Journal of
Water & Climate Change

IWA PUBLISHING

Journal of Water and Climate Change Vol 14 No 7, 2294 doi: 10.2166/wcc.2023.052

Assessing machine learning tools for methane emission prediction from POME treatment in Malaysia

Kashwin Selvanathan[a], Kishaan Ragu[a], Hia Hung Yi[a], Sara Kazemi Yazdi[a,*], Zhiyuan Chen[b] and Reza Godary[a]

[a] Department of Chemical and Environmental Engineering, Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih 43500, Selangor Darul Ehsan, Malaysia
[b] Department of Computer Science, University of Nottingham, 43500 Selangor, Malaysia
*Corresponding author. E-mail: sara.yazdi@nottingham.edu.my

SKY, 0000-0001-9350-2947; ZC, 0000-0002-4915-1593

ABSTRACT

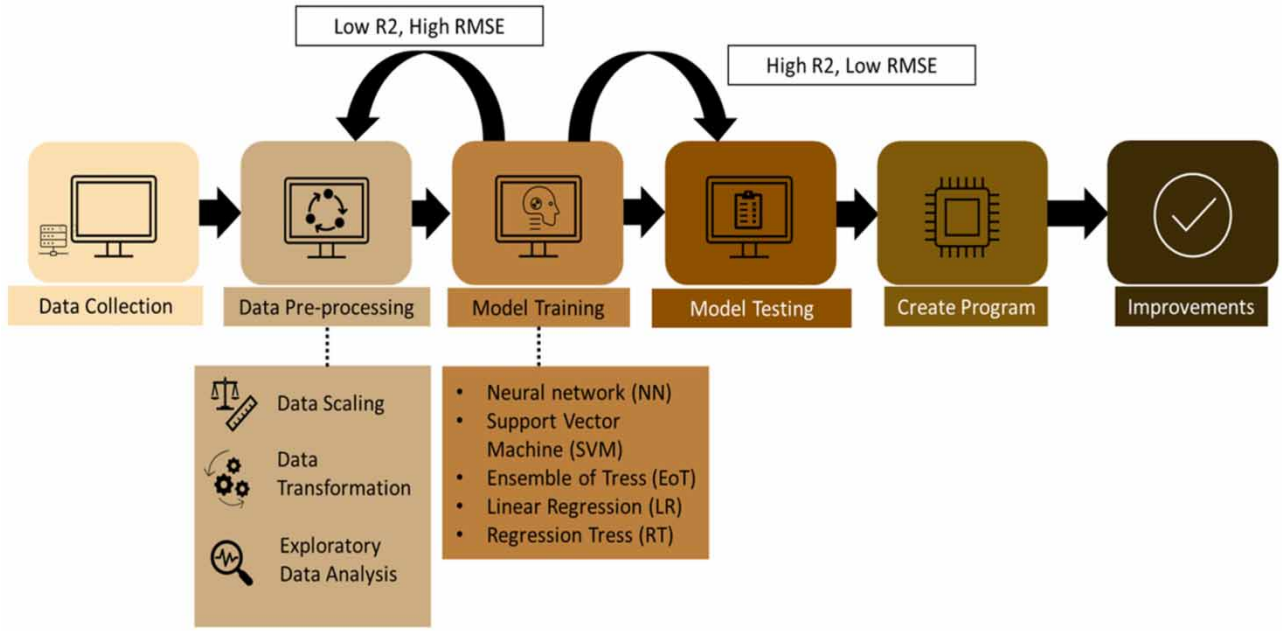
Palm oil mill effluent (POME) treatment is an anthropogenic activity contributing to global warming through methane emission. The inability to address this issue would deem true the catastrophic impacts of global temperatures exceeding 2 °C as was predicted by the Intergovernmental Panel on Climate Change (IPCC) in 2015. Little research and development exist on GHGs monitoring and methane emissions in POME treatment facilities as opposed to research on improving biogas production. A methane emission prediction tool based on machine learning models and tools can address this problem and consequently facilitate the development of efficient carbon neutrality approaches in POME treatment plants. In this study, six regression models were explored alongside their kernels using eight predictors, linking towards methane emission volume. The best model found was support vector machine (SVM), producing performance metrics for R^2 and RMSE with values of 0.45 and 0.749, respectively.

Key words: machine learning, methane, palm oil mill effluent (POME), prediction, support vector machine (SVM), transformation

HIGHLIGHTS

- Selection and utilisation of a suitable machine learning algorithm for the prediction of CH_4 emissions.
- Obtaining raw data on POME treatment for database development and understanding the nature and quality of the dataset using available data from pre-processing methods.
- Integrating the developed database into the CH_4 emission tool.
- Comparing factors influencing CH_4 emissions.
- Determining the highest influencing parameters.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Palm oil, a type of vegetable oil extracted from the fruits of the oil palm tree (*Elaeis guineensis*) (CABI 2019), is labelled as the most widely used vegetable oil on the planet. Based on recent data obtained from Statista (2022b), the current global palm oil production rate is capped at 75 million metric tons, equivalent to about 20% of the global vegetable oil production. Statista (2022a) has also reported that in addition to production, the worldwide consumption of palm oil is currently capped at 73 million metric tonnes, dominating the world vegetable oil market by capturing 35% in market shares. The reason palm oil possesses such great statistics is mainly due to the ability of oil palm fruit as a very efficient crop, capable of producing enormous volumes of oil over relatively small areas of land throughout the year (Abd Ghani 2021). In addition to that, as the global palm oil market achieved a market value of 50.6 billion USD in 2021, experts such as IMARC group strongly believe that by 2027, the global palm oil market is expected to reach 65.5 billion USD, with a CAGR of 4.3% between 2022 and 2027 (IMARC Group 2021). Malaysia, as the world's second largest producers of palm oil after Indonesia, is responsible for 26% of global palm oil production and 34.3% of global palm oil exports (MPOC 2020). In terms of economic potential and growth, it is undeniable that palm oil will continue to provide various benefits to Malaysia in the present, but also in the future. However, despite the benefits and positive projections stated above, from an environmental standpoint, palm oil does have a radical effect on the environment as the refining process of oil palm to palm oil produces a non-toxic wastewater known as palm oil mill effluent (POME) (Zafar 2022).

POME is defined as the effluent formed in palm oil mills during the last phases of palm oil processing that can be categorised as liquid waste (Bachrun & Baskara 2023). POME contains (95–96)% water, (0.6–0.7)% oil, (4–5)% total solids of which (2–4)% contributes to suspended solids (Hassan & Abd-Aziz 2012). In terms of concentration, POME typically consists of total solid concentration of (~40,500 mg/L), oil and grease concentration of (~4,000 mg/L), high chemical oxygen demand (COD) concentration of (~50,000 mg/L), and high biological oxygen demand (BOD) (~25,000 mg/L) (Hassan & Abd-Aziz 2012). Based on the concentrations stated, it can be classified that POME contributes to the death of aquatic life if channelled directly into receiving streams as it contains 100 times higher COD and BOD concentrations than municipal sewage (Kamyab *et al.* 2018). High COD and BOD values translate to low oxygen availability in water which will cause aquatic organisms to suffocate and die (EPA 2012). Based on the book titled 'Palm Oil' by Lai *et al.* (2016), it is estimated that roughly 2.5 tonnes of POME are generated for every tonne of crude palm oil recovered from milling. As a prominent palm oil producer, the amount of POME generated by Malaysia's palm oil industry is approximately 50 million tons per annum (Akhbari *et al.* 2020).

Therefore, to prevent water pollution such as oxygen depletion and eutrophication, the proper treatment of POME is essential (Ng 2017).

Malaysia has taken action to tackle pollution caused by POME by implementing strict discharge limits on industrial wastewater effluents through the Environmental Quality Act 1974 (Department of Environment Malaysia 1979) and further improvised in the Environmental Quality (Sewage) Regulations 2009 (Department of Environment 2010). Based on the environmental regulations, the limits set by the Department of Environment for BOD_5, COD, suspended solids and oil and grease are 50, 200, 100, and 20 mg/L, respectively. As a result, Malaysia has adopted a few treatment methods to ensure that treated POME achieves the target discharge limits. These methods are mainly anaerobic treatment (AD), membrane treatment (MD), and evaporation method (EM). According to Kamyab *et al.* (2018), the most extensively used technique for the treatment of POME is the biological approach, which is based on anaerobic and aerobic ponding systems using bacteria. In the case of Malaysia, AD in the form of open ponding is mainly selected as a primary treatment approach due to its cost-effectiveness (Appels *et al.* 2008) associated with low capital and operating costs (Sarbatly 2020). Despite anaerobic ponding, AD is also applied in equipment such as fluidised bed reactors, closed anaerobic digesters, up flow anaerobic sludge fixed film reactor (UASFF), ultrasonic membrane anaerobic system (UMAS), and many more to treat POME.

A common misconception interpreted by operators working in palm oil mills is that POME is the most expensive and challenging waste to manage (Madaki & Seng 2013). Instead of viewing it as waste, POME contains a very high potential to become a resource if subjected to the right treatment methods such as AD. According to Ward *et al.* (2008), POME can produce a carbon neutral energy source in the form of biogas. Also, according to Tambone *et al.* (2010), after the AD process, the final residue is nutrient-dense and can be utilised as fertiliser in agriculture. By looking at these statements, POME can generate various resources if utilised accordingly.

As understood from the statement made earlier, POME can yield biogas through AD. Loh *et al.* (2017) verified that biogas from AD via open ponding or open digestor tanks comprises (60–70)% methane (CH_4), (30–35)% carbon dioxide (CO_2), and trace amounts of hydrogen sulphide (H_2S). This is due to the bacterial activity breaking down the organic matter present in POME in the absence of oxygen (EPA 2022). Despite the obvious benefits of operating under oxygen free conditions, the biggest downside to this is the release of greenhouse gases (GHGs) into the atmosphere. Methane (CH_4) is the second most abundant manmade GHG after carbon dioxide (CO_2), accounting for roughly 20% of global emissions (EPA 2021). Methane is 25 times more effective than carbon dioxide at trapping heat in the atmosphere (United States Environmental Protection Agency 2021). Therefore, it is crucial to capture biogas generated from POME treatment as it is proven by World Biogas Association (2018) that capturing biogas would enable the reduction of global emissions by (18–20)% which is in line with the Paris Agreement to address climate change (UNFCCC 2015). Malaysia as a signatory to the Paris Agreement has committed to tackling climate change by implementing more biogas capture in the palm oil industry as it contributes to about 23.7% of the total methane emission in Malaysia (Ministry of Environment & Water 2020).

By looking into the state of our global climate, the global warming rate has rapidly intensified, leading to the declaration of a climate emergency state of 2,089 jurisdictions from 38 different countries as of March 2022 (climateemergencydeclaration.org 2022). According to the Intergovernmental Panel on Climate Change (IPCC), the global warming levels of 1.5 and 2 °C will be exceeded in the 21st century if no significant and more intensive measures are taken to combat climate change (Masson-Delmotte *et al.* 2021). As a result, Malaysia vowed to cut its greenhouse gas emission intensity across the economy by 45% based on GDP in 2030 at the recent United Nations Framework Convention on Climate Change (UNFCC) COP 26 held from 31 October to 12 November 2021 in Glasgow (Dalm 2021). And one way to achieve this ambition is by focusing on the installation of methane capturing facilities in new and existing palm oil mills (Ministry of Energy 2017).

Despite the benefits and potentials POME contains, Malaysia has not yet been close to fully capturing the methane emissions and utilising it to generate biogas as most of the palm oil mills are operating on ponding systems. In addition to that, based on a study by Khairul *et al.* (2019), only a total of 50 palm oil mills are currently in the testing phase of biogas recovery under the Clean Development Mechanism (CDM) in Malaysia. Following that, Statista (2020b) has reported that there are 457 existing Malaysian palm oil mills in operation as of the year 2020. This shows that only 11% of the overall Malaysian palm oil industry has adapted the transition phase to biogas capture which is far from ideal.

To ensure the process of capturing methane runs smoothly, identifying the exact amounts of biogas emissions from POME treatment facilities accurately should be regarded as a top priority for Malaysia. This can be achieved through means of prediction or real-time monitoring. An accurate identification method can serve as a design basis for engineers and building

operators to design the capture systems accurately and in accordance with the size of the associated palm oil mills. However, as easy as it may seem, there are a few challenges associated with the creation of a prediction tool or programme to quantify methane emissions.

Firstly, the numerous process parameters produce many instabilities in the anaerobic digestion process, which leads to several unpredictabilities in the system's methane generation. According to studies done by Lam & Lee (2011), Abdurahman *et al.* (2013) methane emission from anaerobic digestion is influenced by a number of parameters, including operating temperature and pH that affects the survivability of methanogens in the organic matter (Zinder *et al.* 1984; Choong *et al.* 2018). Besides that, BOD, COD, and suspended solids affect growth of microorganisms such as methanogens that would impact the digestion of POME which in turn would cause variation in the methane produced (Utami *et al.* 2016; Putro 2022). Hydraulic retention rate (HRT) and organic loading rate (OLR) would impact the efficiency of the digestion of POME, thus making them crucial parameters in methane production (Zainal *et al.* 2022). In addition to that, according to Abdurahman *et al.* (2013), the technology used to pre-process and prepare the digestion for AD processed can alter the amounts of emitted methane from the process, ranging between 36 and 71.9%. When it comes to estimating and monitoring CH_4 emissions, these uncertainties, along with the lack of research and studies on this topic, represent a significant challenge to obtain accurate measurements of CH_4 from POME.

Secondly, conventionally, methane monitoring has been done through the traditional offline grab-sampling methods by physically going to the treatment sites and conducting experiments and tests to verify the amount of methane emitted during the treatment processes. As of 2020, the arrival of the COVID-19 pandemic halted all physical activities in a recent study by Diffenbaugh *et al.* (2020). These activities include the development and transition of renewable energy (RE) by the government of Malaysia such as methane monitoring and capturing in palm oil mills. Due to this, the reliance on fossil fuels as the primary sources of energy has re-emerged since the energy demands keep rising despite the restrictions implemented by the movement control orders (MCOs), resulting in an increase in GHG emissions (Vaka *et al.* 2020). Therefore, the transition to RE has never been more challenging as Malaysia now generates only 8% of its energy from renewable sources. This is small in percentage when compared to the ambitious goal of 20% energy generated from renewable resources by 2025 (Joshi 2019). Moreover, this shows that there is a need to develop a tool to predict and monitor GHGs from POME which can aid the operators and personnel working remotely to determine the plants' methane emissions and facilitate CH_4 capturing. Besides that, according to NST Business (2021), since some palm oil mills have been transitioning to automation via Industrial Revolution 4.0 by implementing Palm Oil Mill Integrated System (POMIS) to facilitate daily operations, incorporating a monitoring or prediction tool implemented in their Internet of Things (IoT) would drive these palm oil mills towards sustainability at a faster rate.

Machine learning (ML) has advanced in recent years, and it may be employed as a predictive tool in microbial ecology and system biology studies (Kazemi Yazdi & Scholz 2010; Witten 2011). The rise of various ML algorithms such as linear regression (LR), artificial neural network (ANN), support vector machine (SVM), and Gaussian process regression (GPR) incorporated into these studies have been successful in predicting various outputs. This also includes the successful prediction of biogas production from AD (Zaied *et al.* 2020; Asadi & McPhedran 2021). However, despite this successful prediction, most of these studies involve optimising the respective palm oil mill's biogas production and their AD processes. There are still very few developments towards a specific programme that can be used by the palm oil mills to incorporate within their IoT and be used to remotely monitor and predict methane emissions. Therefore, the main purpose of this research is to develop a CH_4 prediction tool based on an ML algorithm to aid in the monitoring and control of CH_4 emissions in POME treatment plants. To achieve this, a correlation must be built between critical parameters that will affect CH_4 emission based on collected databases to support this tool using several ML tools.

2. METHODOLOGY

To develop the prediction tool, a methodology roadmap has been constructed to highlight the steps taken to achieve it. The methodology consists of six sections which are data collection, data pre-processing and data preparation, model development, model performance, model testing, and programme development using Figure 1.

2.1. Data collection

To develop the CH_4 emission prediction tool, a dataset is required as a foundation to kickstart the research. Without a dataset, there can be no programme development to fulfil the aims and objectives of the study as stated by Dekker (2006) in an article.

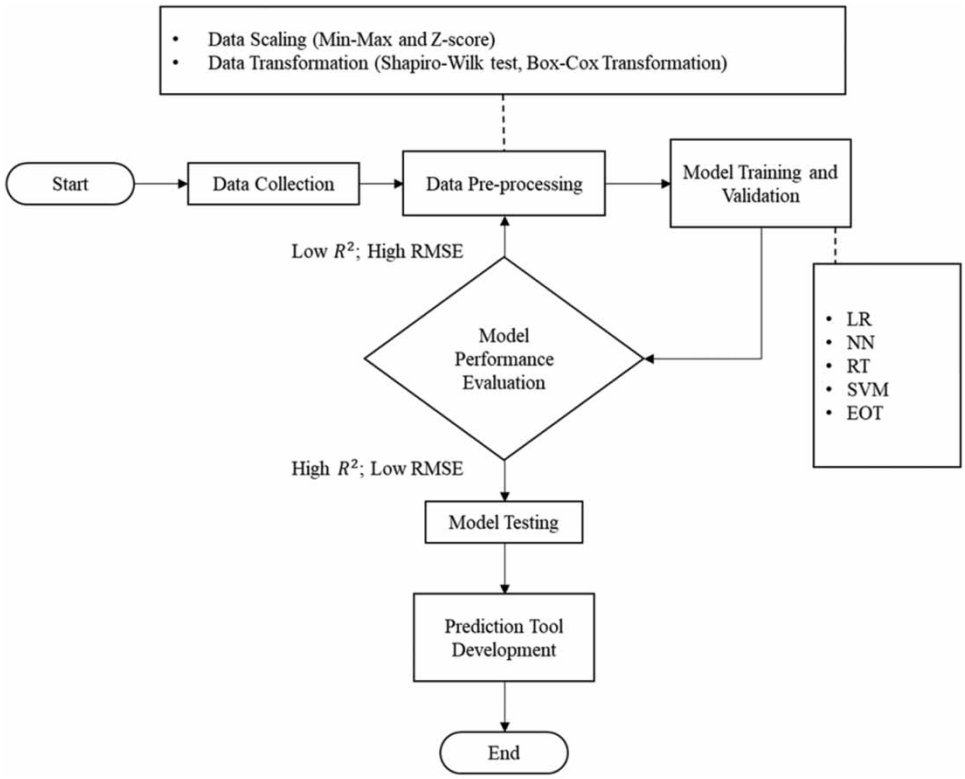


Figure 1 | Machine learning tool roadmap.

Therefore, to foster the development of the prediction tool, datasets were obtained from four different palm oil mills across Malaysia. Each of the datasets obtained primarily focuses on POME treatment and compromises 24-monthly data points which ranged from 2019 to 2021. The datasets obtained contained the main process parameters such as COD, BOD_5, TS, SS, pH, temperature, OLR, and HRT affecting the quality of POME and final quality of biogas produced at the end of each month. All the parameters obtained from the palm oil mills were used as input variables into the models as they have a varying but crucial influence on the biogas production. As mentioned by Loh *et al.* (2017), since the biogas obtained predominantly consists of CH_4, the scope of research is to predict the amount of CH_4 that will be released at the end of POME treatment. An overall summary of the datasets obtained and combined is provided in Table 1.

Table 1 | Summary of obtained dataset

Parameter	Unit	Range
COD	mg/L	53,450–92,844
BOD_5	mg/L	22,500–47,520
TS	mg/L	20,148–56,420
SS	mg/L	12,300–5,7650
pH	–	4.6–5.2
Temperature	°C	46.8–62.4
OLR	kg COD_{in}/m^3 day	0.9–1.8
HRT	days	34–88
CH_4	Nm^3	11,340–295,480

2.2. Data pre-processing and preparation

The process of changing data from one version to another is known as data transformation. The most common data transformations are those that convert raw data into a clean, usable format. An analyst will use exploratory data analysis to analyse the data behaviour, retrieve data from the original source, perform the transformation, and lastly save the data in the proper database throughout the data transformation process. This involves methods such as normalisation, standardisation, and much more.

2.2.1. Normalisation

Normalise, as the name suggests, is the action to give an attribute from any dataset the same weightage (Han *et al*. 2011). Data normalisation can be divided into numerous categories. According to McCaffrey (2020), min–max normalisation, *z*-score normalisation, and constant factor normalisation are the three most prevalent forms of normalisation used in ML. The most popular method among the three types of data normalisation is min–max normalisation as it is simple and straightforward to understand the mathematical concept. Min–max normalisation performs a linear transformation where the minimum value of each dataset is converted to a 0, while the highest value is converted to a 1 (Han *et al*. 2012). The mathematical formula for min–max normalisation is described using Equation (1) (Han *et al*. 2012):

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A}(1.0 - 0) + 0 \tag{1}$$

where v_i' is the normalised value; v_i is the value in dataset; $\min_A$ is the minimum value in dataset; and $\max_A$ is the maximum value in dataset.

As mentioned, the dataset shown in Table 1 was transformed using min–max normalisation using the normalise (x, 'range', [0 1]) syntax provided in MATLAB® 2022a (MathWorks 2022b). The syntax mentioned utilised the equation provided by Equation (3) and was further subjected to the regression learner toolbox provided by MATLAB® 2022a. The performance of each model present within the regression learner is tabulated and provided in the results and discussion section.

2.2.2. Standardisation

Standardisation or *z*-score normalisation is a technique which utilises the mean and standard deviation for each variable across a set of training data (Jayalakshmi & Santhakumaran 2011). To carry out standardisation, it is required for data analysts to compute the mean and standard deviation of a dataset (Saleh & Fleyeh 2022). Once computed, Equation (2) is applied to execute the standardisation process:

$$X_{is} = \frac{X_i - \bar{X}}{\sigma} \tag{2}$$

where X_{is} is the standardised value; X_i is the value in dataset A; $\bar{X}$ is the mean of dataset; and σ is the standard deviation of dataset.

Upon applying Equation (2), the structure of a dataset will be converted into a single, standardised data format where the new mean and standard deviation values are 0 and 1 (Liu 2020). Unlike min–max normalisation, the standardised values of each variable in a dataset can take a positive or negative value (Elen & Avuçlu 2021). This is justified as the absolute value of X_{is} is defined as the distance in standard deviation units between the raw score X_i and mean X. When the raw score is below the mean, X_{is} is negative and positive for vice versa. The advantage of applying standardisation as a method of data transformation is because it can minimise the effects of outliers present in any dataset (Jayalakshmi & Santhakumaran 2011).

Following normalisation**,** the same dataset in Table 1 was also transformed using standardisation using the normalise(x) syntax provided in MATLAB® 2022a (MathWorks 2022b). The syntax obeyed the equation provided by Equation (2) and the standardised data was further subjected to the regression learner toolbox provided by MATLAB® 2022a. Following this, the performance of each model found within the regression learner is tabulated and provided in the Results and Discussion section. By observing the performance of the models obtained using the regression learner, the prediction capabilities of the *tool* can be further improvised by revisiting the pre-processing and preparation of the dataset.

2.2.3. Exploratory data analysis (EDA)

Tukey (1977) defines EDA as numerical detective work. In terms of engineering statistics, EDA refers to the crucial process of doing preliminary data analysis utilising summary statistics and visualisations to find trends. EDA's major goal is to assist data analysts before making any assumptions. EDA is capable of detecting obvious errors, providing a clearer understanding of data patterns, the detection of outliers, and the discovery of variable correlations (IBM Cloud Education 2020). The commonly used methods to carry out EDA on datasets are Quantile-Quantile plots (QQ plot), normality tests, and skewness tests (Seltman 2018).

2.2.4. Shapiro–Wilk test (normality test)

Based on a book titled 'Testing for Normality' by Thode (2002), the Shapiro–Wilk test is recommended to be used as the primary choice for normality testing as it performs well in every aspect and for various applications. The primary function of the Shapiro–Wilk test is to understand if a random sample is derived from a normal distribution or not (King & Eckersley 2019). According to the test, the assumption of normality is driven by the W value following a significance level, α. W values below α suggests that the sample is not evenly distributed while opposite results of W translate that the data is normally distributed (Fang & Yang 2019). The general mathematical formula used to compute the W value is provided in Equation (3) (Looney & Hagan 2011):

$$W = \frac{\left[\sum_{i=1}^{n} a_i x_{[i]}\right]^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{3}$$

where x_i is observations arranged in ascending order; $\bar{x}$ is the mean of distribution; and a_i is constants as a function of n.

2.2.5. Skewness test

Skewness is a statistic in engineering statistics that measures the asymmetry of a random variable's probability distribution around its mean. In other words, skewness is a term used to describe the degree of departure from horizontal symmetry. Skewness takes the form of different types, namely positive, negative, or zero. It is impossible for a dataset to have a skewness value of zero as it resembles perfect symmetry (Klima 2021). The skewness of any dataset can be computed using Equation (4) provided by Weinberg (2016):

$$\text{Skewness} = \frac{\sum (X_i - \bar{X})^3 / N}{\sqrt{\sum (X_i - \bar{X})^2 / N}^3} \tag{4}$$

where X_i is the random variable; $\bar{X}$ is the mean of distribution; and N is the number of variables in dataset.

In statistics, there are a few ways to identify types of skewness. The rule pointed out by Bulmer (1967) says skewness values ranging between 0 and 0.5 indicate a generally symmetrical distribution, values between 0.5 and 1 indicate a moderately skewed distribution, and values greater than 1 indicate a strongly skewed distribution. This rule also applies to the negative skewness values as well. According to Orcan (2020), by utilising this rule, skewness can determine the normality of a dataset. Furthermore, it has been confirmed by IBM (2020) and Lee (2020) that certain transformations may be used to prepare datasets for ML according to the type of data skewness.

Shapiro–Wilk tests for normality and skewness tests were done on each parameter within the dataset using IBM® SPSS® Statistics v26 software. By using the software, the statistical tests can determine if all the parameters should be normalised or standardised for the tool or different modes of transformation methods are required to be used for specific parameters. The results of the statistical tests are provided in the Results and Discussion section.

2.2.6. Quantile-Quantile plot (QQ plot)

The University of Virginia Library (2022) defines the QQ plot as a graphical tool to determine if a set of data obeys theoretical distribution such as normal or exponential distribution. In modern data analysis, if a dataset contains a large number of samples, $N \geq 30$, assumptions that the dataset follows a normal distribution is safe and a good procedure (Toby Mordkoff 2000). QQ plots are useful because they allow data analysts to rapidly assess whether the assumption is reasonable, and if

not, how the assumption is flawed, and which distribution should be used instead. QQ plots are also able to guide data analysts to use appropriate transformations according to the distribution plot observed (Samuels *et al.* 2021). In modern engineering statistics, software such as MATLAB® and SPSS® are frequently used to plot various QQ plots.

Based on the results observed from the statistical tests, parameters that do not obey normality are addressed by referring to the skewness values. Based on the skewness, the types of transformations to be used were then decided. For generally symmetrical skewness values, standardisation was used following the skewness rule under Results and Discussion. QQ plots were also used to verify the transformation using standardisation is deemed valid. Meanwhile, parameters that were found to be moderately and strongly skewed based on the same skewness rule following Results and Discussion were then subjected to QQ plots to determine the type of distribution obeyed. The QQ plot distributions used were normal distribution, lognormal distribution, Weibull distribution, logistic distribution, and Gamma distribution.

Following the QQ plots shown under Results and Discussion, one variable strongly obeys the lognormal distribution while the remaining two variables were seen to obey the logistic distribution. The parameter that obeyed the QQ plot for lognormal distribution was transformed using the natural logarithm function log(x) syntax provided in MATLAB® 2022a (MathWorks 2022a). For the variables that followed a logistic distribution, common transformation methods such as square root, exponential, and inverse were used and verified using the Shapiro–Wilk test. As the Shapiro–Wilk test deemed that these transformations were not proper for the remaining two variables, the Box-Cox transformation was used. The Box-Cox transformation was done using MATLAB® 2022a boxcox (x) syntax.

Following all the transformations carried out for each parameter in the dataset, the parameters that were subjected to transformations aside from standardisation are standardised once more before being used for training in the regression learner. The results obtained based on the performance of each model are provided in Results and Discussion.

3. RESULTS AND DISCUSSION

3.1. Normalisation and standardisation

After normalisation was carried out for the entire dataset and used in the regression learner for training, the best performance for each regression technique with kernel was recorded and tabulated using Table 2.

Next, the performance of the best models with kernels found in the regression learner was recorded based on the standardised dataset. Table 3 displays the recorded performance.

Despite the effort of trying to equalise the range and achieve a consistent trend for each feature using both normalisation and standardisation, the prediction capabilities of ML techniques were inadequate and required to be revisited as seen under

Table 2 | Regression learner result using normalisation

Model	RMSE	R^2	MSE	MAE
SVM	0.157	0.38	2.45×10^{-2}	0.115
Ensemble of trees	0.164	0.32	2.68×10^{-2}	0.128
Linear regression	0.173	0.24	3.00×10^{-1}	0.136
Regression tree	0.195	0.04	3.81×10^{-2}	0.149
Neural network	0.276	-0.93	7.60×10^{-2}	0.217

Table 3 | Regression learner result using standardisation

Model	RMSE	R^2	MSE	MAE
SVM	0.758	0.43	5.74×10^{-1}	0.561
Ensemble of trees	0.823	0.32	6.78×10^{-1}	0.659
Linear regression	0.866	0.25	7.50×10^{-1}	0.683
Regression tree	0.974	0.06	9.49×10^{-1}	0.780
Neural network	1.254	−0.57	1.57	0.933

Tables 2 and 3. It can also be seen that the performance of ML models using standardisation is slightly outperforming normalisation in terms of R^2 values. This is proven through a study carried out by Singh & Singh (2020) on different normalisation methods, stating standardisation utilising mean and standard deviation is much more suited to be selected as a first choice when compared to min–max normalisation for prediction purposes. Secondly, when comparing RMSE, MAE, and MSE values for both forms of transformations, normalisation produces significantly fewer errors compared to standardisation. This does not necessarily mean normalisation produces lesser error than standardisation. The values based on errors are such because the way the data was transformed is different. Normalisation as mentioned ensures the dataset in between number ranges of [0–1] but standardisation fits the data according to the mean and standard deviation of the dataset. The key difference between both forms of transformation is that standardisation scales the features of the dataset into a common, flexible version without distorting the differences in a range of values while normalisation distorts the data by restricting the scale. To justify the statement previously, standardisation can address outliers and noise unlike normalisation which is sensitive to outliers and able to promote distortion in the dataset (Chanal *et al.* 2022). To avoid confusing the ML algorithms and achieve the overall goal of developing a CH_4 prediction tool, standardisation should be chosen as the primary choice for normalising data in developing the model as it has much more superiority in terms of data representation.

Despite the comparison between both forms of transformation methods, the overall performance of ML models cannot be classified as a high performing model as mentioned in the paragraph earlier. This is primarily due to several reasons such as the parameters exhibiting different behaviour and properties which cannot be addressed using both normalisation and standardisation purely (Singh & Singh 2020). In addition to that, some parameters found within the dataset may be transformed into a better version to ease the training process for ML models while the remaining transformed parameters may inhibit the performance capabilities instead. As a result, the performance metrics would return a low R^2 and high RMSE value. Therefore, as a solution, Singh & Singh (2020) proposed that transformations applied towards different parameters found in datasets should address their properties instead of generalising a single transformation method to be used in every scenario.

3.2. Exploratory data analysis and transformations

As mentioned in the sections earlier, Shapiro–Wilk test and skewness tests were done using IBM® SPSS® Statistics v26 software to understand the behaviour of each parameter following the findings using normalisation and standardisation transformations. The results of each parameter after undergoing the tests are tabulated below:

Based on the information shown in Table 4, only three parameters, pH, OLR, and HRT, are normally distributed as they returned a '0' value after undergoing the Shapiro–Wilk test. Following that, the Shapiro–Wilk test is determined by W (p-value) with respect to the significance level, α. In data science and hypothesis testing, the common significance level used in every statistical software is 5% ($\alpha = 0.05$). This translates that there exists a 5% risk that the null hypothesis would be rejected (Frost 2020). Based on the Shapiro–Wilk test, an α value of 0.05 also means that the null hypothesis is 95% confident that the data follows a normal distribution and would require strong evidence to reject it. This shows the p-value is crucial to understand the behaviour of each parameter as it reflects the compatibility towards normality (Di Leo & Sardanelli 2020).

Table 4 | Shapiro–Wilk and skewness test result

Parameter	SW	Skewness	*p*-value
COD IN	1	−0.065	0.001
BOD_5 IN	1	−0.029	0.000
pH	0	0.382	0.370
TEMP	1	−0.576	0.000
SS IN	1	1.374	0.000
TS IN	1	0.092	0.000
OLR	0	0.070	0.506
HRT	0	−0.160	0.151
CH_4	1	−0.773	0.004

From observing the table of results, the *p*-value of the three variables is found to be greater than 0.05, which accepts the null hypothesis stating the data is normally distributed (Ramachandran & Tsokos 2021). Therefore, the method selected to transform the three parameters before being used in the regression learner was standardisation. Next, the remaining six parameters failed to exhibit normality as they returned a number '1' following the Shapiro–Wilk test. This is because the *p*-value of each parameter yields a result which is lower than 0.05, translating that the remaining parameters are not normally distributed according to King & Eckersley (2019).

To address the remaining six variables that are non-normal, the skewness values from Table 4 were used in reference to EDA, in order to determine the type of skewness the dataset for each parameter possessed. It can be said that from the six, three parameters which are COD IN, BOD_5 IN, and TS IN can be classified as generally symmetrical due to their skewness values being closer to zero. According to Brown (2022), skewness values close to zero indicate that the three datasets are close to being normally distributed. Therefore, standardisation can be used to transform these parameters.

The use of QQ plot will be able to verify the type of distribution obeyed for the parameters, TEMP, SS IN, and CH_4, which are moderately and greatly skewed. The distribution types obeyed by these parameters are summarised using Table 5 while visual representation using QQ plot was also provided using Figures 2–4.

As observed from Table 5, both moderately distributed parameters, TEMP and CH_4, followed a logistic distribution. As there are no transformations within the MATLAB® 2022a software to address this distribution type, Box-Cox transformation was chosen as the transformation technique to address this issue since power-based transformations can reduce left skewness

Table 5 | Distribution of parameter

Parameter	QQ plot distribution
TEMP	Logistic
SS IN	Lognormal
CH_4	Logistic

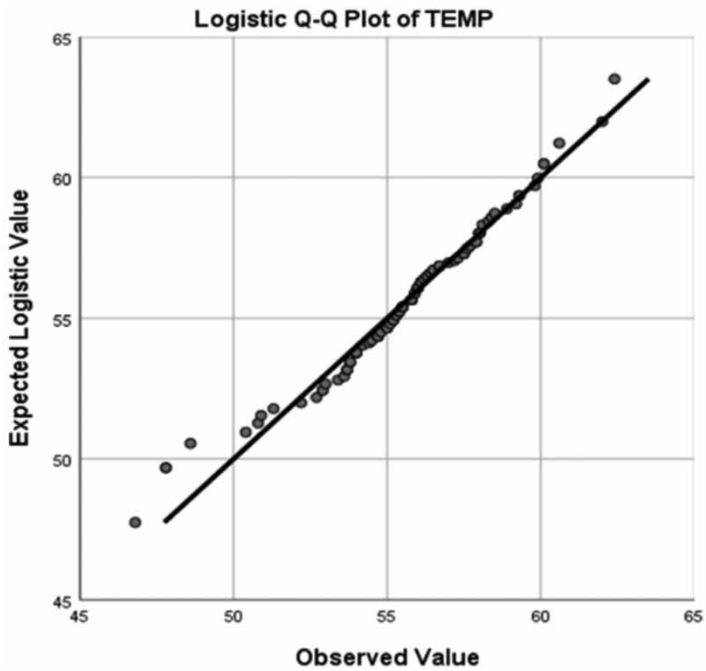


Figure 2 | QQ plot for temperature.

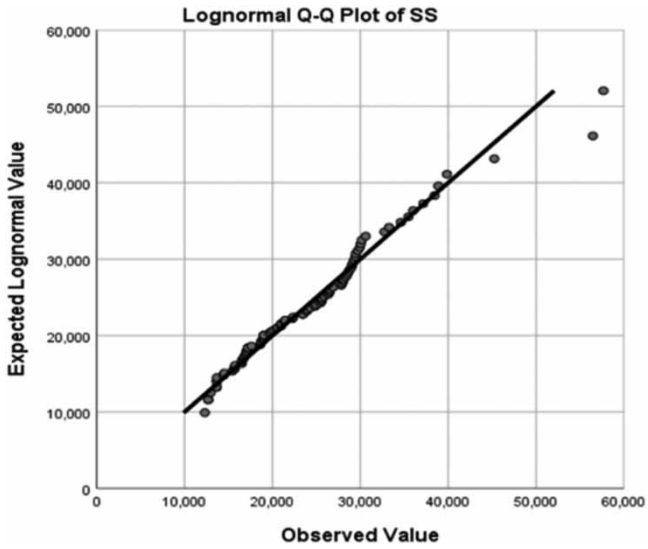


Figure 3 | QQ plot for SS.

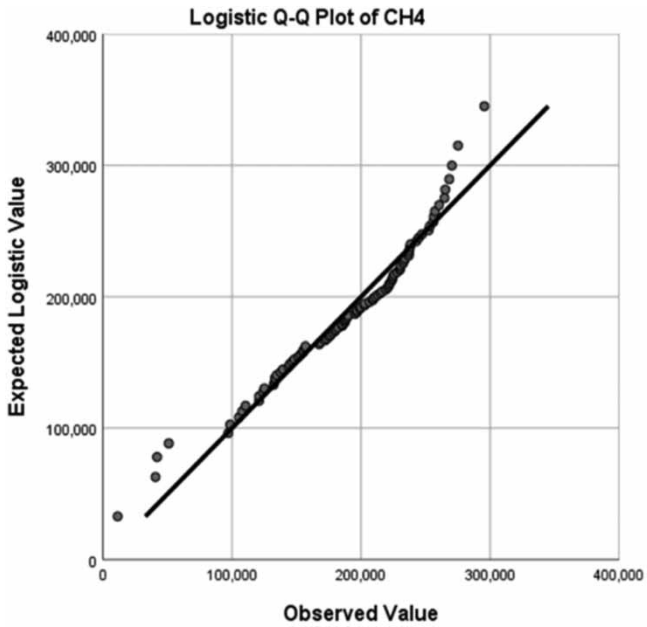


Figure 4 | QQ plot for methane.

(Watthanacheewakul 2021). In addition to that, using Box-Cox transformations can improve forecasting abilities of time series-based datasets, which is an advantage when considering methods to improve the performance of the ML model.

Lastly, for greatly skewed data on parameter SS IN, the QQ plot result observed in Figure 2 shows that it follows the log-normal distribution strongly. To support the fact that the data follows lognormal distribution, the SS IN parameter has a skewness value of 1.374, meaning that it is skewed to the right and matches the description made. Therefore, the transformation suited to address this distribution is without doubt the natural logarithm (Watthanacheewakul 2021). With the use of natural logarithm, it reduces the right skewness of the dataset into a normalised version.

Following all these transformations, the transformed dataset was sent to the regression learner to obtain the performance metrics of the best models with kernels and tabulated below:

Table 6 | Regression learner result using EDA

Model	RMSE	R^2	MSE	MAE
SVM	0.749	0.45	5.61×10^{-1}	0.566
Ensemble of trees	0.858	0.27	7.35×10^{-1}	0.667
Linear regression	0.877	0.24	7.69×10^{-1}	0.683
Regression tree	0.996	0.02	9.92×10^{-1}	0.773
Neural network	1.183	−0.38	1.40	0.909

When compared to both Tables 2 and 3, Table 6 shows minimal improvement in terms of R^2 for SVM-based models even after exhausting different forms of data analysis, statistical tests, and transformation carefully. Although the MSE and RMSE have improved for the SVM model using Medium Gaussian kernel, the differences observed are not significant and the model still requires further improvement to be classified as a high performing model considered for the prediction tool. The prime reason contributing to the inability to produce a high performing model is the lack of more data points in each parameter within the dataset. As obtaining new datasets by visiting the palm oil mills and carrying out tests is unlikely due to COVID-19 restrictions, data augmentation was seen as a promising alternative. Naturally, having more data can aid the ML model to capture a better pattern to improve the prediction capabilities of CH_4. Therefore, data augmentation enables the possibility of achieving this by generating new modified, synthetic data from the limited observed data obtained from the POME plants, while minimising the occurrence of overfitting (Maharana *et al*. 2022).

This study is the first in a series of parallel research currently being conducted by the authors of this paper on developing a ML-assisted prediction tool to measure and monitor methane emissions from POME treatment facilities. The studies following the presented project in this paper are focusing on (1) validating the final methane emission prediction model with special emphasis on its potential in biogas capturing and recovery and (2) addressing different species of emissions generated and released through anaerobic digestion processes in POME treatment. The findings from the above-mentioned studies together with the outcomes of this research will facilitate the development of an advanced emissions prediction tool to be used for the accurate measurement of the total carbon footprint of POME treatment process.

4. CONCLUSION

In this research, a prediction tool was developed with the intention to aid the palm oil industry to monitor CH_4 emissions and provide a potential pathway towards CH_4 capture. SVM, ensemble of trees, LR, neural network, and regression tree were the various ML models explored. Transformation techniques in the form of standardisation, normalization, and transformation were used. The models were found to give the best performance when exploratory data analysis and transformation were done. SVM was the best performing model with an R^2 and RMSE of 0.45 and 0.75. Further studies on the subject matter should focus on exploring data augmentation to increase the number of data available for training and testing the prediction model. Besides that, other ML models such as GPR can also be explored to improve the model performance.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abd Ghani, M. 2021 *8 Things to Know About Palm Oil | WWF*. Available from: https://www.wwf.org.uk/updates/8-things-know-about-palm-oil (accessed 4 April 2022).

Abdurahman, N. H., Rosli, Y. M. & Azhari, N. H. 2013 The performance evaluation of anaerobic methods for palm oil mill effluent (POME) treatment: a review. *International Perspectives on Water Quality Management and Pollutant Control*. doi:10.5772/54331.

Akhbari, A., Kutty, P. K., Chuen, O. C. & Ibrahim, S. 2020 A study of palm oil mill processing and environmental assessment of palm oil mill effluent treatment. *Environmental Engineering Research* **25** (2), 212–221. doi:10.4491/EER.2018.452.

Appels, L., Baeyens, J., Degrève, J. & Dewil, R. 2008 Principles and potential of the anaerobic digestion of waste-activated sludge. *Progress in Energy and Combustion Science* **34** (6), 755–781. doi:10.1016/J.PECS.2008.06.002.

Asadi, M. & McPhedran, K. 2021 Biogas maximization using data-driven modelling with uncertainty analysis and genetic algorithm for municipal wastewater anaerobic digestion. *Journal of Environmental Management* **293**, 112875. doi:10.1016/J.JENVMAN.2021.112875.

Bachrun, R. & Baskara, S. 2023 Greywater flow characteristics for closed channel maintenance. *Civil Engineering Journal (Iran)* **9** (1), 29–40. doi:10.28991/CEJ-2023-09-01-03.

Brown, S. 2022 *Measures of Shape: Skewness and Kurtosis*. Available from: https://brownmath.com/stat/shape.htm (accessed 22 April 2022).

Bulmer, M. 1967 *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York.

CABI 2019 *Invasive Species Compendium*. Available from: https://www.cabi.org/isc/datasheet/20295 (accessed 4 April 2022).

Chanal, D., Yousfi Steiner, N., Petrone, R., Chamagne, D. & Péra, M.-C. 2022 Online diagnosis of PEM fuel cell by fuzzy C-means clustering. *Encyclopedia of Energy Storage* 359–393. doi:10.1016/B978-0-12-819723-3.00099-8.

Choong, Y. Y., Chou, K. W. & Norli, I. 2018 Strategies for improving biogas production of palm oil mill effluent (POME) anaerobic digestion: a critical review. *Renewable and Sustainable Energy Reviews* **82**, 2993–3006. doi:10.1016/j.rser.2017.10.036.

climateemergencydeclaration.org. 2022 *Climate Emergency Declarations in 2,089 Jurisdictions and Local Governments Cover 1 Billion Citizens – Climate Emergency Declaration*. Available from: https://climateemergencydeclaration.org/climate-emergency-declarations-cover-15-million-citizens/ (accessed 6 April 2022).

Dalm, N. 2021 *Msia Intends to Reduce Greenhouse gas Emission by 45 pct by 2030*. Available from: https://www.nst.com.my/news/nation/2021/10/735618/msia-intends-reduce-greenhouse-gas-emission-45-pct-2030 (accessed 6 April 2022).

Dekker, R. 2006 *(1) The Importance of Having Data-sets*. Available from: https://www.researchgate.net/publication/42581175_The_importance_of_having_data-sets (accessed 17 April 2022).

Department of Environment 2010 Environmental Requirements: A Guide for Investors. Department of Environment Ministry of Natural Resources and Environment Wisma Sumber Asli, Precinct 4 Federal Government Administrative Centre 62574 PUTRAJAYA.

Department of Environment Malaysia 1979 Environmental Quality (Sewage).

Diffenbaugh, N. S., Field, C. B., Appel, E. A., Azevedo, I. L., Baldocchi, D. D., Burke, M., Burney, J. A., Ciais, P., Davis, S. J., Fiore, A. M., Fletcher, S. M., Hertel, T. W., Horton, D. E., Hsiang, S. M., Jackson, R. B., Jin, X., Levi, M., Lobell, D. B., McKinley, G. A., Moore, F. C., Montgomery, A., Nadeau, K. C., Pataki, D. E., Randerson, J. T., Reichstein, M., Schnell, J. L., Seneviratne, S. I., Singh, D., Steiner, A. L. & Wong-Parodi, G. 2020 The COVID-19 lockdowns: a window into the earth system. *Nature Reviews Earth & Environment* **1** (9), 470–481. doi:10.1038/s43017-020-0079-1.

Di Leo, G. & Sardanelli, F. 2020 Statistical significance: *p* value, 0.05 threshold, and applications to radiomics – reasons for a conservative approach. *European Radiology Experimental* **4** (1), 1–8. doi:10.1186/S41747-020-0145-Y/METRICS.

Elen, A. & Avuçlu, E. 2021 Standardized variable distances: a distance-based machine learning method. *Applied Soft Computing* **98**, 106855. doi:10.1016/J.ASOC.2020.106855.

EPA 2012 *5.2 Dissolved Oxygen and Biochemical Oxygen Demand | Monitoring & Assessment | US EPA*. Available from: https://archive.epa.gov/water/archive/web/html/vms52.html (accessed 4 April 2022).

EPA 2021 *Importance of Methane | US EPA*. Available from: https://www.epa.gov/gmi/importance-methane (accessed 6 April 2022).

EPA 2022 *How Does Anaerobic Digestion Work? | US EPA*. Available from: https://www.epa.gov/agstar/how-does-anaerobic-digestion-work (accessed 6 April 2022).

Fang, K. & Yang, M. 2019 A practical approach to model validation. *Model Engineering for Simulation* 123–161. doi:10.1016/B978-0-12-813543-3.00007-X.

Frost, J. 2020 *Hypothesis Testing: An Intuitive Guide for Making Data Driven Decisions*.

Han, J., Kamber, M. & Pei, J. 2011 *Data Mining. Concepts and Techniques*, 3rd edn. The Morgan Kaufmann Series in Data Management Systems, Burlington, MA.

Han, J., Kamber, M. & Pei, J. 2012 Data preprocessing. *Data Mining* 83–124. doi:10.1016/B978-0-12-381479-1.00003-4.

Hassan, M. A. & Abd-Aziz, S. 2012 Waste and environmental management in the Malaysian palm oil industry. *Palm Oil: Production, Processing, Characterization, and Uses* 693–711. doi:10.1016/B978-0-9818936-9-3.50026-5.

IBM 2020 *Transforming Variable to Normality for Parametric Statistics*. Available from: https://www.ibm.com/support/pages/transforming-variable-normality-parametric-statistics (accessed 11 April 2022).

IBM Cloud Education 2020 *What is Exploratory Data Analysis? | IBM*. Available from: https://www.ibm.com/cloud/learn/exploratory-data-analysis (accessed 10 April 2022).

IMARC Group 2021 *Palm Oil Market Analysis, Size, Trends and Forecast 2022–2027*. Available from: https://www.imarcgroup.com/palm-oil-processing-plant (accessed 4 April 2022).

Jayalakshmi, T. & Santhakumaran, A. 2011 Statistical Normalization and Back Propagation for Classification. doi:10.7763/IJCTE.2011.V3.288.

Joshi, D. 2019 Evaluating the Performance of the Sustainable Energy Development Authority (SEDA) and Renewable Energy Policy in Malaysia.

Kamyab, H., Chelliapan, S., Fadhil, M., Din, M., Rezania, S., Khademi, T. & Kumar, A. 2018 Palm oil mill effluent as an environmental pollutant. *Palm Oil* doi:10.5772/INTECHOPEN.75811.

Kazemi Yazdi, S. & Scholz, M. 2010 Assessing storm water detention systems treating road runoff with an artificial neural network predicting fecal indicator organisms. *Water, Air, and Soil Pollution* **206** (1–4), 35–47. doi:10.1007/S11270-009-0084-Y/FIGURES/6.

Khairul, M., Sarwani, I., Fawzi, M., Osman, S. A. & Nasrin, A. B. 2019 Bio-methane from palm oil mill effluent (POME): transportation fuel potential in Malaysia. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences Journal Homepage* **63**, 1–11. Available from: www.akademiabaru.com/arfmts.html (accessed 6 April 2022).

King, A. P. & Eckersley, R. J. 2019 Inferential statistics IV: choosing a hypothesis test. *Statistics for Biomedical Engineers and Scientists* 147–171. doi:10.1016/B978-0-08-102939-8.00016-5.

Klima, K. 2021 *Normality Testing – Skewness and Kurtosis | The GoodData Community*. GoodData Corporation, San Francisco, CA. Available from: https://community.gooddata.com/metrics-and-maql-kb-articles-43/normality-testing-skewness-and-kurtosis-241 (accessed 11 April 2022).

Lai, O.-M., Tan, C. & Akoh, C. C. 2016 *Palm Oil Production, Processing, Characterization, and Uses*.

Lam, M. K. & Lee, K. T. 2011 Renewable and sustainable bioenergies production from palm oil mill effluent (POME): win-win strategies toward better environmental protection. *Biotechnology Advances* **29** (1), 124–141. doi:10.1016/J.BIOTECHADV.2010.10.001.

Lee, D. K. 2020 Data transformation: a focus on the interpretation. *Korean Journal of Anesthesiology* **73** (6), 503. doi:10.4097/KJA.20137.

Liu, C. 2020 *Data Transformation: Standardization vs Normalization – Kdnuggets*. Available from: https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html (accessed 14 April 2022).

Loh, S. K., Nasrin, A. B., Azri, M., Adela, N., Muzzammil, N., Jay, D. & Eleanor, S. 2017 Biogas capture – a means of reducing greenhouse gas emissions from palm oil mill effluent. *Oil Palm Bulletin* **75**, 27–36.

Looney, S. W. & Hagan, J. L. 2011 Statistical methods for assessing biomarkers and analyzing biomarker data. *Essential Statistical Methods for Medical Statistics* **27**, 27–65. doi:10.1016/B978-0-444-53737-9.50005-0.

Madaki, Y. S. & Seng, L. 2013 *(1) (PDF) Palm Oil Mill Effluent (POME) from Malaysia Palm Oil Mills: Waste or Resource*. Available from: https://www.researchgate.net/publication/308539738_Palm_oil_mill_effluent_POME_from_Malaysia_palm_oil_mills_Waste_or_resource (accessed 5 April 2022).

Maharana, K., Mondal, S. & Nemade, B. 2022 A review: data pre-processing and data augmentation techniques. *Global Transitions Proceedings* doi:10.1016/J.GLTP.2022.04.020.

Masson-Delmotte, V., Zhai, P., Chen, Y., Goldfarb, L., Gomis, M. I., Matthews, J. B. R., Berger, S., Huang, M., Yelekçi, O., Yu, R., Zhou, B., Lonnoy, E., Maycock, T. K., Waterfield, T., Leitzell, K. & Caud, N. 2021 *Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Available from: www.ipcc.ch (accessed 6 April 2022).

MathWorks 2022a *Natural Logarithm – MATLAB log*. Available from: https://www.mathworks.com/help/matlab/ref/log.html (accessed 18 April 2022).

MathWorks 2022b *Normalize Data – MATLAB Normalize*. Available from: https://www.mathworks.com/help/matlab/ref/double.normalize.html (accessed 17 April 2022).

McCaffrey, J. 2020 *Data Prep for Machine Learning: Normalization – Visual Studio Magazine*. Available from: https://visualstudiomagazine.com/articles/2020/08/04/ml-data-prep-normalization.aspx (accessed 12 April 2022).

Ministry of Energy, G.T. and W. 2017 Green-Technology-Master-Plan-Malaysia-2017-2030.

Ministry of Environment and Water 2020 Malaysia Third Biennial Update Report to the UNFCCC.

MPOC 2020 *Malaysian Palm Oil Industry – MPOC*. Available from: https://mpoc.org.my/malaysian-palm-oil-industry/ (accessed 4 April 2022).

Ng, I. 2017 The impacts of logging and palm oil on aquatic ecosystems and freshwater sources in Southeast Asia. *EnviroLab Asia* **1**, 17. Available from: http://scholarship.claremont.edu/envirolabasiahttp://scholarship.claremont.edu/envirolabasia/vol1/iss3/3 (accessed 5 April 2022).

NST Business 2021 *66 out of 67 FGV Palm oil Mills Using POMIS for Enhanced Productivity*. Available from: https://www.nst.com.my/business/2021/04/682878/66-out-67-fgv-palm-oil-mills-using-pomis-enhanced-productivity (accessed 7 April 2022).

Orcan, F. 2020 Parametric or non-parametric: skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education* **2020** (2), 255–265. doi:10.21449/ijate.656077.

Putro, L. H. S. 2022 Emissions of CH_4 and CO_2 from wastewater of palm oil mills: a real contribution to increase the greenhouse gas and its potential as renewable energy sources. *Environment and Natural Resources Journal* **20** (1), 61–72. doi:10.32526/ENNRJ/20/202100149.

Ramachandran, K. M. & Tsokos, C. P. 2021 Categorical data analysis and goodness-of-fit tests and applications. *Mathematical Statistics with Applications in R* 461–490. doi:10.1016/B978-0-12-817815-7.00011-7.

Saleh, R. & Fleyeh, H. 2022 Using supervised machine learning to predict the status of road signs. *Transportation Research Procedia* **62**, 221–228. doi:10.1016/J.TRPRO.2022.02.028.

Samuels, P., Marshall, E. & Lahmar, J. 2021 *Community Project Encouraging Academics to Share Statistics Support Resources All STCP Resources Are Released Under a Creative Commons Licence Checking Normality for Parametric Tests in SPSS Checking Normality in SPSS*. Available from: www.statstutor.ac.uk.

Sarbatly, R. H. 2020 *Membrane Technology for Water and Wastewater Treatment in Rural Regions*. ICI Global, Hershey, PA.

Seltman, H. J. 2018 *Experimental Design and Analysis*. Illinois Institute of Technology, Chicago, IL.

Singh, D. & Singh, B. 2020 Investigating the impact of data normalization on classification performance. *Applied Soft Computing* **97**, 105524. doi:10.1016/J.ASOC.2019.105524.

Statista 2022a *Global Vegetable Oil Consumption, 2019/20 | Statista*. Available from: https://www.statista.com/statistics/263937/vegetable-oils-global-consumption/ (accessed 4 April 2022).

Statista 2020b *Malaysia: Number of Palm Oil Mills in Operation 2020 | Statista*. Available from: https://www.statista.com/statistics/1093045/malaysia-number-of-palm-oil-mills-in-operation/ (accessed 6 April 2022).

Tambone, F., Scaglia, B., D'Imporzano, G., Schievano, A., Orzi, V., Salati, S. & Adani, F. 2010 Assessing amendment and fertilizing properties of digestates from anaerobic digestion through a comparative study with digested sludge and compost. *Chemosphere* **81** (5), 577–583. doi:10.1016/J.CHEMOSPHERE.2010.08.034.

Thode, H. C. 2002 *Testing for Normality*. CRC Press, Boca Raton, FL.

Toby Mordkoff, J. 2000 *The Assumption(s) of Normality*.

Tukey, J. W. 1977 John W. Tukey – Exploratory Data Analysis-Addison Wesley (1977).

Unfccc 2015 *Adoption of the Paris Agreement – Paris Agreement Text English*.

United States Environmental Protection Agency 2021 *Importance of Methane | US EPA*. Available from: https://www.epa.gov/gmi/importance-methane (accessed 23 May 2022).

University of Virginia Library 2022 *Understanding Q-Q Plots | University of Virginia Library Research Data Services + Sciences*. Available from: https://data.library.virginia.edu/understanding-q-q-plots/ (accessed 11 April 2022).

Utami, I., Redjeki, S., Astuti, D. H. & Sani 2016 Biogas production and removal COD–BOD and TSS from wastewater industrial alcohol (vinasse) by modified UASB bioreactor. *MATEC Web of Conferences* **58**, 01005. doi:10.1051/matecconf/20165801005.

Vaka, M., Walvekar, R., Rasheed, A. K. & Khalid, M. 2020 A review on Malaysia's solar energy pathway towards carbon-neutral Malaysia beyond COVID'19 pandemic. *Journal of Cleaner Production* **273**, 122834. doi:10.1016/J.JCLEPRO.2020.122834.

Ward, A. J., Hobbs, P. J., Holliman, P. J. & Jones, D. L. 2008 Optimisation of the anaerobic digestion of agricultural resources. *Bioresource Technology* **99** (17), 7928–7940. doi:10.1016/J.BIORTECH.2008.02.044.

Watthanacheewakul, L. 2021 *Proceedings of the World Congress on Engineering 2021*.

Weinberg, S. L. 2016 Sharon Lawner Weinberg_ Sarah Knapp Abramowitz – Statistics Using Stata_An Integrative Approach-Cambridge University Press (2016).

Witten, I. H. 2011 *Data Mining Third Edition*. Morgan Kaufmann, Burlington, MA.

World Biogas Association 2018 *How Can Biogas Help Mitigate Climate Change?*. Available from: www.worldbiogasassociation.org (accessed 6 April 2022).

Zafar, S. 2022 *What is POME | BioEnergy Consult*. Available from: https://www.bioenergyconsult.com/tag/what-is-pome/ (accessed 4 April 2022).

Zaied, B. K., Rashid, M., Nasrullah, M., Bari, B. S., Zularisam, A. W., Singh, L., Kumar, D. & Krishnan, S. 2020 Prediction and optimization of biogas production from POME co-digestion in solar bioreactor using artificial neural network coupled with particle swarm optimization (ANN-PSO). *Biomass Conversion and Biorefinery* **2020**, 1–16. doi:10.1007/S13399-020-01057-6.

Zainal, B. S., Gunasegaran, K., Tan, G. Y. A., Danaee, M., Mohd, N. S., Ibrahim, S., Chyuan, O. H., Nghiem, L. D. & Mahlia, T. M. I. 2022 Effect of temperature and hydraulic retention time on hydrogen production from palm oil mill effluent (POME) in an integrated up-flow anaerobic sludge fixed-film (UASFF) bioreactor. *Environmental Technology and Innovation* **28**, 102903. doi:10.1016/j.eti.2022.102903.

Zinder, S. H., Cardwell, S. C. & Anguish, T. 1984 Methanogenesis in a thermophilic (58°C) anaerobic digestor: methanothrix sp. as an important aceticlastic methanogen. *Applied and Environmental Microbiology* **47** (4), 796–807. doi:10.1128/aem.47.4.796-807.1984.

First received 30 January 2023; accepted in revised form 29 May 2023. Available online 9 June 2023