Multi-modal approach for prediction of skin toxicity in small molecules

Dinabandhu Behera, Sujit Tangadpalliwar, Nitin Singhal, Chaith Kondragunta

AIRA Matrix Pvt. Ltd., Mumbai, India

Introduction

- Calculating precomputed descriptor values is a common method for identifying molecular data, however it is ineffective for large molecular sizes.
- To accurately characterise skin hazardous data, we devised a multimodal technique.
- We have also created a tool that reveals the class designation as well as the essential substructure found in skin hazardous compounds.

Objectives

A. To create a multimodal machine learning system for skin toxicity classification. The issue falls into the domain of binary classification.

B. To see how performance differs depending on the type of molecular dataset used.

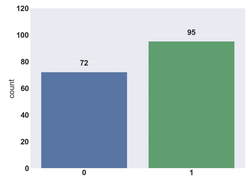
Materials

A. Skin hazardous data was gathered from public datasets such as TOXREF and SIDER, with a total of 197 molecular data samples in the dataset.

B. For tabular data, image data, and graph data production, we used Padel descriptor, Rdkit, and pytorch geometric, respectively.

C. We used 17536 features from the padel 2d descriptor as well as fingerprints. The image data used was 200 by 200 pixels.

D. We used sklearn, xgboost, and pytorch libraries for modelling, and optuna for model hyperparameter tuning.



Count Plot of Target

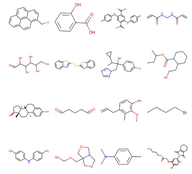
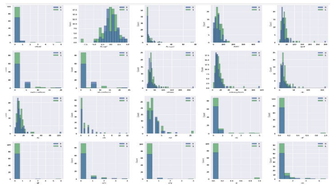


Image Data Samples



Distributions of Features

Data Visualization

Methods

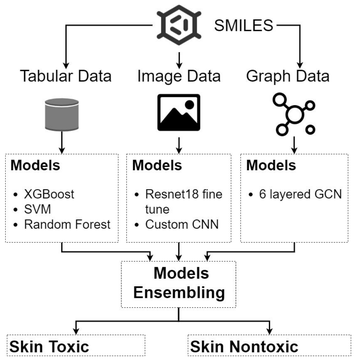
A. For all modalities, we employed the same fivefold stratified kfold and tuned model hyperparameters with optuna TPEsampler.

B. **Tabular data**: For feature engineering, we employed median imputation, winsorizor, IsolationForest, HSIC-Lasso, random-forest, mutual information based feature selection, Quantile transform, and conventional scaler approaches, as well as training xgboost, svm, and random forest models on training data.

C. **Image data**: We employed the fine-tuned ResNet18 network and a bespoke four-layer CNN for classification and the vertical horizontal flip augmentation approach.Graph data:

D. We generated node and edge features with the help of rdkit, and then developed a custom 6 layered GNN for graph classification

E. To increase the total AUC score, we employed weighted average ensemble.



Prediction Pipeline

Results

A. Using the tabular data, the AUC score of xgboost, Random Forest (RF) and SVM are 0.7285, 0.7267 and 0.7248 respectively.

B. The AUC score of the CNN model using ResNet18 pretraining was 0.7428, and the AUC score of the custom CNN model was 0.6831.

C. The AUC score for GNN model was 0.7833.

D. The ensemble weighted average of all models had an AUC of 0.8014.

Table 1 – AUC Score of Models

	Tabular data			Image data		Graph data
Models	**XGB**	**RF**	**SVM**	**ResNet18**	**Custom CNN**	**GCN**
	72.85	72.67	72.48	74.28	68.31	78.33
Ensemble	80.14					

Conclusions

A. **Tabular data** has a high dimensionality, which necessitates a larger number of molecular samples to achieve a satisfactory result. Also, for SMILES lengths greater than 200, the padel description takes significant amount of time. As a result, applying tabular models to larger molecules is extremely difficult.

B. Because of the line structure and irregular size of the molecular structure, working with **image data** is extremely difficult.

C. **The GNN models** rely on the molecular graph structure itself, which contains the individual atomic-node and edge information, making this method the most natural way of classifying molecular structure and outperforming other models.