International Journal ofApplied Power Engineering (IJAPE)
Vol. 9, No. 2, August 2020, pp. 113~126
ISSN: 2252-8792, DOI: 10.11591/ijape.v9.i2.pp113-126

Power system fault identification and localization using multiple linear regression of principal component distance indices

Alok Mukherjee[1], Palash Kr. Kundu[2], Arabinda Das[3]
[1]Govt. College of Engineering and Ceramic Technology, India
[2,3]Departement of Electrical Engineering, Jadavpur University, India

Article Info

Article history:

Received Jul 25, 2019
Revised Feb 19, 2020
Accepted Mar 3, 2020

Keywords:

Multiple linear regration
PCA scores
PCDI
Ratio analysis
Ratio error analysis
Ratio indices

ABSTRACT

This paper is focused on the application of principal component analysis (PCA) to classify and localize power system faults in a three phase, radial, long transmission line using receiving end line currents taken almost at the midpoint of the line length. The PCA scores are analyzed to compute principal component distance index (PCDI) which is further analyzed using a ratio based analysis to develop ratio index matrix (R) and ratio error matrix (RE) and ratio error index (REI) which are used to develop a fault classifier, which produces a 100% correct prediction. The later part of the paper deals with the development of a fault localizer using the same PCDI corresponding to six intermediate training locations, which are analyzed with tool like multiple linear regression (MLR) in order to predict the fault location with significantly high accuracy of only 87 m for a 150 km long radial transmission line.

Corresponding Author:

Arabinda Das,
Departement of Electrical Engineering,
Jadavpur University,
188, Raja S. C. Mullick Road, Kolkata 700032, India.
Email: adas_ee_ju@yahoo.com

1. INTRODUCTION

Electrical power transmission system is one of the most spatially extended technical systems, directly exposed to the environment and fairly often subjected to atmospheric hazards leading to different types of faults. Hence, power system stability, reliability, protection as well as regulated power flow has been prime topics of research. Identification and classification and localization of faults have been under in depth research since long. Prediction of fault location, especially in long transmission systems with high and very high voltage and large power systems is one of the most challenging works in the research area for the development of a robust power system protection algorithm. Hence, prompt detection of faults and classification along with precise fault location determination has been practiced by scientists in order to ensure system safety and stability. Supervised learning algorithms like the artificial neural network (ANN) along with probabilistic neural network (PNN) have a great impact in the area of identification and localization of the fault [1-4]. ANN sometimes is combined with other topologies like fuzzy logic in fault treatment [5]. Wavelet transformation and wavelet entropy has been extensively implemented successfully in fault analysis [6-7]. Wavelet transformation has often been combined with other methods like Adaptive neuro fuzzy inference system [8], genetic algorithm (GA) [9], principal component analysis (PCA) [10] etc.

Other analytical techniques include support vector machines which have significant contribution to the design of power system protection algorithm [11-12]. Dynamic phasors is another approach used for the analysis of faults in power system [13]. Principal component analysis (PCA), on the other hand, is a

usefulstatistical and analytical technique for multivariate statistical analysis. It reduces multi a dimensional data set to a set of directions, called principal components (PC), in the decreasing order of importance, retaining variability of the data and their mutual variations, highlighting broadly the similarities and differences [14-17]. Hence, PCA has been used extensively in power system analysis, especially in fault detection, classification and distance prediction where multiple dimensional data are obtained regarding voltage, current, power, frequency etc and/or a combination of these parameters [18-22]. Hence, fault analysis becomes an important issue in power system research. Classification and detection of fault is thus absolutely essential to save vital time and effort of the working personnel. Close approximation of the fault location makes it easier to detect and remove of the fault. Hence, faults are required to be detected fast, and located accurately to restore normal power flow at the earliest.

The proposed work is intended to develop a simple PCA based power system protection algorithm suitable for the classification and localization of different types of power system faults in a three phase radial long transmission system, using pattern indices and fault signatures developed by application of PCA, leading to the development of principal component distance index (PCDI) [22]. Similarity analysis of the PCDI based fault signatures identifies the maximum proximity of the test data with any of the fault prototypes using minimum square error (MSE) criteria, thus classifying the unknown fault. Fault localization has been carried out using statistical analysis like multiple linear regression (MLR) [23] over the three phase PCDI. This helps in developing a general regression model which is further used to test unknown data for fault localization. A transmission line prototype has been modeled in EMTP-ATP simulation [24], followed by analysis of quarter cycle pre-fault and half cycle post-fault receiving end current waveforms in the MATLAB environment, using ten different fault prototypes conducted at varying fault locations along the line span and healthy condition, using PCA based proposed protection scheme

2. SYSTEM DESIGN

A single end fed 400 kV, 150 km long, single circuit, three phase, radial, overhead transposed AC transmission line has been designed in electromagnetic transient programming (EMTP) joining fifteen three phase line cable constants (LCC) blocks, each of 10 km in cascade and is shown in Figure 1. Ten different types of faults viz., SLG-A, SLG-B, SLG-C, DL-AB, DL-BC, DL-CA, DLG-AB, DLG-BC, DLG-CA, and LLL have been conducted at fifteen different locations, 10 km apart, throughout the entire length of 150 km. The resistance and inductance of the bundled line are taken as 0.0585 Ω/km (DC resistance) and 0.2 mH/km. The inter-spacing of the three lines is kept at of 17.5 cm between two adjacent horizontal lines.

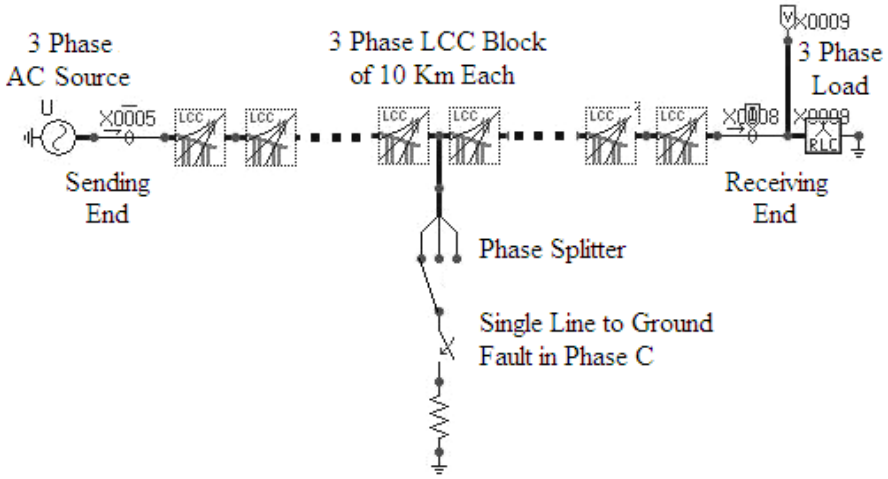

Figure 1. Simulation model of the radial, single end fed, long transmission line

3. DATA PREPARATION

The proposed algorithm is trained with only one set of training fault data of ten different types of faults conducted at almost the midpoint of the line, i.e., at 70 km from sending end of the 150 km long transmission line and healthy condition data. Quarter cycle pre-fault and half cycle post-fault receiving end

line fault current is collected at a sampling frequency of 10 kHz, i.e., 2000 samples/cycle, thus the sample vector becomes an array containing 1500 data points for each type and for each phase. Hence, the three phases training data matrix corresponding to each fault type and carried out at 70 km, takes the dimension of 1500×3, i.e., with 1500 rows and 3 columns, for one type of training fault and illustrated as:

$$X_i = [Ia_{i\,1}\ Ib_{i\,1}\ Ic_{i\,1};\ Ia_{i\,2}\ Ib_{i\,2}\ Ic_{i\,2};\ldots;\ Ia_{i\,1500}\ Ic_{i1500}]_{\ 1500\times3}$$

where Ia, Ib and Ic are receiving end line currents under fault condition for three different phase and *i* is the index defining fault prototype, which counts up to twelve. *i*=1 represents healthy or no fault condition, *i*=2 to 11 represent ten different fault prototypes and, *i*=12 stands for the test data or unknown type. Hence, the total data matrix takes the dimension [1500×(3×12)] i.e. 1500×36. Hence,

$$X = [X_1\ \ X_2\ \ X_3 \ldots X_{12}]_{\ 1500\times36}$$

Further modification has been carried out by grouping the individual phases to construct three individual phase identified matrices [Xa], [Xb], and [Xc] given by:

$$Xai = \begin{bmatrix} Ia_{1\,1} & Ia_{2\,1}\ldots & Ia_{12\,1}; \\ Ia_{1\,2} & Ia_{2\,2}\ldots & Ia_{12\,2}; \\ \ldots & \ldots & \ldots\ ; \\ Ia_{1\,1500} & Ia_{2\,1500}\ldots & Ia_{12\,1500}] \end{bmatrix}_{1500\times12}$$

Xbi and Xci are also constructed in a similar way, each of which is a 1500×12 matrix Hence the modified data matrix takes the form as:

$$Xi = [Xai\quad Xbi\quad Xci]_{\ 1500\times36}$$

[Xa], [Xb] and [Xc] for each protype *i*, are processed by the PCA algorithm separately to obtain a the PCA scores of each of phase individually, hence producing a Principal Component Distance Indices (PCDI) matrix of the order 1×3. Twelve such prototypes are analyzed in sequence to form the complete PCDI matrix of the dimension 12×3, denoted by P, each row of which correspond to the twelve fault cases and test condition and each column represents three individual phases. As mentioned before, PCA reconstructs a data set in the ascending order of importance and for the sake of ease of analysis, only two most important directions (PCs) and the corresponding score data are considered for the present purpose, hence used to construct PCDI matrix. These PCDI values are approximate estimation of the deviation of each fault current from healthy condition. The directions of variation is given by the eigenvectors obtained from the covariance matrix of the transformed data points or scores and the magnitudes of deviation from the origin (origin is assigned to the no fault condition) are given by the corresponding eigenvalues.

4. PCA ALGORITHM

4.1. Generalized PCA algorithm

Input: N x d data matrix X (each row contain a d dimensional data point)

- Comput mean : $\mu = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$
- Subtract mean from rows of X: $\tilde{X} = X - \mu$
- Compute covariance matrix: $\sum = \frac{1}{N}\tilde{X}^T\tilde{X}$
- Calculate eigenvalues and eigenvectors of $\sum$

Pick few eigenvectors (d'<d) corresponding to the largest eigenvalues and put them in the column of A in descending order of eigenvalues i.e. $A = [V_1, V_2, \ldots . V_{d"}]$, where V_1, V_2 are the 1st , 2nd PCs and so on.

- Compute the new data matrix (PC scores) in reduced dimension: $\tilde{X} = A^T X$

4.2. PCA algorithm applied for the proposed work

Step 1: Assign input data: the input data is taken as the phase identified matrices: [Xa], [Xb], and [Xc] each phase data in computed individually, say, denoted by j_k where k takes the indices a,b, and c.

Step 2: Comput mean of X_k for each of the columns individually as: $(\mu_{k)n} = [\frac{1}{N}\sum_{i=1}^{N} j(n)^{(i)}]$, where *i* indexes rows and takes the values 1 to 1500 and n indexes columns and takes the values 1 to 12.

Step 3: Subtract mean of each corresponding column from each of the rows of J_k for each column independently to form the modified joint matrix as: $J_{k\ MOD\ n} = J_{k\ n} - (\mu_k)_n$. Hence the dimension retains the same as 1500×12.

Step 4: Compute covariance matrix: $\sum = \frac{1}{N}\ (J_{k\ MOD\ n})^T (J_{k\ MOD\ n})$

Step 5: Calculate eigenvalues and eigenvectors of $\sum$

Step 6: Pick few eigenvectors (d' <d) corresponding to the largest eigenvalues and put them in the column of A in descending order of eigenvalues i.e. $A = [V_1, V_2, ...V_{d"}]$, where V_1, V_2 are the 1st, 2nd PCs and so on. For the proposed work, we have taken only the two largest eigenvector, hence, V_1 andV_2.

Step 7: Compute the new data matrix (PC scores) in reduced dimension: $J_{k\ MOD\ n\ (new)} = A^T J_{k\ MOD\ n}$. Hence, the dimension of the score matrix $J_{k\ MOD\ n\ (new)}$ should become the same as 12×1500. The proposed work uses only the two most significant directions. Hence, the working $J_{k\ MOD\ n\ (new)}$ dimension reduces to a 12×2 which acts as the PC score matrix for the proposed work.

Step 8: Forming PCDI matrix: PCA distance is formed by finding out the vector distance of each of the training and the test score (2D) from the no-fault score (2D) which is the origin, thus forming PCDI matrix for each phase and producing 12×1 PCDI vector for each phase and the total PCDI 12×3 matrix considering all the three phases, say, denoted by $S_{12\times3}$. The top eleven rows of S correspond to the eleven different training conditions and each column represents the three individual phases and the twelfth row indicates that of the test condition, given by,

$$S=[PCDI\text{-}A_i \quad PCDI\text{-}B_i \quad PCDI\text{-}C_i]_{12\times3}$$

where i=1 to 12 in the sequence as NO-FLT (healthy), SLG-A, SLG-B, SLG-C, DL-AB, DL-BC, DL-CA, DLG-AB, DLG-BC, DLG-CA, LLL, and TEST. The proposed algorithm discussed so far and the formation of PCDI follows the flowchart as given in Figure 2.

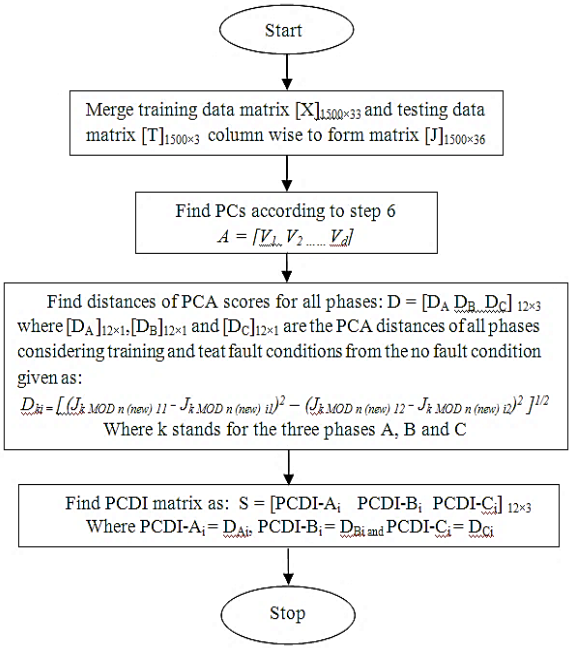

Figure 2. Flowchart illustrating the formation of PCDI matrix

Further, the total PCDI matrix S is segmented into two matrices, viz. training PCDI matrix (denoted by P) and test PCDI matrix or vector (denoted by Q), hence reducing the 12×3 matrix into two matrices as given here:

$$P_i=[\text{PCDI-A}_i \quad \text{PCDI-B}_i \quad \text{PCDI-C}_i]_{11\times 3}$$

$$Q_i=[\text{PCDI-A}_{TEST} \ \text{PCDI-B}_{TEST} \ \text{PCDI-C}_{TEST}]_{1\times 3}$$

$$S=[P \ ; \ Q \ ]_{12\times 3}$$

Further, similarity analysis has been carried out in order to compare the experimental data (Q vector) with the training fault signatures (P matrix) for each individual phases and find out the maximum similarity with any of the eleven different patterns, thus classifying the fault. It is observed that PCDIs vary following a certain pattern when computed for faults conducted at increasing geometric distance from the sending end, but the pattern for the three individual phases of PCDI remain identical. E.g., for DL-AB fault, the magnitudes of PCDI of phase A and B are very high, in comparison to phase C which remains almost zero for being the undisturbed phase. This pattern remains almost the same even with changing fault location. Besides, this rate of change in magnitude of PCDI of each phase is very much identical with increasing or decreasing fault locations. In order to establish the above inference mathematically, the PCDI of each phase is divided with the PCDI of the other phase which should remain almost the same regardless of the geometric distance of the fault as all the 3 phase PCDI vary almost in the same ratio on changing fault distances. The 3D Ratio Matrix (R) is hence formed using the 3D PCDI vector thus formed for each type of training fault and the test data, the elements of which are formed as follows [22]:

$$[R]=[(\text{PCDI-A}_i/\text{PCDI-B}_i) \quad (\text{PCDI-B}_i/\text{PCDI-C}_i) \quad (\text{PCDI-C}_i/\text{PCDI-A}_i)]_{12\times 3}$$
$$=[\text{Ratio } 1_i \text{Ratio } 2_i \text{Ratio } 3_i]_{11\times 3}$$

where *i* represent the same indexing pattern. It is to be noted here that for a no-fault condition, PCDI of all the phases are zero and assigned as origin. Hence no-fault condition is identified by comparison of the PCDI directly with a very low constant value as mentioned later and the rest are used for the ratio analysis purpose. [R] is further segmented into training and test matrices, as given by:

$$[\text{Ratio}_{TRAINING}]_{10\times 3}=[R]_{(i = 2 \text{ to } 11)\times 3} \text{ and}$$
$$[\text{Ratio}_{TEST}]_{1\times 3}=[R]_{(i =12)\times 3}$$

The [Ratio $_{TEST}$]vector will be similar to any of the ten fault prototypes defines by the ten rows of [Ratio $_{TRAINING}$].In order to model this inference mathematically, a 3D ratio error matrix (RE) is formed using the [Ratio$_{TRAINING}$]and [Ratio$_{TEST}$] as:

$$[RE]=[(\text{Ratio}_{\text{TRAINING } 1\, i} \sim \text{Ratio}_{\text{TEST } 1}) \ (\text{Ratio}_{\text{TRAINING } 2\, i} \sim \text{Ratio}_{\text{TEST } 2}) \ (\text{Ratio}_{\text{TRAINING } 3\, i} \sim \text{Ratio}_{\text{TEST } 3})]_{11\times 3}$$

Finally, a column vector of ratio error index (REI) is found comparing the ratio error values of each type, the elements of which is given as:

$$\text{Ratio error index (REI)}_i=\text{Ratio Error } 1_i+\text{Ratio Error } 2_i+\text{Ratio Error } 3_i;$$

Quite understandably, the $[REI]_i$ will be minimum when the test and the corresponding training pattern match identically and this matrix is used to classify the fault by identifying the index *i* with the minimum possible REI value.Apart from these, two other threshold values ε_1 and ε_2 are selected, one being the upper threshold and the other being the lower one, based on the test data set found. The no fault condition is detected by direct comparison PCDI summation of the test data with the lower threshold as follows:

$$\text{PCDI}_{TEST} \text{ sum}=\text{PCDI-A}_{TEST}+\text{PCDI-B}_{TEST}+\text{PCDI-C}_{TEST};$$

If PCDI_{TEST} sum is less than the lower threshold ε_1, it is identified as no fault due to the absence of any major disturbance in any phase, thus detecting no-fault. On the other way, a fault is detected for the same PCDI_{TEST} sum being higher than ε_1. DL faults are similarly found by comparing the ratio error index with that of the upper threshold ε_2 followed by direct analysis of [PCDI] and [R]. The entire analysis is well understood from the case study discussed in the next section. It is further observed that for DL faults,

the directly unaffected phase remains almost undisturbed, whereas in case of DLG faults, some disturbance occur even in the directly unaffected line due to the involvement of ground and flow of zero sequence current through the ground and the grounded neutral of the transmission system, thus making a differentiation between the two types very clear. This inference is also observed from the PCDI matrix discussed in the case study model. Figure 3 elaborates the proposed algorithm in detail.

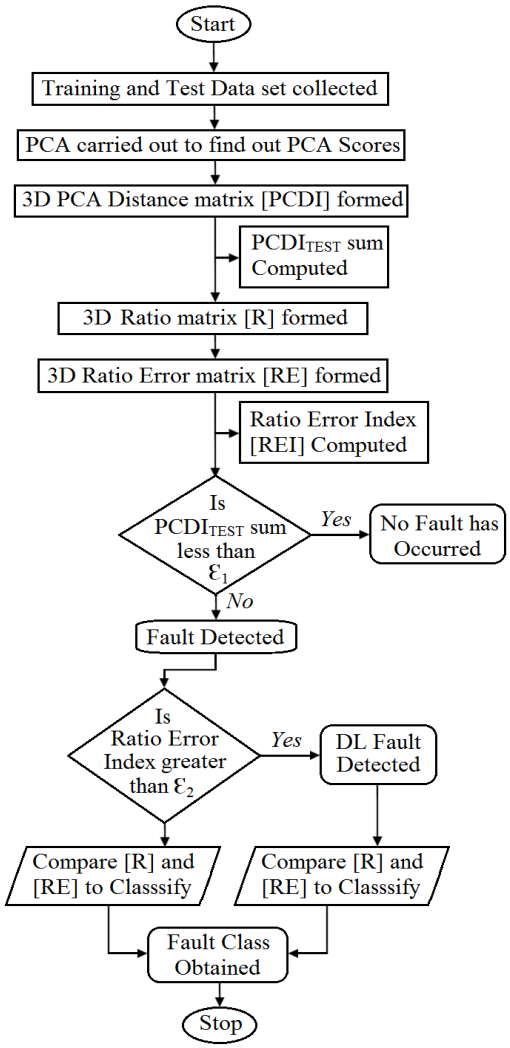

Figure 3. Flowchart of the proposed PCA based fault classifier algorithm

5. RESULTS AND ANALYSIS

A sample data set for any arbitrary fault is taken here for the purpose of case study and the same is processed through the proposed PCA algorithm to produce PCDI matrix as shown in the initial columns of Table1 which is a combined view of the [PCDI], [R], and [RE]. The [PCDI] is further represented graphically in the form of a three dimensional plot in Figure 4.Close observation of Figure 4 reveals that the PCI vector of the unknown type i.e. legend 9 is closest to the SLG-BG fault i.e., legend 3 compared to any other type with minimum Euclidian distance, which is further ascertained by forming [R] as shown in the middle columns of the same Table 1. [R] is again represented graphically in Figure 5.Close observation of [PCDI] and [R] reveal a certain distinguishing feature for each particular type of fault, i.e., the test fault PCDI values closely resemble that of SLG-B and this similarity is further boosted from Ratio values of the same, marked in bold letters. The same is observed from Figure 5 as well where the Euclidian distance between legend 3 and 9 is much less as compared to the same in Figure 4, thus ascertaining the test pattern to be SLG-B with MSE criteria. Thus, formation of R greatly emphasizes on the similarity between the test data and any one of the eleven sets of fault prototypes and this is also tested with varying fault location.

Table 1. PCDI, ratio matrix and ratio error matrix formed from the dataset

Fault type	[PCDI]			[R]			[RE]			Ratio error index (REI)
	PCDIA	PCDI-B	PCDI-C	Ratio 1	Ratio 2	Ratio 3	Ratio error 1	Ratio error 2	Ratio error 3	
HEALTHY	0	0	0	NaN	NaN	NaN	NA	NA	NA	NA
SLG-A	16.10	3.51	3.50	4.59	1.00	0.22	4.325	2.781	0.782	7.888
SLG-B	3.87	14.87	3.87	0.26	3.84	1.00	0.004	0.058	0.001	0.063
SLG-C	4.31	4.31	14.66	1.00	0.29	3.40	0.737	3.489	2.399	6.625
DL-AB	13.99	13.10	3.6E-15	0.99	3.9E+15	2.6E-16	0.735	3.877e+15	0.999	3.877E+15
DL-BC	3.6E-15	11.44	11.43	3.1E-16	1.00	3.2E+15	0.265	2.782	3.165e+15	3.165E+15
DL-CA	13.95	2E-15	13.95	6.9E+15	1.5E-16	1.00	6.839e+15	3.783	0.001	6.84E+15
DLG-AB	15.27	16.58	3.55	0.92	4.67	0.23	0.657	0.882	0.767	2.306
DLG-BC	2.86	12.46	15.14	0.23	0.82	5.29	0.035	2.960	4.288	7.283
DLG-CA	17.97	3.18	13.55	5.66	0.23	0.75	5.089	3.548	0.202	8.839
LLL	17.00	14.28	14.22	1.19	1.00	0.84	0.926	2.778	0.163	3.867
TEST DATA	3.51	13.28	3.51	0.26	3.78	1.00	NA	NA	NA	NA

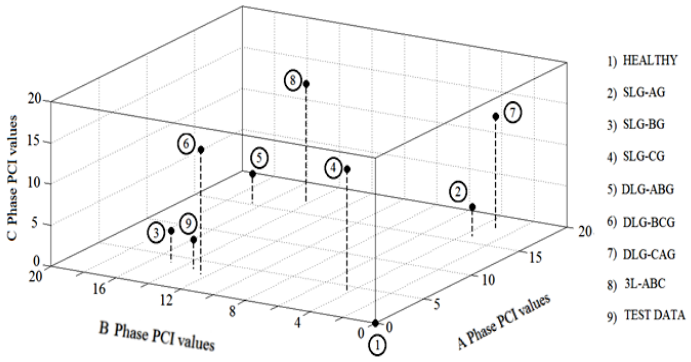

Figure 4. 3D plot of three phase PCDI values for training (ten different types of faults and healthy condition) and test data

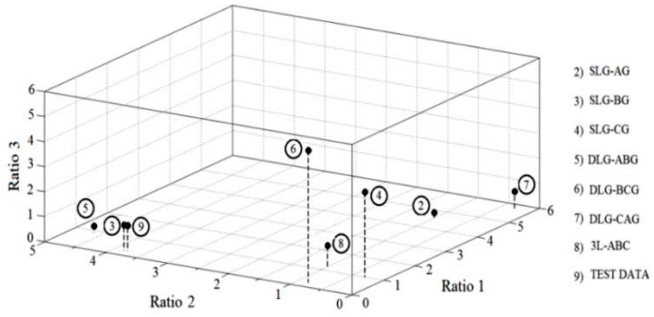

Figure 5. 3D plot of the three phase Ratio Indices for seven different types of faults (DL fault excluded) and test data

It is further observed that since the unaffected phase is least disturbed in case of a DL fault, accordingly indicated in the corresponding PCDI values, hence, the Ratio index of any one of the ratios is abruptly high for the DL faults. This is readily observed from Table 1 that, e.g., PCDI-C for DL-AB fault is very much low since phase C is the unaffected phase here, which when is used to form [R], Ratio 2 becomes abruptly high and this is reflected in [RE] as well as in the Ratio Error Index (REI) so formed and is shown in the final column of Table I. It shows that REI is hugely larger for DL faults in comparison to all the other prototypes. This key feature is used effectively to identify the DL faults from the rest and the upper threshold value ε_2 is set comparing all other fault types. For the given set of PCDI, it is well observed that ratio error index for faults other than DL faults is well below 100 and that for DL faults is way above it. Hence, ε_2 for this case can safely be set at 100. Hence, for the same reasons listed above, DL faults are not included to

form fault signatures in Figure 5. It is further observed that even on varying the geometric fault distance from the sending end, the PCI vary following a particular pattern as described by the PCDI of Table2 where a typical fault data for SLG-BG fault, for example, is taken at different distances 10 km apart all throughout the entire span of 150 km long line. More importantly, it is observed that their mutual ratio remains very much similar, even with varying geometric distance (km) over the entire span of the transmission line as described by the three RI vector values of the same table. The above fact is also represented in Figure 6 which is constructed using the three phase [PCDI] and [R] values of Table 1 where, as described earlier, SLG-B fault is taken for example for different fault locations.

Table 2. Ratio matrix formed by the PC distances with variation in geometric fault distances

Fault location(km)	PCI-A	PCI-B	PCI-C	Ratio 1	Ratio 2	Ratio 3
10	1.4382	5.0673	1.438	0.2838	3.5239	0.9999
20	2.3356	8.4395	2.3345	0.2767	3.6151	0.9995
30	2.8854	10.64	2.8834	0.2712	3.6902	0.9993
40	3.2509	12.159	3.2484	0.2674	3.7432	0.9992
50	3.5135	13.282	3.511	0.2645	3.7828	0.9993
60	3.7141	14.158	3.7119	0.2623	3.8141	0.9994
70	3.8766	14.872	3.8754	0.2607	3.8375	0.9997
80	4.0063	15.442	4.0055	0.2594	3.8552	0.9998
90	4.1147	15.934	4.1137	0.2582	3.8734	0.9998
100	4.2109	16.378	4.2098	0.2571	3.8905	0.9998
110	4.2983	16.783	4.2976	0.2561	3.9051	0.9998
120	4.3756	17.154	4.3748	0.2551	3.9211	0.9998
130	4.45	17.502	4.4498	0.2543	3.9333	1
140	4.526	17.832	4.5269	0.2538	3.9391	1.0002

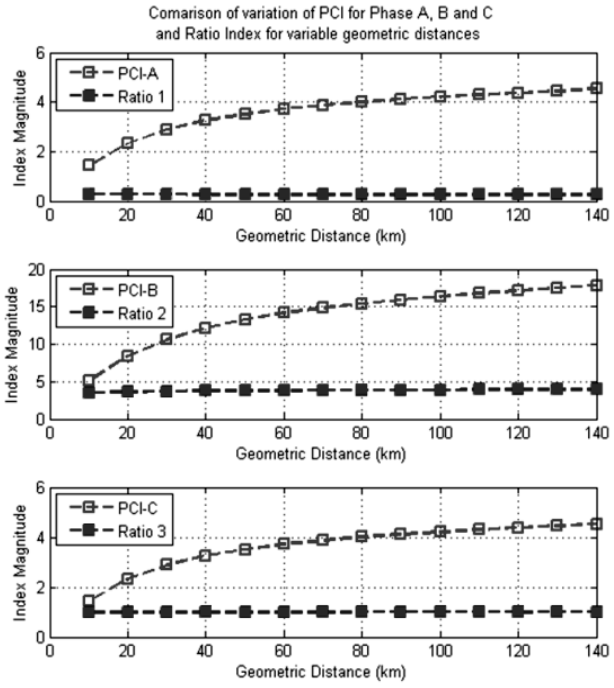

Figure 6. Variation of three phase PCDI and ratio indices with different geometric fault distances

It is well observed form Figure 6 that the variations of [R] values are remarkably lesser than that of [PCDI values for different geometric distances over the line span for all the three phases, rightly justifying the role of [R] and its usefulness in finding out the unknown fault pattern, therefore, is taken as the Fundamental Governing Factor in the determination of the unknown fault pattern. This conclusion is also justified from Table3 which shows the results of several unknown fault patterns for faults conducted at various distances along the 150 km line span and classifier accuracy is 100% using the proposed PCA-Ratio based classifier algorithm.

Table 3. Fault classifier results with only one set of training data

Fault type	PURE	AG	BG	CG	AB	BC	CA	ABG	BCG	CAG	ABC
PURE	13	0	0	0	0	0	0	0	0	0	0
AG	0	13	0	0	0	0	0	0	0	0	0
BG	0	0	13	0	0	0	0	0	0	0	0
CG	0	0	0	13	0	0	0	0	0	0	0
AB	0	0	0	0	13	0	0	0	0	0	0
BC	0	0	0	0	0	13	0	0	0	0	0
CA	0	0	0	0	0	0	13	0	0	0	0
ABG	0	0	0	0	0	0	0	13	0	0	0
BCG	0	0	0	0	0	0	0	0	13	0	0
CAG	0	0	0	0	0	0	0	0	0	13	0
ABC	0	0	0	0	0	0	0	0	0	0	13
Overall accuracy: 100 %											

6. FAULT DISTANCE ESTIMATION

The later and another vital section of the proposed research is prediction of the fault location. The proposed fault distance predictor algorithm is designed using multiple linear regression (MLR) analysis. MLR takes into account the trends and curvatures of more than one data set and effectively compute one primary direction of variation using the multiple data set. The proposed work utilizes this important feature of MLR and uses the three phase features in terms of PCDI to form one key curvature, incorporating the features of all the PCDI. For this purpose, six intermediate non-equidistant locations at 10, 20, 50, 90, 130, and 140 km distance from the sending end of the 150 km long line have been chosen as the six training points for the proposed fault localizer algorithm. Ten different types of faults have been conducted at these six training locations and receiving end current waveforms have been recorded as the training data, each of which is fed to undergo the proposed fault classifier algorithm discussed in the previous section and the three phase PCDI are found for each of the six training points. This 3D training data set for each fault prototype is saved as a look up table and is scaled to unity for generalization and providing uniformity. Hence, the training data matrix, for each fault pattern takes the dimension of 6×3, called as training distance PCDI matrix afterwards and is given by D*i* as:

$$D_i=[\text{PCDI-}A_{ij}\ \ \text{PCDI-}B_{ij}\ \ \text{PCDI-}C_{ij}]_{6\times3}$$

where, i=1 to 10 define each of the ten training fault prototypes mentioned before and j=1 to 6 defines the six training geometric distances at 10, 20, 50, 90, 130, and 140 km respectively. Hence for the ten types of faults, there are ten such training distance PCDI matrices, together which forms the total training distance PCDI matrix given by $D_{TRAINIING}$ as:

$$D_{TRAINIING}=[D_1\ \ D_2\ \ D_3\ \ \ldots.\ \ D_{10}]_{6\times30}$$

Post classification of the fault, the test PCDI matrix Q as found in the earlier section is saved. Next the D_i matrix corresponding to the particular identified type with index *i*is taken up from $D_{TRAINIING}$, followed by interpolation of the test Q vector from the corresponding Di using the Multiple Linear Regression (MLR) method in order to predict the geometric distance of the corresponding fault.

7. CASE STUDY AND ANALYSIS

A case study is shown here with SLG-A fault. The variation of receiving end line currents with varying geometric fault distance for SLG-A fault is shown in Figure 7. The same data is processed through the PCA algorithm to produce [PCDI] and consequence calculations. Table4 describes the absolute PCDI values and the corresponding scaled values for SLG-A fault at six training locations. The D SLG-A matrix is formed using the PCDI values as recorded in Table 4 using values from column 2, 3, and 4.

Similarly, D scaled SLG-A matrix is formed using values from column 5, 6, and 7 which on plotting against the respective fault geometric locations, reveal a curvilinear nature as shown in Figure 8. It is observed that each of the fault types show difference in curvature for three individual phases. Hence, the proposed scheme has been designed with multiple linear regression (MLR) for each prototype individually, which takes into account all the three phase PCDIs to produce a fairly accurate estimate of the fault location. The mathematical analysis of the MLR scheme adopted here is explained first following its application in designing the fault location prediction algorithm [23].

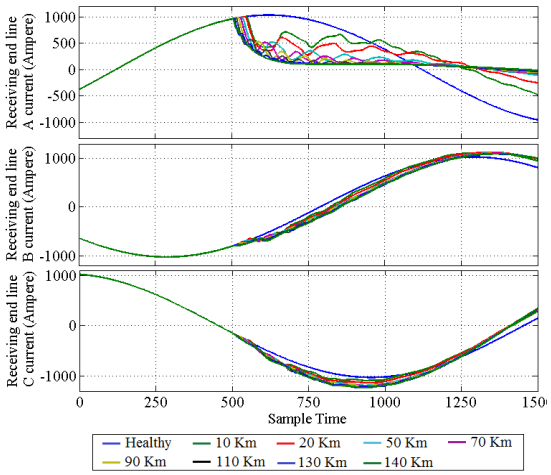

Figure7. Receiving end line current vs. sampled time plot for different geometric fault locations for SLG-A fault

Table 4. PCDI for A phase to ground fault at six different locations

Fault location (km)	PCDI			PCDI (scaled)		
	Phase A	Phase B	Phase C	Phase A	Phase B	Phase C
10	7.449	1.6804	1.6816	0	0	0
20	11.2046	2.4365	2.4365	0.3421	0.3034	0.3039
50	14.9931	3.2307	3.2307	0.6873	0.6221	0.6223
90	16.7925	3.7172	3.7172	0.8512	0.8174	0.8172
130	18.1095	4.0781	4.0781	0.9712	0.9622	0.9622
140	18.4261	4.1723	4.1723	1	1	1

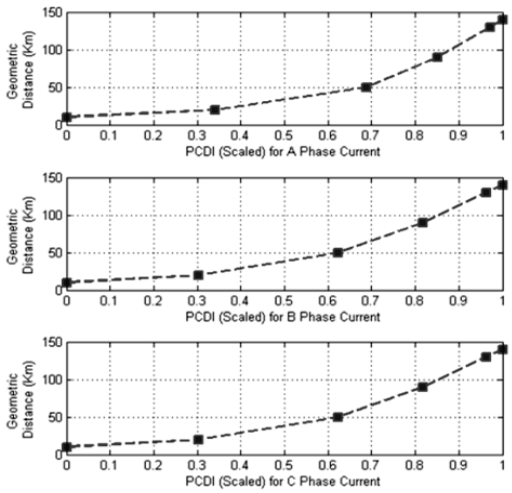

Figure 8. Geometric fault distance vs. PCDI (scaled) plot for three phase receiving end line currents for SLG-A fault at six training locations

8. APPLICATION OF MULTIPLE LINEAR REGRESSION (MLR)

Principal component analysis (PCA) as explained so far, itself is an important and effective tool in order to reduce a large number of multivariate data to a few primary directions of major variation. The three different phases of PCDI have difference in curvature which is well observed from Figure 10. This is further extended for all ten different fault patterns. The three phase PCDI for each pattern, is processed by the proposed MLR based scheme to achieve a single computed direction of variation, taking into account all the three curvatures from the three phases which is finally taken as the training data for the proposed fault distance predictor algorithm. Regression analysis is an important statistical tool to determine the relationship, called the regression function, between a dependent variable 'y', and a single or several independent variables 'xi'. Regression function also involves a set of unknown parameters 'bi', called the regression coefficients. A simple linear regression model is described as:

$$Y=b_0+b_1x_1 \quad (1)$$

Linear regression models with multiple independent variables are referred to as multiple linear models, a model of such representation is given as:

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+\ldots+b_nx_n \quad (2)$$

where n is the total number of independent variables.

The proposed algorithm uses scaled PCDI for the three phases as the input variable x_i. Figure 10 reveals that the three phase scaled PCDI shows a curvilinear nature, rather than a straight line trend. Hence, to take into account this curvature of PCDI, the proposed algorithm is extended to multiple orders of these primary inputs xi depending on the Minimum Square Error (MSE) criteria. It is also to be mentioned here that the no of independent variables have clearly been taken depending upon the MSE criteria, and more so, the number of such variables vary from one fault type to another and also the order and type and interdependence, if any, among the variables. Thus, the proposed scheme is constructed as:

The primary input variable is defined as,

$$X_1=D_{scaled\ i}=[PCDI\text{-}A_{scaledi}PCDI\text{-}B_{scaledi}PCDI\text{-}C_{scaledi}]_{6\times3} \quad (3)$$

and the elements are ordered as,

$$X_1=[x_{1\,11}\ \ x_{1\,12}\ \ x_{1\,13};\ x_{1\,21}\ \ x_{1\,22}\ \ x_{1\,23;}\ \ldots\ldots\ \ldots;\ \ x_{1\,61}\ \ x_{1\,62}\ \ x_{1\,63}]_{6\times3} \quad (3a)$$

where D scaled i the training matrix of any particular type of fault containing three phase PCDI corresponding to six different training locations, hence taken as the primary input variables and i takes the index of the fault class identified by ratio analysis. Further, MLR has been adopted here with multiple inputs of several orders of the primary input defined by X1. Hence, the complete regression equation for one training pattern takes the form:

$$Y=X_1B_1+X_2B_2+\ldots\ldots+X_kB_k \quad (4)$$

where Y is the output vector, and in the proposed case, Y is formed by the six training geometric fault locations, defined as:

$$Y=[y_1\ \ y_2\ \ y_3\,y_4\ \ y_5\ \ y_6\,]^T_{6\times1} \quad (5)$$

where y_1, $y_{2,}$ …, y_6 etc. takes the training fault locations taken as 10, 20, 50, 90, 130, and 140 respectively which is fixed for the proposed scheme and X_k is the k-th order polynomial expression of the primary input X_1, i.e.

$$X_k=[X_1]^k_{6\times3} \quad (6)$$

The idea is to train the proposed MLR based fault localization algorithm with the best fit arrangement, taking together all the three phases, although the maximum variation occurs in case of the directly affected line. The maximum order of X_1i.e., index k has been assumed 12, i.e., twice the number to training locations, only to reduce computational complexity and the intermediate orders, i.e., the values of index k is set according to the MSE criteria. Thus, the complete input matrix is described by,

$$[X]=[X_1\ \ X_2\ \ldots\ldots X_k]_{6\times3k} \quad (6a)$$

and the coefficient matrix Bi for each variable Xi obtained on regression analysis as 1×3 vectors described by,

$$B_i=[b_{i1}b_{i2}b_{i3}]^T_{3\times1} \quad (7)$$

and the complete coefficient matrix for each training pattern is defined by B as:

$$\begin{aligned}B&=[B_1;\ B_2;\ \ldots\ldots B_k]_{3k\times1}\\&=[b_{11}b_{12}\ \ b_{13};\ b_{21}\ b_{22}\ b_{23};\ \ldots\ b_{k1}b_{k2}\ b_{k3}]^T_{3k\times1}\end{aligned} \quad (8)$$

which is a 3k×1 vector. In general, the coefficients of B are described as:

$$B=[b_1\ \ b_2\ \ b_3\ \ \ldots\ \ b_{3k}]^T_{3k\times1} \quad (9)$$

The maximum number of input variables and each particular order are different for each ten training patterns and have been chosen depending in MSE criterion, producing different coefficient matrix for each training patterns. In a word, the non-linear nature of the PCDI has been scaled using MLR analysis. Equations (4) to (9) describe the MLR analysis for each fault pattern only, the complete equation of which can be given in matrix for as obtained from (4) as follows:

$$Y=X\,B \tag{10}$$

In order to estimate the regression coefficients, a least square approach has been adopted; thus, the algorithm minimizes each of the errors described as:

$$\varepsilon_i=\sum(y_i-b_1x_{i1}-b_2x_{i2}-b_3x_{i3}-\ldots\ldots-b_{3n}x_{ik}) \tag{11}$$

which is found with all possible training values and this is minimized by setting

$$\underline{[B]}=([X]^T[X])^{-1}([X]^T[Y]) \tag{12}$$

where$([X]^T[X])$as well as$([X]^T[X])^{-1}$ are k×k dimensional symmetric matrix and $([X]^T[Y])$ is a k dimensional vector. Hence the fitted values are,

$$\underline{[Y]}=[X]\,\underline{[B]} \tag{13}$$

and the residuals are given by,

$$\underline{[R]}=[Y]-\underline{[Y]} \tag{14}$$

These residuals have been minimized following MSE criteria and the corresponding orders of polynomials for each type of training set has been achieved and stored in a look up table. Thus, each training pattern has different B vector having difference both in magnitudes, as well as in dimensionality. In order to test any unknown fault current, the proposed fault classifier algorithm based on the ratio analysis is applied first to identify the exact type of fault followed by fault distance prediction analysis using MLR as described.The three phase PC Indices corresponding to the experimental waveform have been analyzed using the same location prediction algorithm using the regression coefficient matrix (B) corresponding to the exact predicted fault type, as determined by the classifier, and the predicted location has been derived. The proposed algorithm is described in Figure 9 in the form of flowchart.

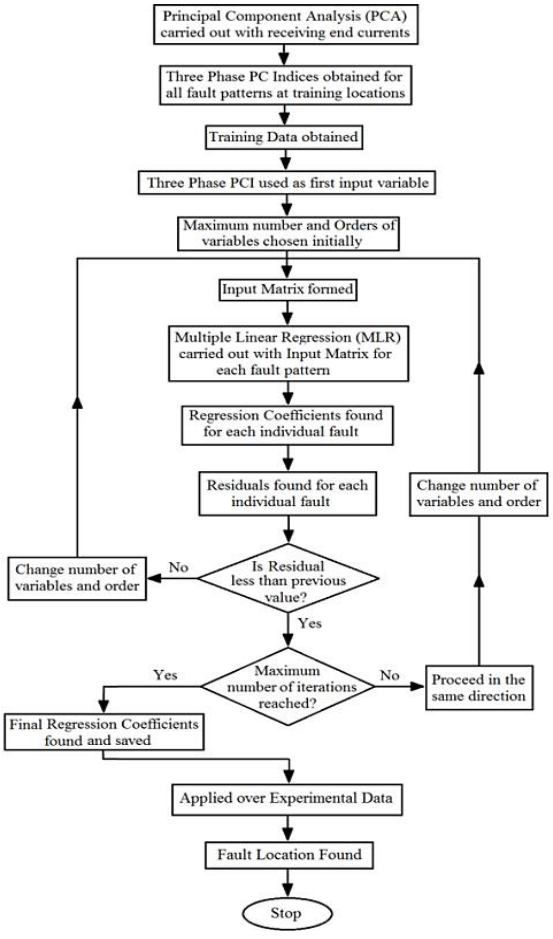

Figure 9. Flowchart illustrating fault location predictor algorithm

9. RESULTS OF FAULT LOCATIONS PREDICTOR

Table5 shows a summary of results by the proposed fault location predictor algorithm for ten different types of faults occurred at different locations, which shows that the proposed scheme produces an average deviation of 0.0871 km for the 150 km long transmission line which is well beyond satisfactory margin.

Table 5. Results showing the fault distance predictor algorithm performance with varying fault location

Type of fault	Average deviation with all distances (km)	Average deviation (km)
SLG-A	0.1409	0.0871
SLG-B	0.1045	
SLG-C	0.0113	
DL-AB	0.1301	
DL-BC	0.0391	
DL-CA	0.1786	
DLG-AB	0.0287	
DLG-BC	0.0447	
DLG-CA	0.12	
LLL	0.0736	

10. CONCLUSION

A simple and effective power system protection scheme for classification and distance prediction of long transmission line has been proposed here for a single end fed 400 kV, 50 Hz, 150 km long radial transmission line. Principal component analysis and multiple linear regression analysis has been adopted here to realize, design and implement the proposed protection scheme. Quarter cycle pre-fault and half cycle post-fault receiving end three phase fault current waveforms have been fed as the only input to the algorithm. PCA scores thus computed analyzing the input data have been used to construct principal component distance indices (PCDI) which are used to develop a ratio based algorithm to identify and classify faults. Results show that the classifier shows 100% accuracy using only one set of training data taken almost at the midpoint of the line.

Thus the low training data is one of the key features of the proposed fault classifier. The scheme used PCA based analysis only instead of ANN aor Wavelet transform based approaches. ANN requires large training data and hence the training time is also very high. Wavelet analysis, on the other hand is computationally heavily burdened. Most of the other methods too have further complex analysis, which require higher time of computation. Simplicity of the scheme compared to some other existing methods and less computation time are other key features of the scheme. The proposed protection scheme is further extended to develop fault localizer algorithm. The average deviation of predicted fault location is only about 87.1 m. Hence, the proposed algorithm has high accuracy in determining power system fault locations as well. Accurate fault localization helps the personnel to identify the fault point fast and saves valuable time and effort to restore normal operation at the earliest. Thus the proposed protection scheme has all the qualities for the development of reliable transient-based power system protection unit.

REFERENCES

[1] M. Kezunovic, I. Rikalo, D. J. Sobajic, C. W. Fromen and D. R. Sevcik, "Automated fault analysis using neural network," *9th Annual Conference for Fault and Disturbance Analysis, Texas A&M University,*1994.

[2] M. Kezunovic and I. Rikalo, "Detect and classify faults using neural nets," *IEEE Computer Applications in Power*, vol. 9, no. 4, pp. 42-47, 1996.

[3] A. Jain, A. S. Thoke and R. N. Patel, "Fault classification of double circuit transmission line using artificial neural network," *International Journal of Electrical Systems Science and Engineering,*vol. 1, no. 4, pp. 750-755, 2008.

[4] M. Sanaye-Pasand and H. Khorashadi-Zadeh, "Transmission line fault detection & phase selection using ANN," *International Conference on Power Systems Transients*, pp. 1-6, 2003.

[5] S. Vasilic and M. Kezunovic, "Fuzzy ART neural network algorithm for classifying the power system faults," *IEEE Transactions on Power Delivery*, vol. 20, no. 2, pp. 1306-1314, 2005.

[6] S. C. Shekar, G. Kumar and S. V. N. L. Lalitha, "A transient current based micro-grid connected power system protection scheme using wavelet approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 1, pp. 14, 2019.

[7] A. R. Adly, R. A. El Sehiemy, M. A. Elsadd and A. Y. Abdelaziz, "A novel wavelet packet transform based fault identification procedures in HV transmission line based on current signals," *International Journal of Applied Power Engineering,*vol. 8, no. 1, pp. 11-21, 2019.

[8] K. S. Swarup, N. Kamaraj and R. Rajeswari, "Fault diagnosis of parallel transmission lines using wavelet based ANFIS," *International Journal of Electrical and Power Engineering,*vol. 1, no. 4,pp. 410-415, 2007.

[9] J.Upendar, C. P. Gupta and G. K.Singh, "Discrete wavelet transform and genetic algorithm based fault classification of transmission systems," *National Power Systems Conference*, pp. 323-328, 2008.
[10] M. A. Beg, M. K. Khedkar, S.R. Paraskar and G.M.Dhole, "Classification of fault originated transients in high voltage network using DWT-PCA approach," *International Journal of EngineeringScience and Technology*, vol. 3, no. 11, pp. 1-14, 2011.
[11] V. Malathi and N. S. Marimuthu, "Multi-class support vector machine approach for fault classification in power transmission line," 2008 *IEEE International Conference on Sustainable Energy Technologies*, Singapore, pp. 67-71, 2008.
[12] B. Ravikumar, D. Thukaram and H. P. Khincha, "Application of support vector machines for fault diagnosis in power transmission system," *IET Generation, Transmission & Distribution*, vol. 2, no. 1, pp. 119-130, 2008.
[13] A. M. Stankovic and T. Aydin, "Analysis of asymmetrical faults in power systems using dynamic phasors," *IEEE Transactions on Power Systems*, vol. 15, no. 3, pp. 1062-1068, 2000.
[14] I. T. Jolliffe, "Principal Component Analysis: Principal components in regression analysis," *Springer Series in Statistics*,*Springer*, New York, pp. 129-155, 1986.
[15] S. M. Holland, "Principal components analysis (PCA)," Department of Geology, University of Georgia, pp. 30602-2501, 2008.
[16] L. I. Smith, "A tutorial on principal components analysis," Department of Computer Science, University of Otago, pp. 1-26, 2002.
[17] L. H. Chiang, E. L. Russell and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 50, no. 2, pp. 243-252, 2000.
[18] Q. H. Alsafasfeh, I. Abdel-Qaderand and A. M. Harb, "Fault classification and localization in power systems using fault signatures and principal components analysis," *Energy and Power Engineering*, vol. 4, no. 6, pp. 506-522, 2012.
[19] J. Mina and C. Verde, "Fault detection for large scale systems using dynamic principal components analysis with adaptation," *International Journal of Computers Communications and Control*, vol. 2, no. 2, pp. 185-194, 2007.
[20] Y. Ma and J. Zhang, "Fault diagnosis based on PCA and D-S evidence theory," *2009 Asia-Pacific Power and Energy Engineering Conference*, Wuhan, pp. 1-5, 2009.
[21] C. Zhang, G. He and S. Liang, "PCA-based analog fault detection by combining features of time domain and spectrum," *2009 International Workshop on Intelligent Systems and Applications*, Wuhan, pp. 1-4, 2009.
[22] A. Mukherjee, P. Kundu and A. Das, "Identification and classification of power system faults using ratio analysis of principal component distances," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 11, pp. 7603-7612, 2014.
[23] D. C. Montgomery, E. A. Peck and G. G. Vining, "Introduction to linear regression analysis," John Wiley & Sons, 2012.
[24] N. Watson and J. Arrillaga, "Power systems electromagnetic transients simulation," *IET Power and Energy Series*, vol. 39, 2003.