



## Development of disaster severity classification model using machine learning technique

Lee, Seungmin<sup>a</sup> · Baek, Seonuk<sup>b</sup> · Lee, Junhak<sup>c</sup> · Kim, Kyungtak<sup>d</sup> · Kim, Soojun<sup>e</sup> · Kim, Hung Soo<sup>f\*</sup>

<sup>a</sup>Master's Course, Program in Smart City Engineering, Inha University, Incheon, Korea.

<sup>b</sup>Master's Course, Program in Smart City Engineering, Inha University, Incheon, Korea.

<sup>c</sup>Master's Course, Program in Smart City Engineering, Inha University, Incheon, Korea.

<sup>d</sup>Senior Research Fellow, Department of Hydro Science and Engineering Research, Korea Institute of Civil Engineering and Building Technology (KICT), Goyang, Korea.

<sup>e</sup>Professor, Department of Civil Engineering, Inha University, Incheon, Korea.

<sup>f</sup>Professor, Department of Civil Engineering, Inha University, Incheon, Korea.

Paper number: 23-012

Received: 13 February 2023; Revised: 22 March 2023; Accepted: 24 March 2023

### Abstract

In recent years, natural disasters such as heavy rainfall and typhoons have occurred more frequently, and their severity has increased due to climate change. The Korea Meteorological Administration (KMA) currently uses the same criteria for all regions in Korea for watch and warning based on the maximum cumulative rainfall with durations of 3-hour and 12-hour to reduce damage. However, KMA's criteria do not consider the regional characteristics of damages caused by heavy rainfall and typhoon events. In this regard, it is necessary to develop new criteria considering regional characteristics of damage and cumulative rainfalls in durations, establishing four stages: blue, yellow, orange, and red. A classification model, called DSCM (Disaster Severity Classification Model), for the four-stage disaster severity was developed using four machine learning models (Decision Tree, Support Vector Machine, Random Forest, and XGBoost). This study applied DSCM to local governments of Seoul, Incheon, and Gyeonggi Province. To develop DSCM, we used data on rainfall, cumulative rainfall, maximum rainfalls for durations of 3-hour and 12-hour, and antecedent rainfall as independent variables, and a 4-class damage scale for heavy rain damage and typhoon damage for each local government as dependent variables. As a result, the Decision Tree model had the highest accuracy with an F1-Score of 0.56. We believe that this developed DSCM can help identify disaster risk at each stage and contribute to reducing damage through efficient disaster management for local governments based on specific events.

**Keywords:** Natural disaster, Standard warning criteria, Machine learning, Disaster severity classification model

## 머신러닝 기법을 이용한 재해강도 분류모형 개발

이승민<sup>a</sup> · 백선욱<sup>b</sup> · 이준학<sup>c</sup> · 김경탁<sup>d</sup> · 김수준<sup>e</sup> · 김형수<sup>f\*</sup>

<sup>a</sup>인하대학교 스마트시티공학과 석사과정, <sup>b</sup>인하대학교 스마트시티공학과 석사과정, <sup>c</sup>인하대학교 스마트시티공학과 석사과정,

<sup>d</sup>한국건설기술연구원 수자원하천연구본부 선임연구위원, <sup>e</sup>인하대학교 사회인프라공학과 교수, <sup>f</sup>인하대학교 사회인프라공학과 교수

### 요 지

최근 급격한 도시화와 기후변화에 따라 재난에 의한 피해가 증가하고 있다. 국내 기상청에서는 표준 경보(주의보, 경보)를 전국적으로 통일된 표준 경보 기준(3시간 및 12시간 최대 누적강우량)에 따라 발령하여 재해에 따른 지역별, 재난 사상별 특성이 고려되지 않은 문제점이 있다. 따라서 본 연구에서는 서울특별시, 인천광역시, 경기도의 호우·태풍에 대한 재해 피해액 및 누적강우량을 활용하여 대상지역별 재해강도에 따른 단계별 기준을 설정하고, 강우에 따라 발생할 수 있는 재해의 강도를 분류하는 모형을 개발하고자 하였다. 즉, 본 연구에서는 호우·태풍에 의한 재해 피해액 누적 분포 함수의 분위별로 재해강도의 범주(관심, 주의, 경계, 심각 단계)를 분류하였고, 재해강도의 범주에 따른 누적강우량 기준을 대상 지자체별로 제시하였다. 그리고 지자체별 재해강도 분류모형 개발을 위해 4가지(의사결정나무, 서포트 벡터 머신, 랜덤 포레스트, XGBoost)의 머신러닝 모형을 활용하였는데 강우량, 누적강우량, 지속시간 최대 강우량(3시간, 12시간), 선행강우량을 독립변수로 이용하여 종속변수인 지자체별 재해강도를 분류하였다. 각 모형별 F1 점수를 이용한 정확도 평가 결과, 의사결정나무의 F1 점수가 0.56으로 가장 우수한 정확도를 보였다. 본 연구에서 제시한 머신러닝 기반 재해강도 분류모형을 활용하면 호우·태풍에 의한 재해에 대한 지자체별 위험 상태를 단계별로 파악할 수 있어, 재난 담당자들의 신속한 의사결정을 위한 기초 자료로 활용될 수 있을 것으로 판단된다.

**핵심용어:** 자연재난, 표준 경보 기준, 머신러닝, 재해강도 분류모형

\*Corresponding Author. Tel: +82-32-860-7572

E-mail: sookim@inha.ac.kr (Kim, Hung Soo)

## 1. 서론

최근 기후변화로 인해 태풍 및 집중호우 등과 같은 자연재난의 발생빈도와 강도가 증가하고 있다. 또한, 도시화 등으로 불투수 면적이 증가함에 따라 내수침수로 인한 인명 및 재산 피해도 증가하고 있다(AON, 2021; Kim, 2022; Shin et al., 2014). 우리나라에서는 지속적으로 발생하는 호우·태풍 피해를 최소화하기 위해, 재난 및 안전관리 기본법에 의거하여 재난관리를 4단계(예방, 대비, 대응, 복구)로 세분화하여 대처하고 있다. 현재 국내 기상청에서 발령하고 있는 표준 경보 기준 재난관리 4단계 중 대응단계에 해당하며, 표준 경보 기준에서 사용하고 있는 호우주의보와 호우경보는 2단계로 구분하여 운영되고 있다. 호우주의보는 3시간 강우량이 60 mm 이상이거나 12시간 강우량이 110 mm 이상일 때 발령되며, 호우경보는 3시간 누적강우량이 90 mm 이상, 12시간 누적강우량이 180 mm 이상으로 예상될 때 발령된다(KMA, 2022). 그러나 현재 우리나라는 전국 지자체별로 지역별 재해특성이 고려되지 않은 상태로 표준 경보 기준이 일괄 적용되고 있다. 또한, 호우주의보와 호우경보도 재난 위기경보 4단계(관심, 주의, 경계, 심각)와 다르게 2단계를 기준으로 운영되기 때문에 재난 및 안전관리 기본법과 연동하기 어렵다는 문제점이 있다. 따라서, 지역별 재해특성과 재난의 분류기준을 고려하여 표준 경보 기준을 재설정할 필요가 있다.

지역별 재해특성을 고려한 호우 예·경보 기준에 대한 연구를 진행하기 위해 선행연구를 살펴보았다. Kim et al. (2011)에서는 5년간의 강우 자료를 이용하여 강우분석을 실시하고 강우자료와 피해의 관계를 조사하였다. 이후 표준 경보 기준과 호우피해 발생과의 연계성을 분석하여 호우에 의한 피해가 가장 빈번했던 강우량(30 mm/1 hr, 60 mm/3 hr, 70 mm/6 hr, 110 mm/12 hr)을 찾아냈다. 이는 현재 기상청에서 제시하는 호우주의보·경보 기준이 되었다.

Choi et al. (2018b)에서는 3가지의 머신러닝 기법(의사결정나무, 서포트 벡터 머신, 랜덤 포레스트)과 선형회귀분석을 이용하여 수도권 지역의 호우피해 예측함수를 개발하였다. 1994년부터 2011년까지 자료를 학습범위로 설정하고 2012년부터 2016년까지 자료를 평가하여 예측력을 확인하였다. 그 결과 해당 연구에서는 피해 발생 2일 전의 기상관측 자료를 사용했을 때 서포트 벡터 머신 기반의 함수가 가장 높은 예측력을 보였다. 일반적으로 기상관측 자료를 많이 이용하여 학습할수록 머신러닝 기반 모형의 예측력이 높아질 것이라고 예상하지만, 많은 기상관측 자료를 이용할수록 재해와 독립변수 간의 상관성이 떨어지면서 예측력 향상에 악영향을 미치는

것으로 확인되었다.

Kim (2021)에서는 모든 행정구역에 동일한 표준 경보 기준을 적용하는 대신 호우피해 특성이 유사한 지역을 군집화하고, 다양한 독립변수(공간적 강우분포를 고려한 국지성 호우, 침투조건을 고려한 유효우량 등)들을 반영하여 피해를 유발하는 최적 경계 도출에 대한 강우지수를 개발하였다. 연구개발 결과 기존 사용되고 있는 모형과 비교했을 때 예측력 등에선 성능 향상을 확인하였으나 기상청의 예측강우의 신뢰도에 따라 예측성능이 크게 바뀐다는 문제점이 있다.

또한 머신러닝을 이용한 연구에는 수력발전소의 잠재발전량 예측, 호우피해 예측, 수위 예측 등의 연구들이 있다(Choi et al., 2018a, 2019; Choi et al., 2019; Kim et al., 2022b; Jung et al., 2021). 선행연구 조사 결과, 우리나라는 표준 경보 발령 기준이 전국적으로 동일한 기준으로 운영되어 지역별 재해특성이 반영되지 않았기 때문에 표준 경보 기준과 각 지역의 재해피해 간의 상관성이 떨어지는 것으로 확인되었다(Kim et al., 2011; Choi et al., 2018b; Kim, 2021). 기존에 사용되던 자기회귀의 경험적 모형들은 통계적 관점에서 입력·출력 관계를 추정하는 모형으로 연구 초기에는 자동회귀·이동평균 모형(autoregressive moving average model, ARMA Model), 자기회귀·누적이동평균 모형(autoregressive integrated moving average model, ARIMA Model)과 같은 선형 모형을 통해 강우량을 예측하였지만, 비선형 시계열 자료인 강우량을 선형 모형으로 해석하기엔 어려운 부분들이 있었다(Kang, 1998; Kim et al., 2000, 2020; Han et al., 2009; Kim, 2010; Lee et al., 2021; Montanari et al., 1997). 최근 컴퓨터 알고리즘의 발전으로 머신러닝(machine learning) 기반의 모형(artificial neural network, support vector machine, random forest, extreme gradient boosting, adaptive neuro fuzzy inference system)이 비선형성을 고려한 모형으로 기상, 호우 및 홍수 등을 예측하는데 사용되고 있다(Abrahart et al., 2004; Assem et al., 2017; Bae et al., 2019; Choi, 2019; Ghuman et al., 2011; Granata et al., 2016; Mosavi et al., 2018; Riad et al., 2004; Shoaib et al., 2016; Yan et al., 2018).

따라서, 본 연구에서는 호우피해 발생시 영향력이 가장 큰 수도권의 지자체 단위(경기도, 서울특별시, 인천광역시) 행정구역을 대상으로 강우자료와 호우피해 자료를 사용하여 지역적 재해특성을 고려한 호우피해 발생 기준을 설정하였다. 기존 머신러닝 기법을 이용한 예측모형 개발에 대한 연구에 자료의 불균형 문제 해결을 위한 방법론을 추가하고, 모형 개발에 필요한 자료를 최소화하기 위하여 독립변수로 관측된 강우자료만을 활용하였다. 또한 설정된 기준으로 시도별 강우량에

따른 호우·태풍 피해의 강도를 분류하고자 의사결정나무, 서포트 벡터 머신, 랜덤 포레스트, XGBoost 등 4가지의 모형을 적용하여 재해강도 분류모형(Disaster Severity Classification Model, DSCM)을 개발하였다.

## 2. 자료 및 연구 방법

### 2.1 연구 대상지역 및 자료 수집

#### 2.1.1 연구 대상지역

전국적으로 통일된 표준 경보 기준 적용으로 발생하는 문제점을 보완하고자 대상지역별 재난 위기경보 4단계에 따른 누적강우량 기준을 제안하고자 한다. 연구 대상지역은 국내에서 건물 및 인구의 밀집도가 높은 수도권 지역(서울특별시, 인천광역시, 경기도)으로 선정하였다. 과거 대상지역에 발생한 호우 사상은 대표적으로 '10.09.21~'10.09.22 집중호우, '11.07.13 집중호우, '11.07.26~'11.07.29 집중호우 등이 있으며, 하천 급류로 인한 침수피해가 가장 많이 발생하였다.

#### 2.1.2 자료 수집

기상청에서는 기온, 강수량, 풍향, 풍속 등의 기상 관측 자료를 제공하며, 이는 우리나라 수자원 및 재난 분야 연구 등에서 기초 자료로 활용되고 있다. 일반적으로 국내의 기상관측소는 크게 종관기상관측장비(Automated Synoptic Observing System, ASOS)와 자동기상관측장비(Automatic Weather System, AWS) 관측소로 구분한다. ASOS 관측소는 주로 기상관서에 설치되며, 설치된 지역의 기상자료를 실시간으로 제공하고 해당 자료는 실제 기상예보에 활용되고 있다. 전국에 103개 지점이 있으며, 1904년부터 관측을 시작하여 신뢰도 높은 기상자료를 보유 및 제공하고 있다. AWS 관측소는 과거 사람이 직접 관측하던 자료를 자동으로 관측할 수 있도록 설계한 장비로, 현재 전국적으로 510개의 지점이 있다. 관측은 무인으로 진행되며, 1997년부터 관측을 개시하여 수집한 기상자료들을 보유 및 제공하고 있다(KMA, 2022).

본 연구에서는 수도권의 3개 시도를 대상으로 지역 내 8개 ASOS 관측소와 99개 AWS 관측소에서 수집된 강우자료를 활용하여 분석을 진행하였다. 총 107개의 강우관측소에 수집된 강우자료들은 재해강도 분류모형의 독립변수로 활용되었으며, 수집한 강우자료의 기초통계량을 Table 1과 같이 나타내었다. 또한, 대상지역과 분석에 반영된 기상청의 ASOS 및 AWS 관측소의 위치는 Fig. 1과 같이 도식화하였다.

시도별 강우량의 경우, 대상지역 내 불규칙적으로 분포된 관측지점에 가까운 면적을 가중치로 하여 면적단위 강우량으로 변환하는 방법인 티센다각형법(Thiessen polygon method)으로 산정하였다. 그리고 호우·태풍 피해의 기준을 설정하기 위해, 강우자료와 NDMS (National Disaster Management System) 자료 기반 지역별, 호우사상별 피해액 자료를 연계하여 기준을 설정하였다.

### 2.2 의사결정나무

의사결정나무(Decision Tree, DT)는 자료를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내며, 그 모양이 나무와 같아 의사결정나무라 불린다. 의사결정나무는 분류 규칙을 기반으로 자료를 분할하여 예측을 수행하는 방법론 중 하나이다(Quinlan, 1987; Song and Chae, 2008).

의사결정나무의 기본 알고리즘은 CHAID, CART, ID3,

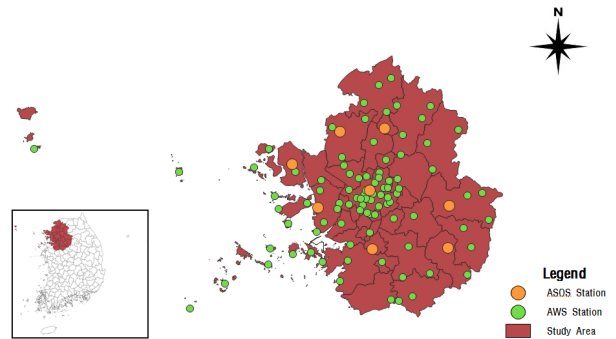


Fig. 1. Location of study area and weather observation station

Table 1. Basic statistics of variables

Classification	Variables	Min	Max	Mean	Standard Deviation	Coefficient of Variation
Gyeonggi Province	Rainfall (mm)	0.00	23.46	0.15	0.77	5.21
	Amount of Damage (1,000 won)	79	3,037,521	39,170	104,699	3.59
Seoul City	Rainfall (mm)	0.00	46.59	0.15	1.05	7.24
	Amount of Damage (1,000 won)	6,750	169,258,080	10,296,525	36,774,229	3.57
Incheon City	Rainfall (mm)	0.00	35.04	0.13	0.81	6.42
	Amount of Damage (1,000 won)	750	3,565,714	670,760	1,100,388	1.64

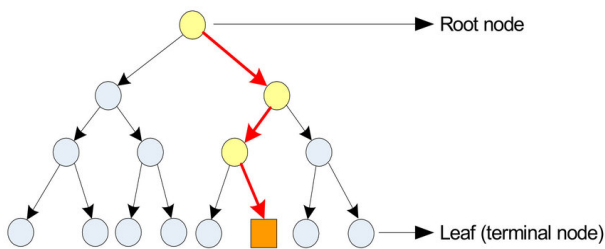


Fig. 2. Conceptual diagram of decision tree model

C4.5, C5.0 등이 있으며, 이들의 장점을 결합한 다양한 알고리즘이 존재한다(Kass, 1980; Breiman and Ihaka, 1984; Quinlan, 1986). 분석을 위한 매개변수로는 가지치기를 위한 복잡성을 의미하는 CP가 있다(Prakash *et al.*, 2022).

분석 과정을 나무구조로 표현 가능하기 때문에 판별분석(discriminant analysis), 회귀분석(regression analysis), 신경망(neural networks) 등과 같은 방법들에 비해 분석 과정을 쉽게 이해하고 설명할 수 있다는 장점이 있으나, 로지스틱 회귀(logistic regression)와 같이 각 예측 변수의 효과를 파악하기 어렵고 새로운 자료에 대한 예측이 불안정할 수 있다는 한계점이 있다(Choi *et al.*, 2019; Kang, 2022).

의사결정나무에 대한 개념도를 Fig. 2와 같이 나타내었다.

### 2.3 서포트 벡터 머신

서포트 벡터 머신(Support Vector Machine, SVM)은 패턴 인식, 자료 분석을 위한 지도 학습 모형으로 주로 분류 및 회귀 분석에 사용한다. 신경망보다 사용이 간결하며, 회귀분석에서 사용이 가능하나 분류에서 주로 사용하고 있는 기법이다(Cortes and Vapnik, 1995).

다양한 연산이 필요하고 입력 자료가 많을 경우에 학습 속도가 크게 느려진다는 단점이 있으나 가중치 벡터 범주, 수치 예측 문제에 사용 가능하며 오류 자료에 대한 영향을 적게 받는다는 장점이 있다(Choi *et al.*, 2018b).

서포트 벡터 머신은 비확률적 선형분류모형으로 서로 다른 특성을 가진 데이터들이 하나의 집합에 있을 때 최적의 경계 기준을 찾아 마진(margin)의 폭이 최대화가 되도록 범주를 구분하는 방법이다. 범주가 분류된 후에 신규 자료가 입력되었을 때, 해당 자료가 어느 범주에 속하는지 판단할 수 있는 모형이다. 매개변수로는 대표적으로 자료 표본의 영향력을 의미하는  $\sigma$ , 분류 경계를 의미하는  $C$ 가 있다(Karatzoglou *et al.*, 2006).

모형 특성상 자료의 크기가 방대할수록 예측에 소요되는 시간이 길어지긴 하지만, 분류 및 회귀분석에 모두 사용이 가능하다. 대용량의 자료를 효과적으로 처리할 수 있으며, 과적

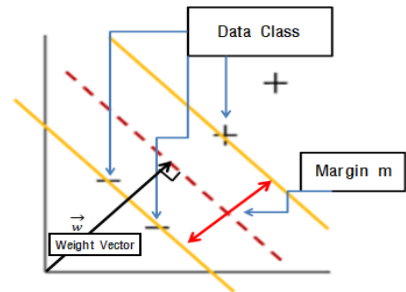


Fig. 3. Conceptual diagram of support vector machine model

합 문제를 최소화하여 모형의 정확도가 높다는 장점이 있다(Kim *et al.*, 2019; Kim *et al.*, 2022b).

서포트 벡터 머신에 대한 개념도를 Fig. 3과 같이 나타내었다.

### 2.4 랜덤 포레스트

랜덤 포레스트는 정확성, 단순성 및 유연성으로 인해 가장 많이 사용되는 알고리즘 중 하나로, 분류 및 회귀분석 등에 사용되는 앙상블(ensemble) 학습 방법 중 하나이다(Breiman, 2001).

랜덤 포레스트는 무작위로 특징을 선정해 나무의 마디로 지정하고, 분기할 때마다 의사결정나무와 같은 방식으로 불순도를 파악하여 분기해 나가는 가치를 여러개 만들어 비교한다. 이후 그 안에서 최적의 분류 및 예측을 하는 모형이다. 전체 자료에서 무작위 자료의 추출을 통해 여러개의 학습 자료 표본을 추출(bootstrapping)하여 표본들의 특성을 무작위로 선정하여 다수의 분류기를 일괄적으로 학습(bagging)하는 알고리즘으로, 매개변수로는 대표적으로 무작위 분류기의 개수를 의미하는  $mtry$ 가 있다(Liaw and Wiener, 2002).

모형 특성상 자료의 크기가 방대할수록 예측에 소요되는 시간이 길어지나, 분류 및 회귀분석에 모두 사용이 가능하다. 대용량의 자료를 효과적으로 처리할 수 있으며, 과적합 문제를 최소화하여 모형의 정확도가 높다는 장점이 있다(Kim *et al.*, 2019; Kim *et al.*, 2022b).

랜덤 포레스트에 대한 개념도를 Fig. 4와 같이 나타내었다.

### 2.5 XGBoost

여러개의 약한 분류기(weak classifier)를 순차적으로 학습하여 잘못된 자료에 가중치를 부여하는 방식으로 오류를 개선하는 모형(gradient boost)이 있다. XGBoost는 gradient boost 모형이 가진 단점을 간차를 이용하여 보완하는 새로운 모형을 순차적으로 결합한 뒤, 이를 선형 결합하여 만든 모형을 생성하는 지도학습 알고리즘인 Gradient Tree Boosting에 과적합 방지를 위한 기법이 추가된 지도학습 알고리즘이다(Chen

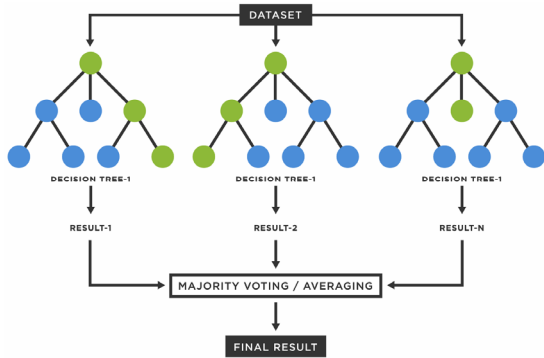


Fig. 4. Conceptual diagram of random forest model

and Guestrin, 2016; Go *et al.*, 2020).

XGBoost는 실제값과 예측값의 오차를 훈련자료에 투입하고 gradient를 이용하여 오류를 보완하는 방식으로 사용되며, 매개변수로는 부스팅 반복횟수를 의미하는 nrounds, 패널티 비율을 의미하는 lambda와 alpha, 학습률을 의미하는 eta가 있다(Ryu *et al.*, 2020).

적은 자료를 활용할 시에 과적합 가능성이 높고 모형의 정확도가 낮다는 단점이 있으나, 충분한 양의 자료가 있을 때 과적합이 잘 일어나지 않아 예측 성능이 좋다는 장점이 있다(Kim *et al.*, 2022b; Kim *et al.*, 2019).

XGBoost에 대한 개념도를 Fig. 5와 같이 나타내었다.

### 2.6 정확도 평가

본 연구에서 제시한 재해강도 분류모형의 정확도를 평가하기 위한 지표로는 정확도(accuracy), 재현율(recall), 정밀도(precision), F1 점수를 활용하였다. F1 점수는 분류모형의 재현율(recall)과 정밀도(precision)를 고려한 지표로, 0~1의 값을 가지며 숫자가 높을수록 정확도가 높다는 것을 의미한다(Kulkarni *et al.*, 2020; Sharma *et al.*, 2022). 각 지표의 계산시에 사용하는 변수의 의미는 Table 2와 같으며, 정확도, 재현율, 정밀도, F1 점수를 산정하기 위한 수식을 Eqs. (1)~(4)와 같이 나타내었다.

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + TN} \quad (3)$$

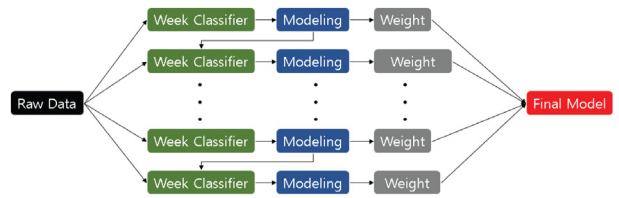


Fig. 5. Conceptual diagram of XGBoost model (Kim *et al.*, 2019)

Table 2. Confusion matrix for performance evaluation of the model (Karimi, 2021)

Division		Observation	
		True	False
Classification	Positive	TP (True Positive)	FP (False Positive)
	Negative	TN (True Negative)	FN (False Negative)

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2TP}{TP + (FP + FN)} \quad (4)$$

여기서 TP는 모형이 재해가 발생할 것이라 분류했을 때 실제로 발생한 경우, TN은 모형이 재해가 발생하지 않을 것으로 분류했을 때 실제로 발생한 경우, FP는 모형이 재해가 발생할 것으로 분류했을 때 실제로는 발생하지 않은 경우, FN은 모형이 재해가 발생하지 않을 것으로 분류했을 때 실제로도 발생하지 않은 경우를 의미한다.

## 3. 재해강도 분류모형 개발

### 3.1 호우·태풍의 사상 선정

호우·태풍 사상을 선정하기 위해 수도권 지역의 NDMS자료에서 침수 피해가 발생한 호우·태풍 사상을 확인하였다. NDMS는 국가안전관리시스템을 의미하며, 재해 예방 및 재해 피해로부터 국민을 보호하기 위한 목적으로 구축된 재난관리 기반정보시스템이다(Hong *et al.*, 2005).

NDMS에서는 피해 사상별로 발생한 피해액 및 복구액을 시설단위로 제공한다. 따라서 본 연구에서는 최적의 호우·태풍 피해에 대한 재해강도 분류모형을 개발하고자 NDMS에서 제공하는 피해액 자료를 활용하였다. 대상지역 내에 2008년부터 2020년까지 발생한 117개의 호우·태풍 사상에 대한 자료 분포를 확인 해보면 하위 25% 금액에 해당하는 피해사상이 약 97.3%인 것을 확인할 수 있었다. 이는 NDMS 피해액



자료의 분포가 불균형하다는 것을 의미하며, 해당 문제를 개선하기 위하여 누적 분포 함수(Cumulative Distribution Function, CDF)을 이용하여 기준을 설정하였다.

호우·태풍의 사상을 분리하기 위하여 무강우 지속시간(Interevent Time Definition, IETD)를 결정하였다(Kim *et al.*, 2022a; Choi, 2016; Kim, 2018; Lee, 2021). 티센다각형법을 이용하여 산정한 지역별 강우량에 자기상관계수(Autocorrelation, AC), 변동계수(Coefficient of Variation, CV), 연평균

강우사상 발생개수 분석으로 나온 무강우 지속시간을 비교하였다. 자기상관계수 및 연평균 강우사상 발생 개수 분석 시 무강우 지속시간이 특정 값에 수렴하지 않아 본 연구에서는 변동계수 분석법을 통한 무강우 지속시간을 활용하였다. 지역별 변동계수 분석법을 통한 무강우 지속시간은 경기도에서 11시간, 서울특별시에서 13시간, 인천광역시에서 11시간으로 산정되었으며, 이를 통해 분류한 호우사상은 NDMS에서 분류한 호우사상과 유사한 것으로 확인되었다.

**Table 3.** Basic statistics of independent variables for Gyeonggi province

Classification	Max	Mean	Standard Deviation	Coef of Variation
Rainfall (mm)	35.04	7.26	5.42	0.75
Cumulative Rainfall (mm)	403.77	49.95	56.93	1.14
3 hour Maximum Rainfall (mm)	75.95	16.07	12.30	0.77
12 hour Maximum Rainfall (mm)	283.41	62.70	60.69	0.97
Antecedent Rainfall (5 Days) (mm)	164.31	31.89	26.55	0.83

### 3.2 머신러닝 기반 모형의 입력자료

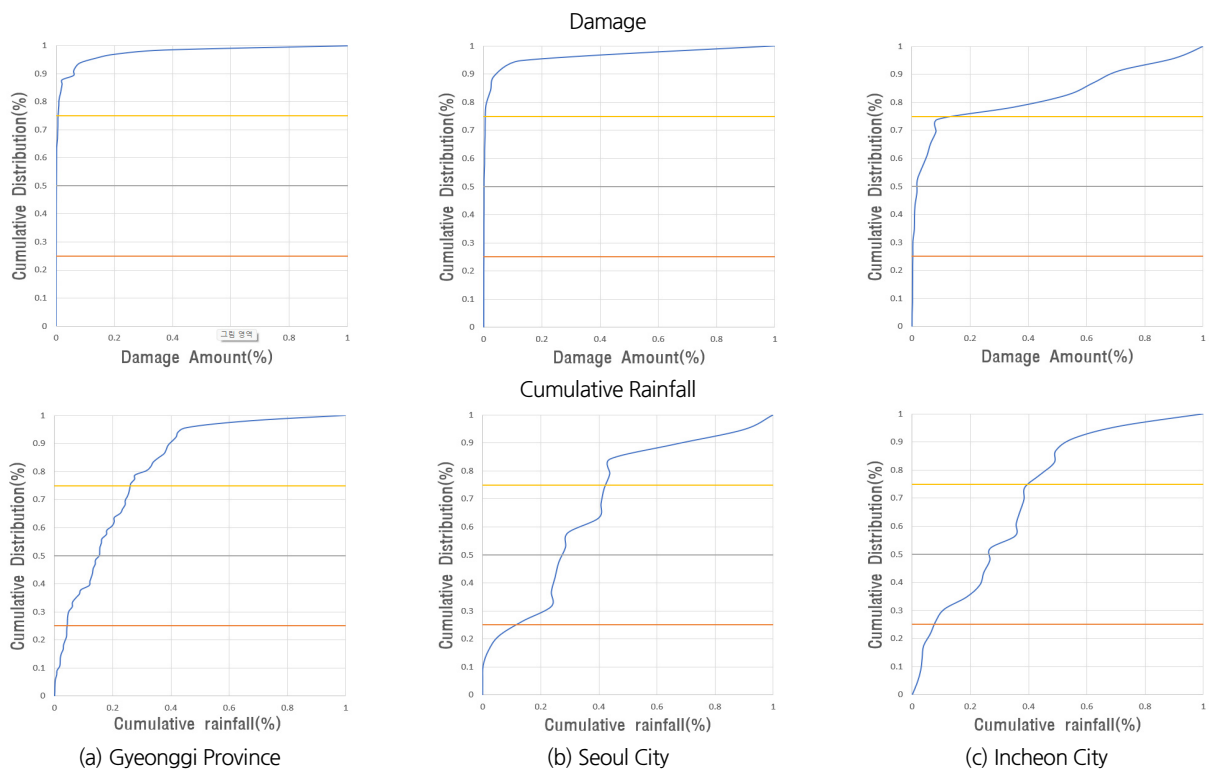
#### 3.2.1 강우량 자료 기반의 입력자료 구축

호우·태풍 피해에 대한 최적의 머신러닝 기반 재해강도 분류 모형을 개발하기 위하여 독립변수(independent variable)로 ① 시간당 강우량, ② 사상별 누적강우량, ③ 3시간 최대강우량, ④ 12시간 최대강우량, 홍수량 산정시 사용하는 선행토양함수 조건(AMC)의 기준이 되는 ⑤ 5일 선행강우량을 이용하였다.

피해가 가장 많이 발생한 경기도 지역에서 사용된 독립변수의 기초 통계량을 Table 3과 같이 나타내었다.

#### 3.2.2 NDMS 자료 기반의 입력자료 구축

본 연구에서는 종속변수로 NDMS의 피해액 자료를 이용하



**Fig. 6.** Cumulative distributions of damage and rainfall for each local government

였다. 호우·태풍 피해액 자료에 대한 확률 밀도 함수  $f(x)$ 와 구간  $[a, b]$ 에 대해서 확률 변수  $X$ 가 포함될 확률  $P(a \leq X \leq b)$ 를 산정하였다. 또한, 주어진 확률변수가 특정 값보다 작거나 같은 확률을 산정하고자 누적 분포 함수를 이용하였다. 이를 바탕으로 호우·태풍 피해액의 범주를 재난 위기경보 4단계(관심, 주의, 경계, 심각)에 따라 25%, 50%, 75% 경계로 분류하여 종속변수(1~4)로 활용하고, 재해강도 분류를 위한 강우 기준을 지역별로 설정하였다.

경기도, 서울특별시, 인천광역시에서 호우 피해액과 누적 강우량의 분류 기준을 설정하기 위해 산정한 누적 분포함수 곡선은 Fig. 6과 같으며, 재해강도를 분류하기 위해 설정한 누적 강우량 기준을 Table 4에, 경기도 지역을 대표로 하여 기준 표준 정보 기준과 본 연구에서 제시하는 기준을 적용하였을 때 발령되는 표준 정보를 비교한 결과를 Table 5에 정리하였다.

3.2.3 자료 불균형 문제 해결을 위한 보정

일반적으로 국내 강우자료의 경우 연간 강우 발생 일수보다 무강우 일수가 압도적으로 긴 것이 현실이다. 그리고 호우·태풍 피해액 자료의 경우에도 사상별 발생 피해액의 편차가 심해 자료의 불균형이 발생하였다.

머신러닝 모형은 학습 자료가 불균형할 경우 예측 자료의 정확도가 높아도 재현율이 급격히 작아지는 현상이 발생하게 된다. 본 연구에서 분석에 사용한 자료 특성상 자료 불균형 문제는 피할 수 없었기 때문에 이를 보완하고자 학습구간에서 SMOTE (Synthetic Minority Over-sampling Technique) 방법을 적용하였다.

Table 4. Classification of cumulative rainfall criteria for each local government

Classification	Gyeonggi Province	Seoul City	Incheon City
Criteria 1 (mm)	25	70	35
Criteria 2 (mm)	90	163	121
Criteria 3 (mm)	153	252	181

Table 5. Comparison of standard criteria and disaster risk stages from cumulative distributions

Classification	3 hour Maximum Rainfall (mm)	12 hour Maximum Rainfall (mm)	Cumulative Rainfall (mm)	Amount of Damage (1,000 won)	Existing Standard Warning Criteria	Modified Standard Warning Criteria
1	48.90	119.26	237.16	5,799,888	Watch	Red
2	27.05	46.45	71.11	959	-	Yellow
3	65.31	146.93	149.86	1,738,190	Watch	Orange
4	76.50	175.47	322.72	37,152,530	Watch	Red
5	36.99	93.34	142.12	422,025	-	Orange
∴	∴	∴	∴	∴	∴	∴

SMOTE 방법은 자료의 불균형을 해결하기 위한 방법론들 중 하나로, 자료의 개수가 적은 범주의 표본을 가져온 뒤 해당 범주 내 임의의 값을 추가하여 새로운 샘플을 만들어 자료에 추가하는 오버샘플링(oversampling) 방식(Chawla et al., 2002; Han et al., 2005)이며, 그 개념도를 Fig. 7과 같이 나타내었다.

SMOTE 방법을 미적용했을 때와 비교하여 적용했을 때 대표 정확도 평가 지표인 F1 점수가 약 20% 개선된 것을 확인할 수 있었다. 이에 따라 모형의 평가에서 SMOTE 방법을 적용하여 산출한 결과를 사용하였으며, 경기도, 서울특별시, 인천광역시에 대하여 SMOTE 방법 적용 전과 후 지표의 평균을 Table 6과 같이 비교하였다.

3.3 재해강도 분류모형 개발

머신러닝에서 분류(classification)는 독립변수에 따라 가장 연관성이 큰 종속변수(피해액 범주)를 예측하는 것을 의미한다. 이는 특정 범주에 속한 자료가 주어졌을 때 자료를 기반

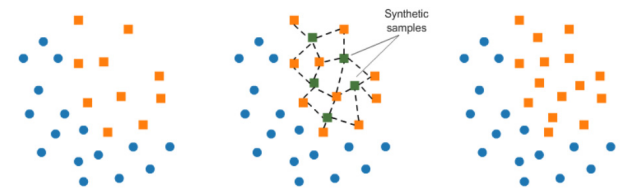


Fig. 7. Conceptual diagram of SMOTE method (Alencar, 2018)

Table 6. Comparison of F1-Score for SMOTE Application in the Learning Process of each Model

Model	F1-Score before SMOTE	F1-Score after SMOTE	Improvement (%)
Decision Tree	0.49	0.56	13.94
Support Vector Machine	0.38	0.45	18.31
Random Forest	0.36	0.45	25.01
XGBoost	0.39	0.48	23.11
Average	0.41	0.48	19.62

으로 도출되는 임의의 결과가 어느 범주가 어디에 속하는지를 알아내는 방법이다. 본 연구에서는 다양한 유형별 강우자료와 피해액 자료를 이용하여 특정 강우에 따라 발생하는 피해의 규모를 파악하고자 하였으며, 최적의 분류모형을 찾기 위해 의사결정나무, 서포트 벡터 머신, 랜덤 포레스트, XGBoost 모형을 적용하였다. 또한, 각 모형별로 분류모형을 학습할 때 임의의 탐색(random search) 방법을 통해 각 모형의 매개변수를 최적화하였다.

본 연구는 R언어로 진행되었으며 2008년부터 2015년까지 총 503개의 호우사상 자료를 학습구간, 2016년부터 2018년까지 총 171개의 호우사상 자료를 평가구간으로 구분하였다. 또한, 2019년부터 2020년까지 총 110개의 호우사상 자료를 이용하여 모형의 분류 성능을 평가하였다.

호우-태풍 피해에 대한 재해강도 분류모형의 학습구간은 2008년부터 2015년까지의 강우 및 NDMS 피해액 자료를 이용하였다. 최적의 호우-태풍 피해에 대한 재해강도 분류모형을 개발하고자 SMOTE 방법을 학습구간에 적용하였으며, 방

범론 적용 후 총 10,970개의 자료를 학습자료로 활용하였다.

강우자료의 경우 수도권 내에 위치한 ASOS 관측소의 자료에 티센다각형법을 적용한 면적단위 강우자료를 활용하였다. 강우자료에서는 각 시간에 발생한 누적강우량, 지속시간 최대강우량 및 선행강우량을 계산하여 입력자료로 활용하였다. 또한, 모형의 성능을 평가하기 위해서 분류된 실제 재해강도(occurrence damage class) 범주와 모형의 분류(classification) 범주를 이용한 혼동행렬(confusion matrix)을 이용하였다.

학습구간 학습 후 2016년부터 2018년까지의 입력자료를 바탕으로 모형의 평가를 진행하였으며, 피해 규모가 가장 큰 경기도 지역을 대표로 하여 구축한 혼동행렬을 Table 7과 같이 나타내었다. 또한, 각 지역별·모형별 학습구간 성능 평가 결과를 Table 8과 같이 나타내었다.

혼동행렬을 살펴보면, Class 1에서 높은 정확도를 보이고, 나머지 Class에선 상대적으로 낮은 정확도를 가졌다. 이는 자료의 표본이 Class 1에 가장 많이 분포하는 불균형 형태를 띠기 때문으로 확인되었다.

서울특별시의 경우 다른 지자체와 비교하여 F1 점수가 낮은걸 확인할 수 있다. 이는 호우피해와 상관성이 있는 지표들이 강우자료 이외 다수 존재하며, 서울특별시가 그 지표들의

Table 7. Confusion matrix of performance evaluation for the learning process of each model

(1) Decision Tree		Occurrence Damage Class			
		1	2	3	4
Classification	1	36	1	2	0
	2	7	1	2	0
	3	0	1	1	0
	4	1	1	1	2
(2) Support Vector Machine		Occurrence Damage Class			
		1	2	3	4
Classification	1	11	0	0	0
	2	33	4	6	2
	3	0	0	0	0
	4	0	0	0	0
(3) Random Forest		Occurrence Damage Class			
		1	2	3	4
Classification	1	24	1	1	0
	2	20	2	3	0
	3	0	1	1	0
	4	0	0	1	2
(4) XGBoost		Occurrence Damage Class			
		1	2	3	4
Classification	1	39	0	3	0
	2	4	2	0	0
	3	0	0	1	1
	4	1	2	2	1

Table 8. Performance comparison for the learning process of models in each local government

Region	Classification	Accuracy	Recall	Precision	F1 Score
Gyeonggi Province	Decision Tree	0.71	0.48	0.56	0.52
	Support Vector Machine	0.27	0.54	0.31	0.40
	Random Forest	0.52	0.54	0.55	0.55
	XGBoost	0.77	0.48	0.51	0.50
Seoul City	Decision Tree	0.95	0.95	0.25	0.40
	Support Vector Machine	0.92	0.48	0.32	0.38
	Random Forest	0.97	0.50	0.33	0.40
	XGBoost	0.97	0.50	0.33	0.40
Incheon City	Decision Tree	0.91	0.63	0.73	0.67
	Support Vector Machine	0.95	0.95	0.25	0.40
	Random Forest	0.93	0.32	0.25	0.28
	XGBoost	0.93	0.25	0.25	0.25
Average	Decision Tree	0.86	0.69	0.51	0.53
	Support Vector Machine	0.71	0.66	0.29	0.39
	Random Forest	0.81	0.45	0.38	0.41
	XGBoost	0.89	0.41	0.36	0.38



영향을 다른 대상지역에 비해 가장 크게 받기 때문이라고 판단하였다.

각 모형의 지자체별 학습구간의 F1 점수 평균을 비교한 결과 의사결정나무 모형이 평균 0.53으로 가장 높은 정확도를 가지는 것으로 확인되었다.

### 3.4 재해강도 분류모형의 적용성 평가

본 연구에서는 총 4가지 모형을 개발하여 발생 강우에 따른 피해 강도 분류 기술을 개발하였다.

개발한 기술의 적용성을 검토하기 위하여 모형을 구축할 때 학습에 활용되지 않은 2019년부터 2020년까지의 자료를 이용하여 모형별 정확도를 평가 및 비교 하였고 SMOTE 적용 전과 후의 F1 점수를 비교하였다.

**Table 9.** Confusion matrix of performance evaluation for the evaluation process of each model

(1) Decision Tree		Occurrence Damage Class			
		1	2	3	4
Classification	1	22	1	2	0
	2	4	1	1	1
	3	0	0	1	0
	4	1	0	0	2
Accuracy : 0.72		F1 Score : 0.61			
(2) Support Vector Machine		Occurrence Damage Class			
		1	2	3	4
Classification	1	6	0	0	0
	2	21	2	4	3
	3	0	0	0	0
	4	0	0	0	0
Accuracy : 0.22		F1 Score : 0.39			
(3) Random Forest		Occurrence Damage Class			
		1	2	3	4
Classification	1	18	1	0	1
	2	9	0	3	1
	3	0	0	0	0
	4	0	1	1	1
Accuracy : 0.53		F1 Score : 0.31			
(4) XGBoost		Occurrence Damage Class			
		1	2	3	4
Classification	1	22	0	2	2
	2	4	1	0	0
	3	0	0	0	0
	4	1	1	2	1
Accuracy : 0.67		F1 Score : 0.41			

학습구간의 학습 및 평가 후 경기도 지역을 대표로 하여 모형별 정확도를 평가·비교한 혼동행렬을 Table 9와 같이 나타내었다.

피해가 가장 많이 발생한 경기도 지역을 대표로 하여 4가지 모형을 비교 평가 한 결과, Class 1에서는 의사결정나무와 XGBoost, Class 2에서는 서포트 벡터 머신, Class 3 및 Class 4에서는 의사결정나무 모형의 정확도가 우수한 것으로 확인하였다.

학습에 활용되지 않은 2019년부터 2020년까지의 자료를 이용하여 재해강도 분류모형의 적용성을 검토하였다. 각 지자체별로 가장 좋은 F1 점수를 가진 모형은 경기도에서 의사결정나무 0.61, 서울특별시에서 랜덤 포레스트 0.65, 인천광역시에서 의사결정나무 0.68로 확인되었으며, 지역별·모형별 정확도평가 결과를 Table 9와 같이 나타내었다.

F1 점수를 대상 지자체별로 평가하여 평균을 Table 10과 같이 비교하였다. 비교한 결과 의사결정나무는 0.56, 서포트 벡터 머신은 0.45, 랜덤 포레스트는 0.45, XGBoost는 0.48로 의사결정나무의 정확도가 가장 우수한 모형으로 평가되었다.

**Table 10.** Performance comparison for the evaluation process of models in each local government

Region	Classification	Accuracy	Recall	Precision	F1 Score
Gyeonggi Province	Decision Tree	0.72	0.67	0.56	0.61
	Support Vector Machine	0.22	0.53	0.31	0.39
	Random Forest	0.53	0.41	0.25	0.31
	XGBoost	0.67	0.42	0.41	0.41
Seoul City	Decision Tree	0.89	0.32	0.47	0.38
	Support Vector Machine	0.95	0.49	0.50	0.49
	Random Forest	0.95	0.95	0.50	0.65
	XGBoost	0.95	0.95	0.50	0.64
Incheon City	Decision Tree	0.89	0.66	0.71	0.68
	Support Vector Machine	0.92	0.47	0.49	0.48
	Random Forest	0.86	0.31	0.46	0.37
	XGBoost	0.86	0.31	0.46	0.37
Average	Decision Tree	0.84	0.55	0.58	0.56
	Support Vector Machine	0.70	0.50	0.43	0.45
	Random Forest	0.78	0.56	0.40	0.45
	XGBoost	0.83	0.56	0.46	0.48

## 4. 결론

본 연구에서는 재난 및 안전관리 기본법과 연계가 가능한 강우 기준을 제시하고, 호우-태풍 피해에 대한 재해강도 분류 모형을 개발하고자 하였다.

표준 경보 기준은 발생할 것으로 예측되는 3시간 누적강우량과 12시간 누적강우량을 기준으로 발령 기준이 설정되어 있으며, 전국적으로 동일한 기준을 일괄 적용하여 실제로 호우피해가 발생하고 있지만 특보는 발령되지 않는 선례가 많은 게 실정이다. 이에 따라 본 연구에서 제안하는 표준 경보 기준을 적용하여 비교했을 때, 각 지역별로 표준 경보 발령이 안되거나 재해의 강도가 낮은 것으로 표준 경보가 발령되는 문제의 상당수가 해결된 것을 확인할 수 있었다. 그러나 행정경계에 인접하는 지역의 경우 가까운 지역임에도 불구하고 서로 상이한 누적강우량 기준이 적용된다는 한계점이 있으며, 이는 추후 연구에 있어 중요한 해결과제가 될 것이다.

본 연구에서 제시하는 지역별 재해특성에 맞춘 분류 기준을 사용하면 호우피해 발생이 예측될 경우, 재난 및 안전관리 기본법에서 제시하는 단계별로 적합한 조치를 올바르게 수행할 수 있다. 그리고 본 연구에서 제시한 머신러닝 기반 재해강도 분류모형은 특정 강우가 발생함에 따른 피해 범주를 분류할 수 있어 홍수 대응단계에서도 신속한 의사결정을 지원할 수 있다. 다만 호우피해 발생에는 지역내 하천의 수위, 지역의 개발 정도, 인구밀도 등과 같이 영향을 미치는 요소가 다수 존재하며, 본 연구에서 제시하는 모형은 강우량만을 독립변수로 활용하기 때문에 도출되는 결과에 불확실성은 존재한다. 이러한 문제를 해결하기 위하여 추후 연구에서 신뢰할 수 있는 자료의 확보와 더불어 호우피해의 규모와 상관성이 있는 독립변수를 추가로 구축하여 모형을 보완해야 할 것으로 판단된다.

## 감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 물 관리연구사업의 지원을 받아 연구되었습니다(21AWMP-B121100-06).

## Conflicts of Interest

The authors declare no conflict of interest.

## Reference

- Abrahart, R., Kneale, P.E., and See, L.M. (2004). *Neural networks for hydrological modeling*. CRC Press, Bock Raton, FL, U.S., pp. 1-13.
- Alencar, R. (2018). Resampling strategies for imbalanced datasets, accessed 9 March 2023, <<https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets>>.
- AON (2021). *Weather, climate & catastrophe insight*. 2018 Annual Report, London, UK.
- Assem, H., Ghariba, S., Makrai, G., Johnston, P., Gill, L., and Pilla, F. (2017). "Urban water flow and water level prediction based on deep learning." *ECML PKDD 2017*, Springer, Skopje, Macedonia, Part III, No.10, pp. 317-329.
- Bae, Y.H., Kim, J.S., Wang, W.J., Yoo, Y.H., Jung, J.W., and Kim, H.S. (2019). "Monthly inflow forecasting of Soyang River dam using VARMA and machine learning models." *Journal of Climate Research*, Vol. 14, No. 3, pp. 183-198.
- Breiman, L. (2001). "Random forests." *Machine Learning*, Vol. 45, No. 1, pp. 5-32.
- Breiman, L., and Ihaka, R. (1984). *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California, CA, U.S.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357.
- Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system." *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco, CA, U.S., pp. 785-794.
- Choi, C., Kim, J., Han, H., Han, D., and Kim, H.S. (2019). "Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea." *Water*, Vol. 12, No. 1, pp. 93-110.
- Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y., and Kim, H.S. (2018a). "Development of heavy rain damage prediction model using machine learning based on big data." *Advances in meteorology*, Vol. 2018, 5024930.
- Choi, C.H. (2016). *Mega flood simulation occurred by consecutive extreme storm event and typhoon*. Master Thesis, Inha University, pp. 29-31.
- Choi, C.H. (2019). *Development of combined heavy rain damage prediction models using machine learning and effectiveness of disaster prevention projects*. Ph.D. Dissertation, Inha University, pp. 1-12.
- Choi, C.H., Kim, J.S., Kim, D.H., Lee, J.H., Kim, D.H., and Kim, H.S. (2018b). "Development of heavy rain damage prediction functions in the seoul capital Area using machine learning techniques." *Journal of The Korean Society of Hazard Mitigation*, Vol. 18, No. 7, pp. 435-447.
- Cortes, C., and Vapnik, V. (1995). "Support-vector networks." *Machine Learning*, Vol. 20, No. 3, pp. 273-297.

- Ghumman, A.R., Ghazaw, Y.M., Sohail, A.R., and Watanabe, K. (2011). "Runoff forecasting by artificial neural network and conventional model." *Alexandria Engineering Journal*, Vol. 50, No. 4, pp. 345-350.
- Go, C.M., Jeong, Y.Y., Jee, Y.G., Lee, Y.M., Kim, B.S. (2020). "A study on hydrological rainfall adjustment using machine learning and probability matching method during heavy rainfall season." *Journal of Climate Research*, Vol. 15, No. 4, pp. 257-267.
- Granata, F., Gargano, R., and De Marinis, G. (2016). "Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model." *Water*, Vol. 8, No. 3, 69.
- Han, H., Wang, W.Y., and Mao, B.H. (2005). "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *ICIC 2005*, Springer, Hefei, China, Part 1, pp. 878-887.
- Han, J.W., Kwon, H.H., and Kim, T.W. (2009). "Reliability evaluation of parameter estimation methods of probability density function for estimating probability rainfalls." *Journal of the Korean Society of Hazard Mitigation*, Vol. 9, No. 6, pp. 143-152.
- Hong, J.H., Sin, T.G., Yun, U.J., Lee, T.S., and Jo, W.C. (2005). "Roadmap of NDMS Facility DB Joint Utilization System." *In Proceedings of the Korean Institute of Industrial Safety Conference*, KSS, pp. 179-184.
- Jung, J., Han, H., Kim, K., and Kim, H.S. (2021). "Machine learning-based small hydropower potential prediction under climate change." *Energies*, Vol. 14, No. 12, pp. 3643-3653.
- Kang, D.G. (2022). *A decision tree for estimating mode of the response variable*. Master Thesis, Korea University, pp. 6-11.
- Kang, T.H. (1998). *Study on the development of forecasting method for rainfall, runoff and water quality in urban stream*. Ph. D. Dissertation, Kyonggi University, pp. 1-21.
- Karatzoglou, A., Meyer, D., and Hornik, K. (2006). "Support vector machines in R." *Journal of Statistical Software*, Vol. 15, pp. 1-28.
- Karimi, Z. (2021). Confusion matrix, research gate, accessed 23 February 2023, <[https://www.researchgate.net/publication/355096788\\_Confusion\\_Matrix](https://www.researchgate.net/publication/355096788_Confusion_Matrix)>.
- Kass, G.V. (1980). "An exploratory technique for investigating large quantities of categorical data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 29, No. 2, pp. 119-127.
- Kim, B.J., Sohn, K.T., Oh, J.H., Baik, J.S., Lee, Y.H., and Baek, H.J. (2000). "Analysis of the long-term change and extreme events of daily summer rainfall over Korea." *Journal of the Korean Data Analysis Society*, Vol. 20, No. 1, pp. 37-44.
- Kim, D., Lee, J., Kim, J., Lee, M., Wang, W., and Kim, H.S. (2022a). "Comparative analysis of long short-term memory and storage function model for flood water level forecasting of Bokha stream in NamHan River, Korea." *Journal of Hydrology*, Vol. 606, 127415.
- Kim, D.H. (2018). *Development of consecutive storm event based (conseb) rainfall-runoff model for short term runoff simulation and its applicability under climate change*. Ph. D. Dissertation, Inha University, pp. 1-6.
- Kim, D.H. (2022). *Development of flood water level forecasting and flood damage risk assessment method for river basin using AI-based hybrid model*. Ph. D. Dissertation, Inha University, pp. 1-173.
- Kim, D.H., Kim, J.W., Kwak, J.W., Necesito, I.V., Kim, J.S., and Kim, H.S. (2020). "Development of water level prediction models using deep neural network in mountain wetlands." *Journal of Wetlands Research*, Vol. 22, No. 2, pp. 106-112.
- Kim, D.H., Lee, K.S., Hwang-Bo, J.G., Kim, H.S., and Kim, S.J. (2022b). "Development of the method for flood water level forecasting and flood damage warning using an AI-based model." *Journal of the Korean Society of Hazard Mitigation*, Vol. 22, No. 4, pp. 145-156.
- Kim, J.S. (2021). *Development of prediction and warning technique of heavy rain damage risk based on ensemble machine learning and risk matrix*. Ph. D. Dissertation, Inha University, pp. 238-242.
- Kim, J.S., Lee, J.H., Kim, D.H., Choi, C.H., Lee, M.J., and Kim, H.S. (2019). "Developing a prediction model (Heavy rain damage occurrence probability) based on machine learning." *Journal of the Korean Society of Hazard Mitigation*, Vol. 19, No. 6, pp. 115-127.
- Kim, K.S. (2010). *A study on the real time forecasting for monthly inflow Daecheong dam using hydrologic time series analyses*. Master Thesis, Seokyeong University, pp. 1-27.
- Kim, Y.H., Choi, D.Y., Jang, D.E., Yoo, H.D., and Jin, G.B. (2011). "An improvement on the criteria of special weather report for heavy rain considering the possibility of rainfall damage and the recent meteorological characteristics." *Atmosphere*, Vol. 21, No. 4, pp. 481-495.
- Korea Meteorological Administration (KMA) (2022). Special weather reports standards, accessed 27 December 2022, <<https://www.weather.go.kr/w/weather/warning/standard.do>>.
- Kulkarni, A., Chong, D., and Batarseh, F.A. (2020). *Foundations of data imbalance and solutions for a data democracy*. Academic Press, Cambridge, MA, U.S., pp. 83-106.
- Lee, H., Kim, H.S., Kim, S., Kim, D., and Kim, J. (2021). "Development of a method for urban flooding detection using unstructured data and deep learning." *Journal of Korea Water Resources Association*, Vol. 12, No. 54, pp. 1233-1242.
- Lee, J.S. (2021). *Development and application of artificial intelligence based model for real time flood*. Ph. D. Dissertation, Inha University, pp. 40-41.
- Liaw, A., and Wiener, M. (2002). "Classification and regression by randomForest." *R News*, Vol. 12, No. 3, pp. 18-22.
- Montanari, A., Rosso, R., and Taquq, M.S. (1997). "Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation." *Water Resources Research*, Vol. 33, No. 5, pp. 1035-1044.
- Mosavi, A., Ozturk, P., and Chau, K.W. (2018). "Flood prediction using machine learning models: Literature review." *Water*, Vol. 10, No. 11, 1536.
- Prakash, D.B., Kumar, K.A., and Kumar, R.P. (2022). "Hyper-

- parameter optimization using metaheuristic algorithms.” *CVR Journal of Science and Technology*, Vol. 23, No. 1, pp. 37-43.
- Quinlan, J.R. (1986). “Induction of decision trees.” *Machine Learning*, Vol. 1, pp. 81-106.
- Quinlan, J.R. (1987). “Simplifying decision trees.” *International Journal of Man-Machine Studies*, Vol. 27, No. 3, pp. 221-234.
- Riad, S., Mania, J., Bouchaou, L., and Najjar, Y. (2004). “Predicting catchment flow in a semi-arid region via an artificial neural network technique.” *Hydrological Processes*, Vol. 18, No. 13, pp. 2387-2393.
- Ryu, S.E., Shin, D.H., and Chung, K. (2020). “Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization.” *IEEE Access*, No. 8, pp. 177708-177720.
- Sharma, D.K., Chatterjee, M., Kaur, G., and Vavilala, S. (2022). *Deep learning applications for disease diagnosis*. Academic Press, Cambridge, MA, U.S., pp. 31-51.
- Shin, J.Y., Lim, S.M., Kim, J.H., and Kim, T.W. (2014). “Analysis of urban flood damage characteristics using inland flood scenarios and flood damage curve.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 14, No. 1, pp. 291-302.
- Shoaib, M., Shamseldin, A.Y., Melville, B.W., Khan, M.M. (2016). “A comparison between wavelet based static and dynamic neural network approaches for runoff prediction.” *Journal of Hydrology*, Vol. 535, pp. 211-225.
- Song, Y.S., and Chae, B.G. (2008). “Development to prediction technique of slope hazards in gneiss area using decision tree model.” *The Journal of Engineering Geology*, Vol. 18, No. 1, pp. 45-54.
- Yan, J., Jin, J., Chen, F., Yu, G., Yin, H., and Wang, W. (2018). “Urban flash flood forecast using support vector machine and numerical simulation.” *Journal of Hydroinformatics*, Vol. 21, No. 1, 016111.