Computers in Biology and Medicine 35 (2005) 475–493

Computers in Biology and Medicine

http://www.intl.elsevierhealth.com/journals/cobm

An algorithmic approach to understand trace elemental homeostasis in serum samples of Parkinson disease

M.B. Sanjay Pande[a], P. Nagabhushan[a,*], Murlidhar L. Hegde[b], T.S. Sathyanarayana Rao[c], K.S. Jagannatha Rao[b]

[a] *Department of Studies in Computer Science, Manasa Gangothri, University of Mysore, Mysore 570 006, India*
[b] *Department of Biochemistry and Nutrition, Central Food Technological Research Institute, Mysore 570 013, India*
[c] *Department of Psychiatry, JSS Medical College Hospital, Mysore 570 004, India*

Received 17 October 2003; accepted 5 April 2004

Abstract

A classical problem in neurological disorders is to understand the progression of disorder and define the trace elements (metals) which play a role in deviating a sample from normal to an abnormal state, which implies the need to create a reference knowledge base (KB) employing the control samples drawn from normal/healthy set in the context of the said neurological disorder, and in sequel to analytically understand the deviations in the cases of disorders/abnormalities/unhealthy samples. Hence building up a computational model involves mining the healthy control samples to create a suitable reference KB and designing an algorithm for estimating the deviation in case of unhealthy samples. This leads to realizing an algorithmic cognition–recognition model, where the cognition stage establishes a reference model of a normal/healthy class and the recognition stage involves discriminating whether a given test sample belongs to a normal class or not. Further if the sample belongs to a specified reference base (normal) then the requirement is to understand how strong the affiliation is, and if otherwise (abnormal) how far away the sample is from the said reference base. In this paper, an exploratory data analysis based model is proposed to carry out such estimation analysis by designing distribution and parametric models for the reference base. Further, the knowledge of the reference base in case of the distribution model is expressed in terms of zones with each zone carrying a weightage factor. Different distance measures are utilized for the subsequent affiliation analysis (City block with distribution model and Doyle's with Parametric model). Results of an experimental study based on the database of trace elemental analysis in human serum samples from control and Parkinson's neurological disorder are presented to corroborate the performance of the computational algorithm.

© 2004 Published by Elsevier Ltd.

[*] Corresponding author. Tel./fax: +91-821-510789.
E-mail address: pnagabhushan@hotmail.com (P. Nagabhushan).

0010-4825/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.compbiomed.2004.04.003

Keywords: Parkinson's disorder; Trace elemental knowledge base; Parametric model; Doyle's distance computations; Distribution model; Zonalization; Weightage factors; Affiliation analysis

1. Introduction

Parkinson's disease (PD) is a complex and progressive neurodegenerative disorder that affects the control of body movement. The degeneration occurs predominantly in dopaminergic neurons in a small brain area called the substantia nigra [1], though the cause of this selective degeneration is still obscure. Epidemiological investigations suggest that both environmental factors such as metals, pesticides, etc. and genetic factors could predispose people to PD. Inter-relationships among the trace metals play an important role in normal and pathological state of a cell. Limited data is available concerning the levels of metals in serum during pathological conditions of PD. Moreover, most of the available information is limited to few selected elements [2–4] and there is no study, which examines inter-elemental relationships with regard to severity of PD, which could have clinical and diagnostic significance. Since several elements interact metabolically, the understanding of the concurrent metal levels and their inter-relationship pattern is very essential for clinical correlation [5]. Further, more than an individual metal, the comprehensive metal homeostasis and its inter-relationships with other elements are known to play a significant role in the biological system. Our recent investigations on trace elemental concentrations in PD serum first time evidenced a clear imbalance in certain metal levels and element-to-element interrelationships or homeostasis in early and severe PD patients compared to control [6]. Based on this novel observation we feel that mapping of trace metal homeostasis in serum samples could be of diagnostic importance. Keeping this fact in view, we have proposed in this research communication that an algorithmic model of the comprehensive database on serum elemental homeostasis with the progression of PD may have diagnostic applications. In the present paper, an exploratory data analysis based computation model has been developed to carry out such analysis through distribution and parametric models representing reference knowledge bases (KB) on healthy control samples in the context of PD.

Computational models provide a better means of modeling complex systems (such as the nervous system, neuro-disorders, etc.). The development of computational model can be made richer and more robust by consideration of clinical data, just as pathology has always enriched the understanding of physiology. There are several compelling reasons for attempting to develop models and simulation of these diseases not as a replacement for other approaches, but as a means to gain new insight into pathogenesis and treatment. In some sense, the function of the brain is computation, thus the underlying computational processes must be uncovered to understand the basis of disease. To achieve therapeutic goals, many separate studies are required, from the first step in translating basic research advances, animal testing, preliminary safety studies in human patients, and clinically large trials with the purpose of evaluating the effectiveness of a therapy. Biochemical studies of Parkinsonism have had the most profound impact on understanding and management of chronic neurological disorder (PD). For biochemical study to be meaningful there must be a high degree of sensitivity, specificity and a strong correlation with the clinical picture is necessary to ensure that biochemical abnormalities reflect the clinical findings. To arrive at a good algorithmic approach for analysis, we need to design the KB and ensure accuracy in the classification of data samples. Creation of a reference

KB implies consolidating a huge database of control/training set into knowledge parameters, which provide meaningful and useful comprehension for later analysis. The problem of assigning a test sample into one of the two classes as healthy or unhealthy, demands the creation of a reference KB of only healthy samples in the context of PD. It should be understood that a reference base will contain the relevant knowledge summaries mined from normal/control samples, since the reference base for various varieties of abnormalities can neither created nor will it be a feasible approach. In this sense it is a specific case of knowledge mining [7] from the databases of normal samples particularly in the context of PD. (The group of samples that support the creation of the reference KB in one context may not help the creation of the reference KB in some other context.)

Often real-world data sets are predominantly composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. Data are a set of facts (e.g., cases in database) and pattern is an expression in some language describing a subset of the data or a model applicable to the subset [8]. The pattern should be valid on new data with some degree of certainty, if the new data belongs to the subclass, else the pattern model should efficiently refute the data. The anticipated severity or exhaustiveness in the distance computations, can perhaps be reduced at least in a specific instance, but often encountered in situations such as the sample which needs to be identified either as belonging to healthy (desired) group or otherwise (in so defined two class problem), provided the healthy class is properly modeled to represent the desired reference base. The creation of such reference KB for healthy class is the phase of cognition or learning [9]. The next phase is the recognition which involves labeling a given sample as healthy or unhealthy, which is technically defined as classification [9,10] in cognition–recognition literature.

In a crude way, classification means assigning an unknown sample to a known label. The unknown sample is compared with KB. Estimation of the sample distribution is based on a training set whose classification is known beforehand (e.g., assigned by human experts). Classifying data into normal (Healthy) or abnormal (Unhealthy), depicts the closeness or distinctness (Farness/Nearness) of features in the samples. There are different distance measures available which deal these problems with different concepts. But there is no single general distance measure [11], which can be applied on all types of data sets. If a classifier is efficiently designed it will be able to perform well on new patterns. In preprocessing stage, we first identify the relevant features and then use a feature extractor to measure them. These measurements are then passed to a classifier, which performs the classification [12]. The importance of 2-way classifier, such as recognizing a sample with a healthy sample set or otherwise can be very much appreciated in the field of medical diagnosis. A physician has to label a patient at the first level as healthy/unhealthy. Generally the diagnosis is based on the features collected by various tests. The complexity of the problem is because of more and more data generated through various pathological/clinical tests, and the features could also exhibit overlapping features making diagnosis a difficult problem. If a person is to be declared unhealthy then the confidence level from feature-wise analysis also should be high, where a feature is an observation from a diagnosis test. Further in case a person is unhealthy, the physician has to estimate the degree of ill-health too. Keeping this in view such type of typical 2-way discrimination problems, some simple but effective models are proposed in this paper.

This research communication is organized in the following way: Section 2 describes the clinical methodology followed to collect the control and PD affected samples and two different computational models designed for analysis, the role of ratio features as a means to enhance the performance analysis is studied. Section 3 provides a detailed picture of the computational models. Section 4 includes some useful discussions and recommendations. Section 5 provides the summary.

2. Methodology

2.1. Patients

Blood samples from 25 each controls (normal/healthy samples) early PD and severe PD patients were collected from Sri Venkateswara Institute of Medical Science Tirupathi, India, and JSS Medical College and Hospital Mysore, India. The PD patient group was graded into early PD and severe PD according to clinical severity. All the patients met the commonly accepted diagnostic criteria for PD [13] and were evaluated by the Unified Parkinson's Disease Rating Scale (UPDRS) [14] and the Hoen and Yahr staging [15]. The first stage of the Hoen and Yahr staging of PD was considered as Early PD while the latter stages of Hoen and Yahr staging were graded into Severe PD. A 10 ml volume of venous blood sample was collected from each PD patient/control and serum was separated by centrifugation. All the precautions were taken to eliminate metal contamination while collecting , storage and analysis of the samples in accordance with NCCLS criteria [16].

2.2. Ethical issue

Ethical approval for collecting blood samples from patients with PD and control humans were obtained from research ethical committee of JSS Medical College and Hospital and Sri Venkateswara Institute of Medical Science, India. A written consent was obtained from the patients/caretakers prior to the collection of blood samples.

2.3. Instrumentation and elemental analysis

Elemental analysis was carried out using Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES) either by sequential or simultaneous mode depending on the elements to be analyzed. The optimization of ICP-AES was evaluated by line selection and detection limits for each element. The validation of the analysis was tested by analyzing serum matrix match multi-element synthetic standard and certified standard reference material (Bovine liver 1577a) obtained from the National Bureau of Standards, USA [17]. The lines were selected for each element in such a way that interference from other elements were minimized. Table 1 provides the summary of laboratory observations made on 25 healthy samples, 25 early PD affected and 25 severely affected samples. The concentration of trace elements (identified as features for algorithmic purpose) are presented in mean ± standard deviation format. The second row gives the range of concentration values for each trace element in minimum , maximum format. A perusal through the table indicates that it is in practice just difficult to discriminate the healthy and affected samples. This is the motivation to devise algorithmic model for the purpose.

Table 1
Data base containing trace—element concentrations

	Na	S	P	Ca	Mg	Cu	Zn	Fe
Control	135.4 ± 4.1	36.6 ± 3.7	3.2 ± 0.4	2.2 ± 0.2	0.9 ± 0.09	0.014 ± 0.003	0.009 ± 0.001	0.023 ± 0.009
	126.5–141.3	31.1–44.5	2.3–4.0	1.8–2.5	0.78–1.1	0.009–0.019	0.006–0.01	0.016–0.047
Early PD	142.6 ± 11.4	32.0 ± 5.0	4.12 ± 0.9	2.41 ± 0.3	1.05 ± 0.1	0.022 ± 0.008	0.008 ± 0.002	0.02 ± 0.004
	125.0–164.5	25.3–45.6	2.7–6.8	1.8–3.0	0.82–1.3	0.007–0.035	0.006–0.012	0.01–0.028
Severe PD	142.9 ± 9.8	31.1 ± 3.3	3.65 ± 0.5	2.23 ± 0.2	1.05 ± 0.08	0.02 ± 0.006	0.007 ± 0.001	0.017 ± 0.007
	127.2–168.5	26.5–38.0	2.8–4.5	2.18–2.69	0.86–1.19	0.011–0.035	0.005–0.009	0.004-0.035

Values represented in second line indicate Feature values in range format.

2.3.1. Computational-analysis model

The main theme of this research paper is to devise a computational approach for critical analysis of an input (a test sample). At the outset, it is required to label whether the sample is normal or abnormal and if abnormal it could be early or severe, which requires a further refinement in the computational model using clinical observations (Explained in Section 5). Pattern recognition methods advocate the use of transformed feature space [18] in place of original feature space with the aim of improved performance. In this work we have explained the analysis with original feature space of trace elements (Na, S, P, etc.) and the ratio type transformed features (Na/S, Na/Fe, etc.). In case of large number of features, Feature reduction [18] is also advocated in pattern recognition studies, but this aspect is kept beyond the scope of this paper.

As bought out in Section 2, two phases are involved in computational modeling: (i) Learning Phase (ii) Recognition Phase. Learning phase involves building up a reference KB using healthy samples (Controls). Two different models are suggested for representing the knowledge: (i) Parametric Model (ii) Distribution Model. In the recognition phase a test sample is contrasted with the reference KB to decide the label (Normal/Abnormal), and further the degree of belongingness (affinity) of the sample with that class. This is accomplished using different distance measures [19]. These details are brought out in the following sections.

2.4. Parametric model for affiliation analysis

The efficacy of affiliation estimation of a new sample either as healthy (desired) or not depends on the learning of healthy (desired) set of data. A simple procedure for learning the healthy group is proposed here, which consists of expressing the feature-wise knowledge parameters of healthy samples in terms of mean and standard deviation of N control samples constituting the reference base. The recognition stage involves testing whether a presented test case (sample) belongs to the healthy (desired) group or not. The most important objective, in either case is to estimate the amount/degree of affiliation of the sample to the reference base (in case it is healthy) or to estimate the amount/degree of being away from the reference base. In this work, we have employed Doyle's distance [20] measure to carry out the affiliation analysis. The specific recommendation to opt for Doyle's distance measure is due to the fact that it captures the essence of both mean and standard deviation parameters, which are defined as the knowledge parameters in our work. Doyle's distance is given by $\sqrt{(\mu-\mu^1)^2+(\sigma-\sigma^1)^2}$. Here $\mu=$ Mean value of the reference base (healthy/desired), $\mu^1=$ New mean value treating the test sample also as an additional member in the reference base. Similarly σ & σ^1 represent standard deviation values. The procedure is to compute (μ,σ) parameters for every feature in the cognition (learning) phase. In the recognition phase, Doyle's distance is computed for every feature and the total distance is the sum of all feature-wise distances computed.

The way it is put above indicates that μ^1 and σ^1 are computed using $N+1$ sample where N is the size during the learning and $N+1$ is due to the inclusion of the test sample. If the test sample belongs to a healthy group then the computation of μ^1 and σ^1 based on $(N+1)$ samples does not pose any computation pitfalls, however it could result in incorrect affiliation estimation if the test sample does not belong to the healthy (desired) set. To alleviate this problem, it is suggested to replicate the test sample N times to keep the size $=N+N=2N$ for the computation of μ^1 and σ^1, so that the 'weightage factors' for the healthy group and the test group are made uniform. The Doyle's distance is computed feature-wise and the sum of Doyle's distances over all features gives an index for affiliation estimation if the sample is healthy or otherwise.

2.5. Distribution model

The assignment of variables into two groups must always be motivated by the nature of the response variables and never by an inspection of data. Normal distribution serves as a prototype of a benchmark to test statistical significance. If an observed difference markedly departs from the model of all possible difference, it is interpreted as unique, significant and meaningful [21,22]. Majority of physical and mental traits tend to be distributed as to approximate normal distribution stretching from $-\infty$ to $+\infty$ and covering the unit area interpreted as probability of occurrence of the universe of traits or events which describes, the normal distribution as an ideal [23].

In this work we have proposed to decompose the normal distribution into several mutually exclusive zones. Subsequent to zonalization weightage factor is assigned to each zone. Later, affiliation analysis of a test sample is performed by a suitable distance measure such as a City block [19]. The specific recommendation to opt for zonalization is due to the fact that it captures the essence of distribution parameters, which are defined as the knowledge parameters in our work, and the reason to choose the City block for distance computation is for its computational ease. The discrimination efficacy of the new samples either as healthy or not, depends on the learning done on the healthy (control) data set. The strength of knowledge derived mainly depends on distribution pattern of each feature.

2.5.1. Zonalization

Zonalization of feature-wise distribution into different regions provides a clear indication of domain knowledge. This notion of zonalization is explained below [24]. Let $\prod_1 \ldots \prod_m$ be m populations with density functions $P_1(x)$, $P_m(x)$, respectively. We wish to divide the distribution space based on the observations of training data set into m mutually exclusive and exhaustive regions. $R_1 \ldots R_m$. If the region falls into R_j we shall say that it comes from $\prod_j$. The procedure assigns the point x

to one of the R_j based on which regions $R_1 \ldots R_m$ is defined. The discrimination procedure requires classifying an observation as coming from $\prod_j$ if it falls in R_j.

2.5.2. Attribution of weightage factors for zones

Consider a pattern $x = (x_1, x_2, \ldots x_n)^{\mathrm{T}}$ in R^n. The measurements x_i, $1 \leqslant i \leqslant n$ which represents the sample and by which one is supposed to classify the pattern are usually not equally important. Clearly measurements of less importance should be assigned smaller weights [24]. The stage of creation of a KB consists of representing the healthy samples with a distribution pattern, and zonalizing the distribution into regions (such as Intrinsic, Safe, Permissible and Acceptable zones, respectively). The recognition stage involves identifying whether a presented test case (sample) belongs to the healthy group or not. The most important observation in either case is to estimate the closeness of the test sample to the reference base (in case it is healthy) or to estimate the farness of being away from the reference base.

2.6. A typical normal distribution, zonalization and weightage factors

The pictorial bell shaped representation of a typical normal distribution is shown in Fig. 1.

Computation of $\underline{R}$, $\overline{R}$ is performed using below formulae:

$\underline{R} = [\mathrm{Min} - (\mathrm{Min} * 1/100)]$, $\bar{R} = [\mathrm{Max} + (\mathrm{Max} * 1/100)]$, where Min, Max are the lowest and highest value in feature of training samples.

The representation structures $(\_)$, $(^{-})$ indicate, respectively, upper and lower limit of the range values.

The zonalization proposed is demonstrated in Fig. 1 and more explicitly in Table 2. The distribution model is split into Intrinsic, Safe, Acceptable, and Permissible regions. These names are so chosen that they could be close to real life observations. If a test sample lies in the intrinsic zone, then it can be declared to be intrinsically healthy. Similarly a test sample can be declared to be safely healthy, acceptably healthy or permissibly healthy. If a test sample lies beyond the permissible zone then it is to be considered unhealthy. To express these qualitative discriminations on a number scale, suitable weightage factors, as illustrated in Table 2, are attributed to different zones. The cut off points and the weightage factors for different zones are so chosen that they provide proper meaning to the words Intrinsic, Safe, Acceptable, and Permissible and Beyond Permissible regions. There is ample scope for research to decide these weightage factors and cutoff points.

Table 2 explains Fig. 1 more vividly. For instance the so-called safe region is defined as being present in between the cutoff points $(\underset{\sim}{r}, \underset{\bullet}{r})$ on the left swing portion of the distribution and in between the cutoff points $(\overset{\bullet}{r}, \tilde{r})$ in the right swing portion. The range for this zone is from 90% to 70% $(\mathrm{Peak}/\sqrt{2})$ of the peak of the corresponding normal distribution. This implies that test sample is quite safe (healthy) if it lies in this range and is not a PD affected sample. The corresponding weightage factor 0.01 is assigned while computing the discriminating distance for the sample in the specific algorithm presented later in Section 3.3.

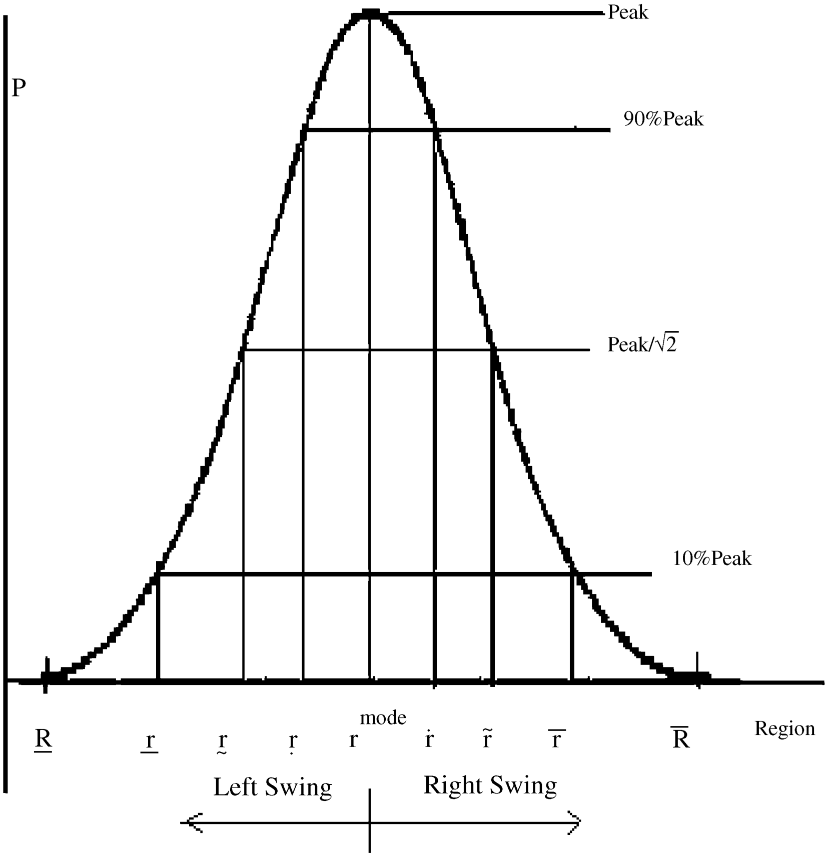

Fig. 1. Illustrating the Distribution Pattern for Normal Samples, with Left and Right swing.

2.7. Ratio features

Concepts are inventions of the human mind which are used to construct a model of the world. They package reality into discrete units for further processing, they support powerful mechanisms for doing logic, and are indispensable for both precise and extended chains of reasoning but concepts and percepts cannot form a perfect model of the world as they are abstractions that select features that are important for one purpose, but ignore details and complexities that may be just as important for some other purpose [25]. Most of the studies related to trace elements were mainly concentrated to certain elemental (Feature-wise) analysis. Our team started to look at the problem in elemental to elemental interrelations (Ratio-wise) as a novel idea for understanding the in-depth hidden information present in CSF of both normal and AD [26], which has been extended in the present research work for PD data set.

Analysis with a very large number of features is difficult. Probably the simple alternative approach would be to combine the feature values of multidimensional data by performing simple operations like addition, subtraction rationing. In fact, these operations are very common with remotely sensed

Table 2
Zonalization of normal distribution

Proposed zones or regions	Suggested cut off point for zonalization	Proposed zonal weightage factor	Zones or (Regions) [Left swing, Right swing]
Intrinsic	90% Peak	0.001	$[(\underset{\bullet}{r}, r^{\text{mode}}), (r^{\text{mode}}, \overset{\bullet}{r})]$
Safe	Peak/$\surd 2$	0.01	$[(\underset{\sim}{r}, \underset{\bullet}{r}), (\overset{\bullet}{r}, \tilde{r})]$
Acceptable	10% Peak	0.1	$[(\underline{r}, \underset{\sim}{r}) ,(\tilde{r}, \bar{r})]$
Permissible	Nearest neighbor	1.0	$[(\underline{R}, \underline{r}), (\bar{r}, \bar{R})]$
Beyond permissible	Far off	10.0	$< \underline{R} \ > \bar{R}$

multispectral data [27,28]. Parametric and partitioned distribution model have to be now devised using transformed ratio features and distance computations and affiliation analysis have to be done on new transformed ratio features. However, there is a disadvantage that the total number of distinct ratio features will be more than the original number of features. If $f_1,\ f_2 \dots f_n$ are the n observations in the original features space, then ratio feature space have $n(n-1)/2$ number of distinct transform features, and $n(n-1)$ all possible ratio features as given in the following two matrix representations for a n-d original feature space of $f_1,\ f_2 \dots f_n$. This results in the increase of dimension leading to curse of dimensionality [18]. Various dimensionality reduction procedures [29] are suggested in pattern recognition literature. But this aspect is not investigated in this research work.

$$\begin{pmatrix} f_1/f_2 & f_1/f_3 & f_1/f_4 & \cdot & \cdot & \cdot & f_1/f_n \\ & f_2/f_3 & f_2/f_4 & \cdot & \cdot & \cdot & f_2/f_n \\ & & f_3/f_4 & \cdot & \cdot & \cdot & f_3/f_n \\ & & & & & & \vert \\ & & & & & & f_{n-1}/f_n \end{pmatrix},$$

Matrix 1: $n(n-1)/2$ distinct ratio feature space

$$\begin{pmatrix} - & f_1/f_2 & f_1/f_3 & f_1/f_4 & \cdot & \cdot & \cdot & f_1/f_n \\ f_2/f_1 & - & f_2/f_3 & f_2/f_4 & \cdot & \cdot & \cdot & f_2/f_n \\ & \cdot & & & & & & \\ & \cdot & & & & & & \\ f_n/f_1 & f_n/f_2 & f_n/f_3 & f_n/f_4 & \cdot & \cdot & \cdot & - \end{pmatrix}.$$

Matrix 2: $n(n-1)$ all possible ratio feature space.

Table 3
Feature-wise knowledge parameters for normal samples: control set size $N = 25$

Feature	Element	Mean (μ)	SD (σ)
F1	Na	134.496	4.919
F2	S	36.918	3.816
F3	Ca	2.236	0.228
F4	Mg	0.919	0.097
F5	P	3.202	0.391
F6	Fe	0.019	0.004
F7	Cu	0.013	0.003
F8	Zn	0.009	0.002

3. Results

3.1. Elemental concentration

The elements chosen for the present study have biological importance and are involved in brain function. Elemental concentration (μmol/ml) for control, early PD and severe PD serum are given in Table 1. The data are presented in μmole concentration in order to calculate mole ratio of elements and also to determine inter-elemental correlation. The results clearly showed that serum levels of K, Mg, Cu, Co and P were higher ($p < 0.01$) in both early PD and severe PD compared to control. S and Al were significantly low ($p < 0.01$) in both early and severe PD, while Fe and Zn were decreased significantly ($p < 0.01$) in only severe PD compared to control, which may reflect the severity of PD. Interestingly, in early PD serum the concentrations of P, Cu, K and Ca were higher than control and severe PD. However, there was no significant change in the total concentration (μmol/ml) of elements among control and PD groups.

3.2. Parametric model based analysis

The knowledge parameters of the reference base are computed using the training/learning set (Control/Healthy samples) with $N = 25$. Table 3 illustrates the feature-wise knowledge parameters (μ, σ) derived from control samples. Similarly ratio feature-wise knowledge parameters are also computed.

The affiliation indices (Doyle's distance values) feature-wise were computed and significant variations of the affiliation index (Doyle's distance value) because of early/severe samples indicated that the sample was refuted by the reference base as illustrated in Table 4. In this table results are presented in minimum–maximum format, the minimum value corresponding to the smallest Doyle's distance computed over all the test samples presented and similarly the maximum corresponding to the largest Doyle's distance.

From the table the following observations can be made

- Doyle's distance component shows a quick rise in value with the samples drawn for early PD, and the distance drops with the progression of disorder from early to severe PD.

Table 4
Feature-wise distance computations using Parametric Model (Test samples are drawn from all three types)

Distance of a test samples from the base reference (Healthy)				
Features	Element	Control samples [Min, Max]	Early PD samples [Min, Max]	Severe PD samples [Min, Max]
F1	Na	0.10298, 0.19499	0.1016, 2.93454	0.10598, 1.22425
F2	S	0.0772, 0.20571	0.11391, 0.7156	0.10248, 0.5917
F3	Ca	0.00462, 0.01558	0.00462, 0.05797	0.00462, 0.02163
F4	Mg	0.00199, 0.00908	0.0021, 0.0205	0.0021, 0.01615
F5	P	0.00794, 0.03894	0.00795, 0.23341	0.00794, 0.08804
F6	Fe	0.00008, 0.00014	0.00008, 0.00079	0.00009, 0.00372
F7	Cu	0.00007, 0.00016	0.00007, 0.00376	0.00007, 0.00225
F8	Zn	0.00003, 0.00009	0.00003, 0.00015	0.00004, 0.00021
Net Doyle’s distance	D	0.22818, 0.40601	0.48647, 3.16591	0.46954, 1.41341

Table 5
Summary of net Doyle’s distance

Net Doyle’s distance			
Ratio type	Control samples [Min, Max]	Early PD samples [Min, Max]	Severe PD samples [Min, Max]
Distinct ratio features	208, 429	505, 1113	502, 663
All ratio features	206, 447	505, 1130	503, 714

- The range of distances for samples from severe cases cannot be separable from the range of distances computed for samples drawn from early PD cases with Na, S, Mg, etc.

Thus we cannot clearly understand the progression of disorder by just analyzing selective elements. But the net Doyle’s distance

$$D = \sum_{i=1}^{n} d_i,$$

where d_i is Doyle’s distance component for the ith feature shows a clear discrimination between the healthy samples and abnormal samples, which consists of test samples from both Early and severe PD.

This observation is true even with ratio features. Although it appears that there is an open scope for advanced mining [30] with ratio features to discover hidden knowledge, it is not included in this research communication, since we are still researching on this issue. The summary of net Doyle’s distance on ratio features is given in Table 5.

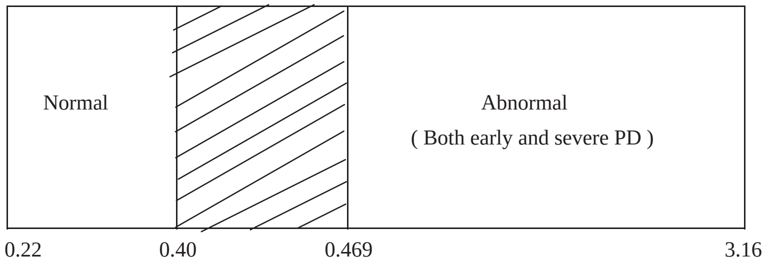

Fig. 2. Schematic diagram illustrating separation of normal and abnormal samples with Doyle’s distance computations.

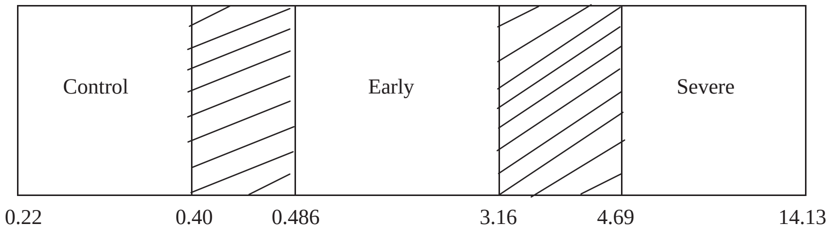

Fig. 3. Schematic diagram illustrating separation of samples between the normal and abnormal (Early PD and Severe PD) after Doyle’s distance is refined with W.

Overall we prefer to stick to the following observations,

1. Healthy samples and abnormal samples are distinctly separable in terms of net Doyle’s distance Tables 4 and 5.
2. It is generally difficult to establish a clear line of separation between early PD and severe PD samples from Doyle’s distance analysis of the samples identified as abnormals.
3. Supervised analysis shows that the conventional truth that severe PD cases should generally exhibit higher distance than early PD cases cannot also be distinctly observable.
4. The observation (3) implies that Doyle’s distance index should be refined with some deeper observations made for the samples which are labeled as abnormals. Such a supporting clue can be drawn from clinical observations made on the PD affected samples [14].

Clinical observations have reported the total UPDRS scale in the order of $57.6 \pm 15.6 *$ UPDRS scale for severe PD case and $23.1 \pm 11.4 *$ UPDRS scale for early PD case [14]. Even without a detailed clinical analysis, a neuro-physician declares a case to be severe if motor movements are uncontrollably high. With this supporting evidence Doyle’s distance is refined with a refining weightage factor W as

Refined Doyle’s Distance $=$ W times the corresponding Doyle’s distance.

With $W = 1$ for early PD cases indicated by a value less than 30 in UPDRS scale and $W = 10$ for severe PD cases indicated by a value greater than 50 in UPDRS scale.

With this refinement the parametric model based analysis with Doyle’s distance measure shows the clear separation of classes and satisfies the conventionally understood truth of distances as indicated in the schematic diagrams shown in Figs. 2 and 3, drawn with reference to Table 4.

Table 6
Net Doyle's distance for ratio features with and without refinement

Ratio type	Net Doyle's distance (without Refinement)	
	Control samples [Min, Max]	Affected samples [Min, Max]
Distinct ratio features	208, 429	502, 663
All ratio features	206, 447	503, 714

	Net Doyle's distance (with Refinement)		
	Control Samples [Min, Max]	Early PD Samples [Min, Max]	Severe PD Samples [Min, Max]
Distinct ratio features	208, 429	505, 1113	5020, 6630
All ratio features	206, 447	505, 1130	5030, 7140

Table 6 gives a concise representation of the computations carried out employing (i) distinct ratio features and (ii) all ratio features. Doyle's distance for a healthy test sample is in the range of 200 –400, while the affected samples maintain a larger distance in the range of 500–700. The results are presented again with the refinement done on the computed distances for affected samples, which show pronounced discrimination between early and severe PD samples.

3.3. Distribution model based analysis

Frequency distribution table is computed using the feature values. If f_i is the number of samples for x_i, then $p(x_i) = f_i/n$, where $n = 25$. Graph $p(x_i)$ versus x_i has been observed to simulate bell shaped distribution Table 7 shows the results of zonalization on eight individual features of this reference data set. Similar computations can be made on ratio features. These regions symbolize the knowledge of the reference base. (Theoretical details are covered in 2.5 and 2.6.)

The individual control samples representing *healthy* were presented as test samples to simulate healthy cases and test samples either from *early* or *severe* obviously simulate unhealthy/abnormal cases. Large distance measures or farness because of unhealthy sample indicates that the sample is refuted by the reference base.

The distance computations are based on City block formula [19]. General City block finds distance in the ith feature for two samples A and B using the formula $d_i = |A_i - B_i|$ and overall distance between A and B is $D = \sum d_i$ for all features. City block distance computations are modified with zonalization weightage factor as explained below.

Suppose F_i is the modal value of the ith feature in reference base, and f_i is the ith feature value corresponding to the test sample, then $d_i = W_i|F_i - f_i|$, where d_i is the City block distance component for ith feature and W_i is the weightage assigned to the zone, where f_i falls. The net distance over all features is $D = \sum d_i$. The feature-wise City block distances and overall City block distance computed, as explained above are given in Table 8. Net distance for feature-wise analysis

Table 7
Computed knowledge parameter for normal samples in distribution model based on feature-wise analysis: control set size $N = 25$

	Modal point	Wing[a]	Intrinsic	Safe	Acceptable	Permissible
Na	135	L	133.7–135	131.6–133.7	126.8–131.6	124.7–126.8
		R	135–136.7	136.7–137.6	137.6–141.8	141.8–143.4
S	36.62	L	36–36.62	35.0–36.0	31.12–35.0	30.6–31.12
		R	36.62–37.5	37.5–38.87	38.87–44.37	44.37–45.45
Ca	2.18	L	2.12–2.18	2.24–2.12	1.92–2.24	1.83–1.92
		R	2.18–2.28	2.28–2.38	2.38–2.57	2.57–2.62
Mg	0.925	L	0.88–0.925	0.83–0.88	0.82–0.83	0.79–0.82
		R	0.925–0.95	0.95–1.02	1.02–1.1	1.10–1.20
P	3.125	L	3.02–3.125	2.91–3.02	2.53–2.91	2.47–2.53
		R	3.125–3.27	3.27–3.43	3.43–3.90	3.90–4.04
Fe	0.0195	L	0.018–0.0195	0.016–0.018	0.016–0.016	0.0158–0.016
		R	0.0195–0.022	0.022–0.0243	0.0243–0.0278	0.0278–0.0282
Cu	0.0135	L	0.0123–0.0135	0.0109–0.0123	0.009–0.0109	0.0089–0.009
		R	0.0135–0.0151	0.0151–0.0165	0.0165–0.0187	0.0187–0.0191
Zn	0.00825	L	0.0080–0.00825	0.0072–0.008	0.0062–0.0072	0.0059–0.0062
		R	0.00825–0.0086	0.0086–0.0091	0.0091–0.0105	0.0105–0.0106

[a] L: left swing; R: right swing.

Table 8
Feature-wise distance computations using distribution model (Test samples are drawn from all three types)

Ele	Features	Distance of a sample from the base reference (Healthy)		
		Control samples [Min, Max]	Early PD samples [Min, Max]	Severe PD samples [Min, Max]
Na	F1	0.015, 0.84	0.45, 29.56	0.00087, 17.17
S	F2	0.40, 0.79	0.39, 7.80	0.359, 9.361
Ca	F3	0.02, 0.03	0.0008, 0.018	0.0018, 0.513
Mg	F4	0.00002, 0.14	0.00061, 0.31	0.01, 0.227
P	F5	0.00008, 0.037	0.000043, 1.175	0.0001, 0.916
Fe	F6	0.0000025, 0.0009	0.0000015, 0.0000025	0.0155, 0.0354
Cu	F7	0.0000025, 0.0035	0.00035, 0.0055	0.00025, 0.0055
Zn	F8	0.00000025, 0.00025	0.00000025, 0.00025	0.00000025, 0.0032
D	Net distance	0.474–1.612	8.29–31.45	7.29–17.93

is summarized in Table 9. The computed distance are presented in the minimum–maximum format (giving the range) as explained earlier with the previous model.

Similar calculations are made with ratio features and the results are summarized in Table 10. Analysis of these Tables 8–10 show perfect resemblance to the observations made in the Section 3.2 above. As recommended earlier, the discrimination is done in two stages. In first stage normal and

Table 9
Net distance for feature-wise analysis with and without refinement (distribution model)

Feature	Net distance		
	Control samples [Min, Max]	Early PD samples [Min, Max]	Severe PD samples [Min, Max]
Without refinement	0.474–1.612	8.29–31.45	7.29–17.93
With refinement	0.474–1.612	8.29–31.45	72.9–179.3

Table 10
Net distance for distinct ratio type features with and without refinement (distribution model)

Distinct ratio type	Net distance (in $\times 10^3$ units)		
	Control samples [Min, Max]	Early PD samples [Min, Max]	Severe PD samples [Min, Max]
Without refinement	0.545–3.85	4.57–19.41	18.34–109.89
With refinement	0.545–3.85	4.57–19.41	183.4–1098.9

abnormal cases become separable. In the second stage, further refining of distance with a weightage factor $W = 1$ for early and $W = 10$ for severe PD classifies the abnormal samples also into clear two groups. The distance values after refinement process are given in Tables 9 and 10 which show clear separation of all three classes.

3.4. Comparative analysis

In this section we present the computational analysis of the two algorithms suggested. A Parametric model requires a simple structure for creating the KB, since only two parameters (μ, σ) are required for each feature. Here the memory required is proportional to $2n$ where n is the number of features. Feature-wise Doyle's distance computation involves two subtractions, two squaring operations, one addition and one square root computing operation. Computing the net Doyle's distance involves summing up all individual Doyle's component, which is proportion to $O(n)$, where n is the number of Doyle's components and n is number of features in feature-wise analysis.

The Distribution model requires a slightly complex structure for creating a KB since four zonal details and corresponding weightage factors have to be stored for each feature. Thus memory requirement is in proportion to $5n$, where n is the number of features. However, the notion of zonalization is close to real life conventions (Intrinsic, Safe, Acceptable, and Permissible). Feature-wise City block distance computation involves a nested If structure to decide the weightage factor, one subtraction and one multiplication operation. The net city block distance involves summing up all individual City block distance components, which is proportion to $O(n)$ in feature-wise analysis. Both the models have shown similar discrimination efficiency. In fact the magnitude of separation between the classes is relatively higher with distribution based computations than parametric computations. In summary both the models are computationally efficient as well as performance wise efficient.

4. Discussion and recommendation

Essential trace elements are required at very low concentrations for the proper functioning of human biological systems, yet at a higher concentration they are toxic. The controversy over metal levels present in PD serum has lead the problem to be viewed in different avenues. The effect of change in metal levels is not restricted to the analyzed metals alone, but it leads to the large effect on total summation of metals present in the serum samples. Further, the inter-relations of these metals give a clear indication on the homeostasis of metals. Thus, we have carried out this work by considering eight metals and also computed the inter-relations of this by taking ratio features.

Both Doyle's distance and Zonalization based city block components show a quick rise in value with the samples drawn for early PD, and the distance drops with the progression of disorder from early to severe PD. The range of distance for samples from severe cases cannot be separable from the range of distance computed for samples drawn from early PD cases with Na, S, Mg, etc. We also have observed that distance computations are of a very low order in Zn, Fe.

We can observe that it is generally difficult to discriminate early PD cases and severe PD cases, in accordance with the conventional expectation that distance indices in early PD cases should be lower than that of severe PD cases. The progression of the disorder could be known by clinical diagnosis. Thus the design of the models was refined with the weightage factors (criteria for discriminating factor for the separation for early and severe disorder) 1, 10 for early and severe disorder, respectively.

From the detailed algorithmic computations presented in the previous section, it is evident that the ranges of distances for healthy and affected samples can be properly calibrated to read directly the severity of the progression of the PD, when the distance is computed with the feature set of trace elemental values or with their ratios for a test sample.

The research outcome clearly demonstrates that it is possible to devise a directly interpretable calibrated scale to quantitatively assess the progression of PD, but such a calibration can be very effective with a pretty large number of samples in the control/training set; the lack of which is the only major limitation in the work presented.

To the best of our knowledge, this is the first attempt in studying in general a medical database by applying zonalization to distribution and Doyle's distance measure to parametric models to understand progression of neurological disorder by taking serum as a diagnostic medium. The present computation model is able to distinguish the metal homeostasis between normal to Early PD to Severe PD hence it has diagnostic value. We preempt that more rigorous analysis could be suggested by medical professionals to identify contributing features, which could enhance further both computational efficiency and effective diagnosis.

5. Summary

Given a test sample, under diagnosis for Parkinson's disease, classifying it into a normal or affected (early/severe) is a complex issue. In the present paper, we have proposed two different algorithmic models namely Parametric method (which incorporates mean and variance features) and Distribution method (which involves partitioning the normal distribution of every feature and assigning weightage factors to different partitions made on the normal distribution) which have been processed on the trace elemental values (concentration) present in serum samples of both normal and PD affected (early/severe) population. The explicit need was to cognize a reference KB based on control samples drawn from normal/healthy set, which could be used in the recognition stage, for an accurate discrimination of the test sample as being healthy or affected. The value of the distance computed can be read on a calibrated scale to recognize the severity of the progression of the disorder.

In summary, a pattern recognition inspired algorithmic approach to understand trace elemental homeostasis in serum samples of Parkinson disease is presented in this research paper.

Acknowledgements

The authors wish to thank Vice Chancellor, University of Mysore, Director, CFTRI, Mysore, Principal, JSS Medical College and Hospital, Mysore and Sri Venkateswara Institute of Medical Science, India for their encouragement. Sanjay Pande is thankful to ICMR for awarding SRF and Murlidhar L. Hegde is thankful to CSIR for awarding JRF.

References

[1] J.M. Fearnley, A.J. Lees, Aging and Parkinson’s disease: substantia nigra regional selectivity, Brain 114 (1991) 2283–2301.

[2] F.J. Jimenez-Jimenez, P. Fernandez-Calle, M. Martinez-Vanaclocha, E. Herrero, J.A. Molina, A. Vazquez, R. Codoceo, Serum levels of zinc and copper in-patients with Parkinson’s disease, J. Neurol. Sci. 112 (1992) 30–33.

[3] F.J. Jimenez-Jimenez, J.A. Molina, M.V. Aguilar, I. Meseguer, C.J. Mateos-Vega, M.J. Gonzalez-Munoz, F. Bustos, A.M. Salio, M.O. Pareja, M. Zurdo, M.C. Martinez-Para, Cerebrospinal fluid levels of transition metals in patients with Parkinson’s disease, J. Neural Transm. 105 (1998) 497–505.

[4] F.C. Valdivia, F.J. Jimenez-Jimenez, J.A. Molina, P.F. Calle, A. Vazquez, F.C. Liebana, S.L. Lobalde, L.A. Peralta, M. Rabasa, R. Codoceo, Peripheral iron metabolism in patients with Parkinson’s disease, J. Neurol. Sci. 125 (1994) 82–86.

[5] Uitti, A.H. Rajput, B. Rozdilsky, W.K. Yuen, Regional distribution of metals in human brain, Clin. Invest. Med. 10 (1987) 10–13.

[6] M.L. Hedge, P. Shanmugavelu, B. Vengamma, T.S. Sathyanarayana Rao, R.B. Menon, R.V. Rao, K.S. Jagannatha Rao, Serum trace elemental levels and complexity of inter-elemental relationships in Patients of Parkinson’s disease, Under communication.

[7] W.J. Frawley, G. Piatetsky-Shapiro, Knowledge Discovery in databases, An Overview in Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA, 1991, 1–27.

[8] J. Han, Knowledge discovery in databases: an attribute-oriented approach, Proceedings of the 18th VLDB Conference Vancouver, British Columbia, Canada, 1992.

[9] R.J. Schalkoff, Pattern Recognition: Statistical, Structural and Neural Approaches, Wiley, New York, 1992.

[10] O.D. Richard, E.H. Peter, G S. David, Pattern Classification, EDS-2, Wiley, New York, 2000.

[11] M.T. Evangelia, Supervised and Unsupervised Pattern Recognition, Feature Extraction and Computational Intelligence, CRC Press, Boca Raton, London, New York, Washington, DC, 2000.

[12] S.I. Weiss, C. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems, Morgan Kaufmann, San Francisco CA, 1991.

[13] W.J. Weiner, A.E. Lang, Parkinson ’s Disease. Movement Disorder: A Comprehensive Survey, Futura Publishing Co., Mount Kisco, NY, 1989, pp. 23–115.

[14] S. Fahn, R.L. Elton, and members of the UPDRS Development Committee, Unified Parkinson's disease rating scale, In: S. Fahn, C.D. Marsden, M. Goldstein, D.B. Calne (Eds.), Recent Developments in Parkinson's Disease, Vol. 2, UPDRS Publishers, New Jersey, 1987, pp. 153–163.

[15] M.M. Hoen, M.D. Yahr, Parkinsonism: onset, progression and mortality, Neurology 17 (1967) 19–423.

[16] National Committee for Clinical Laboratory Standard, Approved Guidelines: Control of Pre-analytical Variation in Trace Element Determination, Vol. 17(13), National Committee for Clinical Laboratory Standard, International Clinical Laboratory Standard, Washington DC, 1997, pp. 1–30.

[17] M.T. Rajan, K.S. Jagannatha Rao, M.B. Mamatha, R.V. Rao, P. Shanmugavelu, R.B. Menon, M.V. Pavithran, Quantification of trace elements in normal human brain by Inductively Coupled Plasma Atomic Emission Spectrometry, J. Neurol. Sci. 146 (1998) 153–163.

[18] A.K. Jain, P.W. Robert, Duin Janchang Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000).

[19] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[20] A.C. Copeland, M.M. Trivedi, Models and metrics for signature strength, Evaluation of camouflaged targets, SPIE Proc. 3070 (21) (1997) 582–591.

[21] Introduction to Data Mining and Knowledge Discovery by Two Crows Corporation, 3rd Edition, Two Crows Corporation, Maryland, 1999, ISBN: 1-892095-02-5.

[22] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, Berlin, 1996.

[23] B. Robert, Learning Theory, Holt, Rinchart and Winston, New York, 1979.

[24] M. Fnedman, A. Kandel, Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy logic approaches, Series in Machine Perception Artificial Intelligence, Vol. 32, Imperial College Press, London, 1999.

[25] J. Sowa, Conceptual Structures, Information Processing in Mind and Machine, Addison-Wesley System Programming Series, Addison-Wesley, Reading, MA, 1984, p. 344.

[26] M.B. Pande, P. Nagabhushan, K.S. Jagannatha Rao, Computer model to understand trace elemental inter-relationship in cerebrospinal fluid of normal and Alzheimer's disease: a diagnostic approach, The Fourth Keele Meeting on Aluminum 24–27 Brichall Centre for Inorganic Chemistry and Material Science, Keele University, Staffordshire, UK, 2001.

[27] RRSSC Lecture Notes, Digital Image Processing—National Natural Resources Management Systems, Department of Space, Indian Space Research Organisation, India, 1988.

[28] S. Prakash, Classification Analysis of Remote Sensed Data: Some New Approaches, Ph.D. Thesis, University of Mysore, Mysore, India, 1998.

[29] R.O. Duba, P.E. Hart, Pattern Classification and Scene Analysis, Wiley-Inter Science Publication, Wiley, New York, 1990.

[30] H. Sano, Biochemistry of the extra pyramidal system Parkinsonism and Related Disorders, Biochemistry (English Translation), Vol. 6, Russia, 2000, pp. 3–6.

P. Nagabhushan a Fellow of Institution of Engineers (FIE) is a professor in Department of Studies in Computer Science, University of Mysore, Mysore, India. He holds BE (1980), MTech (1983) and Ph.D. (1988). He has been a visiting professor and invited researcher at USA, Japan, France and many universities in India. He has been actively associated with many funded research projects. His areas of research interest include Cognition–Recognition, Pattern Recognition, Image Analysis, Dimensionality Reduction, Data Mining, Document Image Analysis, Advanced Exploratory Data Analysis and related problems. He has been actively involved as a referee/reviewer and resource person in his areas of research interest for journals, conferences, workshops, funding agencies and statutory bodies of the Government.

M.B. Sanjay Pande is a Senior Research Fellow of Indian Council of Medical Research, India and currently a doctoral student jointly in Department of Studies in Computer Science, University of Mysore, Mysore and Central Food Technological Research Institute, Mysore, India. He holds Diploma in Computer Science(1989), BE in Computer Science (1996) and MTech in Biomedical Engineering (1999). His areas of research interest include Bioinformatics, Pattern Recognition, Psychiatric Disorders.

Muralidhar L. Hegde M.Sc. Biochemistry (2000) and currently doing doctoral program in the area Genomics in Parkinson disease in University of Mysore, India. He is a CSIR fellow at CFTRI, Mysore, India.

T.S. Sathyanarayana Rao is MD(1983) in Psychiatry. He is heading the Department of Psychiatry at JSS Medical College and Hospital, Mysore, India. His areas of research are Bipolar disorders, Sexology, Genetics and other neuropsychiatry disorders. He is the Chief Editor for Indian Journal of Psychiatry. He is the Fellow of Indian Psychiatric Society.

K.S. Jagannatha Rao M.Sc. (1979) and Ph.D. in Zoology (1984). He is a senior scientist at Department of Biochemistry and Nutrition, CFTRI, Mysore, India. His area of research is in Toxicogenomics and toxic protein–nucleic acid interaction studies in Neurological disorders and computational neuroscience. He has published papers in several leading international journals and has several international collaborations on Genomics. He has been actively involved with many funded research projects.