Tia Tate | tate.tia@epa.gov | 0000-0002-7359-1660

Addressing Data Bias in Machine Learning for Hepatotoxicity Predictions Using Targeted Transcriptomics

T. Tate[1], G. Patlewicz[1], I. Shah[1]

[1]Center for Computational Toxicology and Exposure, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

PO11

www.epa.gov

INTRODUCTION

The development of new approach methods (NAMs) to inform chemical hazards and risks is presently a key emphasis in both national and international initiatives to drastically reduce animal testing. As a result, there is a renewed recognition that *in silico* methods can offer practical alternatives to bridge the gap between chemical characteristics, biological activities, and potential hazards to the environment and human health. Machine learning techniques are commonly used in the construction of Quantitative Structure-Activity Relationships (QSARs) classification, and regression models that link chemical molecular structural features to physical, chemical, or biological properties. Several studies have shown that machine learning-based models can successfully predict substance toxicity using chemical structure descriptors, bioactivity descriptors from High-Throughput Screening (HTS) tests, and hybrid mixes of the two.

Using targeted high throughput transcriptomic and chemical structural data, we use several machine learning classification approaches (Artificial Neural Networks (ANN), Gradient Boosting, K-Nearest Neighbor (KNN), Logistic Regression (LR), Naïve Bayes(NB), Random Forest, and Support Vector Machine Classification (SVC) algorithms) to create prediction models of chemically induced hepatotoxicity. We compare these approaches to the Generalized Read Across (GenRA) approach we applied in Tate et al (2021) and evaluate the F1, sensitivity, and precision of these supervised classification models with varying features and feature combinations using targeted HTTr descriptors along with chemical structure and a hybrid combination of both for predicting liver toxicity.

METHODS & MATERIALS

Data applied in the Analysis

- Chemicals and References Chemicals derived from the ToxCast library
- Individual Descriptors:
  - 2048 Chemical structure (Morgan) descriptors generated by python's RDKit library.
  - 93 Gene Hit Calls Descriptors from metabolically competent HepaRGTM cells LTEA assay of ToxCast HTS data from multiple concentration level 5 TCPL package in R.
- Hybrid Descriptors:
  - 2143 Morgan + Gene (BC)
  - 4825 Morgan + Torsion Topological + ToxPrints* (CA)
  - 2918 Gene + Morgan + Torsion Topological + ToxPrints (CBA)
- Toxicity Outcomes extracted from ToxRefDB v2:
  - 922 target effects
  - 5 Liver specific target effects

Classification Approaches

- Generalized Read Across (GenRA) [genra-py]
- Artificial Neural Networks (ANN)
- Gradient Boosting
- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)
- Naïve Bayes (NB)
- Random Forest
- Support Vector Machine Classification (SVC)

*ToxPrints are a set of 729 features as described in Yang et al (2015)

Balancing Methods

- **Under-Sampling**
  - Selecting from Majority Class:
    - Condensed Nearest Neighbor (CNN)
    - Near-Miss-1 (NM)
  - Removing from Majority Class:
    - Random
    - Tomek Links(TL)
    - Edited Nearest Neighbors (ENN)
- **Over-Sampling**
  - Synthetic Minority Oversampling Technique (SMOTE)

Machine Learning

- **Cross Validation**
  - 5-fold
- **Performance Evaluation**
  - F1-Score
  - Recall (sensitivity)
  - Precision (positive predictive value)

Figure 1: Workflow for classification process

❖ Morgan structural fingerprints, transcriptomic "hit call" descriptors, and a hybrid mix of chemical and biological descriptors were used to represent 1060 compounds.

❖ Several classification approaches were evaluated using F1, recall (sensitivity), precision scores with five-fold cross-validation followed by one way ANOVA and Tukey's HSD for multiple comparisons of mean differences.

U.S. Environmental Protection Agency
Office of Research and Development

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA.

RESULTS

Figure 2: Target Distribution of Positive and Negative Chemicals

Chr:Chronic; Sub:Sub-Chronic; Dev:Developmental; Mgr: Multigenerational Reproductive; Sac:Sub-Acute; N+/- Number of positive/negative chemicals

There's an evident unequal distribution of positive and negative chemicals for each of our target effects that may cause potential bias in performance scores.

Liver Endpoints	No. Pos Chemicals	No. Neg Chemicals
Chr	236	128
Dev	43	219
Mgr	115	86
Sac	58	41
Sub	356	158

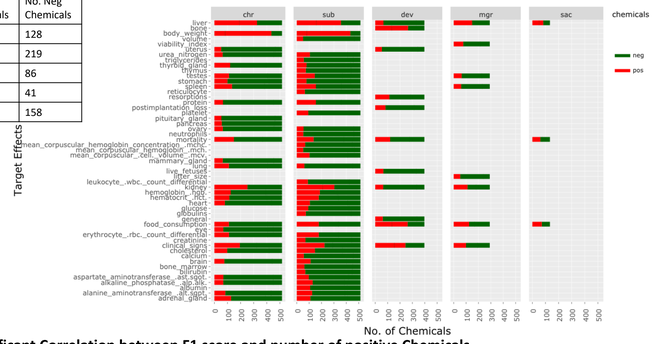


Figure 3: Significant Correlation between F1 score and number of positive Chemicals

❖ For data pooled by all toxicity endpoints and descriptors, Spearman's correlations for F1 performance score and number of positive compounds revealed a positive correlation between these factors.

❖ Similarly, a higher quantity of negative chemicals resulted in poorer performance ratings, according to the same study.

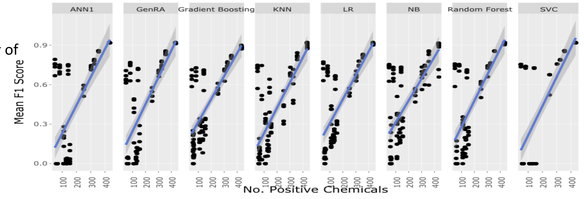
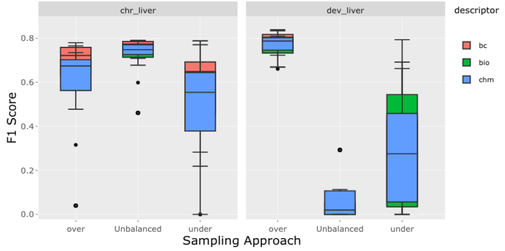


Figure 4: Case Example of Chr_ and Dev_ Liver toxicity endpoints comparing an unbalanced approach to under- and over-sampling balance techniques. Bc:Mrgn+Gene; Bio:Gene Hit Calls; Chm: Morgan Structural Fingerprints.



❖ For chr liver (N+>N-), our unbalanced method outperformed the oversampling technique for all descriptors, while the under-sampling strategy (aggregating all under-sampling approaches tested) exhibited varied performance.

❖ The unbalanced method, along with the variable under-sampling strategy and the over-sampling approach, proved to perform the worst for dev liver (N<N-).

❖ Note: The hybrid descriptor appears to improve performance ratings for each sampling technique using chr liver.

Figure 5: Significant Correlations between F1 Score and number of Positive Chemicals using various sampling approaches

❖ The number of positive chemicals appears to have an impact on the performance of unbalanced and under-sampling techniques like TL, CNN, and ENN, but the over-sampling strategy, SMOTE, appears to be unaffected.

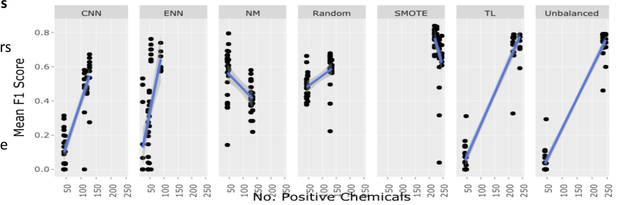
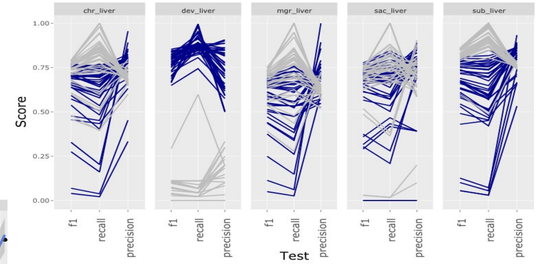


Figure 6: Evaluation of SMOTE over-sampling approach vs unbalanced approach for all liver endpoints.



❖ The unbalanced method appeared more sensitive (recall) than when balanced with SMOTE in situations when there were more positives than negative compounds. However, there is no discernible difference in accuracy or F1 score.

SUMMARY & CONCLUSIONS

We sought to address machine learning performance bias in the prediction of liver toxicity in this work by utilizing a targeted transcriptomic hit-call data set, chemical structural data, and a hybrid mix of both for several classification algorithms. We determined that:

❖ The performance of our classifiers were influenced by the quantity of positive compounds (better performance with more positive chemicals). As a result, we explored a variety of under-sampling and SMOTE over-sampling approaches to examine how they affected classifier performance.

❖ Like the imbalanced method, the initial number of positive chemicals influenced classifier performance with certain under-sampling approaches.

❖ SMOTE, on the other hand, produced relatively consistent results in a variety of balancing situations, and increased performance by 62% when the initial negative chemical balance outweighed the positive.

❖ Additionally, we determined that using hybrid descriptors can enhance hepatoxicity predictions (0.63 ± 0.16), when performance bias owing to imbalanced data is addressed (0.73 ± 0.03).

Ongoing analysis are applying feature selection approaches to determine how classifier performance is impacted and identifying relevant genes to build putative adverse outcome pathways (AOPs).

REFERENCES

Tate, T ., et al 2021 Computational Toxicology 19, 100171

Shah, I., et al., 2021 Bioinformatics Application Note.

Yang, C. et al., 2015 J. Chem. Inf. Model. 55, 510-528.

Helman, G., et al., 2019. Regul Toxicol Pharmacol. 109, 104480.