Journal of Wetlands Research
Vol. 25, No. 4, November 2023, pp. 306-314

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2023.25.4.306

기계학습 분류모델을 이용한 하천퇴적물의 중금속 오염원 식별

반민정[*] · 신상욱[**] · 이동훈[*] · 김정규[***] · 이호식[****] · 김영[*****] · 박정훈[******] · 이순화[*******] · 김선영[********] · 강주현[*†]

[*]동국대학교
[**]GS건설
[***]고려대학교
[****]국립한국교통대학교
[*****]고려대학교 세종캠퍼스
[******]전남대학교
[*******]영남대학교
[********]국립환경과학원

Identifying sources of heavy metal contamination in stream sediments using machine learning classifiers

Min Jeong Ban[*] · Sangwook Shin[**] · Dong Hoon Lee[*] · Jeong-Gyu Kim[***] · Hosik Lee[****] · Young Kim[*****] · Jeong-Hun Park[******] · ShunHwa Lee[*******] · Seon-Young Kim[********] · Joo-Hyon Kang[*†]

[*]*Dongguk University-Seoul*
[**]*GS E&C*
[***]*Korea University*
[****]*Korea National University of Transportation*
[*****]*Korea University-Sejong*
[******]*Chonnam National University*
[*******]*Yeungnam University*
[********]*National Institute of Environmental Research*

(Received : 18 August 2023, Revised : 17 October 2023, Accepted : 17 October 2023)

요 약

하천퇴적물은 유역내 다양한 오염원으로부터 발생하는 중금속, 유기물 등 오염물질의 수용체일 뿐만 아니라 수질 오염 및 수생태 악영향을 유발할 수 있는 2차적 오염원이기에 중요한 관리대상이라고 할 수 있다. 오염된 하천퇴적물의 효과적인 관리를 위해서는 오염원에 대한 식별과 이와 연계된 관리대책의 수립이 우선되어야 한다. 본 연구는 하천퇴적물내 측정된 다양한 이화학적 오염항목 분포 특성에 기반하여 퇴적물의 주요 오염원을 식별하기 위한 방법으로서 기계학습모델의 적용성을 평가하였다. 기계학습 모델의 성능 평가를 위해 전국 4대강 수계내 주요 폐금속광산 및 산업단지 인근에서 수집된 총 356개의 하천퇴적물에 대한 중금속 10개 항목(Cd, Cu, Pb, Ni, As, Zn, Cr, Hg, Li, Al)과 토양항목 3개(모래, 실트, 점토 비율) 수질항목 5개(함수율, 강열감량, 총유기탄소, 총질소, 총인)를 포함한 총 18개 오

[†]To whom correspondence should be addressed.
Department of Civil and Environmental Engineering, Dongguk University-Seoul
E-mail : joohyon@dgu.ac.kr

- **Min Jeong Ban** Dongguk University-Seoul / Graduate student (mjban99@dgu.ac.kr)
- **Sangwook Shin** GS E&C / Leader (sangwookshin@gsenc.com)
- **Dong Hoon Lee** Dongguk University-Seoul / Research fellow (leedonghoon@dgu.ac.kr)
- **Jeong-Gyu Kim** Korea University / Professor (lemonkim@korea.ac.kr)
- **Hosik Lee Korea** National University of Transportation / Professor (hlee@ut.ac.kr)
- **Young Kim** Korea University-Sejong / Professor (kimyo@korea.ac.kr)
- **Jeong-Hun Park** Chonnam National University / Professor (Parkjeo1@jnu.ac.kr)
- **ShunHwa Lee** Yeungnam University / Professor (leesh@yu.ac.kr)
- **Seon-Young Kim** National Institute of Environmental Research / Researcher (ecosun0@korea.kr)
- **Joo-Hyon Kang** Dongguk University-Seoul / Professor (joohyon@dgu.ac.kr)

염항목에 대한 분석자료를 활용하였다. 기계학습 분류 모델로서 선형판별분석(linear discriminant analysis, LDA)과 서포트벡터머신(support vector machine, SVM) 분류기를 사용하여 폐금속광산('광산')과 산업단지('산단') 인근에서의 하천퇴적물 시료의 분류 성능을 평가한 결과, 채취 지점 및 시기별 4가지 경우(비강우시 광산, 강우시 광산, 비강우시 산단, 및 강우시 산단)에 대한 퇴적물 시료의 분류 성능이 우수하였으며, 특히 비선형 모델인 SVM(88.1%)이 선형모델인 LDA(79.5%) 보다 퇴적물을 분류하는데 있어 보다 우수한 성능을 나타냈다. SVM 앙상블 기반 비배타적 다중라벨분류기 모델을 이용하여 각 시료채취 지점 상류 유역 1km 반경 내 지배적인 토지이용 및 오염원을 다중 타겟값으로 다중분류 예측을 수행한 결과, 폐금속광산과 산업단지의 분류는 비교적 높은 정확도로 수행하였으나, 도시와 농업지역 등 다른 비점오염원에 대한 분류정확도는 56~60%범위로 비교적 낮게 나타났다. 이는 다중라벨 분류모델의 복잡성에 비해 데이터셋의 크기가 상대적으로 작아서 발생한 과적합에 기인한 것으로 향후 보다 많은 측정자료가 확보될 경우 기계학습 모델을 적용한 오염원 분류의 정확도를 보다 향상시킬 수 있을 것으로 판단된다.

핵심용어 : 하천퇴적물, 기계학습, 중금속, 오염원, 분류모델, 토지이용

Abstract

Stream sediments are an important component of water quality management because they are receptors of various pollutants such as heavy metals and organic matters emitted from upland sources and can be secondary pollution sources, adversely affecting water environment. To effectively manage the stream sediments, identification of primary sources of sediment contamination and source-associated control strategies will be required. We evaluated the performance of machine learning models in identifying primary sources of sediment contamination based on the physico-chemical properties of stream sediments. A total of 356 stream sediment data sets of 18 quality parameters including 10 heavy metal species(Cd, Cu, Pb, Ni, As, Zn, Cr, Hg, Li, and Al), 3 soil parameters(clay, silt, and sand fractions), and 5 water quality parameters(water content, loss on ignition, total organic carbon, total nitrogen, and total phosphorous) were collected near abandoned metal mines and industrial complexes across the four major river basins in Korea. Two machine learning algorithms, linear discriminant analysis (LDA) and support vector machine (SVM) classifiers were used to classify the sediments into four cases of different combinations of the sampling period and locations (i.e., mine in dry season, mine in wet season, industrial complex in dry season, and industrial complex in wet season). Both models showed good performance in the classification, with SVM outperformed LDA; the accuracy values of LDA and SVM were 79.5% and 88.1%, respectively. An SVM ensemble model was used for multi-label classification of the multiple contamination sources inlcuding landuses in the upland areas within 1 km radius from the sampling sites. The results showed that the multi-label classifier was comparable performance with sinlgle-label SVM in classifying mines and industrial complexes, but was less accurate in classifying dominant land uses (50~60%). The poor performance of the multi-label SVM is likely due to the overfitting caused by small data sets compared to the complexity of the model. A larger data set might increase the performance of the machine learning models in identifying contamination sources.

Key words : Freshwater Sediment, Machine Learning, Heavy metal, Pollutant sources, Classifier, Landuse

1. 서 론

퇴적물은 오염원에서 장기간 오염물질이 축적되는 경우 유해물질의 저류지 역할을 하며 수역 환경에 지속적인 악영향을 미친다(Kang et al., 2009). 수질 및 퇴적물의 오염은 저서 수생태계의 오염을 야기하며 오염 퇴적물 준설시 육상생태계에도 큰 악영향을 초래한다. 오염된 퇴적물은 다양한 유기물질과 영양물질, 중금속 등이 축적될 수 있으며 특히 하천 퇴적물은 중금속 오염에 매우 취약하다(Liber et al., 2019). 많은 중금속과 유기화합물은 생태독성과 생물농축 등으로 수생생물과 사람에게 유해한 영향을 미치므로 퇴적물은 저서 수생태계 관리를 위한 중요한 요소이다(Ali et al., 2019; Zhang and Mei, 2015).

최근 수생환경에서 중금속은 주로 인위적인 오염원의 영향으로 검출되고 있다(Jiang et al., 2021). 산업활동과 광산활동 등 토지 개발 과정에서의 인위적 활동들은 중금속의 주요 발생원으로 평가되고 있다(Wu et al., 2021; Navarro et al., 2008). 광산활동이 중단된 폐금속광산 또한 폐광산 인근 폐기물의 집중강우나 강풍으로 인한 유출과 갱내수 유출을 통해서도 토양 및 수질오염을 야기할 수 있으므로, 오염원 중심의 지속적인 관리가 필요하다(Jung et al., 2006). 일반적으로 유역내 명확한 오염원이 존재하는 경우 퇴적물에는 고농도의 중금속이 검출되므로 퇴적물 조사를 통해 오염원 특정 및 관리방안 수립이 가능하다(Guan et al., 2018; Sojka et al., 2022).

환경부는 수계 퇴적물에 대한 오염 현황 모니터링과 관리를 위하여 퇴적물측정망을 운영하고 있으며 이 외에도 광산 및 산업단지 등 오염우심지역을 중심으로 한 다양한 조사 사업을 통하여 퇴적물에 포함된 중금속류와 유기물, 영양염류 등 다항목 측정 데이터를 지속적으로 생산·관리하고 있다. 이러한 측정 데이터가 퇴적물 및 수질관리를 위한 실효적인 대책 수립을 위한 기초자료로서 활용되기 위해서는

유역내 오염원 및 퇴적물 오염특성 간의 상호 관련성에 대한 정량·정성적 분석과 함께, 이를 토대로 한 퇴적물의 주요 오염 원인에 대한 식별이 필요하다. 이와 관련하여 기존의 연구사례를 살펴보면 유역내 오염물질의 발생 및 퇴적물까지의 도달과정에 대한 물리적 모델링 기술에 기반하여 오염원과 퇴적물간의 정량적 관계를 해석하거나 다중회귀모델과 같은 통계적 기법에 기반하여 퇴적물 오염에 대한 주요원인을 파악하는 연구가 주를 이루어 왔다(Stoichev et al., 2020; Chueh et al., 2021). 유역모델과 같은 물리적 모델의 경우 수많은 매개변수를 포함하고 있기 때문에 모델의 보정과정에서 큰 불확실성이 존재하며(dos Santos Simoes et al., 2008), 통계적 모델과 같은 데이터 기반 모델은 해당지역에서 충분한 데이터의 수집이 이루어져야 하므로 다양한 변수를 적용하기 위해 많은 시간과 비용이 소모된다(Srinivasan et al., 2010). 그러나 상술한 바와 같이 환경측정망 운영 등 퇴적물을 비롯한 환경의 질에 대한 상시적인 모니터링 프로그램의 운영을 통하여 데이터의 크기가 지속적으로 증가하고 있으므로, 다양한 종류의 환경 변수와 오염원에 대한 특성을 수용하고, 복잡하고 비선형적인 오염원의 영향을 반영할 수 있는 기계학습과 같은 데이터 기반 모델링 연구의 중요성이 점차 증가하고 있다(Zhang et al., 2022; Zhu et al., 2022; Cha & Kim, 2020; Chung et al., 2020).

본 연구에서는 하천 퇴적물의 중금속 오염원 식별 및 오염특성 비교 분석을 위한 방안으로 기계학습 모델의 적용성을 평가하였다. 4대강 수계에 분포한 주요 폐금속광산 및 산업단지 인근에서 수집된 356개 하천퇴적물 시료의 10개 중금속 항목, 5개 수질항목, 그리고 3개 토양 항목 등 총 18개 항목 분석 자료를 활용하였으며, 기계학습 모델로서 선형판별분석(linear discriminant analysis, LDA)와 서포트벡터머신(support vector machine, SVM) 분류기를 사용하여 퇴적물 시료 채취시의 기상조건(강우 및 비강우) 및 인근 유역특성(폐금속광산 및 산업단지)에 따른 퇴적물 시료의 분류 성능을 평가하였다. 나아가 비배타적 다중라벨분류 모델을 사용하여 상류 방향 영향권내 주요 토지이용(도시 및 농업), 폐금속광산 유무, 산업단지 유무 및 산업폐수처리장 유무 등 총 5개 오염원 분류를 통해 유역내 잠재적 오염원의 식별 가능성을 평가하였다.

2. 연구방법

2.1 데이터 확보

본 연구를 위하여 2018-2019년 기간 동안 4대강 수계 18개 중권역내(Fig. 1) 폐금속광산 및 산업단지 인근 총 205개 하천지점에서 채취한 퇴적물 시료 총 356개에 대한 분석데이터를 국립환경과학원으로부터 제공받았다. 퇴적물 데이터가 수집된 지점들은 폐광산 6개소, 산업단지 8개소, 산업단지 및 폐광산 복합지역 2개소 등 총 16 지역 인근에

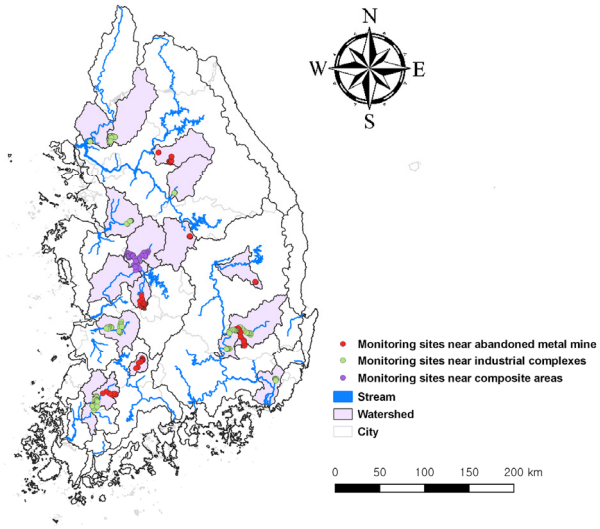

Fig. 1. Sampling Locations of the Stream Sediments

위치한 하천 지점들이다. 연구지역내 폐광산은 과거 금속광산으로 활용되었던 지역들로 중금속 오염 우려가 높으며, 산업단지의 경우 기계, 조립금속, 섬유, 의복, 전기, 전자, 자동차 부품 제조 등 여러 업종이 혼재된 복합 단지들로서 다양한 중금속 배출이 우려되는 지역들이다.

총 각 하천퇴적물 데이터셋은 모래, 실트, 점토, 함수율, 강열감량(LOI), 총유기탄소(TOC), 총질소(TN), 총인(TP), Cd, Cu, Pb, Ni, As, Zn, Cr, Hg, Li, Al 등 총 18개 항목에 대한 농도값을 포함하고 있다(Table 1). 각각의 비강우시 및 강우시 데이터셋에 대하여 금강유역을 제외한 퇴적물 채취지점은 인근의 특성이 폐금속광산 또는 산업단지로 확연히 구분되므로 해당 퇴적물 데이터셋은 인근 유역특성에 따라 각각 '광산(MM)'과 '산단(IC)'으로 분류하였다. 금강유역에서의 퇴적물 데이터는 모두 비강우시 채취된 시료에 대한 것으로 폐금속광산 및 산업단지가 복합적으로 존재하는 특성을 가지기 때문에 '혼합지역'으로 분류하였다. 혼합지역 데이터의 경우, 상대적으로 산업단지가 광범위하게 분포된 '혼합지역-A(Composite-A)'와 상대적으로 폐금속광산지점이 많은 '혼합지역-B(Composite-B)'로 구분하였다.

2.2 통계분석

퇴적물과 토양 오염도의 오염원별(산업단지, 폐금속광산) 및 시기별(강우시, 비강우시) 오염특성을 비교하여 각 토지이용을 특징짓는 오염항목과 특이적 농도변화를 갖는 오염항목을 도출하여 하천퇴적물 오염원 관리를 위한 우선순위 항목을 제시하기 위해 다변량 분산분석(Multi-variate Analysis of Variance, MANOVA)을 수행하였다. MANOVA는 강우유무와 퇴적물 오염원이 퇴적물에 미치는 영향 특성에 대해 통계소프트웨어 SPSS를 이용하였다. MANOVA 수행 후 일원배치 ANOVA로 측정항목별 사후검증(post-hoc test)을 수행하여 각 조건별 농도 수준 차이

Table 1. Summary Statistics of the Sediment Quality

Parameter	Metal Mine (MM)						Industrial Complex (IC)						Composite Area (CA)		
	Dry (N=108)			Wet (N=31)			Dry (N=130)			Wet (N=22)			Dry (N=65)		
	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std
Clay(%)	6.97	2.80	9.50	4.91	2.10	4.18	7.61	3.40	16.19	7.01	7.07	5.48	3.58	2.10	5.61
Silt(%)	9.90	2.80	17.59	2.95	2.20	3.74	10.47	4.98	16.46	2.50	1.47	2.00	10.24	4.52	14.20
Sand(%)	83.18	91.10	21.19	90.74	93.00	6.71	81.69	90.48	24.40	90.49	90.78	4.88	86.17	92.80	18.14
w(%)	27.97	21.84	19.93	23.11	24.20	13.06	27.35	25.75	8.56	28.17	26.94	8.04	20.79	20.48	6.43
LOI(%)	5.75	3.61	6.48	2.93	2.45	1.52	2.33	1.57	2.21	1.92	1.57	1.39	4.15	1.94	4.50
TOC(%)	1.68	1.31	1.57	1.68	0.56	1.62	0.86	0.59	0.79	0.79	0.53	0.79	1.23	1.00	1.03
TN (mg/kg)	1.60	0.56	2.42	0.51	0.33	0.57	0.58	0.31	0.67	0.32	0.23	0.30	0.76	0.85	0.17
TP (mg/kg)	966.01	377.50	1094.76	1046.71	530.00	1326.82	799.31	590.00	649.46	520.37	490.00	506.92	1267.31	1200.00	935.11
Cd (mg/kg)	7.30	1.00	12.00	0.53	0.01	1.66	11.27	7.95	12.22	5.75	4.65	5.19	20.67	13.00	19.31
Cu (mg/kg)	15.25	0.98	69.01	2.69	0.01	4.59	0.88	0.01	2.45	6.67	1.80	14.26	0.13	0.01	0.23
Pb (mg/kg)	523.61	67.50	1178.41	64.25	54.70	25.71	196.33	74.85	666.91	551.63	212.60	1353.81	49.91	43.40	28.76
Ni (mg/kg)	85.12	43.51	119.81	83.91	47.80	89.73	46.47	34.90	34.31	59.18	43.00	41.74	30.08	30.60	17.91
As (mg/kg)	24.65	21.00	17.32	31.94	24.70	23.85	42.50	23.60	75.68	170.55	38.60	410.43	50.41	46.34	28.88
Zn (mg/kg)	96.52	12.25	309.77	45.89	39.20	44.98	10.49	8.60	8.85	13.24	10.35	12.64	6.96	6.68	3.88
Cr (mg/kg)	2939.19	238.10	16148.04	179.90	118.40	252.15	585.61	320.75	654.04	1589.27	1419.05	1516.17	199.68	167.90	137.96
Hg (mg/kg)	43.23	35.10	24.59	182.25	121.30	176.48	73.76	48.20	96.13	124.97	74.65	150.76	120.00	102.20	71.96
Li (mg/kg)	0.08	0.05	0.10	11.24	0.29	15.49	0.17	0.08	0.22	0.37	0.17	0.41	0.05	0.05	0.03
Al (mg/kg)	36.19	32.70	19.20	28.63	25.70	11.30	42.80	37.95	20.44	42.87	39.40	17.37	48.23	47.93	28.11

N = number of sediment samples, Med = median, Std = Standard deviation

를 비교 검토하였다. 분석시 정규성 확보를 위해 모든 환경변수를 로그변환하였고 통계적 유의성은 $p<0.05$를 기준으로 하였다.

2.2 통계분석

퇴적물과 토양 오염도의 오염원별(산업단지, 폐금속광산) 및 시기별(강우시, 비강우시) 오염특성을 비교하여 각 토지이용을 특징짓는 오염항목과 특이적 농도변화를 갖는 오염항목을 도출하여 하천퇴적물 오염원 관리를 위한 우선순위 항목을 제시하기 위해 다변량 분산분석(Multi-variate Analysis of Variance, MANOVA)을 수행하였다.

MANOVA는 강우유무와 퇴적물 오염원이 퇴적물에 미치는 영향 특성에 대해 통계소프트웨어 SPSS를 이용하였다. MANOVA 수행 후 일원배치 ANOVA로 측정항목별 사후검증(post-hoc test)을 수행하여 각 조건별 농도 수준 차이를 비교 검토하였다. 분석시 정규성 확보를 위해 모든 환경변수를 로그변환하였고 통계적 유의성은 $p<0.05$를 기준으로 하였다.

2.3 기계학습 모델

2.3.1 단일분류 모델

퇴적물의 오염 특성을 반영하여 효율적인 오염원 식별을 목적으로 기계학습 분류모델로서 선형모델인 LDA와 비선형 모델인 SVM을 적용하여 퇴적물의 분류성능을 비교 평가하였다. LDA는 데이터 특성 추출 및 시각화에 용이하며, SVM은 더 적은 파라미터의 최적화로 일반화 능력이 우수하여 수계 오염원 분류에 많이 사용하는 모델이다(Liu & Lu, 2014, Kabilian & Selvi, 2016). 특히 본 연구에서는 고차원 특징 추출에 유용한 LDA와 데이터의 개수가 상대적으로 작은 경우 우수한 성능을 보이는 것으로 알려진 SVM을 선정하였다(Belousov et al., 2002; Yu & Yang, 2001). 분류모델에는 Table 1에 명시된 18개 분석 항목을 입력변수로 하여 오염원 분류를 수행하였으며, 모든 입력변수는 상용로그변환 후 표준화하였다. 기계학습모델은 Python을 이용하여 오염원과 시기에 따라 총 4개의 class (강우시 광산, 비강우시 광산, 강우시 산단, 비강우시 산단)로 구분하는 분류모델로 구축하였다. 각 모델은 오염원이 특정된 금강 이외 지역의 데이터(291개)를 이용하여 학습-검증 과정을 수행하고, 오염원이 복합적으로 존재하는 금강권역의 데이터(65개)를 최종 테스트용 데이터로 사용하여 오염원이 특정되지 않은 미지의 퇴적물 시료에 대하여 오염항목 분석 결과만을 이용한 오염원 분류 가능성을 평가하였다. 금강권역의 오염원 평가는 상대적으로 산업단지가 광범위하게 분포된 '혼합지역-A(Composite-A)'를 "비강우시 산단"으로, 상대적으로 폐금속광산지점이 많은 '혼합지역-B(Composite-B)'를 "비강우시 광산"으로 임의 분류하여 모델에서 출력된 확률값을 이용하여 정확도를 평가하였다. 모델의 학습-검증 과정에서 오염원이 특정된 금강 이외 지역의 데이터를 80:20으로 나누어 5-fold 교차검증을 이용하였으며(Joseph et al., 2022), 격자탐색 기법으로 하이퍼파라미터를 최적화하였다. SVM은 Gaussian 커널함수를 이용하여 구성하였으며, 하이퍼파라미터는 일반적으

로 조정되는 C, gamma를 이용하여 최적화하였다(Syarif et al., 2016).

2.3.2 다중분류 모델

비배타적 다중라벨분류기(Non-exclusive multi-label classification)를 이용하여 퇴적물의 중금속 및 유기물 농도에 따른 유역 내 잠재 오염원의 영향과 기여도 평가 모델로의 이용 가능성을 추가적으로 평가하였다(Fig. 2).

비배타적 다중라벨분류기는 SVM 알고리즘을 기반('다중분류 SVM')으로, python 라이브러리 sklearn툴의 Multioutput_classification 기능을 사용하여 구축하였다. 다중분류 SVM 입력 자료에는 ArcGIS 공간분석을 통해 지점별 영향반경 1 km 이내의 집수구역에 대한 토지이용비율(도시 및 농업), 폐광산 유무, 산업단지 유무 및 산업폐수처리장 유무 정보를 추출하여 '광산(MM)', '산단(IC)', '도시(UB)', '농업(AG)', '산업폐수처리장(IW)' 등 총 5개의 class로 구분하였다(Fig. 2 및 Table 2). 즉 각 데이터의 샘플링 지점에 대응하는 유역내에서 1km 반경 이내의 영향권역을 설정한 후 해당 권역내에 광산지역이나 산단지역, 또는 산업폐수처리장이 존재하는 경우 각각 '광산', '산단', 또는 '산업폐수처리장' 라벨을 부여하였고 해당 영향권역내 도시지역이나 농업지역이 차지하는 비율이 전체 데이터의 Q2 (50% 백분위수) 이상에 해당되는 경우 해당 토지이용의 영향이 클 것으로 판단하여 각각 '도시' 또는 '농업' 라벨을 부여하였다. 결과적으로 최종평가지역인 금강지역

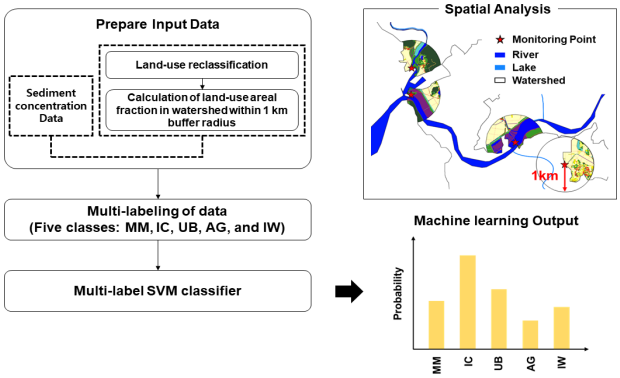

Fig. 2. Procedure of the multi-label classification modeling

65개 데이터는 '광산'에 35개, '산단'에 48개, '도시'에 43개, '농업'에 42개, '산업폐수처리장'에 46개 데이터로 할당되었다. 각각의 퇴적물 샘플에 두 개 이상의 라벨이 부여될 수 있으므로 다중분류 SVM은 각각의 데이터에 대하여 5개 잠재적 오염원에 대한 이진 분류를 수행하게 된다.

3. 연구결과

3.1 지점별 및 시기별 퇴적물 성상 비교 분석

MANOVA결과 시기별 및 지점별 퇴적물의 오염도 성상에 유의한 차이가 나타났다($F = 14.4$, $p < 0.05$). 사후분석결과 '산단'의 경우 Cd, Cu, Pb, Ni, Zn, Cr, Li 항목에서 강우시 유의하게 높았으며 TOC, TP, TN, As, Hg, Al 항목에서는 유의한 차이가 없었다. 유기물 항목에서 유의한 차이가 없었던 것은 산단지역의 경우 유기물의 오염원은 미미하기 때문으로 보이며, 중금속의 경우 As, Hg, Al를 제외하고 모든 중금속 항목에서 강우시 증가하므로 비점오염원의 작용으로 볼 수 있다(Shen et al., 2012). 반면 As와 Hg의 경우 광산지역에서는 강우발생시 증가하였으나 산단지역에서는 유의한 변화가 적으므로 이 금속들의 주요 오염원은 광산지역임을 유추할 수 있다. '광산'의 경우 강우시와 비강우시 유기물, 입도 항목을 비롯한 대부분의 항목에서 유의한 차이를 보였으나 TP, Cd, Pb, Al는 유의한 차이가 없었다. 유의한 차이를 보인 항목들 중 대부분이 강우시 농도가 증가했지만 Cu와 Zn는 예외적으로 강우시 농도가 감소하였다. Cu와 Zn는 대표적인 도시비점오염원에서 기인하는 중금속으로 알려져 있어(Rodríguez-Seijo et al., 2017) 광산지역에서 강우가 발생하는 경우 희석에 의한 원인으로 추정된다.

3.2 단일분류 기계학습모델을 이용한 퇴적물의 분류

Fig. 3a 및 3b는 LDA를 이용한 퇴적물 시료 분류 결과에 따른 요인득점 biplot을 나타낸 것이다. 학습 및 검증 데이터에 대한 퇴적물의 평균 분류 정확도는 79.5%, F1 score는 68.0% (정밀도 69.5%, 67.0%)로서 높은 분류 성능을 나타냈다. 산점도와 95% 신뢰타원을 비교한 결과 폐금속광산과 산업단지, 그리고 강우시 및 비강우시 퇴적물

Table 2. Criteria of the data classes for the multi-label classifier

Class	Criteria for class label
Metal Mines (MM)	Sampling sites with presence of mines in the corresponding watershed within 1.0 km buffer radius
Industrial complexes (IC)	Sampling sites with presence of industrial complexes in the corresponding watershed within the 1.0 km buffer radius
Urban (UB)	Sampling sites where the fraction of urban areas in the corresponding watershed within 1.0 km buffer radius; the fraction value is above Q2 of all the data
Agricultural (AG)	Sampling sites where the fraction of agricultural areas in the corresponding watershed within 1.0 km buffer radius; the fraction value is above Q2 of all the data
Industrial Wastewater Treatment Plant (IW)	Sampling sites with presence of industrial wastewater treatment plants in the corresponding watershed within 1.0 km buffer radius

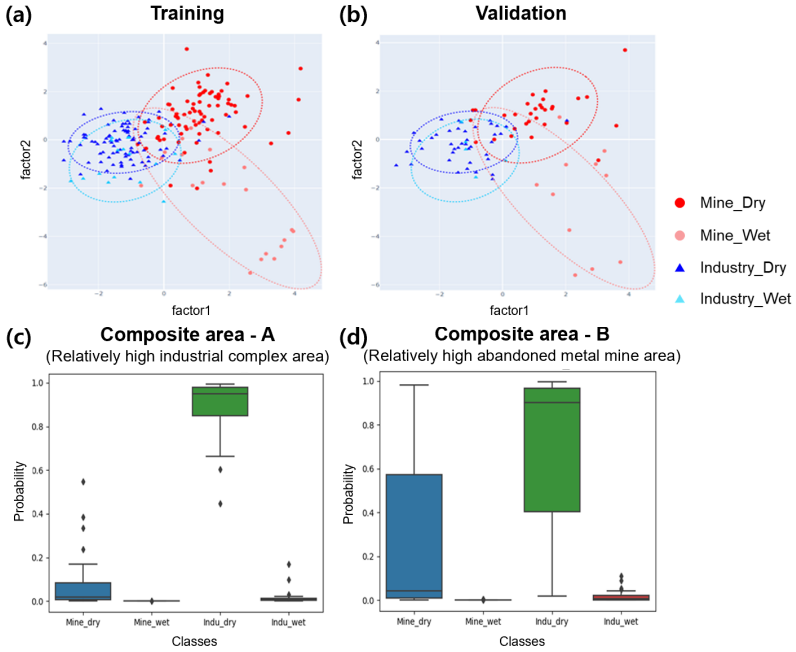

Fig. 3. (a) LDA score biplot for the training dataset; (b) LDA score biplot for the validation dataset; (c) Classification results by LDA for Composite area-A; (d) Classification results by LDA for Composite area-B

시료가 전반적으로 유의하게 분류되고 있음을 확인할 수 있었다. 다만 산업단지 인근의 퇴적물의 경우에 강우영향에 따른 퇴적물 요인특점에 성상에 차이가 적었는데 산업단지에서 발생하는 중금속 등의 오염물질의 경우 비점오염원의 형태 보다는 산업폐수처리장 등 점오염원의 형태로 주로 배출되기 때문에 강우 유무에 따른 퇴적물의 성상차이가 상대적으로 적기 때문으로 풀이된다. 반면 폐금속광산지역 인근 퇴적물은 주로 강우에 의해 광산 주변에 축적되어 있던 중금속을 포함한 오염물질이 비점오염원의 형태로 주로 배출될것으로 예상되며 LDA에서도 강우영향에 따라 퇴적

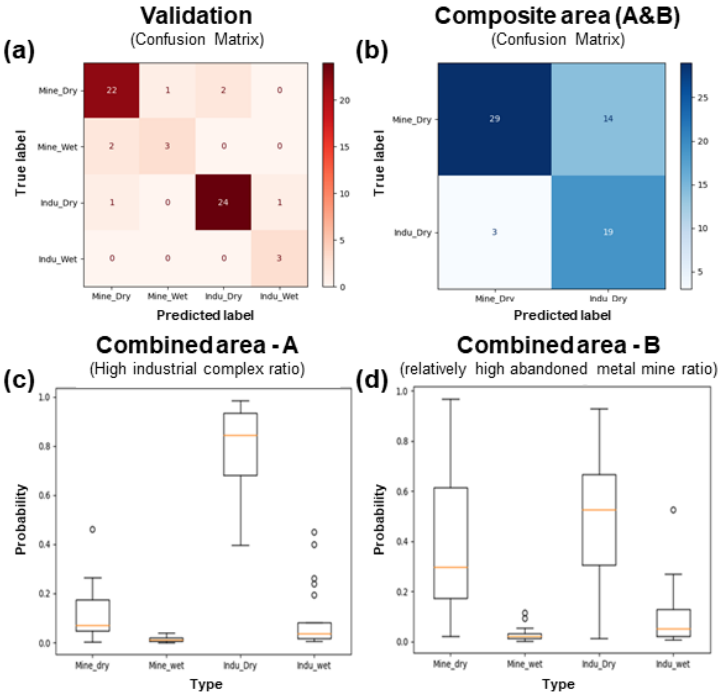

Fig. 4. (a) Confusion matrix of the validation dataset in SVM; (b) Confusion matrix of the test data (composite area) predicted by the trained SVM; (c) Classification results by SVM for Composite area-A; (d) Classification results by SVM for Composite area-B

물 성상이 확실하게 구분되었다.

학습과 검증에 사용되지 않은 '혼합지역-A(Composite-A)'와 '혼합지역-B(Composite-B)' 데이터를 이용하여 LDA의 분류 성능을 평가해 보았다. LDA를 이용한 혼합지역 데이터 분류결과 산업단지가 광범위하게 분포한 혼합지역-A 데이터의 경우 모델에서도 대부분 산업단지로 분류되었으며(Fig. 3c) 광산이 상대적으로 많은 혼합지역-B 데이터의 경우에도 산업단지로 분류될 확률이 광산지역으로 분류될 확률보다는 전반적으로 높았지만 혼합지역-A에 비해서는 광산으로 예측하는 비율이 증가하였다(Fig. 3d). 따라서 모델을 통하여 혼합지역에서도 우세한 오염원의 특성을 반영하여 분류가 적합하게 이루어짐을 확인할 수 있었다. 또한 두 유역의 모든 오염원에서 강우조건은 비강우일 것으로 예측하여 강우영향에 대한 진단을 적절히 수행할 수 있는 것으로 판단하였다.

Fig. 4는 비선형모델인 SVM 역시 강우영향 및 오염원에 대한 진단이 적절히 수행됨을 보여준다. 혼합지역을 제외한 지점을 대상으로 SVM을 학습 및 검증한 결과(Fig. 4a) 분류 정확도는 88.1%, F1 score는 79.0% (정밀도 80.6%, 재현율 77.7%)로 LDA에 비해 높은 정확도를 보였다. Fig 4b는 학습된 SVM 모델에서 전체 혼합지역(Composite area A 및 B)의 분류 결과를 나타낸 도표로, LDA와 유사하게 산업단지와 폐광산지역 모두 모델에서 산업단지로 분류된 지점이 많았다. SVM 모델에서 산업단지로 분류된 폐금속광산 지점이 많은 것은 폐금속광산으로 임의 분류한 Composite area B 지점 또한 산업단지가 공존하고 있기 때문으로 사료된다. Fig. 4c 및 4d는 혼합지역에서 퇴적물에 대한 SVM 분석결과를 나타낸 도표로, LDA와 유사하게 실제 산업단지 비율이 높은 지역(Composite area A)은 대부분 산업단지로 분류하였고(Fig. 4c), 폐금속광산이 비교적 많이 포함된 지역(Composite area B)은 광산보다는 산업단지로 분류하는 데이터의 개수가 많았다(Fig. 4d). 그러나 Composite area B에서는 LDA보다 SVM에서 폐금속광산으로 분류할 확률이 증가하므로 SVM모델에서 토지이용의 비율을 더 잘 반영할 수 있는 것으로 나타났다.

3.3 다중분류 기계학습모델을 이용한 퇴적물의 분류

다중분류 SVM은 여러 오염원의 복합적 영향을 고려하기 위해 수행하였다. 다중분류 결과 학습은 90.7%로 모델적합이 잘 됨을 확인하였으나, 테스트 데이터셋에 대한 예측정확도가 산업폐수처리장 0.5077, 광산지역의 경우 0.7231, 도시지역은 0.6615, 농업지역은 0.5692, 산단지역은 0.6308로 정확도는 대부분 학습 정확도에 비해 다소 낮게 나타났다(Fig. 5). 단일분류기와 마찬가지로 광산과 산업단지의 분류는 비교적 높은 정확도로 수행하였으나, 도시와 농업지역 등 다른 비점오염원에 대한 분류정확도는 비교적 낮았다. 이는 데이터셋의 크기와 분류모델의 한계로 인한 과적합으로 나타난 현상인 것으로 사료된다. 그러나 광산과 산업단지의 영향이 혼재하는 유역에서 단일분류 SVM과

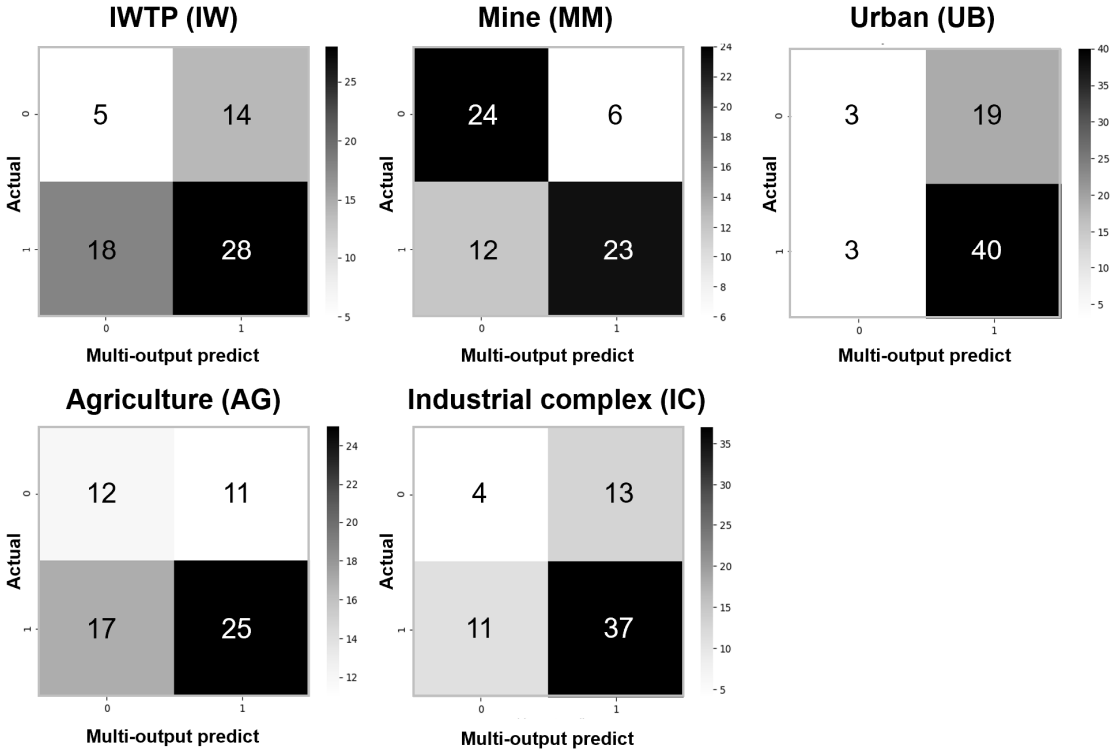

Fig. 5. Confusion matrix of the multi-label classification test result

비교하여 광산의 영향을 동시에 명시할 수 있음을 확인하였다. 따라서 향후 전국 측정망자료 등 보다 많은 자료를 활용하고, 토지이용에 따른 복합적 영향이 명확히 존재하는 데이터셋을 추가할 수 있을 경우 다중분류 SVM의 정확도를 향상시킬 수 있을 것으로 예상된다.

4. 결론

본 연구에서는 퇴적물 오염물질의 특성 차이에 기반한 오염원 식별 방법으로 기계학습모델의 적용성을 평가하였다. 기계학습모델로 선형판별분석(LDA)과 서포트벡터머신(SVM)을 이용하였으며, 폐금속광산지역과 산업단지, 강우영향별 퇴적물 시료의 식별이 가능함을 보였다. 결과적으로 두 모델 모두 강우영향 판별시에는 높은 정확도를 보였고 LDA에 비해 비선형 모델인 SVM에서 실제 유역내 오염원(광산 및 산단)의 영향을 더 잘 반영하는 결과를 도출하는 것을 확인하였다. 다중분류 SVM은 개별 오염원에 대한 정확도는 데이터 수의 한계로 분류정확도가 낮았으나, 복합지역에서 다수 오염원의 영향 예측이 가능함을 확인하였다. 향후 보다 많은 데이터를 활용할 경우 도시, 농업, 교통 등 보다 상세한 오염원의 식별이 가능할 것으로 판단된다.

사 사

본 연구는 환경부의 재원으로 국립환경과학원의 지원을 받아 수행하였습니다(NIER-2020-04-02-126).

References

Ali, H., Khan, E., Ilahi, I., (2019). Environmental chemistry and ecotoxicology of hazardous heavy metals: environmental persistence, toxicity, and bioaccumulation. J. Chem. Article ID 6730305.

Belousov, A. I., Verzakov, S. A., & Von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: support vector machines. Chemometrics and intelligent laboratory systems, 64(1), 15-25.

Cha, Y., Shin, J., & Kim, Y. (2020). Data-driven modeling of freshwater aquatic systems: Status and prospects. Journal of Korean Society on Water Environment, 36(6), 611-620.

Chueh, Y. Y., Fan, C., & Huang, Y. Z. (2021). Copper concentration simulation in a river by SWAT-WASP integration and its application to assessing the impacts of climate change and various remediation strategies. Journal of environmental management, 279, 111613.

Chung, S., Kim, S., Park, H., & Seo, D. (2020). Future Development Direction of Water Quality Modeling Technology to Support National Water Environment Management Policy. Journal of Korean Society on Water Environment, 36(6), 621-635.

dos Santos Simoes, F., Moreira, A. B., Bisinoti, M. C., Gimenez, S. M. N., & Yabe, M. J. S. (2008). Water quality index as a simple indicator of aquaculture effects on aquatic bodies. Ecological indicators, 8(5), 476-484.

Guan, Q., Wang, F., Xu, C., Pan, N., Lin, J., Zhao, R., ... & Luo, H. (2018). Source apportionment of heavy metals in agricultural soil based on PMF: A case study in Hexi Corridor, northwest China. Chemosphere, 193, 189-197.

Jiang, Y., Gui, H., Chen, C., Wang, C., Zhang, Y., Huang, Y., ... & Qiu, H. (2021). The Characteristics and Source Analysis of Heavy Metals in the Sediment of Water Area of Urban Scenic: A Case Study of the Delta Park in Suzhou City, Anhui Province, China. Polish Journal of Environmental Studies, 30(3).

Jung, M. C., & Jung, M. Y. (2006). Evaluation and management method of environmental contamination from abandoned metal mines in Korea. Journal of the Korean Society of Mineral and Energy Resources Engineers, 43(5), 383-394.

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(4), 531-538.

Kabilan, N., & Selvi, M. S. (2016, April). Surveillance and steering of irrigation system in cloud using Wireless Sensor Network and Wi-Fi module. In 2016 International Conference on Recent Trends in Information Technology (ICRTIT) (pp. 1-5). IEEE.

Kang, J.H., Lee, Y.G., Lee, K.Y., Cha, S.M., Cho, K.H., Lee, Y.S., Ki, S.J., Yoon, I.H., Kim, K.W., Kim, J.H., (2009). Factor affecting metal exchange between sediment and water in an estuarine reservoir: a spatial and seasonal observation. J. Environ. Monit. 11, 2058-2067.

Liber, Y., Mourier, B., Marchand, P., Bichon, E., Perrodin, Y., & Bedell, J. P. (2019). Past and recent state of sediment contamination by persistent organic pollutants (POPs) in the Rhône River: Overview of ecotoxicological implications. Science of the Total Environment, 646, 1037-1046.

Liu, M., & Lu, J. (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?. Environmental Science and Pollution Research, 21, 11036-11053.

Navarro, M. C., Pérez-Sirvent, C., Martínez-Sánchez, M. J., Vidal, J., Tovar, P. J., & Bech, J. (2008). Abandoned mine sites as a source of contamination by heavy metals:

a case study in a semi-arid zone. Journal of Geochemical exploration, 96(2-3), 183-193.

Rodríguez-Seijo, A., Andrade, M. L., & Vega, F. A. (2017). Origin and spatial distribution of metals in urban soils. Journal of soils and sediments, 17, 1514-1526.

Shen, Z., Chen, L., Liao, Q., Liu, R., & Hong, Q. (2012). Impact of spatial rainfall variability on hydrology and nonpoint source pollution modeling. Journal of Hydrology, 472, 205-215.

Sojka, M., & Jasku ł a, J. (2022). Heavy metals in river sediments: contamination, toxicity, and source identification—a case study from Poland. International Journal of Environmental Research and Public Health, 19(17), 10502.

Srinivasan, R., Zhang, X., & Arnold, J. (2010). SWAT ungauged: hydrological budget and crop yield predictions in the Upper Mississippi River Basin. Transactions of the ASABE, 53(5), 1533-1546.

Stoichev, T., Coelho, J. P., De Diego, A., Valenzuela, M. G. L., Pereira, M. E., de Chanvalon, A. T., & Amouroux, D. (2020). Multiple regression analysis to assess the contamination with metals and metalloids in surface sediments (Aveiro Lagoon, Portugal). Marine Pollution Bulletin, 159, 111470.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA (Telecommunication Computing Electronics and Control), 14(4), 1502-1509.

Wu, H., Xu, C., Wang, J., Xiang, Y., Ren, M., Qie, H., ... & Lin, A. (2021). Health risk assessment based on source identification of heavy metals: A case study of Beiyun River, China. Ecotoxicology and Environmental Safety, 213, 112046.

Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data—with application to face recognition. Pattern recognition, 34(10), 2067-2070.

Zhang, X., Mei, X., (2015). Effects of benthic algae on release of soluble reactive phosphorus from sediments: a radioisotope tracing study. Water Sci. Eng. 8 (2), 127-131.

Zhang, Z., Huang, J., Duan, S., Huang, Y., Cai, J., & Bian, J. (2022). Use of interpretable machine learning to identify the factors influencing the nonlinear linkage between land use and river water quality in the Chesapeake Bay watershed. Ecological Indicators, 140, 108977.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health, 1, 107-116