# Prediction of Biological Activities of Volatile Metabolites Using Molecular Fingerprints and Machine Learning Methods

Azian Azamimi Abdullah[1] and Shigehiko Kanaya[2]
[1]*Biomedical Electronic Engineering Programme, School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia.*
[2]*Computational Systems Biology Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, 630-0192, Nara, Japan*
*azamimi@unimap.edu.my*

***Abstract*— Volatile metabolites are small molecules, comprise a diverse chemical group with various biological activities and have high vapor pressures under ambient conditions. It is crucial to determine the biological activities of volatile metabolites as they play important roles in chemical ecology and human healthcare. In this study, we have accumulated 341 volatiles emitted by biological species associated with 11 types of biological activities and deposited the data into our database, which is called KNApSAcK Metabolite Ecology Database. Using this dataset, we have developed 72 classification models to predict biological activities of volatile metabolites by using various machine learning methods. Eight types of molecular fingerprints were used to represent the molecules, which are PubChem (881 bits), CDK (1024 bits), Extended CDK (1024bits), MACCS (166 bits), Klekota-Roth (4860 bits), Substructure (307 bits), Estate (79 bits), and atom pairs (780 bits). A new type of fingerprint was also proposed by combining all features of these eight fingerprints (Combine, 9121 bits). The best classification model was developed by our proposed fingerprint (Combine, 9121 bits) trained with gradient boosting method algorithm (GBM) with predictive accuracy at 94.43%. The results indicated that molecular fingerprints and machine learning methods could be useful for predicting biological activities of volatile metabolites.**

***Index Terms*—Biological Activities; Fingerprints; Machine Learning; Volatile Metabolites.**

## I. Introduction

Metabolomics is the scientific study of quantification of low mass compounds profiles and analysis of chemical processes involving metabolites in a comprehensive fashion. In general, metabolites can be divided into two groups: primary and secondary metabolites. Primary metabolites are directly involved in the normal growth, development and reproduction. On the other hand, secondary metabolites are not directly involved in these processes, but usually have important ecological functions, such as inter- or intra-species communication, antifungal, antimicrobial activities and also as a defense against pests and pathogens. Small proportions produced by these secondary metabolites are volatile metabolites or also known as volatile organic compounds (VOCs) that play important roles in chemical ecology and human healthcare.

VOCs can be defined as small compounds ranging in between C5 to C20 carbon count with a molecular weight in the range of 50 to 200 Daltons [1]. They comprise a diverse chemical group of organic compounds with various biological activities and have high vapor pressures under ambient conditions. All living organisms including human, animals, plants and microorganisms produce VOCs naturally. The naturally produced VOCs play important roles in communication between plants and they also serve as signaling molecules by passing information between organisms [2]. For human and other animals, VOCs are important as scents and flavor of food [3]. Recently, an increased number of researchers are utilizing VOCs as a biomarker to identify various kinds of diseases [4]-[12]. Hence, the importance of VOCs for living organisms specifically in chemical ecology, agriculture and human healthcare need to be further explored.

Here, we investigate the relationships between chemical structures of VOCs and biological activities by applying four types of machine learning methods, which are deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM) as classification models for predicting the biological activities of VOCs based on their chemical structures.

## II. Materials And Methods

This section discusses the datasets used for this study, molecular fingerprints, machine learning methods and evaluation of model performance.

### *A. Datasets*

In this study, we have accumulated 341 volatiles emitted by various biological species associated with 11 types of biological activities and deposited the data into our database, which is called KNApSAcK Metabolite Ecology Database [13].This database is available and can be accessed freely at http://kanaya.naist.jp/MetaboliteEcology/top.jsp. From our accumulated data, 57.3% of the activities belong to chemical ecology such as antifungal, antimicrobial, attractant, defense, enhance plant growth, inhibit root growth and repellent activities. On the other hand, 42.7% of the activities belong to human health-related activities such as disease biomarker, odor, anticholinesterase and antioxidant as shown in Figure 1. There are many VOCs, which have several biological activities. Figure 2 shows the relative frequencies of VOCs, which have several biological activities. There are 239 VOCs (about 70%), which have only one specific biological activity. 28 VOCs have 2 biological activities, 52 VOCs have

3 biological activities, 17 VOCs have 4 biological activities, 3 VOCs have 5 biological activities and only 2 VOCs have 6 biological activities. For simplicity, we empirically select the most relevant biological activity to each particular compound.

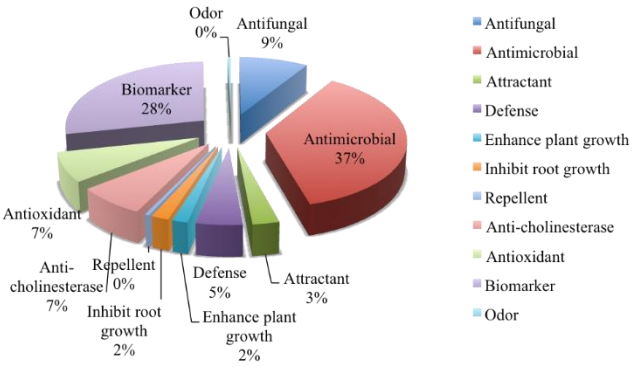


Figure 1: Pie chart showing the relative frequencies belonging to 11 biological activities.

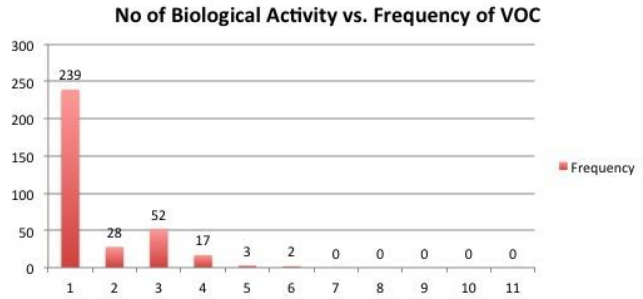


Figure 2: The relative frequencies of VOCs, which have several biological activities.

### B. *Molecular Fingerprints*

The fingerprint of a chemical compound is a binary vector indicating the substructures it contains. In this study, eight types of molecular fingerprints are used to represent the molecules, which are PubChem (PubChem, 881 bits), CDK (CDK, 1024 bits), Extended CDK (Extended, 1024bits), MACCS (MACCS, 166 bits), Klekota-Roth (KR, 4860 bits), Substructure (Sub, 307 bits), Estate (Estate, 79 bits), and atom pairs (AP, 780 bits). We also proposed a new type of fingerprint, by combining all features and substructures obtained by these fingerprints (Combine, 9121 bits). The reason why we use many types of fingerprints, is that we want to investigate which fingerprint method can generate the best prediction model. We converted the SDF files of all 341 VOCs into binary fingerprints using ChemDes software [14]. After we obtained the binary matrix of fingerprints, we performed the data-processing method by removing all columns that contain "0". This is because it might be not relevant for the classification of VOCs based on substructures. The features or substructures displayed in a binary matrix, was used as input to the classification models. There are 11 classes of biological activities, which have been used as outputs for the classification model.

The VOC-Substructure-Biological activities relations can be represented as a matrix, shown in Table 1 where rows represent VOCs and columns represent substructures of molecular fingerprints. We added one additional column to represent biological activities for each of VOCs.

Table 1
Representation of VOCs, Substructures and Biological Activities as a Two-Dimensional Matrix.

| VOCs | Substructures | | | | | | Biological Activities |
|---|---|---|---|---|---|---|---|
| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | ... | $S_M$ | |
| $VOC_1$ | 1 | 0 | 1 | 1 | ... | 0 | Antimicrobial |
| $VOC_2$ | 1 | 1 | 0 | 0 | ... | 0 | Biomarker |
| $VOC_3$ | 0 | 1 | 0 | 1 | ... | 0 | Defense |
| ... | ... | ... | ... | ... | ... | ... | |
| $VOC_N$ | 1 | 0 | 0 | 0 | ... | 1 | Odor |

### C. *Machine Learning Methods*

Machine learning algorithms are generally developed in computer science or adjacent disciplines and find their way into chemical modeling by process of diffusion. Recently, machine learning methods are popular in chemoinformatics and quantitative structure–activity relationships (QSAR), which usually predicting the unknown property values of a test set of molecules based on the known values for a training set [15]-[17]. We implemented four types of supervised machine learning methods for predicting biological activities of VOCs, which are a deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM) using H2O package in R program [18]. DNN was one of the increasingly popular methods in the machine learning community in the past years and produce a good performance in many applications such as machine vision, speech processing, drug discovery and other artificial intelligence fields [19]-[23]. One of the main differences between DNN and the conventional artificial neural networks is that DNN has more than one hidden layer and more neurons in each layer, thus making the learning process become more "deeper" and "wider" [24]. It is difficult and time-consuming to find the best parameters for DNN due to a large number of adjustable parameters. Hence we took the approach by choosing the best parameter by using the multi-dimensional hyper-parameter optimization method. We selected the best parameter and then, compared with the default parameter. Table 2 shows the DNN parameter used in this study. We used the default setting for DNN1; Rectifier activation function, 200 neurons in both hidden layer 1 and hidden layer 2 and epochs were set to 10. We varied the parameter for DNN2 and DNN3 by using the Tanh and Maxout activation function. For DNN4, we selected the best parameter based on multi-dimensional hyper-parameter optimization method; Rectifier activation function with dropout, 5 hidden layers, 200 neurons in every hidden layer, 20% dropout rate in the input layer and each of hidden layer and the epoch was set to 10000. For DNN5, we used the Maxout activation function, 5 hidden layers, and 200 neurons in every hidden layer and the epoch were set to 10000.

Other than DNN, we also compared the classification performance of GBM, RF and GLM methods. GBM is a family of powerful machine-learning techniques for regression and classification problems, which produce a prediction model in the form of an ensemble of weak prediction models, typically decision trees [25]. This algorithm also produces good performance in many applications including cheminformatics [26]-[28]. RF is an ensemble method that consists of many decision trees for classification and regression tasks. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual

trees [29]. In statistics, the GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value [30].

Table 2
List of DNN Parameters Used in this Study.

| Parameter list | DNN1 (default) | DNN2 | DNN3 | DNN4 | DNN5 |
|---|---|---|---|---|---|
| Activation function | Rectifier | Tanh | Maxout | Rectifier with Dropout | Maxout |
| Input dropout ratio | | | | 20% | |
| Hidden dropout ratio | | | | 20%, 20%, 20%, 20%, 20% | |
| Hidden layer 1 | 200 | 200 | 200 | 200 | 200 |
| Hidden layer 2 | 200 | 200 | 200 | 200 | 200 |
| Hidden layer 3 | | | | 200 | 200 |
| Hidden layer 4 | | | | 200 | 200 |
| Hidden layer 5 | | | | 200 | 200 |
| Epoch | 10 | 10 | 10 | 10000 | 10000 |

### *D. Evaluation of Model Performance*

The performance of multi-classification models was measured by mean squared error (MSE) value and accuracy (%). We conducted two sets of experiments: (1) Using all datasets as training, and (2) Using 10-fold cross-validation technique. In this technique, the compounds were randomly divided into ten parts, where nine parts were used for training and remaining part was used for testing. This process is carried out ten times in such a way that each part was used once for testing.

## III. Results And Discussion

In this study, we have developed 72 classification models to predict biological activities of VOCs by nine types of molecular fingerprints trained with four types of supervised machine-learning methods, which are DNN, GBM, RF and GLM. We conducted two types of experiments; (1) Using all datasets as training, and (2) Using 10-fold cross-validation technique. Table 3 shows the list of 72 models and their performances for both experiments. Mean squared error (MSE) and accuracy (%) was used as the performance indicator.

Table 3
Performance of 72 Classification Models using Different Fingerprints (FP) and Machine Learning (ML) Methods.

| Model No. | FP+ML | (1) 100% training | | (2) 10-fold CV | |
|---|---|---|---|---|---|
| | | MSE | Accuracy (%) | MSE | Accuracy (%) |
| 1 | Combine+DNN1 | 0.1053 | 87.39 | 0.4840 | 48.92 |
| 2 | Combine+DNN2 | 0.1072 | 87.68 | 0.4943 | 46.91 |
| 3 | Combine+DNN3 | 0.2447 | 74.78 | 0.5023 | 47.20 |
| 4 | Combine+DNN4 | 0.5051 | 91.49 | 0.5064 | 44.44 |
| 5 | Combine+DNN5 | 0.1619 | 83.28 | 0.4514 | 53.69 |
| 6 | Combine+RF | 0.4213 | 57.77 | 0.4232 | 57.95 |
| 7 | Combine+GBM | 0.3953 | 94.43 | 0.3983 | 57.67 |
| 8 | Combine+GLM | 0.4319 | 76.83 | 0.4323 | 58.66 |
| 9 | KR+DNN1 | 0.1582 | 80.65 | 0.4840 | 52.46 |
| 10 | KR+DNN2 | 0.1411 | 81.82 | 0.4862 | 47.31 |
| 11 | KR+DNN3 | 0.1656 | 81.82 | 0.4679 | 50.68 |
| 12 | KR+DNN4 | 0.0542 | 92.08 | 0.5382 | 40.42 |
| 13 | KR+DNN5 | 0.0805 | 91.20 | 0.5000 | 48.40 |
| 14 | KR+RF | 0.4104 | 54.25 | 0.4173 | 56.73 |
| 15 | KR+GBM | 0.1267 | 88.56 | 0.4144 | 53.76 |
| 16 | KR+GLM | 0.3484 | 70.09 | 0.4397 | 58.08 |
| 17 | PubChem+DNN1 | 0.1775 | 80.94 | 0.4472 | 51.24 |
| 18 | PubChem+DNN2 | 0.1265 | 81.82 | 0.5219 | 43.82 |
| 19 | PubChem+DNN3 | 0.1768 | 79.77 | 0.4604 | 52.58 |
| 20 | PubChem+DNN4 | 0.0587 | 91.20 | 0.5767 | 35.31 |
| 21 | PubChem+DNN5 | 0.0816 | 90.33 | 0.4764 | 49.68 |
| 22 | PubChem+RF | 0.4074 | 55.43 | 0.4083 | 57.81 |
| 23 | PubChem+GBM | 0.1214 | 88.86 | 0.3931 | 55.39 |
| 24 | PubChem+GLM | 0.3679 | 65.98 | 0.4595 | 56.47 |
| 25 | CDK+DNN1 | 0.2230 | 74.19 | 0.4918 | 46.06 |
| 26 | CDK+DNN2 | 0.2205 | 71.85 | 0.5494 | 40.97 |
| 27 | CDK+DNN3 | 0.2025 | 76.83 | 0.4981 | 46.99 |
| 28 | CDK+DNN4 | 0.1089 | 85.04 | 0.5754 | 35.53 |
| 29 | CDK+DNN5 | 0.1698 | 90.32 | 0.5091 | 44.37 |
| 30 | CDK+RF | 0.4555 | 57.77 | 0.4635 | 52.28 |
| 31 | CDK+GBM | 0.1498 | 83.87 | 0.4328 | 51.45 |
| 32 | CDK+GLM | 0.3724 | 66.28 | 0.4731 | 51.86 |
| 33 | Extended+DNN1 | 0.2230 | 81.53 | 0.4707 | 46.06 |
| 34 | Extended+DNN2 | 0.2205 | 74.49 | 0.5169 | 40.97 |
| 35 | Extended+DNN3 | 0.2025 | 70.97 | 0.5196 | 46.99 |
| 36 | Extended+DNN4 | 0.1089 | 86.51 | 0.6431 | 35.53 |
| 37 | Extended+DNN5 | 0.5051 | 83.28 | 0.4914 | 44.37 |
| 38 | Extended+RF | 0.4555 | 52.79 | 0.4361 | 52.28 |
| 39 | Extended+GBM | 0.1498 | 86.22 | 0.4171 | 51.45 |
| 40 | Extended+GLM | 0.3724 | 68.62 | 0.4461 | 51.86 |
| 41 | AP+DNN1 | 0.4246 | 52.19 | 0.5482 | 39.69 |
| 42 | AP+DNN2 | 0.4150 | 53.08 | 0.5460 | 39.74 |
| 43 | AP+DNN3 | 0.4606 | 50.15 | 0.5729 | 40.61 |
| 44 | AP+DNN4 | 0.3413 | 59.53 | 0.5734 | 41.49 |
| 45 | AP+DNN5 | 0.3742 | 56.89 | 0.5278 | 42.51 |
| 46 | AP+RF | 0.4948 | 49.56 | 0.4963 | 50.13 |
| 47 | AP+GBM | 0.3831 | 59.53 | 0.4964 | 49.25 |
| 48 | AP+GLM | 0.4787 | 52.79 | 0.5104 | 51.68 |
| 49 | Sub+DNN1 | 0.3793 | 60.12 | 0.5411 | 42.45 |
| 50 | Sub+DNN2 | 0.3491 | 61.58 | 0.4998 | 44.33 |
| 51 | Sub+DNN3 | 0.3132 | 65.39 | 0.5515 | 40.49 |
| 52 | Sub+DNN4 | 0.2179 | 73.90 | 0.5396 | 39.71 |
| 53 | Sub+DNN5 | 0.2655 | 68.33 | 0.5062 | 44.51 |
| 54 | Sub+RF | 0.4480 | 51.61 | 0.4541 | 51.38 |
| 55 | Sub+GBM | 0.4497 | 68.33 | 0.4492 | 50.69 |
| 56 | Sub+GLM | 0.4502 | 58.94 | 0.4845 | 55.38 |
| 57 | Estate+DNN1 | 0.3621 | 58.65 | 0.5249 | 43.22 |
| 58 | Estate+DNN2 | 0.3775 | 58.94 | 0.4919 | 47.89 |
| 59 | Estate+DNN3 | 0.4585 | 46.33 | 0.5830 | 37.53 |
| 60 | Estate+DNN4 | 0.2531 | 68.33 | 0.5230 | 42.79 |
| 61 | Estate+DNN5 | 0.3149 | 65.10 | 0.4762 | 48.04 |
| 62 | Estate+RF | 0.4561 | 53.08 | 0.4558 | 52.78 |
| 63 | Estate+GBM | 0.2974 | 66.86 | 0.4527 | 51.56 |
| 64 | Estate+GLM | 0.4667 | 55.13 | 0.4956 | 51.39 |
| 65 | MACCS+DNN1 | 0.2418 | 71.85 | 0.5004 | 46.25 |
| 66 | MACCS+DNN2 | 0.2353 | 73.31 | 0.4853 | 45.32 |
| 67 | MACCS+DNN3 | 0.1849 | 77.13 | 0.5010 | 45.65 |
| 68 | MACCS+DNN4 | 0.0780 | 88.56 | 0.5599 | 41.63 |
| 69 | MACCS+DNN5 | 0.5128 | 87.39 | 0.5103 | 45.14 |
| 70 | MACCS+RF | 0.4293 | 53.08 | 0.4327 | 52.28 |
| 71 | MACCS+GBM | 0.3997 | 87.09 | 0.3998 | 56.29 |
| 72 | MACCS+GLM | 0.3989 | 60.70 | 0.4732 | 55.35 |

Figure 3 shows the distribution of 72 classification models (MSE value) by using all datasets as training and 10-fold cross-validation technique. For the first experiment, by using all datasets as training, the best classification model was developed by Klekota-Roth fingerprint trained with the DNN4 method, with MSE value 0.05420784. Second best classification model was developed by PubChem fingerprint with MSE value 0.05871162, followed by MACCS fingerprint with MSE value 0.07807859. Both fingerprints

were also trained with DNN4. The best parameter for deep learning was obtained by using rectifier activation function with dropout rate at 20%. A number of the hidden layers was set to 5 and 200 neurons for each of hidden layer. Estate and atom pair fingerprint did not perform well in the classification model. This is because the length of the Estate fingerprint is only 79 bits, which is too short to characterize molecules. Too much information loss led to the bad prediction.

For the second experiment, we adopted the 10-fold cross-validation technique to evaluate the performance of our models. The lowest MSE error was obtained by using PubChem fingerprint trained by GBM method at 0.39318013, followed by Combine fingerprint also trained by GBM method. The obtained MSE error was 0.39837325. MACCS fingerprint trained by GBM method also gave good MSE value at 0.39979038 compared to other models. The worst performance was obtained using Extended fingerprints trained with DNN4 and Estate fingerprint trained with DNN3.

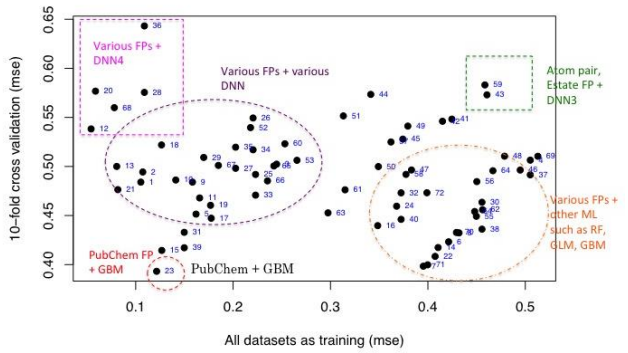


Figure 3: Performance of 72 classification models by using all datasets as training and 10-fold cross-validation technique (MSE value).

Based on Figure 3, it seems that all data are distributed randomly and there is no correlation between the performance obtained by using all datasets as training and 10-fold cross-validation technique. We observed that there are two types of models: 1) the left side is affected by over-fitting problem, and 2) the right side is not changed for both experiments. The left side points, which most of the combination of fingerprint types and DNN methods suffered from over-fitting problems due to the many parameters of DNN. The performance of DNN is good when using all datasets as training, however it becomes worst when we used 10-fold cross-validation technique, such as model No 12 (Klekota-Roth fingerprint trained with DNN4 method) and model No 36 (Extended fingerprint trained with DNN4 method). The small number of our sample data and many parameters of DNN might cause this over-fitting problem. DNN always requires a large amount of data to be trained, usually more than 50,000 samples. In our study, we only have 341 VOC data for the classification task. In theory, over-fitting is a major problem for DNN and we have proved this experimentally. Moreover, the Klekota-Roth and Extended fingerprints have many substructures or features (more than 1000), which need to be trained and as a result, they are suffering from over-fitting problem too. The right side points did not change much for both experiments. For example, the classification model No 43 (atom pair fingerprint trained with DNN3 method) and model No 59 (Estate fingerprint trained with DNN3 method) performed poorly in both experiments. From this result, we can understand two things; 1) Atom pair and Estate fingerprint did not perform well in model building, 2) DNN3 is the worst, compared to other DNN models. Atom pair fingerprint are a structural descriptor type that is defined by the shortest paths among the non-hydrogen atoms in a molecule. Each path is described by the types of atoms in a pair, the length of their shortest bond path, the number of their *pi* electrons and the non-hydrogen atoms bonded to them. The number of atom pairs describing a molecule grows with its number of atoms. The fingerprints provided by PubChem are a binary representation of the presence and absence of a library of 881 substructure features. Compared to atom pairs, the PubChem fingerprints are a knowledge-based system that stores less information than the much more complex and unbiased atom pair concept. PubChem fingerprints are also less sensitive than atom pair descriptors. The length of the Estate fingerprint is only 79 bits, which is too short to characterize molecules and some of the information might be loss, which cause the bad prediction. It is also observed that hyperparameters of DNN can affect the overall performance. The reason why DNN3 performed poorly for both experiments, is because the Maxout activation function and a small number of epochs. Rectifier activation function is a better choice for this classification task.

Also, based on Figure 3, we can observe that the classification model No 23 (PubChem fingerprint trained with GBM method) gives good results in both experiments. This model obtained MSE value = 0.1214795 when using all datasets as training and MSE value = 0.39318013 in case of 10-fold cross-validation technique. The results show that GBM method is good at predicting biological activities of VOCs. GBM appears to be a very effective and efficient machine-learning method. It is efficient because it achieves these results with much less computational effort than either of those methods and produces much smaller models. Overall, GBM results somehow are contrary with DNN results.

We also evaluated the performance of all 72 models in term of classification accuracy. Classification accuracy is the ratio of correct predictions to total predictions made and often presented as a percentage by multiplying the result by 100. Figure 4 shows the performance of 72 classification models in term of accuracy value (%) by using all datasets as training and 10-fold cross-validation technique. Also, it can be seen that all data are distributed randomly and there is no correlation between the performance obtained by using all datasets as training and 10-fold cross-validation technique. Similarly to MSE result, we observed that there are two types of models: 1) the right side is affected by the over-fitting problem, and 2) the left side is not changed for both experiments. The right side models, such as model No 12 (Klekota-Roth fingerprint trained with DNN4 method), model No 20 (PubChem fingerprint trained with DNN4 method) and model No 36 (Extended fingerprint trained with DNN4 method) give good classification result when using all datasets as training, however it becomes worst when we used 10-fold cross-validation technique. The small number of our sample data, many parameters of DNN and a large number of features need to be trained might cause this problem, which we have explained previously.

Contrarily, there are few models, which performed poorly in both experiments. The classification model No 43 (atom pair fingerprint trained with the DNN3 method) and model No 59 (Estate fingerprint trained with the DNN3 method) performed poorly in case of using all datasets as training and 10-fold cross-validation technique. This is due to the small number of substructures for Estate fingerprint, which is too

short to characterize molecules. The atom pair fingerprint is also known as a very sensitive fingerprint and this is the reason why it performed poorly in both experiments. Based on Figure 4, we observed that the classification model No 7 (Combine fingerprint trained with GBM method) gives good results in both experiments. This model obtained accuracy value of 94.43% when using all datasets as training and 57.67% in case of 10-fold cross-validation technique. The results show that GBM method is good at predicting biological activities of VOCs. This result somehow is aligned with our previous result shown in Figure 3, where we proved that GBM appears to be a very effective and efficient algorithm, compared to other machine learning methods.

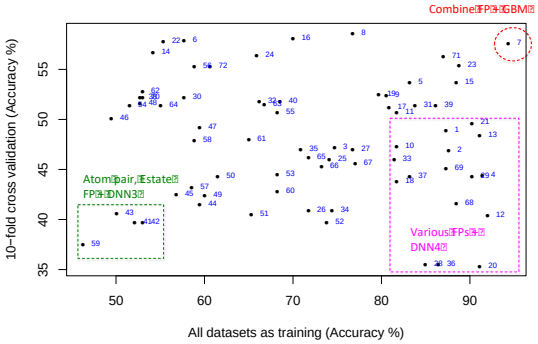


Figure 4: Performance of 72 classification models by using all datasets as training and 10-fold cross-validation technique (accuracy).

## IV. CONCLUSIONS

This study is conducted in order to further investigate the relationships among organisms, volatile metabolites and their corresponding biological activities. We employed supervised machine learning methods to predict biological activities of VOCs based on chemical structures. We have developed 72 classification models for the prediction of biological activities of VOCs by 9 types of fingerprints and trained by the deep neural network (DNN), gradient boosting machine (GBM), random forest (RF) and generalized linear model (GLM). Based on the computational results, PubChem and Combine fingerprints were recommended as the input for the prediction model. Gradient boosting machine (GBM) method can outperform deep neural network (DNN) in term of classifying VOCs, in our case. GBM method has an advantage in term of computational speed and requires less parameter for optimization. Hence, we highly recommend using GBM for the prediction of biological activities of VOCs based on chemical structures.

In future, more VOCs can be accumulated, and comprehensive analysis can be performed in the context of human healthcare and chemical ecology. The prediction outcome may be useful for the discovery of novel agricultural tools and also for the non-invasive identification of biomarkers in the medical diagnostic field.

## ACKNOWLEDGMENT

This work was funded by Universiti Malaysia Perlis (UniMAP) research grant 9009-00053.

## REFERENCES

[1] D. D. Rowan, "Volatile Metabolites," *Metabolites*, vol. 1, no. 1, pp. 41–63, Nov. 2011.

[2] C. N. Kanchiswamy, M. Malnoy, and M. E. Maffei, "Chemical diversity of microbial volatiles and their potential for plant growth and productivity," *Front. Plant Sci.*, vol. 6, no. March, p. 151, 2015.

[3] G. V. P. Reddy and A. Guerrero, "Interactions of insect pheromones and plant semiochemicals," *Trends Plant Sci.*, vol. 9, no. 5, pp. 253–261, 2004.

[4] M. Phillips, R. N. Cataneo, R. Condos, G. a Ring Erickson, J. Greenberg, V. La Bombardi, M. I. Munawar, and O. Tietje, "Volatile biomarkers of pulmonary tuberculosis in the breath.," *Tuberculosis (Edinb).*, vol. 87, no. 1, pp. 44–52, Jan. 2007.

[5] M. Phillips, R. N. Cataneo, C. Saunders, P. Hope, P. Schmitt, and J. Wai, "Volatile biomarkers in the breath of women with breast cancer.," *J. Breath Res.*, vol. 4, no. 2, p. 26003, Jun. 2010.

[6] C. Wang, B. Sun, L. Guo, X. Wang, C. Ke, S. Liu, W. Zhao, S. Luo, Z. Guo, Y. Zhang, G. Xu, and E. Li, "Volatile Organic Metabolites Identify Patients with Breast Cancer, Cyclomastopathy, and Mammary Gland Fibroma," *Sci. Rep.*, vol. 4, pp. 1–6, 2014.

[7] A. Wilson, "Advances in Electronic-Nose Technologies for the Detection of Volatile Biomarker Metabolites in the Human Breath," *Metabolites*, vol. 5, no. 1, pp. 140–163, 2015.

[8] K. E. Pijls, A. Smolinska, D. M. A. E. Jonkers, E. J. C. Moonen, J. W. Dallinga, A. A. M. Masclee, G. H. Koek, and F. J. Van Schooten, "Volatile organic compounds in exhaled air as potential non-invasive biomarker for liver cirrhosis," *J. Hepatol.*, vol. 1), no. November 2015, p. S415, 2014.

[9] R. P. Arasaradnam, M. McFarlane, E. Daulton, J. Skinner, N. O'Connell, S. Wurie, S. Chambers, C. Nwokolo, K. Bardhan, R. Savage, and J. Covington, "Non-invasive exhaled volatile organic biomarker analysis to detect inflammatory bowel disease (IBD)," *Dig. Liver Dis.*, vol. 48, no. 2, pp. 148–153, 2016.

[10] T. Feinberg, J. Herbig, I. Kohl, G. Las, J. C. Cancilla, J. S. Torrecilla, M. Ilouze, H. Haick, and N. Peled, "Cancer metabolism: the volatile signature of glycolysis-in vitro model in lung cancer cells," *J Breath Res*, vol. 11, no. 1, p. 16008, 2017.

[11] M. Hertel, E. Schuette, I. Kastner, S. Hartwig, A. M. Schmidt-Westhausen, R. Preissner, S. Paris, and S. Preissner, "Volatile organic compounds in the breath of oral candidiasis patients: a pilot study," *Clinical Oral Investigations*, pp. 1–11, 2017.

[12] M. Ashrafi, M. Bates, M. Baguneid, T. Alonso-Rasgado, R. Rautemaa-Richardson, and A. Bayat, "Volatile organic compound detection as a potential means of diagnosing cutaneous wound infections," *Wound Repair and Regeneration*, 2017.

[13] A. A. Abdullah, M. Altaf-Ul-Amin, N. Ono, T. Sato, T. Sugiura, A. H. Morita, T. Katsuragi, A. Muto, T. Nishioka, and S. Kanaya, "Development and Mining of a Volatile Organic Compound Database," *Biomed Res. Int.*, vol. 2015, 2015.

[14] J. Dong, D. S. Cao, H. Y. Miao, S. Liu, B. C. Deng, Y. H. Yun, N. N. Wang, A. P. Lu, W. Bin Zeng, and A. F. Chen, "ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation," *J. Cheminform.*, vol. 7, no. 1, pp. 1–10, 2015.

[15] J. B. O. Mitchell, "Machine learning methods in chemoinformatics," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 4, no. October, p. n/a-n/a, 2014.

[16] A. Lavecchia, "Machine-learning approaches in drug discovery: Methods and applications," *Drug Discovery Today*, vol. 20, no. 3. pp. 318–331, 2015.

[17] H. Khani, M. B. Sepehrifar, and S. Yarahmadian, "An improvement on the prediction power of the 3D-QSAR CoMFA models using a hybrid of statistical and machine learning methods: a case study on γ−secretase modulators of Alzheimer's disease," *Med. Chem. Res.*, vol. 26, no. 6, pp. 1184–1200, 2017.

[18] M. Intelligence, "High Performance Machine Learning in R with H2O," no. October, 2015.

[19] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, 2015.

[20] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Mol. Inform.*, vol. 35, no. 1, pp. 3–14, 2016.

[21] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[22] T. C. Kietzmann, P. Mcclure, and N. Kriegeskorte, "Deep Neural Networks in Computational Neuroscience," *bioRxiv*, vol. 133504, 2017.

[23] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, "Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data," *J. Cheminform.*, vol. 9, no. 1, 2017.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[25] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[26] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. Gifford, "Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships," *J. Chem. Inf. Model.*, p. acs.jcim.6b00591, 2016.
[27] J. Lu, D. Lu, X. Zhang, Y. Bi, K. Cheng, M. Zheng, and X. Luo, "Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine," *Biochim. Biophys. Acta - Gen. Subj.*, vol. 1860, no. 11, pp. 2664–2671, 2016.
[28] L. Zhang, H. Ai, W. Chen, Z. Yin, H. Hu, J. Zhu, J. Zhao, Q. Zhao, and H. Liu, "CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods," *Sci. Rep.*, vol. 7, no. 1, p. 2118, 2017.
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[30] J. Nelder and R. Wedderburn, "Generalized Linear Models," *J. R. Stat. Soc. Ser. A*, vol. 135, pp. 370–384, 1972.