CHAPTER 4

DATASET DESCRIPTION

This chapter presents a detailed description about the datasets used for experimentation in this research work. The following three datasets were used for experimentation: Hepatitis, Thrombosis and Parkinson's datasets. Hepatitis and Thrombosis datasets were released in Principles and Practice of Knowledge Discovery in Databases (PKDD) challenge for a data mining contest that was held on 2005 and 1999 respectively (Hepatitis 2005; Thrombosis 1999). These datasets were collected from Chiba hospital, which contains clinical records stored from 1981 to 2001 and 1980 to 1999 respectively. The Parkinson's Disease (PD) dataset contains data collected in the Unit of the Tel-Aviv Sourasky medical center at the laboratory for gait & neuro dynamics, movement disorders (Gait Database 2005). The gait database includes Parkinson's disease data acquired from three research studies (Yogev et al. 2005; Frenkel-Toledo et al. 2005b; Hausdorff et al. 2007).

4.1 HEPATITIS DATASET

Hepatitis dataset contains laboratory examination reports of Hepatitis B and C patients. Hepatitis B and C are virus infections that damage the liver of a patient. These viruses are highly infectious since they have a high risk of causing liver cirrhosis or Hepatocarcinoma. It has been observed that all the patients who are prone to Hepatitis B or C does not develop the risk of causing liver cirrhosis or Hepatocarcinoma. Hence the differences in temporal patterns between Hepatitis B and C remain undiscovered. Table 4.1

shows the summary of Hepatitis and Thrombosis dataset. Hepatitis data set consists of 771 patient's laboratory test reports of Hepatitis B and C. Each patient has undergone 983 laboratory tests. It has to be noted that not all the laboratory tests taken are related to Hepatitis. The expert guidance and the dataset descriptions given by Ohsaki (2002) were considered and 29 suggested laboratory tests have been selected for experimentation with Hepatitis dataset. The average missing value percentage in Hepatitis dataset is 11.

4.2 THROMBOSIS DATASET

The formation of a blood clot inside a blood vessel is commonly referred as Thrombosis. Thrombosis prevents the flow of blood through the circulatory system thereby causing severe medical complications or death. Thrombosis mainly arises due to collagen diseases, which is known as auto-immune disorder. The patients who are prone to collagen disease tend to generate antibodies that attack their own bodies. Thrombosis is considered as a major complication of collagen disease. Thrombosis data set consist of laboratory test report pertaining to 1000 patients. Each patient has undergone 564 laboratory tests. The expert's knowledge and dataset descriptions given in Tsumoto 1999; Zytkow & Gupta 2001) were considered and 33 suggested tests have been identified for experimentation with Thrombosis dataset. The average missing value percentage in Thrombosis dataset is 8.

The observations of patients whose EHR reports contain more than 30% of incomplete data were not included for experimentation. Accordingly, 499 patients of Hepatitis datasets and 770 patients of Thrombosis datasets were considered for experimentation. The Hepatitis and Thrombosis dataset was used in our previous work Khanna et al. (2007). Currently access to the datasets is unavailable. The contributions presented in Chapter 5, 7, 8 and 9 have used Hepatitis and Thrombosis dataset for experimentation. Table 4.2 shows a detailed description on Hepatitis and Thrombosis datasets.

Table 4.1 Hepatitis and Thrombosis Dataset Summary

Dataset	Total Records	Expert Suggested Lab test	Total Patients	Average Missing value (%)
Hepatitis	1565876	29	771	11
Thrombosis	57543	33	1000	8

Table 4.2 Hepatitis and Thrombosis Dataset Description

S.No	Laboratory Tests	Meaning	Normal Range	Unit	Hepatitis Significance	Thrombosis Significance
1.	GOT	Glutamic Oxaloacetic Transaminase	7< N< 40	IU/l	Yes	No
2.	GPT	Glutamic Pylvic Transaminase	7< N< 40	IU/l	Yes	No
3.	LDH	Lactate Dehydrogenase	216<N<450 *a	IU/l	Yes	No
4.	ALP	Alkaliphophatase	72<N<206 *a	IU/l	Yes	No
5.	TP	Total Protein	6.5<N<8.2	g/dl	Yes	No
6.	ALB	Albumin	3.9 < N < 5.1	g/dl	Yes	No
7.	UA	Uric Acid	3.4<N <7 (Male) 2.4<N <6 (Female)	mg/dl	Yes	No
8.	UN	Urea Nitrogen	8<N < 20	mg/dl	Yes	No
9.	CRE	Creatinine	0.6<N < 1.3	mg/dl	Yes	Yes
10.	T-BIL	Total Bilirubin	0.2<N < 1.2	mg/dl	Yes	No
11.	T-CHO	Total Cholesterol	125<N < 220	mg/dl	Yes	No
12.	TG	Triglyceride	35<N<150	mg/dl	No	Yes
13.	CPK	Creatinine Phosphokinase	40<N < 180	IU/l	No	Yes

Table 4.2 (Continued)

S.No	Laboratory Tests	Meaning	Normal Range	Unit	Hepatitis Significance	Thrombosis Significance
14.	HGB	Hemoglobin	12 < N < 18	g/dl	No	Yes
15.	HCT	Hematoclit	36< N < 50	%	No	Yes
16.	GLU	Blood Glucose	80<N < 120	mg/dl	No	Yes
17.	WBC	White Blood Cell	4< N < 9	E03	No	Yes
18.	RBC	Red Blood Cell	3.75 < N < 5	E06	Yes	Yes
19.	F-A/G	Liver specific F antigen	1.3<N<1.9	%	Yes	No
20.	PLT	Platelet	120 < N < 350	E03	No	Yes
21.	PT	Prothrombin Time	10.5<N < 12.5	sec	No	Yes
22.	APTT	Activated Partial Prothrombin Time	28<N < 38	sec	No	Yes
23.	FG	Fibrinogen	150 < N < 350	mg/dl	Yes	Yes
24.	A2PI	Marker Of DIC	70 < N < 130	%	No	Yes
25.	U-PRO	Proteinuria	0<N<1000	mg/dl	Yes	Yes
26.	IGG	Ig G	900 < N < 1800	mg/dl	No	Yes
27.	IGA	Ig A	90 < N < 350	mg/dl	No	Yes
28.	IGM	Ig M	70< N <240	mg/dl	No	Yes
29.	CRP	C-Reactive Protein	0< N < 0.3	mg/dl	No	Yes
30.	F-A1.GL	Iron metabolism	2.5<N<3.9	%	Yes	No
31.	RA	Rhuematoid Factor	N<40	U/ml	No	Yes
32.	RF	Raha	0 < N < 40		No	Yes
33.	RNP	Anti-Ribonuclear Protein	20 < N < 26	U	No	Yes
34.	SM	Anti-SM	15 < N < 25	U/ml	No	Yes
35.	SCl70	Anti-Scl70	16 < N < 20	EU/mL	No	Yes
36.	RVVT	Measure Of Degree Of Coagulation	-	-	No	Yes
37.	LAC	Measure Of Degree Of Coagulation	-	-	No	Yes

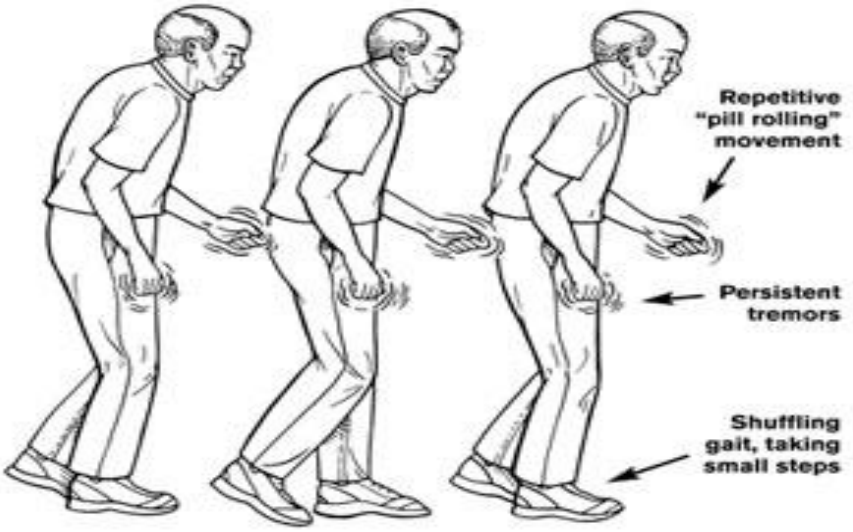
Table 4.2 (Continued)

S.No	Laboratory Tests	Meaning	Normal Range	Unit	Hepatitis Significance	Thrombosis Significance
38.	CENTROMEA	Anti-Centromere	29 < N < 41	AU/mL	No	Yes
39.	DNA	Anti-DNA	30.0<N<75.0	IU/mL	No	Yes
40.	aCLIgG	Anti-Cardiolipin Antibody (Igg)	14<N<80	GPL	No	Yes
41.	ANA	Anti-Nucleus Antibody Concentration	N<=1.0	-	No	Yes
42.	aCL IgA	Anti-Cardiolipin Antibody (Iga)	11<N<80	GPL	No	Yes
43.	KCT	Measure Of Degree Of Coagulation	N<0.15	-	No	Yes
44.	SSA	Anti-SSA	20 < N < 80	mL	No	Yes
45.	SSB	Anti-SSB	20 < N < 80	mL	No	Yes
46.	AMY	Alpha-Amylase	35<N<110*b	IU/l	Yes	No
47.	CHE	Cholinesterase	180<N<430	IU/l	Yes	No
48.	CL	Chloride Test	96<N<106	mEq/L	Yes	No
49.	D-BIL	Bilirubin, Direct	0<N<0.3	mg/dl	Yes	No
50.	F-ALB	Formaldehyde-Treated Bovine Serum Albumin	63.1<N<74.5	%	Yes	No
51.	F-CHO	Cholesterol, Free	36<N<77*d	mg/dl	Yes	No
52.	G_GL	Gamma-Globulin	7.3<N<17.3*c	%	Yes	No
53.	G-GTP	Gamma-Glutamyltranspeptidase	0.1<N<60*e	IU/l	Yes	No
54.	I-BIL	Bilirubin, Indirect	0.2<N<0.9*f	mg/dl	Yes	No
55.	LAP	LeucineAminopeptidase	35<N<100*d	IU/l	Yes	No
56.	TTT	Thymol Turbidity Test	0<N<5*g	U	Yes	No
57.	ZTT	Zinc Sulfate Turbidity Test	4<N<12*g	U	Yes	No
58.	U-SG	Urine- Specific Gravity	-	-	Yes	No

**a, *b, *c, *d, *e, *f, *g denotes lab scale of 200,150,30,50,60,1 and 10 respectively.*

4.3 PARKINSON'S DISEASE DATASET

Parkinson's disease (PD) affects the nervous system of the patients by destroying neurons in the brain that produce a chemical named dopamine. The dopamine is responsible for sending messages to the brain for movement co-ordination (Davie 2008; Ahlrichs & Lawo 2013; Yogev et al. 2005). Thus, most of the PD affected patient's exhibit movement disorders resulting in the postural instability or walking disturbances as shown in Figure 4.1 (Parkinson's Disease 2015, Ahlrichs & Lawo 2013).

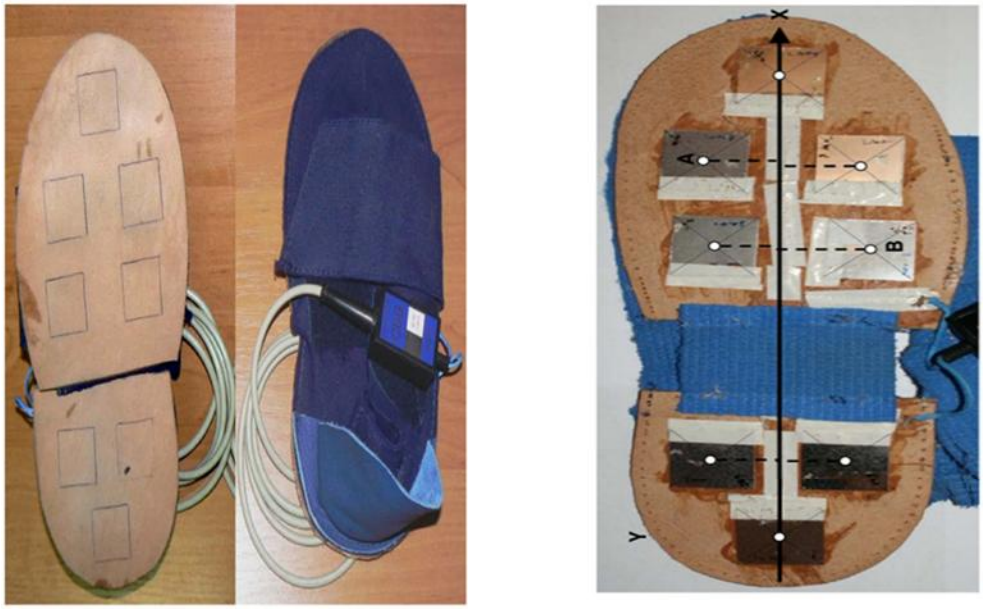


Source : Parkinson's Disease (2015)

Figure 4.1 Gait disturbances in Parkinson's Disease

In the data acquisition, a computerized force-sensitive wearable sensor from Ultraflex Computer DynoGraphy, Infotronic Inc. shown in Figure 4.2 measures the stride-to-stride variations and gait of a subject (Jelen et al. 2008). A gait cycle also referred as stride contains one stance phase and one swing phase. The stance phase represents the period at which the foot strikes on the ground. The swing phase represents the period at which the same foot lifts up the floor. The wearable sensor consists of a pair of shoes

each of which contains eight sensors that is placed in the insole. The subjects were asked to wear those shoes and walk using different styles such as treadmill walking, unassisted walking on a ground level, walking on a ground level using walker, dual-task walking. The Vertical Ground Reaction Force (VGRF) from each sensor measured in newton is recorded in the attached memory card.



(a) **(b)**

Source : Jelen et al. (2008)

Figure 4.2 Wearable Sensors: (a) Wearable Sensors Attached Shoes; (b) Eight Sensors in the Shoe's Insole

These walking (gait) patterns of the PD subjects and normal subjects were observed for 2 minutes. The sensor generates output for every 0.01 sec and for each subject 12000 observations were considered. The detailed mode of the study and statistical analysis were discussed in (Yogev et al. 2005; Frenkel-Toledo et al. 2005b; Hausdorff et al. 2007) which provide detail descriptions of the PD data considered in this work. The Table 4.3 presents an overview of the data that were used by the authors in their study

(Yogev et al. 2005; Frenkel-Toledo et al. 2005b; Hausdorff et al. 2007). Each person involved in the study is referred as a subject. Totally, this database stores 93 PD subjects and 73 control subjects. Table 4.4 provides the descriptions of the attributes in the PD datasets.

Table 4.3 Parkinson's Dataset Summary

Dataset Study	Subjects	Total Subjects	Female	Male
Yogev et al. (2005)	PD	29	9	20
	CO	18	8	10
Frenkel-Toledo et al. (2005b)	PD	35	13	22
	CO	29	11	18
Hausdorff et al.(2007)	PD	29	13	16
	CO	26	14	12

Table 4.4 Parkinson's Dataset Description

Column Number in Dataset	Description	Units
1	Time	Seconds
2 to 9	Measured Vertical ground reaction force (VGRF) from each of eight sensors (L1-L8) in left leg.	Newton
10 to 17	Measured Vertical ground reaction force (VGRF) from each of eight sensors (R1-R8) in Right leg.	Newton
18	Total force under the left leg	Newton
19	Total force under the Right leg	Newton

The time stamped data in Table 4.4 and few non time-stamped data, such as the medical identity, gender, age, height (meters), Weight (Kg), HoehnYahr were considered for experimentation. The contribution presented in Chapter 6 has used PD dataset.