THE APPLICATION OF EXTENSIVE FEATURE EXTRACTION AS A COST STRATEGY IN CLINICAL DECISION SUPPORT SYSTEM

Odikwa, Henry[1], Ugwu, Chidiebere[2], Inyiama Hycinth[3]

[1,2]Department of Computer Science, University of Port Harcourt, Nigeria
[3]Department of Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria

ABSTRACT

Patients waste great deal of resources in the cause of identification of pathogens that caused their ailments; this calls for concern, hence the need to develop a veritable tool for minimizing the cost involved in classification of disease pathogens without compromising accuracy. In this paper, we developed a feature extraction model which reduces the clinical markers for prostate cancer and diabetes. The feature extraction, in the form of principal component analysis (PCA), was used to extract relevant components from prostate cancer and diabetes datasets. The simulation and experiment of the system were done with matlab. The system was able to extract 3 relevant features out of 4 prostate cancer clinical markers and 4 relevant features out of 5 diabetes clinical markers. The result showed that when trained in a multilayer neural network it yielded better classification accuracy with the extracted relevant features with 80% and 75% component analysis in prostate cancer and diabetes datasets respectively.

KEYWORDS

Feature extraction algorithm, principal component analysis, prostate cancer, diabetes, clinical markers.

1. INTRODUCTION

Feature extraction is a powerful tool used for reducing high dimensional data, especially in image processing [5]. Consequently, it can be applied in extracting relevant features to support decision-making in clinical markers. It is not far-fetched, that medical personnel must go into several consultations based on the numerous tests presented to them as a result of numerous clinical markers at their disposal.

Moreover, on the part of the patient, financial involvement is a constraint as he is faced with many clinical tests to be conducted which also cost much money and in turn time consuming and energy. The essence of using feature extraction is to minimize cost of analyzing feature vectors and reduce them to fewer samples without compromising the accuracy of classification [1]. If the features extracted are cautiously selected, it is assumed that the feature sets will definitely extract the related information from the input data in order to execute the preferred mission using this abridged representation instead of the full-size input and dimension (feature vectors). Feature extraction on its own involves simplifying the total sum of resources necessary to illustrate a huge set of data correctly [4]. It has been discovered that when analyzing a complex or intricate data, one of the major setbacks is the number of variables involved in the analysis [12] [15]. This paper is vital in solving the problem of employing too many clinical markers (features) in classification of diseases, which thereby reduces the cost of running too many clinical tests.

2. REVIEW OF RELATED WORK

Feature extraction has been widely employed in face recognition, image compression and also in finding patterns in high dimensional data to reduce cost of analyzing huge sum of data to smaller components [2]. Recently, in the health sector, it has been adopted as a tool for extracting relevant

DOI: 10.5121/ijdkp.2017.7603

medical tests without compromising the validity of the information. Fourier Transform Raman Spectroscopy (FTRS) was used as a diagnostic tool for the detection of oral cancer by using auto-associative network to analyze the spectra region of squamous cell [8]. This method was able to select 1556cm^3 of squamous cells datasets for analysis in neural network training. [3], it used feature selection methods to improve the classification of position emission tomography (PET) images to diagnose Alzheimer's disease. A textual feature extraction that is capable of extracting gray level run length matrix (GLRLM) was developed to differentiate segmented region in a TRUS test [9]. This method segmented images from trans-rectal ultra sound (TRUS) samples into few components. Auto-associative was employed to detect multiple cardiovascular diseases in a patient suffering from heart disease [14]. In [10], authors proposed feature selection method for diagnosing of prostate cancer from ultrasound imaging using digitized images from prostate TRUS, and the performance outweighs the method developed previously [7]. These methods were successfully employed with auto associative feature extraction model in pattern and image processing. This paper presents a different approach of extending feature extraction in reducing clinical markers for disease classifications, which results in cost reduction in clinical decision support system.

3. Materials And Method

Prostate biomarkers of 500 patients from Federal medical center repository, comprising four (4) prostate biomarkers which included prostate specific antigen (PSA), digital rectal examination (DRE), prostate weight and prostate volume were analyzed with a feature extraction model depicted in Fig.1. Furthermore, 100 dataset of diabetes disease with five biomarkers which included glycated hemoglobin (AIC) test, body mass index (BMI), random blood sugar (RBS), fasting blood sugar (FBS) and oral glucose test (OGT) were also analyzed and the relevant features extracted. The Fig1 shows the feature extraction model deployed to achieve the cost reduction in clinical decision support system.

3.1 Feature Extraction Model

The Fig 1 shows the model architecture of feature extraction that is designed to extract relevant features from diseased clinical markers.

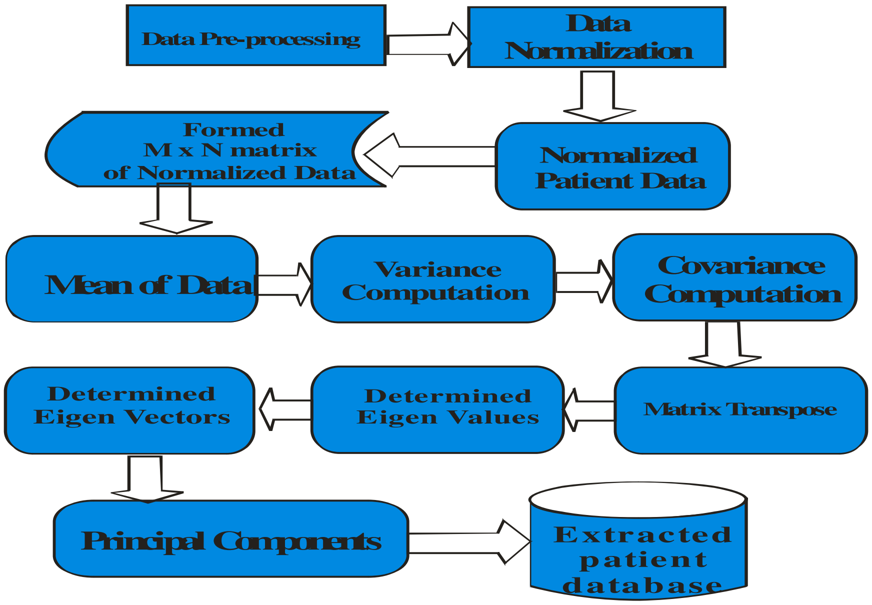


Fig. 1: The Model Architecture

3.2 THE DATA NORMALIZATION

Table 1: Diabetic Data before Normalization

S/N	Age (yrs)	AIC (%)	BMI (g)	RBS (Mmol/L)	FBS (Mmol/L)	OGT (Mmol/L)
1	45	6.30	25.00	11.80	7.00	8.80
2	50	5.30	26.00	12.00	8.00	11.30
3	75	6.80	26.00	11.90	5.70	11.50
4	48	7.00	29.00	13.00	5.80	10.50
5	89	9.00	30.00	11.20	9.00	11.90
6	90	6.00	21.00	13.00	7.00	12.00
7	56	8.10	27.00	14.00	8.00	13.00
8	75	5.30	28.00	12.00	6.50	13.00
9	59	5.20	27.00	11.00	7.30	13.80
10	45	8.00	20.00	11.70	5.90	10.50
11	89	5.90	29.00	11.80	8.00	14.00
12	50	6.30	30.00	11.10	10.00	14.00
13	49	6.00	34.00	15.70	5.70	11.98
14	60	6.90	56.00	15.00	7.00	12.00
15	55	6.80	26.00	11.80	6.00	13.00
16	78	5.80	27.00	11.90	6.80	13.80

The Table 1 shows the clinical markers of diabetes patients having five (5) features, namely; glycated hemoglobin (AIC) measured in percentage, body mass index (BMI) measured in grammes, random blood sugar (RBS) measured in mol per litre, fasting blood sugar (FBS) measured in milli mole per litre and oral glucose test measured in milli mole per litre.

Table 2: Prostate Cancer Data before Normalization

S/N	Age (yrs)	Prostate Weight (ng/mL)	DRE	Prostate Weight (g)	Prostate Volume (mL)
1	68	20.10	3.00	80.00	90.00
2	78	20.50	3.00	80.00	160.00
3	83	15.50	0.00	200.00	70.00
4	85	22.60	2.00	80.00	50.00
5	71	24.00	1.00	55.00	70.00
6	65	1.50	1.00	90.00	40.00
7	61	34.00	0.00	120.00	15.00
8	76	64.00	1.00	70.00	45.00
9	70	18.00	1.00	55.00	70.00
10	81	39.00	0.00	60.00	80.00
11	64	14.00	2.00	80.00	50.00
12	82	23.00	0.00	78.00	60.00
13	64	34.50	2.00	60.00	90.00
14	73	21.30	0.00	70.00	160.00
15	64	34.00	0.00	90.00	70.00
16	73	33.20	0.00	80.00	50.00
17	64	10.20	0.00	80.00	70.00
18	53	71.00	1.00	200.00	40.00
19	72	54.00	0.00	80.00	15.00

Table 2 shows the sample of prostate cancer clinical markers of patients before normalization with four (4) features namely; .prostate specific antigen (PSA), measured in nanogram per milli mole; DRE which has the values from 0 to 3; 0 value means soft, the value 1 means nodular, the value 2 means firm while the value 3 means hard. The prostate weight is measured in grams whereas prostate volume is measured in milli litre.

The normalization of the data was done and the values range from 0 to 1. Since the various attributes (feature vectors) employed in this paper have different variable value range and to reduce the differences in training results, normalization is necessary. The normalized data was subjected to min-max normalization method.
The data normalization is determined with equation (2), using the min-max value method [11].

$$\frac{\left(VAL.\frac{A}{F}.NORM - 0\right)}{(1-0)} = \frac{\left(VAL.\frac{B}{F}NORM. - MIN\right)}{(MAX. - MIN.)}$$

$$\frac{\left(VAL.\frac{A}{F}.NORM - 0\right)}{(1)} = \frac{\left(VAL.\frac{B}{F}NORM. - MIN\right)}{(MAX. - MIN.)}$$

$$\left(VAL.\frac{A}{F}.NORM\right) = \frac{\left(VAL.\frac{B}{F}NORM. - MIN\right)}{(MAX. - MIN.)}$$

$$X^1 = \frac{(X - MIN)}{(MAX - MIN)} \tag{1}$$

Where X represents each data entry, Min is the minimum value from each row entry and max denotes the maximum value from each row entry.

By applying equation (1), the maximum value of AIC is 9 and the minimum value is 4.6, the maximum and minimum values of BMI are 20 and 56 respectively. Also, the minimum and maximum values of RBS are 11 and 28.1 respectively. For FBS, the minimum value is 5.6 and the maximum value is 18, and OGT has a minimum value of 8.7 and maximum value of 20from table 1.

$$X^1(AIC) = \frac{(6.30 - 4.60)}{(9.00 - 4.60)} = 0.40$$

$$X^1(BMI) = \frac{(25.00 - 20.00)}{(56.00 - 20.00)} = 0.10$$

$$X^1(RBS) = \frac{(11.00 - 11.00)}{(28.10 - 11.00)} = 0.00$$

$$X^1(FBS) = \frac{(7.00 - 5.60)}{(18.00 - 5.60)} = 0.10$$

$$X^1(OGT) = \frac{(8.80 - 8.70)}{(20.00 - 8.70)} = 0.00$$

These calculations represent normalization of data in Table 1 and the results of the normalized data in Table 3 for the first column and the other normalized values follow using the same process.

Table 3: Determined Normalization Result for Diabetes

S/N	Age (yrs)	AIC(%)	BMI (g)	RBS (Mmol/L)	FBS (Mmol/L)	OGT (Mmol/L)
1	45	0.40	0.10	0.00	0.10	0.00
2	50	0.20	0.20	0.10	0.20	0.20
3	75	0.50	0.20	0.10	0.00	0.20
4	48	0.50	0.30	0.10	0.00	0.20
5	89	1.00	0.30	0.00	0.30	0.30
6	90	0.30	0.00	0.10	0.10	0.30
7	56	0.80	0.20	0.20	0.20	0.40
8	75	0.20	0.20	0.10	0.10	0.40
9	59	0.10	0.20	0.00	0.10	0.50
10	45	0.80	0.00	0.00	0.00	0.20
11	89	0.30	0.30	0.00	0.20	0.50
12	50	0.40	0.30	0.10	0.40	0.50
13	49	0.30	0.40	0.30	0.00	0.30
14	60	0.50	1.00	0.20	0.10	0.30
15	55	0.50	0.20	0.00	0.00	0.40
16	78	0.30	0.20	0.10	0.10	0.50
17	60	0.30	0.20	0.10	0.10	0.60
18	89	0.20	0.10	0.20	0.20	0.40
19	56	0.30	0.10	0.30	0.10	0.30
20	64	0.50	0.10	0.40	0.00	0.50

Consequently, the maximum value of PSA is 1237 and the minimum value is 0.1; the maximum and minimum values of DRE are 3 and 0 respectively. Also, the minimum and maximum values of Prostate weight are 250 and 25 respectively. For Prostate volume, the maximum value is 160 while the minimum value is 15 from Table 2.

$$X^1(PSA) = \frac{(20.10 - 0.10)}{(1237 - 0.10)} = 0.00$$

$$X^1(DRE) = \frac{(3.00 - 0.00)}{(3.00 - 0.00)} = 1.00$$

$$X^1(PW) = \frac{(80.00 - 25.00)}{(250.00 - 25.00)} = 0.20$$

$$X^1(PV) = \frac{(90.00 - 15.00)}{(160.00 - 15.00)} = 0.50$$

The determined normalization values of PSA, DRE, PW and PV are represented in table 4 for the first column in table 2 and the calculations followed thus for the other data entries normalized in the range of 0 to 1.

Table 4: Determined Normalization Result for Prostate Cancer

S/N	Age (yrs)	Prostate Weight (ng/mL)	DRE	Prostate Weight(g)	Prostate Volume (mL)
1	68	0.0	1.0	0.2	0.5
2	78	0.0	1.0	0.2	1.0
3	83	0.0	0.0	0.8	0.4
4	85	0.0	0.7	0.2	0.2
5	71	0.0	0.3	0.1	0.4
6	65	0.0	0.3	0.3	0.2
7	61	0.0	0.0	0.4	0.0
8	76	0.1	0.3	0.2	0.2
9	70	0.0	0.3	0.1	0.4
10	81	0.0	0.0	0.2	0.4
11	64	0.0	0.7	0.2	0.2
12	82	0.0	0.0	0.2	0.3
13	64	0.0	0.7	0.2	0.5
14	73	0.0	0.0	0.2	1.0
15	64	0.0	0.0	0.3	0.4
16	73	0.0	0.0	0.2	0.2
17	64	0.0	0.0	0.2	0.4
18	53	0.1	0.3	0.8	0.2
19	72	0.0	0.0	0.2	0.0
20	86	0.0	1.0	0.1	0.2

3.3 FEATURE EXTRACTION PROCESS

The computation and analysis of the data was achieved with; matrix formation, calculations of the mean, variance and covariance, then the determination of the eigen values and the eigen vectors.

3.3.1 MATRIX FORMATION

Two feature vectors (attributes) of the dataset from prostate cancer and diabetes datasets are employed to construct an m x n matrix.

$$A = \left\{ \begin{matrix} x_{11} x_{12} \ldots x_1 n \\ x_{21} x_{22} \ldots x_2 n \\ . \quad . \quad \ldots \\ x_{m_1} x_{m_2} \ldots x_m n \end{matrix} \right\}$$

3.3.2 MATRIX SUMMATION

The sum of the matrixes were determined for each entity attribute or feature vector with equation (2).

$$X = \sum_{i=n} N \tag{2}$$

Where N is the total number of feature vectors in each row.

3.3.3 MEAN, VARIANCE AND COVARIANCE *DETERMINATION*

We determined the covariance matrix of the feature vectors. To determine the covariance, we first determine the mean from equation (3)

$$\bar{X} = \sum \left(_{i=1} \right) \frac{X_i}{N} \tag{3}$$

Where xi denotes the summation of the data entries and N is the total number of entries.
Dimensional features were constructed to extract the relevant features from the normalized data, Table 5 shows that the datasets are from x_1 to x_n representing X feature vectors (attributes) and y_1 to y_n denoting Y feature vectors (attributes).

.

Table 5: A 2-dimensional Feature Vectors

X	Y
X_1	Y_1
X_2	Y_2
-	-
X_n	Y_n

The mean was subtracted from each of the data entry. When the mean is subtracted from each of the data sets it produces a dataset whose mean was zero [13]. The adjusted mean table is denoted with X_A and Y_A in Table 6 from the result of the mean using equation (3).

Table 6: Adjusted Mean

X_A	Y_A
$\bar{X} - x_1$	$\bar{Y} - y_1$
$\bar{X} - x_2$	$\bar{Y} - y_2$
-	-
$\bar{X} - x_n$	$\bar{Y} - y_n$

$\bar{X} and \bar{Y}$ are the mean of the data.

The Table 7 shows the variance and covariance, which showcases how the data points vary and correlate with each other. This is a vital aspect of feature extraction process prior to determining the eigen values and eigen vectors. The covariance of X,Y is determined in equation (9).

Table 7: Covariance and Variance Calculation

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
x_1	y_1	$x_1-\bar{X}$	$y_1-\bar{Y}$	$(x_1-\bar{X})(y_1-\bar{Y})$
x_2	y_2	$(x_2-\bar{X})$	$(y_2-\bar{Y})$	$(x_2-\bar{X})(y_2-\bar{Y})$
.	.	.	.	.
.	.	.	.	.
x_n	y_n	$x_n-\bar{X}$	$y_n-\bar{Y}$	$(x_n-\bar{X})(y_n-\bar{Y})$
$\sum x$	$\sum Y$	$\sum X-\bar{x}$	$\sum Y-\bar{Y}$	$\sum(X-\bar{X})(Y-\bar{Y})$

Covariance X and Y are determined with equation (4)

$$Cov(X,Y)=\sum\frac{((X-\bar{X})(Y-\bar{Y}))}{(n-1)} \quad (4)$$

$$\sum X = X_1 + X_2 + \ldots + Xn \quad (5)$$

$\sum X$ is the sum of the data entries in row X

$$\sum Y = y_1 + y_2 + \ldots + yn \quad (6)$$

$\sum y$ is the sum of the data entries in row Y.

$$\sum(X-\bar{X}) = (x_1-\bar{X}) + (x_2-\bar{X}) + \ldots + (xn-\bar{X})\ldots(7)$$

$\sum(x-\bar{X})$ is the sum of entries in that row to determine variance of X applying eqn (7).

$$\sum(y-\bar{Y}) = (y_1-\bar{Y}) + (y_2-\bar{Y}) + \ldots + (yn-\bar{Y})\ldots(8)$$

Also $\sum(y-\bar{Y})$ denotes the sum of entries in row $y-\bar{Y}$ applied in calculating variance of Y using eqn (8).

$$\sum(x-\bar{X})(y-\bar{Y}) = (x_1-\bar{X})+(y_1-\bar{Y})+(x_2-\bar{X})(y_2-\bar{Y})\ldots+(xn-\bar{X})(yn-\bar{Y})\ldots(9)$$

Whereas Var(X,X) is the same as Cov(X,X), more so, Var(Y,Y) is same as Cov(Y,Y) which is determined in table 7.

From table 7, the covariance matrix A of M x N is formed;

Where the covariance matrix is determined in equation (9) and the transpose of the matrix was done.

$$Cov(Matrix) = \frac{1}{(n-1)} \sum (x-\mu)(x-\mu)^{T} \ldots (10)$$

Where n is the number of datasets, x is the corresponding vector values, U is the mean of the data and T is the transpose of the matrix

$$A = \begin{bmatrix} \text{cov (X)} & \text{cov(X,Y)} \\ \text{cov(Y,X)} & \text{cov(Y)} \end{bmatrix}$$

where

$$A^{T} = \begin{Bmatrix} x_{1_1}\, x_{2_1}\, x_{3_1} \ldots x_{m_1} \\ x_{1_2}\, x_{2_2}\, x_{3_2} \ldots x_{m_2} \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ x_{m_1}\, x_{2_n}\, x_{m_n} \ldots x_{m_n} \end{Bmatrix} \begin{matrix} \\ \\ . \\ . \\ \\ \end{matrix}$$

3.3.4 Eigen Value And Eigen Vector Determination

The eigen values and the corresponding eigen vectors were determined from the covariance matrix with equation (11):

$$AX = \lambda X \Rightarrow \lambda_1 \lambda_2 \ldots (11)$$

A is the transpose matrix and X is the covariance matrix, λ_1, λ_2 are the corresponding eigen values.

The other step is to determine the eigen vectors and eigen values from the covariance matrix A. To determine the eigen values, we set Matrix A which is the covariance matrix to det(A-λ_i) and equate it to zero.

$$A = \begin{bmatrix} \text{cov (X)} & \text{cov(X,Y)} \\ \text{cov(Y,X)} & \text{cov(Y)} \end{bmatrix} \Rightarrow \det(A - \lambda) = 0 \qquad (12)$$

Also,

$$A\vec{V} = \lambda \vec{V} \qquad (13)$$

Then with eqn(13), we set the eigen values as follows:

$$A = \begin{bmatrix} \text{cov (X) - A} & \text{cov(X,Y)} \\ \text{cov(Y,X)} & \text{cov(Y) - }\lambda \end{bmatrix}$$

The values of λ which are the eigen values of the covariance matrix were obtained and substituted in matrix A, then multiplied by orthogonal matrix x_1 and x_2 as follows:

$$\begin{bmatrix} \mathrm{cov}(X) - A & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y) - \lambda \end{bmatrix} \begin{bmatrix} x1 \\ x2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Then solve for x_1 and x_2 which are the eigen vectors of the matrix and determine the principal components by associating the eigen values with the corresponding highest eigen vectors.

3.3.5 Principal Component Analysis Judging Criteria

To form the principal component analysis, the eigen vectors with the highest eigen values determined become the first principal component of the data set and the order follows.

4. Experiments And Results

In this paper, feature extraction model was constructed to extract relevant features for a clinical decision support system. The experiment was carried out with prostate cancer and diabetic clinical markers.

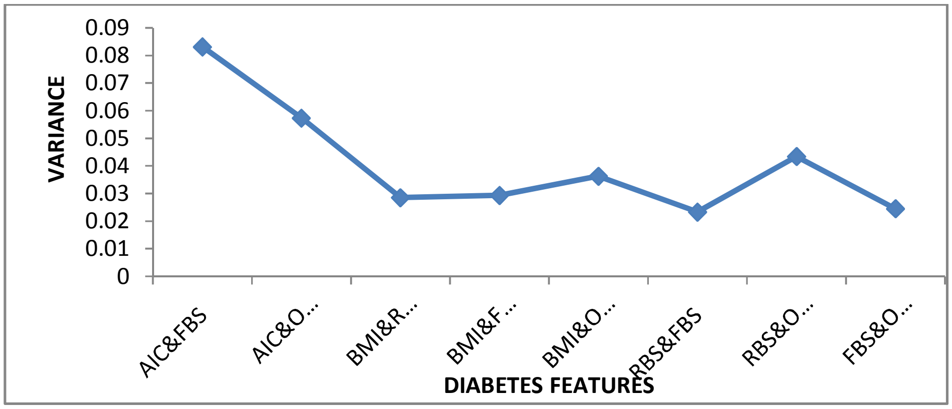


Fig. 2: Graphical representation of determined variance

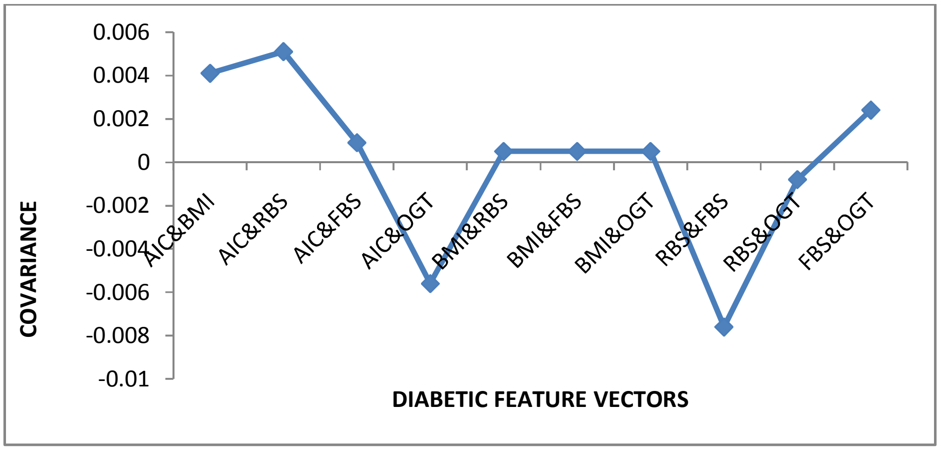


Fig. 3: Graphical representation of diabetes covariance result.

4.1 Diabetes Dataset

Using equation (1), the total sum of data entries for each row was determined. The determined sum of data entries on each row for diabetes datasets for the various features were AIC 47.4, BMI 31.7, RBS 26.2, FBS 13.9 and OGT 33.5.
The mean is determined with equation (3).

The results were 0.1302, 0.4248, 0.2556 and 0.3774 for the 500 dataset of prostate consisting four (4) features of prostate cancer; 0.4740, 0.3170, 0.2620, 0.1390 and 0.3350 for the 100 diabetes dataset comprising five (5) features. The adjusted mean is done with Table 6 and the summation gives a zero mean. Fig 2 shows the graph of determined variance plotted against the features of diabetes.

From Table 7, the determined covariance matrix is done using equation (8). Fig 3 shows the determined covariance matrix values formed by combining the features of diabetes.

To determine the eigen values and eigen vectors of the covariance matrix, matlab was deployed as a simulation tool as the number of diabetes dataset was too big to be analyzed manually. The results are shown in Table 8.

4.2 Prostate Cancer Dataset

With equation (1), the total sum of data entries for each row was determined for the four (4) features of prostate cancer. The sum of PSA is 20.7, DRE is 214, PW is 214 and PV is 188.1.
The mean was determined with equation (2) and we got the following results; PSA 0.1302, DRE 0.4248, PW 0.2556 and PV 0.3774.
The determined variance is done with equations (7) and (8)

The determined eigen values in table 8 for diabetes datasets are the right diagonal matrixes.
The Fig 4 shows the determined variance and Fig 5 shows the result of the covariance by applying equation (9).

The determined eigen values and eigen vectors are shown in Table 9.

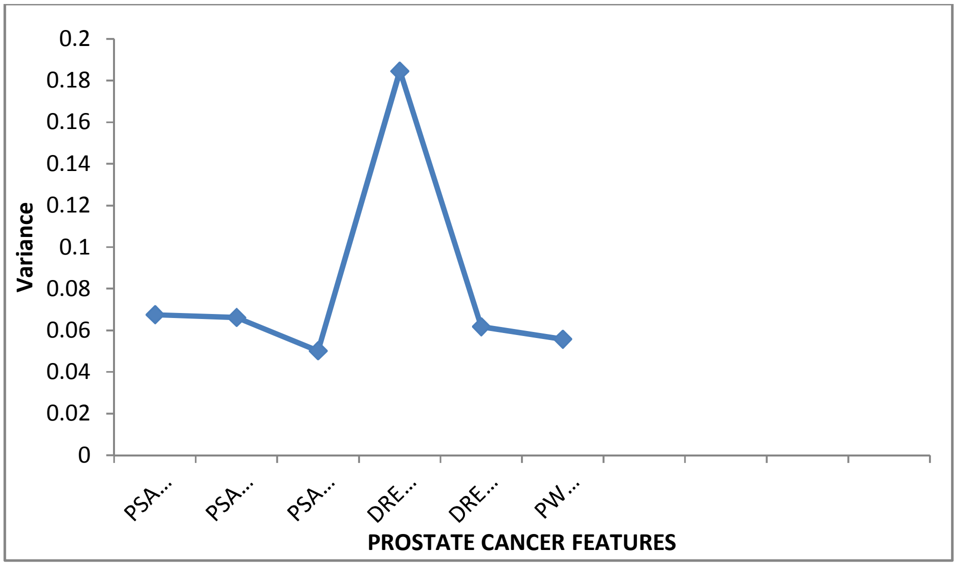
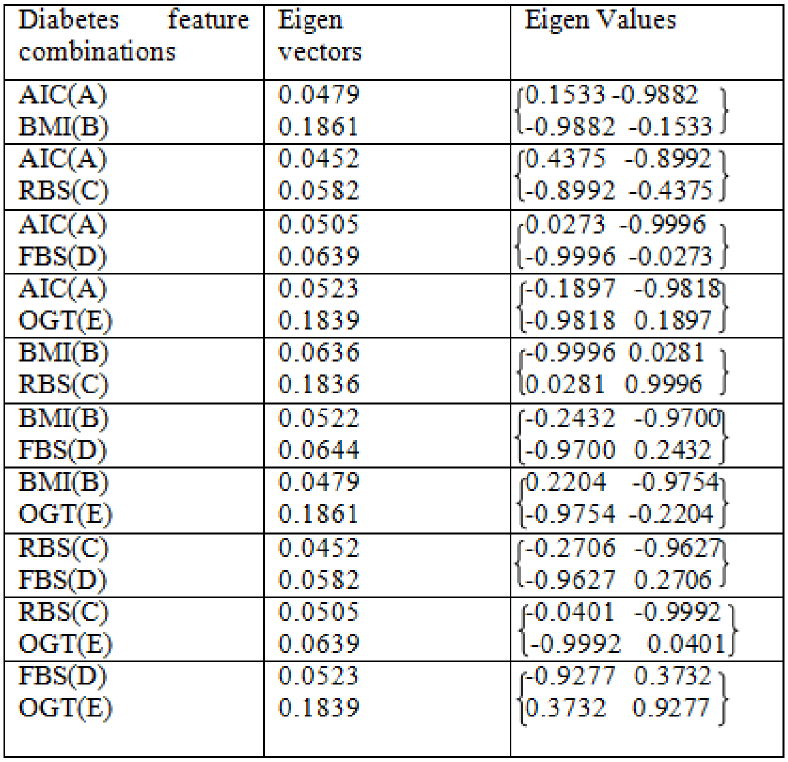


Fig. 4: Graphical representation of determined prostate cancer variance

Table 8: Results of Diabetes Eigen Vectors and Eigen Values



Diabetes feature combinations	Eigen vectors	Eigen Values

AIC(A)
BMI(B)
	0.0479
	0.1861
		$\begin{Bmatrix} 0.1533 & -0.9882 \\ -0.9882 & -0.1533 \end{Bmatrix}$

AIC(A)
RBS(C)
	0.0452
	0.0582
		$\begin{Bmatrix} 0.4375 & -0.8992 \\ -0.8992 & -0.4375 \end{Bmatrix}$

AIC(A)
FBS(D)
	0.0505
	0.0639
		$\begin{Bmatrix} 0.0273 & -0.9996 \\ -0.9996 & -0.0273 \end{Bmatrix}$

AIC(A)
OGT(E)
	0.0523
	0.1839
		$\begin{Bmatrix} -0.1897 & -0.9818 \\ -0.9818 & 0.1897 \end{Bmatrix}$

BMI(B)
RBS(C)
	0.0636
	0.1836
		$\begin{Bmatrix} -0.9996 & 0.0281 \\ 0.0281 & 0.9996 \end{Bmatrix}$

BMI(B)
FBS(D)
	0.0522
	0.0644
		$\begin{Bmatrix} -0.2432 & -0.9700 \\ -0.9700 & 0.2432 \end{Bmatrix}$

BMI(B)
OGT(E)
	0.0479
	0.1861
		$\begin{Bmatrix} 0.2204 & -0.9754 \\ -0.9754 & -0.2204 \end{Bmatrix}$

RBS(C)
FBS(D)
	0.0452
	0.0582
		$\begin{Bmatrix} -0.2706 & -0.9627 \\ -0.9627 & 0.2706 \end{Bmatrix}$

RBS(C)
OGT(E)
	0.0505
	0.0639
		$\begin{Bmatrix} -0.0401 & -0.9992 \\ -0.9992 & 0.0401 \end{Bmatrix}$

FBS(D)
OGT(E)
	0.0523
	0.1839
		$\begin{Bmatrix} -0.9277 & 0.3732 \\ 0.3732 & 0.9277 \end{Bmatrix}$

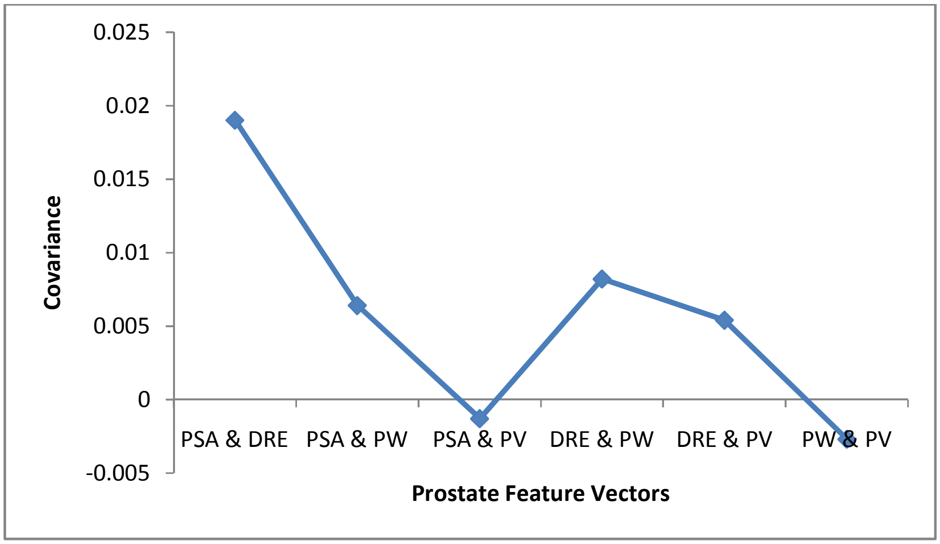


Fig. 5: Graphical Representation of covariance result

Table 9: Results of Prostate Eigen Vectors and Eigen Values

Prostate Feature Combination	Eigen vectors	Eigen Values
PSA(P) DRE (Q)	0.0479 0.1861	-0.9903 0.1389 0.1389 0.9903

PSA(P)
PW(R)
	0.0452
	0.0582
		-0.7647 0.7647
		0.6444 0.7647

PSA(P)
PV(S)
	0.0505
	0.0639
		-0.9953 -0.0971
		-0.0971 0.9953

DRE(Q)
PW(R)
	0.0523
	0.1839
		0.0624 -0.9980
		-0.9980 -0.9980

DRE(Q)
PV(S)
	0.0636
	0.1836
		0.0450 -0.9990
		-0.9990 -0.0450

PW(R)
PV(S)
	0.0522
	0.0644
		0.9741 -0.2262
		-0.2262 0.9741

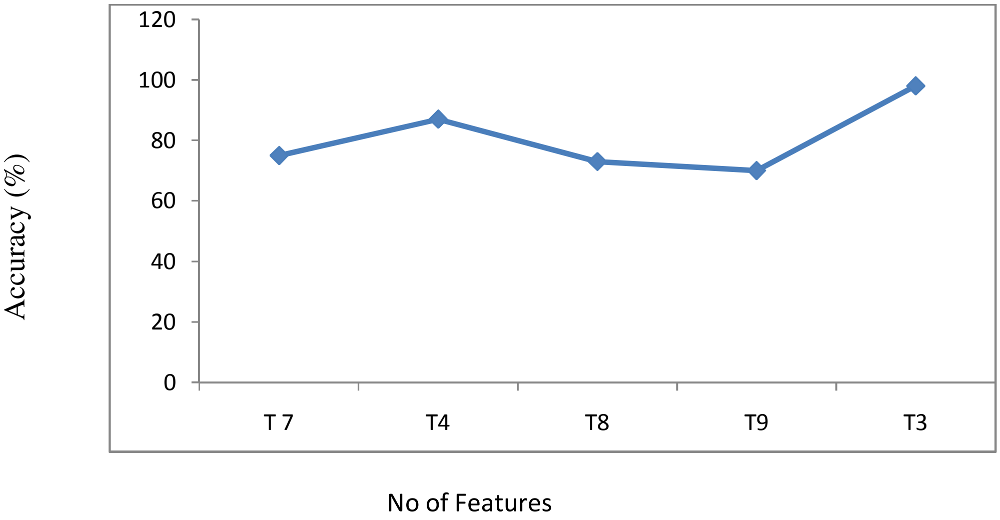


Fig. 6: Graphical representation of accuracy against no of features in neural network

To test the accuracy of the model developed, the absolute error (AE) was used [6],

$$Ab = Y - Y_{est} \tag{14}$$

Where Y is the actual prediction and Y_{est} is the estimated prediction

Table 10: Derived Principal Components

Principal Component Features	Prostate Data	Eigen Values	Diabetic Data	Eigen Values
1st Principal Component	PSA and PW	0.6444	FBS and OGT	0.3732
2nd principal component	PSA and DRE	0.1389	BMI and RBS	0.0281

5. DISCUSSION OF RESULTS

Several experiments and analyses were conducted to extract relevant features out of the prostate cancer and diabetes datasets having four (4) and five (5) features respectively. From the analysis, in Fig 2, AIC and FBS were found having the highest variance while the least were RBS and FBS. This implies that during the feature extraction one of them must be extracted as a relevant feature to form principal component. The covariance as shown in Fig 3 shows AIC and RBS having the highest covariance, this denotes that the two features are closely related. The eigen values and vectors were determined in Table 8, and going by the criteria for selecting the feature components as extracted relevant features, the eigen vector corresponding to the highest eigen value as 0.3732, which corresponds to the eigen vectors of 0.0252 an 0.0322 with the features FBS and OGT. This forms the first principal component for diabetes. The second eigen vector corresponding to the highest eigen value has the value of 0.0281 with eigen vectors of 0.0333 and 0.0511corresponding to the diabetic features of BMI and RBS. Therefore, these features formed the second component for the diabetes clinical data. Hence, this model extracted four features as relevant in the classification of diabetes diseases; this is affirmed in Table 10. For prostate cancer, the analysis is done and the variance of the various combined features determined with DRE and PW having the highest variance and PSA and PV having the least values as shown in Fig 4. This means that between DRE and PW they have equal chances of making the principal components as relevant features; the covariance were determined and Fig 5 shows the graphical representation with PSA and DRE having the highest covariance. In determining the eigen values and eigen vector, PSA and PW had 0.6444 as the highest eigen value and the corresponding eigen vectors were 0.0452 and 0.0582 respectively. Therefore, PSA and PW are the first principal component while the second eigen vector corresponding to the highest eigen value had the value of 0.1389 with corresponding eigen vectors of 0.0479 and 0.1861 respectively, having the features of PSA and DRE and is the second principal component. Thus, for prostate cancer the extracted features are PSA and PW, PSA and DRE, but since PSA has occurred before as the second principal component, therefore its existence in the second component becomes irrelevant. Thus, the relevant features for prostate cancer vital for classification accuracy are PSA, PW and DRE, whereas PV is considered irrelevant, thereby saving cost of running the clinical test by the patients. This is affirmed in Table 10 of the derived principal component. The Fig 6 gives a clearer view when tested the number of features in a neural network classification as training inputs, as the result yielded 98% and 94% accuracies for both prostate cancer and diabetes respectively. While absolute error from equation (14) was determined to be 0.04, this is an improvement on the model that enhances its validity.

6. CONCLUSION

In this paper, feature extraction has been successfully applied to extract relevant features as principal components in clinical markers, as a cost strategy in clinical decision support system by constructing a feature extraction model that applied the processes of determining the mean, variance, covariance, eigen values and vectors of datasets. The model was able to extract relevant features from clinical datasets and thus enhances accuracy of classification of diseases in a clinical decision support system and used prostate cancer and diabetes clinical datasets to validate the model. Therefore this model reduces cost to patients in running too many clinical markers on patients as we have seen that, prostate clinical marker tests reduced from four to three and diabetes marker tests reduced from five to four features respectively. Thus this study suggests that, for every clinical marker deployed to determine the outcome of a patient’s status, relevant features from the markers could be extracted without compromising the accuracy of the disease classification and in turn reduces the great deal of resources wasted by patients in running several of these clinical tests.

7. ACKNOWLEDGMENTS

We wish to acknowledge the efforts of Dr. O.O. Jegede of the Pathological Department, Federal Medical Center, Umuahia, and Dr. C.C. Chukwuegbo of PathConsult Medical Diagnostic Center, Umuahia. We also acknowledge Professor E.O. Nwachukwu for his untiring efforts towards the accomplishment of this paper.

REFERENCES

[1] Addison, J., Wermter, S., & Arevian, G. (2010). A Comparative Feature Extraction and Selection Technique, *Journal of electrical and computer engineering.* 4(102), 2-4.

[2] Alin, G., Leon J.M., Jacek, C., & Wojdel, P. (2010). Comparison between Different Feature Extraction Techniques for Audio-visual Speech Recognition*, Journal on multimodal user interfaces*. 4(2), 1-15.

[3] Bougioukos, P., Cavouras, D., Daskalakia, A., Kossida, S., Nikiforidia, G., & Bezezerianos, A. (2006). Feature Extraction and Analysis of Prostate Cancer Proteomic Mass Spectra for Biomarkers Discovery, *General secretariat for Research and technology* , Greece: 1-6.

[4] Chi-Hua, C. & Semir, Z. (2011). Frontoparietal Activation Distinguishes Face and Space from Artifact Concepts. *Journal of Cognitive Neuroscience, 23(*.9), 258-256.

[5] Chulhee, L., & David, L. (1992). Feature Extraction and Classification Algorithms for High Dimensional Data, *International journal of electrical and electronics engineering*, 8(200), 5-10.

[6] Hong, C., Zhibin, P., Luoqing, L., & Yoanyan, T. (2014). Error Analysis of Coefficient-based Regularized Algorithm for Density Level Detection. *Journal of MIT*, *25*(4), 1107-1121

[7] Lilian, R. H., Farid, H., & Donald, B. R. (2003). Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum, *Journal of computational Biology*, 10(5), 25-46

[8] Manavalan, R., & Thangavel, K. (2012). Evaluation of Textual Feature Extraction from GRLM for Prostate Cancer TRUS Medical Images, *International* journal of advances in engineering and technology, 4(11), 111-117.

[9] Oliveira, A., Mart, A. A., Silveira, L., Zangaro, R. A., & Zampieri, M. (2003). Application of Principal Components Analysis to Diagnosis Hamster Oral Carcinogenesis, *Journal of Biomedical vibrational spectroscopy*, 5321(10)

[10] Radharkrishnan, M., & Kuttiannam, T. (2012). Comparative Analysis of Feature Extraction Method for the Classification of Prostate Cancer from TRUS in Medical Images, *International journal of computer science* 9(1), 171-179.

[11] Rudong, L., Xiao D., Chengchery, M.,. & Lei, L. (2014). Computational Identification of Surrogate Genes for Prostate Cancer Phases Using Machine Learning and Molecular Network Analysis, *Biomed central journal of Shanghai, 5*(8), 320-325.

[12] Santhanam, T.,& Padmavathi, K. (2015). Application of K-means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis, *Procedia of Computer Science journal: Elsevier, 10*(2), 134-136.

[13] Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, C.T., … Golub, T.R. (2011). Diffuse Large b-Cell Lymphonia Outcome Prediction by Gene-expression Profiling and Supervised Learning, *Journal of Nature Medicine*, *13*(19), 68-74.

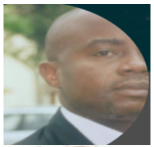
[14] Sreelakshmi, T. G., & Seethal, P. (2010). An accurate ECG Feature Extraction Method for Detecting Multiple Cardiovascular Disease, *Journal of Biomedics*, 123(4),112-110.

[15] Takehiko, I., Shinichi, O., Toshiaki, S., & Mamoru, J. (2013). Application of Artificial Intelligence and Genetic Algorithm in Physical Distribution. *International Journal of Electronics and Information System Division*, *2(*11), 23-24.

AUTHORS

Odikwa Ndubuisi is pursuing his doctorate degree in computer science at University of Portharcourt, Rivers State, Nigeria. He has M.Sc in infomation technology from National Open University of Nigeria and B.Tech in computer science and mathematics from Federal University of Technology, Owerri, Imo State, Nigeria. His research interest is on machine learning algorithms.

Ugwu Chidiebere is an associate professor of computer science in the University Of Portharcourt. He has doctorate degree in computer science from University of Portharcourt, M.SC and B.SC from the same university and presently he is the ICT director of the university. His research interest is on natural language processing.



Inyiama Hycinth is a professor of electronics and computer engineering at Nnamdi Azikiwe University, Awka, Anambra State, Nigeria. His research interest is on Pattern recognition and image processing.