Jurnal Kejuruteraan 36(5) 2024: 1825–1838
https://doi.org/10.17576/jkukm-2024-36(5)-04

Cosmogenic Radionuclide Beryllium-7 Skewed Data Preprocessing for Northeast Monsoon Forecasting in Malaysia

Mohd Fauzi Haris[a,b], Norita Md. Norwawi[a*], Mohd Hafez Mohd Isa[c], Saaidi Ismail[b] & Muhammad Rawi Mohamed Zin[b]

[a]*Cybersecurity and Systems Research Unit, Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia*

[b]*Malaysian Nuclear Agency, 43000 Kajang, Selangor, Malaysia*

[c]*Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia*

**Corresponding author: norita@usim.edu.my*

Received 4 September 2023, Received in revised form 28 April 2024
Accepted 28 May 2024, Available online 30 September 2024

ABSTRACT

The onset of the Northeast Monsoon (NEM) in Malaysia, as defined by The Malaysian Meteorological Department (MET Malaysia), relies on the sustained easterly wind component for at least seven days, with at least one day featuring a speed greater than 5 knots (2.5m/s). While meteorological parameters have historically been crucial in predicting these kinds of events, new research, like that carried out in Kerala, India, has demonstrated the potential to use non-traditional indicators, such as the concentration of the cosmogenic radionuclide Beryllium-7 (7Be) in the north and south hemispheres for monsoon prediction. This article zooms in on a fundamental aspect of monsoon forecasting: raw data preprocessing. It looks into using R statistical tools to refine and recalibrate datasets using this recently introduced parameter, 7Be, for NEM forecasting focusing on Malaysia. By meticulously adjusting and cleansing raw data, this preprocessing stage aims to align the data with the specific requirements of NEM prediction models, thus enhancing their accuracy and reliability. The significance of robust data preprocessing cannot be overstated, particularly in the context of NEM forecasting, where the accuracy of predictions holds profound implications for various sectors such as agriculture, tourism, and infrastructure planning. Potential biases, anomalies, and inconsistencies in the data can be eliminated with careful preparation, resulting in more reliable projections and well-informed decision-making. As such, this article underscores the critical role of data preprocessing in laying the groundwork for reliable and actionable NEM forecasts, ultimately contributing to the resilience and adaptability of Malaysia's socio-economic landscape.

Keywords: Data preprocessing; R statistical software; Beryllium-7

INTRODUCTION

In recent years, data science has seen significant growth with the explosion of big data, advances in machine learning, and the widespread availability of data from various sources. As the field of data science has matured, the importance of data preprocessing has become increasingly recognized. The quality and reliability of the data used for analysis are critical factors in the accuracy and effectiveness of data-driven decision-making.

The gathered raw data should be appropriately processed to yield the best results. Therefore, data preprocessing is a crucial stage in the data analysis pipeline, where raw data is conditioned to be suitable for analysis. Data preprocessing includes cleaning, transforming, and integrating data from various sources to ensure that the

data is consistent, accurate, and complete. This step is crucial as it ensures that the data used for analysis is reliable and trustworthy. Moreover, feature engineering, which is a critical aspect of data preprocessing, is essential for machine learning as it involves selecting and transforming the most relevant features that will be used to train the model.

Numerous tools are available to streamline this task, ranging from open-source software like Python and R for statistical computing and graphic design to commercial alternatives such as MATLAB, SPSS, Tableau, and similar software products. Various tools are available to streamline the task, encompassing open-source software like Python and R Statistical Computing alongside graphic software. Additionally, commercial alternatives such as MATLAB, SPSS, Tableau, and similar products offer further options for simplification. These tools are employed in various domains, including research, data analysis, and modeling, to name a few, and are chosen based on various criteria, such as functionality, ease of use, cost, and suitability for the task at hand. Consequently, selecting an appropriate tool is critical for the task's success and necessitates careful consideration of various factors to ensure optimal outcomes.

Data preprocessing presents challenges, such as handling missing data, selecting relevant features, and normalizing data for meaningful comparisons. However, these challenges can be overcome by employing suitable techniques and tools to ensure accurate, reliable, and meaningful analysis. Data preprocessing is essential in machine learning to clean, transform, and format data, enabling effective learning by machine learning algorithms. However, due to a lack of high-quality training data, machine learning models may produce erroneous or inconsistent results if appropriate data preprocessing is not done. A significant worry in machine learning classification problems is the issue of class imbalance, which has been addressed in the works of (Shamsudin et al. 2023), (Werner de Vargas et al. 2023), and (Felix and Lee, 2019).

This article, however, will not touch on the issue of imbalanced data; instead, it will focus on correctly resolving anomalies and preprocessing skewed data in the context of a study aimed toward northeast monsoon forecasting. This article looks into the usage of R, a statistical software. Notably, the paper explores the application of machine learning algorithms to predict selected meteorological zs such as rainfall patterns, wind speeds, and 7Be intensity concentration associated with the northeast monsoon.

BACKGROUND

Malaysia's Northeast monsoon season (NEM) is annual, typically from November to March, as mentioned by (Ishak, Tauhid Ahmad, and Jit Singh, 2021) and Bohari N.Z.I, (2021) ; however, Tang (2019) claims that NEM occurs from October to March. On the other hand, Tan and Santo (2018) note that the NEM season starts in December and lasts until March. It also differs from Tangang et al. (2012), who claim that the NEM season usually begins in November and ends in February the following year. Despite debates and varying opinions on the start and end dates of Malaysia's Northeast monsoon (NEM) season, there is a consensus that this climatic period brings heavy rainfall that can result in flooding. Furthermore, the NEM season is a significant climatic period for Peninsular Malaysia, as a large portion of the peninsula receives up to 70% of its total annual rainfall (Fakaruddin et al. 2020). Therefore, stakeholders are advised to closely monitor this issue to minimize or lessen the loss and casualties brought on by flooding.

The India Meteorological Department (IMD) began making monsoon forecasts as early as 1886, relying on observations of wind patterns, pressure gradients, and rainfall data (Mohanty et al. 2019 and Rao et al. 2019). This was possible because the IMD had been collecting meteorological data since its establishment in 1875. In contrast, it is unclear when formal monsoon prediction began in Malaysia. That being stated, it is generally agreed upon that the Malaysian Meteorological Department was established in 1958. (Jabatan Meteorologi Malaysia, 2019), Coffee Table 60 Tahun MET Malaysia. Retrieved from Bohari N.Z.I et. al. (2021) and has since been providing weather forecasts and warnings to the public in Malaysia. While the historical development of monsoon prediction in Malaysia is not widely known, it is clear that the IMD's early adoption of forecasting methods was made possible by collecting and analyzing meteorological data over several years, emphasizing the importance of long-term data collection in weather prediction.

Numerous articles delve into the prediction of monsoon-related phenomena. Geen (2021) conducted a study on forecasting the onset of the South China Sea Monsoon, while Saha et al. (2017) utilized deep learning to predict the monsoon over a homogenous region in India. Another approach was introduced by Saha et al. (2021), employing a stacked encoder and ensemble regression model for the summer monsoon prediction. Rajan and Desamsetti (2021) focused on the onset of the Indian summer monsoon, employing a high-resolution model, and Kumar and Singh (2021) reviewed Indian summer monsoon rainfall prediction using machine learning

techniques. On the other hand, (Terzi et al. 2019) have come up with a new approach and parameter to forecast summer monsoon onset and withdrawal for Kerala, India. Beryllium-7 (7Be) is a cosmogenic radionuclide collected from the air surface and, through a specific process, yields raw data with its intensity concentration.

Moreover, a new approach called Trans-equatorial has been deployed, which reflects particulate movement following wind from the north and south hemispheres. This method successfully forecasted monsoon onset with an unprecedented accuracy of ±3 days, almost two months in advance, and an average of 42 ± 7 days before the monsoon withdrawal date. Alternatively, in their investigation, Haris et al. (2023) scrutinized studies concerning weather-related cosmogenic radionuclides, thereby providing greater confidence in the potential for further exploration and utilization of this new parameter in forecasting rainfall during the Northeast Monsoon (NEM) in Malaysia.

The study of monsoon forecasting in Malaysia presents significant opportunities for further exploration, particularly concerning traditional methods and meteorological parameters. Introducing new parameters and investigating unique calculating techniques are required to improve the current forecasting procedures. These advancements are hoped to be adopted and adapted to effectively support stakeholders in preparing for the NEM season, which occurs annually. By doing this, the timely and reliable information delivery for NEM-related decision-making processes would be improved.

Using appropriate tools such as R statistical software may help to speed up the data preprocessing task. (Hackenberger 2020) states that R's strength resides in the abundance of packages containing functions, methods, and procedures for many forms of data processing. The package's flexibility in terms of the programming languages in which their functions are written, such as C/C++, Fortran, Java, and Python, and the redundancy of packages for statistical methods are additional benefits of R. According to Hair et al. (2021), it is recommended to use it within the convenience of an integrated development environment (IDE) like RStudio. In simple words, R is a free software environment for statistical computing and graphics.

METHODS

Data cleaning involves finding and fixing mistakes in the data; data transformation, which entails transferring data from one format to another; and data integration, which requires merging data from several sources, are all examples of data preprocessing procedures. Other strategies include data normalization, which entails scaling data to a standard range to permit meaningful comparisons, and feature selection and extraction, where pertinent features are chosen and retrieved from the data set.

DATA SOURCE AND DATA ACQUISITION METHOD

The data utilized in the present study are not publicly available; however, access to the virtual Data Exploration Centre (vDEC) can be obtained by the Comprehensive Nuclear Test-ban Treaty Organization (CTBTO) through a confidentiality agreement that does not incur any costs. To access the vDEC data, interested parties obtained the requested data after undergoing an approval process from CTBTO. CTBTO provided the concentration data for Beryllium-7 (7Be) intensity from selected radionuclide stations in the Hadley-Ferrel Zone (30° – 60° South and North) from 2010 to 2022. The Malaysian Meteorological Department (MET Malaysia) supplied the onset and withdrawal dates of the Northeast monsoon. Even though this radionuclide does not indicate nuclear activity directly, the detection and existence of this radionuclide are used by CTBTO to confirm that the station is functioning well.

IDENTIFICATION OF SELECTED RELEVANT RADIONUCLIDE STATIONS

CTBTO has played its role as an international body that monitors nuclear tests and has built almost 80 radionuclide stations in addition to several other stations with different technologies for the same purpose. Therefore, selecting an adequate and related station is essential to avoid wasting time processing the unrelated and unnecessary abundance of 7Be raw data. For that, articles related to the northeast monsoon in Malaysia are collected and skimmed through to identify the related stations. In the beginning, all stations located in the equatorial region were selected. In this case, CTBTO radionuclide stations near the equator or in the tropics (latitude 23.5S to 23.5N) have been listed 23 stations. Since, Terzi *et al.* (2019) explained the Trans-equatorial approach, several stations near the station suggested in the articles were chosen. These include RUP54, RUP59, and RUP61 in the northern hemisphere, and all stations located in Australia and near to it that represent the southern hemisphere were also shortlisted.

Based on several other articles, including from MET Malaysia, During the winter season, the central Siberian and Kazakhstan region experiences the dominance of a semi-permanent high-pressure system known as the Siberian High (SH). This system, which remains relatively shallow, confines its influence within the lower 1.5

kilometers of the atmosphere from November to March. A cold surge during the NEM triggers mass convergence over the South China Sea (SCS), leading to enhanced convergence and possible heavy rainfall in the particular region (Yip *et al.* 2016). For that, stations such as Mongolia, Okinawa in Japan, Beijing, Lanzhou, and Guangzhou in China, located along the paths that bring the wind to Malaysia, were also selected. Simultaneously, the occurrence of another wind surge during the Northeast Monsoon (NEM) is commonly referred to as the Easterly surge (Dong *et al.* 2020, Akasaka *et al.* 2018 and Hai *et al.* 2017). The wind from the North Pacific Ocean could pass through several stations, such as Wake Island, Oahu, Hawaii, Midway Islands, and the Philippines. The list of total selected stations is presented in Table 1 below.

TABLE 1. List of selected CTBTO radionuclide stations – Stage 1

Station ID	Description	Latitude	Longitude	Elevation	Date Begin*
AUP04	Melbourne, VIC, Australia	-37.730000	145.100000	50.100000	2000-07-27
AUP06	Townsville, QLD, Australia	-19.249444	146.766111	6.260000	2001-12-06
AUP08	Cocos Islands, Australia.	-12.188222	96.83447	4.570000	2003-06-02
AUP09	Darwin, NT, Australia	-12.430000	130.890000	25.690000	2002-04-08
AUP10	Perth, WA, Australia	-31.928888	115.981944	16.310000	2000-07-27
CNP20	Beijing, China	39.950000	116.400000	0.040000	2017-09-26
CNP21	Lanzhou, China	36.000000	104.200000	1886.730000	2016-10-15
CNP22	Guangzhou, China	23.000000	113.300000	66.000000	2017-05-04
GBP66	BIOT/Chagos Archipelago, UK	-7.303889	72.400556	1.850000	2004-08-02
JPP37	Okinawa, Japan	26.502972	127.904278	83.000000	2006-09-13
MYP42	Tanah Rata, Malaysia	4.480000	101.370000	1472.620000	2009-03-01
MNP45	Ulaanbaatar, Mongolia	47.890000	106.330000	1565.020000	2002-11-26
NZP47	Kaitaia, New Zealand	-35.070000	173.290000	79.110000	2000-05-31
PHP52	Tanay, Philippines	14.581861	121.369600	635.240000	2005-11-01
RUP54	Kirov, Russian Federation	58.586000	49.413361	121.230000	2007-12-02
RUP59	Zalesovo, Russian Federation	53.935806	84.789500	218.390000	2007-06-26
RUP61	Dubna, Russian Federation	56.740556	37.251667	116.580000	2008-09-11
USP77	Wake Island, USA	19.292278	166.610861	8.140000	2007-09-20
USP78	Midway Islands, USA	28.220000	-177.370000	-0.200000	2009-02-02
USP79	Oahu, Hawaii, USA	21.522444	-157.994972	393.810000	2005-03-16
USP80	Upi, Guam, USA	13.569917	144.928300	162.810000	2007-09-05

* Date Begin - pertains to the initial day of commencement or operation

Due to the delayed development of its radionuclide monitoring facility, all stations from China (highlighted with red font) in the analysis are excluded due to insufficient data availability. Moreover, any station exhibiting inadequate data or a dearth of values exceeding 20% throughout the entire study period would similarly be excluded from the analysis.

REMOVING OUTLIER AND DUPLICATE DATA

Much scientific research has come across the term outliers. It is a standard process to avoid misinterpretation of our results. Agreed also in biomedical science (Gress et.al 2018), hormonal research done by Pollet and van der Meij (2017) stressed that outlier handling could substantially impact significance testing. However, psychological research, such as that done by Bakker and Wicherts, (2014), investigated whether removing outliers in psychology papers is related to weaker evidence. The results also revealed that the classification accuracy had significantly improved after the outlier photos were taken out of the dataset, the VGG-16 model had been retrained, and the number of incorrect classifications had decreased (Perez and Tah, 2020). Please exercise caution when considering the removal of outliers, as they may contain valuable information about the situation at hand. Unless outliers can be confidently attributed to experimental errors,

indiscriminately excluding them from the analysis is not recommended (Altman & Krzywinski 2016)

Interquartile range (IQR) was utilized by Dash et al. (2023) and Perez and Tah (2020) to identify outliers. Machine learning algorithms with varied criteria of 10, 15, and 20% were employed to identify outliers from data on shale gas output (Yehia et al. 2022) and suggested that combining outlier remover could improve production forecasting and reserve estimation.

The 7Be raw data obtained from CTBTO consists of several parameters. However, for this study, only two parameters were chosen for further analysis. These parameters are the "Collection Stop," which provides the timestamp indicating the end time of daily data collection, and "Activity.concentration.µBq/m3," which represents the numerical value indicating the intensity concentration of 7Be for the corresponding day. Generating a boxplot visualization can aid in detecting outliers within the data.

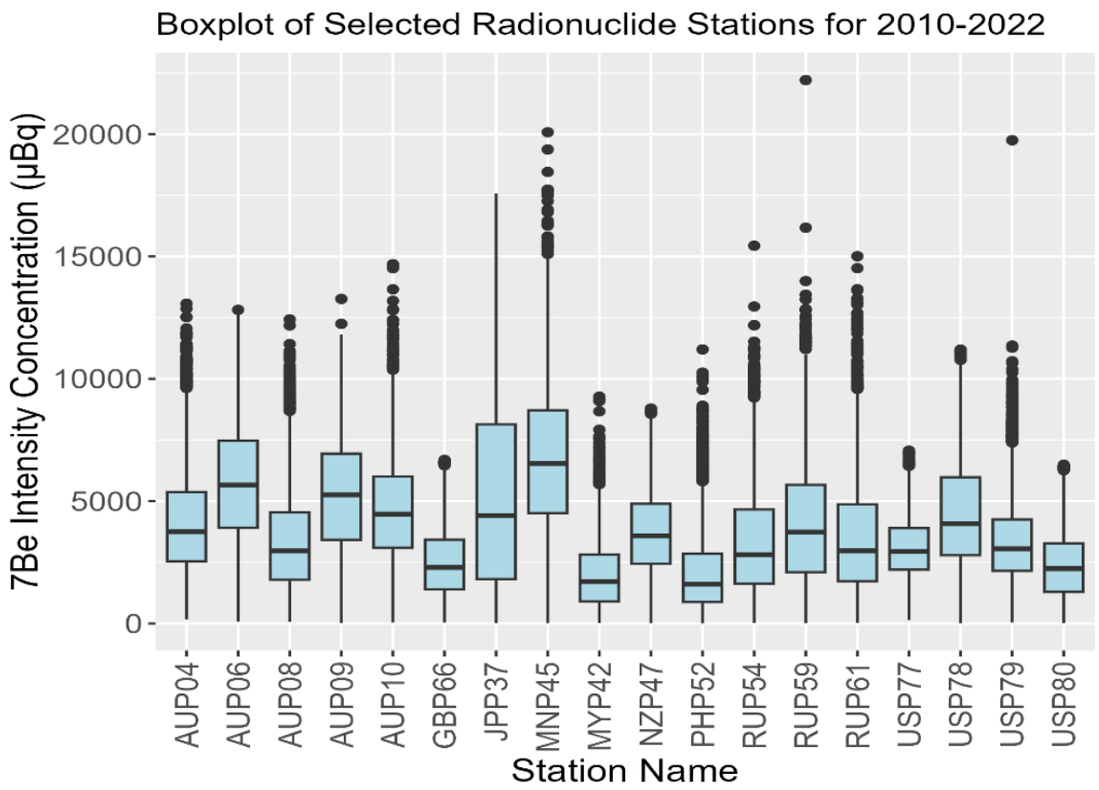

FIGURE 1. Boxplot of all selected radionuclide stations for the study

Based on the depicted plot in Figure 1, it is evident that outliers are present in most stations except for the JPP37 station located in Okinawa, Japan. In order to tackle this issue, the IQR method is utilized to remove outliers. The IQR, calculated as the difference between the 75th and 25th quartiles, is a statistical measure used to assess data spread. To identify outliers, the upper limit is determined by adding 1.5 times the IQR to the 75th percentile, while the lower limit is obtained by subtracting 1.5 times the IQR from the 25th percentile. Data points exceeding the upper limit or falling below the lower limit are deemed outliers, signifying values that significantly deviate from the central distribution of the data. This approach aids in pinpointing and potentially mitigating the influence of exceptional observations in the dataset.

On the other hand, the significance of duplicate data should not be underestimated as it can detrimentally impact data quality. Within the context of this specific time series data, duplications are anticipated to primarily concern dates. Given the criticality of the date in this analysis, the dataset should ideally contain only one entry per day. Consequently, if multiple data points exist for the same date, retaining the most recent entry is essential while discarding the remaining duplicates. This practice aligns with the data collection, processing, and transmission procedures implemented for Vienna's International Data Center (IDC). The case study detailed herein illustrates the removal of both duplicate data and outliers as integral steps in the analysis process.

HANDLING MISSING VALUE

According to Nugroho and Surendro (2019), missing data frequently results in erroneous predictive analytics. Peng et al. (2020) concur on the significance of addressing this challenge in line with acknowledging missing values as a crucial and widespread issue. The presence of missing values substantially impacts data analysis and may lead to biased results and compromised inference where incorrect conclusions are drawn. Recognizing missing values as an essential concern emphasizes the need for robust methods and techniques to handle missing data effectively. As a result, they proposed and presented evidence supporting the higher performance of a Monte Carlo likelihood strategy in removing bias from parameter estimates. Given the research context, it is essential to acknowledge that the patterns of missing values discussed are not exhaustive. The causes of missing values depend on various factors and conditions specific to each study or analysis. The occurrence of missing values invariably leads to unobserved data points, which is also supported by Seu et al. (2022), resulting in a loss of information. However, developing a definitive and comprehensive solution for effectively handling missing values remains an ongoing challenge. Finding a suitable imputation technique is often the first step in more formal time series analysis (Beck et al. 2018).

Missing data in datasets where the data was not recorded is rather prevalent. In addition to missing values in this time series data resulting from equipment malfunction or related data collection device issues, excluding outliers also contributes to the occurrence of days without available data. When solving the missing value problem, the incomplete dataset was divided into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976). At the same time, numerous approaches have been devised to address missing data in datasets, aiming to enhance prediction accuracy and mitigate biased outcomes (Gupta and Gupta 2018).

7BE MISSING VALUES

Multiple CTBTO radionuclide stations are geographically dispersed and sometimes in remote areas, although they effectively monitor the presence of radioactive gases and particles from nuclear detonations within the environment. Moreover, a limited number of uncrewed stations are in operation. Consequently, restoring or substituting any defective or malfunctioning equipment at these work sites may require considerable time. A malfunctioning device may take over a week to repair, so missing 7Be data in this study is considered a Missing Not at Random (MNAR) type. The reason is that the missingness, in this case, depends on the unobserved values, which are the measurements the malfunctioning device should have recorded. The malfunctioning device introduces a non-random pattern of missing data, as the missingness is directly related to the unobserved values.

Consequently, the missing data cannot be considered as Missing Completely at Random (MCAR) or Missing at Random (MAR), as the device malfunction introduces a systematic bias in the missing data mechanism. Understanding the type of missing data is essential for determining how to manage missing data and how to impute it. One method commonly used to handle missing values is case deletion, a traditional approach widely employed (Gupta & Gupta 2018).

Figure 2 presents the percentage of missing values for the selected stations from 2010 to 2022. Three stations from China, as previously mentioned, were excluded at an early stage. The graph indicates that all stations exhibit less than 20% missing values over the period. However, upon closer examination of the raw data yearly, a subset of stations exhibits missing values exceeding 20% and, notably, reaching as high as 97% in 2010. This pattern is visually depicted in Figure 3 and Figure 4, providing a clear illustration of the observed missing data. Imputing data in such a situation needs extra opinions from domain experts to avoid misinterpretation. Case deletion could also be the best choice in this study.

7BE DATA IMPUTATION

In the epidemiology study by Madley-Dowd *et al.* (2019), multiple imputations effectively reduce bias for missing at random (MAR) data, even in cases with a high percentage of missing data. While the applicability of this approach may vary across different fields, it can generally be considered sound advice. In their rainfall prediction study, (Sharma et al. 2021) chose Multivariate Imputation by Chained Equations (MICE) imputation and the Wrapper method to fill out the missing data. Missing data, or missingness, has the potential to introduce bias and reduce efficiency, and Tawn et.al., (2020) have proved that there are significant negative consequences in the presence of missing data. Fortunately, the “mice” package is one of several R software packages and functions that facilitate the implementation of multiple imputations.

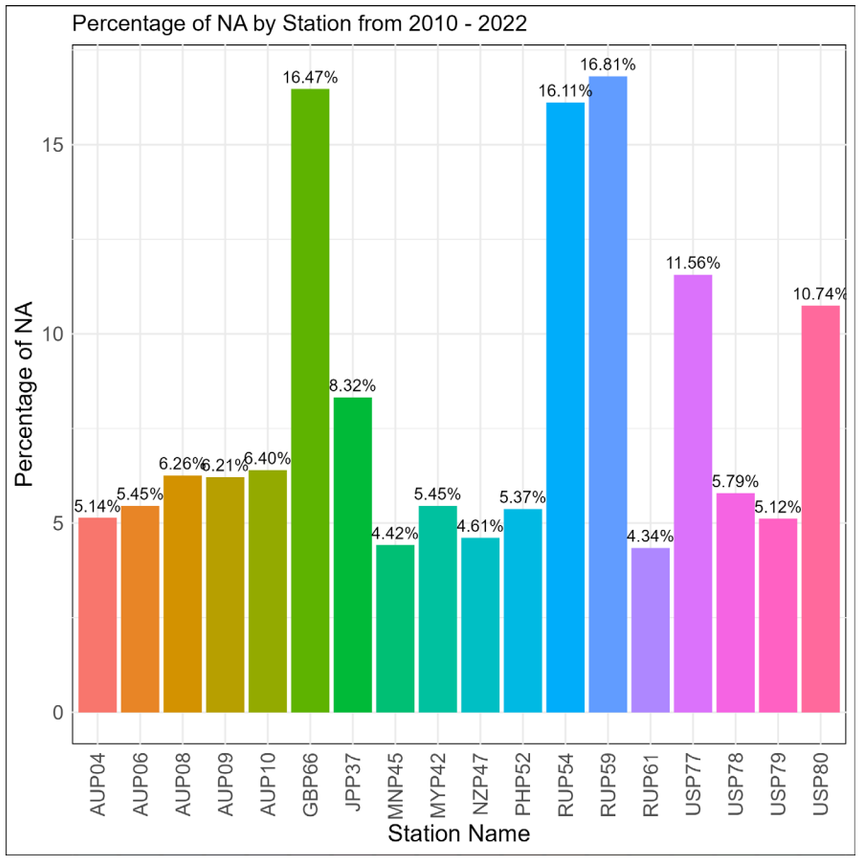

FIGURE 2. Total percentage of missing value for the whole period of study (2010 – 2022)

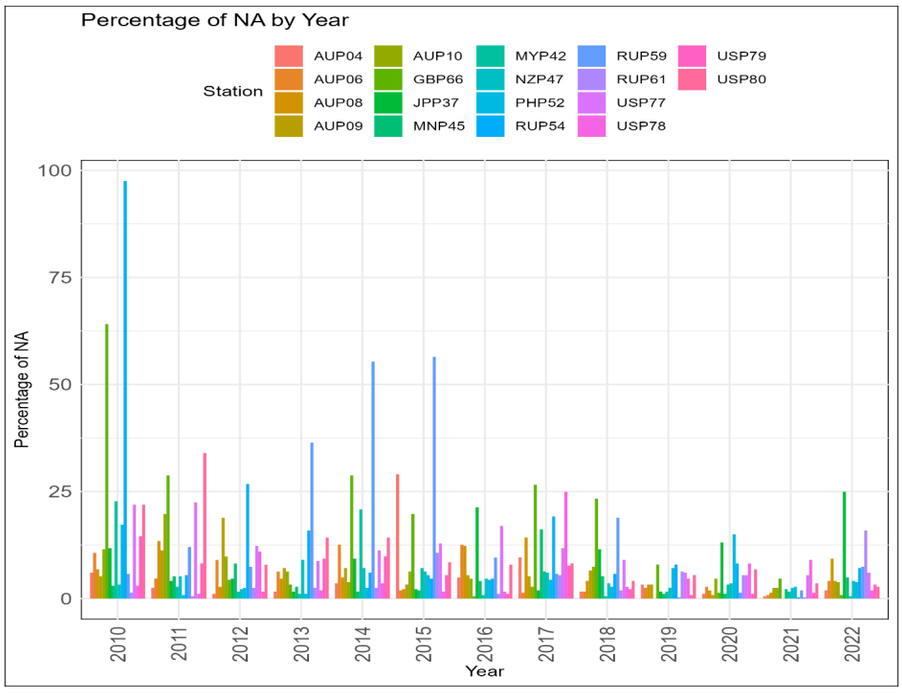

FIGURE 3. Percentage of missing value by stations every year

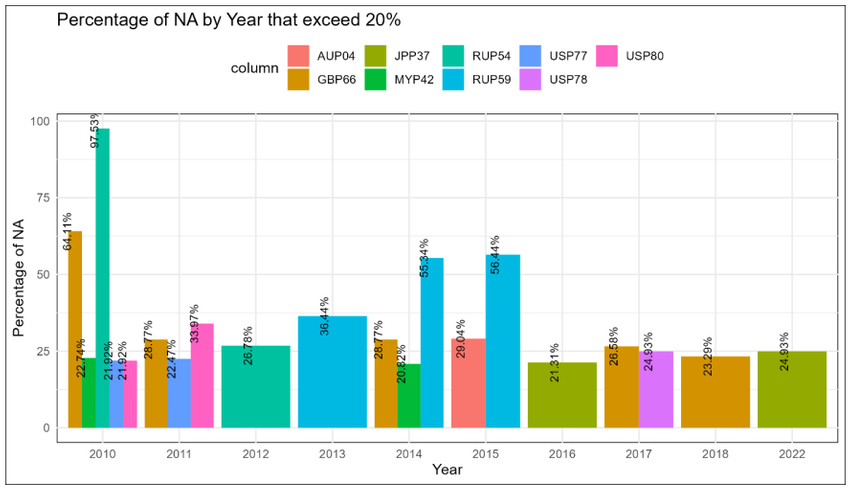

FIGURE 4. Percentage of missing value by stations every year that exceed 20%

Zhang et al. (2013) employed a clustering analysis methodology to impute missing values by assigning distributed weights through MATLAB simulation and measuring the monthly account number of electronic banking products from eighteen areas. Seu et al.(2022) did a review on intelligent missing data imputation techniques. Reviewed several methods of well-known artificial intelligence algorithm-based missing data imputation, aiming to evaluate which method performs perfectly to impute the missing data in the dataset. The comparison result of these methods indicates that KNNimputer and MICE perform the most excellent approach to imputing missing values.

For this investigation, several imputation methods have been selected. The "DMWR" library is used by Knnimputation, the missRanger library by missRanger, and the MICE library by MICE. The first step of imputation using MICE employs predictive mean matching (pmm), with a default value of 5 imputations, 0.001 convergence criteria, and 20 iterations. Results for all chosen imputation strategies are shown in a single graphic in Figure 5 below for a test using data from the RUP54 station in 2010 since it shows the most extended missing value in one calendar year.

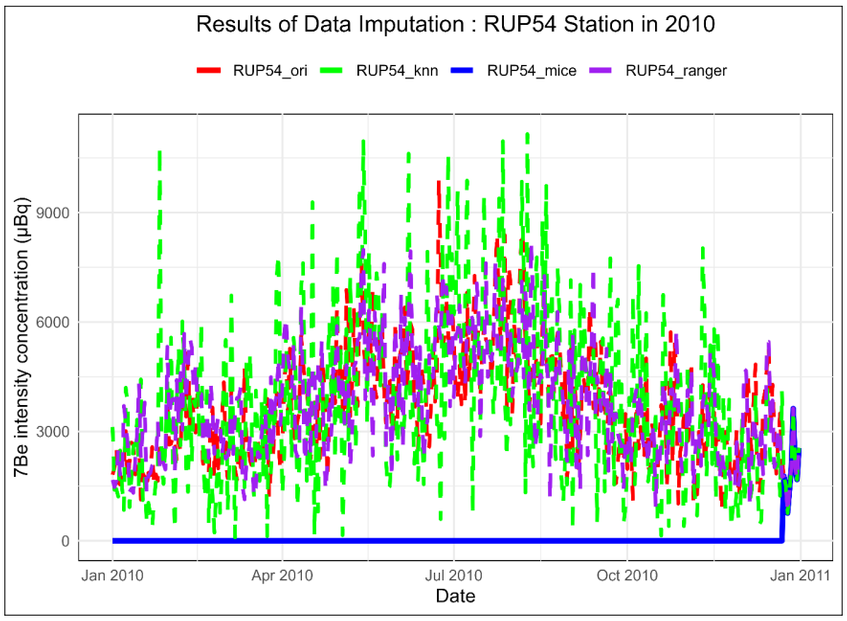

FIGURE 5. Simple plot of results for selected imputation methods

A visual inspection of the provided plot shows that the "knnimputation" and "missRanger" imputation methods exhibit a relatively minimal disparity when compared to the imputation performed using the "mice" method. Given this finding, either "knnimputation" or "missRanger" will be selected for subsequent dataset analysis. However, "mice" could also be further explored since it has many features and advantages compared to other methods. For example, mice use predictive mean matching (pmm) to handle missing numerical values by default.

EVALUATION OF IMPUTATION METHODS

The stations with the fewest missing values have been selected for further consideration when evaluating imputation methods. In Figure 2-4, the radionuclide stations RUP61, MNP45, and NZP47 have less than 5% missing values, with an annual percentage of NA not exceeding 20%. Therefore, RUP61 is randomly selected from 3 shortlisted stations and further tested for calculating the imputation performance. Imputation performance was further observed by calculating the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between original and imputed data. Results are shown in Table 2 below. In the comparative analysis of three data imputation methods, KNN imputation, MISSRANGER, and MICE, missranger emerged as the most effective technique, demonstrating superior performance based on key evaluation metrics. With a Mean Absolute Error (MAE) value of 265.85 and a Root Mean Squared Error (RMSE) value of 881.81, MISSRANGER outperformed both KNN imputation and MICE in terms of imputation accuracy. This suggests that MISSRANGER 's imputation strategy resulted in predictions that were, on average, closer to the true values, as indicated by the lower MAE, and exhibited less variability, as reflected in the lower RMSE. Moving forward, it is advisable to incorporate MISSRANGER as the preferred imputation method in similar data imputation tasks. Additionally, further exploration and experimentation with various datasets and scenarios could provide valuable insights into the robustness and generalizability of MISSRANGER, ensuring its suitability for a broader range of applications. Continual evaluation and comparison with emerging imputation techniques will also contribute to maintaining optimal imputation strategies in data preprocessing pipelines.

TABLE 2. Result of an imputation performance evaluation using MAE and RMSE

	MAE	RMSE
KNN_IMPUTATION	287.08	939.71
missRanger	265.85	881.81
MICE	318.84	1104.86

RESULTS AND DISCUSSIONS

NORMALITY OF THE DATA

The selection of an inappropriate representative value for a dataset and the subsequent calculation of the significance level based on this representative value can potentially lead to erroneous interpretations. Consequently, it becomes crucial to first assess the normality of the data before determining whether the mean is a suitable representative value. If the data exhibits normality, parametric tests are employed to compare means. Conversely, non-parametric methods are employed if normality assumptions are violated, utilizing medians as the representative values for group comparisons. This sequential approach ensures the appropriate selection of representative values and the application of appropriate statistical tests based on the nature of the data. Normally distributed data are best, and most truthfully, represented through mean and standard deviation.

It is crucial to understand that choosing the incorrect test (such as parametric for irregularly distributed data or non-parametric for normally distributed data) can lead to utterly incorrect conclusions. This could misrepresent statistically significant results as non-significant or statistically insignificant results as significant. Therefore, recognizing the crucial role of the normal distribution assumption in ensuring that the various regression approaches are used correctly is essential in regression analysis. Specifically, utilizing skewed data poses a challenge and renders applying most regression methods infeasible. Hence, it becomes imperative to consider the normality assumption when undertaking regression analysis, as it establishes a fundamental prerequisite for conducting robust and reliable regression models.

The non-normal data can be analyzed in two different ways. First, use tests that do not require normality, such as non-parametric tests, or alter the data to fit a normal distribution by employing the proper function. For this, 7Be data was tested using the Shapiro-Wilk test, and the results are shown in Table 3 below. Several criteria, including the well-known p-value, are used in this test to assess the data and determine whether its distribution significantly deviates from the normal distribution. The distribution departs significantly from the normal distribution if the p-value is less than 0.05.

TABLE 3. Data normality check for each station using the Shapiro-Wilk test

	Raw data with NA		Imputed with missRanger	
Station	p_value	Normality	p_value	Normality
AUP04	4.80E-19	Not Normal	2.71E-33	Not Normal
AUP06	7.74E-06	Not Normal	1.05E-12	Not Normal
AUP08	4.81E-21	Not Normal	8.13E-39	Not Normal
AUP09	3.01E-08	Not Normal	1.00E-16	Not Normal
AUP10	1.28E-13	Not Normal	1.54E-25	Not Normal
GBP66	5.66E-20	Not Normal	1.84E-31	Not Normal
PHP52	2.96E-29	Not Normal	1.09E-50	Not Normal
MYP42	7.22E-22	Not Normal	3.98E-42	Not Normal
MNP45	2.08E-07	Not Normal	1.61E-15	Not Normal
NZP47	5.67E-13	Not Normal	1.41E-21	Not Normal
JPP37	8.89E-26	Not Normal	5.15E-42	Not Normal
USP80	1.49E-16	Not Normal	1.00E-30	Not Normal
USP79	1.79E-18	Not Normal	2.00E-38	Not Normal
USP78	1.52E-17	Not Normal	1.72E-32	Not Normal
USP77	6.35E-16	Not Normal	1.40E-28	Not Normal
RUP54	2.04E-22	Not Normal	4.29E-39	Not Normal
RUP59	7.17E-16	Not Normal	1.19E-34	Not Normal
RUP61	3.56E-23	Not Normal	4.85E-44	Not Normal

Table 4 presents the outcomes of applying the logarithmic transformation technique to address the skewed data. This transformation method aims to convert the data into a normally distributed form. The results include each variable's skewness, kurtosis, and Shapiro-Wilk test p-value. The skewness and kurtosis values provide insights into the shape of the transformed distribution, while the Shapiro-Wilk test p-value indicates the level of conformity to a normal distribution. The table allows for a comprehensive evaluation of the effectiveness of the logarithmic transformation in achieving normality in the data.

TABLE 4. Results for Logarithmic transformation to convert non-normally distributed data into normally distributed

	Logarithmic transformation			
Station	Skewness	Kurtosis	Shapiro-Wilk p-value	Normality
AUP04	-0.83265	4.301266	2.37E-32	Not Normal
AUP06	-1.74749	8.226549	2.02E-51	Not Normal
AUP08	-0.8862	4.16997	7.47E-35	Not Normal
AUP09	-1.87314	8.357962	2.32E-55	Not Normal
AUP10	-1.31042	7.522323	1.58E-41	Not Normal
GBP66	-0.90611	4.260678	7.80E-36	Not Normal
PHP52	-0.74361	4.197455	1.35E-29	Not Normal

continue ...

... cont.

MYP42	-0.80553	3.731622	1.01E-33	Not Normal
MNP45	-2.39837	14.50765	1.03E-56	Not Normal
NZP47	-1.49942	8.105874	8.56E-46	Not Normal
JPP37	-0.67956	2.964684	5.10E-38	Not Normal
USP80	-0.93028	4.617571	8.75E-37	Not Normal
USP79	-0.92133	5.643169	1.92E-33	Not Normal
USP78	-1.15362	7.341058	1.44E-38	Not Normal
USP77	-0.96666	5.6901	2.53E-35	Not Normal
RUP54	-1.29489	6.159858	2.42E-43	Not Normal
RUP59	-1.15647	5.737928	2.22E-40	Not Normal
RUP61	-1.43078	3.878671	6.36E-30	Not Normal

DATA NORMALIZATION

Such changes aim to reduce the value of each feature in the studied dataset to some value determined within a specific interval while maintaining the overall data distribution. How well the data are normalized significantly impacts how well machine learning algorithms perform (Izonin et al. 2022). Singh and Singh (2020) investigated the impact of data normalization on classification performance and concluded that choosing a normalization method is challenging because a single method cannot tackle all the problems. Normalization helps improve machine learning algorithms' performance by preventing certain variables from dominating the analysis due to their larger scales. In this scenario, it is recommended to apply data normalization techniques to the 7Be data obtained from multiple stations; furthermore, the station's location is in various latitudes. Normalization brings the data points onto a standardized scale, thus mitigating the influence of disparate minimum and maximum values across stations. Doing so can minimize any inherent biases or distortions arising from variations in the data range.

This approach facilitates fair comparisons and more accurate temperature data analyses across stations. Nonetheless, the necessity of data normalization becomes even more pronounced when integrating the temperature data with other meteorological parameters, such as rainfall, humidity, and wind speed, which inherently possess diverse scales or ranges. Normalizing the data ensures a consistent and equitable representation of the various parameters, enabling robust analyses and meaningful comparisons across meteorological factors.

Normalizing these data enables fair comparisons of all parameters across multiple stations, enhancing the discernment of meaningful patterns, trends, or anomalies. Additionally, it facilitates the utilization of machine learning algorithms or statistical models that rely on normalized or standardized data, ensuring optimal performance and accuracy in the analyses.

SELECTION OF AN APPROPRIATE METHOD FOR DATA HANDLING

Selecting an appropriate method for data imputation is also essential, as predictive analysis relies on high-quality input to yield quality results. Three R libraries, namely DMWR, missRanger, and MICE, were employed to address the missing values through imputation. The imputation methods were evaluated using visual inspection and performance evaluation metrics, specifically Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). After careful consideration, missRanger was identified as the most suitable imputation method for the given dataset based on its superior performance in terms of accuracy and precision. This task also reveals that the 7Be data for all selected stations are non-normally distributed. Subsequent analyses must consider this information, indicating that the mean cannot be used to explain the data.

Raw data from all selected stations also show they were skewed. Therefore, the processing of these data should also be taken care of carefully, as parametric tests cannot be done on non-normally distributed data and may result in entirely wrong findings. There is a risk of misrepresenting statistically significant findings as non-significant or non-significant findings as statistically significant, which can lead to erroneous conclusions. Several techniques, including Logarithmic, Box-Cox, Square Root, and Quantile, can transform non-normally distributed data into regularly distributed data to improve interpretability or get insight into the relationship between variables. However, for this study, the Logarithmic method has been tested for the data but is still not applicable to the 7Be data in hand.

Fortunately, several machine learning techniques can effectively handle skewed data and accommodate non-linear relationships. Examples of such methods include Support Vector Machines (SVM), Neural Networks, and Decision Tree algorithms like Random Forests and Classification and Regression Trees (CART). These algorithms can handle skewed data distributions and capture complex patterns and relationships within the data. However, it is essential to remember that some machine learning algorithms might work better or converge more quickly when the data is roughly regularly distributed or when specific conditions are satisfied. In these circumstances, data can be transformed or normalized to resemble a normal distribution before being fed into the algorithm.

CONCLUSIONS

Data on 7Be intensity concentrations were preprocessed and analyzed from 18 chosen radionuclide stations. Therefore, it is crucial to comprehend the anomalies in the data and those that are missing to prevent incorrect interpretation. Identifying outliers must be carefully examined, as it can also lead to errors in interpreting the output. Missing data is a crucial issue that requires significant attention. The radionuclide station RUP54 in Kirov, Russian Federation, exhibits substantial data for 2010, with 97% missing from January 1st to December 22nd. Failing to address this issue diligently may lead to disparate outcomes or erroneous interpretations. Consequently, if the 2010 data holds significant importance, assessing the performance of imputation methods is imperative. Otherwise, case deletion would be considered one of the most suitable approaches to handle this missing value.

In conclusion, data preprocessing is vital in preparing and optimizing data for analysis and modeling. Improving the accuracy and reliability of our findings can be achieved by effectively addressing missing values, handling outliers, normalizing variables, and managing other data quality concerns. Adequate data preprocessing techniques enable us to extract valuable insights, improve predictive models, and make more informed decisions. It is an essential step in the data analysis pipeline that should not be overlooked. With a well-preprocessed dataset, researchers and practitioners can unlock the true potential of their data and drive meaningful discoveries in various fields.

ACKNOWLEDGMENT

We sincerely appreciate the Government of Malaysia's Department of Public Service for the financial support provided through the scholarship grant.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Akasaka, I. et al. 2018. Seasonal march patterns of the summer rainy season in the Philippines and their long-term variability since the late twentieth century. *Progress in Earth and Planetary Science* 5(1): p. 20. doi: 10.1186/s40645-018-0178-5.

Altman, N. and Krzywinski, M. 2016. Analyzing outliers: Influential or nuisance? *Nature Methods* 13(4): pp. 281–2. doi: 10.1038/nmeth.3812.

Bakker, M. and Wicherts, J. M. 2014. Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE* 9(7): pp. 1–9. doi: 10.1371/journal.pone.0103360.

Beck, M. W. et al. 2018. R Package imputeTestbench to Compare Imputation Methods for Univariate Time Series. *The R Journal* 10(1): 218–233. http://www.ncbi.nlm.nih.gov/pubmed/30607263.

Bohari N.Z.I et al. 2021. *Post-Mortem of Northeast Monsoon 2019/2020* Jabatan Meteorologi Malaysia. https://www.met.gov.my/data/research/researchpapers/2021/RP04_2021.pdf

Dash, C. S. K. et al. 2023. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal* 6(May 2022): 100164. doi: 10.1016/j.dajour.2023.100164.

Dong, S. et al. 2020. The most predictable patterns and prediction skills of subseasonal prediction of rainfall over the Indo-Pacific regions by the NCEP Climate Forecast System. *Climate Dynamics* 54(5–6): 2759–2775. doi: 10.1007/s00382-020-05141-5.

Fakaruddin, F. J. et al. 2020. Occurrence of meridional and easterly surges and their impact on Malaysian rainfall during the northeast monsoon: A climatology study. *Meteorological Applications* 27(1): 1–12. doi: 10.1002/met.1836.

Felix, E. A. and Lee, S. P. 2019. Systematic literature review of preprocessing techniques for imbalanced data. *IET Software* 13(6): 479–496. doi: 10.1049/iet-sen.2018.5193.

Geen, R. 2021. Forecasting South China Sea Monsoon onset using insight from theory. *Geophysical Research Letters* 48(6): 1–10. doi: 10.1029/2020GL091444.

Gress, T. W., Denvir, J. and Shapiro, J. I. 2018. Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine* 4(2). doi: 10.18590/mjm.2018.vol4.iss2.9.

Gupta, S. and Gupta, M. K. 2018. A Survey on different techniques for handling missing values in dataset. *NCRACIT International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT* 4(1): 295–301. www.ijsrcseit.com.

Hackenberger, B. K. 2020. R software: unfriendly but probably the best. *Croatian Medical Journal* 61(1): 66–68. doi: 10.3325/cmj.2020.61.66.

Hai, O. S. et al. 2017. Extreme rainstorms that caused devastating flooding across the east coast of Peninsular Malaysia during November and December 2014. *Weather and Forecasting* 32(3): 849–872. doi: 10.1175/WAF-D-16-0160.1.

Hair, J. F. et al. 2021. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R*, *Practical Assessment, Research and Evaluation*.

Haris, M. F. et al. 2023. Cosmogenic Radionuclide-Beryllium 7 (7Be) for monsoon rainfall forecasting in Malaysia: A systematic literature review. *Malaysian Journal of Science Health & Technology* 9(1): 46–55. doi: 10.33102/mjosht.v9i1.344.

Ishak, A. N., Tauhid Ahmad, N. H. and Jit Singh, M. S. 2021. The diurnal variation of rain intensity in Malaysia for monsoon region using TRMM satelit data. *Jurnal Kejuruteraan* 33(3): 719–731. doi: 10.17576/jkukm-2021-33(3)-30.

Izonin, I. et al. 2022. A two-step data normalization approach for improving classification accuracy in the medical diagnosis domain. *Mathematics* 10(11): 1942. doi: 10.3390/math10111942.

Kumar, A. and Singh, S. 2021. A review on Indian summer monsoon rainfall prediction using machine learning techniques. *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*. IEEE, pp. 524–528. doi: 10.1109/ICSCCC51823.2021.9478104.

Madley-Dowd, P. et al. 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology* 110: 63–73. doi: 10.1016/j.jclinepi.2019.02.016.

Mohanty, U. C. et al. 2019. A review on the monthly and seasonal forecast of the Indian summer monsoon. *Mausam* 70(3): 425–442. doi: 10.54302/mausam.v70i3.223.

Nugroho, H. and Surendro, K. 2019. Missing data problem in predictive analytics. *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. New York, NY, USA: ACM, pp. 95–100. doi: 10.1145/3316615.3316730.

Peng, J., Hahn, J. and Huang, K.-W. 2020. Handling missing values in information systems research: A review of methods and assumptions. *SSRN Electronic Journal* (May). doi: 10.2139/ssrn.3560070.

Perez, H. and Tah, J. H. M. 2020. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. *Mathematics* 8(5): 662. doi: 10.3390/math8050662.

Pollet, T. V. and van der Meij, L. 2017. To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. *Adaptive Human Behavior and Physiology* 3(1): pp. 43–60. doi: 10.1007/s40750-016-0050-z.

Rajan, D. and Desamsetti, S. 2021. Prediction of Indian summer monsoon onset with high resolution model: A case study. *SN Applied Sciences* 3(6): 645. doi: 10.1007/s42452-021-04646-w.

Rao, S. A. et al. 2019. Seasonal prediction of indian summer monsoon in India: The past, the present and the future. *Mausam* 70(2): pp. 265–276.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika* 63(3): 581–592. doi: 10.1093/biomet/63.3.581.

Saha, M. et al. 2021. Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *International Journal of Forecasting*, 37(1): 58–71. doi: 10.1016/j.ijforecast.2020.03.001.

Saha, M., Mitra, P. and Nanjundiah, R. S. 2017. Deep learning for predicting the monsoon over the homogeneous regions of India. *Journal of Earth System Science* 126(4): 54. doi: 10.1007/s12040-017-0838-7.

Seu, K., Kang, M.-S. and Lee, H. 2022. An intelligent missing data imputation techniques: A review. *JOIV : International Journal on Informatics Visualization* 6(1–2): 278. doi: 10.30630/joiv.6.1-2.935.

Shamsudin, H. et al. 2023. An optimized support vector machine with genetic algorithm for imbalanced data classification. *Jurnal Teknologi* 85(4): 67–74. doi: 10.11113/jurnalteknologi.v85.19695.

Sharma, A. et al. 2021. Rainfall prediction: Analysis of machine learning algorithms and ensemble techniques. *2021 7th International Conference on Signal Processing and Communication (ICSC)*. IEEE, pp. 234–240. doi: 10.1109/ICSC53193.2021.9673275.

Singh, D. and Singh, B. 2020. Investigating the impact of data normalization on classification performance. *Applied Soft Computing* 97(xxxx): 105524. doi: 10.1016/j.asoc.2019.105524.

Tan, M. L. and Santo, H. 2018. Comparison of GPM IMERG, TMPA 3B42 and PERSIANN-CDR satellite precipitation products over Malaysia. *Atmospheric Research* 202(July 2017): pp. 63–76. doi: 10.1016/j.atmosres.2017.11.006.

Tang, K. H. D. 2019. Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. *Science of the Total Environment* 650: 1858–1871. doi: 10.1016/j.scitotenv.2018.09.316.

Tangang, F. et al. 2012. Climate change and variability over Malaysia : Gaps in science and research information climate change and variability over Malaysia : Gaps in science and research information. *Sains Malaysiana* (November).

Tawn, R., Browell, J. and Dinwoodie, I. 2020. Missing data in wind farm time series: Properties and effect on forecasts. *Electric Power Systems Research* 189: 106640. doi: 10.1016/j.epsr.2020.106640.

Terzi, L. et al. 2019. How to predict seasonal weather and monsoons with radionuclide monitoring. *Scientific Reports* 9(1): 2729. doi: 10.1038/s41598-019-39664-7.

Werner de Vargas, V. et al. 2023. Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowledge and Information Systems* 65(1): 31–57. doi: 10.1007/s10115-022-01772-8.

Yehia, T. et al. 2022. Removing the outlier from the production data for the decline curve analysis of shale gas reservoirs: A comparative study using machine learning. *ACS Omega* 7(36): 32046–32061. doi: 10.1021/acsomega.2c03238.

Yip, B. et al. 2016. Review of the November 2015 – March 2016 Northeast Monsoon in Malaysia, Jabatan Meteorologi Malaysia. https://www.researchgate.net/publication/325314002_REVIEW_OF_THE_NOVEMBER_2015_-_MARCH_2016_NORTHEAST_MONSOON_IN_MALAYSIA

Zhang, C. et al. 2013. The nearest neighbor algorithm of filling missing data based on cluster analysis. *Applied Mechanics and Materials* 347–350: 2324–2328. doi: 10.4028/www.scientific.net/AMM.347-350.2324.