사용되는 교통사고 데이터셋은 총 34,146건을 사용한다. 이 데이터셋을 통해 모델의 성능을 비교하기 위해 70%의 훈련 데이터와 30% 테스트 데이터셋으로 나누어 실험을 진행하였다.

본 논문에서 제안하는 TATI 모델의 하이퍼 파라미터는 1000개의 epoch을 사용하며, scikit-learn으로 구현되었다. 모델의 오버 피팅을 막기 위해 patience를 10, validation loss를 검증 지표로 사용한 early stopping함수를 구성하였다. 해당 세팅 값은 10번 동안 계속해서 validation loss가 증가하면 모델의 학습을 중지시키는 의미를 가지고 있다. Table 3은 제안 모델에 설정한 최적화 알고리즘과 하이퍼 파라미터를 나타낸다. 해당 하이퍼 파라미터는 실험을 통해 가장 최적의 성능을 나타내는 값으로 설정하였으며, 실험 결과 Adam 최적화 알고리즘, 배치 크기는 128, 학습률은 0.01이 실험 모델의 가장 좋은 파라미터인 것으로 확인되었다.

4.2 실험 결과 및 분석

가중치 계산을 위해 XGBoost를 이용하여 15개 독립요소의 변수 중요도를 분석하고, 교통사고 데이터셋과 결합하기 위해 0-1 사이의 값으로 정규화하였다. Fig. 3은 변수 중요도 분석 실험 결과를 보여준다. 아래부터 위로 올라갈수록 교통사고 심각 정도와 더 관련이 없는 독립변수임을 의미한다. 실험 결과 수치가 가장 높은 변수 week와 두 번째로 높은 *perpetrator*는 다른 변수들과의 큰 차이를 보여, 교통사고 심각 정도에 매우 큰 영향을 미치는 변수임을 알 수 있다. 다음으로 *borough, ac_large, ac_medium*, *mon*, *y*, *time-*

Table 3. Hyper-parameters Setting for TATI Model

Hyper-parameter	Value
Batch size	128
Loss function	categorical cross-entropy
Optimizer	Adam
Learning rate	0.01

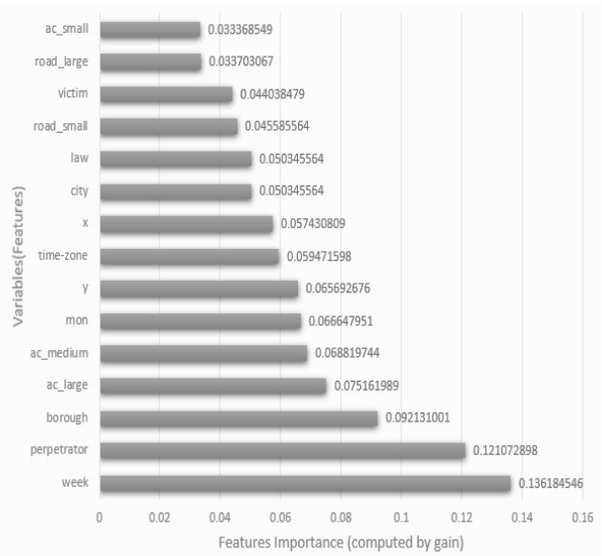

Fig. 3. Computes Feature Importance Results of 15 Independet Variables Normalizes between 0-1

zone 순으로 교통사고 심각 정도에 미친 영향이 크다. *time-zone* 요소 위부터는 변수 중요도가 0.5 이하로 떨어지는 것으로 보아 완전히 영향을 주지 않는다고 볼 수는 없지만, 교통사고 심각 정도와 관계성이 미약한 것으로 나타났다. *time-zone* 이후로는 *x*, *city*, *law*, *road_small*, *victim*, *road_large*, *ac_small* 순으로 결과가 나타났다. 본 실험의 결과를 토대로, Definition 1에서 정의한 *PI*와 결합하여 *CFI*를 생성한다.

딥러닝 모델은 여러 겹의 레이어와 모듈을 쌓아서 구성된다. 레이어가 몇 층으로 쌓이는지, 어떤 식으로 구성되는지에 따라 모델의 성능과 예측 결과가 달라진다. 따라서 각기 다른 레이어와 모듈 구조를 만들어 성능을 평가하는 것은 더 정확한 딥러닝 모델을 만들기 위한 중요한 실험이다. Table 4는 TATI 모델에 적용할 각기 다른 레이어의 구조를 표시하였고, Fig. 4는 각 구조로 학습했을 때의 테스트 데이터에 대한 정확도이다. 10번의 실험 중 가장 높은 값을 표기하였다. 깊이가 4인 Model 1은 테스트 데이터셋의 정확도가 0.8401이다. Model 1과 같은 구조에 dropout 모듈을 추가한 깊이 5인 Model 2의 정확도는 0.8646으로 Model 1보다 높은 수치를 보였다. Model 3은 Model 1의 구조에서 합성곱 레이어를 하나 추가한 모델로 0.76349의 정확도를 가진다. Model 4는 Model 3의 구조에서 dropout 모듈을 추가한

Table 4. TATI Model at Different Structures

	Structures of TATI model
Model 1	conv2d-flatten-dense-dense
Model 2	conv2d-flatten-dense-dropout-dense
Model 3	conv2d-conv2d-flatten-dense-dense
Model 4	conv2d-conv2d-flatten-dense-dropout-dense

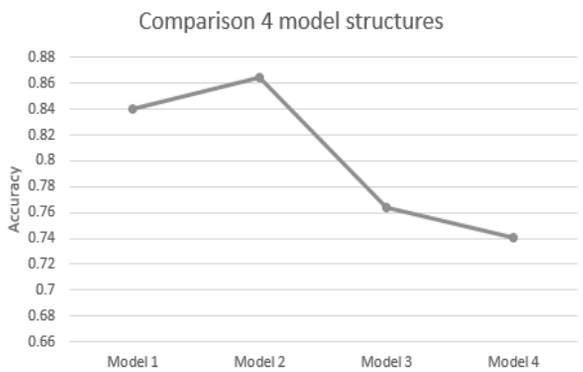

Fig. 4. Highest Accuracy under TATI Model at Different Structures ⟨Table 4⟩

깊이 6의 모델로 0.74099의 정확도를 보이며, 정확도가 점점 떨어지는 추세를 보였다.

실험 결과를 토대로 모델의 구조가 깊어질수록 정확도가 높아지다가 특정 지점에서 낮아지는 것을 관측할 수 있었다. 이는 이론상 모델이 깊어지고 복잡해지면 성능이 좋아지지만, 데이터셋의 크기나 종류에 맞지 않게 구조가 너무 복잡해지면 오히려 성능이 떨어진다는 것을 확인할 수 있었다. 따라서 Model 4보다 깊고 복잡한 모델에 대해서는 더 이상 실험을 진행하지 않았고, 0.8646으로 가장 높은 정확도를 보여준 Model 2의 구조를 TATI 모델에 적용하였다.

모델의 성능을 판단할 때에는 다양한 방법이 있다. 기본적으로 모델 정확도, F1 스코어, Precision과 Recall, 오차 범위 등 다양한 성능 평가 지표를 통해 모델을 평가한다[24]. 본 논문에서는 TATI 모델 성능의 우수성을 입증하기 위해 SVM(Support Vector Mechine), NN(Nueral Network), RNN(Recurrent Neural Network)의 딥러닝 신경망 모델을 사용하여 평균 정확도와 오차 범위를 비교한다.

총 10번의 실험을 진행하였고, 각 실험에 TATI 모델을 제외한 다른 모델들은 이미지로 변환하기 전 텍스트 데이터셋을 사용하였다. 본 실험에서 비교할 모델들의 구조와 하이퍼파라미터의 값은 다음과 같다. SVM 모델은 슬랙 변수 가중치를 10으로 설정하였다. 슬랙 변수 가중치는 값이 작아지면 SVM 모델의 마진이 커지고, 커지면 마진이 작아지는 값으로 SVM 모델 정확도의 지표이다. NN 모델은 두 개의 합성곱 레이어를 두었고, 회귀 모델로 구성하였다. 또한 손실 함수는 MSE, 최적화 함수는 Adam, 학습률은 0.01로 설정하였다. RNN 모델은 5개의 뉴런을 설정하였고, 타임 스탭마다 크기

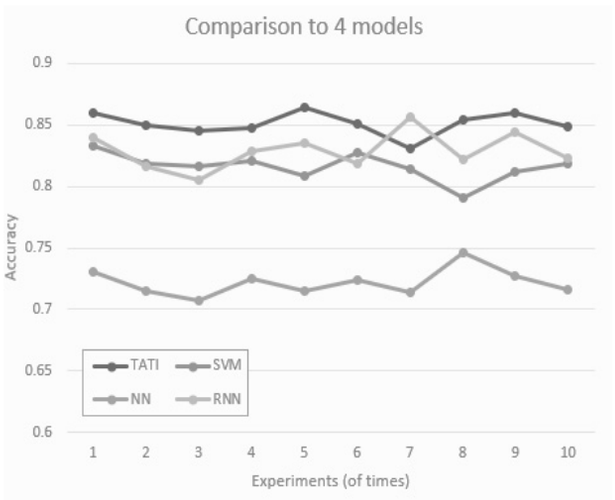

Fig. 5. Accuracy Under 4 Models in Ten Experiments

15의 입력을 받으며, 두 개의 타임 스탭에 대해 작동하도록 구성하였다. RNN 모델의 구조는 relu 활성화 함수로 설정된 하나의 레이어를 쌓고, 두 개의 dense 레이어를 쌓은 형태이다. 최적화 함수는 다른 모델들과 동일하게 Adam을 사용하였고, 배치 사이즈와 에폭 설정 또한 Table 3의 하이퍼 파라미터와 동일하게 진행하였다.

Fig. 5는 각 4개 모델들의 실험 결과를 그래프로 나타낸 그림이다. 10번의 실험 중 1번의 실험을 제외하고는 TATI 모델의 정확도가 가장 높은 것을 확인할 수 있다. NN 모델은 10번의 실험에서 모두 0.8 이하의 정확도를 보이며, 다른 3개의 모델과 차이가 크게 났다. 본 실험에서 NN 모델은 교통사고 심각 정도를 예측하여 적용하기에 적절하지 않은 것으로 드러났다. SVM과 RNN 모델은 대부분의 실험에서 0.8 이상의 정확도를 보였지만 RNN 모델의 6번째 실험을 제외한 실험의 정확도가 TATI 모델보다 낮다는 것을 볼 수 있다. Table 5는 그래프에서 표현한 4개의 모델의 10번의 실험에 대한 정확도와 평균 정확도를 나타내는 표이다. 평균 정확도를 볼 때, TATI 모델이 0.85123로 가장 높고, RNN, SVM, NN 모델 순서로 높은 것으로 나타났다. 교통사고 심각 정도 예측은 생명이 좌우될 수 있는 문제이다. 따라서 본 논문에서는 정확도 뿐만 아니라 모델의 정확도가 얼마나 일정하게 나오는지, 즉 정확도의 오차 범위가 중요한 성능 판단의 지표이다. Table 5를 통해 모델의 최대 정확도와 최소 정확도의 차를 계산하여 오차 범위를 확인하였다. 계산 결과 NN 모델이 0.0337, TATI 모델 0.0391, SVM 모델 0.0418, RNN 모델 0.0516 순으로 오차 범위가 좁은 것으로 나타났다. RNN과 SVM 모델을 비교하였을 때, 평균 정확도는 RNN 모델이 더 높지만 SVM보다 오차 범위가 더 크다. 이는 정확도보다 오차 범위를 더 중요하게 봐야 하는 교통사고 심각 정도에 대한 예측을 위해서는 SVM 모델을 적용하는 것이 더 합리적이라는 것을 의미한다. NN 모델은 4개 모델 중에서 오차 범위가 제일 좁지만, 정확도가 다른 모델들에 비해 현저히 떨어지기 때문에 안정적인 성능을 보이더라도 실제로 적용되기엔 어려운 것으로 분석할 수 있다.

Table 5. Experiments Result TATI Model at Different Structures

	NN	SVM	RNN	TATI
Experiment 1	0.7303	**0.833**	0.8403	0.8603
Experiment 2	0.7145	0.8181	0.8159	0.8493
Experiment 3	**0.7072**	0.8164	**0.805**	0.8455
Experiment 4	0.7245	0.8205	0.8289	0.8476
Experiment 5	0.7145	0.8081	0.8351	**0.8646**
Experiment 6	0.7244	0.8275	0.8186	0.8512
Experiment 7	0.7138	0.8141	**0.8566**	**0.8309**
Experiment 8	**0.7463**	**0.7912**	0.8223	0.8543
Experiment 9	0.7274	0.8123	0.8437	0.8598
Experiment 10	0.7166	0.8191	0.8231	0.8488
Average accuracy	0.72195	0.81603	0.82895	0.85123

비교 모델 간의 성능 비교 실험을 통해 TATI 모델은 약 0.85의 가장 높은 정확도를 보였고, 두 번째로 작은 오차 범위를 확인할 수 있었다. 위의 결과를 통해 본 논문에서 제안된 방법의 우수성을 입증하였다.

5. 결론 및 향후 방향

교통사고 사망자 감소를 위한 많은 연구와 노력이 이루어지고 있지만, 아직까지 뚜렷한 성과는 없는 실황이다. 교통사고 심각 정도를 예측하는 선행 연구들은 대부분 구체적인 심각 정도를 나누는 기준과 방법에 대한 한계점이 존재한다. 본 논문에서는 심각 정도를 보다 자세하게 예측하기 위해 교통사고 사망자와 중상자의 비율을 예측한다. 모델이 예측한 값이 높을수록 사망자가 발생할 위험이 높은 교통사고임을 의미한다.

본 논문에서는 독립요소들의 상관관계와 교통사고 데이터셋을 결합하여 컬러 이미지로 변환한다. TATI 모델의 구조에 따른 성능을 비교하기 위한 실험에서 합성곱 레이어의 개수와 dropout 모듈의 유무에 차이를 두고 분석하였다. 실험 결과에 따르면 하나의 합성곱 레이어와 dropout이 존재하는 구조가 가장 높은 성능을 보였다. 두 번째 실험에서 NN, SVM, RNN, TATI 모델들을 평균 정확도와 오차 범위를 통해 성능을 비교하였다. 실험 결과에 따르면, TATI 모델이 다른 모델보다 0.85로 정확도가 가장 높고, 오차 범위가 두 번째로 좁아 안정적인 모델임을 입증하였다.

본 논문에서는 실험에 사용되는 데이터셋을 TAAS에서 제공하는 정제된 교통사고 데이터셋 하나만을 사용하였고, 15개의 항목을 이용하여 실험을 진행하였다. 이는 컬러 이미지

로 변환하면 이미지의 크기가 작아지기 때문에 딥러닝 모델이 패턴을 학습하고 예측하기에 한계가 있었다. 따라서 날씨와 같은 환경 정보나 운전패턴, 주변 소리 등의 데이터셋 항목을 추가한다면 이미지로 변환하는 알고리즘을 통해 모델의 정확도가 더 높아질 것으로 예상되며, 추후 연구과제가 될 것이다. 또한 본 선행 연구를 기반으로 실시간으로 교통사고 심각 정도를 예측하고, 예측 모델을 해석하고, 데이터셋을 분석하여 다양하게 활용할 수 있는 교통안전 시스템을 고안하려 한다. 심각 정도 예측 결과를 실시간으로 관제센터 등에 송신함으로써 적절한 응급인력 배분 등의 빠른 조치를 취하여 치사율을 낮출 수 있고, 유사도 검색 등의 방안으로 활용함으로써 추후 교통사고 발생 예방의 효과를 가져올 것으로 기대한다.

References

[1] J. Gan, L. Li, D. Zhang, Z. Yi, and Q. Xiang, "An alternative method for traffic accident severity prediction: Using deep forests algorithm," *Journal of Advanced Transportation*, 2020.

[2] J. Orlovska, F. Novakazi, B. Lars-Ola, M. Karlsson, C. Wickman, and R. Söderberg, "Effects of the driving context on the usage of Automated Driver Assistance Systems (ADAS)-Naturalistic Driving Study for ADAS evaluation," *Transportation Research Interdisciplinary Perspectives,* Vol.4, 2020.

[3] S. J. Han, S. Y. Hwang, D. H. Ko, K. J. Eom, Y. S. Oh, and S. Y. Lee, "Research roadmap development for the big data based road accident cause analysis and policy," *The Korea Transport Insttitue(KOTI), Issuepaper*, Vol.17, No.5, 2017.

[4] G. Fountas and P. C. Anastasopoulos, "Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances," *Analytic Methods in Accident Research*, Vol.20, pp.30-45, 2018.

[5] K. El-Basyouny and T. Sayed, "Collision prediction models using multivariate Poisson-lognormal regression," *Accident Analysis & Prevention*, Vol.41, No.4, pp.820-828, 2009.

[6] T. K. Bahiru, D. K. Singh, and E. A. Tessfaw, "Comparative study on data mining classification algorithms for predicting road traffic accident severity," In *Proceedings of 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp.1655-1660, 2018.

[7] V. Rovšek, M. Batista, and B. Bogunović, "Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree," *Transport*, Vol.32, No.3, pp.272-281, 2017.

[8] H. T. Abdelwahab and M. A. Abdel-Aty, "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections," *Transportation Research Record*, Vol.1746, No.1, pp.6-13, 2001.

[9] M. M. Kunt, I. Aghayan, and N. Noii, "Prediction for traffic accident severity: Comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods," *Transport*, Vol.26, No.4, pp.353-366, 2011.

[10] M. I. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," *Applied Sciences*, Vol.7, No.6, pp.476-493, 2017.

[11] M. Zheng, T. Li, R. Zhu, J. Chen, Z. Ma, M. Tang, and Z. Wang, "Traffic accident's severity prediction: A deep-learning approach-based CNN network," *IEEE Access*, Vol.7, pp.39897-39910, 2019.

[12] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Analysis & Prevention*, Vol.43, No.5, pp.1666-1676, 2011.

[13] A. B. Parsa, R. S. Chauhan, H. Taghipour, S. Derrible, and A. Mohammadian, "Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data," arXiv preprint arXiv:1912.06991, 2019.

[14] Y. Chung, "Injury severity analysis in taxi-pedestrian crashes: An application of reconstructed crash data using a vehicle black box," *Accident Analysis & Prevention,* Vol.111, pp.345-353, 2018.

[15] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic accidents classification and injury severity prediction," In *Proceedings of 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pp.52-57, 2018.

[16] M. M. Chen and M. C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," *Information*, Vol.11, No.5, 2020.

[17] L. Y. Chang and H. W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, Vol.38, No.5, pp.1019-1027, 2006.

[18] R. O. Mujlli and J. De Oña, "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks," *Journal of Safety Research*, Vol.42, No.5, pp.317-326, 2011.

[19] S. S. Dhaliwal, A. A. Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost," *Information*, Vol.9, No.7, 2018.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp.785-794, 2016.
[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, pp.1189-1232, 2001.
[22] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.1, No.1, pp.14-23, 2011.
[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.
[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, Vol.27, No.8, pp.861-874, 2006.

추 민 지
https://orcid.org/0000-0001-6320-085X
e-mail : minchu96@sm.ac.kr
2015년 가천대학교 소프트웨어학과(학사)
2020년 ~ 현 재 숙명여자대학교 IT공학과 석사과정
관심 분야 : 빅데이터, 데이터분석, 머신러닝, 인공지능

박 소 현
https://orcid.org/0000-0003-1022-1790
e-mail : shpark@sm.ac.kr
2014년 숙명여자대학교 기악학부(학사)
2014년 ~ 2016년 숙명여자대학교 멀티미디어과학과(석사)
2016년 ~ 2020년 숙명여자대학교 IT공학과(박사)
2020년 ~ 현 재 숙명여자대학교 빅데이터활용 연구센터 책임연구원
관심 분야 : 데이터 마이닝(Data Mining), 스마트 교육(Smart Education), 추천 시스템(Recommendation System), IT 융합(IT Convergence)

박 영 호
https://orcid.org/0000-0002-5284-9589
e-mail : yhpark@sm.ac.kr
1992년 동국대학교 컴퓨터공학과(석사)
2005년 한국과학기술원 전산학과(박사)
1993년 ~ 1999년 한국전자통신연구원 교환전송연구단 선임연구원
2005년 ~ 2006년 한국과학기술원 첨단정보기술연구센터 연구원
2005년 ~ 2006년 동국대학교 컴퓨터멀티미디어학과 겸임교수
2006년 ~ 2010년 숙명여자대학교 멀티미디어학과 조교수
2010년 ~ 2015년 숙명여자대학교 멀티미디어학과 부교수
2015년 ~ 현 재 숙명여자대학교 IT공학과 교수
관심 분야 : 데이터베이스, XML, IR(정보검색), 멀티미디어 데이터베이스, Bio정보공학, 영상미디어, 예술& 공학인터페이스, 데이터베이스 관리시스템, 머신러닝, 빅데이터, 데이터분석, Telecommunication System