Poster Reprint

ASMS 2020
WP 165

Classifying the pesticides in foods between GC-amenable and LC-amenable using the prediction model with molecular descriptors

Sadao Nakamura [1], Takeshi Serino [1, 2],
Takeshi Otsuka [1], Yoshizumi Takigawa [1],
Tarun Anumol [3], Shigehiko Kanaya [2]

[1] Agilent Technologies, Hachioji, Tokyo, Japan

[2] Nara Institute of Science and Technology,
Ikoma, Nara, Japan

[3] Agilent Technologies, Wilmington, DE,
United States

Introduction

One of the frequently asked questions by food analysis chemists who are currently using GC/MS or LC/MS for residual pesticides in foods is whether that pesticide is “GC-amenable” or "LC-amenable”. This is because neither LC/MS nor GC/MS can analyze all pesticides by any single technology, comparisons of the pesticides with both LC/MS and GC/MS have been researched[1,2]. There are several guidelines for the selection between LC-amenable and GC-amenable for pesticides based on the physical and chemical properties[3], and experienced chemists can predict the answer to this question based on the experiences for some degree. A prediction model for classifying the amenabilities of pesticides between GC-amenable and LC-amenable is developed by the quantitative structure-property relationship (QSPR) approach for answering to this question.

Experimental

Preparation of the pesticide list by validated report

Pesticide information for classification model were obtained from two validation reports of residual pesticide analysis in foods[4,5] as below. Details of the pesticides and technologies are listed in the Table 1.

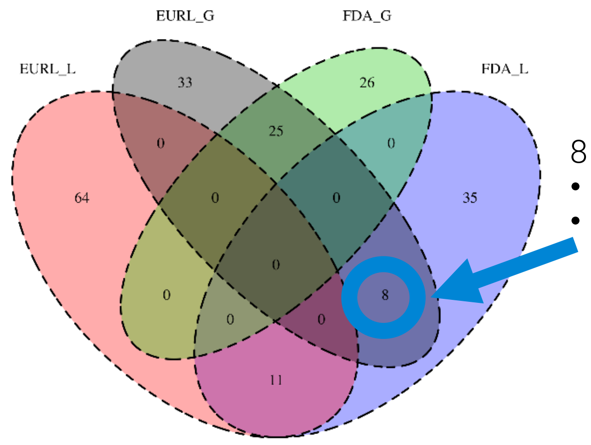
- U.S. Food and Drug Administration(FDA) List[4]

The validation report of 136 pesticides analysis in Avocado using both LC/MS and GC/MS.

- EU Reference Laboratories for Residues of Pesticides(EURL)[5]

The validation report of 127 pesticides analysis in Olive Oil using both LC/MS and GC/MS.

202 pesticides in total are included in both literatures. For improving the classification capability of machine learning, 8 pesticides were excluded from the machine learning which were analyzed differently between both, i.e. by GC/MS in EURL while by LC/MS in FDA as shown in Figure 1. 194 pesticides listed in the Table 1 were used.



8 pesticides were excluded
- LC/MS used in FDA
- GC/MS used in EURL

Figure 1. Venn diagram to describe the number of pesticides by the list(FDA or EURL) and the technology used for analysis (L:LC/MS and G:GC/MS).

Experimental

Molecular descriptors of the pesticides

The canonical SMILES of 194 pesticides were obtained from the PubChem website as listed in the Table 2. 224 molecular descriptors (MDs) of these pesticides were obtained by rcdk package of R program. The MDs with the zero variance among 194 pesticides were removed in order to avoid the errors in machine learning execution, 176 MDs were eventually obtained for machine learning. Each molecular descriptor was standardized for comparison as expressed by the Equation (Eq.1), where z_i is the standardized value to be used for machine learning, x_i is the raw value from rcdk, μ_i is the average of 194 pesticides and σ_i is the standard deviation of 194 pesticides for *ith* molecular descriptor.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \qquad \text{(Eq.1)}$$

Classification of pesticides by the machine learning

Either G(GC/MS) or L(LC/MS) of technology flag is assigned on the 194 pesticides based on the literatures as Table 1. 119 machine learning methods of the classification in caret package listed in the Table 3[6] are evaluated in the present study.

Table 1. Pesticides and technologies used in the list. Technology "L" is analyzed by LC/MS, "G" is GC/MS. List of "E" is EURL list, "F" is FDA list and "Both" is both EURL and FDA list.

Pesticide	Tech	List	Pesticide	Tech	List	Pesticide	Tech	List	Pesticide	Tech	List
alpna-BHD	G	Both	Deltamethrin	G	E	Flusilazole	G	E	Pendimethalin	G	E
alpha-endosulfan	G	Both	desmedipham	L	F	utolanil	L	F	pentachloroaniline	G	F
acetamiprid	L	Both	Desmethyl Pirimicarb	L	E	Flutriafol	L	E	pentachlorobenzene	G	F
Aldicarb	L	E	dichloruanid	L	F	Fluvalinate	G	Both	permethrin	G	F
Aldicarb Sulfone	L	E	dichlorvos	L	F	Forchlorfenuron	L	E	Pethoxamid	L	E
Aldicarb Sulfoxide	L	E	Dicloran	G	E	Furalaxyl	G	E	Phenthoate	G	E
ametryn	L	F	dicrotophos	L	Both	heptachlor epoxide	G	F	phosalone	G	Both
aminocarb	L	F	dieldrin	G	F	hexachlorobenzene	G	F	phosmet	L	Both
amitraz	G	F	difenoconazole	L	Both	hexaconazole	L	F	Picolinafen	G	E
azinphos-methyl	L	F	Diufenican	G	E	Hexythiazox	L	E	Picoxystrobin	L	E
Azoxystrobin	L	E	Dimefuron	L	E	imazalil	L	F	piperonyl butoxide	L	F
b-endosulfan	G	Both	Dimethachlor	L	E	Imidacloprid	L	E	Piridafenthion	G	E
Benalaxyl	G	E	Dimethenamid	L	E	iprodione	G	Both	pirimiphos-methyl	G	F
bendiocarb	L	F	dimethoate	L	Both	Iprovalicarb	L	E	prochloraz	L	F
Benuralin	G	Both	dimethomorph	L	Both	Isocarbophos	G	E	procymidone	G	Both
Bifenox	G	E	Dimoxystrobin	L	E	Isofenphos-Methyl	G	E	profenofos	G	Both
bifenthrin	L	F	Diniconazole	L	E	linuron	L	Both	prometryn	L	F
boscalid	L	F	dinitramine	G	F	Malaoxon	L	E	pronamide	G	Both
bromopropylate	G	Both	dioxacarb	L	F	Mepanipyrim	G	E	propachlor	L	F
Bupirimate	G	E	Dmst	L	E	Metalaxyl	G	E	propanil	G	F
cadusafos	G	F	endosulfan sulphate	G	Both	Metamitron	L	E	propargite	L	F
Carbendazim	L	E	endrin	G	F	Metconazole	L	E	Pymetrozine	L	E
Carbofuran	L	E	EPN	G	F	methamidophos	L	F	Pyraclostrobin	L	E
Carbofuran 3-Oh	L	E	epoxiconazole	L	Both	Methidathion	G	E	Pyrazophos	G	E
Carfentrazone Ethyl	L	E	Etaconazol	L	E	Methiocarb	L	E	Pyridaben	G	E
Chlofenvinphos	G	E	ethiolate	L	F	Methiocarb Sulfone	L	E	pyriproxifen	G	Both
chlordimeform	L	F	ethofumesate	L	F	Methiocarb Sulfoxide	L	E	quinalphos	G	Both
Chlorfenapyr	G	E	Ethoprophos	G	E	Methomyl	L	E	Tebuconazole	G	E
Chloridazon	L	E	Etridiazole	G	F	methyl parathion	G	Both	Teuthrin	G	E
chlorothalonil	G	F	Fenamiphos	L	E	metolachlor	L	F	Terbufos	G	E
Chloroxuron	L	E	Fenamiphos Sulfone	L	E	metolcarb	L	F	Terbutryn	L	E
chlorpyrifos-methyl	G	Both	Fenamiphos Sulfoxide	L	E	Metosulam	L	E	Tetraconazole	G	E
Chlorthiophos	G	E	fenarimol	G	Both	mevinphos	L	F	tetradifon	G	F
Chlozolinate	G	E	fenbuconazole	L	F	MGK-264	G	F	Thiabendazole	L	E
Clortoluron	L	E	Fenitrothion	G	E	monocrotophos	L	Both	Thiacloprid	L	E
Clothianidin	L	E	Fenobucarb	L	E	monolinuron	L	F	Thiamethoxam	L	E
coumaphos	L	F	fenoxycarb	L	F	napropamide	G	F	tolclofos-methyl	G	Both
cyanazine	L	F	Fenpropathrin	G	E	Neburon	L	E	Triadimefon	L	E
cycluron	L	F	fenpropimorph	L	F	o-phenylphenol	G	Both	Triadimenol	L	E
Cyuthrin	G	E	Fenpyroximate	L	E	o,p-methoxychlor	G	F	triallate	G	F
cyhalothrin	G	Both	Fenuron	L	E	omethoate	L	Both	Triazophos	G	E
Cymoxanil	L	E	fenvalerate	G	Both	oxadixyl	G	F	Trioxystrobin	L	E
cypermethrin	G	Both	Flazasulfuron	L	E	Oxamyl	L	E	Triumizole	L	E
cyproconazole	L	F	udioxinil	L	F	Oxyuorfen	G	E	Triuralin	G	Both
dacthal	G	Both	Flufenacet	L	E	Paclobutrazole	L	E	Triticonazole	L	E
DDE(4,4')	G	F	Fluopicolide	L	E	Paraoxon Methyl	L	E	vinclozolin	G	Both
DDT(2,4')	G	F	Fluoxastrobin	L	E	parathion	G	F	Zoxamide	L	E
DDT(4,4')	G	F	uquinconazole	L	Both	penconazole	L	F			
DEF	G	F	Flurtamone	L	E	Pencycuron	L	E			

These machine learning methods are expected to classify the 194 pesticides between G(GC/MS) or L(LC/MS) using the 176 molecular descriptors. Prediction performance of classification is measured by the accuracy of resamples from the 10-fold cross-validation(CV10) iterations and execution time. Execution time is obtained by the "System.time()" command of R package.

Table 2. 176 molecular descriptors in present study

Descriptor Class	Descriptor (Description)
ALOGP Descriptor (2)	ALogP (Ghose-Crippen LogKow), ALogP2 (Square of ALogP)
APol Descriptor (1)	Apol (Sum of the atomic polarizabilities (including implicit hydrogens)
Aromatic Atoms Count Descriptor (1)	naAromAtom (Number of aromatic atoms)
Aromatic Bonds Count Descriptor (1)	nAromBond (Number of aromatic bonds)
Atom Count Descriptor (2)	nAtom (Number of atoms), nB (Number of boron atoms)
Autocorrelation Descriptor Charge (5)	ATSc1, ATSc2, ATSc3, ATSc4, ATSc5 (ATS autocorrelation descriptor, weighted by charges)
Autocorrelation Descriptor Mass (5)	ATSm1, ATSm2, ATSm3, ATSm4, ATSm5 (ATS autocorrelation descriptor, weighted by scaled atomic mass)
Autocorrelation Descriptor Polarizability (5)	ATSp1, ATSp2, ATSp3, ATSp4, ATSp5 (ATS autocorrelation descriptor, weighted by polarizability)
BCUT Descriptor (6)	BCUTw.1l (nhigh lowest atom weighted BCUTS), BCUTw.1h (nlow highest atom), BCUTc.1l (nhigh lowest partial charge), BCUTc.1h (nlow highest partial charge) BCUTp.1l (nhigh lowest polarizability), BCUTp.1h (nlow highest polarizability)
BPolDescriptor (1)	bpol (Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens))
Carbon Types Descriptor (9)	C1SP1 (Triply bound carbon bound to one other carbon), C2SP1 (Triply bound carbon bound to two other carbons), C1SP2 (Doubly hound carbon bound to one other carbon), C2SP2 (Doubly bound carbon bound to two other carbons), C3SP2 (Doubly bound carbon bound to three other carbons), C1SP3 (Singly bound carbon bound to one other carbon), C2SP3 (Singly bound carbon bound to two other carbons), C3SP3 (Singly bound carbon bound to three other carbons), C4SP3 (Singly bound carbon bound to four other carbons)
Chi Chain Descriptor (10)	SCH.3-7 (Simple chain, orders 3-7), VCH.3-7 (Valence chain, orders 3-7)
Chi Cluster Descriptor (8)	SC.3-6 (Simple cluster, orders 3-6) , VC.3-6 (Valence cluster, orders 3-6)
Chi Path Cluster Descriptor (6)	SPC.4-6 (Simple path cluster, orders 4 to 6), VPC.4-6 (Valence path cluster, orders 4-6)
Chi Path Descriptor (16)	SP.0-7 (Simple path, orders 0-7), VP.0-7Valence path, orders 0-7
Eccentric Connectivity Index Descriptor (37)	ECCEN (A topological descriptor combining distance and adjacency information), khs.sCH3 (Count of atom-type E-State: -CH3), khs.dCH2 (=CH2), khs.ssCH2 (-CH2-), khs.tCH (#CH), khs.dsCH (=CH-), khs.aaCH (:CH:), khs.sssCH (>CH-), khs.tsC (#C-), khs.dssC (=C<), khs.aasC (:C:-), khs.aaaC (::C:), khs.ssssC (>C<), khs.sNH2 (-NH2), khs.ssNH (-NH2-+), khs.aaNH (:NH:), khs.tN (#N), khs.dsN (=N-), khs.aaN (:N:), khs.sssN (>N-), khs.ddsN (-N<<), khs.aasN (:N:-), khs.sOH (-OH), khs.dO (=O), khs.ssO (-O-), khs.aaO (:O:), khs.sF (-F), khs.ssssSi (>Si<), khs.dsssP (->P=), khs.dS (=S), khs.ssS (-S-), khs.aaS (aSa), khs.dssS (>S=), khs.ddssS (>S==), khs.sCl (-Cl), khs.sBr (-Br)
Fragment Complexity Descriptor (1)	fragC (Complexity of a system)
Ghose Crippen Molecular Refractivity Descriptor (1)	AMR (Molar refractivity)
H Bond Acceptor Count Descriptor (1)	nHBAcc (Number of hydrogen bond acceptors)
H Bond Donor Count Descriptor (1)	nHBDon (Number of hydrogen bond donors)
KappaShape Indices Descriptor (3)	Kier1-3 (First, Second, Third kappa (κ) shape indexes)
Largest Chain Descriptor (1)	nAtomLC (Number of atoms in the largest chain)
Longest Aliphatic Chain Descriptor (1)	nAtomLAC (Number of atoms in the longest aliphatic chain)
Mannhold LogP Descriptor (1)	MLogP (Mannhold LogP)
MDEDescriptor (19)	MDEC.11 (Molecular distance edge between all primary carbons), MDEC.12 (between all primary and secondary carbons), MDEC.13 (between all primary and tertiary carbons), MDEC.14 (between all primary and quaternary carbons), MDEC.22 (between all secondary carbons), MDEC.23 (between all secondary and tertiary carbons), MDEC.24 (between all secondary and quaternary carbons), MDEC.33 (between all tertiary carbons), MDEC.34 (between all tertiary and quaternary carbons), MDEC.44 (between all quaternary carbons), MDEO.11 (between all primary oxygens), MDEO.12 (between all primary and secondary oxygens), MDEO.22 (between all secondary oxygens), MDEN.11 (between all primary nitrogens), MDEN.12 (between all primary and secondary nitrogens), MDEN.13 (between all primary and tertiary niroqens), MDEN.22 (between all secondary nitroqens), MDEN.23 (between all secondary and tertiary nitrogens), MDEN.33 (between all tertiary nitrogens)
PetitjeanNumberDescriptor (1)	PetitjeanNumber (Petitjean number)
RotatableBondsCountDescriptor (1)	nRotB (Number of rotatable bonds, excluding terminal bonds)
RuleOfFiveDescriptor (1)	LipinskiFailures (Number failures of the Lipinski's Rule Of 5)
TPSADescriptor (19)	TopoPSA (Topological polar surface area)
VAdjMaDescriptor (1)	VAdjMat (Vertex adjacency information (magnitude))
WeightDescriptor (1)	MW (Molecular weight)
WeightedPathDescriptor (5)	WTPT.1 (Molecular ID), WTPT.2 (Molecular ID / number of atoms), WTPT.3 (Sum of path lengths starting from heteroatoms), WTPT.4 (Sum of path lengths starting from oxygens), WTPT.5 (Sum of path lengths starting from nitrogens)
WienerNumbersDescriptor (2)	WPATH (Weiner path number), WPOL (Weiner polarity number)
XLogPDescriptor (1)	XLogP (XLogP)
ZagrebIndexDescriptor (1)	Zagreb (Sum of the squares of atom degree over all heavy atoms i)
Petitjean Shape Index Descriptor (1)	topoShape (Petitjean topological shape index)
Others (16)	nBase (Basic group count descriptor), nSmallRings (the number of small rings from size 3 to 9), nAromRings (the number of aromatic rings), nRingBlocks (total number of distinct ring blocks), nAromBlocks (total number of "aromatically connected components"), nRings3, 5, 6, 7 (individual breakdown of small rings), tpsaEfficiency (Polar surface area expressed as a ratio to molecular size), VABC (Atomic and Bond Contributions of van der Waals volume), HybRatio (the ratio of heavy atoms in the framework to the total number of heavy atoms in the molecule.), tpsaEfficiency.1 (Polar surface area expressed as a ratio to molecular size), TopoPSA.1 (Topological polar surface area), topoShape.1(A measure of the anisotropy in a molecule)

Table 3. Machine Learning methods for regression analysis used in present study

Algorithm	Methods in caret
(a) Ordinary learning methods	
Kernel (17)	dwdPoly, dwdRadial, gaussprRadial, kernelpls, lssvmRadial, stepQDA, svmLinear, svmLinear2, svmLinear3, svmLinearWeights, svmLinearWeights2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmRadialWeights, widekernelpls
Simple Linear (12)	bayesglm, CSimca, glm, glmStepAIC, multinom, ordinalNet, plr, pls, regLogistic, rrlda, RSimca, simpls
Sparse modeling (2)	glmnet, sdwd
Neural Network (11)	avNNet, dnn, mlp, mlpML, mlpWeightDecay, mlpWeightDecayML, monmlp, msaenet, nnet, pcaNNet, rbfDDA
Decision Tree (18)	C5.0, C5.0Cost,C5.0Rules, C5.0Tree, ctree, ctree2, deepboost, evtree, J48, JRip, LMT, OneR, PART, rpart, rpart1SE, rpart2, rpartCost, rpartScore
Centroid,kNN (6)	knn, kknn, lvq, ownn, pam, snn
Spline (4)	earth, gamLoess, gamSpline, gcvEarth
Naive Bayes (2)	naive bayes, nb
Others (13)	dwdLinear, fda, hdda, null, pda, pda2, rda, rocc, sda, slda, sparseLDA, stepLDA, xyf
(b) Ensemble learning methods	
Decision Tree (26)	ada, AdaBag, adaboost, AdaBoost.M1, blackboost, bstTree, cforest, extraTrees, gbm, nodeHarvest, ORFpls, ORFridge, ORFsvm, parRF, ranger, Rborist, rf, rFerns, rfRules, rotationForest, rotationForestCp, RRF, RRFglobal, treebag, wsrf, xgbTree
Simple Linear (3)	glmboost, LogitBoost, xgbLinear
Spline (4)	bagEarth, bagEarthGCV, bagFDA, xgbDART

Results and Discussion

Classification Performance (Accuracy of CV10 resample)

The box plot in the Figure 2. shows the distribution of accuracy for each machine learning method category. The overall accuracy of CV10 is calculated by the Eq. 2.

$$CV10\ Accuracy(\%) = \frac{\sum_{i=1}^{10}(Number\ of\ accurate\ classification\ of\ ith\ test)}{\sum_{i=1}^{10}(Total\ number\ of\ ith\ test)} \times 100 \quad \text{(Eq. 2)}$$

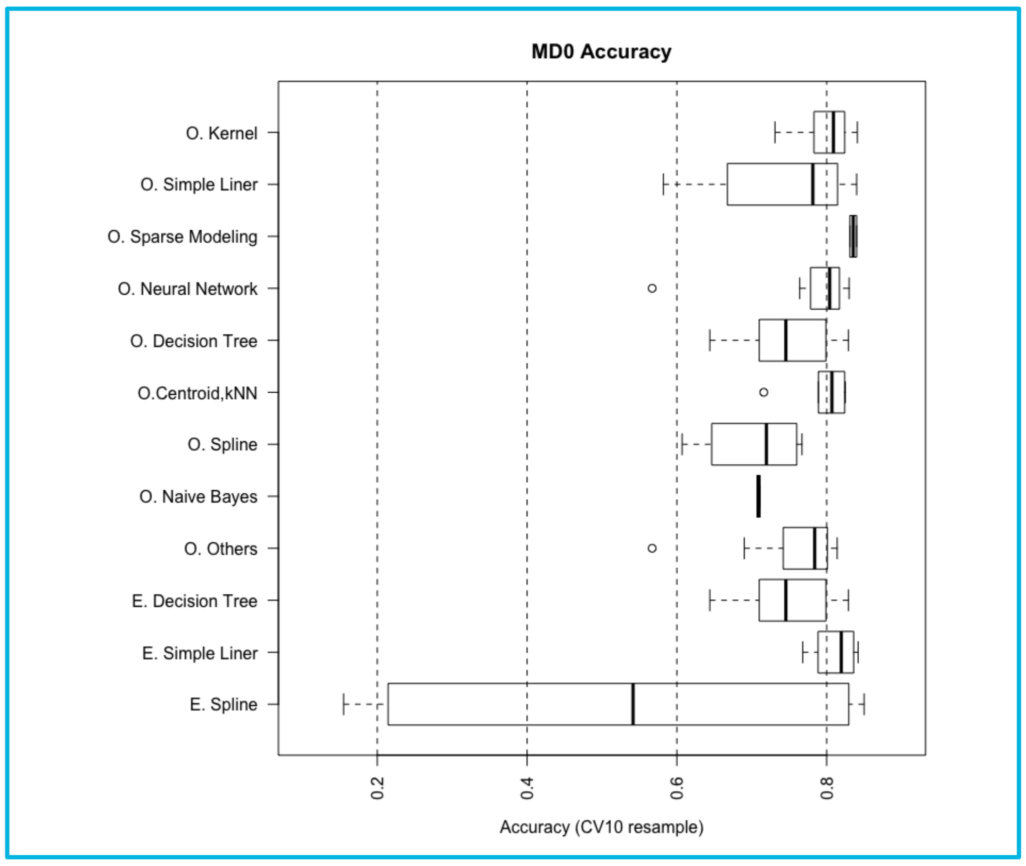


Figure 2. Accuracy of classification (CV10 resample) for 119 machine learning methods

Overall accuracy across the 119 methods was 77%. According to Figure 3, machine learning methods in the ensemble spline method category show larger variability in accuracy than the others. Four machine learning methods, bagEarth (Bagging Earth, 27%), bagEarthGCV (Bagging Earth generalized cross validation, 16%), bagFDA (Bagging flexible discriminant analysis, 81%) and xgbDART (eXtreme Gradient Boosting Dropouts Additive Regression Trees, 85%) were included on this category. According to this result, two methods of bagging earth were not suitable in classifications for this data set.

Execution time(ET)

The result of ET of each the machine learning method is shown in the Figure 3. Methods of ordinary neural network(ranged LogET 1.08 to 2.36) and ensemble spline categories(LogET 1.63 to 2.71) require more execution time than the other categories. The machine learning method with the maximum ET is glmStepAIC(Generalized Linear Model with Stepwise Feature Selection) with the LogET 4.12, i.e. 3 hours and 41 minutes.

Results and Discussion

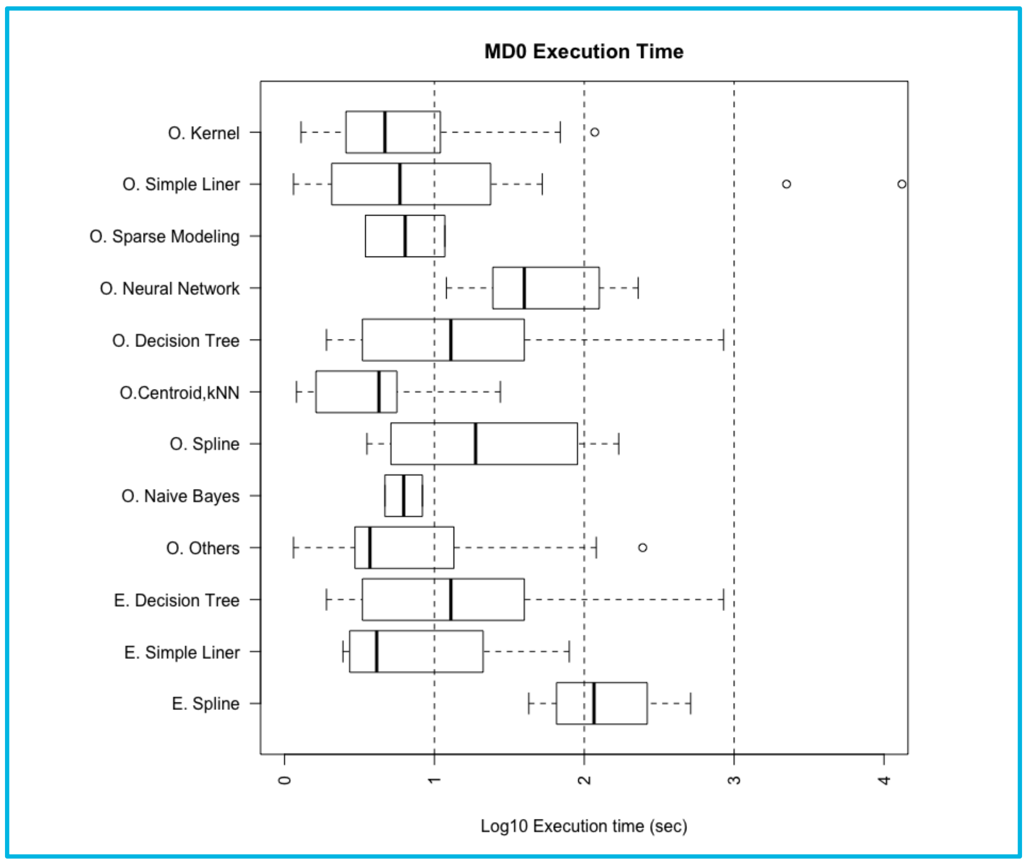


Figure 3. Execution Time for 119 machine learning methods

Total performance - both Accuracy and Execution Time

The results of Accuracy and ET by the machine learning method are shown in the Figure 4.

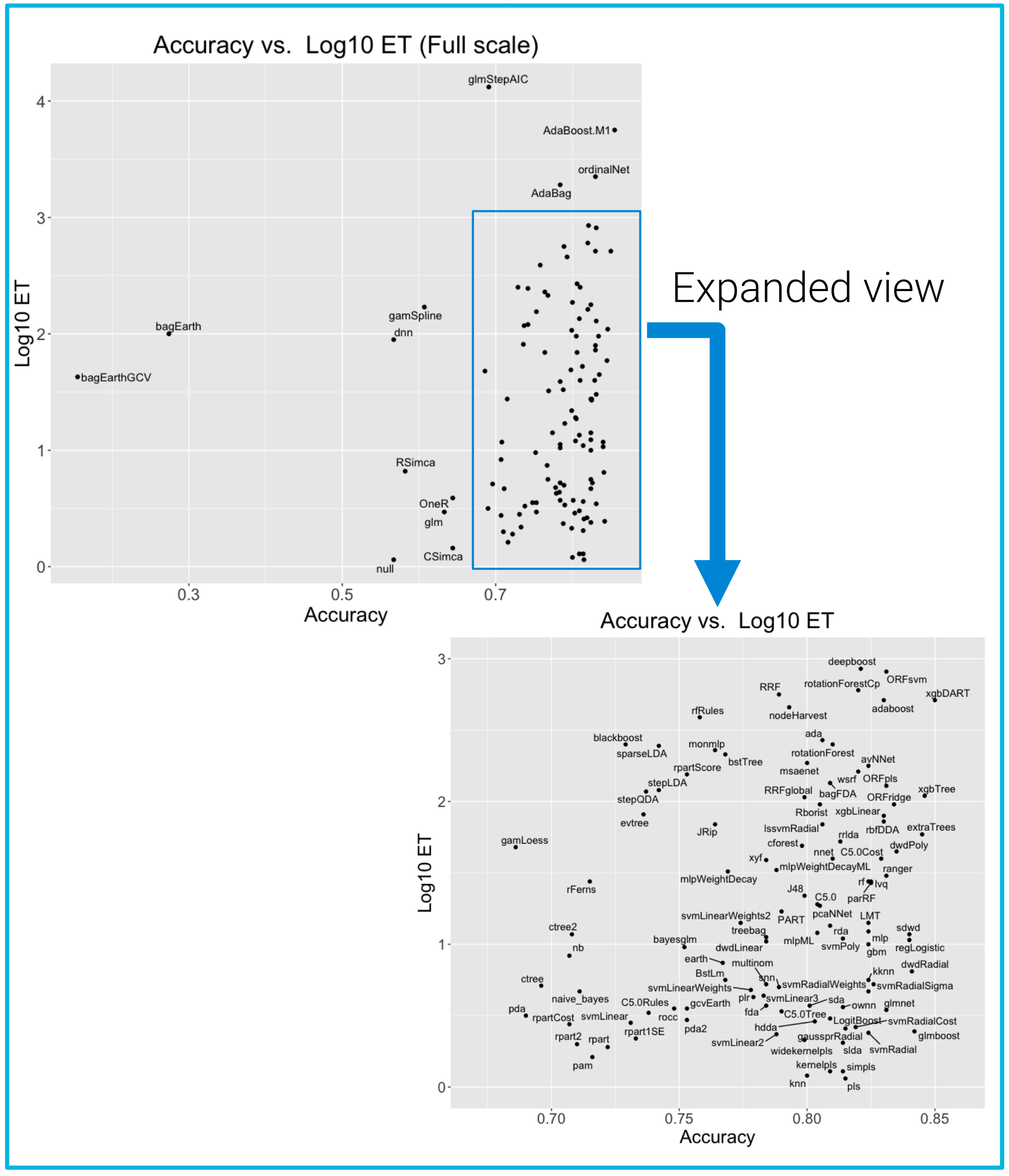


Figure 4. Accuracy and Execution time for 119 machine learning methods

Best 20 machine learning methods in accuracy ranged from 85.5%(AdaBoost.M1) to 83% (svmRadialSigma). Six methods of Ensemble Decision Tree showed higher accuracy for the present data set of GC/MS and LC/MS amenability. The best machine learning method of accuracy is AdaBoost.M1, but it requires 5,600 seconds (1 hour and 34 minutes). The method with higher accuracy with shorter ET is xgbTree, 84.6% within 2 minumes. xgbDART (85.0% accuracy with 8 minutes 33 seconds) was higher accuracy with the moderate ET. xgbDART is highly recommended among 119 methods for the present study with higher accuracy and reasonable execution time for classification.

Conclusions

The classification method of pesticides amenability between LC and GC is developed using the QSPR approach, 119 machine learning methods for classification using 176 molecular descriptors obtained by the 194 pesticides of two validation reports. Prediction accuracy and execution time are the measure of the machine learning method performance.

The recommended machine learning method for the present study is xgbDART with 85.0 % accuracy that requires less than 9 minutes for execution.

References

[1] Z. Barganska, P. Konieczka and J. Namiesnik. 2018. Comparison of Two Methods for the Determination of Selected Pesticides in Honey and Honeybee Samples. *Molecules* **23**: 2582.

[2] C. Anagnostopoulos and G.E.Miliadis. 2013. Development and validation of an easy multiresidue method for the determination of multiclass pesticide residues using GC–MS/MS and LC–MS/MS in olive oil and olives. *Talanta* **1121**: 1-10.

[3] Pesticide Analytical Manual Vol. I, Appendix II, Food and Drug Administration. 1999.

[4] N. Chamkasem, L. W. Ollis, T. Harmon, S. Lee and Greg Mercer. 2013. Analysis of 136 Pesticides in Avocado Using a Modified QuEChERS Method with LC/MS/MS and GC/MS/MS. *J. Agric. Food Chem.* **61**: 2315 - 2329.

[5] EURL-FV(2012-M6) Validation Data of 127 Pesticides Using a Multiresidue Method by LC/MS/MS and GC/MS/MS in Olive Oil, EU Reference Laboratories for residues of pesticides. 2012.

[6] M. Kuhn. 2008. Building Predictive Models in R Using the caret Package. *J. Statistical Software* **28**: 1 - 26.

This information is subject to change without notice.

Published in USA, June 1, 2020

Agilent
Trusted Answers