Journal of Universal Computer Science, vol. 31, no. 4 (2025), 363-382
submitted: 27/2/2024, accepted: 27/11/2024, appeared: 28/3/2025 CC BY 4.0

EBAR: A Novel Machine Learning Model for Quantifying Chemical Concentrations using NIR Spectroscopy

Phan Minh Nhat
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0009-0008-3861-9892, nhat0299@gmail.com)

Ngo Le Huy Hien
(Leeds Beckett University, Leeds, England
https://orcid.org/0000-0002-1439-8289, ngolehuyhien@gmail.com)

Dinh Minh Toan
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0009-0000-3405-4638, toandinh6501@gmail.com)

Le Viet Hung
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0009-0007-5164-4761, hungle.dut@gmail.com)

Phan Binh
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0009-0002-8278-4218, binhphan77373@gmail.com)

Phung Thi Anh
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0009-0000-6018-2684, anhhna17710110@gmail.com)

Nguyen Thi Hoang Phuong
(Pham Van Dong University, QuangNgai, Viet Nam
https://orcid.org/0009-0009-6755-604X, nthphuong@pdu.edu.vn)

Nguyen Van Hieu
(The University of Danang - University of Science and Technology, Danang, Vietnam,
https://orcid.org/0000-0001-5311-806X, nvhieuqt@dut.udn.vn)

Abstract: The examination of Near Infrared Reflectance Spectroscopy (NIR) in cattle and poultry fertilizers provides a viable solution for determining optimal fertilizer composition for crop growth while mitigating adverse impacts on soil and groundwater quality. In recent studies, conventional machine learning models combined with spectral analysis have been used to ascertain cattle and poultry fertilizer concentrations. However, these traditional machine learning models encounter challenges in achieving data generalization, resulting in suboptimal prediction accuracy. To address this issue, this study proposes a synthesized machine learning model named EBAR (Error Based Accumulation Regression), which exhibits a commendable coefficient of determination, with an average $R^2 = 0.865$ across 7 chemical substances, surpassing the performance of existing traditional machine learning models. Additionally, a Backward Elimination technique is designed to identify crucial wavelength ranges for monitoring component concentrations. The research

outcome is promising and acts as a novel benchmark for later models in determining component concentrations through NIR spectroscopy. Future research gears toward expanding datasets and increasing samples of fertilizers, extending examined wavelength, and improving the model's efficiency to apply to various types of foods, including seafood, vegetables, and fruits.

Keywords: Ensemble Learning, Error Based Accumulation Regression, Backward Elimination, Machine Learning, Cattle Manure, Poultry Manure, Near Infrared Reflectance (NIR) spectroscopy.
Categories: I.0, I.2, I.3, I.4, I.5, J.3
DOI: 10.3897/jucs.121757

1 Introduction

The utilization of animal-derived fertilizers is one of the conventional agricultural practices for bolstering soil fertility and facilitating the proliferation of plant organisms by [Rahman et al. 2016]. Animal-based fertilizers contain a plethora of essential components, including nitrogen (N), phosphorus (P), and potassium (K). A majority of these components exist in oxide forms; for instance, P is typically found in the manifestation of phosphorus pentoxide (P_2O_5), K in the guise of potassium oxide (K_2O), and calcium (Ca) in calcium oxide (CaO).

Analyzing the concentration of animal fertilizer plays a vital role in ensuring the judicious dispensation of essential nutrients for plant growth, preventing excessive fertilization that can harm the environment and water sources according to [Jiaying et al. 2022, MacDonald 2010]. In the US, nearly 1.4 billion tons of fertilizer are produced from over 9.8 billion cattle animals every year according to [Pagliari et al. 2020], which underscores the imperative of proper fertilizer management procedures. However, to establish such procedures, the system ought to be based on a rigorous chemical analysis of fertilizer specimens as determined by [Kacprzak et al. 2023].

Currently, research laboratories frequently perform chemical composition analysis of specimens using the Kjeldahl technique [Fuquay et al. 2011], conforming to established benchmarks. Since animal fertilizers are heterogeneous samples, and given the volatile nature of certain chemical constituents, the chemical analysis of these samples requires fast, cost-effective, and highly accurate methodologies. The Kjeldahl method, however, is expensive, time-consuming, and short in meeting these requisites. In light of these challenges, the deployment of Near-Infrared Reflectance Spectroscopy (NIR) has emerged as a highly efficacious alternative as claimed by [Feng et al. 2022]. NIR spectroscopy is widely acclaimed for its reliability in qualitative and quantitative analysis in various industries, including agriculture, food, and petroleum. Notably, NIR spectroscopy distinguishes itself by virtue of its non-destructive nature upon the sample and has become increasingly popular due to recent advancements in mathematics and computing. Nevertheless, the extraction and interpretation of data from NIR spectra remain a formidable challenge as stated by [Roggo et al. 2007].

To address this research gap, machine learning approaches are commonly invoked, including Principal Component Analysis (PCA) for qualitative analysis of NIR data and the application of Partial Least Squares (PLS) regression models to predict the quantitative attributes of interest in the specimens [Kumaravelu et al. 2015]. Moreover, regression models such as Least Absolute Shrinkage and Selection Operator (LASSO), RIDGE, and Random Forest, have been used in NIR data analysis in recent studies [Horf et al. 2022]. Alongside machine learning models, data preprocessing techniques are also commonly applied, such as multiplicative scatter correction and Savitzky-Golay

smoothing [Feng et al. 2022]. Furthermore, some other techniques can be utilized in this domain, such as standard normal variable (SNV) and de-trending (DT) as claimed by [Huang et al. 2008, Devianti et al. 2022, Devianti et al. 2021].

This research has two primary novel contributions:

1. It proposes a synthetic machine learning model architecture, combining traditional machine learning models to determine the chemical composition of animal-derived fertilizer specimens. The examined chemicals include dry matter (DM), NH_4, N, P_2O_5, CaO, MgO, and K_2O.
2. It introduces a novel technique for determining strong absorption wavelength ranges for corresponding chemicals. If the concentration of a specific chemical changes, the absorption at these wavelengths changes the most across the entire wavelength range. This method identifies the correlation coefficient between the absorption at each wavelength and the content of the chemical in the specimen, thereby eliminating wavelength ranges with increasing correlation to pinpoint the desired absorption wavelength ranges.

The proposed model yields promising outcomes, attaining an average root mean square error (RMSE) of 0.726 and a coefficient of determination (R^2) of 0.865 for the 7 examined chemicals, thereby outperforming current traditional machine learning models. The Backward Elimination method also effectively identifies the maximum absorption wavelength ranges for 5 chemicals. Hence, this model will serve as an innovative benchmark for future models in the assessment of component concentrations using NIR spectroscopy.

2 Related Work

2.1 Application of Machine Learning Models in NIR Spectrum-Related Problems

Recent studies in the field of NIR spectroscopy encompass a variety of chemical analysis techniques, including but not limited to quality component analysis, chemical analysis, qualitative and classification analysis, and waveshape analysis [Huy et al. 2023]. Alongside these conventional approaches, the application of machine learning models in NIR spectral data analysis is becoming increasingly popular due to their reliability and effectiveness. [Cozzolino et al. 2007] have combined Principal Component Analysis (PCA) and Partial Least Squares (PLS) models to investigate the influence of temperature on the NIR spectra of wine, as well as the predictive proficiency of the model in quantifying various chemical constituents within wine. The research demonstrated the significance of temperature effects and the model's ability to predict alcohol concentration, sugar content, pH level, and tartaric acid (TA) concentration in white and red wine samples. However, the model proposed in the study did not attain an optimal level of refinement in the characterization of wine's chemical parameters according to [Zhang et al. 2022].

Another study by [Özdemir et al. 2018] implemented PLS models in conjunction with chemometrics to rapidly analyze various synthesized sterols found in olive oil. Nonetheless, the limitation of this study was its suboptimal performance in the case of individual fatty acids according to [Zhang et al. 2022].

In research by [Jiang and Chen 2019], a refined model was proposed to directly quantify doping levels in olive oil through its spectrum. This study harnessed the bootstrapping soft shrinkage (BOSS) algorithm to select relevant spectral features and applied

the PLS model with these chosen features. Ultimately, the performance of the BOSS-PLS model was compared with that of three other models: CARS-PLS, MCUVE-PLS, and IRIV-PLS. The model proposed by this research demonstrated its superiority in solving the problem, and the BOSS technique showed promise as a potential method for selecting critical spectral features. However, the limitation of this study was the limited diversity of the data samples as stated by [Zhang et al. 2022]. In another finding of [Olivos-Trujillo et al. 2015], various regression models were used to determine the oil content and quality characteristics of different seeds. However, it is imperative to recognize that this research did not address the issue of selecting relevant spectral features for the predictive model as claimed by [Zhang et al. 2022].

2.2 Techniques for Eliminating Non-essential Wavelengths in NIR Spectroscopy

The determination of the wavelength range of maximum absorbance for a substance provides a clearer understanding of how that substance interacts with light or electromagnetic waves as determined by [Pavia et al. 2015]. Essentially, this process entails selecting relevant spectral features, serving as a pivotal technique to reduce the dimensionality of input data. Consequently, it simplifies the model and contributes to mitigated computational costs, accelerated training, and enhanced model efficiency [Chandrashekar and Sahin 2014].

In an investigation by [Yu et al. 2023], a combined technique was proposed to enhance the model's interpretability and predictive capability by directly identifying the features of diesel fuel. Empirical results indicated that their proposed method offered better interpretability and reliability compared to traditional machine learning models like PLS and Elastic Net according to [Hien et al. 2021].

In a study conducted by [Mamouei et al. 2020], genetic algorithms were used to eliminate non-essential wavelengths when inputting data into the prediction model. Unlike other methods, the genetic algorithm exclusively used NIR spectral data from the samples without the need for chemical analysis data. However, this technique only eliminated a minimal number of non-essential wavelengths and did not achieve sufficiently high convergence. This limitation remains a salient aspect of the study.

2.3 Traditional Machine Learning Models

Traditional machine learning methods have been widely utilized in various domains for predictive modelling and data analysis. Among these, Support Vector Regression (SVR) is a powerful technique that uses the principles of Support Vector Machines to perform regression analysis, particularly effective in high-dimensional spaces [Almaspoor et al. 2021]. LASSO (Least Absolute Shrinkage and Selection Operator) and RIDGE regression are popular linear models that apply regularization to prevent overfitting, with LASSO focusing on feature selection by enforcing sparsity [Tibshirani et al. 1996] and RIDGE emphasizing coefficient shrinkage [Wieringen 2015]. Elastic Net combines the strengths of both LASSO and RIDGE, providing a balanced approach to feature selection and regularization [Zou and Hastie 2005]. Beyond linear models, ensemble methods like Random Forest [Hien et al. 2024], XGBoost, and LightGBM have gained prominence for their ability to handle complex, non-linear relationships in data. Random Forest aggregates multiple decision trees to improve predictive accuracy and robustness, while XGBoost and LightGBM are gradient boosting techniques known for their efficiency, scalability, and superior performance in handling large datasets and high-dimensional

feature spaces. These methods have proven to be essential tools in the arsenal of machine learning practitioners, offering diverse approaches to model complex phenomena and enhance predictive capabilities.

Despite substantial successes in data exploration, traditional machine learning methods encounter challenges in handling intricate data, such as imbalanced data, data noise, and high-dimensional data. This stems from the inherent limitations of traditional models in capturing data characteristics and underlying structures. As a resolution to these issues, a study by [Dong et al. 2020] introduced a strategy involving the construction of ensemble learning models, integrating data combinations, modelling techniques, and data mining into a unified framework. Ensemble Learning amalgamates valuable insights from the results of traditional machine learning models, thus leading to improved data exploration and enhanced performance.

In spite of the high performance achieved by Support Vector Machine (SVM) models, which offer support for both non-linear and linear data, these models encounter difficulties when confronted with large datasets and are sensitive to noise [Hien 2022]. On the other hand, models such as LASSO, RIDGE, and Elastic Net (ENET) are straightforward and efficient, yet they operate effectively only with linear data. Partial Least Squares (PLS) has the advantage of handling high-dimensional data and correlated features, but it is prone to overfitting and is challenging to interpret. Random Forest, XGBoost, and LightGBM are ensemble models capable of handling heterogeneous data, minimizing overfitting, automatically selecting features, and processing high-dimensional data. Therefore, the use of Ensemble Learning techniques to combine these models creates a composite model that achieves high performance, excels with both linear and non-linear data, handles high-dimensional data and correlated features, and overcomes the limitations of individual models. Because the concentration of chemicals must be determined with a high degree of precision through NIR spectroscopy, traditional models cannot be used independently; instead, they should be integrated.

3 Materials and Methods

3.1 Error Based Accumulation Regression (EBAR)

Inspired by real-life scenarios involving learning from inadequacies and limitations, this study proposes a novel technique that leverages prediction results exhibiting substantial errors derived from traditional machine learning models to create input features for the final model. This proposed model, named ‘EBAR’, is synthesized from traditional machine learning models to enhance its predictive capabilities.

The proposed model (depicted in Figure 1) aims to augment the performance of 10 traditional techniques, including Support Vector Regression with Linear Kernel, Support Vector Regression with Polynomial Kernel, Support Vector Regression with Radial Basis Function Kernel, LASSO Regression, RIDGE Regression, Elastic Net Regression, Partial Least Squares Regression, Random Forest Regression, XGBoost Regression, and Light Gradient Boosting Machine Regression.

The ’EBAR’ model consists of two fundamental processes: training and prediction. In the training process, a meta-model is chosen to amalgamate results from 9 base models, employing out-of-folds predictions from the base models to construct features for training the meta-model.

The training process is described as follows:

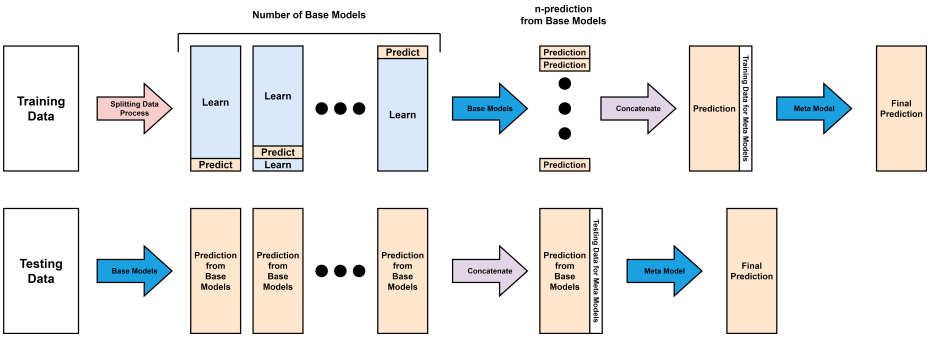

Figure 1: Error Based Accumulation Regression Overall Architecture

1. Utilize k-folds cross-validation to partition the training dataset into k non-overlapping subsets, with each subset:
   (a) Serving as the holdout set.
   (b) The remaining data is used for training.
   This data splitting process is depicted in Figure 2.
2. Train the base models on each training subset and evaluate their performance on the corresponding holdout subset.
3. Use the predictions from the evaluation (out-of-folds predictions) as input, and the target values as output, to train the higher-order model (meta-model).

The prediction process is carried out as follows:

1. Collect predictions from all base models on the testing data and use them as meta-features.
2. Employ these meta-features with the meta-model to make the final predictions.

The training and prediction processes are described in Algorithm 1 and Algorithm 2, and respectively illustrated in Figures 3 and 4.

3.2 Elimination of Unimportant Wavelengths (Backward Elimination)

The absorbance data of the samples at various wavelengths are discretized into absorbance values at specific wavelengths denoted as $x_1, x_2, \ldots, x_n$, which serve as the features in the feature selection problem. Then, the correlation coefficient between each feature and the target variable is computed. Specifically, for $\mathbf{X}_i$ representing the absorbance of samples at wavelength *i*, $\mathbf{Y}$ representing the quantity of a substance in the sample, and E denotes the expected value, the correlation coefficient is calculated using the formula (1):

$$r_i = \frac{E[\mathbf{X}_i\mathbf{Y}] - E[\mathbf{X}_i]E[\mathbf{Y}]}{\sqrt{E[\mathbf{X}_i^2] - (E[\mathbf{X}_i])^2}\sqrt{E[\mathbf{Y}^2] - (E[\mathbf{Y}])^2}} \tag{1}$$

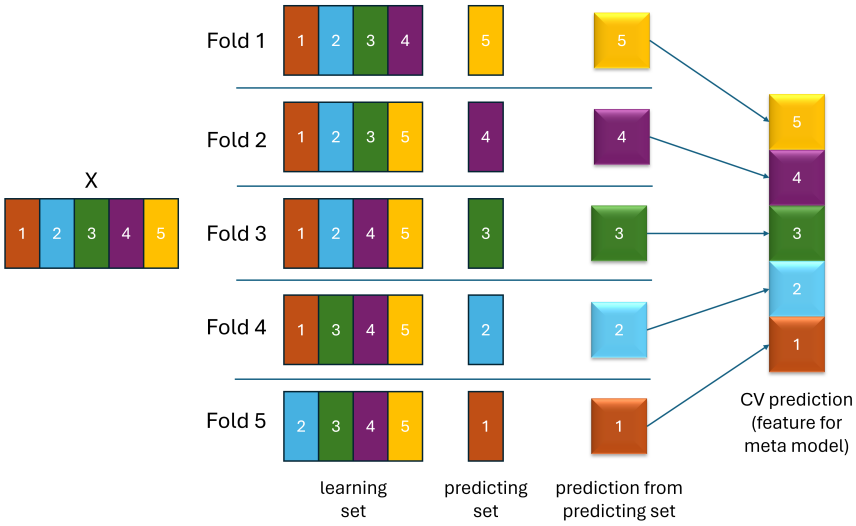

Figure 2: Splitting Data Process with $k = 5$

Algorithm 1 Training Strategy

Input:
X, y: data used for training
$Bases$: list of base models in EBAR model
$Meta$: meta model in EBAR model
Parameters: $n_{folds} \leftarrow$ number of folds
Output: *Parameters of models*

Let $features = []$
$KFold \leftarrow$ k-folds cross validation
for $model$ in $Bases$ **do**
  for $idx_{train}, idx_{holdout}$ in $KFold(X, y)$ **do**
    $model.fit(X[idx_{train}], y[idx_{train}])$
    $pred \leftarrow model.predict(X[idx_{holdout}])$
    $features.append(pred)$
  end for
end for
$Meta.fit(features, y)$

Subsequently, the wavelengths are ranked in descending order based on their absolute correlation coefficient values. Then, wavelengths with low correlation are systematically removed, and following each removal, the EBAR model in combination with cross-validation is applied to assess the impact of the elimination. The evaluation process unfolds as follows:

1. Determine the Root Mean Square Error value $Error_0$ of the EBAR model on the dataset without feature elimination.

2. Calculate the Root Mean Square Error value *Error* of the EBAR model on the dataset with feature elimination.

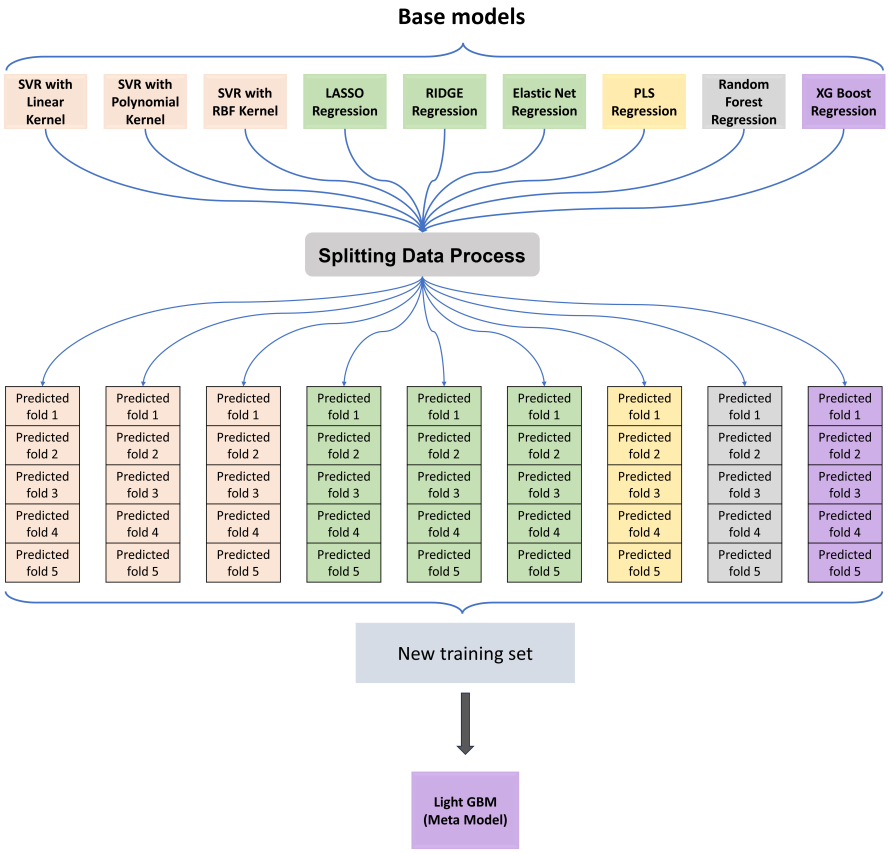

Figure 3: Training Strategy

3. With std_y as the standard deviation of the target variable, calculate the significance level of feature elimination for the model using the formula:

$$p = \frac{|Error - Error_0|}{std_y}$$

4. Compare this significance level with the threshold α:
   (a) If $p < \alpha$, continue removing the next feature;
   (b) If $p \geq \alpha$, stop the process and retain the last removed feature.

The remaining set of features consists of the wavelengths with the strongest absorbance characteristics. This process is depicted through Algorithm 3.

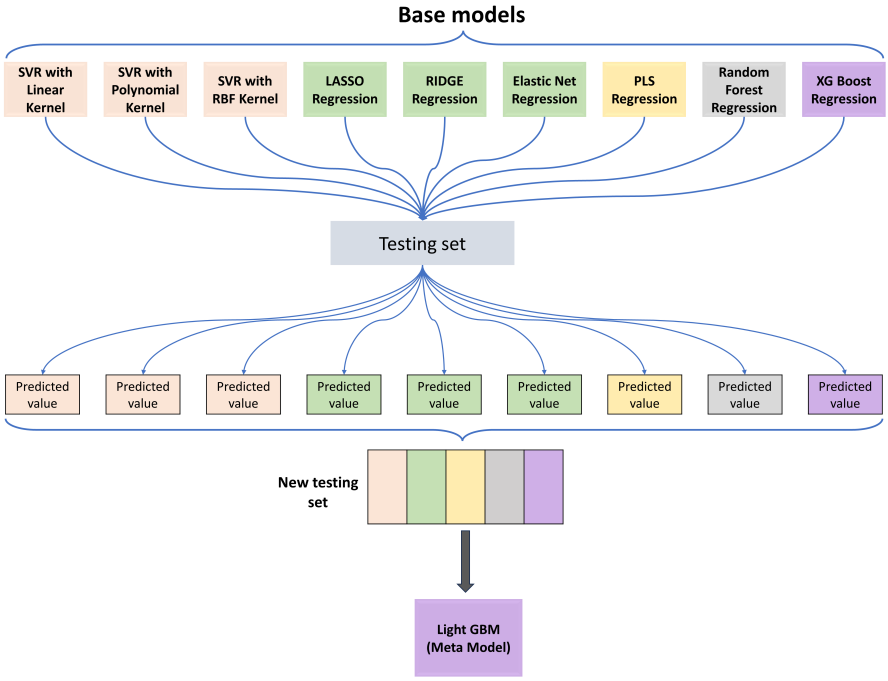

Figure 4: Testing Strategy

Algorithm 2 Testing Strategy

Input:
X, y: data used for testing
$Bases$: list of base models in EBAR model
$Meta$: meta model in EBAR model
Output: *Predictions*

Let $features = []$
for $model$ in $Bases$ **do**
    $pred \leftarrow model.predict(X)$
    $features.append(pred)$
end for
$Predictions \leftarrow Meta.predict(features)$

4 Implementation and Results

4.1 Dataset

The dataset used in this study was collected by Gogé et al. [Gogé et al. 2021], consisting of 196 samples of cattle manure and 136 samples of poultry manure, totaling 332 samples from farming regions in France and Réunion Island. The initial manure samples underwent immediate cryopreservation upon collection to minimize microbial activity affecting NH_4 content. Subsequently, these samples were ground using a Blixer Dito K45 grinder. The samples were then dried at 40°C for four days and ground to a 1 mm particle size using a Cyclotec 1093 mill.

Algorithm 3 Backward Elimination

Input:
X: feature data of samples
Y: target data of samples
$n_{features}$: number of features
$S = \{x_1, x_2, ..., x_{n_{features}}\}$: set of features
$model$: model used to eliminate feature
Parameters: $\alpha \leftarrow$ threshold to eliminate feature
Output: *Set of significant features*

Let r = dictionary()
Let $E_0 = []$
Let $E = []$
KFold $\leftarrow$ *k-folds cross validation*
for $i = 0$ to $n_{features}$ **do**
    $r[x_i] \leftarrow \rho_{X_i,Y}$
end for
for idx_{train}, idx_{test} in $KFold(X, y)$ **do**
    $model.fit(X[idx_{train}], y[idx_{train}])$
    $Y_{pred} \leftarrow model.predict(X[idx_{test}])$
    $E_0.append(RMSE(Y_{pred}, Y))$
end for
$Error_0 \leftarrow mean(E_0)$
$r \leftarrow r.sort(ascending)$
for x in $r.keys()$ **do**
    $S \leftarrow S - x$
    Delete data column of X corresponding to feature x
    for idx_{train}, idx_{test} in $KFold(X, y)$ **do**
        $model.fit(X[idx_{train}], y[idx_{train}])$
        $Y_{pred} \leftarrow model.predict(X[idx_{test}])$
        $E.append(RMSE(Y_{pred}, Y))$
    end for
    $Error \leftarrow mean(E)$
    $p \leftarrow |Error - Error_0|/std_y$
    if $p >= \alpha$ **then**
        $S \leftarrow S + x$
        break
    end if
end for

Chemical analysis was undertaken on the manure samples, encompassing dry matter (DM), ammonium nitrogen (NH_4), nitrogen (N), phosphorus pentoxide (P_2O_5), calcium oxide (CaO), magnesium oxide (MgO), and potassium oxide (K_2O. The chemical analysis was performed using an Inductively Coupled Plasma machine (Element XR Thermo Scientific). To reduce research costs, the researchers conducted exhaustive chemical analysis only on DM, N, and NH_4 in all samples, while for the other compounds, they conducted analysis on 158 samples. A detailed exposition of the statistical characteristics of the compound concentrations is provided in Table 1.

During the near-infrared (NIR) spectroscopy analysis, the specimens underwent scanning using one of three spectrophotometers: XDS-CIRAD, XDS-ARVALIS, and NIRFlex-LDAR, spanning a spectral range from 1100 to 2500 nm. Each measurement represented the average of 32 scans and was recorded as absorbance (Abs) using the formula: $Abs = -log(Reflectance)$. NIR spectra for the 196 cattle manure samples and 136 poultry manure samples are presented in Figure 5. Furthermore, Figure 6 illustrates

Compound	n	Mean	SD	Min	Max	Range
DM	332	37.28	20.26	11.25	82.48	71.22
NH_4	332	0.262	0.276	0.001	1.086	1.085
N	332	1.36	1.09	0.25	4.15	3.90
CaO	158	0.57	0.56	0.09	3.11	3.01
K_2O	158	1.02	0.65	0.19	3.84	3.66
MgO	158	0.23	0.19	0.06	1.05	0.99
P_2O_5	158	0.48	0.58	0.09	3.02	2.93

Table 1: Statistical Analysis of Compound Concentrations in Samples (% of mass)

the distribution of chemical compound data.

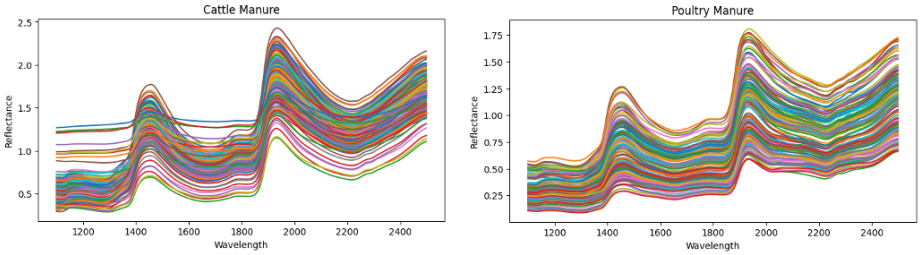

Figure 5: NIR Spectra of Cattle Manure

4.2 Data Preprocessing

The dataset was standardized to achieve a mean of 0 and a standard deviation of 1. NIR spectra of the cattle and poultry manure samples are delineated in Figure 7, while Figure 8 presents the distribution of chemical compound data after standardization.

4.3 Implementation Details

Hardware: The machine learning models were trained on a system with a CPU (Intel Xeon @2.30 GHz), 13 GB RAM, a Tesla K80 accelerator with 12GB GDDR5 VRAM, and cloud-based TPUs with a computational power of 180 teraflops.

Training Strategy: The training process was orchestrated employing 9 base learner models, including SVR (Linear kernel), SVR (Polynomial kernel), SVR (Radial Basis Function kernel), RIDGE Regression, LASSO Regression, ENET Regression, Random Forest Regression, PLS Regression, and XGBoost Regression. The LightGBM Regression model was used as the meta-model for the EBAR model.

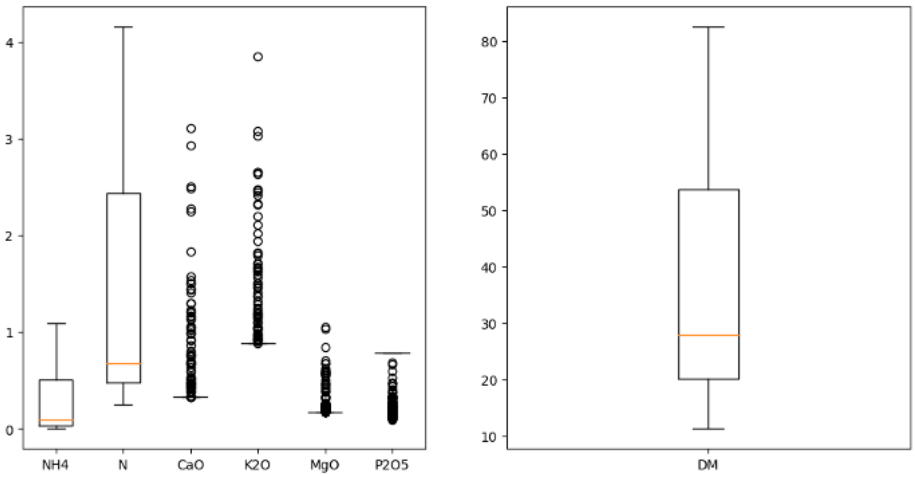

Figure 6: Distribution of Chemical Compound Data

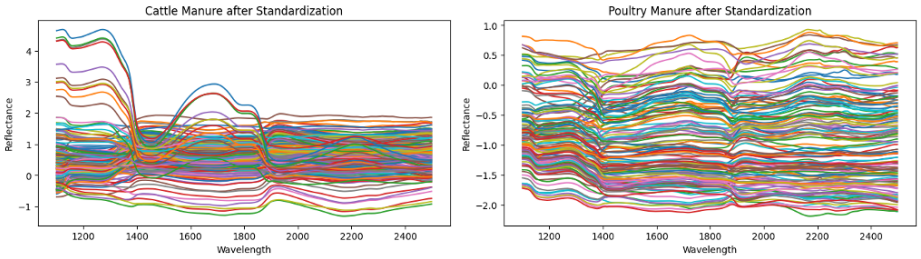

Figure 7: Standardized NIR Spectra of Cattle Manure

Hyperparameter tuning was performed to determine optimal hyperparameter sets for each model using the Gridsearch method [Liashchynskyi and Liashchynskyi 2019], the optimal hyperparameters are showed in Table 2. The effectiveness of the 11 models was assessed through k-fold cross-validation [Hieu et al. 2023] with $k = 5$. The entire dataset is divided equally into 5 non-overlapping subsets. When running an evaluation, each of the 5 subsets is used as the testing set, and the rest are used as the training sets. The evaluation results of the model are averaged after 5 evaluations.

4.4 Metric

The evaluation was conducted using 2 metrics: Root Mean Square Error (RMSE) and R-squared (R^2):

RMSE measures the mean error of predicted values $\hat{y}_i$, in comparison to actual values y_i in the test dataset with n data samples. The RMSE calculation is delineated by the formula (2):

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \tag{2}$$

Algorithm	Parameters	Minimum	Maximum	Optimal
SVR (Linear)	cost (C)	9.77×10^{-4}	32	0.01
	margin (epsilon)	0	0.2	0.05
SVR (Polynomial)	cost (C)	9.77×10^{-4}	32	0.01
	degree	1	5	4
	scale_factor (scale)	10^{-10}	0.1	0.01
	margin (epsilon)	0	0.2	0.05
SVR (RBF)	cost (C)	9.77×10^{-4}	32	0.01
	rbf_sigma (sigma)	10^{-10}	0.1	0.05
	margin (epsilon)	0	0.2	0.05
LASSO	penalty	10^{-10}	1	0.63
	mixture (alpha)	1	1	1
RIDGE	penalty	10^{-10}	1	0.65
	mixture (alpha)	0	0	0
ENET	penalty	10^{-10}	1	0.62
	mixture (alpha)	0.05	1	0.7
PLS	predictor_prop	0	1	0.6
	num_comp	1	4	3
Random Forest	tree_depth	1	15	12
	n_trees	1	2000	1850
	min_rows	1	40	32
XGBoost	tree_depth	1	15	12
	n_trees	1	2000	1800
	learn_rate (eta)	0.001	0.4	0.05
	early_stop	3	20	5
LightGBM	tree_depth	1	15	12
	n_trees	1	2000	1800
	learn_rate (eta)	0.001	0.4	0.05
	early_stop	3	20	5

Table 2: Ranges of hyperparameters used in tuning of best results for various machine learning technique

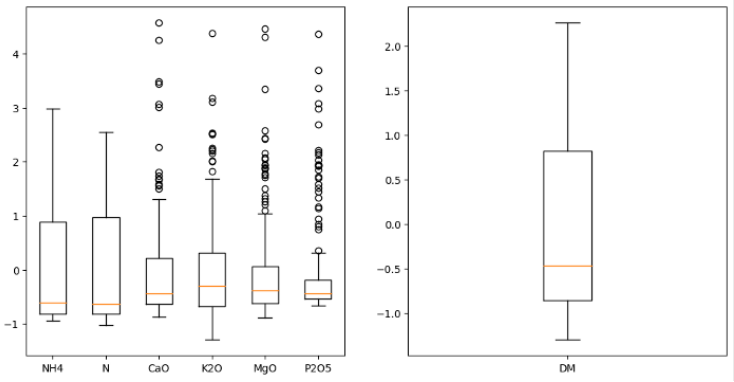

Figure 8: Distribution of Chemical Compound Data after Standardization

R^2, also known as the coefficient of determination, serves to evaluate the congruence between a model and the data. Given y_i as the actual value, $\hat{y}_i$ as the predicted value, and $\bar{y}$ representing the mean value of the target variable, the R^2 value is calculated using the following formula (3):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3}$$

The R^2 value spans a range from 0 to 1, with:

- $R^2 = 0$: indicating that the model does not explain any variation in the data and is entirely uncorrelated to it;
- $R^2 = 1$: indicating that the model completely explains all the variations in the data and perfectly fits the data;
- $R^2 \epsilon (0, 1)$: signifying the relative degree of fit of the model to the data; the higher the R^2 value, the better the model is at explaining variations in the data.

4.5 Evaluation

Tables 3 and 4 demonstrate that the EBAR model outperforms traditional machine learning models across different compounds: DM, NH_4, N, P_2O_5, CaO, and K_2O. Particularly for DM and NH_4, the proposed model exhibits significantly superior outcomes ($RMSE_{DM} = 3.785, RMSE_{NH_4} = 0.053, R^2_{DM} = 0.962, R^2_{NH_4} = 0.968$). However, for MgO, the SVR model with the RBF kernel provides slightly better results with $RMSE_{SVR(RBF)} = 0.078$ and $R^2_{SVR(RBF)} = 0.837$.

Figure 9 illustrates the wavelength ranges at which each constituent in the sample exhibits maximum absorption. Specifically, the wavelength range for DM falls within 1404-1498 and 1856-1892 nm, NH_4 falls within 1386-1506 nm, N falls within 1386-1528 nm, CaO ranges from 1348-1414 and 1496-1626 nm, K_2O ranges from 1336-1634 nm, P_2O_5 ranges from 1308-1678 nm, and MgO ranges from 1370-1420 and 1496-1534 nm. Table 5 demonstrates that this proposed method successfully identified the absorption maxima wavelengths for DM, NH_4, N, K_2O, and P_2O_5. In the case of CaO and MgO, the identified wavelength ranges did not converge, indicating that their

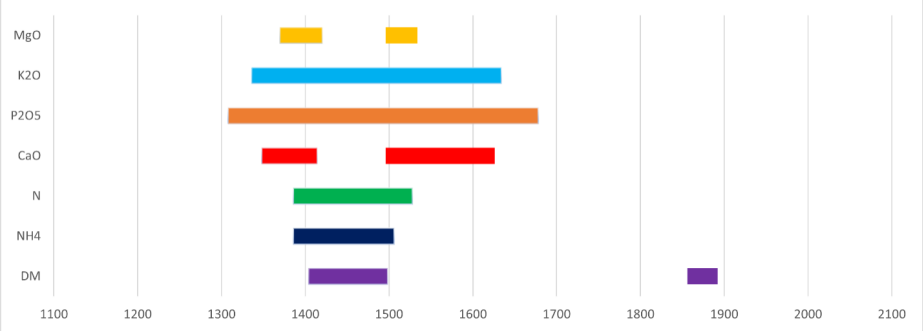

Figure 9: Wavelength Ranges of Maximum Absorption for DM, NH4, N, P2O5, CaO, MgO, and K2O

absorption spectra do not fall within the observed spectral region, which is in accordance with theoretical expectations.

Model	DM	NH_4	N	P_2O_5	CaO	MgO	K_2O	**Average**
SVR (Linear)	6.909	0.075	0.343	0.306	0.332	0.096	0.399	1.209
SVR (Polynomial)	5.158	0.078	0.346	0.307	0.410	0.091	0.398	0.970
SVR (RBF)	5.005	0.091	0.252	0.275	0.373	**0.078**	0.374	0.921
LASSO	7.154	0.082	0.368	0.317	0.335	0.091	0.412	1.251
RIDGE	9.307	0.102	0.505	0.390	0.394	0.107	0.472	1.611
ENET	7.103	0.083	0.370	0.317	0.335	0.090	0.412	1.252
PLS	8.647	0.097	0.449	0.320	0.349	0.092	0.441	1.485
Random Forest	5.987	0.091	0.339	0.449	0.380	0.121	0.471	1.120
XGBoost	5.642	0.082	0.279	0.458	0.407	0.133	0.414	1.059
LightGBM	4.607	0.092	0.345	0.325	0.392	0.114	0.414	0.898
Proposed Model - EBAR	**3.785**	**0.053**	**0.215**	**0.264**	**0.304**	0.089	**0.372**	**0.726**

Table 3: RMSE of different models across DM, NH_4, N, P_2O_5, CaO, MgO, and $K2O$

5 Discussion

In this study, the EBAR model is proposed, demonstrating a notably improved predictive performance compared to traditional machine learning methods for various chemical components, namely DM, NH_4, N, CaO, MgO, P_2O_5, and K_2O, in both cattle and poultry fertilizers. Furthermore, when compared to the findings of a previous study by [Cobbinah et al. 2022], the EBAR model exhibits marginally better results, with $RMSE_{mean} = 0.726$ and $R^2_{mean} = 0.865$ for the EBAR model, as opposed to

Model	DM	NH_4	N	P_2O_5	CaO	MgO	K_2O	**Average**
SVR (Linear)	0.919	0.944	0.928	0.770	0.673	0.716	0.656	0.801
SVR (Polynomial)	0.948	0.946	0.924	0.772	0.470	0.766	0.658	0.783
SVR (RBF)	0.950	0.896	0.951	0.852	0.564	**0.837**	0.689	0.820
LASSO	0.892	0.924	0.900	0.781	0.649	0.761	0.624	0.790
RIDGE	0.812	0.868	808	0.660	0.527	0.678	0.517	0.696
ENET	0.894	0.923	0.899	0.781	0.648	0.771	0.624	0.791
PLS	0.839	0.894	0.851	0.742	0.611	0.748	0.557	0.749
Random Forest	0.915	0.897	0.913	0.568	0.676	0.709	0.677	0.765
XGBoost	0.924	0.916	0.937	0.658	0.552	0.612	0.729	0.761
LightGBM	0.943	0.901	0.908	0.778	0.532	0.752	0.729	0.792
Proposed Model - EBAR	**0.962**	**0.968**	**0.966**	**0.878**	**0.741**	0.799	**0.738**	**0.865**

Table 4: R-squared of different models across DM, NH_4, N, P_2O_5, CaO, MgO, and $K2O$

Compound	**Experimental Wavelength Ranges**	**Theoretical Wavelength Ranges**
DM	1404-1498 and 1856-1892	1400-1500 and 1800-1900
NH_4	1386-1506	1300-1500
N	1386-1528	1300-1550
CaO	1348-1414 and 1496-1626	550-650
K_2O	1336-1634	1050-1640
MgO	1308-1678	1050-1640
P_2O_5	1370-1420 and 1496-1534	260-330

Table 5: Comparison of Experimental and Theoretical Wavelength Ranges

$RMSE_{mean} = 0.772$ and $R^2_{mean} = 0.863$ for the Stacked Regression model. Particularly, in contrast to research employing the PLS model for chemical content prediction, the EBAR model outperforms its predecessor in terms of accuracy.

The proposed Backward Elimination technique narrows down the wavelength range where chemicals in the sample exhibit strong absorption. This determination holds significance not only from a chemical perspective in explaining the chemical interactions but also contributes to refining the machine learning model. This method reduces feature space without compromising prediction accuracy, thus saving training time and model complexity.

In the case of MgO and CaO, while the EBAR model accurately predicts their content, Backward Elimination falters in explaining the precise wavelength range of maximal absorption. This arises from the fact that the observed fertilizer samples are natural products and have not undergone chemical interventions by humans. Table 1 reveals

that the ratio of MgO and CaO to P_2O_5 is approximately 1:3 and 1:1, respectively, in all criteria. Therefore, the EBAR model can predict MgO and CaO based on this ratio. This constraint can be addressed by expanding the surveyed spectral range and examining fertilizer samples that have undergone human chemical interventions.

Further optimization of the EBAR model can be achieved with an extended exploration of hyperparameter selection through Gridsearch Cross Validation. Additionally, the proposed model would exhibit enhanced performance through the augmentation of sample diversity, data distribution, and amount of samples. Henceforth, this study exhibits considerable potential for prospective avenues of future expansion.

6 Conclusions and Future Work

The utilization of machine learning techniques has been demonstrated as a reliable method for qualitative and quantitative analysis of NIR spectra in various industrial and agricultural domains. Although substantial prior endeavors have been conducted such as [Feng et al. 2022], [Huang et al. 2008], and [Devianti et al. 2022], this study proposes a synthetic machine learning model named EBAR, derived from traditional machine learning models, specifically designed for the chemical analysis of fertilizer samples. The model's effectiveness is evaluated using RMSE and R-squared metrics, demonstrating its superiority over conventional machine learning models, particularly the PLS model commonly employed in NIR spectroscopy, with $RMSE_{mean} = 0.726$ and $R^2_{mean} = 0.865$. Furthermore, this study introduces a method for determining the wavelength range of maximum absorption for various chemical components, providing a deeper understanding of their interaction with light and electromagnetic waves and furthering model development.

The results of this study are promising for numerous potential real-world applications. In the future, research efforts will be concentrated on the expansion of the dataset to encompass a broader data distribution, enriching fertilizer samples with human-induced chemical interactions, and extending the surveyed wavelength range to ensure accurate identification of maximum absorption wavelengths for each component. Additionally, the research will concentrate on the enhancement of model efficiency. The scope of this research is not limited to fertilizer samples and can be applied to various other types of products such as meat, fish, vegetables, fruits, or any other food items.

Acknowledgements

This study is funded and implemented for the project with number "24/HĐ-SKHCN,2023". This work is supported by the People's Committee, Da Nang, and the University of Science and Technology, University of Danang.

References

[Almaspoor et al. 2021] Almaspoor, M. H., Safaei, A., Salajegheh, A., Minaei-Bidgoli, B.: "Support Vector Machines in Big Data Classification: A Systematic Literature Review"; (2021).

[Chandrashekar and Sahin 2014] Chandrashekar, G., Sahin, F.: "A survey on feature selection methods"; (2014), Computers & Electrical Engineering, 40(1), 16–28.

[Cobbinah et al. 2022] Cobbinah, E., Generalao, O., Lageshetty, S. K., Adrianto, I., Singh, S., Dumancas, G. G.: "Using Near-Infrared Spectroscopy and Stacked Regression for the Simultaneous Determination of Fresh Cattle and Poultry Manure Chemical Properties"; (2022), Chemosensors, 10(10), 410.

[Cozzolino et al. 2007] Cozzolino, D., Liu, L., Cynkar, W. U., Dambergs, R. G., Janik, L., Colby, C. B., Gishen, M.: "Effect of temperature variation on the visible and near infrared spectra of wine and the consequences on the partial least square calibrations developed to measure chemical composition"; (2007), Analytica Chimica Acta, 588(2), 224–230.

[Devianti et al. 2021] Devianti, D., Yusmanizar, Y., Syakur, S., Munawar, A. A., & Yunus, Y: "Organic fertilizer from agricultural waste: determination of phosphorus content using near infrared reflectance"; (2021), IOP Conference Series: Earth and Environmental Science, 644, 012002. IOP Publishing.

[Devianti et al. 2022] Devianti, D., Sufardi, S., Mustaqimah, M., Munawar, A. A.: "Near Infrared Technology in Agricultural Sustainability: Rapid Prediction of Nitrogen Content from Organic Fertilizer"; (2022), Transdisciplinary Journal of Engineering & Science, 13.

[Dong et al. 2020] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: "A survey on ensemble learning"; (2020), Frontiers of Computer Science, 14, 241–258.

[Feng et al. 2022] Feng, X., Larson, R. A., Digman, M. F.: "Evaluation of Near-Infrared Reflectance and Transflectance Sensing System for Predicting Manure Nutrients". (2022), Remote Sensing, 14(4), 963.

[Fu et al. 2015] Fu, Y., Li, Z., Zhang, H., Xu, P.: "Using support vector machine to predict next day electricity load of public buildings with sub-metering devices"; (2015), Procedia Engineering, 121, 1016–1022.

[Fuquay et al. 2011] Fuquay, J. W., McSweeney, P. L. H., Fox, P. F.: "Encyclopedia of dairy sciences"; (2011), Academic Press.

[Gogé et al. 2021] Gogé, F., Thuriès, L., Fouad, Y., Damay, N., Davrieux, F., Moussard, G., Morvan, T.: "Dataset of chemical and near-infrared spectroscopy measurements of fresh and dried poultry and cattle manure"; (2021), Data in Brief, 34, 106647.

[Hien 2022] Hien, N., L., H.: "Deep Learning Analysis and Optimization for Different Data Sources in Autonomous Vehicles"; (2022), Luleå University of Technology, Department of Computer Science, Electrical and Space Engineering, 130.

[Hien et al. 2021] Hien, N. L. H., Minh, H. H., Van Tien, N., Van Hieu, N.: "Keyphrase Extraction Model: A New Design and Application on Tourism Information"; (2021), Informatica, 45(4), 563–569.

[Hien et al. 2024] Hien, N. L. H., Kor, A.-L., Ang, M. C., Rondeau, E., Georges, J.-P.: "Image Filtering Techniques for Object Recognition in Autonomous Vehicles"; (2024), Journal of Universal Computer Science, 30(1), 49–92.

[Hieu et al. 2023] Van Hieu, N., Hien, N. L. H., Van Huy, L., Tuong, N. H., Thoa, P. T. K.: "PlantKViT: A Combination Model of Vision Transformer and KNN for Forest Plants Classification"; (2023), JUCS: Journal of Universal Computer Science, 29(9).

[Horf et al. 2022] Horf, M., Vogel, S., Drücker, H., Gebbers, R., Olfs, H.W: "Optical Spectrometry to Determine Nutrient Concentrations and other Physicochemical Parameters in Liquid Organic Manures: A Review"; (2022), Agronomy, 12(2), 514.

[Huang et al. 2008] Huang, G., Han, L., Yang, Z., Wang, X.: "Evaluation of the nutrient metal content in Chinese animal manure compost using near infrared spectroscopy (NIRS)"; (2008), Bioresource Technology, 99(17), 8164–8169.

[Huy et al. 2023] Van Huy, L., Hien, N. L. H., Phuong, N. T. H., Van Hieu, N.: "Deep Learning Model with Hierarchical Attention Mechanism for Sentiment Classification of Vietnamese Comments"; (2023), Cybernetics and Physics, 12(2), 111–120.

[Jiaying et al. 2022] Jiaying, M., Tingting, C., Jie, L., Weimeng, F., Baohua, F., Guangyan, L.: "Functions of nitrogen, phosphorus and potassium in energy status and their influences on rice growth and development"; (2022), Rice Science, 29(2), 166–178.

[Jiang and Chen 2019] Jiang, H., Chen, Q.: "Determination of adulteration content in extra virgin olive oil using FT-NIR spectroscopy combined with the BOSS–PLS algorithm"; (2019), Molecules, 24(11), 2134.

[Kacprzak et al. 2023] Kacprzak, M., Malińska, K., Grosser, A., Sobik-Szołtysek, J., Wystalska, K., Dróżdż, D., Meers, E.: "Cycles of carbon, nitrogen and phosphorus in poultry manure management technologies–environmental aspects"; (2023), Critical Reviews in Environmental Science and Technology, 53(8), 914–938.

[Kumaravelu et al. 2015] Kumaravelu, C., Gopal, A.: "A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries"; (2015), 2015 IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR), 8–12.

[Liashchynskyi and Liashchynskyi 2019] Liashchynskyi, P., Liashchynskyi, P.: "Grid search, random search, genetic algorithm: a big comparison for NAS"; (2019), arXiv Preprint arXiv:1912. 06059.

[MacDonald 2010] MacDonald, J. M.: "Manure use for fertilizer and for energy: Report to Congress"; (2010), DIANE Publishing.

[Mamouei et al. 2020] Mamouei, M., Budidha, K., Baishya, N., Qassem, M., Kyriacou, P.: "Comparison of wavelength selection methods for in-vitro estimation of lactate: a new unconstrained, genetic algorithm-based wavelength selection"; (2020), Scientific Reports, 10(1), 16905.

[Olivos-Trujillo et al. 2015] Olivos-Trujillo, M., Gajardo, H. A., Salvo, S., González, A., Muñoz, C.: "Assessing the stability of parameters estimation and prediction accuracy in regression methods for estimating seed oil content in Brassica napus L. using NIR spectroscopy"; (2015), 2015 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 25–30. IEEE.

[Özdemir et al. 2018] Özdemir, İ. S., Dağ, Ç., Özinanç, G., Suçsoran, Ö., Ertaş, E., Bekiroğlu, S.: "Quantification of sterols and fatty acids of extra virgin olive oils by FT-NIR spectroscopy and multivariate statistical analyses"; (2018), LWT, 91, 125–132.

[Pagliari et al. 2020] Pagliari, P., Wilson, M., He, Z.: "Animal manure production and utilization: Impact of modern concentrated animal feeding operations"; (2020), Animal Manure: Production, Characteristics, Environmental Concerns, and Management, 67, 1–14.

[Pavia et al. 2015] Pavia, D., Donald, L.P., Gary, M.L., George, S.K., James, R.V.: "Introduction to spectroscopy"; (2015), SERBIULA (Sistema Librum 2. 0).

[Rahman et al. 2016] Rahman, F., Rahman, M. M., Rahman, G. M., Saleque, M. A., Hossain, A. S., Miah, M. G.: "Effect of organic and inorganic fertilizers and rice straw on carbon sequestration and soil fertility under a rice–rice cropping pattern"; (2016), Carbon Management, 7(1–2), 41–53.

[Roggo et al. 2007] Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., Jent, N.: "A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies"; (2007), Journal of Pharmaceutical and Biomedical Analysis, 44(3), 683–700.

[Tibshirani et al. 1996] Tibshirani, R.: "Regression shrinkage and selection via the lasso"; (1996), Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267–288.

[Wieringen 2015] van Wieringen, W. N.: "Lecture notes on ridge regression"; (2015), arXiv Preprint arXiv:1509. 09169.

[Yu et al. 2023] Yu, H., Li, Y., Du, W., Yang, M., Peng, X., Wang, X., Long, J.: "A Novel Interpretable Ensemble Learning Method for NIR-based Rapid Characterization of Petroleum Products"; (2023), IEEE Transactions on Instrumentation and Measurement.

[Zhang et al. 2022] Zhang, W., Kasun, L. C., Wang, Q. J., Zheng, Y., Lin, Z.: "A Review of Machine Learning for Near-Infrared Spectroscopy"; (2022), Sensors, 22(24), 9764.

[Zou and Hastie 2005] Zou, H., Hastie, T.: ”Regularization and variable selection via the elastic net”; (2005), Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301–320.