석사학위논문

인공지능 기반 상염색체 우성 다낭성 신장 질환 자동 분류 및 분석 연구

부 선 아

제주대학교 대학원

에너지응용시스템학부 전자공학과

2024년 2월

인공지능 기반 상염색체 우성 다낭성 신장 질환 자동 분류 및 분석 연구

이 논문을 공학석사 학위논문으로 제출함

부 선 아

제주대학교 대학원

에너지응용시스템학부 전자공학과

지도교수 도 양 희

부선아의 공학 석사 학위논문을 인준함

2023년 11월

심사위원장 고 석 준 ㊞

위 원 도 양 희 ㊞

위 원 김 영 우 ㊞

Thesis for the degree of Master of Engineering

Artificial Intelligence-based Automated Classification and Analysis of Individuals with Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Seonah Bu

Department of Electronic Engineering
The Graduate School
Jeju National University

February 2023

Artificial Intelligence-based Automated Classification and Analysis of Individuals with Autosomal Dominant Polycystic Kidney Disease (ADPKD)

A Thesis submitted to the graduate school of
Jeju National University in partial fulfillment of
the requirements for the degree of Master of Engineering
under the supervision of Yang-Hoi Doh

The thesis for the degree of Master of Engineering by
Seonah Bu
has been approved by dissertation committee.

February 2023

Chair Seok-Jun Ko

Member Yang-Hoi Doh

Member Youngwoo Kim

목 차

LIST OF FIGURES ·····

LIST OF TABLES ·····

초 록 ·····

Ⅰ. 서 론 ····· 1

Ⅱ. 관련 연구 ····· 5

1. 의료영상 분석 연구 현황 ····· 5

2. 인공지능 기반 컴퓨터보조진단(CAD) 연구 현황 ····· 6

3. 인공지능 기반 상염색체 우성 다낭성 신장 질환 연구 현황 ····· 9

Ⅲ. 연구 방법 ····· 11

1. 데이터 수집 및 전처리 ····· 11

2. 데이터 증강(Data Augmentation) ····· 14

3. 모델 설명 ····· 15

1) ResNet(Residential Network) ····· 15

2) ViT(Vision Transformer) ····· 19

4. 전이 학습 ····· 21

5. 평가 지표 ····· 22

6. 자동 분류 결과 확률 도출 ····· 24

7. 설명가능 인공지능 구현 방법 ····· 26

8. 실험 환경 ····· 27

Ⅳ. 연구 결과 ····· 28
1. 상염색체 우성 다낭성 신장 질환 자동 분류 결과 ····· 28
2. 확률 도출 결과 ····· 33
3. 설명 가능 인공지능 구현 결과 ····· 36

Ⅴ. 고 찰 ····· 38

Ⅵ. 결 론 ····· 40

참 고 문 헌 ····· 41

ABSTRACT ·····

LIST OF FIGURES

Fig. 1. Example MR images of autosomal dominant polycystic kidney disease 3
Fig. 2. Mayo imaging classification of ADPKD ····· 4
Fig. 3. Pictorial illustration of the preprocessing of MR images ····· 13
Fig. 4. Data augmentation examples ····· 15
Fig. 5. Difference between plain networks and residual learning ····· 17
Fig. 6. Differences between VGG-19, 34-layer plain, 34-layer residual architecture ····· 18
Fig. 7. Structural diagram of vision transformer architecture ····· 19
Fig. 8. Confusion matrix of automated classification of ADPKD with models 29
Fig. 9. ROC curves and AUCs of models ····· 31
Fig. 10. Visual plots of training time and accuracy by differed ResNet layers 32
Fig. 11. Three class 1 MR images examples correctly classified with higher probability ····· 33
Fig. 12. Four class 1 MR images that were correctly classified with relatively lower probability values ····· 34
Fig. 13. Three class 1 MR images mis-classified to be class 2 ····· 35
Fig. 14. Two class 2 MR images that were correctly classified with relatively lower class 2 classification probabilities ····· 35
Fig. 15. Visual representation of explainable artificial intelligence procedures 36
Fig. 16. Result of applying explainable artificial intelligence to class 2 MR images ····· 37

LIST OF TABLES

Table. 1. Training and test dataset of ADPKD cases used in this study ····· 11
Table. 2. Confusion Matrix ····· 22
Table. 3. Experimental Environment ····· 27
Table. 4. Classification accuracy of trained networks with models ResNet-18, ResNet-34, ResNet-50 and ViT ····· 28
Table. 5. Classification precision, recall and F1-scores of trained networks with models ResNet-18, ResNet-34, ResNet-50 and ViT ····· 30

인공지능 기반 상염색체 우성 다낭성 신장 질환 자동 분류 및 분석 연구

부 선 아

제주대학교 대학원 에너지응용시스템학부 전자공학과

요약

상염색체 우성 다낭성 신장질환(Autosomal dominant polycystic kidney disease, ADPKD)은 신장에 다수의 낭종이 생기는 질환으로, 한 번 발병하면 신장의 기능을 극도로 저하시켜 생명을 위협하는 유전 질환 중 하나이다. 이 질환의 근본적인 치료방법은 아직까지 없기 때문에, 환자의 영상 분석을 통한 정량적 예후 진단과 위험도 예측은 임상 관리 및 임상 시험에 매우 중요하다. ADPKD를 예측하는 기준 중 하나인 Mayo 영상 분류는 키 보정 총 신장 부피(height-adjusted total kidney volume)와 연령을 기반으로 위험도를 측정한다. 그러나 이는 전형적인 ADPKD 환자(typical case)인 class 1에만 적용할 수 있으며, 눈에 띄는 신장 외 낭종(exophytic cyst)이 있는 전형적이지 않은 환자의 경우(atypical case)인 class 2는 제외시켜야 한다. Class 1과 class 2의 분류는 일반적으로 전문의에 의해 수동으로 이루어지기 때문에 시간과 높은 집중력을 요구하고, 컨디션에 따른 판독자 내 차이 및 숙련도에 따른 판독자 간 차이 등으로 인해 판독 결과가 달라질 수 있다.

본 연구에서는 딥러닝을 기반으로 하여 ADPKD 환자의 MR 영상을 class 1과 class 2로 자동 분류하고, 그 결과에 대한 확률 및 설명 가능한 인공지능(Explainable Artificial Intelligence, XAI)을 활용하여 시각적으로 근거를 제시하는 방법을 제안한다. 이를 위해 HALT-PKD 연구에 참여한 486명의 ADPKD 환자의 MR 영상을 전처리 및 증강하여 데이터를 준비하였다. 우리는 자동 분류를 위해 ResNet(Residual Network)-18, 34, 50 및 ViT(Vision

Transformer)를 사용하여 딥러닝 모델을 훈련 및 테스트하고, 모델의 성능을 높이기 위해 ImageNet-1K 데이터세트에서 사전훈련 된 가중치를 사용한 전이학습을 적용하였다. 결과적으로 성능이 가장 우수한 ResNet-50에서 얻은 결과를 기반으로 소프트맥스(softmax) 함수를 사용하여 확률을 도출하고, XAI 기술을 활용하여 모델의 분류 결정에 대한 근거를 MR 영상 내에서 시각적으로 강조하여 나타내어 분류 결과의 진단 신뢰도를 얻을 수 있었다. 모델의 성능 평가를 위해서 오차 행렬(confusion matrix)과 ROC(Receiver Operating Characteristic, ROC)곡선 및 AUC(Area Under the Curve)를 활용하였다.

자동 분류 결과, ResNet-50 모델이 class 1을 97.7%, class 2를 100%, 평균 98.01%의 정확도를 나타내었고, class 1의 예측 정밀도, 재현율, F1-점수는 각각 1, 0.98, 0.99, class 2는 각각 0.87, 1, 0.93, 그리고 AUC는 0.99로 자동 분류에서 가장 우수한 성능을 보였다.

본 연구에서 제안한 완전 자동화된 분류 방법과 그에 대한 확률은 의사의 1차 판독 후, 2차 판독을 위한 객관적인 지표로 활용될 수 있으며, 확률이 모호한 경우 의사가 의료 영상을 다시 검토함으로써 진단의 정확성과 효율성을 높일 수 있다. 또한, XAI 기술을 기반으로 모델의 분류 결정에 대한 근거를 MR 영상 내에서 시각적으로 강조하여 나타내어 모델의 자동 결정에 대한 신뢰도를 향상시킬 수 있었다. 이러한 접근 방식은 ADPKD 환자의 임상 관리 및 임상 시험에서 효과적으로 사용될 수 있으며, 실제 의료 현장에서 진단 과정을 보다 간편하고 신뢰성 있게 지원함으로써 의료 전문가와 환자들에게 많은 도움을 줄 것으로 기대된다.

Ⅰ. 서 론

4차 산업의 주요 핵심 기술 중 하나인 인공지능은 머신러닝, 특히 딥러닝 기술을 활용한 자연어 처리, 음성인식, 시각인식 등 첨단기술을 개발하는 방향으로 발전해오고 있으며, 안전, 의료, 자동차 등 다양한 분야에서 적용되고 있다[1]. 딥러닝은 특히 이미지, 음성, 텍스트와 같은 복잡한 데이터를 처리하고 분석하는데 우수한 성능을 보이며, 이러한 능력은 의료 분야의 영상 분석에서 중요한 역할을 하고 있다.

의료분야에서 사용되는 X-ray, MR 영상, CT 등의 의료영상은 임상 분석, 의료 진단 및 수술 등을 위해 인체의 내부를 시각화한 데이터로, 환자의 건강상태를 판단하고 질병을 진단하는데 필수적인 정보를 제공한다. 이러한 의료영상은 여러 가지 특징을 가지고 있으며, 명도(brightness), 대조도(contrast), 공간주파수 (spatial-frequency), 균질성(homogeneity), 곡률(curvature), 길이(length) 등의 데이터를 통해 정량적으로 나타낼 수 있다. 각 병변은 앞서 설명한 특징으로 서로를 구분 짓게 하는 고유의 특징을 띄기 때문에 이 특징을 기반으로 전문가가 분석 및 진단을 하게 된다[2]. 그러나 의료영상 분석 및 진단은 많은 시간과 높은 집중력을 요구할 뿐 아니라, 주관적 판단과 전문가의 경험에 의존하므로 일관성과 정확성 면에서 한계를 가지고 있으며, 컨디션에 따른 판독자 내 차이(intra-rater variability) 및 숙련도에 따른 판독자 간 차이(inter- rater variability)로 판독 결과가 달라질 수 있다[3].

이러한 문제점을 해결하기 위해 컴퓨터 보조진단(Computer Aided Diagnosis, CAD) 기술이 등장하였다. 초기 의료영상에 컴퓨터 보조진단을 사용한 경우, 의료영상에서 의심되는 영역을 하나라도 놓치지 않기 위해 과도한 마킹을 사용하는 경향이 있었다(over-estimation). 그러나 이러한 접근 방식은 오히려 컴퓨터 보조진단을 사용하기 전 의사의 판독 정확도보다 낮은 결과를 초래하였다[3, 4]. 과도한 마킹, 즉 민감도(sensitivity)만 높고 특이도(specificity)는 낮은 컴퓨터 보조진단은 영상 진단에 더 많은 주관성을 부여하고, 실제 의료진의 판단을 방해한다. 뿐만 아니라 오히려 의료진으로 하여금 과도한

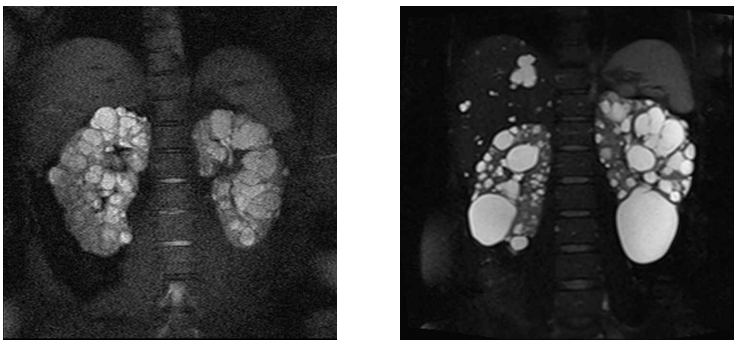
마킹을 확인하기 위한 추가적인 시간을 할애해야 하기 때문에 진단의 효율성을 떨어뜨리는 경우가 많았다. 이러한 문제를 완화하기 위해 다양한 딥러닝 기술이 도입되었다. 딥러닝을 통한 컴퓨터의 시각 인지 능력이 향상되면서[5], 의료영상에서의 컴퓨터 보조진단 기술은 점차 더 빠른 발전을 이루고 있다.

딥러닝 기반의 의료영상 분석은 주로 영상을 입력데이터로 활용하며, 이러한 분석에 특화된 심층 합성곱 신경망(Deep convolutional neural network, DCNN)이 가장 많이 활용되는 모델 중 하나이다[6]. 합성곱 신경망은 이미지의 공간적인 특징을 학습하고 이를 기반으로 판단 및 분류를 하기 때문에 의료영상에서 중요한 정보를 자동으로 추출하고, 질병이나 이상을 탐지하는데 도움을 준다. 이러한 딥러닝 기반 의료영상 분석은 다양한 의료영상에서 병변을 감지하고 진단하며, 의사들에게 빠르고 정확한 정보를 제공하여 진단과 치료에 있어 더 나은 결과를 얻을 수 있도록 도와주고 있다. 이러한 발전은 의료영상 분석 분야에서의 혁신적인 기술과 서비스를 가능하게 하며, 환자와 의료진 모두에게 보다 나은 혜택을 제공할 것으로 기대된다.

상술한 바와 같이 딥러닝 모델은 매우 복잡한 구조와 다양한 가중치 및 매개변수를 기반으로 모델 성능향상을 통한 의료영상 분석 분야의 발전에 기여하고 있지만, 한편으로 모델의 결정을 이해하기 어려운 블랙박스(Black-box) 문제를 야기한다. 특히 의료 분야에서는 의사의 판단이 환자의 건강과 생명에 직결되기 때문에, 모델의 판단 근거를 이해하는 것이 매우 중요하다. 이러한 이유로 설명 가능 인공지능(Explainable Artificial Intelligence, XAI)의 중요성이 더욱 강조되고 있다. XAI는 인공지능 모델의 의사결정 과정을 해석하고 설명할 수 있는 기술을 의미한다. 이는 모델이 왜 특정 결정을 내렸는지 또는 환자와 관련한 어떤 요소를 고려하였는지에 대한 명확한 설명이 가능하며, 이로써 의사와 환자 모두 모델의 판단을 이해하고 신뢰할 수 있게 된다. 이에 따라 의료 분야에서 XAI는 의료진들이 환자의 진단 및 치료 결정을 하는 모델을 신뢰하고 활용할 수 있도록 도움을 준다. 더불어, 이러한 기술은 의사들의 작업을 보조하고 의료진의 부담을 줄여주는 중요한 역할을 하므로 의료분야에서의 인공지능은 XAI를 통해 질병 진단 및 치료에 더 나은 결과를 제공할 것으로 기대된다. 이러한 발전은 환자의 안전과 의료 진행에 긍정적인 영향을 미칠

것이다.

본 논문에서 다루는 질환은 상염색체 우성 다낭성 신장 질환(Autosomal dominant polycystic kidney disease, ADPKD, Fig. 1. (a))으로 한쪽 또는 양쪽 신장에 다수의 낭종이 생기는 질환이다. 이는 한 번 발병하면 신장의 기능을 극도로 저하시켜 생명을 위협하는 유전 질환 중 하나로, 대략 1,000명 당 1명꼴로 발생하는 유전성 신장질환이다[7]. 정상적인 신장은 혈액 내 노폐물과 수분을 제거하는 역할을 하지만, ADPKD로 인해 신장의 실질조직(renal parenchyma)이 수많은 낭종으로 대체되면서 정상 신장 조직이 점차 줄어들게 된다. 이는 신장 기능의 저하를 일으켜 몸에 쌓이는 노폐물을 제대로 배설하지 못하는 만성 신부전 상태(End-Stage Renal Disease, ESRD)로 이어질 수 있다. ADPKD는 신부전의 네 번째 주요 원인으로, ADPKD 환자의 50% 이상이 50세 이전에 신부전이 발생하며, 결과적으로 투석이나 신장이식이 불가피하게 된다[8, 9].

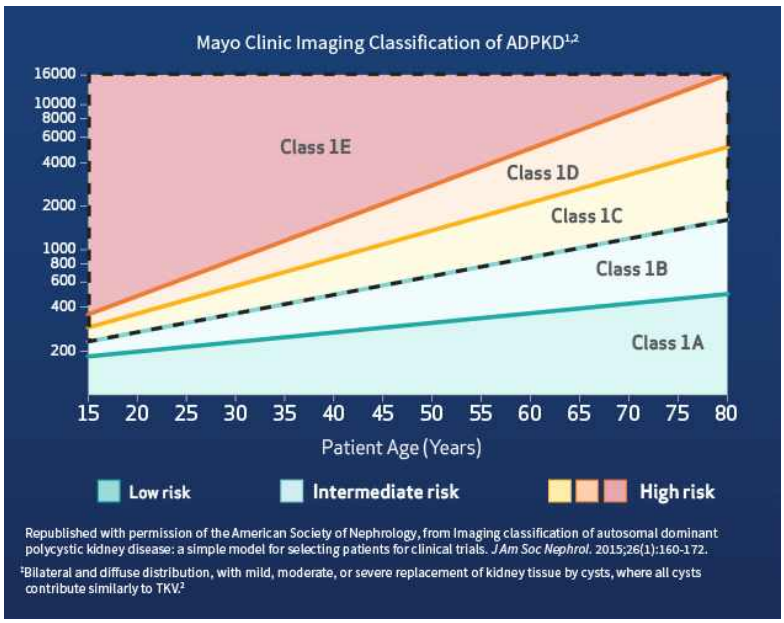


(a) (b)

Fig. 1. Example MR images of autosomal dominant polycystic kidney disease. (a) Typical case(class 1) and (b) atypical case(class 2)

ADPKD의 근본적인 치료법은 아직까지 없기 때문에, 질병 진행 평가부터 임상 관리 및 치료 시험에 이르기까지 ADPKD환자의 영상 분석을 통한 정량적 예후 진단은 중요한 역할을 한다. 특히 영상 분석을 통해 측정되는 총 신장의 부피(Total kidney volume, TKV)는 성별, 나이, 유전자형, 신장 기능 등 다양한

예후 인자 중에서도 ADPKD의 중증도 및 진행 정도를 평가하는 주요 지표이다. ADPKD를 예측하는 기준 중 하나인 Mayo 영상분류(Fig. 2)는 키 보정 총 신장부피(height-adjusted TKV, htTKV)와 연령을 기반으로 총 5개의 위험도를 측정할 수 있다[10]. 이는 전형적인 ADPKD환자의 경우(typical case)에만 해당되며, class 1에 해당된다(Fig. 1. (a)). 그러나 눈에 띄는 신장 외 낭종(exophytic cyst)이 있는 ADPKD환자는 전형적이지 않은 경우(atypical case)로, class 2에 해당되며(Fig. 1. (b)), 이런 환자의 경우에는 눈에 띄는 신장 외 낭종을 제외하고 재계산된 htTKV를 사용해야 되므로 Mayo 영상분류에 적용하기 전 추가 분류가 필요하다[11]. 이러한 과정은 일반적으로 판독 전문의가 주관적으로 판단하기 때문에 판독 결과의 오차에 따른 부정확성과 더불어 추가 시간이 소요되어 효율성에 문제가 생긴다. 따라서 본 논문에서는 딥러닝을 기반으로 하여 눈에 띄는 신장 외 낭종이 발생한 class 2를 자동으로 분류하고, 자동 분류 결과에 대한 확률 및 XAI를 활용하여 시각적으로 분류 결과에 대한 근거를 제시하는 방법을 제안한다.



- 출처 https://www.jynarquehcp.com/measuring-tkv-for-ckd-staging

Fig. 2. Mayo imaging classification of ADPKD

Ⅱ. 관련 연구

1. 의료영상 분석 연구 현황

X-ray, 초음파, CT, MRI, PET 등의 의료영상은 현대 의학에서 환자의 진단과 치료에 필수적인 도구 중 하나로 사용되며, 인공지능과 딥러닝이 등장하기 전부터 질병의 진단 및 치료 계획 수립에 폭넓게 활용되어 왔다. 이전에는 주로 전통적인 컴퓨터 비전과 영상 처리 기술을 바탕으로 의료영상 분석이 이루어졌다. 2000년대 이전 의료영상 분석 분야에서는 특정 질병의 검출 및 분류, 장기의 세부 구조 분할(segmentation), 이종의 의료기기에서 촬영된 의료영상 간의 정합(registration), 유사한 영상 검색(retrieval) 등과 같은 작업들이 주로 전통적인 방법에 의존하였다. 이러한 방법은 인체에 대한 해부학적 지식과 임상적 경험을 기반으로 하며, 비정형 데이터인 의료영상으로부터 특징을 추출하고 기본적인 패턴을 인식하기 위해 설계되었다[12, 13].

특정 질병의 검출 및 분류를 위해 주로 사용된 알고리즘으로는 서포트 벡터 머신(SVM), k-근접 이웃 알고리즘(k-nearest neighborhood algorithm, k-NN) 그리고 인공 신경망(Artificial neural networks, ANN)과 같은 기계학습 알고리즘 등이 일반적으로 활용되었다. k-NN 알고리즘은 새로운 데이터가 입력으로 들어올 때 해당 데이터와 가장 가까운 k개의 이웃 데이터들을 찾아서 클래스를 예측하는 방식이다. Park[14] 등은 이를 사용하여 유방 촬영 영상의 부분 영상에서 유방 종괴의 유무를 분류하는 시스템 개발하였으며, 양양정[15] 은 경동맥 플라그 의료영상 분석 및 텍스쳐 특징을 기반으로 뇌 중풍 예후를 분류하였다.

또한, Park[16] 등은 인공 신경망을 의료영상에 적용하여 CT영상에서 간질성 폐질환(Interstitial Lung Disease, ILD)을 초기에 검출하여 적절한 치료가 이루어지도록 진료를 돕는 시스템을 개발하였다. 인공 신경망은 입력과 결과로 이루어지는 쌍의 데이터를 이용하여 학습함으로써 전역최적함수(Global optimal

functional)에 근접하는 학습 방법이다[12]. 이는 데이터에서 자동으로 특징을 추출하고 패턴을 학습하므로 많은 작업을 자동화하고 빠르게 처리할 수 있지만, 대규모의 레이블링 된 데이터가 필요하며, 과적합(over fitting)으로 인해 새로운 데이터에 대한 적절한 일반화가 어려울 수 있는 문제점이 있다.

기계학습 알고리즘 중, 가우시안 혼합 모델(gaussian mixture model, GMM)은 가우시안 분포 여러 개를 혼합하여 복잡한 확률 분포를 나타내는 알고리즘이다. 이 모델은 주로 영상 내에서 픽셀의 화소 밝기 분포를 여러 개의 가우시안 함수로 모델링하여 사용되며, 영상 내에서 서로 다른 픽셀 그룹 또는 영역을 식별하고 분할할 수 있다. Woo[17] 등은 이를 사용하여 의료영상 분석에서 특정 병변이나 장기의 밝기 특징정보를 모델링하고 강화하여 영역을 분할하였고, 우상근[18] 등은 심장 극성지도(cardiac polarity map)에서의 경색 영역과 정상 영역을 분할 방법을 제안하였다.

이처럼 의료 분야에서 첨단 의료기기와 기술을 사용한 진단 및 치료는 오래전부터 이루어져 왔으며, 이러한 기술은 의료 현장에서 중요한 역할을 해왔다. 전통적인 기계 학습 기술은 의료영상에서 사용할 특징을 사전에 전문가가 수동으로 정의해야 하므로 직관적으로는 이해하기 쉽지만, 모든 특징을 포함시키기 어렵다는 한계가 있다. 또한 이러한 방식으로 학습된 모델은 새로운 데이터에 대한 일반화 능력이 제한될 수 있다[19]. 그러나 최근 딥러닝 기반의 기술이 부상함에 따라 이러한 한계를 극복하고 있다.

2. 인공지능 기반 컴퓨터 보조진단(CAD) 연구 현황

최근 몇 년 동안 의료 데이터의 양과 복잡성이 급격하게 증가하면서 기존의 방식으로는 데이터를 효과적으로 분석하고 활용하는 것이 더욱 어려워졌다. 특히 의료영상 데이터의 양이 급증함에 따라 기존의 의료 인력만으로 의료영상을 처리하기에는 상당한 시간과 노력이 요구되고 있으며, 의료영상 처리의 수작업으로 인한 오진율이 증가하는 등 여러 문제점이 발생하였다[20].

이러한 상황에서 인공지능 알고리즘, 특히 딥러닝 알고리즘은 의료영상 분석 분야에 큰 변화를 가져왔다. 딥러닝 개념은 여러 층으로 이루어진 인공

신경망으로 1970년대에 제안되었으나, 학습 계산의 복잡성 등으로 인해 활발한 연구가 어려웠다. 그러나 최근 몇 년 동안, 병렬처리를 위한 하드웨어인 그래픽카드(Graphic Processing Unit, GPU)의 성능 향상과 효율성 극대화를 위한 다양한 연구를 통해 딥러닝 아키텍처를 활용한 학습의 시간이 크게 줄어들고 성능이 대폭 향상되면서, 특히 영상 및 음성 인식 분야에서 광범위하게 활용되고 있다. 딥러닝은 전통적인 기계학습과 다르게 학습 과정에서 자체적으로 특징을 추출하고 학습하기 때문에, 영상의 종류에 상관없이 일반적인 모델링을 할 수 있으며, 단순한 특징 뿐만 아니라 상위 수준의 특징도 학습할 수 있다[19].

이를 활용한 컴퓨터 보조진단은 의사가 육안으로 의료영상을 분석해야 하는 부담을 줄여주며, 진료 시간을 단축시켜 환자에게 빠르고 효과적인 의료 서비스를 제공할 수 있다. 또한 이렇게 얻어진 정확한 분석 결과는 환자의 진단 및 치료에 대해 중요한 정보를 제공하여 의료 분야에 혁신적인 발전이 이뤄지고 있다.

김동현 등은 Inception 모듈 기반의 딥러닝 모델을 사용하여 위 내시경 영상을 이용한 위 병변 컴퓨터 보조 진단을 제안하였다. 137개의 비정상 위내시경 이미지와 112개의 정상 내시경 이미지를 포함하여 총 249개의 이미지를 취득하고, 의료 데이터 특성상 많은 데이터 확보가 어려운 문제를 해결하기 위해 이미지 회전을 통해 영상의 양을 증강하여 데이터를 구성하였다. 테스트를 위해 사용한 정상 이미지 68개 중 55개를 정상으로 분류하였고, 비정상 이미지 56개 중 38개를 비정상 이미지로 분류하여 민감도(sensitivity)와 특이도(specificity) 값은 각각 0.81과 0.68로 나타났으며, AUC 값은 0.832로 Inception 모듈의 기반으로 구성된 딥러닝 모델의 병변 진단 컴퓨터 보조 진단 시스템은 중등도의 성능을 보인다고 보고하였다[21].

이한성 등은 조기위암과 정상의 분류 뿐만 아니라 위염, 용종 등의 질환을 포함한 위 질환과 정상을 분류하고, 딥러닝 모델의 관심 영역을 시각화할 수 있는 Grad-CAM(Gradient-weighted Class Activation Map)을 활용한 컴퓨터 보조진단 시스템 연구를 수행하였다. 분류를 위해서는 EfficientNetV1을 백본으로 한 EfficientNetV2 모델과 성능 비교를 위해 Xception 네트워크를 사용하였으며, 조기 위암과 정상 분류, 비정상과 정상 분류 테스트에서

EfficientNetV2가 모든 평가 지표면에서 우수한 성능을 보였다. Cifar10 증강 기술을 적용하여 조기위암과 정상의 분류에서는 민감도가 0.878에서 0.966으로 약 10% 향상되었고, 비정상과 정상의 분류에서는 0.69에서 0.817로 약 18%의 성능 향상이 이뤄졌으며, AUC는 0.873에서 0.93으로 7%의 성능향상 결과를 보였다. 그러므로 증강기술이 딥러닝 모델의 분류 성능 향상에 효과적이며, Grad-CAM을 활용하면 의사의 라벨링이 없더라도 병변의 위치를 파악할 수 있다고 보고하였다[22].

이경윤 등은 위내시경에서 많이 발견되는 질환인 위궤양을 분류하기 위해 ResNet-50 딥러닝 모델을 활용하고, 딥러닝 모델의 관심 영역을 시각화할 수 있는 CAM(Class Activation Map)을 활용한 컴퓨터 보조진단 시스템 연구를 수행하였다. 정상 데이터 총 175장 중 170장을 정확히 분류하여 97.14%의 정확도를 보였으며, 위궤양 데이터는 총 130장 중 105장을 위궤양으로 정확히 분류하여 80.77%의 분류 정확도를 보였다. 또한 정밀도는 95.45%, 재현율은 80.77%, 그리고 F1-Score는 0.87이었으며, 0.97의 높은 AUC를 보였다. 정밀도에 비해 재현율이 다소 떨어졌으며, 오검출된 영상을 분석한 결과, 거품으로 인한 잡음, 위장의 주름, 그리고 미세한 위궤양으로 인해 육안으로도 구분이 어려운 경우로 확인되었다. 저자는 향후 추가 데이터 수집과 영상 개선을 한다면 모델의 성능을 더욱 높일 수 있을 것이라고 보고하였다[23].

Bressem 등은 CheXpert와 COVID-19 이미지 데이터 세트를 분류하기 위해 16개의 다양한 CNN 아키텍처를 사용하였다. CheXpert 데이터 세트에서는 심장비대(cardiomegaly), 부종(edema), 경화(consolidation), 무기폐(atelectasis), 흉막(pleural) 총 다섯 가지로 분류되었고, COVID-19 데이터 세트에서는 COVID-19와 일반 폐렴으로 분류되었다. 모든 모델은 CheXpert 데이터 세트에서는 0.83에서 0.89, COVID-19 데이터 세트에서는 0.983에서 0.998까지 높은 AUROC 성능을 보였다. 이 결과는 인공 신경망의 복잡성과 깊이를 증가시키는 것이 높은 데이터 분류 성능을 위한 필수적이지 않음을 보여준다[24].

3. 인공지능 기반 상염색체 우성 다낭성 신장 질환 연구 현황

상염색체 우성 다낭성 신장 질환(이하, ADPKD)은 신장에 무수한 낭종이 자라고 시간이 지남에 따라 TKV가 점진적으로 증가하는 유전적 질환으로, 이로 인해 다양한 합병증이 발생할 수 있다. 현재 ADPKD의 치료 목표는 낭종의 성장을 늦추는 것이기 때문에 조기 발견과 정확한 진단이 중요하다. ADPKD는 신장의 의료영상 검사를 통해 전문의가 판단해야 하므로 의료 전문가의 경험과 전문 지식이 필요하다. 그러므로 의사의 부담을 줄이기 위해 인공지능을 기반으로 한 ADPKD 관련 연구가 꾸준히 진행되고 있으며, 신장 MR 영상에서 다양한 응용 분야로 떠오르고 있다[25]. 이러한 연구는 의료영상을 분석하고 낭종을 식별하는데 도움을 주며, 초기 진단 및 모니터링 과정에서 의사들의 작업을 보조할 수 있다.

Bevilacqua 등은 MR 영상을 기반 ADPKD 분할을 할 때, 수작업으로 특징을 추출할 필요 없이 이미지의 의미론적 분할을 위해 CNN을 사용하는 딥러닝 구조기반 두 가지 방식을 제안한다. 첫 번째 방식은 입력 이미지의 전처리 없이 자동분할을 수행하는 방식이고, 두 번째 방식은 CNN이 자동으로 관심영역(Regions of interest, ROI)을 검출하여 분류기가 이에 대해 의미론적(semantic) 분할을 수행하는 방식이다. 이는 MR 영상에서 신장과 배경을 자동으로 분할해 주며, 두 가지 방법 모두 85% 이상의 정확도를 보여주었다. 저자는 제안한 방법이 신장 영상 분석의 복잡성을 줄이고 분할 작업을 수행하는데 필요한 작업의 부담을 덜어준다고 보고하였다[26].

Goel 등은 신장을 자동으로 분할하여 ADPKD의 중증도 척도인 TKV를 결정하기 위한 딥러닝 모델을 개발하였다. 이 모델은 ADPKD 환자 129명에 대한 213개의 복부 MR 영상을 사용하여 개발된 EfficientNet 인코더가 포함된 U-Net을 기반으로 하였다. 제안한 인공지능 파이프라인이 ADPKD 신장의 TKV 추정을 위한 자동분할을 정확하게 수행하였지만 체액이 찬 위, 방광 팽창, 출혈성 신장 낭종, 간과 오른쪽 신장 경계의 간 낭종과 같은 경우에서는 오류가 발생하였다. 그럼에도 불구하고 제안한 모델을 사용하면 수동으로 윤곽을 그리는 것에 비해 모델을 사용한 분할이 방사선 전문의의 시간을 51% 감소시켰다고

보고하였다[27].

ADPKD 환자의 신장 자동 분할과 더불어 8년 후 ADPKD 예후를 예측하기 위해 Raj 등은 MR 영상에서 신장을 분할하여 총 신창 부피를 계산하는 딥러닝 모델을 개발하고, 분할된 신장에서 이미지 특징을 추출해 이를 기존 바이오마커와 결합하여 8년 후 신장 기능 저하를 예측하였다. ADPKD 신장을 분할하기 위해 Attention U-Net을 사용하고, 예후를 예측하기 위해서는 CNN과 MLP(Multi-layer Perceptron)을 결합하여 사용하였다. 8년 후 환자의 만성 신장 질환(Chronic kidney disease, CKD) 단계를 명확하게 예측하도록 훈련된 모델의 정확도는 0.851, AUC는 0.972로 나타났다. 이러한 분류는 단순히 특정 CKD 단계에 도달할지 여부를 예측하는 것이 아니라 시간이 지남에 따라 CKD 단계의 변화와 그에 따른 신장 기능 저하를 정확하게 예측할 수 있으므로 ADPKD 환자의 진단을 더 잘 지원할 수 있다고 보고하였다[28].

Class 2에 해당하는 ADPKD 환자는 눈에 띄는 신장 외 낭종이 있으므로 TKV 계산 시 질병 진행 위험을 과대평가하지 않기 위해 이러한 낭종을 제외해야 한다[11]. 그러므로 Kim 등은 ADPKD 환자의 신장 외 낭종 부피를 제외하는 딥러닝 기반의 완전 자동화된 총 신장 부피 계산 방법을 개발하고 검증하였다. 3D U-Net을 사용하여 자동 분할을 수행하고, k-fold cross -validation을 통해 네트워크를 훈련하였다. 오른쪽 및 왼쪽 신장, 신장 외 낭종의 DSC(Dice similarity coefficient) 표준편차는 각각 0.9630±018, 0.963±0.018, 0.951±0.016이었으며, 평균은 0.962±0.018로 나타났다. 저자는 이 방법이 ADPKD의 진행을 평가하기 위해 TKV를 자동으로 계산해야 하는 임상 연구에 유용하다고 보고하였다[29].

이와 같이 ADPKD의 진단, 치료 및 관리를 위한 신장 및 낭종 분할과 정량적 평가 지표 등에 대한 연구들은 활발하게 이루어지고 있지만, 아직까지 ADPKD에서 class 1과 class 2를 완전 자동으로 분류할 수 있는 방법은 전무한 실정이다. 따라서 본 연구에서는 딥러닝 기반 ResNet과 ViT 모델을 활용하여 ADPKD 환자들의 MR 영상을 class 1과 class2로 완전 자동 분류를 하고자 한다. 더불어 모델의 결정에 대한 확률과 XAI를 기반으로 진단의 신뢰도 (Diagnostic Confidence)와 설명도(Explainability)를 높이는 방법을 제안한다.

Ⅲ. 연구 방법

1. 데이터 수집 및 전처리

본 연구에서 사용된 데이터는 HALT 연구에서 사용된 ADPKD 환자들의 MR 영상 데이터를 기반으로 하였다[30, 31]. 이 데이터는 신장을 포함한 복부를 앞(anterior)에서 뒤(posterior)로 촬영한 MR 영상으로, 각 데이터마다 슬라이스의 수가 다른 AVW 파일 형태를 가지고 있으며, 전문가들에 의해 class 1과 class 2로 분류되어 있다. 라벨링 된 전체 486개의 데이터 중에서 class 1은 426개, class 2는 60개로 나누어져 있으며, 이 중 학습용 이미지로는 각각 295개와 40개, 테스트용 이미지로는 각각 131개, 20개로 나누어 사용하였다. 본 연구에서는 class 2에 해당하는 훈련 및 테스트 데이터 세트가 상대적으로 적기 때문에, 3개의 독립된 훈련 및 테스트 데이터 세트로 나누어 진행하였다.

Table. 1. Training and test dataset of ADPKD cases used in this study

분류	훈련 데이터 세트 (Train dataset)	테스트 데이터 세트 (Test dataset)	합계
Class 1(Typical)	295	131	426
Class 2(Atypical)	40	20	60
합계	354	155	486

우선 AVW 파일을 NIfTI 파일로 변환하고, 변환한 NIfTI 파일 헤더에 기존 AVW 파일 헤더의 복셀(voxel)을 추가하였다. 그 후, 모델 훈련 시 입력으로 사용될 2D 이미지를 추출하기 위해 3D NIfTI 파일에서 앞쪽을 기준으로 65%에 해당하는 슬라이스 한 장을 선택하여 PNG 파일로 변환하였다. 이때, 신장이나 신장 외 낭종이 영상 내에 제대로 보이지 않는 경우에는 영상을 수동으로 확인하고 적절한 이미지를 선택하였다. 그다음으로 모든 데이터의 픽셀(pixel) 간격을 높이 0.5mm, 너비 0.5mm로 정규화하였다. 단순히 픽셀의 간격만

변경하면 이미지가 왜곡되거나 해상도가 손실될 수 있으므로, 기존 이미지 해상도에 픽셀 간격이 바뀌는 비율(기존 픽셀/목표 픽셀)을 곱하여 해상도를 비율에 맞게 조정하였다. 또한, 딥러닝 모델을 훈련시킬 때 기본적으로 입력 영상을 일괄적으로 같은 크기의 영상으로 변환하는데, 이 과정에서 위치정보에 대한 상대적 차이가 소실될 수 있다. 따라서, 입력 데이터의 크기를 표준화하기 위해, 입력 데이터 중 가장 큰 데이터(1240×1240)를 기준으로 제로 패딩(zero-padding)하였다. 만약 작은 데이터 크기를 기준으로 제로 패딩을 하면 큰 데이터의 정보가 부분적으로 손실될 우려가 있기 때문에, 가장 큰 데이터 크기를 기준으로 제로 패딩 하였다.

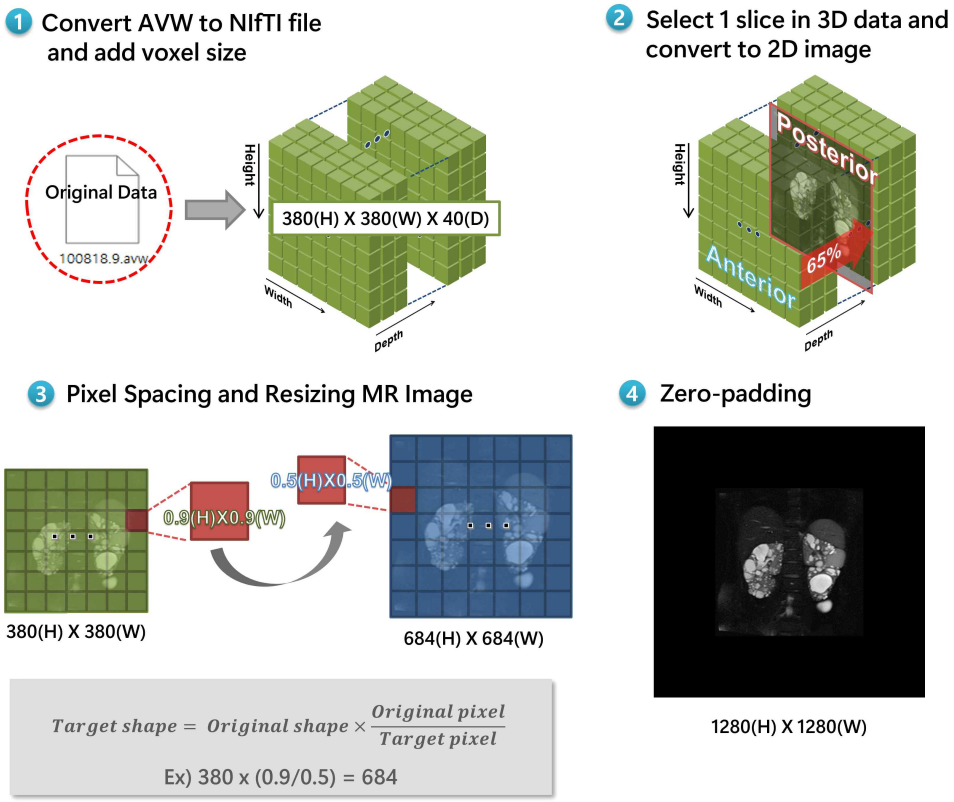


Fig. 3. Pictorial illustration of the preprocessing of MR images. A single representative mid-slice image is selected at 65% of the distance from the anterior to the posterior of the MR images covering kidneys. The selected MR images were resized using the pixel spacing and slice thickness. Zero padding was applied to the resized MR image.

2. 데이터 증강(Data Augmentation)

딥러닝은 대량의 데이터와 그에 대응하는 정답(label)을 동시에 활용하여 모델을 훈련시키는 방식으로 이루어지며, 데이터의 양이 증가함에 따라 모델의 일반화 성능이 향상된다. 이때, 훈련 데이터는 다양한 특징들을 포함하며, 모델이 학습해야 하는 특징들이 균형 있게 분포되어야 한다. 그러나 의료영상 데이터의 경우 개인 정보 문제와 같은 이유로 다량의 데이터를 수집하기 어렵고, 레이블 생성에는 의료 전문가들의 노력과 시간이 필요하기 때문에 많은 비용과 어려움이 따른다[32]. 뿐만 아니라 의료 데이터는 종종 특정 클래스나 결과에 대한 데이터가 다른 클래스에 비해 부족한 불균형한 형태를 지닐 수 있다. 이로 인해 모델이 훈련 데이터가 적은 클래스에 대해 적절하게 학습하지 못하고, 훈련 데이터가 많은 클래스의 성능만 향상되는 경향이 있다.

이처럼 다량의 데이터를 구하기 어려운 상황에서 데이터 증강(data augmentation)은 기존 데이터를 다양한 방법으로 변형하여 더 많은 훈련 데이터를 생성함으로써 모델의 성능 향상을 도와준다. 각 데이터 증강 알고리즘은 영상 내 특징들을 유지한 채로 데이터를 확장해야 한다. 데이터 증강은 일반적으로 훈련 데이터세트에만 적용되며, 검증 및 평가 데이터에는 적용하지 않는다.

본 연구에서는 수평 반전(horizontal flip), 수직 반전(vertical flip), 회전(rotation) 그리고 이미지의 시각적 특성을 변형시키는 Color jitter와 가우시안 블러(Gaussian blur)를 사용하여 데이터를 증강하였다. Color jitter는 이미지의 밝기, 대비, 채도를 조절하고, 색상을 변형하여 이미지를 다양하게 만드는 역할을 한다. 가우시안 블러는 이미지의 선명도를 줄이거나 노이즈를 제거하여 이미지를 부드럽게 만드는 데 사용된다. 실제로는 제로 패딩 전처리 과정을 거친 MR 영상에 데이터 증강을 적용하였으나, 데이터 증강을 더 명확하게 보여주기 위해 제로 패딩을 하지 않은 원본 MR 영상 Fig. 4. (a)를 기준으로 데이터 증강 적용 예시를 나타내었다. Fig. 4의 (b)와 (c)는 0부터 180도 범위에서 랜덤하게 회전을 적용한 결과이고, (d)는 수평반전, (e)는 수직반전, (f)와 (g)는 각각 color jitter와 가우시안 블러를 적용한 결과이다.

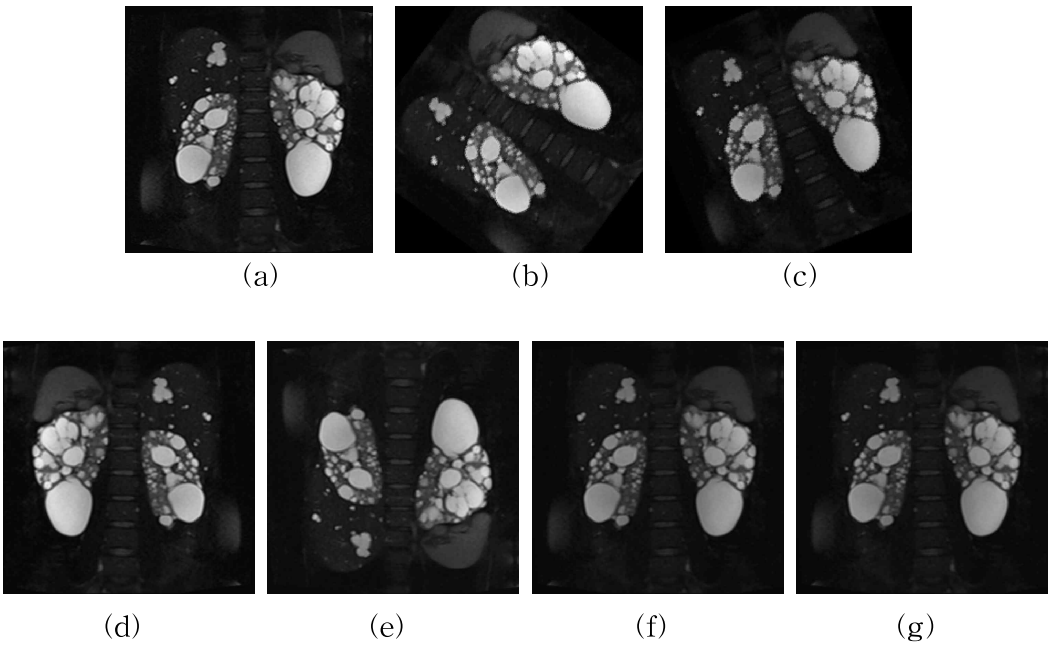


Fig. 4. Data augmentation examples. (a) original, (b) - 45 degrees of rotation, (c) - 150 degrees of rotation, (d) horizontal flip, (e) vertical flip, (f) Color jitter, and (g) Gaussian blur

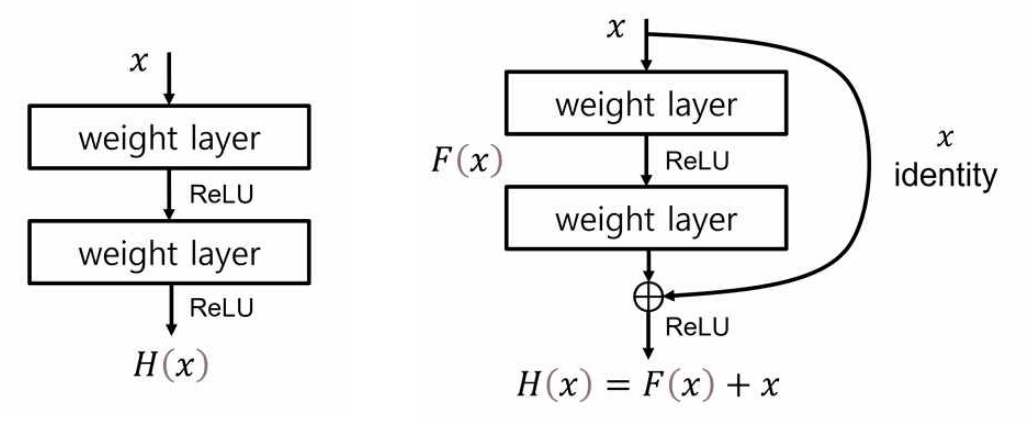
3. 모델 설명

1) ResNet(Residential Network)

딥러닝에서는 일반적으로 CNN(Convolutional Neural Network)의 깊이가 증가할수록 더 나은 성능을 기대할 수 있다. 그러나 '단순히 레이어를 늘리기만 하면 항상 성능이 향상되는가?'라는 의문이 제기되면서 ResNet(Residual Network)이 개발되었다. 이전에는 신경망을 깊게 쌓을수록 항상 더 좋은 성능이 기대되었지만, 20개와 56개의 레이어를 단순히 쌓은 신경망을 사용하여 실험한 결과, 깊이가 더 깊을수록 성능이 향상될 것이라는 예상과 달리 56개의 레이어를 가진 신경망에서 더 높은 오차가 발생한 것이 확인되었다[33]. 과적합(over-fitting)으로 인한 문제라면, 일반적으로 더 깊은 신경망이 상대적으로 낮은 훈련오차(training error)를 가져야 한다. 그러나 56개의 레이어를 가진 더 깊은

신경망이 훈련오차와 테스트오차 모두 더 낮은 결과를 보이고 있으므로 과적합의 문제가 아닌 것으로 볼 수 있다. 이는 기울기 소실 (gradient vanishing)과 폭주(exploding) 문제로, 심층 신경망의 깊이가 증가함에 따라 기울기가 사라지거나 너무 커지는 현상을 나타내며, 이로 인해 훈련이 어려워지게 된다.

이러한 문제점을 해결하기 위해 딥러닝 신경망의 깊이를 증가시키면서 기울기 소실 문제를 줄이고 더 효율적인 훈련을 가능하게 해주는 잔차 학습(Residual learning)이 제안되었다[33]. 기존의 방식(Fig. 5. (a))은 입력 데이터x가 각 레이어를 순서대로 통과하여 최종 출력 $H(x)$를 생성하는 것을 목표로 한다. 하지만 잔차 학습(Fig. 5. (b))은 입력 데이터 x가 신경망의 각 레이어를 통과한 결과인 $F(x)$간의 차이를 학습하는 방법으로, 이 출력에 원래 입력 데이터인 x를 단순히 더함(identity mapping)으로써 최종 출력을 $H(x)=F(x)+x$로 만든다. 이를 통해 신경망의 각 레이어를 통과하면서도 중요한 특징이 소실되지 않고 보존되며, 기울기 소실 문제를 크게 완화시키고 훈련을 더 효율적으로 만들어준다. 즉, 잔차는 $F(x)=H(x)-x$로 표현되며, 이 잔차를 학습하는 것을 잔차 학습이라 한다. ResNet은 이 방식을 통해 2015년 ILSVRC(ImageNet Large Scale Visual Recognition Challenge) 대회에서 3.6%의 오류율로 우수한 성능을 보여주며 우승을 차지하였다. 이로써 ResNet은 딥러닝 분야에서 심층 신경망의 잠재력을 입증하였으며, 깊은 신경망을 구축하고 학습하는 방법에서 혁신적인 모델로 주목받고 있다.



(a) (b)

Fig. 5. Difference between plain networks and residual learning. (a) plain networks and (b) residual learning

Fig. 6은 각각 (a) VGG-19, (b) 34층 일반 네트워크 그리고 (c) 34층 레이어를 가진 ResNet-34 구조를 나타낸다. VGG-19는 ResNet이 등장하기 이전, 2014년 ILSVRC에서 심층 신경망으로서 우승한 신경망으로 이미지 분류 작업에서 우수한 성능을 제공하였지만, 많은 가중치로 인해 학습 시간이 오래 걸린다는 단점이 있다. 반면 ResNet-34는 VGG-19보다 신경망도 깊고, 각 레이어 사이에 잔차 연결(Residual connection)을 통해 모델이 효율적으로 학습할 수 있게 한다. 또한, 이 잔차 연결은 특징이 레이어를 통과하는 동안 손실되는 것을 방지하고, 기울기 소실 문제를 완화시켜 준다. 34층의 일반 네트워크는 간단한 구조를 가지고 있지만, 모델의 복잡성과 성능 사이의 트레이드오프(trade-off)로 인해 기울기 소실과 폭주문제가 발생할 수 있다. 따라서 ResNet은 이러한 문제들을 극복하면서도 높은 성능을 제공하는 혁신적인 딥러닝 모델로 주목받고 있다.

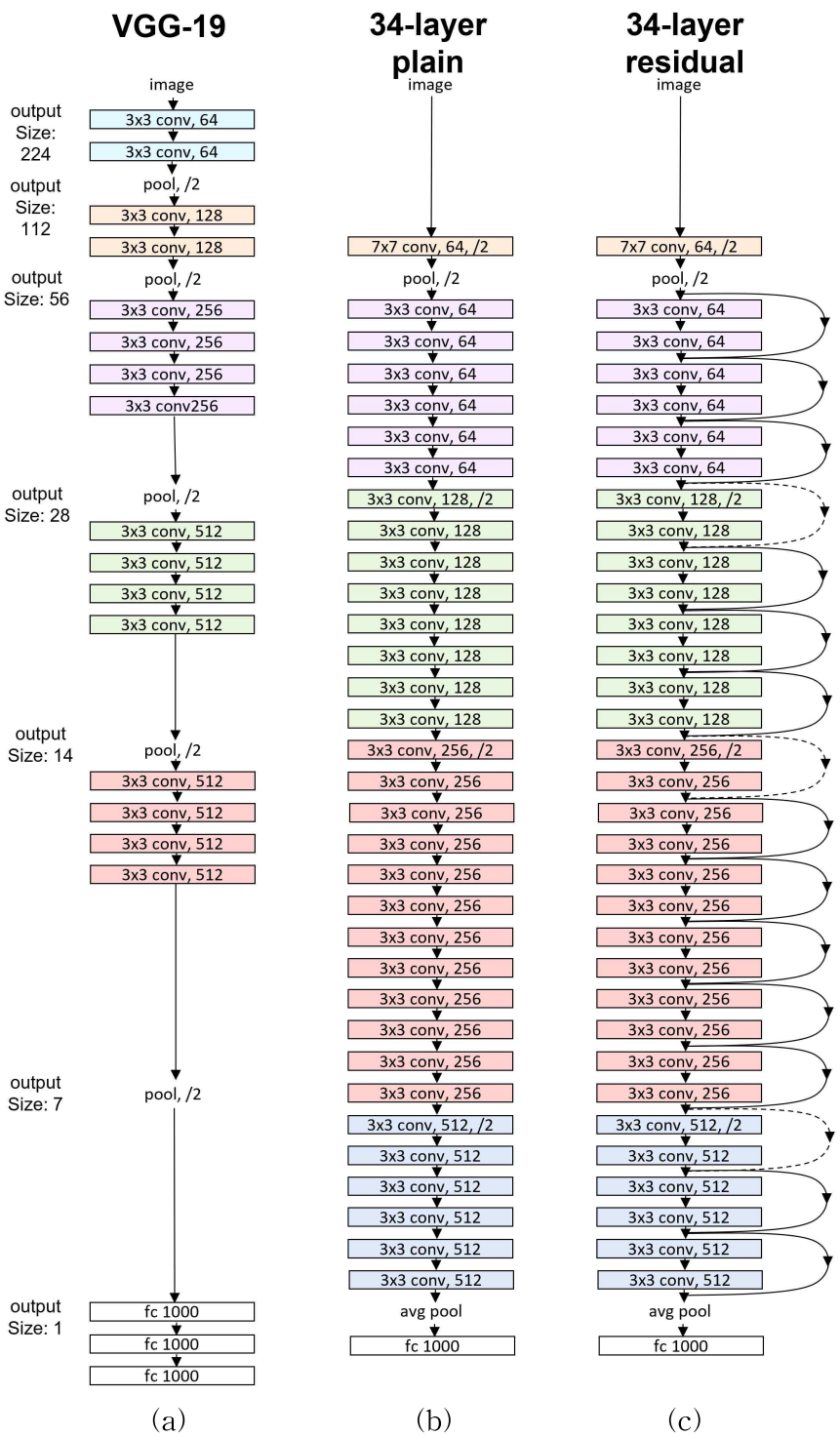
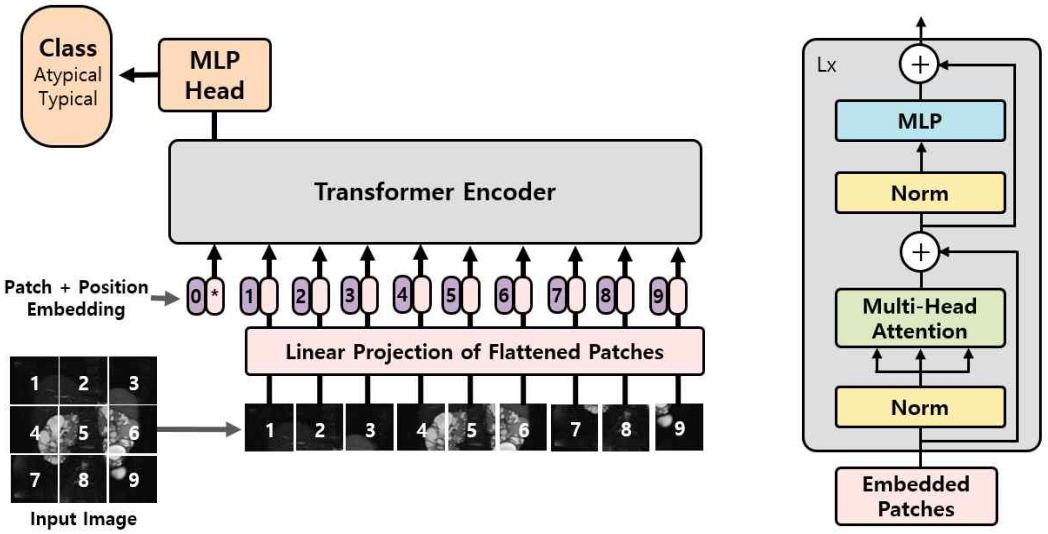


Fig. 6. Differences between VGG-19, 34-layer plain, 34-layer residual architecture: (a) VGG-19, (b) 34-layer plain, and (c) ResNet-34

(2) ViT(Vision Transformer)

트랜스포머(Transformer)[34]는 주로 자연어 처리 및 기계 번역과 같은 NLP(Natural Language Processing) 작업에서 활용되어 왔으며, 연산 속도가 느린 RNN(Recurrent Neural Network)의 한계를 극복하여 초반에는 자연어 처리 분야에서 높은 성과를 보였다. 이후에는 GPT(Generative Pre-trained Transformer), BERT(Bidirectional Encoder Representations from Transformers) 등과 같은 다양한 언어 모델로 확장되어 다양한 응용분야에 사용되고 있다. 더 나아가 최근에는 트랜스포머를 컴퓨터비전(computer vision)에 적용한 ViT (Vision Transformer)[35]가 등장하여, 여러 이미지 인식 벤치마크에서 SOTA (stat-of-the-art)를 달성하였다[36].



(a) (b)

Fig. 7. Structural diagram of vision transformer architecture. (a) Vision Transformer and (b) Transformer Encoder

ViT는 입력 이미지를 텍스트 시퀀스(sequence)와 유사한 방식으로 다룬다. 먼저, 입력된 이미지는 동일한 크기의 작은 패치들(patch)로 나누어진다. 이러한 패치들은 임베딩 레이어를 통해 벡터로 변환된다(patch embedding). 이때,

이미지의 위치정보가 누락되기 때문에 각 패치에 위치 임베딩(pose embedding)이 추가된다. 이러한 방식으로 패치 임베딩과 위치 임베딩이 결합되어 1차원 시퀀스로 평탄화(flatten)된다. 최종적으로 생성된 임베딩 벡터 시퀀스는 트랜스포머의 인코더에 입력으로 전달되며, 이 과정에서 정규화, Multi-Head Attention, 그리고 MLP층을 거쳐 MLP Head에 저장된 정보를 활용하여 이미지를 분류한다[35]. 이러한 시퀀스 구조는 텍스트 데이터를 다루기 위해 설계된 BERT[37]와 유사한 구조를 가지고 있다. 다시 말해, BERT가 트랜스포머를 학습할 때 문서 전체를 활용하는 것처럼, ViT는 트랜스포머를 학습하기 위해 이미지 전체를 활용한다.

초기에는 ViT가 이미지 처리를 위한 귀납적 편향(inductive bias)이 부족하기 때문에 작은 규모의 데이터로 사전 훈련을 할 경우, CNN보다 낮은 성능을 보일 수 있다. 그러나 충분히 큰 규모의 데이터로 사전 훈련된 ViT는 작은 데이터세트에서도 우수한 일반화 성능을 보여준다[35].

4. 전이 학습

전이 학습(Transfer learning)은 이미 훈련된 모델을 새로운 작업에 활용하는 개념으로, 특히 컴퓨터 비전 분야에서 높은 성과를 보이고 있다[38]. 전이 학습을 사용하지 않고 모델을 처음부터 훈련(training from scratch)하려면 상당한 양의 데이터가 필요하며, 이로 인해 계산 비용과 시간이 증가한다. 그러나 전이 학습은 이미 학습된 모델을 활용하여 계산적으로 효율적이기 때문에 학습 속도가 빠르고 적은 데이터로도 더 나은 결과를 얻을 수 있다.

일반적으로 신경망은 깊어질수록 다양한 특징(feature)을 학습하며, 점점 복잡한 특징을 습득한다. 낮은 층에서는 간단하고 기본적인 시각 패턴 같은 낮은 수준 특징(low-level features)을 추출하며, 이는 이미지의 색, 경계(edge), 질감과 같은 특징을 포함한다. 반면 높은 층에서는 낮은 층에서 습득한 특징을 조합하여 더 복잡하고 의미 있는 패턴이나 정보와 같은 높은 수준 특징(high-level features)을 추출하여 물체 인식과 같은 작업을 수행한다[39]. 이러한 특징을 신경망에 학습시키려면 많은 양의 데이터 세트를 사용하여 모델을 훈련시켜야 한다.

특히, 의료영상 데이터 분석 분야에서는 일반적인 데이터에 비해 데이터를 확보하는데 한계가 있기 때문에, 전이 학습은 이와 같은 데이터 부족의 문제를 해결할 수 있는 중요한 방법 중 하나이다. 본 연구에서는 대규모 데이터 세트인 ImageNet-1K 데이터 세트로 학습한 가중치를 사용하여 미세 조정(fine-tuning)을 수행하였다[40]. 미세 조정은 전이 학습의 한 부분으로, 사전 훈련된 모델을 기반으로 하여 분류하고자 하는 데이터에 맞게 모델을 학습시키는 과정이다. 일반적으로 전이 학습된 모델의 초기 레이어는 고정시키고, 이후 나머지 레이어를 새로운 작업에 맞게 다시 훈련시킨다. 본 연구에서는 사전 훈련된 가중치를 사용하여 초기 두 개의 레이어를 고정시킨 후, 나머지 레이어를 새로운 입력 데이터에 따라 가중치를 업데이트하며 모델을 훈련시켰다.

5. 평가 지표

클래스 간의 데이터 양이 불균형한 경우, 특히 의료 분야와 같이 희귀한 질병을 진단하는 분류 모델을 고려할 때, 정확도만(accuracy, M_{acc})을 사용하여 평가하는 것은 한계가 있다. 예를 들어, 모델의 데이터 세트에서 환자의 95%가 해당 질병에 걸리지 않았고, 5%만 실제 질병에 걸린 상태라고 가정할 때, 모든 환자를 정상 상태로 예측하는 단순한 모델을 사용해도 정확도는 95%로 높게 나타난다. 이러한 모델은 실제로 질병을 찾는데 아무런 도움을 주지 못할 것이다. 그러므로 데이터 양이 불균형한 경우 모델의 성능을 더 정확하게 평가하기 위해 오차 행렬(Table. 2)을 활용하고, 정밀도(precision, M_{pre}), 재현율(recall, M_{rec}), F1-점수(F1-score, M_{F1})와 같은 평가지표를 사용한다. 오차행렬을 활용한 정확도는 식 (1)과 같이 나타낼 수 있다.

Table. 2. Confusion Matrix

		예측 클래스 (Predicted class)	
		Positive	Negative
실제 클래스 (True class)	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

$$M_{acc} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

정밀도는 해당 클래스를 얼마나 정확하게 식별하는지를 측정하는 지표로, 높은 정밀도는 모델이 해당 클래스를 높은 정확도로 식별하는 능력이 높다는 것을 의미한다. 특히 의료분야와 같이 잘못된 양성 진단이 중요한 경우, 정밀도는 중요한 평가 지표로 간주된다. 정밀도는 식(2)와 같이 나타낼 수 있다.

$$M_{pre} = \frac{TP}{TP+FP} \qquad (2)$$

재현율(또는 민감도, sensitivity)은 모델이 실제 양성을 얼마나 효과적으로 찾아내는지 측정하는 지표로, 재현율이 높을수록 모델이 실제 양성을 놓치는 경우를 줄이고 더 많은 양성을 식별할 수 있음을 의미한다. 그러므로 재현율은 의료분야와 같이 실제 양성 환자를 놓치는 것이 치명적인 경우 중요한 평가 지표로 간주된다. 재현율은 식(3)과 같이 나타낼 수 있다.

$$M_{rec} = \frac{TP}{TP+FN} \qquad (3)$$

F1-점수는 정밀도와 재현율의 조화 평균으로 계산되며, 모델의 성능을 종합적으로 평가하는 데 사용된다. 이 지표는 모델이 정확한 양성 예측과 실제 양성을 놓치지 않는 두 가지 중요한 측면을 모두 고려한다. 정밀도와 재현율은 서로 상충 관계로, 둘 중 하나를 높이면 다른 하나는 낮아진다. 이러한 관계에서 F1-점수는 정밀도와 재현율 사이의 균형을 나타내며, 모델의 성능을 균형 있게 평가하는 데 도움을 준다. F1-점수는 식(4)와 같이 나타낼 수 있다.

$$M_{F1} = \frac{2}{\frac{1}{M_{pre}}+\frac{1}{M_{rec}}} = \frac{2\times M_{pre}\times M_{rec}}{M_{pre}+M_{rec}} = \frac{TP}{TP+\frac{FN+FP}{2}} \qquad (4)$$

오차 행렬과 같이, 분류 모델의 성능을 더 자세하게 평가하고 시각화하기 위해 ROC(Receiver Operating Characteristic) 곡선과 AUC(Area Under the Curve)를 평가 지표로 활용하였다. ROC 곡선은 다양한 분류 임계값에서 모델의 성능을 나타내며, 민감도와 특이도(specificity) 간의 상충 관계를 도식화한다. 이 곡선의 x축은 실제 음성 중 모델이 양성으로 잘못 예측한 비율(False Positive Rate, FPR)로, '1-특이도'를 의미하며 식 (5)와 같이 나타낼 수 있다. y축은 실제 양성을 정확하게 양성으로 예측한 비율(True Positive Rate, TPR)로 위의 식

(3)에서 확인할 수 있듯이 민감도를 의미한다[41].

$$FPR = 1 - \frac{TN}{FP + TN} = \frac{FP}{FP + TN} \qquad (5)$$

ROC 곡선의 아래 면적을 AUC라 하며, AUC 값은 0과 1 사이의 범위에서 나타난다. AUC가 0.5 이상인 모델의 경우, 모델은 유의미한 분류 능력이 있다고 판단되며, 1에 가까울수록 이상적인 모델에 해당된다. 의료진단에 있어서 AUC가 0.9를 넘는다면 분류 정확도가 우수한 것으로 판단할 수 있으며[41], 0.5라면 모델은 분류 능력이 없음을 의미하고, 0.5 미만인 경우에는 모델의 라벨링이나 알고리즘이 잘못되었을 가능성이 있다.

본 연구에서는 Python 3.9.17에서 PyTorch 1.12.1과 scikit-learn 1.3.0 라이브러리를 사용하여 오차행렬과 ROC 곡선을 생성하였으며, 이를 시각화하기 위해 Matplotlib 3.7.3, Pandas 2.0.3 및 Seaborn 0.12.2 라이브러리를 활용하였다.

6. 자동 분류 결과 확률 도출

인공신경망에서 활성화 함수(activation function)는 각 노드에 들어오는 입력 신호의 총합을 출력 신호로 변환하여 활성화하는 역할을 한다. 특히 딥러닝 모델에서 다중 클래스 분류를 수행할 때, 출력 층의 활성화 함수는 주로 소프트맥스 함수(softmax function)가 사용된다. 소프트맥스 함수는 출력을 0에서 1사이의 실수 값으로 변환하여 준다. 이때, 소프트맥스 함수를 통과한 각 클래스의 출력은 해당 클래스에 속할 확률을 나타내며, 이러한 확률 값들의 총합은 항상 1이 된다. 따라서 소프트맥스 함수의 출력은 확률로 해석할 수 있다. 소프트맥스 함수는 식(6)과 같이 나타낼 수 있다.

$$y_k = \frac{\exp(a_k)}{\sum_{i=1}^{n} \exp(a_i)} \qquad (6)$$

여기서, n은 출력 층의 뉴런 수로 클래스의 총개수를 나타내며, y_k는 k번째 출력 값, a_k는 출력 층의 입력 신호를 나타낸다. 소프트맥스 함수는 지수함수가 사용되기 때문에 입력 값이 조금이라도 크면 지수함수의 값이 급격하게 증가한다. 이러한 특성은 모델의 출력 중 하나의 클래스를 강조하면서 다른 클래스의 확률을 낮추어 가장 확실한 예측을 출력하도록 도와준다. 하지만 이는 컴퓨터 계산에서 오버플로(overflow) 문제를 일으킬 수 있다. 오버플로 문제란 컴퓨터의 숫자 표현 한계를 초과하여 발생하는 문제로, 컴퓨터는 이 값을 적절하게 표현하지 못하고 무한대로 처리하게 되며, 이로 인해 결과가 부정확해질 수 있다. 이러한 문제를 해결하기 위해 소프트맥스의 함수에 C 라는 임의의 정수를 분자와 분모에 곱하여 식 (7)과 같이 변형할 수 있다.

$$y_k = \frac{\exp(a_k)}{\sum_{i=1}^{n}\exp(a_i)} = \frac{C\exp(a_k)}{C\sum_{i=1}^{n}\exp(a_i)} = \frac{\exp(a_k+\log C)}{\sum_{i=1}^{n}\exp(a_i+\log C)} = \frac{\exp(a_k+\log C')}{\sum_{i=1}^{n}\exp(a_i+C')} \quad (7)$$

이러한 변형은 지수함수에 어떤 정수를 더하거나 빼도 결과 값이 동일하다는 증가함수의 특성을 활용한 것이다. 오버플로를 문제를 해결하기 위해서는 입력 신호 중 최댓값을 빼는 방식이 일반적으로 사용된다. 또한 이 증가함수의 특성으로 인해 소프트맥스 함수를 적용시켜도 각 클래스의 대소 관계는 변하지 않는다. 따라서 출력 층에서 소프트맥스 함수를 생략하고 가장 큰 출력을 가지는 클래스를 선택하는 방식으로 데이터를 분류할 수 있다.

본 연구에서는 딥러닝 모델을 기반으로 하여 ADPKD MR 영상을 class 1과 class 2로 자동 분류한 후, 이러한 진단 결과를 확률로 표현하여 의료진과 환자들에게 진단 결과를 제공하고자 하였다. 이를 위해 PyTorch 1.12.1을 사용하여 딥러닝 모델의 출력 결과에 소프트맥스 함수를 적용하여 각 클래스 별 확률을 계산하고 진단 결과를 제공하였다.

7. 설명 가능 인공지능 구현 방법

인공지능이 발전함에 따라 인공지능이 일상생활에 빠르고 광범위하게 도입되고 있으며, 인공지능 시스템이 제공해 주는 서비스가 증가하고 있다[42, 43]. 그러나 이러한 인공지능 시스템은 '블랙박스 문제'를 가지고 있으며, 이를 해결하고자 XAI가 등장하였다. 블랙박스로 작동되는 딥러닝 모델은 중간과정이 어떻게 이루어지는지 명확하게 이해하기 어렵고, 입력에 대한 결과만을 확인할 수 있다. 이로 인해 신뢰성, 윤리적, 그리고 규제 및 법적 책임과 관련된 문제들이 발생한다[44].

신뢰성 문제는 모델이 결정을 내릴 때 어떠한 기준으로 결정을 내리는지 알 수 없기 때문에 발생한다. 윤리적 문제는 인공지능 시스템이 훈련데이터에 내포된 편향을 학습하거나 공정하지 않은 결정을 내릴 가능성이 있다는 점에서 발생한다. 규제 및 법적 책임 문제로는 인공지능 시스템이 내린 결정에 대한 책임은 누구에게 있는지에 대한 문제가 있다. 인간과 달리 인공지능은 도덕적 판단 능력을 가지고 있지 않으므로 판단 결과에 대한 책임과 규제가 모호해질 수 있다. 특히 의료 진단과 같은 중요한 결정에 있어서, 인공지능의 의사결정에 대한 근거를 명확히 알 수 없다면 환자에게 심각한 영향을 미칠 수 있다.

이러한 문제들을 극복하기 위해서 XAI의 기술이 점차 중요해지고 있다. XAI는 쉽게 말해 '누가', '무엇을', '언제', '어디서'와 같은 질문에 더 많은 신뢰도를 부여할 수 있는 모델을 의미하며, 의학 및 보건의료 분야에서는 더 나은 진단적 및 치료적 의사결정을 내리는데 도움을 줄 수 있다[45]. 설명 가능한 인공지능은 크게 '설명 가능한 모델 구축'과 사용자를 위한 '이유 설명 인터페이스'로 나뉘며, 이 중 설명 가능한 모델 구축은 심층 설명학습(deep explanation), 해석 가능한 모델(interpretable models), 모델 귀납(model induction) 등을 통해 개발된다[46, 47].

본 연구에서 사용하는 LIME(Local Interpretable Model-Agnostic Explanation)은 블랙박스 모형을 설명 가능한 모델로 추론해 주는 방법으로 모델 귀납에 해당된다. LIME[48]은 국소적 대리 분석(local surrogate)의 한 형태로, 데이터 하나에 대해 블랙박스 모델이 어떤 부분에 집중하여 특정 예측을 내렸는지를

설명하는 알고리즘이다. LIME의 전제는 데이터 공간의 전체에서 모델 설명은 어려울 수 있지만 국소적인 데이터 공간에서는 의미 있는 모델로 설명할 수 있다는 것이다. 이러한 설명은 모델에서 사용되는 실제 특성과는 관계없이 사람이 이해할 수 있는 표현으로 나타내어져야 한다. 본 연구에서는 LIME 라이브러리의 버전 0.2.0.1을 사용하여 설명을 생성하고 분석에 활용하였다.

이미지 분류에서 해석가능한 표현은 유사한 픽셀을 그룹화하고 슈퍼픽셀(superpixel)의 존재 또는 부재를 나타내는 이진 벡터로 구성된다. 슈퍼픽셀은 이미지에서 유사한 특징을 가진 픽셀을 그룹화한 것으로, 이미지 분할 및 물체 검출과 같은 컴퓨터 비전에서 유용하다. 본 연구에서는 입력 이미지를 슈퍼픽셀로 분할하기 위해 SLIC(Simple linear iterative clustering)을 사용하였다[49]. 이때 scikit-image 0.21.0 라이브러리를 사용하였으며, SLIC의 매개변수인 슈퍼픽셀 수(n_segment)와 compactness를 각각 200과 10으로 설정하여 사용 하였다. 제로 패딩된 데이터세트를 입력으로 사용하였으므로, 테스트 데이터세트도 동일하게 제로 패딩된 데이터세트를 사용하였고, 최종 결과에서는 제로 패딩된 부분을 제거하여 나타내주었다.

8. 실험 환경

본 연구에서 사용된 모델은 Python 3.9.17, PyTorch 1.12.1 및 Ubuntu 16.04.7 LTS로 구축되었으며, 구축된 실험 환경 하드웨어는 Table. 3과 같다. 이 외 XAI 알고리즘 구현을 위해 lime 0.2.0.1을 사용하였다.

Table. 3. Experimental Environment

규격	Version
OS	Linux Ubuntu 16.04.7 LTS
CPU	Intel Xeon Silver 4110
RAM	128GB
GPU	1 Nvidia Titan Xp

Ⅳ. 연구 결과

1. 상염색체 우성 다낭성 신장 질환 자동 분류 결과

테스트 데이터로는 class 1에 해당하는 데이터 131개와 class 2에 해당하는 데이터 20개의 MR 영상을 사용하였다. 본 연구에서는 ResNet과 ViT 모델을 사용하였고, ResNet의 경우 레이어의 개수에 따라 18, 34, 50 모델을 사용하여 총 4가지 모델을 사용하였다. 각 ResNet 모델의 분류 정확도는 86.75%, 94.04%, 98.01%이며, ViT는 94.7%의 정확도를 달성하였다. ResNet-50이 가장 높은 정확도를 보여주고 있지만, Table. 4에서 확인할 수 있듯이 전반적으로 유사한 정확도를 보이고 있으며, ResNet-34, ResNet-50, ViT는 class 2를 완벽하게 분류하여 100%의 정확도를 보이고 있다.

Table. 4. Classification accuracy of trained networks with models ResNet-18, ResNet-34, ResNet-50, and ViT

Model / Accuracy(%)	ResNet-18	ResNet-34	ResNet-50	ViT
Class 2 (atypical)	95.0	100	**100**	100
Class 1 (typical)	85.5	93.1	**97.7**	93.9
평균	86.75	94.04	**98.01**	94.7

Scikit-learn 패키지를 활용하여 오차 행렬을 생성하였으며(Fig. 8), 오차 행렬에서 세로축은 실제 클래스, 가로축은 모델이 예측한 클래스를 나타낸다. Class 2 자동 분류에서는 총 20개의 데이터 중 ResNet-18이 1개의 데이터를 잘못 분류하였으며, 나머지 모델의 경우 잘못 분류된 경우는 나타나지 않았다. Class 1의 테스트 데이터 세트 총 131개 중에서 ResNet-18, ResNet-34, ResNet50, ViT 모델 별 잘못 분류한 데이터의 개수는 각각 19, 9, 3, 8개씩 나타났다.

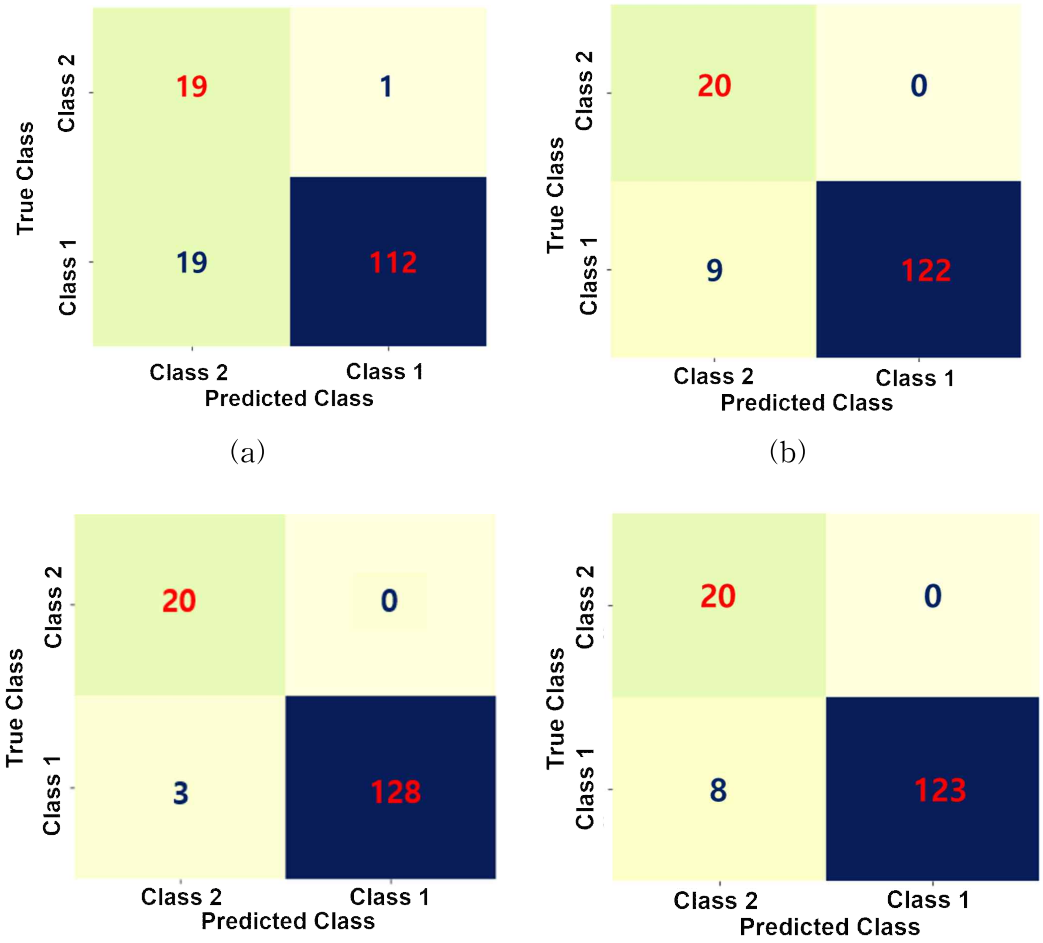


Fig. 8. Confusion matrix of automated classification of ADPKD with models: (a) ResNet-18, (b) ResNet-34, (c) ResNet-50, and (d) ViT

오차 행렬을 활용하여 정확도뿐만 아니라 정밀도, 재현율 그리고 F1-점수를 계산하여 모델의 성능을 비교하였다(Table. 5). 이때, 평가지표에서 ResNet-50 의 성능이 가장 우수하고, ViT, ResNet-34, ResNet-18 순으로 성능이 좋게 나온 것을 확인할 수 있다.

모델의 성능을 비교하고 객관적인 평가를 하기 위해 Scikit-learn 패키지를 사용하여 ROC곡선 그래프와 AUC를 나타내었다. 본 연구에서 모델 별 ROC곡선 그래프의 AUC는 각각 0.90, 0.97, 0.99, 0.97로 본 연구에서 사용한 모든 모델의 AUC는 0.9 이상이었고, ResNet-50의 AUC가 가장 높게 나타났다(Fig. 9).

Table. 5. Classification precision, recall and F1-scores of trained networks with models ResNet-18, ResNet-34, ResNet-50, and ViT

Model	평가지표 / Type	정밀도 (Precision)	재현율 (Recall)	F1-점수 (F1-Score)
ResNet-18	class 2	0.50	0.95	0.66
	class 1	0.99	0.85	0.92
	macro average	0.75	0.90	0.79
	weighted average	0.93	0.87	0.88
ResNet-34	class 2	0.69	**1**	0.82
	class 1	**1**	0.93	0.96
	macro average	0.84	0.97	0.89
	weighted average	0.96	0.94	0.94
ResNet-50	class 2	**0.87**	**1**	**0.93**
	class 1	**1**	**0.98**	**0.99**
	macro average	**0.93**	**0.99**	**0.96**
	weighted average	**0.98**	**0.98**	**0.98**
ViT	class 2	0.71	**1**	0.83
	class 1	**1**	0.94	0.97
	macro average	0.86	0.97	0.90
	weighted average	0.96	0.95	0.95

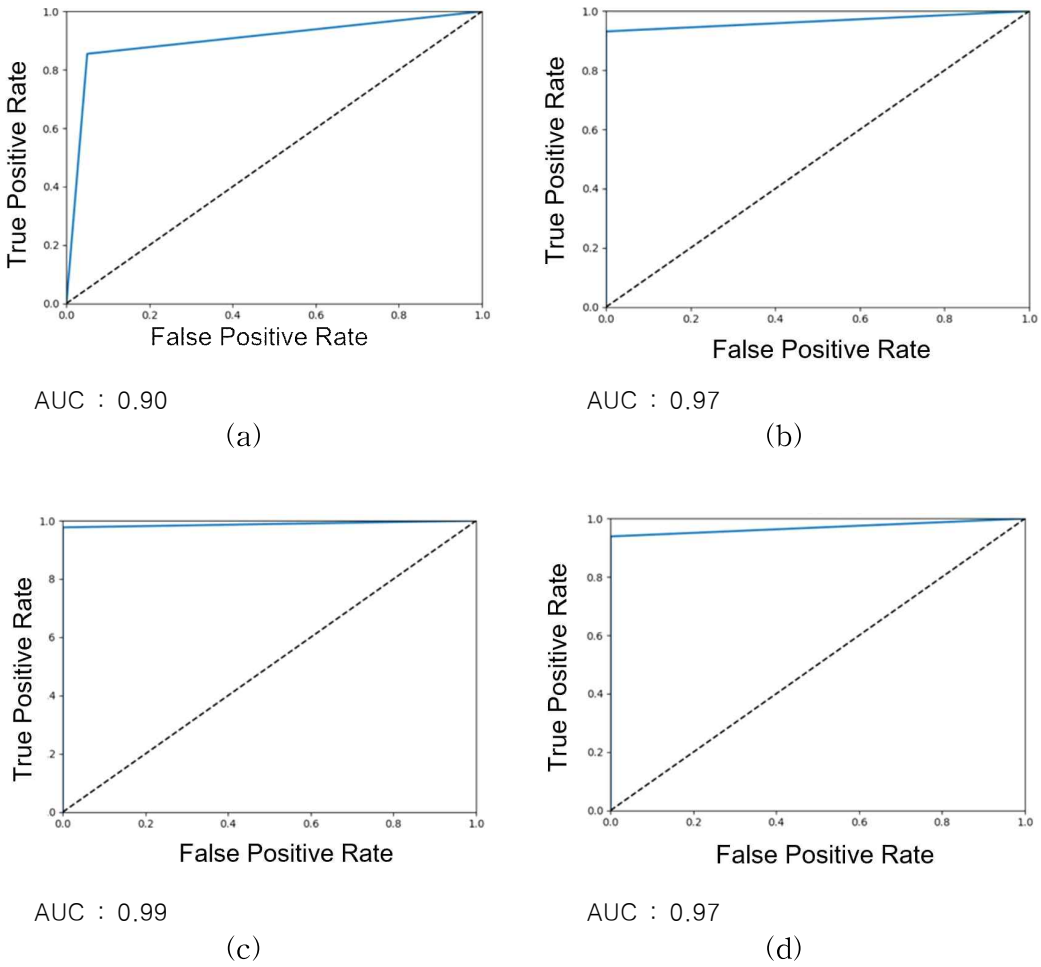


Fig. 9. ROC curves and AUCs of models: (a) ResNet-18, (b) ResNet-34, (c) ResNet-50, and (d) ViT

추가적으로 ResNet의 레이어별 학습 시간과 정확도를 확인하기 위해 ResNet-18, ResNet-34, ResNet-50 외에도 ResNet-101과 ResNet-152 모델을 학습하였다. Fig. 10은 ResNet 모델 레이어별 학습 시간 대비 정확도를 시각적으로 보여주는 그래프이다. 이 그래프에서는 레이어가 깊어질수록 학습 시간이 증가하며 최대 3시간 23분 16초에 달하는 시간이 소요되었다. 정확도는 ResNet-18에서부터 ResNet-34, ResNet-50까지는 86.75%에서부터 지속적으로 증가하여 98.01%의 정확도까지 도달하였다. 그러나 ResNet-101과 ResNet-152에서는 정확도가 96.69%로 감소하였으며, ResNet-50을 기준으로 보았을 때 오히려 학습 시간은 각각 대략 23분, 46분이 더 소요되었다.

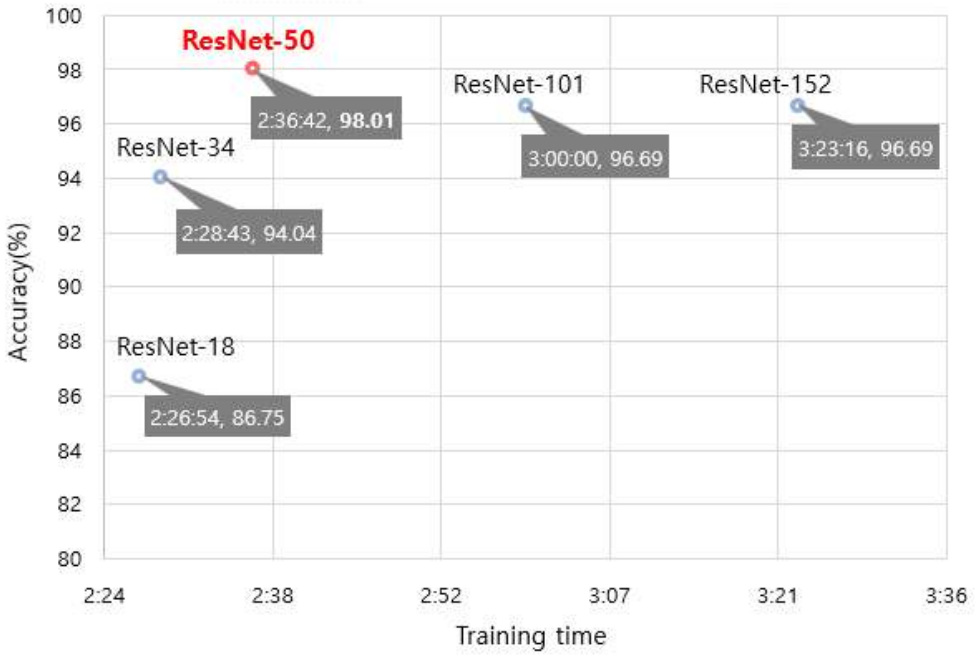
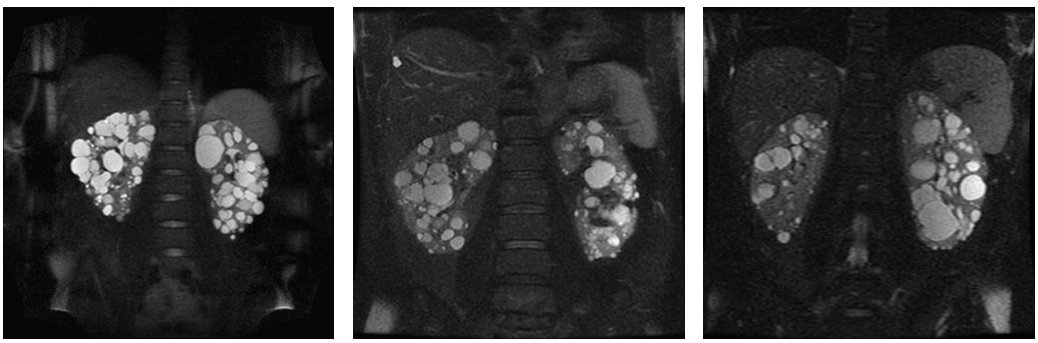


Fig. 10. Visual plots of training time and accuracy by differed ResNet layers

2. 확률 도출 결과

결과가 가장 좋게 나타난 ResNet-50 모델의 분류결과를 기준으로, 소프트맥스 함수를 적용하여 각 클래스로 분류된 확률을 확인하였다. 총 131 개의 class 1 테스트 데이터 중 126 개의 데이터가 class 1 로 정확하게 분류되었으며, 대부분이 99%에서 100%의 높은 확률로 분류되었다. Fig. 11 은 높은 확률로 class 1 로 알맞게 분류된 대표적인 class 1 의 MR 영상이다.



(a) (b) (c)

Fig. 11. Three class 1 MR images examples correctly classified with higher probability. The probability values of most correctly classified class 1 data were above 99%.

이 중 상대적으로 낮은 확률로 분류된 케이스는 Fig. 12 에서 볼 수 있으며, 확률은 각각 (a) 93.7%, (b) 90.9%, (c) 65.6%, (d) 56.2%였다. Fig. 12 (a)와 (b)의 경우, 미세한 여러 개의 우성 낭종이 신장 내부에 흩어져있어 정확한 분석에 한계가 있는 것으로 보인다. Fig. 12 (c)에서는 신장에 상대적으로 큰 우성 낭종이 다수 존재하고, (d)에서는 왼쪽 신장 상부에 상대적으로 두드러지는 낭종이 신장 외 낭종처럼 보이기 때문에 분류 확률이 상대적으로 낮게 나온 것으로 보인다.

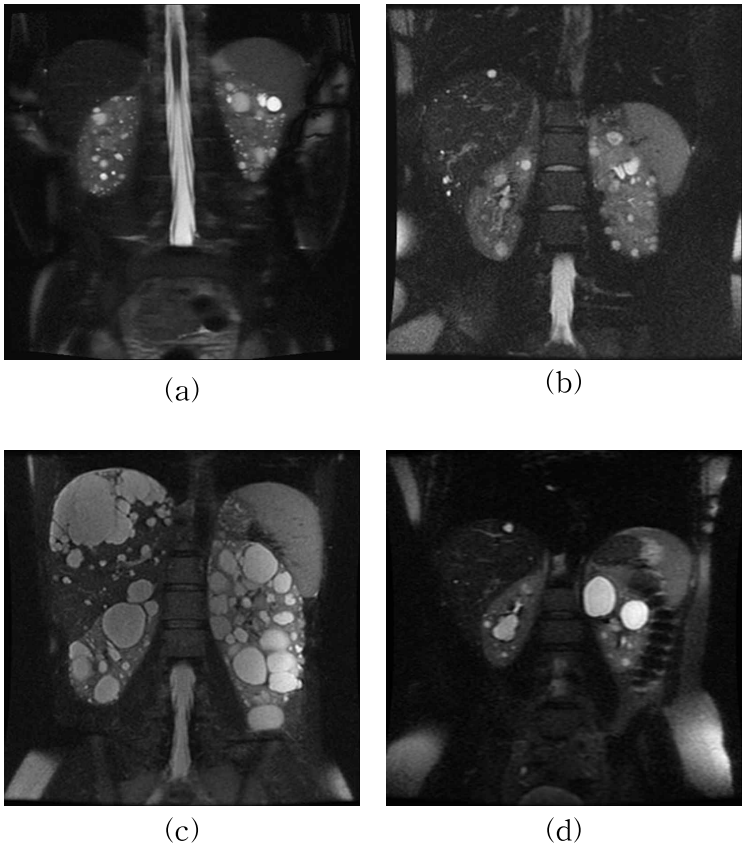
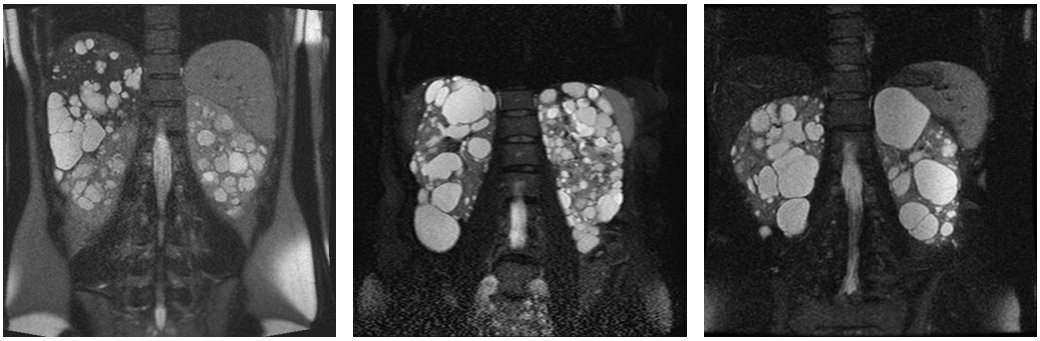


Fig. 12. Four class 1 MR images that were correctly classified with relatively lower probability values. The class 1 classification probabilities were (a) 93.7%, (b) 90.9%, (c) 65.6% and (d) 56.2%, respectively, which was lower than the 99% probability of most correctly classified class 1.

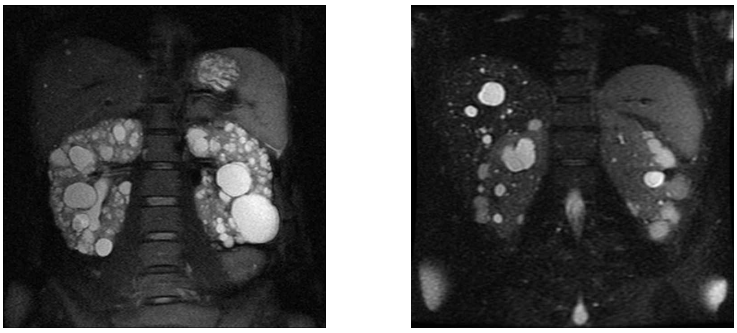
Class 1이 class 2로 잘못 분류된 경우는 Fig. 13으로, class 1로 분류될 확률은 각각 (a) 42.2%, (b) 39%, (c) 26%로 나타났다. 이러한 MR 영상들은 공통적으로 신장에 미세한 여러 우성 낭종들이 전체적으로 분포되어 있고, (a)는 신장의 오른쪽 상부, (b)는 신장의 오른쪽 하부, 그리고 (c)는 신장의 왼쪽 상부에 상대적으로 큰 우성 낭종으로 인해 class 2로 잘못 분류된 것으로 보인다.



(a) (b) (c)

Fig. 13. Three class 1 MR images mis-classified to be class 2. The class 1 classification probabilities were (a) 42.2%, (b) 39%, and (c) 26%

Class 2는 3개의 독립된 훈련 및 테스트 데이터세트로 나누어 진행하여 총 60개를 사용하였으며, 모든 데이터가 class 2로 정확하게 분류되었다. 대부분이 99% 이상의 높은 확률로 분류되었지만, 그중에서 Fig. 14의 (a)는 93.7%, (b)는 98.2%로 비교적 낮은 확률로 분류되었다. (a)의 경우 신장 왼쪽 하부에 두드러지는 우성 낭종이 존재하지만 이와 비슷한 크기의 여러 우성 낭종들도 분포하고 있으며, (b)의 경우 비교적 두드러지는 신장 외 낭종이 보이지 않아 분류 확률이 상대적으로 낮게 나타난 것으로 보인다.

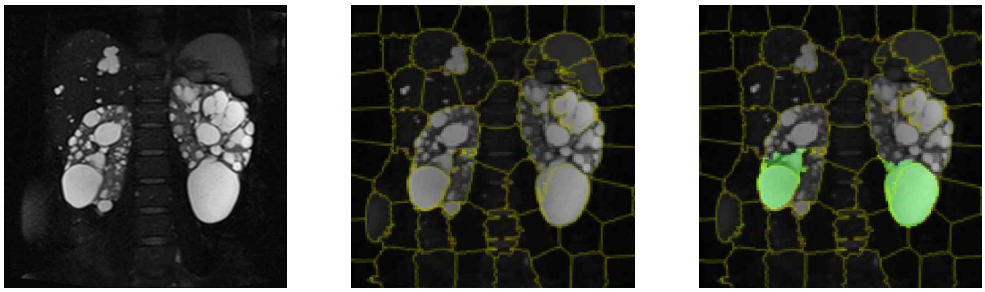


(a) (b)

Fig. 14. Two class 2 MR images that were correctly classified with relatively lower class 2 classification probabilities. The class 2 classification probabilities were (a) 93.7% and (b) 98.2%.

3. 설명 가능 인공지능 구현 결과

자동 분류 성능이 가장 좋게 나온 ResNet-50 모델의 분류 결과를 기반으로 설명 가능 인공지능을 적용하였다. Fig. 15는 설명 가능 인공지능을 적용한 과정을 나타내며, (a)는 신장 외 낭종을 포함하고 있는 class 2 MR 영상으로 원본 입력 이미지에 해당된다. (b)는 입력 이미지를 슈퍼픽셀로 분할하는 SLIC 알고리즘을 적용한 결과이며, 여기서 노란색 선은 슈퍼픽셀의 경계를 나타낸다. 그 후, LIME을 적용하면 자동 분류 결과에 가장 크게 영향을 미친 슈퍼픽셀을 강조한 결과가 (c)와 같이 나타나게 되고, 이때, 강조된 슈퍼픽셀은 녹색으로 표시된다.



(a) (b) (c)

Fig. 15. Visual representation of explainable artificial intelligence procedures. (a) Input the original MR images, (b) apply a simple linear iterative clustering(SLIC) algorithm to segment it into superpixels, and (c) apply a locally interpretable model-independent explanatory (LIME) algorithm to highlight the superpixels that contribute significantly to automatic classification.

Fig. 16은 모델의 자동 분류의 설명력을 향상시키기 위해 모델의 예측 결정에 영향력이 높은 영역을 추출하는 설명 가능 인공지능을 적용하여 추출된 class 2 MR 영상을 보여준다. 상단부분은 class 2의 원본 MR 영상이고, 하단은 해당 영상에 설명 가능 인공지능을 적용한 결과이다. 각 하단에서 녹색으로 강조된 슈퍼픽셀은 뚜렷한 신장 외 낭종을 포함한 슈퍼픽셀 영역을 강조하며, 이는

class 2로 분류할 때 가장 중요한 형태적 특징을 나타내는 영역에 해당된다. (a), (b), (c)에서는 뚜렷한 신장 외 낭종이 정확하게 녹색으로 표시되었지만, (d), (e), (f)에서는 뚜렷한 신장 외 낭종을 포함한 주변 영역까지 약간 과장되어 강조된 것을 확인할 수 있다.

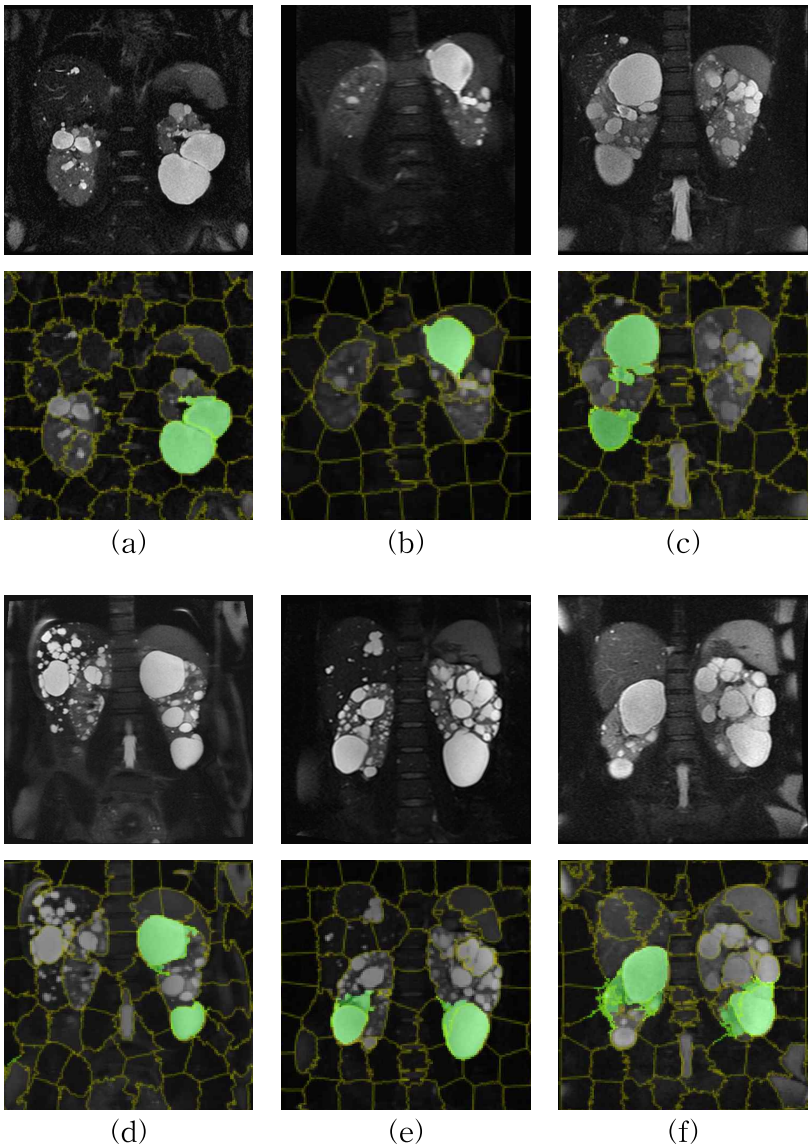


Fig. 16. Result of applying explainable artificial intelligence to class 2 MR images. The top row is the original MR images, the bottom row show the boundaries of the superpixels with yellow lines, and the highly contributing superpixels are highlighted in green.

Ⅴ. 고찰

이미지 처리 분야에서는 주로 합성곱 신경망이 사용되어 왔지만, 최근에는 자연어 처리 분야에서 주로 사용되던 트랜스포머 구조가 이미지 처리 분야로 확장되고 있다. 이에 따라 합성곱 신경망과 트랜스포머 구조의 성능을 비교하기 위해 ResNet과 ViT를 사용하였고, ResNet은 레이어가 18, 34, 50층인 세 가지 심층 합성곱 신경망 구조를 사용하여 레이어의 증가에 따른 모델 별 성능을 비교하였다. 테스트 데이터세트로는 131개의 class 1과 20개의 class 2 MR 영상을 사용하였다. Class 1에 대해 ResNet-18은 19개, ResNet-34는 9개, ResNet-50은 3개, 그리고 ViT는 8개를 class 2로 잘못 분류하였으며, class 2에 대해서는 ResNet-18 모델만 1개를 잘못 분류한 결과를 보였다. 결론적으로 본 연구에서는 ResNet-50의 성능이 가장 좋게 나타났으며, 이때 class 1 자동분류의 정확도, 정밀도, 재현율, F1-점수는 각각 98.1%, 1, 0.98, 0.99, class 2의 자동분류에서는 각각 100%, 0.87, 1, 0.93, 그리고 AUC는 0.99로 여러 모델들 중 가장 뛰어난 성능을 보였다. 전반적으로 네 가지 모델 모두 우수한 성능을 나타냈지만, ResNet-18, 34, 50을 비교한 결과, 모델의 레이어가 증가함에 따라 모델의 성능이 향상되는 것을 확인할 수 있었으며, ViT는 ResNet-50에 비해 성능이 낮게 나타났다. 이러한 결과는 ResNet-50이 이미 오랜 기간 동안 다양한 이미지 분류 작업에 사용되어 왔으며, 대규모 데이터세트로 훈련된 반면 ViT는 상대적으로 최근에 등장한 모델이기 때문에 성능 상 차이가 나타난 것으로 볼 수 있다. 그러나 ResNet-18 및 ResNet-34 대비 ViT가 더 뛰어난 성능을 보여주었다.

또한, ResNet의 레이어 별 성능을 비교하기 위해 ResNet-101과 ResNet-152 모델을 추가로 사용하여 다섯 가지 다른 레이어 구조의 ResNet 모델을 학습하고 시간대비 성능을 비교하였다. 레이어의 깊이가 증가함에 따라 학습 시간은 증가하였으며, ResNet-18부터 ResNet-34, ResNet-50 모델까지는 정확도가 94.04%에서부터 98.01%까지 지속적으로 향상되었다. 하지만 ResNet-101과

ResNet-152 모델은 레이어가 더 깊었음에도 불구하고 정 모델 모두 정확도가 96.69%로 감소하였으며, 학습 시간은 ResNet-50에 비해 각각 23분, 46분가량 더 소요되었다. 이러한 결과는 네트워크가 깊어짐에 따라 정확도가 일정 수준 이상으로 계속해서 상승하지 않고 오히려 감소한다는 경향을 보여준다. 이는 깊은 네트워크가 모델의 복잡성을 증가시켜 오히려 과적합을 유발하고 일반화 성능이 감소하는 것을 나타낸다.

자동 분류 결과, 가장 우수한 성능을 보인 ResNet-50의 분류 결과를 기반으로 소프트맥스 함수를 적용하여 확률을 추론하였다. 모델이 class 1을 class 2로 잘못 분류한 MR 영상은 단 3개에 불과하였다. 이 경우, 미세한 여러 우성 낭종이 신장의 경계에 인접해 있고, 상대적으로 두드러지는 큰 우성 낭종이 여러 개 존재하여 모델이 명확하게 판단하기 어려울 수 있다. 자동 분류 확률에서 분류가 불명확한 영역(gray zone)에 해당하는 MR 영상은 class 1 중 56.2%의 확률로 class 1로 알맞게 분류된 Fig.12 (d)와 class 1의 해당 확률이 각각 42.2%, 39%로 나타나 class 2로 잘못 분류된 Fig. 13의 (a)와 (b)가 해당된다. 이러한 경우, 육안으로 보기에도 상대적으로 두드러지는 우성 낭종으로 인해 의료 전문가도 class 1과 class 2로 분류하기 어려울 수 있다. 따라서 본 연구에서는 이러한 어려운 케이스를 확률로 표현해 줌으로써, 실제 의료 현장에서 의료 전문가들이 이러한 케이스를 재검토하고 다시 판단하는 데 도움이 될 수 있는 근거를 제시한다.

본 연구에서는 분류 결과에 대해 신뢰성을 높이기 위해 확률과 더불어 설명 가능 인공지능을 적용하여 앞서 모델이 자동 분류한 결과에 대해 MR 영상에서 가장 큰 영향을 준 슈퍼픽셀 영역을 시각적으로 강조하여 나타내고자 하였다. 정확하게 분류된 class 2에서 SLIC 및 LIME을 적용시켰을 때, class 2로 분류하는 가장 큰 특징인 눈에 띄는 신장 외 낭종이 강조되어 나타났으며, 이는 자동 분류된 결과에 대한 근거로 충분히 사용될 수 있다. 하지만 강조된 영역이 눈에 띄는 신장 외 낭종을 포함한 주변 영역까지 약간 과장하여 강조되는 부분이 있었다. 그러나 본 연구에서 설명 가능한 인공지능을 적용하는 목적은 신장과 낭종의 정확한 경계 분할이 아닌, 모델이 내린 결정에 대한 근거를 시각적으로 보여주는 것이기 때문에 이는 충분히 감안할 수 있는 부분이다.

Ⅵ. 결론

본 연구에서는 상염색체 우성 다낭성 신장 질환 환자들의 신장이 포함된 복부 MR 영상을 활용하여 ADPKD를 class 1(Typical case)과 class 2(Atypical case)로 자동 분류하는 인공지능 기반의 모델을 학습하고 평가하였다. 이를 통해 완전 자동화된 분류 방법을 제안하였고, 자동 분류 결과에 대한 확률을 제시하였다. 이러한 확률을 통해 인공지능이 독립적으로 진단을 내리는 것이 아니라 의사의 판독 이후에 객관적인 진단 지표 즉, 정량적인 근거를 제공하는 것으로, 이는 2차 판독에 유용하게 활용될 수 있다. 또한, 설명 가능 인공지능 기술을 활용하여 모델의 분류 결정에 대한 근거를 MR 영상 내에서 시각적으로 강조하여 나타내어 모델의 자동 분류 결정에 대한 신뢰도를 향상시킬 수 있었다. 이러한 접근 방식은 ADPKD 환자의 임상관리 및 임상 시험에서 효과적이고 효율적으로 활용될 수 있으며, 실제 의료 현장에서 의료 보조 진단 과정을 보다 간편하고 신뢰성 있게 지원함으로써 의료 전문가와 환자들에게 많은 도움을 줄 것으로 기대된다.

참고 문헌

1. 국경완, *인공지능 기술 및 산업 분야별 적용 사례.* 정보통신기획평가원 주간기술동향, 2019(1888): p. 15-27.

2. Hong, J.-Y., S.H. Park, and Y.-J. Jung, *Artificial intelligence based medical imaging: An Overview.* Journal of radiological science and technology, 2020. 43(3): p. 195-208.

3. 백승옥, *엄청난 데이터+딥러닝 기술 한국, 의료영상 진단의 성지 될 수 있다*, in *동아비즈니스리뷰DBR*. 2015.

4. Fenton, J.J., et al., *Influence of computer-aided detection on performance of screening mammography.* New England Journal of Medicine, 2007. 356(14): p. 1399-1409.

5. 이주열, *인공지능 이미지 인식 기술 동향.* 한국정보통신기술협회, 2020: p. 44-51.

6. Umer, M., S. Sharma, and P. Rattan. *A survey of deep learning models for medical image analysis*. in *2021 International Conference on Computing Sciences (ICCS)*. 2021. IEEE.

7. Bennett, W.M., *Autosomal dominant polycystic kidney disease: 2009 update for internists.* The Korean journal of internal medicine, 2009. 24(3): p. 165.

8. 삼성서울병원. *[신장/비뇨기계 질환] 상염색체 우성 다낭성신질환*. 2015.

9. PKD Foundation, *What is ADPKD?* 2021

10. Radhakrishnan, Y., P. Duriseti, and F.T. Chebib, *Management of autosomal dominant polycystic kidney disease in the era of disease-modifying treatment options.* Kidney Research and Clinical Practice, 2022. 41(4): p. 422.

11. Bae, K.T., et al., *Expanded imaging classification of autosomal dominant polycystic kidney disease.* Journal of the American Society of Nephrology: JASN, 2020. 31(7): p. 1640.

12. 박상철, et al., *의료영상 분석을 위한 기계학습.* 정보과학회논문지: 소프트웨어 및 응용, 2012. 39(3): p. 163-174.

13. 이명교, *인공지능 기반 의료영상 분석 기술 동향.* 정보통신기술진흥센터, 2018: p. 2-13.

14. Park, S.C., X.-H. Wang, and B. Zheng, *Assessment of performance improvement in content-based medical image retrieval schemes using fractal dimension.* Academic radiology, 2009. 16(10): p. 1171-1178.

15. 양양정, *초음파 영상 경동맥 플라그 분류를 위한 통계기반 텍스쳐 분석.* 2007.

16. Park, S.C., et al., *Computer-aided detection of early interstitial lung diseases using low-dose CT images.* Physics in Medicine & Biology, 2011. 56(4): p. 1139.

17. Woo, D. and S. Jeong, *Pixel-wise Intensity Feature Enhancement for Improving Deep Learning based Medical Imaging Segmentation.* 전자공학회논문지, 2022. 59(2): p. 51-58.

18. 우상근, 이용진, 김경민, 이원호, 박지애, 강주현, 김민환, 임상무, *다중가우시안혼합모델을 이용한 심장 극성지도에서의 경색 크기 계측 장치 및 방법*. 2011. p. 14.

19. 이한상, 박민석, 김준모, *의료영상에서의 딥 러닝.* 대한의학영상정보학회지, 2014. 20(1): p. 13-18.

20. 이관용, 김진희, 김현철, *의료 인공지능 현황 및 과제.* 보건산업브리프, 2016: p. 1-28.

21. 김동현 조현종, *위 내시경 영상을 이용한 병변 진단을 위한 딥러닝 기반 컴퓨터 보조 진단 시스템.* 전기학회논문지, 2018. 67(7): p. 928-933.

22. 이한성, 조현종, *딥러닝 기반 위 질환 컴퓨터 보조 진단 시스템 개발 및 Grad-CAM 을 활용한 시각화.* 전기학회논문지, 2023. 72(2): p. 234-240.

23. 이경윤, 김영재, 김광기, *위내시경 디지털 영상에서 정상과 위궤양 딥러닝 분류 모델.* Journal of Digital Art Engineering & Multimedia Vol, 2019. 6(2): p. 133-140.

24. Bressem, K.K., et al., *Comparing different deep learning architectures for classification of chest radiographs.* Scientific reports, 2020. 10(1): p. 13590.

25. Zöllner, F.G., et al., *Assessment of kidney volumes from MRI:*

acquisition and segmentation techniques. American Journal of Roentgenology, 2012. 199(5): p. 1060-1069.

26. Bevilacqua, V., et al., *A comparison between two semantic deep learning frameworks for the autosomal dominant polycystic kidney disease segmentation based on magnetic resonance images.* BMC Medical Informatics and Decision Making, 2019. 19(9): p. 1-12.

27. Goel, A., et al., *Deployed deep learning kidney segmentation for polycystic kidney disease MRI.* Radiology: Artificial Intelligence, 2022. 4(2): p. e210205.

28. Raj, A., et al., *Automated prognosis of renal function decline in ADPKD patients using deep learning.* Zeitschrift für Medizinische Physik, 2023.

29. Kim, Y., et al., *A deep learning approach for automated segmentation of kidneys and exophytic cysts in individuals with autosomal dominant polycystic kidney disease.* Journal of the American Society of Nephrology, 2022. 33(8): p. 1581-1589.

30. Chapman, A.B., et al., *The HALT polycystic kidney disease trials: design and implementation.* Clinical journal of the American Society of Nephrology: CJASN, 2010. 5(1): p. 102.

31. Schrier, R.W., et al., *Blood pressure in early autosomal dominant polycystic kidney disease.* New England Journal of Medicine, 2014. 371(24): p. 2255-2266.

32. 김민규, 배현진, *딥러닝 기반 의료영상 분석을 위한 데이터 증강 기법.* Journal of the Korean Society of Radiology (Taehan Yŏngsang Ŭihakhoe chi), 2020. 81(6): p. 1290.

33. He, K., X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition.* in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.

34. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. 30.

35. Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.

36. Han, K., et al., *A survey on vision transformer.* IEEE transactions on pattern analysis and machine intelligence, 2022. 45(1): p. 87-110.

37. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

38. Rawat, W. and Z. Wang, *Deep convolutional neural networks for image classification: A comprehensive review.* Neural computation, 2017. 29(9): p. 2352-2449.

39. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. 25.

40. Russakovsky, O., et al., *Imagenet large scale visual recognition*

challenge. International journal of computer vision, 2015. 115: p. 211-252.

41. Zhu, W., N. Zeng, and N. Wang, *Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations.* NESUG proceedings: health care and life sciences, Baltimore, Maryland, 2010. 19: p. 67.

42. 한지연, 최재식, *설명가능 인공지능.* 소음· 진동, 2017. 27(6): p. 8-13.

43. Adadi, A. and M. Berrada, *Peeking inside the black-box: a survey on explainable artificial intelligence (XAI).* IEEE access, 2018. 6: p. 52138-52160.

44. 류한석, *[IT칼럼]AI의 복잡성과 블랙박스 문제*, in *주간경향*. 2023.

45. 한형진, *의료/헬스케어 분야에서의 설명 가능 인공지능 (Explainable AI) 연구 동향.* BRIC View 동향리포트, BRIC View, 2021: p. T13.

46. 최재식, *설명가능 인공지능 연구동향.* 정보과학회지, 2019. 37(7): p. 8-14.

47. DW, G.D.A., *DARPA's explainable artificial intelligence program.* AI Mag, 2019. 40(2): p. 44.

48. Ribeiro, M.T., S. Singh, and C. Guestrin. *" Why should i trust you?" Explaining the predictions of any classifier*. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

49. Achanta, R., et al., *SLIC superpixels compared to state-of-the-art superpixel methods.* IEEE transactions on pattern analysis and machine intelligence, 2012. 34(11): p. 2274-2282.

Artificial Intelligence-based Automated Classification and Analysis of Individuals with Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Seonah Bu

Department of Electronic Engineering
The Graduate School
Jeju National University

Abstract

Autosomal dominant polycystic kidney disease (ADPKD) is a genetic disorder characterized by the development of multiple cysts in the kidneys. To date, there is no fundamental treatment for ADPKD, so quantitative classification of patients and risk prediction through image analysis is crucial for clinical management and trials. Mayo imaging classification is one of the widely accepted quantification tool exploiting height-adjusted total kidney volume and age. Nevertheless, this tool can be applied to the patients with class 1 (typical) only, and class 2 (atypical) with prominent exophytic cysts are excluded. Manual classification of class 1 and 2 is performed as pre-step procedure, but this is time-consuming, onerous, and subject to intra-rater and inter-rater variability.

In the era of the Fourth Industrial Revolution utilizing an artificial intelligence techniques, a deep learning plays a significant role in healthcare applications. In this study, we propose a deep learning-based fully automated methodology to automatically classify class 1 and 2 in individuals affected by ADPKD using MR images showing diagnostic confidence of the automatic

classification results. An explainable artificial intelligence (XAI) method is also applied to increase the explainability of the automated classification results. To this end, MR images from 486 ADPKD patients participating in the HALT-PKD study were utilized for data preprocessing and augmentation. We trained and tested the deep learning models with residual network (ResNet)-18, 34, 50 and vision transformer (ViT) and applied transfer learning using pre-trained weights trained on the ImageNet-1K dataset. Accordingly, we utilized the output from ResNet-50, showing best performance, to obtain the diagnostic confidence of the classification results using the softmax function, followed by the superpixels were generated in the MR image so as to highlight contributing regions using the XAI methode. For the performance evaluation, we utilized confusion matrices, the receiver operating characteristic (ROC) curves, and area under the curve (AUC).

Experimental results showed that ResNet-50 model performed the best in automated classification, with class 1 at 97.7%, class 2 at 100%, and an average test accuracy of 98.01%. The precision, recall and F1-score for predicting the class 1 were 1, 0.98 and 0.99, respectively, while those for predicting the class 2 were 0.87, 1, and 0.93, respectively.

The automated classification method proposed in this study and its probabilities can be utilized as an objective indicator for a second opinion after the physicians' interpretation, allowing the physician to thoroughly examine the medical images again if the diagnostic confidence values are ambiguous, enhancing the accuracy and effectiveness of the diagnosis. In addition, the rationale for the model's classification decision based on XAI was visually highlighted within the MR image to improve the confidence in the model's automated decision. The proposed method could effectively facilitate in the clinical management and trials for patients with ADPKD, and is expected to benefit healthcare professionals and patients by making the diagnosis process simpler and more reliable in daily clinical routines.