EPA United States Environmental Protection Agency

Development of skin sensitization, skin irritation, and eye irritation models using online data sources and Python-based machine learning

Abstract ID: 3585791

Ramsland[1], C; Sinclair[1], G.; Martin[2], T.; Williams[3], A.
[1] ORAU, RTP, NC
[2] CCTE, Cincinnati, OH
[3] CCTE, RTP, NC
Fall ACS Meeting, August 22-26, 2021, Atlanta, GA

https://orcid.org/0000-0003-1818-8235

Objective

- To develop QSAR models that predict chemicals with adverse toxicological effects to the skin and eyes
- Harvest and curate publicly available data for model training and validation
- Distinguish between chemicals that are corrosive to skin and eyes from those that are irritating but not corrosive
- Contribute to the effort of reducing the frequency of animal testing by making new computational approaches available to regulators

Skin Sensitization endpoint

- Skin sensitization is characterized by an inflammatory skin reaction caused by dermal exposure to a substance
- Tested using the murine local lymph node assay (LLNA) according to OECD guideline No. 429
- The allergic response is an occupational and environmental health concern, though it is typically not lethal
- Reported as a binary score of either sensitizing or not sensitizing
- Animal testing is the definitive method for classifying compounds as skin sensitizers
- *In silico* predictions can alter the process of determining whether a given chemical is to be regulated as a skin sensitizer
    - chemicals suggested to be very unlikely to cause skin sensitization may not be tested on animals.

Data Collection

The data used to create skin sensitization models was gathered from publicly available sources

- NIH's NICEATM dataset
- OECD QSAR Toolbox
- ECHA REACH dossier data via Echemportal

Data Processing

- Chemical structures were curated by EPA staff
  - Salt compounds were included in the data
- From a standardized chemical structure identifier, EPA T.E.S.T. descriptor software is used to calculate numerical representations of a molecule
- A random forest script was used to separate the data into representative training and test prediction sets.
- Features of final dataset contains nearly 800 calculated 2D T.E.S.T. descriptors that contain information from among following:
  - Constitutional Descriptors
  - Chi Connectivity Indices
  - Kappa Shape Indices
  - E-State Indices
  - Topological Descriptors
  - 2D Autocorrelation Descriptors

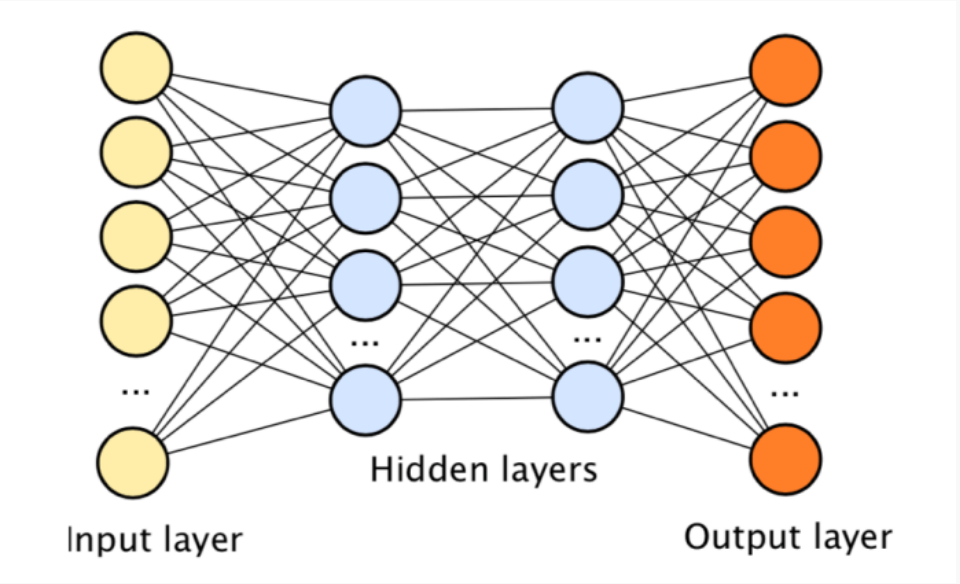
QSAR Modeling Techniques

6 Different modeling techniques were used for classifying skin sensitization

- A deep neural network (DNN)
- Random forest
- Support Vector Machines (SVM)
- XGBoost
- Implementations of k-nearest neighbors and hierarchical clustering algorithms from an existing Java EPA code base

Deep Neural Network

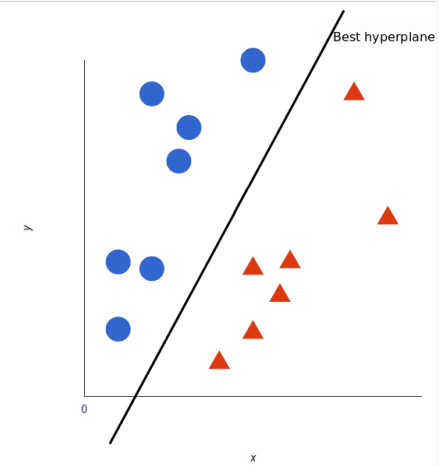
- Deep learning approach using the Keras python package with TensorFlow backend
- Feedforward network with three hidden layer implementation
- Trained by adjusting the weights and biases of network nodes whenever a compound is classified correctly or incorrectly

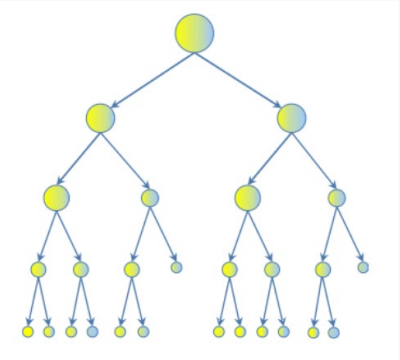


(Dickinson 2018)

Support Vector Machines (SVM) and Random Forest

- Both methods written in Python using the scikit-learn package
- SVM training algorithm relies on the construction of hyperplanes between data belonging to two different classes
- Random forest is an ensemble decision tree approach to classification or regression

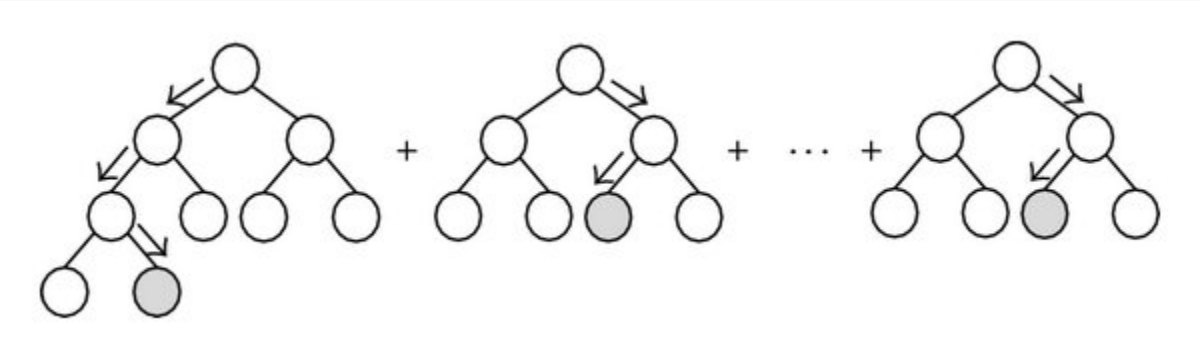




(Stecanella 2017 and Tierney 2018)

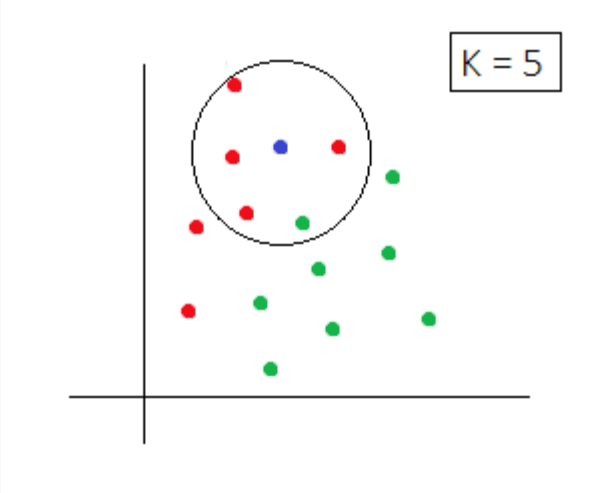
XGBoost

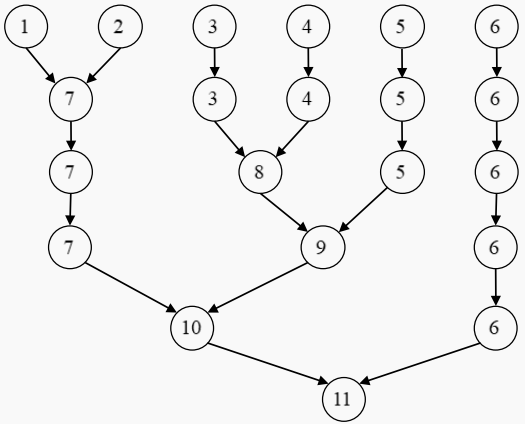
- XGBoost is another ensemble decision-tree based method that adds new models to correct for mistakes made in previous models
- XGBoost Python library is used to create the model



k-Nearest Neighbors and Hierarchical Clustering

- The nearest neighbors approach uses a genetic algorithm based on geometric distance of influential descriptors to find k similar samples in the training space, the averages from those groupings are used to make predictions
- Hierarchical clustering makes predictions using weighted average of several MLR models





(Rajeev 2019)

Developing the Consensus Model

- The predictions for all seven models are reported as log probabilities of observing skin sensitizing behavior for a given compound and the numerical average of these predictions is used to create a consensus prediction for a compound
- A different approach to consensus with an applicability domain is to perform the same mean calculation just described but to remove predictions between 0.40 and 0.60
    - Predictions with lots of uncertainty from the final consensus are removed but the model is not able to predict on as many compounds

Individual Model Results

QSAR Method	Balanced Accuracy	Sensitivity	Specificity	Coverage
kNN	0.739	0.677	0.801	1.0 (all chemicals)
Hierarchical Clustering	0.692	0.619	0.764	1.0
XGBoost	0.718	0.646	0.790	1.0
Random Forest	0.709	0.635	0.783	1.0
SVM	0.713	0.661	0.764	1.0
Deep Neural Network	0.713	0.651	0.775	1.0

Consensus Model Results

Consensus Method	Balanced Accuracy	Sensitivity	Specificity	Coverage
Average of all individual model predictions	0.760	0.693	0.826	1.0 (all chemicals)
AD strategy (0.40-0.60 predictions removed)	0.774	0.693	0.855	0.841

Further Study

- Skin sensitization is the first of three toxicological endpoints we have sourced data from and intend to develop models for
- Following the curation of our skin and eye irritation data, we intend to use a similar approach to modeling those endpoints as well.
- Skin and eye corrosion are closely related properties that we are interested in differentiating from the non corrosive compounds in our irritation data.
- Try out other applicability domain restrictions to see whether we can get better performance with less of a coverage tradeoff

Related Presentations

- *Development of models to predict physicochemical properties of PFAS*, presented by Dr. Todd Martin
- *Systematic development of QSAR data sets from online data*, presented by Gabriel Sinclair

Full Abstract

In 2018, US EPA released a draft policy to reduce animal testing for skin sensitization. The goal of this study was to assemble experimental data from online data sources and develop QSAR (quantitative structure activity relationship) models to predict skin sensitization, skin irritation, and eye irritation. Data was extracted from a variety of online data sources including eChemPortal, NICEATM, QSAR Toolbox, and the open literature. Using Java code, the data was converted to a consistent data format and stored in an SQLite database. Each record was mapped to a unique substance ID in EPA's Distributed Structure-Searchable Toxicology Database. The substance ID allows one to associate each record with a "QSAR-ready" SMILES string which is then used to generate molecular descriptors. Data set records consist of an ID value, a property value, and the molecular descriptor values. Records which contained the same two-dimensional inChiKey were merged. Discordant records were omitted and the data sets were randomly split into a training and prediction sets. For the skin irritation models, to account for corrosive behavior, two layers of binary classification were employed from intervals of the primary irritation index endpoint: distinguishing active vs. inactive substances, and then within the active set, distinguishing irritant vs. corrosive substances. Models were built using methods including random forest, support vector machines (SVM), XGBOOST, Deep Neural Networks (DNN), and k nearest neighbors (kNN). We optimized the hyperparameters for each model by selecting the set which performed best for internal cross validation of the training set or among many different external validation sets. We optimized the classification error, gamma and nu parameters for the SVM method and the learning rate, estimator count, and maximum depth for the XGBoost method. The SVM Consensus models averaging the results from the other approaches were also evaluated.

Acknowledgements

EPA-ORD-CCTE-CCED-CCCB

- Dr. Todd Martin
- Dr. Antony Williams
- Dr. Leora Vegosen
- Dr. Ann Richard, Dr. Chris Grulke, & ChemReg Project

Oak Ridge Associated Universities

- Gabriel Sinclair

References

1. Rajeev, Srijith K-nearest neighbors commonlounge.com (2019) [Image]. Retrieved 7 August 2021 at https://www.commonlounge.com/discussion/946c2b1e406942efbf3919be99ec9c37/history
2. Application of Boosting Regression Trees to Preliminary Cost Estimation in Building Construction Projects - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Gradient-boosted-decision-tree-ensemble_fig4_281513259 [accessed 9 Aug, 2021]
3. Tierney, Brendan Random Forest Machine Learning in R, Python and SQL - Part 1 blog.toadworld.com (Aug 31, 2018) Retrieved 7 August 2021 at https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1
4. Stecanella, Bruno An Introduction to Support Vector Machines (SVM) (June 22, 2017) [Image]. Retrieved 7 August 2021 at https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
5. Dickinson, Ben The limits and challenges of deep learning bdtechtalks.com (February 27, 2018) [Image]. Retrieved 7 August 2021 at https://bdtechtalks.com/2018/02/27/limits-challenges-deep-learning-gary-marcus/