Development of an expert tool using Principal Component Analysis (PCA) approach: for identifying human diseases & its severity

P.Santosh Kumar Patra
Assistant Professor in ISE,
SaIT, Bangalore, India

Dr.Seetharam K
Professor in CSE, MeRITS,
Andhra Pradesh, India

Dipti Prava Sahu,
Assistant Professor in CSE,
RLJIT, Bangalore, India

ABSTRACT
Detecting diseases at early stage can enable to overcome and treat them appropriately. Identifying the treatment accurately depends on the method that is used in diagnosing the diseases. An expert tool can help a great deal in identifying those diseases and describing methods of treatment to be carried out taking into account the user capability in order to deal and interact with expert tool, easily and clearly. Present expert system uses inference rules and plays an important role that will provide certain methods of diagnosis for treatment.
In this paper to analyze a disease we consider three important factors. One analysis with age factor, one analysis with pregnancy factor (in case of a women), and one analysis with duration factor (in case of chronically illness) finally considering all the three factor association in the diagnosis of a human diseases[4]. The analysis of variables is to identify the dimension that is latent, it means finding the severity of the diseases and explaining the appropriate treatment as per Indian Pharmacopoeia Standards . That can be considered in the phenomena of performance correlation. This is to study the effects in the developed principal components analysis (PCA) approach.

Keywords
Expert System, Fuzzy logic, Eigen value, Eigen vector, PCA.

INTRODUCTION
Computer-based methods are increasingly used to improve the quality of medical services. Artificial Intelligence (AI) is the area of computer science focusing on creating expert machines that can engage on behaviors that humans consider intelligent[1]. The proposed tool is for dealing with the problem of a disease diagnosis using an expert tool. An expert system uses expert tool in the system that employs human knowledge captured in a computer to solve problems[9]. that ordinarily require human expertise [2]. Expert system seeks and utilizes relevant information from their human users and from available knowledge bases in order to make recommendations [3].

2.1 Logical concept:
Using PCA covariance matrix: The data and knowledge of to be used in the system are collected from different sources[11]. The first primary source is the medical knowledge of expert doctors. The second source is from specialized databases, books and a few electronic websites data used for analysis [6]. Three different groups are considered. The predictor variables are

Determining Factors:

x[1]: Age. S1, S2, S3, S4 & S5
x[2]: Pregnancy status S1, S2, S3, S4 & S5
x[3]: Chronically ill S1, S2, S3, S4 & S5

Where S1= 1-20%, S2= 17-40%, S3= 35-60%, S4= 55-80%, S5= 75-100%.

Five different levels are identified separately by means of fuzzy logic application for each variable. Four analysis is carried out. For each analysis, fuzzy values are used for respective variables[11]. The variables are standardized using Z distribution principles. Covariance matrix is generated[7]. And from the above five different levels will able to identify the severity with the five different levels like Very Low, Low, Medium, High, Very High and can suggest require treatment with medicine[12].

Table A.1 **Age, S1, S2, S3, S4 & S5:**
Results are: Principle components analysis matrix coefficients

Age	S1	S2	S3	S4	S5
-0.5279	0.0958	-0.1165	0.297	0.7622	-0.171
0.2747	0.5417	-0.6055	0.1781	-0.1434	-0.46
-0.4709	-0.2011	-0.3091	0.5614	-0.5075	0.26
-0.3849	-0.4036	-0.3846	-0.6081	-0.1263	-0.39
0.3303	-0.5986	0.1609	0.4414	0.0313	-0.55
0.4086	-0.3684	-0.592	-0.0186	0.3521	0.47

Rows correspond to observations, columns to components Hotelling's T-squared statistic for each observation[12]

TableA.2 Principal component **Scores**

Age	**s1**	**S2**	**S3**	**S4**	**s5**	**Table A3 TSquare**
-0.4957	-0.1906	0.1258	0.407	-0.1597	0.0442	6.023
-0.7252	0.2425	0.401	-0.0387	0.0713	-0.0191	3.975
1.0191	-0.3603	-0.3267	0.0146	0.064	0.0359	4.5165
0.6961	-0.2878	0.3406	-0.1125	-0.0707	0.498	4.5897
0.1143	-0.0284	0.0531	0.2399	0.2705	-0.0764	5.7638
0.3943	-0.4317	-0.0438	-0.2134	0.147	-0.0398	4.0384
0.6783	0.5618	0.0205	-0.2019	0.1407	-0.048	5.5745
0.0453	0.4413	-0.3056	0.1796	0.0411	0.0132	4.0682
-2.5055	-0.1157	-0.2029	-0.1938	-0.0496	0.0087	7.4958
0.7789	0.1688	-0.062	-0.0809	-0.4546	-0.0644	7.955

The Eigen values of the covariance matrix in latent & Percentage of variance ***Table A4.***

Component	Eigen Value	% Variance
1.Age	2.687	44.7836
2.S1	1.5578	25.9637
3.S2	0.8341	13.9019
4.S3	0.5756	9.5939
5.S4	0.2984	4.9733
6.S5	0.047	0.7837

Residuals obtained by retaining the principal components by the 10-by-6 data matrix. Rows correspond to observations, columns to variables.
Table A5: Principle component Residuals:

Age	S1	S2	S3	S4	S5
0.0016	0.0178	-0.0074	0.0115	0.0338	-0.0173
-0.0007	-0.0077	0.0032	-0.005	-0.0146	0.0075
0.0013	0.0144	-0.006	0.0093	0.0275	-0.0141
0.0018	0.02	-0.0084	0.0129	0.0381	-0.0195
-0.0027	-0.0307	0.0128	-0.0199	-0.0585	0.03
-0.0014	-0.016	0.0067	-0.0103	-0.0305	0.0156
0.0017	0.0193	-0.0081	0.0125	0.0368	-0.0188
0.0005	0.0053	-0.0022	0.0034	0.0101	-0.0052

0.0003	0.0035	-0.0015	0.0023	0.0066	-0.0034
-0.0023	-0.0259	0.0108	-0.0167	-0.0494	0.0253

2.2 Pregnancy as one main component, S1, S2, S3, S4, S5

Principle components analysis matrix with Duration

Table B.1 Principal Component coefficients

Pregnancy	S1	S2	S3	S4	S5
-0.2745	-0.9616	0	0	0	0
0.43	-0.1228	-0.0018	0.8644	-0.227	0.0369
0.43	-0.1228	-0.6074	-0.31	-0.3672	-0.4473
-0.43	0.1228	0.2083	0.245	-0.0846	-0.8303
0.43	-0.1228	0.7648	-0.311	-0.3268	-0.1075
0.43	-0.1228	0.0527	0.0017	0.8365	-0.3124

Principal component **Scores.** Rows correspond to observations, columns to components. Hotelling's T-squared statistic each observation

Table B.2Principal component scores Hotelling's T-squared statistic

Pregnancy	S1	S2	S3	S4	S5
-1.2026	0.18 78	0	0	0	0
0.7282	0.348	0	0	0	0
-1.2454	0.0343	0	0	0	0
-0.8176	0.012	0	0	0	0
-0.2737	0.0039	0	0	0	0
-0.6599	-0.1172	0	0	0	0
0.6167	-0.0364	0	0	0	0
1.1348	-0.0477	0	0	0	0
1.9279	-0.0256	0	0	0	0
-0.2085	-0.3591	0	0	0	0

Table B.3

T-squared
5.0171
4.9187
6.4120
8.1000
4.3345
4.4595
3.1064
6.1498
6.2950
5.2070

Table B.4 The eigen values of the covariance matrix in latent & Percentage of variance, Principle component Latent

Component	Eigen Value	% Variance
1. Pregnancy	5.355	89.25
2.S1	0.645	10.75
3.S2	0	0
4.S3	0	0
5.S4	0	0
6.S5	0	0

Table B.5 Residuals of principal components 1.0e-015 *

Pregnancy	S1	S2	S3	S4	S5
0	0.222	0	-0.0555	0.0069	0.0416
0.222	0.2776	0.0833	-0.1665	0.0347	0.111
-0.4441	0.111	0.0278	-0.0555	0	0.0278
-0.3331	0.0555	0	0	0	0.0278
-0.0555	0	0	0	0	0
-0.222	-0.111	-0.0278	0.0555	-0.0139	-0.0278
0	0	0	0.0555	0	0
0.4441	-0.111	0	0	0	0

0.6661	-0.111	0	0.111	-0.0139	0
-0.111	-0.222	-0.0555	0.111	-0.0278	-0.055

Chronically ill, S1, S2, S3, S4 & S5

Table C.1.Principal component coefficients

Chronically ill	S1	S2	S3	S4	S5
-0.2647	-0.9643	0	0	0	0
0.4313	-0.1184	-0.4359	0.6951	0.2997	-0.1924
0.4313	-0.1184	0.1103	-0.3634	-0.1723	-0.7913
-0.4313	0.1184	0.0756	-0.1132	0.8049	-0.3656
0.4313	-0.1184	0.7971	0.147	0.2923	0.2311
0.4313	-0.1184	-0.3959	-0.5919	0.3782	0.387

Table C2 Principle component LATENT Values

Components	Eigen values	% Variance
1.Chronically ill	5.326	88.7659
2.S1	0.674	11.2341
3.S2	0	0
4.S3	0	0
5.S4	0	0
6.S5	0	0

Table C3. Principal component scores Hotelling's T-squared statistic

Table C.4

Chronically ill	S1	S2	S3	S4	S5		TSQUARED
-0.3711	0.2905	0	0	0	0		5.4337
-0.2147	0.2346	0	0	0	0		2.3666
-1.0217	0.0668	0	0	0	0		4.7236
-0.7907	0.0195	0	0	0	0		5.6540
-0.4862	-0.0193	0	0	0	0		3.4211
-0.4777	-0.0925	0	0	0	0		4.0184
0.6423	-0.0367	0	0	0	0		5.6540
0.4283	-0.1356	0	0	0	0		7.3144
2.6609	0.0492	0	0	0	0		7.3144
-0.3693	-0.3764	0	0	0	0		8.1000

Table C5. Principal Components Analysis Residuals Residuals obtained by retaining the principal components by the 10-by-5 data matrix . Rows correspond to observations, columns to components.

	Chronically ill	S1	S2	S3	S4	S5
1.0e-015 *	-0.222	-0.111	-0.0555	0.111	-0.0139	-0.0555
	0	-0.111	-0.0555	0.111	-0.0139	0
	0.222	0	0	0	0	0
	0.111	0	0	0	0	0
	0.0555	0	0	0	0	0
	-0.111	0	0	0	0	0
	-0.0555	0	0	0	0	0.0555

	0	0	0	0	0	0
	..0555	0.222	0	0	0	0.0555

Table D.1.Principal component coefficients

Age	Pregnancy	Chronically ill	S1	S2	S3	S4	S5
-0.4003	0.2739	-0.1376	0.247	0.5664	-0.1104	0.0065	-0.5931
-0.4427	-0.2665	0.0121	-0.0876	0.1556	-0.5047	-0.5544	0.3729
-0.456	-0.0246	0.1862	-0.0779	0.3531	0.6062	0.2408	0.4478
0.3316	0.3951	-0.4996	0.573	0.2314	0.3074	-0.517	0.2565
-0.3478	0.0377	-0.4439	0.6372	-0.4315	-0.0304	0.2014	0.2164
-0.3526	-0.2032	-0.4513	-0.5392	-0.3284	0.3012	-0.1452	-0.347
0.1387	-0.6521	0.1304	0.462	0.0688	0.3739	-0.3503	-0.2405
0.2476	-0.4789	-0.5282	-0.1088	0.4218	-0.1993	0.4278	0.1336

Table D2 Principle component LATENT Values

Component	Latent Eigen values	%Variance
1.Age	4.203	52.5379
2.Pregnancy	1.7861	22.3258
3.Chronically ill	0.8713	10.8908
4.S1	0.5867	7.3332
5.S2	0.4132	5.165
6.S3	0.0936	1.1699
7.S4	0.0281	0.3514
8.S5	0.0181	0.2259

Table D3 Principal component Score

Age	Pregnancy	Chronically ill	S1	S2	S3	S4	S5
0.6328	-0.1663	0.0171	-0.4219	-0.1445	-0.1213	0.013	-0.0165
-0.7315	-0.0109	0.8922	0.1672	-0.0755	0.0367	0.0723	-0.0247
1.8902	-0.1439	-0.3564	0.0179	0.2819	0.0876	0.0548	-0.0493
1.3449	-0.1335	0.0171	0.264	-0.2435	-0.0163	-0.127	-0.0333
0.5204	-0.1267	0.2723	-0.0475	0.2422	-0.1798	-0.0422	0.0784
0.9154	-0.2636	-0.218	0.3492	0.1283	0.012	0.0368	0.0304
-0.4061	1.0985	-0.3049	-0.0181	-0.1936	-0.3249	0.0475	-0.0076
-0.9001	0.8683	0.1598	-0.2504	0.2824	0.1051	-0.0653	-0.0305
-4.0697	-0.5775	-0.3602	0.0982	0.0028	0.0651	-0.0108	0.0023
0.8036	0.4556	-0.119	-0.1587	-0.2805	0.3358	0.0209	0.0506

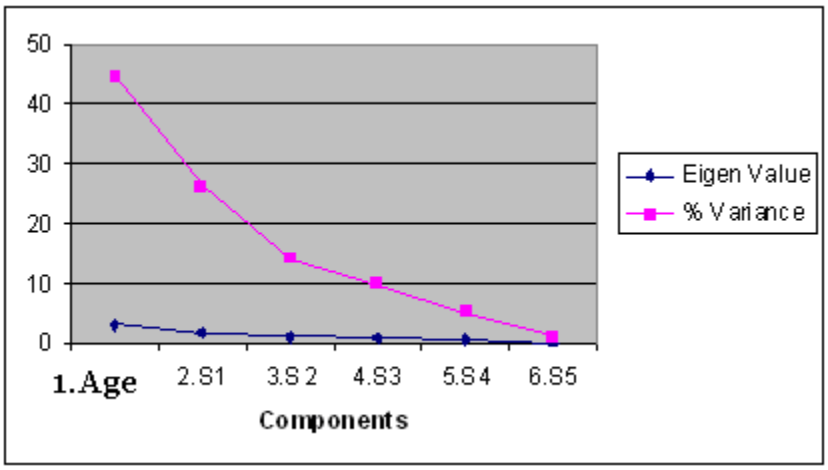
Table D.4 Hotelling's T-squared statistic

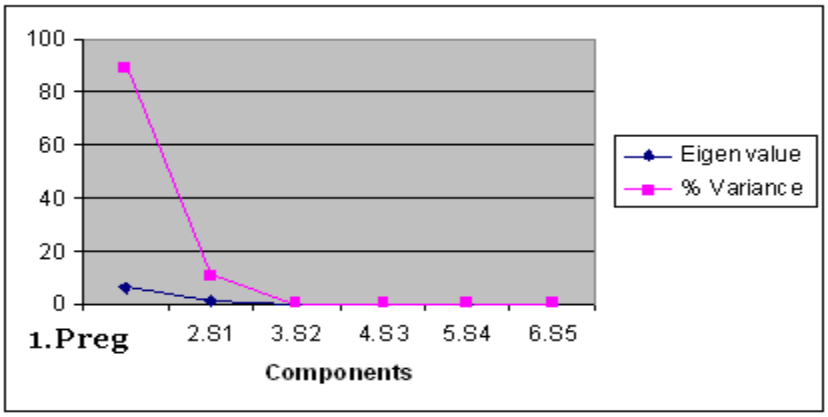
TSQUARE
7.5236
8.0496
6.2913
8.0111
7.0339
4.2299
8.0949

6.9429
7.8061
8.0166

Table D5. Principal Components Analysis Residuals obtained by retaining the principal components by the 10-by-8 data matrix . Rows correspond to observations, columns to components.

1.0e-014 *	Age	Pregnancy	Chronically ill	S1	S2	S3	S4	S5
	0.1665	0.0222	-0.0555	-0.0444	-0.025	0.0128	0.0222	0.0278
	-0.0333	-0.0555	0.0472	-0.0132	-0.0694	-0.0167	0.0035	-0.0555
	0.0111	0.0666	0.0666	0.0069	0.0333	0.0222	0.0049	0.0444
	0.333	0.111	0.0555	-0.0222	0.022	-0.028	0.0035	-0.0389
	-0.139	0.167	0.0444	0.0139	-0.0305	0.025	-0.009	0.0028
	0	0.0999	0.0333	-0.0028	0.0167	0	-0.0007	0.0187
	0.0555	-0.0555	-0.0999	0	0.0028	-0.0167	0.0111	-0.0222
	0.0638	-0.0888	-0.0222	-0.0056	-0.0194	-0.0056	0.0007	0.0194
	-0.1776	-0.111	-0.222	-0.0028	-0.0028	-0.0056	0.0333	0.05
	0.0444	0.0555	0.0278	0.0069	0.0389	-0.0167	0.0028	-0.0097





Details of factors:

Severity of the diseases:

Normally, determinant factors for a disease are age of the patient, pregnancy (in case of women) and duration of diseases (in case of chronic illness). And after imposing the Fuzzy values the severity levels along with the suggested treatment are:
Very low- Take care your self, One check up, Dose as per requirement.
Low- Minimum precaution, Normal dose.
Medium- Precaution, Dose as per sign & symptoms.
High-Strict Precaution, Dose as per IP, Regular check up.
Very High- Immediate treatment, High alert observation, Reports should be collected periodically, Treatment followed by exact doses of medicine[4].

CONCLUSION

An expert tool can be used in consultation since it shows quickly the diagnosis and in addition, it offers explanations of the obtained results, being very helpful to the professional. With the expert system, the user can interact with a computer to solve a certain problem. Use of expert tool can help in diagnosis ,i.e. identifying the human diseases and an appropriate treatment to it[6].
The proposed system performs many functions. It will conclude the diagnosis based on answers of the user to specific question that the system asks the user[9]. The questions provide the system for explanation for the symptoms of the patient that helps the expert system for diagnosis the disease by inference engine[8]. It stores the facts and the conclusion of the inference of the system, and the user, for each case, in database. It processes the database in order to extract rules, which completes the knowledge base[7].

Several properties of this model remain to be investigated. It should be tested on several more databases. Unfortunately databases are typically proprietary and difficult to obtain. Future prospects for medical databases should be good since some hospitals are now using computerized record systems instead of traditional paper-based[10]. It should be fairly easy to generate data for machine diagnosis.

One important aspect of automated diagnosis is the accompanying explanation for the conclusion, a factor that is important for user acceptance. A trained expert would evaluate the quality of the diagnosis performed by the system, followed by adjustment of the utilities.

REFERENCES

[1] Russell, S. and P. Norvig, 2002. Artificial Intelligence: A Modern Approach, Prentice Hall, Second Edition.
[2] Beverly G. Hope, Rosewary H. Wild, « AnExpertSupport System for Service Quality Improvement», Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Science, 1994.

[3] Azaab S., Abu Naser S., and Sulisel O.,2000. A proposed expert system for selecting exploratory factor analysis procedures, Journal of the college of education, 4(2):9-26.
[4] Knowledge Representation For The Nursing Diagnosis By Means Of An Expert System by M. Lourdes Jiménez, José M. Santamaría, L.González1. Á.L. Asenjo, L.M. Laita, M. Beamud
[5] Analysis and design of information systems by V.Rajaraman, 5th print, PHI, pp 113-137
[6] An expert diagnostic tool for engineering systems by A K Verma & K Seetharam (journal of scientific & Industrial research, vol 53, pp 601- 603)
[7]Discrete mathematical structures with applications to computer science by J.P.Tremblay & R. manohar, TMH, pp 8-19.
[8] Conte, S.,Dunsmore, H.and Shen,V, Software Engineering Metrics and models, Benjamin Cummings, Menlo Park CA, 1986.
[9] Guanrong Chen and etal. Introduction to fuzzy sets, fuzzy logic and fuzzy control systems. CRC Press, 2000.
[10] E.M.Hall Managing risk: Methods for software system Development Addson-Wesley;1998.
[11] Katrina, D.M. Applied statistics for software managers.Prentice Hall PTR, 2002.
[12] Daniel T Larose Data Mining: methods and models. WIley interscience 2006.