Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 6 (2024) 2058

Journal of the Nigerian Society of Physical Sciences

Identifying heterogeneity for increasing the prediction accuracy of machine learning models

Paavithashnee Ravi Kumar[a], Majid Khan Majahar Ali[a,*], Olayemi Joshua Ibidoja[a,b]

[a]*School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia*
[b]*Department of Mathematics, Federal University Gusau, Gusau, Nigeria*

Abstract

In recent years, the significance of machine learning in agriculture has surged, particularly in post-harvest monitoring for sustainable aquaculture. Challenges like heterogeneity, irrelevant variables and multicollinearity hinder the implementation of smart monitoring systems. However, this study focuses on investigating heterogeneity among drying parameters that determine the moisture content removal during seaweed drying due to its limited attention, particularly within the field of agriculture. Additionally, a heterogeneity model within machine learning algorithms is proposed to enhance accuracy in predicting seaweed moisture content removal, both before and after the removal of heterogeneity parameters and also after the inclusion of single-eliminated heterogeneity parameters. The dataset consists of 1914 observations with 29 independent variables, but this study narrows down to five: Temperature (T1, T4, T7), Humidity (H5), and Solar Radiation (PY). These variables are interacted up to second-order interactions, resulting in 55 variables. Variance inflation factor and boxplots are employed to identify heterogeneity parameters. Two predictive machine learning models, namely random forest and elastic net are then utilized to identify the 15 and 20 highest important parameters for seaweed moisture content removal. Evaluation metrics (MSE, SSE, MAPE, and R-squared) are used to assess model performance. Results demonstrate that the random forest model outperforms the elastic net model in terms of higher accuracy and lower error, both before and after removing heterogeneity parameters, and even after reintroducing single-eliminated heterogeneity parameters. Notably, the random forest model exhibits higher accuracy before excluding heterogeneity parameters.

DOI:10.46481/jnsps.2024.2058

Keywords: Heterogeneity, Machine learning, Seaweed, Variable selection, Agriculture

Article History :
Received: 31 March 2024
Received in revised form: 11 May 2024
Accepted for publication: 27 May 2024
Published: 16 June 2024

Communicated by: B. J. Falaye

1. Introduction

Machine learning has gained increasing popularity in current technology, being recognized as a branch of artificial intelligence (AI) where AI can comprehend and create systems exhibiting intelligent characteristics [1]. In fact, machine learning is often linked to the concept of big data. Recently, the significance of machine learning has markedly increased in the field of agriculture. Given its pivotal role in our daily lives and the global economy, agriculture has embraced the integration of smart farming practices. These practices heavily rely on machine learning algorithms, utilizing various components such as sensors, drones and remote sensing to address challenges, including the estimation of food production, contributing to the

*Corresponding Author Tel. No.: +6-014-954-3405;
Email address: majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)

promotion of sustainable agricultural practices [2].

Machine learning has been applied by many authors in the field of agriculture. For instance, Ibidoja *et al*. [3], used machine learning algorithms such as boosting, bagging, random forest and support vector machine to determine the significant of drying parameters of seaweed. Furthermore, Arjasakusuma *et al*. [4], applied variable selection using machine learning techniques in their study to estimate forest heights. Lim *et al*. [5], utilized ridge regression to determine the drying parameters of fish.

The consumption of seaweed is increasing due to its beneficial nutritional properties and its potential uses in the food, cosmetics and pharmaceutical industries [6]. Primarily cultivated for carrageenan, seaweed has diverse uses in food, pet food production, and cosmetics [7]. The carrageenan-bearing seaweed sector is based on a few processes, such as cultivation, harvesting, post-harvesting (drying), and marketing. Research by Lomartire *et al*. [8], highlighted the seaweed's significant nutritional value, rich in vitamins, minerals, dietary fibers, proteins, essential amino acids, and polyphenols with antioxidant and anti-inflammatory properties. The use of seaweed in agriculture is especially emphasized since it ensures chemical-free soils and crops, which is beneficial to human health. Consequently, studying seaweed is crucial to discover new bioactive substances that will be beneficial to both human and animal health.

Over the past few decades, the method of preserving food products through drying has gained significant popularity. This practice plays a crucial role as a post-processing technique, serving as an essential means of preserving agricultural crops and marine harvest [5]. Surprisingly, seaweed undergoes the drying process without any negative impact on its quality before it can be utilized for other purposes. The high-water content of fresh seaweed, ranging from 75% to 85%, makes drying a pivotal step that hinders the growth of microorganisms, thereby preventing spoilage and mold formation. Seaweed generally exhibits hydrophilic properties and the presence of hydrophilic surface groups like hydroxyl, carboxyl, and sulphate groups enables effective interaction with water molecules [9]. The solar drier stands out as the optimal method for seaweed drying, facilitating the rapid reduction of water content. Various types of solar driers with exclusive technical performances have been designed and developed worldwide. Among these, the application of Internet of Things (IoT)-based solar drying system employing the v-Groove Hybrid Solar Drier (v-GHSD), as proposed by Ali *et al*. [7], demonstrated enhanced effectiveness in monitoring the drying process [5].

Post-harvest monitoring systems play a vital role in ensuring the sustainability of aquaculture production. Nevertheless, challenges persist in implementing smart monitoring systems for post-harvest management, with issues such as heterogeneity, irrelevant variables, and multicollinearity posing significant obstacles. Heterogeneity emerges as a prominent issue within the field of big data in agriculture. Heterogeneity denotes the extent of variability present in a dataset, referring to the degree to which a system differs from complete uniformity [10]. Various factors contribute to heterogeneity, including differences in parameters and variations in units for temperature, relative humidity, wind, and solar radiation, as well as variability in variances. Examining this variability is essential to prevent inaccuracies in findings and conclusions, as it can result in inconsistent estimations and distort the overall findings [11]. Nunes *et al*. [10], have also discovered that if heterogeneity is not accurately measured, then it is impossible to precisely determine its impact.

In addition to heterogeneity, prediction models can become problematic when incorporating an excessive number of variables, especially if irrelevant ones are included, hence negatively impacting the overall model performance. Obstacles such as multicollinearity also arise in the agricultural sector. Multicollinearity is a statistical phenomenon that occurs when there is a strong linear correlation among two or more independent variables within a dataset [12]. Addressing multicollinearity is a crucial task that should be tackled before initiating the data modelling process since it increases the standard error of coefficients in the model, leading to unstable estimates of parameters in the regression models and a decrease in their precision. In order to tackle the issues of irrelevant variables and multicollinearity, variable selection will be performed using machine learning algorithms. Variable selection is crucial in any statistical research study, especially in the context of big data. As described by Chan *et al*. [12], variable selection helps prevent issues related to multicollinearity, aiming to obtain a more accurate parameter estimate.

While numerous studies have investigated problems related to irrelevant variables and multicollinearity within the field of agriculture, there is a noticeable gap in the literature addressing the issue of heterogeneity, despite its evident presence in real-life agricultural data. Note that limited attention has been given to heterogeneity especially in the context of seaweed drying. For instance, Marenya *et al*. [13], conducted a study in the field of agriculture that addressed the issue of heterogeneity, but the effects of heterogeneity before and after the removal of heterogeneity variables were not thoroughly discussed. Furthermore, no studies have explored the analysis of the inclusion of all single-eliminated heterogeneity parameters back into the model. It is essential to explore the application of big data in agriculture, with a particular focus on addressing heterogeneity within agricultural datasets. Failure to resolve the heterogeneity issue may result in inaccurate findings and flawed scientific conclusions [14]. A study by Wang *et al*. [15], discussed heterogeneous ensemble learning, which leverages the diversity among various machine learning algorithms and has gained growing interest in research on predicting building energy usage. This approach enhances the predictive accuracy of machine learning by effectively combining several predictive models.

Therefore, the primary focus of this study is to identify the significant drying parameters of seaweed that exhibit heterogeneity and to evaluate the impacts of heterogeneity on the removal of moisture content of seaweed. This evaluation is conducted both before and after excluding heterogeneity parameters, as well as once all the single-eliminated heterogeneity parameters are added back into the model. In this study, two ma-

chine learning models, namely random forest and elastic net, are proposed for identifying the significant parameters that determine the moisture content removal of seaweed. Evaluation metrics are then employed to assess the performance and accuracy of the machine learning models.

2. Methodology

2.1. Flowchart of study

Figure 1 provides a summary overview of the entire research presented in this study. This project begins with the data collection from the process of seaweed drying using a hybrid solar drier in Semporna, Sabah. Note that computations are carried out on all potential models, taking into account interactions up to second order. The inclusion of interaction variables is crucial for modelling heterogeneity in the relationships, capturing and accounting for variations in these relationships within the analysis [16].

Two proposed machine learning algorithms, random forest and elastic net, will be applied in the R software as variable selection techniques to identify the significant parameters that determine the moisture content removal of the seaweed. Consequently, both of the mentioned machine learning techniques will independently select the 15 and 20 highest ranking variables, respectively. Moving on, VIF and boxplot analysis are the techniques that will be utilized to identify the significant parameters exhibiting heterogeneity. Ibidoja *et al.* [11], have employed the methods of variance inflation factor (VIF) and boxplot in their study to determine the heterogeneity parameters. Hence, these methods will be considered in this study as well.

Note that the VIF is computed by utilizing the *vif* function from the *car* library in the R software. This computation involves the original dataset and considers only the main effects of the independent variable. Once the VIF values are obtained, the R-squared values for the main drying parameters can be calculated using equation (1).

$$R^2 = 1 - \frac{1}{VIF}. \tag{1}$$

After determining the lowest and highest R-squared values, the study calculates the average R-squared value for the five main drying parameters. This average R-squared value will be used as a benchmark to identify heterogeneity parameters. If the R-squared value of the main drying parameters falls below this benchmark, it indicates the presence of heterogeneity. Evaluation metrics are then performed to assess the performance and accuracy of the model. The evaluation metrics used in this study are Mean Square Error (MSE), Sum of Square Error (SSE), Mean Average Percentage Error (MAPE) and R-squared. Hence, the impacts of heterogeneity both before and after removing heterogeneity parameters can be determined using the evaluation metrics calculated from the stated machine learning techniques. The next step is to add back all the single-eliminated parameters that exhibit heterogeneity to the model, and the accuracy of this model will be determined using the metric validations. Therefore, the optimal model is chosen by comparing the impact of the two different variable selection techniques (random forest and elastic net) in three different scenarios: the model before identifying heterogeneity parameters, the model after removing heterogeneity parameters, and the model that includes the individually eliminated heterogeneity parameters.

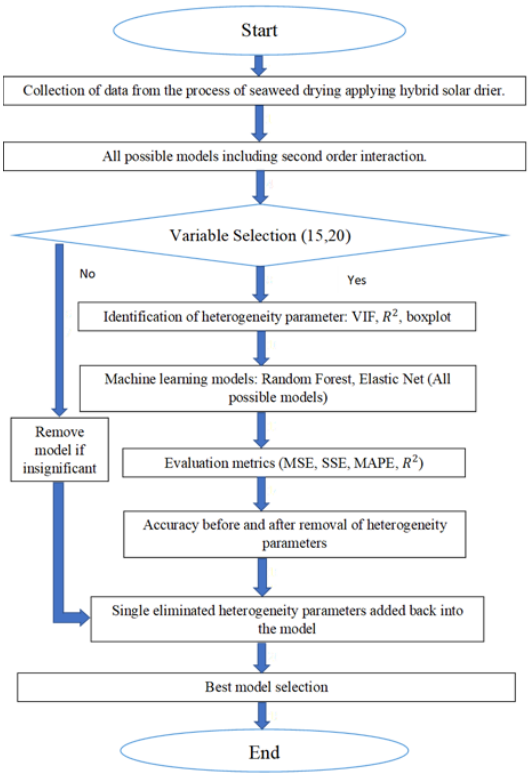

Figure 1. Methodology flowchart.

2.2. Data description

The seaweed drying data in Semporna, Sabah, were gathered from 8^{th} April 2021 to 12^{th} April 2021, between the time frame of 8:00 a.m. to 5:00 p.m. Sabah is the preferred location for seaweed cultivation due to its favourable environmental and geographical conditions. The v-GHSD, which operates as a forced convection indirect type, was used as the smart farming technology in the seaweed drying process. The placement of the sensors is designed to capture data for the drying parameters, which are hourly solar radiation, temperature, humidity, and moisture content. The data is stored in an IoT cloud database, where it is continuously computed every second and then converted into thirty-minute intervals for data analysis [5].

In order to collect the drying parameter data, a total of 29 sensors were strategically placed within the drier but this study focuses on five crucial ones: Temperature (T1, T4, T7), Humidity (H5), and Solar Radiation (PY). Table 1 provides more information on the drying parameters. The selected drying parameters are crucial due to the large number of sensors.

Table 1. Representation of Parameters.

Symbols	Factors	Definitions
Y	Dependent	Moisture Content
H5	Independent	Relative Humidity Chamber
PY	Independent	Solar Radiation
T1	Independent	Temperature (℃) ambient
T4	Independent	Temperature (℃) before entering solar collector
T7	Independent	Temperature (℃) of solar collector

The data in this study consists of 1914 data points with five independent variables and one dependent variable. The impact of interaction variables, including their second-order interactions, will be examined. For example, T1T4 represents the interaction between T1 and T4, and T1T7 means the interaction between T1 and T7. Next, the combination of the two second-order interaction variables, such as T1T4*T1T7, will be studied. Hence, the data includes the main effects of five variables and the interaction effects of 55 variables, resulting in 60 independent variable models influencing the moisture content removal of seaweed. Appendix A provides details of all variables used and their second-order interactions.

2.3. Multiple Linear Regression

Multiple linear regression is a commonly employed statistical technique because of its inherent simplicity and easily understandable interpretation [17]. A multiple linear regression model is a regression technique that suggests a linear relationship between a dependent variable y_i and a set of explanatory variables $x_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)$ for p explanatory variables. Consider a multiple regression model for n observations:

$$y = X\beta + \varepsilon, \tag{2}$$

where y represents a n× 1 vector of dependent variables, X represents a design matrix of order n × p, β represents a p × 1 vector of unknown parameters and ε is a n × 1 vector of error term that has a normal distribution with zero mean, homoscedastic and that the errors are uncorrelated [18].

Ordinary least squares (OLS) is a method used in regression analysis to estimate β that minimize the sum of squared differences between the observed and predicted dependent variable, y. The ordinary least squares estimator of β, denoted $\widehat{\beta}$, is obtained by minimizing $\varepsilon\varepsilon^{'}$ as follows:

$$\begin{aligned} \varepsilon\varepsilon^{'} = (y - X\beta)^{'}(y - X\beta) = \; & y^{'}y - 2\beta^{'}X^{'}y + \beta^{'}X^{'}X\beta \\ \partial\left(\varepsilon\varepsilon^{'}\right)/(\partial\beta) = & -2X^{'}y + 2X^{'}\beta = 0, \\ & X^{'}X\beta = X^{'}y, \\ & \widehat{\beta} = \left(X^{'}X\right)^{-1}X^{'}y. \end{aligned} \tag{3}$$

2.4. Heterogeneity Identification and Variance Inflation Factor

According to Ibidoja *et al.* [11], consider the following multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 T_{i,1} + \beta_2 T_{i,2} + \cdots + \alpha_j + \varepsilon_i. \tag{4}$$

Here, Y_i for $i = 1, 2, \ldots n$ represents the response value for the i^{th}case moisture content and estimates $\beta^{'}s$ represents the regression coefficients for the explanatory variables, specifically the drying parameter ($T^{'}s$). Meanwhile, α_j indicates the parameters that exhibit heterogeneity for $j = 1, 2, \ldots f$ and ε is the random error. Excluding an important variable during the calculation of this regression equation leads to biased and inconsistent estimate of β. Not only that, it is also possible that certain variables are correlated with the error term, thereby violating the assumption of regression.

The VIF serves as the most frequently used and straightforward measure to indicate the presence of multicollinearity [12]. The VIF is defined as:

$$VIF = \frac{1}{1 - R^2}, \tag{5}$$

hence,

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

In this study, the average R^2 will serve as the benchmark for identifying the heterogeneity parameters. If the R^2 of the main drying parameters falls below this benchmark, it indicates the presence of heterogeneity. Jiehong *et al.* [19], discovered that as the strength of the linear relationship between variables increases, the corresponding R-squared value increases, leading to a gradual rise in VIF_i. In other words, a higher VIF indicates a more serious presence of multicollinearity among variables. A VIF value exceeding 10 signifies that multicollinearity exists.

2.5. Machine Learning Algorithms

2.5.1. Random Forest

Random forest is a commonly used supervised machine learning technique that solves both classification and regression problems. Random forest is an ensemble-based learning algorithm that utilizes the concept of bootstrap aggregation. In random forest, predictions are made by averaging the outputs of multiple trees for regression tasks, while classification tasks are based on computing the majority votes of predicted values [20].

According to Louppe [21], the learning set, referred to as $\mathcal{L}$, consists of $\mathbb{N}$ pairs of input vectors and their corresponding output values, denoted as $(x_1, y_1), \ldots, (x_N, y_N)$, where $x_i \in X$ and $y_i \in Y$. A collection of p-input vectors, x_i (for $i = 1, ..., N$) can be represented by a $N \times p$ matrix X. In this matrix X, the rows $i = 1, ..., N$ relate as input vectors x_i, while the columns $j =$

1, ..., p represent the input variables X_j. Likewise, the response can be expressed as a vector $y = (y_1, \ldots, y_N)$.

Within this context, the task of supervised learning involves learning a function φ: X→ Y from the learning set $\mathcal{L}$ = (X, y). The goal is to discover a model that produces predictions $\varphi(\mathbf{x})$, denoted as $\hat{Y}$, as accurate as possible. In this case, Y must be continuous thus the learning task is a regression problem. Hence, the results of the models can be explained as follows. The regressor can be defined as a function of φ**:** $\mathbf{X}\rightarrow \mathbf{Y}$ where $\mathbf{Y} \in \mathbf{R}$.

One advantage of random forest is that it is considered a simple and straightforward machine learning method that is robust to the noise of the target data. Random forest also excels at handling large datasets with both quantitative and qualitative variables [20]. However, its speed can be reduced with a higher number of trees, as increased accuracy requires more time for computation [22].

2.5.2. Elastic Net

Elastic net, a frequently employed regularization algorithm, is often associated with estimating supervised generalized linear models through penalized maximum likelihood. It is a combination of Ridge regression and Lasso regularization, drawing its appealing qualities from the integration of the ℓ_1 and ℓ_2 norms. This integration provides the technique with the ability to select variables while taking into consideration their correlations [23]. Moreover, it effectively handles the issue of multicollinearity among predictor variables [24]. The elastic net loss function for any fixed nonnegative penalty parameters λ_1 and λ_2 is defined as:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X\beta})^T (\mathbf{y} - \boldsymbol{X\beta}) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\boldsymbol{\beta}\|^2, \quad (6)$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ and $\|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \beta_j^2$. The estimator $\widehat{\beta}$ for the elastic net minimized the equation:

$$\widehat{\beta} = \underset{\beta}{arg\min} \{L(\lambda_1, \lambda_2, \beta)\}. \quad (7)$$

The technique employed in this case is the method of least squares with penalties. Assume $\alpha = \lambda_2/(\lambda_1+\lambda_2)$, then resolving for $\widehat{\beta}$ is equivalent to resolving the optimization problem:

$$\widehat{\beta} = arg \underset{\beta}{min} \|y - X\beta\|^2,$$

subject to,

$$(1-\alpha)\|\beta\|_1 + \alpha\|\beta\|^2 \leq t \; for \; some \; t. \quad (8)$$

The penalty of the elastic net is represented by the function $\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$, which combines the penalties of Ridge and Lasso regression. Ridge and Lasso regression can be expressed using the parameter α. The values $\alpha = 0$ and $\alpha = 1$ correspond to Ridge and Lasso regression, respectively [25]. A study by Schreiber-Gregory *et al.* [26], mentioned that the benefits of elastic nets are their ability to impose sparsity and their lack of a restriction on the number of chosen variables. It also promotes a grouping effect for strongly correlated variables. However, the biggest drawback of this method is the potential for double shrinkage in naive elastic nets; thus, caution must be exercised while using it.

Table 2. Range of R-squared values.

Range of R-squared Values	Description
$85\% \leq R^2 \leq 100\%$	Very Good
$70\% \leq R^2 < 85\%$	Good
$50\% \leq R^2 < 70\%$	Reasonably Good
$30\% \leq R^2 < 50\%$	Reasonably Bad
$15\% \leq R^2 < 30\%$	Bad
$0\% \leq R^2 < 15\%$	Very Bad

2.6. Model evaluation

Evaluating the precision and performance of a model is a crucial step in any regression analysis. MSE, SSE, MAPE and R-squared are the model evaluation metrics used in this study to evaluate the model's reliability and accuracy. These metrics help in comparing and identifying the best regression model that best fits the data and achieves the desired prediction accuracy. Generally, a higher level of prediction accuracy of a model is indicated by lower values of MSE, SSE and MAPE. However, a higher value of R-squared suggests a better fit of the model to the data. As indicated by Moreno *et al.* [27], if the MAPE value is below 10, the forecast is highly accurate; however, exceeding 50 indicates an inaccurate forecast. Based on the provided source from Arsad [28], this study will utilize the following range of R-squared values, as shown in Table 2, to evaluate the quality of regression models: The formulas of the evaluation metrics used are displayed in Table 3, where the variable y_i represents the actual observations, $\hat{y}_i$ represents the predicted values, $\bar{y}$ denotes the mean of all the observations and n represents the total number of observations.

3. Result and discussion

3.1. Identification of Heterogeneity parameters

The VIF and R-squared values for the main drying parameters T1, T4, T7, H5 and PY are presented in Table 4. Once the lowest and highest R-squared values have been identified, the average R-squared value of the five main drying parameters can be calculated, and these results are tabulated in Table 5. In this study, the average R-squared value of 0.4843 will be used as the benchmark for identifying the heterogeneity parameters. If the R-squared of the main drying parameters falls below this benchmark, it indicates the presence of heterogeneity. Hence, the parameters T7 and PY exhibit heterogeneity since the R-squared values for both of these parameters, 0.0670 and 0.4298 respectively as shown in Table 4, are lower than the benchmark value. In addition, the boxplot can also be used as supporting evidence to identify the existence of heterogeneity within the drying parameters. This is because the boxplot is useful for examining symmetry and variability as well as for identifying potential outliers [30]. Therefore, the variability of the five single seaweed drying parameters can be shown by the boxplot in Figure 2. Based on the box plot in Figure 2, the variables T7, H5 and PY exhibit heterogeneity. Note that the illustration from the boxplot for variables T7 and PY coincides with the results of the average R-squared obtained earlier, indicating that these two

Table 3. Formulas for evaluation metrics.

Range of R-squared Values	Description	References
Mean Square Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.	[29]
Sum of Square Error (SSE)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.	[29]
Mean Average Percentage Error (MAPE)	$MAPE = \left[\frac{100}{n}\right]\sum_{i=1}^{n}\left\|\frac{(y_i-\hat{y}_i)}{y_i}\right\|$.	[27]
R-squared	$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$	[29]

Table 4. VIF and R-squared values for the main drying parameters.

Parameters	VIF	R^2
T1	10.1707	0.9107
T4	9.1189	0.8903
T7	1.0718	0.0670
H5	2.3982	0.5830
PY	1.7539	0.4298

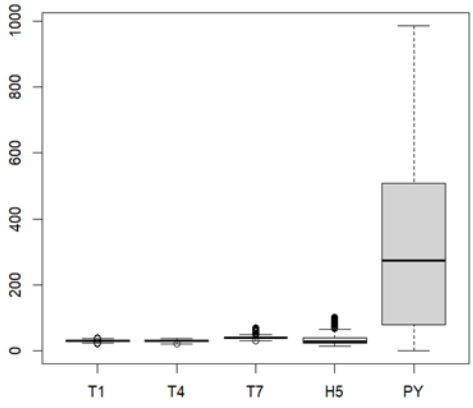

Figure 2. Boxplot for seaweed drying parameters.

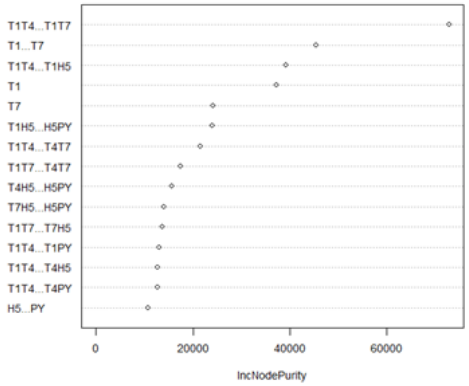

Figure 3. The 15 highest important variables for random forest.

variables exhibit heterogeneity. However, the R-squared value of variable H5 does not indicate the presence of heterogeneity based on the findings obtained earlier since the R-squared value of H5, which is 0.5830, exceeds the benchmark value (0.4843).

Although data visualization increases the speed and quality of decisions [31], it involves summarizing complex data into visual representation, potentially resulting in a partial loss of detail. Meanwhile, according to Almeida *et al*. [32], quantitative results obtained from numerical operations and statistical analysis have the potential to produce predictions that are considered more reliable and accurate. Therefore, in this case, the average R-squared value will be considered to indicate the presence of heterogeneity among the drying parameters, suggesting that parameters T7 and PY show heterogeneity.

On top of that, the presence of multicollinearity among the main drying parameters was assessed using VIF values, as shown in Table 4. The VIF values between 1 and 5 for the parameters T7, H5 and PY indicate a moderate level of correlation among these parameters. However, the VIF of parameter T4 falls into the range between 5 and 10, suggesting a possible presence of multicollinearity. Meanwhile, parameter T1 exhibits a VIF of 10.1707, just slightly exceeding 10, raising concerns about potential multicollinearity. Importantly, no serious multicollinearity issues were identified among the main drying parameters before variable selection. Given that the primary focus of this study is on heterogeneity, an in-depth analysis of multicollinearity after variable selection is not conducted.

3.2. *Results of selected parameters before the removal of Heterogeneity parameters*

Figures 3 and 4 display plots representing the 15 and 20 highest important variables ranked by the random forest algorithm, while Figures 5 and 6 display the plots representing the 15 and 20 highest important variables ranked by the elastic net algorithm using R software. The rankings are based on the importance scores computed by the machine learning techniques.

Based on the findings in Figures 3, 4, 5 and 6, it is evident that most of the important variables for the random forest and elastic net model consist of interaction variables. Both regression models represent the entire drying process of the seaweed as the selected variables - temperature, humidity, and solar radiation, collectively define the whole drying process.

3.3. *Results of selected parameters after the removal of Heterogeneity parameters*

As discussed in subsection 3.1, the parameters T7 and PY exhibit heterogeneity. These two heterogeneity parameters, including their second-order interaction, are then eliminated from the model. As a result, there are only nine parameters

Table 5. Heterogeneity identification.

Lowest VIF	Highest VIF	Lowest R^2	Highest R^2	Average R^2	Heterogeneity Parameters
1.0718	10.1707	0.0670	0.9017	0.4843	T7, PY

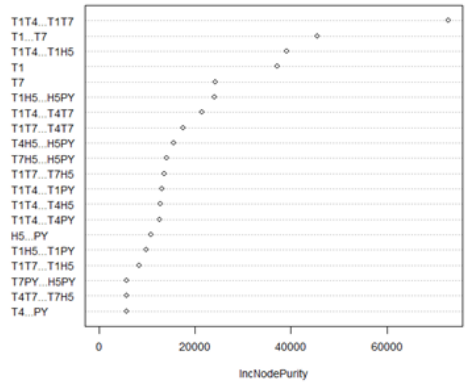

Figure 4. The 20 highest important variables for random forest.

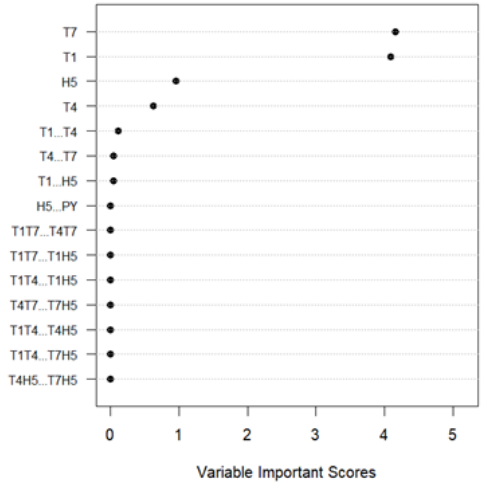

Figure 5. The 15 highest important variables for elastic net.

left, which are T1, T4, H5, T1T4, T1H5, T4H5, T1T4*T1H5, T1T4*T4H5 and T1H5*T4H5, to determine the moisture content removal of the seaweed. Therefore, variable selection is no longer needed since there are only nine parameters, but the proposed machine learning algorithms, random forest and elastic net, will be applied for model evaluation purposes in order to ascertain the impact before and after the removal of heterogeneity parameters.

3.4. Once all the Single-Eliminated parameters that exhibits Heterogeneity are added back into the model

It is important to consider that when a model includes the interaction of single variables, the main effects of those variables themselves should be taken into account. This is because the

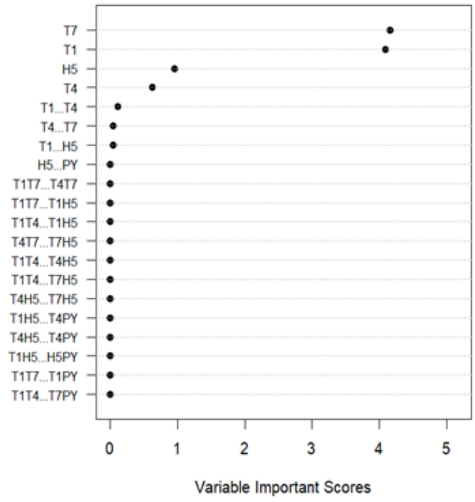

Figure 6. The 20 highest important variables for elastic net.

selection of interaction effects can only occur once the main effects have been chosen [33]. The inclusion of interaction terms is vital in achieving the best possible final model and ensuring unbiased results [5]. As a result, the parameters T7 and PY, which were previously excluded due to their heterogeneous nature, will be added back into the model. Consequently, the current analysis will involve a total of 11 variables: T1, T4, T7, H5, PY, T1T4, T1H5, T4H5, T1T4*T1H5, T1T4*T4H5 and T1H5*T4H5.

4. Comparison of analysis before and after the removal of heterogeneity parameters and after the inclusion of all Single-Eliminated Heterogeneity parameters in the model

Table 6 presents a summary of the comparison of analysis in terms of evaluation metrics conducted before and after the removal of heterogeneity parameters, as well as upon including all single-eliminated heterogeneity parameters into the model for both machine learning models. Notably, the random forest model demonstrates superior performance over the elastic net model for all three scenarios mentioned. This is supported by the smaller values of MSE, SSE and MAPE, along with the remarkably higher R-squared value achieved by the random forest model.

The random forest model consistently outperforms the elastic net model, as demonstrated by significantly lower MSE and SSE values for the 15 and 20 highest important variables before the removal of heterogeneity parameters as compared to the two other analyses. Additionally, the MAPE values of the

Table 6. Comparison of analysis before and after the removal of heterogeneity parameters and inclusion of Single-Eliminated Heterogeneity Parameters.

Machine Learning Models		Random Forest				Elastic Net			
Metric Validation	Highest Ranking Variables	MSE	SSE	MAPE	R^2	MSE	SSE	MAPE	R^2
Before Heterogeneity	15	33.9760	65029.95	4.7012	0.8758	89.8925	172054.3	13.4913	0.6714
	20	33.7274	64554.18	4.5300	0.8767	81.8947	156746.5	12.6896	0.7006
After Heterogeneity	9	88.1500	168719.00	10.0844	0.6778	121.0011	231596.2	15.5954	0.5577
Inclusion of Single-Eliminated Heterogeneity Parameters	11	39.5606	75718.91	5.6415	0.8554	117.8621	225588.1	15.2571	0.5692

random forest model before removing the heterogeneity parameters for the 15 highest important variables (4.7012) and the 20 highest ranking variables (4.5300) are notably below 10, implying the model's forecast is highly accurate. Despite the MAPE value (5.6415) being below 10 for the random forest model with single-eliminated heterogeneity parameters, the model performs even better when the heterogeneity parameters are not excluded, indicating higher prediction accuracy for seaweed moisture content removal. Moreover, the R-squared value for the random forest model before the removal of heterogeneity parameters is notably higher than the two other analyses, implying that a large proportion of variability in determining the moisture content removal of the seaweed can be explained by the random forest model, particularly before removing the heterogeneity parameters, in comparison to the elastic net model. Not only that, the quality of the random forest model for the 15 and 20 highest ranking variables before removing the heterogeneity parameters is very good since the R-squared value is greater than 85%.

However, there is an unexpected reduction in the accuracy of the regression model following the removal of heterogeneity parameters, as reflected in the increased MSE, SSE and MAPE values along with lower R-squared values. The elimination of parameters from the model poses a risk of information loss, potentially introducing specification bias as it may overlook crucial factors that impact the investigated phenomenon. Consequently, the performance of the model may be adversely affected, leading to less accurate outcomes.

Notably, the inclusion of all single-eliminated heterogeneity parameters results in smaller MSE, SSE and MAPE values, along with a higher R-squared value compared to the analysis after the removal of heterogeneity parameters. These findings indicate the superior performance of the random forest model when incorporating these parameters. However, the model's overall peak performance is observed before the exclusion of heterogeneity parameters, as evidenced by smaller MSE, SSE and MAPE values, as well as a slightly higher R-squared value. This suggests a better fit of the random forest model to the data before the removal of heterogeneity parameters.

There is a possibility that the single-eliminated parameters that exhibit heterogeneity might be influenced by confounding factors that are not adequately accounted for in the analysis. A confounding variable, distinct from the one being studied, is a factor correlated with both the dependent variable and the independent variable under investigation [34]. These confounding variables can introduce bias into the study results, consequently affecting the analysis and leading to poorer performance [35]. Furthermore, it is also possible that the data for the single-eliminated parameters exhibiting heterogeneity contains more errors compared to the other parameters.

Therefore, the performance of the random forest model in terms of higher accuracy and lower error before removing the heterogeneity parameters indicates its ability to produce more accurate predictions, reduce errors, and explain a greater proportion of the variance. This makes the random forest a better predictive model for analysing the 15 and 20 highest-ranked variables prior to the removal of heterogeneity parameters. This outcome aligns with the well-known capability of random forest models to offer precise predictions and interpretability when determining the ranks of important variables [36]. Not only that, the findings presented by Callens *et al.* [37], also highlight the advantage of tuning hyperparameters in the random forest algorithm, allowing for the optimization of model performance.

5. Conclusions

This study aims to investigate the presence of heterogeneity among drying parameters and proposes a heterogeneity model within machine learning algorithms to enhance the accuracy of predicting moisture content removal. Utilizing random forest and elastic net for variable selection, the performance of these prediction models is quantitatively assessed using metrics such as MSE, SSE, MAPE and R-squared. It can be concluded that the predictive performance of the random forest model is significantly stronger than the elastic net model, both before and after removing the heterogeneity parameters as well as after the inclusion of the single-eliminated heterogeneity parameters. The random forest model demonstrates higher accuracy, minimal errors and exceptional performance in forecasting the moisture content removal of seaweed, particularly before the removal of heterogeneity parameters. These advantages make it a preferable option over the elastic net model. This conclusion, emphasizing the superior performance of the random forest model, aligns with the findings of several studies conducted by Sharma

et al.[2], Ibidoja *et al*. [11], Ibidoja *et al*. [18], Mukhtar *et al*. [38] and Yesilkanat [39].

This study is crucial to seaweed drying, and the outcomes of this study will assist seaweed farmers in processing seaweed into high-quality products and reduce post-harvest losses. For future studies, it is recommended to explore other machine learning algorithms such as ridge, support vector machine, bagging, boosting and LASSO for variable selection. These algorithms can also be employed to investigate the impact of heterogeneity both before and after the removal of heterogeneity parameters. Furthermore, as this study did not address the issue of outliers, robust regression techniques, including M Huber, M Hampel, M Bi Square, MM and S estimators, can be considered to effectively handle this matter.

Acknowledgment

The authors would like to thank the School of Mathematical Sciences, Universiti Sains Malaysia for their continuous support in this research.

References

[1] T. Panch, P. Szolovits & R. Atun, "Artificial intelligence, machine learning and health systems", Journal of Global Health **8** (2018) 020303. https://doi.org/10.7189/jogh.08.020303.

[2] A. Sharma, A. Jain, P. Gupta & V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review", IEEE Access **9** (2021) 4843. https://doi.org/10.1109/access.2020.3048415.

[3] O. J. Ibidoja, F. P. Shan, M. E. Suheri, J. Sulaiman & M. K. M. Ali, "Intelligence system via machine learning algorithms in detecting the moisture content removal parameters of seaweed big data", Pertanika Journal of Science & Technology **31** (2023) 2783. http://dx.doi.org/10.47836/pjst.31.6.09.

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

[5] H. Y. Lim, P. S. Fam, A. Javaid & M. Ali, "Ridge Regression as Efficient Model Selection and Forecasting of Fish Drying Using V-Groove Hybrid Solar Drier", Pertanika Journal of Science and Technology **28** (2020) 1179. https://doi.org/10.47836/pjst.28.4.04.

[6] J. Echave, P. Otero, P. Garcia-Oliveira, P. E. Munekata, M. Pateiro, J. M. Lorenzo, J. Simal-Gandara & M. A. Prieto, "Seaweed-Derived Proteins and Peptides: Promising Marine Bioactives", Antioxidants **11** (2022) 176. https://doi.org/10.3390/antiox11010176.

[7] M. K. M. Ali, J. Sulaiman, S. Md Yasir & M.H. Ruslan, "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** (2017) 159. https://mjms.upm.edu.my/fullpaper/2017-February-11(S)/Ali,%20M.%20K.%20M.-159-172.pdf.

[8] S. Lomartire, J. C. Marques & A. C. Gonçalves, "An Overview to the Health Benefits of Seaweeds Consumption", Marine Drugs **19** (2021) 341. https://doi.org/10.3390/md19060341.

[9] J. Venkatesan, S. Anil, S. Kim, & M. S. Shim, "Seaweed Polysaccharide-Based Nanoparticles: Preparation and Applications for Drug Delivery", Polymers **8** (2016) 30. https://doi.org/10.3390/polym8020030.

[10] A. Nunes, T. Trappenberg & M. Alda, "The definition and measurement of heterogeneity", Translational Psychiatry **10** (2020) 299. doi:10.1038/s41398-020-00986-0.

[11] O. J. Ibidoja, F. P. Shan, J. Sulaiman & M. K. M. Ali, "Detecting heterogeneity parameters and hybrid models for precision farming", Journal of Big Data **10** 130 (2023). https://doi.org/10.1186/s40537-023-00810-8.

[12] J. Y. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z. Hong & Y. Chen, "Mitigating the multicollinearity problem and its machine learning approach: A Review", Mathematics **10** (2022) 1283. https://doi.org/10.3390/math10081283.

[13] P. Marenya, G. G. Gebremariam, M. Jaleta & D. B. Rahut, "Sustainable intensification among smallholder maize farmers in Ethiopia: adoption and impacts under rainfall and unobserved heterogeneity", Food Policy **95** (2020) 101941. https://doi.org/10.1016/j.foodpol.2020.101941.

[14] K. M. Rhodes, R. M. Turner, J. Savovi?, H. E. Jones, D. Mawdsley & J. P. T. Higgins, "Between-trial heterogeneity in meta-analyses may be partially explained by reported design characteristics", Journal of Clinical Epidemiology **95** (2018) 45. https://doi.org/10.1016/j.jclinepi.2017.11.025.

[15] Z. Wang, Z.Liang, R. Zeng, H. Yuan & R. S. Srinivasan, "Identifying the optimal heterogeneous ensemble learning model for building energy prediction using the exhaustive search method", Energy and Buildings **281** (2023) 112763. https://doi.org/10.1016/j.enbuild.2022.112763.

[16] G.Lamberti, *Modelling with Heterogeneity*, Ph.D dissertation, Facultat de Matemàtiques i Estadística Universitat Politècnica de Catalunya, Barcelona, Spain, 2015. https://upcommons.upc.edu/bitstream/handle/2117/95733/TGL1de1.pdf.

[17] G. Cağil, S. N. Güler, A. Ünlü, Ö. Böyükdibi & G. Tüccar, "Comparative analysis of Multiple linear Regression (MLR) and Adaptive Network-Based fuzzy Inference Systems (ANFIS) methods for vibration prediction of a diesel engine containing NH3 additive", Fuel **350** (2023) 128686. https://doi.org/10.1016/j.fuel.2023.128686.

[18] O. J. Ibidoja, F. P. Shan, Mukhtar, J. Sulaiman & M. K. M. Ali, "Robust M-estimators and Machine Learning Algorithms for Improving the Predictive Accuracy of Seaweed Contaminated Big Data", Journal of Nigerian Society of Physical Sciences **5** (2023) 1137. https://doi.org/10.46481/jnsps.2023.1137.

[19] C. Jiehong, J. Sun, K. Yao, X. Min & C. Yan, "A variable selection method based on mutual information and variance inflation factor", Spectrochimica Acta Part A: Molecular and Biomolecular spectroscopy **268** (2022) 120652. https://doi.org/10.1016/j.saa.2021.120652.

[20] K. Kirasich, T.Smith & B. Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets", SMU Data Science Review **1** (2018) 9. https://scholar.smu.edu/datasciencereview/vol1/iss3/9.

[21] G. Louppe, *Understanding Random Forests: From Theory to Practice*, Ph.D. dissertation, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, Universite de Liege, Belgium, 2014. https://doi.org/10.13140/2.1.1570.5928.

[22] N. Donges, "Random Forest: A Complete Guide for Machine Learning". [Online]. Available: on the World Wide Web: https://builtin.com/data-science/random-forest-algorithm#procon.

[23] J. C. Laria, L. K. H. Clemmensen & B. K. Ersbøll, "A Generalized Linear Joint Trained Framework for Semi-Supervised Learning of Sparse Features", Mathematics **10** (2020) 3001. https://doi.org/10.3390/math10163001.

[24] A. S. Al-Jawarneh, M. T. Ismail & A. M. Awajan, "Elastic Net Regression and Empirical Mode Decomposition for Enhancing the Accuracy of the Model Selection", International Journal of Mathematical, Engineering and Management Sciences **6** (2021) 564. https://doi.org/10.33889/ijmems.2021.6.2.034.

[25] M. K. Mukhtar, B. M. Ali, A. Javaid, M. T. Ismail & A. Fudholi, "Accurate and hybrid regularization - robust regression model in handling multicollinearity and outlier using 8SC for big data", Mathematical Modelling of Engineering Problems **8** (2021) 547. https://doi.org/10.18280/mmep.080407.

[26] N. Deanna, Schreiber-Gregory, Jackson Foundation & Karlen Bader, *Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets*, in Proceedings of the SAS Conference Proceedings: Western Users of SAS Software, 2018, pp. 1–23. https://api.semanticscholar.org/CorpusID:189925961.

[27] J. Moreno, A. L. P. Pol, F. García-Labiano & B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy", PubMed **25** (2013) 500. https://doi.org/10.7334/psicothema2013.23.

[28] Z. Arsad, " Multiple Linear Regression", *Regression Analysis*, School of Mathematical Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia, 2023, pp. 10–31.

[29] H. Pham, "A New Criterion for Model Selection", *Mathematics* **7** (2019) 1215. https://doi.org/10.3390/math7121215.
[30] C. S. Morales, R. Giraldo & M. E. Torres, "Boxplot fences in proficiency testing", Accreditation and Quality Assurance **26** (2021) 193. https://doi.org/10.1007/s00769-021-01474-8.
[31] K. Eberhard, "The effects of visualization on judgment and decision-making: a systematic literature review", Management Review Quarterly **73** (2021) 167. https://doi.org/10.1007/s11301-021-00235-8.
[32] F. Almeida, D. Faria & A. Queirós, "Strengths and Limitations of Qualitative and Quantitative Research Methods", European Journal of Education Studies **3** (2017) 369. https://doi.org/10.5281/zenodo.887089.
[33] N. Hao & H.H. Zhang, "A Note on High-Dimensional Linear Regression with Interactions", The American Statistician **71** (2017) 291. https://doi.org/10.1080/00031305.2016.1264311.
[34] T. H. Tulchinsky & E. A. Varavikova, Measuring, Monitoring, and Evaluating the Health of a Population, Elsevier eBooks, 2014, pp. 91–147. https://doi.org/10.1016/b978-0-12-415766-8.00003-3.
[35] J. Frost, "Confounding Variables Can Bias Your Results", Statistics by Jim. [Online]. Available: on the World Wide Web: https://statisticsbyjim.com/regression/confounding-variables-bias/.
[36] H. Khanum, A. Garg & M. I. Faheem, "Accident severity prediction modeling for road safety using random forest algorithm: an analysis of Indian highways", F1000Research **12** (2023) 494. https://doi.org/10.12688/f1000research.133594.1.
[37] A. Callens, D. Morichon, S. Abadie, M. Delpey & B. Liquet, "Using Random Forest and Gradient boosting trees to improve wave forecast at a specific location", Applied Ocean Research **104** (2020) 102339. https://doi.org/10.1016/j.apor.2020.102339.
[38] M. K. Mukhtar, B. M. Ali, M. T. Ismail, F. M. Hamundu, Alimuddin, N. Akhtar & A. Fudholi, "Hybrid model in machine learning–robust regression applied for sustainability agriculture and food security", International Journal of Electrical and Computer Engineering **12** (2022) 4457. https://doi.org/10.11591/ijece.v12i4.pp4457-4468.
[39] C. M. Yesilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm", Chaos Solitons & Fractals **140** (2020) 110210. https://doi.org/10.1016/j.chaos.2020.110210.

APPENDIX A.

Variables used in this study

data.frame(T1, T4, T7, H5, PY, T1T4, T1T7, T1H5, T1PY, T4T7, T4H5, T4PY, T7H5, T7PY, H5PY, T1T4*T1T7, T1T4*T1H5, T1T4*T1PY, T1T4*T4T7, T1T4*T4H5, T1T4*T4PY, T1T4*T7H5, T1T4*T7PY, T1T4*H5PY, T1T7*T1H5, T1T7*T1PY, T1T7*T4T7, T1T7*T4H5, T1T7*T4PY, T1T7*T7H5, T1T7*T7PY, T1T7*H5PY, T1H5*T1PY, T1H5*T4T7, T1H5*T4H5, T1H5*T4PY, T1H5*T7H5, T1H5*T7PY, T1H5*H5PY, T1PY*T4T7, T1PY*T4H5, T1PY*T4PY, T1PY*T7H5, T1PY*T7PY, T1PY*H5PY, T4T7*T4H5, T4T7*T4PY, T4T7*T7H5, T4T7*T7PY, T4T7*H5PY, T4H5*T4PY, T4H5*T7H5, T4H5*T7PY, T4H5*H5PY, T4PY*T7H5, T4PY*T7PY, T4PY*H5PY, T7H5*T7PY, T7H5*H5PY, T7PY*H5PY)