Health Monitoring of Gas Turbine Engine using Principal Component Analysis Approach

Jerry E. Eyoh, Imoh Jerry Eyoh, Uduak A. Umoh, Edward N. Udoh

Department of Turbo Machinery (Reliability and Maintenance), Exxon Mobile, QIT, Eket, AKS, Nigeria

Department of Computer Science, University of Uyo, Uyo, AKS, Nigeria

Corresponding Author: Uduak A. Umoh

Abstract
Monitoring health condition of Gas Turbine Engines (GTE), using data-based method requires a large amount of systems’ parameter variables (data). These data are nonliner and of high dimension. To analyse these data, dimensionality reduction techniques are required to transform the high-dimensional data to a lower dimensional space. This study adopts Principal Component Analysis (PCA) technique to analyse the GTE data to extract features from the data which are then compared with feature from the normal operating condition of the GTE for change detection. GTE generates a huge amount from the numerous measured variables. We consider twelve measured variables and seven operating conditions from three-hundred gas turbines. Nearest-neighbour classification of the training data is explored for GTE fault diagnosis. We achieve visualization of the low-dimensional data in two-dimension using scatter plot. M-fold cross-validation is employed to test the performance of our model. The model is implemented in Matlab and C++ programming tool. This model serves as predictive maintenance strategy with efficiency and cost effective. It also helps to minimize the downtime of gas turbine engines, improves safety plant operations. Thus enhances system reliability and availability.

Keywords: health monitoring, principal component analysis, fault diagnosis, nearest-neighbour classification, m-fold cross-validation, efficiency.

INTRODUCTION
The advances in technology in the last few decades have led to the development and manufacturing of complex and more reliable systems and equipment that are relevant to our lives. Undeniably, our reliance on these devices, services, systems is continually increasing and we rarely think out about failure and consequences of failure of these systems or equipment. In most cases, failure of these systems directly or indirectly affects millions of people emotionally, psychologically, physically or may even result in fatalities. Regrettably, we are only being reminded of the fallibility of these sophisticated, highly reliable systems on which we depend, when such a failure occurs.

Gas turbine engines have proven to be very efficient and are widely used in many industrial and engineering systems. They are used in systems such as Aircrafts, Electrical power generation systems, Trains, Marine vessels, as drivers to industrial equipment such as high capacity compressors and pumps. In most cases, areas of application of gas turbine engines are safety critical which require very high reliability and availability of these systems. To maintain high system reliability and availability, critical system parameter variables such as engine vibration, bearing temperature, lube oil pressure, etc, must be continuously monitored for prompt detection of deviation from normal operation values. To design a system for high reliability means, increasing the cost of the system and its complexity [Ghoshal et al, 1999]. More so, monitoring, control and protection subsystems of the Gas Turbine Engines further add more cost and complexity to the overall system. The application of a classical maintenance approaches has been proven over the years, to be unsuitable for Engineering systems such as Gas turbine engines [Kadirkamanathan, 2008]. The health state of a GTE is determined by its functional state or characteristics of the parameter variables. Depending on the characteristics of these parameter variables, the GTE health state can be in a particular state [Kadirkamanathan, 2008]. In PCA, data can be transformed to a new set of coordinates or variables that are a linear combination of the original variables. [Martenicz, 2004]

The nearest-neighbours are among the simplest methods of statistical pattern recognition that achieves consistently high performance. The nearest-neighbour algorithm requires a set of labelled (classified) data and attempts to classify a set of unlabelled (unclassified) data from the knowledge of the labelled data set, with Euclidean distance measurement between labelled and unlabelled data.
Cross validation is a model evaluation technique for estimating the accuracy of a classifier and normally performed either using a number of random test/train partition of the data, or using M-fold cross-validation. Basic fault models are due [Isermann, 2006] [Kadirkamanathan, 2008] [Chiang & Braatz, 2001] [Yang, et al, 2005] [Greitzer, et al, 1999].

This study focuses on maintaining high reliability and availability of GTE, by reducing or eliminating erratic failure of the system, extending the phase 2 – the operating period to TN, thus enhancing the reliability of the turbine engine during wear out phase, through predictive maintenance strategies. To achieve our objective, we continually monitored the system operating condition to detect any changes in the normal operating condition of the Gas turbine engine. Gas turbine engines are complex systems that generate significant amount of data through a number of measured variables, such as temperature, pressure, flow, vibrations, velocities, etc. We considered twelve measured variables and seven operating conditions from three hundred gas turbine engines. Principal Component analysis (PCA) as a data based model approach is employed to analyse these data. Nearest-neighbour classification of the training data is explored for GTE fault diagnosis. We achieve visualization of the low-dimensional data in two-dimension using scatter plot. M-fold cross-validation is employed to test the performance of our model. The model is implemented in Matlab. This model serves as predictive maintenance strategy with efficiency and cost effective.

This paper outlines the related literature of the system. Later sections discuss the research methodology, the model experiment, the results and discussion of findings, and finally, some recommendations and our conclusion.

RELATED LITERATURE
System Reliability Concept
In system reliability assessment, the characteristic of the gas turbine engine and its related subsystems and components critical to its normal operation, must be estimated in design, controlled during manufacturing, measured during testing and most importantly, sustained in the field during operation [Nichlass, 2005; Martinez, 2004]. Ideally, systems’ reliability decreases with time and it is complementary with unreliability, that is:

\[ R(t) + F(t) = 1. \]

Where \( R(t) \) is the reliability of the gas turbine engine \( F(t) \) is the unreliability of gas turbine engine. From the analysis of system’s failure rate with time, the life cycle of a given system is divided into three phases as shown in figure 1.1.

Phase 1: The infant mortality period. This phase is characterised by high failure rate, which is caused by manufacturing flaws, wrong installation procedures, damage received during shipping and handling, etc.

Phase 2: The operating period. This phase is characterised by small failure rate, the failure rate during this phase is ideally constant. Adequate system monitoring and right choice of maintenance strategy can help extend the operating period from T2 to TN. This project is aimed at developing a data-based model that would monitor the health condition of the Gas Turbine engine (GTE), to facilitate predictive maintenance strategy, thereby improving and maintaining the systems’ reliability.

Phase 3: This is the wear out period due to age and it is characterised by rapid increase in failure rate due to changes in material properties of the system’s components. Thus failure here is as a result of gradual degradation of some properties of the system’s components that are essential to normal operation of the system [Nichlass, 2005].

PRINCIPAL COMPONENT ANALYSIS
According to [Russell et al, 2001] [Chiang & Braatz, 2001] [Nabney, 2002; Isermann, 2006], PCA is a linear dimensionality technique that is efficient in capturing the variability of the data. It determines a set of orthogonal vectors called loading vectors, ordered by the amount of variance explained in the loading vector directions.

The general task of PCA analysis consists of transforming a data matrix with m variables \( \chi(k) \), where,

\[ \chi = [x_1, x_2, \ldots, x_m] \]

With N measurements, \( k = 1, 2, \ldots, N \), into a new data matrix

\[ T = [t_1, t_2, \ldots, t_r] \]

With also N measurements but smaller dimension, \( r < m \).
This can be achieved through a transformation matrix \( P \),
\[
T[Nr] = X[Nx] P[mx] \tag{4}
\]
\[P = [P1, P2, ..., P_r] \]
As this transformation is a rotational matrix or orthonormal, the following holds;
\[PTP = I \tag{5}\]
So that, \( X = TP \)
In multivariable statistics, \( T \) is called the score matrix and \( P \), the loading matrix.
Equation 3.6 can also be written as;
\[X = t_1P_1 + t_2P_2 + ... + t_rP_r = tPT \tag{6}\]
To find the element \( P_j \) of the transformation matrix \( P \) which leads to maximal variances, a stepwise optimization has to be solved [Isermann.]. For each step \( j \) with;
\[t_j = XP_j \tag{7}\]
a maximal variance of data \( t_j \) means,
\[
\max t_j^T t_j = \max(XP_j)^T (XP_j) \tag{8}\]
(max under the constraint of \( P = [p_1, p_2, ..., pr] \))
\[PTP = I \]
This means that the components are orthonormal.
The standard approach for this optimization problem is to use the method of Lagrange multipliers. If the function \( f(P_j) \) has to be maximized under the condition; \( g = PTjP_j - I = 0 \), the loss function becomes;
\[V = f(P_j) - \lambda_j g(P_j) \tag{10}\]
Where \( \lambda_j \) is the Lagrange multiplier. This gives;
\[V = PTjXT XP_j = \lambda_j(PTjP_j - I) \tag{11}\]
And;
\[2XT XP_j = 2 \lambda_j P_j = 0 \tag{12}\]
Or;
\[XT X - \lambda_j I P_j = 0 \tag{13}\]
With;
\[A = XT X, \]
The following holds;
\[A - \lambda_j I P_j = 0 \tag{13}\]
This is a classical eigenvalue problem. \( A \) is proportional to the correlation matrix or covariance matrix for zero mean variance of the measured data, \( \lambda_j \) is an eigenvalue and \( P_j \) an eigenvector of the matrix \( A \). From (13), it follows that;
\[PTj AP_j = P_j \lambda_j P_j \tag{14}\]
Inserting (14) in \( \max t_j^T t_j = \max P_j XT XP_j \) gives the maximal variance, that is;
\[
\max t_j^T t_j = \max P_j \lambda_j P_j \tag{15}\]
This maximal eigenvalues \( \lambda_j \) give maximal variance for coordinates \( t_j \).
The procedure to determine the transformation matrix \( P \) and the new variable \( T \) is as follows;
Calculate the correlation matrix; \( A = XT X \), with zero mean variables \( E\{xi(k)\} = 0 \) and \( E\{xi2(k)\} = 1 \)
Calculate the eigenvalues \( \lambda_j \) of the matrix \( A \) and the eigenvectors \( P_j \) of \( (A - \lambda_j I)P_j = 0 ; j=1, 2, ..., m \).
Select the largest (most significant) eigenvalues \( \lambda_j \) and corresponding eigenvectors \( P_j \), \( j = 1, 2, ..., r \), leading to the approximation; \( X' = t_1P_1 + t_2P_2 + ... + t_rP_r \).
Determine the transformation matrix \( P; P = [p_1 p_2 ... pr] \).
Calculate the new data matrix, \( T = X'P[t1 t2 ... tr] \), with \( tj = X'Pj \).
The result is a new data matrix \( T \) with all original data but a reduced number \( r < m \) of coordinates or variables, that is the principal components. The PC data matrix, \( Xc \) carries approximately the same information on the variances of the variables as the original data matrix \( X \).
Back-transformation in the original data coordinate system gives,
\[X* = X'P[t1 t2 ... tr] = X[Nx] P[mx] \tag{16}\]
By back-transformation in the original data coordination system, one obtains the original variables with only significant variances, that is, significant noise effects have been removed [Isermann.].

To optimally capture the variations of the data while minimizing the effect of random noise corrupting the PCA representation, the loading vectors corresponding to the a largest singular values are typically retained. When the columns of the loading matrix \( P \) \( Rmx \) are selected to correspond to the loading vectors associated with the first a singular values, the projection of the observation \( \chi \) into the lower-dimensional space are contained in the score matrix,[Russell et al, 2001]
\[T = XP \tag{17}\]
And the projection of \( T \) back into the m-dimensional observation space,
\[\chi = TP^T \tag{18}\]
The residual matrix \( E \) is given by;
\[E = X - \chi \tag{19}\]
The residual matrix captures the variations in the observation space spanned by the loading vectors associated with the \( m - a \), smallest singular values. The sub spaces spanned by \( E \) are called the score space and residual space, respectively. The subspace contained in the matrix \( E \) has a small signal-to-noise ratio, and the removal of this space from \( X \) gives a more accurate representation of the system, [Chiang & Braatz, 2001]
A new observation vector in the Test data set, \( x \ \text{Rm} \), can be projected into the lower-dimensional score space \( ti = xTi \), where \( Ti \) is the ith loading vector, The transformed variable \( ti \) is also called the ith Principal Component of \( x \). The difference between
the transformed variables and the transformed observation is that the transformed variables shall be called the Principal Components and the individual transformed observations will be called the scores [Chiang & Braatz, 2001].

**RESEARCH METHODOLOGY**

Twelve parameter variables of interest are gotten from different Gas Turbine Engines with different operation conditions. These variables are obtained from three hundred Gas Turbine Engines, with seven operating conditions, sixteen different features, and thirteen classes. Before the EngData Training set are standardised, it was re-structured to seven operating conditions, fourteen features and only three classes are considered. These are structured to 300 by 98 matrixes and the EngData Test set are structured to 86 by 98 matrixes. The data classes or labels are structured to 300 by 3 matrixes, then to a vector of 300 by 1.

Data standardization is achieved using standard deviation method. Data standardization using standard deviation and mean, which may be referred to as z-score, is given by;

\[ Z = \frac{x - \mu}{\sigma} \]  

(20)

where \( \sigma \) is the standard deviation of the observed data, \( \chi \).

The EngData Training set was split into two set; Training data set, 200x98. Test data set, 100x98.

The same mean and standard deviation computed from the Training data set standardization are applied on the Test data set.

Let the Training data set of n observations and m systems’ parameter variables stacked into a matrix X be given by;

\[
X = \begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1m} \\
X_{21} & X_{22} & \cdots & X_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nm}
\end{bmatrix}
\]  

(21)

X is the centre data matrix with dimension n x m. This matrix contains observations that are centred about the mean; that is, the Training data set has been subtracted from each row. X in the work is called the Training data set, with n = 300 and m = 98

The Training Data set having been standardized using mean (zero mean) and standard deviation (unity standard deviation), as shown in equation 20, was projected to lower dimensional space. The results of the analysis are shown below; Mean of the standardized Training Data set = 1.0 x 10-13, Standard deviation of the standardized data = 1.00;

PCA data is applied on the standardized data as shown in Table 1.

Reduction order is determined using Scree Test method as shown in Figure 2 In scree test, the knee point on the scree plot is used to determine the number of PCs to be retained for further analysis, indicating that the knee point appear at 12 PCs mark.

Data Visualisation in 2-dimensional space with PCA is carried out on the training data set as presented on . The Labels or the Training target of the EngData set is classified into three classes, namely; Class 1: good GTE with blue cross sign Class 2: average GTE with green circle, and Class 3: bad GTE with red plus sign

PCA and Nearest-Neighbour Classification of GTE Training Data Set for Fault Diagnosis are performed. PCA is not always an optimal dimensionality-reduction procedure for classification purposes especially when it comes to nonlinear data as in the case of this work. To extract features from the EngData Training set, we adopt nearest-neighbour classification method for pattern classification. The new sample is classified by calculating the distance to the nearest training cases; the sign of that point determines the classification of the sample. In this work we take \( K = 1 \). For optimally robust classification, \( K = 3 \) or 5 can be used.

The following inputs are used in this work for 1-NN classification;

- Number of Training cases = 200
- Number of test cases = 100
- Number of classes = 3, (1 = good GTE, 2 = average GTE, 3 = bad GTE).
- Training set = 200 by 12
- Test set = 100 by 12
- Training labels = 200 by 1 (vector)
- Test labels = 100 by 1 (vector).

We adopt 10-fold cross validation of the Training data set to evaluate the accuracy of the nearest-neighbour classification of the GTE data analysed with PCA and the results are as shown in Table 4.

**MODEL EXPERIMENT**

The application of PCA to the Standardized data produces 15 of the 98 eigenvalues presented in Table 1. Although several methods exist for determining the number of PCs to be retained or the reduction order, Scree test technique is adopted for this study to determine the reduction order. The Scree plot is presented in Figure 2. Figure 2 shows the Scree plot used to determine the reduction order. In scree test, the knee point on the scree plot is used to determine the number of PCs to be retained for further analysis, indicating that the knee point appear at 12 PCs mark.
### Table 1 showing 15 PCs, Eigenvalues

<table>
<thead>
<tr>
<th>Principal Components (PCs)</th>
<th>Eigval (latent)</th>
<th>Eigval (%) Cusum of Eigval (%)</th>
<th>Eigval ≥ 1</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC#1</td>
<td>36.6533</td>
<td>37.4013</td>
<td>37.4013</td>
<td>YES</td>
</tr>
<tr>
<td>PC#2</td>
<td>13.9509</td>
<td>14.2356</td>
<td>51.6369</td>
<td>YES</td>
</tr>
<tr>
<td>PC#3</td>
<td>8.5086</td>
<td>8.6822</td>
<td>60.3191</td>
<td>YES</td>
</tr>
<tr>
<td>PC#4</td>
<td>7.2647</td>
<td>7.4129</td>
<td>67.7321</td>
<td>YES</td>
</tr>
<tr>
<td>PC#5</td>
<td>6.4723</td>
<td>6.6044</td>
<td>74.3365</td>
<td>YES</td>
</tr>
<tr>
<td>PC#6</td>
<td>4.8586</td>
<td>4.9577</td>
<td>79.2942</td>
<td>YES</td>
</tr>
<tr>
<td>PC#7</td>
<td>3.7902</td>
<td>3.8675</td>
<td>83.1617</td>
<td>YES</td>
</tr>
<tr>
<td>PC#8</td>
<td>3.2723</td>
<td>3.3390</td>
<td>86.5008</td>
<td>YES</td>
</tr>
<tr>
<td>PC#9</td>
<td>2.3849</td>
<td>2.4438</td>
<td>88.9445</td>
<td>YES</td>
</tr>
<tr>
<td>PC#10</td>
<td>1.6638</td>
<td>1.6977</td>
<td>90.6423</td>
<td>YES</td>
</tr>
<tr>
<td>PC#11</td>
<td>1.2393</td>
<td>1.2646</td>
<td>91.9069</td>
<td>YES</td>
</tr>
<tr>
<td>PC#12</td>
<td>0.9210</td>
<td>0.9398</td>
<td>92.8467</td>
<td>NO</td>
</tr>
<tr>
<td>PC#13</td>
<td>0.8787</td>
<td>0.8966</td>
<td>93.7434</td>
<td>NO</td>
</tr>
<tr>
<td>PC#14</td>
<td>0.7817</td>
<td>0.7977</td>
<td>94.5410</td>
<td>NO</td>
</tr>
<tr>
<td>PC#15</td>
<td>0.7240</td>
<td>0.7388</td>
<td>95.2799</td>
<td>NO</td>
</tr>
</tbody>
</table>

The knee point indicates the number of PCs retained, thus 12 PCs are chosen for further analysis. Which is given by $\text{TrainSet}_Xc = \text{score}(:,1:12)$, or $Xc * p_c(:,1:12)$.

Figure 3 gives a two-dimensional plot of data visualisation in 2-dimensional space with First and Second PC.

### RESULT AND DISCUSSION

In Health Monitoring of Gas Turbine Engine Implementation, EngData Training set was split into training and test data sets, with the capacity of 200x98 and 100x98 respectively. The results of the analysis are showed Mean of the standardized Training Data set of 1.0 x 10^{-13} and Standard deviation of the standardized data of 1.00. The application of PCA to the Standardized data produced 15 of the 98 eigenvalues presented in Table 1. The first column shows the first fifteen principal components of the observed variables. The third column shows the cumulative sum of the eigenvalues of the observed variables, which indicates that about 95% of the information contained in the data set can be represented with only 15 of 98 principal components. The remaining 83 components can be seen as trivial or noisy components. The principal component scores are the representation of Training data set in the principal component space with dimension 200 by 98, which is the back transformation in the original data variables with only significant variances. Each column contains the coefficients for one PC. The columns are in order of decreasing component variance. The size or dimension of the PC is 98 by 98. The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables, That is the first PC accounts for 37.4013 percent of the information of the entire data. Under typical conditions, this means that the first component will be correlated with at least some of the observed variables. The second component extracted will account for the variance in the data set that is not accounted for by the first principal component. Also the second component will not be correlated with the first principal component. The remaining components that are extracted in the analysis display the same two characteristics stated above. Each component accounts for the amount of variance in the observed variables that was not accounted for by the preceding components, and is...
not correlated with all of the preceding components. From the table, it can be seen that over 90 percent of the information contained in the data is captured by just 10 PCs. The Latent, which is a vector containing the eigenvalues of the covariance matrix of the Training data set, and its dimension is 98 by 1.

In reality, the number of components extracted in a principal component analysis is equal to the number of observed variables being analyzed, in this case 98. However, in most analyses, only the first few components account for meaningful amounts of variance, so only these first few components are retained, interpreted, and used in subsequent analyses and the rest account for trivial amounts of variance. Figure 2 shows the Scree plot used to determine the reduction order. In scree test, the knee point on the scree plot is used to determine the number of PCs to be retained for further analysis, indicating that the knee point appear at 12 PCs mark.

Figure 3 gives a two-dimensional plot of data visualisation in 2-dimensional space with 3 labels good GTE with blue cross sign, average GTE with green circle, and bad GTE with red plus sign. The table 2 shows the output of 1-NN classification for GTE fault diagnosis. Due to nonlinear nature of the data under analysis, and the fact that linear method of dimensionality reduction technique has been used to analyse the GTE data, 1-NN classification of the analysed data does not give a very impressive classification. The result as shown in Table 3 indicates that the error rate or percentage misclassification of these data is high as.

The total number of test labels = 100, Total number of classified cases = 74, Total number of misclassification = 26, giving the performance of the classification as 74%. Error rate or percentage misclassification 26% is achieved. Good turbine engine is donated by 1, average turbine engine is donated by 2 and bad turbine engine is donated by 3. Also, Table 3 shows that 12 good GTE out of 15 were classified as good GTE, 3 good GTE out of 15 were classified as average GTE and no good GTE was classified as bad GTE. It can be observed that no bad GTE was classified as good GTE. This is very reasonable for safety critical system such as GTE.

Table 4 shows the result of 10-fold Cross-validation of GTE Training Data set – PCA with average error rate of 17%. It is also observed that M-fold cross validation gives much better result as the percentage misclassification or error rate decreases to 17%.

CONCLUSION
PCA as one of the attractive data-based techniques is simple and cost effective method of monitoring the health condition of a system, as part of the predictive maintenance strategy that seeks to improve and extend the reliability and life of the system. The data-based model performance evaluation indicates that PCA is very suitable in analysing high-dimensional data set, as the type used in this work. It is recommended that this data-based method of systems’ health monitoring be automated and integrated to the GTE. This can be done by using application software to capture important systems parameters, analyze the data, and generate fault management information for predictive maintenance decision to be taken.

REFERENCES


APPENDIX

Table 3 presents NN classification result of Test data set with PCA.

<table>
<thead>
<tr>
<th>S/N</th>
<th>Test Labels</th>
<th>NN Predict</th>
<th>Classified/ Misclassified (CL / MCL)</th>
<th>S/N</th>
<th>Test Labels</th>
<th>NN Predict</th>
<th>Classified/ Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>CL</td>
<td>35 2 2 CL</td>
<td>69</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>CL</td>
<td>36 2 2 CL</td>
<td>70</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>CL</td>
<td>37 1 2 MCL</td>
<td>71</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>CL</td>
<td>38 1 2 MCL</td>
<td>72</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>MCL</td>
<td>39 1 1 CL</td>
<td>73</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>CL</td>
<td>40 2 2 CL</td>
<td>74</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>CL</td>
<td>41 3 2 MCL</td>
<td>75</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>CL</td>
<td>42 3 2 MCL</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>CL</td>
<td>43 2 2 CL</td>
<td>77</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>CL</td>
<td>44 3 2 MCL</td>
<td>78</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>CL</td>
<td>45 3 2 MCL</td>
<td>79</td>
<td>2</td>
<td>1</td>
<td>MCL</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>CL</td>
<td>46 2 2 CL</td>
<td>80</td>
<td>2</td>
<td>1</td>
<td>MCL</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>MCL</td>
<td>47 2 1 CL</td>
<td>81</td>
<td>2</td>
<td>1</td>
<td>MCL</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>MCL</td>
<td>48 2 2 CL</td>
<td>82</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>MCL</td>
<td>49 2 2 CL</td>
<td>83</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>MCL</td>
<td>50 1 2 MCL</td>
<td>84</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>CL</td>
<td>51 2 2 CL</td>
<td>85</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>CL</td>
<td>52 2 1 MCL</td>
<td>86</td>
<td>2</td>
<td>1</td>
<td>MCL</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>CL</td>
<td>53 1 1 CL</td>
<td>87</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>CL</td>
<td>54 2 2 CL</td>
<td>88</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>MCL</td>
<td>55 2 2 CL</td>
<td>89</td>
<td>3</td>
<td>3</td>
<td>CL</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>CL</td>
<td>56 2 2 CL</td>
<td>90</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>CL</td>
<td>57 2 2 CL</td>
<td>91</td>
<td>2</td>
<td>1</td>
<td>MCL</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>MCL</td>
<td>58 2 1 MCL</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>CL</td>
<td>59 3 3 CL</td>
<td>93</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>CL</td>
<td>60 3 2 MCL</td>
<td>94</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>CL</td>
<td>61 2 2 CL</td>
<td>95</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>CL</td>
<td>62 2 1 MCL</td>
<td>96</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>CL</td>
<td>63 1 1 CL</td>
<td>97</td>
<td>1</td>
<td>1</td>
<td>CL</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>CL</td>
<td>64 2 2 CL</td>
<td>98</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>MCL</td>
<td>65 2 2 CL</td>
<td>99</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>CL</td>
<td>66 3 3 CL</td>
<td>100</td>
<td>2</td>
<td>2</td>
<td>CL</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>MCL</td>
<td>67 3 3 CL</td>
<td></td>
<td>Percentage Classified = 74%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>2</td>
<td>CL</td>
<td>68 3 3 CL</td>
<td></td>
<td>Percentage misclassified = 26%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 presents NN classification result of Test data set with PCA.

<table>
<thead>
<tr>
<th>KNOWN CLASSIFICATION</th>
<th>PREDICTED CLASSIFICATION</th>
<th>Good GTE (class 1)</th>
<th>Average GTE (class 2)</th>
<th>Bad GTE (class 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good GTE (class 1)</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Average GTE (class 2)</td>
<td>11</td>
<td>37</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Bad GTE (class 3)</td>
<td>0</td>
<td>12</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
Total number of test cases = 100
Total number of Good GTE = 15; percentage of good GTE classification = 80%
Total number of Average GTE = 48; percentage of average GTE classification = 77%
Total number of Bad GTE = 37; percentage of bad GTE classification = 67.6%

Table 4 shows evaluation of the accuracy of the nearest-neighbour classification of the GTE data analysed with PCA.

<table>
<thead>
<tr>
<th>Partition #1</th>
<th>Partition #2</th>
<th>Partition #3</th>
<th>Partition #4</th>
<th>Partition #5</th>
<th>Partition #6</th>
<th>Partition #7</th>
<th>Partition #8</th>
<th>Partition #9</th>
<th>Partition #10</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL</td>
<td>P</td>
<td>C</td>
<td>/</td>
<td>M</td>
<td>TL</td>
<td>P</td>
<td>C</td>
<td>/</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>C</td>
<td>2</td>
<td>3</td>
<td>M</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>C</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4: Result of 10-fold Cross-validation of GTE Training Data set - PCA.

TL = TEST LABELS
Pd = PREDICTION
C = CLASSIFIED
M = MISCLASSIFIED

Average $E_R = \frac{1}{M} \sum_{i=1}^{M} E_{Ri}$

Average $E_R = \frac{1}{10} (0.25 + 0.1 + 0.15 + 0.1 + 0.25 + 0.20 + 0.15 + 0.0 + 0.25 + 0.25)$

Average $E_R = 0.17 = 17$

M-fold cross validation gives much better result as the percentage misclassification or error rate decreases to 17%.