Science World Journal Vol. 20(No 1) 2025 https://dx.doi.org/10.4314/swj.v20i1.17
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

ROBUST PEARSON CORRELATION COEFFICIENT FOR IMBALANCED SAMPLE SIZE AND HIGH DIMENSIONAL DATA SET

*[1,2]Friday Zinzendoff Okwonu, [3]Owoyi Mildred Chiyeaka, [1,4]Nor Aishah Ahad and [5]Olimjon Sharipov

[1]Institute of Strategic Industrial Decision Modeling, School of Quantitative Science, Universiti Utara Malaysia,06010 UUM Sintok, Kedah Malaysia
[2]Department of Mathematics, Faculty of Science, Delta State University, P.M.B.1, Abraka, Nigeria
[3]Department of Mathematics, Faculty of Science, Dennis Osadebay University, Asaba, Nigeria
[4]School of Quantitative Sciences, College of Arts and Sciences Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[5]Department of Probability Theory and Mathematical Statistics, Institute of Mathematics, National University of Uzbekistan, Tashkent

*Corresponding Author Email Address:fokwonu@gmail.com

ABSTRACT

Conventionally, datasets of practical applications often vary in terms of sample sizes and dimensions; for example, undersampling or oversampling techniques are often applied to solve the minority sample size problems. However, formulating the Pearson correlation for imbalanced sample size and high dimensional data poses impracticable challenges. This study addressed the imbalance sample size problem and proposed a new method that could be used as a dual enabler to solve correlation problems for high dimensional data sets. The mean variance cloning technique (MVCT) would be applied to solve the imbalance sample size problem and the absolute variance variable selection technique (AVVS) would be applied as transpose enabler to enhance the computation of the Pearson correlation. This study aimed at revealing how strong or weak the relationship of an imbalanced sample size and high dimensional data set between two objects could be determined. The comparative results showed that the MVCT and the AVVS Pearson correlation demonstrated robust performance for the imbalanced sample size and high dimensional data set. Therefore, the simulation results have shown that the two preprocessing techniques (MVCT and AVVS) are enabler to enhance robust performance of the Pearson correlation. This study concluded that the enhanced Pearson correlation coefficient (AVVS-PCC, MVCT-AVVS-PCC, MVCT-PCC) indicated robust association and potentially suitable to perform different practical tasks that are aimed at solving complex practical problems.

Keywords: Imbalanced sample size, Pearson correlation coefficient, Heteroscedasticity, Homoscedasticity

INTRODUCTION

The outbreak of Covid-19 pandemic, and online transactions during the global lockdown has vigorously enhanced the generation of variants of data especially in medical practices and commerce. For instance, medical laboratory scientists and big pharmaceutical firms have enormous data on this effect especially during the COVID-19 vaccine trial in epicentre countries of the epidemic (Okwonu *et al*, 2020). The emergence of such data has rendered classical statistical procedures impracticable both in theory and practice (Ahad, *et al*,. 2020). The skyrocketing in the high dimensional data set and imbalanced sample size has led to the development of different over-sampling techniques such as synthetic minority oversampling technique (SMOTE) (Elreedy and Atiya, 2019) quartile oversampling and mean-variance cloning technique (MVCT) (Okwonu *et al,* 2024). The medical sciences and neuroscience have witnessed numerous imbalanced sample sizes curation with multidimensions which alters classical statistical techniques due to the curse of singularity and dimensionality during computational procedures (Wang *et al*, 2020).

The high dimensional data set(p) and small sample size data set (n_i) for different groups categorizations frequently emerge in medical research such as brain imaging (Efron, 2010), behavioural metal disorder (Allen *et al,* 2012), neuroimaging (Poldrack and Gorgolewski, 2014) and epidemiological study (Miller *et al,* 2016). Therefore, data sets from the above studies often lead to $p > n_i$ problems and subsequent imbalanced sample sizes (Smith and Nicholas, 2018; Bzdok et al, 2012). The critical pitfall of the imbalance sample size problem on the Pearson correlation is the heteroscedasticity condition which is a vital reason for weak negative or positive association between two objects (Apanapudor et al,2023, Okwonu and Othman, 2013). Novelly, imbalance sample size euphorically creates heterogeneity of the variance. Meanwhile when the imbalanced sample size is converted to a balanced sample size via under-sampling or over sampling, homoscedasticity is attained. Therefore, homoscedasticity robustify the performance of classical statistical methods such as Pearson correlation and least squares procedures (Ahad and Okwonu, 2023). The majority sample size and minority sample size data set can be used to compute the sample means, variance, standard deviations, and covariance but the majority group would be biased towards the minority group. This bias tends to create very weak positive or negative relationships for the objects. Hence, there is need to solve the bias created by imbalance sample size by either applying under sampling or over sampling techniques. As such, different oversampling procedures have been proposed to enhance the sample size of the minority group to equal the sample size of the majority group (Elreedy and Atiya, 2019). Several studies on solving imbalanced sample size problems by regenerating new data could be traced to the classification and regression domains where SMOTE has extensive applications in different fields of study (Chawla *et al*, 2002; Tang *et al*, 2009; Fernandez *et al,* 2018).

The Pearson correlation was coined with the concept of $n > p$ and for balanced sample size data in other to determine the relationship between two groups of objects. But due to emerging high dimensional data and imbalanced sample size problems in modern scientific discoveries with regards to the field of medical sciences, cancer studies, credit card fraud, medical imaging, commerce and host of others, the Pearson correlation (PC) becomes impracticable. This modern scientific data enrichment encumbers the formation and analysis of conventional statistical concepts like

the PC. Therefore, it becomes impregnable to analyse the degree of associations of two groups imbalanced sample size and $p > n$ problems. Different versions of the PC such as the canonical correlation analysis (Sherry and Henson, 2005) Spearman correlation coefficient (Qin *et a*l, 2020; Ali and AL-Hameed, 2022), Multivariate Pearson correlation (Okwonu *et al*, 2021 Najdi *et al*, 2022; Okwonu *et al*, 2022) point biserial(Ahad *et al*, 2023) have been proposed with numerous modifications to the classical PC to solve different association problems in different fields of study. However, the aforementioned methods could not solve the bias created by the imbalanced sample size due to the heteroscedasticity of the variance (Okwonu and Othman, 2013,Okwonu, 2015).

Conventionally, the Pearson correlation coefficient (PCC) is applied to investigate the linear relationship or association between two variables or objects of interest. The PCC values range between ± 1. The sign directions describe the numerical strength of one variable against the other variable. The variables of the classical Pearson correlation coefficient have been extensively studied and applied to different fields. It has been applied to investigate the paired association of Covid-19 infection, recovery, and death rate (Okwonu *et al*, 2021). It has been applied to investigate software engineering practice in Mexico(Aguilar-Calderon *et al*, 2019), detection of storage environment (Apanapudor *et. al,* 2020; Yang, *et al*, 2021), application to urban water supply(Zhu and Yuan, 2015), antibiotic waste water(Miao *et al,* 2024), gingival thickness(Yiwei, *et al,* 2024) rheological and rheo-fermentation properties(Jiang *et al*, 2024) , non-residual fractions based on ph level (Dong *et al*, 2024), the Pearson correlation has been applied to determine the relationship between pelvic muscle strength and spatiotemporal(Yongjie *et al*, 2024) sunspot number and detected sunspot (Veeramani and Sudhakar, 2024) oxygen and transparency and rotifers density(Tufia *et al,* 2024), lumbar spine and hip development(Zuo *et al,* 2024), parasubiculum and hippocampal subregion(Shi *et al*, 2024), aging peak and SOH(Li *et al*, 2023), Th1/Th2 cytokines in patients(Xia *et al,* 2023), sand and silt content(Scudiero *et al*, 2024). Some of these applications have rendered the classical PCC impracticable as such the principal component analysis has been applied to reduce the dimension of the dataset (Ye and Jia, 2023; Cheng *et al,* 2023). To determine the degree of association between the diverse dimensions of the data set, the canonical correlation analysis (CCA) was applied (Wang *et al,* 2020). Due to the computational complexity of the CCA for imbalanced sample size problems and resultant inability to regularize the sample size may pose potential computational problems. (Okwonu*, et al*, 2022), Reviewed the effect of imbalanced sample size on linear discriminant analysis. The imbalanced sample size problems in high dimensional data set may hinder the determination of the degree of association which render conventional procedures impracticable (Miller *et al*, 2016). Therefore, this paper addressed the problem of imbalance sample size, heterogeneity of the variance, and dimensionality problems in computing and formulating the coefficient of the Pearson correlation. The practicability of the enhanced coefficient would enable researchers to determine the strength of the relationship between two groups of objects. To solve these problems, it is infeasible to compute the value of the standard Pearson correlation with the following conditions (i) imbalanced sample size for two groups problem (that is, the numerator of PCC cannot be computed due to sample size difference); (ii) dimensionality problem $(p > n)$. The sample mean of condition (i) may portray a high degree of biased toward the majority group of the data set in comparison to the minority group. In addition, the imbalanced sample size creates heterogeneity of the variance. This condition may lead to a weak correlation. Condition (ii) is practically impossible due to course of dimensionality problem. Therefore, to solve the above problems which are associated to conditions (i-ii), we applied the mean variance cloning technique (MVCT) an oversampling method which is similar to SMOTE to transform the minority group to majority group. Secondly, the dimensionality problem could be solved by using the absolute variance variable selection technique (AVVS). These two preprocessing techniques are applied as plug-in and transpose methods to compute the coefficient of an enhanced PCC variant (AVVS-PCC, MVCT-AVVS-PCC, MVCT-PCC).

This article is structured as follows. The Mean Variance cloning technique (MVCT), the absolute variance variable selection technique (AVVS), and the variants of the Pearson correlation coefficient (PCC) was described in Section 2. Data collection, results and analysis was presented in Section 3. Conclusions followed in Section 4 respectively.

MATERIALS AND METHODS

This section describes the plug-in and transpose procedures applied to solve the imbalanced sample size and dimensionality problems in formulating the coefficient of the enhanced Pearson correlation (r_{XYc}). The first part of this section focused on the over sampling technique followed by the transpose procedure otherwise called absolute variance variable selection.

Mean variance cloning technique (MVCT)

The mean variance cloning technique (MVCT) is an over sampling technique used to clone data from the original imbalanced data in other to transform minority sample size to majority sample size (Okwonu *et al*, 2024). The MVCT has been applied to solve the problem of heteroscedasticity created by the majority group, thereby improving the strength of association between the two objects. In this case, when the two groups have the same sample size (balanced sample size) the homoscedasticity condition tends to be satisfied thereby enhancing the performance of the Pearson correlation. The MVCT procedure is derived as follows, the first step involves the computation of the sample mean and standard deviation of the minority group, that is:

$$\bar{X}_{rm} = \frac{\sum_{i=1}^{k} X_{rn}}{n_{rn}}, \quad (1)$$

where X_{rn} denote the minority group data points, n_{rn} denote the sample size and $\bar{X}_{rm}$ is the sample minority group mean and S denote the sample minority standard deviation,

$$S_{rn}^2 = \frac{(X_{rn} - \bar{X}_{rm})^2}{n_{rn} - 1}, \quad (2)$$

$$S = \sqrt{S_{rn}^2} \quad (3)$$

where $n_{rn} - 1$ denotes the unbiased sample size. Equation (1) and equation (3) are used to generate the $k = N_1 - n$ sample size. The new sample size is merged with the minority sample size to form the new majority sample size, that is $k + n = N_2$.Therefore, $N_1 = N_2$ which implies that the two groups have equal sample size. Then, the sample majority group from the above is given as

$$X_{srm} = X_{rn} + N(\bar{X}_{rm}, S_{rn}^2)[k = N_1 - n, p] = X_{rn} + \omega(k, p) \quad (4)$$

Meanwhile, the initial group with the largest sample size (majority) is represented by X_{MAJ}. Hence $X_{MAJ}(k + n, p)$ and $X_{srm}(k +$

$n, p)$ has equal sample sizes and dimensions. Thus, the MVCT majority group comprises of the generated data plus the original minority group.
Thus,

$$\bar{X}_{MAJ(1\times P)} = \frac{\sum_{j=1}^{N_1} X_{MJjp}}{N_1} \quad (5)$$

$$\bar{X}_{srm(1\times P)} = \frac{\sum_{j=1}^{N_2} X_{srm}}{N_2} \quad (6)$$

Where X_{MJjp} is the $N_1 \times p$ and X_{srm} is the $N_2 \times p$ data matrix for the two groups. Equation (5) is the sample group mean generated from X_{MJjp}while Equation (6) is the sample group mean generated from X_{srm}, respectively. The significant impact of the MVCT is that it resolves the problem of heterogeneity generated by the majority group over the minority group.

Absolute variance variable selection (AVVS)
Conventionally, the principal component analysis (PCA) (Dash and Liu, 1997) is often applied to perform dimension reduction. However, in practice, the PCA leads to information loss due to the number of eigenvalues selected to reduce the dimension of the data set. Apart from the PCA, other methods such as feature or variable selections (Sun *et al*, 2006) and a host of other dimension reductions have been proposed. In this study, a unique transpose procedure is proposed to solve the dimensionality problems for Pearson correlation (PC). The proposed method utilizes the equal sample size concept in which a $(p > n) \times n$ data matrix is transformed into $n > p$ data matrix. Therefore, the $n > p$ data set is applied to compute the sample estimates used as a plug-in for the PC. The AVVS is described as follows.

$$AVVS = \begin{cases} \leq 1, variables\ retained \\ \geq 1, variable\ deleted \end{cases} \quad (7)$$

$$AVVS = \left| \frac{(S^2_{majority}) - (S^2_{MVCTmajority})}{(S^2_{majority})} \right|, AVVS \leq 1. \quad (8)$$

where $S^2_{majority}$ denote the sample variance from the majority group and $S^2_{MVCTmajority}$ is the MVCT merged majority group. From equation (8), the values that satisfies $AVVS \leq 1$ are used to compute the coefficient of the Pearson correlation value. For $AVVS > 1$, lead to dimension reduction where the irrelevant variables are expunged. Whereas, if all the values from AVVS are less than one, it implies that all the data points in the variables are relevant. In this case, all the values associated to the variables are applied to compute the coefficient of the Pearson correlation. Table 1 is an illustration of the process of variable deletion based on equation (7). The data used in Table 1 are reported in Table 1 of ((Okwonu *et al*, 2024).

Table 1: AVVS procedure and decision analysis

$AVVS_i(i = 1,2,3)$	Decision
0.8334	accept
7.4923	delete
0.0661	accept

From Table 1, we observed that the MVCT balanced data revealed that the second variables exceeded the benchmark value hence only two variables are retained which are suitable to perform the computation of the PC. Hence the values associated to variables 1 and 3 are considered as relevant variables whereas the second variable is classified as redundant or irrelevant variable. In this case, variable 2 is expunged from the data set thereby reducing the three-dimensional data set to two-dimensional data set. This is a classical dimension reduction technique.

MVCT-Pearson correlation coefficient
The MVCT is an oversampling technique for solving imbalanced sample size problem for both $p > n, n > p$ data sets. The MVCT is similar to the synthetic minority over sampling technique (SMOTE) (Sun *et al*, 2006; Chawla *et al,* 2002) .The data set from the MVCT is used as an input to compute the plug-in sample estimate of the Pearson correlation. The Pearson correlation coefficient (PCC) has been studied extensively and had been applied to several fields of studies (Okwonu *et al*, 2020, Okwonu *et al*, 2023). The PCC conventional applications are often based on $n > p$ problems. The PCC is traditionally defined as

$$r_{XYc} = \frac{SS_{(XY_{Imb})}}{\sqrt{S_{(XX)}}\sqrt{S_{(YY_{imb})}}}, \quad (9)$$

where y_{Imb} is the MVCT generated majority sample size and x_i is the majority sample size, therefore their respective plug-in sample estimates are computed as follows

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k},$$
$$\bar{y}_{Imb} = \frac{\sum_{i=1}^{k} y_{iImb}}{k}.$$
$$SS_{(XY)} = UV = \sum_{i=1}^{k}(x_i - \bar{x})(y_{Imb} - \bar{y}_{Imb}), U = \sum_{i=1}^{k}(x_i - \bar{x}),\ V = \sum_{i=1}^{k}(y_{Imb} - \bar{y}_{Imb}), \quad (10)$$

$$S_{(XX)} = \sum_{i=1}^{k}(x_i - \bar{x})^2$$

$$S_{(YY_{Imb})} = \sum_{i=1}^{k}(y_{iImb} - \bar{y}_{Imb})^2$$

Therefore, equation (10) is the main reason for applying the MVCT to obtain balanced sample size. For instance, if the variances of U and V are heterogenous, the output from the PCC tends to be extremely weak. On the other, if the variances of U and V are homogeneous, the value of the PCC would be very strong in either sign directions. Therefore, the heterogeneity of variance problem was carefully ameliorated by invoking the MVCT to transform the unequal sample size group to equal sample size group. From the above, equation (9) is the MVCT-Pearson correlation coefficient (MVCT-PCC).

AVVS-Pearson correlation coefficient
The AVVS-PCC is a transposition procedure which require the transformation of $(p > n) \times n$ into $n > p$. The AVVS is a dimension-based transposition procedure. Therefore, the AVVS transposed data set is used as input to compute the sample estimates of the classical PCC. The AVVS-PCC could be applied for $p > n$ problems. On the other hand, if imbalanced sample size is observed, then the MVCT is applied first to the imbalanced data sample size before invoking the AVVS transposition to compute the coefficient of the Pearson correlation.

RESULTS AND DISCUSSION
Three data sets were used to investigate the comparative performance of the above PCC methods. The data set was culled from reputable data repositories, namely UCI machine learning and

Science World Journal Vol. 20(No 1) 2025 https://dx.doi.org/10.4314/swj.v20i1.17
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Kaggle. The data sets are the breast cancer, the Parkinson disease, and the Pima Indian diabetes. These data sets consist of imbalanced sample sizes and fixed dimensions.

Data set

1. Breast cancer data set(https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset). This data set consists of benign and malignant tumors. The focus is to investigate whether there is a strong or weak relationship between non-cancerous tumor and cancerous tumor. Or alternatively if a cancerous tumour has a strong or weak relationship with non-cancerous tumor. The non-cancerous tumor consists of 357 sample sizes and the cancerous tumor consists of 212 sample sizes. The data set has 30 attributes, that is $n_i > p, i = 1,2$ ...The data set consist of $n_1 = 357, n_2 = 212, p = 30$, while the AVVS PCC approach expunged the irrelevant variables(Equation (9)) and retained the sample sizes of each groups, that is, $n_1 = 357, n_2 = 212, p = 14$, while the MVCT-AVVS approach balanced the sample sizes, that is $n_1 = n_2 = 357, p = 14$ and the MVCT-PCC approach has $n_1 = n_2 = 357, p = 30$. The performance analysis of the variants of the Pearson correlation coefficient is reported in Table 2.
2. The second data set is well researched and applied, the Parkinson disease data set (https://archive.ics.uci.edu/dataset/489/parkinson+data set+with+replicated+acoustic+features). This data set consist of 240 sample sizes with healthy group $n_1 = 144$, $and\ unhealthy\ group\ n_2 = 96$. This data consists of 44 features respectively. In this study, we want to investigate whether there is a strong or weak relationship between the healthy group and the unhealthy group. That is, to determine whether there is a strong or weak possibilities for a healthy person to become unhealthy (Parkinson disease). The result is reported in Table 3.
3. The third data set has been extensively studied, that is the Pima India diabetes data set (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). This data set consists of 500 sample sizes in group one and 268 in group two with 8 attributes. The objective of this study is to determine whether there is a strong or weak relationship between a healthy person to have diabetes in the near future based on lifestyle change. Or a diabetic person to become a healthy person in the future based on lifestyle change. The result is reported in Table 4.

The results in Table 2 were designed to determine whether there is a weak or strong relationship between cancerous tumour and a non-cancerous tumour. The implication of this is that whether a cancerous tumour can become non-cancerous and a non-cancerous tumour becomes cancerous. It is the strength of these relationships we intend to determine. The results demonstrated strong possibilities for either to happen, meaning that a cancerous tumor if treated can become non-cancerous. The different PCC methods suggested strong relationship for either to happen in the future.

Table 2: Determining the relationship between the Benign and the Malignant tumor group.

Methods	AVVS-PCC	MVCT-AVVS-PCC	MVCT-PCC
Correlation value	0.9993	0.9999	0.9772
Sample size	$n_1 = 357, n_2 = 212, p = 14$	$n_1 = 357, n_2 = 357, p = 14$	$n_1 = 357, n_2 = 357, p = 30$

From the results in Table 3, the different methods suggested that there is a strong possibility for a healthy person to have Parkinson disease and also a strong possibility that an unhealthy person becomes healthy due to lifestyle adjustment. Meaning that treatment possibilities could transform an unhealthy person to a healthy person. Therefore, all the methods demonstrated strong possibilities for either to happen in the future.

Table 3: Determining the relationship between the health group and the Parkinson disease group

Methods	AVVS-PCC	MVCT-AVVS-PCC	MVCT-PCC
Correlation value	0.9996	0.9990	0.9990
Sample size	$n_1 = 144, n_2 = 96, p = 44$	$n_1 = 144, n_2 = 144, p = 44$	$n_1 = 144, n_2 = 144, p = 44$

The results in Table 4 demonstrated a very strong possibilities for a healthy person to become diabetic and diabetic person to become healthy. Therefore, the outcome of this study suggests that a healthy person today could be unhealthy tomorrow due to lifestyle changes. The study also suggested that improved lifestyle is a strong possibility that the individual may remain healthy, the contrary is true.

Table 4: Prima Indian diabetic data

Methods	AVVS-PCC	MVCT-AVVS-PCC	MVCT-PCC
Correlation value	0.9999	0.9999	0.9999
Sample size	$n_1 = 500, n_2 = 268, p = 8$	$n_1 = 500, n_2 = 500, p = 8$	$n_1 = 500, n_2 = 500, p = 8$

Previous studies such as the state of health of lithium-ion battery and aging peak with two varying charging rates as reported in Table 5 demonstrated strong degree of association. The results obtained from our study demonstrated similar results reported in previous research such as in Table 1 in (Li *et al,* 2023). The comparative analysis demonstrated that the variants of the PCC performed similar based on the above data set.

Table 5: Pearson correlation results reported in Table 1 in[41]

Charge rate	Peak A	Peak B	Peak C	Peak D
0.2C	0.9723	0.9895	0.9115	0.9873
0.1C	0.9715	0.9872	0.9566	0.9896

Conclusion

The classical Pearson correlation coined to determine how strong or weak the sign directions between two objects has been applied to different fields of study. The imbalanced sample size, heterogeneity of the variance, and dimensionality pitfalls of the classical PCC has been resolved by applying the plug-in and transposition procedures. The results from this study demonstrated that there is a strong association between the objects of study based on the data sets. The comparative performance showed that the plug-in, and transposition PCC affirmed strong association for all the data sets used. In general, the results reported in this paper strongly aligned with the robustness of the Pearson correlation coefficient for determining the relationship between two objects been investigated. The analysis demonstrated that AVVS-PCC, MVCT-AVVS-PCC, and MVCT-PCC methods could be applied to solve imbalanced sample size, heterogeneity of the variance, and dimensionality data set problem for determining the strength of the sign directions of any objects. This study affirmed that the plug-in and the transposition Pearson correlation methods are robust enough to determine how strong or weak the relationships between two objects could be described. This study concludes that the variants of the Pearson correlation could be applied to solve association problems in complex practical problems.

REFERENCES

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, D., Elliott, P., Gallacher,J., Green, J., Matthews, P., Pell, J., Sprosen, T., and Collins R. (2012). UK Biobank: Current status and what it means for epidemiology," *Health Policy Technol*,1(3), 123–126

Ali K., and Al-Hameed, A. (2022). Spearman's correlation coefficient in statistical analysis," *Int. J. Nonlinear Anal. Appl*, 13(1), 3249-3255

Aguilar-Calderon, J., Zaldivar-Colado, A., Tripp-Barba, C., Espinoza-Oliva, R. and Zurita-Cruz, C.E.(2019). "A Pearson Correlation Analysis of the Software Engineering Practice in Micro and Small-Sized Software Industry of Sinaloa, Mexico," *IEEE Latin America Transactions*, 17(2), 210–218.

Ahad, N.A., Okwonu, F.Z., Apanapudor, J.S., and Arunaye, F.I.. (2023). Chi-square and Adjusted Standardised Residual Analysis, *ASM Science Journal*, 18, 1–11.

Ahad, N.A., Okwonu, F.Z., and Siong, P.Y., (2020)"COVID-19 Outbreak in Malaysia: Investigation on Fatality Cases," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 20(1), 1-10.

Apanapudor, JS; Umukoro, J; Okwonu, FZ; Okposo, N(2023)Optimal solution techniques for control problem of evolution equations,Science World Journal,18(3):503-508

Apanapudor, Joshua S; Aderibigbe, FM; Okwonu, FZ,(2020),An optimal penalty constant for discrete optimal control regulator problems, Journal of Physics: Conference Series,1529(4): 042073

Bzdok,D., Schilbach, L., Vogeley, K., Schneider, K.., Laird, A.R.., Langner, R., and Eickhoff, S.B. (2012).Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy, *Brain Struct Funct*, 217(4), 783–796.

Chawla, N.V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. , (2002).SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321–357

Cheng, Y., Wang, H., Li, L. and Liang, J.(2023). A Multi-scale Study of SBS Modified Asphalt-Aggregate Adhesion Under Moisture Damage, *Arab J Sci Eng*, 48(10).

Dash M., and Liu,H. (1997). Feature selection for classification," *Intelligent Data Analysis*, 1(4), 131–156

Dong,Y., Lu, H. and Lin, H.(2024) . Comprehensive study on the spatial distribution of heavy metals and their environmental risks in high-sulfur coal gangue dumps in China," *J Environ Sci (China)*, 136.

Efron, B. (2010). "The Future of Indirect Evidence," *Statistical Science*, 25, (2), doi: 10.1214/09-STS308.

Elreedy, D., and Atiya A.F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf Sci (N Y)*, 505(C), 32–64, doi: 10.1016/j.ins.2019.07.070.

Fernández, A., García, S., Herrera, F., and Chawla, N.V. (2018). "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, 61,

Jiang, Q., Wei, X., Lui, Q., and Zhang, T. (2024).Rheo-fermentation properties of bread dough with different gluten contents processed by 3D printing," *Food Chem*, 433.

Li, Y., Luo, L., Zhang, C. and Liu, H. (2023). "State of Health Assessment for Lithium-Ion Batteries Using Incremental Energy Analysis and Bidirectional Long Short-Term Memory," *World Electric Vehicle Journal*, 14(7).

Miao, S., Zhang, Y., Men, C., Mao,Y. and Zuo, J. (2024). A combined evaluation of the characteristics and antibiotic resistance induction potential of antibiotic wastewater during the treatment process, *Journal of Environmental Sciences*, 138, 626–636.

Miller, K.L., Almagro, F.K., Bangerter,N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R.,Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu,L., Garrett,S., Hudson, S., Collins, R.., Jenkinson, M., .Mathews, P.M., and Smith S.M (2016).Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat Neurosci*, 19(11), 1523–1536.

Najdi, N., Ahad, N.A., and Okwonu, F.Z.(2022).Application of Pearson Correlation Technique to Analyze COVID-19 Pandemic during Eid al-Fitr Period in Malaysia, in *AIP Conference, Malaysia*: AIP. 1-6

Okwonu, F. Z., Ahad, N. A., Apanapudor, J. S., Arunaye, I. F., and Olijom, S.(2024). Application of Mean-Variance Cloning Technique to Investigate the Comparative Performance Analysis of Classical Classifiers on Imbalance and Balanced Data, vol. 978 (1), 1–17.

Okwonu, Friday Zinzendoff; Ahad, Nor Aishah; Okoloko, Innocent Ejiro; Apanapudor, Joshua Sarduana; Kamaruddin, Saadi Ahmad; Arunaye, Festus Irimisose(2022),Robust hybrid classification methods and applications, Pertanika J. Sci. & Technol. 30 (4): 2831 - 2850.

Okwonu, F. Z., Ahad, N. A., Apanapudor, J. S., and Arunaye, F.I.(2021).Robust multivariate correlation techniques: A confirmation analysis using covid-19 data set," *Pertanika J Sci Technol*, 29(2), doi: 10.47836/pjst.29.2.16.

Okwonu, F.Z., Arunaye, F.I. and Ahad, N.A., (2020)"Mathematical Model For Social Distancing In Mitigating The Spread Of Covid-19," *Nigerian Journal Of Science And Environment*,18(1), 173–182.

Okwonu F.Z and Othman, A.R. (2013). Comparative Performance of Classical Fisher Linear Discriminant Analysis and Robust Fisher Linear Discriminant Analysis," *Matematika,* 213–220

Okwonu, F. Z.,; Ahad, N. A.; Hamid, H.; Muda, N.; and Olimjon S.S.(2023), Enhanced robust univariate classification methods for solving outliers and overfitting problems, Journal of Information and Communication Technology,22(1):1-30.

Okwonu, F. Z.(2015), A chi square approach to determine the effect of students' residence on academic performance, Journal of Basic and Applied Research International:15(4): 280-286.

Poldrack .R.A., and Gorgolewski, K.J.(2014).Making big data open: data sharing in neuroimaging," *Nat Neurosci*, 17(11), 1510–1517.

Qin F., Song, Y., Nassis G., Zhao L., Dong Y., Zhao.C., Feng, Y., and Zhao, J. (2020).Physical activity, screen time, and emotional well-being during the 2019 novel coronavirus outbreak in China," *Int J Environ Res Public Health*, 17(14).

Scudiero, E., Corwin, D., Markley, P., Pourreza, A., Rounsaville, T., Bughici, T., and Skaggs T. H. (2024) "A system for concurrent on-the-go soil apparent electrical conductivity and gamma-ray sensing in micro-irrigated orchards," *Soil Tillage Res*, 235, doi: 10.1016/j.still.2023.105899.

Sherry.A., and Henson, R. K. (2005).Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer," *Journal of Personality Assessment*, 84(1).

Shi, J., Li, X., Liu, Q., Liu, J., Yuan, X. and Chen, S. (2024). "Effect of electronic moxibustion on the volume of hippocampal subregion in patients with amnestic mild cognitive impairment," *Chinese Journal of Tissue Engineering Research*, 28(20)

Smith, S.M., and Nichols, T.E.(2018). Statistical Challenges in 'Big Data' Human Neuroimaging," *Neuron*, 97(2), 263–268.

Sun, P., Chawla, S., and Arunasalam, B., (2006).Mining for outliers in sequential databases," in *Proceedings of the Sixth SIAM International Conference on Data Mining*.

Tang, Y., Zhang, Y., Chawla, N. V., and Krasser, S. (2009). SVMs Modeling for Highly Imbalanced Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 281–288.

Tufail,S. Liaqat, I., Saleem, S., Bibi, A. Mubin, M. and Nisar, B. ("Population dynamics of pelagic rotifers in Marala Headworks (Pakistan)," *Brazilian Journal of Biology*, 84.

Veeramani and Sudhakar, M.S. (2024). Automatic detection of sunspots on solar continuum HMI images blending local–global threshold," *New Astronomy*, 105, 2024, doi: 10.1016/j.newast.2023.102089.

Wang, H.T., Smallwood, J., Miranda, J.M., Xia, C.H. Satterthwaite, T.D., Bassett, D.S., and Bzdok, D. (2020).Finding the needle in a high-dimensional haystack: Canonical correlation anxalysis for neuroscientists, *Neuroimage*, 216, doi: 10.1016/j.neuroimage.2020.116745.

Xia, T., Zeng, K., Peng, Q.,Wu, X. and Lei, X,(2023) .Clinical significance of serum Th1/Th2 cytokines in patients with endometriosis, *Women Health*, 63(2)

Yang, Q., Kang, Q., Huang, Q., Cui, Z., Bai, Y., and Wei, H.(2021). Linear correlation analysis of ammunition storage environment based on Pearson correlation analysis, *J Phys Conf Ser*, 1948(1), 012064.

Yiwei, S., Xiangxiang, L.., Jianan, Z.., Jue, C., and Haiping, L. (2024). Consistency of gingival thickness measurement based on cone-beam CT imaging and cone-beam CT superimposed intraoral scan imaging," *Chinese Journal of Tissue Engineering Research*, 28(4), 569-573

Yongjie,L.., Shenyu,F., Yuan, X., Dakuan, Z., and Hongju, L.(2024).Correlation of knee extensor muscle strength and spatiotemporal gait parameters with peak knee flexion/adduction moment in female patients with knee osteoarthritis," *Chinese Journal of Tissue Engineering Research*, 28(9), 1354-1358.

Zhu, B., and Yuan, J.(2015). Pressure transfer modeling for an urban water supply system based on Pearson correlation analysis," *Journal of Hydroinformatics*, 17(1), 90–98.

Zuo, W., Gang, L., Huizhong, B., Lin, X.,Yi, Z., Jingpei, R., Chuanyu, H., and Xiaohong.(2024) .Relationship between lumbar spine development and hip development in children with spastic cerebral palsy," *Chinese Journal of Tissue Engineering Research*, 28(8), 1247-1252.