https://doi.org/10.31925/farmacia.2022.2.17 **ORIGINAL ARTICLE**

IMPROVING STRUCTURE BASED MODELS FOR PREDICTING CHEMICAL FUNCTIONS AND WEIGHT FRACTIONS IN COSMETIC PRODUCTS USING ENSEMBLE SUPPORT VECTOR MACHINE

TANJA VOJINOVIC [1]*, LJILJANA KASCELAN [2], ZORICA POTPARA [1], NATASA RADONJIC [1], ZORKA KASCELAN [1]

[1]*University of Montenegro, Faculty of Medicine, Pharmacy study program, Krusevac bb, 81 000 Podgorica, Montenegro*
[2]*University of Montenegro, Faculty of Economics, 37 Bulevar Jovana Tomasevica, 81000 Podgorica, Montenegro*

**corresponding author: tanjavojinovic88@gmail.com*

Manuscript received: December 2021

Abstract

Through usage of a large number of cosmetic products, consumers are very often exposed to toxic chemicals. This paper aimed to propose a model for the prediction of chemical functions and weight fractions in these products based on the structural and physico-chemical properties of the substances. Due to the imbalance of classes we used Support Vector Machine (SVM) method, which can complement a smaller class with the examples that are most similar to it and identify the examples that are most different. The generality of the SVM method was additionally enhanced by combining it with ensemble Bootstrap Aggregation (Bagging). The research results showed that the proposed bagging SVM method can overcome the disadvantages of previously applied methods. Furthermore, it can help address the lack of information needed to assess exposure to risk from the use of cosmetic products containing toxic chemicals in their composition. The proposed models can be applied to predict whether a certain chemical may be a substitute for a function performed by another possibly toxic chemical in a cosmetic product, as well as to determine the quantity proportion of a certain dangerous chemical on the basis of its chemical structure and physico-chemical properties.

Rezumat

Prin utilizarea unui număr mare de produse cosmetice, consumatorii sunt foarte des expuși la substanțe chimice toxice. Această lucrare a avut ca scop propunerea unui model de predicție a funcțiilor chimice și ponderii în aceste produse, pe baza proprietăților structurale și fizico-chimice ale substanțelor. Datorită structurilor chimice variate, am folosit metoda *Support Vector Machine* (SVM), care poate completa o clasă mai mică cu exemplele care sunt cel mai asemănătoare cu aceasta și poate identifica exemplele care sunt cele mai diferite. Generalitatea metodei SVM a fost îmbunătățită prin combinarea acesteia cu ansamblul *Bootstrap Aggregation* (*Bagging*). Rezultatele cercetării arată că metoda SVM propusă poate depăși dezavantajele metodelor aplicate anterior. Aceasta poate contribui la abordarea lipsei de informații necesare pentru a evalua expunerea la riscuri prin utilizarea produselor cosmetice care conțin substanțe toxice în compoziția lor. Modelele propuse pot fi aplicate pentru identifica dacă o anumită substanță chimică poate fi un substitut pentru o altă substanță chimică posibil toxică într-un produs cosmetic, precum și pentru a determina proporția cantitativă a unei anumite substanțe periculoase ținând cont de structura chimică și de proprietățile fizico-chimice ale acesteia.

Keywords: Support Vector Machine (SVM), chemical functions, cosmetic products, toxic chemicals

Introduction

Instead of the expected beneficial effect, adverse and harmful reactions may occur at the site of application of the cosmetic product. The main cause of side effects and toxic reactions can be a cosmetically active substance or excipient [31]. Allergic reactions are more common than irritants. Many substances found in cosmetic products are both irritants and sensitizers (e.g. perfumes) that and can cause allergic dermatitis [14]. Every single case of hypersensitivity, contact allergies, caused by a cosmetic product is harmful [32]. Some low molecular weight substances are converted into primary irritants or allergens only after UV-A and UV-B radiation and short waves of the visible spectrum. Photoallergic contact dermatitis can develop after using sunscreen products [20]. Consumers of cosmetic products are exposed to the toxic chemicals used in the manufacture of these products on a daily basis. In most cases, manufacturers provide neither a complete information about their qualitative composition, i.e. the set of chemicals they contain, nor the quantitative composition, i.e. the weight fraction of chemicals. The main goals of green chemistry are to reduce the use of toxic chemicals (cancerous as well as those that affect the reproductive and nervous system) while preserving the functionality of the chemical ingredients and the efficiency of the product. Classification of the chemicals by the chemical function that they perform may contribute to finding less toxic substitutes among chemicals that

have the same functionality [26, 33, 38]. The risk of exposure is highly dependent on the quantity of toxic chemical used, and those information are often unavailable or incomplete due to the lack of adequate regulations and business policies of the manufacturer [7, 8, 11]. There are a large number of chemicals on the market while no high-quality *in vivo* toxicity data exist for most of them, and for many there are no toxicity data whatsoever [39]. The standard method for toxicity assessment is high-throughput screening (HTS) applied by the United States Environmental Protection Agency (US EPA). This method involves assessing the potential harmfulness of a chemical *in vitro* by quantifying its bioactivity. The EPA evaluated about 8,000 chemicals using this method. However, there are approximately 100,000 chemicals in use in the US market, which are difficult to test individually [26]. EPA is currently working to develop methods for the substitution of chemicals that have been identified as hazardous with safe chemicals aimed at providing formulations that are safe for humans and the environment [34, 35]. Such methods involve identifying alternatives from available chemical databases. They take an individual chemical to evaluate and return multiple possible alternatives (low-throughput approach). Such an approach is not efficient enough for many chemicals, therefore, the methods are used to automatically identify groups of chemicals with appropriate function in large databases (high-throughput-HT access) for further, more accurate testing. Recently, there have been a number of studies proposing quantitative structure – use relationship (QSUR) methods for classifying and predicting the functionality of chemicals based on their chemical structure and/or physical chemical properties [19, 26].

Isaacs *et al*. [19] proposed a model for prediction of functions and weight proportion of chemicals in cosmetic products based on the physical-chemical properties and use of the chemicals in production, based on the random forest method. Due to the imbalance of classes (the number of examples corresponding to one function is much smaller than the number of examples corresponding to all other functions), they applied under-sampling so that the smaller class equals the greater one. Phillips *et al.* [26] applied the same balanced random forest method for the prediction of chemical function in a broader set of consumer products. The main drawback of the method used in this study is random under-sampling within the ensemble method, which leads to higher number of misclassified examples of negative class i.e. small precision for the positive class. Such a model classifies well known examples i.e. chemicals in a training set, but on an unknown set there is less potential for accurate classification of positive class examples (chemicals that could have a certain function based on their structure and physical chemical properties).

Therefore, the aim of our research is to overcome the problem of class imbalance more efficiently, i.e. to generate the QSUR models with a higher precision of the positive class and thus with a more accurate prediction by targeting chemicals that could act as potential functional substitutes for use in cosmetic products.

It has been confirmed in the literature that the SVM (Support Vector Machine) method could successfully solve the problem of class imbalance by eliminating data noise that leads to class overlaps, which means that this method can be used as a pre-processor that refines data [1, 9, 13]. The classifiers applied to SVM output (on the refined dataset) significantly improve their predictive performance [24, 25, 27]. Farquad and Bose [13] tested SVM as a pre-processor for highly unbalanced data (94%: 6%) and showed that SVM can balance data better than commonly used standard techniques such as under-sampling (taking a subset of a larger class) or oversampling (supplementing of minor class with new examples). SVM provides more minor class examples by associating the most similar major class examples to the minor class. Additionally, the authors showed that when SVM refines data and balances classes, other classification methods (such as neural network, logistic regression, and Random Forest), significantly reduce the misclassification of minor classes on such a refined dataset. Recently, the SVM method has been increasingly used in biochemistry and drug design researches to generate quantitative structure-activity relationship (QSAR) models that predict the activity of molecules based on their structural and other properties [3, 4, 10, 16, 23]. Unlike QSUR models that predict the function of a chemical based on its chemical structure, QSAR models predict its biological activity. Furthermore, QSAR models are mainly focused on organic matter whereas QSUR models are focused on toxic chemicals with an aim to replace them with the less hazardous ones that have similar functions [26]. The Ensemble Bootstrap Aggregation (Bagging) method enhances predictive performances by generating multiple models using a random sampling (with replacement) of the training set and selecting random subsets of predictors when creating models. Prediction is performed by the ensemble average of all models, and the number of models that "vote" for a result determines the probability of accuracy or confidence of the result.

Taking into consideration prior research, this paper proposes the bagging SVM method as a pre-processor of data to generate QSUR models aiming at enhancing their predictive performances. Moreover, to solve the problem of chemical high-fraction class misclassification highlighted by Isaacs *et al.* [19], a multi-class classification was proposed using the LibSVM implementation of the SVM method [6, 12] which applies the one-against-one approach for k-class learning problems by solving

$\frac{k(k-1)}{2}$ binary SVMs. Hsu & Lin [18], have shown that the one-against-one approach is better than the one-against-all for large datasets and the training time for this method is shorter.

Materials and Methods

This section aims at defining the data and their sources, as well as the predictive methods that will be applied.

Data

This paper uses two publicly available datasets that were formed in the research described in the previous section. Using publicly available database of the chemical ingredients of cosmetics with the function that they have in a product (CosIng), Isaacs *et al*. [19] formed the Functional Use (FUse) database. They have reduced many functions by harmonizing similar functions into a common one. Each chemical in this base has its own unique CASRN code, while one chemical can have multiple functions (in different products). Data on the weight fractions of chemicals in individual cosmetic products were obtained from the Consumer Product Chemical Profile Database (CPCPdb) and on-line product information provided by manufacturers (Material Safety Data Sheets – MSDS sheets). The products are assigned to the appropriate category, and each chemical within a product has an appropriate harmonized function taken from the FUse database and merged based on a unique CASRN code of the chemical. Thus, they obtained a data set containing a chemical as an ingredient in each row in a product of a certain category with a unique harmonized function and a weight fraction in that category (the unique identifier for the chemical, the product category and the function of the chemical is CASRN). This dataset is coupled *via* CASRN to the corresponding physical and chemical properties of the chemicals and to their use in manufacturing. The combined data set includes 828 chemicals and 17,103 of their functional uses (35 harmonized functions) in 4,115 cosmetic products, i.e. 66 categories of these products.

Phillips *et al*. [26] expanded the FUse database from previous research with new chemicals and their functions in consumer products, using product composition information from the manufacturers' web pages. The functions are harmonized so that each chemical has a unique function. This set of chemicals is coupled to sets of chemicals that have structural and physical-chemical descriptors *via* a unique CASRN code. Structural descriptors are publicly available and taken from the EPA's Distributed Structure-Searchable Toxicity – DSSTox database, and physical-chemical descriptors (molecular weight, vapour pressure, water solubility, Henry's Law constant, log of the octanol – air partition coefficient - log(Koa), the log of the octanol – water coefficient - log(Kow), half-life of a chemical in soil, sediment, water, and air and the persistence of a chemical in the environment) were obtained using the US EPA's Estimation Program Interface (EPI) Suite. Thus, a training set of 4,791 unique chemicals was obtained with 729 structural properties and 11 physical-chemical properties. Retaining only harmonized functions that include at least 10 chemicals, 49 remained in in the training set.

Our first dataset was taken from the FUse database created in the research of Isaacs *et al*. [19] covering functional use and weight fraction of chemicals in cosmetic product categories (17,103 functional uses). The data were merged (*via* a unique CASRN) with structural descriptors (729 descriptors in total) and physical-chemical properties of chemicals (11 descriptors in total) taken from supplementary materials in a study of Phillips *et al*. [26]. Structural descriptors are binary (dummy) variables that have a value of 1 if a chemical has a structural property defined by that descriptor (e.g.: atom.element_main_group, atom.element_metal_group_I_II, bond.CC..O.C_ketone_alkane_cyclic_.C4., etc.), while if it does not have such a property, the value of the variable is 0. Physical descriptors are numerical variables such as molecular weight, vapour pressure, water solubility, Henry's Law constant, the log of the octanol – air partition coefficient - log(Koa), the log of the octanol – water coefficient - log(Kow), half-life of a chemical in soil, sediment, water and air, and the persistence of a chemical in the environment. After merging and eliminating the functions that have less than 10 uses, there are 11,240 functional uses remaining in this dataset, thus representing the final training set (Fuse_Str_Pc) (Table I). According to Isaacs *et al*. [19] weight fractions were generalized to three categories: low (0.0001 - 0.01), medium (0.01 - 0.3) and high (0.3 - 1) so as to apply the classification predictive method. In this way, the dataset was prepared for prediction of weight fractions of chemicals in cosmetic products based on structural and physical-chemical descriptors, functional uses and categories of cosmetic products, using the multi-class SVM bagging method.

The second dataset was taken from the expanded FUse database created in the study by Phillips *et al*. [26], which contains data on chemicals and their harmonized functions in consumer products. As in the previous case, the data were merged (*via* a unique CASRN) with structural and physical-chemical descriptors taken from the same work, and functions that included less than 10 chemicals were removed. Thus, a final training set (HFunc_Str_Pc) was obtained, comprising 4,665 unique chemicals with 729 structural features, 11 physical-chemical properties and 43 harmonized functions (Table II). This set was used to generate the QSUR model (for prediction of functional uses based on structural and physical-chemical descriptors) by the SVM bagging method.

Table I

Training set *Fuse_Str_Pc*

casrn	chem_name	max_WF	category	function	atom.element_main_group	atom.element_metal_group_I_II	ring.polycycle_tricyclo_benzvalene	molecular_weight	vapor_pressure units (Pa)	logKoa_unitless	logP_unitless
50-00-0	Formaldehyde	0.001	eye makeup, other	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.001	eye makeup, other	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.01	face cream/ moisturizer	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.001	hair styling, gel	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.001	hair styling, gel	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.001	hair styling, gel	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.002	eye makeup, other	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.002	eye makeup, other	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.0005	fragrance	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.009	eye makeup, other	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.00007	face wash, acne	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.01	hand/body lotion	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.00005	face cream/ moisturizer	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-00-0	Formaldehyde	0.01	body wash	Preservatives	0	0	0	30.03	465000	5.211	0.35
50-14-6	Ergocalciferol	0.01	nail products, other	Skin Conditioners	0	0	0	396.66	8.46E-08	12.4	10.44
50-14-6	Ergocalciferol	0.01	nail products, other	Skin Conditioners	0	0	0	396.66	8.46E-08	12.4	10.44
50-14-6	Ergocalciferol	0.01	hair styling	Skin Conditioners	0	0	0	396.66	8.46E-08	12.4	10.44
50-21-5	Lactic acid	0.001	body wash	Skin Conditioners	0	0	0	90.08	3.81	4.758	-0.65
50-21-5	Lactic acid	0.001	face wash	Skin Conditioners	0	0	0	90.08	3.81	4.758	-0.65
50-21-5	Lactic acid	0.05	foot care	Skin Conditioners	0	0	0	90.08	3.81	4.758	-0.65
50-21-5	Lactic acid	0.05	foot care	Skin Conditioners	0	0	0	90.08	3.81	4.758	-0.65
50-21-5	Lactic acid	0.01	toner	Skin Conditioners	0	0	0	90.08	3.81	4.758	-0.65

Table II

Training set *HFunc_Str_Pc*

Casrn.	chem_name	harmonized_function	atom.element_main_group	atom.element_metal_group_I_II	ring.polycycle_tricyclo_benzvalene	molecular_weight	logKoa_unitless	logP_unitless
50-14-6	Ergocalciferol	skin conditioner	0	0	0	396.66	12.4	10.44
50-69-1	D-Ribose	skin conditioner	0	0	0	150.13	8.767	-2.43
50-89-5	Thymidine	skin conditioner	0	0	0	242.23	14.376	-0.64

Casrn.	chem_name	harmonized_ function	atom.element_ main_group	atom.element_ metal_group_ I_II	ring.polycycle_ tricyclo_ benzvalene	molecular_ weight	logKoa_ unitless	logP_ unitless
54-47-7	Pyridoxal phosphate	skin conditioner	0	0	0	247.15	17.792	0.37
56-65-5	Adenosine triphosphate	skin conditioner	0	0	0	507.19	37.869	-3.61
56-86-0	L-Glutamic acid	skin conditioner	0	0	0	147.13	8.531	-3.83
57-00-1	Creatine	skin conditioner	0	0	0	131.14	8.446	-3.72
58-55-9	Theophylline	skin conditioner	0	0	0	180.17	10.123	-0.39
58-61-7	Adenosine	skin conditioner	0	0	0	267.25	19.233	-1.38
59-23-4	Galactose	skin conditioner	0	0	0	180.16	6.189	-2.43
59-30-3	Folic acid	skin conditioner	0	0	0	441.41	28.028	-2.81
60-27-5	Creatinine	skin conditioner	0	0	0	113.12	8.245	-1.21
60-81-1	Phlorizin	skin conditioner	0	0	0	436.42	26.606	0.72
60-92-4	cAMP	skin conditioner	0	0	0	329.21	22.963	-2.36
61-19-8	Adenosine-5-phosphate	skin conditioner	0	0	0	347.23	27.708	-1.68
51-78-5	4-Aminophenol hydrochloride	hair dye	0	0	0	145.59	12.034	-3.09
55-55-0		hair dye	0	0	0	280.3	14.081	2.34
50-01-1	Guanidine mono-hydrochloride	buffer	0	0	0	95.53	10.006	-6.13
56-18-8	Iminobis-3-propylamine	buffer	0	0	0	131.22	9.5	-1.15
50-00-0	Formaldehyde	antimicrobial	0	0	0	30.03	5.211	0.35
54-11-5	Nicotine	antimicrobial	0	0	0	162.24	8.081	1

SVM

To classify linearly non-separable classes that appear in case of class overlaps and noise in the data, Vapnik [37] proposed the SVM method that maps data (it views as n-dimensional vectors) from original space to feature space, where the classes can be separated using hyper-alignment. Finding such a hyper-alignment minimizes the distance between its end position (so that the gap between classes, i.e. the margins, is as large as possible) and the closest points (support vectors). Instead of an explicit mapping function to a greater-dimension space, kernel function that allows the calculation of the scalar product of the vector in the original space (kernel trick) is used. Maximizing the margin in a greater-dimension space is reduced to the quadratic programming optimization problem in the original space, using the kernel function. Different kernel functions can be used, but it is often the most efficient and the most widely used RBF (Radial Basis Function) [28]. Training of SVM classifiers is being realized by choosing the optimal values of the gamma parameter for the RBF kernel, and the parameter C that represents the boundary for the margin i.e. the empty space between the classes. Choosing smaller values for parameter C reduces over-fitting and increases the generality of the SVM model i.e. its predictive performance.

Grid-search and k-fold cross-validation. A grid-search technique combined with k-fold cross-validation is used to select the optimal SVM parameters (C and γ). Grid-search sets appropriate rankings and steps for parameters (i.e. defines grid) and then tests their combinations so that the best predictive learner performances are reached. The k-fold cross-validation process is implemented by dividing the training set into k subsets, of which k-1 is used to train the model and the one remaining to test the predictive performances of the model using unknown examples. The procedure is iteratively repeated so that each of the k sub-sets serves as a test set. The final predictive performances are the averages of the model performances obtained in these k iterations.

Results and Discussion

In order to predict the functions of chemicals in cosmetic products based on their structural and physical-chemical properties, i.e. to generate QSUR models, on the training set HFunc_Str_Pc (numerical data are normalized 0-1 rank transformation), the bagging SVM learners were trained. Model training involves optimizing the C and γ parameters using 5-fold cross-validation so that maximum predictive learner performances are achieved. Using the grid-search technique i.e. the setting appropriate rankings for the parameters, optimal combinations of these parameters were found for the 43 bagging SVM models to predict each of the 43 harmonized functions. Figure 1 shows the predictive performances of the generated models.

As can be seen from Figure 1, the accuracy of the model based on 5-fold cross-validation ranges from 81.63% to 99.98%. All models are valid in contrast to the results obtained by Phillips *et al*. [26] where 8 of the 46 models had a balanced accuracy of less than 75%. For these very important functions such as masking agent, solvent, viscosity, controlling agent and perfumer, models with accuracy greater than 96.40% were obtained. However, it should be taken into consideration that this is not a balanced accuracy, so much of the accuracy falls on the major (negative) class (true negative rate i.e. specificity averages 98.94%, while true positive rate i.e. sensitivity averages 43.99%). The lower sensitivity of the models indicates that the models are not over-fitted, i.e. too dependent on the training set for the positive class, which gives them the potential for better performance on an unknown dataset. The specificity of the model is high because the negative class is much larger than the positive class. The small precision of the positive class, which averages 49.34%, indicates that a number of examples of the negative class that are the least distant from the positive class (chemicals most similar in structure and physical-chemical properties to negative class chemicals), declared as a positive (minor) class through the SVM classifier.

Applying the resulting bagging SVM models to the training set, a prediction was generated for 43 functions. Each bagging prediction generates multiple SVM models and accordingly each bagging prediction has its own probability. Predictions with high probability mean that the greatest number of SVM models thus generated voted for a chemical to have/does not have a certain function.

The goal now is to increase the precision and sensitivity of the model (precision of a positive class and true positive rate) by taking from a large number of negative class representatives only those chemicals that are farthest from the chemicals belonging to the positive class (those having the structure and properties differing to the greatest extent). Members of the negative class to be selected were determined by bagging SVM through assigning them the highest probability of belonging to the negative class.

Therefore, the next step is to determine the optimal threshold (minimum likelihood) of Pr, for predictions that will be accepted. For this purpose, the DT method with the gini_index measure for partitioning [37] and 5-fold cross validation were used. Specifically, the optimal threshold for the probability of an SVM prediction was determined for each of the 43 models as follows. A number of chemicals whose major class prediction was less than the Pr value were excluded from the training set and 5-fold DT predictive performances were tested. The parameter Pr is determined to obtain the maximum predictive performance of the DT learner. Figure 2 shows the predictive performances of the QSUR models thus obtained for each of the 43 functions.

It can be seen in Figure 2 that the average accuracy of the final QSUR models is 95.95%, the precision averages 88.49%, while the average sensitivity and specificity are 80.57% and 97.40%, respectively. Thus, after removing the example of a negative class whose probability of belonging to the class is less than the Pr threshold from the training set, precision increased on average by 39.15%, sensitivity on average by 36.58%, while the sensitivity decreased on average by 1.54%.

In the study of Phillips *et al.* [26] for 49 harmonized function the 49 balanced random forest models were generated, of which 41 were valid (with balanced accuracy of > 75%). For 8 functions (which include some of the most important ones such as perfumer and solvent) no valid models were obtained i.e. random balanced under-sampling was not a method effective enough to predict significant differences in the structure and physical-chemical properties of these chemicals compared to the others. Most models have well recognized the chemicals that make up the positive class in the training set (sensitivity models average about 85%), however the average precision (positive classes) is only about 14%, which means that the predictive power of the model is weak. This is due to the large number of false positives (chemicals that do not have a specific function, but are misclassified by the model as having them). To identify the chemicals that could be functional substitutes, the generated models were applied to 6,356 chemicals in the Tox21 library for which there are structural and physical-chemical descriptors available. Consistent with the small precision of the positive class of obtained models, about 88% of the predictions were invalid (with a probability of less than 50%).

Comparing the performances of the QSUR models thus obtained with the results from the confusion matrix obtained by Phillips *et al.,* [26] (Table III) it can be seen that their accuracy averaged 91.81%, sensitivity 84.62%, specificity 91.83%, and precision 13.73%. The precision of our QSUR models is significantly improved over theirs, while the other 3

indicators were similar. The foregoing could lead to a conclusion that the predictive potential of our QSUR models for the positive class (chemicals having some function) is increased.

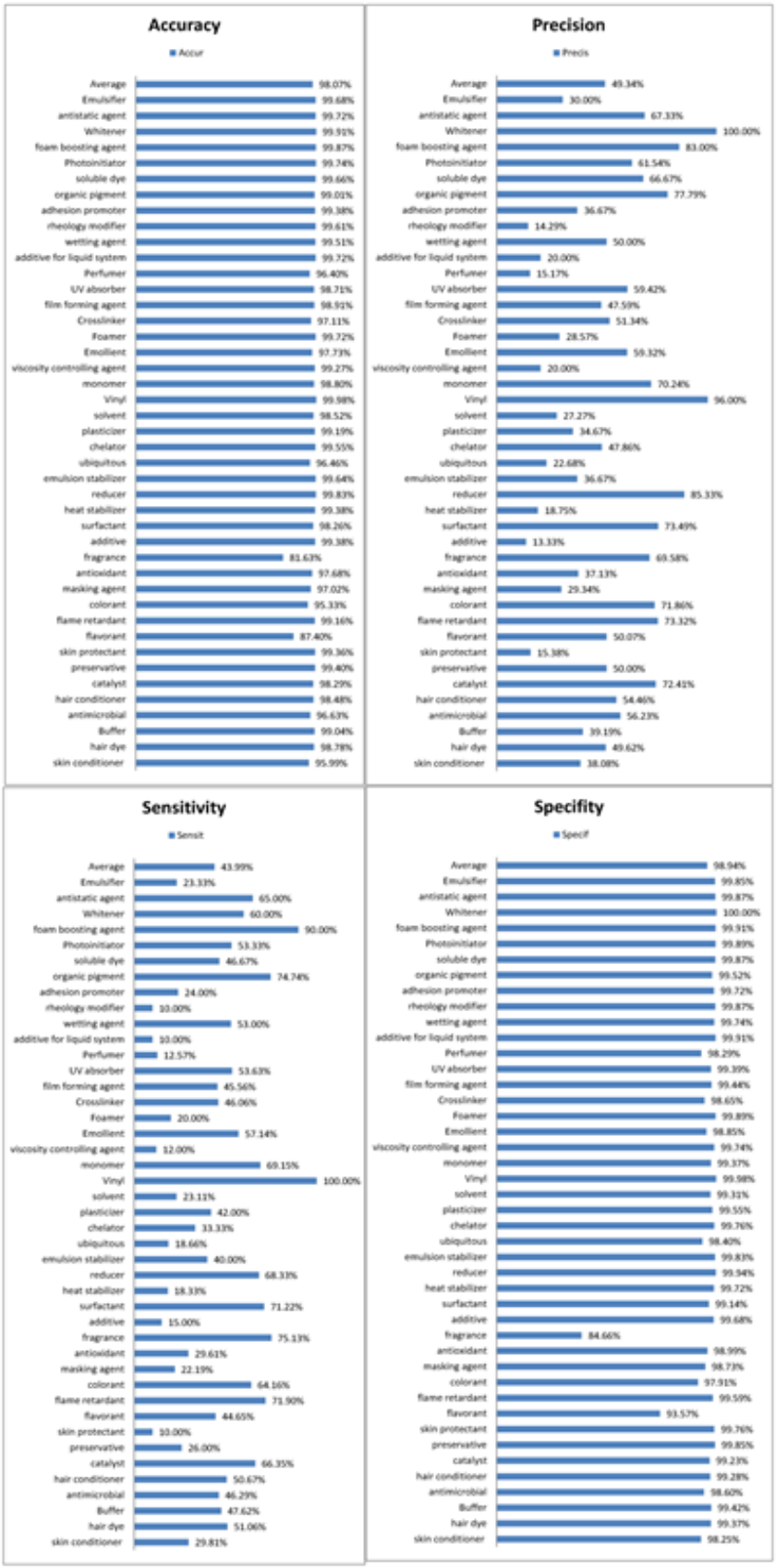

Figure 1.
Predictive performances of the bagging SVM models (5- fold cross-validation, positive class: 1)

Figure 2.
Predictive performance of DT models on bagging SVM output (5- fold cross-validation, positive class: 1)

Table III

Confusion matrix for QSUR models from the research Phillips *et al*. [26]

Harmonized function	tr_neg	tr_pos	fls_neg	fls_pos	Accur	Precis	Sensit	Specif
Additive	4549	15	2	225	95.26%	6.25%	88.24%	95.29%
rubber additive	5572	7	3	224	96.09%	3.03%	70.00%	96.14%
adhesion promoter	4659	19	2	111	97.64%	14.62%	90.48%	97.67%
Antimicrobial	4175	133	42	441	89.92%	23.17%	76.00%	90.45%
Antioxidant	4072	70	18	631	86.45%	9.99%	79.55%	86.58%
antistatic agent	4366	19	0	406	91.53%	4.47%	100.00%	91.49%
Buffer	4269	30	5	487	89.73%	5.80%	85.71%	89.76%
Catalyst	4250	104	30	407	90.88%	20.35%	77.61%	91.26%
Chelator	5417	33	3	353	93.87%	8.55%	91.67%	93.88%
Colorant	4063	288	69	371	90.82%	43.70%	80.67%	91.63%
Crosslinker	4315	108	30	338	92.32%	24.22%	78.26%	92.74%
Emollient	4338	104	22	327	92.72%	24.13%	82.54%	92.99%
Emulsifier	4368	11	0	412	91.40%	2.60%	100.00%	91.38%
emulsion stabilizer	5040	19	2	745	87.13%	2.49%	90.48%	87.12%
film forming agent	4251	37	9	494	89.50%	6.97%	80.43%	89.59%
flame retardant	4480	62	9	240	94.80%	20.53%	87.32%	94.92%
Flavorant	3968	539	105	1194	77.63%	31.10%	83.70%	76.87%
foam boosting agent	5630	19	1	156	97.30%	10.86%	95.00%	97.30%
Foamer	5560	9	2	235	95.92%	3.69%	81.82%	95.94%
Fragrance	2635	1291	193	672	81.95%	65.77%	86.99%	79.68%
hair conditioner	4184	66	11	530	88.71%	11.07%	85.71%	88.76%
hair dye	5213	95	9	489	91.42%	16.27%	91.35%	91.42%
heat stabilizer	4309	14	5	463	90.23%	2.94%	73.68%	90.30%
Humectant	5422	7	5	372	93.51%	1.85%	58.33%	93.58%
lubricating agent	5248	8	2	548	90.53%	1.44%	80.00%	90.55%
Monomer	4554	83	5	149	96.79%	35.78%	94.32%	96.83%
organic pigment	4502	94	1	194	95.93%	32.64%	98.95%	95.87%
Oxidizer	5190	11	1	604	89.58%	1.79%	91.67%	89.58%
Photoinitiator	4537	12	3	239	94.95%	4.78%	80.00%	95.00%
Plasticizer	4355	22	7	407	91.36%	5.13%	75.86%	91.45%
preservative	4947	40	8	811	85.89%	4.70%	83.33%	85.92%
Reducer	5177	25	1	603	89.60%	3.98%	96.15%	89.57%
rheology modifier	5536	7	6	257	95.47%	2.65%	53.85%	95.56%
skin conditioner	4050	117	37	587	86.98%	16.62%	75.97%	87.34%
skin protectant	4332	14	7	438	90.71%	3.10%	66.67%	90.82%
soluble dye	5447	19	0	340	94.14%	5.29%	100.00%	94.12%
Surfactant	4333	136	10	312	93.28%	30.36%	93.15%	93.28%
UV absorber	5093	70	16	627	88.93%	10.04%	81.40%	89.04%
Vinyl	4697	19	0	75	98.43%	20.21%	100.00%	98.43%
wetting agent	5645	22	2	137	97.61%	13.84%	91.67%	97.63%
Whitener	5640	10	1	155	97.31%	6.06%	90.91%	97.33%
Average					91.81%	13.73%	84.62%	91.83%

Bagging SVM prediction did the pre-processing of data for the DT learner and significantly improved its predictive performances even in the case of strong class imbalance. Instead of random under-sampling used in the study by Phillips *et al*. [26] to ensure class balancing, an under-sampling of the major class was made here based on bagging SVM prediction, i.e. only prediction with a higher probability of belonging to the major class was included in the sample. Thus, valid models were obtained for all 43 features with high predictive performances.

Based on the generated DT models, it can be concluded which structural and physical-chemical properties are most responsible for distinguishing the chemicals belonging to the positive class from the others. Thus, chemicals that have these properties can be identified and tested for potential substitutes. For example, the rules for the positive class derived from the DT model for the fragmentation function in Table IV show the structural and physicochemical properties that should satisfy the potential substitutes for this function in cosmetic products.

Rule 5 is the most important as it covers the largest number (1347) of positive examples with an accuracy of 95.73%. There are a number of studies/reports stating that 95% of the chemicals used in cosmetic products as fragrance components are of synthetic origin, derived from petroleum. The odour components

either of synthetic or of natural origin are the other allergens in frequency of causing reactions, accounting for 12.5% of all reactions [30]. The chemicals can cause symptoms such as: respiratory irritation, increased asthma, allergic reactions, mucosal irritations, migraines, headaches, skin problem, cognitive problem, gastro-intestinal problem, contact dermatitis, urticarial and photosensitivity. Some of these chemicals are lyral (synthetic lily scent), nitro and polycyclic musks, amyl cinnamal (usually of synthetic origin, though it may be of natural origin, having a floral jasmine-like scent), etc. These compounds are widely used as fragrances in various personal care products such as cosmetics and perfumery. In recent years, a large number of preparations have been marketed, which are labelled as odourless preparations, with the presence of vegetable ingredients or oils. Concealed allergens include rose oil, vanilla and sweet almondoil. Lilal - Butyl-phenyl-methyl-propanal causes contact dermatitis and is often found as a fragrant ingredient in perfumes, shampoos, bath preparations and lotions [2].

Table IV

Positive class rules derived from DT models

DT rules
Rule 1: 1 {0 = 2, 1 = 14}, Accuracy = 87.50%
logKoa_unitless > 0.199
logKoa_unitless > 0.229
bond.C..O.O_carboxylicEster_aromatic = true
logKoa_unitless ≤ 0.286
Rule 2: 1 {0 = 0, 1 = 6}, Accuracy = 100%
logKoa_unitless > 0.199
logKoa_unitless ≤ 0.229
chain.aromaticAlkane_Ph.C1_acyclic_generic = false
bond.COH_alcohol_ter.alkyl = false
chain.alkaneCyclic_ethyl_C2_.connect_noZ. = true
ring.hetero_.5._N_pyrrole_generic = false
persistence_units (hr) ≤ 0.084
Rule 3: 1 {0 = 0, 1 = 8}, Accuracy = 100%
logKoa_unitless > 0.199
logKoa_unitless ≤ 0.229
chain.aromaticAlkane_Ph.C1_acyclic_generic = false
bond.COH_alcohol_ter.alkyl = true
Rule 4: 1 {0 = 4, 1 = 45}, Accuracy = 91.83%
logKoa_unitless > 0.199
logKoa_unitless ≤ 0.229
chain.aromaticAlkane_Ph.C1_acyclic_generic = true
bond.C..O.O_carboxylicAcid_generic = false
water_solubility_units (mg/L) ≤ 0.129
bond.X.any._halide = false
bond.CN_amine_aliphatic_generic = false
Rule 5: 1 {0 = 60, 1 = 1347}, Accuracy = 95.73%
logKoa_unitless ≤ 0.199
logP_unitless > 0.375
bond.X.any._halide = false
atom.element_metal_metalloid = false
bond.OZ_oxide_peroxy = false
bond.CN_amine_aliphatic_generic = false
bond.C..O.O_carboxylicEster_alkenyl = false
bond.CS_sulfide = false
molecular_weight ≤ 0.229
Rule 6: 1 {0 = 0, 1 = 20}, Accuracy = 100%
logKoa_unitless ≤ 0.199
logP_unitless > 0.375
bond.X.any._halide = false
atom.element_metal_metalloid = false
bond.OZ_oxide_peroxy = false
bond.CN_amine_aliphatic_generic = false
bond.C..O.O_carboxylicEster_alkenyl = true
chain.alkeneLinear_diene_1_2.butene = true
Rule 7: 1 {0 = 2, 1 = 10}, Accuracy = 83.33%
logKoa_unitless ≤ 0.199
logP_unitless ≤ 0.375
bond.C..O.O_carboxylicEster_alkyl = false
bond.CC..O.C_ketone_generic = false
ring.hetero_.5._O_oxolane = false
logP_unitless > 0.359
persistence_units (hr) ≤ 0.020
chain.alkeneLinear_mono.ene_ehtylene_terminal = false
vapor_pressure_units (Pa) > 0.000
Rule 8: 1 {0 = 2, 1 = 25}, Accuracy = 92.59%
logKoa_unitless ≤ 0.199
logP_unitless ≤ 0.375
bond.C..O.O_carboxylicEster_alkyl = false
bond.CC..O.C_ketone_generic = true
molecular_weight ≤ 0.113
bond.C..O.O_carboxylicAcid_alkyl = false
chain.alkeneCyclic_ethene_C_.connect_noZ. = false
logP_unitless > 0.335
Rule 9: 1 {0 = 0, 1 = 24}, Accuracy = 100%
logKoa_unitless ≤ 0.199
logP_unitless ≤ 0.375
bond.C..O.O_carboxylicEster_alkyl = true
air_half_life_units (hr) > 0.000

It is most commonly obtained synthetically *via* cross-aldol condensation between para-terc-butylbenzaldehyde and propanal, followed by hydrogenation of the intermediate alkene. It is the clear, viscous liquid with a strong floral scent. In addition to causing contact dermatitis, it is suspected to have an effect on the endocrine system and oestrogen activity [2]. Citronellol is a colourless oily liquid with a floral scent on rose which is used in cleansers, hair care products, lip-sticks, perfumes. It is a known skin allergen, causes eczema, and often causes complications in people with psoriasis [5]. Nitro- and polycyclic musks are two common and important synthetic musks currently in use [36]. In addition, due to their strong photo-chemical toxicity, [29] carcinogenicity [17] and neuro-toxic properties, as well as endocrine dysfunction, nitro-musks (e.g. musk xylene), their use is being monitored in Japan, in the EU they are also under scrutiny, and further research is being conducted on their potential adverse effects on human health.

The process of high-throughput screening of a set of unknown chemicals using the generated QSUR models would consist of the following steps: (1) prediction of the chemical function using the bagging SVM model;

(2) elimination of chemicals whose non-function is determined with probability less than Pr; (3) application of the DT model to predict function on the purified set of chemicals from step 2; (4) for each chemical, each of the 41 QSUR models will generate the result; (5) the result that has the highest confidence will determine the function for which that chemical can be the substitute; and (6) if two or more results have the same confidence, then the result obtained by the model with the highest precision of positive class will be the winning one.

The next step is to generate a model for the prediction of weight fractions, on the training set Fuse_Str_Pc, in which the nominal variables are transformed into numerical dummy variables, while the numerical are normalized by a 0-1 rank transformation. By training bagging multi-class SVM learner, i.e. by applying grid-search procedure in combination with 5-fold cross-validation, the optimal combination of parameters SVM.C = 750.0 and SVM.gamma = 0.1 was obtained and the following predictive performance was achieved: accuracy – 88.47%, classification error – 11.53%, mean class recall – 83.44% and mean class precision – 85.82%.

For the purpose of predicting the fractions of chemicals in the product category, Isaacs *et al*. [19] generated a classification model, by generalizing fractions into three categories: low, medium and high fractions. They used structural and physical-chemical properties of chemicals as well as their functional uses as predictors. Using the random forest method, a model with a 5-fold cross-validation error of 16.7% was generated. The potential for misclassification of the obtained model is highest for high fractions (about 22%), while for low and medium fractions the class precision is about 84%. Accordingly, applying the model to an unknown dataset, less than 1% of chemicals are predicted to have high fractions (30% - 100% of total weight), while 35% and 65% of chemicals are predicted to have medium and low fractions. Therefore, the predictive performances of the model for the high-fraction chemicals that make up the majority of cosmetic product composition are not satisfactory.

With our model, the precision error for the high class is 15%, which is a better result than the one achieved by Isaacs *et al.* [19] having amounted to 22%, and the 5-fold classification error of the model is decreased by about 5%. This indicates that the multi-class bagging SVM model has a better predictive potential for a class of chemicals with a high participation in cosmetics than the random forest model used in the study mentioned.

As with the QSUR model, by generating a DT model on bagging SVM output, taking high Pr results (those voted for by the great number of SVM models obtained by bagging) for each class should further improve the prediction performances. However, since most of the results on the training set had a Pr greater than 0.99 (meaning that 99% of the generated SVM models voted for that result), the DT on the refined dataset had almost the same predictive performance as the SVM.

Nevertheless, the DT model generated on SVM output is useful as it provides explicit classification rules for low, medium and high fractions from which it can be concluded what are the chemical properties that cosmetic products contain to the greatest extent. Table V shows some of the most significant rules generated by DT model for all three classes (with the highest accuracy and cover).

Table V

Rules derived from DT models for weight factions

DT rules
Rule 1: Medium {Low = 135, Medium = 4193, High = 60}, Acc = 95.55%
logKoa_unitless > 0.086
bond.S..O.O_sulfonate = false
bond.C.O_carbonyl_generic = false
chain.aromaticAlkane_Ar.C_meta = false
logP_unitless > 0.352
chain.alkaneBranch_t.butyl_C4 = false
henrys_law_constant_units (atm-m^3/mol) ≤ 0.000
bond.X.any._halide = false
Rule 2: Low {Low = 174, Medium = 22, High = 1}, Acc = 88.32
logKoa_unitless > 0.086
bond.S..O.O_sulfonate = false
bond.C.O_carbonyl_generic = false
chain.aromaticAlkane_Ar.C_meta = true
Rule 3: Medium {Low = 24, Medium = 1301, High = 16}, Acc = 97.01%
logKoa_unitless > 0.086
bond.S..O.O_sulfonate = false
bond.C.O_carbonyl_generic = true
logP_unitless > 0.523
bond.CC..O.C_ketone_alkene_cyclic_2.en.1.one_generic = false
logP_unitless ≤ 0.685
bond.C.O_carbonyl_1_2.di = false

DT rules
Rule 4: High {Low = 0, Medium = 0, High = 114}, Acc = 100%
logKoa_unitless > 0.086
bond.S..O.O_sulfonate = false
bond.C.O_carbonyl_generic = true
logP_unitless ≤ 0.523
bond.C..O.O_carboxylicEster_acyclic = false
air_half_life_units (hr) > 0.001
Rule 5: Low {Low = 595, Medium = 167, High = 0}, Acc = 78.08%
logKoa_unitless > 0.086
bond.S..O.O_sulfonate = true
Rule 6: High {Low = 12, Medium = 209, High = 830}, Acc = 78.97%
logKoa_unitless ≤ 0.086
logP_unitless ≤ 0.466
bond.C..O.O_carboxylicAcid_alkyl = false

Thus, for example, it can be concluded from Table V, based on Rule 3 that cosmetic formulations can contain up to 30% of chemicals whose octanol-water partitioning coefficient is between 0.523 and 0.685. This means that water pollutants can be significant due to poor water solubility. Also, this rule shows that cosmetic products can contain up to 30% of chemicals with a carbonyl group to which they belong and some that are dangerous to human health. A substance that has a carbonyl group in it and is a common ingredient in various cosmetic products including liquid soaps, shampoos and shower creams/lotions is formaldehyde [21]. According to the International Agency for Research on Cancer, formaldehyde belongs to a group of human carcinogens because there is enough evidence that is causes cancer in humans. This fact is based on the fact that formaldehyde can lead to nasopharyngeal cancer in humans after inhalation and to squamous cell carcinoma of the nasal passages in rats [22]. This is why formaldehyde and paraformaldehyde are used as preservatives at concentrations of up to 0.1% in products used in cosmetics for oral hygiene (not to be used in aerosol products) and up to 0.2% in other products [15].

The procedure for predicting the unknown fraction of a chemical in a cosmetic product using the generated multi-class SVM model implies that the model input provides information on the product category, the function the chemical has in the product, the structural descriptors of the chemical and its physicochemical properties. Based on this information, the model will predict whether the chemical in the specified category is represented by a low, medium or high weight fraction.

Conclusions

An assessment of the toxicity exposure of chemicals in consumer products involves knowledge of the qualitative and quantitative composition of these products. Namely, on the basis of knowledge of the structural properties and amount of chemicals used in the product, the negative impact of the product on the consumer and the environment can be assessed.

This paper proposes methods that, rest on the available information on the functional and quantitative use of chemicals in thousands of real consumer products, generate predictive chemical classification models based on the function and weight fractions that chemicals have in cosmetic products. With these models, the composition of products with unavailable information can be assumed. These methods clearly define the approach by which great libraries of chemicals can be screened to identify potential substitutes for toxic chemicals without impairing the functions that the original chemicals have in the product. Equally, a clear procedure is defined on the manner in which the weight fraction of a chemical in a cosmetic product can be estimated using the generated predictive model. Thus, one can implicitly expose the chemical composition of cosmetic products that is inaccurate or completely inaccessible for many products.

The research results show that the proposed bagging SVM method can overcome the disadvantages of previously applied methods, i.e. increase the precision of prediction.

The proposed methods can help address the lack of information needed to assess exposure to risk from the use of cosmetic products containing toxic chemicals in their composition.

Conflict of interest

The authors declare no conflict of interest.

References

1. Barakat N, Bradley AP, Rule extraction from support vector machines: A review. *Neurocomp.*, 2010; 74(1-3): 178-190.
2. Breiman L, Friedman JH, Olshen RA, Stone CJ, Classification and regression trees. Taylor & Francis: New York, USA, 1984; 155-163.
3. Briem H, Günther J, Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem.*, 2005; 6(3): 558-566.
4. Byvatov E, Schneider G, SVM-based feature selection for characterization of focused compound collections. *J Chem Inf Comput Sci.*, 2004; 44(3): 993-999.

5. Cajkovac M, Kozmetologija. Naklada Slap: 10450 Jastrebarsko Dr Franja Tuđmana 33 Zagreb, Hrvatska, 2005; 319-326, (available in Croatian).
6. Chang CC, LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol.*, 2011; 2(3): 1-27.
7. Consumer Exposure Model United States Environmental Protection Agency; 2015. www.epa.gov/sites/production/files/2015-09/documents/cemuser guide beta test.pdf.
8. Delmaar C, Bokkers B, Burg W, Schuur G, Validation of an aggregate exposure model for substances in consumer products: a case study of diethylphthalate in personal care products. *J Expo Sci Environ Epidemiol.*, 2015; 25(3): 317-323.
9. Diederich J, Studies in Computational Intelligence. Rule Extraction from Support Vector Machines: An Introduction, Diederich J; Springer Verlag: Berlin Heidelberg, Germany, 2008; 80; 3-31.
10. Doucet JP, Barbault F, Xia H, Panaye A, Fan B, Nonlinear SVM approaches to QSPR/QSAR studies and drug design. *Curr Comput Aided Drug Des.*, 2007; 3(4): 263-289.
11. Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, Cohen Hubal EA, The exposure data landscape for manufactured chemicals. *Sci Total Environ.*, 2012; 1(414): 159-166.
12. Fan RE, Chen PH, Lin CJ, Working set selection using second order information for training support vector machines. *JMLR.*, 2005; 6(63): 1889-1918.
13. Farquad MAH, Bose I, Preprocessing unbalanced data using support vector machine. *Decis Support Syst.*, 2012; 53(1): 226-233.
14. Fisher AA, Patch testing with perfume ingredients. *Contact Dermatitis*, 1975; 1(3): 166-168.
15. Halla N, Fernandes IP, Heleno SA, Costa P, Boucherit Otmani Z, Boucherit K, Rodrigues AE, Ferreira ICFR, Barreiro MF, Cosmetics Preservation: A Review on Present Strategies. *Molecules*, 2018; 23(7): 1571: 1-41.
16. Heikamp K, Bajorath J, Support vector machines for drug discovery. *Expert Opin Drug Discov.*, 2014; 9(1): 93-104.
17. Hopkins ZR, Blaney L, An aggregate analysis of personal care products in the environment: identifying the distribution of environmentally-relevant concentrations. *Environ Int.*, 2016; 92(93): 301-316.
18. Hsu CW, Lin CJ, A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw.*, 2002; 13(2): 415-425.
19. Isaacs KK, Goldsmith MR, Egeghy P, Phillips K, Brooks R, Hong T, Wambaugh JF, Characterization and prediction of chemical functions and weight fractions in consumer products. *Toxicol Rep.*, 2016; 3: 723-732.
20. Kaminester LH, Allergic reaction to sunscreen products. *Arch Dermatol.*, 1981; 117(2): 66: 1-7.
21. Karschuk N, Tepe Y, Gerlach S, Pape W, Wenck H, Schmucker R, Wittern KP, Schepky A, Reuter HA, Novel *in vitro* method for the detection and characterization of photosensitizers. *PLoS One*, 2010; 5(12): e15221: 1-10.
22. Lv C, Hou J, Xie W, Cheng H, Investigation on formaldehyde release from preservatives in cosmetics. *Int J Cosmet Sci.*, 2015; 37(5): 474-478.
23. Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM, Advances with support vector machines for novel drug discovery. *Expert Opin Drug Discov.*, 2019; 14(1): 23-33.
24. Martens D, Baesens B, Gestel TV, Vanthienen J, Comprehensible credit scoring models using rule extraction from support vector machines. *EJOR.*, 2007; 183(3): 1466-1476.
25. Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B, Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. *SCI.*, 2008; 80: 33-63.
26. Phillips KA, Wambaugh JF, Grulke CM, Dionisio KL, Isaacs KK, High-throughput screening of chemicals as functional substitutes using structure-based classification models. *Green Chem.*, 2017; 19(4): 1063-1074.
27. Rogic S, Kascelan L, Customer value prediction in direct marketing using hybrid support vector machine rule extraction method. New trends in databases and information systems, Welzer IT, Eder J, Podgorelec V, Springer: Cham Switzerland, 2019; 1064; 283-294.
28. Regulation (EC) No 1223/2009 of the European Parliament and of the Council. Cosmetic Products Official Journal of the European Union, 2009.
29. Rudbaeck J, Hagvall L, Boerje A, Nilsson U, Karlberg AT, Characterization of skin sensitizers from autoxidized citronellol-impact of the terpene structure on the autoxidation process. *Contact Dermatitis*, 2014; 70(6): 329-339.
30. Sanderson M, Christopher DM, Prabhakar R, Hinrich S, Introduction to Information Retrieval. Cambridge University Press: Cambridge, England, 2008; 100-103.
31. Stivić I, Cajkovac M, O nekim pojavama nepoželjnog i štetnog djelovanja kozmetika na kožu. *Farm Glas.*, 1972; 28: 341, (available in Croatian).
32. Stüttgen G, Benefit and Risk der kosmetischen Mittel. *J Soc Cosmet Chem.*, 1981; 32: 231-245.
33. Tickner JA, Schifano JN, Blake A, Rudisill C, Mulvihill, MJ, Advancing safer alternatives through functional substitution. *Environ Sci Technol.*, 2015; 49(2): 742-749.
34. U. S. Environmental Protection Agency. Safer Choice. www.epa.gov/saferchoice.
35. U. S. Environmental Protection Agency. Program for Assisting the Replacement of Industrial Solvents, 2016.
36. Usta Hachem Y, El-Rifai O, Bou-Moughlabey Y, Echtay K, Griffiths D, Nakkash-Chmaisse H, Makki RF, Liral Fragrance chemicals lyral and lilial decrease viability of HaCat cells' by increasing free radical production and lowering intracellular ATP level: protection by antioxidants. *Toxicol in Vitro*, 2013; 27(1): 339-348.
37. Vapnik VN, The nature of statistical learning theory. Springer-Verlag: New York, USA, 2010; 138-167.
38. Voiculescu DI, Ostafe V, Isvoran A, Computational assessment of the pharmacokinetics and toxicity of the intensive sweeteners. *Farmacia*, 2021; 69(6): 1032-1041.
39. Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, Judson RS, High-throughput models for exposure-based chemical prioritization in the ExpoCast project. *Environ Sci Technol.*, 2013; 47(15): 8479-8488.