J. Korea Water Resour. Assoc. Vol. 54, No. 12 (2021), pp. 1255-1263
doi: 10.3741/JKWRA.2021.54.12.1255
pISSN 1226-6280
eISSN 2287-6138

Developing a regional fog prediction model using tree-based machine-learning techniques and automated visibility observations

Kim, Daeha[a]*

[a]Assistant Professor, Department of Civil Engineering, Jeonbuk National University, Jeonju, Korea

Paper number: 21-072
Received: 23 September 2021; Revised: 1 November 2021; Accepted: 1 November 2021

Abstract

While it could become an alternative water resource, fog could undermine traffic safety and operational performance of infrastructures. To reduce such adverse impacts, it is necessary to have spatially continuous fog risk information. In this work, tree-based machine-learning models were developed in order to quantify fog risks with routine meteorological observations alone. The Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), and Random Forests (RF) were chosen for the regional fog models using operational weather and visibility observations within the Jeollabuk-do province. Results showed that RF seemed to show the most robust performance to categorize between fog and non-fog situations during the training and evaluation period of 2017-2019. While the LGB performed better than in predicting fog occurrences than the others, its false alarm ratio was the highest (0.695) among the three models. The predictability of the three models considerably declined when applying them for an independent period of 2020, potentially due to the distinctively enhanced air quality in the year under the global lockdown. Nonetheless, even in 2020, the three models were all able to produce fog risk information consistent with the spatial variation of observed fog occurrences. This work suggests that the tree-based machine learning models could be used as tools to find locations with relatively high fog risks.

Keywords: Fog occurrence, Tree-based machine-learning, Regional prediction

시정계 자료와 기계학습 기법을 이용한 지역 안개예측 모형 개발

김대하[a]*

[a]전북대학교 토목환경자원에너지공학부, 조교수

요 지

안개는 대체수자원이 될 수 있으나 교통사고 위험을 높이고 공항 운영에 제약을 가하는 사회적 영향이 큰 기상현상이다. 본 연구에서는 1 km 미만 가시거리(시정)로 정의되는 안개 발생을 기상자료로 예측하는 지역 기계학습모형을 개발하고 그 예측력을 평가하였다. 전라북도 지역의 10개 기상청 지상관측소의 2017-2019년 시정 및 기상관측자료로 앙상블 분류기법인 Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), Random Forests (RF)를 학습시켜 지역 안개 모형을 개발하였고 독립적인 2020년 자료로 모형의 사용성을 평가하였다. 그 결과, 학습·검증기간(2017-2019)에는 True Skill Score를 기준으로 가장 높은 예측력을 보인 방법은 LGB 기법이었지만 다른 두 모형에 비해 False Alarm Ratio가 컸다. RF 모형과 XGB 방법 역시 기존 연구에 상응하는 예측성능을 보이는 것으로 확인되었다. 2020년 자료를 입력해 안개 발생을 모의했을 때 세 모형의 예측성능은 2017-2019년 기간보다 떨어졌지만 모두 관측 안개일수의 공간분포와 일관되는 안개 위험을 예측했다. 세 기계학습 모형은 안개위험이 상대적으로 높은 지역을 추출하는 기법으로 사용이 가능할 것으로 보인다.

핵심용어: 안개 발생예측, 의사결정나무 기반 기계학습, 지역예측모형

***Corresponding Author.** Tel: +82-63-270-2426
E-mail: daeha.kim@gmail.com (D. Kim)

1. 서 론

안개는 가시거리(시정)를 줄여 교통수단의 운영을 제한하고 사고위험을 높여 폭우, 폭염과 같은 직접적인 피해를 주는 기상재해보다 사회에 더 큰 영향을 줄 수 있는 기상현상이다(Lee and Suh, 2019; Bari and Bergot, 2018; Rebollo and Balakrishnan, 2014; Haeffelin *et al*., 2010). 미연방 고속도로국(Federal Highway Administration, FHWA)에 의하면 안개로 인한 사고는 미국에서 연간 25,500건에 이르고 이에 의한 사망건수는 460명이 넘는 것으로 집계되었다(FHWA, 2018). 2015년 한국 영종대교에서 발생한 106중 추돌사고는 안개의 치명적인 영향을 보여주는 대표적인 예이고 교통 문제 외에도 안개는 사람의 건강에 까지 영향을 미치는 것으로 알려져 있다(Bartoková *et al*., 2015; Gultepe *et al*., 2007; Niu *et al*., 2010). 수자원 분야에서도 안개는 인구증가를 대비할 대체수자원으로 관심을 받고 있는 재산림화(Reforestration)의 재원으로 전망되고 있다(Domen *et al*., 2014).

세계기상기구(World Meteorological Organization, WMO)는 응결된 수증기 입자로 인해 수평시정거리가 1 km 미만으로 감소한 경우를 안개로 정의하고 있고 시정거리가 1 km 이상 5 km 미만인 경우도 미스트(Mist)로 분류하여 위험이 심각해지기 전 단계를 구분한다. 안개는 지면이 이슬점온도 이하로 빠르게 냉각해 발생하는 복사안개(Radiation fog), 습윤한 공기가 차가운 지면 위로 이동해 발생하는 이류안개(Advection fog), 전선면 부근에서 약한 비가 내릴 때 발생하는 강수안개(Precipitation fog) 등으로 분류되고 한국 내륙지역의 짙은 안개는 주로 복사과정에 발생하는 것으로 알려져 있다(Lee and Suh, 2019).

안개발생 예측을 위해 수치기상모형(e.g., Skamarock *et al*., 2019)을 이용할 수 있으나 안개형성 과정은 다양할 뿐 아니라 미세한 기상변화에도 매우 민감하기 예측에 상당한 어려움이 따른다(Steeneveld *et al*., 2015). 미세 기상현상을 정확히 모의하기 위해서는 매우 높은 해상도로 수치모의를 수행해야 하기 때문에 상당히 큰 계산비용이 따르게 된다(Herman and Schumacher, 2016). 수치기상모형과 간단한 single-column 모형을 결합하는 방식으로 이를 극복하려는 노력이 있긴 했지만(e.g., Fedorova *et al*., 2013), 작은 초기조건의 변화에 매우 민감한 수치기상모형의 고유한 불확실성을 피하기는 여전히 어렵다(Lorenz, 1965).

이러한 수치기상모형의 단점을 보완하기 위해 최근에는 기계학습(Machine-learning) 기법이 안개예측에 자주 사용되고 있다. 예를 들어 Bari and Ouagbi (2020)는 의사결정나무(Classification And Regression Tree (CART); Breiman *et al*., 1984) 기반 앙상블(ensemble) 기법, 신경망(Neural network) 모형, 일반선형이론(Generalized linear method)을 이용해 모로코 전역에 적용할 수 있는 지역 시정예측모형을 개발하였고 모형의 뛰어난 성능을 확인하였다. 기상예측 결과를 입력자료로 사용해 가시거리를 예측한 경우로 앞서 수치기상모형의 단점을 크게 개선한 사례이다. 특히 CART 기반 기법들은 기상자료로 안개발생유무를 구분해 내는데 상당히 우수한 성능을 보이는 것으로 확인되고 있다(e.g., Ortega *et al*., 2019; Cornejo-Bueno *et al*., 2017; Bartoková *et al*., 2015).

그러나 여전히 기계학습 모형의 1시간 선행 예측력은 통상 70% 이하이고 선행시간이 길어질수록 예측력은 떨어지기 때문에 예측 자료를 직접 인프라 운영에 적용하는 것은 섣부른 판단일 수 있다. 대신 기계학습 모형은 안개발생 확률을 비교적 높은 신뢰도로 추정하기 때문에 안개위험지도 작성과 같은 지역화의 목적으로는 사용이 가능해 보인다. 미계측지역의 기상변수를 기계학습 기법으로 공간 예측한 예는 상당히 흔하고 그 정확도 역시 매우 높다. 설치·운영 비용이 매우 높아 관측밀도가 높지 않은 에너지 플럭스타워 자료를 기계학습 모형으로 학습시켜 전지구 에너지 플럭스를 추정한 경우가 대표적이다(e.g., Zeng *et al*., 2020; Tramontana *et al*., 2016; Jung *et al*., 2011). 대기와 지면의 상호작용으로 발생하는 안개 역시 동일한 방식으로 위험을 추정할 수 있을 것으로 판단되나 이를 적용한 연구는 아직 찾기 어렵다.

공간적으로 연속적인 안개 위험을 산정하기 위해서는 먼저 지상관측지점의 시정자료와 기상자료 사이의 경험적 관계를 추출해야 한다. 이를 위해 본 연구에서는 1 km 이하 시정발생을 구분할 수 있는 세 가지 CART 기반 앙상블 기계학습 기법을 개발하였다. 한국에서 안개발생 위험이 비교적 높은 것으로 평가되는 전라북도 지역의 지상관측지점 자료를 이용해 모형을 학습시켰고 독립적인 시정관측자료를 이용해 모형의 사용성을 평가하였다.

2. 방법 및 자료

2.1 대상지점 및 자료

지역 안개예측 기계학습 모형 개발을 위해 먼저 전라북도에 위치한 10개 기상청 종관 지상관측소(Automated Synoptic Obseravation System, ASOS)의 1시간 단위 시정자료와 기상자료를 수집하였다(Fig. 1). 전라북도 동남부는 가을철에 안개 사상이 20회 이상 관측되는 지역으로 안개 위험이 한국의

다른지역에 비해 비교적 높다(Lee and Suh, 2019). 기계학습 안개발생 예측에 사용되는 통상적인 기상인자는 복사형 안개발생에 영향을 주는 요소(대기온도, 습도, 지면온도, 운고, 운량, 토양수분 등)와 이류형 안개발생과정에 영향을 주는 요소(풍향, 풍속 등)를 포함한다(Bari and Ouagabi, 2020; Cornejo-Bueno *et al*., 2017; Bartoková *et al*., 2017). 이에 따라 ASOS 지점들의 시간별 기온, 강수, 상대습도, 지면온도, 풍속, 풍향, 대기압, 일조시간 자료를 수집하였고 시정계 관측자료가 배포되기 시작한 2017년 이후 자료를 이용해 분석을 수행하였다.

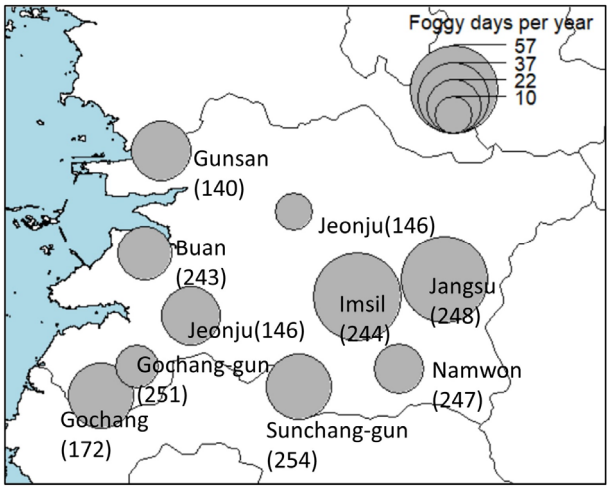

Fig. 1. The locations of the meteorological stations and mean annual foggy days for 2017-2020. The foggy days are defined as days with at least a single hourly record of visibility < 1 km. The numbers in the parentheses are the station identifiers of the Korea Meteorological Administration

수집된 시정관측자료를 이용해 지점별 안개발생 통계특성을 먼저 확인하였는데 안개발생기준인 1 km 미만 시정이 하루에 한 차례 이상이라도 발생한 경우를 안개발생일로 정의했을 때 최대 안개발생 지점은 임실(244)지점으로 2017-2020년 4년 평균 57.0일이 발생하였다. 인근에 위치한 장수(248) 지점과 고창(254) 지점도 연평균 55.0일의 안개일수가 나타나 높은 고도로 인해 상대적으로 기온이 낮고 강수량이 많은 곳에서 안개위험이 높은 것이 확인되었다. 계절적으로는 우기 동안 토양수분이 증가한 후 일교차가 커지는 가을철(9-11월)에 안개일수가 많았고(Fig. 2(a)) 대기가 냉각되는 오전 2시에서 6시에 안개 발생빈도가 증가하며 일출 후 지면이 따뜻해지기 시작하면 안개 위험이 빠르게 감소하는 것으로 나타났다(Fig. 2(b)).

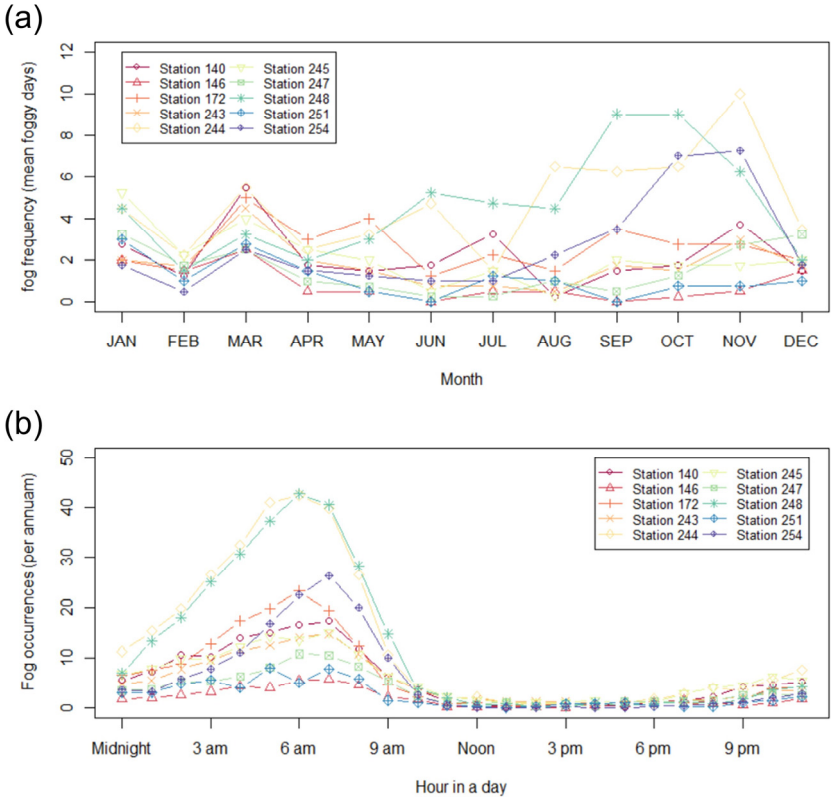

Fig. 2. (a) Monthly mean occurrences of foggy days, and (b) diurnal variations of fog occurrences at each stations during 2016-2020

2.2 기계학습 기법

안개발생(1 km 미만 시정발생) 예측을 위해 사용된 기계학습 기법은 Extreme Gradient Boosting (XGB; Chen and Guestrin, 2016), Light Gradient Boosting (LGB; Ke *et al*., 2017), Random Forests (RF; Breiman, 2001)이다. 세 모형은 모두 다수의 의사결정나무(CART)를 학습자료에 적합시킨 후 투표나 평균으로 종합하여 예측 오차를 줄이는 방법으로 환경재해 분야에 다양한 목적으로 사용된다(Rhee *et al*., 2020; Fan *et al*., 2019).

CART는 분류나 회귀예측을 위해 사용되는 가장 기본적인 기계학습 모형으로 같은 종류로 분류된 자료들의 동질성은 다른 종류로 분류된 자료들의 이질성은 커질도록 의사결정나무를 성장시키는 방법이다. 자료분류는 노드(node) 내 자료 불순도(impurity)를 최소화하는 방향으로 통상 이루어지며 적합된 결정나무는 직관적 해석으로 해석이 가능하기 때문에 분류기준을 파악하는데 매우 유용한 방법이다. 하지만 CART 모형은 새로운 자료가 추가될 때마다 분류기준이 쉽게 달라져 학습자료에 과적합(overfitting) 되는 경향이 있고 이에 따라 예측력이 학습자료에 따라 민감하게 변하는 단점이 있다(Krzywinski and Altman, 2017).

Gradient Boosting (Friedman, 2002)은 예측력이 낮은 초기 CART 모형(weak learner)의 오차를 반복 예측하는 기법으로 손실함수(loss function)의 negative gradient를 최소화시키는 방향으로 반복 학습시킨다. 모형의 구조상 이 기법은 과적합의 위험이 큰데 XGB 방법은 손실함수에 Regularization 항을 추가하여 복잡한 모형에 페널티를 주는 방식으로 이를 방지하는 특징을 가지고 있다. Regularization 항에 추가적으로 요구되는 계산시간 때문에 XGB는 일반적인 Gradient boosting보다 더 긴 학습시간이 요구되지만 과적합 위험은 낮아진다. LGB는 XGB가 가지는 학습의 비효율성을 Leaf-wise loss function으로 개선시킨 기법으로 level 단위로 결정나무를 성장시키는 XGB에 비해 학습과정에서의 손실을 줄이고 계산시간을 크게 단축시킨다(Ke *et al*., 2017). 높은 학습의 효율성과 예측의 정확성을 가지고 있어 수문분야에서도 그 사용이 점차 늘고 있는 방법이다(e.g., Fan *et al*., 2019).

RF모형은 CART 모형의 단점을 개선하는 전통적인 기법으로 학습데이터를 임의추출하여 여러 개의 CART 모형을 만든 후 각 모형의 예측 모아 투표나 평균으로 최종 예측을 수행하는 기법이다. 예측인자를 변화시키면서 CART 모형을 만들어내기 때문에 예측인자를 고정시키는 Bagging (Breiman, 1996) 방법과는 차이가 있으며 앙상블 효과로 예측력이 개별 CART 모형보다 높아지게 된다. 여러 모형의 평균을 취하기 때문에 단일 CART모형이 가지는 과적합의 위험을 크게 낮추지만 (Breiman, 2001), 오차를 줄이는 방향으로 개별 CART 모형을 생산하지 않기 때문에 학습오차는 0으로 수렴하지 않는다.

2.3 안개발생 모형 개발 및 예측력 평가

CART 기반 앙상블 기계학습 모형에 사용된 안개예측인자는 기온, 상대습도, 지면-기온 간 온도차, 5시간 기온변화량, 5시간 강수유무, 24시간 누적 일조시간, 풍속, 풍향, 현지기압, 위도, 경도, 표고, 관측시간, 지면분류(inland or coastal)이다. 강수유무는 Lee and Suh (2019)의 안개분류 방법인 5시간 누적강수량으로 판단하였고 기온변화율 역시 한 시점과 5시간 이전의 온도의 차이로 산정하였다. 강우강도는 시정에 영향을 주기는 하지만 5 km 이내로 시정이 줄어들게 할 가능성은 상당히 낮기 때문에(Bari and Ouagabi, 2020) 누적강수량은 강수유무의 판단 지표로만 사용하였다. 또한 복사안개의 발생은 대기의 맑은 정도에 영향을 받기 때문에 24시간 누적 일조시간 합계치를 이용해 이를 나타내는 지표로 사용하였다. 운량이나 운고 관측자료가 맑은 정도를 나타내는 더 직관적인 자료이기는 하지만 계측되고 있는 지점이 많지 않아 간접적인 예측인자로 누적 일조시간을 사용하였다. 아울러 Lee and Suh (2019)의 분류를 그대로 적용해 10개 지점 중 군산(140) 지점을 해안(Coastal) 지점으로 나머지 9개 지점은 모두 내륙지점(inland)으로 분류하였다. 강수유무와 지면분류를 제외한 모든 연속(continuous) 예측인자는 절대수치 범위가 모형의 예측력에 영향을 주지 않도록 각각의 최대값과 최소값으로 표준화하였다.

모형의 학습과 검증에는 2017-2019년 시정자료와 기상자료가 사용되었고 2020년 자료는 독립적인 기간으로 설정하여 모형의 사용성을 평가하였다. 학습·검증자료는 수문기상연구에서 통상적인 8:2룰을 적용해(Yang *et al*., 2016) 전체 자료의 80%를 임의추출해 각 모형을 학습시켰고 나머지 20% 자료로 예측력 검증을 수행하였다. 전라북도 지역의 안개 발생일수는 통상 연간 60일을 넘지 않고 오전시간에 주로 발생하기 때문에 관측시정이 5 km 이상인 경우(Normal)가 대부분(98.3%)을 차지하게 되는데 Class 불균형(imbalance) 문제를 보완하기 위해 Synthetic Minority Oversampling Technique (SNOTE)을 적용하였다. Fog, Normal의 경우의 수가 균형잡힐 수 있도록 Fog의 경우를 imbalance ratio를 이용해 55배 많아지도록 Nearest Neighbor 기법으로 oversampling 하여 학습자료를 구성하였다. 모형의 예측성능은 Marzban (1998)의 평가지표인 bias (BIAS), probability of detection (POD), false alarm ration (FAR), percent correct (PC), probability of false detection (POFD), true skill score (TSS)를 사용하였다.

$$BIAS = \frac{b+d}{c+d} \tag{1a}$$

$$POD = \frac{d}{c+d} \tag{1b}$$

$$FAR = \frac{b}{b+d} \tag{1c}$$

$$PC = \frac{a+d}{a+b+c+d} \tag{1d}$$

$$POFD = \frac{b}{a+b} \tag{1e}$$

$$TSS = POD - POFD \tag{1f}$$

여기서, a는 안개가 예측되지 않았을 때 일어나지 않은 경우(correct negatives), b는 안개가 예측됐을 때 안개가 일어나지 않은 경우(false alarms), c는 안개가 예측되지 않았을 때 안개가 일어난 경우(misses), d는 안개가 예측됐을 때 안개가 일어난 경우의 수(hits)이다.

3. 결과 및 고찰

3.1 지역 안개기계학습모형 예측력 평가

Table 1은 XGB, LGB, RF 세 모형의 2017-2019 기간의 20% 검증자료로 안개발생을 예측한 후 관측자료와 비교한 평가결과이다. PC 기준으로는 세 기계학습 모형 모두 96% 이상의 분류 정확도를 보였다. TSS로 평가했을 때 XGB, LGB, RF 세 모형의 성능은 최근 연구에 비해 비교적 높게 나타났다. 예를 들어 Bari and Ouagabi (2020)는 30개가 넘는 예측인자를 사용했지만 본 연구의 세 기계학습 모형보다 TSS 값이 낮았다. 물론 이 연구에서는 5 km 시정을 기준으로 안개예측성능을 평가했기 때문에 이에 따른 예측력 저하를 고려해야 한다. 다른 연구로 Bartoková *et al*. (2015)의 두바이 지역 CART기반 안개 Nowcasting 모형(선행시간 최대 6시간)을 들 수 있는데 기계학습 모형과 WRF 기상모의 자료를 결합했을 때 안개예측성능은 POD 0.88 정도였다. 하지만 선행 visibility 자료를 예측인자로 사용한 경우로 본 연구의 세 모형과 직접 성능을 비교하기는 어려운 경우이다.

유사하게 Miao *et al*. (2020)은 Long short-term memory (LSTM) 모형을 이용해 중국 Anhui 지역의 안개예측을 수행했는데 짧은 선행시간(2시간 이하) 예측성능은 TSS 0.60 정도였다. 예측선행시간이 길어질 때 성능이 줄어드는 정도가 AdaBoost 기법(Friedman *et al*., 2000)을 비롯한 다른 기계학습 모형보다 작은 것이 장점인 모형이다. 하지만 Miao *et al*. (2020)에서도 선행 시정거리가 안개 예측에 직접 사용되어 본 연구의 실시간 예측의 경우와 직접 비교하기는 어렵다. 반면 Cornejo-Bueno *et al*. (2017)은 Multi-layer Perceptron와 Gaussian Process 모형으로 본 연구와 유사하게 기상자료만으로 1 km 이하 가시거리 구분하였는데 1시간 선행 예측성능은 TSS 0.69와 0.79 정도였다. 다만 이 경우는 지역 모형이 아닌 스페인 Valladolid 공항자료만을 이용한 지점 모형이다. 유사한 방법으로 Ortega *et al*. (2019)는 플로리다주 올랜도 지역 Visibility 분류모형을 개발했는데 PC 기준 최대 0.89 정도의 성능을 보였다. Nowcasting을 목적으로 진행된 기존 연구와 예측성능을 직접 비교하기는 어렵지만 개발된 모형들의 예측 선행시간에 따른 성능 변화를 고려했을 때 XGB, LGB, RF 세 모형의 성능은 기존 모형 수준이거나 그 이상인 것으로 보인다.

Fig. 3는 XGB, LGB, RF 모형의 각 지점별(TSS) 성능 분포이다. 전반적으로 안개발생일이 많은 지역에서 예측력이 높은 것을 확인할 수 있다. 이는 안개 위험이 상대적으로 높아 데이터 불균형이 적은 지역일수록 모형의 성능이 높음을 의미한다. 유일하게 해안(Coastal) 지점으로 분류된 군산(140)지점에서는 Boosting 모형(XGB와 LGB)의 성능이 RF보다 높고 내륙지역에서는 RF모형의 성능이 비교적 견고(robust)하게 나타나는 것을 알 수 있다. LGB 모형의 상대적으로 높은 FAR을 고려했을 때 전통적인 RF의 예측능력은 최신 기계학습 기법보다 낮지 않은 것으로 보인다.

Table 1. Performance metrics of RF, XGB, and LGB fog prediction models

	Confusion Matrix Comp. (%)				BIAS	POD	FAR	PC	POFD	TSS
	a	b	c	d						
XGB	97.65	0.51	0.72	1.13	0.888	0.612	0.311	0.988	0.005	0.607
LGB	94.65	3.5	0.31	1.54	2.733	0.833	0.695	0.962	0.036	0.797
RF	98.13	0.29	0.42	1.16	0.916	0.732	0.201	0.993	0.003	0.729

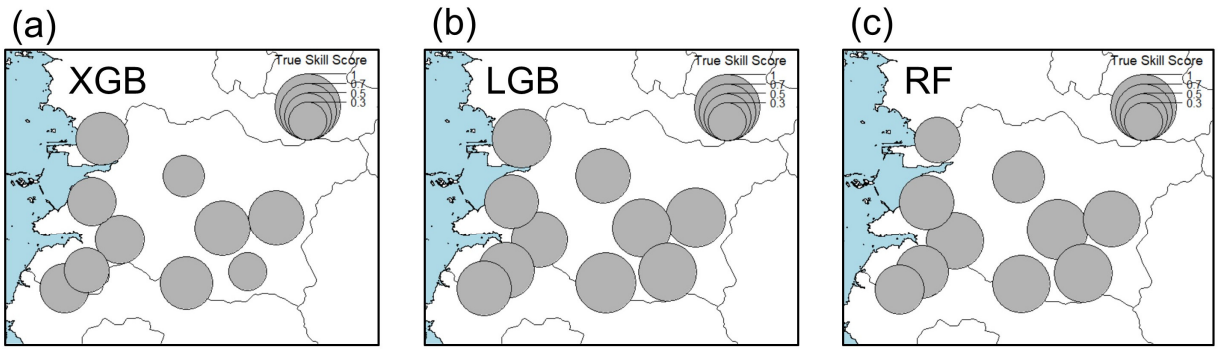

Fig. 3. True skill scores (POD minus POFD) at each station by (a) XGB, (b) LGB, and (c) RF

Table 2. Relative importance of predictors to fog occurrence forecasts for each model measured by information gain

Rank	Important predictors (Percentage of importance)		
	XGB	LGB	RF
1	Relative humidity (50.8)	Relative humidity (52.3)	Relative humidity (32.5)
2	Air temperature (11.6)	Air temperature (11.6)	Air temperature (10.5)
3	Sunshine hours (7.0)	Sunshine hours (7.2)	Hour (9.8)
4	Atmospheric pressure (5.5)	Surface-air temp. diff. (5.2)	Sunshine hours (7.2)
5	Surface-air temp. diff. (5.4)	Atmospheric pressure (4.5)	Wind speed (7.0)
6	Elevation (3.6)	Elevation (4.4)	Surface-air temp. diff. (6.4)
7	5-hour temp. change (3.6)	5-hour temp. change (2.9)	Atmospheric pressure (5.2)

3.2 예측인자 중요도 평가

CART 기반 모형은 Information gain이나 오차 감소량를 이용해 예측인자의 중요도를 평가할 수 있는 장점을 가지고 있다. 각 모형의 15개 예측인자 중 중요도가 높은 상위 7개 변수를 Table 2에 나타냈다. 동시간 안개예측에서 가장 중요한 인자는 세 모형에서 모두 상대습도로 나타났으며 상대적 중요도는 최대 52.3% 였다. 상대습도는 이슬점온도와 응결 수증기의 양을 동시에 결정하는 요소이기 때문에 이슬점에만 영향을 주는 기온보다 훨씬 중요도가 더 높은 것은 당연한 결과이다. 풍속과 풍향의 중요도는 상대적으로 낮았는데 이는 때 전북지역에서 이류안개의 발생이 많지 않음을 의미한다. 9개 지점이 내륙지점으로 분류되고 한국 내륙지역에서 발생하는 안개의 71.3%가 복사안개인 것으로 판단했을 때(Lee and Suh, 2019) 풍속과 풍향이 주요 예측인자로 나타나지 않은 것은 예상된 결과이다. 이에 반해 복사안개 발생에 영향을 상당히 미치는 지면-대기 온도차와 24시간 누적 일조시간은 세 모형 모두에서 중요도가 높게 나타났다. 예측인자의 중요도로 판단했을 때 지역 안개모형이 주로 예측하는 안개의 종류는 복사안개로 보인다.

XGB, LGB, RF 안개모형 중요도(Table 2)는 Bari and Ouagabi (2020)의 모로코 지역 연구와는 상당히 다른 결과이다. Bari and Ouagabi (2020)의 연구에서는 XGB 모형의 경우 태양복사량과 해면기압이 가장 중요한 인자로, RF 모형의 경우는 표고와 경도가 가장 중요한 예측인자였다. 이는 모로코 지역에서 가시거리는 대기에 수분을 공급하는 바다와 모래바람을 일으키는 사하라사막에 인접한 정도에 따라 변하기 때문으로 판단된다. 다시말하면 위치정보 자체가 안개예측에 중요한 역할을 한 것이다. 그러나 상대습도는 안개 분류에 가장 중요한 조건 중 하나이고(Lee and Suh, 2019; Tardif and Rasmunssen, 2007), Bari and Ouagabi (2020)의 XGB 모형이 이를 주요 예측인자로 포함하지 않은 것은 상당히 예외적이라 할 수 있다. 이류형 안개가 많이 발생하는 지역에서 긴 선행시간으로 안개를 예측할 때 상대습도의 중요도는 떨어질 수 있지만, 주요 예측인자로 전혀 포함되지 않는 것은 이례적이다.

Table 2의 중요도는 지상 관측자료가 아닌 기상 재분석 자료(Hersbach *et al*., 2020)를 사용할 때 안개 예측인자 선택에 도움을 줄 수 있다. 예를 들어 일조시간은 날씨가 맑은 정도를 나타내는 간접인자로 사용되었는데 상대적 중요도가 상당히 높게 나타났기 때문에 재분석 자료를 사용할 경우 Cloud cover와 같은 직접적인 인자로 대체될 수 있다. 유사하게 현열과 잠열자료로 지면-대기 온도차를 대신해 재분석자료의 현열자료를 사용할 수도 있다.

Table 3. Performance evaluation of fog predictions by XGB, LGB, and RF models for the year of 2020

	Confusion Matrix Comp. (%)				BIAS	POD	FAR	PC	POFD	TSS
	a	b	c	d						
XGB	97.33	1.22	0.92	0.54	1.204	0.368	0.694	0.979	0.012	0.356
LGB	94.30	4.25	0.61	0.84	3.500	0.580	0.834	0.951	0.043	0.537
RF	97.76	0.79	1.06	0.39	0.812	0.271	0.666	0.982	0.008	0.263

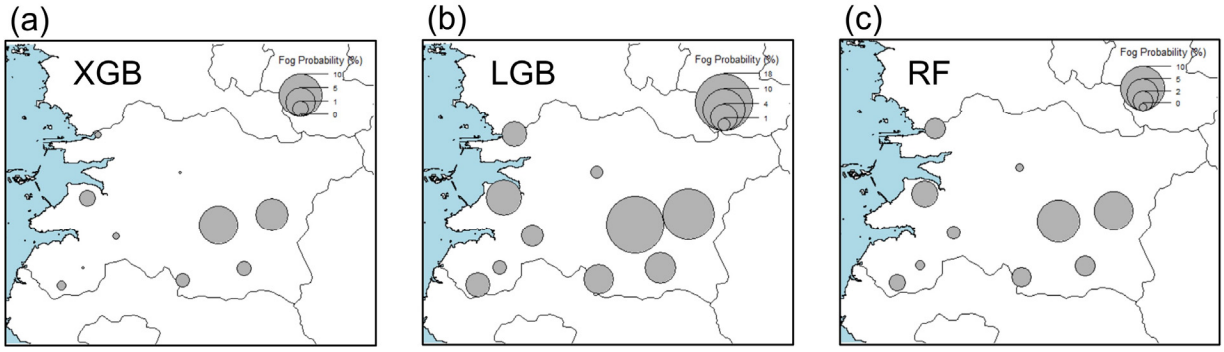

Fig. 4. Fog occurrence probabilities at each stations predicted by the (a) XGB, (b) LGB, and (c) RF

3.3 2020년 안개발생 예측결과

Table 3은 XGB, LGB, RF 모형으로 학습과 검증에 사용되지 않았던 2020년 기상자료를 이용해 안개 발생을 모의한 후 예측성능을 평가한 결과이다. 2017-2019 기간 검증자료로 평가한 결과에 비해 PC는 크게 감소하지는 않았지만 세 모형의 POD가 상당히 줄어든 것이 확인되었다. 테스트 자료를 학습·검증자료와 시간적으로 완전히 분리시켰을 때는 특히 RF 모형의 예측성능이 Boosting 기법에 비해 크게 떨어졌다. 여전히 세 모형 중 LGB모형이 가장 높은 TSS 값을 보였지만 여전히 FAR이 다른 모형보다 높아 공항 Traffic과 같은 False Alarm 위험비용이 높은 운영 목적으로는 적절하지 않을 것으로 보인다. 코로나 바이러스 Lockdown으로 대기질이 크게 개선된 2020년에는 수증기 응결이 덜 발생했을 가능성이 있고(Liu *et al*., 2021; Seo *et al*., 2020; Klemm and Lin, 2016) 이는 기상변동성만을 고려하는 세 기계학습 모형의 성능을 낮출 수 있는 요인이다. 실제로 기상청 자료에 의하면 전북지역의 2020년 평균 안개일수는 1.4일로 2017년-2019 평균보다 낮다.

그럼에도 불구하고 세 모형의 성능을 확률이 50% 이상일 때 무조건 안개가 발생할 것으로 판단한 Tables 1 and 3의 평가결과 만으로 그 사용성을 모두 평가할 수는 없다. Fig. 4는 세 모형으로 예측된 2020년 평균 안개확률을 나타낸다. FAR이 높은 LGB가 가장 높은 안개확률을 예측했고 XGB와 RF는 좀 더 낮게 안개확률을 예측하고 있다. 각 모형의 안개발생 확률은 다르지만 모두 도시지역(전주)의 안개확률은 낮게 고도

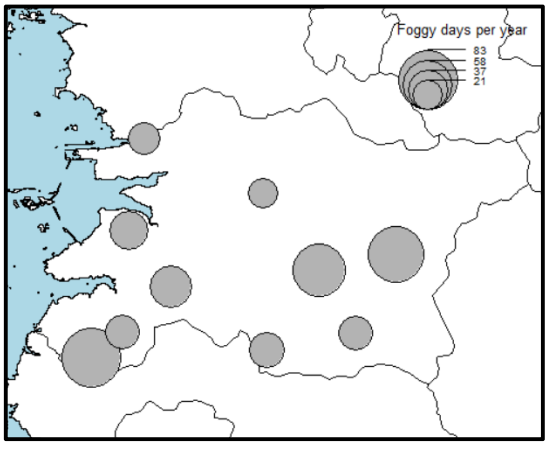

Fig. 5. The distribution of foggy days in 2020. The definition of foggy days are the same as in Fig. 1

가 높은 지점과 해안선에 가까운 지점의 안개확률은 높게 추정하는 것이 쉽게 확인된다. 도시지역의 열섬현상은 안개위험을 크게 낮추고(Gautam and Singh, 2018) 강수량과 토양수분이 많은 곳은 지면-대기 상호작용으로 상대습도도 높아(Qualls and Crago, 2020) 안개 위험이 높게 나타날 수 있다. 다시 말하면 지형 특성과 인구통계학적 특성 역시 온도와 상대습도 자료에 영향을 주기 때문에 기상외적 요소 역시 모형에 어느 정도는 반영되어 있다고 볼 수 있다. 아울러 대기질에 큰 변화가 있었다 하더라도, 2020년 안개확률의 공간분포는 실제 안개발생일 분포(Fig. 5)와 일관됨을 확인할 수 있다. 따라서 안개

위험의 공간적 분포를 파악하는 정도의 목적으로 세 기계학습은 충분히 사용될 수 있을 것으로 판단된다.

4. 결 론

본 연구에서는 2017-2020년 전북지역 10개 기상청 ASOS 지점의 기상관측자료와 시정관측자료를 이용하여 의사결정나무 기반 Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), Random Forests (RF) 기계학습 기법의 동시간 안개예측 성능을 평가하였다. 세 모형의 성능 평가결과는 다음과 같이 요약된다.

1) 2017-2019년 자료기간을 8:2로 나눠 세 기계학습 모형을 학습시키고 성능을 평가한 결과 True Skill Score를 기준으로 가장 높은 예측력을 보인 방법은 LGB 기법이었다. 하지만 LGB 모형은 다른 두 모형에 비해 False Alarm Ratio가 상당히 컸고 RF 모형의 성능이 비교적 견고하게 나타났다.
2) 안개발생 예측에 가장 중요한 인자는 세 모형 모두에서 상대습도로 확인되었고 다른 예측인자보다 월등히 그 중요도가 높았다. 10개 지점 중 9개 지점이 복사안개가 지배적으로 발생하는 내륙에 위치하기 때문이고 이류형 안개의 예측인자인 풍향과 풍속은 10개 지점에 대해서는 크게 중요한 인자가 아닌 것으로 확인되었다.
3) 대기질이 크게 개선된 2020년 기상자료로 안개 발생을 모의했을 때 세 모형의 예측성능은 상당히 떨어지는 것으로 확인되었다. 하지만 세 모형 모두 관측 안개일수 일관된 안개확률 분포를 예측했기 때문에 기상학적 안개위험이 상대적으로 높은 지역을 추출하는 기법으로는 충분히 사용이 가능할 것으로 보인다.

감사의 글

본 연구는 국토교통과학기술진흥원 국토교통지역혁신기술개발사업(21RITD-C162665-01)로 수행되었으며, 이에 관계자 분들께 감사드립니다.

References

Bari, D., and Bergot, T. (2018). “Influence of environmental conditions on forecasting of an advection-radiation fog: A case study from the Casablanca region, Morocco.” *Aerosol and Air Quality Research*, Vol. 18, pp. 62-78.

Bari, D., and Ouagabi, A. (2020). “Machine-learning regression applied to diagnose horizontal visibility from mesoscale NWP model forecasts.” *SN Applied Sciences*, Vol. 2, No. 4, pp. 1-13.

Bartoková, I., Bott, A., Bartok, J., and Gera, M. (2015). “Fog prediction for road traffic safety in a coastal desert region: Improvement of nowcasting skills by the machine-learning approach.” *Boundary-Layer Meteorology*, Vol. 157, No. 3, pp. 501-516.

Breiman, L. (1996). “Bagging predictors.” *Machine Learning*, Vol. 24, No. 2, pp. 123-140.

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

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*. Chapman and Hall/CRC, Boca Raton, FL, U.S.

Chen, T., and Guestrin, C. (2016). “Xgboost: A scalable tree boosting system.” *Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining, The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, San Francisco, CA, U.S., pp. 785-794.

Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J., Cerro-Prada, E., and Salcedo-Sanz, S. (2017). “Efficient prediction of low-visibility events at airports using machine-learning regression.” *Boundary-Layer Meteorology*, Vol. 165, No. 2, pp. 349-370.

Domen, J.K., Stringfellow, W.T., Camarillo, M.K., and Gulati, S. (2014). “Fog water as an alternative and sustainable water resource.” *Clean Technologies and Environmental Policy*, Vol. 16, No. 2, pp. 235-249.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., and Zeng, W. (2019). “Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data.” *Agricultural Water Management*, Vol. 225, 105758.

Federal Highway Administration (FHWA) (2018). Road weather management program, “How do weather events impacts roads?”, Accessed 4 September 2021, <https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm>.

Fedorova, N., Levit, V., Da Silva, A.O., and dos Santos, D.M.B. (2013). “Low visibility formation and forecasting on the northern coast of Brazil.” *Pure and Applied Geophysics*, Vol. 170, No. 4, pp. 689-709.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). “Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors).” *The Annals of Statistics*, Vol. 28, No. 2, pp. 337-407.

Friedman, J.H. (2002). “Stochastic gradient boosting.” *Computational Statistics & Data Analysis*, Vol. 38, No. 4, pp. 367-378.

Gautam, R., and Singh, M.K. (2018). “Urban heat island over Delhi punches holes in widespread fog in the Indo Gangetic Plains.” *Geophysical Research Letters*, Vol. 45, No. 2, pp. 1114-1121.

Gultepe, I., Tardif, R., Michaelides, S.C., Cermak, J., Bott, A., Bendix,

J., Müller, M.D., Pagowski, M., Hansen, B., Ellrod, G., and Jacobs, W. (2007). "Fog research: A review of past achievements and future perspectives." *Pure and Applied Geophysics*, Vol. 164, No. 6, pp. 1121-1159.

Haeffelin, M., Bergot, T., Elias, T., Tardif, R., Carrer, D., Chazette, P., Colomb, M., Drobinski, P., Dupont, E., Dupont, J.C., and Gomes, L. (2010). "PARISFOG: Shedding new light on fog physical processes." *Bulletin of the American Meteorological Society*, Vol. 91, No. 6, pp. 767-783.

Herman, G.R., and Schumacher, R.S. (2016). "Using reforecasts to improve forecasting of fog and visibility for aviation." *Weather and Forecasting*, Vol. 31, No. 2, pp. 467-482.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., and Simmons, A. (2020). "The ERA5 global reanalysis." *Quarterly Journal of the Royal Meteorological Society*, Vol. 146, No. 730, pp. 1999-2049.

Jung, M., Reichstein, M., Margolis, H.A., Cescatti, A., Richardson, A.D., Arain, M.A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., and Gianelle, D. (2011). "Global patterns of land atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations." *Journal of Geophysical Research: Biogeosciences*, Vol. 116, G00J07.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.Y. (2017). "Lightgbm: A highly efficient gradient boosting decision tree." *Advances In Neural Information Processing Systems*, Vol. 30, pp. 3146-3154.

Klemm, O., and Lin, N. (2016). "What causes observed fog trends: Air quality or climate change?" *Aerosol and Air Quality Research*, Vol. 16, No. 5, pp. 1131-1142.

Krzywinski, M., and Altman, N. (2017). "Classification and regression trees." *Nature Methods*, Vol. 14, No. 8, pp. 757-758.

Lee, H.K., and Suh, M.S. (2019). "Objective classification of fog type and analysis of fog characteristics using visibility meter and satellite observation data over South Korea." *Atmosphere Korean Meteorological Society*, Vol. 29, No. 5, pp. 639-658.

Liu, F., Wang, M., and Zheng, M. (2021). "Effects of COVID-19 lockdown on global air quality and health." *Science of the Total Environment*, Vol. 755, 142533.

Lorenz, E.N. (1965). "A study of the predictability of a 28-variable atmospheric model." *Tellus*, Vol. 17, pp. 321-333.

Marzban, C. (1998). "Scalar measures of performance in rare-event situations." *Weather and Forecasting*, Vol. 13, No. 3, pp. 753-763.

Miao, K.C., Han, T.T., Yao, Y.Q., Lu, H., Chen, P., Wang, B., and Zhang, J. (2020). "Application of LSTM for short term fog forecasting based on meteorological elements." *Neurocomputing*, Vol. 408, pp. 285-291.

Niu, F., Li, Z., Li, C., Lee, K.H., and Wang, M. (2010). "Increase of wintertime fog in China: Potential impacts of weakening of the Eastern Asian monsoon circulation and increasing aerosol loading." *Journal of Geophysical Research: Atmospheres*, Vol. 115, D00K20.

Ortega, L., Otero, L.D., and Otero, C., (2019). "Application of machine learning algorithms for visibility classification." *2019 IEEE International Systems Conference (SysCon)*, IEEE, Orlando, FL, U.S., pp. 1-5.

Qualls, R.J., and Crago, R.D. (2020). "Graphical interpretation of wet surface evaporation equations." *Water Resources Research*, Vol. 56, No. 10, e2019WR026766.

Rebollo, J.J., and Balakrishnan, H. (2014). "Characterization and prediction of air traffic delays." *Transportation research part C: Emerging technologies*, Vol. 44, pp. 231-241.

Rhee, J., Park, K., Lee, S., Jang, S., and Yoon, S. (2020). "Detecting hydrological droughts in ungauged areas from remotely sensed hydro-meteorological variables using rule-based models." *Natural Hazards*, Vol. 103, pp. 2961-2988.

Seo, J.H., Jeon, H.W., Sung, U.J., and Sohn, J.R. (2020). "Impact of the COVID-19 outbreak on air quality in Korea." *Atmosphere*, Vol. 11, No. 10, 1137.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Liu, Z., Berner, J., Wang, W., Powers, J.G., Duda, M.G., Barker, D.M., and Huang, X.Y. (2019). *A description of the advanced research WRF model version 4*. National Center for Atmospheric Research, Boulder, CO, U.S.

Steeneveld, G.J., Ronda, R.J., and Holtslag, A.A.M. (2015). "The challenge of forecasting the onset and development of radiation fog using mesoscale atmospheric models." *Boundary-Layer Meteorology*, Vol. 154, No. 2, pp. 265-289.

Tardif, R., and Rasmussen, R.M. (2007). "Event-based climatology and typology of fog in the New York City region." *Journal of Applied Meteorology and Climatology*, Vol. 46, No. 8, pp. 1141-1168.

Tramontana, G., Jung, M., Schwalm, C.R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M.A., Cescatti, A., Kiely, G., and Merbold, L. (2016). "Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms." *Biogeosciences*, Vol. 13, No. 14, pp. 4291-4313.

Yang, T., Gao, X., Sorooshian, S., and Li, X. (2016). "Simulating California reservoir operation using the classification and regression tree algorithm combined with a shuffled cross validation scheme." *Water Resources Research*, Vol. 52, No. 3, pp. 1626-1651.

Zeng, J., Matsunaga, T., Tan, Z.H., Saigusa, N., Shirai, T., Tang, Y., Peng, S., and Fukuda, Y. (2020). "Global terrestrial carbon fluxes of 1999-2019 estimated by upscaling eddy covariance data with a random forest." *Scientific Data*, Vol. 7, No. 1, pp. 1-11.