Research Article

Improved Defined Approaches for Predicting Skin Sensitization Hazard and Potency in Humans

Haojian Li [1], Jing Bai [1], Guorui Zhong [1], Haosi Lin [1], Changsheng He [1], Renke Dai [1], Hongli Du [1] and Lizhen Huang [1,2]

[1]School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China; [2]Guangdong Provincial Key Laboratory of Fermentation and Enzyme Engineering, South China University of Technology, Guangzhou, China

Abstract

Since the EU banned animal testing for cosmetic products and ingredients in 2013, many defined approaches (DA) for skin sensitization assessment have been developed. Machine learning models were shown to be effective in DAs, but the predictivity might be affected by data imbalance (i.e., more sensitizers than non-sensitizers) and limited information in the databases. To improve the predictivity of DAs, we attempted to apply data-rebalancing ensemble learning (bagging with support vector machine (SVM)) and a novel and comprehensive Cosmetics Europe database. For predicting human hazard and three-class potency, 12 models were built for each using a training set of 96 substances and a test set of 32 substances from the database. The model that predicted hazard with the highest accuracy (90.63% for the test set and 88.54% for the training set, named hazard-DA) used SVM-bagging with combinations of all variables (V6), while the model that predicted potency with the highest accuracy (68.75% for the test set and 82.29% for the training set, named potency-DA) used SVM alone. Both DAs showed better performance than LLNA and other machine learning-based DAs, and the potency-DA provided more in-depth assessment. These findings indicate that SVM-bagging-based DAs provide enhanced predictivity for hazard assessment by further data rebalancing. Meanwhile, the effect of imbalanced data might be offset by more detailed categorization of sensitizers for potency assessment, thus SVM-based DA without bagging could provide sufficient predictivity. The improved DAs in this study could be promising tools for skin sensitization assessment without animal testing.

1 Introduction

Allergic contact dermatitis (ACD) is a clinically relevant condition induced by contact with an allergen. 15%-20% of the population suffers from ACD at some point in their life (Thyssen et al., 2007). To assess skin sensitization, animal methods such as the local lymph node assay (LLNA) and the guinea-pig maximization test (GMPT) have been adopted as the "benchmark" methods in many countries (Daniel et al., 2018). However, in the context of growing concern about animal welfare, the EU banned animal testing for cosmetic products and their ingredients in 2013 (EU, 2009). Thus, there is now an urgent need to develop non-animal methods that can fully reflect skin sensitization potency.

Substantial progress has been made in this regard by developing *in vitro* assays addressing different key events (KEs) in the skin sensitization adverse outcome pathway (AOP) (OECD, 2015, 2018a,b). The KEs in the AOP include the molecular initiating event (covalent binding to skin proteins) and the cellular response (activation of keratinocytes and dendritic cells) to the sensitizers. To evaluate these KEs, *in vitro* methods such as the direct peptide reactivity assay (DPRA), KeratinoSens™, and h-CLAT have been developed and accepted by the Organisation for Economic Co-operation and Development (OECD) (Emter et al., 2010; Gerberick et al., 2004; Sakaguchi et al., 2006). However, no single *in vitro* method can comprehensively represent the complexity of the processes involved in skin sensitization (Osamu et al., 2015). Therefore, as the components of the integrated approach to testing and assessment (IATA), many defined approaches (DA), which cover complementary characteristics of the *in vitro* methods and further take physicochemical properties and structure into consideration, have greatly improved the predictivity of skin sensitization (Worth and Patlewicz, 2016).

Received September 19, 2018; Accepted January 15, 2019;
Epub January 23, 2019; © The Authors, 2019.

ALTEX 36(3), 363-372. doi:10.14573/altex.1809191

Correspondence: Lizhen Huang, PhD, School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China
(huanglzh@scut.edu.cn)

Many algorithms are used in the data interpretation procedure for the defined approaches development, including some simple rule-based strategies and machine learning models. In machine learning models, the parameters could improve with increasing input data due to their being data-driven, resulting in quantitative prediction of potency being achieved in those DAs, but not in simple rule-based DAs (Jaworska et al., 2015; Kleinstreuer et al., 2018). Among the available machine learning models, the SVM model showed higher predictivity for human hazard, with an accuracy of up to 81.70%, compared with other machine learning models, i.e., artificial neural network (ANN) and Bayesian network (BN) (Kleinstreuer et al., 2018), and thus was considered a promising model.

However, the databases used to develop the available DAs usually included far more sensitizers than non-sensitizers (i.e., imbalanced data) (Jaworska et al., 2015; Hoffman et al., 2018), which might affect the predictivity of the DAs. Previous work has shown that the majority classes in the database are often correctly predicted by SVM, whereas minority classes tend to be misclassified (López et al., 2013; Li et al., 2016). Many studies in other fields have shown that the combination of bagging with SVM could significantly improve the predictivity of DAs developed on the basis of imbalanced data (Mordelet and Vert, 2014; Yu et al., 2018; Zararsiz et al., 2012). The bagging method could be applied to generate new, balanced training subsets suitable for building an SVM model by sampling with replacement from a given imbalanced training set (Guo et al., 2017). Therefore, the combination of SVM and bagging may be an effective option to develop DAs to assess skin sensitization.

Due to the complexity of the skin sensitization process, it is essential to apply a database that consists of comprehensive data from test methods together with animal and human reference data to develop a DA. Human potency was lacking in previous databases such as the LLNA EC3 or LLNA binary data (Hirota et al., 2015; Strickland et al., 2016). However, animal data is not a precise prediction target as some substances with high potency in humans were misclassified as non-sensitizers in the LLNA, e.g., 6-methyl-3,5-heptadien-2-one and tea leaf absolute (Api et al., 2017; Hoffmann et al., 2018). Besides, previous databases contained less data from newly OECD-accepted *in vitro* assays and only a limited number of cosmetic substances, which might reduce the predictivity for skin sensitization assessment of the cosmetic substances. Thus, to improve DA predictivity, the Cosmetics Europe database was established by compiling existing and newly generated data of the test methods together with LLNA and human reference data for 128 substances (Hoffmann et al., 2018). This new database provides more detailed information for skin sensitization, including human potency categories 1-6, LLNA data, continuous data from *in vitro* assays, and physicochemical properties of the chemicals in cosmetics. The human potency categories 1-6 (1,2: high potency; 3,4: low potency; 5,6: non-sensitizer) based on clinical patch test data with specific cut-offs for exposure and incidence are now considered as the best potency targets for assessment (Api et al., 2017; Basketter et al., 2014). Moreover, the continuous data from *in vitro* assays also can produce higher performance DAs than binary data (Zang et al., 2017). Thus, developing DAs using the Cosmetics Europe database could enable improved assessment of the human skin sensitization potency of chemicals.

Against this background, with the aim of further improving the predictive performance regarding human hazard and potency, we attempted to develop novel DAs by applying ensemble learning (SVM-bagging) and the newly established Cosmetics Europe database. All substances were divided into a training set of 96 substances and a test set of 32 substances. The predictivity of the novel DAs was validated by using the test set and results were compared with the LLNA and published DAs (Kleinstreuer et al., 2018).

2 Materials and methods

2.1 Substance database

We applied the Cosmetics Europe database, which has excluded metals and other substances used less frequently in cosmetics. For the 128 substances in the database, data on human potency class, DPRA, h-CLAT, KeratinoSens™, and six physicochemical properties relevant to skin exposure and penetration were collected. The six physicochemical properties were the octanol:water partition coefficient, water solubility, vapor pressure, melting point, boiling point, and molecular weight, see the supplementary file[1] (Hoffmann et al., 2018).

2.2 Characterization of the substances

Substances were assigned to six human potency classes in the database according to data from human maximization tests (HMT), human repeat insult patch tests (HRIPT), and diagnostic patch tests (DPT) (Hoffmann et al., 2018; Kleinstreuer et al., 2018): Classes 5 and 6 are non-sensitizers, classes 3 and 4 are low potency sensitizers, and classes 1 and 2 are high potency sensitizers. In this study, low potency and high potency sensitizers were classified as 1 and 2, respectively, and non-sensitizers were classified as 0. Of the 128 substances, 68.75% (88/128) were classified as positive for sensitization in humans and 31.25% (40/128) were classified as negative. Among the sensitizers, 58 substances were low potency sensitizers, while the others were high-potency sensitizers. Skin sensitizers may require oxidation (pre-haptens) and/or metabolism (pro-haptens) in order to produce a skin sensitization reaction. Among the 88 sensitizers, 10 were pre-haptens, 3 were pro-haptens, and 9 were pre/pro-haptens.

2.3 Model variables

DPRA

DPRA is an *in chemico* test that assesses the ability of a substance to form a hapten-protein complex, which is KE1 in the skin sensitization AOP. It measures the reactivity of a test substance towards two model synthetic peptides, one containing lysine (mixed at a ratio of 1:50 with the test substance) and the other containing cysteine (mixed at a ratio of 1:10 with the test sub-

[1] doi:10.14573/altex.1809191s

stance). The depletion of the peptides after incubation for 24 h with the test substance is measured using high performance liquid chromatography. Data used from the DPRA included average cysteine peptide depletion (Cys), average lysine peptide depletion (Lys), average depletion of cysteine and lysine peptides (Avg.Lys.Cys) and sensitizer/non-sensitizer outcome based on a decision tree. Data of Cys, Lys, and Avg.Lys.Cys were used as the model variables. The negative peptide depletion values were set to zero. The mean values of the Cys and Lys of 126 substances were applied. Cys and Lys values for dextran and 2-hexylidene cyclopentanone were not available.

KeratinoSens™

KeratinoSens™ assesses the ability of substances to activate and induce the expression of cytoprotective genes in keratinocytes based on activation of the Keap1-Nrf2 pathway of KE2 (keratinocyte activation) in AOP. This assay measures the antioxidant response element (ARE)-induced luciferase expression in a stable human keratinocyte cell line. The luciferase expression and cell viability are measured after 48 h of incubation with the test substance. The test substance is classified as a sensitizer if luciferase expression is activated over 1.5-fold compared with that in vehicle control and cell viability is over 70%. The EC1.5 value, i.e., the concentration at which luciferase expression is activated 1.5-fold, was used as the model variable. This value was available for all 128 substances.

h-CLAT

The h-CLAT assesses the ability of substances to activate and mobilize dendritic cells in the skin based on KE3 in the AOP. The assay measures the change of expression of CD86 and CD54 surface marker expression in human THP-1 cells by flow cytometry. Substances are classified as sensitizers if the relative fluorescence intensity is at least 150% of the baseline level for CD86 or at least 200% of the baseline level for CD54 at concentrations where cell viability is $\geq$ 50% of the control in at least two of three independent tests. There were 45 missing data for CD86 EC150 and 71 missing data for CD54 EC200 in the database. Thus, the binary outcomes of h-CLAT were used as the model variable and were available for 127 of the substances, while the missing h-CLAT outcome of 2-hexylidene cyclopentanone was imputed to be positive in accordance with the DPRA and KeratinoSens™.

Physicochemical properties

We adopted the data on octanol:water partition coefficient, water solubility, vapor pressure, molecular weight, melting point, and boiling point as variables. These were available for 122 substances. The mean value of the physicochemical properties of the 122 substances was applied for the 6 substances for which the corresponding data was not available.

2.4 Data processing

Selection of training set and test set

The 128 substances in the database were divided into 75% training and 25% external test sets. Based on the human potency classes assigned in the database, all substances were first classified as sensitizers or non-sensitizers, after which the sensitizers were classified as having high or low potency. Each substance in each potency class was randomly assigned to the training set or the test set. This process yielded a training set containing 96 sub-

Tab. 1: Six variable sets used to build models for predicting human hazard and potency

Variable	Variable set[a]					
	V1	V2	V3	V4	V5	V6
DPRA Cys		X	X	X		X
DPRA Lys		X	X	X		X
Avg.Cys.Lys		X	X	X	X	X
KeratinoSens™ EC1.5			X	X	X	X
h-CLAT Binary Result			X	X	X	X
Molecular Weight	X	X			X	X
Melting Point	X	X			X	X
Boiling Point	X	X			X	X
Log S	X	X		X	X	X
Log P	X	X		X	X	X
Log VP	X	X			X	X

[a] X denote the input variables included in each variable set.

DPRA Cys, depletion of cysteine peptide of the direct peptide reactivity assay; DPRA Lys, depletion of lysine peptide of the direct peptide reactivity assay; Avg.Cys.Lys, average depletion for cysteine and lysine; h-CLAT Binary Result, the binary result of human cell line activation test; Log S, log water solubility; Log P, log octanol:water partition coefficient; Log VP, log vapor pressure.

stances (75% of the 128), which consisted of 66 human sensitizers (68.75% or 66/96; 44 (45.83%) low potency substances and 22 (22.92%) high-potency substances) and 30 human non-sensitizers (31.25% or 30/96). The external test set consisted of the remaining 32 substances (25% of the 128), with 22 human sensitizers (68.75% or 22/32), 14 (43.75%) low potency substances and 8 (25%) high-potency substances), and 10 human non-sensitizers (31.25% or 10/32). The training and test sets were similar to one another and to the full 128-substance set with respect to the distributions of human potency.

Building predictive models

We used the training set of 96 substances to build models for predicting human outcomes using the SVM model, and using the trained SVM as the base model for the bagging method. The bagging method is a re-sampling technique applied by generating variable training subsets by sampling with replacement from a standard training set. Model building was implemented using packages in the scikit-learn in Python (Pedregosa et al., 2011). Prediction models were initially developed using SVM and each of six variable sets based on different combinations of the 11 collected variables, and were then integrated with the bagging method to improve performance. Thus, 12 models each were built. Table 1 defines the six variable sets.

Evaluation of model performance

Model performance for hazard assessment was evaluated by calculating the sensitivity, specificity and accuracy for predicting human outcomes using Cooper statistics by the formulae below (Strickland et al., 2016). The selection of the best model for hazard assessment was based on the accuracy and the mean value of sensitivity and specificity in both test set and training set. Model performance for potency assessment was evaluated by calculating accuracy, over-predicted rate and under-predicted rate by the formulae below (Kleinstreuer et al., 2018). The selection of best model for potency was based on the accuracy and the mean

Tab. 2: Performance of 12 models for predicting human hazard

Model[a]	Variable set[b]	Data set[c]	Sensitivity (%)	Specificity (%)	Average of sensitivity and specificity (%)	Accuracy (%)
SVM	V6	Training	95.45	73.33	84.39	88.54
		Test	90.91	80.00	85.46	87.50
SVM-Bagging	V6	Training	95.45	73.33	84.39	88.54
		Test	90.91	90.00	90.46	90.63
SVM	V5	Training	100.00	63.33	81.66	88.54
		Test	95.45	40.00	67.72	78.13
SVM-Bagging	V5	Training	100.00	73.33	86.66	91.67
		Test	95.45	40.00	67.72	78.13
SVM	V4	Training	93.94	36.67	65.30	76.04
		Test	100.00	80.00	90.00	93.75
SVM-Bagging	V4	Training	98.48	63.33	80.90	87.50
		Test	95.45	80.00	87.72	90.63
SVM	V3	Training	93.94	36.67	65.30	76.04
		Test	100.00	80.00	90.00	93.75
SVM-Bagging	V3	Training	93.94	36.67	65.30	76.04
		Test	100.00	80.00	90.00	93.75
SVM	V2	Training	100.00	66.67	83.34	89.58
		Test	63.64	20.00	41.82	50.00
SVM-Bagging	V2	Training	100.00	66.67	83.34	89.58
		Test	68.18	20.00	44.09	53.13
SVM	V1	Training	98.48	46.67	72.58	82.29
		Test	86.36	0.00	43.18	59.38
SVM-Bagging	V1	Training	98.48	50.00	74.24	83.33
		Test	81.82	0.00	40.91	56.25

[a] SVM, support vector machine; SVM-Bagging, bagging method with SVM as its base model. [b] Variables in each variable set are shown in Table 1. [c] The training set of 96 substances contains 30 non-sensitizers and 66 sensitizers. The test set of 32 substances contains 10 non-sensitizers and 22 sensitizers.

value of over-predicted and under-predicted substances in both test set and training set. Any candidate model that incorrectly predicted high-potency substances as non-sensitizers could not qualify as the best model.

$$Sensitivity = \frac{(True\ Positives)}{(True\ Positives + False\ Negatives)}$$

$$Specificity = \frac{(True\ Negatives)}{(True\ Negatives + False\ Positives)}$$

$$Accuracy\ of\ model\ for\ hazard\ assessment = \frac{(True\ Positives + True\ Negatives)}{(True\ Positives + True\ Negatives + False\ Positives + False\ Negatives)}$$

$$Over\text{-}predicted = \frac{(Over\text{-}predicted\ substances\ amount)}{(Total\ substances\ amount)}$$

$$Under\text{-}predicted = \frac{(Under\text{-}predicted\ substances\ amount)}{(Total\ substances\ amount)}$$

$$Accuracy\ of\ model\ for\ potency\ assessment = \frac{(Correct\text{-}predicted\ substances\ amount)}{(Total\ substances\ amount)}$$

3 Results

3.1 Performance of machine learning models for predicting human hazard

Predictivity of the machine learning models

The performance of the 12 models for predicting human hazard is shown in Table 2. The accuracy for the training set ranged from 76.04% to 98.48%, while that for the test set ranged from 50.00% to 93.75%. Overall, the bagging method mainly improved the accuracy by improving the corresponding specificity, which increased either in the training set or the test set for 3 of the 6 SVM models (i.e., SVM V1, V5, and V6). The difference of specificity between training set and test set tended to decrease with more input variables.

The best model for assessing human hazard was SVM-bagging-V6 (bagging model using V6 variable set and SVM as the base model), which did not misclassify any high-potency substance as a non-sensitizer. For the test set, it had 90.63% accuracy and an average of 90.46% for specificity and sensitivity. For the training set, it had 88.54% accuracy and an average of 84.39% for specificity and sensitivity. SVM-bagging-V6 is referred to as hazard-DA in the following.

Substances misclassified by hazard-DA in the training set

Hazard-DA misclassified 11 substances in the training set, with 3 false negatives and 8 false positives (Tab. 3). The false-negative substances were isocyclogeraniol, benzyl alcohol and benzyl cinnamate, which were all low potency substances and not pro/pre-haptens. h-CLAT was the only *in vitro* method that correctly identified isocyclogeraniol and benzyl alcohol as sensitizers, and KeratinoSens™ was the only *in vitro* method that correctly identified benzyl cinnamate as a sensitizer.

Substances misclassified by hazard-DA in test set

Hazard-DA misclassified 3 substances in the test set, with 2 false negatives and 1 false positive (Tab. 4). The false-negative substances were benzoyl peroxide and resorcinol, which were all low potency substances. Resorcinol was the only pro-hapten that was misclassified. DPRA was the only *in vitro* method that correctly identified benzoyl peroxide as a sensitizer, and h-CLAT was the only *in vitro* method that correctly identified resorcinol as a sensitizer.

Tab. 3: Misclassified substances of hazard-DA in the training set

Name	Cas	Hazard[a]	Prediction[b]
isocyclogeraniol	68527-77-5	1	0
benzyl alcohol	100-51-6	1	0
benzyl cinnamate	103-41-3	1	0
diethyl phthalate	84-66-2	0	1
octanoic acid	124-07-2	0	1
propyl paraben	94-13-3	0	1
anethole	104-46-1	0	1
benzyl benzoate	120-51-4	0	1
pentachlorophenol	87-86-5	0	1
α-methyl-1,3-benzodioxole-5-propionaldehyde	1205-17-0	0	1
α-iso-methylionone	127-51-5	0	1

[a] "1" means sensitizer, "0" means non-sensitizer. [b] human hazard prediction of the models.

Tab. 4: Misclassified substances of hazard-DA in the test set

Name	Cas	Hazard[a]	Prediction[b]	Pre/pro-hapten
benzoyl peroxide	94-36-0	1	0	NA
resorcinol	108-46-3	1	0	pro-hapten
salicylic acid	69-72-7	0	1	NA

[a] "1" means sensitizer, "0" means non-sensitizer. [b] human hazard prediction of the models.

Tab. 5: Performance of 12 models for predicting human potency

Model[a]	Variable set[b]	Data set[c]	Over-predicted (%)	Under-predicted (%)	Accuracy (%)
SVM	V6	Training	10.42	7.29	82.29
		Test	21.88	9.38	68.75
SVM-Bagging	V6	Training	10.42	9.38	80.21
		Test	25.00	9.38	65.63
SVM	V5	Training	11.46	6.25	82.29
		Test	15.63	12.50	71.88
SVM-Bagging	V5	Training	9.38	6.25	84.38
		Test	18.75	12.50	68.75
SVM	V4	Training	22.92	11.46	65.63
		Test	25.00	6.25	68.75
SVM-Bagging	V4	Training	22.92	11.46	65.63
		Test	25.00	6.25	68.75
SVM	V3	Training	23.96	12.50	63.54
		Test	25.00	6.25	68.75
SVM-Bagging	V3	Training	27.08	9.38	63.54
		Test	28.13	6.25	65.63
SVM	V2	Training	20.83	8.33	70.83
		Test	28.13	12.50	59.38
SVM-Bagging	V2	Training	20.83	6.25	72.92
		Test	28.13	9.38	62.50
SVM	V1	Training	15.63	14.58	69.79
		Test	28.13	43.75	28.13
SVM-Bagging	V1	Training	13.54	16.67	69.79
		Test	28.13	37.50	34.38

[a] SVM, support vector machine; SVM-Bagging, bagging method with SVM as its base model. [b] Variables in each variable set were shown in Table 1. [c] The training set of 96 substances contains 30 non-sensitizers, 44 low potency sensitizers and 22 high potency sensitizers. The test set of 32 substances contains 10 non-sensitizers, 14 low potency sensitizers and 8 high potency sensitizers.

3.2 Performance of machine learning models for predicting human potency

Predictivity of the machine learning models

The performance of the 12 models for predicting human potency is shown in Table 5. The accuracy for the training set ranged from 63.54% to 84.38%, while that for the test set ranged from 28.13% to 71.88%. Overall, the bagging method improved accuracy of two SVM models (i.e., SVM V1 and V2) either in the training set or in the test set, while it did not improve the other SVM models. SVM-V5 was not considered to be improved by bagging because its accuracy increased in the training set but reduced in the test set.

The SVM-V5 had the highest accuracy in the test set but misclassified two high potency sensitizers as non-sensitizers. Together with SVM-V5, SVM-V5-bagging misclassified two high potency sensitizers as non-sensitizers. For the V6 variable set, bagging did not improve the SVM model, and the SVM model showed a slightly higher accuracy. Therefore, the best model for human potency was SVM-V6 (SVM model using the V6 vari-

able set), which had 68.75% accuracy, 21.88% over-predicted rate and 9.38% under-predicted rate for the test set, and 82.29% accuracy, 10.42% over-predicted rate and 7.29% under-predicted rate for the training set. The SVM-V6 model is referred to as potency-DA in the following.

Substances misclassified by potency-DA in the training set
Potency-DA misclassified 17 substances in the training set, with 7 under-predicted substances and 10 over-predicted substances (Tab. 6). None of the high-potency substances were misclassified as a non-sensitizer, and none of the non-sensitizers were misclassified as a high-potency substance.

Substances misclassified by potency-DA in the test set
Potency-DA misclassified 10 substances in the test set, with 3 under-predicted substances and 7 over-predicted substances (Tab. 7). None of the high-potency substances were misclassified as a non-sensitizer, and none of the non-sensitizers were misclassified as a high-potency substance.

4 Discussion

Given the bans on the testing of cosmetics products and their ingredients in animals in many parts of the world, feasible and accurate DAs as alternatives to such testing are needed urgently. Here, novel DAs for predicting human hazard and potency, hazard-DA and potency-DA, were developed. The hazard-DA was generated by the combination of the bagging method and the SVM model, while the potency-DA was generated by the SVM model alone. Both hazard-DA and potency-DA showed higher predictivity than the other machine learning based DA and the LLNA (Kleinstreuer et al., 2018) (Tab. 8, 9).

The bagging method improved the accuracy of three SVM models (i.e., SVM V1, V5, and V6) by improving the predictivity of non-sensitizers and thus improving the specificity for hazard assessment. The hazard-DA (SVM-bagging-V6) had an improved accuracy of 90.63% in the test set, which was contributed by the increase of specificity from 80.00% to 90.00%, while the accuracy of the corresponding single SVM model was only 87.50%. Furthermore, the hazard-DA showed higher predictivity than the other validated machine learning based DAs and LLNA. This indicates that bagging indeed helped to rebalance the imbalanced hazard data and thus could improve the model performance, which is consistent with previous work in another field (Yu et al., 2018). However, for assessment of the three-class potency, the bagging method improved accuracy of two SVM models (i.e., SVM V1 and V2) either in the training set or in the test set, while it did not improve the other SVM models with more input variables. A potential explanation for this unexpected phenomenon is that the effect of imbalanced data on predicting potency had been offset by more detailed categorization of sensitiz-

Tab. 6: Misclassified substances of potency-DA in the training set

Name	Cas	Potency[a]	Prediction[b]	Pre/Pro-hapten[c]
isoeugenol	97-54-1	2	1	pre-Michael Acceptor
3-dimethylaminopropyla-mine	109-55-7	2	1	pro-Schiff base
lyral	31906-04-4	2	1	NA
2-hexylidene cyclopentanone	17373-89-6	2	1	NA
6-methyl-3,5-heptadien-2-one	1604-28-0	2	1	NA
2-methoxy-4-methylphenol	93-51-6	2	1	pro/pre-Michael Acceptor
benzyl alcohol	100-51-6	1	0	NA
diethyl phthalate	84-66-2	0	1	NA
octanoic acid	124-07-2	0	1	NA
propyl paraben	94-13-3	0	1	NA
anethole	104-46-1	0	1	NA
benzyl benzoate	120-51-4	0	1	NA
pentachlorophenol	87-86-5	0	1	NA
α-methyl-1,3-benzodioxole- 5-propionaldehyde	1205-17-0	0	1	NA
α-iso-methylionone	127-51-5	0	1	NA
citral	5392-40-5	1	2	NA
treemoss	90028-67-4	1	2	NA

[a] "2" means high potency, "1" means low potency, "0" means non-sensitizer. [b] "1" means sensitizer, "0" means non-sensitizer.
[c] NA, not applicable.

Tab. 7: Misclassified substances of potency-DA in the training set

Name	Cas	Potency[a]	Prediction[b]	Pre/Pro-hapten[c]
propyl gallate	121-79-9	2	1	pre-Michael Acceptor
benzoyl peroxide	94-36-0	1	0	NA
resorcinol	108-46-3	1	0	pro-Michael Acceptor
benzalkonium chloride	8001-54-5	0	1	NA
2-mercaptobenzothiazole	149-30-4	1	2	NA
tetramethylthiuramdisulfide	137-26-8	1	2	NA
1,4-dihydroquinone	123-31-9	1	2	pre-Michael Acceptor
iodopropynyl butylcarbamate	55406-53-6	1	2	NA
2-hydroxyethyl acrylate	818-61-1	1	2	NA
ethyleneglycol dimethacrylate	97-90-5	1	2	NA

[a] "2" means high potency, "1" means low potency, "0" means non-sensitizer. [b] human potency prediction of the models.
[c] NA, not applicable.

Tab. 8: Performance of DAs and LLNA for predicting human hazard[a]

DA	hazard-DA	BASF 2/3 (DKH)	Kao STS	Kao ITS	ICCVAM SVM	Shiseido ANN (D_hC)	Shiseido ANN (D_hC_KS)	P&G BN ITS-3	LLNA
Accuracy (%)	90.63	77.20	80.20	85.00	81.70	78.60	78.60	75.60	74.20
Sensitivity (%)	90.91	79.30	97.70	93.80	86.40	95.40	100.00	81.30	85.20
Specificity (%)	90.00	72.50	41.00	66.70	71.80	41.00	30.80	64.10	50.00

[a] Kleinstreuer et al., 2018

Tab. 9: Performance of DAs and LLNA for predicting human potency[a]

DA	potency-DA	Kao STS	Kao ITS ITS	Shiseido ANN (D_hC)	Shiseido ANN (D_hC_KS)	P&G BN ITS-3	LLNA
Accuracy (%)	68.75	63.50	69.20	61.10	62.70	54.80	59.40
Over-predicted (%)	21.88	22.20	13.30	22.20	25.40	20.00	19.50
Under-predicted (%)	9.38	14.30	17.50	16.70	11.90	25.20	21.10

[a] Kleinstreuer et al., 2018

ers in the database, resulting in the bagging method being useless in those models.

The potency-DA had an accuracy of 68.75% in the test set, which was higher than for LLNA. In the LLNA data, one human non-sensitizer had been classified as a high-potency substance (benzalkonium chloride), and two human high-potency sensitizers had been classified as non-sensitizers (6-methyl-3,5-heptadien-2-one and tea leaf absolute). This might have resulted from species differences in skin sensitization (Natsch and Emter, 2015). In the potency-DA, high-potency substances and non-sensitizers were not misclassified between these groups. Consequently, the potency-DA developed on the basis of the Cosmetics Europe database including data from human cell lines (KeratinoSens™ and h-CLAT) could assess the sensitization potency more accurately than the animal test.

At present, further *in vitro* assays for skin sensitization are being developed, so the robustness of such assays may be an important factor in the DA's performance. For example, address-

ing KE2, KeratinoSens™ showed good predictivity for hazard assessment, but its robustness may be affected by random integration (i.e., the plasmids of the ARE-luciferase reporter cassette insert into the genome randomly) or its inability to detect other skin sensitization regulation factors (Lai et al., 2016; Uemura et al., 2016; Soldner et al., 2016). Our group recently developed the EndoSens assay by precise knock-in of a reporter gene into the HMOX1 expression cassette. This assay was more robust and could be a better choice as an input variable for a DA (Zhong et al., 2018). *In vitro* assays addressing KE3 (i.e., the IL-8 Luc and U-SENS) have also been validated and accepted by OECD (OECD, 2018a). Thus, there is still great potential to optimize the DA's performance by using data from more robust *in vitro* assays addressing different KEs of the AOP when more testing data are available.

In conclusion, hazard-DA and potency-DA developed in this study are promising DAs for predicting human hazard and potency. Further work should focus on testing the models with an expanded set of substances, and applying them to data obtained from other validated and accepted assays to develop a more accurate DA for skin sensitization hazard and potency assessment.

Supplementary file

The supplementary file[1] contains the data from the Cosmetics Europe database and the prediction results of the models used in this study.

References

Api, A. M., Parakhia, R., O'Brien, D. and Basketter, D. A. (2017). Fragrances categorized according to relative human skin sensitization potency. *Dermatitis 28*, 299-307. doi:10.1097/DER.0000000000000304

Basketter, D. A., Alepee, N., Ashikaga, T. et al. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis 25*, 11-21. doi:10.1097/DER.0000000000000003

Daniel, A. B., Strickland, J., Allen, D. et al. (2018). International regulatory requirements for skin sensitization testing. *Regul Toxicol Pharmacol 95*, 52-65. doi:10.1016/j.yrtph.2018.03.003

Emter, R., Ellis, G. and Natsch, A. (2010). Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol Appl Pharmacol 245*, 281-290. doi:10.1016/j.taap.2010.03.009

EU – European Union (2009). EC 1223/2009 Regulation (EC) No 1223/2009 of the European Parliament and of the Council 30 November 2009 on cosmetic products. *OJ L 342*, 59-209. http://data.europa.eu/eli/reg/2009/1223/oj

Gerberick, G. F., Vassallo, J. D., Bailey, R. E. et al. (2004). Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci 81*, 332-343. doi:10.1016/j.tiv.2012.11.006

Guo, H., Li, Y., Shang, J. et al. (2016). Learning from class-imbalanced data: Review of methods and applications. *Exp Syst Appl 73*, 220-239. doi:10.1016/j.eswa.2016.12.035

Hirota, M., Fukui, S., Okamoto, K. et al. (2015). Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization. *J Appl Toxicol 35*, 1333-1347. doi:10.1002/jat.3105

Hoffmann, S., Kleinstreuer, N., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (I): The Cosmetics Europe database. *Crit Rev Toxicol 48*, 344-358. doi:10.1080/10408444.2018.1429385

Jaworska, J. S., Natsch, A., Ryan, C. et al. (2015). Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: A decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch Toxicol 89*, 2355-2383. doi:10.1007/s00204-015-1634-2

Kleinstreuer, N. C., Hoffmann, S., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Crit Rev Toxicol 48*, 359-374. doi:10.1080/10408444.2018.1429386

Lai, S., Wei, S., Zhao, B. et al. (2016). Generation of knock-in pigs carrying Oct4-tdTomato reporter through CRISPR/Cas9-mediated genome engineering. *PloS One 11*, e0146562. doi:10.1371/journal.pone.0146562

Li, Y., Guo, H., Liu, X. et al. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowl Based Syst 94*, 88-104. doi:10.1016/j.knosys.2015.11.013

López, V., Fernández, A., García, S. et al. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf Sci 250*, 113-141. doi:10.1016/j.ins.2013.07.007

Mordelet, F. and Vert, J. (2014). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit Lett 37*, 201-209. doi:10.1016/j.patrec.2013.06.010

Natsch, A. and Emter, R. (2015). Reporter cell lines for skin sensitization testing. *Arch Toxicol 89*, 1645-1668. doi:10.1007/s00204-018-2287-8

OECD – Organization for Economic Cooperation and Development (2015). Test No. 442C: In Chemico Skin Sensitisation: Direct Peptide Reactivity Assay (DPRA). *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264229709-en

OECD (2018a). Test No. 442E: In Vitro Skin Sensitisation: In Vitro Skin Sensitisation assays addressing the Key Event on activation of dendritic cells on the Adverse Outcome Pathway for Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264264359-en

OECD (2018b). Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264229822-en

Osamu, T., Shiho, F., Kenji, O. et al. (2015). Test battery with the human cell line activation test, direct peptide reactivity assay and derek based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. *J Appl Toxicol 35*, 1318-1332. doi:10.1002/jat.3127

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in Python. *J Mach Learn Res 12*, 2825-2830. http://www.jmlr.org/papers/v12/pedregosa11a.html

Sakaguchi, H., Ashikaga, T., Miyazawa, M. et al. (2006). Development of an in vitro skin sensitization test using human cell lines; human cell line activation test (h-CLAT). II. An inter-laboratory study of the h-CLAT. *Toxicol In Vitro 20*, 774-784. doi:10.1016/j.tiv.2005.10.014

Soldner, F., Stelzer, Y., Shivalila, C. S. et al. (2016). Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature 533*, 95-99. doi:10.1038/nature17939

Strickland, J., Zang, Q., Kleinstreuer, N. et al. (2016). Integrated decision strategies for skin. sensitization hazard. *J Appl Toxicol 36*, 1150-1162. doi:10.1002/jat.3572

Thyssen, J. P., Linneberg, A. and Johansen, J. D. (2007). The epidemiology of contact allergy in the general population – Prevalence and main findings. *Contact Dermatitis 57*, 287-299. doi:10.1111/j.1600-0536.2007.01220.x

Uemura, T., Mori, T., Kurihara, T. et al. (2016). Fluorescent protein tagging of endogenous protein in brain neurons using CRISPR/Cas9-mediated knock-in and in utero electroporation techniques. *Sci Rep 6*, 35861. doi:10.1038/srep35861

Worth, A. P. and Patlewicz, G. (2016). Integrated approaches to testing and assessment. Validation of alternative methods for toxicity testing. *Adv Exp Med Biol 856*, 317-342. doi:10.1111/bcpt.13018

Yu, L., Zhou, R., Tang, L. and Chen, R. (2018). A DNB-based resampling SVM ensemble learning. paradigm for credit classification with imbalanced data. *Appl Soft Comput 69*, 192-202 doi:10.1016/j.asoc.2018.04.049

Zang, Q., Paris, M., Lehmann, D. M. et al. (2017). Prediction of skin sensitization potency using machine learning approaches. *J Appl Toxicol 37*, 792-805. doi:10.1002/jat.3424

Zararsiz, G., Elmali, F. and Ozturk, A. (2012). Bagging support vector machines for leukemia classification. *Int J Comput Sci Issues 9*, 355-358. https://bit.ly/2DtSOqK

Zhong, G., Li, H., Bai, J. et al. (2018). Advancing the predictivity of skin sensitization by applying a novel HMOX1 reporter system. *Arch Toxicol 92*, 3103-3115. doi:10.1007/s00204-018-2287-8

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgement

This work was funded by Science and Technology Planning Projects of Guangzhou City (201704020176; 201803030014) and the National Natural Science Foundation of China (31871292).