ViT 모델을 활용한 토마토 병해 탐지 연구

(Tomato Disease Detection Using Vision Transformer Model)

김정기*, 이명훈**
(Jeong Ki Kim, Meong Hun Lee)

요 약

농업 분야 또한 중요 산업으로 주목받고 있으며, AI를 활용한 작물 관리, 시설관리 등이 진행되고 있지만 작물 병해 관리는 아직까지 AI를 활용한 데이터 기반의 분석보다는 경험과 직관에 의존하여 한계점이 보이고 있는 상황이다. 본 논문에서는 ViT 모델을 활용한 토마토 병해 탐지 연구를 진행한다. 적합한 토마토 이미지 데이터셋을 구성하고 AI 기반 이미지 데이터 학습을 진행하고, 모델을 활용한 학습 결과를 확인하고자 한다. 그 후 CNN 모델과 ViT모델의 정확도, 정밀도 등의 성능 비교를 통해 연구를 진행한다. ViT 모델을 통한 학습 결과 정확도 88.9%, 정밀도 86.7%, 재현율 84.2% F1-score 85.4%이며, AUC-ROC 0.905 Log loss 0.276으로 CNN 모델과 비교하였을 때 우수한 성능을 보여주고 있다. 해당 연구를 통해 ViT모델 또한 이미지 데이터 분석 분야에서 경쟁력있는 모델임을 확인하였다. 또한, 스마트팜 환경에서의 병해를 조기 탐지하여 작물 피해를 최소화할 수 있으며, 노동력 절감 및 농가의 생산성 향상에 도움이 될 것으로 기대한다.

■ 중심어 : ViT 모델 ; CNN 모델 ; 병해 탐지 ; 인공지능 ; 이미지 데이터

Abstract

The agricultural sector is also attracting attention as an important industry, and crop management and facility management using AI are in progress, but crop disease management is still showing limitations by relying on experience and intuition rather than data-based analysis using AI. In this paper, we conduct a study on tomato disease detection using the ViT model. We would like to construct an appropriate tomato image dataset, conduct AI-based image data learning, and confirm the learning results using the model. After that, the research is conducted by comparing the performance of the CNN model and the ViT model such as accuracy and precision. The accuracy of the learning results through the ViT model is 88.9%, the precision is 86.7%, the reproduction rate is 84.2% F1-score 85.4%, and the AUC-ROC 0.905 log loss is 0.276, showing excellent performance compared to the CNN model. Through this study, it was confirmed that the ViT model is also a competitive model in the field of image data analysis. In addition, it is expected that it can minimize crop damage by early detection of diseases in the smart farm environment and help reduce labor and improve productivity of farmers.

■ keywords : Vision Transformer model ; Convolutional Neural Networks; Disease Detection ; Artificial intelligence ; Image Data

Ⅰ. 서 론

다분야의 기술이 발전됨에 따라 농업 분야 또한 중요 산업으로 주목받고 있으며, 스마트팜을 활용한 스마트농업 또한 다양한 AI 기술들과 접목하여 연구개발이 이루어지고 있다[1-3]. AI를

* 준회원, 국립순천대학교 스마트융합학부
** 종신회원, 국립순천대학교 융합바이오시스템기계공학과
이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임 (IITP-2025-RS-2020-II201489)
접수일자 : 2025년 03월 10일
수정일자 : 2025년 03월 27일
게재확정일 : 2025년 05월 30일
교신저자 : 이명훈 e-mail : leemh777@scnu.ac.kr

활용한 작물관리, 시설관리 등이 진행되고 있지만 기후변화로 인한 작물 병해 관리는 아직까지 연구개발단계에 머물러 있다[4,5]. 매년 작물의 병해로 인한 피해 규모와 금액은 증가하고 있지만 농민들은 AI를 활용한 데이터 기반의 분석보다는 경험과 직관에 의존하여 작물 병해 문제를 해결하고 있는 경우가 대부분이다[6]. 지구온난화로 인한 지구의 기온은 매년 상승 중이며, 그로 인한 작물 병해의 피해 규모와 금액 또한 상승 중인 추세를 보이고 있기 때문에 기존의 방안만으로 문제점들을 해결하기에는 한계점이 보이는 상황이다[7]. 이러한 문제를 해결하기 위해 AI 기반 병해 탐지 모델을 활용하여 연구를 진행하고자 한다.

본 연구에서는 토마토 작물을 대상으로 이미지 데이터를 활용한 병해 탐지 연구를 통해 매년 증가하는 병해로 인한 피해를 줄여보고자 한다. 그 과정에서 ViT(Vision Transformer) 모델을 활용한 AI 분석을 진행하고, 기존에 주로 이미지 분석에 사용된 CNN(Convolutional Neural Networks) 모델과 병해 탐지 성능을 비교하여 연구를 진행하고자 한다. 본 논문의 순서는 다음과 같다. 2장에서 사용하는 모델과 관련된 연구를 소개하고 3장에서는 ViT 모델을 활용한 토마토 병해 탐지 연구 진행 과정을 소개하며, 기존의 CNN 모델과 성능 평가 및 비교를 통해 연구를 진행한다. 마지막 4장에서는 결론 및 추후 연구를 통해 본 논문을 마무리하고자 한다.

Ⅱ. 관련 연구

1. ViT(Vision Transformer) 모델

ViT 모델은 2021년 Google research에서 발표한 모델로, Transformer 아키텍처를 이미지 처리에 적용한 모델이다[8].

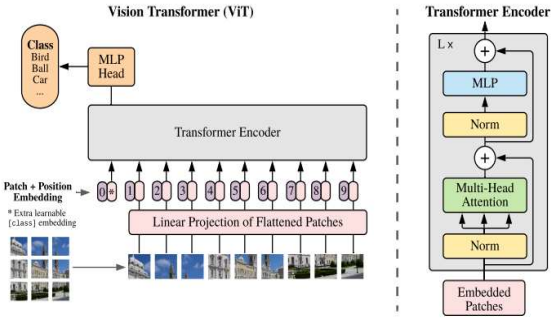

그림 1. ViT 모델 구조

<그림 1>은 ViT의 모델 구조로 기존의 이미지 데이터 분석에서 자주 사용되는 CNN 모델은 합성곱 연산을 사용하여 이미지 데이터의 국소적인 특징을 학습하는 반면, ViT 모델은 입력 이미지를 작은 패치로 나누고 각 패치를 1차원 토큰으로 변환한다. 패치에는 Linear embedding을 적용 후 시퀀스로 만들어준 형태로 Transformer 입력에 들어가게 되며, 이는 이미지 전체의 전역적인 특징의 학습을 진행한다는 기능이 있다. 일반적으로 적은 양의 데이터셋 보다는 대용량 데이터셋을 활용하여 학습을 진행하였을 때 정확도 높은 학습이 가능하다는 특징이 있다[9]. 지속적인 연구를 통해 최근 연구에서 Swin Transformer, Data-dfficient ViT 모델 등 기존보다 경량화된 ViT 모델들이 등장하면서, 상대적으로 적은 데이터를 보유하고 있어도 CNN 모델을 대체할 수 있다는 전망을 보인다.

2. CNN(Convolutional Neural Networks) 모델

CNN 모델은 이미지 분류 및 객체 탐지에서 강점을 보여 주로 사용되는 신경망 모델로, 합성곱 필터를 사용하여 이미지 데이터의 국소적인 특징을 학습하여 효과적으로 추출한다[10].

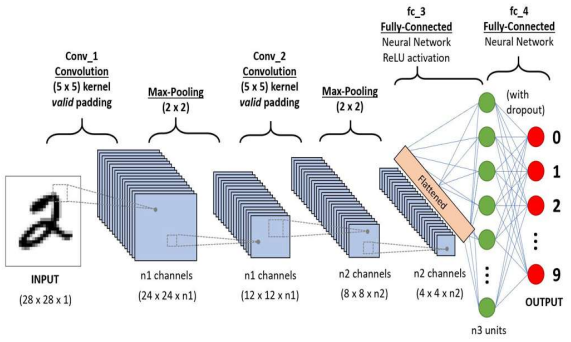

그림 2. CNN 모델 시퀀스 과정

<그림 2>는 CNN 모델의 학습 과정으로 CNN의 기본구조는 입력계층, 합성곱 계층, 풀링 계층, 완전연결층으로 구성되며, 층을 거치며 이미지 데이터의 특징을 순차적으로 학습한다. CNN의 합성곱 계층은 작은 크기의 필터를 사용하여 이미지에서 특징을 추출하며, 특정 지역을 학습하는 데 강점이 있는 모델이다. CNN의 대표적인 모델로는 AlexNet, VGGNet, ResNet, Efficient Net 등이 있으며, 이미지 데이터를 계층적으로 특징을 추출하여 분석하는 방식을 통해 이미지 데이터 분석에 강점이 있다[11]. CNN은 상대적으로 작은 데이터셋에서 높은 정확도의 성능을 발휘지만 이미지 데이터의 전체적인 관계를 학습하거나 빅데이터인 대용량의 데이터셋을 분석하는 과정에서는 한계가 있다[12,13].

III. 토마토 병해 탐지

3장에서는 ViT 모델을 활용한 토마토 병해 탐지 연구를 진행하기 위해 적합한 토마토 이미지 데이터셋을 구성한다. Google Colab을 통해 AI 기반 이미지 데이터 학습을 진행하였고, <표1>과 같이 GPU 사양 및 프레임워크 등의 설정을 통해 환경을 구성하였다. 그 후 연구를 통해 CNN 모델과 ViT 모델의 정확도, 정밀도 등의 모델 평가 지표를 비교하여 연구 내용을 제시하고자 한다.

표 1. 이미지 데이터 분석 환경 구성

항목	적용버전
GPU	NVIDIA Tesla T4(Google Colab)
Python	3.11
PyTorch	2.5.1
Torchvision	0.20.1
Timm	1.0.15
OpenCV	4.11.0.86
Scikitlearn	1.6.1
Matplotlib	3.10.0

1. 토마토 병해 탐지 연구

그림 3. 토마토 이미지 데이터셋

<그림 3>은 병해 탐지 연구에 사용할 이미지 데이터로 모델을 구성하는데 앞서 작물은 토마토, 병해는 흰가루병으로 선정하여 이미지 데이터를 구성하였다. 정상 이미지와 병해 이미지는 본 연구를 진행하기 위해 직접 촬영한 이미지와 AI-Hub 등의 공공데이터를 활용하여 구성하였다. 각각 1,000장의 정상 토마토 이미지 데이터와 700장의 병해 토마토 이미지 데이터를 확보하였다. 그 후 ViT 모델을 사용하기에 적합한 기존보다 대용량의 데이터셋 구성과 정상 및 병해 이미지 간의 데이터 불균형 해소를 위한 데이터 증강작업을 진행하였다.

표 2. 데이터 전처리 및 증강

변환기법	적용된 값
Resize	(224, 224)
Random Horizontal Flip	p=0.5
Random Rotation	±45°
Color Jitter	brightness=0.3, contrast=0.3, saturation=0.3, hue=0.1
Random Affine	translate=(0.2, 0.2)
Random Resized Crop	scale=(0.8, 1.0), size=224
ToTensor	–
Normalize	mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5]

<표 2>는 ViT 모델 기반 학습에 적합한 데이터셋을 구성하기 위한 데이터 전처리 및 증강 설정값을 나타낸 표이다. Resize는 이미지 크기를 조정하기 위한 데이터 전처리 항목으로 ViT와 CNN 모델은 입력 크기가 일정해야 하므로, 모든 이미지를 (224x224) 크기로 변환하여 일관된 데이터 전처리를 진행하였다. Random Horizontal Flip은 데이터 증강기법의 하나로 p=0.5의 확률로 이미지를 랜덤 좌우 반전을 진행하여, 데이터 다양성을 증가시킬 수 있도록 하였다[14]. Random Rotation은 데이터 증강기법 중 하나인 랜덤 회전 방법으로 다양한 각도에서 병해를 인식할 수 있도록 ±45° 각도로 회전하게끔 하였다. Color Jitter 또한 데이터 증강기법 중 하나로 이미지의 색상 변화를 진행하여 밝기, 대비, 채도, 색조 등을 무작위로 변경하는 기법으로 적용된 값의 범위 내에서 랜덤하게 조정하여, 다양한 조명 조건에서도 병해를 인식할 수 있도록 하였다. Random Affine은 데이터 증강기법 중 하나인 랜덤 이동 변환으로 이미지를 0.2의 비율로 이동시켜 모델이 병해 위치와 관계없이 잘 탐지할 수 있도록 하였다. Random Resized Crop은 이미지별 랜덤한 영역을 크롭한 후 다시 224x224 크기로 리사이즈하는 데이터 증강기법으로 특정 병해 패턴이 항상 동일한 위치에 존재하지 않기 때문에 다양한 이미지 영역을 학습하여 모델이 병해 탐지를 더 원활히 수행하도록 하였다[15]. ToTensor은 이미지 데이터 증강을 Pytorch에서 처리할 수 있도록 Tensor 형태로 변환시키기 위한 기법이며, Normalize는 정규화를 통해 모델학습과정을 안정화시키는 기법이다. 해당 기법들을 통해 데이터 전처리 및 증강을 진행하였다. 증강 후 3,000장의 정상 토마토 이미지 데이터와 2,900장의 병해 이미지 데이터를 확보하여, 대용량 데이터 분석에 강점이 있는 ViT 모델에 적합한 데이터셋을 구성하였다. 전체 데이터를 9:1로 훈련 세트와 테스트 세트를 나누고, 훈련 세트에서 다시 8:2로 분할하여 검증세트를 구성하였으며, 72:18:10의 비율로 학습할 데이터셋을 구성하였다.

데이터셋 구성 후 ViT 모델 중 비교적 적은 데이터의 양으로도 유의미한 정확도를 나타낼 수 있는 Swin Transformer 모델을 활용하여 학습을 진행하도록 설정하였다.

```
for epoch in range(num_epochs):
    model.train()
    running_loss = 0.0

    for images, labels in train_loader:
        images, labels = images.to(device), labels.to(device)

        optimizer.zero_grad()
        outputs = model(images)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()

        running_loss += loss.item()

    print(f"Epoch [{epoch+1}/{num_epochs}],
    Loss: {running_loss/len(train_loader):.4f}")
```

그림 4. Swin Transformer 모델 구성 코드

<그림 4>에서는 사용할 ViT 모델을 활용한 Swin Transformer 모델 구성 코드를 작성하여 구성하였다. ViT 모델 중 비교적 적은 데이터의 양으로도 유의미한 정확도를 나타낼 수 있는 Swin Transformer 모델을 활용하여 학습을 진행하도록 설정하였고, 현재 보유한 데이터셋에 적합한 학습방법, 학습횟수 등을 설정하여 조정하였다. 학습 정확도의 향상을 위해 Criterion 손실함수와 AdamW 옵티마이저를 최적화 알고리즘으로 지정하였고, 기본으로 부착되어있는 GELU 활성화 함수를 활용해 모델 학습 루프를 통해 반복적으로 학습할 수 있도록 설정하였다. 학습 진행 과정에서는 학습횟수와 손실값 등의 세부값을 확인하며, 세부조정을 할 수 있도록 모델을 구성하였다. 학습한 모델을 활용하여 실제로 정확한 예측을 하는지에 대한 결과값을 확인하였다. 학습에는 사용되지 않은 병해 및 정상 이미지 데이터를 활용하여, AI가 정상 이미지로 예측한다면 healthy 병해 이미지로 예측한다면 diseased로 예측하도록 진행하였고, 예측결과는 diseased로 해당 결과를 통해 탐지 모델이 정상적으로 작동하여 예측을 진행한다는 것을 확인하였다.

2. 모델 성능 평가

표 2. CNN 모델과 VIT 모델의 성능 평가

모델	정확도	정밀도	재현율	F1-score	AUC-ROC	Log loss
CNN (ResNet-50)	87.4%	85.3%	83.4%	83.7%	0.882	0.315
VIT (Swin Tranformer)	88.9%	86.7%	84.2%	85.4%	0.905	0.276

<표 2>에서는 각각 모델의 성능 평가값을 정리하였다. ViT 모델의 성능 비교를 위해 이미지 데이터 분석에서 주로 사용되는 CNN 기반의 ResNet-50 및 ReLU 활성화 함수를 활용한 모델과 본 실험에서 사용한 ViT 기반 Swin Transformer 모델을 비교하였다. CNN 모델의 정확도(Accuracy)는 87.4%, 정밀도(Precision)는 85.3%, 재현율(Recall)은 83.2%, F1은 83.7%로 평가되었고 ViT 모델의 정확도는 88.9%, 정밀도는 86.7%, 재현율은 84.2%, F1은 85.4%로 측정되었다. CNN 모델의 AUC-ROC는 0.882 ViT 모델은 0.905이며, Log Loss값 또한 CNN 모델은 0.315 ViT 모델은 0.276으로 최적화가 CNN 모델보다 더 잘 되어있음을 확인하였다. 성능 비교 통해 ViT 모델의 성능 평가값이 전체적으로 높게 측정이 되었으며, 이러한 결과는 ViT 모델이 기존의 이미지 데이터에서 주로 사용되는 CNN 모델과 비교하였을 때도 경쟁력 있는 모델이라는 것을 확인할 수 있다.

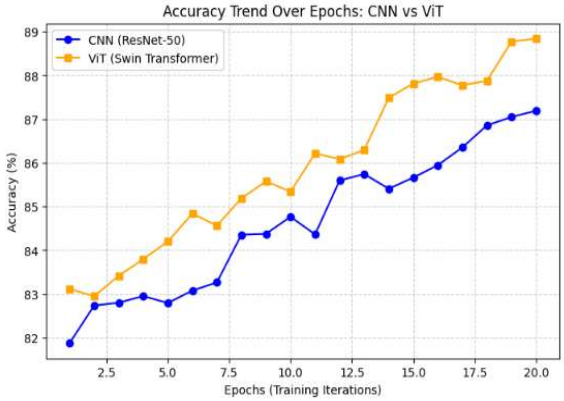

그림 6. 모델별 정확도 학습 성능

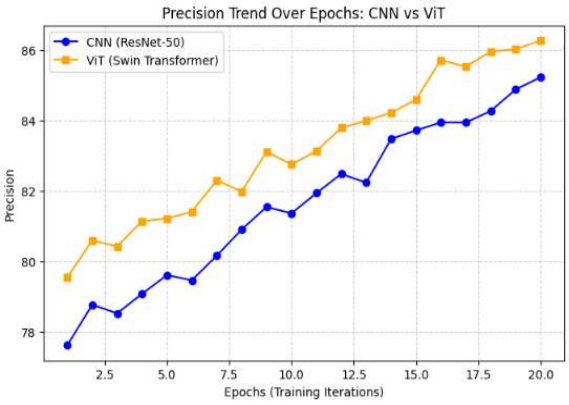

그림 7. 모델별 정밀도 학습 성능

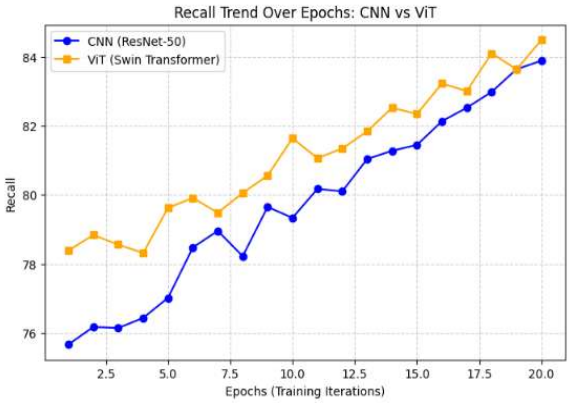

그림 8. 모델별 재현율 학습 성능

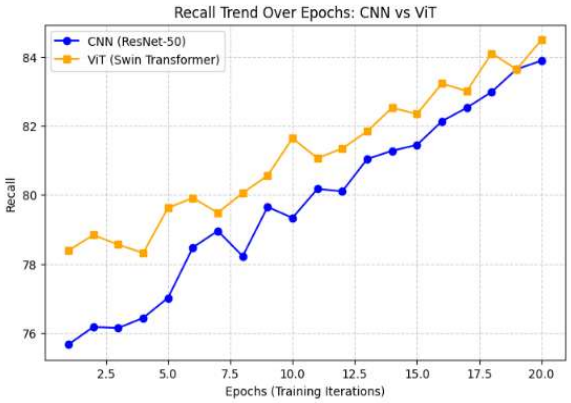

그림 9. 모델별 F1-socore 학습 성능

<그림 6,7,8,9>는 모델별 학습횟수에 따른 분류 모델 평가 지표인 정확도, 정밀도, 재현율, F1-score를 순서대로 시각화하여 그래프로 나타내었다. 4개의 항목을 비교하였을 때 전체적으로 비슷한 경향의 추이를 보이고 있으며, CNN 모델의 성능지표보다 ViT의 성능지표가 학습을 진행할수록 더 높게 측정되어 변화하고 있는 것을 확인할 수 있다.

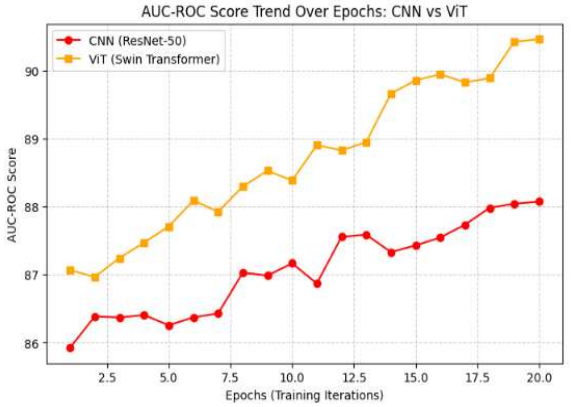

그림 10. 모델별 AUC-ROC 학습 성능

<그림 10>은 AUC-ROC 값을 시각화하였으며, AUC는 값이 100에 가깝게 측정될수록 모델이 정상과 병해를 명확하게 구별할 수 있음을 확인하는 지표이다. 해당 그래프를 통해 ViT 모델이 CNN 모델보다 병해 탐지 구별이 더 우수하게 학습되었음을 보여준다.

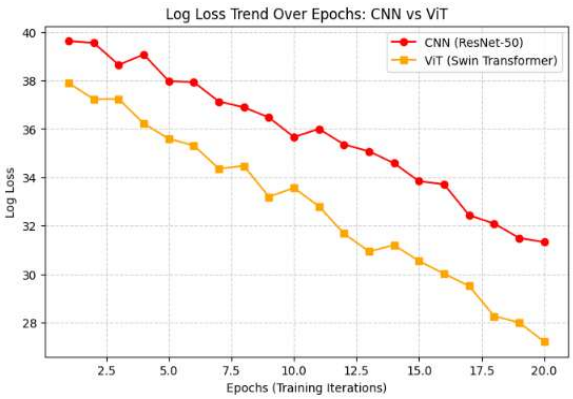

그림 11. 모델별 Log loss 학습 성능

<그림 11>은 Log Loss 값을 시각화하였으며, Log loss는 값이 낮을수록 모델이 더 정확한 확률값을 제공한다는 것을 확인하는 지표이다. 해당 그래프를 통해 ViT 모델이 CNN 모델보다 병해 탐지 및 예측 결과값을 더 신뢰할 수 있도록 학습되었음을 보여준다.

Ⅳ. 결론 및 추후 연구

본 논문에서는 AI와 ViT 모델을 활용하여 이미지 데이터 기반 토마토의 병해 탐지 연구를 진행하였다. 연구를 진행하기 위해 기존의 이미지 데이터에서 주로 사용되는 CNN 모델과 ViT 모델을 연구하였으며, 학습을 위한 토마토의 흰가루병 이미지 데이터와 정상 이미지 데이터를 통해 데이터셋을 구성하였다. ViT의 Swin Transformer 모델을 활용하여 토마토의 흰가루병 병해를 탐지할 수 있도록 모델을 구성하였고, 이를 통해 결과값을 확인하고 CNN 모델과 ViT 모델 간의 성능 비교 순으로 연구를 진행하였다. AI를 활용한 병해 탐지 연구를 통해 스마트팜 환경에서의 병해를 조기 탐지하여 작물 피해를 최소화할 수 있

으며, 연구단계에서 멈추는 것이 아니라 실제 농가에서까지 적용되었을 경우 노동력 절감 및 농가의 생산성 향상에 도움이 될 것으로 기대된다. 추후 연구를 통해 본 연구에서 구성하였던 데이터셋보다 더욱 많은 빅데이터를 활용하여 학습을 진행하게 된다면, VIT 모델은 CNN과 비교하였을 때 대용량의 데이터 처리에 강점이 있는 모델이기 때문에 더욱 유의미한 성능 지표를 얻을 수 있을 것으로 예상한다, 또한 데이터셋 전처리 과정에서도 작물뿐만 아니라 사람 손이 포함된 이미지가 일부 사용되었기 때문에, 향후 연구에서는 ROI 기반 라벨링 등을 적용하여 작물 및 병해 부위 중심의 명확한 이미지 데이터셋을 구성한다면 더욱 신뢰성 높은 모델별 학습 결과를 기대할 수 있을 것이다. 추가로 ViT의 ViT-B/16, DeiT-S 등 다양한 모델을 통해서도 연구 결과값을 얻을 수 있기 때문에 추후 연구를 확장하여 진행할 계획이다.

REFERENCES

[1] 박경섭, 유용권, 박현준, and 손정익, "한국형 스마트팜 연구 기술 개발 현황," *in 한국원예학회 학술발표요지,* 51-52쪽, 2022년

[2] 천예원, "스마트팜 기술 현황과 표준화 동향 분석 : 국내·외 비교분석을 중심으로", *석사학위논문, 중앙대학교,* 서울, 2023년

[3] 남기정, "농협의 스마트농업 추진현황과 향후 발전 방안에 관한 연구," *협동조합경영연구,* 제55권, 45-69쪽, 2021년

[4] 나명환, 조완현 and 김상균. "딥러닝 알고리즘을 이용한 토마토에서 발생하는 여러가지 병해충의 탐지와 식별에 대한 웹응용 플랫폼의 구축," *품질경영학회지,* 제48권, 제4호, 581-596쪽, 2020년

[5] 손경자, 박경운, 이승재, and 한경석, "데이터경제 활성화를 위한 농업경영 빅데이터 플랫폼 구축 방안에 관한 연구," *한국IT정책경영학회 논문지,* 제12권, 제5호, 1967-1973쪽, 2020년

[6] 김주희, 최민경, 문형철, and 전형권, "전북지역 토마토와 박과류 무농약재배지의 주요 병해 발생 현황," *환경생물,* 제38권, 제3호, 486-495쪽, 2020년

[7] 송석호, 이주형, 박달주, and 이명원, "농업용수 기후변화 취약성 평가 및 농업생산기반시설 기후위기 적응대책에 대한 연구," *한국기후변화학회지,* 제14권, 제6-2호, 1005-1011쪽, 2023년

[8] 최수정, "Vision Transformer를 이용한 Metric-based Meta-Learning", *석사학위논문, 동국대학교,* 서울, 2024년

[9] 이민주, 채명호, and 임완수, "Vision Transformer를 이용한 자동변조인식 기술," *한국통신학회논문지,* 제49권, 제8호, 1074-1081쪽, 2024년

[10] 김일, "영역 기반 convolutional neural network를 이용한 과수 영상에서의 병해 탐지", *석사학위논문, 세종대학교,* 서울, 2020년

[11] 이창준, 심춘보, 김진성, 김준영, 박준, 박성욱, 정세훈. "EfficientNet 활용한 딸기 병해 진단 서비스," *스마트미디어저널,* 제11권, 제5호, 26-37쪽, 2020년

[12] Phung Van Hiep, "CNN-based Model Design for High Accuracy Image Classification on Cloud and Crop Pest", *박사학위논문, 한밭대학교,* 대전, 2020년

[13] 심준현, 김철진, "시계열 이미지 데이터 기반 상품 추천을 위한 CNN 모델 성능 비교 연구," *한국지식정보기술학회 논문지,* 제18권, 제5호, 1253-1264쪽, 2023년

[14] 오주현, "딥러닝을 위한 이미지 처리 기반의 데이터 증강", *석사학위논문, 성신여자대학교,* 서울, 2024년

[15] 이현조, 구현정, 이경철, and 채철주, "스마트팜 AI 모델 활용을 위한 농작물 생육 데이터 증강 기술에 관한 연구," *in 한국콘텐츠학회 종합학술대회 논문집,* 309-310쪽, 2024년

저 자 소 개

김정기(준회원)

2023년 국립순천대학교 정보통신학과 공학사 졸업.

2025년~현재 국립순천대학교 스마트융합학부 석사 재학.

<주관심분야 : 인공지능, 스마트 인포메이션, 인포메이션 시스템>

이명훈(종신회원)

2004년 국립순천대학교 정보통신공학 학사 졸업.

2006년 국립순천대학교 정보통신공학 석사 졸업.

2010년 한국전자통신연구원 선임연구원

2011년 국립순천대학교 정보통신공학 박사 졸업.

<주관심분야 : 모바일 네트워크, 인공지능, ICT융합, 표준, 스마트팜>