고급기계학습 TERM-PROJECT

Voice Pathology Detection:
딥러닝 모델을 이용한 장애 음성 인식 분류

3조
12192960 김승준
12202554 김지은
12215366 김기훈

CONTENTS

목차

3조
12192960 김승준
12202554 김지은
12215366 김기훈

1. 팀원 소개 및 역할 분담

김승준

- 관련 선행 연구 조사
- 데이터 전처리, 증강기법 구현
- 적용 코드 작성, 평가
- PPT 제작, 발표 자료

김지은

- 관련 선행 연구 조사
- 데이터 전처리, 증강기법 구현
- 정규화 기법 탐색
- PPT 제작, 발표 자료

김기훈

- 관련 선행 연구 조사
- 데이터 전처리, 증강기법 구현
- 적용 코드 작성, 평가
- PPT 제작, 발표 자료

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

1주차

-데이터 증강 없이 사전학습 모형(pre-trained model) 기반 학습

➔ResNet / VGGNet / Google Net / Transformer

2주차

MFCC -> 1D-CNN

VTLN 정규화

3주차

이미지 단순 선형 증강(1.1배, 1배, 0.9배)

4주차

성별 나누어 학습 후 앙상블

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

1주차 MFCC -〉 1D-CNN
Spectrogram 정규화 (*Spectrogram 형태 변경 X)

1. VGGNet

- 수업에서 다룬 모델 中 단순한 구조 / 빠른 속도의 VGGNet 적용 -〉 70% 정도의 아쉬운 정확도

2. VGGish

- VGGNet을 오디오 정보 처리에 이용하기 위해 개량한 모델
- 오디오 분류에 주로 사용되어, 좋은 정확도 기대 -〉 VGGNet 과 별 다른 차이 X (70% 초반)

3. Resnet 34 / 50 / 101/ 152

- Tutorial 에서의 34 이용 -〉 단순히 복잡도가 높은 50, 101, 152로 변경
- 기존 정확도 60 % 대 -〉 70 % 초반으로 상승
- spectrogram 분석 시, 모델 깊이와의 상관관계 확인

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

2주차 MFCC -〉 1D-CNN
Spectrogram 정규화

1. Spectrogram 정규화

- 발화 내용은 모두 동일하나, 화자의 목소리는 전부 다름을 확인
- Librosa 라이브러리 내장 정규화 파라미터 Normalization = True 를 이용 -〉 해당 변수들을 정규화하여 학습
- 자동 정규화 후, ResNet 에 통과 -〉 70% 를 넘지 못하여 효과 X

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

2주차 MFCC -〉 1D-CNN
Spectrogram 정규화

2. MFCC 자료형 활용 시도

- MFCC: 음성 정보에서 추출한 Melspectrogram에서 다시 특징을 추출한 1차원 형태의 자료
- 최근 연구에 따르면, MFCC를 1D CNN 모델에 학습시켜 음성 분류 작업에 활용하는 사례 증가
- 1D CNN은 시계열 정보 처리에 효과적 -〉 MFCC를 활용한 정확도 개선을 기대했으나, 70% 초중반에 그치는 한계

➔MFCC가 구음장애의 특성을 추출하기에는 지나치게 단순화된 자료형태일 수 있다는 판단

& MFCC의 이전 단계인 Melspectrogram을 직접 활용하는 아이디어 제안

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

3주차 데이터 증강 시도 & Melspectrogram 데이터 형태 이용 확정

1.1. 무작위 추출 후 단순 선형 증강

- Melspectrogram 추출 후, 미리 정해진 확률값(0.5)에 따라 이미지 크기 조절
  - 0.9배 혹은 1.1배 비율로 크기 조정
  - 조정된 이미지를 원래 데이터셋에 추가
- ResNet 모델에 학습 진행
  - 70% 대 중반의 정확도, 랜덤 요소로 인한 시행별 정확도 편차가 큼

1.2. 전체 추출 후 단순 선형 증강

- 전체 데이터셋에 대해 크기 변환(0.9배, 1.1배) 수행 후 원래 데이터셋에 추가
- 랜덤 요소 제거 -> 편차 감소 및 무작위 추출 방식 대비 미세한 정확도 향상(여전히 70% 대 중반)

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

3주차 데이터 증강 시도

2. 좌우/ 상하 반전 이용

- ‘크기 변환’에서 효과 X
- 기본 이미지 데이터 증강 기법 中 하나인 ‘방향 반전’ 시도 -> 오히려 정확도 감소(70% 초반)

➔ 데이터 증강을 하더라도, Melspectrogram 자체가 시계열 데이터이므로 방향 자체는 유지

& 크기, 방향 전환이 아닌 마스킹 기법의 증강 이용을 결정

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

4주차 남성, 여성 데이터로 주파수 기반 분할, 앙상블 학습

1. 음성 데이터 특성 분석
- 남성과 여성의 음역대 차이 인식, 다양한 음역대로 인해 모델의 특징 추출 어려움 발생

2. 데이터 분할 및 전처리
- 160Hz을 기준으로 음성 데이터 분할 (160Hz 미만: 남성, 160Hz: 여성)
- 3개의 ResNet34 모델 구축, 3세트의 train 데이터 구축(남성 / 여성 데이터 / 전체 데이터)

3. 모델 학습 및 앙상블
- 각 ResNet34 모델을 해당 데이터셋으로 학습
- 학습된 모델을 적절한 가중치로 앙상블하여 분류
- 정확도 및 F1 score: 70% 대 후반 -> 이후 소개할 모델의 성능이 해당 모델을 능가하여 폐기

➔각 모델별 특화된 특징 추출 능력 확인, 다양한 모델 학습 후 앙상블 방식 채택 결정

3조
12192960 김승준
12202554 김지은
12215366 김기훈

2. 시도해

각 시도해로부터 유지하기로 한 방식

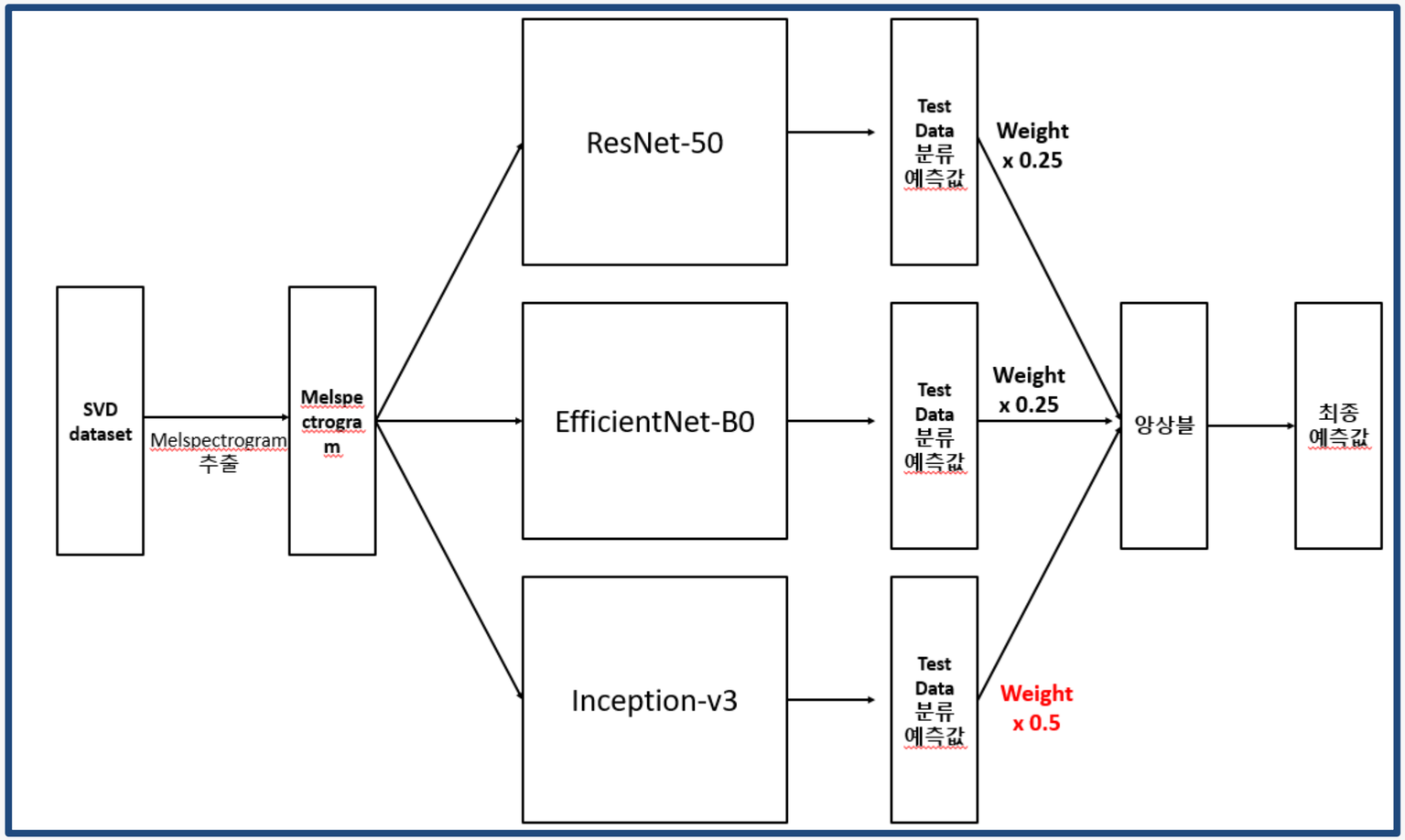
- 데이터 형태 측면
  - Melspectrogram 추출
  - SpecAugment 방식을 이용하여 데이터 증강

- 모델 측면
  - Resnet 34보다 복잡한 구조의 pre-trained 모델들을 이용
  - 여러 음역대의 음성들 있음을 고려하여 다양한 모델을 학습 후 앙상블

3조
12192960 김승준
12202554 김지은
12215366 김기훈

3. 모델 및 학습 방법

모델 구조 도식화



3조
12192960 김승준
12202554 김지은
12215366 김기훈

3. 모델 및 학습 방법

모델 구조 도식화

- 증강된 데이터 셋 준비
- ResNet 50 , EfficientNet-B0 , Inception-v3에 각각 학습
- 각각의 모델에서 예측한 이진 확률값 추출 후, 모델별 중요도에 따라 가중 평균 도출
- 가중평균값 기반으로 Healthy / Pathology 이진 분류

3조
12192960 김승준
12202554 김지은
12215366 김기훈

3. 모델 및 학습 방법

튜토리얼 모델과의 차이점

```
# dataloader
train_dataloader = DataLoader(trainset, batch_size=32, shuffle=True)
test_dataloader = DataLoader(testset, batch_size=16, shuffle=False)
```

```
resnet50_optimizer = optim.Adam(resnet50_model.parameters(), lr=0.0005)
efficientnet_b0_optimizer = optim.Adam(efficientnet_b0_model.parameters(), lr=0.0005)
inception_v3_optimizer = optim.Adam(inception_v3_model.parameters(), lr=0.0005)
```

- 데이터셋 증강 시행 <- 이후 설명
- 데이터셋 크기가 증가함에 따라 학습 메모리 초과를 우려해 배치 사이즈를 높였고,
  큰 배치사이즈와 작은 러닝레이트의 조합은 과적합을 유발할 수 있기에 러닝레이트를 0.0005로 조정하였음

3. 모델 및 학습 방법

튜토리얼 모델과의 차이점

```python
# ResNet50 모델
resnet50_model = resnet50(weights=ResNet50_Weights.DEFAULT)
resnet50_model.fc = nn.Sequential(
    nn.Dropout(p=0.5),
    nn.Linear(2048, 2)
)
# EfficientNet-B0 모델
efficientnet_b0_model = efficientnet_b0(weights=EfficientNet_B0_Weights.DEFAULT)
efficientnet_b0_model.classifier = nn.Sequential(
    nn.Dropout(p=0.5),
    nn.Linear(1280, 2)
)

# Inception-v3 모델
inception_v3_model = inception_v3(weights=Inception_V3_Weights.DEFAULT)
inception_v3_model.fc = nn.Sequential(
    nn.Dropout(p=0.5),
    nn.Linear(2048, 2)
)
```

- 다양한 종류의 모델을 사용하였고, 각 모델들의 형태에 맞게 마지막 선형변환 레이어 입력헤드 변경.

이후 각각 epoch 10으로 학습, 앙상블 진행

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4. 정확도 향상을 위한 적용 기법들

1. 데이터 증강
   - 주파수 마스킹과 시간 마스킹을 사용하여 Melspectrogram 이미지에 데이터 증강을 적용

2. 드롭 아웃
   - 모델의 과적합 문제를 해결하고 일반화 능력을 향상
   - 50%의 확률로 뉴런을 제거하도록 설정하여 모델이 특정 뉴런에 과도하게 의존하지 않도록 설정

3. 앙상블
   - 세가지 모델 ResNet50, Efficient_B0, Inception-v3를 앙상블하여 예측 결과를 결합
   - 각 모델의 예측 확률에 가중치를 적용하여 최종 예측을 수행

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.1. 데이터 증강

Spectrogram -〉 Melspectrogram 변경

- 인간은 500~1000Hz 간 차이는 잘 인지, 10,000Hz~10,500Hz 간 차이는 인지 어려움
- '인간은 주파수 스케일을 선형으로 인식하지 못한다'는 점에 착안

➔주파수를 로그함수로 이루어진 Mel -filter에 통과

➔Mel-scale로 바꾼 후 Spectrogram을 추출하여 Melspectrogram을 획득.

- Librosa와 torch.auio 라이브러리에서 Mel-scale 변환을 해주는 메서드를 활용하여 변환

```
melspectrogram = T.MelSpectrogram(
    sample_rate=sample_rate,n_fft=2048, hop_length=512, n_mels=128)
mel_spec = melspectrogram(waveform)
```

- n_fft: 주파수 해상도, hop_length는 푸리에 변환 시 윈도우가 겹치는 정도, N_mels는 mel filter의 수

```
spectrogram = T.Spectrogram(n_fft=512)
spec = spectrogram(waveform)
```

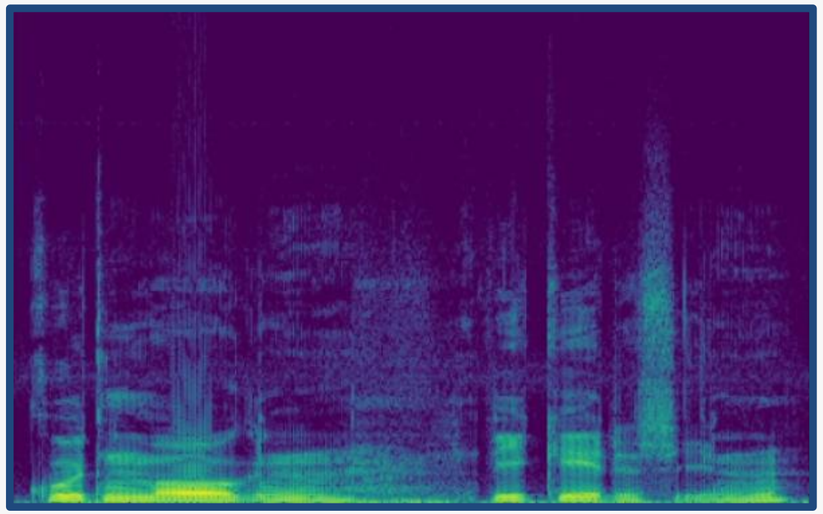
- 기존의 T.Spectrogram에서 melspectrogram을 호출하는 것 만으로도 형 변환 가능

3조
12192960 김승준
12202554 김지은
12215366 김기훈

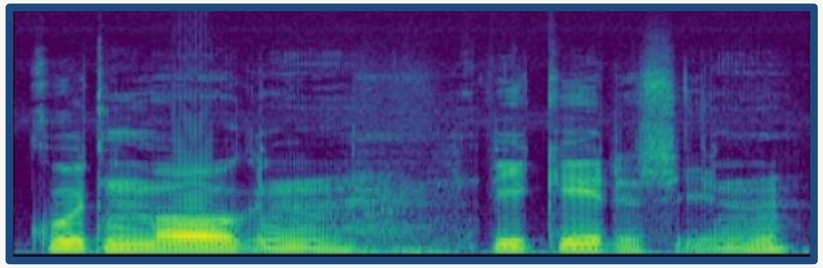
4.1. 데이터 증강

Spectrogram -> Melspectrogram 변경

- 기존 Spectrogram과 Melspectrogram간 형태 차이: 저주파 영역, 고주파 영역의 선명도 차이
- 파라미터는 동일하게 유지



Spectrogram

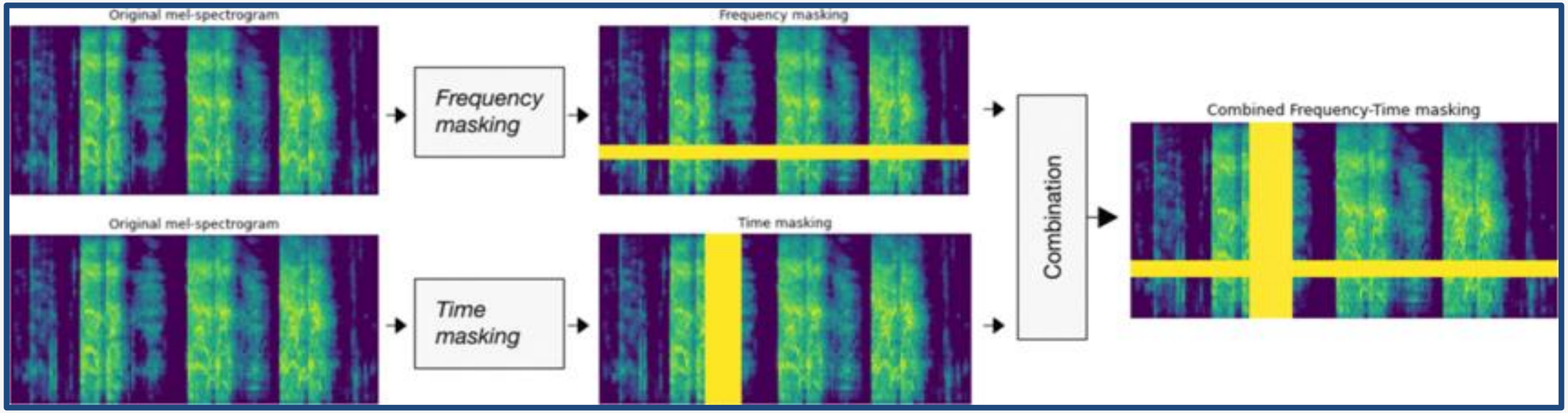


Melspectrogram

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.1. 데이터 증강

마스킹



https://www.researchgate.net/publication/359996671_Attention-based_hybrid_CNN-LSTM_and_spectral_data_augmentation_for_COVID-19_diagnosis_from_cough_sound

- 기존 음성 데이터를 변형하여 새로운 학습 샘플을 생성
- 모델의 일반화 능력을 향상
- 학습 데이터에 과도하게 최적화되는 과적합 문제를 완화

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.1. 데이터 증강

1. 최종 학습 데이터셋 크기

- 증강 전 데이터셋 크기
  - Train set: Healthy 532개, Pathology 762개
  - Test set: Healthy 100개, Pathology 100개
- 증강 후 데이터셋 크기
  - Train set: Healthy 1,064개, Pathology 1,524개
  - Test set: Healthy 100개, Pathology 100개

*증강 기법 적용으로 학습 데이터 크기 약 2배 증가

```
# 파일 개수 확인
train_healthy_images = list(glob.glob('./SVD/melspectrograms/train/healthy/*.png'))
train_pathology_images = list(glob.glob('./SVD/melspectrograms/train/pathology/*.png'))
test_healthy_images = list(glob.glob('./SVD/melspectrograms/test/healthy/*.png'))
test_pathology_images = list(glob.glob('./SVD/melspectrograms/test/pathology/*.png'))
print(f'train healthy   : {len(train_healthy_images)} images')
print(f'train_pathology : {len(train_pathology_images)} images')
print(f'test_healthy    : {len(test_healthy_images)} images')
print(f'test_pathology  : {len(test_pathology_images)} images')
```

```
train healthy   : 1064 images
train_pathology : 1524 images
test_healthy    : 100 images
test_pathology  : 100 images
```

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.2. 드롭 아웃

학습 과정에서 랜덤하게 선택된 뉴런들을 일시적으로 제거하는 기법

- 제거된 뉴런들은 해당 학습 단계에서 출력에 기여하지 않음
- 각 학습 단계마다 제거되는 뉴런들이 랜덤하게 선택됨

모델이 특정 뉴런에 과도하게 의존 하는 것을 방지

- 확률이 높을수록 더 많은 뉴런이 제거되어 규제 효과가 커짐
- 확률이 너무 높으면 학습이 불안정해질 수 있으므로 적절한 값 선택 필요

3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.3. 앙상블

ResNet50

- 50개의 레이어로 구성된 ResNet 아키텍처의 한 종류

Efficient_B0

- 2019년 구글에서 발표한 CNN 아키텍처
- ResNet에서의 잔류 학습과 같은 연산 방법은 변화를 주지 않으면서 채널의 수, 레이어의 수, 해상도에 변화를 줌
- 제한된 자원 내에서 변화를 주어 모델의 성능을 최대화 하는 방법

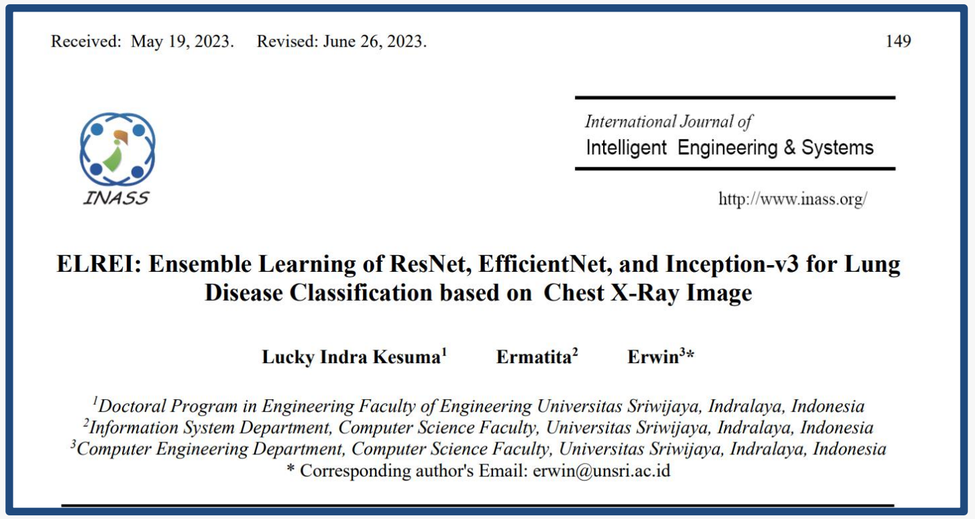
3조
12192960 김승준
12202554 김지은
12215366 김기훈

4.3. 앙상블

Inception-v3

- GoogLeNet에 변형을 가해 나온 모델 (Inception V1이 GooLeNet)
- 비대칭 컨볼루션 필터 사용 (3x1 컨볼루션과 1x3 컨볼루션을 조합, 다양한 형태의 특징으로 효과적으로 추출)
- RMSProp Optimizer 사용, Label Smoothing을 사용

앙상블 기법 도입



Received: May 19, 2023. Revised: June 26, 2023. 149

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

ELREI: Ensemble Learning of ResNet, EfficientNet, and Inception-v3 for Lung Disease Classification based on Chest X-Ray Image

Lucky Indra Kesuma[1] Ermatita[2] Erwin[3]*

[1]Doctoral Program in Engineering Faculty of Engineering Universitas Sriwijaya, Indralaya, Indonesia
[2]Information System Department, Computer Science Faculty, Universitas Sriwijaya, Indralaya, Indonesia
[3]Computer Engineering Department, Computer Science Faculty, Universitas Sriwijaya, Indralaya, Indonesia
* Corresponding author's Email: erwin@unsri.ac.id

https://inass.org/wp-content/uploads/2023/05/2023103114-2.pdf

3조
12192960 김승준
12202554 김지은
12215366 김기훈

5.1. Epoch 진행에 따른 train loss/accuracy 수렴 과정

1. ResNet50 모델

- Epoch 1: Loss 0.6110, Accuracy 63.12%
- Epoch 5: Loss 0.0571, Accuracy 97.81%
- Epoch 10: Loss 0.0248, Accuracy 99.06%
- Loss 감소 및 Accuracy 증가 추세 명확히 관찰

2. EfficientNet-B0 모델

- Epoch 1: Loss 0.0617, Accuracy 63.75%
- Epoch 5: Loss 0.0041, Accuracy 98.44%
- Epoch 10: Loss 0.0024, Accuracy 98.75%
- 초기 Epoch부터 높은 Accuracy 달성, Loss 빠르게 감소

3. Inception-v3 모델

- Epoch 1: Loss 0.6407, Accuracy 62.50%
- Epoch 5: Loss 0.0524, Accuracy 97.81%
- Epoch 10: Loss 0.0144, Accuracy 99.69%
- Epoch 진행에 따라 Loss 감소 및 Accuracy 증가 추세 뚜렷

3조
12192960 김승준
12202554 김지은
12215366 김기훈

5.2. 실험결과_1

```
weights = [0.25, 0.25, 0.5]

[[90 10]
 [21 79]]
accuracy: 0.8450, recall: 0.7900, precision: 0.8876, f1: 0.8360
```

Max Accuracy

Accuracy: 0.8450
Recall (Sensitivity): 0.7900
Precision: 0.8876
F1: 0.8360

0.8210
+- 0.0132
=
80.78~83.42

mean acc +- stdv

1차 Accuracy: 0.8050
2차 Accuracy: 0.8200
3차 Accuracy: 0.8200
4차 Accuracy: 0.8450
5차 Accuracy: 0.8150

```
accuracy = (tp + tn) / (tp + tn + fp + fn)
recall = (tp) / (tp + fn)
precision = (tp) / (tp + fp)
f1 = (2 * recall * precision) / (recall + precision)
```

3조
12192960 김승준
12202554 김지은
12215366 김기훈

5.2 실험결과_2

Confusion Matrix

- 첫 번째 행(Actual Healthy)
  - [90 10]
  - 실제 100개 Healthy 샘플 중,
    올바르게 예측 = 90 / 잘못 예측 = 10
- 두 번째 행(Actual Pathology)
  - [21 79]
  - 실제 100개 Pathology 샘플 중,
    잘못 예측 = 21 / 올바르게 예측 = 79

3조
12192960 김승준
12202554 김지은
12215366 김기훈

5.2. 실험결과_3

1차 Accuracy: 0.8050

2차 Accuracy: 0.8200

3차 Accuracy: 0.8200

4차 Accuracy: 0.8450

5차 Accuracy: 0.8150

```
[[72 28]
 [11 89]]
accuracy: 0.8050, recall: 0.8900, precision: 0.7607, f1: 0.8203
```

```
[[77 23]
 [13 87]]
accuracy: 0.8200, recall: 0.8700, precision: 0.7909, f1: 0.8286
```

```
[[74 26]
 [10 90]]
accuracy: 0.8200, recall: 0.9000, precision: 0.7759, f1: 0.8333
```

```
[[90 10]
 [21 79]]
accuracy: 0.8450, recall: 0.7900, precision: 0.8876, f1: 0.8360
```

```
[[83 17]
 [20 80]]
accuracy: 0.8150, recall: 0.8000, precision: 0.8247, f1: 0.8122
```

```
accuracy = (tp + tn) / (tp + tn + fp + fn)
recall = (tp) / (tp + fn)
precision = (tp) / (tp + fp)
f1 = (2 * recall * precision) / (recall + precision)
```

3조
12192960 김승준
12202554 김지은
12215366 김기훈

6. 결론

- **POINT 01. 모델 성능 요약**
  - ResNet50, EfficientNet-B0, Inception-v3 모델을 앙상블한 딥러닝 모델 제안
  - Mel-spectrogram 이미지를 입력으로 받아 Healthy와 Pathology 음성을 분류
  - 데이터 증강, 드롭아웃, 학습률 조정 등의 기법 적용을 통한 모델 성능 향상

- **POINT 02. 추가 개선 가능성**
  - 여전히 일부 샘플에 대한 오분류 발생
  - 데이터셋의 확장, 다양한 데이터 증강 기법 적용, 모델 아키텍처 최적화
  - 데이터셋 크기에 비해 학습 시 리소스 소모가 큼

- **POINT 03. 향후 연구 방향**
  - 더 큰 규모의 데이터셋 -> 모델의 일반화 능력을 높이는 것이 중요
  - 설명 가능한 AI (XAI: Explainable AI) -> 모델의 예측 근거를 시각화하고 해석
  - 리소스 사용을 줄이고 성능을 유지하도록 모델 개량

3조
12192960 김승준
12202554 김지은
12215366 김기훈

질의 응답

고급기계학습 3조(김승준, 김지은, 김기훈)