chemosensors

MDPI

Article

Using Near-Infrared Spectroscopy and Stacked Regression for the Simultaneous Determination of Fresh Cattle and Poultry Manure Chemical Properties

Elizabeth Cobbinah [1,†], Oliver Generalao [2,†], Sathish Kumar Lageshetty [3], Indra Adrianto [4,5], Seema Singh [6] and Gerard G. Dumancas [1,*]

[1] Department of Chemistry, Loyola Science Center, The University of Scranton, Scranton, PA 18510, USA
[2] Center for Informatics, University of San Agustin, Gen. Luna St, Iloilo City 5000, Philippines
[3] Research and Development Department, CHASM Advanced Materials, 2501 Technology Place, Norman, OK 73071, USA
[4] Department of Public Health Sciences, Henry Ford Health, Detroit, MI 48202, USA
[5] Department of Medicine, Michigan State University, East Lansing, MI 48824, USA
[6] Sandia National Laboratories and Joint Bioenergy Institute, Livermore, CA 94550, USA
* Correspondence: gerard.dumancas@scranton.edu; Tel.: +1-405-730-8752
† These authors contributed equally to this work.

Supplementary figures

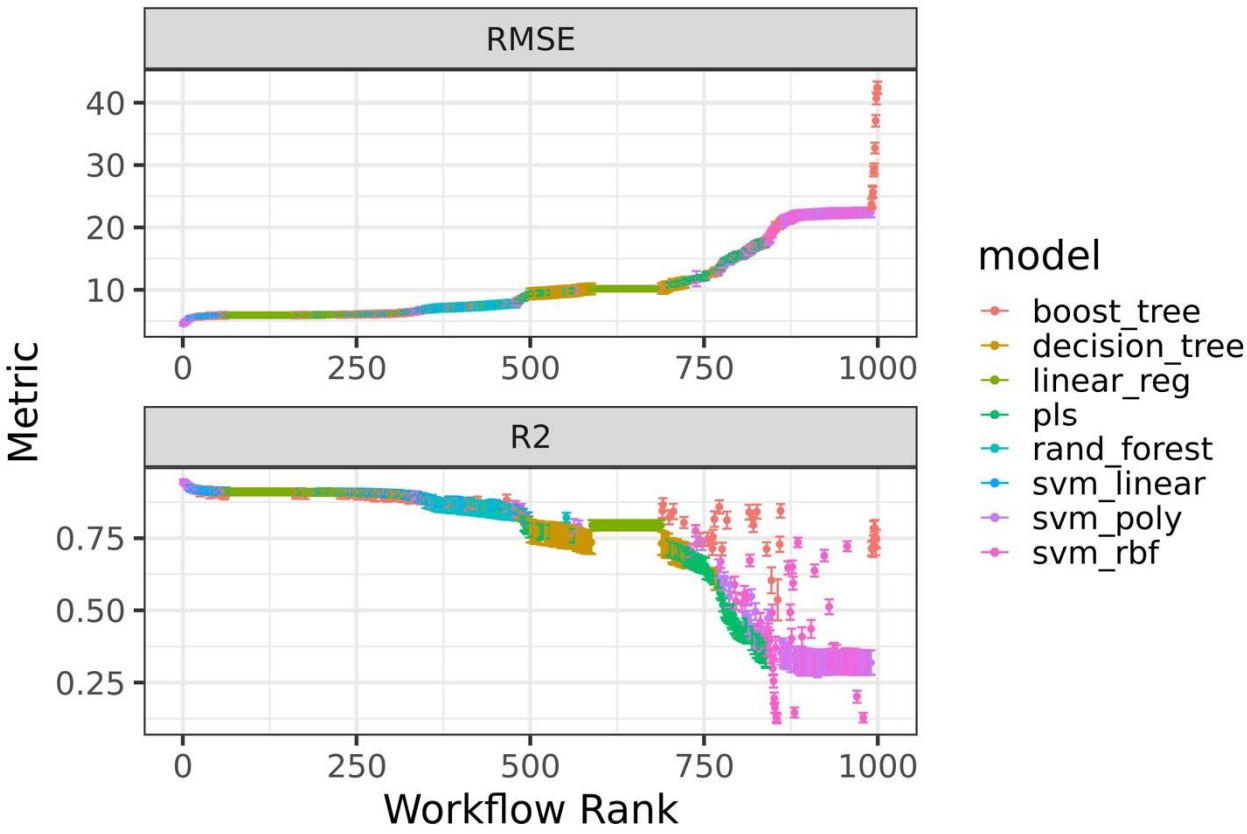

Figure S1. Workflow rank of the machine learning technique used in the dry matter analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following: svm_linear = support vector regression with linear kernel; svm_poly = support vector regression

with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg (includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees ; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

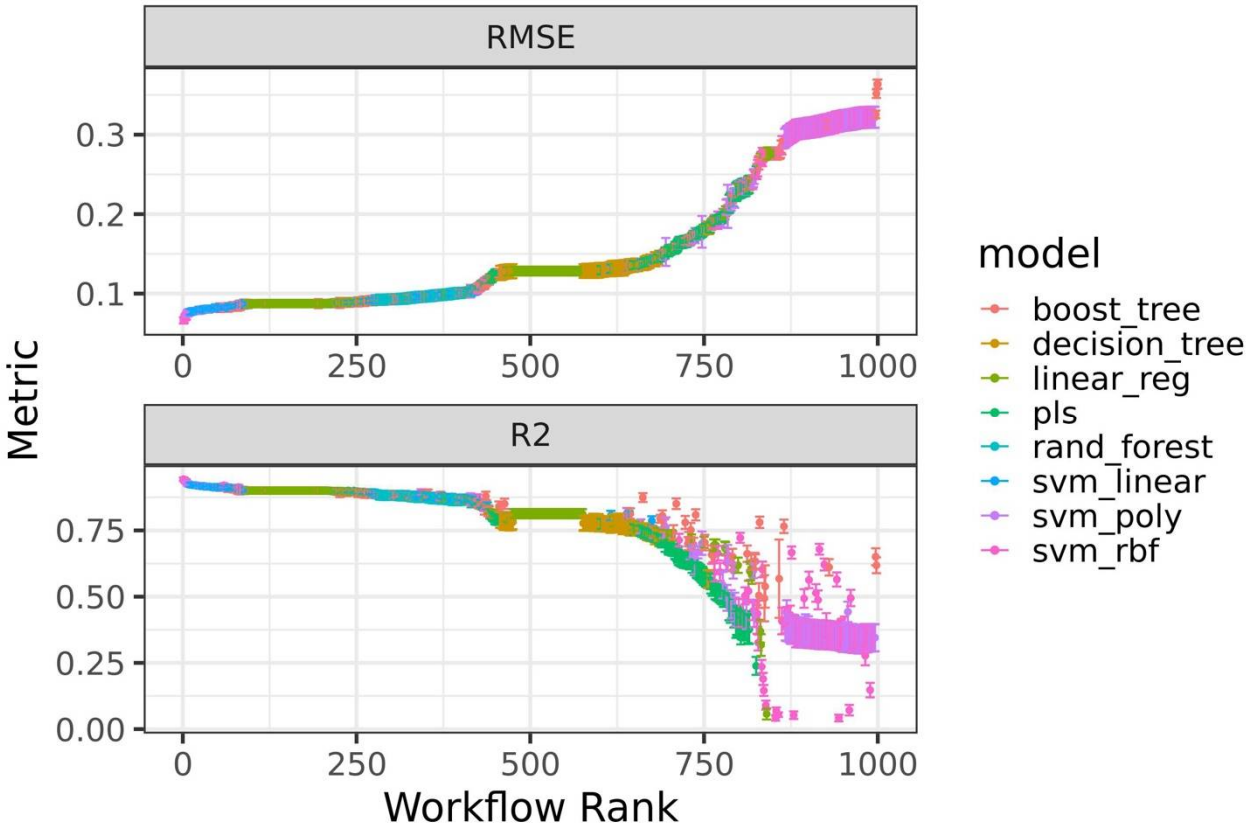

Figure S2. Workflow rank of the machine learning technique used in the NH_4 analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following: svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg (includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

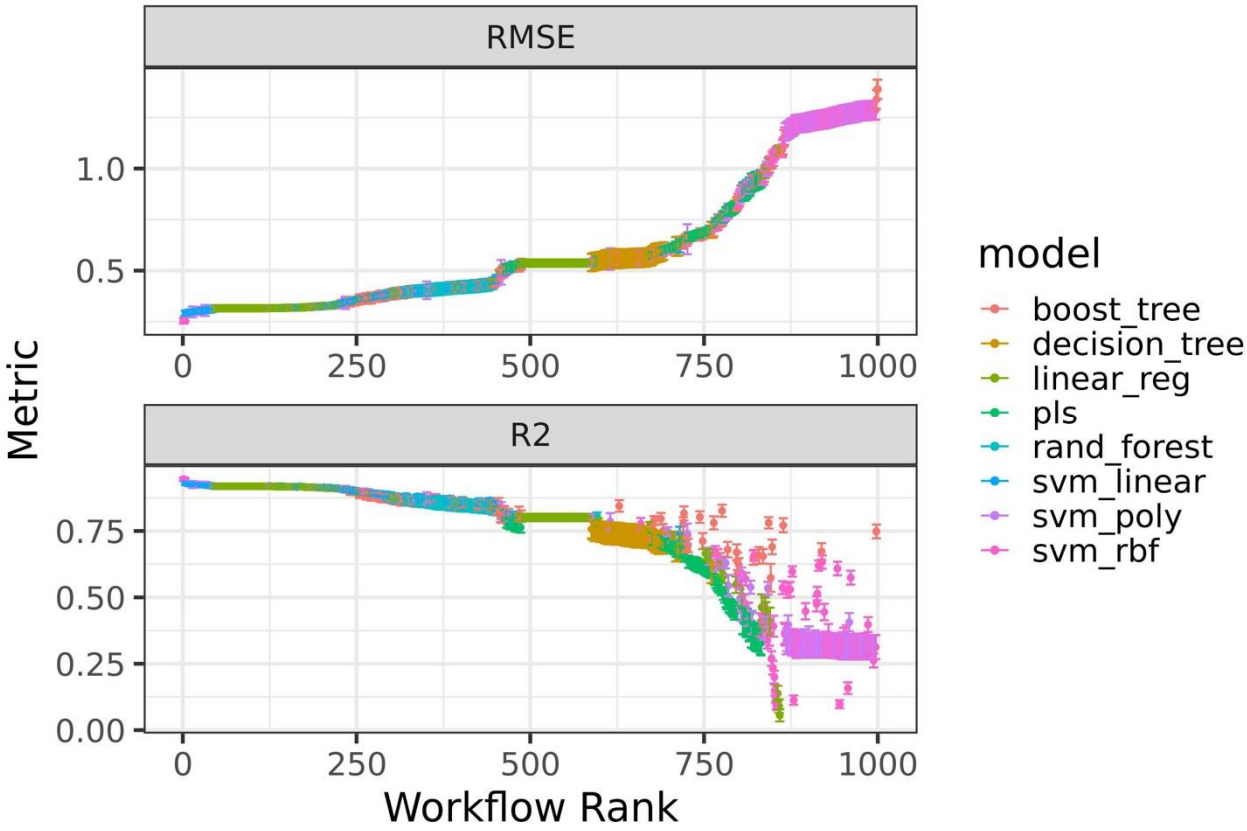

Figure S3. Workflow rank of the machine learning technique used in the total N analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following; svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg (includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

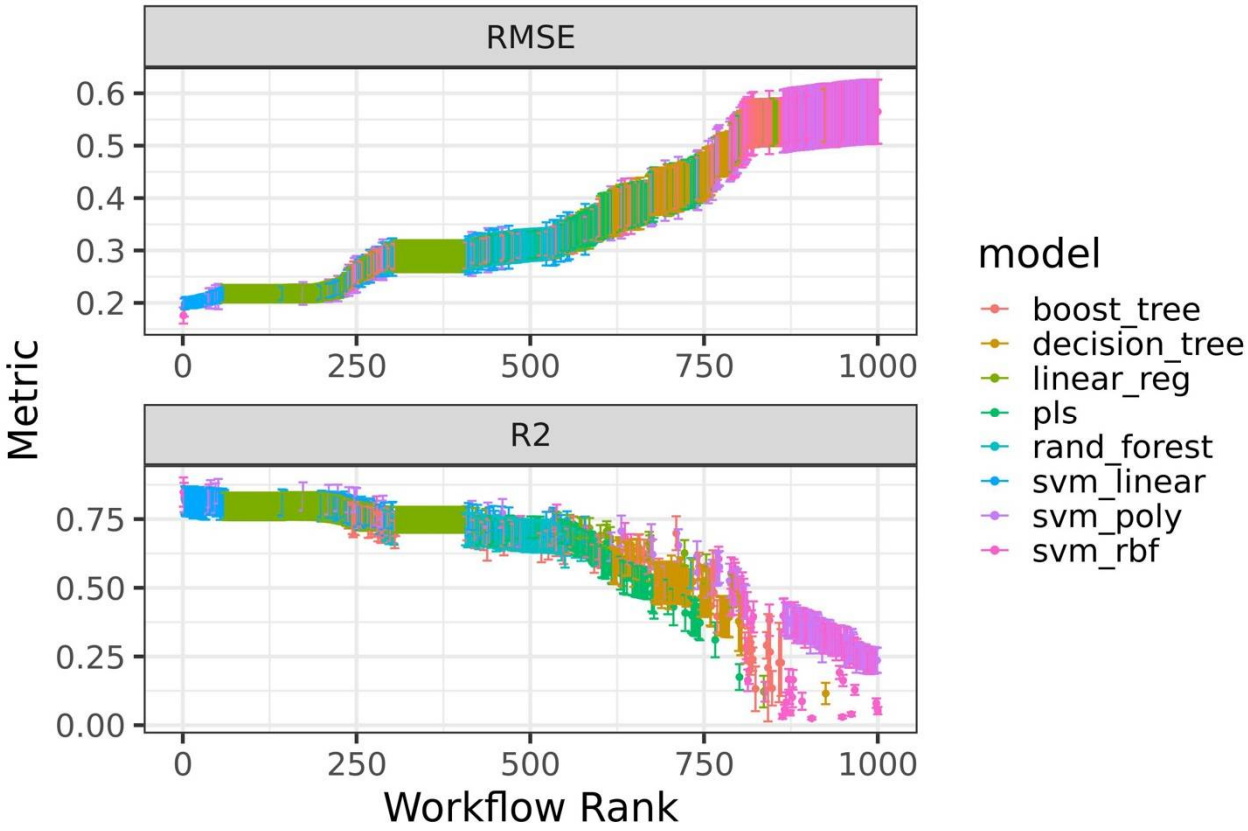

Figure S4. Workflow rank of the machine learning technique used in the P_2O_5 analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following; svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg (includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

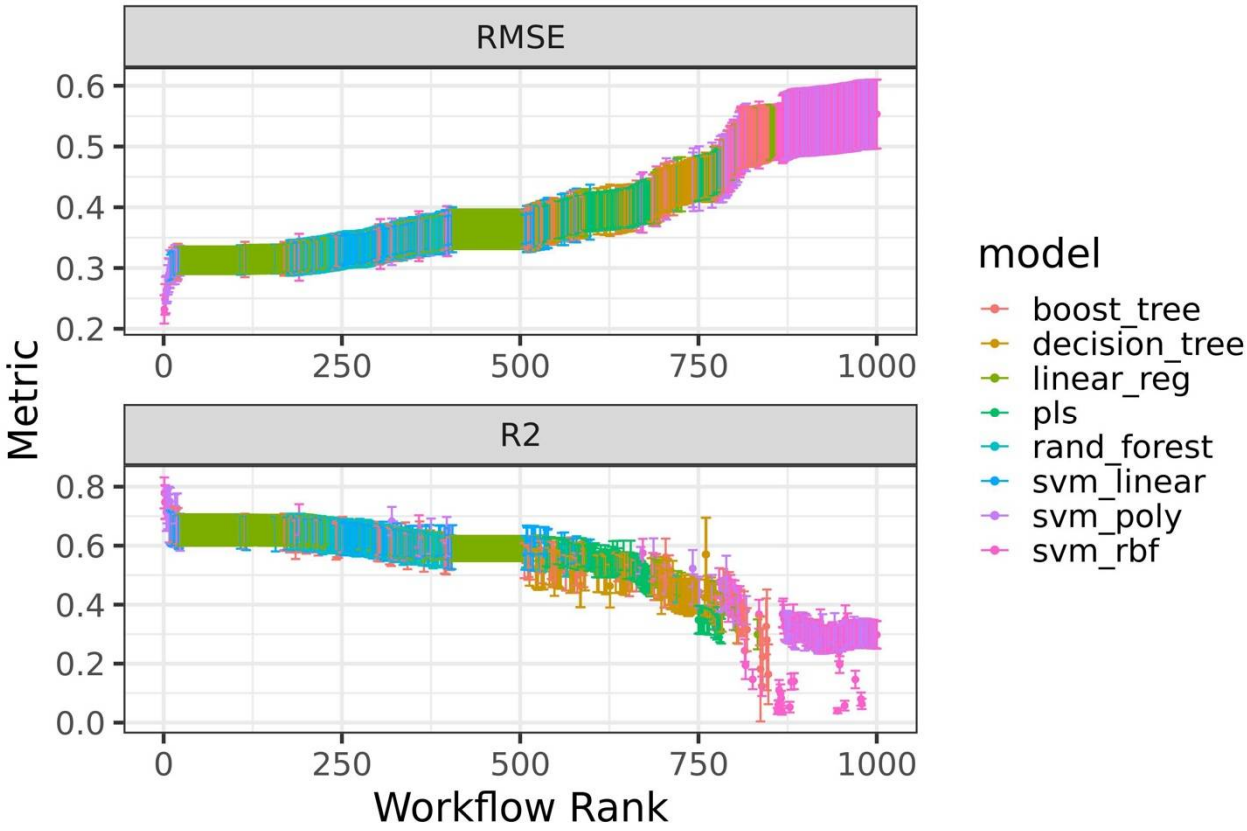

Figure S5. Workflow rank of the machine learning technique used in the CaO analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following; svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg (includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

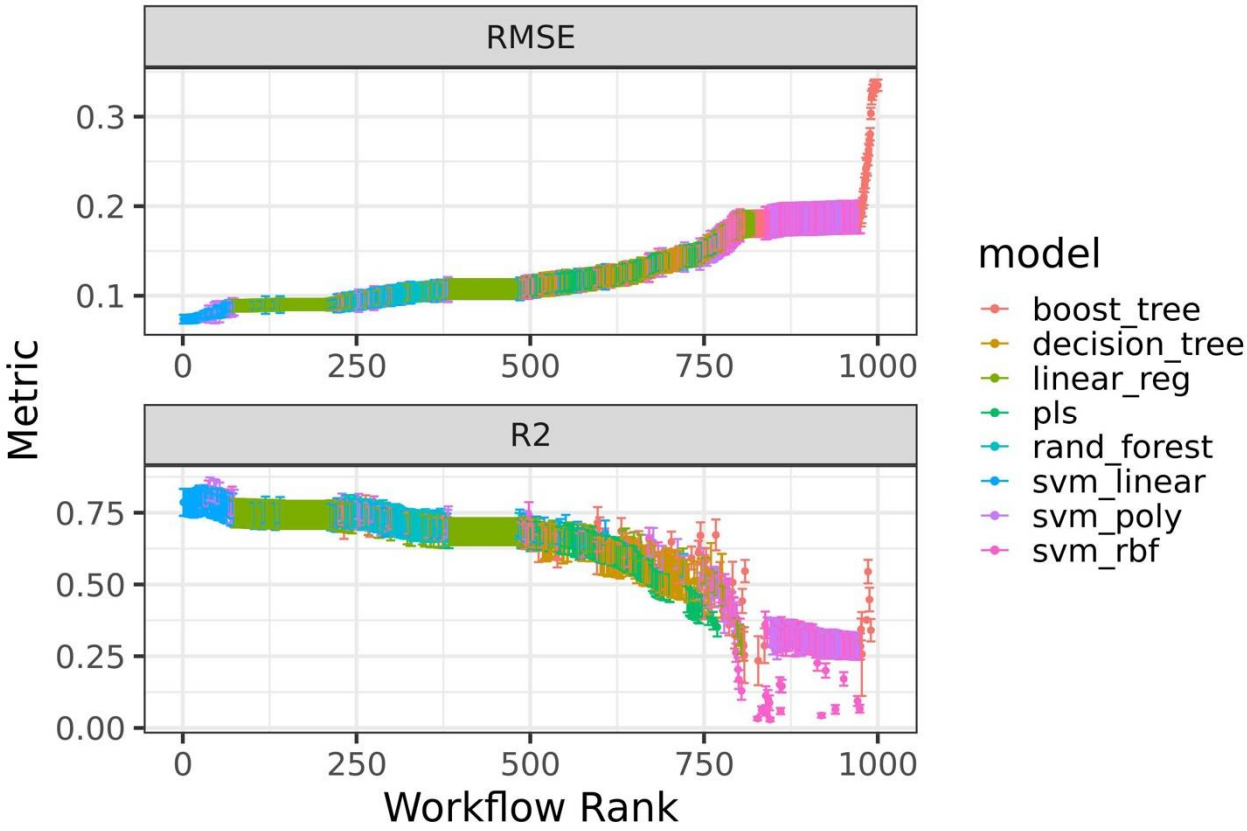

Figure S6. Workflow rank of the machine learning technique used in the MgO analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following; svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg(includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

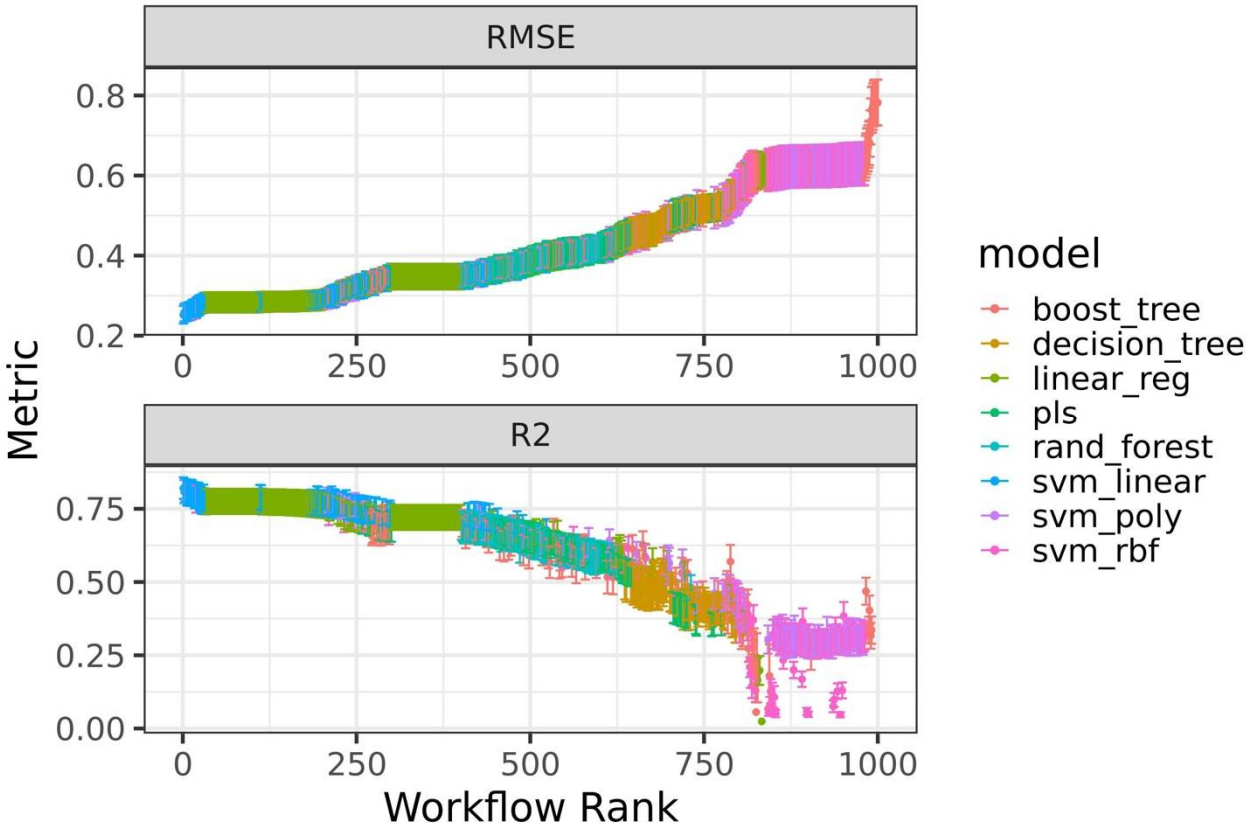

Figure S7. Workflow rank of the machine learning technique used in the K_2O analysis for the stacked regression. The figure shows the model configuration on the horizontal axis with the ranks decreasing from left to right (the value of one being the best) versus the performance metrics (RMSE and R^2) on the vertical axis on the cross-validation sets. The vertical lines for each point represent a 90% confidence bound for each model configuration. The statistical techniques used are the following; svm_linear = support vector regression with linear kernel; svm_poly = support vector regression with polynomial kernel; svm_rbf = support vector regression with radial kernel; linear_reg(includes least absolute shrinkage and selection operator (LASSO), ridge regression and elastic net regression) pls = Partial least squares; rand_forest = Random forests via randomForest; decision_tree = recursive partitioning and regression trees; boost_tree = Boosted trees. The mentioned models with different configurations were used as the Level 1 models in the stack regression workflow.

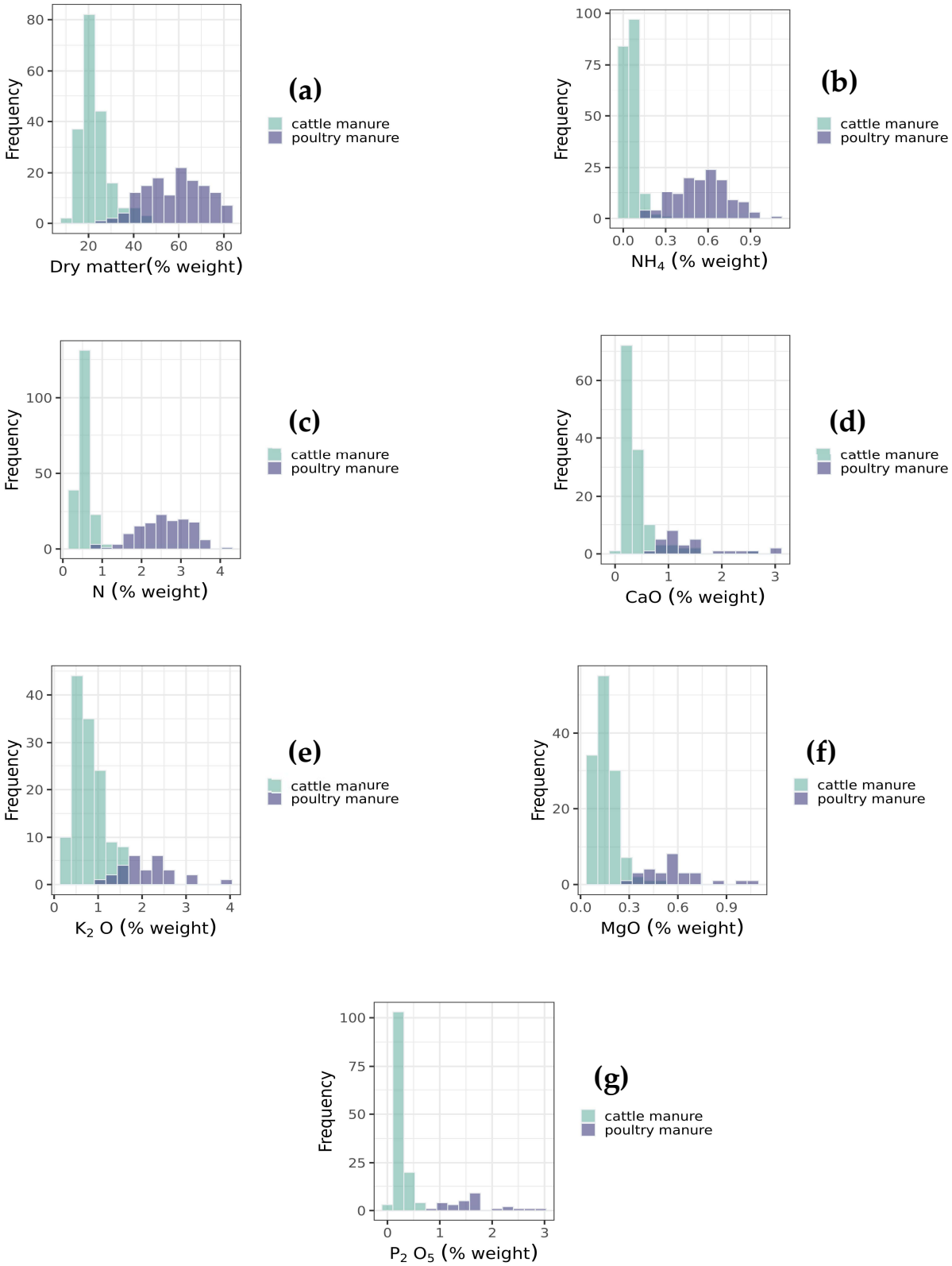

Figure S8. Histograms for the 332 samples of (a) dry matter, (b) total ammonium nitrogen (NH_4), (c) total nitrogen (N), and 158 samples of (d) CaO, (e) K_2O, (f) MgO, (g) P_2O_5 in the entire dataset before splitting.

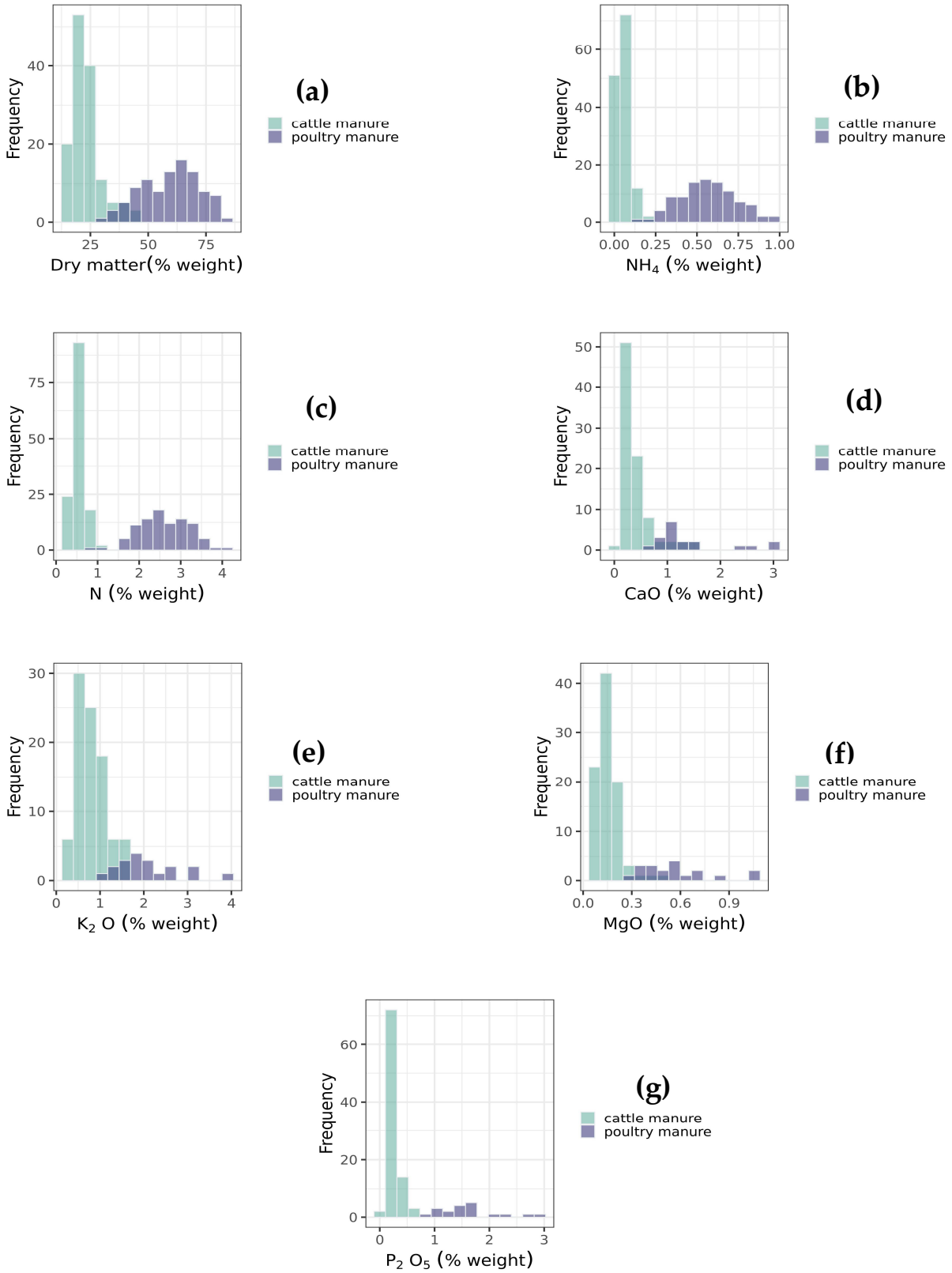

Figure S9. Histograms for the 232 samples of (a) dry matter, (b) total ammonium nitrogen (NH_4), (c) total nitrogen (N), and 110 samples of (d) CaO, (e) K_2O, (f) MgO, (g) P_2O_5 in the training set.

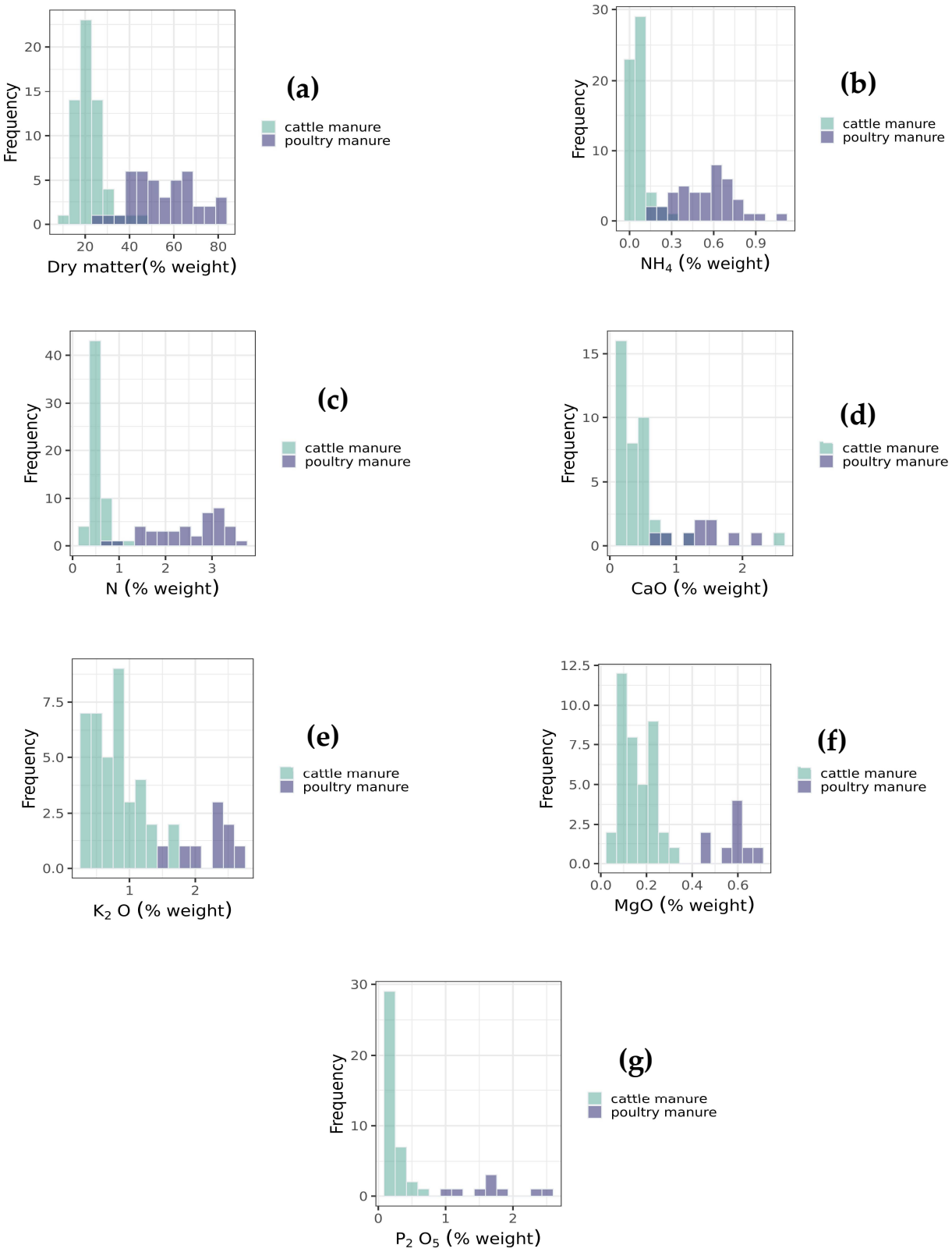

Figure S10. Histograms for the 110 samples of (a) dry matter, (b) total ammonium nitrogen (NH_4), (c) total nitrogen (N), and 48 samples of (d) CaO, (e) K_2O, (f) MgO, (g) P_2O_5 in the testing set.

Supplementary tables

Table S1. Ranges of hyperparameters used in tuning of best results for various machine learning techniques. A space-filling design with a grid number of 100 is used. There are 100 equally spaced values between (including) each hyperparameter's minimum and maximum values that were used for tuning. For hyperparameters that are meaningful only when the values are integers, i.e., the latent variable (LV) in partial least squares (PLS), non-integer values are just skipped during tuning (SVRLin = support vector with linear kernel; SVRPoly = support vector with polynomial kernel; SVRRad = support vector with radial kernel; LASSO = least absolute shrinkage and selection operator; RIDGE = ridge regression; ENET = elastic net; RF= Random forests; RPART = recursive partitioning and regression trees; XGB = Boosted trees).

Hyperparameters, ranges

Algorithm	parameters	minimum	maximum
SVRLin	cost(C)	9.77E-04	32
	margin(epsilon)	0	0.2
SVRPoly	cost(C)	9.77E-04	32
	degree	1	3
	scale_factor(scale)	1.00E-10	0.1
	margin(epsilon)	0	0.2
SVRRad	cost(C)	9.77E-04	32
	rbf_sigma(sigma)	1.00E-10	1
	margin(epsilon)	0	0.2
LASSO	penalty	1.00E-10	1
	mixture (alpha)	1	1
RIDGE	penalty	1.00E-10	1
	mixture (alpha)	0	0
ENET	penalty	1.00E-10	1
	mixture (alpha)	0.05	1
PLS	predictor_prop	0	1
	num_comp(ncomp)	1	4
RF	mtry	1	?*
	trees(ntree)	1	2000
	min_n(min_rows)	1	40
RPART	tree_depth(maxdepth)	1	15
	min_n(min_rows)	2	40
	cost_complexity(cp)	1.00E-10	0.1
XGB	tree_depth(maxdepth)	1	15
	trees(nrounds)	1	2000
	learn_rate(eta)	0.001	0.316
	mtry(colsample_bynode)	1	?*
	min_n(min_child_weight)	2	40
	loss_reduction(gamma)	1.00E-10	31.623
	sample_size(sub_sample)	0.1	1
	stop_iter(early_stop)	3	20

*mtry depends on the number of columns of the predictors and the mode of the model (either regression or classification) which are being computed during the process of tuning.

Table S2. Optimized parameters obtained from different machine learning models (SVRLin = support vector with linear kernel; SVRPoly = support vector with polynomial kernel; SVRRad = support vector with radial kernel; LASSO = least absolute shrinkage and selection operator; RIDGE = ridge regression; ENET = elastic net; PLS = partial least squares; RF= Random forests; RPART = recursive partitioning and regression trees; XGB = Boosted trees).

Algorithm	parameters	DM	NH4	N	P2O5	CaO	MgO	K2O
SVRLin	cost(C)	0.119	0.054	0.216	0.136	0.149	1.064	0.101
	margin(epsilon)	0.062	0.122	0.078	0.122	0.049	0.168	0.015
SVRPoly	cost(C)	0.3249	0.3249	3.3491	3.3491	26.7003	3.3491	3.3491
	degree	2	2	1	1	3	1	1
	scale_factor(scale)	0.0049	0.0049	0.0290	0.0290	0.0015	0.0290	0.0290
	margin(epsilon)	0.0480	0.0480	0.0968	0.0968	0.1323	0.0968	0.0968
SVRRad	cost(C)	21.3698	5.6840	21.3698	21.3698	21.3698	21.3698	21.3698
	rbf_sigma(sigma)	0.00035	0.00112	0.00035	0.00035	0.00035	0.00035	0.00035
	margin(epsilon)	0.14871	0.11354	0.14871	0.14871	0.14871	0.14871	0.14871
LASSO	penalty	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10
	mixture (alpha)	1	1	1	1	1	1	1
RIDGE	penalty	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10	1.133E-10
	mixture (alpha)	0	0	0	0	0	0	0
ENET	penalty	1.312E-10	5.186E-07	2.076E-05	1.312E-10	9.369E-08	1.620E-07	2.076E-05
	mixture (alpha)	0.9616	0.8071	0.9868	0.9616	0.9692	0.4144	0.9868
PLS	predictor_prop	0.8080	0.3335	0.8080	0.9532	0.1933	0.9381	0.2586
	num_comp(ncomp)	7	9	7	9	10	11	9
RF	mtry	90	112	41	112	446	137	137
	trees(ntree)	1799	1741	973	1741	1039	1288	1288
	min_n(min_rows)	3	3	9	3	4	3	3
RPART	tree_depth(maxdepth)	10	11	5	3	4	12	3
	min_n(min_rows)	7	13	5	30	11	29	15
	cost_complexity(cp)	3.12E-05	6.483E-08	5.650E-07	9.964E-05	2.943E-07	1.679E-05	3.906E-04
XGB	tree_depth(maxdepth)	13	10	10	7	10	4	10
	trees(nrounds)	1594	1662	1662	813	1662	1305	1662
	learn_rate(eta)	0.077	0.011	0.011	0.182	0.011	0.042	0.011
	mtry(colsample_bynode)	31	45	45	596	45	235	45
	min_n(min_child_weight)	28	14	14	2	14	8	14
	loss_reduction(gamma)	4.03E-09	5.905E-09	5.905E-09	3.287E-05	5.905E-09	1.401E-07	5.905E-09
	sample_size(sub_sample)	0.7197	0.6952	0.6952	0.4932	0.6952	0.8133	0.6952
	stop_iter(early_stop)	16	4	4	8	4	11	4
Stacked Regression (ENET)	penalty	0.7374	0.0303	0.0202	0.0202	1	0.4848	0.0404
	mixture (alpha)	0.3434	0.8485	1	0.8990	0.0101	0.0101	1

Table S3. The top 10 (or 7) highest weighted (stacking coefficient) members of a stacked ensemble of different models with non-zero coefficients for each of the chemical contents: dry matter (DM), total ammonium nitrogen (NH_4), total nitrogen (N), phosphorus pentoxide (P_2O_5), calcium oxide (CaO), magnesium oxide (MgO), and potassium oxide (K_2O) (SVRLin = support vector with linear kernel; SVRPoly = support vector with polynomial kernel; SVRRad = support vector with radial kernel; LASSO = least absolute shrinkage and selection operator; RIDGE = ridge regression; ENET = elastic net; PLS = partial least squares; RF= Random forests; RPART = recursive partitioning and regression trees; XGB = Boosted trees).

Chemicals	Algorithm	Model members	Stacking coefficients or weight	Description
DM	XGB	XGB_1_108	2.926	The 10 highest weighted members out of 23
	SVRPoly	SVRPoly_1_066	0.110	
	SVRPoly	SVRPoly_1_065	0.108	
	SVRPoly	SVRPoly_1_020	0.084	
	SVRPoly	SVRPoly_1_083	0.079	
	SVRRad	SVRRad_1_064	0.072	
	SVRLin	SVRLin_1_094	0.069	
	SVRLin	SVRLin_1_035	0.056	
	SVRRad	SVRRad_1_100	0.056	
	SVRLin	SVRLin_1_050	0.054	
NH_4	SVRPoly	SVRPoly_1_065	0.158	The 10 highest weighted members out of 30
	SVRRad	SVRRad_1_086	0.092	
	SVRPoly	SVRPoly_1_020	0.091	
	XGB	XGB_1_051	0.078	
	SVRLin	SVRLin_1_087	0.073	
	SVRLin	SVRLin_1_046	0.067	
	SVRRad	SVRRad_1_037	0.062	
	SVRRad	SVRRad_1_100	0.054	
	SVRLin	SVRLin_1_057	0.051	
	SVRRad	SVRRad_1_064	0.033	
N	SVRPoly	SVRPoly_1_066	0.274	The 7 highest weighted members out of 7
	SVRRad	SVRRad_1_087	0.235	
	SVRLin	SVRLin_1_060	0.184	
	SVRRad	SVRRad_1_037	0.179	
	SVRLin	SVRLin_1_054	0.116	
	SVRPoly	SVRPoly_1_083	0.006	
	XGB	XGB_1_051	0.005	
P_2O_5	SVRRad	SVRRad_1_100	0.218	The 10 highest weighted members out of 11
	XGB	XGB_1_087	0.195	
	PLS	PLS_1_086	0.193	
	SVRRad	SVRRad_1_064	0.179	
	SVRPoly	SVRPoly_1_083	0.153	
	PLS	PLS_1_054	0.121	
	XGB	XGB_1_066	0.051	
	PLS	PLS_1_098	0.030	
	XGB	XGB_1_067	0.014	
	XGB	XGB_1_021	0.012	
CaO	SVRRad	SVRRad_1_100	0.039	The 10 highest weighted members out of 152
	SVRRad	SVRRad_1_087	0.038	
	SVRRad	SVRRad_1_086	0.036	
	SVRRad	SVRRad_1_037	0.034	
	SVRPoly	SVRPoly_1_028	0.034	
	SVRPoly	SVRPoly_1_036	0.031	

	SVRPoly	SVRPoly_1_066	0.030	
	SVRPoly	SVRPoly_1_020	0.030	
	SVRPoly	SVRPoly_1_083	0.030	
	SVRPoly	SVRPoly_1_065	0.029	
MgO	XGB	XGB_1_036	6.000	The 10 highest weighted members out of 249
	XGB	XGB_1_006	0.038	
	XGB	XGB_1_087	0.017	
	SVRPoly	SVRPoly_1_020	0.014	
	SVRPoly	SVRPoly_1_028	0.012	
	XGB	XGB_1_022	0.011	
	SVRPoly	SVRPoly_1_083	0.010	
	SVRPoly	SVRPoly_1_010	0.009	
	PLS	PLS_1_023	0.009	
	PLS	PLS_1_083	0.009	
K_2O	SVRLin	SVRLin_1_074	0.568	The 7 highest weighted members out of 7
	XGB	XGB_1_022	0.194	
	XGB	XGB_1_079	0.169	
	SVRRad	SVRRad_1_100	0.081	
	SVRPoly	SVRPoly_1_040	0.021	
	RPART	RPART_1_059	0.018	
	SVRRad	SVRRad_1_064	0.012	

Table S4. Statistical significance table that compares the ratio of the standard errors between two algorithms with that of the critical F-value in the training set ($F_{critical} = 1.242$ at 231 degrees of freedom for DM, NH_4 and N; $F_{critical} = 1.372$ at 109 degree of freedom for P_2O_5, CaO, MgO and K_2O). In calculating the ratio, the best performing algorithm for each of the chemical components was used as the denominator. If the ratio is less than the critical F-value then the two RMSE values are not significantly different.

Algorithm	**DM**		**NH4**		**N**		**P2O5**		**CaO**		**MgO**		**K2O**	
	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?
SVRLin	1.445	No	1.351	No	1.311	No	1.270	Yes	1.669	No	1.000	Yes	1.000	Yes
SVRPoly	1.000	Yes	1.112	Yes	1.356	No	1.384	No	1.295	Yes	1.104	Yes	1.050	Yes
SVRRad	1.050	Yes	1.000	Yes	1.000	Yes	1.000	Yes	1.000	Yes	1.159	Yes	1.138	Yes
LASSO	1.704	No	1.744	No	1.542	No	1.532	No	1.820	No	1.495	No	1.262	Yes
RIDGE	5.030	No	3.765	No	4.456	No	2.696	No	2.457	No	2.127	No	6.131	No
ENET	1.702	No	1.735	No	1.542	No	1.531	No	1.815	No	1.450	No	1.265	Yes
PLS	2.232	No	1.957	No	2.319	No	2.119	No	2.299	No	2.075	No	1.373	No
RF	2.293	No	1.930	No	2.240	No	2.630	No	1.864	No	1.592	No	1.904	No
RPART	4.224	No	3.625	No	4.514	No	4.506	No	2.476	No	2.326	No	3.132	No
XGB	1.565	No	1.544	No	1.854	No	1.901	No	1.710	No	1.549	No	1.670	No

Table S5. Statistical significance table that compares the ratio of the standard errors between two algorithms with that of the critical F-value in the testing set ($F_{critical} = 1.394$ at 99 degrees of freedom for DM, NH_4 and N; $F_{critical} = 1.624$ at 47 degrees of freedom for P_2O_5, CaO, MgO and K_2O). In calculating the ratio, the best performing algorithm for each of the chemical components was used as the denominator. If the ratio is less than the critical F-value then the two RMSE values are not significantly different.

Algorithm	DM		NH4		N		P2O5		CaO		MgO		K2O	
	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?	ratio	ratio < crit. val?
SVRLin	2.857	No	1.843	No	2.499	No	1.294	Yes	1.081	Yes	1.528	Yes	1.140	Yes
SVRPoly	1.592	No	1.991	No	2.542	No	1.303	Yes	1.753	No	1.362	Yes	1.138	Yes
SVRRad	1.499	No	2.703	No	1.351	Yes	1.047	Yes	1.457	Yes	1.000	Yes	1.005	Yes
LASSO	3.063	No	2.220	No	2.875	No	1.387	Yes	1.170	Yes	1.366	Yes	1.216	Yes
RIDGE	5.183	No	3.422	No	5.415	No	2.098	No	1.625	No	1.899	No	1.597	Yes
ENET	3.019	No	2.231	No	2.900	No	1.388	Yes	1.173	Yes	1.330	Yes	1.218	Yes
PLS	4.474	No	3.099	No	4.270	No	1.413	Yes	1.271	Yes	1.400	Yes	1.394	Yes
RF	2.145	No	2.682	No	2.439	No	2.783	No	1.512	Yes	2.407	No	1.591	Yes
RPART	7.619	No	5.544	No	6.793	No	3.546	No	1.488	Yes	3.443	No	1.559	Yes
XGB	1.905	No	2.220	No	1.648	No	2.898	No	1.729	No	2.898	No	1.229	Yes
Stack Reg	1.000	Yes	1.000	Yes	1.000	Yes	1.000	Yes	1.000	Yes	1.399	Yes	1.000	Yes