Statistical Analysis And Preprocessing Of Palm Kernel Oil Extraction Machine Dataset For Application In Optimal Yield Model Development

Emmanuel Udama Odeh[1]

Department of Mechanical and Aerospace Engineering
University of Uyo, Akwa Ibom State, Nigeria
emmanuelodeh@uniuyo.edu.ng

Emem Sunday Ezekiel[2]

Department of Mechanical and Aerospace Engineering
University of Uyo, Akwa Ibom State, Nigeria
ememezekiel@gmail.com

OLALEYE, O. Olukayode[3]

Marine Engineering Department,
Maritime Academy, Oron, Akwa Ibom State, Nigeria
kayola_man@yahoo.com

Abstract*—In this work, statistical analysis and preprocessing of palm kernel oil (PKO) extraction machine dataset for application in optimal yield model development is presented. The aim of this present work is to provide statistical analysis approaches that are used to preprocess and evaluate an original dataset and the corresponding augmented dataset generated from the original dataset, to ensure that the augmented dataset is accurate replica of the original dataset and is suitable for application in the machine learning model development. Specifically, an original 125-records dataset empirically collected from a10-ton palm kernel oil (PKO) extractor machine is considered along with 5000-records dataset generated from the 125-records dataset using Generative Adversarial Network (GAN) model. The results of the exploratory data analysis showed that each of the four variables in the original 125-records dataset has no missing value and no outlier; the main shaft speed has a mean of 18 rpm, the cone gap has a mean of 1.5 mm, the moisture content has a mean of 10 %, and the oil yield has a mean of 38.8792 %. Also, the results show that the maximum oil yield occurred at data point 62 which is at main shaft speed of 18 rpm, cone gap of 1.5 mm and moisture content of 8 % and it has oil yield of 43.4 %. The confidence interval results for the original 125-records dataset and the augmented 5000-records dataset show that there is no significant difference in the mean of the two datasets. This shows that the augmented dataset is an accurate replica of the original small size dataset, and hence the augmented dataset can be used in place of the original dataset to train and validate machine learning models meant for the case study a10-ton palm kernel oil (PKO) extractor machine.

Keywords*— *Statistical Analysis, Palm Kernel Oil Extraction Machine, Data Preprocessing, Machine Learning Models, Data Augmentation

1. Introduction

Nowadays, data driven approaches are increasingly being applied in studying various industrial processes and systems [1,2,3]. Moreover, the widespread use of artificial intelligence (AI) models has added to the demand for data driven approaches in studying systems and processes, especially in the industrial sector where automation and cost cutting measures are in high demand [4,5,6]. In practice, when dataset is grossly inadequate for machine learning-based modeling data augmentation can be used [7,8]. In that case, some approaches are used to synthesize additional data records which are then used for the modeling [9,10]. When such data synthesis is conducted, it is advisable to ensure that the

synthesized dataset is an accurate replica of the original dataset from which it was generated. This can be done through statistical analysis of the original dataset and the synthesized dataset.

Also, when datasets are employed in machine learning model-based studies, the dataset is preprocessed using some statistical approaches to make the dataset more suitable for application in the machine model development and to ensure high performance of the machine learning models [11,12]. Accordingly, in this study, the focus is on the dataset obtained from a palm kernel oil (PKO) extracting machine [13,14]. Notably, the dataset empirically obtained from the case study PKO extractor machine is not big enough for machine learning models training and evaluation. As such, data augmentation is conducted on the dataset which led to a larger dataset with numerous synthesized data records [15,16]. The focus in this work is to conduct statistical data analysis and preprocessing for the case study PKO extractor machine dataset with an initial 125-records which was eventually augmented to 5000-records dataset. The analysis is meant to ascertain the suitability of the augmented dataset for application in machine learning model used to determine the optimal PKO yield of the case study plant.

2. Methodology

The aim of this study is to present statistical analysis and preprocessing that are carried out on the dataset empirically obtained from a 10-ton palm kernel oil (PKO) extractor machine. The statistical analysis and preprocessing are conducted on the original 125-records dataset empirically collected from the 10-ton PKO extractor machine. However, the data was augmented to 5000 data records using Generative Adversarial Network (GAN) model. Specifically, this study presents the exploratory data analysis on the original 125 record dataset. It also presents the data normalization using MinMax method, the outlier determination using the quartile and inter quartile range method, and the correlation matrix using the Karl Pearson's coefficient of correlation method. Furthermore, the confidence interval at 95% confidence level is presented for the original dataset and the augmented dataset.

2.1 Data Normalization using Minmax Method

One of the pre-processing task carried out was data normalization using the minmax approach given in Equation 1 where;

$$d_{xN} = \frac{d_i - d_{min}}{d_{max} - d_{min}} \quad (1)$$

Where d_x denote the x^{th} data to be normalized, d_{min} denotes the minimum value, d_{max} denotes the maximum value and d_{xN} denotes the normalized value for d_x.

2.2 Identification of Outliers using Tukey's Fences Method

The Tukey's Fences method is used to identify the outliers in the dataset. Now, consider an ordered data; $d_1, d_2, d_3, \dots, d_n$ having $d_i \leq d_{i+1}$ then, the quartile, $q\left(\frac{p}{4}\right)$ is given as;

$$q\left(\frac{p}{4}\right) = d_k + \alpha(d_{k+1} - d_k) \quad (2)$$

Where, p =1 in the case of the first quartile, p = 2 in the case of the second quartile and p = 3 in the case of the third quartile, while $\lfloor d \rfloor$ indicates the nearest lower integer part of d.

$$k = \lfloor p(n+1)/4 \rfloor \quad (3)$$

$$\alpha = p(n+1)/4 - \lfloor p(n+1)/4 \rfloor \quad (4)$$

So, in the first quartile given as Q1= $q\left(\frac{1}{4}\right) = q(0.25)$, the second quartile we have $Q2 = q\left(\frac{2}{4}\right) = q(0.5)$ and the third quartile is $Q3 = q\left(\frac{3}{4}\right) = q(0.75)$. The IQR (interquartile range) is given as;

$$\text{IQR} = Q3 - Q1 \quad (5)$$

The possible outliers in the dataset are determined with respect to the Upper Fence and the Lower Fence where;

Upper Fence = $Q3 + 1.5(IQR)$ (6)

Lower Fence = $Q1 - 1.5(IQR)$ (7)

Notably, those data > Q3 or < Q1are outlisers.

2.3 Computations of Correlation Matrix using Karl Pearson's Coefficient of Correlation

The Karl Pearson's coefficient of correlation was used to determine the correlation among the four variables that were considered in the study. The correlation is computed between two variables at a time and the process is repeated until all the variables are paired and the correlation parameter is determined. Now, when the dataset consisting of n data items $d_1, d_2, d_3, \ldots, d_n$ is considered, then the mean of the dataset denoted as $\bar{d}$ is given in Equation 8 as;

$$\bar{d} = \frac{1}{n}\left(\sum_{x=1}^{x=n} d_x\right) \quad (8)$$

The standard deviation of the dataset denoted as s is given in Equation 9 as;

$$s = \sqrt[2]{\left\{\left(\frac{1}{n-1}\right)\left(\sum_{x=1}^{x=n}\left(d_x - \bar{d}\right)^2\right)\right\}} \quad (9)$$

Again, consider n elements dataset having $c_1, c_2, c_3, \ldots, c_n$ data points with mean $\bar{c}$ where;

$$\bar{c} = \frac{1}{n}\left(\sum_{x=1}^{x=n} c_x\right) \quad (10)$$

Then, if the two datasets in d_x and c_x are considered as bivariate dataset which is a data having two variables, then the Karl Pearson's coefficent of correlation, $r_{c,d}$ can be detemined for the bivariate data as given in Equation 11;

$$r_{c,d} = \frac{\sum_{x=1}^{x=n}[(c_x-\bar{c})(d_x-\bar{d})]}{\sqrt[2]{\left(\sum_{x=1}^{x=n}(c_x-\bar{c})^2\right)\left(\sum_{x=1}^{x=n}(d_x-\bar{d})^2\right)}} \quad (11)$$

If there are 4 variables, then possible $r_{c,d}$ are shown in Table 1.

Table 1 The visualization of the coefficient of correlation among the four variables considered in the study (where rx,y = ry,x due to the symmetric property)

	Main Shaft Speed (Denoted as MSS)	Cone Gap(Denoted as CG)	Moisture Content (Denoted as MC)	Oil Yield (Denoted as NOY)
Main Shaft Speed (Denoted as MSS)	r1,1	r2,1	r3,1	r4,1
Cone Gap(Denoted as CG)	r1,2	r2,2	r3,2	r4,2
Moisture Content (Denoted as MC)	r1,3	r2,3	r3,3	r4,3
Oil Yield (Denoted as NOY)	r1,4	r2,4	r3,4	r4,4

2.4 Computation of the Confidence Interval (CI)

The confidence interval is used to check if there is significant difference in the original data set and the augmented dataset. Two methods used in the study are;

a) The Z –test method with the population standard deviation and sample mean
b) The confidence interval computation approach for two-unpaired sample tests when the variances can be pooled

(a) Method I : The Z –test with the population standard deviation and sample mean

In order to assess the effectiveness of the original dataset (denoted as OrData) and the GAN augmented dataset (denoted as GaData), the CI is computed at 95 % confidence level (CL) with the following:

i. the population mean (μ) of OrData

ii. the population standard deviation (σ) of OrDat

iii. the sample mean ($\bar{x}$) of GaData

iv. the sample standard deviation (s) of GaData

Let confidence interval be CI , sample mean be sample mean $be\ \bar{x}$, margin of error be EM, the number of samples be n, the confidence level be CL, then,

$$\text{CI} = \bar{x} \pm \text{EM} \quad (12)$$

$$\text{EM} = Z^* \left(\frac{\sigma}{\sqrt{n}}\right) \quad (13)$$

When the CL is expressed in percentage, then, the significance level, α is given as;

$$\alpha = (100 - \text{CL})/100 \quad (14)$$

$$\alpha/2 = \frac{((100-\text{CL})/100)}{2} \quad (15)$$

$$Z^* = Z_{\alpha/2} \quad (16)$$

The value of $Z_{\alpha/2}$ is read from the standard normal distribution table. In this study, CL = 95% and $Z^* = Z_{0.05/2} = Z_{0.025} = 1.96$.

$$\text{CI} = \bar{x} \pm \text{EM} = \bar{x} \pm Z^* \left(\frac{\sigma}{\sqrt{n}}\right) \quad (17)$$

For this research,

$$\text{CI} = \bar{x} \pm 1.96 \left(\frac{\sigma}{\sqrt{n}}\right) \quad (18)$$

The implication of the results from the calculation of CL is that 95% of the time or we are 95% confident that the population mean, μ lies in the interval of $\bar{x} - 1.96 \left(\frac{\rho}{\sqrt{n}}\right)$ and $\bar{x} + 1.96 \left(\frac{\rho}{\sqrt{n}}\right)$. If the result is true that the population mean, μ lies in the interval then we can say that the there is no significant difference between the mean of the GAN generated dataset, GaData and that of the original dataset, OrData.

(b) Method II : The confidence interval computation approach for two-unpaired sample tests when the variances can be pooled

Consider two independent data samples such that:

i. $\bar{x}_1$ as the mean of dataset 1
ii. n_1 as number of samples dataset 1
iii. s_1 as standard deviation dataset 1
iv. $\bar{x}_2$ as the mean of dataset 2
v. n_2 as number of samples of dataset 2
vi. s_2 as standard deviation of dataset 2
vii. Where $n_1 \neq n_2$

Then, the two samples mean can be compared using the confidence interval, CI as follows;

$$s_L = \text{Maximum}(s_1, s_2) \quad (19)$$

$$s_s = \text{Minimum}(s_1, s_2) \quad (20)$$

$$F = \left(\frac{{S_L}^2}{{S_s}^2}\right) \quad (21)$$

If F <4 then Use this method else use the other method.

Now compute the difference between mean, $\bar{x}_*$.

$$\bar{x}_* = |\bar{x}_1 - \bar{x}_2| \quad (22)$$

Degree of freedom, df

$$df = n_1 + n_2 - 2 \quad (23)$$

The pooled variance, ${s_P}^2$ is given as;

$${s_P}^2 = \frac{(n_1-1)({S_1}^2)+(n_2-1)({S_2}^2)}{df} \quad (24)$$

Standard Error of the difference of means (SED) is given as;

$$SED = \sqrt{\frac{{s_P}^2}{n_1} + \frac{{s_P}^2}{n_2}} \quad (25)$$

$$\alpha/2 = \frac{((100-\text{CL})/100)}{2} \quad (26)$$

$$t^* = t_{\alpha/2} \text{ at the given } df \quad (27)$$

$$\text{EM} = t^*(SED) = t^* \left(\sqrt{\frac{{s_P}^2}{n_1} + \frac{{s_P}^2}{n_2}}\right) \quad (28)$$

$$\text{CI} = \bar{x}_* \pm \text{EM} = \bar{x} \pm t^* \left(\sqrt{\frac{{s_P}^2}{n_1} + \frac{{s_P}^2}{n_2}}\right) \quad (29)$$

3. Results and discussion

3.1 The Results for the Exploratory Data Analysis

The original empirically collected 125-records dataset of the PKO extractor machine is shown in Table 2 while the summary of the key descriptive statistical analysis parameters for the case study 125-records dataset parameters are

presented in Table 3. The exploratory analysis showed that each of the four variables in the dataset has 125 data items, no missing value and no outlier. The main shaft speed has a mean value of 18 rpm, while the cone gap has a mean value of 1.5 mm, the moisture content has a mean value of 10 %, and the oil yield has a mean value of 38.8792 %.

Table 2 The original empirically collected 125-records dataset of the PKO extractor machine

S/No,	Main Shaft Speed (RPM)	Cone Gap(mm)	Moisture Content (%)	Oil Yield (%)	S/No,	Main Shaft Speed (RPM)	Cone Gap(mm)	Moisture Content (%)	Oil Yield (%)
1	14	0.5	6	35.9	63	18	1.5	10	43.1
2	14	0.5	8	37.3	64	18	1.5	12	42.8
3	14	0.5	10	36.6	65	18	1.5	14	42.5
4	14	0.5	12	35.9	66	18	2	6	40.4
5	14	0.5	14	33.8	67	18	2	8	41.7
6	14	1	6	36.6	68	18	2	10	41.4
7	14	1	8	38	69	18	2	12	41.1
8	14	1	10	37.3	70	18	2	14	40.8
9	14	1	12	36.6	71	18	2.5	6	38.9
10	14	1	14	35.9	72	18	2.5	8	40
11	14	1.5	6	37.3	73	18	2.5	10	39.7
12	14	1.5	8	38.7	74	18	2.5	12	39.4
13	14	1.5	10	38	75	18	2.5	14	39.1
14	14	1.5	12	37.3	76	20	0.5	6	39.8
15	14	1.5	14	36.6	77	20	0.5	8	41.3
16	14	2	6	36.6	78	20	0.5	10	41
17	14	2	8	37.3	79	20	0.5	12	40.7
18	14	2	10	36.6	80	20	0.5	14	40.4
19	14	2	12	35.9	81	20	1	6	40.9
20	14	2	14	35.2	82	20	1	8	41.6
21	14	2.5	6	35.2	83	20	1	10	41.3
22	14	2.5	8	35.9	84	20	1	12	41
23	14	2.5	10	35.2	85	20	1	14	40.7
24	14	2.5	12	34.2	86	20	1.5	6	41.4
25	14	2.5	14	33.8	87	20	1.5	8	42.4
26	16	0.5	6	35.9	88	20	1.5	10	42.1
27	16	0.5	8	38.6	89	20	1.5	12	41.8
28	16	0.5	10	36.4	90	20	1.5	14	41.5
29	16	0.5	12	35.5	91	20	2	6	40.8
30	16	0.5	14	34.6	92	20	2	8	42.6
31	16	1	6	36.8	93	20	2	10	42.3
32	16	1	8	37.1	94	20	2	12	42
33	16	1	10	36.4	95	20	2	14	41.7
34	16	1	12	35.7	96	20	2.5	6	40.2
35	16	1	14	35	97	20	2.5	8	42
36	16	1.5	6	37.3	98	20	2.5	10	41.7

37	16	1.5	8	38.3	99	20	2.5	12	41.4
38	16	1.5	10	37.6	100	20	2.5	14	41.1
39	16	1.5	12	36.9	101	22	0.5	6	39.5
40	16	1.5	14	36.2	102	22	0.5	8	39.8
41	16	2	6	36.5	103	22	0.5	10	39.5
42	16	2	8	37.1	104	22	0.5	12	39.2
43	16	2	10	36.4	105	22	0.5	14	38.9
44	16	2	12	35.7	106	22	1	6	40.3
45	16	2	14	35	107	22	1	8	40.8
46	16	2.5	6	35.3	108	22	1	10	40.5
47	16	2.5	8	35.9	109	22	1	12	40.2
48	16	2.5	10	35.5	110	22	1	14	39.9
49	16	2.5	12	34.9	111	22	1.5	6	41.1
50	16	2.5	14	34.3	112	22	1.5	8	42
51	18	0.5	6	38.8	113	22	1.5	10	41.5
52	18	0.5	8	40.1	114	22	1.5	12	41
53	18	0.5	10	40.3	115	22	1.5	14	40.5
54	18	0.5	12	39.8	116	22	2	6	40.1
55	18	0.5	14	39.3	117	22	2	8	40.5
56	18	1	6	39.4	118	22	2	10	40.2
57	18	1	8	41.2	119	22	2	12	39.9
58	18	1	10	40.9	120	22	2	14	39.6
59	18	1	12	40.5	121	22	2.5	6	38.4
60	18	1	14	40.1	122	22	2.5	8	38.8
61	18	1.5	6	42.1	123	22	2.5	10	38.5
62	18	1.5	8	43.4	124	22	2.5	12	38.2
63	18	1.5	10	43.1	125	22	2.5	14	37.9

Table 3 The summary of the key descriptive statistical analysis parameters for the case study 125-records dataset parameters

Groups	Main Shaft Speed (RPM)	Cone Gap (mm)	Moisture Content (%)	Oil Yield (%)
Num of observations	125	125	125	125
Num of missing values	0	0	0	0
Minimum	14	14	6	33.8
Maximum	22	22	14	43.4
Range	8	8	8	9.6
Mean ($\bar{x}$)	18	1.5	10	38.8792
Standard Deviation (S)	2.8398	2.8398	2.8398	2.4607
Q1	16	16	8	36.6
Median	18	18	10	39.4
Q3	20	20	12	40.9
Interquartile range	4	4	4	4.3
Outlier	none	none	none	none

3.2 The Results for the MinMax Normalization of the Original 125-Records Dataset

The result of the MinMax normalization of the four variables in the case study dataset is shown in Figure 1. The MinMax normalization results in Figure 1 showed that after normalizing the variables values to values between 0 and 1 and plotting the line charts on a common axis, the main shaft speed and the cone gaps are each

increased in the steps of 25 % of the range of the each of the given parameters, starting with the minimum value to the maximum value of the given parameter. Also, for each main shaft speed setting and each cone gap setting, the moisture content is varied in the step of 25 % of the range of the moisture content, starting with the minimum value to the maximum value of the moisture content parameter.

For each combination of main shaft speed, cone gap and moisture content the oil yield as obtained from the PKO extractor machine is shown in Figure 1. The graph showed that the maximum oil yield occurred at data point 62 which is at main shaft speed of 18 rpm, cone gap of 1.5 mm and moisture content of 8 % and it has oil yield of 43.4 %. In the normalized chart, the maximum oil yield of 1 at main shaft speed of 0.5, cone gap of 0.5 and moisture content of 0.25 and it has normalized oil yield of 1.0.

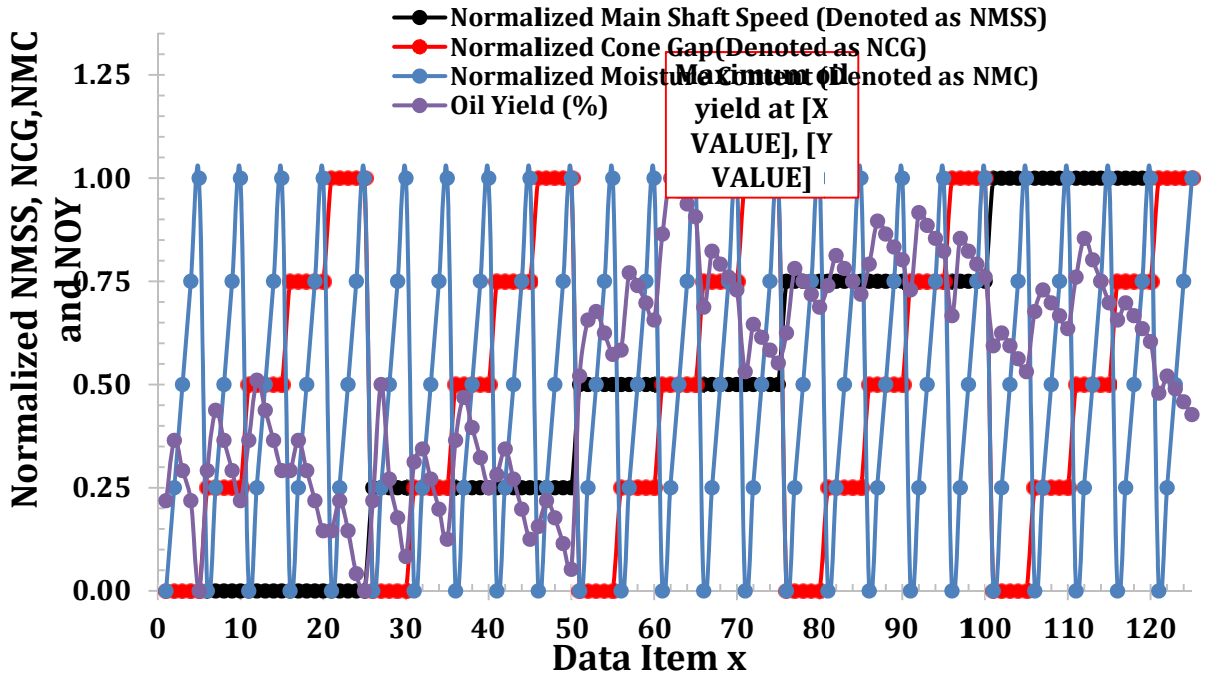

Figure 1 The line chart of the normalized empirically collected dataset from the 10-ton palm kernel oil extraction machine

3.3 The Results of the Correlation Matrix for the Original 125-Records Dataset and the Augmented 5000-Records Dataset

The correlation matrix of the original data is shown in Figure 2 while the correlation matrix of the augmented data is shown in Figure 3. According to Figure 2 and Figure 3, both the original 125-records dataset and the augmented 5000-records dataset have the same correlation results with shaft speed having the highest correlation coefficient of 0.71 with respect to the oil yield in both datasets, the moisture content has correlation coefficient of -0.11 with respect to the oil yield in both datasets, while the cone gap has the least correlation coefficient of -0.056 with respect to the oil yield in both datasets.

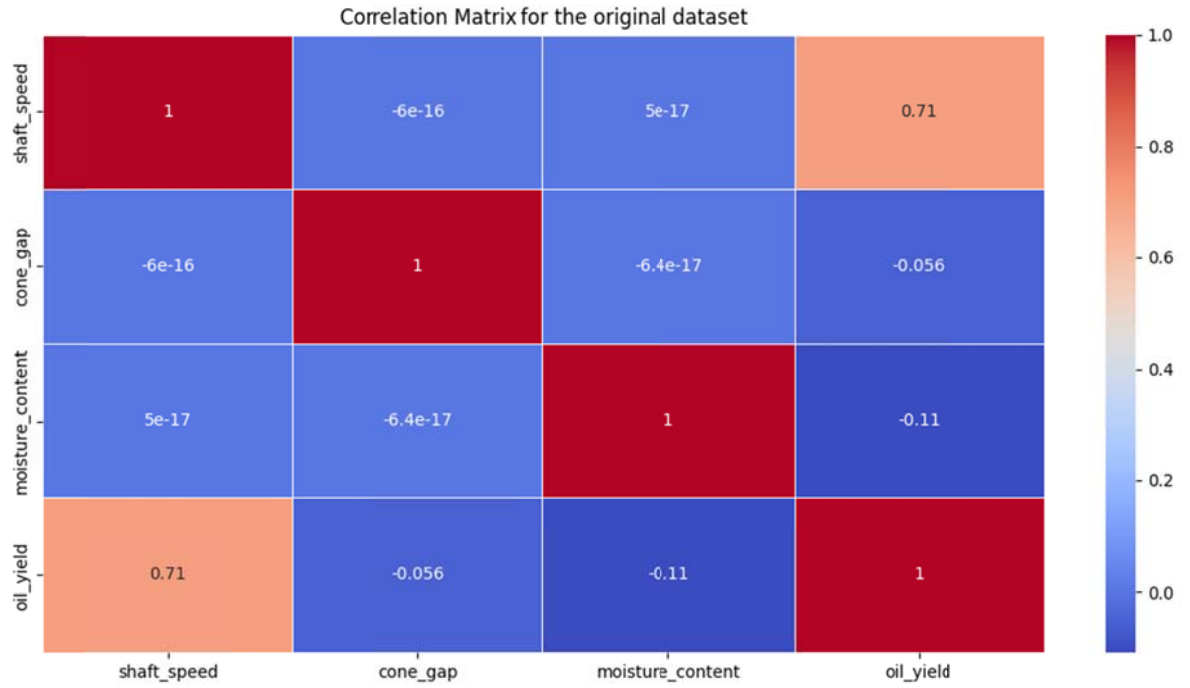

Figure 3: The correlation matrix for the original 125-records dataset

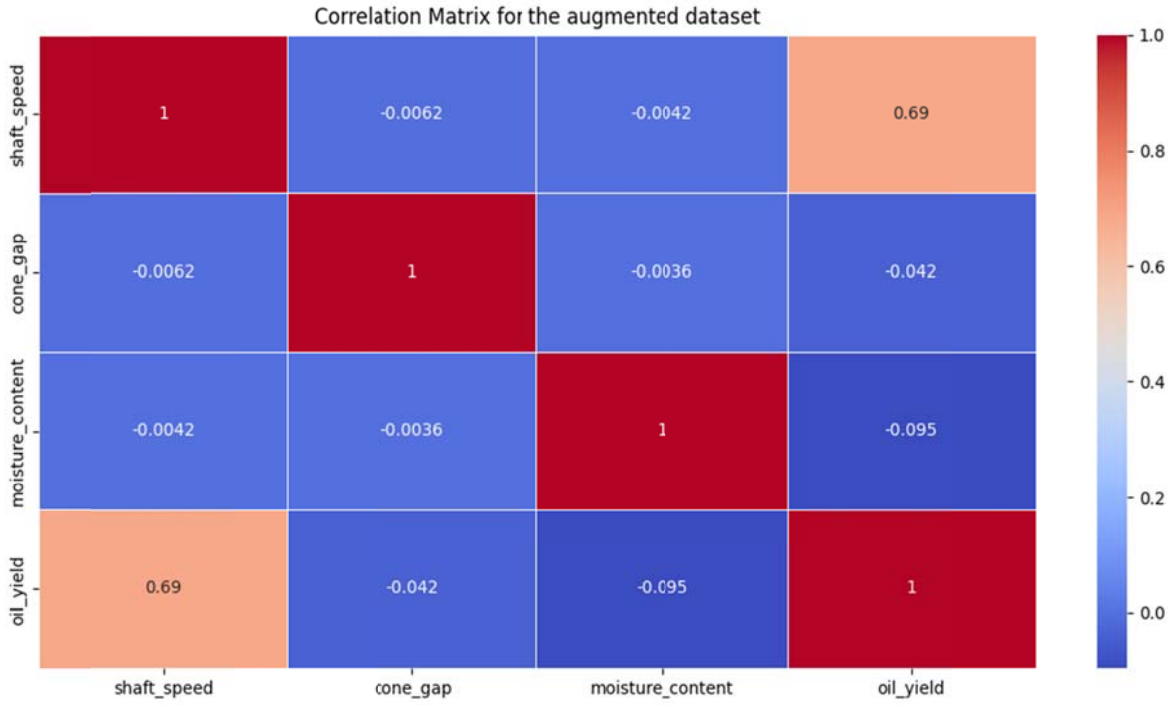

Figure 3: The correlation matrix for the augmented 5000-records dataset

3.4 The Results of the Confidence interval for the Original 125-Records Dataset and the Augmented 5000-Records Dataset

The Confidence Interval (CI) results (Table 2) of the original 125-records dataset and the augmented 5000-records dataset show that for each of the four parameters at 95 % confidence level, there is no significant difference in the mean of the original 125-records dataset and the augmented 5000-records dataset. This is because in all the four parameters the confidence intervals results obtained bracketed the mean value for the original 125-records dataset. For instance, in the case of shaft speed, the original 125-records dataset has mean of 18 rpm and CI of 17.4973 to 18.5027 which bracketed the mean of 18. Also, the augmented 5000-records dataset has CI of 17.9825 to 18.4081 which bracketed the mean of 18.

Similar results are obtained for the four parameters which indicates that the augmented 5000-records dataset maintained the same pattern as the original 125-records dataset.

Table 4 Summary of the Confidence Interval (CI) Analysis Results for the Original 125-Records Dataset and the Augmented 5000-Records Dataset

	Shaft speed	Cone gap	Moisture content	Oil yield
Number of data records in the original dataset	125	125	125	125
Mean of original dataset	18	1.5	10	38.879
Original dataset 95% CI lower value	17.4973	1.3743	9.4973	38.4436
Original dataset 95% CI upper value	18.5027	1.6257	10.5027	39.3148
Number of data records in the augmented dataset	5000	5000	5000	5000
Mean of augmented dataset	18.3648	1.519300	10.02080	39.1501
Augmented dataset 95% CI lower value	17.9825	1.4997	9.9424	38.7837
Augmented dataset 95% CI upper value	18.4081	1.5389	10.0992	39.2164

4. Conclusion

Different statistical analysis and data preprocessing approaches essential for data employed in machine learning model training and validation are presented. The analysis are applied to two datasets, the first dataset is the original 125-records dataset empirically collected from the case study 10-ton palm kernel oil (PKO) extractor machine. The second dataset is a 5000-records dataset generated from the 125-records dataset using Generative Adversarial Network (GAN) model. The statistical analysis and data preprocessing carried out included descriptive statistical analysis, determination of outliers, data normalization, and confidence interval determination for comparison of means of the two datasets. Through the normalized dataset, the maximum PKO yield and the input configuration that gave the maximum PKO yield are identified. In all, the statistical analysis and data preprocessing presented in this study are relevant for data driven model development especially using machine learning or deep learning approaches.

References

1. Balaji, K., Rabiei, M., Suicmez, V., Canbaz, C. H., Agharzeyva, Z., Tek, S., ... & Temizel, C. (2018, June). Status of data-driven methods and their applications in oil and gas industry. In *SPE Europec featured at EAGE Conference and Exhibition?* (p. D031S005R007). SPE.
2. Guo, W., Pan, T., Li, Z., & Li, G. (2020). A review on data-driven approaches for industrial process modelling. *International Journal of Modelling, Identification and Control*, *34*(2), 75-89.
3. Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2021). A review of data-driven decision-making methods for industry 4.0 maintenance applications. *Electronics*, *10*(7), 828.
4. Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial intelligence applications for industry 4.0: A literature-based study. *Journal of Industrial Integration and Management*, *7*(01), 83-111.
5. Adekunle, B. I., Chukwuma-Eke, E. C., Balogun, E. D., & Ogunsola, K. O. (2021). Machine learning for automation: Developing data-driven solutions for process optimization and accuracy improvement. *Machine Learning*, *2*(1).
6. Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial artificial intelligence in industry 4.0-systematic review,

challenges and outlook. *IEEE access*, *8*, 220121-220139.

7. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.
8. Dunphy, K., Fekri, M. N., Grolinger, K., & Sadhu, A. (2022). Data augmentation for deep-learning-based multiclass structural damage detection using limited information. *Sensors*, *22*(16), 6193.
9. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*.
10. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 18.
11. Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99.
12. Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), 1.
13. Ezeoha, S. L., & Akubuo, C. O. (2021). Influence of palm kernel variables on the yield and quality of oil expressed using an expeller. *Research in Agricultural Engineering*, *67*(2).
14. Ezeoha, S. L. (2020). Effects of some kernel factors on palm kernel oil extraction using a screw press. *Agricultural Engineering International: CIGR Journal*, *22*(1), 156-161.
15. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.
16. Khan, A., Hwang, H., & Kim, H. S. (2021). Synthetic data augmentation and deep learning for the fault diagnosis of rotating machines. *Mathematics*, *9*(18), 2336.