

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO CLASSIFY NORMAL BRAIN TISSUE AND BRAIN LESIONS LIKE LOW AND HIGH GRADE GLIOMA, METASTASES AND MULTIPLE SCLEROSIS

ANINDYA GANGULY¹, TAPAN KRISHNA BISWAS^{2*}, RAJIB BANDOPADHYAY²,
AJAY KUMAR DUTTA³

1== College of Health and Human Sciences, Charles Darwin University, Australia,
2=Department of Instrumentation and Electronics Engineering, Jadavpur university, India.
3= Department of Production Engineering, Jadavpur University, India.

*= Corresponding Author- Email-tbiswas52@gmail.com

ABSTRACT

Principal Component Analysis (PCA) an extremely useful method of Statistical techniques is applied when working with a lot of parameters or independent numerical variables to predict the different pathological lesions in the brain like Multiple Sclerosis (MS), Glioma, Glioblastoma of different grades and Metastasis. Statistical techniques such as factor analysis or Principal Component Analysis(PCA) help to overcome such difficulties.

In different brain diseases structural alterations in the normal tissue may be noticed in MR images. It is not so simple to detect the brain lesions correctly even from the MR spectroscopic graph. Enormous data collected from various patients such as – Refractive Index, T2 relaxation values, Apparent Diffusion Coefficient (ADC), Creatine (CR), Choline (CHO), NAA (N-Acetyl Aspartate), ratio of CR/NAA, LIP/LAC (Lipid/lactate), MI (Myoinositol), CHO/CR and T2 value in the periphery of lesion may be confusing. The relationship between each variable may not be clear and that there is a chance of over fitting the data. By reducing the dimension of the feature space by “feature elimination” and “feature extraction”, there may be less chance of over fitting the data. PCA helps identifying the disease condition in doubtful cases by generating a map depicting and classifying the diseases.

Keywords: Principal Component Analysis (PCA); Magnetic Resonance Imaging (MRI); Metabolites of MR Spectroscopy; Refractive Index (RI); Ground Truth Image; Independent Numeric and Dependent Variable ; Prediction.

INTRODUCTION:

Tissue characterization or accurate diagnosis is not possible by observing the structural changes in the MR images without getting the histopathological study after a brain biopsy (Figure1) [1,2]. Sometimes some confusion is created by the images of Glioma in different stages, Glioblastoma, metastasis from primary cancer site and benign diseases like multiple sclerosis (relapsing remitting or tumefactive multiple sclerosis) [2]. Even MR Spectroscopy (MRS) fails to detect the exact character of the lesion from the graph generated by the peak of different metabolites along with the quantity [3, 4]. In this study, **Principal component Analysis (PCA)** had been tried to simplify the physical and chemical data derived from the MR Images [5]. PCA is a sophisticated and extensively used procedure for deciding the composition of recurrent variability.

One of the important problems in MR image processing is classification of diseases and tissue character based on chemical and physical information. Depending on attributes of multiple independent numeric variables extracted from the input images or **ground truth images** prediction of disease/tissue classification can be made by PCA [6].

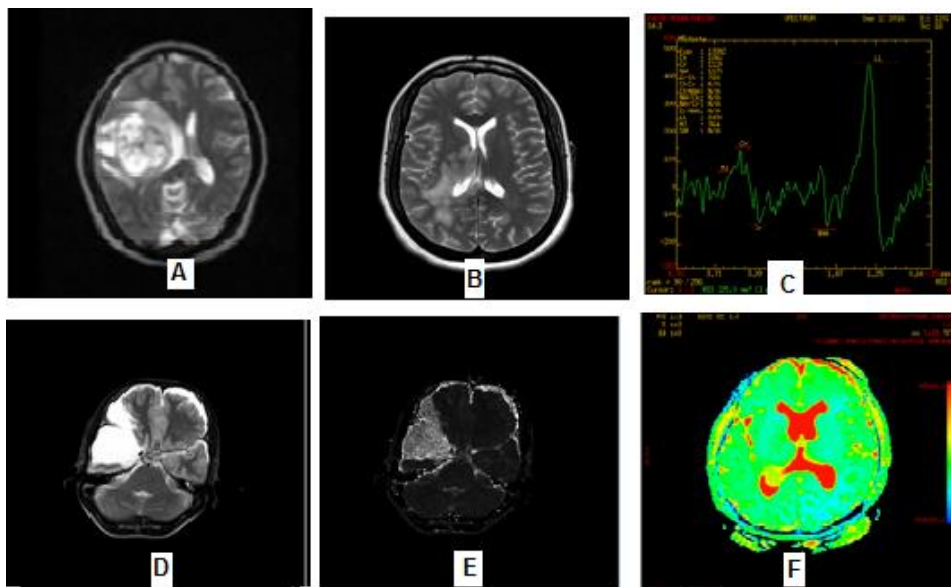


Figure1 A. Glioblastoma B. Multiple Sclerosis (MS), C. MRS-showing metabolites D. Arachnoid Cyst. E. T2 map F. ADC MAP

This classification process of PCA relies on two steps:

- 1) At first step the classifier model is built to discern the predefined group of the image classes [5,6]. The classifier model is created based on a data structure consisting of multiple parts like Independent Numeric Variable of different parameters and Dependent variable of diagnosis of diseases or tissues.
- 2) Second step: The constructed classifier model is used to classify diseases [6].

PRINCIPAL COMPONENT ANALYSIS (PCA)

For data analysis, multiple components or inputs as independent variables were tabulated. Out of them each component is a vector comprising of principal component score derived from each predictor variable of output. The data set may have many variables (31 rows and 12columns) and most of the variables are correlated. Some strategic method or technique to be adopted to reduce the number of variables and to retain some important variables [7].

This PCA technique has the ability to minimize the dimension of the data set in such a way that it becomes effortless to analyze, visualize and interpret. Prediction of the tissue and diseases is also plausible by a statistical procedure that transforms a set of correlated variables or observations into a lesser number of uncorrelated variables of principal components [8].

PCA helps in data compression by **Feature eliminating and Feature Extraction**. The purpose of this method is to diminish dimension of the variables which may not be required for interpretation [9].

Back ground of PCA:

For pattern recognition in the data containing maximum dimension which are identical or dissimilar, PCA has a role to analyze the data set [10]. In images of face recognition and image compression PCA is widely used.

From the different variables, PCA finds a linear combination such that the maximum difference or variance is identified from the variables. During feature eliminating process from the variables there may be a chance of missing the dropped variables which could contribute or could have produced benefit to the prediction [10,11]. In feature extraction, if there is a particular number of independent variables then same number of new independent variables are created and each new independent variable becomes an amalgamation of each of the previously

used “old” independent variables. However, these new independent variables are created in a special order or specific way to predict the dependent variable [10-12].

As an extra benefit, each of the new variables after PCA may be all independent of one another. This is an advantage to fit a linear regression model with these new variables. The *spread* of the data set has different measure or “Variance” [13,14]. In fact it is almost identical to the standard deviation [14].

The formula of variance is depicted here [14]:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n - 1)}$$

This variance is then eliminated and PCA searches for a second linear combination which usually gives details of the greatest magnitude of the remaining variance, and this process repeats almost immediately. PCA is very much significant if there is a data set of a large number of variables with some redundancy in these variables that means some of the variables are correlated with one another [15]. So, a smaller number of principal components (artificial variables) are generated by shrinkage or reducing the observed variables of most of the variance in the observed variables [15].

Multivariate data works with the interaction between numerous random variables. The sets of observations of the random variables are represented by a multivariate data matrix .

Variance is an important parameter to measure *spread* of data of observed values. It is the range to which an allocation is stretched or compressed[16].

The amount of variance is measured by Eigenvalues or the Characteristic roots which explain a given factor or feature measuring a variance or difference of the total sample [16]. Eigenvalues or vectors are characteristic of a matrix. The principal components are orthogonal or right angled to one another, and they are statistically independent of one another [16].

METHODS:

Data collection :

After getting institutional ethics 131 patients of different gender and age (from 9 to 84 years) were studied in a 3 Tesla MR Magnet (SIGNA HDxt, GE, USA). Histo-pathological diagnosis of the materials collected from the Stereotaxic and post surgery biopsies were made and correlated with the following parameters [17-19] :

PARAMETERS

RI VALUES: RI of tissues collected from biopsies of brain materials were determined by Abbe Refractometer (Suprashes Model AAR-33, India) [17-19].

T2 RELAXATION VALUES: T2 mapping was done with the help of multi ECHO read out train (with different echo times 30,60,90,120,150,180ms respectively) in the 3T MR, with a TR of 4000ms. T2 relaxation value of various brain tissue and brain lesions were generated from the map [17]

(Figure1E).

ADC (APPARENT DIFFUSION COEFFICIENT): ADC map was created in the MR magnet and ADC values of the tissues were determined depicting rate of diffusion of water within the tissues mm^2/sec (Figure1F).

METABOLITES QUANTIFICATION OF MR SPECTROSCOPY (MRS): By single and multi voxel Spectroscopy applying PRESS technique, TR- 9602 and TE- 35-144ms quantification of metabolites like CHO,CR,NAA,MI, Lipid, Lactate, CHO NAA,CHO CR and CHO NAA ratio was determined [17-20] (Figure1c).

GROUND TRUTH MR INPUT IMAGE: Therefore a ground truth MR image thus formed, consists of information like RI values (derived from RI mapping), T2 values (from T2 mapping) and ADC values (from ADC mapping) and metabolites from the MRS quantification .All the data were tabulated in Table1 [17].

TABLE 1. Showing all the data collected from the patients and MR magnet:

DISEASE	RI	T2	CHO	ADC	CR	CR/NAA	LIP/LAC	MI	CH/CR	T2peri	CHO/NAA
CSF	1.3333	400	1610	300	1400	0.346	1400	910	1.15	400	0.402
CSF	1.3334	395	1680	320	1800	0.367	1760	1056	1.14	395	0.412
CSF	1.3335	390	1700	330	1967	0.389	1600	1076	1.15	390	0.432
CSF	1.3336	384	1890	340	1989	0.411	1675	1080	1.14	384	0.498
ms	1.3421	340	11750	145	8320	0.557	4160	2912	1.4	240	0.779
ms	1.3439	328	8904	135	2800	0.433	4490	5576	3.15	241	1.39
ms	1.3498	316	7896	124	4560	0.225	3570	3536	1.73	243	0.389
ms	1.3497	304	5947	120	5400	0.7396	6766	4294	1.1	245	0.873
ms	1.3589	249	3448	75	3320	0.7112	5423	2322	1.02	230	0.821
ms	1.3641	245	1610	73	2212	0.941	1440	364	0.495	227	0.465
gmatter	1.3956	130	1601	76	2209	0.938	1441	362	0.491	166	0.461
gmatter	1.3956	125	1601	76	2209	0.938	1441	362	0.491	168	0.461
gmatter	1.3957	123	1589	78	2219	0.941	1467	345	0.491	167	0.459
gmatter	1.3952	121	1458	80	2320	0.878	1443	321	0.494	169	0.456
w matter	1.4251	95	1180	70	2443	0.788	1345	312	0.488	148	0.453
w matter	1.4256	89	1108	71	2435	0.771	1341	320	0.468	146	0.447
w matter	1.4259	85	1098	77	2387	0.774	1211	321	0.467	150	0.445
edema	1.3741	160	1231	84	2216	0.776	1123	325	0.467	246	0.443
edema	1.3823	182	1331	130	2321	0.787	1011	321	0.456	243	0.442
edema	1.3821	182	1298	128	2314	0.781	1009	314	0.454	244	0.441
edema	1.3822	184	1444	131	2310	0.778	1001	313	0.445	245	0.441
GLIOMA	1.4331	90	1443	127	2243	0.766	989	310	0.423	175	0.431
GLIOMA	1.4446	99	1365	177	2254	0.712	917	300	0.343	170	0.341
Gblastma	1.4551	110	2655	156	2112	0.678	900	311	0.311	195	0.332
Gblastma	1.4512	116	2774	142	3280	1.06	2240	312	0.844	190	0.907
Gblastma	1.4562	118	2661	140	3189	1.02	2134	314	0.7881	185	0.89
Gblastma	1.4611	123	1281	139	2998	1.01	2098	316	0.7662	175	0.876
METS	1.4768	135	1321	127	2532	0.654	1011	340	0.432	200	0.432
METS	1.4834	147	1388	139	2211	0.667	1021	341	0.445	219	0.411
METS	1.4911	151	1411	131	2019	0.713	1119	356	0.449	223	0.423

INDEPENDENT VARIABLES AS INPUTS:

RI values, T2 value, ADC value, Quantities of metabolites: (Choline, Creatine, MI, NAA, Lipid/lactate)

Ratio of Choline : NAA, Ratio of Creatine : NAA and Ratio of Cho: Cr.

DEPENDENT VARIABLE: TO LIVE PREDICT (OUTPUT OR DECISION) :

Diseases like MS, Glioma, Glioblastoma (Grade III/IV Astrocytoma), metastasis and tissues like Gray /white matters, CSF are regarded as dependent variables [17].

Principal Component Analysis: Using **XI STAT** (AddinSoft, France) program, PCA (Pearson Type (n)) was run on the data of the table 1. It contains a matrix of 12 X 31 (12 columns and 31 rows). PCA was applied to analyze and decrease the dimensional representation of ground truth images after extracting several features for output or prediction of diseases. The observations and standard deviation of the variables are tabulated in Table 2.

TABLE 2 Showing observations and standard deviation of the variables:

Summary statistics:							
Variable	Observations	Obs, with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
ADC	25	0	25	70.000	340.000	142.600	86.090
CHO	25	0	25	1098.000	0	2784.44	2790.88
CR	25	0	25	1400.000	8320.000	0	1
CH/CR	25	0	25	0.311	3.150	0.824	0.619
CHO/NA	25	0	25	0.332	1.390	0.537	0.239
CR/NAA	25	0	25	0.225	1.060	0.699	0.221
LIP/LAC	25	0	25	900.000	6766.000	2046.52	1562.58
MI	25	0	25	300.000	5576.000	0	4
RI	25	0	25	1.333	1.455	1.384	0.040
T2	25	0	25	36.000	400.000	207.080	121.622

These features or dependent variables were used as input to PCA which in turn determined the correlation between the variables (Table 3).

In classification process, each training data is converted into a vector. The covariance matrix is computed by multiplying several variance or factors by other factors or variance.

TABLE 3: Correlation of Variables of Matrix

Correlation matrix (Pearson (n)):										
Variables	ADC	CHO	CR	CH/CR	CHO/NAA	CR/NAA	LIP/LAC	MI	RI	T2
ADC	1	0.023	0.178	0.296	-0.113	-0.703	-0.085	0.062	-0.485	0.623
CHO	-0.023	1	0.847	0.744	0.640	-0.416	0.737	0.850	-0.406	0.412
CR	-0.178	0.847	1	0.339	0.430	-0.125	0.687	0.586	-0.233	0.212
CH/CR	0.296	0.744	0.339	1	0.738	-0.646	0.625	0.881	-0.602	0.636
CHO/NAA	-0.113	0.640	0.430	0.738	1	-0.041	0.733	0.753	-0.243	0.219
CR/NAA	-0.703	0.416	0.125	-0.646	-0.041	1	-0.247	-0.517	0.643	0.725
LIP/LAC	-0.085	0.737	0.687	0.625	0.733	-0.247	1	0.872	-0.454	0.415
MI	0.062	0.850	0.586	0.881	0.753	-0.517	0.872	1	-0.551	0.542
RI	-0.485	0.406	0.233	-0.602	-0.243	0.643	-0.454	-0.551	1	0.974
T2	0.623	0.412	0.212	0.636	0.219	-0.725	0.415	0.542	-0.974	1

RESULT AND DISCUSSION

The PCA (Pearson N) analysis provides a good model of these data, with 80% of the variance explained in the components. Eigenvalues, Variability % and Cumulative % in relation to factors were extracted and depicted in the Table4.

Table 4. PRINCIPAL COMPONENT ANALYSIS: showing Eigenvalue, variability and cumulative%

Column1	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Eigenvalue	5.560	2.451	0.769	0.576	0.328	0.246	0.036	0.018	0.011	0.005
Variability (%)	55.596	24.514	7.693	5.763	3.281	2.457	0.357	0.181	0.112	0.046
Cumulative %	55.596	80.110	87.803	93.566	96.847	99.304	99.660	99.841	99.954	100.000

Eigenvectors of the variables, factors with correlation among the factors were tabulated in the Table 5 and 6 respectively.

TABLE 5. EIGENVECTORS:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
ADC	0.124	0.532	0.031	0.297	0.707	0.094	0.247	0.078	0.022	0.188
CHO	0.363	0.226	0.258	0.297	0.078	0.352	0.313	0.125	0.647	0.008
CR	0.260	0.322	0.652	0.095	0.289	0.145	0.252	0.096	0.465	0.028
CH/CR	0.383	0.030	0.389	0.221	0.202	0.274	0.007	0.456	0.413	0.397
CHO/NAA	0.290	0.306	0.550	0.058	0.389	0.215	0.395	0.246	0.149	0.284
CR/NAA	0.275	0.395	0.101	0.462	0.342	0.270	0.535	0.099	0.118	0.214
LIP/LAC	0.346	0.256	0.019	0.246	0.161	0.716	0.052	0.401	0.177	0.141
MI	0.398	0.140	0.138	0.069	0.242	0.277	0.516	0.560	0.268	0.105
RI	0.314	0.309	0.115	0.572	0.126	0.179	0.185	0.335	0.183	0.487
T2	0.319	0.363	0.105	0.396	0.014	0.172	0.178	0.323	0.145	0.641

TABLE6: Correlations between variables and factors

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
ADC	0.29 2	0.83 3	- 0.027	0.22 6	0.405	0.047	0.047	0.011	0.002	0.01 3
CHO	0.85 6	0.35 3	- 0.226	0.22 5	- 0.045	-0.175	0.059	-0.017	0.069	0.00 1
CR	0.61 3	0.50 4	- 0.572	0.07 2	0.166	-0.072	-0.048	0.013	-0.049	0.00 2
CH/CR	0.90 3	0.04 7	- 0.341	0.16 8	- 0.116	-0.136	-0.001	-0.061	-0.044	0.02 7
CHO/NA A	0.68 4	0.47 8	- 0.482	0.04 4	0.223	-0.106	-0.075	0.033	0.016	0.01 9
CR/NAA	-	-	-	-	0.196	-0.134	0.101	-0.013	-0.012	-

	0.64	0.61	0.089	0.35						0.01
	9	8		0						5
		-		-						-
LIP/LAC	0.81	0.40		0.18						0.01
	6	1	0.016	7	0.092	0.355	-0.010	-0.054	0.019	0
		-								
MI	0.93	0.22	-	0.05	-					0.00
	8	0	0.121	2	0.139	0.137	0.097	0.075	-0.028	7
	-	-								
RI	0.74	0.48	-	0.43						0.03
	1	4	0.101	4	0.072	0.089	0.035	-0.045	-0.019	3
T2	0.75	0.56		0.30						0.04
	2	8	0.092	1	0.008	-0.085	0.034	-0.043	-0.015	4

As the PCA works, normal brain tissue like gray white matter CSF (Cerebro spinal Fluid) and pathological condition like MS, perilesion edema, low grade glioma, high grade glioma, and metastases can be classified or differentiated out of large data.

A **scree plot** or a simple line segment **plot** is shown (Figure 2) using the fraction of total variance in the data represented by each Principal component or factors fraction explaining the most of cumulative variability and eigenvalues. It represents values in descending order of contribution to total variance.

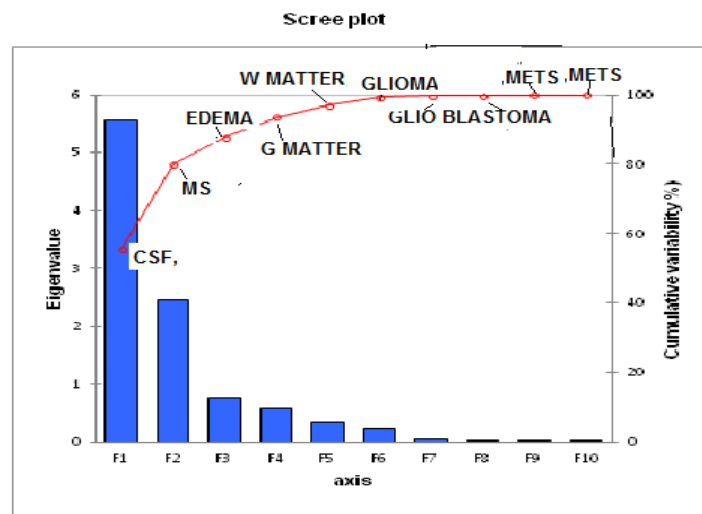


Figure 2. Scree plot derived prediction of diseases utilizing Eigenvalue, cumulative variability % and different factors.

The role of PCA is to check the pattern of relationship among the large number of data of the brain tissues and diseases emphasizing their identical nature or disparity, reducing to one dimension.

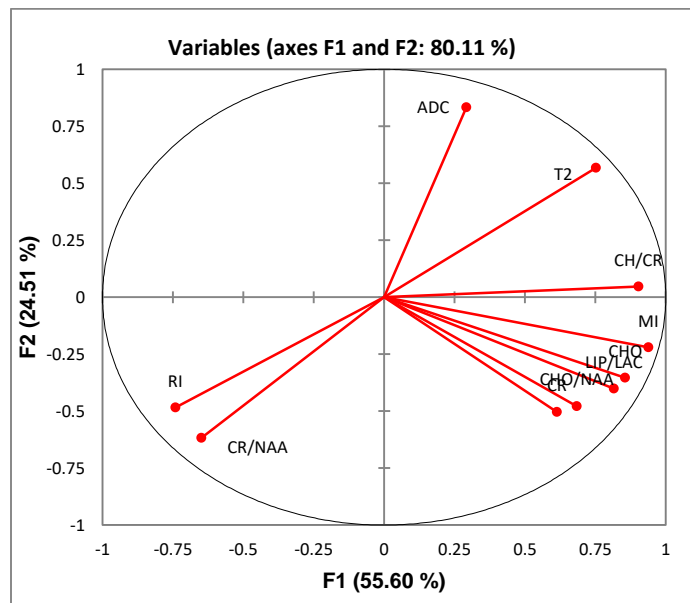


Figure 3. showing variables (Parameters) in Axes F1 and F2 (80.11%) derived from the ground truth images.

By compressing or mapping the data (Figure 3) it converts the large number of dimensions gathered from the ground Truth Image inputs (Table1) into a lower dimensional space. However, the main goal in dimensionality reduction was achieved preserving as much of the significant information as possible.

From the variables shown in Figure 3, a map of the disease or brain tissue can be generated as well (Figure 4). It can be deduced that RI and Cr/NAA ratio can discriminate Glioblastoma III/IV, METS, Glioma, Gray/white matters and perilesional edema much better than other variables. On the other hand ADC value can specify CSF. Metabolites like CH:CR.LIP,LAC and

T2 can detect MS clearly away from other lesions.

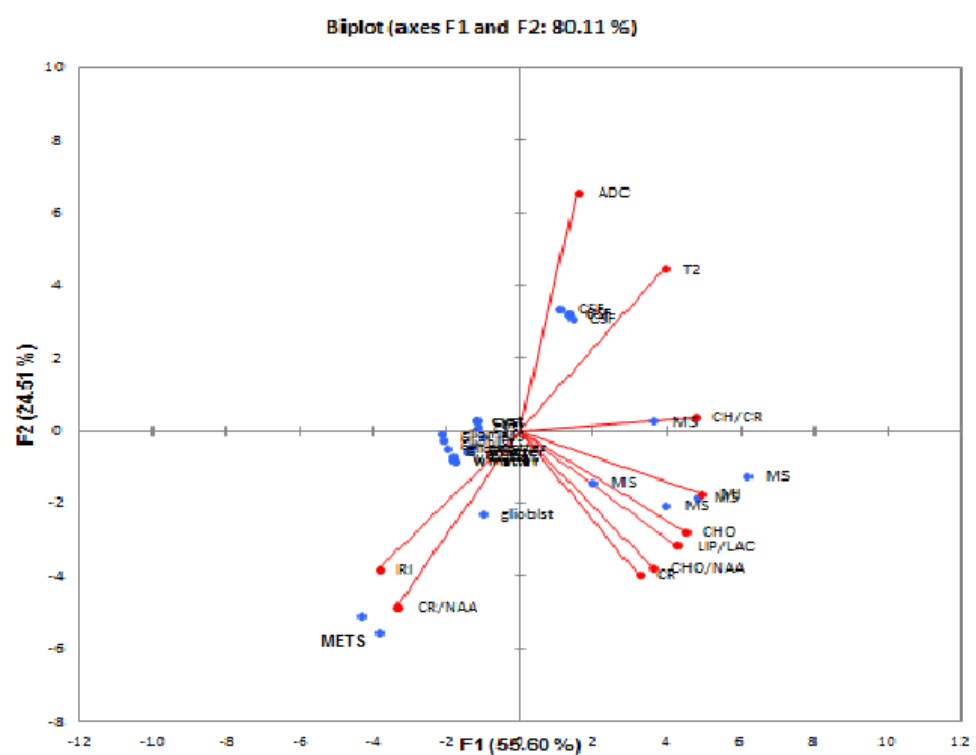


Figure 4. Variables versus disease and tissue discrimination

Further to Figure 4, PCA could extract maps (Figure 5 and 6) of brain tissue and diseases (lesions) from the observations (variables), separating each entity in the space representing F1 and F2 dimension. Glioblastoma Gr II/IV and metastases stand out in the periphery away from the normal tissue of gray and white matters, CSF. Glioma and a fraction MS are placed close to gray/white matters. However Gr I and II Glioblastoma lies in the vicinity of the low grade Glioma.

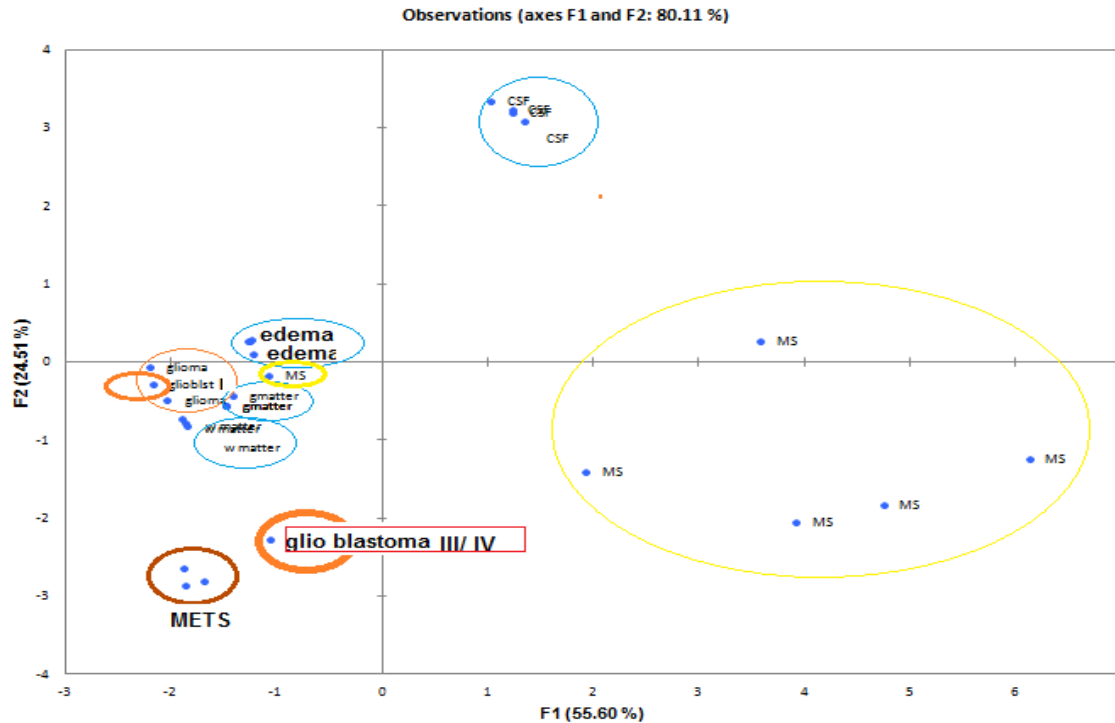


Figure 5. Mapping of the normal brain tissues and lesions by PCA utilizing observations F1 and F2 (80.11%) are shown.

Conclusion: PCA an important statistical device can reduce the number and dimension of complex variables like the parameters extracted from the ground truth MR images and produce a linear relationship in a simple way. Mapping of tissue and diseases can be generated from the depleted variables losing their complex dimensionality.

Thus PCA helps discriminating different disease process and brain tumors. RI is found to be superior to all other parameters (like T2 values, ADC values and important metabolites and their ratio) in differentiating diseases.

REFERENCES:

- [1] Taghpour Zahir SH, Rezaei sadrabadi Dehghani F, Evaluation of Diagnostic Value of CT Scan and MRI in Brain Tumors and Comparison with Biopsy, Iranian Journal of Pediatric Hematology Oncology 2011 ;1.(4):121-125
- [2] Hagen T, Nieder C, Moringlane JR. Feiden W, Konig J, Correlation of preoperative neuroradiologic with postoperative histological diagnosis in pathological intracranial process. Der Radiologe, Nov 1995; 35(11):808-15
- [3] Horská Alena and Barker Peter B., Imaging of Brain Tumors: MR Spectroscopy and Metabolic Imaging, Neuroimaging Clin N Am. 2010 ; 20(3): 293–310.
- [4] Jansen JF, Backes WH, Nicolay K, Kooi ME. 1H MR spectroscopy of the brain: absolute quantification of metabolites. Radiology 2006; 240 (2): 318–32.
- [5] H. Mazien, Proposed system for the diagnosis of skin diseases using Multiwavelet Transform and Decision Tree, M.Sc. Thesis, University of Technology, Department of Computer Science, 2015.
- [6] Jameela Ali, et al., "Red Blood Cell Recognition using Geometrical Features," International Journal of Computer Science Issues 2013 vol. 10, no. 1.
- [7] S. Chandrasiri, et al., Morphology Based Automatic Disease Analysis Through Evaluation of Red Blood Cells, in Department of Information Technology/ Sri Lanka Institute of Information Technology, Fifth International Conference on Intelligent Systems, Modelling and Simulation 2014.
- [8] Mohammed Hussein Miry, Akel A. Alzaiez, Abbas Hussein Miry, Image Authentication Using PCA And BP Neural Network, Eng. & Tech. Journal, 2010; 28,(22):6536-6545.
- [9] K. Hosny, Exact and fast computation of geometric moments for gray level images, Applied Mathematics and Computation Journal, 2007 vol. 189 :. 1214 1222.
- [10] D. N. George, Tumor Type Recognition Using Artificial Neural Networks, M.Sc. thesis ,2013, Iraqi Commission for Computers And Informatics Institute for Postgraduate Studies.
- [11] Tarun Jhaldiyal, Pawan Kumar Mishra : Analysis and Prediction of Diabetes Mellitus Using PCA, REP and SVM, International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-2, Issue-8, August 2014.

- [12] Vanishri Arun, Arunkumar B.V, Padma S.K. and Shyam V, Disease Classification and Prediction using Principal Component Analysis and Ensemble Classification Framework,2017;10(14):107-116.
- [13] Ian T. Jolliffe, Jorge Cadima: Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society, A Mathematical, Physical and Engineering Sciences, 2016; 374: 206514.
- [14] Lindsay.J.Smith A tutorial on Principal Components Analysis 2002, www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- [15] Jackson JE. A User's Guide to Principal Components. New York: John Wiley & Sons; 1991.
- [16] Peres-Neto PR, Jackson DA, Somers KM. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal* 2005, 49:974–997.
- [17] T K Biswas, R Bandopadhyay, A Dutta, Validating The Discriminating Efficacy Of MR T2 Relaxation Value Of Different Brain Lesions And Comparison With Other Differentiating Factors: Use Of Artificial Neural Network And Principal Component Analysis. *The Internet Journal of Radiology*. 2017 Volume 20 Number 1. ISPUB DOI: 10.5580/IJRA.52614
- [18] Biswas TK, Gupta A. Retrieval of true color of the internal organ of CT images and attempt to tissue characterization by refractive index : Initial experience. *Indian Journal of Radiology and Imaging* 2002;12:169-178
- [19] Biswas TK, Luu T In vivo MR Measurement of Refractive Index, Relative Water Content and T2 Relaxation time of Various Brain lesions With Clinical Application to Discriminate Brain Lesions. *The Internet Journal of Radiology* 2009;13 (1).
- [20] T K Biswas, S R Choudhury, A Ganguly, R Bandopadhyay, A Dutta, Refractive Index As Surrogate Biological Marker Of Tumefactive And Other Form Of Multiple Sclerosis And Its Superiority Over Other Methods, *Internet Journal of Radiology*, <https://print.ispub.com/api/0/ispub-article/46167>.