Journal of KIIT. Vol. 23, No. 6, pp. 41-53, Jun. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.6.41

앙상블 기계학습 기반 LncRNA - 질병 연관성 예측 모델 연구

김보훈*, 김상욱**, 하지환***

An Ensemble Machine Learning Framework for Identifying LncRNA - Disease Association

Bohoon Kim*, Sanguk Kim**, and Jihwan Ha***

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00242528)

요 약

LncRNA는 비암호화 RNA(non-coding RNA) 유전자의 한 종류로서, 다양한 질병에서 중요한 역할을 하는 것으로 알려져 있다. LncRNA는 최근까지 다양한 연구와 모델들이 제안되어왔다. 기존 연구에서는 개별 모델을 적용하여 LDA(LncRNA Disease Association)를 분석하였으나, 단일 모델의 한계로 인해 높은 예측 성능을 보장하기 어려웠다. 본 연구는 SVM(Support Vector Machine), GBM(Gradient Boosting Machine), XGBoost (eXtreme Gradient Boosting)을 결합한 앙상블 학습 기법(Ensemble Learning)을 적용하여 LncRNA - 질병 연관성 예측 모델의 성능을 향상시키는 방안을 제안한다. SVD(Singular Value Decomposition) 및 거리 기반 음성 샘플링(distance-based negative sampling) 전략을 활용하여 AUC 0.93의 예측 성능을 달성하였으며, 다양한 앙상블 모델 조합에 대한 비교 실험을 수행하였다. 실험 결과, 제안한 모델이 기존 방법들에 비해 우수한 예측 성능을 보임을 확인하였다.

Abstract

Long non-coding RNAs (LncRNAs) are a class of non-coding RNA genes that have been reported to play critical roles in various human diseases. Numerous studies and predictive models have been proposed to investigate LncRNA - disease associations. However, previous approaches that relied on individual models have shown limited predictive performance due to the inherent constraints of single-model architectures. To address this issue, this study proposes an ensemble learning-based method that integrates Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and eXtreme Gradient Boosting (XGBoost) to enhance the accuracy of LncRNA - disease association prediction. By applying Singular Value Decomposition (SVD) and a distance-based negative sampling strategy, the proposed predictive model achieved an AUC of 0.93, demonstrating a significant improvement in performance. Furthermore, comparative experiments were conducted using various combinations of individual ensemble models. The results indicate that the model proposed in this study outperformed the other methods, confirming its superior predictive capability.

Keywords

bioinformatics, ensemble learning, support vector machine, gradient boosting

* 부경대학교 빅데이터융합학과 석사과정
- ORCID: https://orcid.org/0009-0008-6011-3644
** 부경대학교 데이터공학과 석사과정
- ORCID: https://orcid.org/0009-0007-3765-2212
*** 부경대학교 빅데티어융합전공 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6086-5693

· Received: Apr. 17, 2025, Revised: May 23, 2025, Accepted: May 26, 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

1990년대부터 시작된 인간 게놈 프로젝트(Human Genome Project)를 통해, 우리는 유전체에 대한 정보를 완전히 파악함으로써 생명체의 복잡성을 이해할 수 있을 것으로 기대하였다. 그러나 이후의 분석 결과, 인간 유전체 내에서 단백질을 암호화하는 유전자의 수는 약 2만 개 수준에 불과하며, 이는 선형동물인 예쁜꼬마선충(Caenorhabditis elegans)의 유전자 수와 유사하다는 사실이 밝혀졌다. 이러한 결과는 단백질 코딩 유전자만으로는 인간과 같은 고등생물의 생리적 및 진화적 복잡성을 충분히 설명하기 어렵다는 점을 시사한다. 이에 따라 최근에는 비암호화 유전자 영역의 기능과 역할에 대한 관심이 높아지고 있으며, 이를 규명하기 위한 다양한 연구가 활발히 진행되고 있다[1].

비암호화 RNA(Non-coding RNA)는 전체 유전자의 98% 이상을 차지하며, 그 크기에 따라 short non-coding RNAs, mid-size non-coding RNAs, long non-coding RNAs로 분류된다. 또한, 이들은 역할과 전사 후 변형 과정에 따라 다시 세분화된다. 이 중에서 short non-coding RNA인 microRNA(miRNA)와 long non-coding RNA(LncRNA)는 여러 질병과 관련이 있는 것으로 알려져 있다[1]. 이에 따라 LncRNA와 질병 간의 연관성(LncRNA-Disease Association, LDA)을 예측하고 이해하기 위한 다양한 계산적 접근법이 시도되었다. 기존 연구들은 네트워크 기반 분석, 행렬 분해, 단일 머신러닝 모델 등을 활용하여 LDA 예측을 수행해왔다. 하지만 이러한 접근 방식들은 단일 모델의 내재적 한계, 데이터의 희소성 문제, 특징 표현의 제약 등으로 인해 예측 성능 향상에 어려움을 겪어왔다. 특히, 일부 연구에서 앙상블 기법을 시도하였으나, 사용된 모델 조합이나 특징 추출 방식이 제한적이어서 여전히 성능 개선의 여지가 남아있다.

본 연구는 기존 연구들의 이러한 한계를 극복하기 위해 다음과 같은 차별화된 접근 방식을 제안한다. 첫째, SVM(Support Vector Machine), GBM(Gradient Boosting Machine), XGBoost(eXtreme Gradient Boosting)라는 세 가지 강력한 알고리즘을 조합한 특화된 앙상블 모델을 구축하여 개별 모델의 강점을 활용하고 약점을 보완함으로써 예측 성능을 극대화하고자 한다. 둘째, LncRNA와 질병 데이터를 효과적으로 표현하기 위해 특이값 분해(SVD, Singular Value Decomposition) 기반의 특징 추출 기법을 적용하여 고차원 희소 데이터 문제를 해결하고 잠재 특징을 효과적으로 학습한다. 셋째, 실제 실험적으로 검증된 양성 샘플 외에 신뢰성 있는 음성 샘플을 확보하기 어려운 LDA 데이터의 특성을 고려하여, 거리 기반 음성 샘플링 전략을 도입함으로써 레이블 노이즈 문제를 완화하고 모델의 강건성(Robustness)을 높인다. 이 세 가지 방법을 사용해 기존 단일 모델 및 단순 앙상블 기반 접근 방식의 예측 정확도와 강건성 한계를 극복하고, 보다 신뢰도 높은 LncRNA-질병 연관성 예측 모델을 제시하는 데 기여한다.

Ⅱ. 관련 연구

최근 4차 산업혁명의 영향으로 인공지능을 활용한 연구가 점차 증가하고 있으며, 특히 질병과 관련된 LncRNA 분석에서도 인공지능 모델이 핵심 기술로 자리 잡고 있다[2]-[15].

비암호화 RNA는 다양한 생물학적 및 의학적 연구에서 점점 더 주목받고 있으며, 생물학적, 발달적, 병리학적 과정에서 중요한 역할을 수행하는 것으로 밝혀졌다[16]-[19]. 이에 따라 LncRNA와 질병 간의 연관성을 분석하는 다양한 계산적 접근법이 제안되었다. 대표적인 방법으로 Ping's method가 있으며, 이는 LncRNA-질병 연관성 데이터를 바탕으로 이분 그래프(Bipartite network)를 구축하여 두 요소 간의 잠재적 관계를 예측하는 기법이다. 해당 모델은 기존의 연관성 정보를 활용해 네트워크를 형성하며, 공통 이웃을 공유하는 노드 간의 유사성을 계산하여 예측을 수행한다[20].

J. Zhang et al.[21]이 제안한 LncRDNetFlow 모델은 LncRNA 유사성 네트워크, 단백질-단백질 상호작용 네트워크, 질병 유사성 네트워크 등 이질적인 데이터 네트워크를 통합하는 방식을 채택하였다. 이 모델은 흐름 전파(Flow propagation) 알고리즘을 활용

하여 네트워크의 위상 정보(Topological information)를 반영한 글로벌 거리(Global distance measurement)를 계산함으로써, LncRNA와 질병 간의 숨겨진 연관성을 예측한다. LDAP 모델은 LncRNA의 서열 정보, 단백질-단백질 상호작용 데이터, 질병 유사성 정보를 종합적으로 평가한 후, 이를 Bagging SVM을 이용해 예측하는 방식으로 설계되었다[20]. 한편, MFLDA는 다양한 생물학적 데이터를 기반으로 행렬 분해(Matrix factorization) 기반의 데이터 융합(Data fusion) 기법을 적용하여 LncRNA-질병 연관성을 예측하는 모델이다[23]. WGRCMF 모델은 그래프 정규화(Graph regularization) 기법을 적용하여 협력 행렬 분해(Collaborative matrix factorization) 기반의 예측 성능을 개선하였다[24]. 특히, 가중치 행렬을 도입하여 알려지지 않은 연관성의 영향을 최소화함으로써, 보다 신뢰성 높은 예측이 가능하도록 설계되었다. 또한, SIMCLDA 모델은 유사성 기반 귀납 행렬(Inductive matrix)을 활용하여, 기능적으로 유사한 LncRNA가 동일한 질병과 상호작용을 할 가능성이 높다는 가설을 바탕으로 특징 공간을 정의한다[25]. 이를 통해 질병 유사성을 정량적으로 평가하여 보다 정밀한 예측을 수행할 수 있다.

한편, X. Chen et al.[26]은 앙상블 학습(Ensemble learning) 기법을 활용하여 의사결정 나무(Decision tree) 기반의 새로운 예측 모델을 제안하였다. 해당 모델은 PCA 기법을 적용하여 차원을 축소함으로써 질병과 관련된 RNA를 효과적으로 추출하였다. 의사결정 나무는 학습된 트리 구조를 기반으로 다수결 방식으로 최종 결정을 내리는 방식이지만, 데이터 패턴이 복잡하거나 비선형적인 경우 단일 트리 모델만으로 최적의 의사결정을 내리기 어렵다. 특히, 고차원 데이터에서는 차원의 저주(Curse of dimensionality)로 인해 성능 저하가 발생할 가능성이 높으며, 깊은 트리를 학습할 경우 과적합(Overfitting)의 위험이 커진다. 이를 보완하기 위해 XGBoost나 그래디언트 부스팅(Gradient boosting)과 같은 모델에서는 과적합을 방지하는 다양한 기법을 적용하여 보다 안정적인 성능을 제공한다.

일반적으로 앙상블 학습은 단일 모델보다 성능이 우수할 가능성이 높다. 랜덤 포레스트(Random forest)나 XGBoost와 같은 개별 모델도 강력한 성능을 보이지만, 서로 다른 모델을 조합하면 일반화 성능이 더욱 향상될 수 있다[27]. 단순한 모델들의 앙상블도 강력한 예측 성능을 가질 수 있으나, 개별 모델이 충분히 독립적인 정보를 보유하고 서로를 보완할 수 있을 때 더 효과적이다[28].

III. 앙상블 기계학습 기반 모델링

3.1 데이터

다양한 연구를 통해 질병 발병 인자로서의 RNA 기능들이 다양한 생물학적 근거로 입증되고 있다. 이로 인하여, 다양한 온라인 데이터베이스에서 질병과 LncRNA의 관계 정보를 제공하고 있다.

본 논문에서는 질병 관련 LncRNA 정보를 제공하는 LncRNADisease v3.0 온라인 데이터베이스에서 질병과 LncRNA 관계 정보를 다운로드하여 모델의 훈련 및 검증 데이터로 사용하였다[29]. 검증된 관계 값을 사용하기 위해, 생물학 실험을 통해 증명된 데이터베이스를 사용하였다. LncRNADisease v3.0는 1,759개의 LncRNA와 344개의 질병에 대해 실험적으로 뒷받침되는 6,856개의 알려진 LncRNA-질병 연관성 데이터를 제공하고 있다.

유전자와 질병 관계 예측을 위한 학습 데이터는 이렇게 실험적으로 인과관계가 검증된 유전자와 질병 쌍 6,856개를 양성 샘플로 구성하였다. 반면, 인과관계가 밝혀지지 않은 대부분의 샘플은 잠재적인 관계가 존재할 수 있는 미확정 데이터로 간주된다. 이러한 샘플을 그대로 음성 샘플(Negative sample)로 사용할 경우, 심각한 레이블 노이즈(Label noise)가 발생할 수 있다. 이를 방지하기 위해 본 연구에서는 거리 기반 음성 샘플링 전략을 도입하였다. 구체적으로, 모든 유전자와 질병 쌍 간의 유클리드 거리(Euclidean distance)를 계산하고, 양성 샘플들 간 거리의 평균값을 기준 임계값(Threshold)으로 설정하였다. 이후, 이 평균 거리보다 충분히 멀리 떨어진 유전자와 질병 쌍만을 신뢰할 수 있는 음성 샘플로 선정하였다. 이 방식을 통해 총 2,313개의 음성 샘플(Negative samples)을 생성하였다.

최종적으로 6,856개의 양성 샘플과 생성된 2,313개의 음성 샘플을 결합하여, 총 9,169개의 샘플로 구성된 전체 데이터셋을 구축하였다. 이 데이터셋을 이진 분류(Binary classification) 모델 학습 및 평가를 위한 실험 데이터로 사용하였다. 이와 같은 샘플링 전략은 PU Learning의 한계를 극복하고, 모델 학습 시 음성 레이블로 인한 오분류를 최소화하는 데 활용하였다.

3.2 특이값 분해(SVD) 기반 특징 추출

본 연구에서는 유전자(LncRNA)와 질병(Disease) 간의 연관성을 분석하기 위해 특이값 분해(SVD, Singular Value Decomposition)를 활용한 특징 추출(Feature extraction) 기법을 적용하였다[5]. LncRNA 및 질병 데이터는 범주형 데이터(Categorical data)로 구성되어 있으므로, 이를 수치적으로 표현하기 위해 원 핫 인코딩(One-Hot encoding) 기법을 적용하였다. 이후, 고차원의 희소 행렬을 보다 효율적으로 처리하기 위해 특이값 분해를 수행하여 차원을 축소하였다(그림 1).

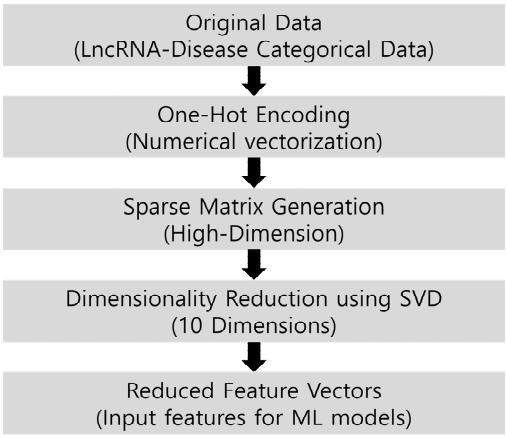

그림 1. 데이터 전처리 흐름도
Fig. 1. Data preprocessing flowchart

특이값 분해는 고차원 희소 데이터의 문제를 해결하는 데 효과적인 기법으로, 선행 단계에서 원본 행렬을 특이값 분해하여(식 (1)), 저차원 표현을 생성한다. 이를 통해 희소한 데이터 구조를 밀집 행렬로 변환하여, 계산 복잡도를 감소시키고 데이터 내 잠재적 관계를 보다 효과적으로 학습할 수 있도록 한다.

특히, 본 연구에서는 유전자-질병 관계를 나타내는 행렬에 대해 특이값 분해를 적용하여, 다음과 같이 분해하였다: 여기서 유전자와 질병의 k차원 저차원 표현이며, 상위 k개의 특이값으로 구성된 대각행렬이다. 이 과정은 데이터의 주요 패턴을 보존하면서 노이즈를 제거하는 효과를 지니며, k 값은 정보 보존과 차원 축소 간의 균형을 고려하여 설정된다. 또한, 유전자와 질병 각각의 특이값 분해 기반 임베딩 벡터를 각각 계산한 후, 이들을 병합(Concatenation)하여 하나의 입력 벡터로 구성하였다. 이는 유전자와 질병 양측의 잠재적 표현을 동시에 활용함으로써, 관계 예측의 성능을 높이는 데 기여한다. 이러한 저차원 병합 벡터를 바탕으로 기계학습 및 앙상블 모델을 적용하여 LDA 관계를 예측하는 연구를 진행하고자 한다.

$$X = U \sum V T^T \qquad (1)$$
$$X_{reduced} = U_k \sum_k V_k^T$$

3.3 기계학습 기반 예측 모델 선정

유전자와 질병 간의 관계 예측은 의학 및 생물정보학 분야에서 중요한 연구 주제로 다루어지고 있다[30]. 기존 연구에서는 기계학습 및 행렬 분해 기반 모델을 활용하여 RNA와 질병 간의 연관성을 분석하였으나, 단일 모델의 한계로 인해 높은 예측 성능을 확보하는 데 어려움이 있었다. 이를 해결하기 위해, 본 연구에서는 SVM, GBM, XGBoost 모델을 결합한 앙상블 기법1을 적용하여 LDA 예측의 성능을 향상시키는 방법을 제안한다.

SVM은 마진 최대화(Margin maximization)를 통해 결정 경계를 정의하므로 일반화 성능이 우수하고, 규제(Regularization)를 통해 과적합 위험을 관리할 수 있다. 이러한 능력 덕분에 SVM은 분류 경계가 명확하지 않거나 복잡한 데이터셋에서도 안정적인 성능을 보이며, 유전자 발현 분석, 단백질 분류 등 다양한 생물정보학 문제에 성공적으로 적용되어 그 효용성을 입증해왔다.

GBM은 여러 개의 약한 학습기(주로 결정 트리)를 순차적으로 학습시키면서 이전 단계의 오류를 보완해나가는 부스팅(Boosting) 계열 알고리즘이다. 이는 데이터 내의 복잡한 상호작용과 비선형 관계를 효과적으로 모델링할 수 있으며, 예측 정확도가 높다는 장점이 있다. 생물학적 시스템은 다수의 유전자나 분자 간의 복잡한 상호작용 네트워크로 구성되는 경우가 많으므로, 이러한 상호작용을 잘 포착할 수 있는 GBM은 LncRNA-질병 연관성과 같은 생물정보학 예측 문제에 유용하게 활용될 수 있다.

XGBoost는 GBM을 기반으로 성능과 속도를 크게 개선한 알고리즘이다. GBM의 장점을 계승하면서도, 규제(Regularization) 항을 손실 함수에 추가하여 과적합을 효과적으로 방지하고 병렬 처리(Parallel processing)를 지원하여 대용량 데이터에서도 빠른 학습 속도를 보인다. LncRNA-질병 데이터와 같은 생물정보학 데이터는 노이즈가 많고 복잡하며 데이터 크기가 커질 수 있는데, XGBoost의 높은 예측 성능, 과적합 방지 기능, 계산 효율성은 이러한 데이터의 분석 및 모델링에 매우 적합하다.

이 세 모델을 앙상블의 기반 학습기로 함께 사용한 이유는 각 모델이 가진 서로 다른 강점과 학습 방식을 결합하여 상호 보완적인 효과를 얻기 위함이다. SVM이 데이터 경계면을 잘 구분하는 데 강점을 보인다면, 부스팅 계열 모델들은 복잡한 패턴과 상호작용을 포착하는 데 뛰어나다. 이처럼 서로 다른 관점에서 데이터를 학습하는 모델들의 예측 결과를 종합함으로써, 단일 모델을 사용할 때보다 예측의 정확성과 안정성, 일반화 성능을 크게 향상시킬 수 있다. 본 연구에서는 유전자와 질병이라는 제한된 특징 두 개를 기반으로 하여, 이러한 모델들의 장점을 최대한 활용하고 생물학적 특성과 데이터 구조를 고려한 최적의 앙상블 예측 모델을 구축하고자 하였다.

또한, 앙상블 모델의 성능을 극대화하기 위해 최종 메타 학습기(Meta-learner)로 로지스틱 회귀(Logistic regression)를 적용하였다. 이를 통해 개별 모델의 예측 결과를 효과적으로 결합하여 예측력을 향상시키고, 보다 정밀한 질병 연관성 분석이 가능하게 하였다. 이하에서는 연구에서 활용한 각 모델의 주요 특징과 기능을 상세히 설명하고자 한다.

3.3.1 로지스틱 회귀 모델

로지스틱 회귀 모델은 이진 분류 문제에서 널리 사용되는 기본적인 통계 기반 기계학습 알고리즘이다. 선형 회귀(Linear regression)와 유사한 형태를 가지지만, 예측값을 0과 1 사이의 확률값으로 변환하기 위해 시그모이드(Sigmoid) 함수를 적용하는 점에서 차이가 있다. 로지스틱 회귀 모델의 기본 구조는 선형 회귀 방식과 동일하게 입력 변수 X에 대해 가중치 w와 편향 b를 적용하여 선형 결합된 값을 계산하는 형태를 따른다. 즉, 예측 함수는 $z = wX + b$ 형태의 선형 방정식으로 정의된다. 이후, 선형 결합된 값 z에 시그모이드 함수를 적용하여 출력값을 확률값으로 변환한다. 이를 통해 모델의 출력값은 항상 0과 1 사이의 확률값이 되며, 일반적으로 0.5를 기준 임계값으로 설정하여 결과를 이진 분류한다. 즉, $P(Y=1 \mid X) \geq 0.5$이면 1로, 미만이면 0으로 분류하는 방식이다(그림 2).

또한, 로지스틱 회귀는 손실 함수(Loss function)로 로그 우도 함수(Negative log-likelihood)를 기반으로 한 교차 엔트로피 손실(Cross-Entropy loss)을 사용하며, 경사 하강법(Gradient descent) 또는 보다 최적화된 변형 기법(Nelder-Mead, L-BFGS 등)을 이용하여 최적의 가중치를 학습한다.

본 연구에서는 로지스틱 회귀를 메타 학습기로 활용하여, 개별 모델(SVM, Gradient Boosting, XGBoost)의 출력을 결합하고 최적의 분류 경계를 학습하는 역할을 수행하도록 설계하였다.

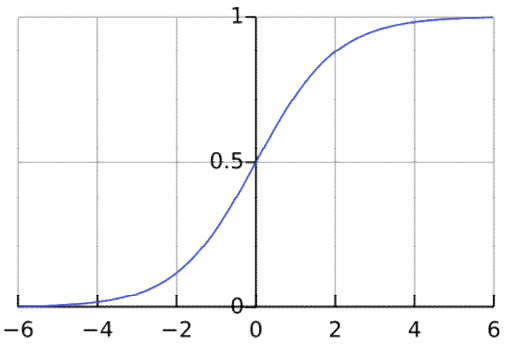

그림 2. 로지스틱 회귀 모델
Fig. 2. Logistic regression

3.3.2 그래디언트 부스팅

그래디언트 부스팅은 여러 개의 약한 학습기(Weak learner)를 결합하여 강력한 예측 모델을 구축하는 부스팅 기법 중 하나이다. 랜덤 포레스트와 같은 배깅(Bagging) 방식과는 달리, 그래디언트 부스팅은 개별 모델을 순차적으로 학습시키며, 각 단계에서 이전 모델의 오류를 보완하는 방식으로 성능을 향상시킨다(그림 3).

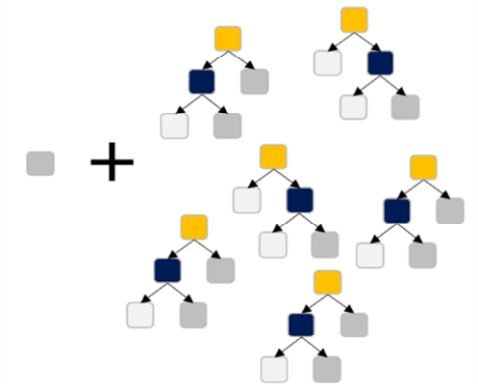

그림 3. 그래디언트 부스팅
Fig. 3. Gradient boosting

본 연구에서 활용한 그래디언트 부스팅 모델은 결정 트리(Decision tree)를 약한 학습기로 사용하며, 새로운 트리를 추가할 때마다 기존 모델이 가진 예측 오류를 줄이는 방향으로 학습을 진행한다. 먼저, 첫 번째 약한 학습기를 훈련하여 초기 예측 모델을 생성한 후, 예측값과 실제값 사이의 오차를 측정하기 위한 손실 함수를 정의한다. 일반적으로 회귀 문제에서는 평균제곱오차(MSE, Mean Squared Error), 분류 문제에서는 로그 손실(Log loss)이 손실 함수로 사용된다. 이후 현재 모델이 예측한 값과 실제값의 차이를 계산하여 잔차(Residual)를 도출하고, 이 잔차를 최소화할 수 있도록 새로운 결정 트리를 학습시켜 기존 모델을 보완한다. 이렇게 생성된 트리는 기존 모델과 결합되어 업데이트되며, 이때 학습률(Learning rate)을 적용하여 모델이 과적합 되지 않도록 조정한다. 이러한 과정을 반복적으로 수행함으로써 점진적으로 예측 성능을 개선할 수 있다. 이와 같은 학습 방식은 개별 모델이 독립적으로 학습하는 배깅 기법과 달리, 이전 모델의 오류를 지속적으로 보완하며 최적의 예측 모델을 구축하는 데 효과적이다. 특히, 그래디언트 부스팅은 각 트리가 예측 오차의 기울기 방향으로 학습을 진행한다는 점에서 기존 부스팅 기법보다 더 효과적인 성능 개선이 가능하다.

본 연구에서는 그래디언트 부스팅을 활용하여 LncRNA와 질병 간의 연관성을 예측하는 데 적용하였으며, 메타 학습기와의 조합을 통해 모델의 성능을 극대화하는 방안을 모색하였다.

3.3.3 익스트림 그래디언트 부스팅

익스트림 그래디언트 부스팅(XGBoost)은 기존 그래디언트 부스팅의 한계를 보완하고 보다 효율적인 학습을 가능하게 하기 위해 개발된 알고리즘이다. 그래디언트 부스팅 기법은 높은 예측 성능을 제공하지만, 연산 비용이 크고 학습 속도가 느리다는 단점이 존재한다. 이러한 문제를 해결하기 위해 XGBoost는 알고리즘의 효율성, 유연성, 정확성을 극대화하는 다양한 최적화 기법을 도입하였다.

XGBoost의 주요 특징 중 하나는 연산 속도의 향상이다. CPU 및 GPU 병렬 연산을 적극 활용하여 학습 속도를 증가시키고, 대규모 데이터셋에서도 효율적으로 학습할 수 있도록 설계되었다. 또한, 압축된 데이터 저장 방식과 메모리 캐싱(Memory caching) 기법을 적용하여 메모리 사용량을 줄이는 동시에 데이터 접근 속도를 개선하였다. 모델의 일반화 성능을 높이기 위해 XGBoost는 L1 및 L2 정규화 항을 포함한 최적화 기법을 활용하여 과적합을 방지하고 보다 강건한(robust) 학습을 유도한다. 이와 더불어, 손실 함수의 유연성을 확대하여 회귀(Regression), 분류(Classification), 랭킹(Ranking) 등 다양한 머신러닝 태스크에 적용할 수 있도록 설계되었다.

이러한 다양한 최적화 기법 덕분에 XGBoost는 기존의 그래디언트 부스팅 알고리즘 대비 학습 속도가 빠르고 계산 성능이 뛰어나며, 일반화 성능이 우수한 모델로 평가된다. 특히, 대규모 데이터셋을 다루는 실무 환경에서 강력한 성능을 발휘하며, 금융, 의료, 추천 시스템, 자연어 처리(NLP) 등 다양한 분야에서 널리 활용되고 있다.

3.4 앙상블 기반 예측 모델링

앙상블 학습은 여러 개의 개별 학습 모델을 결합하여 성능을 향상시키고, 보다 강건한 예측을 가능하게 하는 기법이다. 단일 모델의 한계를 극복하고 일반화 성능을 개선하기 위해 다양한 앙상블 기법이 제안되었으며, 본 연구에서는 Bagging, Stacking, Voting을 활용하여 최적의 예측 모델을 구축하였다.

Bagging(Bootstrap aggregating)은 부트스트랩 샘플링(Bootstrap sampling)을 통해 다수의 학습 데이터를 생성하고, 개별 모델을 독립적으로 학습시킨 후 평균화 또는 투표 방식을 통해 최종 예측을 수행하는 기법이다. Stacking(Stacked generalization)은 서로 다른 유형의 학습 모델을 결합한 후, 최종적으로 메타 학습기를 적용하여 최적의 조합을 학습하는 방법이다. 또한, Voting은 여러 개의 분류 모델에서 도출된 예측 결과를 확률 기반으로 평균 내어 최종 결정을 수행하는 방식으로, 다수결(Hard voting) 또는 가중 확률 평균(Soft voting)을 적용할 수 있다.

본 논문에서는 이러한 앙상블 학습 기법을 조합하여 최종 모델을 구성하였으며, 다양한 모델 간의 상호 보완적 특성을 활용하여 예측 성능을 극대화하고자 하였다.

3.4.1 Bagging: 부트스트랩 샘플링 학습

배깅(Bagging, Bootstrap Aggregating)은 원본 데이터에서 복원 추출(Bootstrap sampling)을 통해 여러 개의 부분 집합을 무작위로 생성한 후, 각 부분 집합을 개별 학습 모델에 독립적으로 학습시키는 기법이다(그림 4).

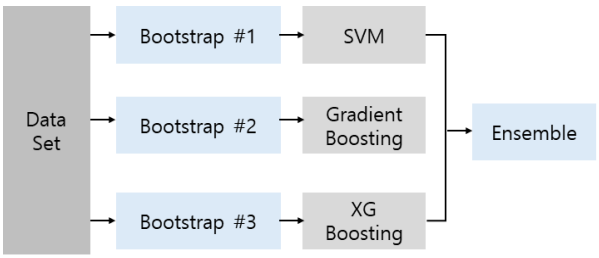

그림 4. 배깅 흐름도
Fig. 4. Bagging flowchart

$$D_b \sim Bootstrap\,Sample\,from\,D,\ b = 1,2,...B \quad (2)$$

수학적으로, 원본 데이터셋에서 b개의 부트스트랩 샘플을 생성하며(식 (2)), 각 샘플은 중복을 허용하여 원본 데이터셋에서 선택된다. 이 과정을 통해 각 개별 모델은 서로 다른 데이터 조합을 학습하게 되며, 이를 통해 편향(Bias)을 유지하면서 분산(Variance)을 감소시키는 효과를 얻을 수 있다. 또한, 부트스트랩 샘플링은 모델이 특정 데이터에 과적합되는 현상을 완화시키는 동시에, 일반화 성능을 향상시키는 역할을 한다. 이 기법은 특히 불균형 데이터셋에서 예측 성능을 개선하고, 다양한 모델이 상호 보완적으로 작용할 수 있도록 한다.

3.4.2 Stacking: 메타 학습기(XGBClassifier)

Stacking은 개별 학습 모델의 예측 결과를 새로운 특징(Feature)으로 활용하여 메타 모델(Meta-learner)을 학습하는 기법이다(그림 5). 즉, 각 기본 학습기(Base learner)가 도출한 예측값을 입력으로 하는 새로운 데이터셋을 구성하며(식 (3)), 이를 기반으로 최종 예측을 수행하는 메타 학습기(XGBClassifier)를 훈련한다. 이러한 구조를 통해 개별 모델의 강점을 결합하고 약점을 보완하여 보다 정교한 예측 성능을 도출할 수 있다.

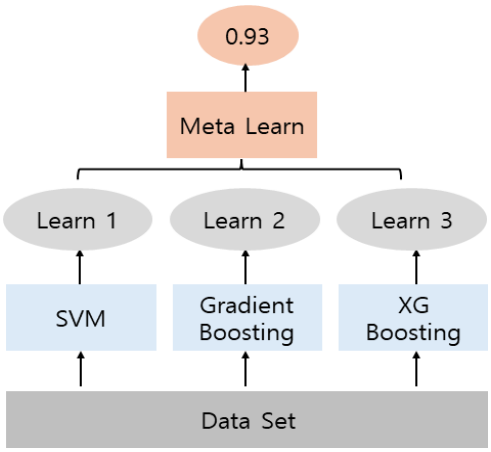

그림 5. 메타학습기 흐름도
Fig. 5. Meta-learning machine flowchart

$$Z = (\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}) | i = 1,...,N \quad (3)$$

3.4.3 Voting: 각 확률 평균으로 결합

Voting은 개별 모델의 예측 결과를 결합하여 최종 예측을 도출하는 기법으로, 단순 평균, 가중 평균 또는 모델의 신뢰도를 반영한 가중 평균 방식이 적용될 수 있다 (그림 6). 특히, 가중 평균 방식에서는 각 모델의 신뢰도에 따라 가중치 Wk를 부여하며, 이를 통해 보다 정교한 예측 성능을 얻을 수 있다(식 (4)). 이러한 방식은 각 모델의 성능 차이를 반영하여 예측의 정확도를 향상시킬 수 있으며, 모델 간의 상호 보완적인 특성을 활용하여 최종 예측의 신뢰성을 증가시킨다.

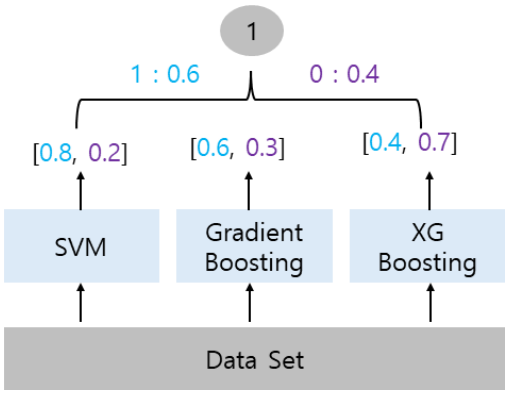

그림 6. 확률 기반 결합
Fig. 6. Probability-based integration

$$\hat{y}_i^{(ensemble)} = \frac{1}{K}\sum_{k=1}^{K}\hat{y}_i^{(k)} \qquad (4)$$
$$\hat{y}_i^{(ensemble)} = \sum_{k=1}^{K} w_k \hat{y}_i^{(k)}$$

3.4.4 앙상블 학습 기대효과

앙상블 학습은 성능 및 예측 측면에서 우수한 이점을 제공하며, 특히 과적합을 방지하고 모델의 일반화 성능을 향상시키는 데 효과적이다. Bagging 기법을 활용하면 데이터의 다양성이 증가하여 보다 균형 잡힌 모델 학습이 가능하며, 강건한 예측 성능을 보장할 수 있다. 또한, 서로 다른 모델을 조합함으로써 개별 모델이 가진 단점을 보완할 수 있으며, 이를 통해 전반적인 예측 정확도를 개선할 수 있다.

Stacking 기법을 적용하면 개별 모델이 도출한 결과를 종합하여 최적의 예측 성능을 도출할 수 있으며, Voting 기법을 활용하면 개별 모델의 편향된 결과를 완화하여 보다 신뢰성 높은 예측이 가능하다. 더불어, 앙상블 기법은 노이즈를 효과적으로 감소시켜 모델의 안정성을 높이는 데 기여한다.

Ⅳ. 실험결과

4.1 교차 검증 실행

본 논문에서는 제안된 모델의 예측 성능을 평가하기 위해 Stratified K-Fold Cross-Validation 기법을 적용하였다. K-Fold 교차 검증(K-Fold CV)은 데이터를 K개의 서브셋(폴드)으로 분할한 후, K−1개의 폴드를 학습에 사용하고, 나머지 1개의 폴드를 테스트에 활용하는 성능 평가 방법이다. 이 기법은 특히 K=5일 경우, Leave-One-Out Cross Validation (LOOCV)과 비교하여 연산 비용을 크게 절감할 수 있다. 또한, Stratified K-Fold는 클래스 불균형 문제를 고려하여 각 폴드에 클래스 비율을 동일하게 유지함으로써, 모델의 일반화 성능을 보다 정확하게 평가할 수 있다.

특히, 신뢰성 있는 성능 평가를 수행하기 위해, 각 폴드가 클래스의 비율을 유지하도록 설계된 Stratified K-Fold 방식을 적용하였다. 즉, 데이터를 K개의 폴드로 분할한 후, 각 폴드 내에서 양성(1)과 음성(0)의 비율을 유지하도록 샘플링을 진행하였으며, 이를 통해 각 폴드에서 모델을 학습 및 평가한 후 평균 AUC 성능을 도출하였다. 이러한 결과는 그림 7에서 확인할 수 있다.

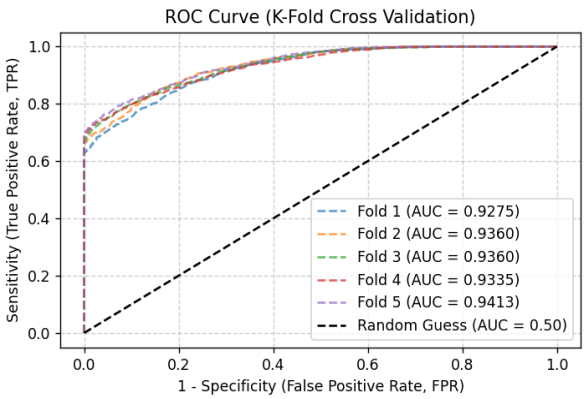

그림 7. K-Fold 교차 검증
Fig. 7. Stratified K-Fold cross-validation

4.2 비교 실험 실행

본 논문에서는 K-Fold 교차 검증을 적용하여 기존의 기계 학습 기반 예측 모델들과 성능을 비교하였다 [17][18]. LncRNA와 질병 간 연관성 예측 성능을 평가하기 위해, AUC(Area Under the Receiver Operating Characteristic Curve) 값을 계산하고, 이를 기존 예측 모델들과 비교하는 실험을 진행하였다(표 1). 이를 통해 제안된 모델의 성능 우수성을 입증하였다.

표 1. 비교 실험 결과
Table 1. Comparative experiments results

Model	Inputs	Database	Methods	Performance (AUC)
MFLDA	LDA	LncRNADisease	k-fold	0.8315
WGRCMF	LDA	LncRNADisease	k-fold	0.8929
Ensemble	LDA	LncRNADisease	k-fold	0.9364

ROC(Receiver Operating Charateristic)는 분류 모델에서 성능을 표현하는 지표로서, AUC 값은 ROC 곡선 아래 영역을 의미한다. 그림 8에서 볼 수 있듯이 본 논문에서 제시하는 앙상블 모델이 AUC = 0.9364 값을 도출하면서 기존 단일 모델인 MFLDA (0.83125), WGRCMF(0.8929) 보다 향상된 성능을 보였다.

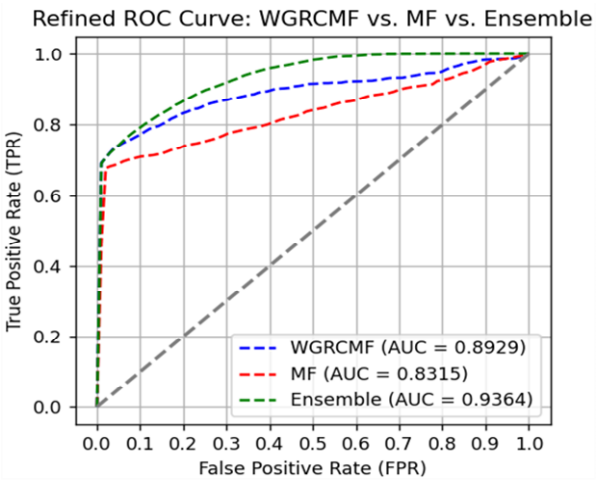

그림 8. ROC 곡선 기반 성능 비교 평가
Fig. 8. ROC curve based performance comparison evaluation

다만, 샘플링 방법에 따라 AUC 값은 일정 수준의 변동이 발생할 수 있다. 또한 제안된 예측 모델의 성능을 보다 심층적으로 검증하기 위하여, 다양한 개별 모델 앙상블 조합을 활용한 비교 실험을 수행하였다(표 2).

표 2. 실험 데이터 세트
Table 2. Experimental data sets

Case	Inputs	Model	Methods	Performance (AUC)
Step 1	LDA	LR+RF+GB	k-fold	0.8335
Step 2	LDA	SVM+KNN+GB	k-fold	0.8879
Step 3	LDA	SVM+GB+XGB	k-fold	0.9364

실험은 총 3단계로 진행되었으며, 각 단계에서 서로 다른 모델 조합을 구성하여 성능을 비교하고 분석하였다. 1단계에서는 로지스틱 회귀, 랜덤 포레스트, 그래디언트 부스팅 모델을 조합하였으며, 2단계에서는 서포트 벡터 머신(SVM, Support Vector Machine), K-최근접 이웃(KNN, K-Nearest Neighbors), 다층 퍼셉트론(MLP, Multi-Layer Perceptron)을 활용하였다. 본 연구에서 제안하는 3단계 모델은 서포트 벡터 머신, 그래디언트 부스팅, 익스트림 그래디언트 부스팅 모델로 구성되었다. 실험 결과, 제안된 3단계 앙상블 모델이 1단계 및 2단계 모델 조합보다 우수한 성능을 보였으며, 기존 모델들보다 예측력이 향상되었음을 확인할 수 있었다. 또한, 본 연구에서는 거리 기반 음성 샘플링 전략을 도입하였다. 이는 모든 유전자와 질병 쌍 간의 유클리드 거리를 계산하고, 양성 샘플들 간 거리의 평균값을 기준 임계값으로 설정하였다. 이후, 이 평균 거리보다 충분히 멀리 떨어진 유전자와 질병 쌍만을 신뢰할 수 있는 음성 샘플로 선정하였다. 이는 양성 샘플과의 구조적 또는 기능적 유사성이 낮은 쌍을 중심으로 음성 데이터를 구성하였다. 최종적으로 정제된 음성 샘플과 양성 샘플을 결합하여 이진 분류 모델 학습을 위한 실험 데이터를 구성하여 앙상블 모델 학습 시 음성 레이블로 인한 오분류를 최소화하고 성능을 향상하고자 하였다.

또한, 제안한 모델의 실효성을 검증하기 위해 유방암(Breast neoplasms)을 대상으로 사례 기반 분석(Case study)을 수행하였다.

이를 위해 문헌 및 생물학적 데이터베이스를 참고하여 유방암과 연관된 LncRNA 유전자를 선정하고, 유클리드 거리 기반 필터링 기법을 적용하여 이들과의 유사성이 낮은 음성 샘플을 구성함으로써 테스트 세트를 구축하였다. 양성 및 음성 샘플을 통합한 데이터셋을 기반으로 각 유전자에 대한 질병 관련성 점수를 예측하고, 해당 점수가 사전 정의한 임계값을 초과하는 경우 관련성이 있는 것으로 판단하였다. 이후, 이 예측 결과를 실제 레이블과 비교하여 모델의 성능을 평가하였다.

그 결과는 표 3에 제시되어 있으며, 본 연구에서 제안한 모델은 AUC(Area Under the Curve), 재현율(Recall), F1-score 등의 주요 성능 지표에서 전반적으로 우수한 결과를 나타냈다. 이러한 결과는 본 연구의 접근 방식이 실제 바이오마커 탐색 및 질병 연관성 예측 과제에 효과적으로 활용될 수 있음을 시사한다.

표 3. 유방암 관련 유전자 샘플링 실험 결과
Table 3. Breast Neoplasms related LncRNA sampling experiments results

Inpusts	Sampling of 10 LDA Cases
Disease	Breast Neoplasms
AUC	0.9375
Recall	0.75
F1-score	0.8571

Ⅴ. 결론 및 향후 과제

LncRNA는 다양한 인간 질병의 발병 및 진행에 중요한 역할을 하는 비암호화 RNA 유전자로서, 이들의 기능과 질병 연관성을 규명하는 연구가 활발히 진행되고 있다. 그러나 기존 LncRNA-질병 연관성(LDA) 예측 모델들은 주로 단일 모델에 의존하여 예측 성능 향상에 한계를 보여왔다. 본 논문에서는 이러한 한계를 극복하고자, SVM, GBM, XGBoost 모델을 결합한 특화된 앙상블 기계학습 기반 LDA 예측 모델을 제안하였다. 제안 모델은 LncRNA와 질병 데이터를 효과적으로 표현하기 위해 특이값 분해(SVD) 기법을 활용하여 특징 벡터를 추출하였고, 거리 기반 음성 샘플링 전략을 통해 학습 데이터의 품질을 개선하였다.

실험 결과, 제안된 앙상블 모델은 AUC 0.93이라는 우수한 예측 성능을 기록하며, 비교 대상이었던 기존 단일 모델 기반 접근법들보다 유의미하게 향상된 결과를 보여주었다. 이러한 높은 성능은 다음의 요인들이 복합적으로 작용한 결과로 해석된다. 첫째, 각기 다른 강점을 가진 SVM, GBM, XGBoost 모델을 앙상블로 결합함으로써 개별 모델의 약점은 보완하고 예측의 정확성과 안정성을 극대화하였다. 다양한 모델 조합에 대한 비교 실험(표 2)을 통해 본 연구에서 최종 선택한 조합의 우수성을 확인하였다. 둘째, SVD를 통해 고차원 희소 데이터를 저차원의 의미 있는 특징 공간으로 효과적으로 변환함으로써, 모델이 LncRNA와 질병 간의 잠재적 연관성을 보다 용이하게 학습할 수 있도록 하였다. 셋째, 신중하게 설계된 거리 기반 음성 샘플링 전략은 레이블 노이즈 문제를 완화하여, 모델이 보다 명확한 데이터를 기반으로 학습함으로써 예측 성능과 강건성을 높이는 데 기여하였다.

결론적으로, 본 연구는 강력한 개별 모델들을 효과적으로 조합한 앙상블 구조를 채택하고, SVD 및 특화된 샘플링 기법을 통해 데이터 신뢰도를 높임으로써, 기존 LDA 예측 연구의 한계를 극복하고 LncRNA와 질병 간의 연관성을 보다 정확하고 신뢰성 높게 예측하는 방법론을 성공적으로 제시하였다. 이렇게 제안된 모델은 수많은 LncRNA 후보 중에서 특정 질병과 연관될 가능성이 높은 후보군을 효과적으로 우선순위화(Prioritization) 할 수 있다. 이는 시간과 비용이 많이 소요되는 생물학적 실험의 대상을 좁혀, 질병 관련 LncRNA 발굴 및 기능 연구의 효율성을 크게 높이는 가이드 역할을 할 수 있다.

향후 연구에서는 개발된 모델의 일반화 성능을 검증하기 위해 더 다양한 종류의 LncRNA-질병 연관성 데이터셋에 적용해 볼 계획이다. 또한, 하이퍼파라미터 최적화(Hyperparameter optimization) 과정을 통해 모델 성능을 추가적으로 개선하고, 최근 주목받고 있는 그래프 신경망(GNN, Graph Neural Network)과 같은 딥러닝 모델을 앙상블에 통합하여 예측 성능을 더욱 향상시키는 방안을 모색할 것이다. 구체적으로 LncRNA 간의 유사성, 질병 간의 유사성, 그리고 알려진 LncRNA-질병 상호작용 정보를

하나의 이종 정보 네트워크(Heterogeneous information network)로 모델링하고, 그래프 신경망을 적용하여 LncRNA와 질병 노드의 풍부한 토폴로지(Topology) 정보를 임베딩 벡터로 학습할 계획이다. 이렇게 GNN을 통해 생성된 고차원 특징 벡터를 기존 앙상블 모델의 추가 입력 특징으로 사용하거나, GNN 모델 자체를 앙상블의 새로운 구성원으로 통합하는 방식을 탐구하여 예측 성능의 향상을 시도할 것이다. 이러한 후속 연구를 통해 인공지능 및 생물정보학 기술을 활용한 질병 관련 유전자 기능 예측 연구의 발전에 기여할 수 있을 것으로 기대한다.

References

[1] Anastasiadou, E.; Jacob, L. S.; Slack, F. J. Non-Coding RNA Networks in Cancer. Nat. Rev. Cancer 2018, 18(1), 5.

[2] J. Ha, C. Park, C. Park, and S. Park, "IMIPMF: Inferring miRNA-disease probabilistic interactions using matrix factorization", Journal of Biomedical Informatics, Vol. 102, pp. 103358, Feb. 2020. https://doi.org/10.1016/j.jbi.2019.103358.

[3] 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.

[4] 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.

[5] 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/1109/ACCESS.2021.3084148.

[6] 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. https://doi.org/10.3390/jpm12060885.

[7] 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. 2022. https://doi.org/10.1109/TCBB.2022.3191972.

[8] J. Ha, "SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association", Knowledge-Based Systems, Vol. 263, pp. 110295, Mar. 2023. https://doi.org/10.1016/j.knosys.2023.110295.

[9] 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.

[10] 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.

[11] J. Ha, "DeepWalk-Based Graph Embeddings for miRNA–Disease Association Prediction Using Deep Neural Network", Biomedicines, Vol. 13, No. 3, pp. 536, Mar. 2025. https://doi.org/10.3390/biomedicines13030536.

[12] S. Kim and J. Ha, "Development of LncRNA-Disease Associations Prediction Model using CNN-RNN Feature Extraction based on LncRNA Sequence Data", JKIIT, Vol. 22, No. 10, pp. 1-11, Oct. 2024. https://doi.org/10.14801/jkiit.2024.22.10.1.

[13] J.-H. Ha, "Autoencoder-based Disease-related miRNA Prediction Research using Deep Learning", JKIIT, Vol. 20, No. 6, pp. 33-40, Jun. 2022. https://doi.org/10.14801/jkiit.2022.20.6.33.

[14] 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, 2025.

[15] K. Kim and J. Ha, "GMFLDA: Improved prediction of lncRNA-disease association via graph convolutional network", IEEE Access, Vol 13, pp. 85330-85341, 2025.

[16] L. J. Core, J. J. Waterfall, and J. T. Lis, "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters", Science, Vol. 322, No. 5909, pp. 1845-1848, Dec. 2008. https://doi.org/10.1126/science.1162228.

[17] M. Esteller, "Non-coding RNAs in human disease", Nature Reviews Genetics, Vol. 12, No. 12, pp. 861-874, Dec. 2011. https://doi.org/10.1038/nrg3074.

[18] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions", Nature Reviews Genetics, Vol. 10, No. 3, pp. 155-159, Mar. 2009. https://doi.org/10.1038/nrg2521.

[19] M.-C. Tsai, et al., "Long noncoding RNA as modular scaffold of histone modification complexes", Science, Vol. 329, No. 5992, pp. 689-693, Jul. 2010. https://doi.org/10.1126/science.1192002.

[20] P. Ping, L. Wang, L. Kuang, S. Ye, M. F. B. Iqbal, and T. Pei, "A Novel Method for LncRNA-Disease Association Prediction Based on an LncRNA-Disease Association Network", IEEE/ACM Trans. Comput. Biol. and Bioinf., Vol. 16, No. 2, pp. 688-693, Mar. 2018. https://doi.org/10.1109/TCBB.2018.2827373.

[21] J. Zhang, Z. Zhang, Z. Chen, and L. Deng, "Integrating Multiple Heterogeneous Networks for Novel LncRNA-Disease Association Inference", IEEE/ACM Trans. Comput. Biol. and Bioinf., Vol. 16, No. 2, pp. 396-406, Mar. 2019. https://doi.org/10.1109/TCBB.2017.2701379.

[22] 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.

[23] 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.

[24] J. X. Liu, Z. Cui, Y. L. Gao, and X. Z. Kong, "WGRCMF: A weighted graph regularized collaborative matrix factorization method for predicting novel LncRNA-disease associations", IEEE J. Biomed. Health Inform., Vol. 25, No. 1, pp. 257-265, Jan. 2021. https://doi.org/10.1109/jbhi.2020.2985703.

[25] 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.

[26] X. Chen, M. X. Liu, and G. Y. Yan, "RWRMDA: predicting novel human microRNA disease associations", Molecular BioSystems, Vol. 8, No. 10, pp. 2792-2798, Oct. 2012. https://doi.org/10.1039/c2mb25180a.

[27] T. G. Dietterich, "Ensemble Methods in Machine Learning", Multiple Classifier Systems, Vol. 1857, pp. 1-15, 2000. https://doi.org/10.1007/3-540-45014-9_1.

[28] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble Selection from Libraries of Models", Proc. of the Twenty-First International Conference on Machine Learning (ICML '04), Banff, Alberta, Canada, pp. 137-144, Jul. 2004. https://doi.org/10.1145/1015330.1015432.

[29] L. Xiao, et al., "LncRNADisease v3.0: an updated database of long non-coding RNA-associated diseases", Nucleic Acids Res., Vol. 52, No. D1, pp. D1365-D1369, Jan. 2024. https://doi.org/10.1093/nar/gkad828.

[30] J. Tan, X. Li, L. Zhang, and Z. Du, "Recent

advances in machine learning methods for predicting LncRNA and disease associations", Front. Cell. Infect. Microbiol., Vol. 12, No. 1071972, Nov. 2022. https://doi.org/10.3389/fcimb.2022.1071972.

저자소개

김 보 훈 (Bohoon Kim)

2013년 8월 : 부산대학교 바이오정보전자학과(공학사)
2024년 3월 ~ 현재 : 부경대학교 빅데이터융합전공(공학석사)
2013년 7월 ~ 현재 : ㈜두산 디지털이노베이션 수석
관심분야 : 기계학습, ERP, PLM, CRM 등 기간계 시스템 통합

김 상 욱 (Sanguk Kim)

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

하 지 환 (Jihwan Ha)

2020년 8월 : 연세대학교 컴퓨터과학과(공학박사)
2020년 9월 ~ 2021년 8월 : 하와이 암센터 포스닥 연구원
2021년 9월 ~ 현재 : 부경대학교 데이터정보과학부 빅데이터융합전공 조교수
관심분야 : 기계학습, 생물정보학, 데이터마이닝, 딥러닝, 추천시스템