KIPS Trans. Softw. and Data Eng.
Vol.10, No.7 pp.257~262
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.7.257

Black Ice Formation Prediction Model Based on Public Data in Land, Infrastructure and Transport Domain

Jeong Ho Na† · Sung-Ho Yoon† · Hyo-Jung Oh††

ABSTRACT

Accidents caused by black ice occur frequently every winter, and the fatality rate is very high compared to other traffic accidents. Therefore, a systematic method is needed to predict the black ice formation before accidents. In this paper, we proposed a black ice prediction model based on heterogenous and multi-type data. To this end, 12,574,630 cases of 46 types of land, infrastructure, transport public data and meteorological public data were collected. Subsequently, the data cleansing process including missing value detection and normalization was followed by the establishment of approximately 600,000 refined datasets. We analyzed the correlation of 42 factors collected to predict the occurrence of black ice by selecting only 21 factors that have a valid effect on black ice prediction. The prediction model developed through this will eventually be used to derive the route-specific black ice risk index, which will be utilized as a preliminary study for black ice warning alart services.

Keywords : Black Ice, Public Data, Data Cleansing, Machine Learning, Prediction

국토 교통 공공데이터 기반 블랙아이스 발생 구간 예측 모델

나 정 호† · 윤 성 호† · 오 효 정††

요 약

매년 동절기 블랙아이스(Black Ice)로 인한 사고는 빈번하게 발생하고 있으며, 치사율은 다른 교통사고에 비해 매우 높다. 따라서 블랙아이스 발생 구간을 사전에 예측하기 위한 체계화된 방법이 필요하다. 이에 본 논문에서는 이질(heterogeneous)·다형(diverse)의 데이터를 활용한 블랙아이스 발생 구간 예측 모델을 제안한다. 이를 위해 국토 교통 공공데이터와 기상 공공데이터 42종의 12,574,630건을 수집하여, 결측값을 처리하고 정규화하는 등의 전처리 과정을 수행한 뒤 최종 약 60만여 건의 정제 데이터셋을 구축하였다. 수집된 요인들의 상관관계를 분석하여 블랙아이스 예측에 유효한 영향을 주는 21개 요인을 선별, 다양한 학습모델을 조합하는 방법을 통해 블랙아이스 발생 예측 모델을 구현하였다. 이를 통해 개발된 예측 모델은 최종적으로 노선별 블랙아이스 위험지수 도출에 사용되어 블랙아이스 발생 경고 서비스를 위한 사전 연구로 활용될 것이다.

키워드 : 블랙아이스, 공공데이터, 데이터정제, 머신러닝, 예측

1. 서 론

동절기의 도로는 다른 계절과 다르게 지상 기온이 영하로 떨어지는 날씨로 인해 도로 결빙이 자주 발생한다. 도로 결빙은 교통사고와 도로 혼잡 및 정체와 같은 교통문제를 야기하므로 겨울철 도로 노면 관리는 중요하며, 사회적으로 많은 관심과 관련 기관의 철저한 관리가 필요하다[1].

특히 블랙아이스(Black Ice)는 일반적인 도로 결빙과는 달리 포장된 도로와 같은 색상인 검은 색을 띄고 있어 운전자가 이를 식별하여 대처하는 것이 어렵다. 실제로 지난 2015-18년 전체 교통사고 건수는 감소한 반면, 블랙아이스로 유발되는 교통사고는 9.1% 증가했다. 특히 블랙아이스의 위험성은 노면상태별 치사율에서 더욱 두드러지게 나타난다. 도로교통공단의 교통사고분석시스템에 따르면 노면상태별 교통사고 치사율은 결빙(4.6%)으로 인한 사고가 건조(1.4%), 젖음(1.9%), 적설(1.2%) 등에 비해 3~4배 높게 나타났다[2].

이렇듯 블랙아이스로 유발되는 교통사고는 높은 치사율을 보여주기 때문에 이러한 사고가 발생하기 전에 블랙아이스 구간을 예측하여 예방하는 것이 필요하다. 본 연구에서는 공공데이터포털을 활용해 이질적(heterogeneous)이고 다양한(diverse) 데이터를 수집하여 국내 고속도로 구간을 대상으로 블랙아이스 발생 여부를 예측하는 모델을 구축하고자 한다(Fig. 1 참조). 이를 위해 먼저 국내 고속도로 노선·위치정보, 고속도로정보, 기상정보, 교통정보 등 여러 데이터를 수집한다. 수집된 데이터는 데이터 전처리 및 가공 과정을 거쳐 고품질의 데이터로 정제하고, 이후 상관관계 분석 및 기계학

※ 이 논문은 2021년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
※ 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A5B8099507).
† 비 회 원 : 전북대학교 기록관리학과 석사과정
†† 정 회 원 : 전북대학교 문헌정보학과 부교수, 문화융복합아카이빙연구소 연구원
Manuscript Received : February 9, 2021
Accepted : March 31, 2021
* Corresponding Author : Hyo-Jung Oh(ohj@jbnu.ac.kr)

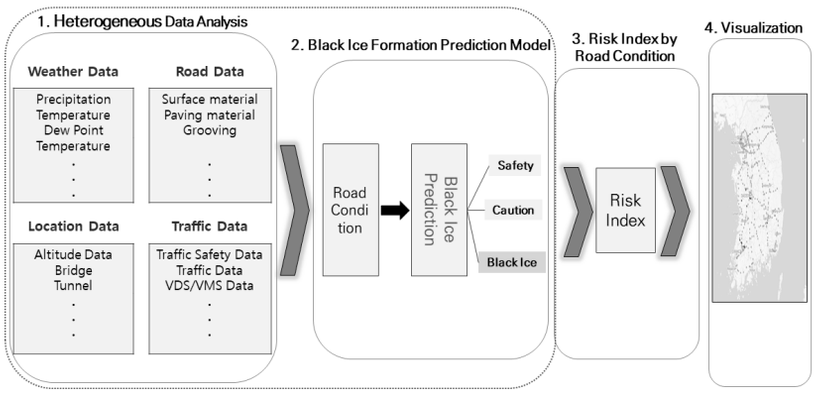


Fig. 1. Overview of Black Ice Warning Alert Service

습에 적합한 형태인 하나의 데이터셋(Dataset)으로 결합하여 각 변수의 단위를 통일하는 정규화를 수행한다. 확보된 데이터셋을 기반으로 모델을 학습한다. 구축된 예측 모델은 성능 평가를 거치며, 특정 일자를 예시로 하는 세부 정확도 및 재현율 검증을 수행하고자 한다.

2. 관련 연구

블랙아이스 발생 예측과 관련한 초기 연구들은 주로 특정 기간 내 기상데이터만을 활용한 경우가 대부분이었다. 그중 박근영 등(2017)의 연구에서는 AWS(Automatic Weather System)에서 제공하는 기상데이터를 활용해 지난 2012년 12월 14일과 21일에 발생한 교통사고 사례를 분석하였다[3]. 사고 발생지점의 지상 기온과 풍량, 강우량 등을 확인하여 두 사례가 블랙아이스 및 어는 비 형성에 유리한 기상조건임을 확인하였다. 이를 토대로 지상 기온이 0°C 이하인 조건에서 아침 최저기온을 보이는 오전 6-8시에 어는 비가 내린다면 블랙아이스 발생 가능성이 높아짐을 유추하였다. 이영미 등(2018)은 분석 지역을 제주도로 특정하고 결빙예측모델인 WRF(Weather Research and Forecasting)를 활용해 결빙 발생을 예측하였다[4]. WRF를 통해 2017년 12월 1일부터 2018년 2월 28일까지의 기상예측자료를 생성하였으며, 결빙예측결과 검증은 제주도청에서 제공하는 동 시기의 도로결빙 및 교통통제 상황자료를 이용했다. 정확도 검증을 수행하여 해당 연구에서 도출된 모듈의 신뢰성을 확보하였음을 평가했다. 연구의 한계점으로는 시간대별로 기상데이터를 활용하지 못한 점을 언급하여, 정시마다 관측한 데이터를 활용한다면 정확도의 향상에 도움이 될 것으로 예상하였다.

김상엽 등(2015)은 상기 연구들이 기상정보를 주로 활용한 점과는 달리 인적, 차량, 도로환경 등, 도시에서 발생하는 도로 결빙사고 발생에 영향을 미치는 다양한 요인을 분석하였다[1]. 회귀분석(Logistic Regression Model)을 이용한 분석 결과, 새벽 혹은 출근 시간에 위험도가 가장 높았으며, 차량 요인에서는 승용차에 의한 사고와 차량 단독사고의 위험도가 높았다. 또한 도로환경 요인에서는 종단경사가 존재하는 구간, 왕복 4차로 이하의 도로, 교량에서 위험성이 높은 것을 도출하였다.

그 외에도 김진영 등(2020)은 차량에 부착된 센서를 통해 블랙아이스를 식별하는 방안을 제시하였다[5]. 블랙아이스가 대기 온도보다 지면 온도가 약 2-3°C 낮은 환경에서 발생확률이 높은 점에 착안하여 차량에 온도센서를 부착해 열복사 에너지를 측정하는 방안을 제시하였다. 또한 넓은 도로의 열복사 에너지를 측정하여 실시간으로 블랙아이스를 식별하는 기술에 대한 연구가 필요하다고 언급하였다.

본 연구는 앞서 살펴본 선행연구에서 주로 활용된 기상데이터를 포함하여 블랙아이스 발생에 영향을 줄 것으로 예상되는 고속도로 위치·노면·고도정보, 도로포장정보 등의 다양한 데이터를 발굴, 수집하고자 한다. 특히 차량이 고속으로 주행하는 고속도로에서 발생하는 블랙아이스는 더 높은 치사율과 위험성을 내포하기 때문에 국내 고속도로 지점 1,000곳을 대상으로 예측 모델을 구축하고자 한다. 또한 블랙아이스 발생 예측 결과의 시의성을 제고하기 위한 방안으로 고속도로 주행 중 제공되는 실시간 기상정보, 교통사고정보, 교통상황정보를 활용하였다. 이를 통해 수집주기가 상대적으로 길어 실시간 정보를 반영하기 어려운 국토 교통 분야 공공데이터의 한계를 보완하고자 한다.

3. 블랙아이스 발생 구간 예측 모델

3.1 블랙아이스 발생 구간 예측 연구 모델

Fig. 1은 본 연구진이 최종적으로 개발하고자 하는 블랙아이스 발생 경고 서비스 흐름도를 도식화한 그림이다. 설계 과정은 크게 4단계로 구성된다. 먼저 이질 데이터 수집 및 분석 과정을 수행하여 고품질의 데이터를 정제하고, 선별된 요인들을 활용하여 블랙아이스 발생 구간 예측 모델을 학습한다. 이어 예측 모델에 의해 도출된 값은 노면상태에 따라 위험지수를 산출한다. 상기 과정을 거쳐 도출된 결과 값을 최종적으로 웹을 통해 각 고속도로 노선별 블랙아이스 위험지수를 시각화하는 서비스로 제공하고자 한다.

본 논문은 전체 설계 과정 중 1-2단계에 해당하는 이질 데이터 분석과정을 거쳐 도출된 요인들을 활용한 블랙아이스 발생 구간 예측 모델 구축에 방점을 두고자 한다. 이러한 예측 모델을 구축하고자 먼저 국토 교통 공공데이터를 수집 및 분석하였다. 국내 고속도로 노선정보, 고속도로별 위치정보, 차량감지시스템(VDS: Vehicle Detection System), 교통상

황정보(VMS: Variable Message Sign), 노면 정보, 기상정보 등 이질적인 공공데이터를 수집하였고, 수집한 데이터를 대상으로 데이터 통합, 결측값(Missing Value) 처리, 단위 통일 등의 전처리 및 정규화 과정을 수행한 뒤 수집된 데이터를 하나의 데이터셋으로 결합하였다.

특히 본 연구의 주안점인 블랙아이스 발생 구간 예측모델의 학습은 Fig. 2와 같이 이질 데이터를 입력받아 노면 상태를 '결빙', '노면 미끄럼', '어는비'등 10개의 상태로 구분하는 다중분류(Multi-label Classification) 모델로 진행하였다. 최종 학습모델은 다양한 기계학습방법을 조합해서 활용하는 앙상블(Ensemble) 기법으로 구현하였다.

본 논문에서 구현한 블랙아이스 발생 구간 예측 모델은 다음과 같은 특징을 지닌다. 첫째, 도로의 포장재, 위치정보, 기상정보, 교통정보 등 국토 교통 데이터를 수집 및 분석하여 도로정보를 파악한다. 둘째, 여러 모델학습 기법을 대상으로 테스트를 수행하여 가장 높은 정확도를 가진 기법을 활용했다. 셋째, 이질적인 데이터를 기반으로 모델학습을 수행해 높은 정확도로 블랙아이스 발생 구간을 예측한다.

3.2 데이터 수집 및 전처리

1) 국토 교통 공공데이터 수집

문재인 정부의 데이터 댐 구축 정책 추진으로 인해 공공영역에서 생산된 다양한 유형의 데이터는 데이터 '포털' 혹은 '플랫폼' 등을 통해 일반인에게 제공되고 있다. 그중 국토 분야의 공공데이터 포털의 대표적인 예로 국토교통부의 국가정보오픈플랫폼[1], 국가공간정보포털[2]과 국토지리정보원의 국토정보플랫폼[3] 등이 있다. 교통 분야 공공데이터 포털로는 국토교통부의 교통정보공개정보서비스, 도로교통공단의 교통사고분석시스템[4], 한국도로공사의 고속도로 공공데이터 포털[5] 등이 있다.

데이터 포털 등을 통해 제공되는 국토 교통 공공데이터는 이용에 제한이 없으며, 개방성이 OECD 국가 중 1위일 만큼 매우 높은 수준이다[6]. 또한 다양한 유형별 데이터와 더불어 이를 효율적으로 활용할 수 있도록 지원하는 Open API(Application Programming Interface) 또한 제공된다. 그러므로 본 연구진은 데이터 수집 대상으로 다형성, 개방성, 접근성이 특징인 국토 교통 공공데이터를 선정하였다.

Table 1은 본 연구에서 활용한 국토 교통 공공데이터를 정리한 것이다. 표에서 나타나듯이 본 연구를 위해 수집된 데이터는 총 42종, 12,574,630건으로, 여러 기관에서 제공되기 때문에 다양한 형식으로 구성되어 있을 뿐 아니라 값의 범위도 매우 상이하다. 데이터 유형별 특징을 살펴보면 다음과 같다. 첫째, '위치'정보는 국내의 고속도로의 설치 위치와 노선별 고도 정보를 포함한다. 둘째, '도로'정보는 고속도로의 포장재와 표층재 정보, 그루빙 설치 현황 및 고속도로 설해취약지역 등의 데이터가 포함되어 있다. 셋째, '실시간 교통' 정보는 VDS, VMS 설치 정보와 실시간 교통데이터가 포함된

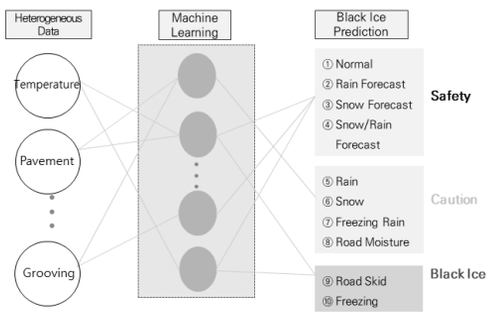


Fig. 2. Black Ice Prediction Model

Table 1. Collection Data

Collection Data		Column	Rows	Used Column
Location Data	Highway Route Data	7	45,954	3
	Elevation Data of Highway Route	3	1068	2
Road Data	Highway Pavement Data	13	53,846	3
	Grooving Data	23	2,295	0
	Snow Damage Vulnerable Section of Highway	8	34	1
Traffic Data	VDS Installation Data	10	6,303	0
	VDS Traffic Data	4	11,680,000	0
	VMS Installation Data	8	1,470	0
	VMS Message Data	8	500,000	1
Weather Data	Monthly Weather Data	23	57,381	10
	VDS Weather Data	8	226,279	1
Total of Collected Data		**114**	**12,574,630**	
Used Data After Pre-Processing		**42**	**604,000**	**21**

다. 넷째, '기상'정보는 기온, 풍속, 이슬점온도, 일조시간, 지면온도 등의 노선별 기상정보와 종관기상의 정보가 포함되어 있다. 마지막으로 '기상'정보, '실시간 교통량'정보로 대표되는 시계열 데이터와 '위치'정보, '도로'정보인 비시계열 데이터는 고속도로 노선명과 도로 좌표를 매핑(Mapping)한 뒤 VDS 설치지점인 콘존(Conzone)을 기준으로 통합하였다.

실시간 교통정보는 2020년 8월 1일 기준으로, 전국 고속도로를 대상으로 데이터를 수집하였다. 특히 기상정보의 경우, 2016년~2019년을 기준으로 블랙아이스 및 결빙이 주로 발생하는 기간(11월~3월)을 대상으로 하였다. 기상정보 중 VDS 기반 기상데이터는 한국도로공사 고속도로 공공데이터포탈[6]에서 수집하였으며, 종관기상 데이터는 기상청 기상자료개방포털[7]에서 수집하였다. 위치정보 중 고도 데이터는 구글의 'Elevation API[8]'를 활용하였다. 수집된 이질·다형의 데이터를 활용하기 위해서는 다음과 같이 정규화를 포함한 전처리 작업이 수반된다.

2) 데이터 정제

Table 1의 수집된 1,260만여 건의 국토 교통 공공데이터의 전처리 과정은 다음과 같다. 우선 각 테이블을 Python 환경에서 Pandas[9]와 Numpy를 이용하여 하나의 테이블로 통

1) https://www.vworld.kr/v4po_main.do
2) http://www.nsdi.go.kr/lxportal/?menuno=2679
3) http://map.ngii.go.kr/mn/mainPage.do
4) http://openapi.its.go.kr/portal/main.do
5) http://data.ex.co.kr/
6) http://data.ex.co.kr/portal/fdwn/view?type=ETC&num=O3&requestfrom=dataset
7) https://data.kma.go.kr/data/grn d/selectAsosRltmList.do?pgmNo=36
8) https://developers.google.com/maps/documentation/elevation/start

합한 후 원시데이터(Raw data)를 정규화하였다. 데이터 전처리 과정을 통해 생성된 구조화 데이터(Structured data)를 기반으로 블랙아이스 발생구간 예측 모델을 학습시키고 확률결과를 Pickle로 저장하였다.

이 과정 중 국토 교통 공공데이터의 전처리를 위해 다음의 과정을 수행하였다. 첫째, 도로정보, 기상정보 등 수집한 데이터는 콘존을 기준으로 하나의 테이블로 통합해 데이터셋을 구축하였다. 둘째, 결측값(Missing value)을 처리하였다. 예컨대 기상정보 중 '평균기온'에 결측값이 있는 경우, '최고기온'과 '최저기온'의 평균값을 구해 결측값을 처리하였다. 마지막으로 VMS의 메시지 중 "결빙주의", "살얼음"과 같은 유사한 keyword는 "결빙"으로 통일하는 등 각 기상변수를 정리 및 통합하였다. 구축된 데이터셋의 범주형 변수는 정수형으로, 수치형 변수는 단위를 통일하고자 정규화하였다. 일련의 과정을 거쳐 최종적으로 전체 1,260만 건 중 42개 요인으로 구성된 총 604,000건의 정제 데이터를 확보하였다. 이러한 수치는 전체 수집 데이터 대비 4.8%에 불과한 것으로, 이는 관련 데이터를 발굴하고 수집하는 과정보다 확보된 데이터로부터 고품질의 데이터를 선별하고 가공하는 전처리 과정의 중요성을 방증하는 결과이다.

4. 블랙아이스 발생 구간 예측 모델 학습

4.1 요인별 상관관계 분석

본격적인 모델 학습에 앞서 수집된 Table 1의 국토 교통 데이터들의 요인별 상관관계를 분석하였다. 이는 블랙아이스 발생과 밀접한 요인들을 선정하고 관련성이 상대적으로 부족한 요인들을 제거함으로써 모델의 정확도뿐만 아니라 학습속도 등의 효율을 높이기 위함이다. 상관관계 분석 결과, 전체 42개 요인 중 상관도가 0 이상인 요인은 21개로, Fig. 3은 관련성이 높은 요인부터 순차적으로 나열한 것이다.

세부 요인을 살펴보면, 위치정보는 '노선번호', '상향선 고도' 등 총 5개이며, 도로정보는 '도로포장재', '결빙취약구역' 등 4개가 있었다. '기상'정보에는 '최저기온', '평균이슬점온도' 등 총 11개고, 실시간 교통정보는 'VMS표출시간' 1개이다. 이 중 가장 큰 영향을 가지는 것은 실시간 교통정보의 'VMS표출시간'임을 확인하였고, 기상정보는 전체 21개 요인 중 11개 요인이 포함되어 예상대로 블랙아이스 발생에 가장 많은 영향을 끼치는 것으로 분석되었다. 또한, 도로와 위치 정보인 노선 '종류'와 '고도', '도로포장재' 또한 블랙아이스 발생 예측에 영향을 주는 것으로 나타났다. 그 외의 21개의 요인 중 위치정보요인은 '노선방향' 등 6개, 도로정보요인은 '그루빙설치여부' 등 6개이며 기상정보요인은 '평균지면온도' 등 8개로 파악되었다.

상관분석을 통해 발견한 특이한 점은 본 연구진이 사전에 예상한 바와 달리 '평균지면온도'가 블랙아이스 발생에 영향을 주지 않는 것으로 파악되었다. 이는 '평균지면온도' 정보가 실제 도로표면온도가 아닌 기상청의 기상관측소에서 수집한 지표면 온도이기 때문에 실제 도로환경에 적용시키기에

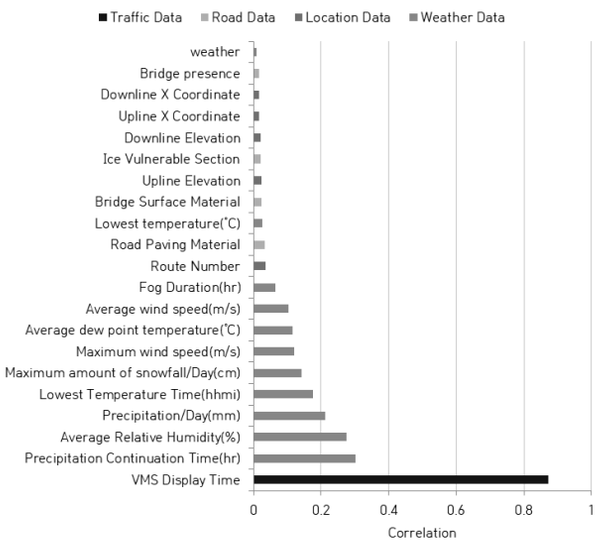


Fig. 3. Factor-Specific Correlation Result

부적합함에서 기인한 것으로 사료된다.

4.2 모델 학습

블랙아이스 발생 예측 모델 학습은 Fig. 2와 같이 주어진 환경요인에 따라 구체적인 노면 상태(10개)를 예측하는 다중분류모델로, 앞서 선별된 21개의 요인데이터를 기반으로 기계학습 기법의 정확도를 비교한 후, 가장 높은 성능을 가진 모델을 조합하여 연구를 진행하였다. 모델에 활용할 기법은 머신러닝과 딥러닝의 대표 기법인 KNN(K-Nearest Neighbor), RF(Random Forest), DT(Decision Tree)와 MLP(Multi-layer Perceptron) 등과 이를 결합한 앙상블 모델을 제안하여 성능평가를 수행했다. 블랙아이스 발생 예측 모델별 정확도는 Fig. 4와 같다.

성능 비교 결과, DT와 MLP 기법이 가장 정확도가 높은 것을 확인할 수 있었다. DT는 주어진 무질서한 훈련 샘플에서 트리 분류 모델을 추출할 수 있는 인스턴스 기반 유도학습모델[7]이고, MLP는 입력 데이터 집합을 적절한 출력 집합에 매핑하는 피드 포워드 네트워크 모델(Feed Forward Network Model)로 최소 3 개의 노드 계층인 입력 계층, 은닉계층, 출력 계층으로 구성되며 입력 노드를 제외한 각 노드는 비선형 활성화 함수(Nonlinear Activation Function)를 사용한다[8][9]. KNN은 입력 데이터와 학습데이터 간의 거리를 바탕으로 근접한 데이터를 찾아 K개의 범주로 분류하는 방법[10]이며, RF는 DT와 같은 트리 기반 예측 모델을 활용한 앙상블 기법이다[11].

복잡도가 높은 모델에 비해 비교적 단순한 DT의 성능이 가장 높은 결과 역시 본 연구에서 사용하는 데이터의 편향성을 반영한 것으로, 특정 자질에 의해 블랙아이스 발생 여부가 결정되고 있음을 반증한다.

Fig. 4의 결과에 따라 DT와 MLP를 조합하여 앙상블 모델 중 Soft Vote 기법[12]을 활용하였으며, 모델 학습에 사용한 파라미터 값은 다음과 같다. 먼저 DT는 Max depth를 15, min samples split와 min samples leaf는 각각 2와 1로 설정했으며, MLP에서는 batch size는 32, max iteration은 15로

9) https://pandas.pydata.org/

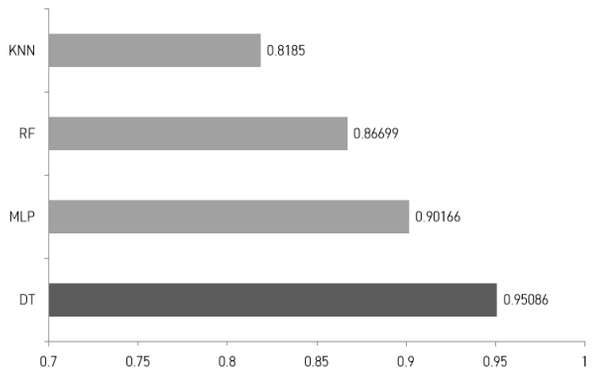


Fig. 4. Accuracy by Model Learning Method

설정하였다. 또한 앙상블 모델의 과적합(overfitting)을 방지하고자 앞서 선별한 21개 요인의 데이터셋을 학습데이터와 검증데이터, 테스트 데이터로 구분하였으며, 그 비율은 5:3:2로 설정하였다. 그 결과 전체 데이터셋 604,000건 중 학습데이터는 302,000건, 검증데이터는 181,200건, 테스트데이터는 120,800건으로 구분되었다.

4.3 블랙아이스 발생 예측 모델 평가

학습된 예측 모델성능을 평가하기 위해 분류성능평가지표(Confusion matrix, [13])를 사용하였고 그 결과는 Table 2와 같다. 평가에 활용한 데이터는 앞서 설명한 바와 같이 12만여 건이다.

Table 2 성능을 비교해보면 먼저, 기존 선행연구와 같이 '기상'정보에 해당하는 18개 요인만을 대상으로 구축한 학습모델의 평가 결과는 정밀도(Precision)가 83.8%, 재현율(Recall)이 85.3%, 정확도(Accuracy)는 85.3%로 나타났다. 이에 반해 본 연구에서 수집, 정제한 전체 42개의 요인을 활용한 학습 모델 평가 결과는 정밀도가 94.1%, 재현율이 94.4%, 정확도가 94.4%로, 기존 기상요인만 활용한 경우에 비해 9%이상의 성능 향상을 보였다.

한편, 마지막으로 선행 절에서 수행한 요인별 상관관계 분석 결과(Fig. 3 참조)에 따라 선별된 21개의 요인을 활용한 학습 모델 평가결과는 정밀도가 95%, 재현율이 95.1%, 정확도 95%로 평가되었다. 이는 수집된 모든 데이터를 활용한 경우보다 더 높은 정확도는 보이는 것으로, 예측 모델에 영향을 미치는 고품질의 선별된 데이터를 활용하는 것이 정확도와 효율성 측면에서 유리함을 의미한다.

Table 3은 구축한 모델을 활용한 예측 예시로, 2019년 12월 30일의 전국 고속도로 1,000개 지점을 대상으로 노면상태를 예측한 결과값이다. Table 3에서 보이는 바와 같이 해당 날짜에는 전체 분류 대상 10개 노면 상태 중 '이상없음', '비', '어는비', '노면습기', '결빙'의 5가지 노면상태만 관측되었고, 실제 예측 결과 역시 대부분 이 5 상태 중 하나로 할당되었다. 학습모델 정확률은 '이상없음(100%)', '결빙(96%)', '어는비(51%)', '노면습기(41%)', '비(27%)' 순으로 나타났다.

한편 세부 성능분석 결과, 노면 상태별로 정확률보다 재현율의 편차가 매우 큰 것으로 나타났는데, 이는 각각의 노면 상태

Table 2. Result of Assessment of Black ice Risk Section

	Precision	Recall	Accuracy
only Weather(18 features)	83.8%	85.3%	85.3%
Proposed Method(42 features)	94.1% (+10.3%)	94.4% (+9.1%)	94.4% (+9.1%)
Proposed Method(21 features)	**95.0% (+11.2%)**	**95.1% (+9.8%)**	**95.0% (+9.7%)**

Table 3. Result of Assessment of Black ice Risk Section By Surface Type

Risk Rank	Road Condition	Sample	Prediction	Correct	Precision	Recall
Safety	Normal	344	344	344	100%	100%
	Rain Forecast	-	-	-	-	-
	Snow Forecast	-	-	-	-	-
	Snow/Rain Forecast	-	-	-	-	-
Caution	Rain	55	166	45	27%	82%
	Snow	-	-	-	-	-
	Freezing rain	55	76	39	51%	71%
	Road surface moisture	26	27	11	41%	42%
Black Ice	Road Skid	0	10	0	-	-
	Freezing	520	377	361	96%	69%
Total		1000				

데이터 간의 불균형 문제(Biased problem) 때문으로 파악되며, 향후 각각 노면상태 데이터가 충분히 축적된다면 해소될 것으로 기대된다. 특히 블랙아이스 발생 예측은 그 문제의 특성상 교통사고를 사전에 예방한다는 차원에서 거짓 알람(False Alarm)이 미탐지(Miss)보다 중요함에도 불구하고, '결빙'에 대한 재현율이 낮은 점은 학습 데이터 자체가 평상시의 '이상없음' 상태의 자료 수가 블랙아이스 발생 상태의 자료 수에 비해 월등히 많음에서 기인한 것으로, 이러한 특성을 감안하여 재현율을 유지하면서 정확률을 높이는 방향으로 연구를 진행한다면 실제 블랙아이스 발생 경고 서비스 구축에 보다 효율적이라고 판단된다.

5. 결 론

본 연구는 전국 고속도로를 대상으로 블랙아이스 발생 구간을 예측하는 모델을 구축하였다. 선제적으로 국토 교통 공공데이터 12,574,630개를 수집해 전처리 과정을 거쳐 604.000개의 정제된 데이터를 확보했다. 이를 하나의 데이터셋으로 결합해 요인별 상관관계 분석을 거쳐 전체 42개 요인 중 상관도가 0 이상인 21개 요인을 선별하였다. 앙상블 모델을 활용해 예측 모델을 평가한 결과, 유의미한 상관도를 가진 요인들을 대상으로 한 모델의 성능이 가장 높은 것으로 확인되었으며, 이는 기존 선행연구에서 수행된 기상정보 기반의 모델보다 높은 성능을 보여줬다. 이를 토대로 2019년 12월 30일을 대상으로 전국 고속도로 1000개 지점의 블랙아이스 발생 예측 예시를 도출해 본 연구진이 구축한 예측 모델의 성능을 확인하였다.

본 연구는 선행연구의 한계점을 해소하고자 기상정보 외에 다양하고 이질적인 데이터를 수집 및 분석 대상으로 수집하

였다. 상관관계 분석을 통해 다양한 요인들이 블랙아이스 발생에 영향을 주는 요인임을 확인하였다는 점에서 의의가 있다. 또한 수집된 데이터를 전처리 및 가공하여 전체 4.8%에 불과한 정제 데이터를 확보하는 과정에 비추어보았을 때, 데이터 수집 및 발굴 과정보다 고품질의 데이터를 선별하는 과정의 중요성이 드러났다. 한계점으론 예측 모델의 결과값에 해당하는 각 노면 상태 데이터 간 불균형 문제가 있어 노면 상태에 따라 정확률 및 재현율이 크게 상이한 점이다. 향후 본 연구 결과는 노선별 블랙아이스 위험지수 도출에 사용되어 최종적으로 웹 기반의 블랙아이스 발생 경고 서비스를 구현하기 위한 사전 연구로 활용될 것이다.

References

[1] S. Y. Kim, S. Y. Kim, Y. S. Jang, S. K. Kim, D. C. Min, H. H. Na, and J. S. Choi, "A study on the effects of factors of traffic accidents caused by frozen urban road surfaces in the winter," *International Journal of Highway Engineering*, Vol.17, No.2, pp.79-87, 2015.

[2] Ilyo News Article, [Internet] https://ilyo.co.kr/?ac=article_view& entry_id=384970, 2020.

[3] G. Y. Park, S. H. Lee, E. J. Kim, and B. Y. Yun, "A case study on meteorological analysis of freezing rain and black ice formation on the load at winter," *Journal of Environmental Science International*, Vol.26 No.7 pp.827-836, 2017.

[4] Y. M. Lee, S. Y. Oh, and S. J. Lee, "A study on prediction of road freezing in Jeju," *Journal of Environmental Science International*, Vol.27, No.7, pp.531-541, 2018.

[5] J. Y. Kim, H. J. Lee, and J. R. Paik, "Survey of distinction of black ice using sensors," *Journal of The Korea Society of Computer and Information*, Vol.28, No.1, pp.78-87, 2020.

[6] Y. H. Kim, "Government 3.0 based consumer oriented big data service activation plan: Busan city service analysis information system and busan public data portal," *Local Information Magazine*, Vol.101, pp.16-21, 2016.

[7] D. Lai, Y. Zhang, X. Zhang, Y. Su, and M. B. Bin Heyat, "An automated strategy for early risk identification of sudden cardiac death by using machine learning approach on measurable arrhythmic risk markers," *IEEE Access*, Vol.7, pp.94701-94716, 2019.

[8] K. V. Sujatha, and S. Meenakshi Sundaram, "A combined PCA- MLP model for predicting stock index," *A2CWiC '10: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*, pp.1-6, Sep. 2010.

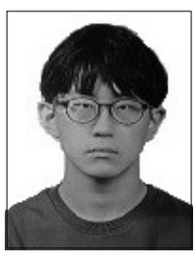
[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation," MIT Press, pp.318-362, Jan. 1986.

[10] S. M. Lee, J. S. Yeon, J. S. Kim, and S. S. Kim, "Semisupervised learning using the AdaBoost algorithm with SVM-KNN," *The transactions of The Korean Institute of Electrical Engineers*, Vol.61, No.9, pp.1336-1339, 2012.

[11] M. W. Lee, Y. G. Kim, Y. J. Jun, and Y. H. Shin, "Random Forest based Prediction of Road Surface Condition Using Spatio-Temporal Features," *Journal of Korean Society of Transportation*, Vol.37, No.4, pp.338-349, 2019.

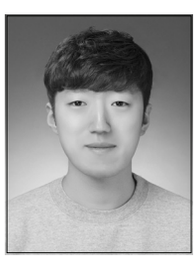
[12] Y. H. Kim, J. Y. Hong, and B. J. Kim, "Performance Comparison of Machine Learning Classification Methods for Decision of Disc Cutter Replacement of Shield TBM," *Journal of Korean Tunnelling and Underground Space Association*, Vol.22, No.5, pp.575-589, 2020.

[13] M. C. Jeong, J. H. Lee, and H. Y. Oh, "Ensemble Machine Learning Model Based Youtube Spam Comment Detection," *Journal of the Korea Institute of Information and Communication Engineering*, Vol.24, No.5, pp.576-583, 2020.



나 정 호

https://orcid.org/0000-0001-5118-3239
e-mail : jhna2012@naver.com
2020년 전북대학교 문헌정보학과(학사)
2020년 ~ 현 재 전북대학교 기록관리학과 석사과정
관심분야 : Electronic Record, Machine Learning, Bigdata



윤 성 호

https://orcid.org/0000-0001-6197-5336
e-mail : tjdgh9410@naver.com
2019년 전북대학교 문헌정보학과(학사)
2019년 ~ 현 재 전북대학교 기록관리학과 석사과정
관심분야 : Electronic Record, Preservation Format, Bigdata

오 효 정

https://orcid.org/0000-0001-8067-2832
e-mail : ohj@jbnu.ac.kr
2008년 한국과학기술원 컴퓨터공학과 (공학박사)
2000년 ~ 2015년 한국전자통신연구원 지식마이닝연구실 책임연구원
2015년 ~ 현 재 전북대학교 문헌정보학과 부교수
관심분야 : Knowledge Extraction, Text mining, Bigdata Analysis