Journal of Big Data

RESEARCH Open Access

Detecting heterogeneity parameters and hybrid models for precision farming

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

*Correspondence:
majidkhanmajaharali@usm.my

[1] Department of Mathematics, Federal University Gusau, Gusau, Nigeria
[2] School of Mathematical Sciences, Universiti Sains Malaysia USM, 11800 Penang, Malaysia
[3] Faculty of Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia

Abstract

Precision farming (PF) plays a crucial role in the field of agriculture to solve the challenges of food shortages in society. Heterogeneity, multicollinearity, and outliers are problems in PF because they can cause bias and lead to incorrect inferences. However, traditional methods typically assume it to be a homogenous model, and in machine learning, data scientists ignore heterogeneity. In this study, the aim is to identify the heterogeneity parameters and develop hybrid models before and after heterogeneity. Data on seaweed is collected using sensor smart farming technology attached to v-Groove Hybrid Solar Drier (v-GHSD). There are 29 drying parameters, and each parameter has 1914 observations. We considered the highest order up to the second order interaction, and the parameters increased to 435 parameters from 29 parameters. In high-dimensional data, the number of observations is less than the number of parameters. The authors proposed a method using the variance inflation factor to identify the heterogeneity parameters. Seven predictive models such as ridge, random forest, support vector machine, bagging, boosting, LASSO and elastic net are used to select the 15, 25, 35 and 45 significant drying parameters for the moisture content removal of the seaweed, and hybrid models are developed using robust statistical methods. For before heterogeneity, the hybrid model random forest M Hampel with 19 outliers is the best, because it performs better when compared to other models. For after heterogeneity, the hybrid model boosting M Hampel with 19 outliers is the best, because it performs better when compared to other models. These results are vital to seaweed precision farming. The study of heterogeneity will not only help us to comprehend the dynamics of the large number of the drying parameters, but also gives a way to leverage the data for efficient predictive modelling.

Keywords: Big data, Precision agriculture, Heterogeneity, Machine learning, Forecast, Parameters

Introduction

Farming involves the growing of crops and the rearing of livestock. It is a source of raw materials for industries. The traditional methods used by farmers are not precise, which leads to manual labour and the consumption of time [1]. Precision farming (PF) plays a vital role in the field of agriculture to solve the challenges of food shortages in society. The PF method is a subset of smart farming technologies (SFTs) that deals with information systems, the internet of things (IoT), precision agriculture

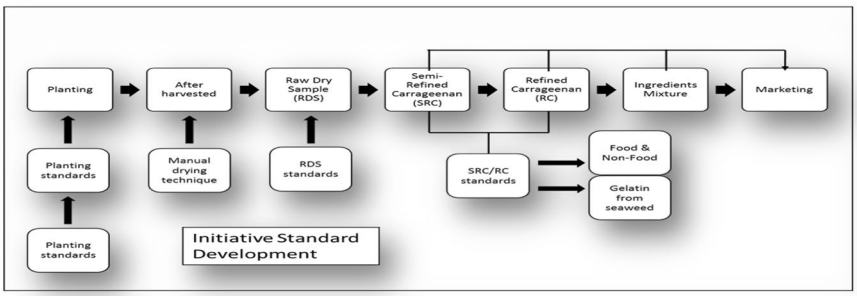

Fig. 1 Seaweed processing application [25]

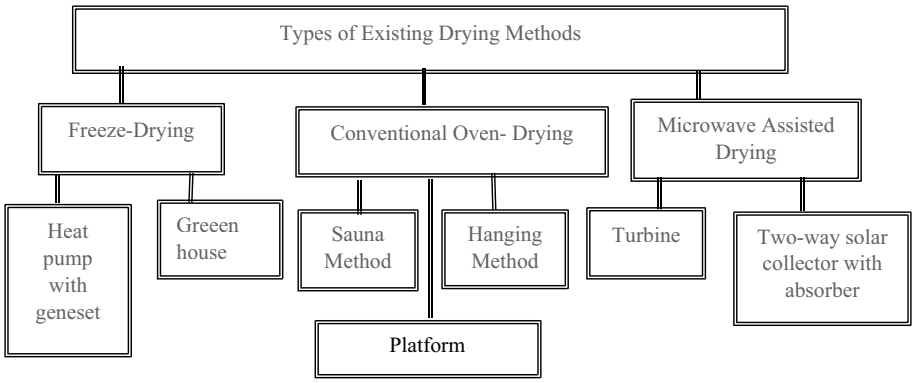

Fig. 2 Types of existing drying methods for seaweed

systems, artificial intelligence, cloud computing, farm management, wireless sensor networks, robotics, and automation of agriculture [2, 3]. The merit of the method is that it boosts farm profits and cuts down the cost of production [6].

Seaweeds are also called macroalgae. They are like plant organisms attached to rocks or rock layers. In addition, they grow in lakes, oceans, rivers, and water bodies [7, 8]. It a crucial source of fat, carbohydrates, vitamins, fibre, and ash, as well as proteins and beta-carotene [9]. For example, seaweed is useful in many forms (for example, powder, fresh, salted, canned, dried or extracts) for eating by humans or as feeds, biofuels, medicines, and fertilisers [10]. (See Fig. 1 for the stages involved in seaweed pre-harvest and post-harvest of seaweed).

One of the post-harvest problems with seaweed is the high moisture content. According to [11], seaweed is easily damaged when it is very fresh. Therefore, this demands that seaweed be dried after harvesting. The drying of seaweed is used to reduce the moisture content [15]. The biomass weight of seaweed during transportation will be decreased, which makes it available for additional processing [12]. Drying also reduces storage, transportation, and processes to prevent losses and increase value [14]. The types of drying are freeze-drying (direct drying method), conventional drying and microwave- assisted drying (solar). See Fig. 2 for details. A solar drier is the most efficient drying method for seaweed and can dry the water content faster

[16]. These authors [13, 17–19] have employed solar driers in their studies. The drying parameters using v-Groove Hybrid Solar Drier (v-GHSD) were monitored effectively by [13, 17]. Furthermore, the internet of things (IoT) based solar drying system using the v-Groove Hybrid Solar Drier (v-GHSD) was more effective in monitoring the drying behaviour [13, 15]. All the parameters involved in solar drying should be studied to reduce the moisture content of seaweed, improve food quality and quantity. However, the methods Density-Based Spatial Clustering of Applications (DBSCAN), Clustering Large Applications (CLARA), Partitioning About Medoids (PAM) and multiple linear regression were used to find the optimal parameters to increase the production of crops [24].

ML algorithms are used to model complicated problems that humans cannot understand because of their complexity. In addition, these algorithms are useful to detect diseases, predict soil parameters, predict crop yield, and detect species [1, 6].

A study conducted by [26] on fish drying investigated the moisture content using ridge regression in conjunction with eight selection criteria. The most significant factors influencing the moisture content and the interaction terms were investigated. From the results, the important drying parameters can be predicted from the moisture content of fish. Research by [27] on the drying parameters that determined the moisture content removal of seaweed was investigated. From the results, bagging performed better than boosting in determining the drying parameters of the seaweed, but heterogeneity was not considered.

Big data analysis comes with many challenges, such as outliers, and multicollinearity. Many studies have been conducted on how to handle these problems. Another problem facing big data is heterogeneity and there is insufficient knowledge about heterogeneity, especially in the field of agriculture using seaweed big data. In addition, the data obtained in big data has varied sources and some are structured and unstructured [28]. All these complexities make the data complicated to analyse. Heterogeneity refers to variation in the data. This variability needs to be investigated to avoid wrong results and inferences.

Heterogeneity is a problem in the field of agriculture. For example, [29] found that there is substantial heterogeneity driving the forces of the rice ecosystem. The results showed that the adoption of each management method has heterogeneity. According to [30], heterogeneity was based on the spatial characteristics and behaviour of the participants, which influenced decision making. In the study of hydrological response to heterogeneity using a variable infiltration capacity model by [31], accounting for heterogeneity in land use gives better responses to hydrology and evapotranspiration. A study on the effects of ignoring heterogeneity showed that ignoring heterogeneity results in overestimation of the technical efficiency and underestimation of the parameters of the models [32]. The study on farmland heterogeneity revealed that under different ecosystem services (ES). The changes in heterogeneity are not the same, there is a need for improvement in the ES to understand the market, especially for pest regulation and crop production [33]. According to [34], the effect of temperature on yield was a significant heterogeneity and it was an eye opener for adaptation between cooler and warmer counties. The study on bird diversity by [35] revealed that the community is affected by cropland heterogeneity and cropland size.

Additionally, there is little research on the parameters influencing the moisture content removal of seaweed. Even in the literature found, few researchers have worked on seaweed big data. Also, few studies considered the interaction terms in seaweed drying. There is no study that compared the outliers before and after heterogeneity. Finally, there is no study on heterogeneity using big data in agriculture, especially on the moisture content removal of seaweed.

A lot of studies have been done on outliers and multicollinearity, but not on heterogeneity. In fact, we do not find any literature in the agricultural field that addresses heterogeneity using drying parameters. Hence, this study focuses on how to detect the heterogeneity of drying parameters and develop hybrid models to determine the significant parameters of the moisture content removal of seaweed. Interaction effects up to the second order for the seaweed big data are incorporated into the model. In addition, hybrid models using seven supervised ML algorithms with robust estimation are utilised to determine the significant parameters that determine the moisture content removal of the seaweed and reduce the number of outliers. The accuracy of the ML algorithms is also investigated via evaluation metrics. Finally, the impact of the errors is also compared before and after heterogeneity.

Materials and methods

Seven supervised machine learning algorithms such as ridge, random forest, support vector machine, bagging, boosting, LASSO and elastic net will be used to determine the significant parameters for the moisture content removal of the seaweed before and after heterogeneity. In addition, robust methods are utilised for the development of the hybrid models. The flowchart in Fig. 3 states the procedure and methodology used in this research.

Data description

The data are collected from 8th April 2021 to 12th April 2021, between the hours of 8:00 am to 5:00 pm during the drying of seaweed by using v-Groove Hybrid Solar Drier (v-GHSD) at Semporna, South-Eastern Coast of Sabah, Malaysia. Some of the parameters are temperature, relative humidity ambient, relative humidity chamber, and solar radiation. Table 1 shows the 29 main parameters, and each parameter has 1914 observations in this study, which is equivalent to 536,870,912 equations. Each observation area is evaluated as a parameter and the region is considered to simplify the system. This is not feasible to deal with because of the time and complexity. The addition of the second order interaction to the main 29 seaweed drying parameters increased all the parameters to 435. Optimization by selecting the first 15, 25, 35 and 45 high-ranking important variables is performed.

Phase I

This involves the addition of all possible models up to second order and testing of assumptions. According to [15], the total number of models can be calculated by using Eq. 1.

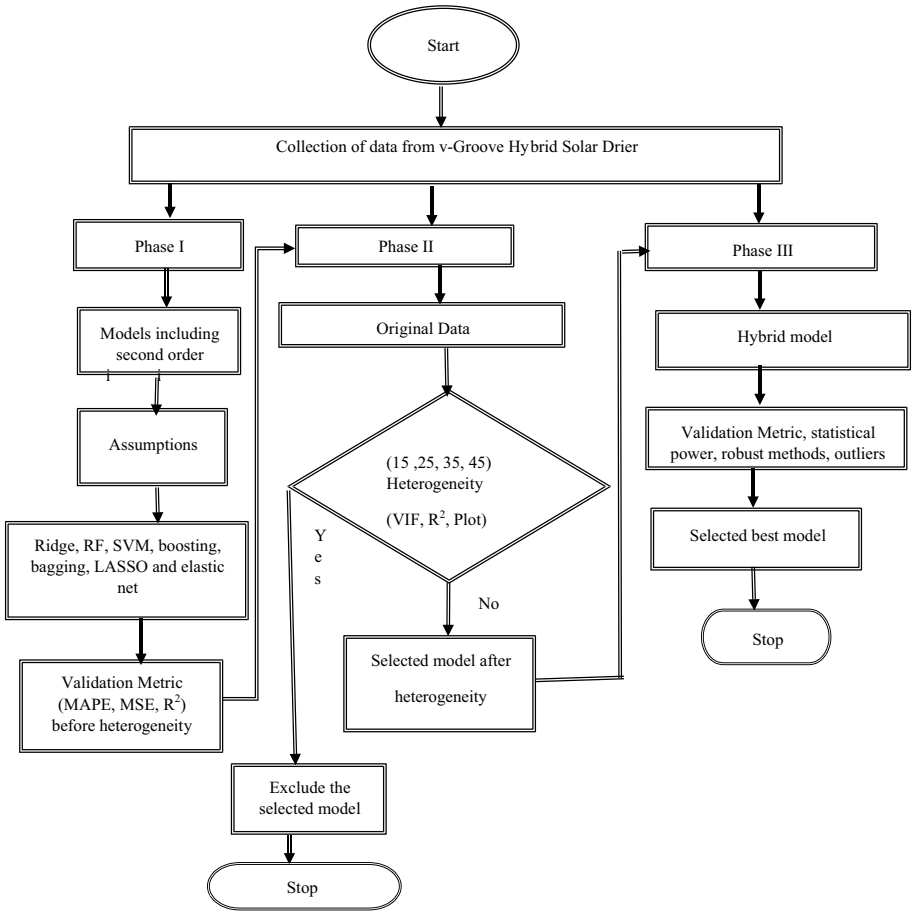

Fig. 3 Flowchart for the study

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)
T7, T14, T15, T16, T21, T22	Independent	Temperature (°C) for the solar collector
T9, T10, T11, T12	Independent	Temperature (°C) behind the inside chamber
T13, T17, T19	Independent	Temperature (°C) in the front of (inside chamber)
T23, T25, T26, T27, T28, 29	Independent	Temperature (°C) from the solar collector to the chamber

$$N = \sum_{j=1}^{k} j \begin{pmatrix} k \\ c \\ j \end{pmatrix} \tag{1}$$

where N represents number of possible models, k is the total number of explanatory variables and $j = 1, 2, 3, \ldots, k$. The assumptions of linearity, errors, observations, independent variables, and heterogeneity are checked in the R programming language. Then ridge, random forest (RF), support vector machine (SVM), boosting, bagging, ridge, LASSO and elastic net are used to select the significant parameters that determine the moisture content removal. The 15, 25, 35 and 45 parameters are selected because features selection can only provide the rank of important variables and does not tell us the number of significant factors [36]. Next, the validation metrics are computed using mean absolute percentage error (MAPE), mean squared error MSE and coefficient of determination (R^2).

Phase II

Next, the computation of VIF is done with *vif* from the *car* library in R using the original data. This gives the range of the values for the variances before we compute the R-squared and 90% confidence interval. If the model has a value that falls below the maximum R-squared, then it exhibits heterogeneity. The models that exhibit heterogeneity are excluded and the models that do not exhibit heterogeneity are included. Then, the ML algorithms in phase I are used to select the 15, 25, 35 and 45 significant parameters.

Phase III

Next, the hybrid models are developed for before and after heterogeneity using robust methods. Data with outliers can be analysed by using robust estimation [37, 38]. The robust methods that are used are M Bi-Square, M Hampel, M Huber, MM and S. Finally, the validation metrics are computed using the 3—sigma limits to identify the number of outliers. The sigma limits are used for quality improvement [41].

The v-Groove Hybrid Solar Drier (v-GHSD)

In this study, v-Groove Hybrid Solar Drier (v-GHSD) was used for drying the seaweed. Solar drier is a used in precision agriculture to dry foods by using solar energy to improve the quality of food and reduce wastage. The v-GHSD drier (Fig. 4) comprises a solar panel, a v-aluminium roof, a drying chamber solar collector, and sensors using the

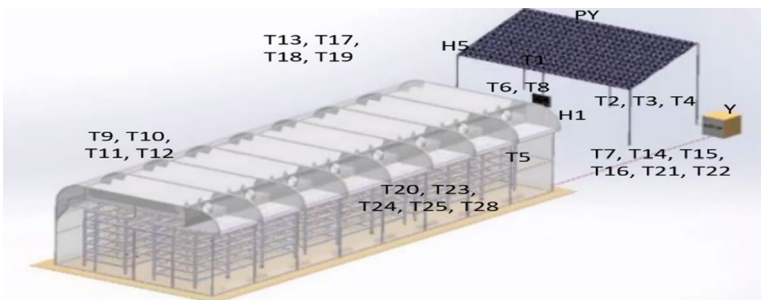

Fig. 4 v-Groove Hybrid Solar Drier (v-GHSD)

internet of things to retrieve data. All the parameters are to receive data from different locations of the drying drier. The sensors are positioned to measure the data for temperature, solar radiation, relative humidity, and moisture content. IoT cloud database was used to understand the performance and the interaction of drying parameters during identified drying period and then, the data are stored in cloud database for every second and later converted to thirty minute intervals for performing analysis and identifying heterogeneity parameters and reduce the multicollinearity and outliers, using the proposed model to determine the moisture content removal.

Heterogeneity identification

Heterogeneity refers to variability of observations. This variability leads to inconsistent estimates and distort conclusion [42]. Suppose we have this multiple linear regression (MLR)

$$Y = \beta_0 + T_1\beta_1 + T_2\beta_2 + \cdots + a_j + \varepsilon \tag{2}$$

where Y is the moisture content, estimates $\beta\prime s$ are the regression coefficients, $T\prime s$ are the drying parameter, a_j denote heterogeneity, that is, the parameters that exhibit heterogeneity and ε is the random error. In Eq. 2, a common problem is the issue of multicollinearity, and this happens when many variables that are correlated and significant not only with dependent variable, but also with each other. Our interest in this equation is a_j. In Eq. (2), if we estimate the regression equation and omit a crucial variable, then the estimate of β will be biased and inconsistent. According to [43], the variance inflation factor in multiple regression is used to quantify the level of severity. It can be computed with

$$VIF_l = \frac{1}{1 - R_l^2} \tag{3}$$

Which means that $R^2 = 1 - \frac{1}{VIF}$.

If the R^2 satisfied certain conditions, then the parameter is said to exhibit heterogeneity.

Evaluation metric

The suitability and accuracy of the models were evaluated using the mean absolute percentage error (MAPE), mean squared error (MSE) and coefficient of determination (R^2). The metrics are stated in Table 2, where y_i is the actual value and $\bar{y}$ is the mean of the actual value and $\widehat{y}_i$ is the forecast value.

Table 2 Evaluation metric

Metrics	Equations	Description
MAPE	$\frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\widehat{y}_i}{y_i}\right\rvert$	It is widely used because it is easy to interpret and due to its scale-independency [44].
MSE	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2$	This is good for given weights to outliers that need to be identified [45].
R^2	$1-\frac{\sum(y_i-\widehat{y}_i)^2}{\sum(y_i-\bar{y})^2}$	This gives the proportion of variance in the dependent variable which can be predicted from the independent variables. R^2 lies between 0 and 1 [45, 46].

Statistical power for percentage change and absolute change

Statistical power is the probability of a test to reject a false null hypothesis. Statistical power $= P(reject\ H_1|H_1\ is\ false)$ where H_1 is the null hypothesis. For a t-test, the equation becomes $P(|t|>t_{\alpha/2}) = P(P_t<\alpha)$ where $t_{\alpha/2}$ represents t-value under the level of significance α and P_t is the t-test p-value.

$$\text{Percentage change } P_c = \frac{B_{MAPE} - A_{MAPE}}{B_{MAPE}} \times 100 \tag{5}$$

$$\text{Absolute change } A_c = |B_{MAPE} - A_{MAPE}| \tag{6}$$

where B_{MAPE} and A_{MAPE} are the MAPE before and after heterogeneity.

To know the best indicator to use between percentage and absolute change, the statistical power must be compared [47]. Statistical power was compared through simulation [48]. According to [49], absolute change was used to study the weight change. Absolute change was used to investigate change in obesity by [50]. Percentage change was used to study change in loss of fat by [51]. A test statistic that compared the maximum likelihood of an absolute change to a percentage change was developed by [52]. According to [53], the percentage change is not affected by the unit of measurement, but the paper did not explain how to choose between absolute and percentage change.

For the evaluation, if $R = \frac{Statistical\ power\ of\ absolute\ change}{Statistical\ power\ of\ percentage\ change} > 1$ [47], then absolute change has a better statistical power than percentage change, then we choose absolute change, otherwise, we choose percentage change.

Results and discussion

In this research, the assumptions of linear regression are verified to understand the data. The heterogeneity parameters among the seaweed drying parameters are identified. To determine the significant factors that determine the moisture content removal of the seaweed, seven popular supervised machine learning algorithms such as ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net are utilized. Furthermore, metric validations were conducted, and hybrid models were developed.

The variability of the 29 main parameters is shown in Fig. 5. Each box-plot represents each drying parameter for the seaweed and helps to understand the heterogeneity among the main parameters. The points outside the box-plot are the outliers. A box-plot uses the 5-number summary of Q1, Q2, Q3, minimum and maximum value to summarise the data. The assumptions of linearity between the dependent and independent variables are checked. No linear relationship exists between them. The assumption of no multicollinearity among the independent variables is not satisfied. The values of the variance inflation factor (VIF) are high, the highest value of the VIF was 75,337.29. It shows the high level of multicollinearity. The assumption that the observations are independent is also checked using the Durbin Watson Test. From the results we obtained, the p-value of 0 is less than the significance level $\alpha = 0.05$, which shows that the residuals are autocorrelated. It means that the observations are not independent. In addition, the normality assumption is also checked with the Kolmogorov–Smirnov test. The the p-value = 2.2e−16 which is less than 0.05 means we have enough evidence to say that the

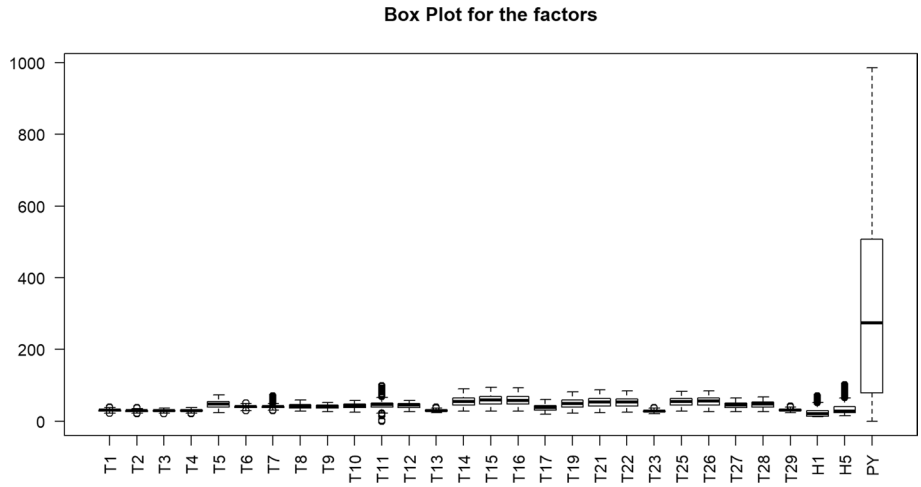

Fig. 5 Box-plot for the seaweed drying parameters

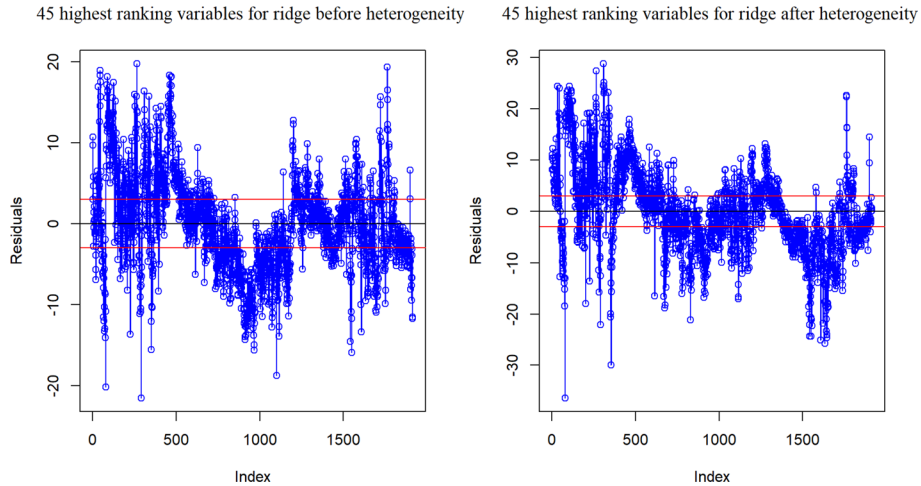

Fig. 6 Comparison between the standardized residuals for 45 highest ranking variables for ridge before and after heterogeneity

residuals do not come from a normal distribution. Figures 6, 7, 8, 9, 10, 11 and 12 show the standardised residual plots for the ridge, RF, SVM, bagging, boosting, LASSO and elastic net for before and after heterogeneity.

Based on these results in Table 3, the parameters T7, T11, H5, T6, T8, H1, and PY exhibit heterogeneity. This is also evident in Fig. 5. After removing the seven parameters that exhibit heterogeneity and including the second order interaction, there are 253 parameters that determine the moisture content removal of the seaweed. The selection of important features was used by [54, 55]. The summary of the assessment results for the ML models is stated in Table 4. However, before the heterogeneity parameters are removed, all validation model measures reveal that random forest outperforms other models in predicting the significant parameters. In addition, evaluation measures with MAPE (2.125891), MSE (7.330011) and R-squared (0.9732063), indicate that significantly

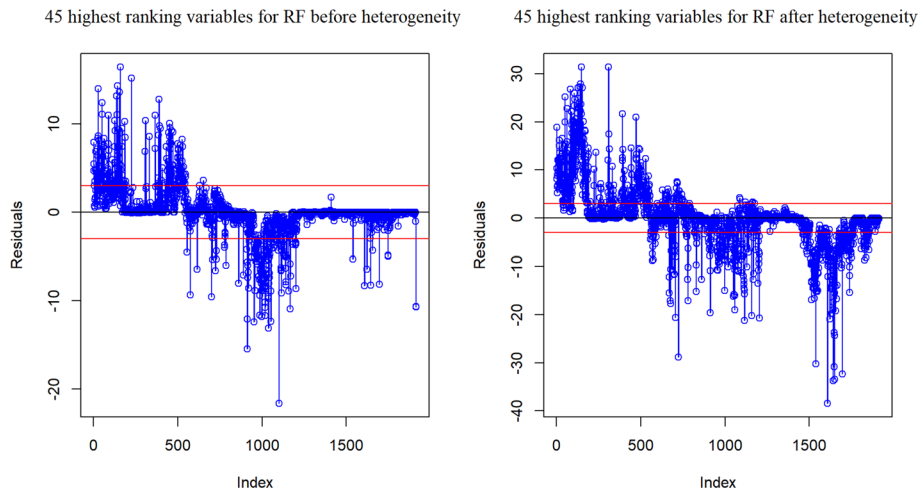

Fig. 7 Comparison between the standardized residuals for 45 highest ranking variables for random forest before and after heterogeneity

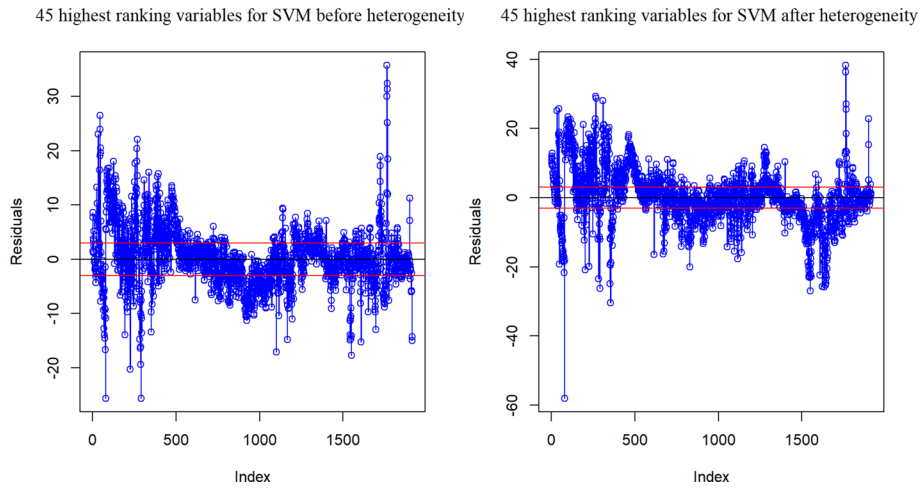

Fig. 8 Comparison between the standardized residuals for 45 highest ranking variables for support vector machine before and after heterogeneity

better results are obtained by random forest for the 45 highest important variables when compared to the 45 highest important variables for other models for significant parameters that determine the moisture content removal. After the heterogeneity parameters are removed, all validation model measures also reveal that random forest outperforms ridge, support vector machine, bagging, boosting, LASSO, and elastic net in predicting the significant parameters that determine the moisture content removal of the seaweed.

In addition, evaluation measures with MAPE (7.588079), MSE (44.39000) and R-squared (0.8377405) indicate that significantly better results are obtained by random forest for the 45 highest important variables when compared to the 45 highest important variables for ridge, support vector machine, bagging, boosting, LASSO, and elastic net significant parameters that determine the moisture content removal. Since the random forest algorithm performed better than the other methods based

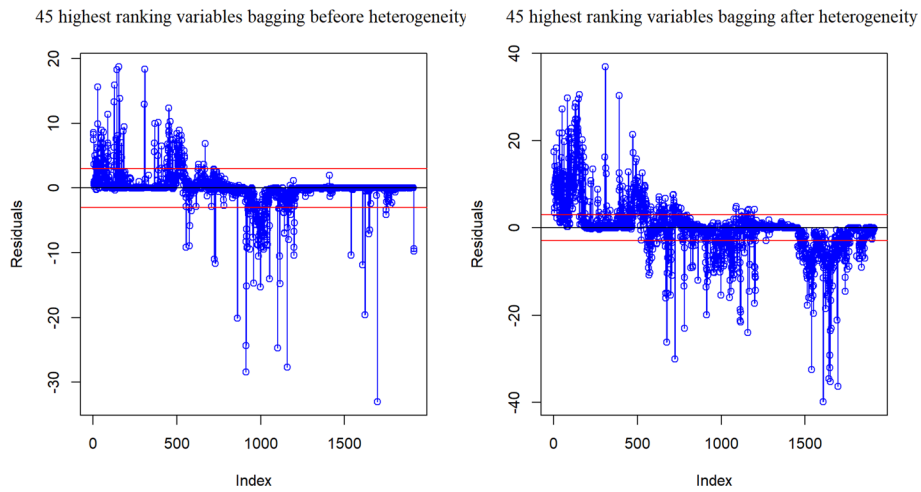

Fig. 9 Comparison between the standardized residuals for 45 highest ranking variables for bagging before and after heterogeneity

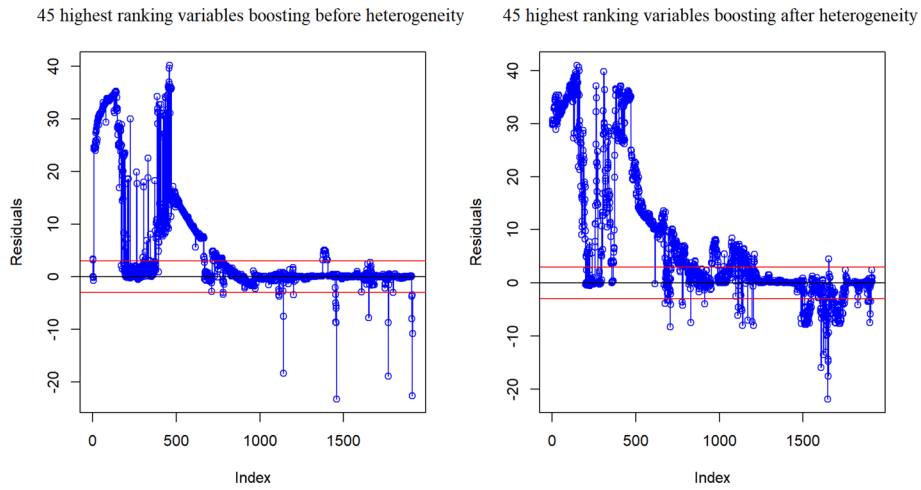

Fig. 10 Comparison between the standardized residuals for 45 highest ranking variables for boosting before and after heterogeneity

on the results of the metrics, the 15, 25, 35 and 45 highest important variables for random forest are the most important parameters that accurately forecast the moisture content removal of the seaweed. This also confirms the results of [27, 54, 59, 60] where random forest absolutely performed better than the other methods. All the values for MAPE random forest are less than 10. It is sufficient to say that this is a high prediction accuracy for the predictive model. This is in line with [61] which claims that if MAPE value is less than 10, it is a high prediction accuracy.

By comparing the metric validation for after and before heterogeneity parameters are removed, generally for ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net in Table 3, the MAPE and MSE after the heterogeneity parameters are removed are higher than the values of MAPE and MSE when the

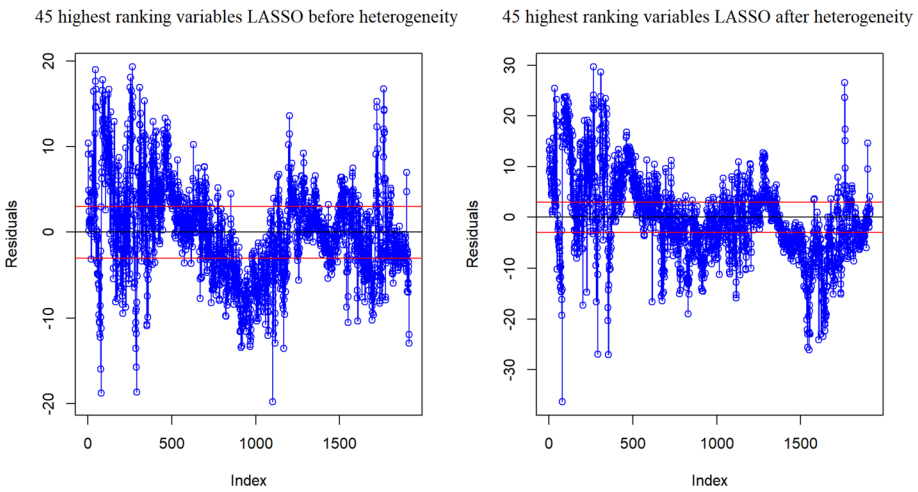

Fig. 11 Comparison between the standardized residuals for 45 highest ranking variables for LASSO before and after heterogeneity

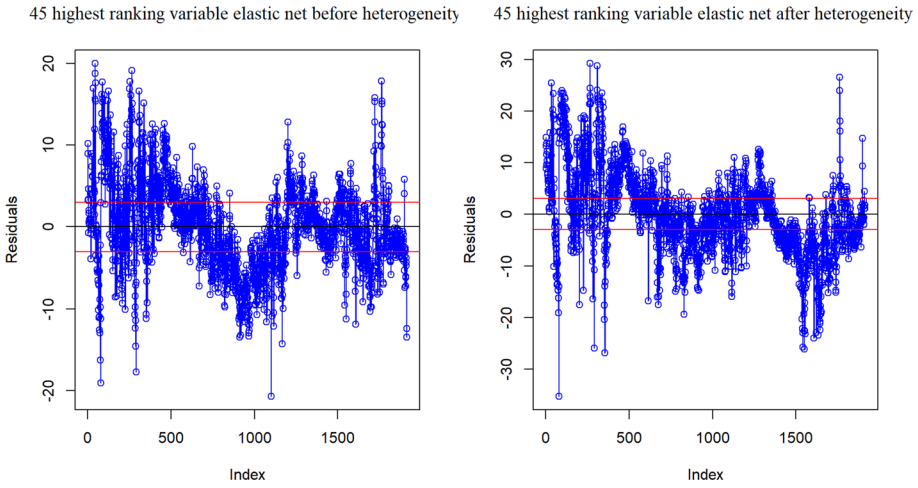

Fig. 12 Comparison between the standardized residuals for 45 highest ranking variables for elastic net before and after heterogeneity

Table 3 Heterogeneity parameters

Lowest VIF	Highest VIF	Lowest R squared	Highest R-squared	90% CI	Heterogeneity parameters
3.067297	75,337.29	0.67398	0.999987	[0.786375, 0.8875918]	T7, T11, H5, T6, T8, H1, PY

heterogeneity parameters have not been removed in the model. Also, the R-squared values after heterogeneity parameters are removed are lower than the R-squared before heterogeneity is removed. The results have shown that the removal of some variables can reduce the accuracy of the model.

The heterogeneity parameters that were removed did not increase the accuracy of the model. According to [62], if an MAPE validation is equal or less after the removal of a

Table 4 Determination of optimal machine learning models before and after heterogeneity

ML models	High ranking variables	Metric validations before heterogeneity			Metric validations after heterogeneity			Absolute change	Percentage change
		MAPE	MSE	R^2	MAPE	MSE	R^2		
Ridge (control)	15	14.64603	83.92337	0.6932309	13.45105	82.84078	0.6971881	1.194980	8.159071
	25	11.50656	56.4660	0.7935971	12.63606	75.31612	0.7246934	1.129500	− 9.81614
	35	10.0306	48.23541	0.8236828	12.00500	70.29434	0.7430497	1.974400	− 19.68380
	45	9.657189	44.48745	0.8373829	11.95927	69.44397	0.7461581	2.302081	− 23.83800
Random forest	15	2.458969	9.910512	0.9637737	9.885843	67.35215	0.7538052	7.426874	− 302.03200
	25	2.337353	9.010273	0.9670644	7.909333	47.21578	0.8274099	5.571980	− 238.38800
	35	2.174667	7.790909	0.9715216	7.663343	45.15805	0.8349317	5.488676	− 252.39200
	45	2.125891	7.330011	0.9732063	7.588079	44.39000	0.8377405	5.462188	− 256.93600
Support vector machine	15	8.614626	45.25618	0.8347612	11.77207	77.48160	0.7169731	3.157444	− 36.65210
	25	7.980399	35.80985	0.8691446	11.15354	71.12697	0.7401082	3.173141	− 39.76170
	35	7.568951	34.00095	0.8757802	10.89938	68.85807	0.7484105	3.330429	− 44.00120
	45	7.351331	32.38644	0.8816661	10.62685	66.33326	0.7575719	3.275519	− 44.55680
Bagging	15	12.25897	74.29053	0.7284423	11.30002	66.52011	0.7568458	0.958950	7.822440
	25	9.778194	47.33173	0.8269861	10.62821	57.44370	0.7900233	0.850016	− 8.69298
	35	8.413645	36.41955	0.8668739	9.417039	48.41542	0.8230248	1.003394	− 11.92580
	45	8.151903	33.65611	0.8769752	8.983211	45.01187	0.835466	0.831308	− 10.19770
Boosting	15	8.168942	142.4542	0.5310293	13.09470	217.8164	0.3416015	4.925758	− 60.29860
	25	8.697362	136.3236	0.5543729	13.16813	215.4273	0.346658	4.470768	− 51.40370
	35	8.183671	140.1463	0.5368431	12.78951	208.6947	0.3629861	4.605839	− 56.28080
	45	8.203304	134.0864	0.5569358	8.228835	135.3237	0.5510545	0.025531	− 0.311230
LASSO	15	14.39656	101.8853	0.6275736	12.27376	74.04000	0.7293580	2.122800	14.74519
	25	10.82264	52.90467	0.806615	11.65852	67.91064	0.751763	0.835880	− 7.72344
	35	8.977735	37.69348	0.8622172	11.60206	67.40559	0.7536091	2.624325	− 29.23150
	45	8.149872	31.57626	0.8845778	11.52189	66.86088	0.7556002	3.372018	− 41.37510
Elastic Net	15	13.12778	78.47416	0.7131497	12.31004	73.09066	0.7328282	0.817740	6.22908
	25	9.485387	41.90456	0.8468243	11.72084	68.13366	0.7509478	2.235453	− 23.56730
	35	9.051548	37.81546	0.8617713	11.64224	67.48376	0.7533234	2.590692	− 28.62150
	45	8.191381	32.53884	0.8810592	11.66734	67.30154	0.7539895	3.475959	− 42.43430

parameter, it does not mean that the parameter has no effect on the response variable. It means that the variability level in the data was not enough to be explained by the model.

The percentage change for ridge 15, bagging 15, LASSO 15 and elastic net is positive. This represents 14.3% of the total number of models and the few cases where MAPE before heterogeneity is higher than MAPE after heterogeneity. The percentage change of 24 models is negative, which means that the MAPE before heterogeneity is lower than the MAPE after heterogeneity. This represents 85.7% of the total number of models. Random forest 15, 25, 35 and 45 models have the highest negative percentage change compared to other models.

In summary, through the validation metrics, the ability of ridge, random forest, support vector machine, bagging, boosting, elastic net, and LASSO is evaluated to accomplish more substantial and significant conclusions. The results are shown in Table 4 for all models. It is observed that random forest shows higher accuracy than other models models. This proves the superiority of random forest before and after heterogeneity over the other models and it leads to higher accuracy with the lowest errors. According to [54] the number of parameters is crucial because it will reduce the training time and avoid the curse of dimensionality.

Table 5 Comparison of statistical power

	Absolute change	Percentage change	Remarks
Test statistic	8.0924	− 3.4367	$\frac{8.0924}{-3.4367} < 1$ Percentage change will be used since the ratio is less than 1
P-value	1.078e−08	0.001921	
Df	27	27	

The comparison of the statistical power is shown in Table 5. The ratio of the test statistic for absolute change to percentage change is less than 1. This shows that percentage change has better statistical power than absolute change to explain the results and draw valid conclusions.

Table 6 shows the results of the hybrid model and the original model before and after heterogeneity for 45 high-ranking variables. The 3-sigma limits are also provided to identify the number of outliers and make comparisons. For the ridge before heterogeneity, the best robust estimator is M Hampel with 16 outliers, while the original has 23 outliers.

For the random forest before heterogeneity, the best robust estimator is M Hampel with 19 outliers, while the original has 45 outliers. For the support vector machine before heterogeneity, the best robust estimator is M Hampel with 23 outliers and the original has 24 outliers. For the elastic net before heterogeneity, the best robust estimator is M Hampel and M Huber with 33 outliers, while the original has 29 outliers. With these results. For before heterogeneity, M Hampel robust estimation performs better than M Bi-Square, M Huber, MM and S.

For the ridge after heterogeneity, the best robust estimators are M Bi-Square and MM with 22 outliers, while the original has 29 outliers. For the random forest after heterogeneity, the best robust estimator is M Hampel with 29 outliers, while the original has 41 outliers. For the support vector machine after heterogeneity, the best robust estimator is M Bi-Square with 27 outliers, while the original has 24 outliers. For the bagging after heterogeneity, the best robust estimator is M Hampel with 21 outliers, while the original has 28 outliers. For the elastic net after heterogeneity, the best robust estimator is M Hampel and M Huber with 23 outliers, while the original has 33 outliers. With these results. For after heterogeneity, the ridge performs better with M Bi-Square and MM. Random forest, bagging and boosting perform better with M Hampel. Support vector machine and LASSO perform better with M Bi-Square. The elastic net performs better with M Hampel and M Huber.

Generally, the outliers using the 3-sigma limits for before and after heterogeneity indicate that for the original model, the number of outliers increases from before heterogeneity to after heterogeneity for ridge, LASSO, and elastic net. It is constant for support vector machine. It decreases for random forest, bagging and boosting.

Conclusions and future work

The heterogeneity parameters are identified, and hybrid models were developed to forecast the significant drying parameters that determine the moisture content removal of the seaweed after drying. Seven predictive models, such as ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net are used for determining the

Table 6 Comparison between the number and percentage of outliers outside the 3-sigma limits for the original and hybrid models for 45 high-ranking variables

ML models	Robust method	$\mu \pm 3\sigma$(%)		Remarks
		Before heterogeneity	After heterogeneity	
Ridge (control)	Original	23(1.20)	29(1.52)	Increase
	M Bi-Square	25(1.31)	22(1.15)	Decrease
	M Hampel	16(0.84)	24(1.25)	Increase
	M Huber	35(1.83)	28(1.46)	Decrease
	MM	63(3.29)	22(1.15)	Decrease
	S	33(1.72)	32(1.67)	Decrease
Random forest	Original	45(2.35)	41(2.14)	Decrease
	M Bi-Square	26(1.36)	34(1.78)	Increase
	M Hampel	19(0.99)	29(1.52)	Increase
	M Huber	25(1.31)	33(1.72)	Increase
	MM	85(4.44)	48(2.51)	Decrease
	S	75(3.92)	30(1.57)	Decrease
Support vector machine	Original	24(1.25)	24(1.25)	Constant
	M Bi-Square	27(1.41)	27(1.41)	Constant
	M Hampel	23(1.20)	28(1.46)	Increase
	M Huber	27(1.41)	28(1.46)	Increase
	MM	80(4.18)	68(3.55)	Decrease
	S	82(4.28)	36(1.88)	Decrease
Bagging	Original	31(1.62)	28(1.56)	Decrease
	M Bi-Square	34(1.78)	28(1.46)	Decrease
	M Hampel	28(1.46)	21(1.10)	Decrease
	M Huber	30(1.57)	26(1.36)	Decrease
	MM	79(4.13)	31(1.62)	Decrease
	S	75(3.92)	29(1.52)	Decrease
Boosting	Original	15(0.78)	14(0.73)	Decrease
	M Bi-Square	33(1.72)	25(1.31)	Decrease
	M Hampel	29(1.52)	19(0.99)	Decrease
	M Huber	31(1.62)	25(1.31)	Decrease
	MM	94(4.91)	26(1.36)	Decrease
	S	81(4.23)	31(1.62)	Decrease
LASSO	Original	26(1.36)	35(1.83)	Increase
	M Bi-Square	37(1.93)	42(2.19)	Increase
	M Hampel	38(1.99)	27(1.41)	Decrease
	M Huber	41(2.14)	33(1.72)	Decrease
	MM	73(3.81)	43(2.25)	Decrease
	S	54(2.82)	33(1.72)	Decrease
Elastic Net	Original	29(1.52)	33(1.72)	Increase
	M Bi-Square	36(1.88)	28(1.46)	Decrease
	M Hampel	33(1.72)	23(1.20)	Decrease
	M Huber	33(1.72)	26(1.36)	Decrease
	MM	74(3.87)	30(2.25)	Decrease
	S	51(2.67)	26(1.36)	Decrease

significant parameters in conjunction with robust methods. These hybrid models are useful for determining the significant parameters that determine the moisture content removal of the seaweed. For before heterogeneity, the hybrid model random forest M

Hampel with 19 outliers is the best, because it performs better when compared to other models. For after heterogeneity, the hybrid model boosting M Hampel with 19 outliers is the best, because it performs better when compared to other models.

For future studies, the traditional statistical methods and machine learning models for predicting the moisture content removal of seaweed can be compared. The number of selected drying parameters can be increased or all the parameters with interaction can be used. Other robust estimators such as least trimmed squares (LTS), least absolute deviation (LAD) and least median of squares (LMS) estimators can be used to develop a hybrid model.

Abbreviations

PF	Precision farming
v-GHSD	V-Groove Hybrid Solar Drier
LASSO	Least absolute shrinkage and selection operator
MAPE	Mean absolute percentage error
MSE	Mean squared error
SSE	Sum of squared error
R-squared	Coefficient of determination
ML	Machine learning
SFTs	Smart farming technologies
IoT	Internet of things
GPS	Global positioning system
RF	Random forest
SVM	Support vector machine
VIF	Variance inflation factor

Acknowledgements
The authors are grateful to the "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2022/STG06/USM/02/13" for their support in this research.

Author contributions
OJI: He conducted the analysis and manuscript development. FPS: She designed the experiment, contributed to the results, discussion, and supervision. JS: He contributed to the writing, logic and editing. MKMA: He designed the experiment, contributed to the manuscript writing and supervision of the work. All the authors approved the final manuscript.

Funding
The authors are grateful to the "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2022/STG06/USM/02/13" for their financial support.

Availability of data and materials
Data is available on request. Materials and methodologies are described in this paper.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 12 November 2022 Accepted: 7 August 2023
Published online: 19 August 2023

References

1. Durai SKS, Shamili MD. Smart farming using machine learning and deep learning techniques. Decis Anal J. 2022;3:100041.
2. Moysiadis V, Sarigiannidis P, Vitsas V, Khelifi A. Smart Farming in Europe. Computer Science Review, 2021;39. https://doi.org/10.1016/j.cosrev.2020.100345.

3. Klerkx L, Jakku E, Labarthe P. A review of social science on digital agriculture, smart farming and agriculture 4.0: new contributions and a future research agenda. NJAS Wageningen J Life Sci. 2019;90–91. https://doi.org/10.1016/j.njas.2019.100315.
4. Rose DC, Chilvers J. Agriculture 4.0: broadening responsible innovation in an era of smart farming. Front Sustain Food Syst. 2018. https://doi.org/10.3389/fsufs.2018.00087.
5. Balafoutis AT, van Evert FK, Fountas S. Smart farming technology trends: Economic and environmental effects, labor impact, and adoption readiness. Agronomy. 2020;10(5). https://doi.org/10.3390/agronomy10050743.
6. Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: a comprehensive review. IEEE Access. 2021;9:4843–73.
7. National Oceanic and Atmospheric Administration. What is seaweed? National Ocean Service. 2017. https://oceanservice.noaa.gov/facts/seaweed.html#:~:text=%22Seaweed%22%20is%20the%20common%20name,Marine%20Sanctuary%20and%20National%20Park.
8. Guiry MD. What are seaweeds? The Seaweed Site. 2014. https://www.seaweed.ie/algae/seaweeds.php.
9. Suwati S, Romansyah E, Syarifudin S, Jani Y, Purnomo AH, Damat D, et al. Comparison between natural and cabinet drying on weight loss of seaweed Euchuema cottonii Weber-van Bosse. Sarhad J Agric. 2021;37(SpecialIssue 1):1–8.
10. Buschmann AH, Camus C, Infante J, Neori A, Israel Á, Hernández-González MC, et al. Seaweed production: overview of the global state of exploitation, farming and emerging research activity. Eur J Phycol. 2017;52(4):391–406.
11. Pradana GB, Prabowo KB, Hastuti RP, Djaeni M, Prasetyaningrum A. Seaweed drying process using tray dryer with dehumidified air system to increase efficiency of energy and quality product. IOP Conf Ser Earth Environ Sci. 2019. https://doi.org/10.1088/1755-1315/292/1/012070.
12. Ali MKM, Sulaiman J, Md Yasir S, Ruslan M. Cubic spline as a powerful tools for processing experimental drying rate data of seaweed using solar drier. Malay J Math Sci. 2017;11:159–72.
13. van Oirschot R, Thomas JBE, Gröndahl F, Fortuin KPJ, Brandenburg W, Potting J. Explorative environmental life cycle assessment for system design of seaweed cultivation and drying. Algal Res. 2017;1(27):43–54.
14. Xiao HW, Mujumdar AS. Importance of drying in support of human welfare. Drying Technol. 2020;38(12):1542–3.
15. Suherman S, Djaeni M, Kumoro AC, Prabowo RA, Rahayu S, Khasanah S. Comparison drying behavior of seaweed in solar, sun and oven tray dryers. MATEC Web Conf. 2018. https://doi.org/10.1051/matecconf/201815605007.
16. Ali MKM, Fudholi A, Sulaiman J, Muthuvalu MS, Ruslan MH, Yasir SMd, et al. Post-harvest handling of eucheumatoid seaweeds. In: Tropical seaweed farming trends, problems and opportunities. Springer International Publishing, Cham; 2017. p. 131–45.
17. Ali MKM, Sulaiman J, Md Yasir S, Ruslan M. Cubic Spline as a Powerful Tools for Processing Experimental Drying Rate Data of Seaweed Using Solar Drier. Malaysian Journal of Mathematical Sciences, 2017;11:159–172.
18. Nimnuan P, Nabnean S. Experimental and simulated investigations of the performance of the solar greenhouse dryer for drying cassumunar ginger (Zingiber cassumunar Roxb.). Case Stud Thermal Eng. 2020;22. https://doi.org/10.1016/j.csite.2020.100745.
19. Lakshmi DVN, Muthukumar P, Layek A, Nayak PK. Drying kinetics and quality analysis of black turmeric (Curcuma caesia) drying in a mixed mode forced convection solar dryer integrated with thermal energy storage. Renew Energy. 2018;120. https://doi.org/10.1016/j.renene.2017.12.053.
20. Pankaew P, Aumporn O, Janjai S, Pattarapanitchai S, Sangsan M, Bala BK. Performance of a large-scale greenhouse solar dryer integrated with phase change material thermal storage system for drying of chili. Int J Green Energy. 2020;17(11). https://doi.org/10.1080/15435075.2020.1779074.
21. Vijayan S, Arjunan TV, Kumar A. Exergo-environmental analysis of an indirect forced convection solar dryer for drying bitter gourd slices. Renew Energy. 2020;146. https://doi.org/10.1016/j.renene.2019.08.066.
22. Hao W, Liu S, Mi B, Lai Y. Mathematical modeling and performance analysis of a new hybrid solar dryer of lemon slices for controlling drying temperature. Energies (Basel). 2020;13(2). https://doi.org/10.3390/en13020350.
23. Nabnean S, Nimnuan P. Experimental performance of direct forced convection household solar dryer for drying banana. Case Stud Thermal Eng. 2020;22. https://doi.org/10.1016/j.csite.2020.100787.
24. Majumdar J, Naraseeyappa S, Ankalaki S. Analysis of agriculture data using data mining techniques: application of big data. J Big Data. 2017;4(1). https://doi.org/10.1186/s40537-017-0077-4.
25. Ali MKM, Critchley AT, Hurtado AQ. The impacts of AMPEP K+ (Ascophyllum marine plant extract, enhanced with potassium) on the growth rate, carrageenan quality, and percentage incidence of the damaging epiphyte Neosiphonia apiculata on four strains of the commercially important carrageenophyte Kappaphycus, as developed by micropropagation techniques. J Appl Phycol. 2020;32(3). https://doi.org/10.1007/s10811-020-02117-0.
26. Lim HY, Fam PS, Javaid A, Ali MKM. Ridge regression as efficient model selection and forecasting of fish drying using v-groove hybrid solar drier. Pertanika J Sci Technol. 2020;28(4):1179–202.
27. Majahar Ali MKM, Tahir Ismail M, Hamundu FM, Akhtar NA, et al. Hybrid model in machine learning–robust regression applied for sustainability agriculture and food security. Int J Electric Comput Eng. 2022;12(4):4457–68.
28. El-Din AMG, Senousy MB. A Solution for Handling Big Data Heterogeneity Problem. In: Lecture Notes in Networks and Systems. Springer, Singapore. 2022;224.https://doi.org/10.1007/978-981-16-2275-5_11.
29. Gouraram P, Goyari P, Paltasingh KR. Rice ecosystem heterogeneity and determinants of climate risk adaptation in Indian agriculture: farm-level evidence. J Agribus Dev Emerg Econ. 2022. https://doi.org/10.1108/JADEE-03-2022-0044.
30. Kanchanaroek Y, Aslam U. Policy schemes for the transition to sustainable agriculture—farmer preferences and spatial heterogeneity in northern Thailand. Land Use Policy. 2018;1(78):227–35.
31. Srivastava A, Kumari N, Maza M. Hydrological response to agricultural land use heterogeneity using variable infiltration capacity model. Water Resour Manage. 2020;34(12):3779–94.
32. Li K, Liu J, Xue Y, Rahman S, Sriboonchitta S. Consequences of ignoring dependent error components and heterogeneity in a stochastic frontier model: an application to rice producers in northern Thailand. Agriculture. 2022;12(8):1078.
33. Botzas-Coluni J, Crockett ETH, Rieb JT, Bennett EM. Farmland heterogeneity is associated with gains in some ecosystem services but also potential trade-offs. Agric Ecosyst Environ. 2021;1:322.

34. Keane M, Neal T. Climate change and U.S. agriculture: accounting for multi-dimensional slope heterogeneity in production functions. Quantitative Economics, 2000;11:1391–1429
35. Liao J, Liao T, He X, Zhang T, Li D, Luo X, et al. The effects of agricultural landscape composition and heterogeneity on bird diversity and community structure in the Chengdu Plain. China Glob Ecol Conserv. 2020;1:24.
36. Drobnič F, Kos A, Pustišek M. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. Electronics (Switzerland). 2020;9(5) https://doi.org/10.3390/electronics9050761.
37. Alma ÖG. Comparison of robust regression methods in linear regression. Int J Contemp Math Sci. 2011;6(9):409–21.
38. Javaid A, Ismail MT, Ali MKM. Efficient model selection of collector efficiency in solar dryer using hybrid of LASSO and robust regression. Pertanika J Sci Technol. 2020;28(1):193–210.
39. Mohamed AE, Almongy HM, Mohamed AH. Comparison between M-estimation, S-estimation, and MM estimation methods of robust estimation with application and simulation. Int J Math Arch. 2018;9(11):55.
40. Mukhtar Ali MKM, Javaid A, Ismail MT, Fudholi A. Accurate and hybrid regularization—robust regression model in handling multicollinearity and outlier using 8SC for big data. Math Model Eng Probl. 2021;8(4):547–56.
41. Wijaya IMS, Sari DI. Quality control of optical fiber disruption with big data using the six sigma method. JURTEKSI (J Teknol Sist Inform). 2022;8(2):125–32.
42. Gormley TA, Matsa DA. Common errors: how to (and not to) control for unobserved heterogeneity. Rev Financ Stud. 2014;27(2):617–61.
43. Cheng J, Sun J, Yao K, Xu M, Cao Y. A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta A Mol Biomol Spectrosc. 2022;5:268.
44. Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast. 2016;32(3):669–79.
45. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci. 2021;7:1–24.
46. Gouda SG, Hussein Z, Luo S, Yuan Q. Model selection for accurate daily global solar radiation prediction in China. J Clean Prod. 2019;1(221):132–44.
47. Stridbeck R, Zhang L, Han K. How to analyze change from baseline: absolute or percentage change? D-level Essay in Statistics. 2009;1–18.
48. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol. 2001. https://doi.org/10.1186/1471-2288-1-6.
49. Waleekhachonloet OA, Limwattananon C, Limwattananon S, Gross CR. Group behavior therapy versus individual behavior therapy for healthy dieting and weight control management in overweight and obese women living in rural community. Obes Res Clin Pract. 2007;1(4):223–32.
50. Neovius M, Rössner S. Results from a randomized controlled trial comparing two low-calorie diet formulae. Obes Res Clin Pract. 2007;1(3):165–71.
51. Kim MK, Tanaka K, Kim MJ, Matuso T, Endo T, Tomita T, et al. Comparison of epicardial, abdominal and regional fat compartments in response to weight loss. Nutr Metab Cardiovasc Dis. 2009;19(11). https://doi.org/10.1016/j.numecd.2009.01.010.
52. Kaiser L. Adjusting for baseline: change or percentage change? Stat Med. 1989. https://doi.org/10.1002/sim.4780081002.
53. Törnqvist L, Vartia P, Vartia YO. How should relative changes be measured? Am Stat. 1985;39(1):43–6.
54. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data. 2020;7(1):1–26.
55. Han Y. Stable feature selection: theory and algorithms. State University of New York at Binghamton. 2012.
56. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. ACM computing surveys (CSUR). 2017;50(6):1–45.
57. Gupta C. Feature selection and analysis for standard machine learning classification of audio beehive samples. (Doctoral dissertation, Utah State University). 2019.
58. Ali MKM, Mukhtar, Ismail MT, Ferdinand MH, Alimuddin. Machine learning-based variable selection: An evaluation of Bagging and Boosting. Turk J Comput Math Educ. 2021;12(13):4343–9.
59. Roell GW, Sathish A, Wan N, Cheng Q, Wen Z, Tang YJ, et al. A comparative evaluation of machine learning algorithms for predicting syngas fermentation outcomes. Biochem Eng J. 2022;1:186.
60. Adugna T, Xu W, Fan J. Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. Remote Sens (Basel). 2022;14(3). https://doi.org/10.3390/rs14030574.
61. Sumari ADW, Charlinawati DS, Ariyanto Y. A simple approach using statistical-based machine learning to predict the weapon system operational readiness. In: The 1st International Conference on Data Science and Official Statistics. 2021. p. 343–51.
62. Jimenez-Marquez SA, Thibault J, Lacroix C. Prediction of moisture in cheese of commercial production using neural networks. Int Dairy J. 2005;15(11):1156–74.

Publisher's Note