|  |  |
| --- | --- |
| Research title | 주요 독성 데이터를 활용한 Binary Classification 모델 구축 및 평가 |
| Name | 고강빈 |

**Abstract**

**Background**

심독성 등 치명적인 독성을 가진 약물이 시장에 공급될 경우, 심각한 결과를 초래할 수 있으며, 신약의 임상실험 단계에서 발견될 시 약품이 전량 폐기 될 수도 있기에, 독성여부는 약물 연구에서 주요하게 다루는 요인 중 하나이다. 따라서 초기에 화합물의 구조로부터 독성을 알아낼 수 있다면 이러한 위험성을 줄임과 동시에 신약 개발의 성공 확률을 높일 수 있을 것이다.

이 글에서는 여러 기계학습 모델을 활용하여 DILI(간독성) 외 3종의 주요 독성의 화학구조로부터 독성을 예측하고, 이를 평가하여 더 나은 예측 방식을 찾아보고자 한다.

학습데이터에는 각 독성과 연관된 화합물에 대하여, 화학구조(Canonical SMILES)와 독성 Label을 사용하게 될 것이며, 이후 새로운 화학 구조로부터 각 독성에 대한 레이블을 예측하게 될 것이며, 해당 모델에 대한 평가는 Sensitivity, Specificity, Balanced Accuracy 총 세 종류의 평가지표로 진행될 것이다.

**Results**

**Conclusions**

**Methods**

Dataset

본 연구에서는 총 4종의 약물 독성 데이터(DILI, Acute Oral Toxicity, AMES Mutagenicity, hERG)를 이용해 진행되었으며, 각각의 수집은 관련 논문, 공개된 데이터베이스 등에서 이루어졌다.

1. DILI

실험에 사용된 데이터들은 최신 논문을 포함하여 다수의 Drug-Induced Liver Injury 논문으로부터 수집되었다. 각각의 데이터는 Less-DILI-Concern을 제외한 후, Most-DILI-Concern을 Positive로, No-DILI-Concern을 Negative로 Re-labeling되었으며[[1]](#footnote-1), Compound의 SMILES 구조가 공개되어 있지 않은 경우에는 PubChemPy API를 이용하여 SMILES를 수집하였다. 이 과정에서 실험의 정확도를 위해 Compound name에 대응되는 SMILES가 하나만 존재하는 화합물만을 채택[[2]](#footnote-2)하여 Dataset을 구성하였다. 전체 Dataset에서 Feature로써 활용하기 위해 SMILES와 DILI Class를 분리하였으며, 각 데이터의 출처를 구분하기 위해 ID를 추가하였다.

수집된 데이터의 SMILES는 RDKit 라이브러리의 StandardizeSmiles 메서드에 의해 Canonicalize되었으며, 처리가 완료된 Canonical SMILES를 기준으로 중복 처리를 진행했다. 하나의 Canonical SMILES에 대응되는 DILI Class가 한 종류인 경우 제일 위에 오는 데이터 하나만을 남기고 나머지를 제거하였으며[[3]](#footnote-3), DILI Class가 같은 Canonical SMILES에 대해 여러 종류로 나타나는 경우에는 평균을 확인하여 처리하였다. 0과 1의 Label이 동일한 비율로 있는 경우는 잘못된 실험 데이터로 간주하여 제거하였으며, 나머지는 반올림을 적용하였다.

최종적으로 수집된 데이터는 2,261개로, 1,037개의 DILI Negative 데이터와 1,224개의 DILI Positive 데이터를 포함한다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **DILI Positive** | **DILI Negative** | **Total** |
| [Ma et al, 2020](https://pubs.acs.org/doi/full/10.1021/acs.chemrestox.0c00322) | 197 | 282 | 479 |
| [DILIst](https://www.sciencedirect.com/science/article/pii/S1359644619303824) [[4]](#footnote-4) | 768 | 511 | 1,279 |
| [Kotsampasakou et al, 2017](https://www.sciencedirect.com/science/article/pii/S0300483X17301701) | 398 | 413 | 811 |
| [Williams et al, 2019](https://pubs.acs.org/doi/abs/10.1021/acs.chemrestox.9b00264) | 44 | 21 | 65 |
| [Kim et al, 2017](https://link.springer.com/article/10.1186/s12859-017-1638-4) | 197 | 280 | 477 |
| [Zhang et al, 2016](https://onlinelibrary.wiley.com/doi/full/10.1002/minf.201500055) | 785 | 532 | 1,317 |
| **Total** | **1,224** | **1,037** | **2,261** |

Table 1. Overview of the datasets (DIILI)

1. Acute Oral Toxicity

해당 독성에 대한 실험에는 Gadaleta et al, 2019에서 공개한 rat acute oral systemic toxicity (LD50) 관련 Dataset을 사용하였다. 해당 Dataset은 8,994개의 Training set과 2,895개의 Evaluation set으로 이루어져 있으며, NICEATM과 NCCT의 Cheminformatics community에서 수집할 수 있다.

해당 데이터는 LD50(mg/kg) 값에 기반하여 4개의 Category로 분류되어있었는데, 이 중 Category에 해당하는 LD50(mg/kg) 값이 없는 데이터의 신뢰성의 확인을 위해 LD50값이 있는 데이터 8,908개에 대하여 해당 값과 매칭되는 class간의 consistency를 확인하였다. 결과적으로 전체 데이터에 대해서 각 Category의 class가 LD50값에 근거한 분류와 일치함을 확인하였고, 나머지 데이터에 대해서도 신뢰성이 높다고 판단, 실험에 사용하기로 하였다. 모든 데이터를 very\_toxic(LD50 < 50mg/kg)과 nontoxic(LD50>2,000mg/kg)으로 분류하였으며 very\_toxic을 Positive(1)로, nontoxic을 Negative(0)로 Re-labeling하였다.

마찬가지로 SMILES의 Canonicalization을 진행했으며, Canonical SMILES를 기준으로 중복 처리를 진행하였다. 해당 데이터의 경우 중복 데이터에 대한 Class의 Variation이 없어 따로 처리를 해주지않았다.

최종적으로 수집된 데이터는 5,812개로, 4,859개의 Acute Oral Toxicity Negative 데이터와 953개의 Acute Oral Toxicty Positive 데이터를 포함한다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | | **Category** | **Class 1** | **Class 2** | **Class 3** | **Class 4** | **Class 5** |
| [Gadaleta et al, 2019](https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0383-2) | | **vT** | 945 | 10,405 | - | - | - |
| **nT** | 6,430 | 4,856 | - | - | - |
| **EPA** | 945 | 2,449 | 5,563 | 2,161 | - |
| **GHS** | 223 | 722 | 1,480 | 3,929 | 4,856 |
| **Total** | **5,812** | **Positive** | | 953 | **Negative** | | 4,859 |

Table 2. Overview of the dataset (Acute Oral Toxicity)

1. AMES Mutagenicity

NIHS 및 AMES Mutagenicity 관련 논문에서 데이터를 수집하였다. Source를 구분하기 위해 추가적으로 ID를 생성하였으며, Mutagenicity의 경우 Positive(class A)를 1, Negative를 0으로 Labeling하였다. 데이터의 통일성을 위해 전체 데이터에서 extraordinary / inorganic molecules 데이터와, DEREK 또는 MultiCASE 내부 데이터와 모순된 결과를 보이는 데이터를 제거하였으며[[5]](#footnote-5), 타 독성 데이터와 같은 기준으로 중복처리를 진행하였다.

최종적으로 수집된 데이터는 8,634개로, 3,800개의 AMES Negative 데이터와 4,834개의 AMES Positive 데이터를 포함한다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **AMES Positive** | **AMES Negative** | **Total** |
| [Hansen et al, 2009](https://pubs.acs.org/doi/full/10.1021/ci900161g) | 3,503 | 3,009 | 6,512 |
| [Li et al, 2021](https://pubmed.ncbi.nlm.nih.gov/33506363/) | 703 | 680 | 1,383 |
| [NIHS](http://www.nihs.go.jpdgmamesqsar.html) | 672 | - | 672 |
| [Xu et al, 2012](https://pubs.acs.org/doi/10.1021/ci300400a) | 4,866 | 3,482 | 8,348 |
| **Total** | **4,834** | **3,800** | **8,634** |

Table 3. Overview of the dataset (AMES Mutagenicity)

1. hERG

해당 데이터는 BindingDB, ChEMBL, PubChem과 Cai et al, 2019에서 수집되었다. 각 Source를 구분하기 위해 마찬가지로 ID가 추가되었으며, IC50의 값이 10uM인 경우를 기준으로, 이보다 적을 경우 1, 이외의 경우 0으로 hERG labeling을 진행하였다[[6]](#footnote-6). 전체 데이터에 대하여 Canonicalization을 진행하였으며, 이를 기준으로 중복을 제거하였다. 또한 class가 서로 다른 데이터 2개를 추가로 제거하여 최종적으로 5,823개의 Negative와 8,464개의 Positive를 합하여 총 14,287개의 데이터를 얻었다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **hERG Positive** | **hERG Negative** | **Total** |
| BindingDB | 633 | 290 | 923 |
| ChEMBL | 3,954 | 1,741 | 5,695 |
| PubChem | 2,205 | 1,922 | 4,127 |
| [Cai et al, 2019](https://pubs.acs.org/doi/full/10.1021/acs.jcim.8b00769)(Training set) | 3,485 | 2,826 | 6,311 |
| **Total** | **8,112** | **5,789** | **13,901** |

Table 4. Overview of the dataset (hERG)

**Feature Engineering**

Dataset의 SMILES에 대하여 Feature Engineering을 수행하여 SMILES를 실수 기반 벡터인 Descriptor와 Binary 벡터인 Topological Fingerprints로 변환 후, 두 가지 벡터를 활용하여 Feature를 생성, 모델 학습에 이용하였다. 해당 과정은 RDkit의 Chem.MolFromSmiles를 이용해 Canonical SMILES로부터 Mol Object를 가져왔으며, AllChem.MorganFingerprintAsBitVect를 이용해 Binary Vector를 가져왔다. 또한 Descriptor.\_descList를 통해 RDkit에서 제공하는 전체 Descriptor 총 208개의 리스트를 가져왔으며, 이를 토대로 MoleculeDescriptors.MolecularDescriptorCalculator()를 통해 인스턴스를 생성, CalcDescriptor로 Descriptor의 tuple을 가져와 저장하였다.

최종적으로 처리된 데이터는 1,024개의 Fingerprints와 208개의 Descriptor를 포함하여 총 1,232개의 Feature를 가진다. Feature 개수 대비 적은 개수의 데이터를 실험에 활용하고 있기 때문에 Overfitting이 우려되어 Feature Selection을 진행하였다.

Hashing 방식으로 만들어지는 Fingerprints의 경우, 잘못된 Feature를 골라내면 임의의 Molecule이 입력되었을 때 On으로 표시된 bit가 제거되어 버릴 수 있으므로, Descriptor에 대해서만 Feature Selection을 진행하였다.

**Classifier evaluation**

총 여섯 종류의 Classifier[[7]](#footnote-7)를 사용하여 Classification을 진행 후, 각각의 성능을 분석하고 비교하였다. Generalize된 classification performance를 측정하기 위해 각 Classifier에 대해 Startified 3-Fold Cross Validation procedure이 수행되었으며, 이 과정은 매번 다른 random fold division을 활용하여 3번씩 수행되었다. 최종 classification performance는 전체 수행결과의 평균값으로 측정되었다.

Evaluation metric으로는 서로 다른 세 가지의 classification performance indicator[[8]](#footnote-8)가 사용되었으며 각각에 대한 정의는 다음과 같다.

1) Sensitivity (True Positive Rate)

Recall이라고도 하며, Observed Positive 대비 True Positive의 rate를 계산한다. 식은 다음과 같다.

Positive를 Positive로 예측한 True Positive(TP)와 Positive를 Negative로 예측한 False Negative(FN) 중 TP의 비율로, 모델이 얼마나 정확하게 Positive 값을 찾느냐를 나타내는 지표이다.

2) Specificity (True Negative Rate)

Predictied Negative 대비 True Negative의 rate를 계산한다. 식은 다음과 같다.

모델이 Negative로 예측한 값들 중 True Negative(TN)의 비율로, 모델의 Negative 예측 정확도를 나타낸다.

3) Balanced Accuracy

위에서 계산한 Sensitivity와 Specificity의 산술평균으로 1에 가까울수록 좋은 성능을 보인다고 할 수 있으며, Dataset이 Imbalance한 경우에 Accuracy보다 더 높은 신뢰성을 가지고 활용할 수 있는 지표이다.

**Logistic Regression**

**Decision Tree**

**Random Forest**

**k-Nearest Neighbor (k-NN)**

**Naive Bayesian**

**Support Vector Machine (SVM)**

1. Ma et al, 2020: Deep Graph Learning with Property Augmentation for Predicting Drug-Induced Liver Injury [↑](#footnote-ref-1)
2. Kim et al, 2017: Prediction models for drug-induced hepatotoxicity by using weighted molecular fingerprints [↑](#footnote-ref-2)
3. Xu et al, 2021 [↑](#footnote-ref-3)
4. Thakkar et al, 2020: Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity [↑](#footnote-ref-4)
5. Hansen et al, 2009 [↑](#footnote-ref-5)
6. Cai et al, 2019 [↑](#footnote-ref-6)
7. Logistic Regression / Decision Tree / Random Forest / k-NN / Naive Bayesian / SVM [↑](#footnote-ref-7)
8. Su et al, 2014 [↑](#footnote-ref-8)