

클래스 조건부 테스트 단계 증강 기반 전립선암 악성도 예측

정주립¹⁰

이민주¹

황성일²

홍헬렌^{1*}

¹서울여자대학교 소프트웨어융합학과

{jjulip, minjulee, hlhong}@swu.ac.kr

²분당서울대학교병원 영상의학과

hwangsi49@gmail.com

Prostate Cancer Aggressiveness Prediction Using Class-conditional Test-Time Augmentation

Julip Jung¹⁰

Min Ju Lee¹

Sung Il Hwang²

Helen Hong^{1*}

¹ Department of Software Convergence, Seoul Women's University

² Department of Radiology, Seoul National University Bundang Hospital

요약

본 연구는 전립선암 악성도 예측을 위한 테스트 단계 증강 기법으로, 클래스 조건에 따라 증강 결과를 집계하는 Class-conditional Test-Time Augmentation(ClassTTA) 프레임워크를 제안한다. 기존 테스트 단계 증강 방식은 모든 증강 결과를 동일한 가중치로 평균하여, 증강 효과의 클래스별 차이를 충분히 반영하지 못하는 한계를 가진다. 제안하는 ClassTTA는 테스트 단계에서 생성된 다수의 증강 예측 결과에 대해 클래스별로 독립적인 가중치를 학습함으로써, 고위험 전립선암과 저위험 전립선암에 대해 차별화된 증강 활용 전략을 적용한다. 전립선 다중 파라미터 자기공명영상(Magnetic Resonance Imaging, MRI) 데이터셋을 이용한 실험 결과, ClassTTA는 기존 traditional TTA 대비 균형 정확도(Balanced Accuracy)를 0.81에서 0.85로, 민감도(Sensitivity)를 0.74에서 0.83으로 향상시키며 전반적인 분류 성능을 일관되게 개선하였다. 또한 SHAP 기반 분석을 통해, 제안 방법이 클래스별로 유의미한 증강만을 선택적으로 활용하는 집계 전략을 학습함을 확인하였다.

Abstract

This study proposes a Class-conditional Test-Time Augmentation(ClassTTA) framework for prostate cancer aggressiveness prediction, which aggregates augmented predictions using class-specific weighting strategies. Conventional test-time augmentation methods rely on simple averaging and do not adequately account for class-dependent effects of different augmentations. The proposed ClassTTA addresses this limitation by learning independent aggregation weights for each class, enabling different augmentation utilization strategies for high-grade and low-grade prostate cancer. Experimental results on multiparametric prostate Magnetic Resonance Imaging (MRI) datasets demonstrate that ClassTTA consistently improves overall classification performance over conventional TTA by increasing balanced accuracy from 0.81 to 0.85 and sensitivity from 0.74 to 0.83. In addition, SHAP-based analysis confirms that the proposed method learns an aggregation strategy that selectively utilizes meaningful data augmentations for each class.

키워드: 전립선암, 예측, 테스트 단계 증강, 다중 파라미터 자기공명영상

Keywords: prostate cancer, prediction, test-time augmentation, multiparametric MRI

*corresponding author: Helen Hong / Department of Software Convergence, Seoul Women's University (hlhong@swu.ac.kr)

Received : 2026.02.09./ Review completed : 1st 2026.04.03. 2nd 2026.06.15./ Accepted : 2026.08.03.

DOI : 10.15701/kcgs.2026.32.4.57

ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

전립선암(Prostate Cancer, PCa)은 남성에서 가장 흔하게 진단되는 암 중 하나로, 암 관련 사망의 주요 원인으로 보고되고 있다[1]. 전립선암은 종양의 악성도에 따라 임상적 경과와 치료 전략이 크게 달라지기 때문에, 질병의 진행 단계에서 악성도를 정확하게 평가하는 것은 환자의 예후 예측과 치료 결정에 있어 핵심적인 요소이다[2]. 현재 임상에서는 조직 생검(biopsy)을 통해 채취된 조직의 형태학적 패턴을 기반으로 글리슨 점수(Gleason Score)를 산출하여 전립선암의 악성도를 평가한다. 글리슨 점수는 가장 우세한 암세포 패턴과 두 번째로 우세한 패턴을 각각 1점에서 5점까지 평가하여 합산한 지표로, 전립선암의 조직학적 등급과 예후를 판단하는 표준 지표로 널리 사용되고 있다[3]. 일반적으로 점수가 높을수록 고위험 전립선암(high-grade PCa, hPCa), 낮을수록 저위험 전립선암(low-grade PCa, lPCa)으로 분류되며, 이는 치료 방향을 결정하는 중요한 기준이 된다. 그러나 조직 생검은 침습적인 절차로 환자에게 신체적 부담을 주며, 국소적인 조직 샘플링에 의존하기 때문에 병변 전체의 특성을 충분히 반영하지 못한다는 한계가 있다. 또한 병리 판독 과정에서 관찰자의 주관적 개입될 수 있어 진단 결과의 변동성이 발생할 수 있으며, 이는 불필요한 반복 생검이나 과잉 치료로 이어질 위험이 있다[4].

이러한 한계를 보완하기 위한 대안으로, 최근에는 자기공명영상(Magnetic Resonance Imaging, MRI)을 활용한 비침습적 전립선암 진단 방법이 주목받고 있다. 전립선 MRI는 해부학적 구조를 제공하는 T2-weighted 영상, 조직의 확산 특성을 반영하는 Diffusion-Weighted Imaging (DWI), 그리고 정량적 확산 지표인 Apparent Diffusion Coefficient (ADC) map 등 서로 보완적인 정보를 제공하는 다중 모달리티로 구성된다. 이러한 다중 파라미터 영상 정보를 종합적으로 활용함으로써 전립선암 악성도 평가에 있어 병변의 형태 및 기능적 특성을 보다 정밀하게 반영할 수 있다[5].

최근 딥러닝 기술은 의료 영상 분석 분야 전반에서 우수한 성능을 보이며, 전립선 MRI를 이용한 PCa 진단 및 악성도 예측에서도 높은 잠재력을 보여주고 있다. 특히 딥러닝 모델은 임상적으로 의미 있는 전립선암을 효과적으로 탐지할 수 있어 기존 진단 방법을 보완하는 도구로 활용될 수 있다. 그러나 이러한 모델은 일반적으로 단일 입력에 대한 단일 예측 결과에 의존하기 때문에, 영상 획득 조건의 변화나 미세한 입력 변동에 민감하게 반응하여 예측 성능이 불안정해질 수 있다. 또한 모델의 내부 의사결정 과정이 명확히 드러나지 않는 블랙박스 특성으로 인해, 실제 임상 환경에서 예측 결과를

신뢰하고 해석하는 데 어려움이 존재한다.

이러한 문제를 완화하기 위한 접근으로 Test-Time Augmentation(TTA)이 제안되었다[6]. TTA는 테스트 단계에서 입력 영상을 다양한 방식으로 변형한 후, 각 변형에 대한 예측 결과를 집계함으로써 모델의 강건성을 향상시키는 기법이다. 이러한 접근은 Krizhevsky 등[7]이 다중 크롭 및 반전 기반의 테스트 앙상블을 통해 ImageNet 분류 성능을 향상시킨 사례를 포함한 초기 연구들에서 그 효과가 관찰된 이후, Shanmugam 등[8]의 TTA 집계 방식 분석 등으로 발전하며 컴퓨터 비전 분야에서 널리 활용되는 성능 향상 기법으로 확립되었다. 또한 Ayhan 등[9]은 TTA를 통해 생성된 예측 분포를 활용하여 모델의 불확실성을 정량화할 수 있음을 보이며, 입력 변동에 대한 분석 도구로서의 가능성을 제시하였다. 특히 의료 영상에서는 영상 획득 조건, 노이즈, 환자 간 해부학적 차이 등으로 인해 입력 데이터의 분포 변동이 크게 발생하기 때문에, 이러한 변동성에 대응하기 위한 방법으로 TTA가 효과적으로 활용되어 왔다. 피부암 분류, 유방암 검출 등 다양한 의료 영상 분석 태스크에서 TTA는 소규모 데이터 환경에서도 일반화 성능을 보완하고, 입력 변화에 따른 예측 불안정성을 완화하는 데 기여하는 것으로 보고되었다[10,11]. 그러나 기존의 전통적인 TTA 방법은 대부분 모든 증강 결과를 동일한 가중치로 단순 평균하는 방식에 기반하고 있어, 서로 다른 증강이 예측 결과에 미치는 영향의 차이를 충분히 반영하지 못한다는 한계가 있다. 일부 증강은 특정 샘플이나 특정 클래스에 대해 유용한 정보를 제공하는 반면, 다른 증강은 오히려 예측을 혼란스럽게 만들 수 있음에도 불구하고, 이들 기여도가 동일하게 취급된다는 점이 지적되어 왔다. 특히 의료 영상에서는 병변의 크기, 위치, 영상 품질 등에 따라 증강의 효과가 크게 달라질 수 있어, 고정된 집계 방식은 성능 안정성 측면에서 한계를 가진다.

이러한 한계를 극복하기 위해 최근에는 증강 결과의 중요도를 학습적으로 반영하는 적응적 TTA 방법들이 제안되고 있다. Kim 등[12]은 각 증강 입력에 대한 예상 손실을 추정하여 가장 유의미한 변환을 선택하는 loss-prediction 기반 TTA를 제안하였으며, Chun 등[13]은 반복적인 최적화 과정을 통해 증강 선택을 점진적으로 개선하는 Cyclic TTA를 제안하였다. 또한 Son 등[14]은 모델의 불확실성을 기반으로 신뢰도가 높은 경우에만 증강을 선택적으로 적용하는 Selective TTA를 통해 성능과 효율성을 동시에 향상시켰다. 이러한 접근들은 단순 평균 방식 대비 성능 향상 가능성을 보여주었으나, 대부분 클래스 불변 방식으로 설계되어 있어 모든 클래스에 대해 동일한 증강 활용 전략을 적용한다는 공통적인 한계를 가진다. 특히 전립선암 진단과 같이 고위험군과 저위

험군의 임상적 중요도가 비대칭적인 상황에서는, 클래스별로 차별화된 증강 전략이 필요하다.

본 연구에서는 이러한 한계를 극복하기 위해 Class-conditional Test-Time Augmentation(ClassTTA)을 제안한다. 제안 방법은 테스트 단계에서 생성된 다수의 증강 예측 결과를 클래스 조건에 따라 상이한 가중치로 집계함으로써, hPCa와 lPCa에 대해 서로 다른 증강 활용 전략을 학습하도록 설계되었다. ClassTTA는 고위험 전립선암과 저위험 전립선암에 대해 서로 다른 증강 가중치를 학습하여 각 클래스의 특성에 최적화된 예측을 수행한다. 이를 통해 임상적으로 중요한 hPCa에 대한 민감도를 향상시키는 동시에, lPCa에서 불필요한 오탐을 억제하여 특이도를 유지함으로써 실제 임상 환경에서의 실용성을 높인다. 또한 증강 결과의 클래스별 기여도를 정량적으로 분석하여 모델의 예측 과정에 대한 해석 가능성을 제공하고, 어떤 영상 변형이 각 클래스 예측에 유용한지에 대한 임상적 통찰을 제공한다. 제안 방법의 성능은 분당서울대병원의 특성이 다른 두 개의 전립선 MRI 데이터셋을 통해 검증되었으며, 기존 TTA 방법 대비 우수한 악성도 분류 성능을 달성하였다.

2. 제안 방법

본 장에서는 전립선암 악성도 예측을 위한 ClassTTA 프레임워크의 구조와 동작 원리를 설명한다. 제안하는 ClassTTA는 테스트 단계에서 생성된 다수의 증강 예측 결과를 단순 평균하는 기존 TTA 방식의 한계를 극복하기 위해, 클래스 조건에 따라 증강 결과의 기여도를 학

습적으로 조절하는 집계 메커니즘을 도입한다. 핵심 아이디어는 모든 증강이 모든 클래스에 동일하게 유용하다는 가정을 전제하지 않고, 특정 증강이 특정 클래스의 예측에는 유리하지만 다른 클래스에는 불리하게 작용할 수 있다는 점을 집계 단계에서 반영하는 것이다. 이를 통해 입력 변동에 대한 평균적인 예측 안정성을 확보함과 동시에, 클래스별 임상적 요구사항을 고려한 보다 유연한 예측 전략을 구현하고자 한다.

그림 1은 제안하는 ClassTTA 프레임워크의 전체 개요를 나타낸다. 본 프레임워크는 전립선 MR 영상으로부터 악성도를 예측하는 기본 예측 네트워크를 기반으로, 테스트 단계에서 입력 변동성을 확장하기 위한 다수의 증강 변형을 적용하고, 이를 통해 생성된 예측 결과들을 클래스 조건부 집계 과정을 통해 통합하는 구조로 구성된다.

입력 영상 x 는 전립선 병변이 포함된 다중 파라미터 MR 영상으로 구성되며, 본 연구에서는 해부학적 정보를 제공하는 T2-weighted 영상, 조직의 확산 특성을 반영하는 DWI, 그리고 정량적 확산 지표인 ADC map을 사용한다. 각 모달리티는 동일한 병변 영역을 기준으로 정렬된 상태로 입력되며, 이들 정보는 기본 예측 네트워크 내부에서 공동으로 처리되어 클래스별 예측 점수를 산출한다. 테스트 단계에서는 입력 영상에 대해 다수의 증강 변형을 적용하여 여러 개의 증강 입력을 생성하고, 각 증강 입력으로부터 얻어진 예측 결과를 클래스 조건부 집계 과정을 통해 통합함으로써 최종 예측을 도출한다.

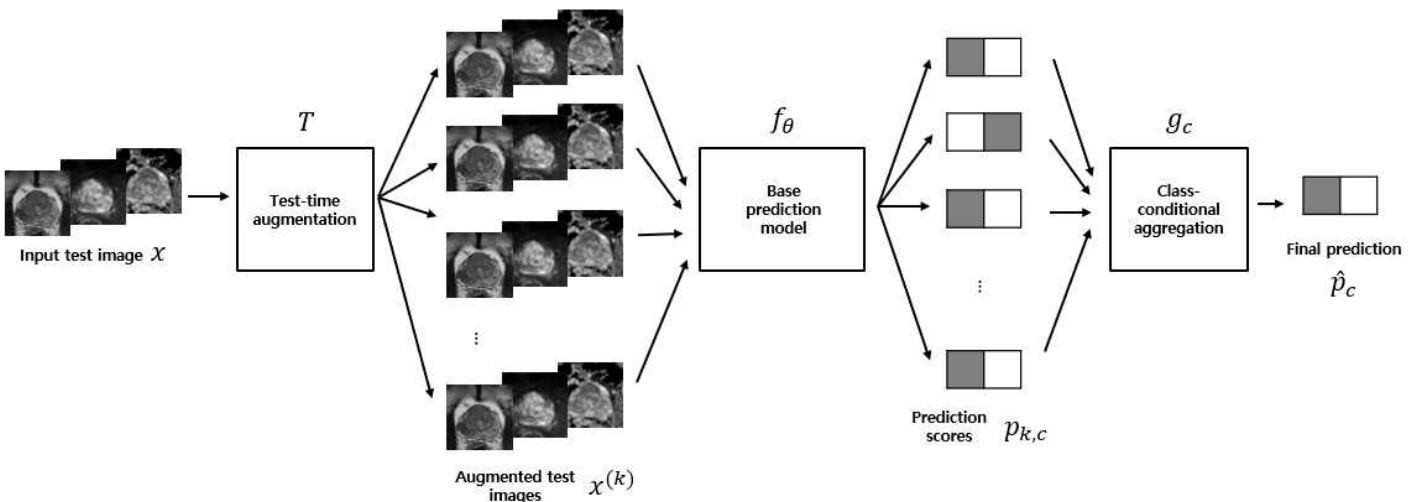


Figure 1. Overview of the proposed Class-conditional Test-Time Augmentation (ClassTTA) framework. Given an input test image X , multiple augmented samples $X^{(k)}$ are generated using a test-time augmentation operator T . Each augmented input is processed by a fixed, pre-trained base prediction model f_θ to produce class-wise prediction scores $p_{k,c}$. The final prediction $\hat{p}_c$ is obtained by aggregating the augmented predictions through a class-conditional aggregation function g_c .

2.1 테스트 단계 증강 전략

테스트 단계에서 입력 영상 x 에 대해 총 K 개의 증강 변형 $\{T_1, T_2, \dots, T_K\}$ 을 적용하여 증강 입력 집합 $\{x^{(k)}\}_{k=1}^K$ 을 생성한다. 각 증강 입력은 다음과 같이 정의된다.

$$x^{(k)} = T_k(x), \quad k = 1, \dots, K$$

본 연구에서는 총 K 개의 증강 변형을 사용하며, 증강 전략은 경미한 변형과 강한 변형의 두 가지 범주로 구성된다.

경미한 변형은 기본 예측 네트워크의 학습 단계에서도 사용된 증강으로, 수평 뒤집기(horizontal flip), 제한된 범위의 회전($\pm 5^\circ$), 그리고 소규모 평행 이동($\pm 20\%$)을 포함한다. 이러한 증강은 영상의 구조적 특성을 유지하면서, 영상 획득 조건의 미세한 변화로 인한 입력 변동성을 모사하는 것을 목표로 한다.

강한 변형은 보다 극단적인 입력 변화를 가정하여 모델의 강건성을 평가하기 위해 포함된 증강으로, 수직 뒤집기(vertical flip), 큰 각도의 회전($\pm 90^\circ$), 대규모 평행 이동($\pm 50\%$), 그리고 밝기 및 대비 조절을 포함한다. 이러한 변형은 영상의 시각적 특성을 크게 변화시키므로, 증강 종류에 따라 예측 특성이 달라질 수 있다.

이와 같이 생성된 증강 입력들은 이후 단계에서 동일한 예측 네트워크를 통해 처리되며, 본 절에서는 증강 입력의 생성 과정까지만을 다룬다. 증강 입력으로부터 얻어진 예측 결과의 통합 방식은 다음 절에서 설명한다.

2.2 클래스 조건부 학습형 가중치 집계

2.1절에서 설명한 테스트 단계 증강을 통해 입력 영상으로부터 K 개의 증강 입력이 생성되며, 이에 대한 클래스별 예측 점수들이 기본 예측 네트워크로부터 얻어진다. 본 절에서는 이러한 증강별 예측 결과를 클래스 조건에 따라 통합하는 학습형 가중치 집계 메커니즘을 설명한다.

각 증강 입력 $x^{(k)}$ 에 대해, 사전 학습된 기본 예측 네트워크 f_θ 는 클래스 c 에 대한 예측 점수 $p_{k,c}$ 를 출력한다. 여기서 $k \in \{1, \dots, K\}$ 는 증강 인덱스를, $c \in \{1, \dots, C\}$ 는 클래스 인덱스를 의미한다. 테스트 단계에서는 기본 예측 네트워크의 파라미터를 고정된 상태로 추론만 수행되며, 모든 증강 입력은 동일한 네트워크를 공유한다. ClassTTA에서는 각 클래스 c 와 증강 k 에 대해 독립적인 집계 가중치 $\alpha_{k,c}$ 를 정의하고, 이를 학습 가능한 파라미터로 설정한다. 이러한 가중치는 증강별 예측 결과가 최종 예측에 기여하는 상대적인 중요도를 클래스 조건부로 모델링한다. 클래스 c 에 대한 최종 예측 점수 $\hat{p}_c$ 는 증강별 예측 점수들을 클래스 조건부 집계 함수 g_c 를

통해 다음과 같이 계산된다.

$$\hat{p}_c = g_c(\{p_{k,c}\}_{k=1}^K) = \sum_{k=1}^K \alpha_{k,c} p_{k,c}$$

여기서 g_c 는 클래스 c 에 대해 정의된 선형 가중 합 형태의 집계 함수로, 테스트 단계에서 생성된 다수의 증강 예측 결과를 클래스 조건에 따라 통합하는 역할을 한다.

이와 같은 클래스 조건부 가중 합 방식은 모든 증강 결과를 동일한 비중으로 평균하는 기존의 테스트 단계 증강 방식과 달리, 클래스별로 유의미한 증강의 기여를 강조하거나 불필요한 증강의 영향을 억제할 수 있도록 한다. 특히 고위험 전립선암과 저위험 전립선암 간의 영상적 특성과 임상적 중요도가 상이한 상황에서, 클래스 조건부 집계는 보다 유연하고 효과적인 예측 전략을 제공한다.

집계 가중치 $\alpha_{k,c}$ 는 검증 데이터셋을 사용하여 학습되며, 이 과정에서 기본 예측 네트워크 f_θ 의 파라미터는 업데이트되지 않는다. 따라서 ClassTTA의 성능 향상은 특정 추출 네트워크의 구조적 변경이 아니라, 테스트 단계에서 생성된 다수의 예측 결과를 클래스 조건에 따라 통합하는 집계 전략에서 비롯된다.

2.3 학습 및 추론

ClassTTA의 학습은 두 단계로 진행된다. 첫 번째 단계에서는 기본 예측 네트워크 f_θ 를 학습 데이터셋을 사용하여 일반적인 지도 학습 방식으로 사전 학습한다 [17]. 이 단계에서는 다양한 데이터 증강 기법을 적용하여 모델의 일반화 성능을 향상시킨다. 두 번째 단계에서는 클래스 조건부 집계 가중치를 학습한다. 검증 데이터셋의 각 샘플에 대해 K 개의 예측 점수 $\{p_{k,c}\}$ 를 얻는다. 이 예측 점수들을 사용하여 집계 가중치를 학습하며, 이 과정에서 기본 예측 네트워크의 파라미터는 고정된다. 테스트 단계에서는 새로운 입력 영상에 대해 동일한 방식으로 K 개의 증강 예측을 생성하고, 학습된 클래스 조건부 집계 가중치를 적용하여 최종 예측 $\hat{p}_c$ 를 산출한다. 이 과정에서는 추가적인 모델 학습이나 파라미터 업데이트가 발생하지 않는다.

2.4 해석 가능성 분석

ClassTTA는 예측 성능 향상뿐만 아니라, 모델의 의사 결정 과정에 대한 해석 가능성도 제공한다. 본 연구에서는 SHAP(SHapley Additive exPlanations) 분석 기법 [15]을 활용하여 각 증강 예측 점수 $p_{k,c}$ 가 최종 예측 $\hat{p}_c$ 에 미치는 기여도를 정량적으로 분석한다. 증강별 예측 점수를 입력 특징으로 간주하여 SHAP 값을 계산함으로써, 특정 증강이 각 클래스의 예측에 긍정적 또는 부정적으

로 기여하는 정도를 평가할 수 있다. 이를 통해 ClassTTA가 클래스별로 차별화된 증강 활용 전략을 학습했는지를 시각적으로 분석한다. 예를 들어, 특정 기하학적 변형은 고위험 전립선암의 불규칙한 형태적 특징을 강조하여 높은 기여도를 보이는 반면, 동일한 증강이 저위험 전립선암에서는 억제된 기여도를 나타낼 수 있다. 이러한 분석은 ClassTTA가 임상적으로 의미 있는 방식으로 증강 정보를 선택적으로 활용하고 있음을 시사한다.

3. 실험 및 결과

본 장에서는 제안하는 ClassTTA의 성능을 검증하기 위해 수행한 실험 설정과 정량적 평가 결과 및 정성적 분석 결과를 제시한다. 제안 방법은 전립선 다중 파라미터 MR 영상을 이용한 악성도 예측 과제에서 기존 TTA 기법들과 비교 평가되었다.

3.1 실험 데이터

본 연구에서는 서로 다른 시점에 수집된 두 개의 전립선 다중 파라미터 자기공명영상(multiparametric MRI, mpMRI) 데이터셋을 사용하였다. Dataset 1은 2011년 1월부터 2018년 4월까지, Dataset 2는 2015년 5월부터 2018년 11월까지 수집되었으며, 두 데이터셋 모두 분당 서울대학교병원에서 획득되었고 기관생명윤리위원회(Institutional Review Board, IRB)의 승인을 받았다. 대상 환자는 전립선 특이 항원(Prostate-Specific Antigen, PSA) 수치(≥ 3), 경직장 초음파 검사, 직장수지검사를 통해 전립선암이 의심되어 조직 생검을 시행한 환자들로 구성되었다. 이 중 mpMRI 촬영을 수행하고 병리학적 정보가 확보된 환자만 데이터셋으로 구성하였다.

Dataset 1에는 256명의 환자로부터 313개의 병변이 포함되었고, Dataset 2에는 306명의 환자로부터 371개의 병변이 포함되었다. 각 병변은 글리슨 점수를 기준으로 저위험 전립선암(IPCa)과 고위험 전립선암(hPCa)으로 이진 분류되었으며 임상주의 의견에 따라 $GS \leq 3+4$ 는 저위험군으로, $GS \geq 4+3$ 는 고위험군으로 정의하였다. 두 데이터셋은 수집 시점의 차이로 인해 일부 환자에서 생검 이후 촬영된 영상에 기인한 출혈 신호(bleeding effect)가 관찰되었으며, 이러한 출혈 신호는 Dataset 2에서 상대적으로 더 빈번하게 나타나는 경향을 보였다.

모든 mpMRI 영상은 병리 정보에 기반하여 정의된 병변 영역을 기준으로 강제 정합(rigid registration)을 수행한 후 정렬되었으며, 이는 선행 연구[16]에서 제안된 전처리 절차를 따른 것이다. 이때, ADC map은 T2-weighted 영상과의 정합을 위해 상호정보(mutual information) 기반의 강제 정합(rigid registration)을 적용

하여 translation 및 rotation 변환을 통해 정렬하였다. 또한, DWI는 ADC map과 동일한 변환 파라미터를 적용하여 T2-weighted 영상에 정렬되도록 하였으며, 이는 두 영상이 동일한 확산기반에서 획득되어 공간적으로 일관된 구조를 가지기 때문이다. 이후 병변 중심을 기준으로 영상 패치를 생성하였으며, 패치 생성 과정은 중앙 중심 패치 및 크기 정규화 패치 생성 방법을 제안한 선행 연구[16]의 방법에 따라 수행되었다. 본 연구에서는 T2-weighted 영상, DWI, ADC map의 세 모달리티에 대해 동일한 정렬 및 패치 추출 절차를 적용하였다.

3.2 실험 설정

전립선암 악성도 예측 실험은 다중 파라미터 MR 영상을 입력으로 하여 저위험 전립선암과 고위험 전립선암을 이진 분류하는 문제로 수행하였다. 기본 예측 네트워크는 EfficientNet-B0을 백본으로 사용하였으며, 학습 단계에서는 일반적인 데이터 증강 기법을 적용하여 모델을 사전 학습하였다[17].

본 연구에서는 기본 예측 네트워크 학습과 클래스 조건부 집계 모듈 학습의 목적이 상이함을 고려하여, 각 단계에 대해 독립적인 데이터 분할 전략을 적용하였다. 전체 악성도 예측 성능 평가를 위한 기본 예측 네트워크 학습에는 hold-out validation 방식(6:2:2)을 사용하였으며, Dataset 1은 189개를 학습, 72개를 검증, 52개를 테스트에 사용하였고, Dataset 2는 223개를 학습, 86개를 검증, 62개를 테스트에 사용하였다. 반면, ClassTTA의 클래스 조건부 집계 가중치 학습 및 평가에는 5-fold cross-validation 방식을 적용하였다. 각 데이터셋은 5개의 fold로 분할되었으며, 각 반복에서 4개의 fold를 집계 가중치 학습에 사용하고 나머지 1개의 fold를 평가에 사용하였다. 이러한 설정은 제한된 데이터 환경에서 특정 데이터 분할에 대한 과적합 가능성을 줄이고, 제안 방법의 일반화 성능을 보다 안정적으로 평가하기 위함이다.

ClassTTA의 집계 가중치는 사전 학습된 기본 예측 네트워크를 고정된 상태에서 각 fold의 학습 데이터에 대해 학습되었다. 테스트 단계에서는 기본 예측 네트워크의 학습 단계에서 사용된 19개의 경미한 변형만을 적용한 설정(mild)과, 여기에 7개의 강한 변형을 추가한 총 26개의 증강 변형을 적용한 설정(severe)의 두 가지 증강 전략을 사용하였으며, 모든 비교 실험에서 동일한 증강 집합과 동일한 기본 예측 네트워크를 사용하였다.

제안하는 ClassTTA의 성능을 평가하기 위해 테스트 단계 증강을 적용하지 않고 단일 입력을 예측한 결과만을 사용하는 경우(baseline), 모든 증강 예측 결과를 동일한 가중치로 평균하는 전통적인 TTA(baseline + traditional TTA) 방식, 증강별 가중치를 학습하되 클래스

스 구분 없이 동일한 가중치를 적용하는 학습형 TTA 방식(Augmentation-conditional TTA, AugTTA)과 비교 실험을 수행하였다. 모든 방법(baseline, traditional TTA, AugTTA, ClassTTA)에 대해 동일한 fold 분할과 평가 절차를 적용하였다.

평가 지표로는 분류 성능을 정량적으로 비교하기 위해 정확도(Accuracy), 균형 정확도(Balanced Accuracy), 민감도(Sensitivity), 특이도(Specificity), 그리고 AUC(Area Under the ROC Curve)를 사용하였다. 또한 제안 방법의 성능 향상이 통계적으로 유의미한지를 검증하기 위해, 각 방법의 5-fold cross-validation 결과에 대해 대응 표본 t-검정(paired t-test)을 수행하였다.

3.3 결과

본 장에서는 제안하는 ClassTTA의 성능을 기존 TTA 기반 방법들과 비교하고, 테스트 단계 변형에 대한 강건성, 데이터셋 간 일반화 성능, 그리고 모델의 의사결정

해석 가능성을 단계적으로 분석한다. 이를 통해 ClassTTA가 단순한 성능 향상뿐만 아니라, 다양한 테스트 조건에서도 안정적으로 동작하는 예측 프레임워크임을 검증하고자 한다.

표 1과 표 2는 각각 Dataset 1과 Dataset 2에서 수행한 5-fold cross-validation 기반 전립선암 악성도 예측 성능 비교 결과를 나타낸다. 두 데이터셋 모두에서 ClassTTA는 baseline, traditional TTA, 그리고 AugTTA 대비 전반적으로 가장 우수하거나 경쟁력 있는 성능을 보였다. 성능 수치는 5-fold cross-validation 결과를 기반으로 평균과 표준편차로 산출하였다. 표에서는 정량적 비교를 위해 성능 지표를 0-1 범위로 제시하였으며, 본문에서는 동일한 값을 가독성을 고려하여 백분율(%)로 환산하여 서술한다.

Table 1. Performance comparison of prostate cancer aggressiveness prediction on Dataset 1 using different test-time augmentation strategies. Results are reported as mean $\pm$ standard deviation over 5-fold cross-validation. ClassTTA achieved the best overall performance and significantly outperformed both the baseline ($p < 0.001$) and AugTTA ($p = 0.038$). The highest values are denoted in bold.

Method	Balanced Accuracy	Accuracy	Sensitivity	Specificity	AUC
Baseline	0.76 $\pm$ 0.01	0.77 $\pm$ 0.01	0.70 $\pm$ 0.07	0.82 $\pm$ 0.06	0.75 $\pm$ 0.01
Baseline + traditional TTA	0.81 $\pm$ 0.03*	0.80 $\pm$ 0.03	0.74 $\pm$ 0.06	0.87 $\pm$ 0.03	0.76 $\pm$ 0.05
Baseline + augTTA	0.83 $\pm$ 0.02*	0.82 $\pm$ 0.03	0.79 $\pm$ 0.05	0.87 $\pm$ 0.03	0.76 $\pm$ 0.04
Baseline + classTTA	0.85 $\pm$ 0.01^{††}	0.85 $\pm$ 0.01	0.83 $\pm$ 0.04	0.87 $\pm$ 0.03	0.80 $\pm$ 0.03

*: $p < 0.05$ vs Baseline

†: $p = 0.038$ vs Baseline + augTTA

‡: $p < 0.001$ vs Baseline

Table 2. Performance comparison of prostate cancer aggressiveness prediction on Dataset 2 using different test-time augmentation strategies. Results are reported as mean $\pm$ standard deviation over 5-fold cross-validation. ClassTTA achieved the best overall performance and showed statistically significant improvements over the baseline and traditional TTA. The highest values are denoted in bold.

Method	Balanced Accuracy	Accuracy	Sensitivity	Specificity	AUC
Baseline	0.70 $\pm$ 0.01	0.69 $\pm$ 0.01	0.66 $\pm$ 0.02	0.75 $\pm$ 0.04	0.76 $\pm$ 0.02
Baseline + traditional TTA	0.77 $\pm$ 0.01*	0.76 $\pm$ 0.01	0.82 $\pm$ 0.10	0.71 $\pm$ 0.09	0.73 $\pm$ 0.04
Baseline + augTTA	0.79 $\pm$ 0.01 [†]	0.79 $\pm$ 0.02	0.83 $\pm$ 0.07	0.75 $\pm$ 0.07	0.75 $\pm$ 0.02
Baseline + classTTA	0.80 $\pm$ 0.03^{**}	0.79 $\pm$ 0.01	0.84 $\pm$ 0.11	0.75 $\pm$ 0.06	0.76 $\pm$ 0.04

*: $p < 0.01$ vs Baseline

†: $p < 0.001$ vs Baseline

‡: $p < 0.05$ vs Baseline + traditional TTA

Dataset 1에서는 테스트 단계에서 증강을 적용하지 않은 기본 모델은 균형 정확도 76%, 정확도 77%, 민감도 70%, 특이도 82%, AUC 0.75를 기록하였다. 전통적인 TTA를 적용한 경우, 균형 정확도는 81%, 정확도는 80%, 민감도는 74%로 향상되었으며, 특이도 역시 87%로 증가하였다. 그러나 AUC는 0.76 수준으로 제한적인 개선만이 관찰되었다. 증강별 가중치만 학습하는 AugTTA는 균형 정확도 83%, 정확도 82%, 민감도 79%를 기록하며 traditional TTA 대비 추가적인 성능 향상을 보였고, 특이도 또한 87%로 유지되었다. 반면, 제한하는 ClassTTA는 균형 정확도 85%, 정확도 85%, 민감도 83%, 특이도 87%, AUC 0.80으로 전반적인 평가 지표에서 가장 우수한 성능을 기록하였다. 특히 임상적으로 중요한 고위험 전립선암 검출 성능을 반영하는 민감도는 baseline 대비 약 13%p 향상되었으며, 특이도 또한 안정적으로 유지되었다. 또한 균형 정확도 기준 baseline 및 AugTTA 대비 통계적으로 유의미한 성능 향상이 확인되었다($p < 0.05$).

Dataset 2에서도 유사한 성능 향상 경향이 관찰되었다. 기본 모델은 균형 정확도 70%, 정확도 69%, 민감도 66%, 특이도 75%, AUC 0.76을 기록하였다. Traditional TTA를 적용한 경우 균형 정확도와 정확도가 각각 77%, 76%로 향상되었으며, 민감도 또한 82%까지 증가하였다. AugTTA는 균형 정확도 79%, 정확도 79%, 민감도 83%를 기록하며 추가적인 성능 향상을 보였다. 제한하는 ClassTTA는 균형 정확도 80%, 정확도 79%, 민감도 84%를 기록하여 가장 우수한 분류 성능을 보였으며, baseline 대비 균형 정확도가 약 10%p 향상되었다.

한편, ClassTTA 적용 후 민감도와 특이도의 균형이 개선된 것에 비해 AUC의 향상은 상대적으로 제한적으로 나타났다. 이는 민감도와 특이도가 특정 결정 임계값에서의 분류 결과를 반영하는 반면, AUC는 모든 임계값에 걸쳐 고위험군과 저위험군의 전반적인 예측 순위 구분 능력을 평가하기 때문이다. ClassTTA의 클래스 조건부 집계는 결정 임계값 주변에 위치한 일부 샘플의 예측 점수를 올바른 방향으로 조정하여 고위험 전립선암의 검출 민감도와 저위험 전립선암의 분류 성능을 개선하였으나, 전체 샘플의 상대적인 예측 순위에는 큰 변화를 발생시키지 않은 것으로 해석된다. 따라서 AUC는 비교적 유사하게 유지된 반면, 임상적으로 사용되는 특정 분류 임계값에서는 민감도와 특이도의 균형이 개선된 것으로 판단된다.

표 3은 Dataset 1에서 테스트 단계에 경미한 변형을 적용한 Mild 조건과 추가적인 강한 변형을 함께 적용한 Mild + Severe 조건의 성능을 비교한 결과를 나타낸다. 전통적인 TTA와 AugTTA 방식은 경미한 변형 조건에서는 일정 수준의 성능을 유지하였으나, 강한 변형

이 적용된 경우 균형 정확도와 특이도가 감소하는 경향을 보였다. 이는 단순 평균 기반 집계 방식이 특정 증강에 의해 왜곡된 예측을 충분히 억제하지 못했기 때문으로 해석된다. AugTTA는 일부 설정에서 강한 변형 조건에서도 성능 향상을 보였으나, 변형 유형에 따라 성능 변동성이 비교적 크게 나타났다. 반면, ClassTTA는 Mild 조건과 Mild + Severe 조건 모두에서 균형 정확도 88%, 정확도 88%, 민감도 93%, 특이도 83%, AUC 0.86으로 동일한 성능을 보였다. 이는 ClassTTA가 추가된 강한 변형을 모두 동일하게 반영하는 것이 아니라, 클래스별 가중치 학습을 통해 불필요하거나 예측을 왜곡하는 강한 변형의 기여도를 억제하고 유의미한 증강 결과만을 선택적으로 반영했기 때문으로 해석할 수 있다. 그 결과 Mild + Severe 조건에서도 최종 분류 결과와 샘플 간 예측 순위가 Mild 조건과 동일하게 유지되어, 소수 둘째 자리까지 제시한 모든 평가 지표가 동일하게 나타났다. 이러한 결과는 ClassTTA가 입력 변형의 강도가 증가하더라도 클래스별로 증강의 기여도를 적응적으로 조절하여 안정적인 예측을 유지할 수 있음을 보여준다.

그림 2는 ClassTTA에서 학습된 클래스 조건부 집계 전략의 특성을 분석하기 위해 수행한 SHAP 기반 해석 결과를 나타낸다. 각 증강 변형을 입력 특징으로 간주하여 SHAP 값을 계산함으로써, 증강별 예측 기여도를 클래스별로 정량적으로 분석하였다. hPCa의 경우, 일부 강한 기하학적 변형 및 특정 강도의 변형 증강이 상위 중요 증강으로 일관되게 나타났으며, 이들 증강은 양의 SHAP 값을 통해 hPCa 예측을 강화하는 방향으로 기여하는 경향을 보였다. 이는 ClassTTA가 병변의 형태적 불규칙성이나 경계 변화가 강조되는 증강을 고위험군 예측에 선택적으로 활용하고 있음을 시사한다. 반면, lPCa에서는 hPCa와는 상이한 증강들이 상위 중요 증강으로 나타났으며, 일부 강한 증강의 경우 SHAP 값이 상대적으로 낮거나 억제되는 경향을 보였다. 이는 저위험군 예측에서 과도한 변형이 오히려 예측에 부정적으로 작용할 수 있음을 ClassTTA가 학습적으로 반영한 결과로 해석할 수 있다. 이러한 결과는 ClassTTA가 모든 증강을 동일하게 활용하는 기존 TTA 방식과 달리, 클래스 조건에 따라 증강의 기여도를 차별적으로 조절하는 집계 전략을 실제로 학습하고 있음을 시각적으로 입증한다. 즉, 제안 방법은 성능 향상뿐만 아니라, 예측 과정에서 어떤 증강이 각 클래스 예측에 유의미하게 작용하는지에 대한 해석 가능성도 함께 제공한다.

Table 3. Performance comparison under different levels of test-time perturbations. “Mild” denotes augmentations used during training, while “Mild + Severe” includes additional strong perturbations uncommon in prostate MR images. ClassTTA maintains stable performance even under severe perturbations, whereas traditional TTA and AugTTA show noticeable performance degradation. The highest values are denoted in bold.

Method	Type of perturbations	Balanced Accuracy	Accuracy	Sensitivity	Specificity	AUC
Baseline + Traditional TTA	Mild	0.71	0.74	0.42	1.00	0.68
	Mild + Severe	0.68	0.70	0.42	0.93	0.68
Baseline + AugTTA	Mild	0.76	0.78	0.58	0.93	0.80
	Mild + Severe	0.82	0.81	0.83	0.80	0.79
Baseline + ClassTTA	Mild	0.88	0.88	0.93	0.83	0.86
	Mild + Severe	0.88	0.88	0.93	0.83	0.86

4. 결론

본 연구에서는 전립선암 악성도 예측을 위한 클래스 조건 테스트 단계 증강 프레임워크를 제안하였다. 제안 방법은 테스트 단계에서 생성된 다수의 증강 예측 결과를 클래스 조건에 따라 차별적으로 집계함으로써, 모든 증강을 동일하게 평균하는 기존 TTA 방식의 한계를 극복하고자 하였다. 실험 결과, ClassTTA는 기존 TTA 및 학습형 TTA 방법 대비 전반적인 분류 성능을 일관되게 향상시켰으며, 특히 임상적으로 중요한 고위험 전립선암 검출에서 민감도를 유의미하게 개선하는 동시에 저위험

전립선암 예측에서 특이도를 안정적으로 유지하였다. 또한 특성이 상이한 데이터셋의 경우와 강한 입력 변형이 적용된 경우에도 안정적인 성능을 보여, 입력 변동성과 데이터 분포 차이에 대한 강건성을 확인하였다. 더 나아가 SHAP 기반 해석 가능성 분석을 통해 ClassTTA가 클래스별로 상이한 증강 활용 전략을 학습함을 확인하였으며, 이는 제안 방법이 성능 향상뿐만 아니라 임상적으로 해석 가능한 방식으로 증강 정보를 활용함을 시사한다.

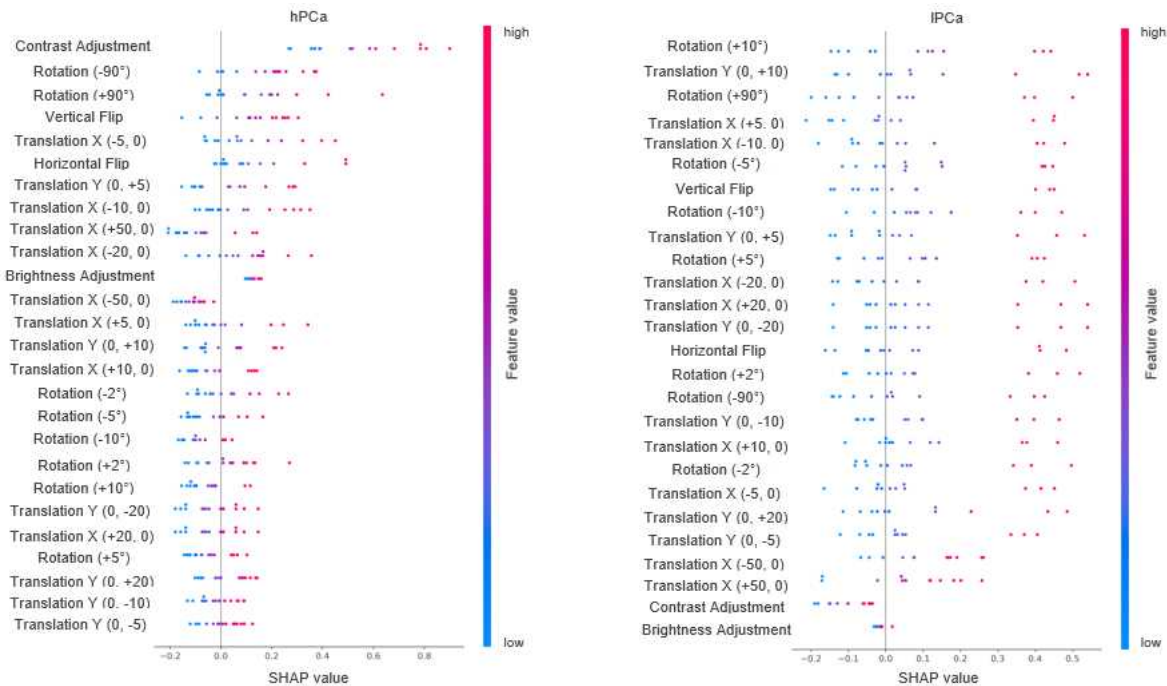


Figure 2. Class-conditional SHAP summary plots for ClassTTA. The left and right panels correspond to high-risk (hPCa) and low-risk (lPCa) prostate cancer, respectively. The plots show that ClassTTA selectively emphasizes different augmentation types depending on the class, highlighting meaningful augmentations for hPCa while suppressing unnecessary contributions for lPCa.

감사의 글

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2023-00207947), 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원(RS-2022-KH129703) 및 서울여자대학교 학술연구비의 지원(2026)을 받아 수행되었습니다.

References

- [1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7-30, 2020.
- [2] J. Cuzick, M. A. Thorat, G. Andriole, O. W. Brawley, W. Chu, A. Coldman, and C. M. Tangen, "Prevention and early detection of prostate cancer," *The Lancet Oncology*, vol. 15, no. 11, pp. e484-e492, 2014.
- [3] J. R. Stark, S. Perner, M. J. Stampfer, J. A. Sinitto, S. Finn, A. S. Eisenstein, and L. A. Mucci, "Gleason score and lethal prostate cancer: Does 3+4 = 4+3?," *Journal of Clinical Oncology*, vol. 27, no. 21, pp. 3459-3464, 2009.
- [4] R. N. Flach et al., "Significant inter- and intralaboratory variation in Gleason grading of prostate cancer: A nationwide study of 35,258 patients in The Netherlands" *Cancers*, vol. 13, no. 21, p. 5378, 2021.
- [5] S. Yoo, "Multiparametric magnetic resonance imaging for prostate cancer," *Korean Journal of Urology*, vol. 56, no. 7, pp. 487-499, 2015.
- [6] P. Wang, W. Chen, and J. H. Lim, "Revisiting test-time augmentation," *IEEE Access*, vol. 8, pp. 214222-214235, 2020.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1106-1114, 2012.
- [8] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1214-1223, 2021.
- [9] M. S. Ayhan and P. Berens, "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Proceedings of the Medical Imaging with Deep Learning (MIDL)*, 2018.
- [10] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 501-512, 2019.
- [11] T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303-312, 2017.
- [12] I. Kim, Y. Kim, and S. Kim, "Learning loss for test-time augmentation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 20519-20529, 2020.
- [13] S. Chun, J. Y. Lee, and J. Kim, "Cyclic test-time augmentation with entropy weight method," in *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, vol. 180, pp. 433-442, 2022.
- [14] J. Son and S. Kang, "Efficient improvement of classification accuracy via selective test-time augmentation," *Information Sciences*, vol. 642, p. 119148, 2023.
- [15] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765-4774, 2017.
- [16] J. Jung, H. Hong, Y. G. Kim, S. I. Hwang, and H. J. Lee, "Prediction of prostate cancer aggressiveness using quantitative radiomic features from multiparametric MRI," *Medical Imaging 2020: Computer-Aided Diagnosis, Proc. SPIE*, vol. 11314, pp. 113143I, 2020.
- [17] Y. Kim, M. Lee, H. Hong, and S. I. Hwang, "Deep regression model with ordinal and triplet loss for the prediction of prostate cancer aggressiveness in multiparametric MR images," *Medical Imaging 2023: Computer-Aided Diagnosis, Proc. SPIE*, vol. 12465, paper 124650P, 2023.

〈 저 자 소 개 〉



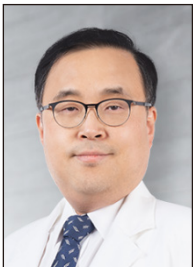
정 주 립

- 2007년 2월 서울여자대학교 멀티미디어학과 졸업(학사)
- 2009년 2월 서울여자대학교 대학원 컴퓨터학과 졸업(석사)
- 2017년 7월 서울여자대학교 대학원 컴퓨터학과 졸업(박사)
- 2016년 9월~현재 서울여자대학교 소프트웨어융합학과 초빙강의교수
- 관심분야 : 인공지능, 딥러닝, 의료영상처리 및 분석
- <https://orcid.org/0000-0001-5882-3956>



이 민 주

- 2026년 2월 서울여자대학교 소프트웨어융합학과 졸업(학사)
- 관심분야 : 인공지능, 딥러닝, 의료영상처리 및 분석
- <https://orcid.org/0009-0007-7423-2997>



황 성 일

- 1996년 2월 서울대학교 의과대학 졸업(학사)
- 2005년 2월 강원대학교 의과대학원 졸업(석사)
- 2009년 8월 서울대학교 의과대학원 졸업
- 2007년~현재 분당서울대병원 영상 의학과 교수(세부 전공 : 비노생식기 영상의학)
- 관심분야 : 전립선 영상, 인공지능 및 조직검사
- <https://orcid.org/0000-0001-7516-5369>



홍 헬 렌

- 1994년 2월 이화여자대학교 전자계산학과 졸업(학사)
- 1996년 2월 이화여자대학교 전자계산학과 졸업(석사)
- 2001년 8월 이화여자대학교 컴퓨터학과 졸업(박사)
- 2001년 9월~2003년 7월 서울대학교 컴퓨터공학부 BK 조교수
- 2006년 3월~현재 서울여자대학교 소프트웨어융합학과 교수
- 관심분야 : 인공지능, 딥러닝, 의료영상처리 및 분석
- <https://orcid.org/0000-0001-5044-7909>