

CNN-GRU와 이중 그래프 컨볼루션 네트워크를 활용한 lncRNA-질병 연관성 예측 연구

김상욱*¹, 윤태현*², 류지열**³, 하지환***⁴

CNN-GRU and Heterogeneous Graph Convolutional Networks for lncRNA - Disease Association Prediction

Sanguk Kim*¹, Taehyeon Yun*², Jee-Youl Ryu**³, and Jihwan Ha***⁴

본 과제(결과물)는 2025년도 교육부 및 부산시의 재원으로 부산RISE혁신원의 지원을 받아 수행된 지역혁신중심
대학지원체계(RISE)의 결과입니다(2025-RISE-02-001-002)

요 약

LncRNA(Long non-coding RNA)는 암을 비롯한 다양한 인간 질환의 발생과 진행 과정에서 핵심적인 역할을 수행하므로, lncRNA와 질환 간의 새로운 연관성을 밝히는 것은 의학적·생물학적 측면에서 중요한 의미를 가진다. 기존 계산 모델들은 주로 네트워크 위상 정보에 의존하여, lncRNA가 지닌 고유한 생물학적 특성을 충분히 반영하지 못하는 한계를 보인다. 이를 해결하기 위해, 본 연구에서는 lncRNA 염기서열 특징과 이중 정보 네트워크 구조를 통합 학습하는 CG-HGCN 모델을 제안한다. CG-HGCN은 CNN과 GRU를 결합하여 염기서열로부터 특징을 추출하고, 이를 lncRNA, miRNA, 질환으로 구성된 HG-GCN의 초기 노드 특징으로 활용한다. 이후 네트워크 상의 복잡한 상호작용을 학습하여 최종적으로 lncRNA-질환 연관 확률을 예측한다. 실험 결과, 서열 정보와 이중 네트워크 정보를 동시에 고려함으로써 연관성을 효과적으로 예측할 수 있음을 확인하였다.

Abstract

Long non-coding RNAs (lncRNAs) play a critical role in the onset and progression of various human diseases, including cancer. Therefore, identifying novel associations between lncRNAs and diseases is of great biological and clinical significance. However, existing computational models mainly rely on network topological information and often fail to adequately capture the intrinsic biological characteristics of lncRNAs. To address this limitation, we propose a novel model, CG-HGCN, which integrates lncRNA sequence features with heterogeneous information network structures. Specifically, CG-HGCN employs a combination of CNN and GRU to extract informative features from lncRNA sequences, which are then used as initial node representations in an HG-GCN composed of lncRNAs, miRNAs, and diseases. By learning complex interactions within the network, the model predicts the association probabilities between lncRNAs and diseases. Experimental results demonstrate that incorporating both sequence information and heterogeneous network data enables CG-HGCN to effectively predict lncRNA - disease associations.

Keywords

long non-coding RNA, lncRNA, disease association prediction, sequence data, heterogeneous graph network, CNN-GRU

* 부경대학교 데이터공학과

- ORCID¹: <https://orcid.org/0009-0007-3765-2212>

- ORCID²: <https://orcid.org/0009-0009-3548-7430>

** 부경대학교 정보통신학과 교수

- ORCID: <https://orcid.org/0000-0003-2343-7056>

*** 부경대학교 빅데이터융합전공 교수(교신저자)

- ORCID: <https://orcid.org/0000-0002-6086-5693>

· Received: Aug. 28, 2025, Revised: Sep. 26, 2025, Accepted: Sep. 29, 2025

· Corresponding Author: Jihwan Ha

Major of Big Data Convergence, Division of Data Information Science,
Pukyong National University, Busan 48513, Korea

Tel.: +82-51-629-4614, Email: jhha@pknu.ac.kr

I. 서 론

lncRNA(Long non-coding RNA)는 단백질을 암호화하지 않는 200 뉴클레오타이드(Nucleotide) 이상의 RNA 분자로, 유전자 발현 조절, 세포 분화 및 발달 등 다양한 생명 현상에서 핵심적인 조절자(Key regulator)로 기능한다[1][2]. 최근 수많은 연구를 통해 이러한 lncRNA의 비정상적인 발현이 암, 심혈관 질환, 신경 퇴행성 질환 등 여러 인간 질병의 발병 및 진행과 밀접하게 연관되어 있음이 밝혀졌다[3][4]. 따라서, 특정 lncRNA와 질병 간의 연관성(LDA, lncRNA-Disease Association)을 규명하는 것은, 질병의 분자적 메커니즘(Molecular mechanism)을 깊이 이해하고 새로운 진단용 바이오마커(Biomarker)나 혁신적인 치료 표적(Therapeutic target)을 발굴하는 데 매우 중요하다.

전통적인 생물학적 실험은 lncRNA와 질병 간의 인과관계를 증명하는 가장 확실한 방법(Gold standard)이지만, 막대한 시간과 비용이 소요되어 수많은 잠재적 연관성 후보군을 대규모로 탐색하기에는 현실적인 한계가 있다. 이러한 실험적 한계는 방대한 생물학적 데이터로부터 잠재적인 LDA를 효율적으로 예측하고 실험 검증의 우선순위를 정해주는 계산 모델의 개발 필요성을 제기한다.

이러한 필요성에 따라, 본 연구에서는 lncRNA와 질병간의 잠재적 연관성을 예측하기 위해 lncRNA의 염기서열 정보를 lncRNA, miRNA, 질병 간의 네트워크(Network) 구조 정보에 통합하여 예측하는 CG-HGCN 모델(Model)를 제안한다. CG-HGCN은 lncRNA, miRNA, 질병을 노드(Node)로 하고 이들 간의 상호작용을 엣지(Edge)로 하는 이중 정보 네트워크(Heterogeneous information network)를 그래프 형태로 구축하고, CNN-GRU로 lncRNA 염기서열의 특징 벡터(Feature vector)를 추출하여 그래프의 lncRNA 노드 초기 특징으로 사용한다. 이렇게 구축된 그래프 구조 정보를 HG-GCN(Heterogeneous Graph convolutional Network)이 학습하여 모든 노드의 최종 임베딩(Embedding)을 생성한다. 마지막으로 연관성 예측 분류기(Classifier)가 lncRNA와 질병 노드 쌍의 최종 임베딩을 기반으로 연관 확률을 예측한다.

II. 관련 연구

2.1 네트워크 기반 접근법

네트워크 기반 접근법은 lncRNA, 질병, miRNA 등으로 구성된 생물학적 네트워크의 위상(Topology) 정보를 활용하여 알려지지 않은 연관성을 추론한다. 특히, RWR(Random Walk with Restart) 기반 모델이 주로 사용되며, RWRHLD, BiWalkLDA, MHRWR 등의 연구가 보고되었다[5]-[9]. 이러한 모델들은 이중 네트워크 내에서 정보 흐름을 시뮬레이션(Simulation)하여 노드 간 근접성을 점수화한다. 그러나 초기 RWR 모델은 알려진 연관 관계를 시작점으로 활용하기 때문에, 연관된 lncRNA 정보가 없는 새로운 질병에 대해서는 예측이 어려운 명확한 한계를 가진다. 이를 극복하기 위해, IRWRLDA[6]는 lncRNA 발현 유사성과 질병 의미 유사성을 통합하여 초기 확률을 설정하였으며, GrwLDA[7]는 각 노드 타입(Node type)에서 개별적으로 RWR을 수행한 후 결과를 통합함으로써 고립된 노드 문제에 대응하고자 하였다. 이러한 접근법은 다중 소스 데이터를 효과적으로 통합하여 노드 표현을 강화한다는 장점을 가지지만, 여전히 신뢰도 높은 네트워크 구축을 위해 알려진 연관성에 의존해야 한다는 단점을 가진다.

2.2 머신러닝 및 행렬 분해 기반 방법

전통적인 머신러닝 기반 방법들은 LDA 예측을 이진 분류(Binary classification)문제로 정의한다. 알려진 연관성은 긍정 샘플(Positive sample)로, 알려지지 않은 관계는 부정 샘플(Negative sample)로 간주하여 lncRNA와 질병의 특징을 추출하고 둘을 구별하는 분류기를 학습시킨다. LRLSLDA[10]는 정규화된 최소 제곱법을 적용한 초기 연구이며, 이후 SVM(Support vector machine)을 사용하는 LDAP[11], Random Forest를 사용하는 RFLDA[12], IPCARF[13], Naive Bayesian을 사용하는 CFNBC[14] 등 다양한 분류 모델들이 제안되었다. 이러한 접근법은 다양한 데이터 소스로부터 계산된 lncRNA와 질병의 유사

성 점수 등 다차원의 특징(Multi-view features)을 유연하게 통합하여 학습 모델을 구축할 수 있다는 장점을 가진다.

행렬 분해 기반 방법들은 LDA 예측을 추천 시스템의 누락된 값 예측 문제로 간주한다. lncRNA-질병 연관 행렬을 저차원(Low-rank) 행렬로 근사화하여 잠재 특징(Latent features)을 추출하고 이를 통해 연관성을 예측한다. MFLDA[15]는 행렬 삼중 분해를, PMFLDA[16]는 확률적 행렬 분해를, SIMCLDA[17]는 유도 행렬 완성을 사용하는 등 다양한 기법이 연구되었다. 이러한 방법들은 부정 샘플 없이도 학습이 가능하며, 다중 소스 데이터를 유연하게 통합하는 장점이 있다. 하지만, 두 방법 모두 얕은(Shallow) 모델에 기반하고 있어, 데이터의 비선형적 특징을 학습하는 데에는 한계가 있다.

2.3 딥러닝 및 그래프 신경망 기반 방법

최근 딥러닝을 포함한 인공지능 모델들이 다양한 분야에서 우수한 성능을 보임에 따라, 질병 관련 lncRNA 탐색 연구뿐만 아니라 다양한 생물정보학 연구 분야에서도 이러한 모델들이 핵심 기술로 활용되고 있다[18]-[32]. 특히, 기존의 얕은(Shallow) 모델들이 지닌 비선형적 특징 학습의 한계를 극복하기 위해, 복잡한 데이터 표현을 효과적으로 포착할 수 있는 딥러닝 기반 방법들이 활발히 도입되고 있다. 초기에는 FCN(Fully Connected Network)를 사용한 NNLDA[33], DMFLDA[34], CNN(Convolutional Neural Networks)을 사용한 CNNLDA[35], LDAPred[36], 오토인코더(Auto-encoder)를 사용한 CNNDLP[37], VADLP[38], GAN(Generative Adversarial Networks)을 사용한 BiGAN[39] 등 다양한 딥러닝 아키텍처를 활용하여 노드의 비선형 특징을 학습하려는 시도가 주를 이루었다.

최근에는 lncRNA-질병 네트워크의 위상 데이터를 가장 효과적으로 활용할 수 있는 GNN(Graph Neural Network) 기반의 연구가 각광받고 있다. GCN(Graph Convolutional Network)를 사용한 GCNLDA[40], HGC-GAN[41], GAT(Graph Attention Network)를 사용한 GANLDA[42], HGATLDA[43],

GAE(Graph Auto-encoder)를 사용한 PANDA[44], VGAE[45], GAMCLDA[46] 등 다양한 GNN 모델들이 제안되었으며, 이러한 방법들은 노드 표현의 품질을 향상시켜 다른 방법들보다 전반적으로 뛰어난 성능을 보였다.

딥러닝과 GNN 기반 방법들은 우수한 예측 성능을 달성했지만, 여전히 해결해야 할 과제가 남아있다. 다수의 훈련 파라미터로 인해 과적합 문제가 발생하기 쉽고, GNN의 경우 층이 깊어질수록 과도한 평활화(Over-smoothing) 문제가 발생할 수 있다. 또한, 대부분의 GNN 모델은 네트워크 구조에 의존하기 때문에, 연결이 적은 노드에 대한 예측 성능이 저하되고 노드 자체의 특징을 직접적으로 활용하는데 한계가 있다.

앞선 한계점들을 해결하기 위해, 본 연구에서 제안한 CG-HGCN 모델은 CNN-GRU, HG-GCN와 분류기로 구성된 End-to-End 접근방법으로 깊은 아키텍처(Deep architecture)를 통해 기존의 얕은 학습 방법으로 포착하기 어려웠던 비선형적인 관계를 효과적으로 학습하고, 그래프 구조 자체가 실제 생물학적 네트워크 안에서 정보가 흐르도록 학습 방향을 제한하여 일종의 정규화(Regularization) 역할을 하여 모델의 과적합을 완화한다. 또한, 초기 lncRNA 노드 특징으로 염기서열로부터 추출한 특징 벡터를 사용함으로써, 노드 자체의 내재된 정보가 부족했던 기존 네트워크 분석의 초기 특징 부재 및 고립된 노드 문제를 극복하여 개선된 연관성 예측을 기대할 수 있다.

III. 데이터셋 및 전처리

3.1 데이터 베이스

본 연구에서 lncRNA와 질병 간의 잠재적 연관성을 예측하는 모델을 개발하기 위해, 표1에 정리한 공개 생물학 데이터베이스(Database)들로부터 데이터를 수집하여 통합 데이터셋(Dataset)을 구축하였다. 이 데이터셋은 lncRNA, miRNA, 질병이라는 세 가지 주요 개체(Entity)와 이들 간의 상호작용으로 구성된 이중 정보 네트워크의 기반이 된다.

모델 학습의 핵심이 되는 lncRNA와 질병 간의

관계는 두 개의 데이터베이스에서 확보하여 보완적으로 사용하였다. 첫 번째는 Lnc2Cancer v3.0[47]으로 암과 lncRNA의 관계에 특화된 데이터베이스이다. 15,000개 이상의 논문을 검토하여 실험적으로 검증된 정보를 제공하며, 2,659개의 lncRNA와 216개의 암 하위 유형에 대한 9,254개의 lncRNA와 암 연관성 데이터를 포함하고 있다. 본 연구에서는 암 관련 연관성에 대한 긍정 샘플을 확보하는데 사용되었다. 두 번째는 lncRNADisease v2.0[48]으로 암을 비롯한 다양한 인간 질병과 lncRNA의 관계를 포괄하는 데이터베이스이다. 약 12,000개의 논문을 검토하여 19,166개의 lncRNA와 529개의 질병에 대해 실험적으로 검증된 연관성 약 10,000개를 포함하고 있다. 모델이 특정 질병군에 과적합되지 않고 범용적인 예측 성능을 갖도록 데이터의 다양성을 확장하는 역할을 한다.

lncRNA가 질병에 영향을 미치는 간접적인 분자 메커니즘을 네트워크에 반영하기 위해 두 가지 데이터베이스에서 얻은 데이터를 활용하였다. 첫 번째 starBase v2.0[49]은 CLIP-Seq 실험 기법으로 검증된 RNA 분자 간 상호작용 정보를 제공하는 데이터베이스로, 실험적으로 검증된 miRNA와 lncRNA의 연관성을 약 10,000개를 제공한다. 본 연구에서는 이중 네트워크 내에서 lncRNA 노드와 miRNA 노드를 연결하는 엣지를 구성하는 데 사용하였다. 두 번째 HMDD v3.2은 실험적으로 검증된 인간 miRNA와 disease 연관성을 수집한 데이터베이스이다. HMDD v3.0[50] 기준으로 17,412편의 문헌에서 수집된 32,281개의 실험적으로 검증된 miRNA와 disease의 연관성을 포함하며, 총 1,102개의 miRNA와 850개의 질병을 제공한다. 네트워크에서 miRNA 노드와 질병 노드를 잇는 엣지를 구축하는 데 사용되었다. 위 네 가지 데이터베이스를 이용해 lncRNA-miRNA-질병으로 구성된 이중 정보 네트워크를 구축할 수 있었다.

본 연구에서는 기존 네트워크 기반 연구의 한계를 극복하기 위해 lncRNA 생물학적 특징을 모델에 활용하였다. 따라서, lncRNA의 생물학적 특징 데이터를 수집하기 위해 NCBI[51] 데이터베이스를 사용하였다. NCBI는 미국 NIH(National Institutes of Health) 산하의 생물정보학 센터로, 분자생물학과 유전체학 관련 방대한 공개 데이터와 검색 도구를 제

공한다. 이러한 NCBI로부터 1,363개의 lncRNA에 대한 서열 데이터를 수집하였다. 이 서열 정보는 본 연구에서 제안하는 모델의 CNN-GRU에 입력되어 각 lncRNA 노드의 고유한 초기 특징 벡터(Initial feature vector)를 생성하는 역할을 한다.

표 1. 연구에 사용된 데이터베이스 메타데이터 요약
Table 1. Summary of database metadata used in the study

Database	Content	Statistics
Lnc2Cancer v3.0	Cancer-specific lncRNA-disease associations	2,659 lncRNAs, 216 cancer types, 9,254 associations
lncRNADisease v2.0	Comprehensive lncRNA-disease associations	19,166 lncRNAs, 529 diseases, ~10,000 associations
starBase v2.0	CLIP-Seq based RNA interactions	~10,000 miRNA-lncRNA associations
HMDD v3.2	Experimentally validated miRNA-disease associations	(Based on v3.0) 1,102 miRNAs, 850 diseases, 32,281 associations
NCBI	Genomic and molecular biology data	Obtained sequence data for 1,363 lncRNAs

3.2 데이터 전처리

여러 데이터베이스에서 수집한 원본 데이터(Raw data)는 모델 학습에 사용하기 위해, 이를 정제하고 통합하여 모델 입력에 적합한 형태로 변환하는 전처리 과정을 수행하였다. 전처리 과정은 크게 데이터 통합 및 정규화, 네트워크 구성, 최종 데이터 구조화, 그리고 모델 학습 데이터셋 생성의 네 단계로 구성된다.

첫 번째 데이터 통합 및 정규화 과정에서는 서로 다른 출처의 데이터를 일관성 있게 통합하기 위해 모든 데이터셋에 포함된 lncRNA, miRNA, 질병의 이름을 소문자로 변환하고 양쪽의 공백을 제거하는 이름 정규화를 수행하였다. 이후, lncRNA-질병 연관성 데이터인 Lnc2Cancer v3.0과 lncRNADisease v2.0는 하나로 통합한 후 중복된 항목을 제거하여 단일화된 연관성 목록을 생성하였다. lncRNA 염기서열은 CNN-GRU의 입력에 적합하도록 전처리를 거쳤

다. 우선, 모든 염기서열의 길이를 2,000으로 표준화하였다. 이는 RNACentral[52] 데이터베이스의 lncRNA 길이 분포 분석에 근거한 결정으로, 전체 서열의 75%가 1,767 뉴클레오타이드 이하인 반면 최대 길이는 346,385에 달하는 극단적인 분포를 보였다. 만약 최대 길이에 맞춰 패딩(Padding)할 경우, 대부분의 짧은 서열에서 정보 희석(Information dilution)문제가 발생하고 계산 비효율성(Computational inefficiency)이 커지게 된다. 따라서 대다수 서열의 정보를 온전히 보존하면서 계산 효율성을 확보할 수 있는 2,000을 기준 길이로 설정하였다. 이 기준에 따라 2,000보다 긴 서열은 잘라내고 짧은 서열은 특수 패딩 문자를 추가하여 길이를 통일하였으며, 최종적으로 'A', 'T', 'G', 'C'와 패딩 문자를 각각 정수 0, 1, 2, 3, 4로 변환하여 모델의 입력 형식에 맞추었다.

두 번째 네트워크 구성 단계에서는 모델의 핵심인 이중 정보 네트워크를 효과적으로 구성하기 위해 다음과 같은 단계의 필터링을 적용하였다. 우선, 염기서열 정보가 없는 lncRNA는 특징 벡터를 생성할 수 없으므로, NCBI에서 염기서열을 확보할 수 있는 lncRNA와 관련된 모든 연관 데이터만 일차적으로 필터링하였다. 이후, 염기서열이 있는 lncRNA를 중심으로 연결 그래프를 구성하기 위해, 1차 필터링을 통과한 lncRNA-질병 및 lncRNA-miRNA 관계에 포함된 lncRNA, miRNA, 질병들만 초기 유효 노드 집합으로 정의하였다. 다음으로, miRNA-질병 연관성 데이터 중 최소 한쪽의 노드가 이 유효 노드 집합에 포함된 경우 연결된 관계의 노드 또한 유효 노드 집합에 추가하여 네트워크의 연결성을 확장하였다. 마지막으로, 확정된 최종 노드 목록에 포함되지 않는 개체와의 관계는 모든 데이터셋에서 제거하였다. 이 과정을 통해 모든 노드가 네트워크 내에서 최소 하나 이상의 연결을 갖도록 보장하여, GCN 학습 과정에서 정보가 전파될 수 없는 고립 노드(Isolated node)의 생성을 방지하였다. 최종적으로 구성된 이중 네트워크는 1,392개의 lncRNA 노드, 958개의 miRNA 노드, 그리고 728개의 질병 노드를 포함하며, 노드 간의 엣지는 5,313개의 lncRNA-질병 연관성, 31,054개의 miRNA-질병 연관성 그리고

2,530개의 lncRNA-miRNA 상호작용으로 구성된다.

세 번째 데이터 구조화 단계에서는 모델 입력을 위해 전처리가 완료된 lncRNA, miRNA, 질병의 최종 목록에 대해 문자열 이름 대신 각각 고유한 정수 인덱스(Index)를 부여하였다. 또한, lncRNA 염기서열 데이터는 lncRNA 인덱스 순서에 맞게 정렬된 리스트로 변환하여, i 번째 서열이 인덱스 i 를 가진 lncRNA에 해당하도록 구조화하였다.

네 번째 모델 학습 데이터셋 생성 단계에서는 모델의 학습을 위해 알려진 연관성, 즉 긍정 샘플과 동일한 수의 부정 샘플을 생성하였다. 이를 위해 최종 유효 노드 집합에 포함된 lncRNA와 질병을 무작위로 조합하여 새로운 쌍을 만들었다. 이때, 생성된 쌍이 기존 긍정 샘플 목록에 존재하지 않는, 즉 아직 알려지지 않은 관계임을 확인하는 과정을 거쳤다. 이러한 방식을 통해 긍정 샘플과 부정 샘플이 1:1 비율을 이루는 균형 잡힌(Balanced) 학습 데이터셋을 최종적으로 구축하였다.

IV. 제안 모델: CG-HGCN

본 연구에서는 lncRNA 염기서열의 특징과 lncRNA, miRNA, 질병 간의 복잡한 상호작용 네트워크를 통합적으로 학습하여 lncRNA-질병 연관성을 예측하는 그림 1과 같은 End-to-End 딥러닝 모델을 제안한다. 제안 모델은 CNN-GRU, HG-GCN, 그리고 연관성 예측 분류기의 세 가지 주요 모듈(Module)로 구성된다.

모델의 전체적인 흐름은 다음과 같다. 먼저, CNN-GRU이 전처리된 lncRNA 염기서열을 입력받아 각 lncRNA에 대한 256차원의 초기 특징 벡터를 생성한다. 이 특징 벡터는 이중 네트워크 그래프에서 lncRNA 노드의 초기값으로 설정된다. 다음으로, HG-GCN이 이중 네트워크 그래프를 입력받아 GCN 연산을 통해 이웃 노드의 정보를 종합한 256차원의 최종 임베딩을 학습한다. 마지막으로, 연관성 예측 분류기가 예측하고자 하는 특정 lncRNA와 질병 노드의 최종 임베딩을 입력받아 둘 사이의 연관 확률을 계산한다.

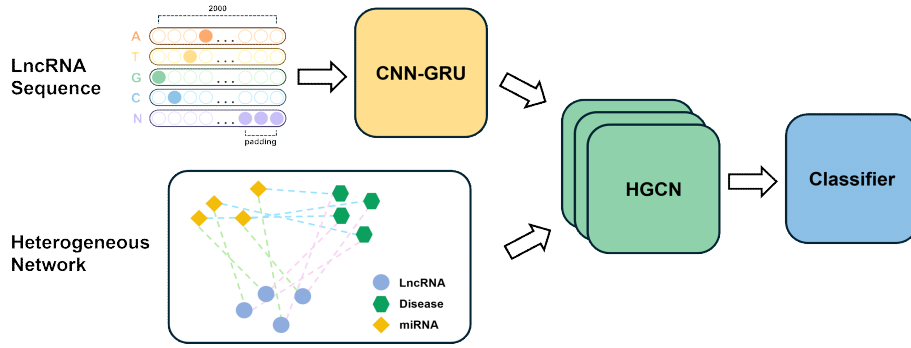


그림 1. CG-HGCN 모델 구조
Fig. 1. CG-HGCN model architecture

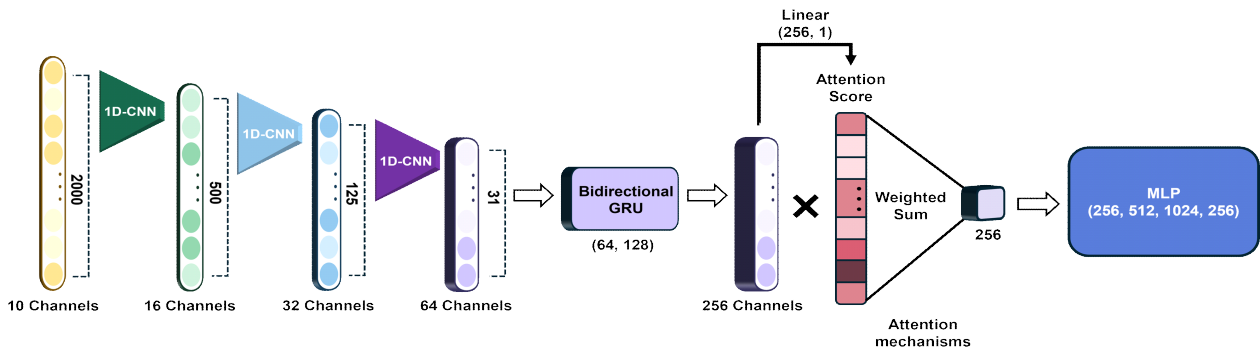


그림 2. CNN-GRU 모듈 구조
Fig. 2. CNN-GRU module architecture

4.1 CNN-GRU

이 모듈은 생물학적 염기서열이라는 원본 데이터를 특징 벡터로 변환하는 특징화(Feature engineering) 역할을 수행한다. 이를 통해, lncRNA의 기능과 관련된 핵심적인 서열 패턴을 추출하여 lncRNA 노드의 생물학적 특징을 정의하는 초기 정보를 제공한다.

모듈의 구조는 그림 2와 같이 3개의 1D-CNN 층, 양방향 GRU(Bidirectional Gated Recurrent Unit), 어텐션 메커니즘(Attention mechanism)을 순차적으로 결합한 구조를 가진다. 입력 데이터를 통해 살펴보면, 정수 인코딩된 염기서열은 먼저 Embedding 레이어를 통과하여 10차원의 밀집 벡터(Dense vector)로 변환된다. 그다음, 이 벡터는 3개의 1D-CNN 층을 통과하며 국소적인 서열 모티프(Motif)를 감지한다. 임베딩된 서열 벡터는 3개의 1D-CNN 층을 통과하며, 각 층 이후에는 ReLU 활성화 함수가 적용된다. 각 1D-CNN 층은 커널 크기와 스트라이드를

4로 설정하였고, 특징 맵(Feature map)의 채널을 16, 32, 64 순으로 확장시킨다. 따라서, 서열 벡터의 길이는 31, 특징 열은 64차원이 된다. 이어서, 1D-CNN을 통해 추출된 벡터는 서열의 순서와 장거리 의존 관계를 학습하기 위해 양방향 GRU 층을 통과한다. 은닉 상태 크기(Hidden state) 128의 양방향 GRU로 벡터의 특징 열은 256차원의 순차적 특징으로 변환된다. 이후, 양방향 GRU의 출력 서열 벡터에 어텐션 메커니즘을 적용하여 서열의 여러 위치 중 중요한 부분에 가중치를 부여하고 가중합(Weighted sum)을 계산하여 256차원의 벡터를 생성한다. 마지막으로, 어텐션 메커니즘을 통해 얻은 벡터는 3개의 완전 연결 레이어(Fully connected layer)를 통과하며, 특징 차원을 512, 1024로 늘렸다가 최종적으로 lncRNA를 대표하는 256차원의 특징 벡터를 출력한다. 각 레이어 사이에는 ReLU 활성화 함수가 사용된다.

4.2 HG-GCN

HG-GCN은 lncRNA 노드의 초기 특징을 네트워크 전체로 전파하고 다른 노드의 정보를 집계하여, 개별 노드의 특징을 생물학적 관계 네트워크 안에서 역할과 위치를 반영한 정교한 임베딩으로 발전시키는 역할을 한다. 이를 위해, 위해 다층(Multi-layer) 이중 그래프 컨볼루션 네트워크(Heterogeneous graph Convolutional network)를 설계하였다. 각 층은 관계형 그래프 컨볼루션 네트워크(Relational Graph Convolutional Network, R-GCN)[53]에 기반하며, 다양한 종류의 노드와 엣지를 효과적으로 모델링(Modeling)한다.

모듈은 총 3개의 HG-GCN 층으로 구성되며, l 번째 층에서 특정 노드 v 의 특징 벡터 $h_v^{(l)}$ 은 연결된 이웃 노드들의 정보를 정규화하여 취합하는 과정을 통해 다음 층의 특징 벡터인 $h_v^{(l+1)}$ 로 업데이트(Update)된다. 이 과정은 다음과 같은 식 (1)로 표현되며, 그림 3과 같이 나타낼 수 있다.

$$h_v^{(l+1)} = \sigma \left(\sum_{r \in R_u \in N_{r(v)}} \frac{1}{c_{v,r}} W_r^{(l)} h_u^{(l)} \right) \quad (1)$$

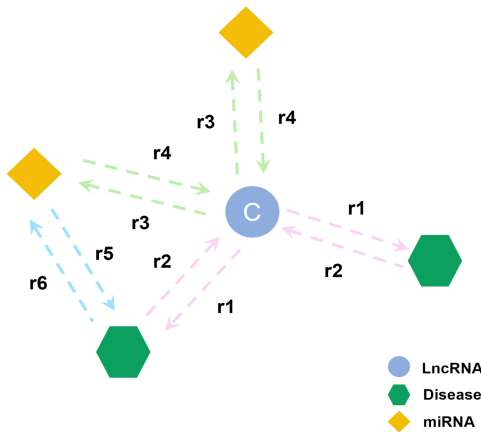


그림 3. HG-GCN의 정보 전파 과정
Fig. 3. HG-GCN message passing process

각 기호의 의미는 다음과 같다. $h_v^{(l)}$ 은 노드 v 의 l 번째 레이어의 은닉 상태 벡터이다. $h_v^{(0)}$ 은 노드의 초기 입력 특징을 의미한다. σ 는 비선형 활성화 함수(Non-linear activation function)로 이 모듈에서는 ReLU를 사용하였다. R 은 그래프에 존재하는 모든 관계(Relation) 타입의 집합이다. $N_{r(v)}$ 는 관계 r 을 통해 노드 v 와 연결된 이웃 노드들의 집합을 나타

낸다. $W_r^{(l)}$ 은 관계 r 에 해당하는 l 번째 레이어의 학습 가능한 변환 행렬(Transformation matrix)이다. 각 관계 타입과 층별로 독립적인 행렬을 사용하여 관계별 특징을 학습한다. $c_{v,r}$ 는 관계 r 에 대한 노드 v 의 정규화 상수(Normalization constant)로, $|N_{r(v)}|$ 를 사용하였다. 이는 특정 관계에 대한 노드의 차수(Degree)를 의미하며, 학습 안정성을 높이는 역할을 한다. 그림 3에서 볼 수 있듯이, 업데이트 대상인 중심 노드는 자신과 여러 종류의 관계로 연결된 이웃 노드들로부터 정보를 전달받는다. 이때 각 이웃 노드의 특징 벡터는 해당 관계 r 에 특화된 변환 행렬 W_r 을 통과하며 고유하게 변환된다. 이는 각 관계가 갖는 생물학적 의미의 차이를 모델이 학습할 수 있도록 한다. 변환된 정보들은 중심 노드로 모여 합산(Aggregation)되고, 비선형 활성화 함수 σ 를 거쳐 중심 노드의 다음 계층 은닉 상태 벡터 $h_v^{(l+1)}$ 로 최종 업데이트된다.

모듈의 입력으로는 앞서 구성한 이중 정보 네트워크 그래프와 lncRNA 노드의 초기 특징 벡터가 사용된다. 이때 lncRNA 노드의 초기 특징은 염기서열 특징 추출기 모델이 추출한 256차원 임베딩을 사용하며, miRNA와 질병 노드의 특징은 256차원의 무작위 벡터(Random vector)로 초기화한다. 입력된 노드 특징 $h_v^{(0)}$ 은 식 (1)에 따라 3개의 층을 순차적으로 통과하며 점진적으로 정제된다. 각 층에서는 여러 관계로부터 계산된 이웃 노드의 정보가 합산(Sum) 방식으로 통합되며, 최종적으로 네트워크의 구조적 정보와 주변 노드의 특징이 반영된 256차원의 노드 임베딩이 생성한다.

4.3 연관성 예측 분류기

연관성 예측 분류 모듈은 최종적으로 의사결정(Decision making)을 수행하는 부분이다. HG-GCN을 통해 생성된 lncRNA와 질병의 특징 벡터를 입력받아, 두 벡터 간의 복잡한 상호작용을 학습하고 이들이 연관될 확률을 최종적으로 계산한다.

모듈의 구조는 4개의 완전 연결 레이어로 구성된 MLP(Multi-Layer Perceptron)이다. 입력으로는 예측하고자 하는 lncRNA 노드의 256차원 최종 임베딩과

질병 노드의 256차원 최종 임베딩을 추출한 후, 두 임베딩 벡터를 연결(Concatenate)한 512차원의 벡터를 사용한다. 이 연결된 벡터는 4개의 완전 연결 레이어를 순차적으로 통과한다. 출력 차원은 256, 128, 64, 1로 작아지며, 각 층 사이에는 ReLU 활성화 함수가 적용된다. 최종 출력값은 시그모이드(Sigmoid) 활성화 함수를 거쳐 0과 1 사이의 확률값으로 변환되며, 이 값이 lncRNA와 질병 간 연관성의 예측 점수로 사용된다. 예측 점수가 1에 가까울수록 해당 lncRNA가 질병과 연관될 가능성이 높음을 의미하며, 이를 통해 각 lncRNA-질병 쌍의 연관성을 정량적으로 평가할 수 있다.

4.4 모델 학습

제안된 전체 모델은 End-to-End 방식으로 한 번에 학습된다. 따라서, 분류기의 예측 확률과 실제 레이블 간의 차이를 최소화하도록 모든 파라미터(Parameters)가 동시에 업데이트되며 구체적인 학습 파라미터는 다음과 같다. 손실 함수(Loss function)는 이진 분류 문제에 표준적으로 사용되는 이진 교차 엔트로피(BCE, Binary Cross-Entropy) 손실 함수를 사용한다. 최적화 알고리즘(Optimizer)으로는 AdamW를 사용하였다. 이때, 학습률(Learning rate)은 0.0001, 가중치 감쇠(Weight decay)는 0.01로 설정하였다. 학습 단위(Batch size)는 32로 모델은 한 번에 32개의 lncRNA-질병 쌍을 처리하여 파라미터를 업데이트한다. 마지막으로 과적합 방지를 위해 각 fold의 학습 과정에서 검증 손실(Validation loss)이 더 이상 개선되지 않으면 학습을 조기 종료하는 조기 종료(Early stopping) 기법을 적용하여 과적합을 방지하였다.

V. 실험 및 결과

5.1 모델의 성능 검증

제안된 모델의 예측 성능을 객관적으로 평가하기 위해 10-겹 교차 검증(10-fold cross-validation)을 수행하였다. 전체 데이터셋을 10개의 상호 배타적인 폴

드(Fold)로 분할하고, 각 폴드를 한 번씩 검증 세트로 사용하며 나머지 9개의 폴드를 훈련 세트로 활용하여 학습과 검증을 10회 반복하였다. 이 과정에서 각 샘플은 정확히 한 번 검증 세트에 포함되도록 구성되어, 데이터 분할에 따른 편향을 최소화하고 모델의 일반화 성능을 신뢰성 있게 평가할 수 있다.

모델의 성능은 두 가지 지표를 사용하여 평가하였다. 하나는 AUC(Area Under the ROC Curve)로 수신자 조작 특성(Receiver Operating Characteristic, ROC) 곡선 아래의 면적으로 모델이 긍정 샘플과 부정 샘플을 얼마나 잘 구별하는지를 나타내는 지표이다. 이때 ROC 곡선은 분류 결정 임계값(Classification threshold)의 변화에 따른 진양성 비율(True Positive Rate, TPR)과 위양성 비율(False Positive Rate, FPR)의 관계를 나타낸다. 하나는 AUPR (Area Under the PR Curve)로 정밀도-재현율(Precision-Recall, PR) 곡선 아래의 면적이다. 이때 PR 곡선은 재현율(Recall) 변화에 따른 정밀도(Precision)의 변화를 보여준다. AUPR은 특히 양성 샘플 수가 음성 샘플에 비해 매우 적은 불균형 데이터셋에서 모델 성능을 평가하는 데 유용하다. lncRNA-질병 연관성 예측과 같은 생물정보학 문제에서는 실제 연관성을 놓치지 않으면서, 예측한 연관성이 실제로 정확할 확률을 함께 고려하는 것이 중요하므로, AUPR은 AUC를 보완하는 핵심적인 성능 지표로 사용된다. 진양성 비율, 재현율, 위양성 비율 및 정밀도의 정의는 식 (2)~(4)와 같다.

$$TPR = Recall = \frac{TP}{TP + FN} \quad (2)$$

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

실험 결과, 10-겹 교차 검증을 통해 각 폴드에서 측정된 AUC 및 AUPR 점수는 표 1에 제시하였으며, ROC 곡선과 PR 곡선 그래프는 각각 그림 4과 그림 5에 나타내었다.

표 1. 10-겹 교차 검증 성능 결과

Table 1. 10-Fold cross-validation performance results

Fold	1	2	3	4	5	6	7	8	9	10	mean
AUC	0.9401	0.9474	0.9245	0.9416	0.9362	0.9527	0.9378	0.9444	0.9366	0.9195	0.9381 $\pm$ 0.0095
AUPR	0.9531	0.9531	0.9323	0.9451	0.9366	0.9458	0.9458	0.9499	0.9389	0.9288	0.9435 $\pm$ 0.0086

제안된 모델은 10회 교차 검증에서 평균 AUC 0.9381, 평균 AUPR 0.9435로 높은 예측 성능을 달성하였다. 또한, 각 폴드별 성능 편차는 AUC의 경우 표준편차 0.0095, AUPR의 경우 0.0086으로 매우 낮게 나타나 모델의 안정성을 확인할 수 있었다.

평균 AUC와 AUPR 점수가 모두 0.93 이상으로 높게 측정된 것은 본 연구에서 제안한 모델이 lncRNA와 질병 간의 연관성을 매우 정확하게 예측함을 시사한다. 이는 lncRNA 염기서열의 특징과 이 중 네트워크의 구조적 정보를 통합하는 접근법이 효과적이었음을 보여준다. 특히, 10개 폴드에 걸쳐 성능의 표준편차가 0.01 미만으로 매우 작게 나타났다. 이는 모델이 학습 데이터의 구성에 크게 좌우되지 않고, 다양한 데이터 분할에 대해 일관되고 안정적인 성능을 보이는 강건함(Robustness)을 가지고 있음을 의미한다. 따라서, 본 모델은 알려지지 않은 lncRNA-질병 연관성을 예측하는데 신뢰성 있게 활용될 수 있는 잠재력을 가지고 있다고 판단된다.

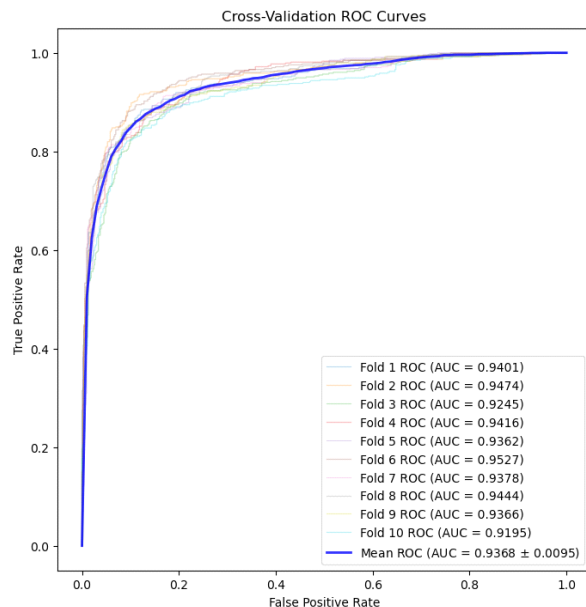


그림 4. 10-겹 교차 검증 ROC 곡선

Fig. 4. 10-fold cross-validation ROC curves

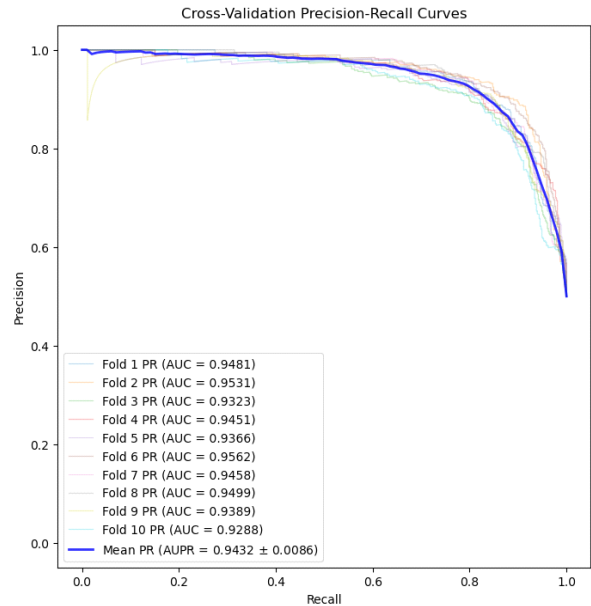


그림 5. 10-겹 교차 검증 PR 곡선

Fig. 5. 10-fold cross-validation PR curves

5.2 염기서열 길이 최적화

본 연구에서 lncRNA 염기서열의 길이를 2,000으로 표준화한 결정의 타당성을 실험적으로 검증하기 위해, 길이를 1,500, 2,500, 3,000으로 변경하며 분석을 수행하였다. 다른 모든 실험 조건은 동일하게 유지하였으며, 그 결과는 표 2와 같다.

표 2. 염기서열 길이에 따른 모델 성능 비교

Table 2. Model performance comparison by sequence length

Sequence length	Mean AUC ($\pm$ std)	Mean AUPR ($\pm$ std)
1,500	0.9336 ($\pm$ 0.0139)	0.9407 ($\pm$ 0.0114)
2,000	0.9381 ($\pm$ 0.0095)	0.9435 ($\pm$ 0.0086)
2,500	0.9386 ($\pm$ 0.0095)	0.9433 ($\pm$ 0.0084)
3,000	0.9336 ($\pm$ 0.0135)	0.9404 ($\pm$ 0.0114)

실험 결과, 모델은 서열 길이 2,000과 2,500에서 높은 예측 성능을 보였다. 구체적으로, AUC는 길이 2,500에서 0.9386으로, AUPR은 길이 2,000에서 0.9435

로 가장 높은 수치를 기록하였으나 두 길이 간의 성능 차이는 매우 근소하여 사실상 동등한 수준으로 평가되었다. 반면, 서열 길이가 1,500 미만이거나 3,000 이상일 경우에는 평균 성능이 하락하고 표준편차가 증가하여 안정성이 저하되는 경향이 나타났다. 이는 서열이 너무 짧으면 lncRNA의 고유 패턴을 학습하기에 정보가 불충분하고, 너무 길면 불필요한 패딩으로 인한 정보 희석 문제가 발생하기 때문으로 분석된다. 따라서, 두 평가 지표에서 높은 성능을 보인 2,000과 2,500 중, 계산 효율성 측면에서 더 유리한 2,000을 최종적인 최적 길이로 결정하였다.

5.3 CNN-GRU 제거 연구

본 연구에서 제안한 모델의 핵심 구성 요소 중 하나인 CNN-GRU의 중요성을 검증하기 위해 제거 연구(Ablation study)를 수행하였다. 이 실험에서는 CNN-GRU를 제거하고 대신 lncRNA 노드의 초기 특징을 무작위 벡터로 초기화하였다. HG-GCN과 연관성 예측 분류기를 포함한 나머지 모델 구조와 학습 파라미터는 모두 기존 모델과 동일하게 유지하였다. 이 모델을 동일한 10-겹 교차 검증 데이터셋으로 학습하고 평가함으로써, 염기서열 정보로부터 추출된 특징이 모델 성능에 얼마나 기여하는지를 정량적으로 분석하고자 하였다.

CNN-GRU 제거 모델의 10-겹 교차 검증 성능 결과는 표 3과 같고 ROC 곡선 그래프는 그림 6, PR 곡선 그래프는 그림 7과 같다. 염기서열 특징 추출 모듈을 제거한 모델은 평균 AUC 0.9329, 평균 AUPR 0.9399를 기록하였다. 이는 네트워크 구조 정보만으로도 높은 수준의 예측이 가능함을 보여준다. 그러나 이를 전체 모델의 성능인 AUC 0.9381, AUPR 0.9435과 비교했을 때, AUC는 약 0.0052, AUPR은 약 0.0036만큼 성능이 하락한 것을 확인할 수 있다. 또한, CNN-GRU 제거 모델의 성능 표준편차는 전체 모델에 비해 증가하였다. 이는 염기서열

정보가 평균적인 예측 정확도를 높일 뿐만 아니라, 데이터 분할에 따른 성능 변동성을 줄여 모델의 안정성에도 기여함을 시사한다.

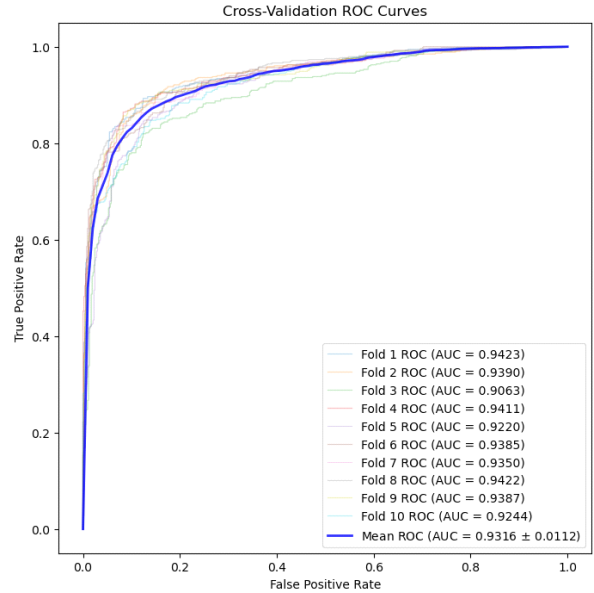


그림 6. CNN-GRU 제거 모델 ROC 곡선
Fig. 6. CNN-GRU ablation model ROC curves

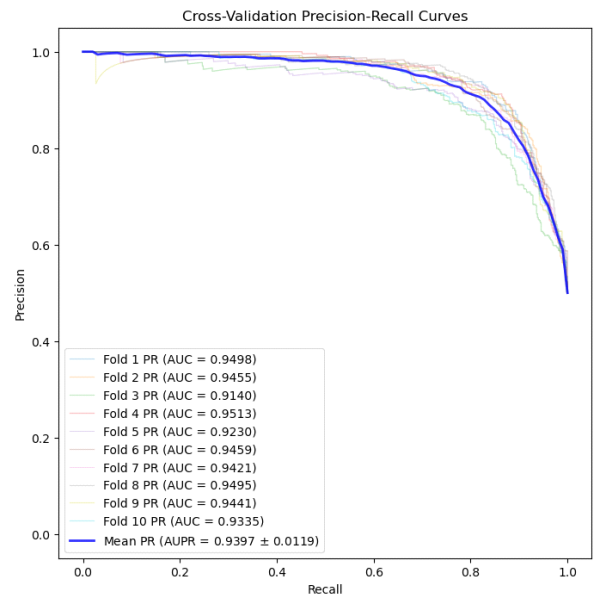


그림 7. CNN-GRU 제거 모델 PR 곡선
Fig. 7. CNN-GRU ablation model PR curves

표 3. CNN-GRU 제거 모델 성능 결과

Table 3. CNN-GRU ablation model performance results

Fold	1	2	3	4	5	6	7	8	9	10	mean
AUC	0.9423	0.9390	0.9063	0.9411	0.9220	0.9385	0.9422	0.9422	0.9387	0.9244	0.9329 ± 0.0112
AUPR	0.9498	0.9140	0.9140	0.9513	0.9230	0.9459	0.9421	0.9495	0.9441	0.9335	0.9399 ± 0.0119

이러한 성능 향상은 lncRNA 염기서열 정보가 단순히 네트워크 위상 정보만으로는 포착할 수 없는 lncRNA 고유의 생물학적 특성을 모델에 제공함으로써, 최종 예측 정확도 향상에 기여했음을 시사한다. 따라서 CNN-GRU는 제안 모델의 성능을 한 단계 더 개선하는 핵심 구성 요소임을 확인할 수 있었다.

5.4 비교 실험

제안된 모델의 우수성을 입증하기 위해, lncRNA-질병 연관성 예측 분야의 연구 모델들과 성능을 비교하였다. 비교 모델은 접근 방식에 따라 GAN 기반 BiGAN[39], 행렬 분해 기반 SIMCLDA[17], MFLDA[15], GCN 기반 HGC-GAN[37], GAT 기반 GANLDA[38], GAE 기반 GAMCLDA[42]으로 다양한 방법론을 포함하도록 선정하였다. 모든 비교 모델은 본 연구에서 사용한 것과 동일한 데이터와 10-겹 교차 검증 방법으로 평가되었다.

각 모델의 평균 AUC 및 AUPR 점수는 표 4와 같으며 비교 모델의 점수는 HGC-GAN[37] 연구로부터 얻을 수 있었다. 비교실험 결과, 본 연구에서 제안한 모델이 다른 모델들과 비교하였을 때 가장 우수한 성능을 보였다. 특히 AUC와 AUPR 두 지표 모두에서 최고 성능을 달성하여 제안된 접근법의 효과성을 입증하였다. GANLDA, GAMCLDA와 비교했을 때, 제안 모델은 AUC에서 더 높은 점수를 기록하였다. 이러한 성능 향상은 네트워크 구조 정보에만 의존하는 것과 달리, 제안 모델이 lncRNA 염기서열로부터 추출한 특징을 네트워크 정보와 효과적으로 통합하여 학습했기 때문으로 분석된다. 이는 lncRNA의 고유한 생물학적 정보가 연관성 예측의 정확도를 높이는 데 중요한 역할을 함을 뒷받침한다. AUPR 성능 격차는 매우 크게 나타났다. 이는 해당 모델들이 알려진 연관성을 제외한 모든 lncRNA-질병 쌍을 부정 샘플로 사용하는 데이터 전처리 방식을 따르기 때문이다. 이러한 방식은 극심

한 데이터 불균형을 야기하여 AUPR을 크게 저하시키는 원인이 된다. 반면, 본 연구에서는 긍정 샘플과 1:1 비율로 부정 샘플을 추출하여 데이터 불균형 문제를 완화하였고, 이는 높은 AUPR 점수로 이어졌다.

가장 성능이 근접한 모델인 HGC-GAN과 비교했을 때, 제안 모델은 AUC에서 약 0.0071, AUPR에서 약 0.0038 더 높은 점수를 기록하였다. HGC-GAN 역시 염기서열 데이터를 활용하지만, 단순한 k-mer 빈도수를 특징으로 사용하는 반면, 제안 모델은 CNN-GRU 기반의 딥러닝 모델을 통해 서열의 국소적, 순차적 패턴을 심도 있게 학습하여 특징을 추출한다. 이러한 결과는 lncRNA 염기서열의 복잡한 정보를 효과적으로 특징화하는 것이 최종 예측 성능 향상에 결정적인 영향을 미침을 보여준다. 결론적으로, 비교실험 결과는 lncRNA 염기서열 특징과 이중 네트워크 구조를 통합하는 본 연구의 접근 방식이 기존의 방법론들보다 더 정확한 lncRNA-질병 연관성 예측을 가능하게 함을 보여준다.

VI. 결론 및 향후 과제

본 연구에서는 lncRNA와 질병 간 잠재적 연관성을 예측하기 위한 새로운 딥러닝 모델을 제안하였다. 기존 연구들은 네트워크 위상 정보 또는 lncRNA의 개별 특징 정보 중 하나에 편중되어 있다는 한계를 가지고 있으며, 이를 극복하기 위해 본 연구에서는 lncRNA 염기서열 정보와 lncRNA, miRNA, 질병으로 구성된 이중 정보 네트워크의 구조적 정보를 효과적으로 통합하는 데 중점을 두었다. 제안된 모델은 CNN, GRU 및 어텐션 메커니즘을 결합하여 lncRNA 염기서열로부터 심층적인 특징을 추출하고, 이를 3계층 HG-GCN의 초기 노드 특징으로 활용하여 네트워크 내 복잡한 상호작용을 학습하였다.

표 4. 비교 실험 성능 결과

Table 4. Comparative experiment performance results

Model	CG-HGCN	HGC-GAN	GANLDA	BiGAN	GAMCLDA	SIMCLDA	MFLDA
AUC	0.9381	0.9310	0.8834	0.8930	0.9071	0.8433	0.8217
AUPR	0.9435	0.9397	0.0284	0.8851	0.0378	0.8824	0.8720

이러한 모델의 성능을 검증하기 위해 10-겹 교차 검증을 수행한 결과, 제안 모델은 평균 AUC 0.9381, AUPR 0.9435라는 우수한 예측 성능과 낮은 표준편차를 보여주며 그 효과성과 안정성을 입증하였다. 또한, CNN-GRU를 제거하는 제거 연구를 통해 해당 모듈이 모델의 정확도와 안정성 향상에 기여함을 확인하였고, 여러 최신 모델과의 비교실험에서는 모든 비교 대상보다 뛰어난 성능을 기록하며 제안된 접근법의 우수성을 증명하였다. 특히, 단순 k-mer 방식이 아닌 딥러닝을 통해 서열 특징을 추출하는 방식이 성능 향상에 결정적이었음을 확인하였다. 결론적으로, 본 연구에서 제시한 CG-HGCN 모델이 알려지지 않은 lncRNA-질병 연관성을 예측하는 데 효과적임을 보여준다. 이는 향후 질병 매커니즘 규명 및 새로운 바이오마커 발굴을 위한 후보군 우선순위를 정하는 데 유용한 계산 도구로 활용될 수 있을 것이다.

본 연구의 성과를 바탕으로 향후 연구를 확장할 수 있는 몇 가지 방향이 있다. 우선, 단백질-lncRNA 상호작용이나 lncRNA 발현 프로파일과 같은 다양한 종류의 오믹스 데이터를 통합하여 현재의 이중 네트워크를 더욱 풍부하게 구성할 수 있다. 또한, 본 연구에서는 lncRNA 노드의 특징만을 염기서열로부터 추출하였으나, miRNA의 염기서열이나 질병 온톨로지(Disease Ontology)를 활용하여 모든 노드 타입의 특징 표현을 고도화할 필요가 있다. 이는 네트워크 내에 월등히 많은 miRNA-질병 엣지로 인해 lncRNA 특징의 영향력이 일부 희석되었을 가능성을 보완하고, 각 개체의 고유한 특성을 모델이 더 잘 학습하도록 도울 것이다. 모델 아키텍처 측면에서는, 이중 그래프 어텐션 네트워크(Heterogeneous Graph Attention Networks)나 트랜스포머(Transformer)와 같은 고급 모델을 탐색하여 관계의 중요도를 동적으로 학습하거나 서열 내의 더 복잡한 의존성을 포착하는 연구를 진행할 수 있다. 마지막으로, 현재 모델의 예측 정확도를 넘어 생물학적 해석을 가능하게 하기 위해, 어텐션 가중치 분석이나 GNNExplainer와 같은 기법을 적용하여 모델의 해석 가능성을 증진시키는 연구는 매우 중요한 후속 과제가 될 것이다.

References

- [1] K. C. Wang and H. Y. Chang, "Molecular Mechanisms of Long Noncoding RNAs", *Molecular Cell*, Vol. 43, No. 6, pp. 904-914, Sep. 2011. <https://doi.org/10.1016/j.molcel.2011.08.018>.
- [2] M. Kazimierczyk, M. K. Kasprzowicz, M. E. Kasprzyk, and J. Wrzesinski, "Human Long Noncoding RNA Interactome: Detection, Characterization and Function", *IJMS*, Vol. 21, No. 3, pp. 1027, Feb. 2020. <https://doi.org/10.3390/ijms21031027>.
- [3] X. Zhang, R. Hong, W. Chen, M. Xu, and L. Wang, "The role of long noncoding RNA in major human disease", *Bioorganic Chemistry*, Vol. 92, pp. 103214, Nov. 2019. <https://doi.org/10.1016/j.bioorg.2019.103214>.
- [4] J. Beermann, M.-T. Piccoli, J. Viereck, and T. Thum, "Non-coding RNAs in Development and Disease: Background, Mechanisms, and Therapeutic Approaches", *Physiological Reviews*, Vol. 96, No. 4, pp. 1297-1325, Oct. 2016. <https://doi.org/10.1152/physrev.00041.2015>.
- [5] M. Zhou, et al., "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network", *Molecular BioSystems*, Vol. 11, No. 3, pp. 760-769, 2015. <https://doi.org/10.1039/C4MB00511B>.
- [6] X. Chen, Z.-H. You, G.-Y. Yan, and D.-W. Gong, "IRWLDA: improved random walk with restart for lncRNA-disease association prediction", *Oncotarget*, Vol. 7, No. 36, pp. 57919-57931, Aug. 2016. <https://doi.org/10.18632/oncotarget.11141>.
- [7] C. Gu, et al., "Global network random walk for predicting potential human lncRNA-disease associations", *Sci Rep*, Vol. 7, No. 1, pp. 12442, Sep. 2017. <https://doi.org/10.1038/s41598-017-12763-z>.
- [8] Y. Gao, J. Hu, and X. Shang, "Identification of

- lncRNA-disease association using bi-random walks", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, pp. 1249-1255, Feb. 2018. <https://doi.org/10.1109/BIBM.2018.8621132>.
- [9] X. Zhao, Y. Yang, and M. Yin, "MHRWR: Prediction of lncRNA-Disease Associations Based on Multiple Heterogeneous Networks", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 18, No. 6, pp. 2577-2585, Jan. 2021. <https://doi.org/10.1109/TCBB.2020.2974732>.
- [10] X. Chen and G.-Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles", *Bioinformatics*, Vol. 29, No. 20, pp. 2617-2624, Oct. 2013. <https://doi.org/10.1093/bioinformatics/btt426>.
- [11] W. Lan, et al., "LDAP: a web server for lncRNA-disease association prediction", *Bioinformatics*, Vol. 33, No. 3, pp. 458-460, Feb. 2017. <https://doi.org/10.1093/bioinformatics/btw639>.
- [12] D. Yao, X. Zhan, X. Zhan, C. K. Kwoh, P. Li, and J. Wang, "A random forest based computational model for predicting novel lncRNA-disease associations", *BMC Bioinformatics*, Vol. 21, No. 1, pp. 126, Mar. 2020. <https://doi.org/10.1186/s12859-020-3458-1>.
- [13] R. Zhu, Y. Wang, J.-X. Liu, and L.-Y. Dai, "IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier", *BMC Bioinformatics*, Vol. 22, No. 1, pp. 175, Apr. 2021. <https://doi.org/10.1186/s12859-021-04104-9>.
- [14] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for lncRNA-disease association prediction based on the Naïve Bayesian classifier", *BMC Bioinformatics*, Vol. 20, No. 1, pp. 396, Jul. 2019. <https://doi.org/10.1186/s12859-019-2985-0>.
- [15] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncRNA-disease associations", *Bioinformatics*, Vol. 34, No. 9, pp. 1529-1537, May 2018. <https://doi.org/10.1093/bioinformatics/btx794>.
- [16] Z. Xuan, J. Li, J. Yu, X. Feng, B. Zhao, and L. Wang, "A Probabilistic Matrix Factorization Method for Identifying lncRNA-Disease Associations", *Genes*, Vol. 10, No. 2, pp. 126, Feb. 2019. <https://doi.org/10.3390/genes10020126>.
- [17] C. Lu, et al., "Prediction of lncRNA-disease associations based on inductive matrix completion", *Bioinformatics*, Vol. 34, No. 19, pp. 3357-3364, Oct. 2018. <https://doi.org/10.1093/bioinformatics/bty327>.
- [18] J. Ha and K. Kim, "Neighborhood-regularized matrix factorization for lncRNA-disease association identification", *International Journal of Molecular Sciences*, Vol. 26, No. 9, pp. 4283, Apr. 2025. <https://doi.org/10.3390/ijms26094283>.
- [19] K. Kim and J. Ha, "GMFLDA: improved prediction of lncRNA-disease association via graph convolutional network", *IEEE Access*, Vol. 13, pp. 85330-85341, May 2025. <https://doi.org/10.1109/ACCESS.2025.3568461>.
- [20] J. Ha, "DeepWalk-based graph embeddings for miRNA-disease association prediction using deep neural network", *Biomedicines*, Vol. 13, No. 3, pp. 536, Feb. 2025. <https://doi.org/10.3390/biomedicines13030536>.
- [21] J. Ha, "Graph convolutional network with neural collaborative filtering for predicting miRNA-disease association", *Biomedicines*, Vol. 13, No. 1, pp. 136, Jan. 2025. <https://doi.org/10.3390/biomedicines13010136>.
- [22] J. Ha, "MDMF: Predicting miRNA-Disease Association Based on Matrix Factorization with Disease Similarity Constraint", *Journal of Personalized Medicine*, Vol. 12, No. 6, pp. 885, May 2022.
- [23] J. Ha and S. Park, "NCMD: Node2vec-based neural collaborative filtering for predicting

- miRNA-disease association", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 20, No. 2, pp. 1257-1268, Mar.-Apr. 2022. <https://doi.org/10.1109/TCBB.2022.3191972>.
- [24] J. Ha, "SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association", KKnowledge-Based Systems, Vol. 263, pp. 110295, Mar. 2023. <https://doi.org/10.1016/j.knosys.2023.110295>.
- [25] J. Ha, "LncRNA Expression Profile-Based Matrix Factorization for Predicting lncRNA- Disease Association", IEEE Access, Vol. 12, pp. 70297-70304, May 2024. <https://doi.org/10.1109/ACCESS.2024.3401005>.
- [26] B. Kim, S. Kim, and J. Ha, "An Ensemble Machine Learning Framework for Identifying LncRNA-Disease Association", The Journal of Korean Institute of Information TechNology, Vol. 23, No. 6, pp. 41-53, Jun. 2025. <http://doi.org/10.14801/jkiit.2025.23.6.41>.
- [27] S. Kim and J. Ha, "Development of LncRNA-Disease Associations Prediction Model using CNN-RNN Feature Extraction based on LncRNA Sequence Data", The Journal of Korean Institute of Information TechNology, Vol. 22, No. 10, pp. 1-11, Oct. 2024. <http://doi.org/10.14801/jkiit.2024.22.10.1>.
- [28] J.-H. Ha, "Autoencoder-based Disease-related miRNA Prediction Research using Deep Learning", The Journal of Korean Institute of Information TechNology, Vol. 20, No. 6, pp. 33-40, Jun. 2022. <http://doi.org/10.14801/jkiit.2022.20.6.33>.
- [29] J. Ha, C. Park, C. Park, and S. Park, "IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization", Journal of Biomedical Informatics, Vol. 102, pp. 103358, Feb. 2020. <https://doi.org/10.1016/j.jbi.2019.103358>.
- [30] J. Ha, C. Park, C. Park, and S. Park, "Improved prediction of miRNA-disease associations based on matrix completion with network regularization", Cells, Vol. 9, No. 4, pp. 881, Apr. 2020. <https://doi.org/10.3390/cells9040881>.
- [31] J. Ha, C. Park, and S. Park, "PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach", BMC Systems Biology, Vol. 13, No. 1, pp. 1-13, Mar. 2019. <https://doi.org/10.1186/s12918-019-0700-4>.
- [32] J. Ha and C. Park, "MLMD: Metric Learning for Predicting MiRNA-Disease Associations", IEEE Access, Vol. 9, pp. 78847-78858, May 2021. <https://doi.org/10.1109/ACCESS.2021.3084148>.
- [33] J. Hu, Y. Gao, J. Li, and X. Shang, "Deep Learning Enables Accurate Prediction of Interplay Between lncRNA and Disease", Front. Genet., Vol. 10, Oct. 2019. <https://doi.org/10.3389/fgene.2019.00937>.
- [34] M. Zeng, et al., "DMFLDA: A Deep Learning Framework for Predicting lncRNA-Disease Associations", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 18, No. 6, pp. 2353-2363, Jan. 2021. <https://doi.org/10.1109/TCBB.2020.2983958>.
- [35] P. Xuan, Y. Cao, T. Zhang, R. Kong, and Z. Zhang, "Dual ConVolutional Neural Networks With Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes", Front. Genet., Vol. 10, May 2019. <https://doi.org/10.3389/fgene.2019.00416>.
- [36] P. Xuan, L. Jia, T. Zhang, N. Sheng, X. Li, and J. Li, "LDAPred: A Method Based on Information Flow Propagation and a ConVolutional Neural Network for the Prediction of Disease-Associated lncRNAs", International Journal of Molecular Sciences, Vol. 20, No. 18, pp. 4458, Jan. 2019. <https://doi.org/10.3390/ijms20184458>.
- [37] P. Xuan, N. Sheng, T. Zhang, Y. Liu, and Y. Guo, "CNNDLP: A Method Based on ConVolutional Autoencoder and ConVolutional Neural Network with Adjacent Edge Attention for

- Predicting lncRNA-Disease Associations", *International Journal of Molecular Sciences*, Vol. 20, No. 17, pp. 4260, Jan. 2019. <https://doi.org/10.3390/ijms20174260>.
- [38] N. Sheng, H. Cui, T. Zhang, and P. Xuan, "Attentional multi-level representation encoding based on conVolutional and variance autoencoders for lncRNA-disease association prediction", *Brief Bioinform*, Vol. 22, No. 3, pp. bbaa067, May 2021. <https://doi.org/10.1093/bib/bbaa067>.
- [39] Q. Yang and X. Li, "BiGAN: lncRNA-disease association prediction based on bidirectional generative adversarial network", *BMC Bioinformatics*, Vol. 22, No. 1, p. 357, Jun. 2021. <https://doi.org/10.1186/s12859-021-04273-7>.
- [40] P. Xuan, S. Pan, T. Zhang, Y. Liu, and H. Sun, "Graph ConVolutional Network and ConVolutional Neural Network Based Method for Predicting lncRNA-Disease Associations", *Cells*, Vol. 8, No. 9, pp. 1012, Sep. 2019. <https://doi.org/10.3390/cells8091012>.
- [41] Z. Lu, H. Zhong, L. Tang, J. Luo, W. Zhou, and L. Liu, "Predicting lncRNA-disease associations based on heterogeneous graph conVolutional generative adversarial network", *PLOS Computational Biology*, Vol. 19, No. 11, pp. e1011634, Nov. 2023. <https://doi.org/10.1371/journal.pcbi.1011634>.
- [42] W. Lan, X. Wu, Q. Chen, W. Peng, J. Wang, and Y. P. Chen, "GANLDA: Graph attention network for lncRNA-disease associations prediction", *Neurocomputing*, Vol. 469, pp. 384-393, Jan. 2022. <https://doi.org/10.1016/j.neucom.2020.09.094>.
- [43] X. Zhao, X. Zhao, and M. Yin, "Heterogeneous graph attention network based on meta-paths for lncRNA-disease association prediction", *Brief Bioinform*, Vol. 23, No. 1, pp. bbab407, Jan. 2022. <https://doi.org/10.1093/bib/bbab407>.
- [44] A. B. O. V. Silva and E. J. SpiNosa, "Graph ConVolutional Auto-Encoders for Predicting Novel lncRNA-Disease Associations", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 19, No. 4, pp. 2264-2271, Jul. 2022. <https://doi.org/10.1109/TCBB.2021.3070910>.
- [45] Z. Shi, H. Zhang, C. Jin, X. Quan, and Y. Yin, "A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations", *BMC Bioinformatics*, Vol. 22, No. 1, p. 136, Mar. 2021. <https://doi.org/10.1186/s12859-021-04073-z>.
- [46] X. Wu, W. Lan, Q. Chen, Y. Dong, J. Liu, and W. Peng, "Inferring lncRNA-disease associations based on graph autoencoder matrix completion", *Computational Biology and Chemistry*, Vol. 87, pp. 107282, Aug. 2020. <https://doi.org/10.1016/j.compbiolchem.2020.107282>.
- [47] Y. Gao, et al., "lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data", *Nucleic Acids Res*, Vol. 49, No. D1, pp. D1251-D1258, Jan. 2021. <https://doi.org/10.1093/nar/gkaa1006>.
- [48] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, "lncRNADisease 2.0: an updated database of long Non-coding RNA-associated diseases", *Nucleic Acids Res*, Vol. 47, No. D1, pp. D1034-D1037, Jan. 2019. <https://doi.org/10.1093/nar/gky905>.
- [49] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data", *Nucleic Acids Res*, Vol. 42, No. D1, pp. D92-D97, Jan. 2014. <https://doi.org/10.1093/nar/gkt1248>.
- [50] Z. Huang, et al., "HMDD v3.0: a database for experimentally supported human microRNA-disease associations", *Nucleic Acids Res*, Vol. 47, No. D1, pp. D1013-D1017, Jan. 2019. <https://doi.org/>

10.1093/nar/gky1010.

- [51] National Center for Biotechnology Information, <https://www.ncbi.nlm.nih.gov/>. [accessed: Jul. 21, 2025]
- [52] The RNAcentral Consortium, "RNAcentral: a hub of information for Non-coding RNA sequences", Nucleic Acids Res, Vol. 47, No. D1, pp. D221-D229, Jan. 2019. <https://doi.org/10.1093/nar/gky1034>.
- [53] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks", The Semantic Web, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., Cham: Springer International Publishing, pp. 593-607, 2018. https://doi.org/10.1007/978-3-319-93417-4_38.

저자소개

김 상 옥 (Sanguk Kim)



2024년 8월 : 부경대학교
냉동공조공학과(공학사)
2024년 8월 : 부경대학교
빅데이터융합전공(공학사)
2024년 9월 ~ 현재 : 부경대학교
데이터공학과 석사과정
관심분야 : 기계학습, 생물정보학,
데이터마이닝, 딥러닝, 추천시스템

윤 태 현 (Taehyeon Yun)



2025년 2월 : 국립부경대학교
해양바이오신소재학과(이학사)
2025년 2월 : 국립부경대학교
빅데이터융합전공(공학사)
2025년 3월 ~ 현재 : 국립부경대학
교 데이터공학과 석사과정
관심분야 : 기계학습, 생물정보학,
데이터마이닝, 딥러닝, 의료인공지능

류 지 열 (Jee-Youl Ryu)

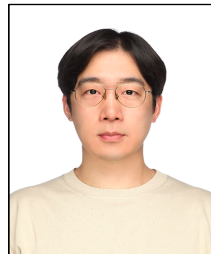


1993년 2월 : 부경대학교
전자공학과(공학사)
1997년 2월 : 부경대학교
전자공학과(공학석사)
2004년 12월 : 애리조나
주립대학교 전기공학과(공학박사)
2009년 3월 ~ 현재 : 부경대학교

정보통신공학과 교수

관심분야 : 시스템-온-칩 설계, 고주파 회로 설계,
임베디드 시스템 설계

하 지 환 (Jihwan Ha)



2015년 8월 : 연세대학교
컴퓨터과학과(공학석사)
2020년 8월 : 연세대학교
컴퓨터과학과(공학박사)
2020년 9월 ~ 2021년 8월 :
하와이 암센터 포스닥 연구원
2021년 9월 ~ 2025년 8월 :

부경대학교 데이터정보과학부 빅데이터융합전공
조교수

2025년 9월 ~ 현재 : 부경대학교 데이터정보과학부
빅데이터융합전공 부교수

관심분야 : 기계학습, 생물정보학, 데이터마이닝, 딥러닝,
추천시스템