**R Code for Skin Sensitization Classification**

**Author:** Dan Zang

**Affiliation:** Integrated Laboratory Systems, Inc.

**Contact:** [dzang@ils-inc.com](mailto:dzang@ils-inc.com)

**Publication:** Zang Q, Paris M, Lehmann DM, Bell S,Kleinstreuer NC, Allen D, Matheson J, Jacobs A, Casey W, Strickland J. Prediction of skin sensitization potency using machine learning approaches. *Journal of Applied Toxicology*;2017, 37(7): 792-805.

**Purpose:** the R code is used for building and validating classification models for predicting the potency of skin sensitization using support vector machine approach.

**Data sets:** there are three data sets. *SkinData-LLNA.txt* is used for LLNA data modeling (including Sensitizers vs Nonsensitizers, Strong Potency vs Weak Potency and three-category classification); *Human-Sensitizers-Nonsensitizers.txt* is used for Sensitizers vs Nonsensitizers classification for human data; *Human-ThreeCategories.txt* is used for Strong Potency vs Weak Potency and three-category classification for human data.

Create a folder named *"C:/SkinSensitizationClassification"*.

Save the data sets (*"SkinData-LLNA.txt"*, *"Human-Sensitizers-Nonsensitizers.txt"* and *"Human-ThreeCategories.txt"*) to the folder *"C:/SkinSensitizationClassification"*.

***LLNA-SkinData.txt*:** there are 120 rows (chemicals) and 16 columns (variables).

Among the 120 rows:

Rows 1-28: strong potency chemicals for training set;

Rows 29-68: weak potency chemicals for training set;

Rows 69-94: nonsensitizers for training set;

Rows 95-101: strong potency chemicals for test set;

Rows 102-113: weak potency chemicals for test set;

Rows 114-120: nonsensitizers for test set.

Among the 16 columns:

The first two columns are chemical name and CASRN;

Columns 3-8 are molecular weight (MW), partition coefficient (LogP), water solubility (LogS), vapor pressure (LogVP), melting point (MP) and boiling point (BP), respectively;

Column “OECD” is OECD QSAR Toolbox prediction values;

Columns “hCLAT” and “hCLAT.CD86.CD54.MIN” represent binary and numeric values for hCLAT assay, respectively;

Columns “avgDepletionLysCys” and “ARE.EC1.5” are DPRA and KeratinoSens assay values, respectively;

Column “PosNeg” represents if LLNA label is positive (sensitizer) or negative (nonsensitizer);

Column “Potency” represents if a chemical is a strong sensitizer (1A), a weak sensitizer (1B) or a nonsensitizer (Neg);

Column “Training.Test” indicates if a chemical is in training or test set.

***Human-Sensitizers-Nonsensitizers.txt*:** there are 96 rows (chemicals) and 17 columns (variables).

Among the 96 rows:

Rows 1-21: nonsensitizers for training set;

Rows 22-72: sensitizers for training set;

Rows 73-81: nonsensitizers for test set;

Rows 82-96: sensitizers for test set;

Among the 17 columns:

The first two columns are chemical name and CASRN;

Columns 3-8 are molecular weight (MW), partition coefficient (LogP), water solubility (LogS), vapor pressure (LogVP), melting point (MP) and boiling point (BP), respectively;

Columns “Lys”, “Cys”, “avg-Lys-Cys” and “DPRA” are DPRA assay values;

Columns “hCLAT” and “Keratino” represent hCLAT and Keratino assays, respectively;

Column “OECD” is OECD QSAR Toolbox prediction values;

Column “PosNeg” represents if a chemical is positive (sensitizer) or negative (nonsensitizer);

Column “Training.Test” indicates if a chemical is in training or test set.

***Human-ThreeCategories.txt*:** there are 87 rows (chemicals) and 14 columns (variables).

Among the 87 rows:

Rows 1-19: strong potency chemicals for training set;

Rows 20-41: weak potency chemicals for training set;

Rows 42-63: nonsensitizers for training set;

Rows 64-70: strong potency chemicals for test set;

Rows 71-79: weak potency chemicals for test set;

Rows 80-87: nonsensitizers for test set.

Among the 14 columns:

The first two columns are chemical name and CASRN;

Columns 3-8 are molecular weight (MW), partition coefficient (LogP), water solubility (LogS), vapor pressure (LogVP), melting point (MP) and boiling point (BP), respectively;

Columns “hCLAT.CE86.CD54.MIN”, “avgDepletionLysCys” and “ARE.EC1.5” represent hCLAT, DPRA and KeratinoSens assays, respectively;

Column “PosNeg” represents if a chemical is positive (sensitizer) or negative (nonsensitizer);

Column “Category” represents if a chemical is strong / weak sensitizer or nonsensitizer;

Column “Training.Test” indicates if a chemical is in training or test set.