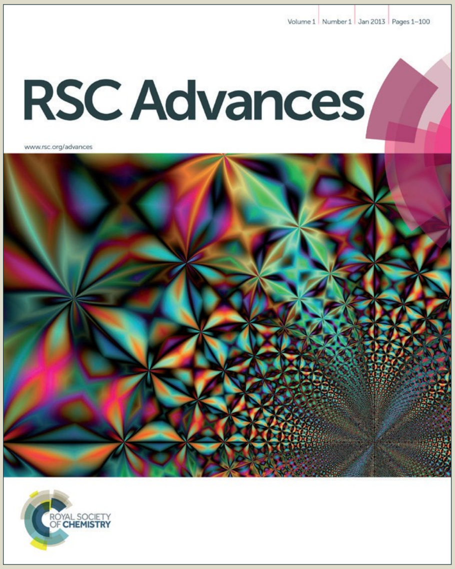
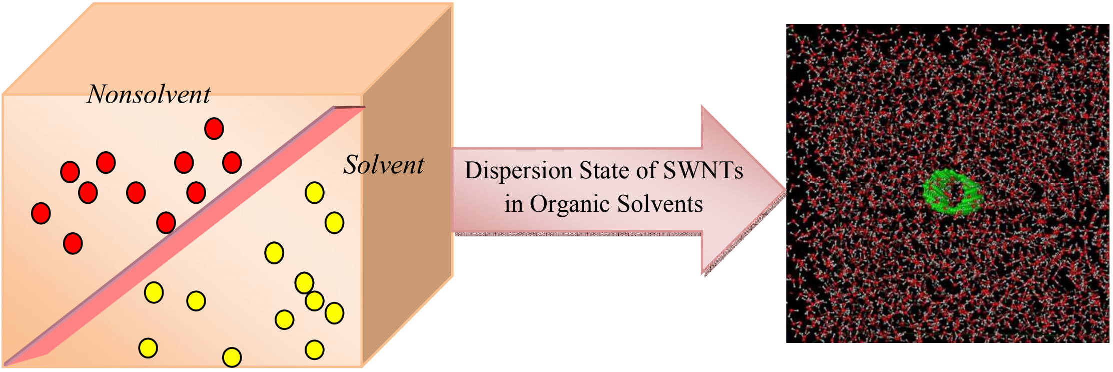
RSC Advances


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
RSC Advances
www.rsc.org/advances
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances


Nonsolvent
Solvent
Dispersion State of SWNTs
in Organic Solvents

Application of Classification Models to Identify Solvents for Single-Walled Carbon Nanotubes Dispersion

M. Salahinejad

Environmental Laboratory, NSTRI, Tehran, Iran

E-mail: salahinejad@gmail.com

Abstract

In this study, a list of classification models was developed to categories organic solvents with respect to their dispersibility of single-walled carbon nanotubes (SWNTs). The organic solvents were classified into solvent and nonsolvent based on the ability to disperse the SWNTs. Various feature selection techniques combined with different classifier algorithms of linear and quadratic discriminate analysis (LDA and QDA), decision trees (random forest and J48), neural networks and support vector machine (SVM) were explored on a data set consisting of the structurally diverse organic solvents. The physicochemical descriptors such as partial charges, volsurf (the volumes and surfaces of grid points at different energy levels), subdivided surface area and some shape descriptors contributed to the classification models. The validation studies using test set, leave-one-out and 10-fold cross-validation methods provide statistical parameters such as specificity, sensitivity, accuracy, Mathew´s correlation coefficient and Kappa index to evaluate the developed classification models. The sum of ranking difference (SRD) procedure reveals that the random forest classifier based on selected descriptors by the wrapper feature selection method is the best classification model, while the SVM, MLP and QDA containing models that are ranked as good models. The structural features along with electrostatic interactions of solvent molecules play the significant role in discriminating good solvents from nonsolvents in SWNTs dispersion.

Keywords: Classification, Single-walled carbon nanotube (SWNT), Organic solvent, Dispersibility

1. Introduction

Single-walled carbon nanotubes (SWNTs) with extraordinary thermal, mechanical, optical and electrical properties [1] have been identified as promising nanomaterials in many fields including: biomedical,[2] drug delivery systems and cancer therapy,[3] energy storage devices,[4] composites fillers,[5] nanoprobes and sensors[6, 7] and catalyst.[8] Because of high polarizability, hydrophobic and smooth surface of SWNTs, they always tend to aggregate into mixture of bundles or ropes of various species, thus preventing their use in many applications.[9, 10] Therefore exfoliation and debundling of SWNTS to isolated ones, usually in a liquid phase, is necessary before use in various areas[11] and attracting great interests as one of the main challenges in carbon nanotube research.

Stable dispersion with the aid of surfactants, biomolecules [12] and organic polymers [13, 14] is the most common "solubilization" method of SWNTs in different aqueous and organic media. However, these procedures tend to degrade of the SWNTs' electronic properties[15] and make difficulties for completely removal of these solubilizing agents from nanotubes.[16] Thus, the direct dispersion of nanotubes into proper organic solvent have the potential advantages involving the ability to remove the solvent through evaporation and to find suitable purification and dispersion methods.

Attempts to identify the optimal solvent properties have been based on solubility parameters such as Hildebrand or Hansen parameters[17-21] or surface energy.[17, 22] Ham et al. explored the relation between the Hansen solubility parameters and the degree of dispersion state of SWNTs in various organic solvents.[21] The dispersion state of organic solvents classified as three groups of dispersed, swollen and sedimented of SWNTS based on the dispersive component of the solubility parameter. Bergin et al. measured the dispersibility of SWNTs in a range of organic solvents and explored the nanotube dispersibility based on the

Hansen and surface energy solubility parameters of the solvents.[23] The organic solvents classified as solvent and nonsolvent. Nonsolvents are defined as solvents with effectively zero of SWNTs dispersibility. However, it was concluded that neither Hansen nor surface energy solubility parameters were fundamental to distinct between solvents and nonsolvents and to evaluate and predict the dispersion state for SWNTs in different organic solvents.

In previous attempt, we investigated the application of quantitative-structure property relationship (QSPR) models to predict the dispersibility of SWNTs in various organic solvents.[24, 25] This work aims to develop in-silico classification models, which can be used to classify the organic solvents based on their SWNTs dispersibility and to explore the important structural features related to the dispersion of carbon nanotubes. Various feature selection and classification techniques were used to compare different chemometrics tools to approach the difficult problems of predicting the dispersion state of organic solvents for SWNTs.

2. Methodology

2.1 Dataset

The data on dispersibility of as-produced HiPCO SWNTs in different organic solvents extracted from the work of Bergin et.al.[17] The dispersions of these nanomaterials were determined by measuring the dispersion absorbance as concentration ($C_{\max}$) of SWNTs after sonication and a mild centrifugation. A detailed description of the method was presented in references 22 and 26.[22, 26] Table 1of supporting information displays a complete list of the simplified molecular input line entry specification (SMILES) and molecular structures of the solvents. The compounds of 1-29, which are able to disperse SWNTs considered as solvent and the others (30-59) where no SWNTs was reliably detected after centrifugation considered as nonsolvets. An additional three new amidine compounds (compounds 60-62), extracted from the work of Barman et al.[27] 1,8-diazabicycloundec-7-ene (DBU) was reported as solvent and the 1,5-diazabicyclo(4.3.0) non-5-ene (DBN) and 1,1,3,3-tetramethylguanidine (TMG) were reported as nonsolvents for SWNTs. This data set contains 30 solvents and 32 nonsolvents for SWNTs.

The Kennard-Stone (KS) algorithm was applied to split the data set into training and test sets. The KS method is usually performed on the matrix of molecular descriptors (X) based on Euclidean distance measure most representative objects. In modified KS, the response vector (y) was added as an additional column to the matrix of descriptors (X). This modified method, KS(Xy), help to evenly distribution of samples within both descriptors and response spaces[28] and can enhance the influence of the response on the splitting results.[28] A training set of 48 compounds was used to build and adjust the parameters of the classification models, and the rest of the molecule (14 compounds) was used to evaluate classification model's prediction ability as test set.

2.2 Molecular descriptor calculation

The MOE (Chemical Computing Group Inc.) v2010.10 was used to generate different kinds of descriptor, including physical properties, potential energies, partial charge, surface area, volume and shape indices, adjacency and distance matrices descriptors and conformation-dependent charge indices of solvent molecules. After elimination of high correlation descriptors and those contained only zero or constant values for all solvents, more than 150 descriptors were initially used in this study before applying feature selection methods. The physicochemical descriptors of the solvents were used as independent variables in the classification studies.

2.3 Feature selection techniques

We used different wrapper and filter feature selection (FS) techniques to select the most relevant molecular descriptors for the classification models of organic solvents based on their ability to disperse SWNTs. We experimented with several evaluators and search methods for finding final set of features within the Weka software[29] including:

- Correlation-based feature selection (Cfs) subset evaluator (CfssubsetEval) with best first search method, which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of descriptors that are highly correlated with the class while having low inter-correlation are preferred. The best first search method searches the space of descriptor subsets by greedy hill-climbing augmented with a backtracking facility.[29]
- Relief-F attribute evaluator (reliefFAttributeEval) uses instance based learning to assign a relevance weight to each feature that each feature's weight reflects its ability to distinguish among the class values.[29]
- Information gain (InfoGain) attributes evaluator (InfoGainAttributeEval) uses information gain to select attributes by measuring information gain with respect to the class.[30]
- Wrapper attributes subset evaluator (WrapperSubsetEval) which uses the method of classification itself to measure the importance of features set.[31]

2.4 Classification methods

The performance of different classification algorithm such as discriminant analysis (DA), decision trees (DT), radial basis function (RBF) and multilayer perceptron (MLP) networks was examined and compared to identify solvent and nonsolvent compounds for SWNTs dispersion.

The DA classification as known and classic method among traditional classifiers, performs dimensionality reduction by maximizing the between-class variance and minimizing the within-class variance.[32] Linear discriminant analysis (LDA), can only consider linear boundaries while quadratic discriminant analysis (QDA) separates the class regions by quadratic boundaries.[33]

The DT classifiers algorithm are effective and powerful tools for classification which are in the form of a tree structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes.[34] Random Forest (RF) classifier uses a collection of decision trees, in order to improve the classification rate while J48 tree algorithm basically uses the divide-and-conquer algorithm by splitting a root tree into a subset of two partitions of child nodes.[35]

The feed forward multilayer networks or multilayer perceptrons (MLPs) and radial basis function networks (RBFN) are two if the most widely used neural network classifiers, which are based on the training procedure by an activation function that associates input vectors with a corresponding target vector.[36] The MLP uses one or more hyper planes to isolate the classes in the input space, while RBFs use a local approach, which model the separate class distributions by localized radial basis functions. Support vector machine learning classification are based on the concept of separating planes that define decision boundaries.[37, 38]

2.5 Statistical significance of classification models

Different variable selection and classifiers techniques were applied on the training set and then the developed models were validated using the test set samples. Some parameters such as true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), extracted from the confusion matrix of the actual class versus predicted class, are used to assess the quality of a binary classification. Goodness-of-fit parameters were estimated for the models on the basis of sensitivity, specificity and accuracy on the training set. Sensitivity (Sn) describes the fraction of solvent molecules correctly classified, specificity (Sp) defines the fraction of nonsolvent molecules correctly classified, and accuracy (ACC) is the fraction of molecules correctly classified. Sensitivity, specificity and accuracy are calculated as follows, respectively:

$$Sn = TP/(TP + FN) \qquad (1)$$

$$Sp = TN/(TN + FP) \qquad (2)$$

$$ACC = (TP + TN)/(TP + TN + FP + FN) \qquad (3)$$

Matthew's correlation coefficient (MCC) and Cohen's kappa values are two statistics used to validate the predictive performance of classification models. The MCC is computed as below:

$$\mathrm{MCC} = ((TP \times TN) - (FP \times FN))/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \qquad (4)$$

The MCC takes a range of values from +*1* to -*1*, where +1 represents a perfect prediction and -1 an inverse prediction. The Kappa statistic is a metric that compares an observed accuracy with an expected accuracy. The expected accuracy is defined as the accuracy that would be expected to be present by chance alone.[39, 40] The equation used for computing Kappa coefficient, k, is expressed as below:

$$k = (P_o - P_e)/(1 - P_e) \qquad (5)$$

where P_o and P_e are the numbers of observed and expected compounds, respectively. the kappa values more than 0.8 are considered as excellent, 0.40-0.8 as fair to good, and less than 0.40 as poor.[41]

In order to evaluate the ability of classification models, leave-one-out cross-validation (LOO-CV) and 10-fold cross-validation was performed.

2.6 Software

Feature selection methods, DTs, RBF, MLP and LibSVM classifications were performed using the Weka software.[29] The LDA and QDA were calculated by means of the classification toolbox[42], available at http://michem.disat.unimib.it/chm/. Calculations were performed using MATLAB 7.6 (MathWorks, Inc., Natick, MA, USA).

2.7 Sum of ranking difference (SRD)

The analysis of ranking of the classification models were performed using the program available at http://knight.kit.bme.hu/CRRN/ .[43, 44] The average vector corresponding to the calculated statistical parameters such as Sp, Sn, ACC, MCC and k-coefficient were used for the SRD analysis of the developed classification models.

3. Results And Discussion

3.1 Classification models

Classification methods are fundamental chemometrics techniques designed to find mathematical models able to recognize the membership of each object to its proper class on the basis of a set of measurements [33]. Feature selection is of considerable importance in classification procedures. These techniques provide three main benefits: reducing the computational complexity, improving the prediction performance of classification models and providing a better understanding of the underlying process.[45]

Classification models with the ability to predict the SWNTs dispersibility of organic solvents were developed based on physicochemical descriptors calculated from MOE software as independent variables. A single model or method will not give the best result for any set of data, hence multiple classification models with different approaches were constructed to compare their results. Three filter feature selection methods, as classifier-independent techniques, based on a specific criteria such as correlation (Cfs), distance (ReliefF) and information (InfoGain) and a wrapper method (Wrapper), which are based on the performance of a particular classifier, were examined for selecting the most effective subset of features. Table 1 displayed the abbreviation and a brief description of the selected features for each method.

The features selected by aforementioned FS methods were applied to different classification techniques, namely DA, DT, RBF, MLP and SVM classifiers. Tables 2-4 gave the statistical performance results of the obtained classification models based on different feature selection techniques and classifiers. The statistical parameters such as Sp, Sn, ACC and MCC were calculated from the confusion matrix of the training and test set and the LOO and 10-fold cross-validation techniques for all developed classification models. Here, Sn is the ability of the classification model to correctly recognize solvent compounds as solvent and Sp is a measure of the classification model to identify nonsolvent compounds as nonsolvent (both in percentage). All the developed classification models were validated by LOO and 10-fold cross-validation and test set methods.

Sum of ranking difference (SRD) values were calculated for all classification models in order to compare the performance of each method. The SRD carried using simulated random numbers in conjunction with the theoretical distribution of the SRD values called comparison of ranks with random numbers (CRRN) procedure. Table 5 provides the SRD and p% interval of the variables of the classification analyses while Fig. 1. displayed the SRD-CRRN test results of the data matrix given in Table 5. As shown in Fig.1 and Table 5, the RF classification based on the variables selected by wrapper method feature selection method, gave the best ranking values with the smallest SRD (the smaller SRD, the better the model). As bolded in Table 3 and labeled in Fig. 1., the developed Wrapper-RF model gave an accuracy more than 95 and 71% in the training and test sets respectively, and reasonable results on cross-validation techniques.

The LDA classifier that can only learn linear boundaries, showed the worst performance in developed classification models, while the QDA, RBF and SVM that also can learn nonlinear boundaries and are therefore more flexible, showed better performance and significant results.

A deeper inspection in the Tables 2-5, reveals that the wrapper method yielded better performance in classification models with significant statistical parameters. In this study, a wrapper FS technique adopted with genetic algorithm as a random search method was used for all classifier procedures. As mentioned above, Wrapper method is based on the performance of a particular classifier to measure the importance of features set; hence they generally result in better performance than filter methods in which the feature subset selected based on a specific criteria.

3.2 Interpretation of the contributing descriptors

Analysis and interpreting the descriptors entered the classification models can provide detailed information on the molecular structure of organic solvents and gain some insight into the factors affecting the dispersibility of SWNTs in organic solvents. Dispersion or solubilization processes of carbon nanotubes strongly depend on the magnitude and sign of the enthalpy of mixing (ΔH_{mix}) and as discussed in details elsewhere[16, 17, 46] The magnitude and quantity of solvent-nanotube interactions have an effective impact to successfully disperse a nanotube in a solvent. A stable dispersion is accomplished when the SWCNT-dispersing media interaction energy is more favorable compared to the media-media or nanotube-nanotube interactions.[47]

As indicated in Table 1, different kind of descriptors involved in the classification models generated to discriminate between solvents and nonsolvents for SWNTs dispersion. Most of selected descriptors are partial charge descriptors, that depends on the partial charge of each atom of a chemical structure, such as fractional polar van der Waals surface area (PEOE_VSA_FPOL), fractional positive van der Waals surface area (Q_VSA_FPOS), total polar van der Waals surface area (PEOE_VSA_POL) and positive charge weighted surface area ($CASA^+$). The selected features of vsurf_CW_1, vsurf_CW_6 and vsurf_$EWmin_1$ are volsurf descriptors. The volsurf descriptors depend on the structural connectivity and the conformation of the molecules.[48, 49] The vsurf_CW descriptor describes the capacity factor of a molecule at different energy levels and reveals the hydrophilicity of the molecules on unit surface area. The vsurf_$EWmin_1$ represents the lowest hydrophilic energy of a molecule. The SMR_VSA_1, as a subdivided surface area descriptor, defined as Van der Waals surface area (VSA) descriptor[49] that characterized as the amount of surface area with molar refractivity and describes the polarizability of a molecule. The BCUT_SLOGP_2 is a BCUT descriptor using atomic contribution to logP (octanol/water). The BCUT descriptor encodes atomic properties relevant to intermolecular interactions and calculated from the distance and adjacency matrices.[50] The std_dim_2 is the second largest standardized dimension and depends on the structure connectivity and conformation of molecule. As a count atom descriptor, a_nO represent the number of oxygen atoms.

The importance of partial charges descriptors, which are dominated by electrostatic interactions, in classification models imply the role of electronic properties of organic solvents in the dispersibility of SWNTs. The impact of volsurf, subdivided surface area and shape descriptors in developed classification models highlight the effect of structural features of solvent molecules on their ability to disperse SWNTs.

4. Conclusion

Different chemometrics approaches were used in order to build the predictive models with the aim of classifying organic solvents as solvents and nonsolvent for SWNTs. These in-silico studies will be useful to reduce the computational cost and time consumption. The partial charges, volsurf, surface area and shape descriptors were chosen based on feature selection methods and used to discriminate between solvents and nonsolvents. The SRD values showed that the developed wrapper-RF model was the superior model in discrimination of solvents and nonsolvents for SWNTs dispersibility with ACC more than 95% and 71% for training and test sets respectively. The structural features along with electrostatic interactions of solvent molecules play an important role in discriminating good solvents from nonsolvents in SWNTs dispersion.

As a first report on the classification of organic solvents based on their SWNTs dispersibility, simple molecular descriptors and freely available classification packages were used to develop classification models. One important challenge in constructing in-silico modeling would be possibility to develop a reliable and predictive model based on a limited number of experimental data on nanomaterials.[51] The influence of training sample size on the classification performance and the hypothesis that variance in classification learning can be expected to decrease as training set size increases were examined and confirmed by many studies.[52-54] We examined different feature selection techniques combined by various classifier algorithm to overcome the sample size effect on classification difficulty. However, the results obtained from this study are significant and can be improved with larger sample size and with sophisticated classifier methods.

Supporting Information: A complete list of the simplified molecular input line entry specification (SMILES) and molecular structures of the organic solvents used for classification model with their dispersion state of SWNTs.

ACKNOWLEDGMENT

The authors would like to *thank the Iranian National Science Foundation* (INSF) for their financial support.

References:

1. M. S. Dresselhaus, G. Dresselhaus, J. C. Charlier and E. Hernandez, *Phil. Trans. R. Soc. Lond. A*, 2004, **362**, 2065–2098.
2. N. Sinha and J. T. Yeow, *IEEE T NanoBioSci.*, 2005, **4**, 180-196.
3. D. J. Lim, M. Sim, L. Oh, K. Lim and H. Park, *Arch. Pharm. Res.*, 2014, **37**, 43-52.
4. P. Simon and Y. Gogotsi, *Nat. Mater.*, 2008, **7**, 845-854.
5. Z. Spitalsky, D. Tasis, K. Papagelis and C. Galiotis, *Prog. Polym. Sci.*, 2010, **35**, 357-401.
6. M. Feng, H. Han, J. Zhang and H. Tachikawa, in *Electrochemical Sensors, Biosensors and their Biomedical Applications*, Academic Press, San Diego, 2008, pp. 459-501.
7. C. Jianrong, M. Yuqing, H. Nongyue, W. Xiaohua and L. Sijiao, *Biotechnol. Adv.*, 2004, **22**, 505-518.
8. R. P. Raffaelle, B. J. Landi, J. D. Harris, S. G. Bailey and A. F. Hepp, *Mater. Sci. Eng. B*, 2005, **116**, 233-243.
9. D. A. Britz and A. N. Khlobystov, *Chem. Soc. Rev.*, 2006, **35**, 637-659.
10. A. Jeffery, C. Harsh, F. Michael, H. Natalie and P. Jordan, in *Handbook of Nanophysics*, CRC Press, 2010, pp. 1-24.
11. S. D. Bergin, V. Nicolosi, S. Giordani, A. d. Gromard, L. Carpenter, W. J. Blau and J. N. Coleman, *Nanotechnol.*, 2007, **18**, 455705.
12. M. Islam, E. Rojas, D. Bergey, A. Johnson and A. Yodh, *Nano Lett.*, 2003, **3**, 269-273.

13. Y. Liu, C. Chipot, X. Shao and W. Cai, *J. Phys. Chem. B*, 2010, **114**, 5783-5789.

14. L. Vaisman, H. D. Wagner and G. Marom, *Adv. Colloid Interface Sci.*, 2006, **128-130**, 37-46.

15. J. W. Park, J. Kim, J. O. Lee, K. C. Kang, J. J. Kim and K. H. Yoo, *Appl. Phys. Lett.*, 2002, **80**, 133-135.

16. B. J. Landi, H. J. Ruf, J. J. Worman and R. P. Raffaelle, *J. Phys. Chem. B*, 2004, **108**, 17089-17095.

17. S. D. Bergin, Z. Sun, D. Rickard, P. V. Streich, J. P. Hamilton and J. N. Coleman, *ACS Nano*, 2009, **3**, 2340-2350.

18. Q. Cheng, S. Debnath, E. Gregan and H. J. Byrne, *J. Phys. Chem. C*, 2008, **112**, 20154-20158.

19. M. L. Usrey, A. Chaffee, E. S. Jeng and M. S. Strano, *J. Phys. Chem. C*, 2009, **113**, 9532-9540.

20. S. Detriche, G. Zorzini, J. F. Colomer, A. Fonseca and J. B. Nagy, *J. Nanosci. Nanotechnol.*, 2008, **8**, 6082-6092.

21. H. T. Ham, Y. S. Choi and I. J. Chung, *J. Colloid Interface Sci.*, 2005, **286**, 216-223.

22. S. D. Bergin, V. Nicolosi, P. V. Streich, S. Giordani, Z. Sun, A. H. Windle, P. Ryan, N. P. P. Niraj, Z.-T. T. Wang, L. Carpenter, W. J. Blau, J. J. Boland, J. P. Hamilton and J. N. Coleman, *Adv. Mater.*, 2008, **20**, 1876-1881.

23. S. D. Bergin, Z. Sun, D. Rickard, P. V. Streich, J. P. Hamilton and J. N. Coleman, *Acs Nano*, 2009, **3**, 2340-2350.

24. M. K. Rofouei, M. Salahinejad and J. B. Ghasemi, *Fuller. Nanotub. Car. N.*, 2013, null-null.

25. M. Salahinejad and E. Zolfonoun, *J. Nanopart. Res.*, 2013, **15**, 1-9.

26. S. Giordani, S. D. Bergin, V. Nicolosi, S. Lebedkin, M. M. Kappes, W. J. Blau and J. N. Coleman, *J. Phys. Chem. B*, 2006, **110**, 15708-15718.

27. S. N. Barman, D. Pan, M. Vosgueritchian, A. P. Zoombelt, G. Galli and Z. Bao, *Nanotechnol.*, 2012, **23**, 344011.

28. T. Puzyn, A. Mostrag-Szlichtyng, A. Gajewicz, M. Skrzyński and A. P. Worth, *Structural Chemistry*, 2011, **22**, 795-804.

29. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *SIGKDD Explor. Newsl.*, 2009, **11**, 10-18.

30. J. Novakovic, 17th Telecommunications forum TELFOR, 2009.

31. A.G Karegowda, M.A.Jayaram and A S Manjunath, *Int. J. Comput. Appl.*, 2010, **1**, 12-17.

32. G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*, Wiley, 1992.

33. D. Ballabio and R. Todeschini, *Infrared spectroscopy for food quality analysis and control. Ed. D. Sun (Academic Press: Burlington, MA) pp*, 2009, 83-104.

34. Y. Zhao and Y. Zhang, *Adv. Space Res.*, 2008, **41**, 1955-1959.

35. D. L. Gupta, A. K. Malviya and S. Singh, *Int. J. Comput. Appl.*, 2012, **55**, 39-44.

36. G. P. Zhang, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2000, **30**, 451-462.

37. C. C. Chang and C.-J. Lin, *ACM Trans. Intell. Syst. Technol.*, 2011, **2**, 27.

38. C. W. Hsu, C. C. Chang and C.-J. Lin, 2003.

39. M. Banerjee, M. Capozzoli, L. McSweeney and D. Sinha, *Can. J. Stat.*, 1999, **27**, 3-23.

40. S. M. Vieira, U. Kaymak and J. M. Sousa, Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, 2010.

41. S. Spycher, M. Nendza and J. Gasteiger, *QSAR Comb. Sci.*, 2004, **23**, 779-791.

42. D. Ballabio and V. Consonni, *Anal. Method.*, 2013, **5**, 3790-3798.

43. K. Héberger and K. Kollár-Hunek, *J. Chemometr.*, 2011, **25**, 151-158.

44. K. Kollár-Hunek and K. Héberger, *Chemom. Intell. Lab. Syst.*, 2013, **127**, 139-146.

45. I. Guyon and A. Elisseeff, *J. Machin. Learn. Res.*, 2003, **3**, 1157-1182.

46. S. Giordani, S. D. Bergin, V. Nicolosi, S. Lebedkin, M. M. Kappes, W. J. Blau and J. N. Coleman, *J. Phys. Chem. B*, 2006, **110**, 15708-15718.

47. S. B. Fagan, A. G. Souza Filho, J. O. G. Lima, J. M. Filho, O. P. Ferreira, I. O. Mazali, O. L. Alves and M. S. Dresselhaus, *Nano Lett.*, 2004, **4**, 1285-1288.

48. G. Cruciani, P. Crivori, P. A. Carrupt and B. Testa, *J. Mol. Struct.*, 2000, **503**, 17-30.

49. P. Labute, *J. Mol. Graph. Model.*, 2000, **18**, 464-477.

50. D. T. Stanton, *J. Chem. Info. Comput. Sci.*, 1999, **39**, 11-20.

51. B. Rasulev, A. Gajewicz, T. Puzyn, D. Leszczynska and J. Leszczynski, in *Towards Efficient Designing of Safe Nanomaterials: Innovative Merge of Computational Approaches and Experimental Techniques*, 2012, p. 220.

52. S. J. Raudys and A. K. Jain, *IEEE Trans. Pattern. Anal. Mach. Intell.*, 1991, **13**, 252-264.

53. D. Brain, G. Webb, D. Richards, G. Beydoun, A. Hoffmann and P. Compton, Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales, 1999.

54. V. Popovici, W. Chen, B. G. Gallas, C. Hatzis, W. Shi, F. W. Samuelson, Y. Nikolsky, M. Tsyganova, A. Ishkin and T. Nikolskaya, *Breast Cancer Res.*, 2010, **12**, R5.

Table 1. Molecular descriptors selected by different feature selection techniques, their abbreviation and description.

Selected descriptors	**Description**	**Subset evaluator**
a_nO	Number of oxygen atoms	ReliefF
BCUT_SLOGP_2	BCUT descriptors using atomic contribution to logP	Wrapper
CASA$^+$	Positive charge weighted surface area	InfoGain, Cfs
GCUT_PEOE_0	GCUT descriptors	Wrapper
PEOE_VSA_FNEG	Fractional negative van der Waals surface area	Wrapper
PEOE_VSA_FPOL	Fractional polar van der Waals surface area	Cfs
PEOE_VSA_FPPOS	Fractional positive polar van der Waals surface area	Wrapper
PEOE_VSA_POL	Total polar van der Waals surface area	ReliefF
PEOE_VSA_PPOS	Total positive polar van der Waals surface area	ReliefF
PEOE_VSA^{+4}	Sum of v_i where q_i is in the range (0.20,0.25)	ReliefF
Q_VSA_FPOS	Fractional positive van der Waals surface area	InfoGain, Cfs
Q_VSA_NEG	Total negative van der Waals surface area	Wrapper
SMR_VSA$_1$	Sum of atomic molar refractivity with polarities in the range 0.11 to 0.26	Cfs
std_dim$_2$	Standard dimension 2	Wrapper
vsurf_CW$_1$	Capacity factor of order 1	Wrapper
vsurf_CW$_6$	Capacity factor of order 6	InfoGain, Cfs
vsurf_EWmin1	Lowest hydrophilic energy	InfoGain

Table 2. Performance of DA classifiers with various feature selection techniques.

Subset evaluator	Classifier	Training set					Test set					LOO-CV		10-fold-CV	
		Sp	Sn	ACC	MCC	k	Sp	Sn	ACC	MCC	k	ACC	MCC	ACC	MCC
Cfs	LDA	70.83	60.71	62.50	0.25	0.25	50.00	50.00	57.14	0.13	0.13	56.25	0.13	56.25	0.13
InfoGain		70.83	65.38	66.67	0.33	0.33	50.00	50.00	57.14	0.13	0.13	56.25	0.13	60.42	0.21
ReliefF		75.00	78.26	77.08	0.54	0.54	83.33	71.43	78.57	0.58	0.57	77.08	0.54	79.17	0.58
Wrapper		75.00	69.23	70.83	0.42	0.42	83.33	50.00	57.14	0.23	0.19	60.42	0.21	66.67	0.33
Cfs	QDA	75.00	78.26	77.08	0.54	0.54	75.00	60.00	62.50	0.26	0.42	52.08	0.04	58.33	0.12
InfoGain		83.33	60.61	64.58	0.31	0.29	70.83	58.62	58.33	0.19	0.57	52.08	0.04	60.42	0.21
ReliefF		79.17	79.17	79.17	0.58	0.58	75.00	75.00	75.00	0.50	0.71	75.00	0.50	77.08	0.54
Wrapper		83.33	90.91	87.50	0.75	0.75	62.50	60.00	60.42	0.21	0.16	60.42	0.21	58.33	0.17

Sp: Specificity (%), Sn: Sensitivity (%), ACC: Accuracy (%), MCC: Matthews Correlation Coefficient, k: Kappa coefficient, LOO-CV: leave-one-out cross-validation

Table 3. Performance of DT classifiers with various feature selection techniques.

Subset evaluator	Classifier	Training set					Test set					LOO-CV		10-fold-CV	
		Sp	Sn	ACC	MCC	k	Sp	Sn	ACC	MCC	k	ACC	MCC	ACC	MCC
Cfs	RF	95.83	79.31	85.42	0.72	0.71	100.00	75.00	85.71	0.75	0.72	79.17	0.58	81.25	0.63
InfoGain		95.83	76.67	83.33	0.69	0.67	100.00	66.67	78.57	0.65	0.59	68.75	0.38	70.83	0.43
ReliefF		70.83	100.00	85.42	0.74	0.71	83.33	71.43	78.57	0.58	0.57	66.67	0.33	72.92	0.46
Wrapper		**100.00**	**92.31**	**95.83**	**0.92**	**0.92**	**100.00**	**60.00**	**71.43**	**0.55**	**0.47**	**68.75**	**0.38**	**70.83**	**0.42**
Cfs	J48	83.33	86.96	85.42	0.71	0.71	83.33	62.50	71.43	0.46	0.44	77.08	0.54	68.75	0.38
InfoGain		100.00	72.73	81.25	0.67	0.63	83.33	50.00	57.14	0.23	0.19	75.00	0.53	72.92	0.48
ReliefF		58.33	100.00	79.17	0.64	0.58	66.67	80.00	78.57	0.56	0.55	70.83	0.44	66.67	0.34
Wrapper		100.00	62.50	81.25	0.67	0.63	66.67	57.14	64.29	0.29	0.29	64.58	0.29	62.50	0.25

Sp: Specificity (%), Sn: Sensitivity (%), ACC: Accuracy (%), MCC: Matthews Correlation Coefficient, k: Kappa coefficient, LOO-CV: leave-one-out cross-validation

Table 4. Performance of SVM, RBF and MLP classifiers with various feature selection techniques.

Subset evaluator	Classifier	Training set					Test set					LOO-CV		10-fold-CV	
		Sp	Sn	ACC	MCC	k	Sp	Sn	ACC	MCC	k	ACC	MCC	ACC	MCC
Cfs	LiBSvm	83.33	90.91	87.50	0.75	0.75	83.33	83.33	85.71	0.71	0.71	81.25	0.63	81.25	0.63
InfoGain		66.67	72.73	70.83	0.42	0.42	83.33	62.50	71.43	0.46	0.44	62.50	0.25	60.42	0.21
ReliefF		70.83	100.00	85.42	0.74	0.71	83.33	62.50	71.43	0.46	0.44	54.17	0.16	75.00	0.50
Wrapper		95.83	92.00	93.75	0.88	0.88	83.33	55.56	64.29	0.34	0.31	58.33	0.17	64.58	0.29
Cfs	RBFN	79.17	100.00	89.58	0.81	0.79	83.33	71.43	78.57	0.58	0.57	77.08	0.54	75.00	0.51
InfoGain		75.00	87.50	79.17	0.59	0.58	100.00	66.67	78.57	0.65	0.59	56.25	0.13	62.50	0.25
ReliefF		70.83	94.44	83.33	0.69	0.67	83.33	83.33	85.71	0.71	0.71	68.75	0.38	66.67	0.33
Wrapper		95.83	100.00	97.92	0.96	0.96	50.00	50.00	57.14	0.13	0.13	58.33	0.17	60.42	0.21
Cfs	MLP	79.17	82.61	81.25	0.63	0.63	83.33	71.43	78.57	0.58	0.57	77.08	0.55	77.08	0.54
InfoGain		65.22	62.50	64.58	0.29	0.29	50.00	60.00	64.29	0.26	0.26	60.42	0.21	62.50	0.25
ReliefF		75.00	85.71	81.25	0.63	0.63	83.33	71.43	78.57	0.58	0.57	56.25	0.46	75.00	0.50
Wrapper		70.83	70.83	70.83	0.42	0.42	66.67	50.00	57.14	0.17	0.16	56.25	0.56	60.42	0.21

Sp: Specificity (%), Sn: Sensitivity (%), ACC: Accuracy (%), MCC: Matthews Correlation Coefficient, k: Kappa coefficient, LOO-CV: leave-one-out cross-validation

Table 5. Sum of ranking difference (SRD) and p% interval of the variables of the classification analyses

	Ranking results		p%	
Model	Model code	SRD	x < SRD > =x	
Wrapper-RF	12	10	7.45E-05	1.73E-04
Cfs-MLP	21	14	3.88E-04	8.44E-04
Wrapper-QDA	8	18	1.78E-03	3.67E-03
Cfs-RBF	25	18	1.78E-03	3.67E-03
Wrapper-SVM	20	20	3.67E-03	7.35E-03
Cfs-QDA	5	22	7.35E-03	1.43E-02
InfoGain-RF	10	22	7.35E-03	1.43E-02
ReliefF-RF	11	22	7.35E-03	1.43E-02
Cfs-J48	13	22	7.35E-03	1.43E-02
ReliefF-SVM	19	24	1.43E-02	2.71E-02
ReliefF-RBF	27	24	1.43E-02	2.71E-02
Wrapper-LDA	4	26	2.71E-02	5.00E-02
Cfs-SVM	17	26	2.71E-02	5.00E-02
Wrapper-J48	16	28	5.00E-02	8.98E-02
InfoGain-SVM	18	28	5.00E-02	8.98E-02
InfoGain-MLP	22	28	5.00E-02	8.98E-02
InfoGain-QDA	6	32	0.16	0.27
Cfs-RF	9	32	0.16	0.27
Wrapper-MLP	24	32	0.16	0.27
ReliefF-QDA	7	34	0.27	0.44
ReliefF-MLP	23	34	0.27	0.44
InfoGain-J48	14	36	0.44	0.72
ReliefF-J48	15	36	0.44	0.72
InfoGain-RBF	26	36	0.44	0.72
Wrapper-RBF	28	36	0.44	0.72
Cfs-LD	1	38	0.72	1.13
InfoGain-LD	2	38	0.72	1.13
ReliefF-LD	3	44	2.60	3.80
	XX1	46	4.61	5.47
	Q1	58	24.45	27.12
	Med	66	48.78	52.08
	Q3	74	73.59	76.22
	XX19	84	94.77	95.59

XX1-frst icosaile (5%), Q1-frst quartile, Med-median, Q3-last quartile, XX19-last icosaile (95%).

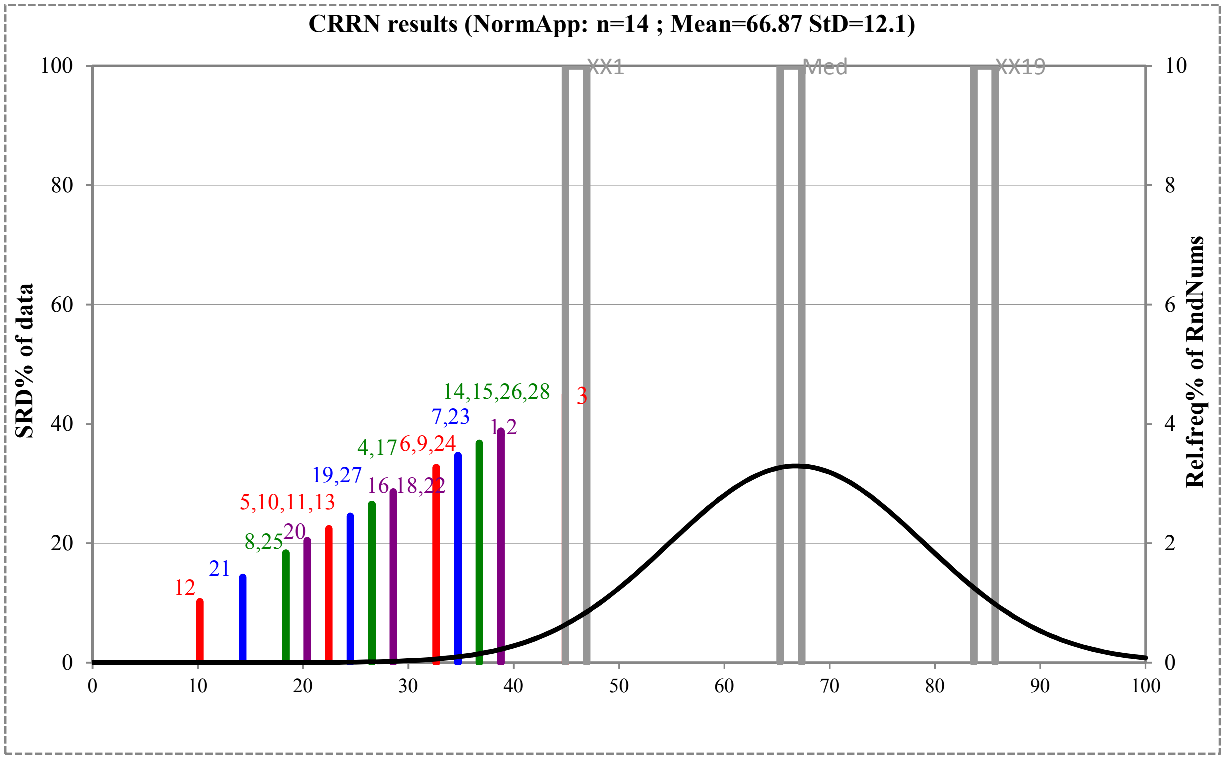


Figure. 1. SRD-CRRN results of the classification models. XX1: first icosaile (5%), Q1:first quartile, Med: median, Q3:last quartile, XX19:last icosaile (95%). The label numbers indicated the model code given in the Table 5.