Pertanika J. Sci. & Technol. 31 (6): 2783 - 2803 (2023)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Intelligence System via Machine Learning Algorithms in Detecting the Moisture Content Removal Parameters of Seaweed Big Data

Olayemi Joshua Ibidoja[1,2], Fam Pei Shan[2], Mukhtar Eri Suheri[3], Jumat Sulaiman[4] and Majid Khan Majahar Ali[2]*

[1]*Department of Mathematics, Federal University Gusau, Gusau, Zamfara State, 234, Nigeria*
[2]*School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*
[3]*Department of Statistics, University of Sulatan Ageng Tirtayasa, Banten, Indonesia*
[4]*Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 UMS, Kota Kinabalu, Sabah, Malaysia*

ABSTRACT

The parameters that determine the removal of moisture content have become necessary in seaweed research as they can reduce cost and improve the quality and quantity of the seaweed. During the seaweed's drying process, many drying parameters are involved, so it is hard to find a model that can determine the drying parameters. This study compares seaweed big data performance using machine learning algorithms. To achieve the objectives, four machine learning algorithms, such as bagging, boosting, support vector machine, and random forest, were used to determine the significant parameters from the data obtained from v-GHSD (v-Groove Hybrid Solar Drier). The mean absolute percentage error (MAPE) and coefficient of determination (R2) were used to assess the model. The importance of variable selection cannot be overstated in big data due to the large number of variables and parameters that exceed the number of observations. It will reduce the complexity of the model, avoid the curse of dimensionality, reduce cost, remove irrelevant variables, and increase precision. A total of 435 drying parameters determined the moisture content removal, and each algorithm was used to select 15, 25, 35 and 45 significant parameters. The MAPE and R-Square for the 45 highest variable importance for random forest are 2.13 and 0.9732, respectively. It performed best, with the lowest error and the highest R-square. These results show that random forest is the best algorithm to decide the vital drying parameters for removing moisture content.

Keywords: Big data, drying, machine learning, seaweed, variable selection

ARTICLE INFO

Article history:
Received: 18 Ocotober 2022
Accepted: 07 March 2023
Published: 08 September 2023

DOI: https://doi.org/10.47836/pjst.31.6.09

E-mail addresses:
ojibidoja@fugusau.edu.ng (Olayemi Joshua Ibidoja)
fpeishan@usm.my; fpeishan@gmail.com (Fam Pei Shan)
mukhtar@untirta.ac.id (Mukhtar Eri Suheri)
jumat@ums.edu.my (Jumat Sulaiman)
majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Globally, the demand for food is increasing every day. The United Nations world population index predicts that by 2050, there will be 9.7 billion people on the planet (Namana et al., 2022). The demand for food will increase due to the rate of population growth and the effect of COVID-19. Rahimi et al. (2022) stated that COVID-19 affected the treatment of animals, while the lockdown affected the production of food and the supply of labour. Bajan et al. (2020) stated that world population growth comes with increased food production demand. It is essential to increase production, which involves increasing energy consumption to meet this demand. The need to meet the demand for food products and the food market is necessary (Safronova et al., 2022). Food is a security to the survival of human beings, and the hunger problem needs to be solved with a breakthrough in biotechnology (Prosekov & Ivanova, 2018). By 2050, the world should be prepared to feed over 9 billion people (Cole et al., 2018). To feed the increasing population, drying or preserving food is an important alternative that can be considered to preserve the nutritional value and quality of food.

Drying food involves the removal of moisture from the food. Solar driers dry many products, especially aquaculture and agriculture (Javaid et al., 2020). Nuroğlu et al. (2019) have used "drying in an oven under the magnetic field" and "drying under the sun and magnetic field" to dry grape samples and chilli pepper. From the results, the chilli pepper was the most contaminated when the traditional drying method was used. In East Africa, the rate of loss of farm produce is high due to the use of sun drying. The authors provided a prototyped modified solar dryer as another option. Multiple metallic solar panels were used to boost the drying performance of the solar drier (Ssemwanga et al., 2020).

The global seaweed industry provides diverse products directly or indirectly for human consumption, with a total value of approximately US$ 10 billion a year (Bixler & Porse, 2011). Malaysia's seaweed production grew from 1,000 metric tons in 1991 to 14,000 metric tons in 2009; in 2012, it was 33,000 metric tons. It is anticipated to continue growing exponentially over the next 30 years. The nursery, cultivating, drying, harvesting, processing, and marketing are a few processes involved in the carrageenan-bearing seaweed sector. Managing harvested seaweed biomass is essential to the carrageenophyte industry's entire value chain. It is important to understand the drying process for foods (Ali, Fudholi et al., 2017; Ali, Sulaiman et al., 2017). Seaweeds are very important to marine resources and are found in coastal waters. Seaweeds are very beneficial to human beings and fish. Seaweeds can be used as food, fertilizer, cosmetics, biofuel and medicine (Echave et al., 2022; Pradhan et al., 2022).

Machine learning variable selection has been used by many authors (Ali et al., 2021; Arjasakusuma et al., 2020; Gunn et al., 2022; Meyer et al., 2019). Application of the important variables will improve accuracy, reduce overfitting, and ensure robustness. Lim et al. (2020) used ridge regression to determine the drying parameters of fish and included the interaction terms.

As the dimensions of datasets in predictive modelling continue to grow, feature selection becomes increasingly practical. Datasets with complex feature interactions and high levels of redundancy still present a challenge to existing feature selection methods. Variable selection is becoming useful, and the dimensions of data, complexity, volume, interactions, heterogeneity, and the number of irrelevant variables make it a problem for traditional methods. A machine-learning algorithm can model complicated patterns (Solyali, 2020).

To fill the gap in the existing literature, a comparison of performance and evaluation of seaweed big data using four machine learning algorithms, such as bagging, boosting, support vector machine and random forest, will be evaluated.

MATERIALS AND METHODS

Model Building

Knowing the important parameters that determine the moisture content of the seaweed after drying is important. We combined the second-order interaction with the main 29 seaweed drying parameters to get 435 drying parameters, to develop an intelligent system. All 435 parameters determine the moisture content of the seaweed after drying. With this limitation, we have 435 parameters to predict the moisture content of seaweed after drying. Therefore, we selected 15, 25, 35, and 45 as the most important variables, and boosting, bagging, random forest, and support vector machines will be used as the algorithms. The system for drying has already been designed and data collected, but no optimization has been done. The flowchart in Figure 1 shows the procedure and the methodology used in this study.

Stage I

It involves the inclusion of all possible models.

$$\frac{n!}{(n-r)!\,r!} + number\ of\ single\ factor \tag{1}$$

where n is the number of single factors, *and* r is the number of orders. Equation 1 can compute the total number of all possible models.

Stage II

It requires simulation using random forest, boosting, support vector machine, and bagging machine learning algorithms to model the data. Each machine learning algorithm is then used to select the 15, 25, 35 and 45 highest important variables to determine the moisture content removal of the seaweed after drying. If the parameters are significant and satisfy the conditions, they will be imposed in the optimization. Otherwise, it will be removed. Features selection can only provide the rank of relevant variables and not the number of significant factors, and there is no rule for determining how many parameters should be used in a prediction model (Chowdhury & Turin, 2020; Drobnič et al., 2020; Kaneko,

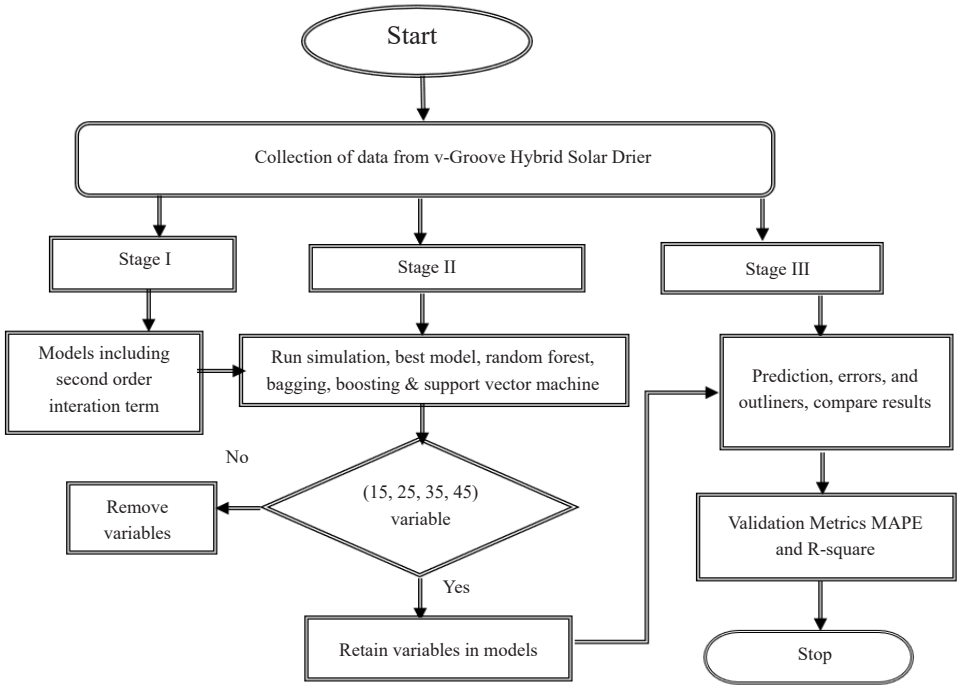

Figure 1. Flowchart of the steps for the study

2021). Hence, if the parameters do not rank among the 15, 25, 35, and 45 highest important variables, they will be removed from the model.

Stage III

The prediction must be made to achieve the objectives further, and the errors must be calculated. The outliers are also computed using 2 - sigma limits. Outliers are observations far from the measures of location (Leys et al., 2019). The best model is selected using mean absolute percentage error (MAPE) and coefficient of determination (R-square) metric validation.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (2)$$

Where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the forecast value.

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

Where y_i is the actual value, and $\hat{y}_i$ is the forecast value.

The metrics were used to assess the model. The formulae for the model metric validations are given in Equations 2 and 3. The lower the MAPE, the better the prediction accuracy. The higher the R-square, the better the prediction accuracy.

Boosting

Boosting is an ensemble machine learning modelling technique that can create a powerful classifier from a huge number of weak classifiers. Boosting can be used to improve the precision of a machine-learning algorithm. Rahman et al. (2020) stated that boosting algorithms use the training observations that end up in misclassifications. Boosting algorithms use the weight of the samples of the weak classifiers depending on the precision of the preceding boosting rounds (Alsahaf et al., 2022). Boosting is to measure the errors in the predecessors of the classifier, which makes it sensitive to outliers. It cannot be scaled- up since the estimator relies on the accuracy of the past predictors. Given the matrix of explanatory variables $X = [X_1, \dots, X_p] \in \mathbb{R}^{n \times p}$ and the dependent variable vector $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$. In the regression coefficients vector $\beta \in \mathbb{R}^p$, the value of the predicted response variable is represented by $\boldsymbol{X\beta}$, and the residuals are denoted by $\boldsymbol{\varepsilon = y - X\beta}$. The LSB **(**$\varepsilon$) denotes the least squares for the boosting. The LSB **(**ε) in regression produces models with interesting statistical properties (Freund et al., 2017). LSB **(**ε) algorithm is expressed as follows according to Freund et al. (2017) and Friedman (2001).

Algorithm: LSB (ε)

Fix the rate of learning $\varepsilon > 0$; the rate of learning is the rate at which the coefficients produced by LSB **(**ε) converge to the group of the unregularized least square solutions, and the iterations number M and initialize the $\hat{\beta}^0 = 0$ and $\hat{r}^0 = y$.

1. Do the following for $0 \leq k \leq M$, do the following:
2. Determine the covariate index j_k and $\tilde{u}_{j_k}$ as shown below:

   $j_k \in \underset{1 \leq m \leq p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\hat{r}_i^k - x_{im}\tilde{u}_m\right)^2$, where

   $\tilde{u}_m = arg\min_{u \in \mathbb{R}} \left(\sum_{i=1}^{n} \left(\hat{r}_i^k - x_{im}u\right)^2 \right)$, for m = 1,2, ... , p.
3. Update the residuals and coefficients of regression as follows:

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon X_{jk}\tilde{u}_{j_k}$$

$$\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \varepsilon\tilde{u}_{j_k} \text{ and } \hat{\beta}_j^{k+1} \leftarrow \hat{\beta}_j^k, j \neq j_k$$

Furthermore, the LSB algorithm at the kth iteration selects the covariate index j_k, leading to the maximal decrease in the fit of the univariate regression to the present residuals. Suppose the $X_{jk}\tilde{u}_{j_k}$ represents the best fit for the univariate regression to the present residuals. In that case, LSB will update the residuals $\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon X_{jk}\tilde{u}_{j_k}$ and the coefficients of the j_k th: $\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \tilde{u}_{j_k}$ to minimize the error.

Bagging

Bagging is an ensemble machine-learning technique that can improve the accuracy and

performance of machine-learning algorithms. A common integration technique known as bagging generates numerous copies of the training set using a bootstrapping procedure, which is then utilized to train several models (Yang et al., 2020). Bagging apply comparable learners to tiny sample populations before calculating the average of all the forecasts (Kabari et al., 2019). Bagging can improve the accuracy of a model, reduce overfitting of data, and deal with data of higher dimensionality effectively. However, bagging is computationally expensive.

Consider a regression setup where the data is represented by $L_i = (Y_i, X_i)\ (i = 1, .., n)$, and Y_i and X_i represent the p-dimensional variables for the ith instance. In the presence of a new independent variable or covariate x, a dependent variable for $\mathbb{E}[Y|X = x] = f(x)$, the response variable that corresponds to x can be represented by:

$$\hat{\theta}_n(x) = h_n(L_1, \dots, L_n)(x)$$

The estimator could be a learning algorithm or complex model, for instance, a linear regression via testing, classification, and regression trees. The h_n denotes the function of the sample n.

Theoretically, bagging is defined as:

i. Build a bootstrap sample $L_i^* = (Y_i^*, X_i^*)\ (i = 1, \dots, n)$ based on the practical distribution of the pairings $L_i = (Y_i, X_i)\ (i = 1, \dots, n)$.
ii. Using the plug-in principle to determine the bootstrapped forecaster $\hat{\theta}_n^*(x)$, which is, $\hat{\theta}_n^*(x) = h_n(L_1, \dots, L_n)(x)$.
iii. $\hat{\theta}_{n;B}(x) = \mathbb{E}^*\left[\hat{\theta}_n^*(x)\right]$ is the bagged predictor.

The bootstrap expectation in step (III) can be applied by Monte Carlo: Start with step (I) for every bootstrap simulation $j \in \{1, \dots, J\}$, to approximate $\hat{\theta}_{n;B}(x) \approx J^{-1} \sum_{j=1}^{J} \hat{\theta}_n^*(x)$, we compute $\hat{\theta}_n^*(x)\ (j = 1, \dots, J)$ as in step II, J is frequently chosen in the range of 50, depending on the sample size and computational cost related with assessing the predictor. The plug-in principle is used in bootstrapping by estimating the population's distribution from the information in the sample distribution.

Random Forest (RF)

RF can be defined as a combination of many classification and regression trees (CARTs), and the aim is to solve the problem of overfitting in individual CART (Georganos et al., 2021). Given that $\mathcal{L}$ is a learning set with a group of $\mathbb{N}$ pairs of features, with the response x_1, x_2, ..., x_n if $x_i \epsilon X$. A group of p-features x_i (*for* i = 1,2, ..., N) is an $N \times p$ matrix X, in which the rows i = 1,2, ..., N relates as x_i, with columns j = 1,2, ..., p as x_j. Similarly, the response can be written as a vector $y = (y_1, y_2, ..., y_N)$. In this scenario, the supervised learning job can be indicated as learning the function $\varphi: X \to Y$ from the learning set $\mathcal{L} = (X, y)$. The goal is to develop a model whose predictions for the variable $\varphi(x)$, represented by $\hat{Y}$, are as precise

as possible. In this situation, the *Y* variable must be continuous. The results of the model can be described as follows, given that the regressor function represents $\varphi: X \to Y$, in which $Y \epsilon \mathbb{R}$. During the statistical learning process, the explanatory variables, X_1, X_2, ... , X_p and the response variable *Y*, are random variables, and the values $X \times Y$ are selected collectively in respect of the joint probability distribution *P* (*XY*), the *X* is the vector of random [X_1], [X_2], ..., [X_p]. RF can handle large data effectively and has high precision over decision tree algorithms. However, more resources are needed for computation, and it takes more time when compared to a decision of the tree algorithm.

Algorithm:
For $b = 1$ to n

1. Create a bootstrapped sample D_b^* from the training set *D*.
2. Grow the tree by using the *m* from the bootstrapped sample D_b^*.

For a specific mode

i. Select *m* variables randomly.
ii. Identify the top split variables and values.
iii. Divide a node using the top divided variables and values.

Replicate steps 1–3 till the stopping conditions are satisfied.

Support Vector Machine (SVM)

SVM is popularly used to solve regression and classification problems. SVM has the capacity to discover nonlinear connections by using kernel function (Rashidi et al., 2019). Given $\{(x_1, y_1), \dots, (x_n, y_n)\} \in X \times \mathbb{R}$, *X* denotes the pattern space of the inputs $(X = \mathbb{R}^d)$.

From Cortes and Vapnik (1995), by comparing the standard Gaussian regression with the squared error loss function is minimized with the loss for observations *i* given as $L\{y_i, f(x_i)\} = \{y_i, f(x_i)\}^2$.
However, in support vector regression, the ε- insensitive loss function is minimized, and any loss lesser than ε is set to 0. Outside that bound, a simple linear loss function is applied in Equation 4:

$$L_\epsilon = f(x) = \begin{cases} 0, & if \ |y_i - f(x_i)| < \epsilon \\ |y_i - f(x_i) - \epsilon|, & otherwise \end{cases} \quad (4)$$

For instance, suppose *f*(*x*) is a linear function $f(x) = \beta_0 + x_i^t \beta$. Equation 5 represents the loss function given as:

$$\sum_{i=1}^{n} max(y_i - x_i^t \beta - \beta_0 - \epsilon, 0) \quad (5)$$

The ϵ is the tuning parameter and can be written as the constrained optimization problem: minimize $\frac{1}{2} \| \|_{\beta} \| \|^2$, subject to the constrain in Equation 6:

$$\begin{cases} y_i - x_i^t \beta - \beta_0 \leq \varepsilon \\ -(y_i - x_i^t \beta - \beta_0) \leq \varepsilon \end{cases} \tag{6}$$

If observations do not lie within the $\mathcal{E}$ band around that regression line, then there is no solution to the problem. The slack variables ζ_i and ζ_i^* are used; this allows the observations to fall outside the ε band around that regression line.

Minimize: $$\frac{1}{2} \| \|_{\beta} \| \|^2 + K \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \tag{7}$$

Subject to $$\begin{cases} y_i - x_i^t \beta - \beta_0 \leq \varepsilon + \zeta_i \\ -(y_i - x_i^t \beta - \beta_0) \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \end{cases} \tag{8}$$

The equation to minimize is shown in Equation 7, and the constraint is provided in Equation 8. However, before, $K > 0$ regulated how strong it is to prevent observations beyond the ε band.

Evaluation Metric

MAPE is widely used because it is easy to interpret and due to its scale-independency (Kim & Kim, 2016).

R^2 is the proportion of variance in the dependent variable that can be predicted from a set of independent variables. R^2 lies between 0 and 1. Moreover, the value determines the performance of the model (Chicco et al., 2021; Gouda et al., 2019; Ibidoja et al., 2016).

The MAPE and R-square are very useful for validating a model. The MAPE is the average difference between predicted and real values. In this study, MAPE is the average percentage error between the moisture content of the seaweed predicted by the model and the real value. The lower the MAPE, the better the model fits the data. R-square is the amount of variance in the dependent variable that the independent variable can describe. The higher the R-square value, the better the dependent variable can be explained. The R-square is also called the coefficient of determination; it is the quantity of variability in a variable explained by changes in the other variable. For a high precision in the model, we expect the error to be small and is the loss function for regression in machine learning (Jierula et al., 2021).

DATA COLLECTION

We used primary data on seaweed drying obtained using a v-groove hybrid solar drier. The data was collected using sensors installed on the v-GHSD (v-Groove Hybrid Solar Drier)

in Semporna, Malaysia, on the southeastern coast of Sabah. The data contains 29 different drying parameters that determine the moisture content of the seaweed, and each parameter has 1914 observations. The dependent variable is the moisture content after drying. We have different parameters that represent different relative humidity. Another parameter is the sun, and the remaining parameters are the temperature.

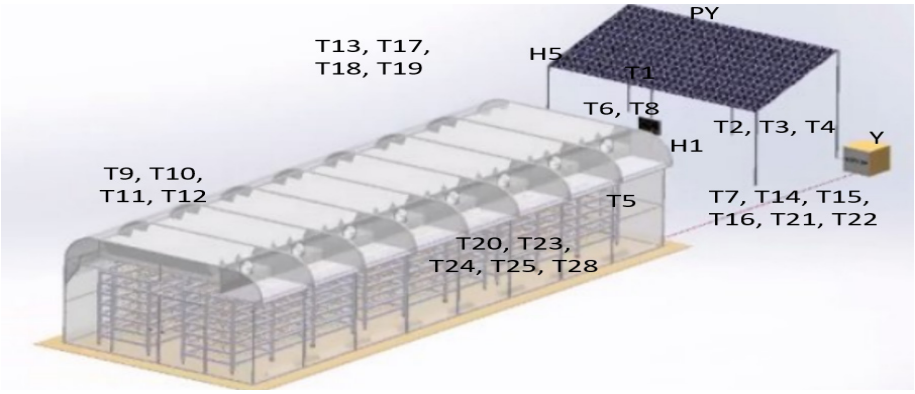

Figure 2. v-GHSD (v - Groove Hybrid Solar Drier)

Figure 2 represents the v-GHSD (v - Groove Hybrid) solar drier used to dry the seaweed and remove moisture. It also shows the positions of the parameters that determine the moisture content removal of the seaweed. The sensors are placed in strategic positions to measure the data. From Table 1, the T's are the temperature, Y is the moisture content, H's are the humidity, and PY is the solar radiation.

Table 1

Representation of parameters

Symbols	Factors	Meanings
Y	Dependent	Moisture Content
H1	Independent	Relative Humidity Ambient
H5	Independent	Relative Humidity Chamber
PY	Independent	Solar Radiation
T1	Independent	Temperature (°C) ambient
T2, T3, T4	Independent	Temperature (°C) prior to entering the solar collector
T5	Independent	Temperature (°C) in opposite the down v-Groove (Solar Collector)
T6, T8	Independent	Temperature (°C) in front of the up v-Groove (Solar Collector)

RESULTS AND DISCUSSION

Tables 2, 3, 4, and 5 show the variables selected for 15, 25, 35 and 45 for bagging, boosting, support vector machine, and random forest, respectively.

Table 2

The 15 highest variable importance

Model	Selected Variables
Bagging	T1, T4, T7, T8, T1*T8, T1*H5, T3*T12, T6*H1, T9*T13, T11*T15, T12*T28, T14*T22, T19*T21, T22*T26, T26*T27
Boosting	T1, T7, H1, H5, T9*PY, T25*T28, H1*H5, T10*H5, T13*T28, T28*T29, T10*H1, T7*H1, T26*T28, T12*H1, T2*T7
Support Vector Machine	T15, T16, T25, T26, T2*T8, T6*T8, T6*T10, T7*T10, T8*T17, T9*H5, T12*PY, T13*T17, T19*H5, T21*H5, T26*H5
Random Forest	T8, T2*T6, T1*T6, T6*T13, T21*H5, T19*H5, T7*T9, T22*H5, T6*T29, H5*PY, T7*H1, T8*H5, T26*H5, T8*H1, T1*T2

Table 3

The 25 highest variable importance

Model	Selected Variables
Bagging	T1,T4,T6,T7,T8,T9,T26,T1*T4,T1*T5,T1*T8,T1*H5,T2*T22,T3*T12,T3*PY,T4*T22,T6*H1,T6*H5,T9*T13,T10*PY,T11*T15,T12*T28,T14*T22,T19*T21,T22*T26,T26*T27
Boosting	T1,H1,H5,T7,T9*PY,T25*T28,H1*H5,T10*H5,T13*T28,T28*T29,T10*H1,T7*H1,T26*T28,T12*H1,T2*T7,T10*PY,T6*T7,T8*H1,T11*PY,T8*PY,T1*T4,T3*T5,T9*T21,T5*T26,T6*H5
Support Vector Machine	T15,T16,T25,T26,T28,T1*T17,T2*T8,T3*T17,T6*T8,T6*T10,T7*T10,T8*T17,T9*H5,T10*H1,T11*PY,T12*PY,T13*T17,T15*T16,T15*T25,T19*T22,T19*H5,T21*H5,T25*T28,T26*H5,T27*H5
Random Forest	T7, T8, H1, T1*T2, T1*T6, T1*T7, T1*T9, T2*T6, T2*T7, T2*T13, T6*T13, T6*T29, T7*T9, T7*H1, T8*H5, T8*H1, T14*H5, T19*H5, T21*H5, T22*H5, T25*H5, T26*H5, H1*PY, H1*H5, H5*PY

Table 4

The 35 highest variable importance

Model	Selected Variables
Bagging	T1, T4, T6, T7, T8, T9, T22, T26, PY, T1*T4, T1*T5, T1*T8, T1*H5, T2*T22, T3*T12, T3*PY, T4*T22, T5*T11, T5*T28, T6*H1, T6*H5, T6*PY, T7*T22, T7*PY, T8*T9, T9*T13, T10*T19, T10*T27, T10*PY, T11*T15, T12*T28, T14*T22, T19*T21, T22*T26, T26*T27

Table 4 *(Continue)*

Model	Selected Variables
Boosting	T1, T7, T8, T9, H1, H5, PY, T9*PY, T25*T28, H1*H5, T10*H5, T13*T28, T28*T29, T10*H1, T7*H1, T26*T28, T12*H1, T2*T7, T10*PY, T6*T7, T8*H1, T11*PY, T8*PY, T1*T4, T3*T5, T9*T21, T5*T26, T6*H5, H5*PY, T6*H1, H1*PY, T23*H5, T4*T5, T14*H5, T2*T28
Support Vector Machine	T12, T15, T16, T25, T26, T28, T1*T17, T2*T5, T2*T8, T2*T17, T3*T17, T3*T23, T4*T17, T5*T19, T6*T8, T6*T10, T7*T10, T7*T16, T7*T17, T8*T17, T8*T28, T9*H5, T10*H1, T11*T14, T11*PY, T12*PY, T13*T17, T15*T16, T15*T25, T19*T22, T19*H5, T21*H5, T25*T28, T26*H5, T27*H5
Random Forest	T2, T7, T8, T9, H1, T1*T2, T1*T6, T1*T7, T1*T9, T1*T13, T2*T6, T2*T7, T2*T13, T5*T9, T6*T8, T6*T9, T6*T13, T7*T9, T7*H1, T7*PY, T8*H1, T8*H5, T8*PY, T9*H5, T10*H5, T11*H5, T14*H5, T15*H5, T19*H5, T21*H5, T22*H5, T25*H5, H1*H5, H1*PY, H5*PY

Table 5

The 45 highest variable importance

Model	Selected Variables
Bagging	T1, T4, T6, T7, T8, T9, T10, T14, T22, T26, PY, T1*T2, T1*T4, T1*T5, T1*T8, T1*T10, T1*H5, T2*T22, T3*T12, T3*PY, T4*T22, T5*T11, T5*T28, T6*T15, T6*H1, T6*H5, T6*PY, T7*T22, T7*PY, T8*T9, T9*T13, T9*T23, T10*T19, T10*T27, T10*PY, T11*T15, T11*T16, T11*T17, T12*T25, T12*T28, T14*T22, T14*T23, T19*T21, T22*T26, T26*T27
Boosting	T1, T7, T8, T9, H1, H5, PY, T1*T4, T2*T7, T2*T28, T3*T5, T4*T5, T5*T25, T5*T26, T6*T7, T6*H1, T6*H5, T7*T12, T7*H1, T7*H5, T8*H1, T8*PY, T9*T21, T9*T26, T9*H5, T9*PY, T10*T23, T10*H1, T10*H5, T10*PY, T11*PY, T12*H1, T12*H5, T13*T28, T14*H5, T23*H5, T25*T28, T25*T29, T26*T28, T26*T29, T26*H5, T28*T29, H1*H5, H1*PY, H5*PY
Support Vector Machine	T12, T15, T16, T25, T26, T28, PY, T1*T5, T1*T6, T1*T17, T2*T5, T2*T6, T2*T8, T2*T17, T3*T5, T3*T17, T3*T23, T4*T5, T4*T17, T5*T13, T5*T19, T6*T8, T6*T10, T7*T10, T7*T16, T7*T17, T8*T17, T8*T28, T9*T23,T9*H5, T10*T23, T10*H1, T11*T14, T11*T22, T11*PY, T12*PY, T13*T17, T15*T16, T15*T25, T19*T22, T19*H5, T21*H5, T25*T28, T26*H5, T27*H5
Random Forest	T2, T7, T8, T9, H1, T1*T2, T1*T6, T1*T7, T1*T9, T1*T13, T1*H1, T2*T6, T2*T7, T2*T9, T2*T13, T3*T8, T3*H1, T4*T8, T4*H1, T5*T9, T6*T7, T6*T8, T6*T9, T6*T13, T7*T9, T7*H1, T7*PY, T8*H1, T8*H5, T8*PY, T9*H5, T10*H5, T11*H5, T13*H1, T14*H5, T15*H5, T16*H5, T19*H5, T21*H5, T22*H5, T23*H1, T25*H5, H1*H5, H1*PY, H5*PY

Table 6 is a four-dimensional results summary of the single factor the machine learning algorithms selected for the 45 high-ranking variables. Bagging and boosting have 4 single factors in common, bagging and SVM have 2 single factors in common, and bagging and random forest have 3 single factors in common for the first 45 highest important significant factors that determine the moisture content of the seaweed. For every machine learning algorithm, 11 (24.4%), 7 (15.6%), 7 (15.6%), and 5 (11.1%) single parameters are selected for bagging, boosting, support vector machine, and random forest, respectively. Bagging selects the highest number of single variables.

Table 6

Number of single variables for the 45 selected single factors for each machine learning algorithm

	Bagging	Boosting	SVM	Random Forest
Bagging	11	4	2	3
Boosting	4	7	-	4
SVM	2	-	7	-
Random Forest	3	4	-	5

Table 7 is a four-dimensional results summary of the second-order factors the machine learning algorithms selected for the 45 high-ranking variables. Bagging and boosting have 4 second-order factors in common, bagging and SVM have 2 second-order factors in common, and bagging and random forest have 2 second-order factors in common. Random forest and boosting have the highest number of second-order factors of 11 in common with the factors selected. For the interaction, bagging, boosting, support vector machine and random forest selected, 34 (75.6%), 38 (84.4%), 38 (84.4%), and 40 (88.9%).

Table 7

Number of similar variables for the 45 selected interaction factors for each machine learning algorithm

	Bagging	Boosting	SVM	Random Forest
Bagging	34	4	2	2
Boosting	4	38	7	11
SVM	2	7	38	6
Random Forest	2	11	6	40

The high number of interaction variables selected showed how important it is to consider the interaction in the moisture content removal of agricultural products. The results of the interaction are also supported by the results of (Javaid et al., 2019).

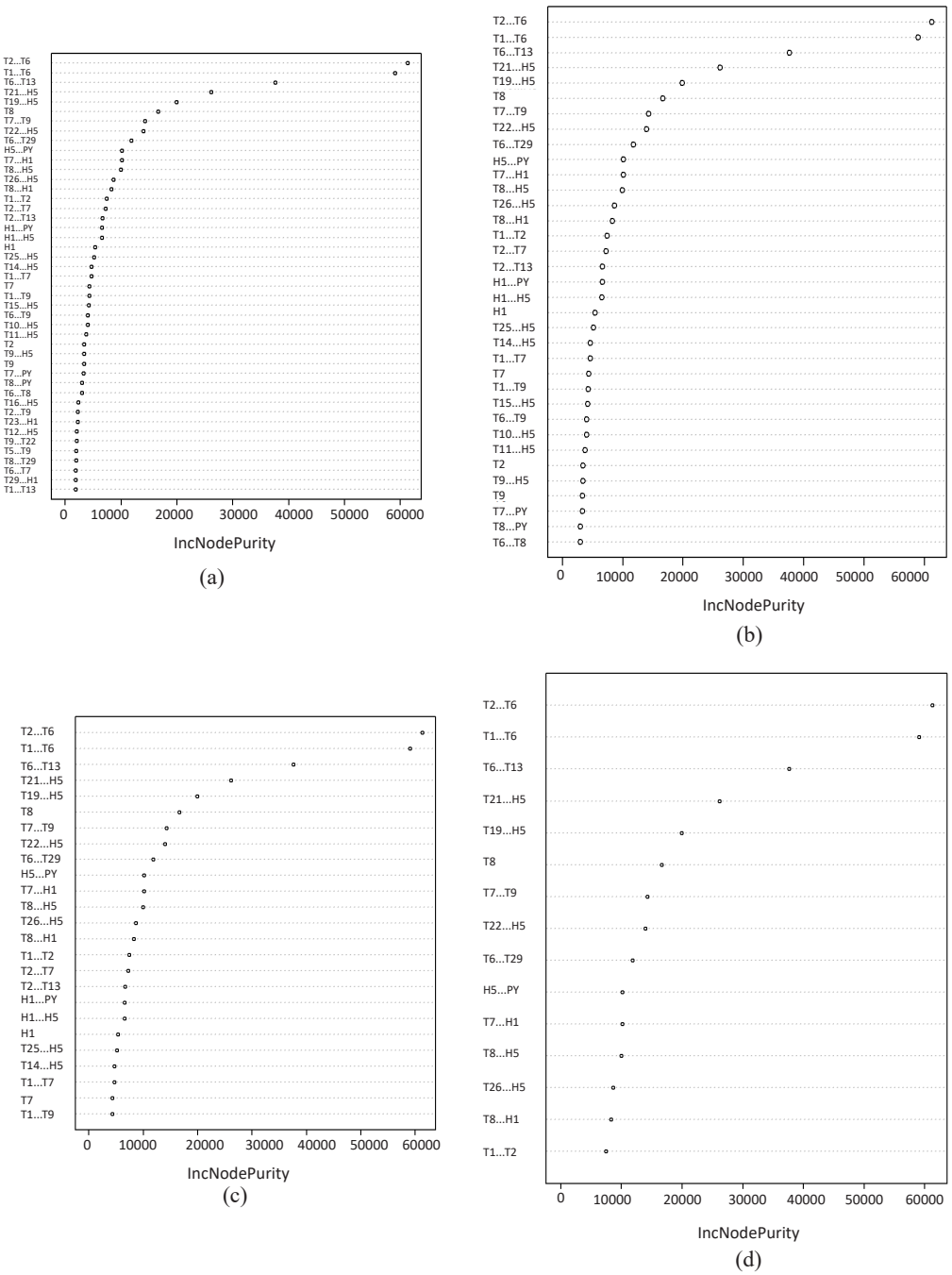

Figure 3. Variable importance for (a) 45, (b) 35, (c) 25, and (d) 15 random forest

Figure 3 shows the plot of the variable importance of the selected 45, 35, 25, and 15 parameters of the seaweed big data using the random forest algorithm. We used the *randomForest* and *caret* packages in the R programming language to achieve this. It reveals how important the variable is in determining the moisture content removal of the seaweed. The five most important factors are the parameters T2*T6, T1*T6, T6*T13, T21*H5, and T19*H5. T2*T6 and T1*T6 rank at the top for the factors.

Figures 4 to 7 show the standardized residual plots of random forest, support vector machine, bagging and boosting, respectively. Each algorithm produced different patterns of plots. Likewise, there were differences when the number of parameters 15, 25, 35 and 45 selected were compared for each algorithm. It also shows the upper and lower control limits to identify outliers. The percentage of outliers is calculated using the number of observations outside the 2-sigma limit.

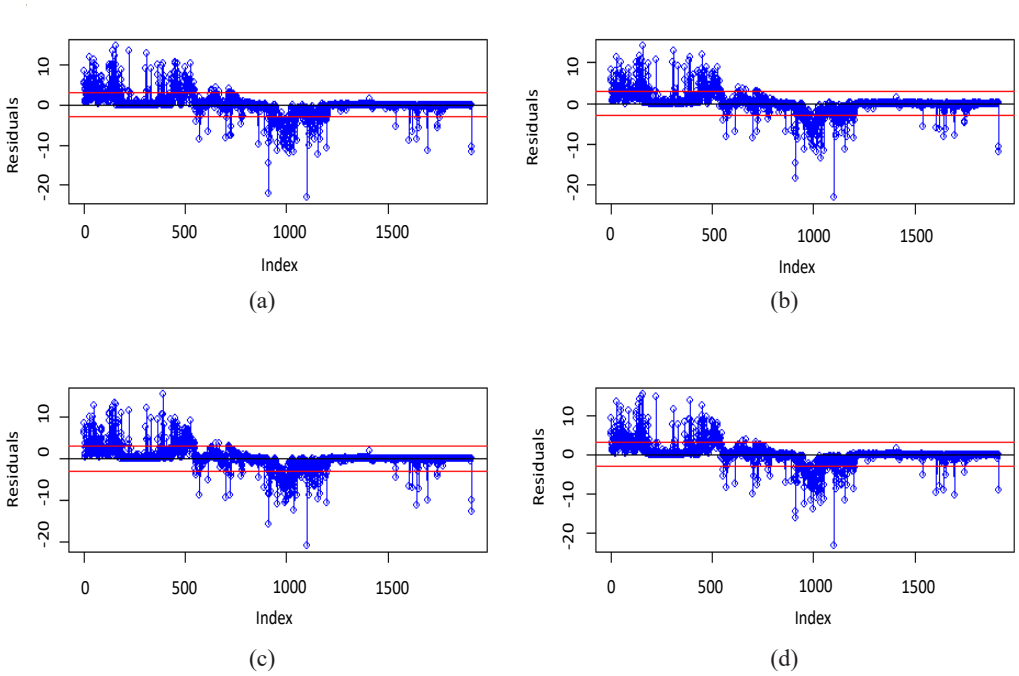

Figure 4. Plot for standardized residuals for (a) 15, (b) 25, (c) 35 and (d) 45 high-ranking variables for random forest

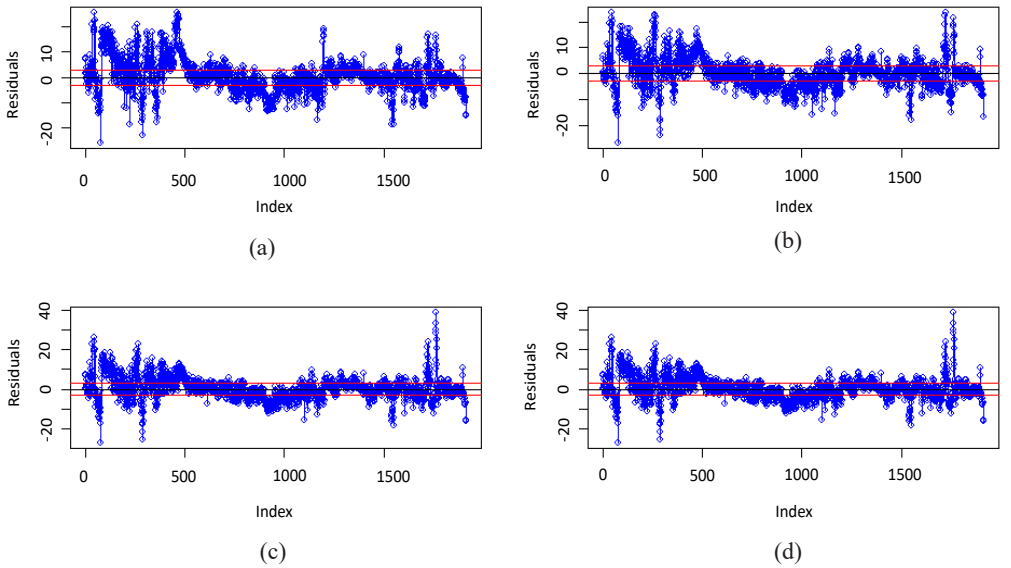

Figure 5. Plot for standardized residuals for (a) 15, (b) 25, (c) 35 and (d) 45 high-ranking variables for support vector machine

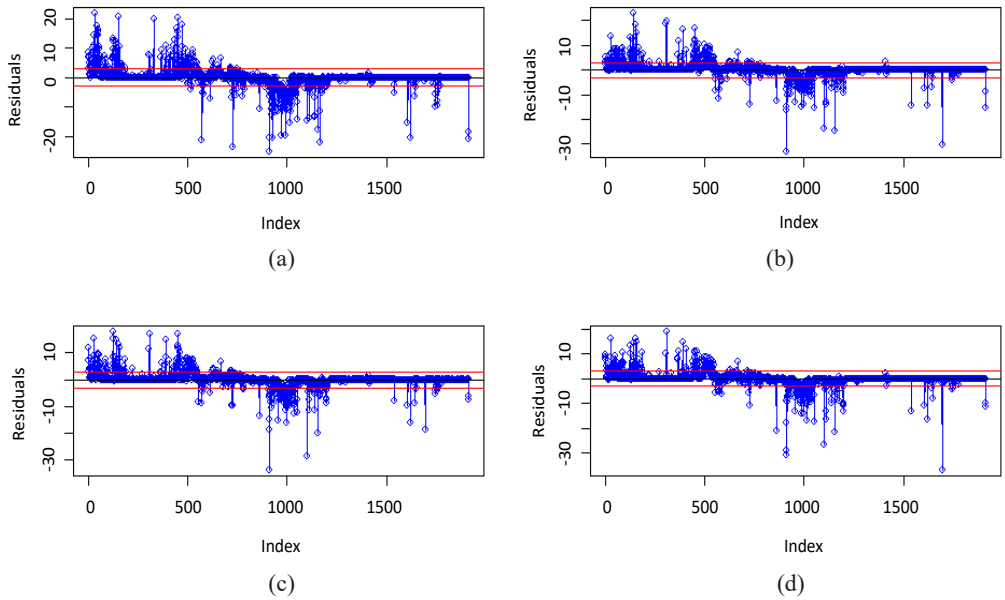

Figure 6. Plot for standardized residuals for (a) 15, (b) 25, (c) 35 and (d) 45 high-ranking variables for bagging

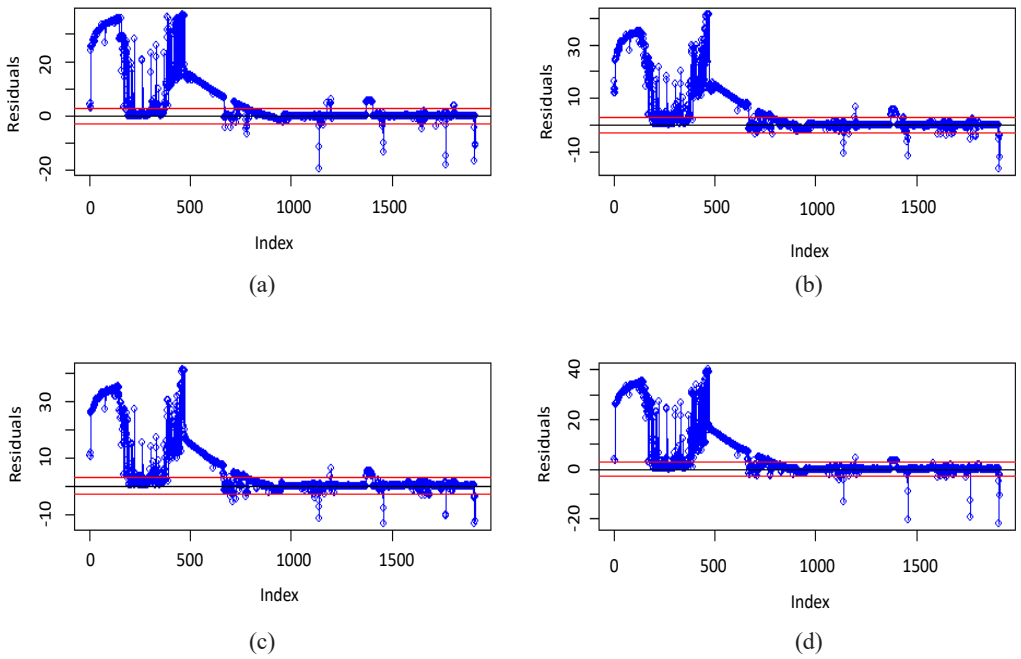

Figure 7. Plot for standardized residuals for (a) 15, (b) 25, (c) 35 and (d) 45 high-ranking variables for boosting

Table 8 shows the evaluation metrics for the 15, 25, 35 and 45 highest important ranking variables. The 45 highest-ranking important variables for random forest have the value of MAPE (2.13) and R-Square (0.9732), which gave the best performance. It is similar to Chen et al. (2020), where random forest performed better than the support vector machine. With these results, random forest is the best to determine the moisture content removal of the seaweed. MAPE (2.13) represents the average percentage error between the moisture content removal of the seaweed predicted by the model and the real value. Sumari et al. (2021) assert that if MAPE is less than 10, it is high prediction accuracy. The value R-square (0.9732) implies that the selected drying parameters can explain 97.32% of the dependent variable moisture content variance.

Table 8

Evaluation metrics

High Ranking Variables	Bagging		Boosting		Support Vector Machine		Random Forest	
	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2
15	12.26	0.7284	8.17	0.5310	8.61	0.8348	2.46	0.9638
25	9.78	0.8270	8.70	0.5544	7.98	0.8691	2.33	0.9671

Table 8 *(Continue)*

High Ranking Variables	Bagging		Boosting		Support Vector Machine		Random Forest	
	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2
35	8.41	0.8669	8.18	0.5368	7.57	0.8758	2.18	0.9715
45	8.15	0.8767	8.20	0.5570	8.61	0.8348	2.13	0.9732

Table 9 shows the number of outliers with their percentages using the 2- sigma limits. Data can have outliers due to factors that cannot be controlled, and these outliers will affect the prediction accuracy (Lim et al., 2020; Rajarathinam & Vinoth, 2014). For 15 important variables, boosting has the highest number of outliers, with 185. It represents less than 10% of the total observations. For the 25 highest important variables, boosting has the highest number of outliers, with 152 observations. It represents less than 8% of the total observations. Of the 35 highest important variables, boosting has the highest number of outliers, with 171 observations. It represents less than 9% of the total observations. Of the 45 highest important variables, boosting has the highest number of outliers, with 165 observations. It represents less than 9% of the total observations. Although the random forest has one of the highest outliers, it performs better. This result also aligns with (Liu et al., 2018), where random forest performed better with uncertainties and variances.

Table 9

Percentage of outliers outside 2 - sigma limits

Model	The Highest Variable Importance			
	15	25	35	45
	$\mu \pm 2\sigma$ (%)	$\mu \pm 2\sigma$ (%)	$\mu \pm 2\sigma$ (%)	$\mu \pm 2\sigma$ (%)
Random Forest	123(6.43%)	119(6.22%)	99(5.17%)	92(4.81%)
SVM	106(5.54%)	89(4.65%)	79(4.13%)	80(4.18%)
Bagging	81(4.23%)	84(4.39%)	77(4.02%)	73(3.81%)
Boosting	185(9.67%)	152(7.94%)	171(8.93%)	165(8.62%)

The number outside and inside the parentheses are the number of outliers and the percentage of outliers for the 2-sigma limit, respectively.

CONCLUSION

This study computed the total number of all possible models to achieve the objectives. Random forest, boosting, support vector machine, and bagging machine learning algorithms were used to model the data. The 15, 25, 35 and 45 highest important variables were selected to determine the moisture content removal of the seaweed big data after the

drying. The errors and outliers were computed using metric validation and 2-sigma limits. The percentage representing the interaction parameters has shown how important it is to determine the moisture content removal of seaweed because the interaction parameters selected by the algorithms are more than the single parameters. From the results, the random forest with 45 highest variable importance gave better results when compared to the bagging, boosting and support vector machine. The values of MAPE (2.13) and R-Square (0.9732) gave the best performance. With these results, an intelligence system based on random forest is the best algorithm to determine the important drying parameters for the moisture content removal of the seaweed with the lowest error.

A few batches of experiments can be used to confirm these results for future work. Missing value implications can also be investigated.

ACKNOWLEDGEMENT

The authors thank the "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2022/STG06/USM/02/13" for their support. We are also grateful to the anonymous reviewers for their comments and suggestions to improve the clarity and quality of the paper.

REFERENCES

Ali, M. K. M., Fudholi, A., Sulaiman, J., Muthuvalu, M. S., Ruslan, M. H., Yasir, S. M., & Hurtado, A. Q. (2017). Post-harvest handling of eucheumatoid seaweeds. In A. Q. Hurtado, A. T. Critchley & L. C. Neish (Eds.), *Tropical Seaweed Farming Trends, Problems and Opportunities* (pp. 131-145). Springer International Publishing. https://doi.org/10.1007/978-3-319-63498-2_8

Ali, M. K. M., Sulaiman, J., Yasir, S. M., Ruslan, M. H., Fudholi, A., Muthuvalu, M. S., & Ramu, V. (2017). Cubic spline as a powerful tools for processing experimental drying rate data of seaweed using solar drier. *Article in Malaysian Journal of Mathematical Sciences*, *11*(S), 159-172.

Ali, M. K. M., Mukhtar, Ismail, M. T., Ferdinand, M. H., & Alimuddin. (2021). Machine learning-based variable selection: An evaluation of bagging and boosting. *Turkish Journal of Computer and Mathematics Education*, *12*(13), 4343-4349.

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, *187*, Article 115895. https://doi.org/10.1016/j.eswa.2021.115895

Arjasakusuma, S., Kusuma, S. S., & Phinn, S. (2020). Evaluating variable selection and machine learning algorithms for estimating forest heights by combining lidar and hyperspectral data. *ISPRS International Journal of Geo-Information*, *9*(9), 1-26. https://doi.org/10.3390/ijgi9090507

Bajan, B., Mrówczyńska-Kamińska, A., & Poczta, W. (2020). Economic energy efficiency of food production systems. *Energies*, *13*(21), 1-16. https://doi.org/10.3390/en13215826

Bixler, H. J., & Porse, H. (2011). A decade of change in the seaweed hydrocolloids industry. *Journal of Applied Phycology*, *23*(3), 321-335. https://doi.org/10.1007/s10811-010-9529-3

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 1-26. https://doi.org/10.1186/s40537-020-00327-4

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, Article e623. https://doi.org/10.7717/peerj-cs.623

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), Article e000262. https://doi.org/10.1136/fmch-2019-000262

Cole, M. B., Augustin, M. A., Robertson, M. J., & Manners, J. M. (2018). The science of food security. *Npj Science of Food*, *2*(1), 1-8. https://doi.org/10.1038/s41538-018-0021-9

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273-297.

Drobnič, F., Kos, A., & Pustišek, M. (2020). On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics*, *9*(5), Article 761. https://doi.org/10.3390/electronics9050761

Echave, J., Otero, P., Garcia-Oliveira, P., Munekata, P. E. S., Pateiro, M., Lorenzo, J. M., Simal-Gandara, J., & Prieto, M. A. (2022). Seaweed-derived proteins and peptides: Promising marine bioactives. *Antioxidants*, *11*(1), 1-26. https://doi.org/10.3390/antiox11010176

Freund, R. M., Grigas, P., & Mazumder, R. (2017). A new perspective on boosting in linear regression via subgradient optimization and relatives. *Annals of Statistics*, *45*(6), 2328-2364. https://doi.org/10.1214/16-AOS1505

Friedman, J. H. (2001). Greedy Function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2021). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *36*(2), 121-136. https://doi.org/10.1080/10106049.2019.1595177

Gouda, S. G., Hussein, Z., Luo, S., & Yuan, Q. (2019). Model selection for accurate daily global solar radiation prediction in China. *Journal of Cleaner Production*, *221*, 132-144. https://doi.org/10.1016/j.jclepro.2019.02.211

Gunn, H. J., Rezvan, P. H., Fernández, M. I., & Comulada, W. S. (2022). How to apply variable selection machine learning algorithms with multiply imputed data: A missing discussion. *Psychological Methods*, *28*(2), 452-471. https://doi.org/10.1037/met0000478

Ibidoja, O. J., Ajare, E. O., & Jolayemi, E. T. (2016). Reliability measures of academic performance. *International Journal of Science for Global Sustainability*, *2*(4), 59-64.

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2020). Comparison of sparse and robust regression techniques in efficient model selection for moisture ratio removal of seaweed using solar drier. *Pertanika Journal of Science and Technology*, *28*(2), 609-625.

Javaid, A., Muthuvalu, M. S., Sulaiman, J., Ismail, M. T., & Ali, M. K. M. (2019). Forecast the moisture ratio removal during seaweed drying process using solar drier. *AIP Conference Proceedings*, *2184*, Article 050016. https://doi.org/10.1063/1.5136404

Jierula, A., Wang, S., Oh, T. M., & Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences*, *11*(5), 1-21. https://doi.org/10.3390/app11052314

Kabari, L. G., Onwuka, U., & Onwuka, U. C. (2019). Comparison of bagging and voting ensemble machine learning algorithm as a classifier. *International Journal of Computer Science and Software Engineering*, *9*(3), 19-23.

Kaneko, H. (2021). Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables. *Heliyon*, *7*(6), 1-12. https://doi.org/10.1016/j.heliyon.2021.e07356

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669-679. https://doi.org/10.1016/J.IJFORECAST.2015.12.003

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1), 1-10. https://doi.org/10.5334/irsp.289

Lim, H. Y., Fam, P. S., Javaid, A., & Ali, M. K. M. (2020). Ridge regression as efficient model selection and forecasting of fish drying using v-groove hybrid solar drier. *Pertanika Journal of Science and Technology*, *28*(4), 1179-1202. https://doi.org/10.47836/pjst.28.4.04

Liu, C., Tang, F., & Bak, C. L. (2018). An accurate online dynamic security assessment scheme based on random forest. *Energies*, *11*(7), Article 1914. https://doi.org/10.3390/en11071914

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications - Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411,* Article 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Namana, M. S. K., Rathnala, P., Sura, S. R., Patnaik, P., Rao, G. N., & Naidu, P. V. (2022). Internet of things for smart agriculture - State of the art and challenges. *Ecological Engineering and Environmental Technology*, *23*(6), 147-160. https://doi.org/10.12912/27197050/152916

Nuroğlu, E., Öz, E., Bakırdere, S., Bursalıoğlu, E. O., Kavanoz, H. B., & İçelli, O. (2019). Evaluation of magnetic field assisted sun drying of food samples on drying time and mycotoxin production. *Innovative Food Science and Emerging Technologies*, *52*, 237-243. https://doi.org/10.1016/j.ifset.2019.01.004

Pradhan, B., Bhuyan, P. P., Patra, S., Nayak, R., Behera, P. K., Behera, C., Behera, A. K., Ki, J. S., & Jena, M. (2022). Beneficial effects of seaweeds and seaweed-derived bioactive compounds: Current evidence and future prospective. *Biocatalysis and Agricultural Biotechnology*, *39*, Article 102242. https://doi.org/10.1016/j.bcab.2021.102242

Prosekov, A. Y., & Ivanova, S. A. (2018). Food security: The challenge of the present. *Geoforum*, *91*, 73-77. https://doi.org/10.1016/j.geoforum.2018.02.030

Rahimi, P., Islam, M. S., Duarte, P. M., Tazerji, S. S., Sobur, M. A., el Zowalaty, M. E., Ashour, H. M., & Rahman, M. T. (2022). Impact of the COVID-19 pandemic on food production and animal health. *Trends in Food Science and Technology*, *121*, 105-113. https://doi.org/10.1016/j.tifs.2021.12.003

Rahman, S., Irfan, M., Raza, M., Ghori, K. M., Yaqoob, S., & Awais, M. (2020). Performance analysis of boosting classifiers in recognizing activities of daily living. *International Journal of Environmental Research and Public Health*, *17*(3), Article 1082. https://doi.org/10.3390/ijerph17031082

Rajarathinam, A., & Vinoth, B. (2014). Outlier detection in simple linear regression models and robust regression-A case study on wheat production data. *International Journal of Scientific Research*, *3*(2), 531-536.

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. *Academic Pathology*, *6*, 1-17. https://doi.org/10.1177/2374289519873088

Safronova, O. V., Polyakova, E. D., Evdokimova, O. V., Demina, E. N., Lazareva, T. N., & Petrova, O. A. (2022). Development of sustainable systems of food production using spirulina platensis dairy technology as a functional filler. *IOP Conference Series: Earth and Environmental Science*, *981*(2), Article 022074. https://doi.org/10.1088/1755-1315/981/2/022074

Solyali, D. (2020). A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus. *Sustainability*, *12*(9), Article 3612. https://doi.org/10.3390/SU12093612

Ssemwanga, M., Makule, E., & Kayondo, S. I. (2020). Performance analysis of an improved solar dryer integrated with multiple metallic solar concentrators for drying fruits. *Solar Energy*, *204*, 419-428. https://doi.org/10.1016/j.solener.2020.04.065

Sumari, A. D. W., Charlinawati, D. S., & Ariyanto, Y. (2021). A simple approach using statistical-based machine learning to predict the weapon system operational readiness. *Proceedings of the International Conference on Data Science and Official Statistics*, *2021*(1), 343-351. https://doi.org/10.34123/icdsos.v2021i1.58

Yang, W., Yuan, T., & Wang, L. (2020). Micro-blog sentiment classification method based on the personality and bagging algorithm. *Future Internet*, *12*(4), Article 75. https://doi.org/10.3390/fi12040075