# Development of Nearest Neighbor Classifiers Identifying Dermal Sensitizers Based on a Local Lymph Node Assay Database

Shengqiao Li[1, 2], Adam Fedorowicz[2]
West Virginia University, Morgantown, WV 26506[1]
Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505[2]

**Abstract**

K Nearest Neighbor classifiers were developed to predict skin sensitization of a new chemical based on a murine local lymph node assay database of 178 organic chemicals. Two filters were compared for pre-selection of molecular descriptors. The Fisher's Discriminant Ratio filter picked a subset of descriptors which turn out to be more discriminatory than those picked by the t-test filter. Then, a step forward search method was implemented to screen out extra descriptors and simplify the classifiers based on leave-one-out accuracy. Euclidean and Mahalanobis distance metrics were also examined and the results showed the Mahalanobis distance was appropriate for this study. The 3-nearest neighbor classifier of 13 descriptors singled out by the above methods has an especially balanced performance with sensitivity of 92% and specificity of 81% for this unbalanced dataset.

**Keywords:** Nearest Neighbor Classification, KNN, QSAR, Skin Sensitization, Sensitivity, Mahalanobis Distance.

## 1. Introduction

Organic and inorganic chemicals are the predominant causative agents of occupational skin diseases[1]. Chemical agents may be either irritants or sensitizers. Contact with irritants may result in irritant contact dermatitis (ICD) and contact with sensitizers may result in allergic contact dermatitis (ACD). Irritant contact dermatitis is a nonimmunologic local inflammatory reaction following single or repeated application of a chemical substance to the same cutaneous site[2]. Allergic contact dermatitis is a cell-mediated antigen-antibody immune reaction to a specific allergen[3]. Many interactive factors including allergen, environmental and host factors contribute to ACD. Prediction of skin sensitization potential from statistical learning models could give an important guide for chemical hazard assessment. Animal based skin sensitization assays, especially the recently developed Local Lymph Node Assay (LLNA) method, produced a database for Quantitative Structure – Activity Relationship (QSAR) study to predict contact allergenicity of a chemical from a set of molecular structure and physical property descriptors. But the available dataset is still limited and many published statistical methods for classifying whether a chemical is a skin sensitizer or not are not very satisfactory. The specificity is usually low, i.e. a non-sensitizer tends to be incorrectly predicted as a sensitizer. Therefore the development of powerful and robust predictive models will be useful for skin sensitizer screening and identification so as to lower the animal experiment cost and speed up the database buildup process.

Thanks to advances in the computational chemistry, many molecule structure quantization software programs are available. A large number of descriptors of organic compound structure can be easily generated. These descriptors characterize different aspects of the molecular structure and physicochemical properties. The goal of QSAR studies is to find a small set of most informative descriptors related to a specific bioactivity. Classical regression analysis, e.g., multiple linear regression and logistic regression can be used in QSAR studies, but is challenged by the large descriptor space. Recently, the machine learning techniques, such as, Decision Tree, Random Forest,[4,5] Support Vector Machine (SVM),[6-8] Artificial Neural Network (ANN)[9-12] and k Nearest Neighbors (kNN)[13] as well as variable selection techniques are developing rapidly and the applications to QSAR studies are quite active. Nearest neighbor searching usually is slow when the dataset is large, but when more data are available for a specific problem, the kNN method will become more accurate. This is an inherent self-learning capability or evolution. In addition, data *condensing* (aka, *editing*) techniques can also be used for complexity reduction by removal of some similar training samples. Optimization algorithms may be applied to the kNN learning process to select a small set of descriptors with maximal performance. For instance, simulated annealing is customized and combined with kNN, it is called SA-kNN.[14] Simulated annealing is a stochastic optimization algorithm. The convergence is slow and the result cannot be exactly repeated because of the randomness involved in the procedure. The

implementation is also complex. The objectives of this paper are to design a simple, repeatable nearest neighbor classifier training method using different variable selection and to construct powerful classifiers that will identify potential dermal sensitizers via the molecular structure and physiochemical characters.

## 2. Dataset

The database used in this study is from the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) evaluation report of LLNA[15]. The report includes 209 chemicals. Inorganic chemicals and mixtures were excluded from this study and 178 organic chemicals form the database for kNN classifier development. In the dataset, 131 are skin sensitizers and 47 are nonsensitizers according to the report. Chemical names and CAS (Chemical Abstracts Service) Numbers are listed in table 1 and table 2. Three software packages: Cerius2 (Accelrys Inc., San Diego, CA), Dragon 4.0 (http://www.talete.mi.it) and Molconn-Z™ (eduSoft, LC, Ashland, VA), were used to compute molecular descriptors. A total of 206, 1204 and 747 descriptors were generated by them respectively. Repeated, constant and almost constant descriptors were deleted. Thus 1273 descriptors are left per chemical in the dataset. It is worth pointing out that this dataset has a large number of descriptors but a small number of observations, furthermore the two classes are unbalanced and low specificity is expected for most classification methods.

Table 1. LLNA Non-Sensitizers

| Chemical Name | CAS Number | Chemical Name | CAS Number |
|---|---|---|---|
| 2-acetamidefluorene | 53-96-3 | 2-hydroxypropylmethacrylate | 923-26-2 |
| 4-aminobenzoic acid | 150-13-0 | isopropanol | 67-63-0 |
| 3-(benzenesulphonyloxymethyl)-5,5-dimethyldihydro-2(3H)-furanone | 154750-24-0 | kanamycin A | 8063-07-8 |
| benzocaine | 94-09-7 | lactic acid | 50-21-5 |
| benzoyloxy-3,5-benzene dicarboxylic acid | 102059-70-1 | 6-methylcoumarin | 92-48-8 |
| 1-bromobutane | 109-65-9 | methyl salicylate | 119-36-8 |
| 1-bromohexane | 111-25-1 | N'-(4-methylcyclohexyl)-N-(2-chloroethyl)-N-nitrosourea | 13909-09-6 |
| 1-bromononane | 693-58-3 | neomycin | 1405-10-3 |
| 4-chloroaniline | 106-47-8 | 2-nitrofluorene | 607-57-8 |
| chlorobenzene | 108-90-7 | octadecylmethane sulphonate | 31081-59-1 |
| 3-(chlorobenzenesulphonyloxymethyl)-5,5-dimethyldihydro-2(3H)-furanone | 154750-28-4 | phenol | 108-95-2 |
| 2-chloroethanol | 107-07-3 | phthalic acid diethyl ether | 84-66-2 |
| 2,4-dichloronitrobenzene | 611-06-3 | propylene glycol | 57-55-6 |
| di-2-furanylethanedione | 492-94-4 | propylparaben | 94-13-3 |
| dimethyl isophthalate | 1459-93-4 | resorcinol | 108-46-3 |
| 5,5-dimethyl-3-(mesyloxymethyl)dihydro-2(3H)-furanone | 154750-22-8 | salicylic acid | 69-72-7 |
| 5,5-dimethyl-3-(methoxybenzenesulphonyloxymethyl)dihydro-2(3H)- | 154750-23-9 | streptomycin | 57-92-1 |
| 5,5-dimethyl-3-(nitrobenzenesulphonyloxymethyl)dihydro-2(3H)-furanone | 154750-29-5 | sulphanilamide | 63-74-1 |
| 5,5-dimethyl-3-(tosyloxymethyl)dihydro-2(3H)-furanone | 154060-50-1 | sulphanilic acid | 121-57-3 |
| ethyl methanesulfonate | 62-50-0 | tartaric acid | 87-69-4 |
| geraniol | 106-24-1 | tixocortol-21-pivalate | 55560-96-8 |
| hexane | 110-54-3 | trimethylammonium-3-tolyl-ε-caprolactamide chloride | 374680-04-3 |
| hydrocortisone | 50-23-7 | α-trimethylammonium-4-tolyloxy-4-benzenesulphonate | 264869-81-0 |
| 4-hydroxybenzoic acid | 99-96-7 | | |

Table 2. LLNA Sensitizers

| Chemical Name | CAS Number | Chemical Name | CAS Number |
|---|---|---|---|
| abietic acid | 514-10-3 | diethyl sulfate | 64-67-5 |
| 2-(N-acetoxy-acetamido)fluorene | 6098-44-8 | diethylenetriamine | 111-40-0 |
| 3-acetylphenyl benzoate | 139-28-6 | 3,4-dihydrocoumarin | 119-84-6 |
| 4-allylanisole | 140-67-0 | dihydroeugenol | 2785-87-7 |
| ammonium thioglycolate | 5421-46-5 | 2,4-dinitrochlorobenzene | 97-00-7 |
| 2-aminophenol | 95-55-6 | 2,4-dinitrofluorobenzene | 70-34-8 |
| 3-aminophenol | 591-27-5 | 2,4-dinitrothiocyanobenzene | 1594-56-5 |
| aniline | 62-53-3 | 7,12-dimethylbenz[a]anthracene | 57-97-6 |
| 1,2-benzisothiazol-3(2H)-one | 2634-33-5 | 5,5-dimethyl-3-(bromomethyl)dihydro-2(3H)-furanone | 154750-20-6 |
| benzopyrene | 50-32-8 | 5,5-dimethyl-3-(thiocyanatomethyl)dihydro2(3H)-furanone | 154750-32-0 |
| 1,4-benzoquinone | 106-51-4 | 5,5-dimethyl-3-methylenedihydro2(3H)-furanone | 29043-97-8 |
| benzoyl chloride | 98-88-4 | N,N-dimethyl-1,3-propanediamine | 109-55-7 |
| benzoyl peroxide | 94-36-0 | dimethyl sulfostearate | 99785-70-3 |
| benzyl bromide | 100-39-0 | dimethyl sulphate | 77-78-1 |
| 12-bromo-1-dodecanol | 3344-77-2 | disodium 1,2-diheptanoyloxy-3,5-benzenedisulfonate | 374678-48-5 |
| 12-bromododecanoic acid | 73367-80-3 | ethylene glycol dimethacrylate | 97-90-5 |
| 1-bromododecane | 143-15-7 | ethylenediamine | 107-15-3 |
| 1-bromoheptadecane | 3508-00-7 | 1-ethyl-3-nitro-1-nitrosoguanidine | 4245-77-6 |
| 1-bromohexadecane | 112-82-3 | N-ethyl-N-nitrosourea | 759-73-9 |
| 1-bromooctadecane | 112-89-0 | eugenol | 97-53-0 |
| 1-bromopentadecane | 629-72-1 | fluorescein isothiocyanate | 25168-13-2 |
| 1-bromotetradecane | 112-71-0 | formaldehyde | 50-00-0 |
| 7-bromotetradecane | 74036-97-8 | glyoxal | 107-22-2 |
| 2-bromotetradecanoic acid | 10520-81-7 | hexadecanoyl chloride | 112-67-4 |
| 1-bromotridecane | 765-09-3 | hexyl cinnamic aldehyde | 101-86-0 |
| 1-bromoundecane | 693-67-4 | hydroquinone | 123-31-9 |
| 2,3-butanedione | 431-03-8 | hydroxycitronellal | 107-75-5 |
| butyl glycidil ether | 2426-08-6 | 2-hydroxyethyl acrylate | 818-61-1 |
| chloramine T | 127-65-1 | imidazolidinyl urea | 39236-46-9 |
| chlorpromazine | 50-53-3 | 1-iodohexadecane | 544-77-4 |
| 2-chloromethylfluorene | 91679-67-3 | 1-iodohexane | 638-45-9 |
| 1-chloromethylpyrene | 1086-00-6 | 1-iodononane | 4282-42-2 |
| 5-chloro-2-methyl-4-isothiazolin-3-one | 26172-55-4 | 1-iodooctadecane | 629-93-6 |
| 1-chlorononane | 2473-01-0 | 1-iodotetradecane | 19218-94-1 |
| 1-chlorooctadecane | 3386-33-2 | isoeugenol | 97-54-1 |
| 1-chlorotetradecane | 2425-54-9 | isononanoyloxybenzene sulphonate | 109363-00-0 |
| cinnamic aldehyde | 104-55-2 | isophorone diisocyanate | 4098-71-9 |
| citral | 5392-40-5 | 2-mercaptobenzothiazole | 149-30-4 |
| clotrimazole | 23593-75-1 | 2-methoxy-4-methylphenol | 93-51-6 |
| cocoamidopryl betaine | 61789-40-0 | 4-methylaminophenol sulfate | 55-55-0 |
| dodecyl methanesulphonate | 51323-71-8 | 3-methylcatechol | 488-17-5 |
| dodecylthiosulphonate | 127089-67-2 | 4-methylcatechol | 452-86-8 |
| 1,2-dibromo-2,4-dicyanobutane | 35691-65-7 | 3-methylcholantrene | 56-49-5 |
| 3-methyleugenol | 186743-26-0 | 3-phenylenediamine | 108-45-2 |
| 5-methyleugenol | 186743-25-9 | 4-phenylenediamine | 106-50-3 |
| 6-methyleugenol | 186743-24-8 | phthalic anhydride | 85-44-9 |
| methyl dodecanesulfonate | 2374-65-4 | picryl chloride | 88-88-0 |
| methyl hexadec-2-ene sulfonate | 54612-23-6 | β-propiolactone | 57-57-8 |
| methyl methanesulfonate | 66-27-3 | propyl gallate | 121-79-9 |
| 2-methyl-4,5-trimethylene-4-isothiazolin-3-one | 82633-79-2 | 1-propyl-3-nitro-1-nitrosoguanidine | 13010-07-6 |
| 3-methoxyphenylbenzoate | 5554-24-5 | p-xylene | 106-42-3 |
| 1-methyl-3-nitro-1-nitrosoguanidine | 70-25-7 | pyridine | 110-86-1 |
| methylene diphenyl diisocyanate | 101-68-8 | sodium 4-(2-ethylhexyloxycarboxy)benzenesulfonate | 264869-77-4 |
| N-nitroso-N-methylurea | 684-93-5 | sodium 4-sulfophenyl acetate | 46331-24-2 |
| nonanoyl chloride | 764-85-2 | sodium benzoyloxy-2-methoxy-5-benzenesulfonate | 159783-19-4 |
| α-naphthoflavone | 604-59-1 | sodium benzoyloxybenzenesulfonate | 56265-04-4 |
| β-naphthoflavone | 6051-87-2 | sodium norbornanacetoxy-4-benzenesulfonate | 374679-08-0 |
| 4-nitrobenzyl bromide | 100-11-8 | tetrachlorosalicylanilide | 1154-59-2 |
| 4-nitrobenzyl chloride | 100-14-1 | tetramethyl thiuram disulfide | 137-26-8 |
| 4-nitroso-N,N-dimethylaniline | 138-89-6 | 1-thioglycerol | 96-27-5 |
| octadecanoyl chloride | 112-76-5 | 2,4,5-trichlorophenol | 95-95-4 |
| octyl gallate | 1034-01-1 | 2,4,6-trichloro-1,3,5-triazine | 108-77-0 |
| oxazolone | 1564-29-0 | trimellitic anhydride | 552-30-7 |
| penicillin G | 61-33-6 | 3,5,5-trimethylhexanoyl chloride | 36727-29-4 |
| pentachlorophenol | 87-86-5 | 4-vinylpyridine | 100-43-6 |
| phenyl benzoate | 93-99-2 | | |

## 3. Methods

K Nearest neighbor classification is a nonparametric classification method. It does not assume any parametric form for the distribution models of the measured random variables from a population. Due to the flexibility of nonparametric model, it is usually a good classifier for many situations where the joint distribution is hard to model parametrically, especially for QSAR studies where it is common that hundreds even thousands of variables are involved initially.

The mechanism of kNN is very straightforward comparing to other methods. kNN is instance based reasoning. The philosophy of nearest neighbor classification is that two subjects in the same class also have some kind of similarity measurable by a distance metric through some variables of different information, and vice versa. An unknown class case is classified into the same category as its first NN or the wining class by the votes of k nearest neighbors, where $k$ is an odd number, such as 1, 3, 5 or 7, etc. This $k$ is usually chosen by cross-validation. The nearest neighbors are found by a distance or dissimilarity measure. Therefore, the design of NN classifiers involves three key elements, a distance metric, $d$, the number of nearest neighbors, $k$, and selection of a set of variables for distance computation. In the following sections, these three components will be discussed.

### 3.1 Distance Metrics

The Euclidean distance is widely used and can be generalized by the Minkowski distance, $L_q$ norm. Let $\mathbf{v}$ be the difference vector between vector $\mathbf{x}$ and $\mathbf{y}$, then $L_q = \| \mathbf{v} \|_q = \left( \sum_{i=1}^{p} | v_i |^q \right)^{\frac{1}{q}}$. $L_1$ is Manhattan distance or City-Block distance and $L_2$ is Euclidean distance. $L_\infty = \| \mathbf{v} \|_\infty = \max_{1 \le i \le p} | v_i |$, is the maximum of the absolute value of each element of $\mathbf{v}$. Minkowski distances make sense when there is commensurability between different variables. In reality, QSAR data involve lots of noncommensurate variables. Data standardization is the common way to reweight the variables of different units to a comparable level. There are also measures of similarity between two cases related to the Pearson correlation coefficient $r$ in the sample space. The transformations of $r$ could be used as distance measures, such as $1 - r$ and $\sqrt{1 - r^2}$ etc.

Aforementioned distance metrics do not take into account the correlation in the variable space. Another type of distance, the Mahalanobis distance, generalizes the Euclidean distance in another way and rectifies the correlations between variables in the distance calculation. The Mahalanobis distance is defined as $Mah(\mathbf{v}) = \sqrt{\mathbf{v}'\mathbf{M}\mathbf{v}}$, where $\mathbf{M}$ is a positive semidefinite $p \times p$ matrix. If $\mathbf{M}$ is an identity matrix, the distance is just the usual Euclidean distance. If the inverse of the variance-covariance matrix $\mathbf{S}$ is taken as $\mathbf{M}$, the Mahalanobis distance results. Actually Mahalanobis distances not only account for the dependency but also standardize the measure units. So sometimes it is preferable. When the dimensionality is high, $\mathbf{S}$ can be singular and thus is noninversible. In this case, a quasi-inverse will be used. The quasi-inverse is obtained by setting all zero eigenvalues to the smallest positive eigenvalue $\lambda_s$ of $\mathbf{S}$. The quasi-inverse $\mathbf{S}^{-1}$ is $\mathbf{VWV}'$, where $\mathbf{V}$ is the matrix composed of eigenvectors and $\mathbf{W}$ is a diagonal matrix with diagonal elements $w_j = 1/\max(\lambda_j, \lambda_s), j = 1, \ldots, p$.

In addition to the distance metrics discussed above, researchers also proposed more complicated measures based on these for nearest neighbor algorithm, such as, local distance measure,[16] weighted distance,[16] etc. In this study, we examined Euclidean and Mahalanobis distances.

### 3.2 Variable Selection

The high-dimensional QSAR data challenges all classification methods. If too many variables are included in the model, the overfitting problem will be severe and the classification rules result in a loss of flexibility. With limited data points, in high-dimensional space, data points will scatter sparsely and the nearest neighbor concept will not make sense. So the variable selection is an important aspect of a nearest neighbor classifier. Variable selection is a research active field for machine learning. There are two types[17]: unwrapped and wrapped methods according to whether the classification algorithm is involved or not. Unwrapped variable selection is also known as a filtering method. This kind of method is independent of classification algorithms, and only some kind of intrinsic discriminative information is used. On the other hand, wrapped methods are supervised by the specific classification algorithm in order to maximize the classification performance, such as accuracy, of the algorithm. Both of these types of methods were used for this study.

Unwrapped methods include statistical hypothesis tests, such as t-test, ANOVA etc. and the so-called class separatability criteria, such as divergence,

Bhattacharyya distance, Fisher's discriminant ratio (FDR) and mutual information etc. Kovatcheva[18] et al applies correlation analysis for descriptor reduction. Unwrapped methods can be applied onto each variable simultaneously and sequentially. In the latter way, the correlation between a new variable and selected variables acts as a penalty term for computing the measure.

The unwrapped method usually is used as a fast screening approach. Through this method, most of the variables that are not useful are removed from the input dataset. The remaining variables are passed to the classifier wrapped variable selection. In this study, we implemented sequential filters as follows:

1. Compute the filter criterion $C_i$ for each descriptor $\mathbf{x}_i$,
2. Select the first descriptor of maximum $C_i$, put it into the selected descriptor list $L$,
3. For each unselected descriptor $\mathbf{x}_i$, compute Pearson correlation coefficient $r_{ij}$ with descriptor $\mathbf{x}_j$ in $L$ and the mean, $r_i$, of absolute $r_{ij}$.
4. Choose the descriptor $\mathbf{x}_j$ which maximize $C_i - a \cdot r_i$, where $a$ is a penalty coefficient and $a \cdot r_i$ act as a penalty term, put into $L$.
5. Repeat step 3 and 4 until the predefined number of descriptors are selected.

Finding the optimal subset of variables for the final classifier is usually very difficult and impractical. Heuristic search methods, such as forward/backward and stepwise selection, are used in the practice of wrapped variable selection as in classical regression analysis. Heuristic search is tractable and usually gives satisfying results. In this study, we implemented step forward variable selection for kNN. The algorithm works in the following way:

1. For each descriptor in the data, conduct an NN classification and assess the performance, choose the best one by the performance criteria.
2. With the descriptor(s) chosen, iteratively add a new descriptor from those remaining and evaluate the classifier on the augmented descriptor set. Keep the new one with the best classification performance.
3. Repeat the forward step until a pre-specified number of descriptors are selected.

### 3.3 Performance Assessment

If the dataset is large, it can be split into two independent parts: a training set and a testing set. In this study, the data set is small and is not class balanced. We used Leave-One-Out-Cross Validation (LOOCV). The mean accuracy, i.e., the average of sensitivity and specificity, is computed. This is LOOCV accuracy and was used as the criterion to compare different NN classifiers.

## 4. Results and Discussion

Two different filters, Welch's t-test and Fisher's Discriminant Ratio (FDR), which is defined as the ratio of squared class mean difference to sum of the class variances, were used to select 200 descriptors and then the step forward selection method was applied to NN classifiers. The number of descriptors for final classifiers is chosen as 13 for this dataset. This is based on the rule that the ratio of number of sensitizers to number of descriptors is around 10 which is conservative compared with the rule that sets the ratio of total number of chemicals to number descriptors to 10. The stopping criterion of no performance improvement is not adopted here due to premature termination of the forward selection procedure resulting in classifiers of too few descriptors.

Euclidean distance and Mahalanobis distance were compared. Euclidean distance was also compared for standardized and nonstandardized descriptors. Data standardization does not have any impact on the two filters since the t statistic, FDR and correlation are invariant to data centering and scaling. The number of neighbors was determined empirically. We found that when 5 or more neighbors were used for classification, the specificity was too low though the sensitivity was very high. This may be caused by the imbalance of the two classes in the dataset. The sensitizers are predominant in the whole dataset and so they are in the 5 or more neighbors of a chemical. Therefore, the appropriate number of neighbors should be less than 5. The results are reported in table 3.

From the table 3, it can be seen that the effects of filter, distance and number of neighbors are interactive. The descriptor standardization for Euclidean distance does increase the specificity for the 1-NN classifiers but not for the 3-NN classifiers. The best classifier is the 3-NN classifier using Mahalanobis distance with FDR descriptor pre-selection. Its sensitivity is >90% and specificity >80%. This performance is highly satisfactory for this study. The descriptors used in this classifier are listed in table 4. In other cases, the Mahalanobis is not better than Euclidean distance on standardized descriptors.

The significance of this 3-NN model is tested by randomization. The responses are reshuffled 1000 times and thus 1000 random datasets are generated. The 3-NN model is then applied these datasets. The sensitivity and specificity are calculated for such random datasets. All the specificities are less than the actual dataset specificity, 80.9%, i.e. p-value = 0.000 and only one sensitivity is greater than the actual dataset sensitivity, 91.6%, i.e. the corresponding p-value is 0.001. So it is shown the 3-NN model is significantly better than a random model.

Table 3. Nearest Neighbor Classification Results

| Filter | Distance | 1 Nearest Neighbor | | | | 3 Nearest Neighbors | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sensitivity | Specificity | False Positive | False Negative | Sensitivity | Specificity | False Positive | False Negative |
| Welch's t-test | Euclidean | 88.5 | 59.6 | 14.1 | 35.0 | 89.3 | 68.1 | 11.4 | 30.5 |
| | Euclidean (Standardized descriptors) | 87.8 | 76.6 | 8.7 | 30.7 | 91.6 | 68.1 | 11.1 | 25.6 |
| | Mahalanobis | 87.0 | 76.6 | 8.8 | 32.1 | 89.3 | 66.0 | 12 | 31.1 |
| FDR | Euclidean | 89.3 | 57.4 | 14.6 | 34.2 | 91.6 | 74.5 | 9.1 | 23.9 |
| | Euclidean (Standardized descriptors) | 86.3 | 74.5 | 9.6 | 33.9 | 90.1 | 61.7 | 13.3 | 30.9 |
| | Mahalanobis | 90.8 | 66.0 | 11.8 | 28.0 | 91.6 | 80.9 | 7.0 | 22.4 |

Table 4. Descriptors Used for the 3-NN Classifier Based on Mahalanobis Distance

| No. | Name | Definition | Class |
|---|---|---|---|
| 1 | Atype_C_2 | Counts of Carbon atom type CH2R2 | Atom type counter |
| 2 | Atype_H_52 | Counts of Hydrogen atom type HC(OR)(R)-CXR2 | Atom type counter |
| 3 | DELS | Molecular electrotopological variation | Topological descriptors |
| 4 | dXv1 | Difference Valence First Order Chi Index | Molecule Connectivity |
| 5 | H0p | H autocorrelation of lag 0 / weighted by atomic polarizabilities | GETAWAY |
| 6 | HATS3e | Leverage-weighted autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities | GETAWAY |
| 7 | R4e+ | R maximal autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities | GETAWAY |
| 8 | LogP | Logarithm of the partition coefficient | Physicochemical properties |
| 9 | Mor06u | 3D-MoRSE - signal 06 / unweighted | 3D-MoRSE |
| 10 | Mor23e | 3D-MoRSE – signal 23 / weighted by atomic Sanderson electronegativities | 3D-MoRSE |
| 11 | nssCH2 | Count of $-CH_2-$ | Atom-type E-State index |
| 12 | nssO | Count of –O– | Atom-type E-State index |
| 13 | RDF020u | Radial Distribution Function - 2.0 / unweighted | Radial Distribution Function |

## 5. Conclusion

The nearest neighbor classifier developed through unwrapped and wrapped variable selection demonstrated balanced and highly accurate prediction based on leave-one-out cross validation. The sequential filter with correlation adjustment could select a subset of descriptors with predictive potential. The correlation adjustment for filters alleviated the redundancy and increased the relevance of the selected descriptors. The filter based on the Fisher's discriminant ratio performed well. The forward search method could find a parsimonious final set of descriptors with best performance heuristically. The best model, a 3-NN classifier, was obtained using Mahalanobis distance,

which takes account the correlation information between descriptors. In summary, the 3-NN classifier, based on Mahalanobis distance using 13 descriptors has satisfactory performance with a sensitivity of 92% and a specificity of 81%. It is a promising classifier for skin sensitization activity of an organic chemical.

**Acknowledgements**

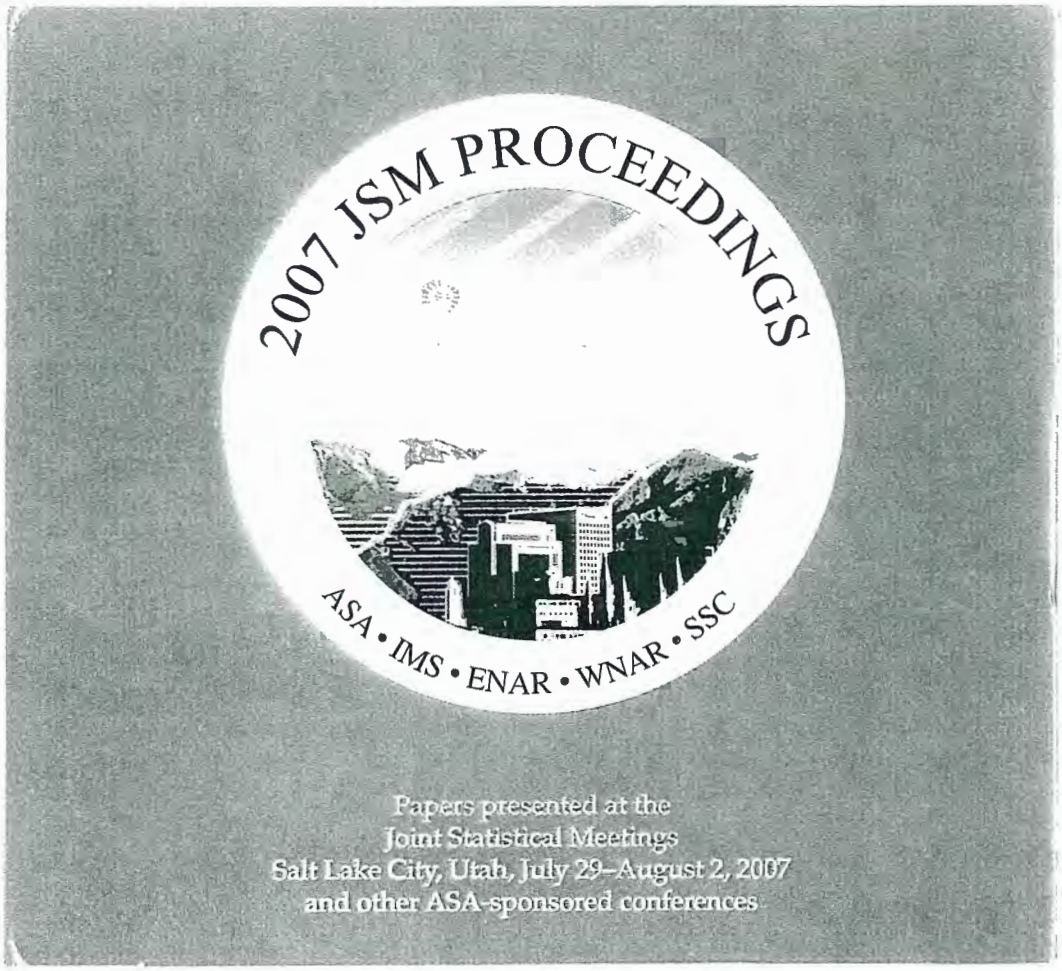
The authors thank Cecil Burchfiel, and Michael Andrew for their valuable comments and helpful discussions which improved the quality of this paper.

**Disclaimer:** *The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.*

**References**

1. *Fundamentals of Industrial Hygiene;* 5th ed.; National Safety Council: Chicago, 2001.
2. Nethercott, J. R.; Holness, D. L. Occupational allergic contact dermatitis. *Clin. Rev. Allergy* **1989,** *7* (4), 399-415.
3. Katz, S. I. Mechanisms involved in allergic contact dermatitis. *J. Allergy Clin. Immunol.* **1990,** *86* (4 Pt 2), 670-671.
4. Breiman, L. Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.
5. Li, S.; Fedorowicz, A.; Singh, H.; Soderholm, S. C. Application of the random forest method in studies of local lymph node assay based skin sensitization data. *J. Chem. Inf. Model.* **2005,** *45* (4), 952-964.
6. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995,** *20* (3), 273-297.
7. Vapnik, V. N. *Statistical learning theory;* Wiley: New York, 1998.
8. Vapnik, V. N. *The nature of statistical learning theory;* 2nd ed.; Springer: New York, 2000.
9. So, S. S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997,** *40* (26), 4360-4371.
10. So, S. S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997,** *40* (26), 4347-4359.
11. So, S. S.; Karplus, M. Genetic neural networks for quantitative structure-activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABAA receptors. *J. Med. Chem.* **1996,** *39* (26), 5246-5256.
12. So, S. S.; Karplus, M. Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.* **1996,** *39* (7), 1521-1530.
13. Cover, T.; Hart, P. Nearest Nieghbor Pattern Classification. *IEEE Transaction on Information Theory* **1967,** *IT-13* (1), 21-27.
14. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002,** *45* (13), 2811-2823.
15. Haneke, K. E.; Tice, R. R.; Carson, B. L.; Margolin, B. H.; Stokes, W. S. ICCVAM evaluation of the murine local lymph node assay. III. Data analyses completed by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods. *Regul. Toxicol. Pharmacol.* **2001,** *34* (3), 274-286.
16. Short, R.; Fukunaga, K. The Optimal Distance Measure for Nearest Neighbor Classification. *IEEE Transaction on Information Theory* **1981,** *IT-27* (5), 622-627.
17. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *Journall of Machine Learning Research* **2003,** *3* (Mar), 1157-1182.
18. Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (2), 582-595.


2007 JSM Proceedings
American Statistical Association
INSTRUCTIONS:
This program will run automatically or run START.htm.
MIRA
DIGITAL PUBLISHING
Papers presented at the
Joint Statistical Meetings
Salt Lake City, Utah, July 29–August 2, 2007
and other ASA-sponsored conferences
IndexSMART™ Manufactured by Mira Digital Publishing. For support, email cdsupport@mirasmart.com or call (866) 341-9590.


2007 JSM PROCEEDINGS
ASA • IMS • ENAR • WNAR • SSC
Papers presented at the
Joint Statistical Meetings
Salt Lake City, Utah, July 29–August 2, 2007
and other ASA-sponsored conferences

## License:

User is granted permission to use the software on a single computer. No additional copies of the CD-ROM in its entirety may be made in any media for archiving, distribution, or publication. Authors maintain the copyright of their individual papers.

## To Run the Software Program:

The program will run automatically or run START.htm.

American Statistical Association
732 North Washington Street
Alexandria, VA 22314-1943
(703) 684-1221 Fax: (703) 684-3410
Email: *asainfo@amstat.org*, Web site: *www.amstat.org*
ISBN 978-0-9791747-4-2