JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 3: 341~346, Mar. 2022

홍수 위험도 판별을 위한 CNN 기반의 분류 모델 구현

조민우[1] · 김동수[1] · 정회경[2*]

Implementation of CNN-based classification model for flood risk determination

Minwoo Cho[1] · Dongsoo Kim[1] · Hoekyung Jung[2*]

[1]Graduate Student, Department of Computer Engineering, Paichai University, Daejeon, 35345 Korea
[2*]Professor, Department of Computer Engineering, Paichai University, Daejeon, 35345 Korea

요 약

지구온난화 및 이상 기후로 인해 홍수의 빈도 및 피해 규모가 늘어나고 있으며, 홍수 취약 지역에 노출된 사람이 2000년도에 비하여 25% 증가하였다. 홍수는 막대한 금전적, 인명적 손실을 유발하며, 홍수로 인한 손실을 줄이기 위해 홍수를 미리 예측하고 빠른 대피를 결정해야 한다. 본 논문은 홍수 예측을 위한 핵심 데이터인 강우량과 수위 데이터를 활용하여 시기적절한 대피 결정이 이루어질 수 있도록 CNN기반 분류 모델을 활용하여 홍수 위험도 판별 모델을 제안한다. 본 논문에서 제안한 CNN 기반 분류 모델과 DNN 기반의 분류 모델의 결과를 비교하여 더 좋은 성능을 보이는 것을 확인하였다. 이를 통해 홍수의 위험도를 판별하여, 대피 여부 판단하며 최적의 시기에 대피 결정을 내릴 수 있도록 하는 초기 연구로서 활용할 수 있을 것으로 사료된다.

ABSTRACT

Due to global warming and abnormal climate, the frequency and damage of floods are increasing, and the number of people exposed to flood-prone areas has increased by 25% compared to 2000. Floods cause huge financial and human losses, and in order to reduce the losses caused by floods, it is necessary to predict the flood in advance and decide to evacuate quickly. This paper proposes a flood risk determination model using a CNN-based classification model so that timely evacuation decisions can be made using rainfall and water level data, which are key data for flood prediction. By comparing the results of the CNN-based classification model proposed in this paper and the DNN-based classification model, it was confirmed that it showed better performance. Through this, it is considered that it can be used as an initial study to determine the risk of flooding, determine whether to evacuate, and make an evacuation decision at the optimal time.

키워드 : DNN, K-Means Clustering, 시계열 데이터, 위험도 판별, 홍수

Keywords : DNN, K-Means Clustering, Time series data, Flood Risk Determination, Flood

Received 23 November 2021, **Revised** 24 November 2021, **Accepted** 29 November 2021

*** Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Professor, Department of Computer Engineering, Paichai University, Daejeon, 35345 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.3.341 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

지구온난화로 인한 이상기후로 세계 각지에서 막대한 양의 폭우 및 폭설로 인한 수해 피해가 지속적으로 발생하고 있다. 우리나라의 경우 2019년에 행정안전부에서 발간된 재해 연보 현황에 따르면 호우 및 태풍으로 인해 발생된 피해 금액이 전체 피해액의 95% 이상을 차지하고 있다. 또한 세계적으로 홍수 위험에 노출된 지역에 거주하는 인구가 2000년에 비해 25% 증가한 8,600만명에 이른다[1].

홍수 피해를 최소화하기 홍수 모니터링, 수위 예측 및 홍수 위험도 분류 등의 다양한 연구가 세계 각지에서 활발히 이루어지고 있으며, 대부분의 연구에서 사용된 핵심 파라미터는 수위 및 강수량을 가장 많이 사용하고 있다[2].

정확한 홍수 위험도 판단 기준에 의해 홍수의 위험도를 판별하고 위험 지역에 위치한 사람들을 대피시키는 것은 매우 중요한 문제이다. 그러나 현재 우리나라의 홍수 위험도 판별을 위한 지표는 규모가 큰 강 유역에만 단순히 고정된 수위값을 사용하여 홍수주의보, 홍수경보, 대홍수경보 등으로 분류한다.

위와 같은 고정된 수치적인 기준을 활용하여 규칙 기반으로 홍수를 판단하는 것 보다 최근 지속적으로 발전하고 다른 많은 분야에서 좋은 성능을 보이는 것이 검증된 기계 학습(Machine Learning) 및 심층 학습(Deep Learning)을 활용하여 홍수 위험도를 분류하는 모델을 본 논문에서 제안한다. 본 논문에서는 울산시 구주교의 수위 및 강수량 데이터를 활용하여 K-Means Clustering를 통해 각 시계열 데이터의 군집을 형성한다. 이후 형성된 군집을 통해 CNN(Convolution Neural Network) 기반의 분류 모델을 적용하여 최종적으로 홍수 위험도 판별 시스템을 제안한다.

Ⅱ. 관련 연구

본 장에서는 위험도 분류를 위한 관련 연구와 본 논문에서 사용된 K-Means Clustering과 CNN 분류 모델에 대해 기술한다. 홍수로 인한 피해는 세계 곳곳에서 발생하고 있으며, 홍수와 관련된 다양한 연구가 진행되고 있다. RNN(Recurrent Neural Network) 기반의 LSTM (Long Short Term Memory) 및 GRU(Gated Recurrent Units)와 같은 모델을 활용하여 수위를 예측하는 시스템, 수문관측용 시스템의 네트워크 트래픽 개선, 홍수 피해지역 표출 모델 등 홍수와 관련된 다양한 분야에서 많은 연구가 진행되고 있다[3-4].

2.1. K-Means Clustering

K-Means Clustering은 기계 학습에 해당하는 알고리즘으로 주어진 데이터를 k개의 군집으로 묶는 알고리즘이다. 비지도 학습의 클러스터링 모델 중 하나로 전체 데이터에 대해 각 군집과 거리 차이의 분산을 최소화하는 방식으로 동작하며, 수식은 다음과 같다.

$$X = S_1 \cup S_2 ... \cup S_k,\ S_i \cap S_j = \Phi \tag{1}$$

$$\arg\min s \sum_{i=1}^{K} \sum_{x_j \in c_i} \| x_j - c_i \|^2 \tag{2}$$

X는 전체 데이터셋을 의미하며, S의 경우 각 군집을 의미한다. k개의 군집으로 이루어져 있으며, 전체 군집의 합집합은 X가 되고, 각 군집의 교집합은 존재하지 않는다.

식 2의 c_i는 S의 중심점을 나타낸다. 따라서 각 데이터와 각 군집의 중심점의 거리가 가장 작은 군집에 속하게 된다. 위 과정이 전체 데이터가 군집에 포함되기까지 반복하여 최종적으로 전체 데이터에 대한 군집이 형성된다.

본 알고리즘은 다양한 분야에서 군집화를 위해 많이 사용되고 있으며 본 논문에는 수위와 강수량으로 이루어진 데이터에 대해 군집을 부여하기 위해 사용된다[5-6].

2.2. CNN

CNN은 분류를 위해 다양한 분야에서 사용되고 있다. 주로 두각을 나타내는 분야는 이미지 분류 및 객체 인식 분야이지만, 시·공간적 특징 추출에 강점을 가지고 있어 시계열 데이터 분류 및 예측 분야에서도 좋은 성능을 보인다. CNN을 활용한 어종 이미지 분류, 미세먼지 예측 모델, 텍스트 분류, 하천 수위 예측 등 다양한 분야에서 사용되고 있으며, 본 논문에서는 CNN 기반의 분류 모델을 구현한다[7-8].

Ⅲ. 학습 데이터 및 모델 설계

본 장에서는 학습 데이터셋을 설명하고 K-Means Clustering을 통한 군집화와 CNN 기반의 분류 모델, 결과 비교를 위한 DNN 분류 모델에 관해 기술한다. 전체 위험도 판별 모델의 구조는 그림 1과 같다.

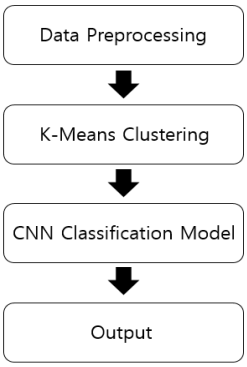

Fig. 1 Structure of the overall flood risk determination model

먼저 데이터셋에 대해 전처리를 진행하고, K-Means Clustering을 활용하여 군집을 형성한다. 다음으로 형성된 군집을 활용하여 CNN 분류 모델을 통해 최종적으로 분류 결과를 확인할 수 있다.

3.1. 학습 데이터 및 데이터 전처리

모델 학습 및 검증을 위해 필요한 데이터는 울산시에 위치한 구수교의 수위 및 울산 지역의 강수량 데이터를 사용하였다. 데이터의 측정 기간은 2015년 1월 1일부터 2021년 11월 19일까지이며, 일자별 시계열 데이터를 활용하였다. 울산의 구수교의 경우 2016년 기준 홍수 피해량이 가장 높게 기록된 지역이고 그 이후로도 잦은 호우 및 홍수 피해가 빈번하게 발생한 지역으로서 해당 지역의 데이터를 활용하여 실험을 진행하였다. 전체 데이터는 물 환경 정보 시스템을 참조하였으며[9], 그림 2, 3을 통해 전체 데이터를 확인할 수 있다.

훈련 데이터와 검증 데이터의 분할은 8:2 비율로 진행하였다. 전체 데이터는 2,106일치 데이터이며, 분할을 진행하여 최종적으로 훈련 데이터의 경우 1,666일치 데이터, 검증 데이터의 경우 417일치 데이터가 사용되었다.

강수량 데이터의 비가 오지 않은 날의 경우 결측값이 존재하여 결측값은 0으로 대체하여 진행하였고, 수위

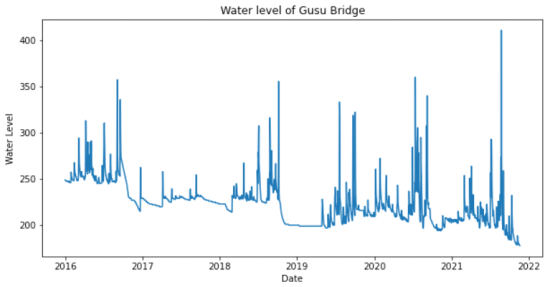

Fig. 2 Water level of Guju Bridge

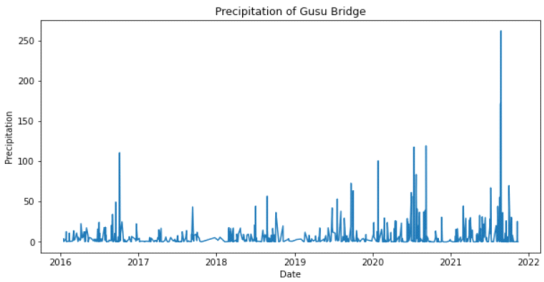

Fig. 3 Water level of Guju Bridge

데이터의 경우엔 단순히 데이터의 평균값 또는 특정값으로 사용하게 되면 시계열 데이터의 특성 상 신뢰성이 떨어지게 되므로 결측값을 과거 2일, 미래 2일의 데이터를 활용하여 4일치 데이터의 평균으로 대체하였다.

또한 MinMaxScaler를 통해 전체 데이터를 0~1 사이 데이터로 정규화 하였다. 이를 통해 데이터의 최솟값과 최댓값의 차이가 작아짐으로써 모델 훈련 시 더 좋은 성능을 얻는 것을 확인하였다.

3.2. K-Means Clustering을 활용한 군집화

데이터 전처리를 진행한 이후 CNN을 활용하여 분류 모델을 학습시키기 위해서 각 데이터의 레이블이 필요하다. 이 레이블을 생성하기 위해 K-Means Clustering을 활용한다. 군집화를 진행한 결과를 시각화한 것은 그림 4와 같다.

군집화를 진행할 때 군집의 수는 4개로 지정하였다. 그래프의 x축과 y축은 각각 수위 및 강수량을 나타내며, x축과 y축의 값이 커질수록 홍수 위험도가 높은 군집으로 판단할 수 있다. 색깔이 다른 4개의 군집을 확인할 수 있고, 검은색 삼각형은 각 군집의 중심을 나타낸다. cluster 3(하늘색)에 해당하는 군집을 보면 다른 군집에 비해 높은 수위 데이터를 가지고 있는 것을 확인할 수 있다.

군집화를 진행한 이후 사이킷런의 StandardScaler를 활용하여 데이터 정규화하였다. StandardScaler의 경우 평균은 0으로, 분산은 1로 조정하여 학습 시 오버피팅을 방지해주는 효과를 얻을 수 있고 정확도 측면에서도 더 좋은 결과가 나오는 것을 확인하였다.

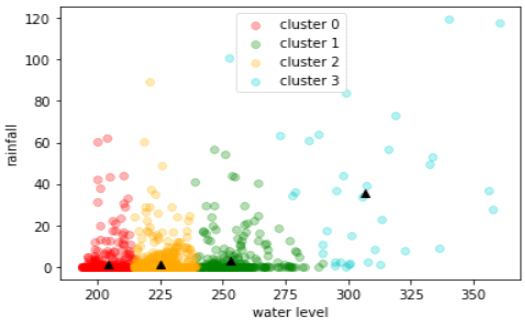

Fig. 4 K-Means Clustering Result

3.3. CNN 기반의 분류 모델

CNN 기반의 분류 모델 학습을 진행하기 위해 K-Means Clustering을 통해 생성된 레이블에 대해 원핫 인코딩(One-Hot Encoding)을 진행하고 입력 데이터의 모양을 3차원으로 변경하였다. 따라서 최종적으로 훈련용 입력 데이터의 모양은 (1666,1,1), 레이블의 모양은 (1666,4)의 형태를 가지게 되고 검증용 입력 데이터의 모양은 (417,1,1), 레이블의 모양은 (417,4)의 모양을 가지게 된다. CNN 모델의 경우 Inception V2 모델을 활용하였고 구성은 그림 5와 같다.

CNN 모델에 사용된 전체 하이퍼파라미터의 수는 16,286개가 사용되었다. 여러 구조로 실험을 진행하였을 때 위와 같은 구조를 사용할 때의 정확도 및 정밀도가 가장 좋은 결과를 얻을 수 있었다.

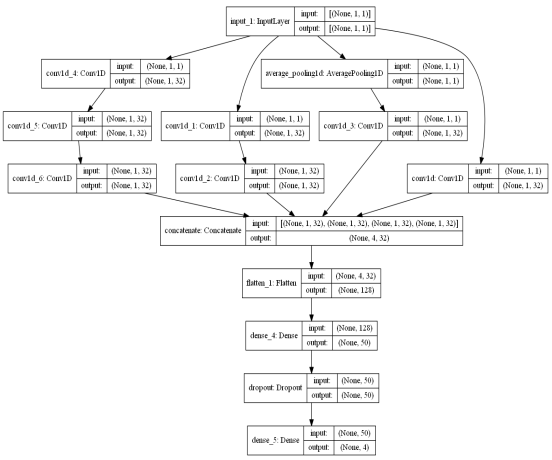

Fig. 5 Structure of CNN Model

3.4. DNN 모델 하이퍼파라미터

본 논문에서 제안하는 CNN 모델과의 성능 평가를 위해 분류 문제에서 많이 사용되는 DNN 분류 모델과 비교하였다. 모델의 구조는 그림 6과 같다.

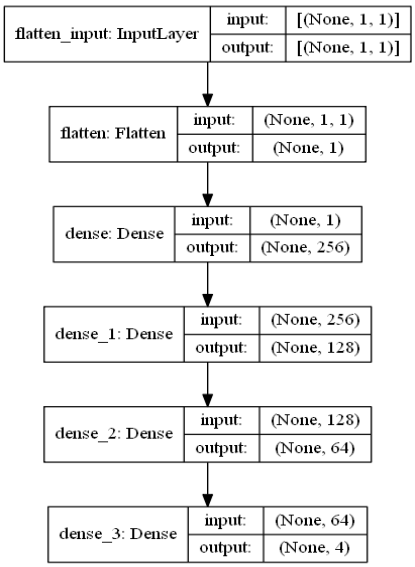

Fig. 6 Structure of DNN Model

Ⅳ. 실험 및 결과

4.1. 모델 비교

본 장에서는 CNN 및 DNN 분류 모델에 대해 실험하였다. 2가지 모델 모두 3.1절에서 언급한 데이터를 사용하였으며 모델의 구성은 그림 5, 그림 6과 같다. 또한 Adam, Adagrad, Momentum, RMSProp 4가지의 최적화 함수 비교를 진행한 결과 2가지 모델 모두 Adam을 사용했을 때 가장 좋은 결과를 얻어 최적화 함수로 Adam을 사용하였다. 손실 함수로는 다중 분류에서 사용되는 Categorical Cross-Entropy Loss를 사용하였다. 2가지 모델 모두 동일한 하드웨어에서 실험을 진행하였다. 훈련 횟수의 경우 300회, 배치 사이즈의 경우 1000으로 실험을 진행하였으며, 표 1은 학습한 하드웨어의 사양이다.

4.2. 실험 결과

그림 7~10은 CNN 및 DNN 모델에 대해 검증 시 손실값, 정확도, 정밀도, 재현율을 비교한 그래프이며, 표 2를 통해 최종 결과 수치를 확인할 수 있다.

Table. 1 System specification

OS	Windows 10
CPU	Intel i7-10700
GPU	Nvidia Geforce RTX 3070
RAM	64GB
Storage	SSD 500GB & HDD 2TB

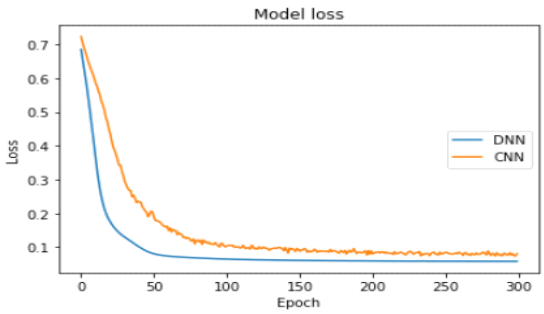

Fig. 7 Model validation loss graph

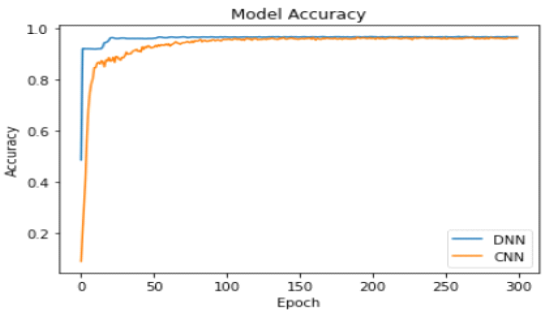

Fig. 8 Model validation accurarcy graph

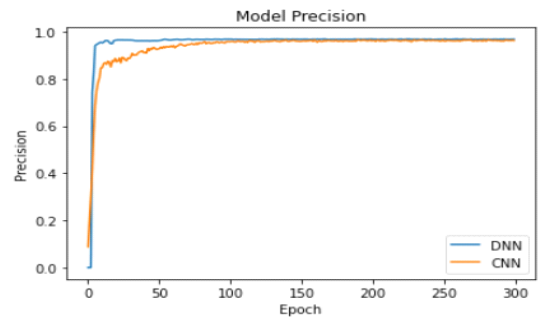

Fig. 9 Model validation precision graph

DNN 기반 분류 모델의 경우 사용된 하이퍼파라미터의 수는 41,924개이며, 훈련을 완료하기까지는 4.24초의 시간이 소요되었다. 검증 시 손실값, 정확도, 정밀도, 재현율의 경우 각 0.0577, 0.9681, 0.9664, 0.9599의 결과를 얻었다.

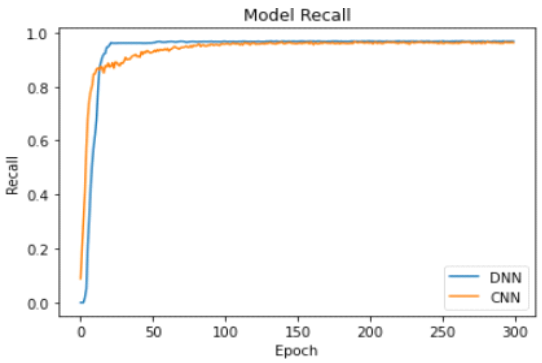

Fig. 10 Model validation recall graph

Table. 2 Model comparison

	DNN	CNN
Hyperparameter	41,924	16,286
Traning time(second)	4.24	7.29
Validation loss	0.0577	0.0782
Validation accuracy	0.9681	0.9712
Validation precision	0.9664	0.9688
Validation recall	0.9599	0.9681

본 논문에서 제안한 CNN 기반 분류 모델의 경우 사용된 하이퍼파라미터의 수는 16,286개이며, 훈련을 완료하기까지는 7.29초의 시간이 소요되었다. 최종적으로 검증 시 손실값, 정확도, 정밀도, 재현율의 경우 각 0.0782, 0.9712, 0.9688, 0.9681의 결과를 얻었다.

CNN 기반 모델이 훈련 시간은 조금 더 소요되었고 손실값은 조금 높은 결과를 보였지만, 분류 문제에서 가장 중요한 정확도를 포함한 정밀도, 재현율의 지표에서 더 좋은 성능을 보이는 것을 확인하였다.

Ⅴ. 결 론

홍수의 피해를 최소화하기 위한 시기적절한 대피 결정은 매우 중요한 문제 중 하나이다. 이를 위해 본 논문에서는 홍수 위험도 판별을 위한 CNN 모델을 제안하였

고, 홍수 위험도 분류에 대해 높은 수준의 정확도로 판별할 수 있는 연구 결과를 얻을 수 있었다.

CNN과 DNN 기반의 모델을 비교하여 CNN의 우수성을 확인할 수 있었고, 96% 이상의 분류 정확도를 보이는 것을 확인할 수 있었다.

그러나 사용된 데이터의 수가 많지 않아 추후 데이터셋을 보완한 추가 연구가 진행될 필요가 있을 것으로 판단된다. 그리고 단순히 군집화를 적용하는 것이 아닌 홍수의 위험도를 판단하는 최적의 기준을 적용하는 연구를 향후 연구로 진행하여 현재 시스템에 적용한다면 현재 제안한 모델보다 더 좋은 결과를 얻어낼 수 있을 것으로 사료된다.

ACKNOWLEDGEMENT

This study was carried out with the support of R&D Program for Forest Science Technology (Project No. 2021340A00-2123-CD01) provided by Korea Forest Service(Korea Forestry Promotion Institute).

REFERENCES

[1] Ministry of Public Administration and Security. 2019 Disaster Yearbook [Internet]. Available: https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do;jsessionid=9q+z++-8qP6PFH1L9NfdGfxr.node20?bbsId=BBSMSTR_000000000014&nttId=81886.

[2] N. A. Maspo, A. N. B. Harun, M. Goto, F. Cheros, N. A. Haron, and M. N. M. Nawi, “Evaluation of Machine Learning approach in flood prediction scenarios and its input parameters: A systematic review,” in *IOP Conference Series: Earth and Environmental Science*, vol. 479, 2020.

[3] S. T. Hong, J. H. Park, and H. K. Jung, “Network traffic analysis of satellite communication system for hydrologic observation,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 23, no. 9, pp. 1139-1145, Sep. 2019.

[4] S. H. Park and H. J. Kim, “Design of Artificial Intelligence Water Level Prediction System for Prediction of River Flood,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 24, no. 2, pp. 198-203, Feb. 2020.

[5] J. Y. Kim, B. S. Kang, and H. K. Jung, “Determination of coagulant input rate in water purification plant using K-means algorithm and GBR algorithm,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 6, pp. 792-798, Jun. 2021.

[6] M. S. Jang, K. W. Nam, and Y. S. Lee, “Analysis and Application of Power Consumption Patterns for Changing the Power Consumption Behaviors,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 4, pp. 603-610, Apr. 2021.

[7] J. H. Park, K. B. Hwang, H. M. Park, and Y. K. Choi, “Application of CNN for Fish Species Classification,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 23, no. 1, pp. 39-46, Jun. 2019.

[8] M. Pan, H. Zhou, J. Cao, Y. Liu, J. Hao, S. Li, and C. H. Chen, “Water level prediction model based on GRU and CNN,” *IEEE Access*, vol. 8, pp. 60090-60100, Mar. 2020.

[9] Water environment information system [Internet]. Available: http://water.nier.go.kr/

조민우(Minwoo Cho)

2021년 배재대학교 컴퓨터공학(공학사)
2021년~현재, 배재대학교 컴퓨터공학 석사과정
※ 관심분야 : Deep Learning, Machine Learning, Big data

김동수(Dongsoo Kim)

2000년 공주교육대학교 초등교육학(학사)
2018년 공주교육대학교 로봇교육 석사과정(석사)
2020년 ~ 현재 배재대학교 컴퓨터공학과 박사과정
2020년 ~ 현재 상곡초등학교 교사
※관심분야 : AI, SW, IoT, 메이킹, 메카트로닉스

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년~현재 배재대학교 컴퓨터공학과 교수
※관심분야 : Machine learning, Big data, Embedded system, U-Healthcare, IoT