Comparative Analysis of PCA and ANOVA for Assessing the Subset Feature Selection of the Geomagnetic Disturbance Storm Time

Ain Dzarah Nafisah M., Muhamad Asraf H., Nooritawati M. T., Nur Dalila K. A., and Mohamad Huzaimy Jusoh

***Abstract*— A Disturbance Storm Time (Dst) index represents the geomagnetic storm strength due to interaction of the Sun towards Earth in the space weather. Formation of the Dst contributed by the total of nine (9) input features namely interplanetary magnetic field (IMF), solar wind density (SWD), solar wind speed (SWS), solar wind input energy (SWIP) and also Earth's magnetic field components comprise of the horizontal intensity component (H), declination component (D), north component (N), east component (E), and vertical intensity component (Z). Large datasets which comprise of 157896 number of data have existed for all features thus require pre-processing and subset feature selection for reducing data dimensionality in order to reduce the data processing time and enhance the performance of the learning algorithm. In this paper two methods of analyzing the features were compared: Principal Component Analysis (PCA) and one-way Analysis of Variance (ANOVA). The main aims for this works are to reduce a large set of input parameters from the Dst index and to compare the subset feature using the proposed methods for acquiring the reduced features. Prior to analyse the features, an independent-samples t-test is used to evaluate if there is a large difference between the mean of two groups that can be correlated with certain characteristics. The results for the features analyzed demonstrated that one-way ANOVA performed better in eliminating seven (7) components out of nine (9) components of features as compared to PCA. This finding was validated with a dendrogram to support that one-way ANOVA outperformed the PCA in reducing the subset features.**

***Index Terms*— Dst index, geomagnetic storm, Principal Component Analysis (PCA), one-way ANOVA, D, space weather**

I. Introduction

SPACE weather is normally referring to the Sun and space conditions that affect the Earth's technological performance. Space weather conditions can affect the Earth's ground technological systems due to the Sun explosion that produces the geomagnetic storm which also known as a solar storm.

This manuscript is submitted on 15th April 2020 and accepted on 8th October 2020. This work was funded by the Faculty of Electrical Engineering, Universiti Teknologi MARA, and the Ministry of Education, Malaysia under the Fundamental Research Grant Scheme; 600-IRMI/FRGS 5/3 (091/2019).

Ain Dzarah Nafisah M., Nooritawati M.T. and Mohamad Huzaimi Jusoh are from the Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia (e-mail: noori425@uitm.edu.my)

Muhammad Asraf H. is with the Faculty of Electrical Engineering, Universiti Teknologi MARA, Pasir Gudang Campus, 81750 Masai, Johor, Malaysia (e-mail: masraf@uitm.edu.my)

Nur Dalila K. A. is with the Faculty of Electrical Engineering, Universiti Teknologi MARA Pasir Gudang Campus, 81750 Masai, Johor, Malaysia (e-mail: nurdalila306@uitm.edu.my)

The changes in the geomagnetic storm carries a negative impact on the Earth's electromagnetism and a space weather condition. The means of measuring the geomagnetic storm can be characterized from the index of the disturbance storm time (Dst) index which comprises of three phases of the Dst index, which are weak (Dst> -50nT), moderate (-100nT < Dst < -50nT) and intense (Dst < -100nT) [1].

The magnetic storm strength of Dst index is formed by the space environment which is derived from the input parameters known as the subset features. This subset features are derived from IMF, solar wind parameters and Earth's magnetic field components [2]. As part of severity indicator, the disturbance storm time shown that when the IMF turns southward, the Dst index level increases and slowly begin to rise back to a quiet time level when the IMF turns northwards [3]. In addition, the usefulness of the Dst index may include a geomagnetic storm forecast to be categorised as mild, moderate, and severe geomagnetic storm.

In this analysis, all of the above features include large datasets containing thousands of raw data, and therefore a reduction in dimensionality may be proposed to extract and reduce features into subset features by eliminating unnecessary and redundant information [4]. The large size of data normally encompasses data similarity which results in diverse dimensionalities in the datasets [5]. This required to reduce the data effectively by proposing the suitable methods; hence in this paper the principal component analysis (PCA) and one-way ANOVA were implemented. These methods are widely used in assessing the subset features and beneficially proven to reduce data dimensionality.

Previous literatures have demonstrated the capability of PCA and ANOVA in their research works. S. Mubarak. and H. Darwis. [6], revealed that there are only four features dominant among the 333 features derived by using pca and manage to optimize the precision of the classification. J. A. Awomeso and S. M. Ahmad [7], stated that the PCA had showed as a very functional tool that reveals a possible source of contamination to the groundwater quality. According to Yuanyuan Sun and Hongtao Shan [8], which also used PCA stated that the dimension is reduced and improved the accuracy of the prediction for a neural network. Harb Hassan [9] stated that the one-way ANOVA results had reduced the data redundancy and extended the network lifetime. According to Z. H. Bohari and M. K. Nor[10],the authors suggested on doing a study of building energy using ANOVA and the result shows that the ANOVA method is very viable

to use for the operation.

In this paper, the objectives are highlighted as follows: firstly, to identify the relevant number of subset features as input to form the Dst index. Secondly to reduce the irrelevant and redundant features using a PCA and one-way ANOVA methods. Finally, the validation outcome will be presented via t-test assessment, in addition to a dendrogram illustration for demonstrating hierarchical relationship between the subset features. The contribution of this paper demonstrates the details procedures of conducting and analyzing the feature selection from the PCA and one-way ANOVA, hence determine the successful reduction of subset features to be achieved. Even though these two methods were profoundly implemented as feature selection, however limited study discovered for space weather study in Malaysia to probe further on data correlation between the features. This paper is organised by brief explanation of feature selection in section II, material and methods in section III includes methodology block diagram. Section IV demonstrates the results and discussion while Section V provides a conclusion of the proposed works.

II. FEATURE SELECTION

Feature selection usually selecting the attributes from the given dataset that are applicable for constructing a model. The purpose of using feature selection is to avoid the curse of dimensionality that often occurs when organizing data in high dimensional space to low dimensional space [11]. This feature selection is important in eliminating unfitting and unnecessary data in order to enhance the performance of the learning algorithms [12]. Popular algorithms like PCA and ANOVA have been used as a feature selection for a dataset with many features.

A. Principal Component Analysis – PCA

PCA has been commonly used to lessen a large number of data dimension into smaller set of new data as result of simplification which reduces the data processing time [13] while keeping the variability present utterly in the dataset [14]. PCA is known as a multivariate control technique that is design to turn the data into a reduced form and retain much of the actual variance in the new data [15]. This PCA allows to determine the differences between the data and hence provide correlations between the features used to obtain a new set of the reduced data [16]. The procedures of implementing the PCA are shown as following:

Step 1: The data will be normalized first to produces a dataset whose mean is zero. The dataset has two dimensions which comprise of X and Y.

Step 2: The covariance matrix will be calculated. Since the dataset is two-dimensional, the result forms a 2-by-2 covariance matrix, refer to equation (1).

$$Matrix(covariance) = \begin{bmatrix} \mathrm{Var}[\mathrm{X}_1] & \mathrm{Cov}[\mathrm{X}_1, \mathrm{X}_2] \\ \mathrm{Cov}[\mathrm{X}_2, \mathrm{X}_1] & \mathrm{Var}[\mathrm{X}_2] \end{bmatrix} \quad (1)$$

Step 3: The eigenvalues and eigenvectors are then calculated for the covariance matrix. λ is an eigenvalue of matrix A if it is for the solution of the characteristic equation as shown in equation (2)

$$\det(\lambda \mathrm{I} - \mathrm{A}) = 0 \quad (2)$$

where I is known as the identity matrix from the same dimension as A which is a required condition for the matrix subtraction. While 'det' is the determinant of the matrix. A corresponding eigenvector ν, for each of the eigenvalue λ can be computed by equation (3)

$$(\lambda \mathrm{I} - \mathrm{A})\nu = 0 \quad (3)$$

Step 4: The eigenvalue is ordered from largest to smallest value. The dimensionality reduction starts here. To decrease the dimensions, the first p eigenvalues is chosen, and the other will be ignored. Next, the eigenvectors as shown in equation (4) will be formed.

$$Featurevector = (\mathrm{eig}_1, \mathrm{eig}_2) \quad (4)$$

Step 5: The principal components is formed in equation (5). The invert of the feature vector will be left-multiplied with the invert of scaled data of the actual dataset.

$$NewData = Featurevector^T \; x \; ScaledData^{\mathrm{T}} \quad (5)$$

where, *NewData* is the matrix that consist the principal components. The *Featurevector* is the matrix that was develop using the eigenvectors and the *ScaledData* is the scaled version of the actual dataset. '*T*' to indicate as the invert of a matrix which is formed by exchange the rows to column and column to rows.

B. One-way Analysis of Variance – ANOVA

Another method to select features known as one-way ANOVA It is used to differentiate the averages of two or more experiments [17]. The comparison is conducted for determining whether the features are significantly difference based of the associated population means. This one-way ANOVA gives a measurable test to see whether the means of several data are equivalent [9]. The study about the variance between the data of the input parameters is an effective way to examine the data redundancy.

The null hypothesis (H_o) in ANOVA expects that the variance between input parameters data is not significant. The test for the one-way ANOVA is known as F-test. The F-test indicates whether any significant difference between two or more populations. The formula used for the F-test is in equation (6) [18],

$$F = \frac{Variance\ between\ the\ sample\ (MSR)}{Variances\ within\ the\ sample\ (MSE)} \quad (6)$$

This F-distribution statistic is analysed by referring to the degrees of freedom n-k and k-1 respectively. Where n is known as the number of data values in all of the data sets while k is the number of populations under consideration [19].

III. Materials and Methods

The block diagram of methodology implemented to assess the subset features selection is depicted in Fig. 1, which comprises of several stages namely data collection, pre-processing via data normalization, and feature selection.

A. Data collection

For this work, total of nine data accumulated mainly consists of IMF, three parameters of solar wind and five components of Earth's magnetic field. The data was collected in the location of Langkawi, Malaysia. The datasets for the IMF and solar wind parameters comprises of SWD,SWS and SWIP were obtained from OMNI web which can be retrieved at NASA's Space Physics Data Facility (SPDF) which is available at *https://omniweb.gsfc.nasa.gov/index.html.* Meanwhile for Earth's magnetic components composes of H, D, N, E, Z can be retrieved at SuperMAG website available at *http://supermag.jhuapl.edu/.* The data sets used in this work was restricted to the years of 2014 and 2016, which were analysed using the SPSS platform with 157,896 data points. The year of 2016 is the most recent year in the database that has received a dataset since the magnetometers installed at the Langkawi station were unable to retrieve Earth's magnetic field data for the other year.

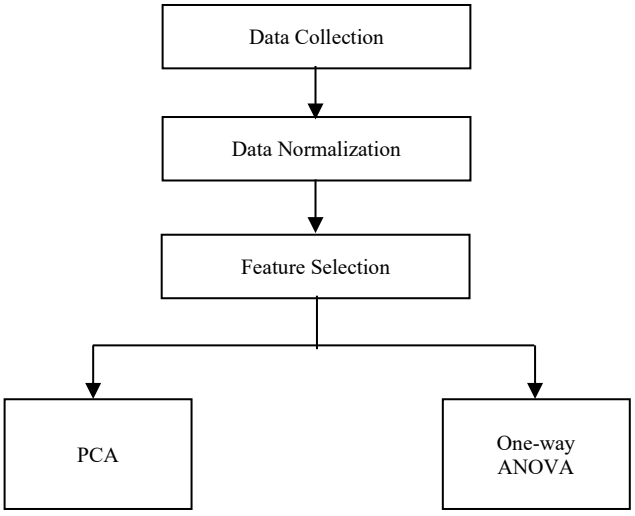


Fig. 1. Block diagram of the methodology

B. Pre-processing data

Data collected for preprocessing to examine whether the data contain outliers that need cleaning for further processing stage. After examining the data, it showed that the missing values in the data were existed. The missing values data then will be replaced by the mean value of the data [20].

Then, prior to the feature reduction process, the whole data need to be normalized within the interval between 0 to 1 by using a formula in equation (7) [21],

$$X = \frac{x\text{-}min(x)}{max(x)\text{-}min(x)} \tag{7}$$

where, x is the initial value of each variable while min(x) and max(x) indicates the maximum and minimum of each of the initial variable values.

C. Feature selection

The feature selection is examined with two techniques namely PCA and one-way ANOVA. PCA is known as a variable-reduction technique capable to reduce a larger set of parameters into a smaller set of new parameters, known as 'principal components', which account for most of the variance in the original variables. The first principal component usually retains the maximum variation that was present in the original components. The steps involved in the PCA is shown in Fig. 2.

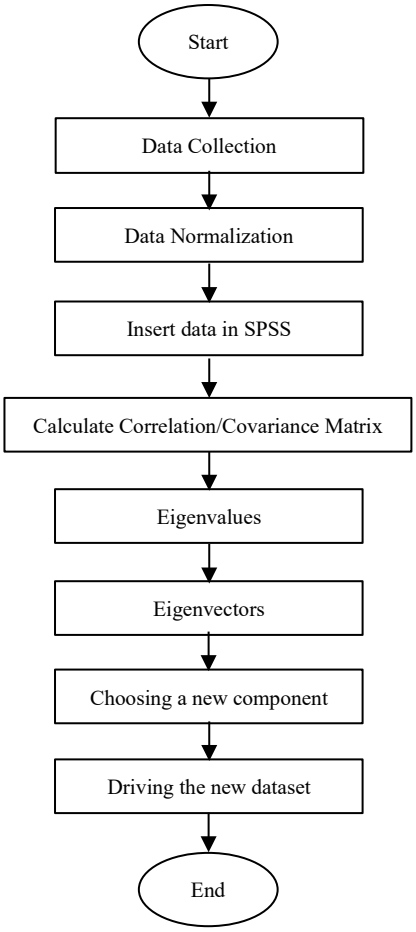


Fig. 2. Flowchart of PCA

The principal components are the eigenvectors of a covariance matrix. A component that had a high eigenvalue usually represent a real underlying factor. The components were selected with at least the value of the Eigenvalue is 1 or more than 1. Other than that, the remaining were considered as scree. Considering another method of one-way ANOVA, it is a parametric test involves statistical evidence by determining whether the input feature in the dataset are significantly difference. The steps involved in the one-way ANOVA is shown in Fig. 3.

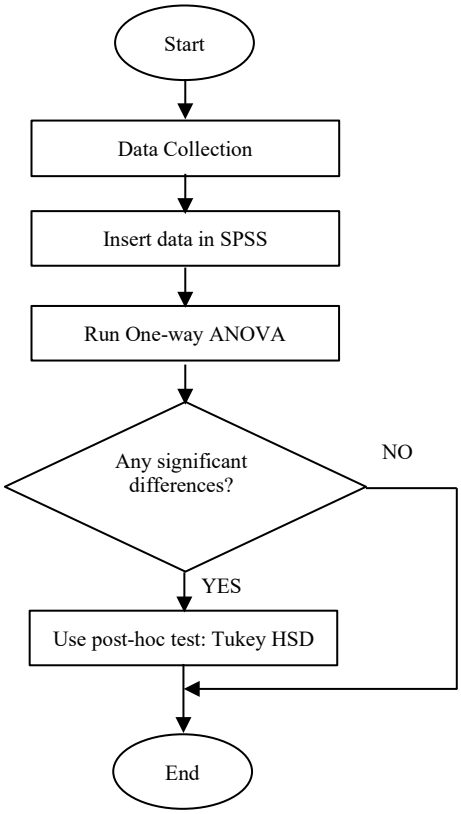


Fig. 3. Flowchart of One-way ANOVA

One-way ANOVA signifies the overall difference between the groups but does not tell which specific groups are differed. By conducting the post hoc test, the confirmation of the differences occurred between the groups can be identified. Post hoc test controlled the experiment wise error rate, which is below than 0.05. In normal case, if the data met the assumption of the homogeneity of variances, the Tukey's HSD (Honestly Significant Difference) post hoc test will be used. The Tukey's will compare the differences between the means of values. If the input feature in the dataset has not demonstrating the significantly difference, the mean will be ignored.

IV. Eperimental Results and Discussion

Comparisons between the PCA and one-way ANOVA were evaluated on the subset features. The results are discussed as follows.

A. Independent-samples t-test

The independent-samples t-test were used if two groups are dependent on each other in contrast [22]. To test the validity of the mean of the random data and the discrepancy between the means of two variables, the t-test is used [23] by comparing the means between the year of 2014 and 2016.

TABLE I
Independent Sample Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
Input Parameter	Month/ Year	F	Sig.	t	df	Sig. (2-tailed)
	Jan14/ Jan16	0.062	0.803	0.609	10419	0.542
	Feb14/ Feb16	0.054	0.816	-0.275	9707	0.783
	Mar14/ Mar16	1.195	0.274	-2.799	10339	0.005
	Apr14/ Apr16	2.193	0.139	-1.893	9665	0.058
	May14/ May16	7.323	0.007	-0.123	10258	0.217
	Jun14/ Jun16	0.099	0.753	-1.163	10128	0.245
	Jul14/ Jul16	57.566	0.000	0.381	9315	0.703
	Aug14/ Aug16	422.160	0.000	8.920	8243	0.000
	Sep14/ Sep16	411.889	0.000	8.240	7910	0.000
	Oct14/ Oct16	17.330	0.000	-2.024	9304	0.043
	Nov14/ Nov16	35.597	0.000	-0.735	6136	0.462
	Dec14/ Dec16	0.831	0.362	0.644	5946	0.519

Observation of these two data to see whether there is a significant difference between the two years and to see which month to assess a performance can be used for the next step. Table I demonstrates that the month January, February, April, May, June, July, and December of the year 2014 and 2016 are not statistically significant difference as the value shown in the Sig. (2-tailed) is greater than 0.05 which is 0.542, 0.783, 0.058, 0.217, 0.245, 0.703, and 0.519. From the Sig. (2-tailed), it shows that the null hypothesis is accepted. It shows that only the dataset from the seven-month stated will be used for performance testing.

B. Principal Component Analysis – PCA

A correlation matrix of the data in Table II shows the correlations between the features or input parameters used to form the Dst index. All the input parameters show a correlation between features except for the input parameters of H, D, N and E.

TABLE II
CORRELATION MATRIX

		IMF	SWD	SWS	SWIP	H	D	N	E	Z
Correlation	IMF	1.000	0.069	0.084	0.538	-0.058	-0.005	-0.057	0.010	0.007
	SWD	0.069	1.000	-0.014	0.028	0.004	0.005	0.004	0.004	0.040
	SWS	0.084	-0.014	1.000	0.130	0.026	-0.038	0.028	0.001	0.148
	SWIP	0.538	0.028	0.130	1.000	-0.134	0.020	-0.133	0.031	0.005
	H	-0.058	0.004	0.026	-0.134	1.000	-0.017	1.000	0.023	0.392
	D	0.005	0.05	-0.038	0.020	-0.017	1.000	-0.017	0.992	0.002
	N	-0.057	0.004	0.028	-0.133	1.000	-0.017	1.000	0.023	0.393
	E	0.010	0.004	0.001	0.031	0.023	0.992	0.023	1.000	0.030
	Z	0.007	0.040	0.148	0.005	0.392	0.002	0.393	0.030	1.000

TABLE III
COMMUNALITIES

	Initial	Extraction
IMF	1.000	0.720
SWD	1.000	0.648
SWS	1.000	0.499
SWIP	1.000	0.731
H	1.000	0.931
D	1.000	0.997
N	1.000	0.932
E	1.000	0.996
Z	1.000	0.426

TABLE IV
TOTAL VARIANCE EXPLAINED

	Initial Eigenvalues			Extraction Sum of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative%
1	2.292	25.470	25.470	2.292	25.470	25.470	2.256	25.066	25.066
2	1.995	22.167	47.637	1.995	22.167	47.637	1.993	22.146	47.212
3	1.575	17.495	65.131	1.575	17.495	65.131	1.606	17.847	65.058
4	1.018	11.308	76.439	1.018	11.308	76.439	1.024	11.381	76.439
5	0.963	10.695	87.134						
6	0.699	7.770	94.905						
7	0.452	5.027	99.932						
8	0.006	0.068	100.00						
9	3.473E-6	3.859E-5	100.00						

From Table II, the value of the correlation matrix has demonstrated the identical value for both H and D components, hence similar feature can be omitted. Table III demonstrates the communalities by examining the Squared Multiple Correlation coefficient (SMC) or also known as R^2 for predicting the variable from the components, which is the proportion of the variance of the variables that has been extracted by the components. The initial value for the input parameters indicated as 1 while as for the extraction, it describes the proportion of variances of each input parameter. The variances should be at least 0.40. Any input parameter with a low R^2 should be discarded. Therefore, the input parameters of D and E have produced the results of a higher value of variance as good selection of feature value.

Meanwhile Table IV indicates the total of variance explained on initial eigenvalue and the rotation sum of squared loadings by each of the principal components. Each of the eigenvalue constitutes the amount of variance by each component. The first principal component indicates the biggest amount of the variance while the second largest amount of variance was dedicated for the second principal component. A component with a high eigenvalue usually represents a real underlying factor. The components to be selected at least with eigenvalue is 1. Only four components from the total show that the eigenvalue was more than 1. Other components were not assumed to represent real traits of underlying factors and the component are considered as scree. The main reason of the principal component analysis is to lessen the amount of input parameter used, hence it needs to have fewer components.

The principle component should have a large value of eigenvalue and many should have a small value of eigenvalue input parameters which share a significant variance. After the rotation, the variance explained is equal to the sums of squared loading (SSL). The value of the principle component reduces to 4 as only 4 principal components that have eigenvalues ≥ 1. The

component of 1 value before rotation is 2.292 while it changes to 2.256 after the rotation. For component 2 the SSL's are 1.995 and 1.993. While for component 3 and 4 the SSL's are 1.575 and 1.606 and 1.018 and 1.024. After the rotation, the total variance of the four components is shown in the equation (8).

$$Total\ variance = \frac{2.292+1.995+1.575+1.018}{9} = 0.764 \quad (8)$$

The total variance for the first four principal components reached 76.4%. Usually one would aim for 100% of total variance which is impossible to gain but often analyses reported that the total variance is between 60% and 70% [24]. As seen from Table IV, the value of the Cumulative % and the % of Variance in the Extraction Sums of the Squared Loadings is the same as the Initial Eigenvalues, but the value of the % of Variance and the Cumulative % is changing in the Rotation Sums of Squared Loadings as the rotation has the effect in optimizing the components structure and make the four components left balanced. Subsequent analysis of the rotated component matrix table in Table V showed a rotation component matrix with a 'Varimax' rotation to find the linear of the constructs [25] and also to redistribute the variance to get a simpler form of the dataset [26]. The factor loading showed was sorted by size. The 0.958 is the strongest factor loading which located on the most top while the weakest factor loading of -0.630 was located at the lowest. As it can be seen the input parameter N, H and Z are load in the first principal component, D and E load in the second principal component, SWIP and IMF load in third principal component and SWD and SWS load in the fourth principle component.

TABLE V
ROTATED COMPONENT MATRIX

	1	2	3	4
N	0.958			
H	0.957			
Z	0.627			
D		0.998		
E		0.997		
SWIP			0.848	
IMF			0.847	
SWD				0.771
SWS				-0.630

The scree plot as shown in Fig. 4 demonstrates a distribution of the variance among the principal components graphically. In this scree plot, those components that are at the bottom is called scree and known as weak factors. For the principal component, the eigenvalue is plotted in the y-axis while the component number is plotted in the x-axis. Only the first four (4) components that have a value of eigenvalue more than one (1). There existed a big drop between component eight (8) and nine (9). As the total variance of the component one (1) until four (4) is 76.4%, only the first four (4) components should be retained.

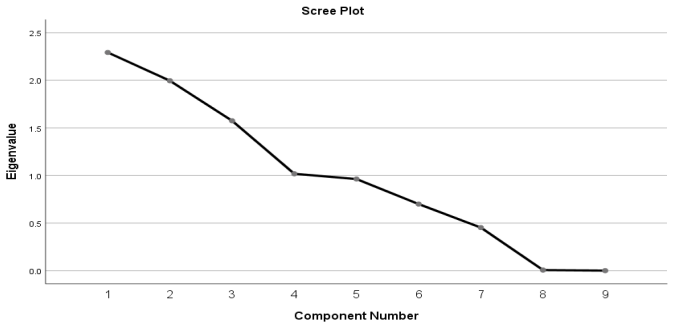


Fig. 4. Scree Plot

Further analysis to verify the reduced subset features can be illustrated using a dendrogram. Dendrogram is a diagram that were illustrated to shows the hierarchical relationship between the components and proved the cluster features similar in PCA analysis. From Fig. 5, the input features that are more similar to each other are grouped together. The vertical line in the dendrogram represents the grouping of the input feature. The vertical line will be located farther to the right side, as the clusters that were being merged become more varied. As for the horizontal lines, it represents the differences in the distances which connect all the principal component that are part of one cluster. It is essential to decide the final number of clusters after the stopping decision is made. The longest horizontal lines represent the largest different between the components. So, the long horizontal lines will show that the components that are dissimilar to each other which are PC1 and PC3 are being combined and discover where the most favorable place to stop the clustering procedure. If the vertical and horizontal lines are close to each other, then it shows that the level of the homogeneity of the clusters combine at those stages is stable.

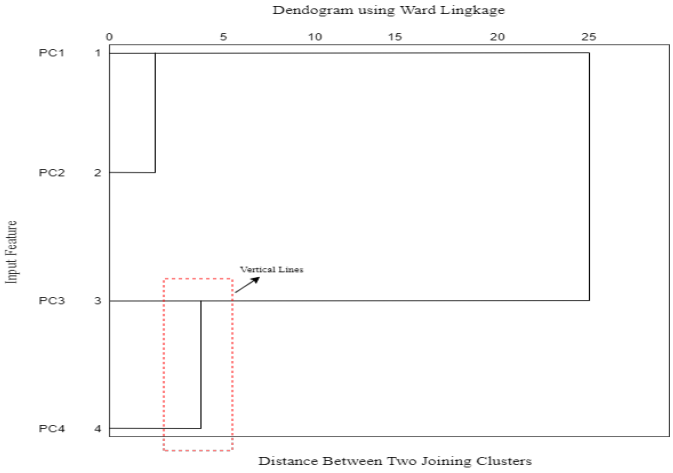


Fig. 5 Dendrogram for PCA

C. *One-way Analysis of Variance – ANOVA*

Table VI below shows the ANOVA results generated by the SPSS. It shows the whether the F-value in the Between Groups row reached significance. Therefore, as it can be the F-value is equal to 4642.823, which reaches significance with a *p*-value of 0.000 which is below than the alpha level, 0.005. This shows that the means of the dissimilar amount of the variable is statistically significance. However, it still did not tell which of the various pairs of means the is significantly difference. The result will be shown in the post hoc Tukey HSD (Honest Significance Difference) test in Table VI.

TABLE VI
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.06E+40	8	2.58E+39	4642.823	0.000
Within Groups	7.23E+40	130281	5.55E+35		
Total	9.28E+40	130289			

This Tukey HSD generally preferred test for one-way ANOVA. This Tukey's is used to assess the contrasts among the sample means to see if there is any significance. The Tukey's will test all the pairwise differences. As from the table VI below, the results shows that all the input feature did not reaches the significance with each other where the *p*-value is 1.000, which is more than the standard 0.05 alpha level except only with the input feature Solar Wind Input Energy (SWIP), which the *p*-value is 0.000. If the significance value below than 0.05, therefore, there is a significant difference between the input feature. The dendrogram is illustrated for ANOVA is shown in Fig. 6. It shows that the input features H, N, IMF, D, E, SWD, SWS, and Z were grouped within the same grouped showing that the input features are more similar to each other except for the SWIP. As the input features H, N, IMF, D, E, SWD, SWS, and Z are similar to each other, only one of the input features will be selected, leaving only two(2) out of nine (9) input features available namely SWIP and IMF.

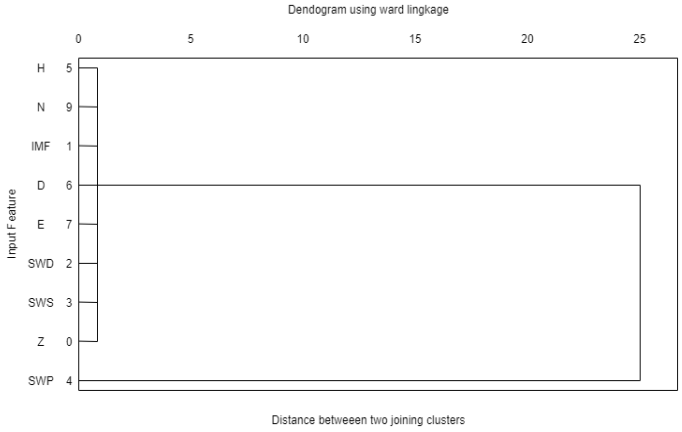


Fig. 6 Dendrogram for ANOVA

TABLE VII
MULTIPLE COMPARISON

		IMF	SWD	SWS	SWIP	H	D	N	E	Z
IMF	Mean (SD)	-	-16.82535	-416.24778	-1.16E+18 *	-41712.423	5.35362	-17586.795	-23090.953	951.41741
	Sig.	-	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
SWD	Mean (SD)	16.82535	-	-399.42242	-1.16E+18 *	-41695.598	22.17897	-17569.969	-23074.127	968.24276
	Sig.	1.000	-	1.000	0.000	1.000	1.000	1.000	1.000	1.000
SWS	Mean (SD)	416.24778	399.42242	-	-1.16E+18 *	-41296.176	421.60140	-17170.549	-22674.705	1367.6652
	Sig.	1.000	1.000	-	0.000	1.000	1.000	1.000	1.000	1.000
SWIP	Mean (SD)	1.16E+18*	1.16E+18	1.16E+18	-	1.16E+18	1.16E+18	1.16E+18	1.16E+18	1.16E+18
	Sig.	0.000	0.000	0.000	-	0.000	0.000	0.000	0.000	0.000
H	Mean (SD)	41712.423	41695.5979	41296.1756	-1.16E+18 *	-	41717.777	24125.629	18621.476	42663.841
	Sig.	1.000	1.000	1.000	0.000	-	1.000	1.000	1.000	1.000
D	Mean (SD)	-5.35362	-22.17897	-421.60140	-1.16E+18*	-41717.777	-	-17592.148	-23096.306	946.06379
	Sig.	1.000	1.000	1.000	0.000	1.000	-	1.000	1.000	1.000
N	Mean (SD)	17586.7946	17569.9693	17170.5469	-1.16E+18*	-24125.629	17592.148	-	-5504.1582	18538.212
	Sig.	1.000	1.000	1.000	0.000	1.000	1.000	-	1.000	1.000
E	Mean (SD)	23090.9528	23074.1275	22674.7050	-1.16E+18*	-18621.471	23096.306	5504.1582	-	24042.370
	Sig.	1.000	1.000	1.000	0.000	1.000	1.000	1.000	-	1.000
Z	Mean (SD)	-951.41741	-968.24276	-1367.6652	-1.16E+18	-42663.841	-946.06379	-18538.212	-24042.370	-
	Sig.	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	-

V. CONCLUSION

This paper shows a result of the independent-samples t-test, principal component analysis and one-way ANOVA using SPSS. The independent-samples t-test result shows that the month of January, February, April, May, June, July, and December of the year 2014 and 2016 are not statistically significant difference as the value shown in the Sig. (2-tailed) is greater than 0.05. Therefore, only the dataset from these months are acceptable for performance testing. As for the PCA result, after extracting the component, there are only four components that need to keep while the other component might be deleted. As for the one-way ANOVA, all the input parameter did not reaches the significance with each other where the *p*-value is 1.000, which is more than the standard 0.05 alpha level except only with the input parameter Solar Wind Input Energy (SWIP), which the *p*-value is 0.000. So, from the one-way ANOVA, only two (2) input features will be selected and the remaining seven (7) input features will be eliminated. The input features that will be selected is Solar Wind Input Energy (SWIP) and only one of the input features in the same group as shown in the dendrogram which is the IMF input features. As for the future, the performance between PCA and one-way ANOVA will be compared to see which methods will give the best accuracy output.

VI. ACKNOWLEDGMENT

The author gratefully acknowledges the Ministry of Education (MOE), Malaysia and Faculty of Electrical Engineering, Universiti Teknologi MARA for granted grant; Fundamental Research Grant

Scheme (Grant No. 600-IRMI/FRGS 5/3 (091/2019)). Author also acknowledge use of OMNI data retrieved at NASA/GSFC's Space Physics Data Facility. In addition, the results presented in this paper rely on data collected at magnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice (www.intermagnet.org).

REFERENCES

[1] T. L. Gulyaeva and A. J. Mannucci, "Echo of ring current storms in the ionosphere," *J. Atmos. Solar-Terrestrial Phys.*, vol. 205, no. February, p. 105300, 2020.

[2] D. Al-feadh and W. Al-ramdhan, "Large Geomagnetic Storms Drives by Solar Wind in Solar Cycle 24 Large Geomagnetic Storms Drives by Solar Wind in Solar Cycle 24," *J. Phys. Conf. Ser.*, 2019.

[3] K. K. Hashimoto *et al.*, "Penetration electric fields observed at middle and low latitudes during the 22 June 2015 geomagnetic storm," *Earth, Planets Sp.*, 2020.

[4] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," vol. 2015, no. 1, 2015.

[5] M. H. ur Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, and S. U. Khan, "Big Data Reduction Methods: A Survey," *Data Sci. Eng.*, vol. 1, no. 4, pp. 265–284, 2016.

[6] S. Mubarak, H. Darwis, F. Umar, L. B. Ilmawan, S. Anraeni, and M. A. Mude, "Feature Selection of Oral Cyst and Tumor Images Using Principal Component Analysis," *2018 2nd East Indones. Conf. Comput. Inf. Technol.*, pp. 322–325, 2018.

[7] J. A. Awomeso, S. M. Ahmad, and A. M. Taiwo, "Multivariate assessment of groundwater quality in the basement rocks of Osun State, Southwest, Nigeria," *Environ. Earth Sci.*, vol. 79, no. 5, pp. 1–9, 2020.

[8] Y. Sun, H. Shan, W. Zhang, L. Ren, and P. Yan, "Reliability prediction of distribution network based on PCA-GA-BP neural network," *AIP Conf. Proc.*, vol. 2154, no. September, 2019.

[9] H. Harb, A. Makhoul, and R. Couturier, "An Enhanced K-Means and ANOVA-Based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks," *IEEE Sens. J.*, vol. 15, no. 10, pp. 5483–5493, 2015.

[10] Z. H. Bohari, R. Ghazali, N. N. Atira, M. F. Sulaima, A. A. Rahman, and M. K. Nor, "Building energy management saving by considering lighting system optimization via ANOVA method," *2018 4th Int. Conf. Comput. Technol. Appl. ICCTA 2018*, vol. 13, no. 13, pp. 216–220, 2018.

[11] B. M. Gayathri and C. P. Sumathi, "Feature selection using Linear Discriminant Analysis for breast cancer dataset," *2018 IEEE Int. Conf. Comput. Intell. Comput. Res.*, pp. 1–5, 2018.

[12] N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification," *2019 Int. Conf. Comput. Intell. Data Sci.*, pp. 1–6, 2019.

[13] S. Bhadauria, "Introduction to Principal Component Analysis in Applied Research," *New Man Int. J. Multidiscip. Stud.*, vol. Vol.1, no. December 2014, pp. 67–75, 2014.

[14] M. M. Than, K. M. Yee, K. Lint, M. Han, and T. W. Hnin, "Determining Spatial and Temporal Changes of Water Quality in Hlaing River using Principal Component Analysis," pp. 1–7, 2020.

[15] A. Hadri, K. Chougdali, and R. Touahni, "Intrusion Detection System using PCA and Fuzzy PCA Techniques," 2016.

[16] E. Banguero, A. Correcher, Á. Pérez-Navarro, E. García, and A. Aristizabal, "Diagnosis of a battery energy storage system based on principal component analysis," *Renew. Energy*, vol. 146, pp. 2438–2449, 2020.

[17] M. Rubessa *et al.*, "Morphometric analysis of sperm used for IVP by three different separation methods with spatial light interference microscopy," *Syst. Biol. Reprod. Med.*, vol. 66, no. 1, pp. 26–36, 2020.

[18] R. De Souza Jacomini, M. Z. Do Nascimento, R. D. Dantas, and R. P. Ramos, "Comparison of PCA and ANOVA for information selection of CC and MLO views in classification of mammograms," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7435 LNCS, no. Cc, pp. 117–126, 2012.

[19] E. Kourd Kaouther, S. Eddine Khelil, and S. Hammoum, "Study with RK4 & ANOVA the location of the tumor at the smallest time for multi-images," *Proc. - Int. Conf. Comput. Vis. Image Anal. Appl. ICCVIA 2015*, 2015.

[20] T. Aljuaid and S. Sasi, "Proper Imputation Techniques for Missing Values in Data sets," *Int. Conf. Data Sci. Eng.*, 2016.

[21] R.-J. Park, B.-S. Kwon, S.-W. Jo, and K.-B. Song, "Analysis of Short-Term Load Forecasting Using Artificial Neural Network Algorithm According to Normalization and Selection ofInput Data on Weekdays," *Asia-Pacific Power Energy Eng. Conf.*, pp. 280–283, 2018.

[22] T. K. Kim, "T test as a parametric statistic," *Korean J. Anesthesiol.*, no. Table 2, 2015.

[23] B. Gerald, "A Brief Review of Independent , Dependent and One Sample," *Int. J. Appl. Math. Theor. Phys.*, vol. 4, no. 2, pp. 50–54, 2018.

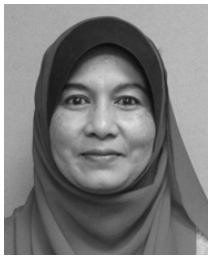
[24] R. Beaumont, "An introduction to Principal Component Analysis & Factor Analysis Using SPSS 19 and R (psych package)," no. April, 2012.

[25] R. Parveen and A. Ahmad, "Public behavior in reducing urban air pollution : an application of the theory of planned behavior in Lahore," no. mundi 2018, 2020.

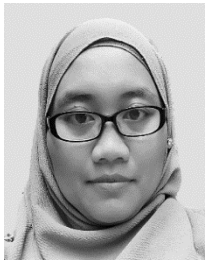
[26] K. Joshi and B. Patil, "Prediction of Surface Roughness by Machine Vision using Principal Prediction of Surface Roughness Machine Analysis Vision using Principal Components based by Regression Components based Regression Analysis," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 382–391, 2020.

Ain Dzarah Nafisah M. was born in Muar, Johor in 1995. She received her Diploma in Electrical Engineering (Electronics) in 2015 and B.Eng (Electronics) in 2019 from Universiti Teknologi MARA. She is currently a Graduate Research Assistant at Universiti Teknologi MARA.

Muhammad Asraf H. received his PhD in Electrical Engineering from Universiti Teknologi MARA, Malaysia in 2014. He is currently a senior lecturer and research interests include image processing, artificial intelligence and deep learning application.



Nooritawati Md Tahir received her PhD in Electrical, Electronics & System Engineering from Universiti Kebangsaan Malaysia. She is currently a Professor at the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia. Her research interests include pattern recognition and artificial intelligence.



Nur Dalila K.A. received the MSc degree from the Department of Electrical Engineering (Computer Control and Automation), National University of Singapore (NUS), Singapore in 2013. Currently she serves as a senior lecturer since 2003 with the Universiti Teknologi MARA Malaysia and her research interests area in control system, instrumentationspace weather and machine learning applications.

M. H. Jusoh is an Associate Professor at Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia. Currently, he is the Director of Center for Satellite Communication (UiTMSAT), a Visiting Professor at Kyushu Institute of Technology, Japan and a Professional Engineer with Board of Engineer Malaysia. He is currently focusing on the integration of research activities in Space Weather and Satellite Applications between Malaysian industries, Government agencies and Universities.