JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 99-107, January 2026
https://doi.org/10.9708/jksci.2026.31.01.099

Early Classification of Parkinson's Disease Using Mel-Spectrogram Voice Analysis via Transfer Learning

Jeong Hun Park*, Young-Kyoon Suh**, Jeeyoung Kim***

*Student, Graduate School of Data Science, Kyungpook National University, Daegu, Korea
**Professor, School of Computer Science and Engineering, Kyungpook National University, Daegu, Korea
***Professor, Graduate School of Data Science, Kyungpook National University, Daegu, Korea

[Abstract]

Parkinson's disease is a progressive neurodegenerative disorder, and early diagnosis is critical for slowing its progression. Voice changes are particularly notable among early indicators, offering a non-invasive pathway for timely detection. However, most existing approaches rely on traditional machine learning methods such as Support Vector Machines (SVM) and Support Vector Regression (SVR), which often fail to capture complex vocal patterns and thus exhibit limited generalization performance. This study proposes a voice-based diagnostic framework for Parkinson's disease and related disorders. Voice recordings were transformed into Mel-spectrogram images and classified using deep learning models, including ResNet152V2 and DenseNet201. The dataset included not only Parkinson's disease but also clinically similar conditions such as essential tremor, multiple system atrophy, and tau-Parkinsonism, alongside healthy controls. Experimental results show that our deep learning models achieve high accuracy (92%) in distinguishing Parkinson's disease and related disorders from healthy individuals. These findings highlight the potential of voice-based deep learning approaches as a non-invasive, cost-effective tool for early diagnosis and clinical support in neurodegenerative disease management.

▸ **Key words**: Parkinson's disease, Voice Signals, Mel-Spectrogram, Deep learning, Transfer learning

[요 약]

파킨슨병은 대표적인 신경퇴행성 질환으로, 병의 진행을 늦추기 위해서는 조기 진단이 중요하다. 특히, 음성 변화는 질병의 초기 징후와 관련이 있어 이를 활용하면 비침습적이고 조기 진단이 가능하다. 그러나 기존 연구들은 지지 벡터 기계(Support Vector Machine), 지지 벡터 회귀(Support Vector Regression) 등과 같은 전통적인 기계학습 기법에 주로 의존하였으며, 복잡한 음성 패턴을 효과적으로 반영하기 어려워 일반화 성능에 한계가 있었다. 이에 본 연구에서는 파킨슨병 및 관련 질환의 음성 데이터를 활용한 음성 기반 진단 모델을 개발하였다. 음성 데이터는 Librosa 라이브러리를 이용하여 Mel-Spectrogram으로 변환한 후, ResNet152V2, DenseNet201 등 다양한 딥러닝 기반 분류 모델을 적용하였다. 데이터에는 파킨슨병 외에도 본태성 진전, 다계통 위축 등 유사 신경계 질환을 포함하였으며, 이를 바탕으로 각 모델의 분류 성능을 평가하였다. 실험 결과, 음성 데이터를 활용한 딥러닝 모델이 정상인과 질환군을 높은 정확도로 구분할 수 있음을 확인하였다. 본 연구는 비침습적이고 비용 효율적인 파킨슨병 조기 진단 시스템 개발 가능성을 제시하며, 향후 신경계 질환의 조기 탐지 및 진단 보조 시스템으로 활용될 수 있다.

▸ **주제어**: 파킨슨병, 음성 데이터, 멜 스펙트로그램, 딥러닝, 전이학습

• First Author: Jeong Hun Park, Corresponding Author: Young-Kyoon Suh, Jeeyoung Kim
*Jeong Hun Park (qkrwjdgns728@knu.ac.kr), Graduate School of Data Science, Kyungpook National University
**Young-Kyoon Suh (yksuh@knu.ac.kr), School of Computer Science and Engineering, Kyungpook National University
***Jeeyoung Kim (jeeyoungkim@knu.ac.kr), Graduate School of Data Science, Kyungpook National University
• Received: 2025. 10. 10, Revised: 2025. 10. 27, Accepted: 2025. 12. 29.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

파킨슨병(Parkinson's Disease)은 중추신경계의 퇴행성 질환으로, 주로 운동 기능 저하, 진전(Tremor), 근육 경직 등의 증상[1]을 유발한다. 특히, 파킨슨병 환자의 약 80~90%는 발화 시작의 어려움, 단조로운 음성, 강도의 변화 등 다양한 음성 이상 증상을 동반[2]하며, 이는 환자의 삶의 질에 직접적인 영향을 미친다. 그러나 현재 파킨슨병의 진단은 주로 신경학적 검사(임상 평가, 영상 검사 등)에 의존하고 있어 고가의 비용, 낮은 접근성, 그리고 조기 진단의 어려움이라는 문제점이 존재한다.

이러한 문제를 해결하는 방법으로, 음성 데이터를 활용한 비침습적(non-invasive) 진단 방식이 최근 주목받고 있다. 음성 데이터는 비교적 손쉽게 구할 수 있으며, 환자의 상태를 실시간으로 반영할 수 있는 특성을 가진다. 특히, 음성 신호를 멜 스펙트로그램(Mel-Spectrogram)과 같은 시각적 이미지로 변환하여 딥러닝 기반의 이미지 분류 모델에 적용하는 방법은 전통적인 음성 분석 방식보다 더 높은 정확도를 보일 가능성이 있다.

기존의 많은 연구는 주로 단순한 음성 특징 기반의 전통적인 기계학습 알고리즘(Support Vector Machine[3], Random Forest[4] 등)에 집중되어 있으며, 데이터셋 또한 파킨슨병 환자와 정상인의 이분법적인 구조에 머무는 경우가 많았다. 본 연구는 이러한 한계를 극복하고자, 파킨슨병 환자뿐 아니라 본태성진전(Essential Tremor), 다발성 신경계 위축(Multiple Systemic Atrophy) 등 유사 질환 데이터를 포함하여, 보다 실용적인 감별 진단 모델을 찾아내고자 한다. 또한 기존 연구에서 상대적으로 부족했던 딥러닝 기반 이미지 분류 기법을 적용하여, 다양한 구조의 CNN(Convolution Neural Networks) 모델(ResNet152, DenseNet201, EfficientNet 등)을 비교·분석하고자 한다.

본 연구는 음성 데이터를 멜 스펙트로그램으로 변환한 후, 다양한 딥러닝 모델을 통해 파킨슨병 및 유사 질환과 정상인을 효과적으로 분류하는 분류 모델을 제안함으로써, 향후 임상 진단 보조 시스템으로의 활용 가능성을 탐색하고자 한다.

본 연구는 다음과 같은 공헌을 갖는다.

- 파킨슨병뿐만 아니라 유사 신경계 질환을 포함한 실용적인 감별 진단 모델 제시하였으며,
- Cosine Annealing과 Focal Loss를 멜 스펙트로그램 기반 CNN 분류 문제에 적용하여, 동일한 입력 조건 하에서 학습 전략이 분류 성능과 안정성에 미치는 영향을 여러 CNN 모델을 통해 실험적으로 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 검토하고, 3장에서 데이터셋과 모델 설계에 관련하여 설명한다. 4장에서는 실험 결과를 제시하고, 5장에서는 본 논문의 결론을 이야기하고 한계점 및 추후 연구를 논한다.

II. Related Work

Loshchilov and Hutter[5]에서 제안된 학습율 스케쥴러인 Cosine Annealing은 학습률의 최댓값과 최솟값을 설정한 후, Cosine 함수를 사용하여 학습률을 조절하는 방법이다. 이 기법은 학습이 진행될수록 학습률이 점진적으로 감소하도록 하여, 모델이 Saddle Point에 빠지지 않게 돕고, 정체 구간을 빠르게 벗어날 수 있게 한다. 이러한 특성 덕분에 모델의 일반화 성능이 향상된다. 본 연구에 따르면, Cosine Annealing을 적용한 Wide Residual Network가 CIFAR-10과 CIFAR-100 데이터셋에 Cosine Annealing을 적용하지 않은 네트워크에 비해 각각 0.5%와 1.0% 높은 성능을 보였다는 실험 결과가 있다. 최근에는 periodic shift 방식을 적용해 학습률 곡선을 조정함으로써 기존 Cosine Annealing의 일반화 성능을 개선한 ps-CALR(Periodic-Shift Cosine Annealing Learning Rate) 기법이 제안[6]되었다.

Focal Loss는 클래스 불균형 문제를 해결하기 위해 RetinaNet 모델에 도입된 손실 함수다. 기존의 객체 탐지 모델들은 객체 영역(Positive)보다[7] 배경 영역(Negative)이 훨씬 많아 클래스 불균형 문제가 존재한다. 이에 따라 학습이 비효율적이며, 모델 성능이 저하될 수 있다. Two-Stage 모델들은 region proposal과 sampling heuristic을 사용하여 불균형 문제를 해결하지만, One-Stage 모델에는 이러한 방법이 적용되지 않았다. 이를 해결하기 위해 Focal Loss는 One-Stage detector에 적용할 수 있는 방식으로 제안되었다. 이 손실 함수는 easy example에는 가중치를 낮추고, hard example에는 가중치를 높여서 학습할 수 있도록 설계되었다.

실험에서는 ResNet 구조와 anchor를 결합한 RetinaNet을 설계하여 Focal Loss의 효과를 검토하였다. Cross Entropy Loss는 모든 샘플에 대해 동일한 가중치를 부여하지만, Focal Loss는 이를 개선하여 easy sample의 가중치를 줄이고, hard sample의 가중치를 강화할 수 있다. 이를 통해 극단적인 클래스 불균형 문제를 해결하고, 모델 성능을 극대화할 수 있었다. 또한, Focal Loss는 modulating factor와 tunable focusing factor

(γ)를 추가하여 더욱 정교한 학습이 가능하게 만든다. γ 값은 0에서 5까지 조정할 수 있으며, γ=0일 때는 기존의 Cross Entropy Loss와 동일한 효과를 가지고, 값이 커질수록 easy sample의 가중치가 줄어든다. 실험 결과, Focal Loss를 적용한 ResNeXt-101 기반 Feature Pyramid Network(FPN)이 γ=2.0, α=0.25일 때 최고의 성능을 보였다는 결과를 얻었다. 이후 연구에서는 Focal Loss의 개념을 확장하여, 객체의 존재 확률뿐만 아니라 바운딩 박스의 품질 정보까지 동시에 반영하는 GFL(Generalized Focal Loss)가 제안[8]되었다. GFL은 분류 점수에 IoU 기반 품질 정보를 포함함으로써 분류와 회귀 간 불일치를 줄이고, 추가적인 연산 비용 없이 다양한 Dense Object Detector에서 mAP(Mean Average Precision)를 향상시키는 성과를 보여주었다.

Gwak and Park[9]에서는 파킨슨병 환자의 음성데이터를 분석하기 위해 음성 신호를 스펙트로그램(Spectrogram) 형식의 이미지로 변환하는 방법을 사용하였다. 이 과정은 음성데이터를 비정형 데이터인 스펙트로그램 형태로 변환하여 딥러닝 모델에 적용할 수 있도록 하는 과정이다. 해당 연구에서는 librosa 라이브러리를 활용하여 음성데이터를 스펙트로그램으로 변환했으며, 이를 통해 파킨슨병 환자의 음성 특징을 시각적으로 표현하였다. 변환된 멜 스펙트로그램은 주파수 축을 멜 스케일(Mel-Scale)로 변환하여 인간의 청각 특성에 맞게 음성을 표현하고, 이를 이미지 형태로 변환하여 CNN과 같은 딥러닝 모델에 입력으로 사용할 수 있게 한다. 멜 스펙트로그램은 시간-주파수 영역에서 음성의 특징을 추출하여 모델이 학습할 수 있도록 하며, 파킨슨병과 같은 질병의 진단 정확도를 높이는 데 중요한 역할을 한다. 또한, 연구에서는 최고 주파수를 설정하는 파라미터인 fmax를 1000으로 설정하여 주파수 범위를 제한하고, 파킨슨병 환자의 음성에서의 특징적인 변화를 더욱 강조할 수 있도록 하였다. 스펙트로그램을 이미지 형식으로 변환한 후, 모델 학습에 적합한 크기로 조정하여 입력 데이터로 사용하였다. 이를 통해 음성신호의 주파수 및 시간적 특성을 시각적으로 표현할 수 있었으며, 환자와 일반인의 음성 차이를 분석하는 데 활용하였다.

최근 Wav2Vec2[10], Whisper[11]와 같은 트랜스포머 기반 음성 모델이 제안되었다. 반면 이러한 모델들은 대규모 사전학습을 기반으로 하며, 특히 사전학습 단계에서는 상당한 연산 자원이 요구된다. 본 연구의 멜 스펙트로그램 기반 CNN 접근법은 비교적 적은 계산 자원으로 학습이 가능하다. 기존 영상 분류용 CNN 구조를 큰 수정 없이 활용할 수 있다. 본 연구는 CNN 기반 분류 방식과 학습 전략의 효과를 검증하는 데 목적이 있다.

III. The Proposed Scheme

3.1. Mel-Spectrogram Transformation

음성 데이터를 멜 스펙트로그램[12]으로 변환하는 과정은 음성 신호를 주파수 영역으로 변환하여 시각적 특징을 추출하고, 이를 딥러닝 모델에 입력할 수 있는 형식으로 변환하는 단계이다. 멜 스펙트로그램은 주파수 축을 멜 스케일(Mel-Scale)로 변환하여 인간의 청각 특성에 맞게 음성을 표현하며, 시간-주파수 영역에서 음성의 시간적 변화와 주파수의 특성을 동시에 파악할 수 있게 해준다.

멜 스펙트로그램 변환 과정은 다음과 같은 단계를 포함한다. 먼저 음성 신호는 프레임 단위로 나누어져, 각 프레임은 STFT(Short-Time Fourier Transform)을 통해 주파수 성분으로 변환된다. 이 과정은 오디오 신호를 시간-주파수 도메인으로 변환하는 과정으로 주파수 분석을 통해 음성의 다양한 특성을 추출할 수 있게 된다. 각 프레임의 크기와 오버랩은 시간 해상도와 주파수 해상도의 균형을 맞추는 데 중요한 요소이다. 다음으로 STFT에서 얻어진 주파수 성분을 Mel-Scale로 변환하는 과정은 인간의 청각 특성에 맞는 주파수 표현을 얻기 위한 것이다. 멜 스케일은 고주파 영역의 주파수 분해능을 줄이고, 저주파 영역의 주파수 분해능을 높여, 인간이 실제로 듣는 소리의 특징을 더 잘 반영하도록 한다.

이 과정은 주파수 축을 압축하여 더 직관적인 음성 분석을 가능하게 한다. 마지막으로 주파수 성분의 진폭을 로그 스케일(Log-Scale)로 변환하는 것은 비선형적 특성을 강조하여, 소리의 크기에 따른 차이를 더 명확하게 반영하기 위함이다. 음성 신호에서 작은 진폭의 차이는 인간의 청각에서 중요하지 않게 여겨지므로, 진폭을 로그 스케일로 변환 함으로써 음성의 특징을 더욱 강조하고, 모델이 학습을 더 잘할 수 있도록 돕는다.

3.2. Data Preprocessing

본 연구에서는 원천 음성 데이터를 활용하여 딥러닝 모델 학습에 적합한 형태로 변환하는 일련의 전처리 과정을 수행하였다. 전처리 과정은 다음과 같은 단계로 진행된다.

첫째, 원천 데이터와 라벨링 데이터를 1:1로 대응시켜 환자별, 성별, 질병으로 분류된 폴더 구조를 기반으로 데이터 경로를 설정하였다. 이를 통해 정상인과 파킨슨병 및 관련 질환군을 분류하고자 하였다.

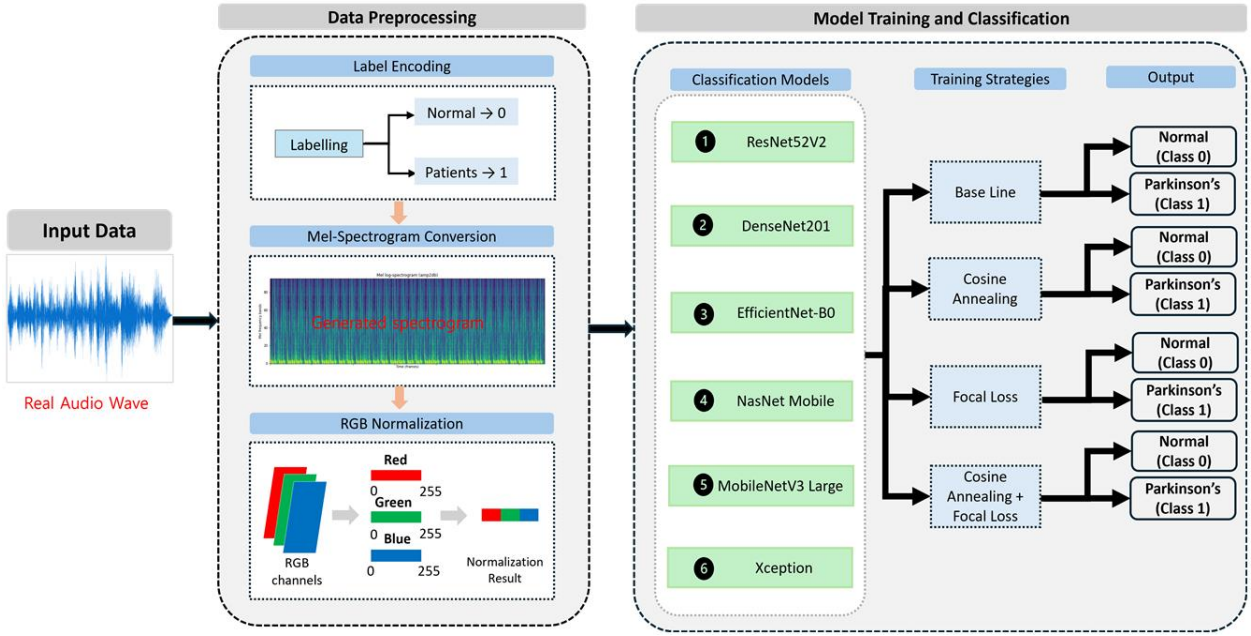

Fig. 1. Overall architecture of the proposed classification scheme

다음으로, json 파일 내 'annotation'필드를 참고하여 음성 파일에서 의미 있는 구간(발화구간)을 추출하였다. 이 과정을 통해 모델 학습 시 불필요한 잡음을 제거하고, 유의미한 음성 특징을 추출하고자 하였다.

이후, 분할된 음성 구간은 멜 스펙트로그램으로 변환하였다. 본 연구에서는 음성 신호의 스펙트럼 정보를 정밀하게 추출하기 위해 멜 스케일 기반 필터를 256개로 설정하였다. 이는 주로 사용되는 64~128개 필터에 비행 더 많은 주파수 구간을 설정하여 미세한 변화와 고주파 대역에서 나타나는 음향 특징을 효과적으로 포착하기 위함이다. 변환된 스펙트로그램은 librosa.display.specshow()함수를 사용하여 시각화한 후, 이미지 파일로 저장하였다. 또한, 이미지 분류 시 분류의 통일성을 위해 CNN(ResNet152V2[11], DenseNet201[12], EfficientNetB0[13], NasNet-Mobile[14], MobileNetV3-Large[15], Xception[16])모델에 입력할 수 있도록 모든 스펙트로그램 이미지는 224 x 224크기로 변환하여 입력 크기를 일정하게 맞추었다. 이 크기는 본 연구에서 활용하는 모델에 일반적으로 사용되는 이미지 크기로, 모델이 효율적으로 학습할 수 있는 크기이다.

본 데이터셋은 클래스 간의 불균형이 심하여, 해당 문제를 해결하고자 클래스별 가중치를 계산하였다. scikit-learn에서 제공하는 compute_class_weight를 사용하여 클래스 '0'(정상인), '1'(환자)에 대한 가중치를 산출하였고, 이후 모델 학습 단계에서 해당 가중치를 적용하여 편향을 최소화하였다. 마지막으로, 전체 데이터를 학습(Train, 21,354개)데이터와 테스트(Test, 5,339개) 데이터를 8:2 비율로 분할 하였으며, 검증(Validation, 1,286개)데이터는 별도로 제공되는 데이터를 활용하였다. 데이터는 모델 입력 전 RGB 채널로 변환 및 정규화를 수행하여 학습 효율성을 높이고자 하였다. 이를 통해 본 연구에서 활용한 CNN 모델이 음성 데이터의 시각적 특징을 안정적으로 학습할 수 있는 조건을 갖추었다.

IV. Experimental Results

본 논문에서 사용한 데이터는 한국지능정보사회진흥원(NIA)에서 운영하는 AI-Hub의 「파킨슨병 및 관련 질환 진단 음성데이터」[19]이다. 해당 데이터셋은 2022년에 구축된 음성 데이터셋으로, 표1과 같이 남성 음성데이터 12,598건(47.2%)과 여성 음성데이터 14,065건(52.8%) 총 26,693건의 음성데이터와 각 음성데이터 파일에 매칭되는 라벨링 데이터(json)로 구성되어 있다. 데이터셋 내 질환 종류는 본태성진전(Essential Tremor, ET, 원인을 모르는 자발성 떨림), 다발성 신경계 위축(Multiple Systemic Atrophy, MSA, 뇌 및 자율신경 기능에 문제가 생긴 퇴행성 신경 질환), 정상인(Normal Condition, NC), 파킨슨병(Parkinson's Disease, PD, 뇌흑질의 도파민계 신경이 파괴되어 나타나는 장애 질환), 타우 파킨슨(Tau-Parkinson's, 타우 단백질이 응집되어 나타나는 신경 질환)과 같이 총 5가지로 구분되어 있다. 각 질환 군별로 성별에 따른 분포 차이가 뚜렷하게 나타나며, 특히 남성과 여성 모두 파킨슨병 환자의 비율이 가장 높게 나타난다.

Table 1. Disease Distribution by Gender

Disease \ Sex	Male	Female
ET	1,817	3,734
MSA	1,625	1,383
PD	2,102	3,042
Tau-Parkinson	5,391	5,061
NC	1,663	875
Total	12,598	14,095

Table 2. Labeling Data Structure


```
{
  "Patient_No": "9-001",
  "TCF_Date": "2020-07-04",
  "Education": "Y",
  "Emotion": "Y",
  "Sex": "F",
  "Date_of_Birth": "1955",
  "Age": 65,
  "Genetic_DM": "X",
  "Speech_case": "N",
  "Severity_case": "M",
  "Underlying_disease_DM": "N",
  "Underlying_disease_Depression": "Y",
  "MoCA_Score": "25",
  "FDG_PET_Date": "",
  "FP_CIT_PET_Date": "",
  "TextData": [
    {
      "category": "A.01",
      "startTime": 0.94,
      "endTime": 6.98,
      "text": "무엇보다도 신체 오른쪽 다리와
              그 부위의 아름다움이 느껴진다"
    }
  ]
}
```


또한, 각 음성 데이터에 매칭되는 라벨링 데이터에는 메타데이터(metadata) 및 발화 정보가 포함된 주석(annotation)을 포함하고 있다. 메타데이터에는 표2와 같이 환자의 성별(Sex), 출생년도(Date_of_Birth), 질환 정보(Genetic_DX)뿐만 아니라 고혈압(HTN), 당뇨(DM), 고지혈증(Dyslipidemia)과 같은 합병증까지 주요 임상 정보가 메타데이터로 기록되어 있다. 어노테이션 정보에는 음성데이터의 번호(Category), 음성 구간(startTime, endTime), 텍스트(labelText)가 포함되어 있으며 환자의 실제 발화내용까지 상세하게 기록되어 있다. 또한, 한 명의 피실험자가 여러 개의 문장을 녹음하여 제공함으로써, 환자 개개인의 다양한 음성 패턴과 특성을 분석할 수 있다.

본 연구에서는 파킨슨병 진단을 위한 음성 데이터 분석에 다양한 딥러닝 모델을 적용하여 분류 성능을 비교하였다. 사용된 모델은 모두 깊은 합성곱 신경망(CNN) 기반의 대표적인 구조들로 다음과 같다. ResNet152V2, DenseNet201, EfficientNetB0, NasNet-Mobile, MobileNetV3-Large, Xception[13-18]이며, 각 모델은 서로 다른 아키텍처 설계 철학을 가지고 있어 음성 기반 파킨슨병 진단에 적합한 특성 추출 능력을 비교 및 검증하기에 적합하다. ResNet152V2는 잔차 학습(Residual Learning)을 통해 매우 깊은 구조에서도 기울기 소실 문제를 완화하여 복잡한 음향 특징을 안정적으로 학습할 수 있다[13]. DenseNet201은 레이어 간 조밀 연결을 기반으로 특징 재사용을 극대화하여 적은 파라미터로도 효율적인 학습이 가능하다[14]. EfficientNetB0와 MobileNetV3-Large는 경량화 모델로서, 적은 계산량으로도 우수한 성능을 제공하여 실제 임상 환경에서의 적용 가능성을 평가할 수 있다[15, 17]. 또한 NasNet-Mobile은 신경망 구조 탐색(NAS)을 통해 최적화된 네트워크 구조를 갖추고 있으며[16], Xception은 깊이별 분리 합성곱을 사용하여 주파수-시간 영역에서 나타나는 음성 특징을 효율적으로 분리하여 학습할 수 있다[18]. 따라서 본 연구에서는 다양한 네트워크 구조의 성능을 비교함으로써, 파킨슨병 진단에 가장 적합한 모델을 도출하고, 모델 선정에 대한 객관적인 근거를 제시하고자 하였다. 이들 모델은 음성 데이터를 멜 스펙트로그램으로 변환한 후, 각 모델의 학습 성능과 예측 정확도를 평가하는데 사용되었다.

학습 환경은 NVIDIA Tesla V100 GPU(32GB RAM)를 사용하여 수행하였고, 소프트웨어 환경은 Ubuntu 22.04, Python 3.11.0, TensorFlow 2.15.0으로 구성하였다. 이를 통해 대규모 이미지 데이터를 효율적으로 처리하고, 모델 학습 속도를 향상할 수 있었다. 학습 시 사용된 주요 하이퍼파라미터는 표3과 같다. 분류 작업은 각 모델의 입력에 적합한 이미지 크기(224 x 224 x 3)로 크기 조정하여 활용하였으며, 전이 학습 기반의 특징 추출을 위해 사전학습 된 가중치를 활용하였다. 이후 분류 작업을 수행하였으며, 평가 지표로는 Accuracy(정확도), Precision(정밀도), Recall(재현율), AUC를 활용하였다. 본 연구에서 Precision 및 Recall은 Parkinson(PD) 클래스를 Positive(1)로 정의하여 산출하였다. 즉, Precision은 모델이 Parkinson으로 판단한 샘플 중 실제로 Parkinson인 비율을 의미하며, Recall은 실제 Parkinson 대상 중 모델이 정확히 탐지한 비율을 나타낸다.

Table 3. Hyperparameters Used in our Model

	Notes
Optimizer	AdamW
Loss	Focal Loss(α=0.25, γ=2.0)
Learning Rate Scheduler	Cosine Annealing
Epochs	100
Batch Size	32
Class Weight	적용

실험은 ① 초기상태(Baseline), ② Cosine Annealing 적용, ③ Focal Loss 적용, ④ Cosine Annealing + Focal Loss 적용을 포함한 4가지 조건에서 이루어졌다(표 4-7). 표 4에서 보는 바와 같이, 초기 상태에서는 DenseNet201이 정확도 0.9024, 정밀도 0.9398, 재현율 0.9429, AUC 0.8672로 전반적인 성능 측면에서 가장 우수한 결과를 나타냈다. 또한, Xception도 정확도 0.8874, 정밀도 0.9285로 뒤를 이었으며 ResNet152V2의 경우 AUC는 다른 모델에 비해 상대적으로 낮았으나, 정밀도와 재현율의 경우 각각 0.9159와 0.9108로 균형 있게 높았다. 반면 MobileNetV3-Large는 정확도 0.7872, 재현율 0.7941로 다른 모델 대비 낮은 성능을 보였으며, 이는 경량화 모델 구조의 한계로 보임을 알 수 있다.

표 5은 Cosine Annealing을 적용한 실험 결과를 나타낸다. 전반적인 정확도 및 재현율의 향상이 관찰되었다. 특히, Xception의 경우 정확도가 0.9030으로 상승하였으며, 재현율 또한 0.9529로 크게 향상되었다. DenseNet201은 정확도 0.9088, 재현율 0.9557, AUC 0.8497로 여전히 가장 안정적인 성능을 유지하였다. 이와 같은 결과는 Cosine Annealing이 학습 후반부에 모델이 local minimum 혹은 saddle point에 머무르지 않고 더 나은 최적점에 도달하도록 도와주는 역할을 수행하였기 때문으로 해석된다. ResNet152V2와 NasNet-Mobile 또한 정확도가 각각 0.8855, 0.8912로 상승하며, Baseline 대비 개선된 성능을 나타내었다.

표 6은 Focal Loss만을 적용한 실험 결과를 나타낸다. 정확도 외에도 재현율과 AUC 값에서 두드러진 개선이 관찰되었다. 특히, DenseNet201은 정확도 0.8867, AUC 0.8828로 전 실험 중 가장 높은 AUC 값을 기록하며, 불균형 데이터 문제에 효과적으로 대응함을 보여주었다. NasNet-Mobile의 정확도가 0.9564까지 향상되었으며, ResNet152V2 역시 재현율이 0.9506까지 향상되었다. 이는 Focal Loss가 일반적인 Cross Entropy Loss보다 학습 초기에 손실을 더 세분화 하여, 클래스 미분류 오류에 더 집중할 수 있도록 설계되었기 때문에, 환자의 발화 특성과 같이 상대적으로 특정 패턴을 잘 반영하여 분석한 것으로 파악된다.

표 7은 Cosine Annealing과 Focal Loss를 동시에 적용한 실험 결과를 보인다. 전반적으로, 가장 고른 성능 향상이 관찰되었다. DenseNet201은 정확도 0.9174, 재현율 0.9523, AUC 0.9034로 매우 높은 성능을 보여주었고, Xception 또한 정확도 0.9215, 정밀도 0.9220, 재현율 0.9501, AUC 0.8582로 종합적으로 최적의 성능을 보여주었다. ResNet152V2 역시 정확도 0.8906, 재현율 0.9529로 초기에 비해 명확한 개선이 이루어졌음을 확인하였다. 이러한 결과는 Cosine Annealing이 학습의 최적 수렴을 유도하고, Focal Loss가 소수 클래스에 대한 분류 능력을 보완한 결과로 볼 수 있으며, 두 기법의 결합이 상호보완적으로 작용했음을 알 수 있다. 종합적으로, 본 연구에서 사용한 Cosine Annealing 스케줄러와 Focal Loss는 독립적으로도 학습 안정성과 분류 성능 향상에 기여하였지만, 두 기법을 동시에 적용했을 때 가장 안정적인 효과를 보여주었다. 특히 DenseNet201은 모든 경우에서 높은 정확도와 재현율을 기록하며, 복잡한 음성 패턴에서도 뛰어난 일반화 성능을 유지하였다. Xception 역시 정밀도와 재현율의 상승 면에서 유의미한 개선을 보여주며 경량성과 성능을 모두 갖춘 현실 응용 측면에서 가능성을 보여주었다. 반면, MobileNetV3-Large는 경량화 구조의 이점을 가지고 있음에도 불구하고 전체적인 분류 성능에서 뚜렷한 한계를 드러냈으며, 이는 의료 데이터와 같이 세밀한 특징 추출이 중요한 환경에서는 경량 모델 활용 시 신중한 접근이 요구된다.

Table 4. Baseline Experiment Results

Model	Acc.	Prec.	Recall	AUC
ResNet 152V2	0.8599	0.9159	0.9108	0.8175
Dense Net201	**0.9024**	**0.9366**	**0.9429**	**0.8672**
Efficient NetB0	0.8758	0.9171	0.9309	0.8591
NasNet Mobile	0.8835	0.9252	0.9308	0.8483
Mobile NetV3 Large	0.7872	0.9321	0.7941	0.8624
Xception	0.8874	0.9285	0.9322	0.8516

Table 5. Result with Cosine Annealing

Model	Acc.	Prec.	Recall	AUC
ResNet 152V2	0.8865	0.9248	0.9355	0.8193
Dense Net201	**0.9088**	**0.9335**	**0.9557**	**0.8497**
Efficient NetB0	0.8803	0.9096	0.9457	0.8237
NasNet Mobile	0.8912	0.9254	0.9418	0.8277
Mobile NetV3 Large	0.8597	0.9028	0.9268	0.7860
Xception	0.9030	0.9293	0.9529	0.8392

Table 6. Result with Focal Loss

Model	Acc.	Prec.	Recall	AUC
ResNet 152V2	0.8786	0.9040	0.9506	0.8627
Dense Net201	**0.8867**	**0.9100**	0.9547	**0.8828**
Efficient NetB0	0.8408	0.8970	0.9070	0.8479
NasNet Mobile	0.8595	0.8807	**0.9564**	0.8245
Mobile NetV3 Large	0.8164	0.9074	0.8617	0.8243
Xception	0.8578	0.9066	0.9194	0.8464

Table 7. Result with Cosine Annealing + Focal Loss

Model	Acc.	Prec.	Recall	AUC
ResNet 152V2	0.8906	0.9156	0.9529	0.8556
Dense Net201	0.9174	**0.9649**	0.9329	**0.9094**
Efficient NetB0	0.8833	0.9154	0.9431	0.8497
NasNet Mobile	0.9026	0.9353	0.9442	0.8756
Mobile NetV3 Large	0.8620	0.8970	0.9364	0.8162
Xception	**0.9215**	0.9220	**0.9901**	0.8582

결론적으로, 파킨슨병 환자의 음성데이터 분류를 위한 딥러닝 모델 설계에 있어 단순한 모델 선택뿐만 아니라, 적절한 학습 스케줄링과 손실 함수 설계가 모델의 성능을 결정짓는 핵심 요인이 될 수 있음을 본 연구는 실험적으로 보여주었다. 특히 DenseNet201 모델에 Cosine Annealing과 Focal Loss를 동시에 적용한 경우와 Xception 모델에 Cosine Annealing과 Focal Loss의 조합은 복잡하고 비선형적인 음성 특성을 효과적으로 분류하며, 향후 조기 진단 시스템 구축에 있어 기반 모델로 활용될 수 있을 것으로 기대된다.

V. Conclusion and Future Work

본 연구에서는 음성 데이터를 멜 스펙트로그램으로 변환한 후, 다양한 CNN 기반 딥러닝 모델과 Focal Loss 및 Cosine Annealing 학습률 스케줄러를 적용하여 파킨슨병 및 유사 질환과 정상인을 분류하는 모델을 제안하였다. 일부 모델에서 정확도, AUC와 같은 주요 지표가 향상된 것을 확인하였으며, 이는 음성 데이터 기반 모델의 파킨슨병 진단 가능성을 암시하는 것으로 사료된다. 따라서 본 연구는 제한적인 환경에서도 음성 데이터를 활용한 신경계 질환 분류 모델 구축이 가능함을 실험적으로 보여주었다는 점에서 그 의의가 있다.

본 연구의 한계점은 다음과 같다. 멜 스펙트로그램 생성 시 멜 필터 수, window size, hop length를 고정하여 사용하였으며, 해당 파라미터 변화에 따른 성능 민감도 분석은 수행하지 않았다. 이는 동일한 입력 조건에서 CNN 구조와 학습 전략의 효과를 비교하는 데 연구의 목적을 두었기 때문이다.

또한 본 연구는 정상인과 질환군을 구분하는 이진 분류 문제에 한정하였다. 질환 간 음성 특성이 유사한 경우에 대한 세부 오류 분석은 포함하지 않는다. 실제 임상 적용에 있어서는 본태성 진전, 다발성 신경계 위축, 타우파킨슨 등 서로 증상이 유사한 질환 간의 감별이 더욱 중요하게 요구될 것으로 판단된다. 이에 각 질환에 대한 다중 분류, 성별, 연령대 및 개인별 발화 특성에 따른 성능 분석은 향후 연구에서 수행할 예정이다.

추후 연구는 다음과 같이 진행될 예정이다. 먼저, 데이터 증강 등의 기법을 고려하여 다중 분류 체계를 도입한 다른 유사 질환 간의 구분 성능까지 본 연구를 확장할 계획이다. 또한, 최근 자연어 처리 및 음성 분야에서 많이 사용되는 트랜스포머(Transformers)[20] 기반 모델에 관한 연구도 계획할 예정인데, 적용 시 특히 시간적 문맥 정보를 반영할 수 있는 구조적 특성을 바탕으로 본 연구에서 수행한 CNN 기반 모델 대비 미세한 음성 변화 감지에 보다 유리할 것으로 판단된다. 마지막으로, 더 많은 실-세계 데이터셋을 활용하여 본 연구 결과의 현실적 적용 가능성을 높이고자 한다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00242528), the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2024-RS-2024-00437756) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and the Basic Research Program funded by the Ministry of Education of Korea and the National Research Foundation of Korea(NRF) (No. RS-2018-NR031059).

REFERENCES

[1] G. Yu, I. Jang, L. Kim, "Voice Handicap Index and Voice-Related Quality of Life in Idiopathic Parkinson's Disease" *Journal of Oriental Neuropsychiatry*, vol. 24, no. 2, pp. 155-162, Jun. 2013. http://dx.doi.org/10.7231/jon.2013.24.2.155

[2] L.A. Mahler, L.O. Ramig, C. Fox, "Evidence-based treatment of voice and speech disorders in Parkinson disease" *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 23, no. 3, pp. 209-215, Jun. 2015. https://doi.org/10.1097/MOO.0000000000000154

[3] R. Prashanth, S. Dutta Roy, P.K. Mandal, S. Ghosh, "High-accuracy detection of early Parkinson's disease through multimodal features and machine learning" *International Journal of Medical Informatics*, vol. 90, pp. 13-21, Jan. 2016. https://doi.org/10.1016/j.ijmedinf.2016.03.001

[4] W. Wu, J. Lee, F. Harrou, Y. Sun, "Early detection of Parkinson's disease using deep learning and machine learning" *IEEE Access*, vol. 8, pp. 147635-147646, Aug. 2020. https://doi.org/10.1109/ACCESS.2020.3016062

[5] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts" *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, pp. 1-12, Toulon, France, Apr. 2017. https://openreview.net/pdf?id=Skq89Scxx

[6] O. V. Johnson, X. Chew, K. W. Khaw, and M. H. Lee, "ps-CALR: Periodic-Shift Cosine Annealing Learning Rate for Deep Neural Networks," *IEEE Access*, vol. 11, pp. 139171-139186, Dec. 2023, doi: 10.1109/ACCESS.2023.3340719.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection" *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, Venice, Italy, Oct. 2017. https://doi.org/10.1109/ICCV.2017.324

[8] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized Focal Loss: Towards Efficient Representation Learning for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3139-3153, May 2023, https://doi.org/10.1109/TPAMI.2022.3180392.

[9] S. Gwak and K. Park, "Designing classification model using phonetic features and images of patients with Parkinsonism" *Proceedings of HCI Korea 2023*, pp. 1055-1060, Incheon, South Korea, Feb. 2023.

[10] A. Baevski, Y. Zhou, A. Mohamed, M. Auli. "Wav2Vec 2.0: A Framework for Self-supervised Learning of Speech Representations". *NeurIPs (Advances in Neural Information Processing Systems)*. 33, pp. 12449-60, 2020.

[11] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-scale Weak Supervision" *International Conference on Machine Learning*, pp. 28492-28518. July, 2023.

[12] Librosa, https://librosa.org, viewed on September 8, 2025.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 630-645, 2016.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700-4708, 2017.

[15] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105-6114, 2019

[16] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8697-8710, 2018.

[17] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314-1324, 2019.

[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251-1258, 2017.

[19] National Information Society Agency(NIA), AI-HUB, https://www.aihub.or.kr, viewed on September 8, 2025.

[20] S. Jeong, S. Kim, E. Lee, H. Kim, "Exploring spectrogram-based audio classification for Parkinson's disease: A study on speech classification and qualitative reliability verification," *Sensors*, vol. 24, no. 9, pp. 4625, May 2024. https://doi.org/10.3390/s24144625

Authors

Jeong Hun Park received his B.S. degree in Computer Science and Engineering from Yeungnam University, Korea, in 2015. He joined the Graduate School of Data Science at Kyungpook National University, Daegu, Korea, in 2023. His research interests lie in AI (Machine Learning, Deep Learning), Big Data Analysis and Visualization.

Young-Kyoon Suh received his M.S. and Ph.D. degrees from KAIST and the University of Arizona, in 2005 and 2015, respectively. He has been an Associate Professor in the School of Computer Science and Engineering at Kyungpook National University (KNU) since 2017. Prior to joining KNU, he was a Senior Researcher at KISTI and a Software Engineering Intern at Teradata Corporation. His research interests include databases, big data, machine learning, and healthcare informatics.

Jeeyoung Kim received her M.S. degree at the School of Computing of the University of Pennsylvania and Ph.D. degree at the School of Computing from University of Florida in 2006 and 2013, respectively. She worked as a project manager at Samsung Electronics Mobile division(2013~2018). She has been working as an assistant professor at the Graduate School of Data Science at Kyungpook National University since 2022, and her research interests include Healthcare AI, Machine Learning, Deep Learning and Human-Centered Data Science.