

데이터과학연구

Journal of Data Science

제15권



2026년 9월

데이터과학융합연구소

Data Science Convergence Research Center

데이터과학융합연구소 임원 및 연구부

- 소 장 : 이주영 교수 (응용통계학과, jooylee@cau.ac.kr)
- 간 사 : 변준영 교수 (응용통계학과, junyoungb@cau.ac.kr)
- 연구부 : 김원국 교수 (응용통계학과)
김정하 교수 (의과대학)
문광수 교수 (심리학과)
박재현 교수 (컴퓨터공학과)
변준영 교수 (응용통계학과)
성병찬 교수 (응용통계학과)
신우영 교수 (의과대학)
윤영철 교수 (의과대학)
이세영 교수 (의과대학)
이재우 교수 (산업보안학과)
이재현 교수 (응용통계학과)
이정규 교수 (의과대학)
이주영 교수 (응용통계학과)
임예지 교수 (응용통계학과)
한수현 교수 (의과대학)
황범석 교수 (응용통계학과)

데이터과학연구 편집위원회

- 편집위원장 : 이주영 교수 (응용통계학과)
편집 위원 : 변준영 교수 (응용통계학과)

Contents

- ◆ A Comparative Study of One-stage and Two-stage Vision-Language Pipelines for Ukrainian Handwritten Document Recognition

김병찬, 김유진, 이승규, 이채승 / 1

- ◆ 흉부 X-ray 이상 소견 탐지를 위한 RT-DETR 기반 2-Stage 객체 탐지 모델 연구

송채린, 김태환, 이송민, 함민아 / 20

- ◆ Multi-label Bioacoustic Classification Using Perch Embeddings and Asymmetric Ensemble for BirdCLEF+ 2026

김시은, 류정찬, 백승준, 이나현 / 34

- ◆ Wildfire Risk Assessment near Power Facilities Based on Meteorological Data and Spatial Information

김도연, 송윤섭, 이은승, 이준구, 이진석 / 48

A Comparative Study of One-stage and Two-stage Vision-Language Pipelines for Ukrainian Handwritten Document Recognition

김병찬¹⁾, 김유진²⁾, 이승규³⁾, 이채승³⁾

Abstract

This study proposes a practical AI solution for Ukrainian handwritten text recognition to accelerate the country's digital transformation and support real-world government workflows. Based on the RUKOPYS Kaggle Challenge, we comparatively analyzed two deep learning pipelines designed to extract and structure text from noisy document images while strictly minimizing the Character Error Rate. Experimental results demonstrated that while a one-stage, end-to-end Vision-Language Model approach (score: 0.780) showed promising structural simplicity and future potential, a two-stage pipeline separating region detection and text recognition achieved higher accuracy (score: 0.816) and better deployment readiness. Ultimately, by robustly processing diverse handwriting styles, the validated architecture is expected to streamline public document management and significantly expedite the rollout of digital government services.

Keywords: Handwritten Text Recognition, Document

1) All authors have equal contribution, Department of Smart Cities, Chung-Ang University, Seoul 06974, Korea

2) Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea

3) Department of Civil Engineering, Chung-Ang University, Seoul 06974, Korea

1. 서론

최근 공공 행정, 교육, 기록 관리 분야에서는 수기 문서를 디지털 데이터로 변환하는 기술의 중요성이 커지고 있다. 특히 기록물, 신청서, 서명 문서 등은 단순 이미지로 보관될 경우 검색 및 자동 처리에 활용하기 어렵다. 따라서 문서 이미지에서 텍스트 영역을 찾고, 유형을 분류하며, 내용을 정확히 전사하는 기술은 디지털 전환의 핵심이다.

본 연구는 Kaggle에서 진행된 RUKOPYS Handwritten Text Recognition Challenge를 대상으로 한다. 이는 우크라이나어 필기체 문서 이미지에서 각 영역의 위치(bounding box), 종류(type), 전사 텍스트를 포함하는 JSON 형태의 구조화 데이터를 생성하는 과제이다. 즉, 단순한 OCR을 넘어 문서 내 여러 영역을 탐지·분류하고 읽기 순서(reading order)에 따라 텍스트를 재구성하는 복잡한 문서 이해 문제이다.

이 문제는 일반 인쇄체 OCR보다 훨씬 어렵다. 첫째, 작성자마다 필체와 잡음 수준이 크게 다르고, 둘째, 우크라이나어 키릴 문자는 기존 라틴 문자 OCR 도구로 처리하기 까다롭다. 셋째, 한 페이지 내에 손글씨, 표, 도장 등 다양한 요소가 혼재되어 문서 구조 파악이 필수적이다. 넷째, 평가 지표에서 page-level CER 비중이 커 개별 OCR뿐 아니라 reading order와 중복 제거가 중요한 성능 요소가 된다.

본 연구에서는 이를 해결하기 위해 두 가지 딥러닝 접근법을 구현하고 비교하였다. 첫 번째는 Qwen2.5-VL 기반의 One-stage End-to-End 방식으로, 전체 페이지 이미지를 VLM에 입력해 결과물(JSON)을 직접 생성한다. 두 번째는 Two-stage 방식으로, YOLO 기반 detector가 영역을 선행 탐지하고 Qwen3-VL 기반 recognizer가 crop된 영역의 텍스트를 인식한다. 두 방식 모두 Qwen 계열 VLM을 사용하지만, 검출 과정을 단일 모델에 통합하는가 혹은 전용 detector로 분리하는가에서 구조적 차이를 갖는다.

본 연구는 RUKOPYS 데이터 특성 및 평가 지표를 분석하고, 두 접근법의 구현 및 성능 향상 과정을 정리하는 것을 목적으로 한다. 나아가 두 방식의 장단점과 병목 요인을 비교하고, 실제 대회 점수를 기준으로 어떤 설계가 더 효과적인지 분석하고자 한다.

2. 데이터 특성 및 평가지표

2.1. RUKOPYS 데이터셋

RUKOPYS 데이터셋은 우크라이나어 필기체 문서를 대상으로 영역 탐지, 영역 타입 분류, 텍스트 전사를 수행하도록 구성된 데이터셋이다. 데이터는 크게 Train(Gold), Silver, Test로 구분된다.

Gold 데이터는 사람이 검수한 annotation이므로 신뢰도가 높다. 반면 Silver 데이터는 수량이 많지만 자동 생성된 annotation이므로 노이즈가 포함되어 있다. 따라서 Silver 데이터를 단순히 추가하는 것이 항상 성능 향상으로 이어지는 않았다. 실제 실험에서도 Silver 데이터 3,000장을 추가한 detector fine-tuning은 기본 Gold 기반 모델보다 낮은 성능을 보였다.

각 문서 이미지는 여러 region으로 구성되며, region schema는 bbox, type, language, legibility, text 등을 포함한다. region type은 handwritten, formula, annotation, printed, table, image, graph 등으로 구성된다. 이 중 image와 graph는 OCR 대상이 아니므로 scorable region에서 제외된다.

[표 1] RUKOPYS 데이터 분할

분할	이미지 수	라벨 특성	용도
Train(Gold)	1,330	사람이 검수한 실제 annotation	학습 및 검증
Silver	8,207	AI가 자동 생성한 annotation	학습 보강
Test	385	정답 비공개	대회 제출

2.2 데이터 출처 및 도메인 다양성

RUKOPYS 데이터셋은 1919년부터 2025년까지 100년 이상을 아우르는 다양한 출처의 문서로 구성되어 있다.

이와 같은 출처 다양성은 모델 학습을 어렵게 만든다. 예를 들어 dictation 문서는 비교적 깔끔한 손글씨가 많아 인식이 쉽지만, school 문서는 글씨가 뾰뾰하고 낙서나 짧은 조각이 많아 탐지와 인식 모두 어렵다. university 문서는 수식과 표가 섞여 있어 일반 OCR과는 다른 처리 전략이 필요하다. archive 문서는 오래된 스캔 문서이기 때문에 인쇄체, 도장, 주석이 함께 나타난다. 따라서 단일 도메인에 최적화된 모델은 전체 데이터셋에서 안정적인 성능을 내기 어렵다.

[표 2] 데이터 출처별 특성

출처	기간	특징
국가 받아쓰기 대회 사진	2020~2025	비교적 깔끔한 손글씨
국가 기록원 스캔 문서	1919~1935	인쇄체, 주석, 도장, 고문서 잡음 혼재
대학교 학생 시험지	최근	텍스트, 수식, 화학식 혼재
초중고 숙제 사진	최근	뽀뽀한 필기, 낙서, 짧은 조각 다수

2.3 평가지표

대회 최종 점수는 Detection F1, Class Accuracy, Region CER, Page CER의 가중합으로 계산된다. 점수 공식은 다음과 같다.

$$Score = 0.15 \times DetectionF1 + 0.05 \times ClassAcc + 0.30 \times RegionCER + 0.50 \times (1 - PageCER) \quad (\text{식1})$$

각 지표의 의미는 다음과 같다.

[표 3] 평가지표 구성

지표	가중치	정의	평가 대상
Page CER	50%	읽기 순서로 region text를 연결한 page-level 문자 오류율	페이지 전체 전사 및 reading order
Region CER	30%	매칭된 scorable region별 text 오류율	region별 OCR
Detection F1	15%	IoU ≥ 0.5 기준 bbox 매칭 후 계산한 F1	영역 탐지
ClassAcc	5%	예측 type과 정답 type의 일치율	영역 타입 분류

이 지표 구조에서 중요한 점은 텍스트 인식 관련 항목이 전체의 80%를 차지한다는 것이다. 그러나 detection의 중요성이 낮다고 볼 수는 없다. 특정 bbox를 탐지하지 못하면 해당 영역은 OCR 입력으로 전달되지 않고, 이는 Region CER과 Page CER에도 직접적인 악영향을 준다. 따라서 이 대회에서는 잘 읽는 모델

뿐만 아니라 읽어야 할 영역을 빠뜨리지 않는 detector와 읽기 순서를 올바르게 재구성하는 후처리가 모두 중요하다.

3. 연구 방법

3.1 전체 접근법

본 연구에서는 동일한 입력과 동일한 목표를 갖는 두 가지 접근법을 비교하였다. 첫 번째는 One-stage 방식이며, 두 번째는 Two-stage 방식이다.

One-stage 방식은 페이지 이미지를 Qwen2.5-VL에 직접 입력하고, 모델이 bbox, type, text를 포함한 JSON을 한 번에 생성하는 구조이다. 이 방식은 파이프라인이 단순하고, page-level 출력을 직접 최적화할 수 있다는 장점이 있다. 그러나 region 수가 많아지면 출력 JSON이 길어지고, 이로 인해 형식 붕괴나 출력 누락이 발생할 수 있다. 또한 VLM이 좌표를 직접 생성해야 하므로 정밀한 bbox 예측이 어렵다.

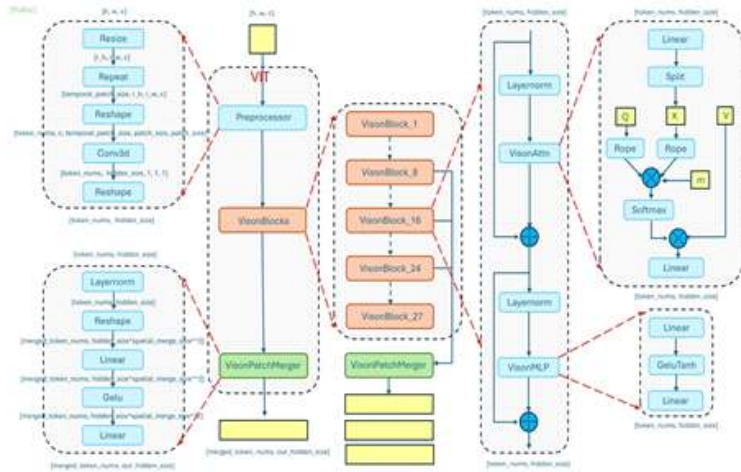
Two-stage 방식은 먼저 YOLO 기반 detector가 문서 내 region을 찾고, 이후 Qwen 기반 recognizer가 crop된 영역을 읽는 구조이다. 이 방식은 detection과 recognition을 분리하여 각 단계의 오류를 독립적으로 진단할 수 있다는 장점이 있다. 또한 전용 detector를 사용하기 때문에 bbox 정확도가 상대적으로 높다. 반면 앞 단계의 오류가 뒤 단계로 전파되는 문제가 있으며, 읽기 순서 조립과 중복 제거를 위한 후처리가 필요하다.

3.2 One-stage End-to-End VLM

One-stage에서는 복잡한 구조의 우크라이나어 필기체 문서에서 region list를 직접 생성하고, 각 영역의 텍스트를 전사하는 end-to-end 작업을 수행하기 위해 Qwen2.5-VL을 기본 VLM으로 채택하였다. Qwen2.5-VL은 이미지와 텍스트를 함께 처리하는 vision-language model로, 문서 이미지 내의 객체 위치 추정, 문서구조 이해, 표·양식·레이아웃 해석, 구조화 정보 추출에 강점을 갖는 것으로 보고되었다 (Bai 등, 2025). 이러한 특성은 RUKOPYS 과제처럼 단순한 OCR 결과가 아니라 bbox, type, text를 포함하는 region-level JSON 출력을 요구하는 문제와 잘 부합한다. 즉, 본 과제에서는 글자를 읽는 능력뿐 아니라 문서 내 영역의 위치와 유형을 함께 판단해야 하므로, 시각적 문서 이해 능력과 구조화 출력 생성 능력을 동시에 갖춘 VLM이 적합하다고 판단하였다.

Qwen 계열 VLM의 주요 장점 중 하나는 다양한 해상도와 비율의 이미지를 처리하기 위한 dynamic resolution 기반 시각 인코딩 구조이다. 이는 문서 이미지를 단일 고정 해상도로만 단순 정규화하는 방식과 달리, 입력 이미지의 크기와 비율에 따라 시각 정보를 유연하게 token화하는 데 초점을 둔다. 특히 Qwen2-VL 계열에서 제안된 Naive Dynamic Resolution은 서로 다른 해상도의 이미지를 서로 다른 수의 visual token으로 변환할 수 있도록 설계되었으며, M-RoPE는 텍스트, 이미지, 비디오의 위치 정보를 통합적으로 반영하기 위한 positional encoding 방식이다 (Wang 등, 2024). 이러한 구조는 RUKOPYS 과제에서 줄 단위 bbox를 예측하고, region 간의 공간적 관계와 읽기 순서를 고려해야 하는 요구사항과 연결된다.

[그림 2] Qwen-VL Vision Module Architecture



[그림 1]은 본 연구에서 One-stage 접근법의 기반 모델로 사용한 Qwen-VL 계열의 Vision Module 구조를 나타낸다. 입력 이미지는 먼저 pre-processor를 거치며, 이 과정에서 resize, reshape, convolution 기반 patch embedding 등의 처리를 통해 visual token으로 변환된다. 이후 visual token은 여러 개의 VisionBlock을 통과한다. 각 VisionBlock은 Layer Normalization, Vision Attention, Vision MLP 및 residual connection으로 구성되며, attention 내부에서는 Q, K, V projection과 RoPE 기반 위치 정보가 함께 사용된다. 이러한 구조는 문서 이미지의 지역적 시각 특징과 공간적 배치를 모델이 반영할 수 있도록

한다.

또한 VisionBlock을 통과한 visual token은 VisionPatchMerger를 통해 병합된다. 이는 고해상도 문서 이미지에서 생성되는 많은 수의 visual token을 보다 효율적인 표현으로 압축하는 역할을 한다. 따라서 [그림 1]의 구조는 다양한 해상도와 비율을 가진 문서 이미지를 처리하면서도, 문서 내 글자 영역의 위치, 줄 단위 구조, 시각적 세부 특징을 언어 모델에 전달할 수 있는 기반을 제공한다. 본 연구의 One-stage 방식은 이러한 Vision Module을 바탕으로 별도의 detector 없이 페이지 이미지로부터 bbox, type, text를 포함하는 region-level JSON을 직접 생성하도록 설계되었다.

구체적으로 본 연구의 One-stage 방식은 OCR을 절차적 인식 문제가 아니라 VLM 기반의 구조화 생성 문제로 정의하였다. 기존 OCR은 일반적으로 영역 검출, crop, 문자 인식, 후처리의 절차적 단계로 구성되는 경우가 많다. 반면 VLM은 페이지 이미지를 입력받아 자연어 또는 JSON과 같은 구조화된 출력을 직접 생성할 수 있다. 따라서 페이지 전체 이미지를 입력하고, prompt를 통해 [{bbox, type, text}, ...] 형식의 region list를 생성하도록 지시하였다.

초기에는 zero-shot 방식으로 Qwen2.5-VL-7B에 전체 페이지를 입력하고 JSON을 출력하도록 하였다. 그러나 이 방식은 두 가지 문제를 보였다. 첫째, region이 많은 페이지에서 출력 JSON이 길어지면서 중간에 출력이 끊기거나 형식이 깨지는 문제가 발생하였다. 둘째, 전체 페이지를 한 번에 처리하는 과정에서 줄 단위 bbox가 부정확하게 생성되거나 지나치게 얇게 잡히는 한계가 나타났다. 이에 따라 prompt-only 방식의 LB 점수는 0.274에 그쳤다. 이후 페이지를 band 단위로 나누는 chunk 전략을 시도하였다. 출력 길이를 줄이면 JSON 형식이 안정될 것으로 기대했지만, chunk 경계에서 텍스트가 잘리거나 region이 중복되는 문제가 발생하여 점수는 0.264로 하락하였다. bbox 후처리를 통해 점수는 0.284까지 개선되었지만, prompt와 후처리만으로는 근본적인 한계가 있었다.

결정적인 성능 향상은 Curriculum LoRA SFT에서 발생하였다. Curriculum Learning은 학습 데이터를 쉬운 예제에서 어려운 예제로 점진적으로 제시함으로써 모델의 학습을 안정화하는 방법으로 알려져 있다 (Bengio 등, 2009). 본 연구에서는 모델이 쉬운 예제부터 어려운 예제까지 단계적으로 학습하도록 구성하였다. Stage 1에서는 짧고 단순한 handwritten 예제를 사용하였다. Stage 2에서는 다양한 region type이 포함된 중간 난이도 예제를 사용하였다. Stage 3에서는 test와 유사한 복잡하고 다타입 region chunk를 사용하였다. 난이도는 region 수, 출력 길이, 수식 또는 표와 같은 특수 타입의 유무를 기준으로 자동 판정하였다. 이 방식은 모델이 먼저 안정적인 출력 형식을 학습하고, 이후 복잡한 페이지 구조로 확장하도록 돕는다.

또한 LoRA와 QLoRA를 활용하여 큰 VLM을 효율적으로 파인튜닝하였다. LoRA는 사전학습 모델의 기존 가중치를 고정된 상태에서 트랜스포머 레이어에 학습 가능한 저랭크 행렬을 삽입하여 다운스트림 태스크에 적응시키는 parameter-efficient 파인튜닝 방법이다 (Hu 등, 2022). QLoRA는 4bit로 양자화된 사전학습 모델을 고정된 채 LoRA adapter를 통해 gradient를 역전파하는 방식으로, 대형 언어 모델의 fine-tuning에 필요한 GPU 메모리 사용량을 크게 줄일 수 있다 (Detrmers 등, 2023). 이를 통해 단일 GPU 환경에서도 대형 VLM을 미세조정할 수 있었다.

마지막으로 OCR Teacher를 활용한 지식 증류를 적용하였다. 지식 증류는 큰 teacher model의 출력 또는 soft target을 활용하여 student model을 학습시키는 방법으로, 모델 압축 및 성능 향상에 널리 활용된다 (Hinton 등, 2015). 본 연구에서는 cyrillic-trocr 기반 줄 단위 손글씨 OCR 모델을 teacher로 사용하고, Silver crop에 pseudo-label을 생성하였다. TrOCR는 Transformer 기반 encoder-decoder 구조를 OCR에 적용한 대표적인 방법으로, 사전학습된 vision-language 구조를 문자 인식에 활용한다 (Li 등, 2023). 이후 이 pseudo-label을 Qwen adapter에 추가로 학습시켜 우크라이나어 필기체 전사 능력을 강화하였다. 이 과정에는 4,324개의 examples가 사용되었다. 최종적으로 chunk overlap에서 발생하는 중복을 제거하여 One-stage 방식의 최고 LB 0.780을 달성하였다.

정리하면, One-stage 방식은 Qwen2.5-VL의 사전학습된 문서 이해 능력을 바탕으로 bbox, type, text를 한 번에 생성하는 통합 구조를 지향하였다. 그러나 zero-shot VLM만으로는 긴 JSON 출력과 정밀 bbox 예측을 안정적으로 수행하기 어려웠으며, RUKOPYS 데이터의 출력 형식과 필기체 전사 특성을 반영한 task-specific fine-tuning이 필요하였다. 본 연구에서는 Curriculum LoRA SFT와 OCR Teacher 기반 지식 증류를 통해 이러한 한계를 보완하였고, 그 결과 prompt-only 방식 대비 큰 폭의 성능 향상을 확인하였다.

3.3 Two-stage Stage 1: Region Detection

Two-stage 방식의 첫 번째 단계는 문서 이미지에서 region bbox와 type을 탐지하는 것이다. 본 연구에서는 YOLO, RetinaNet, CRAFT+ResNet, DBNet+ResNet 등 여러 detector를 비교하였다. zero-shot detector는 거의 성능을 내지 못했으며, RUKOPYS Gold 데이터를 이용한 fine-tuning이 필수적이었다. 모델 비교 결과 YOLO11 fine-tuning이 가장 좋은 성능을 보였다. 특히 CRAFT와 DBNet은 text detection에 특화된 모델이지만, RUKOPYS에서는

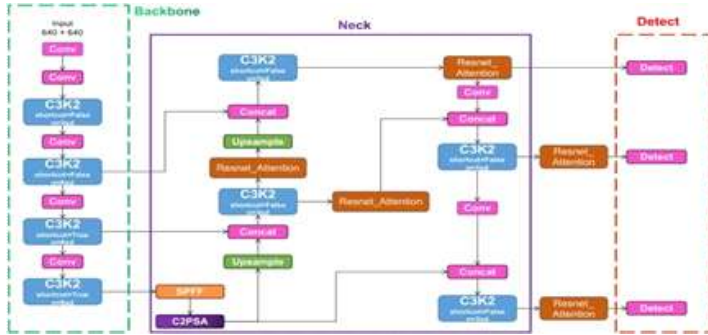
bbox 탐지뿐 아니라 class 예측도 필요했기 때문에 ResNet 기반 분류기를 추가로 학습해야 했다. 반면 YOLO는 객체 탐지와 class prediction을 동시에 수행할 수 있어 대회 구조에 더 적합하였다.

[표 4] Detector 모델 비교

모델	설명	IoU	mAP50	Recall	Precision	F1
YOLO11	zero-shot	0.003	0.000	0.000	0.000	0.000
RetinaNet	zero-shot	0.006	0.000	0.000	0.000	0.000
CRAFT+ ResNet	Gold data 기반 튜닝	0.358	0.011	0.176	0.044	0.071
DBNet+ ResNet	Gold data 기반 튜닝	0.494	0.069	0.214	0.312	0.253
YOLO11 fine-tuning	Gold data 기반 파인튜닝	0.746	0.441	0.736	0.444	0.554
RetinaNet fine-tuning	Gold data 기반 파인튜닝	0.614	0.353	0.603	0.299	0.401

최종적으로 본 연구에서는 우크라이나어 필기체 문서에서 텍스트 영역을 신속하고 정확하게 분할하기 위해, 연산 효율과 정밀도가 뛰어난 YOLO11 객체 탐지 모델을 도입하였다. YOLO11은 C3k2 블록을 통해 적은 파라미터로 깊은 특징을 추출하고, C2PSA 모듈의 공간적 어텐션을 활용해 배경 노이즈로 인한 오탐지를 효과적으로 억제한다. 특히 3가지 해상도의 다중 스케일 예측을 통해 미세한 구두점부터 긴 홀림체 문장까지 다양한 크기의 텍스트 영역을 누락 없이 포착할 수 있으며, 이러한 구조적 우수성은 [표 4]의 실험 결과를 통해 확인되었다.

[그림 3] YOLO v11 model Architecture



클래스별 객체 검출 성능을 분석한 결과, 'handwritten', 'table', 'formula' 클래스는 상대적으로 높은 mAP50 수치를 기록한 반면, 'printed', 'image', 'graph' 클래스의 성능은 다소 저조한 것으로 나타났다. 이러한 결과는 학습 데이터 내의 클래스 불균형에 기인한 것으로 분석된다. 즉, 전체 데이터의 절대다수를 차지하는 'handwritten' 타입과 달리, 희소 클래스들의 경우 표본 수가 상대적으로 부족하여 모델이 유의미한 특징을 충분히 학습하는 데 한계가 존재했다.

다음으로, 모델의 성능 최적화를 위해 다양한 하이퍼파라미터 튜닝 및 학습 전략을 실험하였다. 구체적으로 Silver 데이터 3,000장 추가 학습, DocLayNet Pre-trained weight 적용, 입력 해상도 확대, 그리고 짧은 텍스트 조각이 포함된 페이지의 Oversampling 등을 시도하였으나, 모두 Gold 데이터만을 활용해 학습 시킨 YOLO11m Baseline모델의 성능을 상회하지 못했다. 이에 따라 본 연구에서는 탐지 모델로 YOLO11m 기본 모델을 최종 채택하였으며, Inference단계의 주요 파라미터는 Confidence threshold 0.5, Image size 1280, NMS IoU 0.70으로 고정하였다.

탐지 정확도를 극대화하기 위한 Inference Post-processing 단계에서는 중복 제거(Dedup, IoU 0.55) 및 Small-conf 2-pass 전략을 도입하였다. Dedup 기법은 과도하게 겹치는 바운딩 박스를 제거함으로써 중복 인식으로 인한 페이지 단위의 CER 저하를 방지하는 데 유의미한 효과를 보였다. 또한, Small-conf 2-pass 전략을 통해 1차 검출 단계에서 임계치 미달로 배제된 낮은 신뢰도의 박스들을 재검토함으로써, 누락되기 쉬운 소형 텍스트 영역을 제한적으로 복원할 수 있었다.

추가적인 실험으로서 Custom NMS와 Source-aware policy의 적용 가능성도 검토하였다. Custom NMS는 중형비가 긴 박스에서 미세한 위치 오차만으로도

IoU가 급감하는 현상을 완화하기 위해 x축과 y축의 겹침 비율을 독립적으로 산출하는 기법이다. 아울러 Source-aware policy는 문서의 출처(school, university, archive, dictation 등)에 따라 Confidence 및 NMS IoU 임계치를 차등 적용하는 방식이다. 그러나 다양한 기법을 대조한 결과, 궁극적으로는 기본 YOLO11m 모델에 표준 추론 파라미터를 적용하고, 필수적인 수준의 중복 제거 및 저신뢰도 박스 복원 기법만을 결합한 구성이 성능과 효율 측면에서 가장 실효성 있는 조합으로 판명되었다.

3.4 Two-stage Stage 2: Text Recognition

Two-stage 파이프라인의 두 번째 단계는 Detector가 추출한 Crop의 텍스트를 인식 및 생성하는 과정이다. 초기 연구에서는 ViT(Vision Transformer) 인코더와 트랜스포머 디코더 기반의 인식기를 활용하였으나, 필기체, 수식, 표 등 복잡한 문서 요소를 처리하는 데 한계가 있었다. 이를 보완하기 위해 Fallback 조건을 설계하여 적용하였으나, 파이프라인 구조가 비효율적으로 복잡해지고 전체 인식 품질은 오히려 하락하는 부작용이 발생하였다. 그 결과, 복잡한 분기 처리 대신 텍스트 영역의 이미지를 일괄적으로 Qwen3-VL에 입력하여 전사하고 비텍스트 영역은 빈 문자열로 처리하는 단일화된 아키텍처로 전환하였다. 이 방식은 기존 구조에 비해 단순하면서도 훨씬 안정적인 인식 성능을 보였다.

그러나 생성형 VLM을 형태 보존이 중요한 OCR 작업에 도입할 경우, 모델의 생성적 특성으로 인해 치명적인 단점이 발생한다. 원본에 없는 잘못된 글자를 문맥상 추론하여 채워 넣거나, 언어 번역, 맞춤법 자동 교정, 수식 포맷의 임의 변환 등 Hallucination이 일어날 수 있으며, 이는 결과적으로 CER율을 심각하게 악화시키는 주요 원인으로 작용한다.

이러한 문제를 제어하기 위해 '보이는 그대로의 전사'를 강제하는 Strict OCR prompt를 설계하였다. 텍스트의 임의 완성, 번역, 문자 체계 변환, 부가 설명 생성 등을 원천적으로 금지함으로써, VLM의 생성 능력을 억제하고 OCR 본연의 목적에 부합하도록 유도하였다. 또한, 프롬프트 제어에도 불구하고 VLM 특유의 대화형 접두어나 구조화된 Wrapper) 텍스트가 섞여 나오는 현상을 해결하기 위해, 순수 전사 데이터만을 추출하는 Output Cleanup postprocessing를 도입하였다. 이는 무거운 모델을 재학습하는 비용을 들이지 않고도 CER 하락 요인을 안전하게 제거할 수 있는 실효적인 방안으로 입증되었다.

마지막으로 인식을 극대화를 위해 입력 이미지의 Crop preprocessing 기법을 다각도로 실험하였다. 바운딩 박스의 Margin 조정 및 대비/선명도 보정 등을 시도 하였으나, 여백을 타이트하게 설정하면 글자의 외곽이 손실되고 반대로 넓게

설정하면 인접한 줄의 텍스트가 노이즈로 유입되는 단점이 존재했다. 또한, 과도한 이미지 보정은 원본 필기체의 고유한 획 형태를 왜곡할 위험이 컸다. Validation set 실험 결과, 일괄적인 이미지 전처리는 오히려 성능을 저하시킬 수 있음이 확인되었다. 결론적으로 크롭 전처리는 무조건적인 도입을 지양하고, 문서 내 각 Region type을 고려하여 제한적이고 신중하게 적용하는 것이 가장 최적의 결과를 도출하였다.

3.5 다양한 Metric 기반 병목 진단

Two-stage 방식에서는 단순히 Detector나 Recognizer의 개별 성능 지표보다 대회 점수 기준의 End-to-end 성능이 중요했기에 세 가지 단계로 시스템을 진단했다. 먼저 완벽한 탐지를 가정한 상태로 Qwen OCR만 평가한 Stage 2 진단에서는 점수가 0.792에 그쳐 Recognition 단계 자체의 개선 필요성을 확인한 반면, 예측 Bbox에 GT Text를 주입한 Stage 1 진단에서는 기존 YOLO만으로도 최대 0.934점 달성이 가능해 Detector만이 시스템의 병목 원인이 아님을 파악했다. 이러한 분석을 바탕으로, 전체 평가의 50%를 차지하는 Page CER 지표를 향상시키려면 Reading order 정렬과 중복 텍스트 제거가 핵심이라고 판단하였고, 최종 End-to-end 점수를 기준으로 detector_conf, nms_iou, line_bucket_px, text_similarity_threshold 등 주요 파라미터에 대한 Grid Search를 수행하여 시스템을 최적화했다.

4. 실험 결과 및 분석

4.1 One-stage 실험 결과

One-stage 방식의 성능 변화는 다음과 같다.

[표 5] One-stage 단계별 LB 점수

단계	방법	LB	핵심
prompt-only	Qwen2.5-VL-7B 전체 페이지 → JSON	0.274	형식 붕괴
page chunk	페이지를 band로 분할 추론	0.264	경계 잘림
bbox 보정	텍스트 유지, 박스 후처리	0.284	IoU 일부 개선
Curriculum LoRA SFT	쉬운 예제 → 어려운 예제 단계 학습	0.700	대폭 향상
text-focused boost	OCR Teacher 증류	0.779	전사 강화
중복 제거	chunk overlap dedup	0.780	최종

One-stage 실험에서 가장 중요한 결과는 prompt와 후처리만으로는 0.28 수준의 성능에 머물렀지만, 학습 기반 접근을 적용하자 0.70까지 상승했다는 점이다. 이는 VLM이 단순 prompt 지시만으로 정밀한 region-level JSON 생성을 수행하기 어렵고, 형식, 좌표, 전사 방식을 task-specific하게 학습해야 한다는 것을 보여준다. 또한 OCR Teacher 증류를 통해 0.779까지 점수가 상승한 것은 우크라이나어 필기체 전사 능력이 최종 점수에 큰 영향을 준다는 점을 보여준다. 최종적으로 One-stage 방식은 0.780을 달성하였으며, 통합 구조의 가능성을 확인하였다.

4.2 Stage 1 Detector 분석

Stage 1에서는 YOLO11m 기반 detector가 최종적으로 채택되었다. 세부 튜닝 결과는 다음과 같다. 앞의 표 4가 여러 검출기를 같은 조건에서 비교한 전체 클래스 기준 결과라면, 표 6은 최종 채택한 YOLO11m을 추론 설정까지 정밀 튜닝한 결과다.

[표 6] YOLO 튜닝 성능 비교

방법	상세 설명	mAP50	IoU	Recall	Precision
YOLO11m 기본 모델	Gold data fine-tuning	0.639	0.815	0.886	0.936
데이터 증강	Silver data 3,000장 추가	0.594	-	-	-
Pretrained weight 변경	DocLayNet YOLO 사용	0.625	-	-	-
해상도 조정	1280 → 1536	0.566	-	-	-
해상도 및 데이터 비율 조정	짧은 조각 oversampling	0.559	-	-	-

기본 YOLO11m 모델은 IoU 0.815로 bbox 정확도가 높았으며, precision 0.936으로 잘못된 박스를 많이 생성하지 않았다. 그러나 recall은 0.886으로, 일부 텍스트 region을 아예 탐지하지 못하는 문제가 남아 있었다. 이 미검출 문제는 이후 전체 성능의 중요한 병목으로 작용하였다. 추론 하이퍼파라미터 실험에서는 confidence threshold 0.50, image size 1280, NMS IoU 0.70이 가장 안정적이었다. 이미지 해상도를 1920, 2560, 3200으로 높이면 오히려 recall과 precision이 하락하였다. 이는 단순히 해상도를 키우는 것이 작은 글자 탐지 개선으로 이어지지 않으며, 모델 학습 스케일과 추론 스케일의 불일치가 발생할 수 있음을 시사한다.

4.3 Stage 2 Recognition 및 Two-stage 성능 변화

Two-stage 파이프라인은 Surya 단독 baseline에서 시작하여 최종 LB 0.816까지 개선되었다.

[표 7] Two-stage 단계별 LB 점수

방법	LB	ΔLB	핵심
Surya 단독	0.439	-	오픈소스 baseline
YOLO11m 검출 + Qwen2.5-VL-3B	0.660	+ 0.220	Two-stage 전환
conf0.50 + margin 튜닝	0.719	+ 0.059	검출 임계값 조정
Qwen3-VL-8B QLoRA	0.786	+ 0.067	인식기 스케일
dedup 0.55	0.800	+ 0.013	중복 제거
beam3 디코딩	0.813	+ 0.013	탐색 폭 확대
방법 A	0.816	+ 0.003	저신뢰 박스 회수

가장 큰 향상은 두 지점에서 발생하였다. 첫 번째는 Surya 단독에서 YOLO11m 검출과 Qwen2.5-VL-3B 인식 구조로 전환한 시점이다. 이때 LB가 0.439에서 0.660으로 상승하였다. 두 번째는 Qwen3-VL-8B QLoRA로 인식기 스케일을 확대한 시점이다. 이때 LB가 0.719에서 0.786으로 상승하였다. 인식기별 crop CER을 비교하면 Qwen2.5-VL-3B zero-shot은 CER 1.20으로 매우 낮은 성능을 보였다. 3B QLoRA는 CER 0.155까지 개선되었고, Qwen3-VL-8B QLoRA는 CER 0.110을 달성하였다. 깨끗한 crop만 사용한 경우 CER은 0.099까지 낮아졌다. 이는 인식기의 규모와 fine-tuning이 OCR 성능에 직접적인 영향을 준다는 점을 보여준다.

4.4 병목 진단과 음성 결과

추가 분석 결과, source별 성능 차이가 뚜렷했다. dictation은 0.918로 가장 높은 점수를 보였고, university는 0.762, school은 0.732로 가장 낮았다. school 데이터는 전체의 59%를 차지하면서도 짧은 조각, 뺄뺄한 필기, 낙서가 많아 가장 어려운 도메인이었다. 오라클 분석에서는 짧은 조각을 완벽히 회수하면 점수가 0.8355까지 상승하고, 모든 미탐지 region을 완벽히 회수하면 0.8530까지 상승할 수 있음을 확인하였다. 그러나 이 상한 역시 상위권 0.92대와는 차이가 있어, 단순한 recall 개선만으로는 최상위 성능에 도달하기 어렵다. 음성 결과도 중요한 의미를 갖는다. 융합 실험은 위치를 정확히 알기 어려워 -0.042의 성능 하락을 보였고, 줄분할은 과분할 문제로 -0.036의 하락을 보였다. 1536 해상도 및 oversampling은 scale mismatch로 -0.007, TTA는 가짜 박스 증가로

-0.005, 외부 데이터는 도메인 불일치로 -0.012의 하락을 보였다. 이러한 결과는 무위험으로 적용할 수 있는 후처리나 데이터 확장 전략이 거의 소진되었음을 의미한다. 최종 진단에 따르면 region CER 0.278은 대략 인식 오류 0.15와 미탐지 별점 0.10으로 구성되었다. 즉, 오류의 상당 부분은 recognizer가 잘못 읽은 것이 아니라 입력 crop 자체가 제공되지 않은 데서 발생하였다. 농친 조각의 48%가 5자 이하의 숫자, 날짜, 짧은 단어였다는 점도 detector recall이 현 구조의 중요한 천장임을 보여준다.

4.5 두 접근법의 비교

[표 8] One-stage와 Two-stage 비교

항목	One-stage	Two-stage
구조	Qwen2.5-VL 단일 모델	YOLO 검출 + Qwen3-VL 인식
학습	Curriculum LoRA SFT	검출과 인식 각각 학습
검출	VLM이 bbox 직접 출력	전용 detector 사용
장점	단순, 통합, Page CER 직접 최적화 가능	모듈 진단 가능, 정밀 bbox, 교체 용이
약점	긴 JSON 붕괴, bbox 부정확	단계 오류 전파, 검출 recall 천장
최고 LB	0.780	0.816

현재 성능 기준으로는 Two-stage 방식이 우위에 있었다. 이는 전용 detector를 사용함으로써 bbox 탐지 안정성이 높고, 각 모듈의 오류를 독립적으로 진단할 수 있었기 때문이다. 특히 detector와 recognizer를 분리함으로써 어느 단계에서 점수 손실이 발생하는지 오라클 실험으로 확인할 수 있었다. 반면 One-stage 방식은 구조가 단순하고, 이론적으로는 page-level 출력을 직접 학습할 수 있다는 장점이 있다. 그러나 현재의 VLM은 긴 JSON 출력에서 형식이 깨지고, 정밀한 bbox 좌표를 안정적으로 생성하는 데 한계가 있었다. 따라서 One-stage 방식은 향후 더 강력한 VLM과 구조화 출력 학습 기법이 결합될 경우 발전 가능성이 있지만, 본 연구의 실험 조건에서는 Two-stage 방식이 더 높은 성능을 보였다.

5. 결론

본 연구에서는 RUKOPYS Handwritten Text Recognition Challenge를 대상으로 One-stage와 Two-stage 딥러닝 파이프라인을 구현하고 비교하였다. RUKOPYS 문제는 단순 OCR이 아니라, 문서 이미지에서 region을 탐지하고, type을 분류하고, 텍스트를 전사하며, 페이지 단위 reading order까지 재구성해야 하는 복합적인 문서 이해 문제이다. 특히 평가 지표에서 Page CER과 Region CER이 80%를 차지하기 때문에 텍스트 인식 품질이 중요하지만, detection 실패는 OCR 입력 누락으로 이어져 최종 CER에도 큰 영향을 미친다.

One-stage 방식에서는 Qwen2.5-VL을 사용하여 전체 페이지 이미지로부터 bbox, type, text를 포함한 JSON을 직접 생성하도록 하였다. 초기 prompt-only 방식은 긴 JSON 형식 붕괴와 부정확한 bbox 문제로 0.274의 낮은 점수를 보였다. 그러나 Curriculum LoRA SFT를 적용하자 점수가 0.700까지 크게 상승하였고, OCR Teacher 기반 지식 증류와 중복 제거를 통해 최종 LB 0.780을 달성하였다. 이는 prompt와 후처리보다 task-specific fine-tuning이 훨씬 중요하다는 점을 보여준다. Two-stage 방식에서는 YOLO11m detector와 Qwen3-VL-8B QLoRA recognizer를 결합하였다. Detector 비교 결과 YOLO11m Gold fine-tuning이 가장 안정적인 성능을 보였고, Qwen 기반 recognizer는 ViT 기반 인식기보다 우수한 성능을 보였다. Strict OCR prompt, output cleanup, dedup, beam search, 저신뢰 박스 회수 등의 후처리를 적용하여 최종 LB 0.816을 달성하였다. 이는 본 연구에서 가장 높은 점수이다.

실험 결과, 현재 구조에서는 Two-stage 방식이 One-stage 방식보다 더 높은 성능을 보였다. 그 이유는 전용 detector가 정밀한 bbox를 제공하고, detection과 recognition을 분리하여 오류를 진단하고 교체할 수 있기 때문이다. 그러나 Two-stage 방식은 단계 오류 전파와 detector recall 천장이라는 한계를 갖는다. 특히 짧은 조각이나 school source의 복잡한 필기에서 미검출 문제가 두드러졌다. 본 연구의 주요 학습 내용은 다음과 같다. 첫째, RUKOPYS 데이터셋은 도메인 갭이 커서 zero-shot VLM이나 범용 detector만으로는 충분한 성능을 얻기 어렵다. 둘째, 병목은 단순히 OCR 성능이 아니라 detection recall과 page reconstruction에도 존재한다. 셋째, 학습 기반 접근은 prompt와 후처리의 한계를 크게 넘어설 수 있다. 넷째, LoRA, QLoRA, 지식 증류와 같은 효율적 학습 기법은 제한된 GPU 환경에서 대형 VLM을 운용하는 데 효과적이다. 다섯째, 음성 결과도 연구적으로 의미가 있으며, 외부 데이터, 해상도 증가, 단순 융합과 같은 직관적 전략이 항상 성능 향상으로 이어지지 않음을 확인하였다. 향후 연구에

서는 세 가지 방향을 고려할 수 있다. 첫째, 짧은 조각과 최소 클래스에 특화된 detector 학습이 필요하다. 예를 들어 작은 bbox를 위한 multi-scale training이나 hard example mining을 적용할 수 있다. 둘째, Page CER을 직접 줄이기 위한 reading order 모델링과 layout-aware post-processing이 필요하다. 셋째, One-stage 방식에서는 긴 JSON 출력을 안정적으로 생성하기 위한 constrained decoding 또는 structured output learning을 적용할 수 있다. 이러한 개선을 통해 통합형 VLM의 잠재력과 Two-stage 구조의 안정성을 함께 활용하는 하이브리드 문서 인식 시스템으로 발전시킬 수 있을 것이다.

참고문헌

- Baek, Y., Lee, B., Han, D., Yun, S. and Lee, H. (2019). Character Region Awareness for Text Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems.
- Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations.
- Jocher, G., Qiu, J. and Chaurasia, A. (2024). Ultralytics YOLO11 [Computer software]. <https://github.com/ultralytics/ultralytics>.
- Kaggle (2026). RUKOPYS Handwritten Text Recognition Challenge. <https://www.kaggle.com/competitions/handwritten-to-data>.
- Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z. and Wei, F. (2023). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. Proceedings of the AAAI Conference on Artificial Intelligence.
- Liao, M., Wan, Z., Yao, C., Chen, K. and Bai, X. (2020). Real-time Scene Text Detection with Differentiable Binarization. Proceedings of the AAAI Conference on Artificial Intelligence.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). Focal Loss for Dense Object Detection. IEEE International Conference on Computer Vision (ICCV).
- Qwen Team (2025). Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923.

흉부 X-ray 이상 소견 탐지를 위한 RT-DETR 기반 2-Stage 객체 탐지 모델 연구

송채린¹⁾, 김태환¹⁾, 이송민²⁾, 함민아²⁾

Abstract

Chest X-ray imaging is a representative medical imaging modality widely used for the screening and diagnostic support of various thoracic diseases, including pneumonia, pneumothorax, pleural effusion, pulmonary nodules, and cardiomegaly. However, because chest X-ray images represent three-dimensional anatomical structures projected onto a two-dimensional plane, organs and lesions often overlap. In addition, the size, shape, and contrast of lesions vary considerably, making automatic localization and classification technically challenging. In this study, we addressed the chest X-ray abnormality detection task presented in the Kaggle AMIA Public Challenge 2026 and implemented object detection models for 14 thoracic abnormality classes to analyze potential performance improvements. Specifically, we designed both End-to-End and 2-stage architectures based on the real-time object detection model RT-DETR, incorporating a ResNet-50 classifier pretrained on the chest X-ray domain. We compared the detection performance of the two approaches and analyzed the effects of data augmentation and post-processing strategies, including the removal of Random Erasing and Mosaic augmentation, as well as classifier threshold adjustment. The experimental results showed that the 2-stage model achieved the highest performance, with an mAP of 0.517 at IoU=0.40 on the public leaderboard, ranking 11th in the competition. These results suggest that

1) All authors have equal contribution, Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea

2) Department of Smart Cities, Chung-Ang University, Seoul 06974, Korea

integrating a chest X-ray domain-pretrained classifier into an object detection pipeline can contribute to improving abnormality detection performance in medical imaging.

1. 서론

흉부 X-ray 영상은 폐렴, 기흉, 흉수, 결절, 심비대 등 다양한 흉부 질환을 선별하는 데 널리 사용되는 영상 검사이다. X-ray는 CT나 MRI에 비해 촬영 비용이 낮고 접근성이 높기 때문에 임상 현장에서 가장 보편적으로 활용되는 검사 중 하나이다. 그러나 흉부 X-ray 영상은 해부학적 구조가 겹쳐 나타나며, 병변의 크기와 명암 대비가 다양하기 때문에 자동 판독 모델을 구축하는 데 어려움이 있다.

Kaggle에서 개최된 AMIA Public Challenge 2026(Kaggle, 2026)은 흉부 X-ray 사진에서 14가지 흉부 이상 소견을 자동 탐지 및 분류하는 객체 탐지 모델 구현을 요구한다. 해당 대회는 주어진 흉부 X-ray 샘플 데이터에 대한 클래스 예측, confidence score, 바운딩 박스 예측을 요구한다. 본 연구에서는 YOLO 계열 객체 탐지 딥러닝 모델 RT-DETR 모델을 베이스라인으로 데이터 전처리, augmentation, confidence threshold 조정 등을 적용하여 탐지 성능 개선을 시도하였다.

본 보고서에서 2장은 해당 대회에서 사용한 모델의 아키텍처 구조를 설명한다. 3장에서는 데이터셋, 방법론, 대회에서 사용된 모델 평가지표를 설명한다. 4장에서는 실험 비교 및 결과를 제시한다. 5장에서는 실험 결과에 대한 한계점 및 개선 방안을 논의한다.

2. 선행 연구

2.1 TorchXRayVision

TorchXRayVision(Cohen et al., 2022)은 흉부 X-ray 데이터셋과 사전학습 모델을 함께 제공하는 오픈소스 라이브러리이다. 공개된 흉부 X-ray 데이터셋은 영상 포맷과 라벨 체계가 상이하여 재사용이 용이하지 않으며, 임상 문제를 다루는 연구자가 모델을 매번 처음부터 학습하는 것은 비효율적이다. TorchXRayVision은 이러한 한계를 해소하기 위해, 다수의 데이터셋에 대한 통일된 전처리 인터페이스와 더불어 다양한 데이터 조합으로 사전학습된 모델을 baseline 또는 feature extractor로 제공하도록 설계되었다.

본 연구에서 사용한 resnet50-res512-all 가중치는 512×512 해상도에서 NIH, CheXpert, PadChest, MIMIC-CXR 등 다수의 공개 데이터셋을 통합하여 학습한 ResNet-50 모델로, 18종의 병리 소견에 대한 multi-label classification을 수행하도록 사전학습되어 있다. 흉부 X-ray가 단일 채널(grayscale) 영상이므로 입력 합성곱 계층은 1채널 입력을 받도록 수정되어 있다.

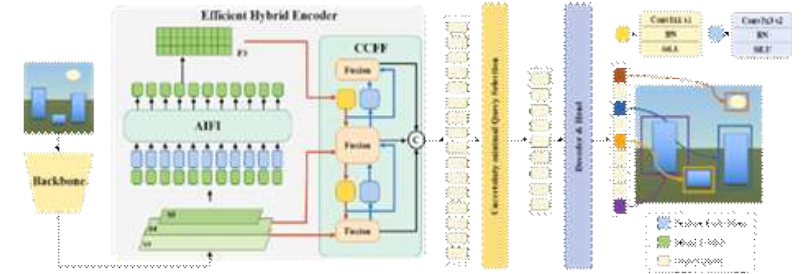
해당 backbone은 본 연구의 Stage 1 분류기에 활용되며, ImageNet 사전학습 가중치와 비교하여 다음과 같은 이점을 가진다. 첫째, 본 대회 데이터와 동일한 흉부 X-ray 도메인에서 학습되어 도메인 정합성이 높다. 본 데이터셋에서 병변을 포함한 이미지는 약 3,398장에 불과하므로 소규모 데이터로 fine-tuning을 수행할 때 도메인 정합 사전학습은 전이학습의 효율을 제고한다. 둘째, 사전학습된 가중치를 활용함으로써 학습 비용을 절감할 수 있다. 셋째, 가중치에 18종 병리에 대한 표현이 이미 인코딩되어 있어, 본 연구가 목표로 하는 병변 유무 판별에 유효한 특징을 재사용할 수 있다.

본 연구에서는 해당 사전학습 backbone을 활용하되, 18종 병리를 출력하는 분류 head를 병변 유무를 구분하는 binary head로 교체하여 fine-tuning하였다.

2.2 RT-DETR(Real-Time DEtector TRansformer)

본 논문은 DETR 구조를 실시간 객체 탐지에 적합하도록 개선한 RT-DETR을 제안한다. RT-DETR(Zhao, Y. et al., 2024)은 backbone, efficient hybrid encoder, Transformer decoder로 구성되며, backbone의 마지막 세 단계 feature map인 S_3 , S_4 , S_5 를 입력으로 사용한다. 기존 DETR 계열 모델은 multi-scale feature를 Transformer encoder에서 함께 처리하면서 계산량이 커지는 문제가 있었기 때문에, RT-DETR은 feature 내부 상호작용과 scale 간 융합을 분리한 efficient hybrid encoder를 설계하였다.

[그림 1] RT-DETR 아키텍처



Efficient hybrid encoder는 Attention-based Intra-scale Feature Interaction(AIFI)와 Cross-scale Feature Fusion(CCFF)로 구성된다. AIFI는 semantic 정보가 풍부한 high-level feature S_5 에만 Self-attention을 적용하여 계산량을 줄이고, CCFF는 CNN 기반 fusion block을 이용해 S_3 , S_4 , S_5 를 결합한다. 즉, Transformer는 의미 정보가 강한 feature의 관계 학습에만 사용하고, scale 간 feature fusion은 CNN으로 처리하여 속도와 정확도의 균형을 맞춘다.

기존 query selection은 classification score만 기준으로 encoder feature를 선택했지만, 객체 탐지는 class뿐 아니라 bounding box 위치도 중요하다. 따라서 uncertainty-minimal query selection을 제안하며 RT-DETR은 classification 예측과 localization 예측 사이의 불확실성(식 1)을 고려하여 더 신뢰도 높은 encoder feature를 decoder의 초기 object query로 사용한다. 이를 통해 decoder가 더 좋은 초기 query에서 탐지를 시작할 수 있어 최종 정확도가 향상된다. 또한 decoder layer 수를 조절하여 별도의 재학습 없이 추론 속도를 조정할 수 있다. 즉 End-to-End DETR 구조를 유지하면서도, 계산량을 줄이면서 query 품질을 높여 실시간 객체 탐지에 적합하도록 설계된 모델이다.

$$U(\hat{x}) = \|P(\hat{x}) - C(\hat{x})\|, \hat{x} \in R^D \quad (\text{식 1})$$

$$L(\hat{\chi}, \hat{Y}, Y) = L_{\text{box}}(\hat{b}, b) + L_{\text{cls}}(U(\hat{\chi}), \hat{c}, c) \quad (\text{식 2})$$

3. 방법론

선행 연구를 통해 TorchXRayVision 오픈소스 라이브러리에서 흉부 X-ray 데이터 도메인으로 Multi-classification task 해결을 위해 학습된 ResNet-50(He, K., 2016) 모델 구조의 가중치가 공개되어 있음을 확인하였다. 그러나 해당 데이터 도메인에서 객체 탐지를 수행하는 모델은 찾을 수 없었다. 본 보고서에 작성된 실험들은 2-Stage와 End-to-End 전략으로 나뉜다.

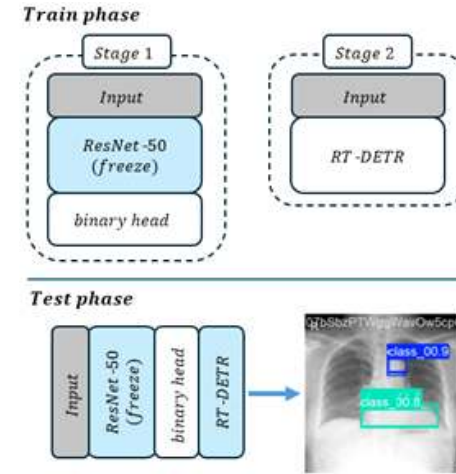
3.1 End-to-End 방법론

<표 4>의 baseline은 RT-DETR의 전체 레이어를 fine-tuning한 실험이다. 이에 이어 TorchXRayVision 모델을 결합한 End-to-End 방법론은 다음과 같다. 본 실험에서 사용된 RT-DETR 모델의 backbone은 ResNet-50 구조로, ImageNet 분류 데이터셋으로 사전학습되었다. ImageNet 분류 데이터셋은 흉부 X-ray 도메인 데이터셋과 도메인이 다르다고 판단하여, 흉부 X-ray 데이터셋으로 사전학습된 backbone 가중치에서 fine-tuning하는 것이 더 좋은 성능을 낼 것으로 기대하였다.

3.2 2-stage 방법론

2-Stage의 경우 Stage 1에서 병변 유무를 탐지하는 binary classifier를 학습하고, Stage 2에서 RT-DETR 모델을 fine-tuning시켰다. Stage 1에서의 입력 이미지는 병변 bounding box가 하나 이상 존재하면 abnormal label을 1로, 병변 box가 없는 이미지는 0으로 정의하였다. 이후 추론과정에서는 병변이 있다고 판단된 이미지들만 detector로 전달하여 object detection task를 수행하도록 하였다. End-to-End는 학습 과정을 나누지 않고, multi-classification과 object detection task를 한꺼번에 해결하도록 학습되었다. 추론 과정에서는 양성 확률을 binary classifier에서 계산한 후 양성 확률이 임계값보다 더 높게 나온 이미지들만 object detection을 수행한다(그림 2). 각 Stage에서 사용한 주요 하이퍼 파라미터들은 <표 1>와 같다.

[그림 2] 2 Stage pipeline



[표 1] Hyper parameter 설정

구분	Stage 1 (Binary Classifier)	Stage 2 (Detector)
Epoch	15	20
Optimizer	AdamW	AdamW
Learning rate	Head: 1e-3 Backbone: 1e-4	1e-4
Batch size	32	8

Stage 1의 classifier는 클래스 불균형을 보정하기 위해 손실 함수로 양성 클래스에 가중치 w 를 둔 Binary Cross-Entropy loss를 사용하였다(식 3). 가중치 w 는 정상 이미지 수와 이상 이미지 수의 비율로 설정한다.

$$L_{WBCE} = -\frac{1}{N} \sum_{i=1}^N [w_+ y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (\text{식 3})$$

Stage 1의 classifier는 흉부 X-ray 이미지들로 사전학습된 ResNet-50 가중

치를 활용하되, 기존의 가중치를 많이 보존하기 위해 점진적으로 unfreezing하는 방법을 사용했다. 첫 3 epoch는 classifier head부분만 학습을 진행했고, 이후 4 epoch 동안은 head와 backbone의 마지막 layer인 layer4를 함께 학습하였다. 그리고 마지막으로 전체 network를 unfreeze해 학습하였다. 또한 사전학습된 표현을 보존하기 위해 backbone의 learning rate는 1e-4로 classifier head의 1e-3보다 낮게 설정하여 적용하였다. 마지막으로 모델 학습시에는 15%의 데이터를 검증 데이터로 사용, early stopping을 적용해 과적합을 방지했다.

3.3 모델 평가지표

mAP(mean Average Precision)는 객체 탐지 모델의 성능을 평가하는 데 널리 사용되는 표준 지표로, 각 클래스별 Average Precision(AP)을 산술 평균하여 산출된다. 이는 영상 내 병변의 위치를 정확히 특정하는 Localization 성능과 병변의 종류를 올바르게 식별하는 Classification 성능을 하나의 값으로 종합하여 반영하므로, 탐지 모델의 전반적인 성능을 일관된 기준으로 비교할 수 있게 한다. 본 연구의 모델 성능 또한 이 지표를 기준으로 평가되었으며, 구체적으로는 PASCAL VOC 2010 표준 프로토콜에 기반한 mAP(IoU=0.40)가 사용되었다.

정답 인정 기준이 되는 IoU(Intersection over Union) 임계값 0.40은 Kaggle AMIA Public Challenge 2026 대회에서 지정한 평가 조건이다. 이는 모델의 예측 경계 상자와 실제 정답 상자의 중첩 면적 비율이 40% 이상일 때 정답으로 인정함을 의미한다. 일반적인 객체 탐지에서 통용되는 0.50보다 완화된 기준으로, 흉부 방사선 영상 내 병변(예: 침윤, 기흉 등)의 물리적 경계가 본질적으로 모호하고 판독 전문의 간에도 라벨링 편차가 존재할 수 있는 의료 영상의 특수성을 반영한 것으로 볼 수 있다. 즉 픽셀 단위의 완벽한 일치보다는 임상적으로 유의미한 병변 부위를 올바르게 포착했는지를 평가하는 데 초점이 있다.

각 질환 클래스별 AP는 Precision-Recall 곡선의 모든 고유 재현율 지점에서 최대 정밀도를 적분하는 All-point 보간법으로 계산된다. 최종 mAP는 14개 이상 징후 클래스 각각의 AP 값을 산술 평균하여 도출되며, 계산식은 다음과 같다.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (\text{식 4})$$

결과적으로 본 평가지표는 특정 혼한 질환에만 과적합되거나 정상 영상에 대해 무분별하게 오탐을 발생시키는 모델을 배제하고, 복잡한 임상 환경에서 전문

의를 안정적으로 보조할 수 있는 범용적이고 안전한 모델을 선별할 수 있도록 가능하다.

4. 실험 결과

4.1 Data description

대회에서 제공한 데이터셋은 실제 흉부 이미지 데이터가 저장된 2개의 폴더와 데이터의 분석 및 제출을 보조하는 4개의 메타데이터(CSV) 파일로 구성된다. 먼저 이미지 데이터는 모델 학습을 위해 사용되는 train 폴더(8,573장)와 모델의 최종 추론 및 평가를 위한 test 폴더(6,427장)로 나뉘며, 각 폴더 내부에는 DICOM 파일에서 변환된 1024x1024 해상도의 PNG 형식 이미지들이 저장되어 있다. 이와 함께 제공되는 4개의 CSV 파일은 이미지와 타겟 라벨 간의 관계를 정의한다. train.csv 파일은 train 폴더 내의 이미지에서 검출된 이상 소견 객체당 하나의 행으로 기록된 메타데이터를 포함한다. 특히 하나의 이미지에 다수의 병변이 존재하거나 여러 판독의가 중복 진단한 경우, 동일한 이미지 식별자(image_id)에 대해 여러 행이 생성되는 다중 행 구조를 띈다. 반면 test.csv 파일은 추론 대상이 되는 테스트 이미지의 파일명 목록만을 단순 포함하고 있다. 이 외에 이미지 리사이즈 전의 실제 치수 정보(세로 높이 dim0, 가로 너비 dim1)를 담은 img_size.csv 파일과 최종 예측 결과물의 제출 규격을 안내하는 샘플 양식인 sample_submission.csv 파일이 함께 구성되어 있다. 모델 학습의 타겟 레이블이 되는 train.csv 파일의 세부 컬럼 구성과 상세 정의는 아래 <표 2>과 같다.

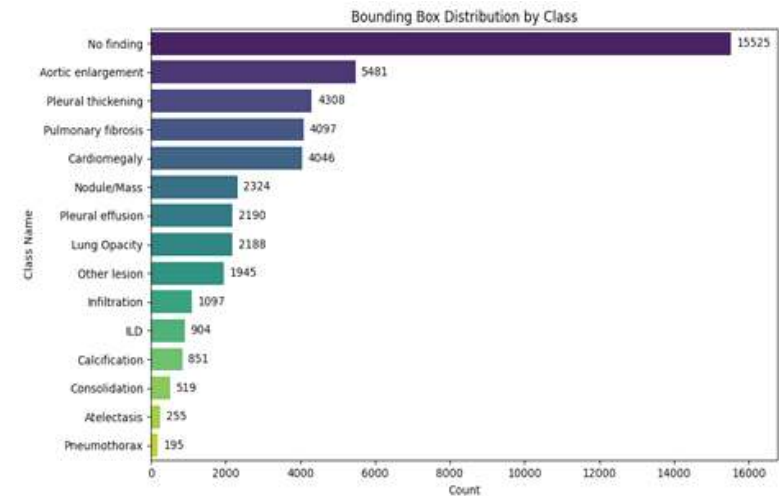
[표 2] train.csv 데이터 컬럼 구성 및 정의

컬럼명	상세 설명
image_id	흉부 이미지의 고유 식별자
class_name	검출된 이상 소견의 명칭(정상인 경우 “No finding”)
class_id	검출된 객체의 클래스 고유 ID(0~13: 이상 소견, 14: No finding)
rad_id	해당 소견을 진단하고 라벨링을 수행한 방사선과 의사 고유 ID
x_min	이상 소견 바운딩 박스 좌상단 X 좌표의 최솟값
y_min	이상 소견 바운딩 박스 좌상단 Y 좌표의 최솟값
x_max	이상 소견 바운딩 박스 우하단 X 좌표의 최댓값
y_max	이상 소견 바운딩 박스 우하단 Y 좌표의 최댓값

[표 3] 이미지 데이터 분포

데이터 구분	이미지 규모	데이터 비율	주요 라벨 특징 요소
전체 이미지	8,573장	100%	이미지당 평균 5.357행의 라벨 정보 포함
정상 이미지	5,175장	60.4%	병변 박스가 없는 정상 데이터(Class 14)
이상 이미지	3,398장	39.6%	최소 1개 이상의 병변 박스 포함(Class 0~13)

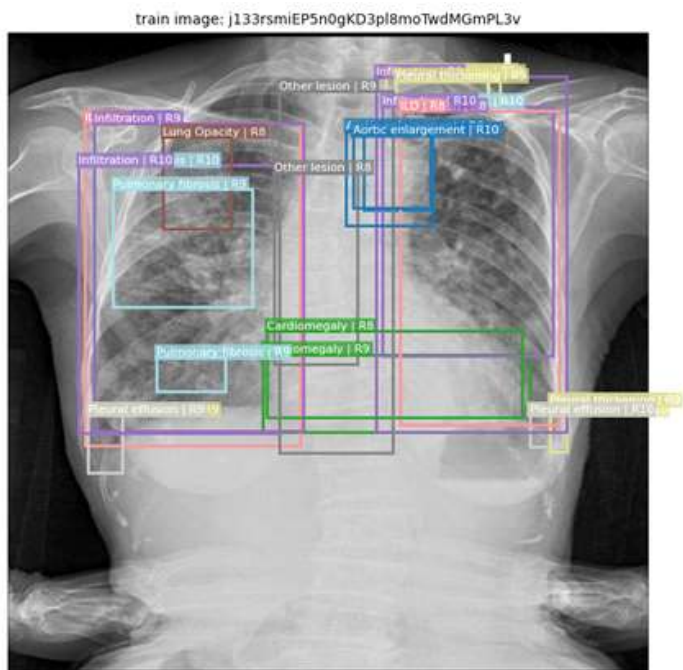
[그림 3] Bounding box 분포



전체 이미지 기준으로는 이상 소견이 있는(비정상) 이미지가 전체 이미지 중에서 39.6%를 차지하며, 나머지는 정상 이미지로 60.4%를 차지한다(표 3). [그림 3]은 Bounding box의 분포를 막대 그래프로 표현한 것이다. 양성 그룹 내에서도 Aortic enlargement에 비해 Pneumothorax 등 특정 클래스의 빈도가 매우 낮아 병변 클래스의 분포가 상당히 불균형한 것을 확인할 수 있다. ○

실제 학습 데이터의 예시인 [그림 4]는 해부학적 구조의 겹침과 병변 특성의 다양성으로 인한 데이터셋의 시각적 복잡성을 보여준다. 3차원인 흉부 구조가 2차원 평면에 투영되는 X-ray 특성상 내부 장기와 병변이 중첩되어 나타나며, 병변의 크기와 명암 대비가 매우 다양하여 일관된 자동 탐지에 난점이 존재한다. 본 데이터셋은 총 17명의 의사 중 이미지당 3명이 배정되어 독립적으로 판독을 수행하였으며, 이미지 내 라벨을 통해 박스 라벨링 전담의인 R8, R9, R10이 참여했음을 확인할 수 있다. 특히 동일 병변에 대해 의사 간 미세한 좌표 편차로 바운딩 박스가 조밀하게 중첩되는 현상과, 한 이미지 내에 다수의 질환 소견이 동시에 발현되는 클래스 동반 출현 패턴이 뚜렷하게 관찰된다.

[그림 4] 다중 판독 및 클래스 동반 출현 예시 이미지



4.2 실험 결과

본 보고서의 모든 실험은 대회 평가지표인 mAP(IoU=0.40)를 기준으로 비교하였다(표 4). Ultralytics 패키지 기반 RT-DETR 학습에는 기본적으로 image translation, random scaling, horizontal flipping 등 다양한 데이터 증강 기법이 적용되며, 이는 입력 영상의 위치, 크기, 방향 변화를 통해 모델의 일반화 성능 향상에 기여한다. 다만 Random Erasing은 영상 일부를 가리는 패치를 삽입하면서도 ground truth box를 유지하기 때문에, 작은 변형 탐지에는 불리하게 작용할 수 있다. 이에 해당 증강을 제거하여 실험하였고, 그 결과 더 높은 성능을 확인하였다.

[표 4] 실험 결과

실험	전략	증강 기법	mAP (IoU=0.4)
RT-DETR (baseline)	End-to-End	default (전체 증강)	0.508
TorchXRayVision & RT-DETR + classifier threshold sweep	End-to-End	erasing=0.0	0.492
TorchXRayVision & RT-DETR	2 Stage	default (전체 증강)	0.510
TorchXRayVision & RT-DETR + classifier threshold sweep	2 Stage	default (전체 증강)	0.510
TorchXRayVision & RT-DETR	2 Stage	erasing=0.0	0.493
TorchXRayVision & RT-DETR	2 Stage	erasing=0.0 + mosaic=0.0	0.493
TorchXRayVision & RT-DETR + classifier threshold sweep	2 Stage	erasing=0.0 + mosaic=0.0	0.499
TorchXRayVision & RT-DETR + classifier threshold sweep	2 Stage	erasing=0.0	0.517

또한 여러 이미지를 한 장으로 합치는 과정에서 구성 이미지를 약 1/4 면적으로 축소하는 mosaic에 대해서도 제거하여 실험하였으나, 낮은 성능을 보여 최종 모델에는 활용하지 않았다. 추론과정에서는 binary classifier의 분류 기준인 threshold를 조절하는 후처리를 시행하였다. 0.1, 0.2, 0.3, 0.4 등의 값으로 실험하여 최적의 결과를 도출하였다. 결론적으로 erasing을 제거한 2-stage 모델의 classifier threshold 실험 결과가 leader board public score 기준 0.517로 가장

우수한 성능을 보였다.

5. 결론 및 한계점

본 프로젝트에서는 AMIA Public Challenge 2026 흉부 X-ray 병변 탐지 대회에 참여하여, RT-DETR 기반의 End-to-End 전략과 ResNet-50 classifier를 결합한 2-stage 전략을 설계하고 비교하였다. 실험결과, 최종적으로는 2-stage 전략이 최종 public score 기준 mAP 0.517로 제일 좋은 성능을 보였다.

주요 성능 개선 요인은 3가지이다. 우선, 2-stage로 전환함으로써, 1차적으로 classifier를 통한 분류 이후 병변이 있는 이상 이미지들에 한해 detection을 적용하였다. 또한, Random erasing 증강을 제거함으로써 무작위로 이미지에 블록을 만들어 가림으로써 작은 병변의 재현율 저하를 완화했다. 마지막으로 classifier threshold shift를 적용하여 최적의 threshold를 설정하였다.

그러나, 한 이미지에 대한 annotator가 여러 명이었던 점과 병변을 포함한 이미지의 개수가 약 3천개 밖에 되지 않아 성능 개선은 미미했다. 측정 기기 혹은 병원별로 데이터셋의 분포가 다를 가능성 또한 배제할 수 없어 일반화를 기대하기는 힘들다. 결론적으로 더 많은 흉부 X-ray 데이터셋으로 모델을 학습시키면 성능 개선을 이루어낼 수 있을 것으로 예상된다.

참고문헌

Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M. et al. (2022) TorchXRyVision: A library of chest X-ray datasets and models, International Conference on Medical Imaging with Deep Learning, PMLR, pp. 231-249.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.

Kaggle (2026) AMIA Public Challenge 2026, Kaggle. Available at: <https://www.kaggle.com/competitions/amia-public-challenge-2026> (Accessed: 23 June 2026).

Zhao, Y. et al. (2024) DETRs beat YOLOs on real-time object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965-16974.

Multi-label Bioacoustic Classification Using Perch Embeddings and Asymmetric Ensemble for BirdCLEF+ 2026

김시은⁵⁾, 류정찬¹⁾, 백승준¹⁾, 이나현²⁾

Abstract

This study presents a deep learning pipeline for automated multi-label classification of 234 species across multiple taxa (birds, amphibians, mammals, reptiles, and insects) in complex natural soundscapes, developed for the BirdCLEF+ 2026 Kaggle competition. Analyzing 60-second audio recordings from diverse tropical ecosystems, we divided each recording into 5-second segments to identify overlapping species vocalizations.

The pipeline is built upon a pre-trained convolutional neural network (Perch_v2) as a frozen feature extractor, providing 1536-dimensional embeddings and species logits for 203 of the 234 target classes. The EoS9 public notebook anchor — incorporating ProtoSSM for temporal sequence modeling, ResidualSSM for residual error correction, and vectorized MLP Probes for 31 unmapped species — served as the dominant component of our ensemble, receiving a blend weight of 0.967. To mitigate calibration discrepancies across diverse architectures, a rank-based blending strategy was applied when integrating auxiliary models including Distilled SED and BirdNET v2.4. An asymmetric three-branch ensemble (EoS9 / PSSM / yukiZ) was finalized with meticulous post-processing, including per-taxon temperature scaling,

1) All authors have equal contribution, Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea

2) Department of Smart Cities, Chung-Ang University, Seoul 06974, Korea

temporal smoothing, and rare-class calibration.

The methodology achieved an Out-of-Fold (OOF) Macro-AUC of 0.992 and a Private Leaderboard score of 0.94141 (rank 788/4,093, top 19.3%), with the underlying EoS9 public notebook recognized with a Silver Notebook Medal by the Kaggle community. These results demonstrate the pipeline's effectiveness in real-world ecological acoustic monitoring, though the gap between OOF and private scores suggests room for improving generalization beyond OOF-tuned post-processing.

1. 서론

1.1 연구의 배경 및 필요성

최근 생물 다양성 보존과 생태계 모니터링 분야에서 자연환경의 소리를 기록한 사운즈케이프(Soundscape) 기반 음향 데이터 분석이 중요한 역할을 하고 있다. 특히 브라질 판타날 습지와 같이 광활하고 접근이 어려운 자연 생태계에서는 직접적인 관찰이 번거로울 뿐만 아니라 접근 자체가 자연 생태계에 영향을 줄 수 있다. 그러므로 음향 기록 장치를 활용한 다양한 생물종의 모니터링이 훨씬 효율적일 뿐만 아니라 판타날 습지의 자연 생태계를 훼손하지 않는 방법이다. 그러나 이렇게 수집된 방대한 오디오 데이터를 조류학자나 생태 전문가가 직접 청취하고 분석하는 것은 시간과 비용 측면에서 명확한 한계가 있다. 따라서 복잡하고 중첩된 자연의 소리 속에서 특정 생물의 소리를 자동으로 식별할 수 있는 딥러닝 기반 모델의 개발이 필요하다.

1.2 문제 정의 및 데이터의 한계

본 연구는 Kaggle BirdCLEF+ 2026 대회를 기반으로 수행되었다. 해당 과제는 브라질 판타날 습지 일대에 설치된 음향 기록 장치의 녹음 데이터를 분석하여 조류·양서류·포유류·파충류·곤충 등 234개 생물종의 소리 포함 여부를 식별하는 다중 레이블 분류(multi-label classification) 문제이다. 제공 데이터는 개별 종의 음향 특징 학습에 활용되는 명확한 생물의 음성만을 담은 35,549개의 단일 개체 녹음(focal recording)과, 실제 자연환경을 그대로 담은 60초 분량의 학습용 사운즈케이프 10,658개로 구성된다. 평가는 60초 오디오를 5초 단위 12개 구간으로 분할하고 각 구간에 대해 234종의 존재 확률을 산출하는 방식으로 이

루어지며, 최종 제출 파일은 각 구간을 행, 234개 종별 확률을 열로 하는 형식을 따른다.

그러나 학습용 사운드시케이프 중 정답 라벨이 제공된 파일은 극히 일부에 불과하다. 또한 실제 테스트 환경과 동일한 조건인 다중 혼재, 환경 소음이 복합된 상황에서의 라벨 데이터는 더욱 희소하다. 이러한 데이터 불균형과 노이즈 문제로 인해 단일 개체 녹음만으로 학습된 모델이 복잡한 자연환경 사운드시케이프에 그대로 적용될 경우, 도메인 불일치가 발생하기 쉽다. 따라서 제한된 라벨 환경에서도 실제 생태 음향 환경의 복잡한 패턴을 효과적으로 일반화할 수 있는 모델링 전략이 요구된다.

1.3 연구 목적 및 평가 지표

본 연구의 목적은 자연 사운드시케이프 환경에서의 다중 레이블 생물음향 분류를 위한 앙상블 파이프라인을 구축하고 그 성능을 분석하는 것이다. 본 연구에서 제안하는 시스템은 대회 공식 평가 지표인 Threshold-free Macro ROC-AUC를 최적화 목표로 삼는다. 이 지표는 임계값 설정 없이 모든 클래스에 걸쳐 예측 순위의 품질을 균등하게 평가한다는 점에서 클래스 불균형이 심한 생태 음향 환경에 적합하다. 본 보고서에서는 음향 특징 추출, 시계열 문맥 보완, 순위 기반 앙상블, 후처리를 아우르는 파이프라인을 제안하고 실제 테스트 데이터에 대한 결과를 분석한다.

2. 관련 연구 및 핵심 기술

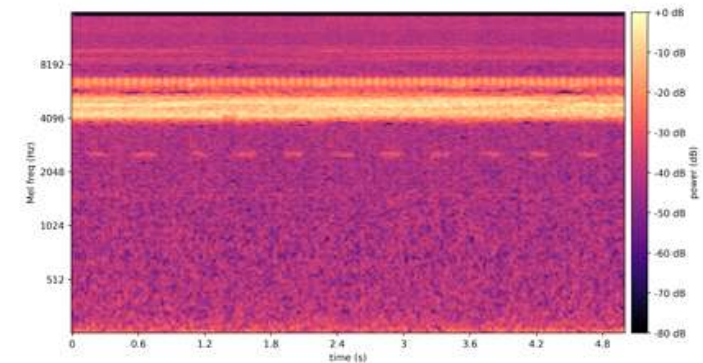
본 장에서는 다중 레이블 기반의 조류 음향 분류 문제의 특성을 정의하고, 전체 파이프라인의 기반이 되는 사전 학습 특징 추출 모델의 구조적 특징과 한계점에 대해 서술한다.

2.1 사운드시케이프 기반 다중 레이블 분류

사운드시케이프는 통제된 환경의 단일 개체 녹음과 달리, 다수의 생물종 소리와 환경 배경 소음이 동시에 중첩된 자연 음향 데이터이다. 1.2절에서 정의한 다중 레이블 분류 문제가 사운드시케이프 환경에서 특히 어려운 이유는 동일한 시간의 구간 안에, 서로 다른 종의 발성이 주파수 및 시간 축에서 겹쳐 나타나기 때문이다. 이때 한 종의 약한 신호가 다른 종의 강한 신호에 가려지거나 환경 소음

과 구분되지 않아 누락되기 쉽다. 또한 5초라는 짧은 구간만을 독립적으로 관찰할 경우, 파편적이거나 미약한 발성은 충분한 단서를 제공하지 못한다. 그림 1은 실제 학습 사운드시케이프의 5초 구간을 Log-mel spectrogram으로 변환한 예시로, 약 4-6kHz 대역에 지속되는 강한 성분과 더 낮은 주파수에서 간헐적으로 반복되는 약한 신호가 동시에 존재함을 보여준다. 따라서 중첩된 신호에서 종별 음향 특징을 분리하고인접 구간의 시간적 맥락을 같이 고려하는 능력이 모델 성능을 좌우한다. 본 연구는 이를 위해 사전 학습된 특징 추출기(Feature extractor)와 시계열 시퀀스 모델을 결합하는 방식을 채택하였다. 그 기반이 되는 특징 추출기를 다음 절에서 설명한다.

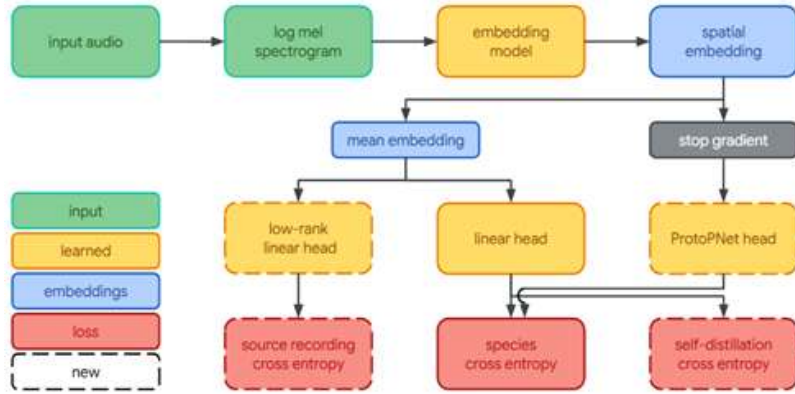
[그림 1] 실제 사운드시케이프 5초 구간의 Log-mel spectrogram.



2.2 Perch 기반 특징 추출기

복잡한 다중 레이블 환경에서 강건한 음향 특징을 포착하기 위해, 본 연구에서는 Google에서 개발하고 사전 학습한 CNN 기반의 조류 음향 분류 모델인 Perch_v2를 고정된 특징 추출기로 도입하였다. Perch_v2는 단순히 종을 직접 분류하는 모델이 아니라, 후속 보정 모델과 앙상블 모델의 입력으로 활용할 수 있는 고차원 음향 표현을 생성한다.

[그림 2] Perch-v2 모델의 구조도



Perch_v2의 연산 과정은 Raw Audio 신호를 Log-mel Spectrogram으로 변환하는 것에서 시작된다. 이후 EfficientNet-B3 기반의 합성곱 신경망을 통과하여 공간 임베딩을 생성하고, 평균 풀링을 거쳐 최종적으로 1536차원의 임베딩 벡터와 기본 중 점수를 도출한다. 이 모델은 중 분류 손실뿐 아니라 프로토타입 기반 분류, 자기 증류, 보조 자기지도 학습 손실을 함께 활용하여 생물 음향 데이터에 적합한 표현을 학습한 모델이다. 따라서 제한된 라벨만 제공되는 사운드시케이프 환경에서도 일반화된 음향 특징을 추출하는 데 유용하다.

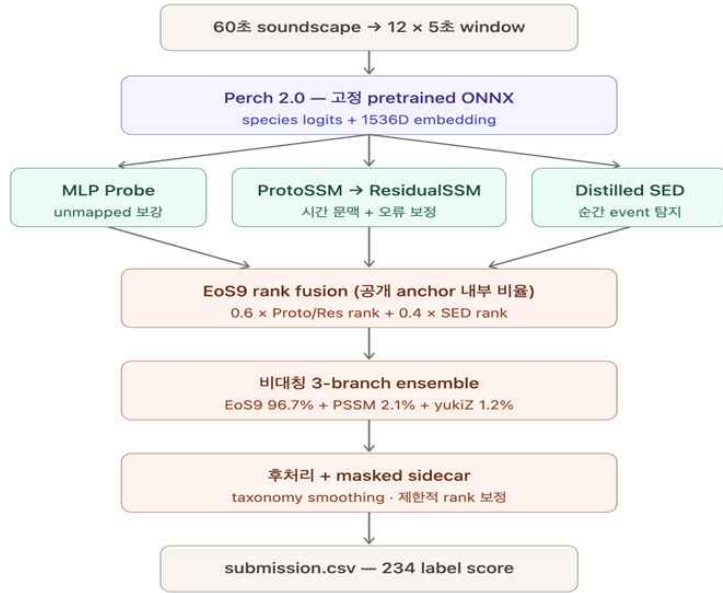
본 연구에서는 Perch_v2의 내부 가중치를 직접 재학습하지 않고, 각 5초 오디오 구간에 대해 1536차원 임베딩과 기본 중 점수를 추출하는 고정된 사전 학습 모델로 사용하였다. 이렇게 추출된 임베딩은 ProtoSSM, ResidualSSM, MLP Probes와 같은 후속 모델의 입력으로 활용되며, 기본 중 점수는 초기 예측 신호로 사용된다. 그러나 Perch_v2를 본 과제에 직접 적용하는 데에는 구조적 한계가 존재한다. 본 대회 타겟 생물종은 총 234종이지만, Perch 모델의 로짓은 이 중 203종과만 직접적으로 매핑되어 있다. 이로 인해 나머지 31개 미매핑 종에 대해서는 사전 학습 모델만으로 충분한 예측 성능을 확보하기 어렵다. 따라서 본 연구에서는 Perch_v2를 단독 분류기로 사용하지 않고, 1536차원 임베딩을 후속 보정 모델과 별도 분류기의 입력으로 활용하는 확장된 파이프라인을 설계하였다. 이러한 설계는 사전 학습 모델의 일반화 성능을 활용하면서도, 대회 타겟 종과 실제 테스트 환경 사이의 차이를 후속 모델을 통해 보완하기 위한 것이다.

3. Methodology

본 장에서는 Perch_v2로 추출한 음향 특징을 바탕으로 시간적 문맥을 보완하고, 미매핑 종과 예측 오차를 보정한 뒤, 여러 예측 브랜치를 결합해 최종 제출 파일을 생성하는 방법론을 설명한다. 전체 추론 파이프라인은 그림 3에 제시하였다.

[그림 3]과 같이 본 연구의 추론 과정은 60초 길이의 사운드시케이프 오디오를 5초 단위의 12개 구간으로 분할하는 단계에서 시작된다. 각 구간은 고정된 Perch_v2 특징 추출기를 통과하여 종별 로짓과 1536차원 임베딩으로 변환된다. 이후 Perch_v2의 출력은 미매핑 종을 보완하기 위한 MLP Probes, 시간적 문맥과 예측 오차를 보정하기 위한 ProtoSSM 및 ResidualSSM, 짧은 음향 이벤트를 탐지하기 위한 Distilled SED의 입력으로 활용된다. 각 모듈에서 생성된 예측값은 먼저 순위 기반 결합을 통해 통합된다. 이후 yukiZ 기반 파이프라인, KarnakbayevPowerOptimization PSSM 파이프라인, KarnakbayevPowerOptimization EoS9 파이프라인이 생성한 예측 파일을 비대칭 가중치로 결합한다. 마지막으로 BirdNET과 같은 보조 모델 및 후처리 기법을 적용하여 예측값을 안정화하고, 각 5초 구간에 대해 234개 생물종의 존재 확률을 담은 submission.csv 형식의 최종 예측 파일을 생성한다.

[그림 3] BirdCLEF+ 2026 다중 레이블 음향 분류 파이프라인의 전체 구조



3.1 ProtoSSM & ResidualSSM

5초 단위로 분할된 오디오 청크를 각각 독립적으로 분석할 경우, 신호가 미약하거나 간헐적으로 발생하는 생물 소리를 놓치기 쉽다. 예를 들어 0초에서 5초 구간에는 매우 희미한 새소리가 포함되고, 5초에서 10초 구간에는 같은 종의 소리가 더 뚜렷하게 나타나며, 10초에서 15초 구간에는 동일한 소리가 반복될 수 있다. 이때 각 구간을 독립적으로 판단하면 첫 번째 구간의 약한 신호를 탐지하지 못할 가능성이 높다. 반면 앞뒤 구간의 정보를 함께 고려하면 같은 생물종의 반복적이고 연속적인 음향 패턴을 파악할 수 있으므로, 미약하거나 간헐적인 소리에 대한 탐지 성능을 향상시킬 수 있다.

이를 위해 본 연구에서는 앞뒤 윈도우의 시간적 맥락을 연결하는 시계열 모델링 기법인 ProtoSSM 및 LightProtoSSM을 도입하였다. ProtoSSM은 Perch 모델에서 추출한 임베딩과 종 점수뿐 아니라, 시간적 위치, 녹음 지점, 시간대 정보

와 같은 메타데이터를 함께 입력으로 사용한다. 또한 클래스 프로토타입 정보를 활용하여 각 종의 대표적인 음향 특징과 현재 구간의 임베딩 사이의 관계를 학습한다. 따라서 ProtoSSM은 개별 구간의 정적인 특징만을 사용하는 모델이 아니라, 연속된 사운드스케이프 안에서 생물 소리가 어떻게 이어지고 반복되는지를 반영하는 시계열 기반 모델로 기능한다.

더불어 ProtoSSM이 도출한 1차 예측 결과의 과대 혹은 과소 추정 오류를 정밀하게 교정하기 위해, 본 연구에서는 2차 예측 보정 모델인 ResidualSSM을 설계하였다. ResidualSSM은 경량화된 BiSSM 기반의 보정 헤드 구조를 사용한다. First-pass 스코어와 Perch 오디오 임베딩을 결합하여 입력하면, BiSSM 레이어와 선형 헤드를 거쳐 오차 보정값을 산출한다. 최종 점수는 식 (1)과 같이 1차 예측값에 보정 가중치 α 를 적용한 잔차 보정값을 더하는 방식으로 미세 조정된다. 여기서 s_{first} 는 ProtoSSM의 1차 예측 점수, Δ 는 ResidualSSM이 산출한 보정값이며, 보정 강도 α 는 [0.10, 0.40] 범위의 그리드 탐색을 통해 검증 성능 기준으로 선택하였다.

$$s_{final} = s_{first} + \alpha \cdot \Delta \quad (\text{식1})$$

3.2 MLP 분류기 (MLP Probes)

2장에서 언급한 바와 같이, 전체 234개 타겟 생물종 중 203종만이 Perch 모델의 로짓과 직접 매핑되며, 나머지 31개 종은 사전 학습 모델만으로는 탐지가 어렵다. 이 31종 중 3종은 약한 proxy 신호를 가지나 28종은 Perch로부터 유효한 신호를 전혀 받지 못한다. 이러한 사각지대를 보완하기 위해, 고밀도 특징 벡터인 Perch 임베딩을 직접 입력으로 받는 독립적인 MLP 분류기(MLP Probes)를 구축하였다.

MLP Probes는 미매핑 종에 국한되지 않고, 학습 데이터 내에 충분한 양성 샘플을 가진 클래스마다 개별적으로 학습된다. 본 실험에서는 총 58개의 클래스별 probe가 학습되었으며, 이는 31개 미매핑 종을 포함해 임베딩 기반의 추가 예측 신호로 보완 가능한 클래스 전체를 포괄한다. 학습 및 추론 효율을 높이기 위해 다수의 클래스별 probe를 단일 텐서 연산으로 동시에 추론하는 벡터화 구현(Vectorized Implementation)을 적용하였다. 이렇게 산출된 probe 신호는 기본 종 점수와 클래스별 가중치로 결합되어, Perch가 직접 다루지 못하는 희귀·미매핑 종에 대한 파이프라인 전체의 커버리지를 보완한다.

3.3 Auxiliary Models & Rank-based Blending

예측의 강건성을 추가로 확보하기 위해 성격이 다른 두 가지 보조 모델을 파이프라인에 사이드카 형태로 결합하였다. 첫째, 스펙트로그램 기반으로 짧고 강렬한 음향 이벤트를 정밀하게 포착하는 'Distilled SED(Sound Event Detection)' 모델을 통해 사운드 이벤트 발생 확률을 생성하고 ProtoSSM의 시계열 정보를 보완하였다. 둘째, 검증된 조류 음향 탐지기인 'BirdNET v2.4'를 활용하여 보수적인 관점의 예측 보정을 수행함으로써 오탐지를 억제하였다.

서로 다른 아키텍처를 가진 모델들의 예측값을 결합할 때, 각 모델이 출력하는 확률 스케일(Calibration)의 차이로 인해 단순 산술 평균을 적용하면 전반적인 성능이 저하될 위험이 있다. 본 연구에서는 이 문제를 해결하기 위해 확률값 대신 각 클래스 내에서의 상대적 순위를 활용하는 'Rank-based Blending' 기법을 제안한다. 이상치와 모델 간 Calibration 차이에 강건한 시스템을 구축하기 위해, 검증 과정을 거쳐 ProtoSSM의 순위 결과와 Distilled SED의 순위 결과를 6:4의 최적 비율로 결합하는 랭크 퓨전(Rank Fusion)을 수행하였다.

3.4 Asymmetric 3-Branch Ensemble

최종 submission.csv를 도출하기 위한 최상단 앙상블 단계에서는 각각 독립적으로 예측 파일을 생성하는 3개의 서브 파이프라인(Branch)을 가중 평균하여 결합하는 전략을 취하였다. 각 브랜치의 구성과 최종 가중치는 [표 1]에서 확인할 수 있다.

본 앙상블 구조는 공개 노트북의 베이스 모델들을 바탕으로 구성되었으며, 메인 브랜치(EoS9)에 0.967의 가중치를 부여하는 비대칭적 형태를 띤다. 이와 같은 비대칭 가중치 설정의 배경은, 단일 성능이 가장 우수한 메인 모델의 내재적 예측 성능을 훼손하지 않으면서, 아주 미량의 보조 가중치(PSSM 0.021, yukiZ 0.012)를 통해 여러 패턴의 다양성(error pattern diversity)만을 주입하기 위함이다. 이러한 비대칭 앙상블 전략은 미지의 평가 데이터(hidden test data)에 대한 모델의 일반화와 안정성을 확보하고자 의도되었다.

[표 1] 최종 예측 생성을 위한 비대칭 3-Branch 앙상블 구성과 Branch별 가중치

Branch	역할	최종 가중치
Karnakbayev_PowerOptimization_EoS9	Main Branch	0.967
yukiZ_Perch_ProtOSSM_ResSSM	Diversity Branch	0.012
Karnakbayev_PowerOptimization_PSSM	Sub Branch	0.021

3.5 후처리 (Post-Processing)

앙상블을 통해 도출된 최종 모델의 출력값을 실제 평가 환경에 맞추어 안정화하고, 불필요한 오탐지를 제어하기 위해 세 가지 주요 후처리 기법을 단계적으로 적용하였다. 우선, 각 생물 분류군이 가지는 고유한 음향적 특성에 맞추어 Temperature 파라미터를 개별 조정하는 스케일링 기법(Per-Taxonomy Temperature Scaling)을 통해 모델이 도출하는 예측 신뢰도(Confidence)의 분포를 정밀하게 교정하였다. 또한, Shift TTA(Test-Time Augmentation) 기법과 앞뒤 윈도우의 예측 정보를 함께 활용하는 스무딩(Temporal & Delta Smoothing) 기법을 적용하여 시계열 데이터가 지닌 시간적 연속성을 끊임 없이 부드럽게 보완하였다. 마지막으로 클래스별로 상이한 임계값(Per-class Threshold)을 최적화하고, 탐지가 어려운 희귀종에 대한 가중치 조정을 수행하는 캘리브레이션(Calibration & Rare-class Adjustment) 과정을 거침으로써 최종 확률값을 안정화하고 오탐지율을 효과적으로 제어하였다.

4. 실험 및 결과 분석

4.1 앙상블 파이프라인 성능 분석

본 연구에서는 BirdCLEF+ 2026 대회의 다중 라벨 분류 문제를 해결하기 위해 Perch 기반의 특징 추출기와 ProtoSSM, ResidualSSM, MLP Probes, Distilled SED, BirdNET 등을 결합한 앙상블 파이프라인을 구축하였다. 최종 성능 평가는 대회에서 제공한 평가 지표인 macro ROC-AUC를 기준으로 수행하였

다.

핵심 모델인 ProtoSSM은 Perch 모델로부터 추출된 임베딩 정보와 메타데이터를 활용하여 시계열적인 문맥 정보를 학습하며, Distilled SED 모델은 짧은 음향 이벤트를 탐지하는 역할을 수행하였다. 두 모델의 출력값은 확률값 자체가 아닌 각 클래스의 순위(rank)를 이용하여 6:4 비율로 결합함으로써 모델 간 calibration 차이에 따른 영향을 줄이고 안정적인 예측 성능을 확보하였다. 또한 ResidualSSM을 통해 1차 예측 결과의 오차를 보정하고, MLP Probes를 이용하여 Perch 모델과 직접 연결되지 않는 31개의 미매핑 종에 대한 분류 성능을 향상시켰다. [표 2]는 주요 파이프라인 구성 요소의 역할과 구성을 정리한 것이다.

[표 2] 제안 파이프라인의 주요 구성 요소와 성능 기여 요약

Pipeline Component	Key Metric/Parameter	역할/구성
ProtoSSM + SED Rank Blend	6:4 rank fusion	calibration 차이에 강건한 시계열 및 이벤트 결합
MLP Probes	31 Unmapped Species (58 probes)	Perch 로짓 미연결 종 커버리지 보완
ResidualSSM	잔차 보정	1차 예측의 추정 보정
3-Branch Ensemble	0.967 / 0.021 / 0.012	메인 예측 유지 및 에러 다양성 주입

4.2 성능 평가 결과

실험 결과 Public Leaderboard에서는 macro ROC-AUC 0.95095를 기록하여 전체 4,093개 팀 중 362위를 달성하였으며 상위 8.8%에 해당하는 성능을 보였다. 최종 평가가 이루어지는 Private Leaderboard에서는 macro ROC-AUC 0.94141을 기록하여 788위를 차지하였으며 상위 19.3%의 성적을 달성하였다. 그리고 저자들이 만든 파이프라인의 핵심 앵커로 사용한 EoS9 공개 노트북은 Kaggle 커뮤니티에서 118표의 추천을 받아 Silver 노트북 메달(Notebook Medal)을 수상하였다. 이는 대회 순위에 따른 메달이 아니라 노트북의 공개 기여도에 대한 커뮤니티 평가이며, 본 연구는 해당 공개 노트북을 주된 기반으로 삼아 앙상블 구성과 후처리 전략을 설계하였다.

4.3 결과 해석 및 분석

본 연구의 높은 성능은 단일 모델이 아닌 다양한 모델들의 상호 보완적인 결합에 의해 달성된 것으로 볼 수 있다. ProtoSSM은 시계열 정보를 학습하여 장기적인 음향 패턴을 반영하였으며, Distilled SED는 짧은 음향 이벤트를 탐지하여 순간적인 신호를 포착하였다. 또한 MLP Probes는 Perch 모델과 직접 연결되지 않는 미매핑 종의 분류를 보완하였고, ResidualSSM은 초기 예측의 잔차를 보정함으로써 최종 예측 정확도를 향상시켰다.

특히 확률값을 직접 평균하지 않고 rank-based blending 기법을 적용함으로써 각 모델의 calibration 차이에 의한 영향을 줄일 수 있었으며, 이를 통해 안정적인 예측 성능을 확보할 수 있었다. 이는 서로 다른 특성을 가진 모델들을 효과적으로 결합한 앙상블 구조가 생태 음향 데이터와 같은 복잡한 다중 라벨 분류 문제에서 높은 성능을 달성할 수 있음을 보여준다.

다만 OOF 기준 macro-AUC(0.992)와 Private 점수(0.94141) 사이에는 약 0.05의 격차가 존재한다. 또한 Public 362위(상위 8.8%)에서 Private 788위(상위 19.3%)로의 순위 하락은, OOF 검증 기준으로 튜닝한 후처리 및 가중치 설정이 미지의 평가 데이터에 완전히 일반화되지 못했을 가능성을 시사한다. 이는 본 파이프라인의 안정성을 과대평가하지 않기 위해 함께 고려되어야 할 지점이며, OOF 최적화에 의존하지 않는 일반화 전략의 개선 여지를 보여준다.

5. 결론

5.1 연구 요약

본 연구에서는 BirdCLEF+ 2026 대회의 다중 라벨 분류 문제를 해결하기 위하여 Perch 기반 특징 추출기와 ProtoSSM, ResidualSSM, MLP Probes, Distilled SED 및 BirdNET을 결합한 앙상블 구조를 구축하였다. 또한 rank-based blending 기법을 적용하여 모델 간 calibration 차이에 따른 영향을 최소화하고 안정적인 예측 성능을 확보하였다.

실험 결과 Public Leaderboard에서 0.95095, Private Leaderboard에서 0.94141의 macro ROC-AUC를 기록하여 전체 4,093개 팀 중 상위 19.3%(788위)의 성적을 달성하였다. 본 파이프라인이 기반으로 삼은 EoS9 공개 노트북은 커뮤니티로부터 Silver 노트북 메달을 받았다. 이러한 결과는 시계열 정보와 음향 이벤트 정보를 결합한 앙상블 구조가 복잡한 생태 음향 데이터 환경에서도 경쟁력 있는 성능을 달성할 수 있음을 보여준다.

5.2 한계점 및 향후 연구

본 연구는 높은 분류 성능을 달성하였으나, 희귀 종에 대한 예측 성능과 계산 효율 측면에서는 여전히 개선의 여지가 존재한다. 또한 여러 모델을 결합한 복잡한 앙상블 구조는 높은 계산 비용과 추론 시간을 요구한다는 한계를 가진다.

향후에는 보다 다양한 음향 이벤트 탐지 모델을 추가적으로 활용하고, 희귀 종에 대한 예측 성능을 향상시키기 위한 연구가 필요할 것으로 기대된다. 또한 모델 경량화와 효율적인 앙상블 구조 설계를 통해 실제 환경에서의 활용 가능성을 더욱 높일 수 있을 것으로 기대된다.

참고문헌

- Ghani, B., Denton, T., Kahl, S. and Klinck, H. (2023). Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, Vol. 13, 22876.
- Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Kahl, S., Denton, T., Sugai, L., Piatti, L., Holbrook, R., Klinck, H. and Oldacre, A. (2026). BirdCLEF+ 2026. *Kaggle*.
<https://www.kaggle.com/competitions/birdclef-2026>
- Kahl, S., Wood, C. M., Eibl, M. and Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, Vol. 61, 101236.
- Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105-6114.
<https://www.kaggle.com/competitions/birdclef-2026>

Wildfire Risk Assessment near Power Facilities Based on Meteorological Data and Spatial Information

김도연¹⁾, 송윤섭¹⁾, 이은승¹⁾, 이준구²⁾, 이진석¹⁾

Abstract

This study developed a wildfire risk assessment framework for areas surrounding power facilities in Gangwon Province, South Korea. The study area was divided into 1 km × 1 km grid cells, and daily meteorological conditions, forest characteristics, power-facility density, and historical wildfire records were integrated at the grid-day level. The input variables included temperature, humidity, precipitation, consecutive dry days, wind speed, forest composition, facility density, and seasonal indicators.

To address the highly imbalanced distribution of wildfire events, all positive observations were retained, while negative observations were sampled from both wildfire occurrence dates and the overall training period. Random Forest, LightGBM, and XGBoost models were trained, and their predictions were converted into relative risk scores ranging from 0 to 100 based on the validation prediction distributions. The three model scores were combined using weights determined from validation Recall at the top 1%, 5%, and 10% risk ranges. Daily risk scores were calculated for each grid during the prediction period, and the mean of the seven highest daily scores was defined as the final grid-level wildfire risk score. Grids with scores of 95 or higher were classified as high-risk grids. Evaluation of actual wildfire

1) All authors have equal contribution, Department of Smart Cities, Chung-Ang University, Seoul 06974, Korea

2) Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea

cases during the test period showed that two of the three wildfire grids were ranked within the top 200 priority grids on their occurrence dates. The proposed framework can support preventive inspection, patrol planning, and risk-based allocation of monitoring resources around power infrastructure.

1. 서론

1.1 연구 배경 및 필요성

산불은 기온, 습도, 강수량, 풍속과 같은 단기적인 기상 조건뿐만 아니라 산림의 구성, 지형적 특성, 인간 활동 및 과거 화재 발생 이력 등이 복합적으로 작용하여 발생한다. 특히 낮은 습도와 장기간의 무강수 상태는 산림 연료의 수분 함량을 감소시키며, 높은 기온과 강한 바람은 화재의 발화 및 확산 가능성을 증가시킬 수 있다. 따라서 산불 위험도를 효과적으로 평가하기 위해서는 특정 시점의 기상 상태뿐만 아니라 누적된 건조 조건과 지역별 공간 특성을 함께 고려해야 한다.

강원특별자치도는 산림 면적의 비중이 높고 산악 지형이 넓게 분포하며, 영동과 영서 지역 간 기상 특성의 차이가 크다. 특히 영동 지역에서는 봄철 건조한 대기와 강풍이 동시에 발생하는 경우가 많아 대형 산불의 위험이 높다. 또한 산림지역을 통과하거나 인접하여 설치된 전력설비는 강풍, 수목 접촉, 설비 이상 등 다양한 요인에 의해 화재 위험과 관련될 수 있다. 이에 따라 전력설비 주변 지역의 기상 및 공간적 특성을 고려하여 화재 위험이 상대적으로 높은 지역을 사전에 식별하는 것은 예방적 설비 점검과 재난 대응 측면에서 중요하다.

기존의 산불위험도 연구에서는 기온, 습도, 풍속, 강수량 등을 이용한 통계적 위험지수와 로지스틱 회귀모형이 주로 활용되었다. 최근에는 변수 간 비선형 관계와 복잡한 상호작용을 학습할 수 있는 Random Forest, XGBoost 및 LightGBM 등의 머신러닝 기법이 산불 발생 예측에 적용되고 있다. 그러나 실제 산불은 전체 관측자료 중 극히 일부에서만 발생하는 사건이므로, 일반적인 정확도만으로 모델을 평가할 경우 대부분의 관측치를 비발생으로 예측하는 모델도 높은 성능을 보이는 것처럼 해석될 수 있다.

본 연구에서는 산불 발생확률 자체를 정확하게 추정하는 것보다, 제한된 점검 및 예방 자원을 화재 위험이 높은 지역에 우선 배분할 수 있도록 상대적인 위험 순위를 산출하는 데 초점을 두었다. 이를 위해 강원특별자치도의 전력설비 주변

지역을 1 km × 1 km 격자로 구분하고, 격자별 기상자료, 산림 특성, 전력설비 분포 및 과거 산불 발생 이력을 결합하였다. 이후 여러 머신러닝 모델의 예측 결과를 비교하고, 상위 위험구간에서 실제 산불을 얼마나 효과적으로 포착하는지를 중심으로 모델을 평가하였다.

1.2 문제 정의

본 연구의 핵심은 강원도 내 가상 전력설비 위치를 1 km 격자로 집계하고, 각 격자에 일별 기상조건과 산림 특성을 결합하여 화재발생 가능성이 상대적으로 높은 격자를 선별하는 것이다. 분석 결과는 화재발생을 확정적으로 단정하는 값이 아니라, 한정된 점검·감시 자원을 어디에 우선 배치할지를 지원하는 위험 우선순위 지표로 정의한다.

1.3 분석 목표

- 전력설비, 기상, 산림 및 과거 산불 이력을 동일한 1 km 격자-일자 단위로 통합한다.
- 기온·습도·강수·풍속의 단기 기상위험과 산림면적·수종·전주 밀집도의 공간 취약도를 함께 반영한다.
- 산불 발생자료가 비발생 자료보다 매우 적은 불균형 구조를 고려하여, 단순 정확도보다 고위험 후보군 내 실제 산불 포착률을 중심으로 모델을 설계한다.
- 여러 모델의 예측값을 0-100점으로 표준화하여 일자별·지역별 위험 우선순위를 제시한다.
- 현장점검, 순찰, 드론 감시 및 예방정비에 활용할 위험도 지도와 후보 목록을 구축한다.

2. 분석 데이터 및 전처리

2.1 공간·시간 기준 통일

서로 다른 형식의 자료를 결합하기 위해 분석 공간을 강원도 전역의 1 km × 1 km 격자로 통일하였다. 전력설비와 산불 발생 위치는 점 자료로 변환하여 해당 격자에 매칭하고, 임상도는 격자와 중첩하여 면적 비율을 계산하였다. 기상자

료는 일 단위로 집계한 뒤 격자자료와 날짜를 기준으로 결합하였다. 최종 데이터는 1km 격자별·날짜별로 구성하였으며, 각 행은 특정 날짜의 특정 격자에 대한 기상 조건, 산림 특성, 전력설비 분포 및 산불 발생 여부를 나타내었다. 분석기간은 2023년부터 2026년까지 3년간의 11월(초겨울) ~ 익년 5월(늦봄)로 제한하였고 예측 시점 이후의 정보가 입력변수에 포함되는 시간 누수를 방지하기 위해 기상변수는 전일 관측값과 과거 누적값을 사용하였다.

2.2 활용 데이터

데이터는 주최측에서 제공한 전력설비 데이터, 기상청에서 제공한 기상 데이터, 산림청에서 제공한 산림데이터와 산불 이력 데이터를 이용해 변수를 만들었다.

[표 1] 데이터에서 생성한 변수들

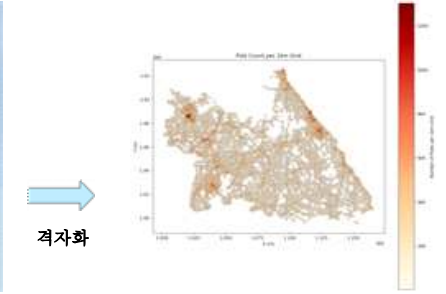
데이터 구분	최종 생성 변수	모델에서의 역할
전력설비	격자내_전주수 $\log(1 + \text{전주수})$	- 전력설비 밀집도를 나타내는 공간 취약도 변수 - 모델에는 $\log(1 + \text{전주수})$를 입력
기상	전일 평균기온 전일 최고기온 전일 평균습도 전일 최저습도 전일 강수량 최근 3일 누적강수량 최근 7일 누적강수량 무강수 지속일수 전일 평균풍속 전일 최대순간풍속	- 고온·건조·강풍 및 최근 강수 부족 등 날씨별 기상 위험을 반영 - 총 10개 입력변수
산림	산림면적비율 침엽수비율 활엽수비율	- 격자별 산림 분포와 수종 구성을 나타내는 고정 공간 특성 - 총 3개 입력변수
산불 이력	산불발생_V1 산불발생_V2 산불발생_V3	- V1: 산불 발생 당일 - V2: 발생일을 포함한 이전 2일 - V3: 발생일을 포함한 이전 7일 - 입력변수가 아닌 정답변수(Label)
계절성	doy_sin doy_cos	11월~5월 사이 반복되는 계절적 위험 패턴을 연속적으로 반영

2.3 전력설비 데이터

[그림 1] 약 138만 건의 전력 설비 위치 데이터



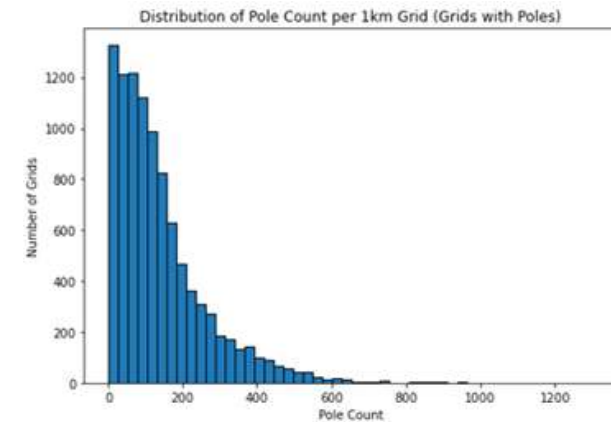
[그림 2] 전력 설비 위치 데이터를 1km 격자로 집계



약 138만 건의 전력설비 위치를 1 km 격자에 공간 집계하여 격자별 전주수를 산출하였다. 총 21,945개 격자 중 9,873개 격자에서 전력설비가 확인되었다.

전주수는 대부분의 격자에서 낮은 값을 보이지만 일부 격자에서 매우 큰 값이 나타나는 오른쪽으로 치우친 분포를 보였다. 이에 분포의 치우침과 큰 값의 영향을 줄이기 위해 모델 학습에는 $\log(1 + \text{전주수})$ 로 변환한 값을 사용하였다.

[그림 3] 격자별 전주수 분포



2.4 기상 데이터

기상자료는 기상청에서 제공하는 방재기상관측자료(AWS)와 종관기상관측자료(ASOS)를 활용하였다. 방재기상관측(AWS) 데이터로 국지적인 강수·풍속·습도 변화를 파악하고, 종관기상관측(ASOS) 데이터로 광역적·장기적 기상 특성을 보완하는 데 활용하였다. 산불과 직접 관련된 기온, 습도, 강수, 풍속을 중심으로 총 10개 변수를 구성하였다.

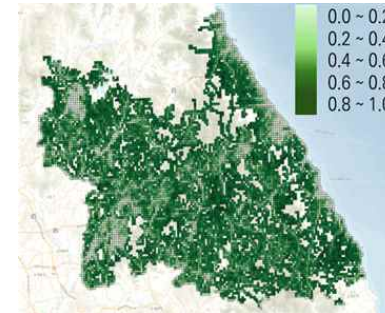
[표 2] 기상데이터 변수

변수군	세부 변수	의미
기온	전일 평균기온, 전일 최고기온	고온과 연료 건조 조건
습도	전일 평균습도, 전일 최저습도	대기·연료의 건조 정도
강수	전일·최근 3일·최근 7일 강수, 무강수 지속일수	수분 공급과 건조 지속성
풍속	전일 평균풍속, 최대순간풍속	불씨 이동과 화재 확산 가능성

2.5 산림 데이터

산림자료는 산림청에서 제공하는 임상도를 활용하였으며, 격자별 산림면적과 수종 구성을 산출하였다. 임상도와 1 km 격자를 중첩하여 격자 내 산림면적을 계산하고, 이를 전체 격자면적으로 나누어 산림면적비율을 산출하였다. 이후 임상도의 수종 정보를 이용해 침엽수림과 활엽수림 면적을 구분하였다. 침엽수림은 수지 함량과 건조한 낙엽층의 영향으로 상대적으로 연소·확산 가능성이 높고, 활엽수림은 수분 함량이 높아 화재 저항성이 상대적으로 크다는 점을 고려하여 두 수종의 비율을 별도 변수로 구성하였다.

[그림 4] 산림 면적 비율 지도



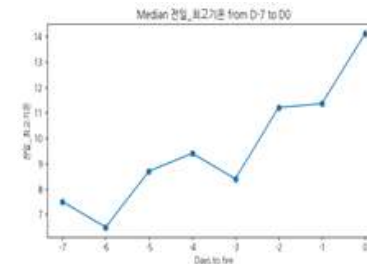
[그림 5] 침엽수, 활엽수 비율 지도



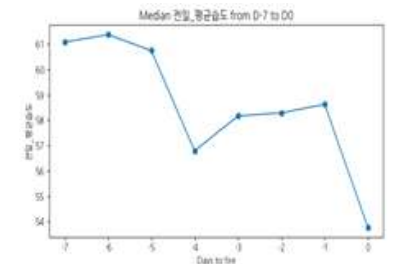
2.6 산불 발생 이력 데이터

산불 발생 이력도 산림청에서 제공한 데이터를 활용하였다. 발생 위치가 주소 형식으로 제공되어 주소를 지오코딩하여 위·경도 좌표로 변환한 뒤 1 km 격자에 매칭하였다.

[그림 6] 산불 발생 일주일 전 최고 기온 중앙값 그래프



[그림 7] 산불 발생 일주일 전 평균 습도 중앙값 그래프



산불 발생 이전 7일간의 기상 변화를 확인한 결과, 발생일에 가까워질수록 전일 평균기온의 중앙값은 전반적으로 상승하였으며, 전일 평균습도의 중앙값은 발생일에 가장 낮게 나타났다. 발생 당일의 기상조건뿐만 아니라 발생 이전부터 누적된 고온·건조 등의 조건에 영향을 받는 것이므로 이를 반영하기 위해 산불 발생 정답변수를 다음과 같이 세 가지 형태로 구성하였다.

[표 3] 산불 데이터 변수

라벨	정의	활용 목적	전체 양성 Grid-day
V1	실제 산불 발생 당일	발생일 판별	36건
V2	발생일 포함 이전 2일	단기 사전위험 탐지	108건
V3	발생일 포함 이전 7일	누적위험 구간·조기경보	288건

* 양성 Grid-day = 산불 발생 라벨이 1인 격자·날짜 자료

2.7 계절성 변수

산불 위험은 연중 일정하지 않으며 겨울철과 봄철에 따라 기상조건과 산림 연료의 건조 상태가 달라진다.

날짜를 단순한 정수로 사용할 경우 12월 31일과 1월 1일이 서로 멀리 떨어진 값으로 표현되는 문제가 있다. 이를 보완하기 위해 연중 날짜를 사인과 코사인 함수로 변환하였다.

$$Season\ Sin_t = \sin\left(\frac{2\pi d_t}{365}\right) \quad (\text{식1})$$

$$Season\ Cos_t = \cos\left(\frac{2\pi d_t}{365}\right) \quad (\text{식2})$$

여기서 d_t 는 날짜 t 의 연중 순번이다.

두 변수를 함께 사용하면 연말과 연초가 서로 인접한 계절로 표현된다.

3. 분석방법

전체 분석은 개별 머신러닝 모델 학습, 검증자료 기반 모델 평가, 모델별 위험 점수 변환, 가중 앙상블 및 최종 위험등급 산정 순서로 진행하였다.

3.1 데이터 분할

무작위 분할은 미래 시점의 기상패턴이 과거 학습자료에 포함되는 시간 누수를 유발할 수 있으므로, 자료를 산불 시점 단위로 구분한 후 시간 순서에 따라 학습 (Train)자료, 검증(Validation)자료 및 시험(Test)자료로 분할하였다.

[표 4] 데이터 분할

구분	기간	Grid-day 수	목적
Train	2023.11 -2024.05	1,737,910	모델 학습
Validation	2024.11 -2025.05	1,900,783	모델 비교·점수변환·가중치 설정
Test	2025.11 -2026.05	1,948,896	Grid별 최종 위험점수 산정

3.2 라벨 스무싱

산불발생 v3를 단순 0/1로만 학습하면 발생 직전 며칠의 위험 변화가 동일한 hard label로 처리되거나 충분히 반영되지 않을 수 있으므로 실제 산불발생 날짜를 기준으로 D-7부터 D0까지 점진적으로 증가하는 soft label을 학습기간에만 생성해준다. 계산식은 다음과 같다.

$$SoftLabel(d) = 0.20 + (1 - 0.20) * (7 - d) / 7 \quad (\text{식3})$$

$$d = \text{산불발생일까지 남은 일수} \quad (0 \leq d \leq 7) \quad (\text{식4})$$

[표 5] 산불발생 날짜까지 7일 동안의 soft label 계산

시점	days_to_fire (d)	v3_soft_label
D-7	7	0.200000
D-6	6	0.314286
D-5	5	0.428571
D-4	4	0.542857
D-3	3	0.657143
D-2	2	0.771429
D-1	1	0.885714
D0	0	1.000000

3.3 사용 모델

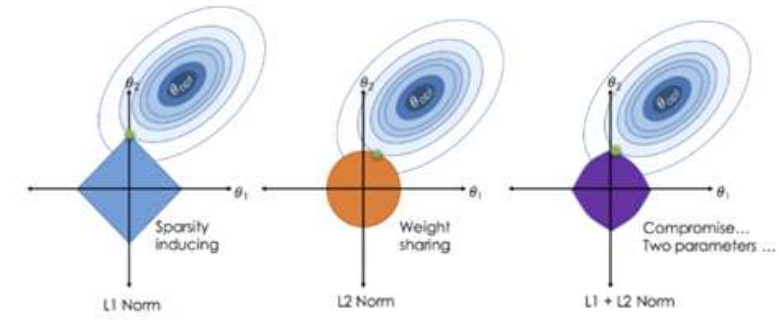
[표 6] 분석에 사용한 모델

모델	역할
Elastic Net	선형 기준모형. L1·L2 정규화로 변수 방향과 불필요 변수 축소 확인
Random Forest	여러 나무를 평균하여 비선형 관계를 안정적으로 반영하고 변수 중요도 산출
LightGBM	Gradient Boosting으로 복잡한 비선형성과 변수 상호작용을 효율적으로 학습
XGBoost	규제항을 포함한 Boosting으로 과적합을 제어하고 안정성 비교
ML Ensemble	RF·LGBM·XGB 점수를 검증 성능에 따라 가중 결합

Elastic Net 로지스틱 회귀는 입력변수의 선형적 영향과 기준 성능을 확인하기 위한 비교모형으로 사용하였다. 최종 위험점수는 비선형 관계와 변수 간 상호작용을 반영할 수 있는 Random Forest, LightGBM 및 XGBoost의 예측결과를 가중 결합하여 산정하였으며, Elastic Net은 최종 앙상블에는 포함하지 않았다.

3.3.1 Elastic Net

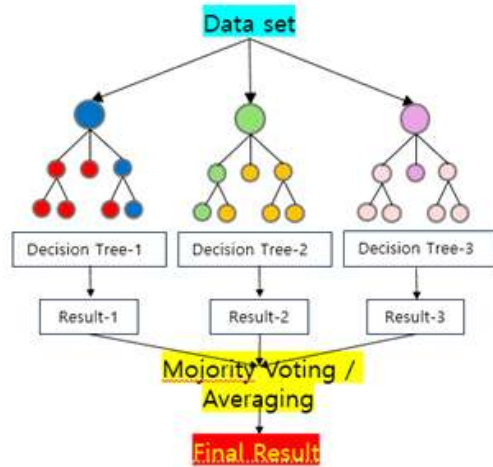
[그림 8] Elastic Net의 분석구조



Elastic Net은 로지스틱 회귀에 L1 규제와 L2 규제를 함께 적용한 선형 분류 모형이다. 기온과 최고기온, 강수량과 누적 강수량처럼 서로 높은 상관관계를 가질 수 있는 변수가 동시에 포함된 경우에도 회귀계수가 지나치게 커지는 것을 방지할 수 있다. 또한 예측에 대한 각 변수의 영향을 회귀계수로 확인할 수 있어 다른 머신러닝 모델보다 결과를 비교적 쉽게 해석할 수 있다. 본 연구에서는 복잡한 머신러닝 모델의 성능을 비교하기 위한 기준모형으로 Elastic Net을 사용하였다.

3.3.2 Random Forest

[그림 9] RandomForest의 분석구조



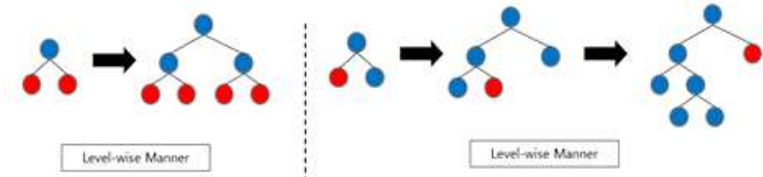
Random Forest는 여러 개의 의사결정나무를 독립적으로 학습한 후 각 나무의 예측결과를 결합하는 앙상블 모델이다. 각 의사결정나무는 원래 학습자료에서 무작위로 추출한 자료와 설명변수의 일부를 이용하여 학습한다. 나무마다 서로 다른 자료와 변수를 사용하기 때문에 각 나무는 조금씩 다른 분류기준을 생성한다.

분류문제에서는 여러 나무의 예측결과를 종합하여 최종 예측을 수행한다. 개별 의사결정나무는 학습자료에 지나치게 적합될 가능성이 있지만, 다수의 나무를 결합하면 개별 나무의 오류가 평균화되어 보다 안정적인 예측결과를 얻을 수 있다.

Random Forest는 변수 사이의 선형관계를 가정하지 않는다. 따라서 낮은 습도와 강한 풍속이 동시에 나타나는 경우나, 강수량이 적고 침엽수림비율이 높은 경우처럼 여러 변수가 결합되어 나타나는 조건을 학습할 수 있다. 또한 모델이 예측에 사용한 변수의 중요도를 산출할 수 있어 산불 위험도 예측에서 어떤 기상 및 공간변수가 중요한 역할을 했는지 확인할 수 있다.

3.3.3 LightGBM

[그림 10] LightGBM의 분석구조

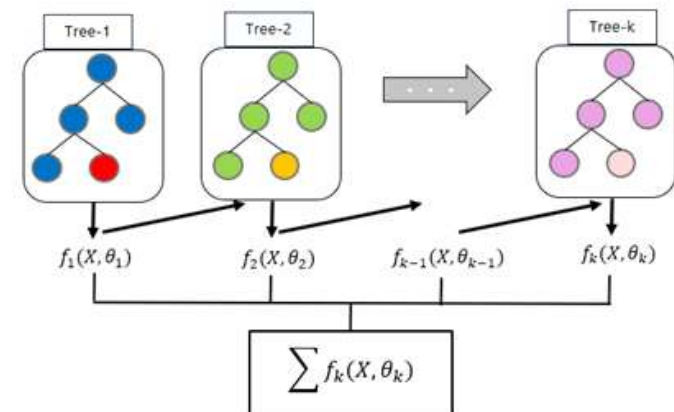


LightGBM은 여러 의사결정나무를 순차적으로 학습하면서 이전 단계에서 발생한 예측오차를 보완하는 부스팅 기반 모델이다.

LightGBM은 연속형 변수값을 일정한 구간으로 나누어 학습하는 히스토그램 기반 방식을 사용하므로 대규모 자료를 비교적 빠르게 처리할 수 있다. 본 연구 자료는 격자와 날짜의 조합으로 구성되어 관측치 수가 매우 많기 때문에, 계산 효율성이 높은 LightGBM을 분석모델에 포함하였다.

3.3.4 XGBoost

[그림 11] XGBoost의 분석구조



XGBoost 역시 여러 의사결정나무를 순차적으로 결합하는 Gradient Boosting 계열의 모델이다.

XGBoost는 이전 나무의 예측결과와 실제값의 차이를 계산하고, 다음 나무가 그 차이를 줄이는 방향으로 학습한다. 초기 모델의 예측이 부정확하더라도 새로운 나무가 반복적으로 추가되면서 전체 예측오차가 점차 감소한다.

LightGBM과 기본적인 부스팅 원리는 유사하지만, XGBoost는 모델의 복잡성을 제한하는 규제기능을 포함한다. 나무의 깊이, 말단노드 수 및 각 노드의 예측값이 지나치게 복잡해지는 것을 제한하여 학습자료에 과도하게 적합되는 문제를 줄인다.

또한 각 나무를 학습할 때 전체 관측치와 변수를 모두 사용하지 않고 일부만 선택하여 학습할 수 있다. 이를 통해 모델 간 다양성을 확보하고 과적합을 완화할 수 있다.

XGBoost는 기온, 습도, 강수량, 풍속 등의 기상변수와 산립비율, 전력설비 수 및 과거 산불 이력 사이의 비선형 관계를 반영할 수 있다. 또한 비교적 안정적인 예측성능을 보여 LightGBM 및 Random Forest와의 성능 비교를 위해 사용하였다.

3.4 불균형 자료 학습 및 평가 전략

본 분석자료는 산불 발생 자료보다 비발생 자료가 압도적으로 많은 불균형 구조를 가진다. 전체 자료를 그대로 학습할 경우 모델이 다수를 차지하는 비발생 자료에 편향되어, 실제 산불 위험이 있는 격자를 충분히 구분하지 못할 수 있다.

이를 보완하기 위해 학습자료에는 산불 발생 라벨이 1인 자료를 모두 포함하였다. 비발생 자료는 산불 발생일과 동일한 날짜의 비발생 격자와 전체 학습기간에서 무작위로 추출한 비발생 격자로 구성하였다. 동일 날짜의 비발생 격자를 포함함으로써 기상조건이 유사한 상황에서 산불 발생 격자와 비발생 격자의 산립 및 전력설비 특성을 비교해 학습하도록 하였다. 또한 무작위로 추출한 비발생 자료를 추가하여 일반적인 비발생 환경도 함께 반영하였다.

이러한 표본추출은 모델 학습단계에만 적용하였다. 검증자료와 시험자료는 표본을 축소하지 않고 해당 기간의 전체 격자·날짜 자료를 사용하여, 실제 운영환경과 동일한 불균형 조건에서 모델의 성능을 평가하였다.

산불은 발생 빈도가 매우 낮으므로 전체 정확도만으로는 모델의 실용성을 판단하기 어렵다. 이에 ROC-AUC와 함께 Top 1%, Top 5%, Top 10% Recall,

Precision 및 Lift를 평가 지표로 사용하였다. Top-k Recall은 위험도가 높은 상위 k% 후보군에 실제 산불 발생 자료가 얼마나 포함되는지를 나타내며, Precision은 해당 후보군 중 실제 산불 발생 자료의 비율을 의미한다. Lift는 상위 후보군의 산불 발생 비율이 전체 자료의 평균 발생 비율보다 몇 배 높은지를 나타낸다.

평가는 전체 시험기간의 격자·날짜 자료를 위험도 순으로 정렬하는 방식과, 날짜별로 모든 격자의 위험순위를 산출하는 방식으로 수행하였다. 전체 기간 평가는 모델의 전반적인 산불 발생 구분능력을 확인하기 위한 것이며, 날짜별 평가는 특정 날짜에 점검해야 할 고위험 격자를 얼마나 효과적으로 선별하는지를 확인하기 위한 것이다.

3.5 모델별 위험점수 및 최종 위험점수 산정

본 연구에서는 Random Forest, LightGBM 및 XGBoost가 산출한 예측값을 동일한 척도의 위험점수로 변환한 후, Validation 자료에서 확인된 모델별 성능을 반영하여 최종 머신러닝 위험점수를 산정하였다.

모델마다 예측값의 범위와 분포가 다르므로 원래 예측값을 직접 비교하거나 단순 평균하는 것은 적절하지 않다. 이에 따라 각 모델의 예측값이 Validation 자료의 예측분포에서 어느 위치에 해당하는지를 계산하고, 이를 0점에서 100점 사이의 상대적 위험점수로 변환하였다.

3.5.1 모델별 위험점수 산정

모델 별 위험점수를 다음과 같이 정의하였다.

$$S_m = 100 \times F_{(m, val)}(f_m(x)) \quad (\text{식5})$$

여기서 각 기호의 의미는 다음과 같다.

- m: 모델의 종류로 Random Forest, LightGBM 또는 XGBoost
- x: 기상, 산립, 전주 수 및 계절성 변수를 포함한 입력자료
- $f_m(x)$: 모델 m이 입력자료 x에 대해 산출한 위험 예측값
- $F_{m, val}$: Validation 자료에서 모델 m의 예측값으로 구성한 누적분포함수
- $S_m(x)$: Validation 예측분포를 기준으로 0~100점으로 변환한 모델별 위험

점수

즉, 모델에 입력변수 x 를 적용하여 예측값 $f_m(x)$ 를 산출한 후, 해당 예측값이 Validation 예측분포에서 몇 번째 백분위에 위치하는지를 계산한다. 그 결과에 100을 곱하여 모델별 위험점수를 산출하였다.

위험 점수에 따라 다음과 같이 고위험 후보군을 분류하였다.

[표 7] 고위험 후보군 분류 기준

위험점수 기준	Validation 기준	하루평균 후보 격자 수	활용 의미
90점 이상	상위 10%	약 920개/일	넓은 조기경보 후보
95점 이상	상위 5%	약 460개/일	고위험 점검 후보
99점 이상	상위 1%	약 92개/일	최우선 점검 후보

3.5.2 모델별 성능점수 산정

세 모델의 양상별 가중치는 Validation 자료에서 계산한 Recall@Top 1%, Recall@Top 5% 및 Recall@Top 10%를 이용하여 결정하였다.

모델 m 의 성능점수 W_m 은 다음과 같이 계산하였다.

$$W_m = 0.25 \times \text{Recall@Top1\%} + 0.50 \times \text{Recall@Top5\%} + 0.25 \times \text{Recall@Top10\%} \quad (\text{식6})$$

상위 5% 구간은 실제 점검 가능한 후보 규모와 산불 포착능력 사이의 균형을 나타내는 주요 관리범위이므로 0.50의 가장 높은 비중을 부여하였다. 상위 1%와 상위 10%에는 각각 0.25의 비중을 부여하여 최우선 점검범위와 넓은 조기경보 범위에서의 성능도 함께 반영하였다.

모델별 Validation 성능과 이를 이용해 계산한 성능점수는 다음과 같다.

[표 8] 모델별 성능 점수

모델	Recall@Top1%	Hit@Top1%	Recall@Top5%	Hit@Top5%
RF	0.1436	26	0.4033	73
LGBM	0.1160	21	0.3481	63
XGB	0.0994	18	0.1934	35

모델	Recall@Top10%	Hit@Top10%	복합 Recall 점수	Ensemble Weight
RF	0.5470	99	0.3743	0.4051
LGBM	0.4475	81	0.3149	0.3408
XGB	0.4530	82	0.2348	0.2541

3.5.3 모델별 가중치 산정

세 모델의 성능점수 합이 1이 되도록 다음과 같이 정규화하여 최종 가중치를 산정하였다.

$$w_m = \frac{W_m}{W_{RF} + W_{LGBM} + W_{XGB}} \quad (\text{식7})$$

세 모델의 성능점수 합은 다음과 같다.

3.5.4 최종 위험점수 산정

최종 머신러닝 위험점수는 Random Forest, LightGBM 및 XGBoost의 모델별 위험점수에 각각의 가중치를 곱한 후 합산하여 계산하였다.

$$RiskScore = w_{RF}S_{RF} + w_{LGBM}S_{LGBM} + w_{XGB}S_{XGB} \quad (\text{식8})$$

실제 산정된 모델별 가중치를 대입하면 최종 위험점수는 다음과 같다.

$$RiskScore = 0.4051S_{RF} + 0.3408S_{LGBM} + 0.2541S_{XGB} \quad (\text{식9})$$

- $S_{RF}(x)$: Random Forest의 0~100점 위험점수
- $S_{LGBM}(x)$: LightGBM의 0~100점 위험점수
- $S_{XGB}(x)$: XGBoost의 0~100점 위험점수

3.6 Grid별 위험 점수 계산

예측기간의 모든 Grid-Day에 대해 위험 점수를 계산한 뒤, 같은 grid에 속한 날짜별 점수 중 가장 높은 7개를 선택해 평균낸 점수를 grid의 최종 위험 점수로 정의한다. 관측일이 7일보다 적은 grid는 존재하는 모든 날짜를 사용한다. 위험 등급은 다음과 같이 설정했다.

[표 9] 위험등급 기준

grid_fire_score	risk_level
< 90	일반
90 이상, 95 미만	관심
95 이상, 99 미만	주의
99 이상	위험

최종 0/1 산불 위험 예측은 grid_fire_score가 95점 이상인 경우를 1, 95점 미만인 경우를 0으로 정의하였다. 따라서 최종 제출값은 각 grid에 대해 주의 또는 위험 등급이면 산불 위험 grid로 예측하고, 일반 또는 관심 등급이면 비위험 grid로 예측한 결과이다.

학습자료에는 9,873개의 grid가 포함되었으나, 최종 위험도 산출은 예측기간에 실제 관측자료가 존재하는 9,194개 grid를 대상으로 수행하였다. 예측기간 입력 자료가 없는 grid는 모델이 판단할 설명변수가 없으므로 기본 위험도 0, 등급 '일반'으로 처리하였다.

4. 결과 및 검증

4.1 실제 산불 발생일 검증

Test 기간에 실제 산불이 발생한 3개 격자를 대상으로, 발생 당일 강원도 전체 9,194개 격자를 위험점수에 따라 정렬하고 각 모델이 산불 발생 격자에 부여

한 점수와 일별 위험순위를 비교하였다. 위험점수는 검증자료의 예측분포를 기준으로 0~100점으로 표준화한 상대적 위험 수준이며, 순위값이 작고 상위에 위치할수록 현장점검 및 순찰 대상에 우선적으로 포함될 가능성이 크다.

분석 결과, 2025년 12월 29일 grid 15519는 모든 모델에서 99점 이상의 매우 높은 위험점수를 기록하였다. 해당 격자는 RF 8위, LightGBM 6위, XGBoost 11위, ML Ensemble 8위로 평가되어, 대부분의 모델에서 전체 9,194개 격자 중 상위 10개 수준의 최우선 위험지역으로 탐지되었다. 특히 ML Ensemble은 개별 모델의 예측결과를 종합하면서도 99.69점과 8위의 높은 순위를 유지하여, 실제 산불 발생지역이 최상위 위험 후보군에 포함된 것을 확인하였다.

2025년 12월 26일 grid 16967도 RF 96.71점, LightGBM 96.62점, ML Ensemble 96.16점으로 높은 위험수준을 보였다. 일별 순위는 각각 160위, 150위, 164위로 나타나 전체 격자의 약 상위 2% 이내에 포함되었으며, 집중점검 범위를 상위 200개 격자로 설정할 경우 실제 산불 발생지역을 사전에 점검 후보로 선별할 수 있는 사례로 확인되었다. XGBoost는 491위로 다른 모델보다 낮았으나, 앙상블 모델은 RF와 LightGBM의 높은 위험 신호를 반영하여 상위권 순위를 안정적으로 유지하였다.

반면 2025년 12월 10일 grid 13280은 RF에서 93.00점과 404위로 상위 500개 격자에 포함되었으나, LightGBM, XGBoost 및 ML Ensemble에서는 상대적으로 낮은 순위를 기록하였다. 이는 산불 발생 사례마다 기상·산림·전력설비 조건에서 나타나는 위험 신호의 강도가 다를 수 있음을 보여준다. 종합하면, 모델은 위험 특성이 뚜렷하게 형성된 두 산불 사례를 상위 10개 및 상위 200개 후보군에서 탐지하였으며, 특히 앙상블 모델은 개별 모델의 예측 편차를 종합하면서 고 위험지역의 점검 우선순위를 제시할 수 있음을 확인하였다.

[표 10] 앙상블 결과

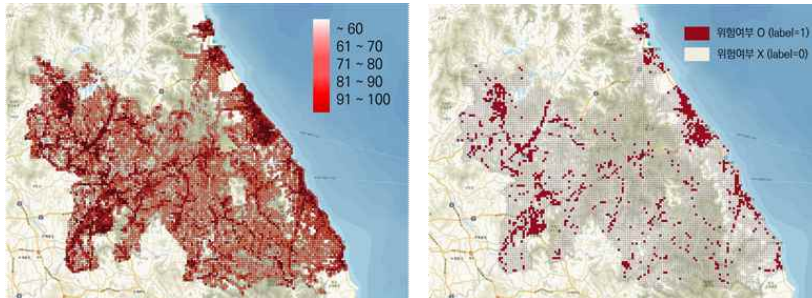
발생일	grid_id	RF	LGBM
2025-12-10	13280	93.00 / 404위	79.11 / 1294위
2025-12-26	16967	96.71 / 160위	96.62 / 150위
2025-12-29	15519	99.71 / 8위	99.52 / 6위

발생일	XGB	ML Ensemble
2025-12-10	58.96 / 2391위	82.71 / 1155위
2025-12-26	90.07 / 491위	96.16 / 164위
2025-12-29	99.13 / 11위	99.69 / 8위

※ 표 안의 값은 위험점수 / 일별 위험순위를 의미함.

4.2 Grid 기반 위험도 시각화

[그림 12] Grid 기반 산불 위험도 점수 [그림 13] Grid 기반 산불 위험 여부 시각화



[그림 12]는 예측된 Grid별 위험점수(Risk Score)를 공간적으로 시각화한 결과이다. 위험도는 0~100점으로 산정하였으며, 붉은색일수록 산불 발생 위험이 높은 지역을 의미한다. 강원도 동해안과 산림 밀집지역을 중심으로 높은 위험도가 분포하는 경향을 확인하였다. [그림 13]는 95% 위험도 기준(pred_95)을 적용하여 전력설비 인근 Grid를 위험(1)과 비위험(0)으로 분류한 결과이다. 위험도 점수가 95% 기준 이상인 Grid를 위험지역으로 정의하였으며, 붉은색은 위험지역(Label=1), 회색은 비위험지역(Label=0)을 나타낸다.

5. 추가 딥러닝 분석 방법

산불은 특정 시점의 기상조건뿐만 아니라 이전 수일간 누적된 건조 상태, 강수 부족, 강풍 등의 영향을 받아 발생한다. 따라서 이러한 특성을 반영하기 위해 단일 시점의 정보뿐만 아니라 과거의 변화 흐름까지 함께 고려하는 시계열 기반 분석을 추가적으로 수행하였다.

이에 본 연구에서는 최근 7일간의 기상·산림·전력설비 정보를 이용하여 산불 위험 상태를 분류하기 위해 LSTM 모델을 적용하였다.

다만 본 분석은 미래 산불 발생 여부를 예측하는 것이 아니라, 최근 기간의 정보를 활용하여 해당 구간이 산불 위험 상태에 해당하는지를 분류하는 것을 목적으로

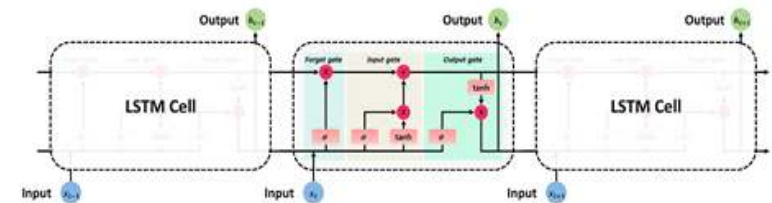
로 수행하였다. 따라서 본 연구의 LSTM 분석은 산불 발생 예측(Prediction)이 아닌 산불 위험 구간 분류(Classification) 문제로 정의하였다.

5.1 LSTM 개요

LSTM(Long Short-Term Memory)은 순환신경망(Recurrent Neural Network, RNN)의 한 종류로, 기존 RNN이 장기 의존성(Long-term Dependency) 학습 과정에서 발생하는 기울기 소실(Vanishing Gradient) 문제를 개선하기 위해 제안된 모델이다.

LSTM은 셀 상태(Cell State)와 게이트 구조(입력 게이트, 망각 게이트, 출력 게이트)를 이용하여 중요한 정보는 유지하고 불필요한 정보는 제거하면서 시계열 데이터를 학습한다. 따라서 시간의 흐름에 따라 변화하는 데이터의 패턴을 효과적으로 학습할 수 있다는 장점이 있다. [그림 14]은 LSTM의 기본 구조를 나타낸 것으로, 입력 데이터와 이전 시점의 정보를 함께 활용하여 현재 상태를 갱신하고 다음 시점으로 전달하는 과정을 보여준다.

[그림 14] LSTM 모델 구조



5.2 데이터 구성 및 전처리

LSTM 모델 학습을 위해 기상 데이터, 산림 데이터, 전력설비 데이터를 통합하여 입력 데이터를 구성하였다. 입력 변수는 기상 변수 10개, 산림 변수 3개, 전력설비 변수 1개로 총 14개 변수로 구성하였다.

LSTM은 시계열 데이터를 입력으로 사용하므로 각 Grid에 대해 최근 7일간의 데이터를 하나의 시퀀스(Sequence)로 생성하였다. 즉, 특정 기준일(Day 0)에 대하여 Day-6부터 Day 0까지의 7일간 데이터를 입력으로 사용하였으며, 각 시점마다 14개의 변수를 포함하도록 구성하였다. 이에 따라 최종 입력 데이터의 형태

는 (7, 14)로 생성되었다.

정답변수(Label)는 산불발생_V3를 사용하였다. 산불발생_V3는 실제 산불 발생 일을 포함하여 이전 7일까지를 위험 상태로 정의한 변수로, 산불 발생 이전에 누적되는 기상 조건의 영향을 반영하기 위해 활용하였다. [그림 15]는 LSTM 모델 학습을 위한 시계열 입력 데이터 구성 예시를 나타낸다.

[그림 15] LSTM 입력 데이터 구성 예시



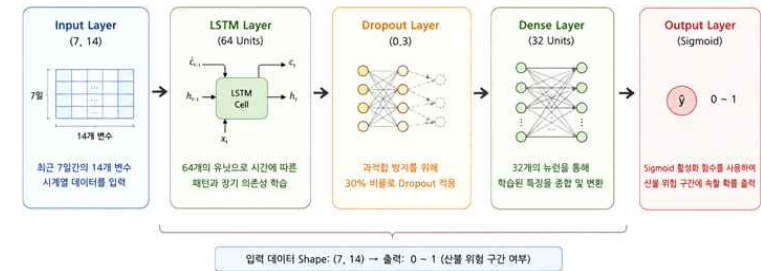
5.3 모델 구조

LSTM 기반 이진 분류 모델은 최근 7일의 14개 변수를 입력받아 시계열 패턴을 학습하고, 해당 기준일이 산불 위험 구간에 속할 확률을 출력하도록 구성하였다.

모델은 64개의 Unit을 갖는 LSTM Layer를 통해 시계열 패턴을 학습하도록 설계하였다. 이후 과적합 방지를 위해 Dropout(0.3)을 적용하였으며, Dense Layer(32 Units)를 통해 학습된 특징을 종합하였다. 마지막 출력층에서는 Sigmoid 활성화 함수를 사용하여 해당 Grid가 산불 위험 구간에 속할 확률을 산출하였다.

[그림 16]은 본 연구에서 사용한 LSTM 기반 산불 위험 구간 분류 모델의 구조를 나타낸다.

[그림 16] LSTM 기반 산불 위험 구간 분류 모델 구조



5.4 학습 및 평가 방법

본 연구의 분석 기간은 2023년 11월부터 2026년 5월까지로 설정하였으며, 실제 산불 발생 이력이 존재하는 Grid를 대상으로 시계열 데이터를 구축하였다. 각 Grid별로 일 단위 관측값을 기반으로 최근 7일간의 연속된 데이터를 하나의 시퀀스로 구성하는 Sliding Window 방식으로 학습 데이터를 생성하였다. 이를 통해 시간에 따른 기상, 산림, 전력설비 변수의 변화 패턴을 반영할 수 있도록 하였다.

본 연구의 목적은 미래 시점의 산불 발생 여부를 직접적으로 예측하는 것이 아니라, 최근 일정 기간의 시계열 데이터를 기반으로 현재 시점이 산불 위험 구간에 해당하는지를 분류하는 것이다. 따라서 모델은 입력된 7일간의 시계열 정보를 바탕으로 해당 시점의 위험 여부를 판단하도록 학습되었다.

모델 성능 평가는 ROC-AUC, Precision, Recall 지표를 활용하여 수행하였다. ROC-AUC(Receiver Operating Characteristic - Area Under the Curve)는 분류 모델의 전반적인 판별 능력을 평가하는 대표적인 지표로, 다양한 임계값(threshold)에 대해 True Positive Rate(재현율)와 False Positive Rate 간의 관계를 나타낸 ROC 곡선 아래 면적을 의미한다. AUC 값은 0과 1 사이의 값을 가지며, 1에 가까울수록 모델이 양성과 음성을 잘 구분함을 의미한다.

Precision(정밀도)은 모델이 위험 상태로 예측한 사례 중 실제로 위험 상태인 비율을 의미하며, 다음과 같이 정의된다.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (\text{식}10)$$

여기서 TP(True Positive)는 실제 위험 상태를 올바르게 예측한 경우, FP(False Positive)는 실제로는 위험하지 않지만 위험으로 잘못 예측한 경우를 의미한다.

Recall(재현율)은 실제 위험 상태인 사례 중 모델이 이를 올바르게 탐지한 비율을 의미하며, 다음과 같이 정의된다.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (\text{식11})$$

여기서 FN(False Negative)은 실제 위험 상태임에도 불구하고 모델이 이를 놓친 경우를 의미한다. 산불과 같은 재난 상황에서는 위험을 놓치는 경우(FN)가 큰 피해로 이어질 수 있기 때문에, Recall은 매우 중요한 평가 지표로 간주된다.

5.5 분석 결과

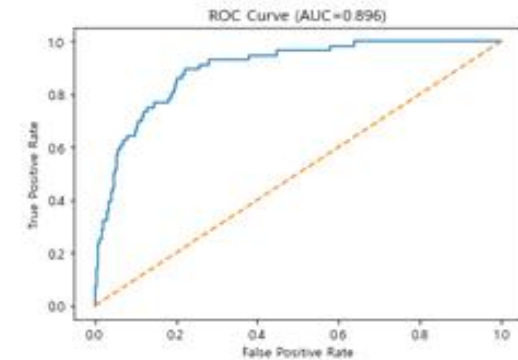
[그림 17]는 LSTM 기반 산불 위험 구간 분류 모델의 ROC Curve를 나타낸다. 실험 결과, 제안된 모델은 ROC-AUC 0.896을 기록하여 높은 수준의 분류 성능을 보였다. 이는 무작위 분류 수준(AUC=0.5)을 크게 상회하는 결과로, 기상·산림·전력설비 정보를 활용한 시계열 패턴 학습이 산불 위험 구간 식별에 효과적임을 의미한다.

또한 Recall은 0.732로 나타났으며, 이는 실제 위험 구간의 약 73.2%를 탐지할 수 있음을 의미한다. 산불 위험도 분석에서는 위험 지역을 놓치지 않는 것이 중요하므로, 높은 Recall 성능은 실제 위험 구간 탐지 측면에서 의미 있는 결과로 판단된다.

다만, 본 분석은 미래 산불 발생 여부를 예측하는 것이 아니라 최근 7일 내 산불 발생 여부를 기준으로 산불 위험 상태를 분류하는 것을 목적으로 수행되었다. 따라서 입력 데이터와 정답변수 모두 최근 7일간의 정보를 기반으로 구성되어 있어, 미래 시점의 산불 발생을 직접 예측하는 모델로 해석하기에는 한계가 존재한다.

그럼에도 불구하고 본 분석은 기상·산림·전력설비 정보의 시계열 패턴이 산불 위험 상태와 밀접한 관련이 있음을 확인하였으며, 산불 위험 구간 탐지 및 위험도 분석을 위한 보조적인 딥러닝 분석 사례로 활용 가능성을 확인하였다.

[그림 17] LSTM 모델의 산불 위험 구간 분류 성능 ROC Curve



6. 활용방안

본 분석에서 산출한 위험점수는 산불 발생 여부를 확정적으로 예측하는 값이 아니라, 강원도 내 전력설비 인근 격자의 상대적인 위험수준을 비교하기 위한 우선순위 지표로 활용할 수 있다. 각 날짜의 기상조건과 격자별 산림·전력설비 특성을 반영하여 위험도가 높은 지역을 선별함으로써, 한정된 인력과 장비를 우선적으로 점검해야 할 지역에 배치할 수 있다.

위험점수는 운영 목적에 따라 단계적으로 적용할 수 있다. 예를 들어 90점 이상은 넓은 범위의 관심지역, 95점 이상은 집중 점검이 필요한 고위험지역, 99점 이상은 최우선 순찰 및 현장 확인이 필요한 지역으로 구분할 수 있다. 이를 통해 평상시에는 넓은 범위의 예방관리를 수행하고, 건조·강풍 등 위험한 기상조건이 예상되는 시기에는 상위 위험 격자를 중심으로 순찰과 설비점검을 강화할 수 있다.

또한 위험점수를 지도 형태로 시각화하면 날짜별 화재위험 분포와 전력설비 밀집지역을 동시에 확인할 수 있다. 관제시스템이나 대시보드에 기상예보자료를 연계할 경우, 다음 날의 격자별 위험도를 사전에 산출하여 점검 대상과 순찰경로를 결정하는 의사결정 지원도구로 활용할 수 있다. 고위험지역에서는 전선 주변 수목 관리, 낙엽 및 가연물 제거, 설비 이상 여부 점검과 같은 예방조치를 우선적으로 시행할 수 있다.

향후에는 과거 산불 발생빈도, 경사도와 표고 등의 지형정보, 도로·주거지와의

거리, 토양수분 및 식생지수 등을 추가하여 위험점수를 고도화할 수 있다. 분석 체계를 다른 지역이나 송전탑·변전소 등 다양한 전력설비에 적용하면, 전국 단위의 화재 예방 및 설비관리 지원체제로 확장할 수 있을 것으로 기대된다.

참고문헌

기상청 국가기후데이터센터, 「방재기상관측(AWS) 자료」, 기상자료개방포털
기상청 국가기후데이터센터, 「종관기상관측(ASOS) 자료」, 기상자료개방포털
날씨 빅데이터 콘테스트 주최기관, 「전력설비 위치자료」, 2026 날씨 빅데이
터 콘테스트 제공자료, 2026.

산림청, 「산불발생통계」, 국가산불정보시스템 및 공공데이터포털
산림청, 「임상도」, 산림공간정보서비스

Cohen, J. D., & Deeming, J. E. (1985). The National Fire-Danger Rating System: Basic Equations. General Technical Report PSW-82. Berkeley, CA: U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station.

Fosberg, M. A. (1978). "Weather in Wildland Fire Management: The Fire Weather Index." Proceedings of the Conference on Sierra Nevada Meteorology, South Lake Tahoe, California. Boston, MA: American Meteorological Society.

Goodrick, S. L. (2002). "Modification of the Fosberg Fire Weather Index to Include Drought." International Journal of Wildland Fire, 11(3), 205-211.

Keetch, J. J., & Byram, G. M. (1968). A Drought Index for Forest Fire Control. Research Paper SE-38. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station.

Lawson, B. D., & Armitage, O. B. (2008). Weather Guide for the Canadian Forest Fire Danger Rating System. Edmonton, Alberta: Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre.

Van Wagner, C. E. (1987). Development and Structure of the Canadian Forest Fire Weather Index System. Forestry Technical Report 35. Ottawa, Ontario: Canadian Forestry Service.

2026년 9월 29일 인쇄

2026년 9월 30일 발행

발행인 이 주 영

편집인 변 준 영

발행처 중앙대학교 데이터과학융합연구소

서울특별시 동작구 흑석동 221

중앙대학교 경영경제대학 응용통계학과

전화 (02) 820-5499 팩스 (02) 814-5498

E-mail : data@cau.ac.kr

Homepage : <http://rcds.cau.ac.kr>

인쇄 유일문화사

논문투고 안내

“데이터과학연구”에 게재할 논문을 연중 접수하고 있습니다. 게재를 원하시는 분은 다음 사항들을 참조하기 바랍니다.

1. 논문분야: 경영, 사회, 교육, 공학, 의·약학 등의 학문 분야에서 데이터 조사 및 분석을 활용한 논문을 폭넓게 게재한다.
2. 논문부수 및 접수방법: 20매 내외의 분량을 연구소 E-mail을 통하여 접수한다. (E-mail : data@cau.ac.kr)
3. 심사 및 수정: 본 논문집에 투고한 논문은 편집위원회의 심사를 거쳐 수정 및 보완과 게재 여부를 결정한다.
4. 발간: 본 논문집은 연간 2회(6월, 12월) 발행을 원칙으로 한다.
5. 논문심사 진행사항 문의: 연구소 조교 (전화 02-820-6352)

데이터과학융합연구소 임원 및 연구부

- 소 장 : 이주영 교수 (응용통계학과, jooylee@cau.ac.kr)
- 간 사 : 변준영 교수 (응용통계학과, junyoungb@cau.ac.kr)
- 연구부 : 김원국 교수 (응용통계학과)
 - 김정하 교수 (의과대학)
 - 문광수 교수 (심리학과)
 - 박재현 교수 (컴퓨터공학과)
 - 변준영 교수 (응용통계학과)
 - 성병찬 교수 (응용통계학과)
 - 신우영 교수 (의과대학)
 - 윤영철 교수 (의과대학)
 - 이세영 교수 (의과대학)
 - 이재우 교수 (산업보안학과)
 - 이재현 교수 (응용통계학과)
 - 이정규 교수 (의과대학)
 - 이주영 교수 (응용통계학과)
 - 임예지 교수 (응용통계학과)
 - 한수현 교수 (의과대학)
 - 황범석 교수 (응용통계학과)

데이터과학연구 편집위원회

편집위원장 : 이주영 교수 (응용통계학과)
편집 위원 : 변준영 교수 (응용통계학과)

