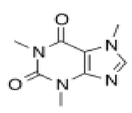
2019 한국 데이터정보과학회 춘계 학술대회

신약 개발을 위한 불균형 데이터 처리와 투과성 예측

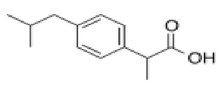
단국대학교 데이터지식서비스공학과/ 이원욱 김민기 황창하 아론티어/고준수
삼성의료원 의생명정보센터 /손인석
인제대학교 통계학과/심주용

Introduction

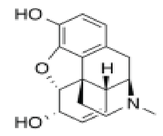
- **BBB**란 Blood-Brain-Barrier의 약자로 뇌 조직을 혈류로부터 분리하여 중추신경계를 보호하는 역할을 한다.
- 보호의 역할로 만들어진 BBB지만, 뇌 질환 치료를 위해 투입한 약물을 뇌에 전달하지 못하게 함으로써, 뇌 질환을 치료 못하게 하는 문제를 발생한다. 따라서 **BBB를 투과(Permeability)할 수 있는 약물을 개발하는 것이 중요하다.**
- 데이터는 약물 구조의 1D 형태인 **'SMILES'** 표기법을 사용한다.



Caffeine
CN1c2ncn(C)c2C(=O)N(C)C1=O



Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(O)=O

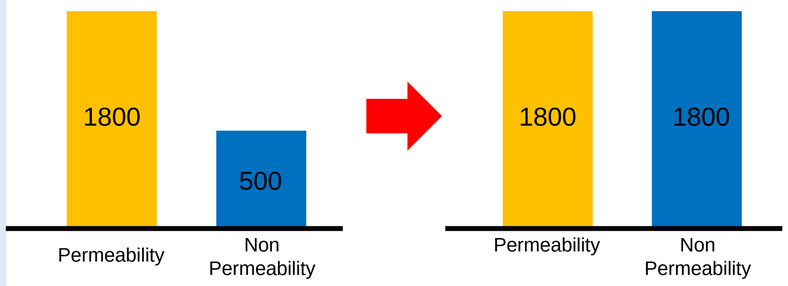


Morphine
[H][C@]12C=C[C@H](O)[C@@H]3Oc4c5c(C[C@H]1N(C)CC[C@@]235)ccc4O

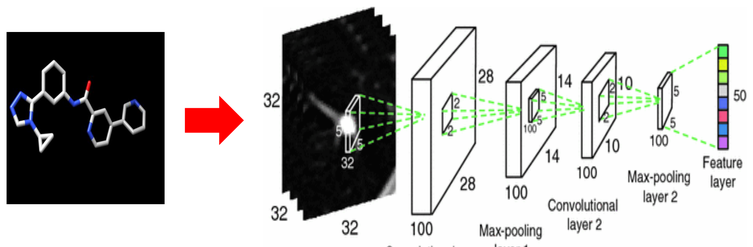
*1차원 형식의 SMILES 표기
(SMILES : Simplified molecular-Input line-entry system)

Objectives

- 최근 몇 년 동안 신약개발과정에서 BBB 투과성이 중요한 문제로 제기되었지만, 선행 연구들은 불균형 데이터를 사용하여 각 Fold 별 AUC 값의 편차가 크거나 값 자체가 낮은 문제들이 있다.
- 또한, 선행연구의 BBB 투과성 예측 모델에는 약리적, 화학적 고려해야 할 사항이 많고 복잡하여 전문지식이 없을 경우 모델을 구축하는 것에 어려움이 있다.
- 따라서 논문의 **1번째 목표는** 생성 딥러닝 모델을 통해 불균형 데이터를 균형데이터로 변환하는 것이다.



- 논문의 **2번째 목표는** 약리적, 화학적 지식이 필요없이, 분자의 2D 그림만으로 BBB 투과성 예측 모델을 만드는 것이다.



Method1

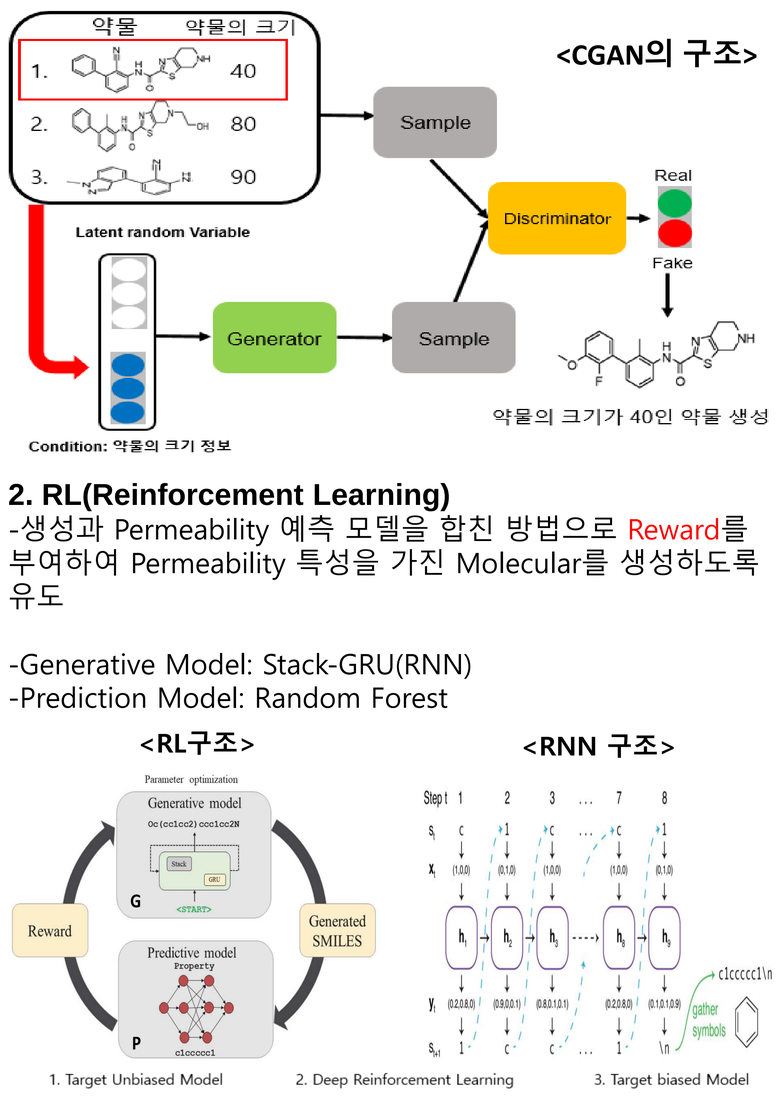
*** 상대적으로 적은 그룹의 데이터를 2가지 생성 딥러닝 모델을 통해 생성하여 불균형 데이터 처리**

1. cGAN(Conditional Generative Adversarial Network)

-GAN에서 Condition을 추가한 모델로 생성시 발생하는 문제를 해결하기 위한 생성 모델.

-538개 Permeability 데이터를 한번에 집어 넣을 경우 약물들의 크기 편차가 커서 생성을 잘하지 못하는 문제가 발생한다.

-따라서 약물의 크기를 Condition으로 주어 비슷한 크기를 가진 약물 데이터를 각각 생성 (5구간)



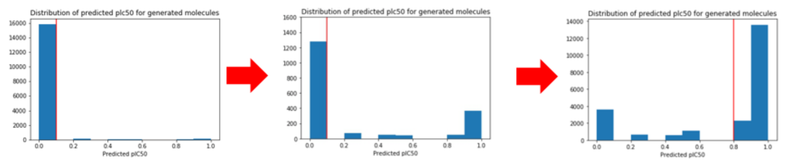
2. RL(Reinforcement Learning)
-생성과 Permeability 예측 모델을 합친 방법으로 Reward를 부여하여 Permeability 특성을 가진 Molecular를 생성하도록 유도

-Generative Model: Stack-GRU(RNN)
-Prediction Model: Random Forest

<RL구조>

<RNN 구조>

1. Target Unbiased Model
2. Deep Reinforcement Learning
3. Target biased Model



Mean value of predictions: 0.0160
Proportion of valid SMILES: 81.29 %
Proportion > threshold: 2.79 %

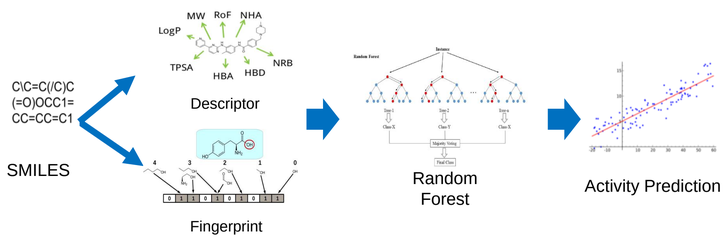
Invalid SMILES Reward: -2
Active SMILES Reward: +1
Inactive SMILES Reward: -1

Mean value of predictions: 0.082
Proportion of valid SMILES: 47.08%
Proportion > threshold: 79.39%

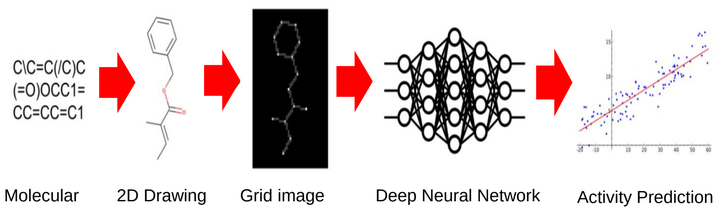
<RL Cycle>

Method2

<Random Forest step >



<CNN step>



Molecular 2D Drawing Grid image Deep Neural Network Activity Prediction

<모델 특징 비교>

약리적, 화학적 지식 필요

QSAR Model	Descriptor 사용	Fingerprint 사용
Random Forest	O	O
CNN	X	X

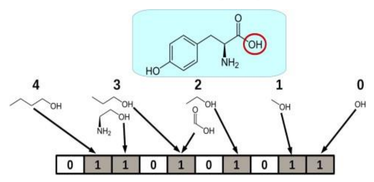
Result1

- 생성한 데이터들 중에도 적은 그룹의 데이터와 구조적으로 유사성이 떨어지는 데이터가 생성될 수 있으며, 이는 모델 성능 저하의 원인이 된다.
- 따라서 약물을 MACCS Fingerprint로 변환하고 Tanimoto Similarity를 0.85 이상으로 잘라서 적은 그룹의 데이터와 구조적으로 유사한 데이터만 남기도록 하였다.
- cGAN의 경우 Similarity가 높은 약물이 생성되는 장점이 있지만, 생성 시간이 오래 걸리는 단점이 있었다.
- 반면 Reinforcement Learning은 생성 시간이 짧은 장점이 있지만, Similarity가 높은 약물을 거의 생성하지 못하였다.

1.Tanimoto Similarity

$$T = \frac{N_{a \cap b}}{N_a + N_b - N_{a \cap b}}$$

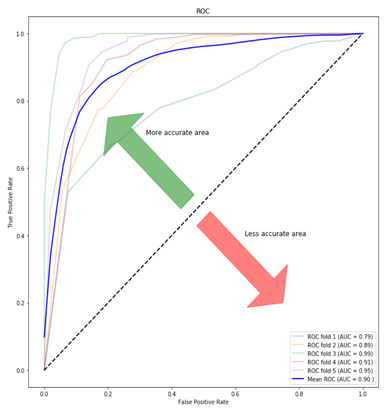
2.MACCS Fingerprint



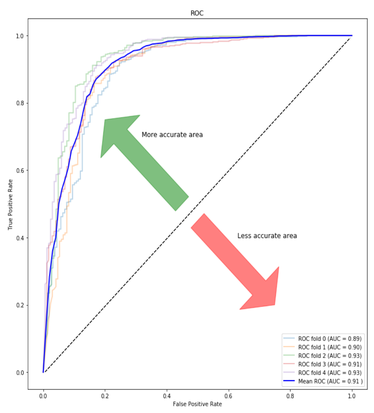
Generative Model	생성 개수	Tanimoto Similarity 0.85 이상의 개수	비율
cGAN	958개	421개	43.94%
RL	600,000개	847개	0.14%

Result2

1. 불균형 데이터 처리 전

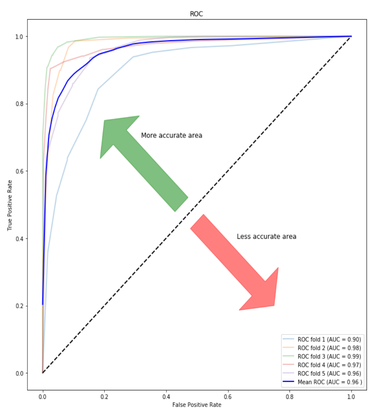


<Random Forest>

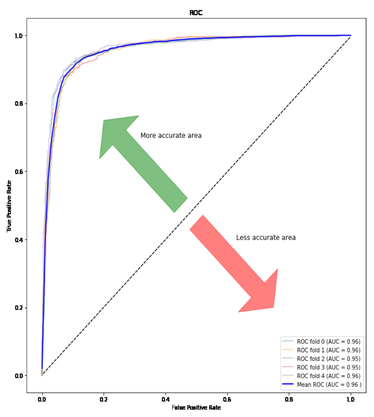


<CNN>

2. 불균형 데이터 처리 후



<Random Forest>



<CNN>

QSAR Model	1 Fold AUC	2 Fold AUC	3 Fold AUC	4 Fold AUC	5 Fold AUC	Mean AUC
Balance CNN	0.96	0.96	0.95	0.95	0.96	0.96
Balance RF	0.90	0.98	0.99	0.97	0.96	0.96
Unbalance CNN	0.89	0.90	0.93	0.91	0.93	0.91
Unbalance RF	0.79	0.89	0.99	0.91	0.95	0.90

<4가지 모델의 AUC 성능 비교 >

-Random Forest의 경우 CNN 보다 각 Fold 별 AUC의 편차가 커서 Overfitting의 가능성이 보였다.

-불균형 데이터를 처리한 후 CNN의 AUC 값이 가장 높고, 편차 또한 가장 적은 것을 확인할 수 있었다.

Conclusion

- 전문 지식이 필요없이 오직 그림으로만 투과성을 예측하는 CNN 모델이 RF모델보다 좋은 성능을 보이는 것을 확인할 수 있었다.
- 불균형 데이터 처리 전에는 AUC 값이 낮을 뿐만 아니라 각 Fold 별로 AUC 편차가 더 큰 것을 확인할 수 있었다.
- 불균형 데이터 처리 후에는 AUC 값이 높아지며, 각 Fold 별 AUC의 편차가 줄어드는 것을 확인할 수 있었다.
- CNN의 Architecture는 Incepction을 사용하였지만, ResNet을 사용할 경우 더욱 높은 성능을 낼 수 있을 것으로 기대된다.