

상수도관 누수 탐지를 위한 머신러닝 및 딥러닝 기법 비교 연구*

이원준¹, 강대희^{1*}, 김광주², 이연창³
조지 메이슨 대학¹, 울산과학기술원², 한국전자통신연구원³
wlee40@gmu.edu, {kangdaehee, yeonchang}@unist.ac.kr, kwangju@etri.re.kr

A Comparative Study of Machine Learning and Deep Learning Methods for Water Pipeline Leakage Classification

Won-June Lee, Dae-Hee Kang, Kwang-Ju Kim, Yeon-Chang Lee
George Mason University, UNIST, ETRI

요약

도시 상수도망의 안정적 운영을 위해서는 누수를 조기에 정확하게 탐지하는 것이 필수적이다. 본 연구에서는 대구광역시 상수도관에 설치된 391개 센서로부터 수집된 주파수 스펙트럼 시계열 데이터를 활용하여, 전통적 머신러닝과 딥러닝 기법의 누수 탐지 성능을 체계적으로 비교하고자 한다. 실험 결과, Random Forest는 전반적으로 가장 안정적이고 해석 가능한 기준선(정확도 약 0.93)을 보였으며, 다층 퍼셉트론 (Multi-Layer Perceptron)은 충분한 데이터 증강이 이루어질 경우 표현학습 기반 접근의 잠재력을 보여주었다. 이러한 결과는 실무형 기준선을 제시함과 동시에, 향후 고도화된 모델 개발 방향을 시사한다.

1. 서론

상수도관의 누수 탐지는 도시 인프라의 안정성과 회복탄력성을 좌우하는 핵심 기술이다[1]. 누수가 장기간 방치될 경우 수자원 손실과 수익 저하를 넘어, 토양 침하·도로 파손·2차 오염으로 이어져 인명 피해를 야기할 수 있다. 이는 노후 관로의 증가, 기후 변화 유동성, 지반 조건의 악화 등이 복합적으로 작용한 결과이다. 이처럼, 상수도관의 누수는 여러 원인으로부터 비롯되는 위험이기에 지속

적인 모니터링과 발 빠른 대응이 요구된다. 따라서, 상수도관의 설치 이후부터 유지 보수로 이어지는 전 주기에서 정확도와 신속성을 동시에 충족하는 자동 누수 탐지 체계의 구축이 필수적이다.

이와 같은 이상 탐지 문제는 전통적인 머신러닝(예: K -Nearest Neighbors (K -NN) [2], Random Forest [3]) 기술이 구축 용이성과 비용 효율성 측면에서 실무적 강점을 보여 왔다[1, 2]. 그러나 실세계 환경에서 수집된 주파수나 압력 시계열 데이터는 펌핑 이벤트나 환경 잡음 등 비정상성(non-stationarity)을 포함한다[4]. 그러므로, 국소 특성에 의존한 분류만으로는 미세 유량, 진동 신호의 누적 변화를 정상 패턴의 추세로부터 안정적으로 분리하기 어렵다. 또한, 고정된 맥락에서 추출한 통계와 주파수 특성만으로는 시간대·요일·계절에 따른 정상 범주의 이동을 일관적으로 설명하는 데 한계가 존재한다.

최근 표현학습 기반 접근은 이러한 제약을 보완하는 데에 효과적인 것으로 알려져 왔다 [1]. 예를 들어, 다층 퍼셉트론(Multi-Layer Perceptron, MLP)은 고정 길이 특징

* 본 연구는 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2020-II201336, 인공지능대학원지원(울산과학기술원)), 한국연구재단의 지원(RS-2025-24533171), 그리고 DGIST 연구개발사업(2025040026)과 한국정부(과학기술정보통신부)가 지원하는 ETRI 연구사업(25ZD1120, 지역산업 IT융합 기술개발 및 지원사업)의 지원을 받아 수행되었습니다.

¹ 공동 1저자

[†] 교신저자

벡터를 입력으로 받아 비선형 변환을 계층적으로 적용함으로써, 입력 데이터 간의 상호작용과 비선형적 경계를 간결한 구조로 학습한다[5]. 합성곱 신경 (Convolutional Neural Network, CNN)은 국소 패턴을 추출하고, 계층적인 필터를 통해 주파수의 대역에 따른 조합적 특징을 점진적으로 구성한다[6].

따라서, 우리는 대구시의 실세계 상수도관 주파수 센서로부터 수집된 데이터를 이용하여, 전통적인 머신 러닝과 딥러닝 방법론들의 누수 탐지 능력을 비교 분석하고자 한다. 이를 통해 입력 표현, 모델 복잡도, 해석 가능성의 균형, 그리고 주파수 신호 내 급작스러운 변화에 대한 대응 전략을 종합적으로 논의하고, 누수 탐지 기술이 나아가야 할 실무적 방향성을 제시하고자 한다.

2. 누수 탐지 방법론

상수도관 누수 탐지 문제는 각 센서로부터 수집된 주파수 스펙트럼 시계열을 입력으로 하여, 해당 센서의 상태 레이블(예: 정상, 일반 누수, 미세 누수)을 예측하는 다중 분류 문제로 간주될 수 있다. 따라서, 우리는 두 가지 전통적 머신 러닝 방법(K-NN, Random Forest)과 두 가지 딥러닝 방법(MLP, CNN)을 활용하고자 한다.

2.1 전통적 머신 러닝 기법

K-NN(K-Nearest Neighbors)[2] 기반 분류는 별도의 학습 과정을 요구하지 않는 방법으로, 먼저 거리 함수(예: Hausdorff 거리 [7])를 통해 주어진 센서 데이터와 나머지 센서 데이터 간 유사도를 계산한다. 이후, K 개의 최근접 이웃 센서 데이터에서 가장 많이 관찰되는 레이블로 해당 센서 데이터의 레이블을 예측한다. **Random Forest**[3] 분류는 여러 개의 의사결정 트리(Decision Tree)를 조합한 모델로, 학습 데이터 $L = (x_i, y_i), i = 1, \dots, n$ 이 있을 때, L 로부터 B 개의 부트스트랩 샘플 L_1, \dots, L_B 을 만들어 낸다. 이후, 개별 의사결정 트리 $\phi(x, L)$ 를 학습시켜 개별 트리로부터 얻은 출력 값을 집계해 최종 예측 레이블을 얻는다. 이를 통해 주파수별 중요도를 계산하는 방식으로 작동한다.

2.2 딥러닝 기법

MLP[5]를 활용하기 위해, 우리는 식 1을 기반으로 각 센서의 초기 입력 값으로 하루치 주파수 스펙트럼을 갖는 벡터 x_i 를 사용한다. 이러한 주파수 입력 값은 배치 정규화를 통해 안정화한 뒤, 주파수 대역 간 상호작용을 은닉층(hidden layer) 내에서 학습한다. 동시에 과적합 방지를 위한 Dropout을 활용한다.

$$h_1 = ReLU(BN(W_1 x_i + b_1)). \quad (1)$$

더 나아가, 보다 밀도 있는 패턴 파악을 위해 식 2를 통해 차원을 감축해 표현 학습을 수행한다. 최종적으로, 식 3과 같이 소프트 맥스를 통해 가장 높은 확률값을 가진 클래스로 주파수 대역의 분류를 예측한다.

$$h_2 = ReLU(W_2 h_1 + b_2), \quad (2)$$

$$\hat{y} = softmax(W_{out} h_2 + b_3). \quad (3)$$

CNN[6]은 인간의 시신경 구조를 모방한 기술로 객체를 인식하기 위한 패턴을 찾는데 특히 유용한 딥러닝 기법이다. 특히, CNN은 필터 혹은 커널이라고 불리는 개념을 활용해 입력에서 고정된 범위만큼 슬라이딩하며 특징을 추출하여, 주어진 입력 데이터의 전반적인 특징이 요약된 특징 맵(feature map)을 생성한다.

주파수 벡터를 입력으로 사용할 때, CNN은 여러 개의 작은 필터가 주파수 축을 따라 슬라이딩하며 국소 구간의 패턴을 감지한다. 각 필터는 피크, 완만한 대역 상승, 하모닉 반복처럼 특정 패턴에 강하게 반응하고, 그 반응을 특징 맵으로 만든다. 이어서 비선형성을 부여하고 풀링(Max/Average, 또는 Global Average Pooling)으로 위치가 몇 개 빈(bin)을 이동해도 반응이 안정적일도록 요약한다. 층을 쌓을수록 수용영역이 넓어져 좁은 피크에서 시작해 대역 전이와 다중 피크의 공존 같은 더 큰 맥락의 조합적 패턴을 포착하게 되며, 마지막에는 요약 벡터를 전결합층(Fully-connected layer)과 소프트 맥스에 통과시켜 정상·일반·미세 클래스 분류에 대한 확률을 출력한다.

3. 실험

3.1 데이터 집합

본 연구에서는 대구광역시 상수도관에 설치된 391개의 주파수 센서로부터 5일간 연속적으로 수집된 데이터를 사용하였다. 각 센서 데이터는 1일 단위로 분할되어 사용되며, 동일한 센서 데이터로부터 수집된 주파수는 동일한 레이블 $y \in \{0: \text{정상}, 1: \text{일반 누수}, 2: \text{미세 누수}\}$ 을 할당하였다. 본 논문의 실험에서 사용한 데이터 집합의 통계는 <표 1>과 같다.

<표 1> 데이터 집합의 통계

센서 인스턴스	주파수	정상 (0)	일반 누수 (1)	미세 누수 (2)
1,955	512	690	705	560

3.2 실험 방법

본 논문에서는 동일한 전처리와 평가 프로토콜 하에서 전통적 머신 러닝과 딥러닝 기법의 누수 탐지 성능을 비교하고자 한다. 이를 위해, 우리는 2장에서 설명한 네 가지 방법에 대해 다음과 같은 평가를 수행한다.

먼저, 각 센서 인스턴스는 512개의 주파수를 가진 1일치의 측정 값을 하나의 1차원 입력 벡터로 가지며, 해당 센서의 정답 레이블은 정상 (0), 일반 누수 (1), 미세 누수 (2) 중 하나가 할당된다. 실험을 위해, 우리는 데이터 집합을 8:1:1의 비율로 학습, 검증, 그리고 평가 데이터로 분할하였다. 이후, 각 방법의 성능은 이상 탐지 문제에서 널리 활용되는 다음 네 가지 지표를 사용하였다: Accuracy, Precision, Recall, F1 Score.

3.3 실험 결과

전통적 머신 러닝과 딥러닝 기법의 적용에 따른 누수 탐지 실험 결과를 <표 2>에 정리하였다. 실험 결과를 통해, 우리는 Random Forest가 전반적으로 가장 우수한 성능(Acc.≈0.934)을 보이는 것을 확인하였다. Random Forest는 정상·일반·미세 클래스 모두에서 F1이 높을 뿐만 아니라, 미세 누수의 경우 Recall=1.000, F1=0.974로 미탐을 거의 발생시키지 않았다. MLP는 두 번째로 안정적이며(Acc.≈0.865), Precision 지표에서는 근소하게 Random Forest 방법을 앞서는 성능을 보였다. 반면, CNN은 국소 패턴 포착에는 강점이 있으나 본 데이터에서는 일반 누수 Recall(≈0.563)이 낮아 미탐이 상대적으로 많았다. 반면에 Hausdorff 기반 K-NN은 전반적으로 가장 취약한 성능으로-(Acc.≈0.345), 고차원 스펙트럼에서 거리 기반 분류는 한계를 보였다. 이는 예외적 아웃라이어의 존재와 클래스 간 경계가 비선형임을 시사한다.

<표 2> 누수 탐지 정확도 측정 결과

방법	Label	평가 지표			
		Prec.	Recall	F1	Acc.
K-NN	정상(0)	0.3889	0.4000	0.3944	0.3450
	일반(1)	0.3279	0.2857	0.3053	
	미세(2)	0.3134	0.3500	0.3307	
Random Forest	정상(0)	0.9385	0.8841	0.9104	0.9337
	일반(1)	0.9167	0.9296	0.9231	
	미세(2)	0.9492	1.0000	0.9739	
MLP	정상(0)	0.9492	0.8000	0.8682	0.8650
	일반(1)	0.8243	0.8714	0.8472	
	미세(2)	0.8358	0.9333	0.8819	
CNN	정상(0)	0.7568	0.8116	0.7832	0.7602
	일반(1)	0.7692	0.5634	0.6504	
	미세(2)	0.7571	0.9464	0.8413	

<표 2>의 결과는 여전히 Random Forest가 실무형 기준선으로 강력함을 보여준다. 다만 MLP가 전반적으로 두 번째로 높은 성능을 기록했다는 점은, 데이터가 충분할수록 딥러닝 기반 표현학습의 잠재력을 기대해볼 수 있음을 시사한다.

4. 결론

본 연구에서는 사회 기반 시설의 주요 사안인 누수 탐지에 있어서 전통적 머신 러닝과 딥러닝 기법의 성능을 분석하였다. 전반적으로 전통적인 머신 러닝 기법인 Random Forest가 가장 안정적인 분류 성능을 보였으며, 클래스별 F1과 정확도에서 균일한 우위를 보였다. 동시에 MLP가 두 번째로 높은 성능을 기록함으로써, 데이터가 충분

할 경우 딥러닝 기반 표현학습의 잠재력을 기대해볼 수 있음을 확인하였다.

향후에는 Graph Transformer와 같은 고도화된 시계열 분석 프레임워크를 기반으로 일반 누수와 미세 누수의 경계를 더 정교하게 구분하고자 한다[8, 9]. 특히 Graph Transformer는 원거리 센서 간 상관성을 직접 모델링할 수 있고, 어텐션 가중치로 의사결정 근거를 시각화할 수 있다는 장점이 있다. 위 파이프라인의 확장을 통해 실시간 탐지 정확도와 해석 가능성을 함께 개선하는 방향의 발전을 기대한다.

참고 문헌

- [1] Basnet, L., Brill, D., Ranjithan, R., Mahinthakumar, K., "Supervised machine learning approaches for leak localization in water distribution systems: Impact of complexities of leak characteristics," Journal of Water Resource Planning and Management 149 (8), 2023.
- [2] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
- [3] Breiman, L., "Random Forests," Machine Learning 45, 5-32, 2001.
- [4] Lauren McMillan, Jawad Fayaz, Liz Varga, "Flow forecasting for leakage burst prediction in water distribution systems using long short-term memory neural networks and Kalman filtering," Sustainable Cities and Society, vol. 99, 2023.
- [5] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408, 1958.
- [6] Yann Lecun, Yoshua Bengio. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 978-0-262-51102-5 1998.
- [7] D. P. Huttenlocher, G. A. Klanderman and W. J. Rucklidge, "Comparing images using the Hausdorff distance," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pp. 850-863, 1993.
- [8] Vaswani, Ashish, et al. "Attention is all you need," Advances in neural information processing systems 30, 2017.
- [9] Yun, Seongjun, et al. "Graph transformer networks," Advances in neural information processing systems 32, 2019.