International Journal of Data Mining and Emerging Technologies
DOI: 10.5958/2249-3220.2014.00004.4

Research Article

Application of Data Mining Models in the Diagnosis of Neuropsychiatric Diseases

Mohit Gangwar[1]*, R. B. Mishra[2], R. S. Yadav[3]

[1]Research Scholar, [3]Professor, Department of Computer Science & Engineering, Motilal Nehru National Institute, Allahabad, Uttar Pradesh, India

[2]Professor, Department of Computer Science & Engineering, Indian Institute of Technology-BHU, Varanasi, Uttar Pradesh, India

*Corresponding author Email id: *mohitgangwar@gmail.com; [2]ravibm@bhu.ac.in; [3]yadavrs64@gmail.com

ABSTRACT

Neuropsychiatry is a complex field and its disease diagnosis depends upon the multiple and overlapping symptoms. Data mining method plays a significant role in the analysis of symptoms for the disease diagnosis. In this paper, we apply different data mining techniques for the diagnosis of five neuropsychiatric diseases. The different data mining techniques that we apply in this paper are based on decision tree and artificial neural network concept. Reduced parameter (Sensitivity analysis) parameters based analysis in combination with decision tree and artificial neural network was also performed. Comparative view of accuracy is performed for reduced and non-reduced parameters.

Keywords: Neuropsychiatric diseases; data mining; sensitivity analysis; decision tree; EEG; FMRI; electroencephalogram; disease diagnosis; intelligent computing; Clementine software.

1. INTRODUCTION

Sign and symptoms plays a very important role in the diagnosis and interpretation of neuropsychiatric diseases. Due to the multiplicity and overlapping of symptoms makes diagnosis complex and confusing. Data mining techniques and methods gives scope to find out the relative importance of symptoms, detect meaningful patterns and relation for decision analysis [1, 2]. Several data mining techniques and its applications applied in medical computing [3–7]. These techniques use different methods i.e. sensitivity analysis, decision tree and artificial neural network etc. Very few application software based analysis available for neuropsychiatric diseases. C5.0 is decision tree developed by Quinlan [8]. It generates either decision tree or rule set. It split the sample based on the field that provides the maximum information gain at each level. C5.0 uses boosting to improve the accuracy. Various types of neural networks used for data mining application like multi-layer perception (MLP) [9] and radial basis function (RBF) [10]. Multi-layer perceptions are feed-forward neural networks trained with back propagation algorithm. MLPs are widely used for pattern classification. They can approximate virtually any input–output map with one or two hidden layers [9]. RBF networks deploy a static Gaussian function for the hidden layer processing element which responds only to a small region of the input space where the Gaussian was centred [10]. Sensitivity analysis provided a view to analyse those parameters that were most important for prediction. Sensitivity analysis was used to reduce network complexity by detecting the variables that have no or less influence on network training. The greater the sensitivity degree, the larger the impact it has to the outcomes of artificial neural networks [11].

Neuropsychiatry is an integrating neuroscience of neurology and psychiatry that aims to investigate the psychiatric symptoms of neurological disorders as well as the neurobiological bases of psychiatric disorders, including organic mental disorders and endogenous psychoses. Additionally, neuropsychiatry aims to prevent or reduce the suffering of individuals with the psychiatric symptoms of cerebral disorders [12].

There were methods depicted in literature that apply data mining techniques based on signal parameters. Taniguchi *et al*. apply data mining analysis for cognitive dysfunction [13]. Maroco, João, *et al*. used data mining techniques for dementia disease prediction [14]. Sigurðsson used data mining method for brain regions with neuroimaging databases [15]. Maia *et al*. used neural network based approach for obsessive compulsive disorder [16].

In this paper we are using SPSS Clementine V11.1 software tool for applying different data mining techniques i.e. C5.0 algorithm for decision tree analysis, three artificial neural network models and sensitivity analysis. Sensitivity analysis and its combination with different ANN methods show the neural network performance with reduced parameters. Data mining problems may involve hundreds or even thousands of fields that can potentially be used as predictors. As a result, a great deal of time and effort may be spent in examining which fields or variables to be included in the model. The combined model helps in choosing the factor that has greater influence on predictions.

The rest of paper was organized as follows. Section 2 covers the disease description. Section 3 describes the application of data mining in neuropsychiatric diseases. Section 4 apply decision tree algorithm and show its results. Section 5 covers different artificial neural network models and shows its comparative results. Section 6 describes the reduced parameter based models. Section 7 shows the comparison of different models. Section 8 describes mid-level analysis. Section 9 explains the decision tree and artificial neural network combination and shows its result. Conclusion was mention in Section 10.

2. DISEASES DESCRIPTION TABLE

The formulation of problem was described in terms of diseases description table (DDT) as shown in Table 1. Diseases description table contains symptoms and their relationship for five neuropsychiatric diseases. The five neuropsychiatric diseases defined in DDT were attention-deficit hyperactivity disorder (ADHD), dementia, mood disorder (MD), obsessive-compulsive disorder (OCD) and schizophrenia (SZ). The description of sign and symptoms of five neuropsychiatric diseases defines in following literature [17–20]. The diseases description table contains 38 symptoms divided into two groups and five categories. First group described by two categories. Category one described by EEG signal parameters i.e. frontal (FL), parietal (PL), occipital (OL) and temporal (TL) brain regions abnormality present/observed in EEG signal; and category two contains neuroimaging (FMRI) parameters i.e. frontal (FL), anterior cingulate cortex (AC), cingulate gyms (CG), parietal (PL), occipital (OL), temporal (TL) and basal ganglion (BG) brain regions observed/present in FMRI. Second group contains three categories i.e. (i) muscular physiology (M Phy) consisting of muscular (M) parameters such as: oversleeping (OS) and muscle weakness (MW); and motor activity (MA) parameters such as: difficulty in movement (MO), difficulty in locomotion (LO) and difficulty in using toilet (UT) etc.; (ii) cognitive parameters are confusion in decision making (CD), hearing disability (HD), forgetting memory (FM), judgment (JG), learning disability (LD), reasoning (RS), speech disability (SD) and vision disability (VD); (iii) psychological parameters are: distraction of work (DW), hallucination (HL), fear (FR), hyper activity (HA), agitation (AG), anxiety (AX), stress (ST), anger (AN), abnormal behaviour (AB), need of perfection (NP), social withdrawal (SW) and delusion (DE). It is observed form Table 1 that group 2 collectively represents physio-psycho (PP) abnormality. Table 1 contains rows and columns. Each row defines disease and columns define symptoms. The sub columns of muscular, motor action, cognitive and psychological parameter contains "Y" if the symptom was present in the disease shown in respective row.

3. APPLICATION OF DATA MINING IN NEUROPSYCHIATRIC DISEASES

The experiment was performed on data mining based tool i.e. IBM SPSS Clementine V11.1. This software provides flexibility and easiness to implement many classification models that we have applied in this paper. This software was easy to understand and its graphical user interface ability make user to apply different types of predictive models. Its numerous capabilities make it best data mining analysis tool as compared to other software.

Table 1: Diseases description table

Diseases	Group 1											Group 2: Physio-psycho (PP)																										
	Category 1: EEG signal				Category 2: Neuroimaging (FMRI)							Category 1: Muscular physiology (M Phy)							Category 2: Cognitive (C)								Category 3: Psychological (P)											
												Muscular (M)		Motor action (MA)																								
	FL	PL	OL	TL	FL	AC	CG	PL	OL	TL	BG	OS	MW	MO	LO	UT	CS	W	CD	HD	FM	JG	LD	RS	SD	VD	DW	HL	FR	HA	AG	AX	ST	AN	AB	NP	SW	DE
ADHD	Y	N	N	N	Y	N	N	Y	N	Y	N	N	Y	Y	N	Y	Y	N	N	Y	N	N	Y	N	N	Y	Y	N	N	Y	N	N	N	N	N	N	N	N
Dementia	N	N	Y	Y	Y	N	N	Y	N	N	Y	Y	Y	Y	N	Y	N	N	Y	N	Y	Y	Y	Y	Y	N	N	Y	N	N	Y	N	N	N	Y	N	N	N
Mood Disorder	Y	N	N	N	Y	Y	N	N	N	Y	N	Y	N	Y	N	N	N	N	Y	N	N	N	Y	Y	N	N	Y	N	N	Y	N	Y	Y	N	N	N	N	N
OCD	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y	Y	N	N	N	N	Y	Y	N	N	Y	N	Y	N	N	N	N	Y	N	N	N	N	Y	N	Y	N	N
SZ	Y	N	Y	Y	Y	Y	N	Y	Y	Y	N	Y	N	N	Y	N	N	N	N	N	Y	N	N	N	Y	N	N	Y	N	N	N	N	N	N	Y	N	Y	Y

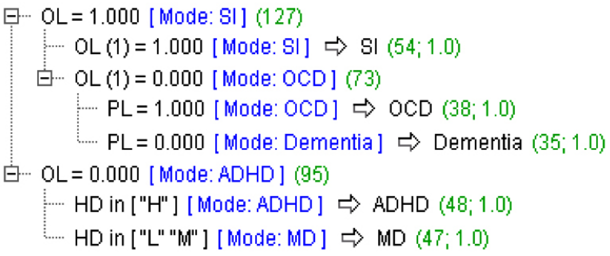
The data set used and created in this study was based on Table 1 information. A total 333 records of five neuropsychiatric diseases involved. The 333 records includes: 71 of ADHD, 55 of Dementia, 66 of MD, 57 of OCD and 81 of SZ. Twenty-seven parameters belong to group 2 define by categorical variable and values associated with each categorical variable was High/Medium/Low. Reaming 11 parameters of group 1 coded as Boolean variable and define by 0/1 value.

Decision tree and three different artificial neural network models was implemented in this experiment. Sensitivity analysis and its combination with decision tree and artificial neural network based on the reduced parameter based concept were also performed. When decision tree based parameters used for constructing three different types of neural network shows best performance of neural network. Comparative view of results analysed in this paper. The data set used for different models applied in this study contains 333 data. In these 333 data, 222 data used for training and construction purpose and 111 data used for testing purpose.

4. C5.0 BASED DECISION TREE

To implement C5.0 based decision tree algorithm, we use 222 data set for generating decision tree and rule set for the diagnosis of five neuropsychiatric diseases. The aberration given in the form (a; b) in Figure 1 and Table 2 can be described in the following manner: In (a; b) denotation, a represents the number of data in the data set to which said disease/rule applies or true and b represent the confidence level i.e. value represent the truth value applicable to number of data represented by value a.

Five rules generate for five neuropsychiatric diseases. One hundred and eleven test data set uses for testing the accuracy of decision tree and rules. The C5.0 shows 100% accuracy for the test data set.



```
⊟ OL = 1.000 [ Mode: SI ] (127)
    OL (1) = 1.000 [ Mode: SI ] ⇨ SI (54; 1.0)
  ⊟ OL (1) = 0.000 [ Mode: OCD ] (73)
      PL = 1.000 [ Mode: OCD ] ⇨ OCD (38; 1.0)
      PL = 0.000 [ Mode: Dementia ] ⇨ Dementia (35; 1.0)
⊟ OL = 0.000 [ Mode: ADHD ] (95)
    HD in ["H" ] [ Mode: ADHD ] ⇨ ADHD (48; 1.0)
    HD in ["L" "M" ] [ Mode: MD ] ⇨ MD (47; 1.0)
```

Figure 1: C5.0 based decision tree

Figure 1 was a decision tree. In this tree each decision branch produced a disease. The two values associated with disease in the bracket interpreted as instances and confidence. The first value was the number of data in the data set to which disease applicable and second value interpreted as confidence involve with the disease.

The above decision tree contains five decision nodes. The decision analysis interpreted as for first rule R1 (if OL = No (95 data)) and HD = High (48 data) then ADHD (48; 1.0) i.e. disease was ADHD for 48 data with full confidence.

Second rule R2 (if OL = Yes (127 data)) and OL (1) = No (73 data) and PL = No (35 data) then dementia (35; 1.0) i.e. disease was dementia for 35 data with full confidence.

Third rule R3 (if OL = No (95 data)) and HD = Low or Medium (47 data) then MD (47; 1.0) i.e. disease was MD for 47 data with full confidence.

Fourth rule R4 (if OL = Yes (127 data)) and OL (1) = No (73 data) and PL = Yes (38 data) then OCD (38; 1.0) i.e. disease was OCD for 38 data with full confidence.

Fifth rule R5 (if OL = Yes (127 data)) and OL (1) = Yes (54 data) then SZ (35; 1.0) i.e. disease was SZ for 35 data with full confidence.

The Table 2 represents Figure 1 in terms of rules. Table 2 contains five rules and each rule contains parameters involve in the diagnosis. The if–then form rule represents parameter and its value relation with disease. The (a; b) values represent for each rule i.e. for first rule applicable to ADHD disease described that there were 48 data set to which rule applies with full (1.0) confidence. In the same manner we can describe other rules. Table 2 contains five rules applicable to five diseases with 100% confidence. The rules shown in Table 2 can be read with respect to Figure 1 in the following sequence i.e. R5, R4, R2, R1 and R3.

Table 2: Five rules for diagnosis

Rules
Rules for ADHD – contains 1 rule(s) Rule 1 for ADHD (48; 1.0) if OL = 0.000 and HD in ["H"] then ADHD
Rules for Dementia – contains 1 rule(s) Rule 1 for Dementia (35; 1.0) if OL = 1.000 and OL (1) = 0.000 and PL = 0.000 then Dementia
Rules for MD – contains 1 rule(s) Rule 1 for MD (47; 1.0) if OL = 0.000 and HD in ["L" "M"] then MD
Rules for OCD – contains 1 rule(s) Rule 1 for OCD (38; 1.0) if OL = 1.000 and OL (1) = 0.000 and PL = 1.000 then OCD
Rules for SZ – contains 1 rule(s) Rule 1 for SZ (54; 1.0) if OL = 1.000 and OL (1) = 1.000 then SZ

5. ANN MODEL

Three ANN methods were Quick, Dynamic, based on error back propagation algorithm and radial basis function network (RBFN) were deployed. All three methods based on 38 input parameters and five output parameters. The input layer neurons calculation was based on 27 categorical and 11 Boolean variables i.e. (27 * 3 + 11 = 92) because categorical variables defined by three values as described in Section 3. ANN has different number of hidden layers based on different methods. So input layers contains 92 neurons and output layer contains five neurons w.r.t. five diseases. All models uses 222 datasets for neural network training and 111 datasets used as test data.

5.1. Quick Model

In Quick model implementation one hidden layer with four neurons for training purpose used. Testing result accuracy is 98.2%. The details of Quick model implementation defined in Table 3. Relative importance of 38 parameters produced by Quick model was shown in Table 4.

5.2. Dynamic Model

In Dynamic model implementation two hidden layers with 15 neurons for training purpose was used. Testing result accuracy was 95.5%. The details of Dynamic model implementation defined in Table 3. Table 4 showed the sensitivity analysis of Dynamic ANN method for 38 symptoms.

Table 3: Parameters used and results analysis

Parameters	Quick Model Values	Dynamic Model Values	RBFN Model Values
Training data set	222	222	222
Correct cases:Accuracy	109:98.2%	106:95.5%	111:100%
Wrong cases:Misclassification	2:1.8%	5:4.5%	0:0%
Neurons in input layer	92	92	92
Hidden layer	1	2	1
Neurons in Hidden layer	4	15	20
Neurons in output layer	5	5	5

5.3 RBFN Model

In RBFN model implementation one hidden layer with 20 neurons for training purpose was used. Testing result accuracy is 100%. The details of Dynamic model implementation defined in Table 3. Table 4 showed the sensitivity analysis of Dynamic ANN method for 38 symptomsand also the results of sensitivity analysis of three ANN methods.

5.4 ANN Models Comparison

When applying test datasets for testing ANN methods accuracy. Test result showed in Table 5 indicated that RBFN model produced best accuracy as compare to quick and dynamic model.

6. REDUCED PARAMETERS BASED MODEL

In this section, we will use and apply sensitivity analysis values defined in Table 4 for the implementation of reduced parameter based models. These reduced parameters based model such as sensitivity analysis with artificial neural network, sensitivity analysis with decision tree and decision tree with artificial neural network. The implementation and result analysis details were given below.

6.1 Sensitivity Analysis Combined with Artificial Neural Network

Table 4 defines sensitivity analysis for all 38 symptoms. Sensitivity analysis described in terms of symptoms' values from highest to lowest. Values for highest to lowest indicate that the importance of symptoms in the design of neural network. For creating model for reduced parameter based symptoms. We were dividing 38 symptoms in to different levels on the basis of values defined in Table 4. Different level contains different number of symptoms for each model. In reduced parameter based analysis we were creating five different levels. Artificial neural network and decision tree model will be implemented for each level. The method to find out the number of symptoms in each level was described as:

D = (Highest Value – Lowest Value)/5, D indicate difference of values between highest value of symptom and lowest value of symptom. Fist level i.e. D1= Lowest value; Second level i.e. D2=D1+2*D; Third level i.e. D3=D1+3*D; Forth Level=D1+4*D; Fifth level i.e. D5= Highest Value. The number of symptoms in each level depends upon the value associated with each level and above.

Reduced Parameters Based Quick Model (A1): The sensitivity analysis values for 38 symptoms given in Table 4 for quick method. These values used to calculate values for different levels i.e. Highest Value = 0.00330597 and Lowest Value = 0.000200075 therefore D = (0.00330597-0.000200075)/5= 0.000621179, D1 = 0.000200075, D2 = 0.001442433, D3 = 0.002063612, D4 = 0.002684791 and D5 = 0.00330597 as shown in Table 6. Based on different levels values and corresponding number of symptoms in each levels. We will construct quick model for each level and test the accuracy of the model as shown in Table 6.

Reduced Parameters Based Dynamic Model (A2): The sensitivity analysis values for 38 symptoms given in Table 4 for dynamic method. These values used to calculate values for different levels i.e. Highest Value = 0.0720821 and Lowest Value = 0.000248337 therefore D = (0.0720821 – 0.000248337)/5 = 0.014366752, D1 = 0.000248337, D2 = 0.028981841, D3 = 0.043348593, D4 = 0.057715345 and D5 = 0.0720821 as shown in Table 6. Based on different levels values and corresponding number of symptoms in

Table 4: Three ANN model generated sensitivity analysis

Quick Model		Dynamic Model		RBFN Model	
Symptoms	**Value**	**Symptoms**	**Value**	**Symptoms**	**Value**
W	0.00330597	H A	7.21E-02	FM	0.0502723
F R	0.00329488	DW	6.78E-02	ST	0.0464501
RS	0.00284828	LD	5.03E-02	DE	0.0442345
NP	0.00246506	MO	4.84E-02	AB	0.0440155
A X	0.00242787	AB	4.03E-02	HD	0.0423901
CD	0.00230833	MW	3.89E-02	SW	0.0418053
AN	0.00218485	H L	3.83E-02	CD	0.0413082
ST	0.00203183	SD	3.70E-02	UT	0.0408744
A G	0.00194363	FM	3.47E-02	NP	0.0391959
LD	0.00186472	DE	3.12E-02	CS	0.0390125
AB	0.00163384	ST	2.49E-02	W	0.038963
CS	0.00160042	TL	2.04E-02	A G	0.0385003
TL (1)	0.00153248	SW	1.96E-02	A X	0.0361186
CG	0.00150896	LO	1.89E-02	H L	0.0360267
DW	0.00141692	A X	1.80E-02	JG	0.0352716
VD	0.00140706	OL	1.77E-02	SD	0.0345426
JG	0.00122909	OL (1)	1.44E-02	OS	0.0341877
MO	0.00121859	RS	1.36E-02	LO	0.0330612
FM	0.0012001	HD	1.35E-02	DW	0.0321607
HD	0.00109114	JG	1.23E-02	MO	0.0316691
SW	0.00108488	CD	1.20E-02	H A	0.0312588
OS	0.00108459	VD	1.10E-02	AN	0.0310293
UT	0.00103656	UT	9.64E-03	F R	0.0304422
DE	0.00101778	CS	8.21E-03	AC	0.0277012
LO	9.59E-04	A G	8.21E-03	MW	0.0266596
SD	8.12E-04	AN	7.43E-03	FL (1)	0.0259189
H L	7.67E-04	F R	5.39E-03	LD	0.025278
H A	6.91E-04	W	4.97E-03	RS	0.0223607
OL	6.86E-04	AC	4.69E-03	OL	0.0222577
MW	6.43E-04	NP	4.57E-03	PL	0.0217966
BG	6.38E-04	BG	3.93E-03	VD	0.0202265
PL (1)	6.17E-04	CG	3.46E-03	OL (1)	0.015403
PL	4.88E-04	PL (1)	3.30E-03	BG	0.01347
OL (1)	3.40E-04	OS	3.14E-03	TL (1)	0.0133759
AC	2.41E-04	TL (1)	3.10E-03	CG	0.01301
FL (1)	2.10E-04	PL	1.58E-03	TL	0.0118652
TL	2.01E-04	FL	1.35E-03	FL	0.0093346
FL	2.00E-04	FL (1)	2.48E-04	PL (1)	0.009179

Table 5: Test result

Method	Test Result (%)
RBFN	100
Quick	98.2
Dynamic	95.5

each levels. We will construct dynamic model for each level and test the accuracy of the model as shown in Table 6.

Reduced Parameters Based RBFN Model (A3): The sensitivity analysis values for 38 symptoms given in Table 4 for RBFN method. These values used to calculate values for different levels i.e. Highest Value = 0.0502723 and Lowest Value = 0.009179 therefore D = (0.0502723 – 0.009179)/5 = 0.00821866, D1 = 0.009179, D2 = 0.02561632, D3 = 0.03383498, D4 = 0.04205364 and D5 = 0.0502723 as shown in Table 6. Based on different levels values and corresponding number of symptoms in each levels. We will construct rbfn model for each level and test the accuracy of the model as shown in Table 6.

In the Table 6, there were three rows corresponding to each model. Each row was further divided into two parts: the first part contains values for different levels and the second part contains accuracy with the number of symptoms given in brackets at level.

The test dataset result shown in Table 6 for quick method indicate that level D1 and D2 produce almost equal accuracy i.e. 98.2% and 95.5%, respectively. Remaining D3, D4 and D5 levels shows test dataset accuracy <90%. It means symptoms <D2 were not producing relevant diagnosis accuracy. Similarly, level D1 and D2 in dynamic method indicate almost equal accuracy i.e. 95.5% and 97.3% it means symptoms <D2 levels not producing acceptable accuracy for diagnosis. Level D1 and D2 give equal accuracy (100%) for RBFN method. In RBFN model level D3 and D4 produced testing accuracy i.e. 97.3% and 94.59%, respectively and acceptable for diagnosis. Level D5 give accuracy <50%.

6.2 Reduced Parameter Based Analysis with Decision Tree

The reduced parameter based concept for decision tree analysis was same as that we have performed for different artificial neural networks. We have divided 38 symptoms in to different levels for all three neural networks. Now we will input different levels parameters to decision tree for all three neural networks. The model A4 was based on the quick method based sensitivity analysis and decision tree was implemented for each level as shown in Table 7. The model A5 was based on the dynamic method based sensitivity analysis and decision tree was implemented for each level as shown in Table 7. The model A6 was based on the RBFN method based sensitivity analysis and decision tree was implemented for each level as shown in Table 7. There are different numbers of symptoms/parameters in each level for three different types of neural network based sensitivity analysis. In Table 7, there were three rows corresponding to each model. Each row was further divided into two parts: the first part contains value and the second part contains accuracy with the number of parameters given in brackets at level.

The test dataset results shown in Table 7 for A4 decision tree model based on quick method based sensitivity analysis indicate that level D1 and D2 produce nearly similar results i.e. 100% and 97.3%. The results for level D3, D4 and D5

Table 6: Test dataset accuracy for different levels

Level Model	D1	D2	D3	D4	D5
A1	0.000200075	0.001442433	0.002063612	0.002684791	0.00330597
	98.2% (38)	95.5% (14)	73.87% (7)	56.76% (3)	36.04% (1)
A2	0.000248337	0.028981841	0.043348593	0.057715345	0.0720821
	95.5% (38)	97.3% (10)	59.46% (4)	40.54% (2)	44.14% (1)
A3	0.009179	0.02561632	0.03383498	0.04205364	0.0502723
	100% (38)	100% (26)	97.3% (17)	94.59% (5)	45.95% (1)

Table 7: Test dataset accuracy for decision tree models

Level Model	D1	D2	D3	D4	D5
A4	0.000200075	0.001442433	0.002063612	0.002684791	0.00330597
	100% (38)	97.3% (14)	79.28% (7)	62.16% (3)	42.34% (1)
A5	0.000248337	0.028981841	0.043348593	0.057715345	0.0720821
	100% (38)	91.89% (10)	54.95% (4)	38.74% (2)	44.14% (1)
A6	0.009179	0.02561632	0.03383498	0.04205364	0.0502723
	100% (38)	92.79% (26)	95.5% (17)	97.3% (5)	45.05% (1)

not in acceptable limit for diagnosis so there was no need to consider these levels. Similarly, A4 model based on dynamic method produced acceptable accuracy for level D1 and D2 i.e. 100% and 91.89% but level D3, D4 and D5 results are not useful for diagnosis because accuracy was <60%. In RBFN based A5 model, level D1 produced 100% accuracy and level D2, D3 and D4 accuracy were >90%. So, Level D2, D3 and D4 can consider for diagnosis but level D5 accuracy was not in acceptable limit i.e. <50%.

6.3 Comparative View of The Parameters

Table 8 showed comparative view of symptoms/parameters belongs to different levels. Level D1 contains all 38 symptoms for all the models. In level D2, RBFN contains maximum number of symptoms i.e. 26 and dynamic and quick contains 10 and 14 parameters, respectively. Decreasing trend observed for D3, D4 and D5 level. Table 8 also showed the common parameters between two or more methods. Dynamic-RBFN and RBFN-Quick for level D2 showed almost 10 common parameters. All three methods commonness shows only one common parameter at D2 level.

Table 8: Number of symptoms in different level

Level Model	D1	D2	D3	D4	D5
Quick	38	14	7	3	1
Dynamic	38	10	4	2	1
RBFN	38	26	17	5	1
Quick-Dynamic	38	2	-	-	-
Dynamic-RBFN	38	9	-	-	-
RBFN-Quick	38	10	4	-	-
Quick-Dynamic-RBFN	38	1	-	-	-

· It is observed from Table 8 that level D1 and D2 shows significant involvement in the design and accuracy of all the models. Only RBFN shows acceptable accuracy for D3 and D4 levels.

· The number of parameters commonness can be observed for only for level D2 but difference between their types was also observed.

7. COMPARISON OF DIFFERENT MODELS

Table 9 described the comparative view of different models for all the levels. Comparison was based on criteria as shown in second column of Table 9. Model A1, A2 and A3 represent neural network based analysis and model A4, A5 and A6 represent decision tree based analysis. The criterion difference between neural network and decision tree model was in terms of architecture.

In general, number of symptoms play very important role in the accuracy of a particular model. Due to the different numbers of parameters at different levels, it affects the neural network architecture and number of neurons in input and hidden layer. Reduce architecture shows down trend in accuracy and increasing trend in misclassification. Only model A3 and A6 are producing acceptable accuracy for level D3 and D4. Level D5 showed more than 50% misclassification for all the models therefore it was useless to consider for the diagnosis.

In Figure 2, X-axis represented five different levels and Y-axis for accuracy. The six different graphs plotted for all the models. From the Figure 2, we can observed that level D1 and D2 showing all most straight line accuracy i.e. 100%

Table 9: Parameters and results of ANN and decision tree methods

Model	Criteria	D1	D2	D3	D4	D5
A1	Architecture	92-4-5	38-3-5	21-3-5	9-3-5	3-3-5
	Correct cases	109	106	82	63	40
	Wrong cases	2	5	29	48	71
	Training set	222	222	222	222	222
	Accuracy	98.2%	95.5%	73.87%	56.76%	36.04%
	Misclassification	1.8%	4.5%	26.13%	43.24%	63.96%
A2	Architecture	92-5-10-5	30-5-7-5	12-4-5-5	6-2-2-5	3-2-2-5
	Correct cases	106	108	66	45	49
	Wrong cases	5	3	45	66	62
	Training set	222	222	222	222	222
	Accuracy	95.5%	97.3%	59.46%	40.54%	44.14%
	Misclassification	4.5%	2.7%	40.54%	59.46%	55.86%
A3	Architecture	92-20-5	74-20-5	51-20-5	15-20-5	3-3-5
	Correct cases	111	111	108	105	51
	Wrong cases	0	0	3	6	60
	Training set	222	222	222	222	222
	Accuracy	100%	100%	97.3%	94.59%	45.95%
	Misclassification	0%	0%	2.7%	5.41%	54.05%
A4	Correct cases	111	108	88	69	47
	Wrong cases	0	3	23	42	64
	Training set	222	222	222	222	222
	Accuracy	100%	97.3%	79.28%	62.16%	42.34%
	Misclassification	0%	2.7%	20.72%	37.84%	57.66%
A5	Correct cases	111	102	61	43	49
	Wrong cases	0	9	50	68	62
	Training set	222	222	222	222	222
	Accuracy	100%	91.89%	54.95%	38.74%	44.14%
	Misclassification	0%	8.11%	45.05%	61.26%	55.86%
A6	Correct cases	111	111	106	108	50
	Wrong cases	0	8	5	3	61
	Training set	222	222	222	222	222
	Accuracy	100%	92.79%	95.5%	97.3%	45.05%
	Misclassification	0%	7.21%	4.5%	2.7%	54.95%

for all six models. A downtrend can be observed after level D2. The graph showed exception for RBFN method for model A3 and A6 because the down trend started for those models after level D4. The graphs converged at level D5 with almost 50% accuracy therefore level D5 was not producing and significant contribution for diagnosis.

8. MID-LEVEL ANALYSIS

Parameters involved at mid-level for the diagnosis can be obtained from applying formula in Table 4. The mid-level values for each ANN methods was calculated from the formula i.e. $D_{(Mid)}$ = (Highest Value – Lowest Value)/2. The

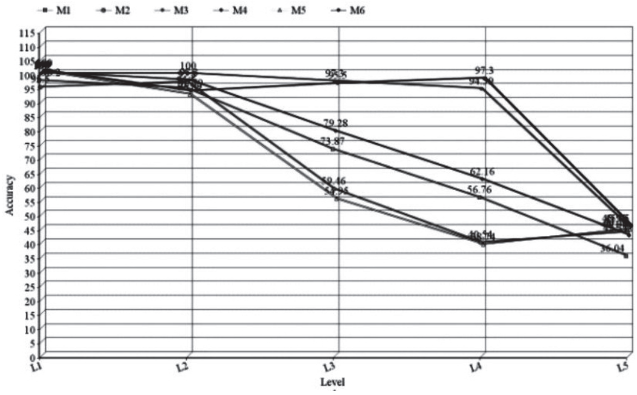


Figure 2: Level vs model graph

mid-level value for each ANN method is shown in Table 10.

Table 10: Mid-level values

Quick	Dynamic	RBFN
0.001552947	0.03591688	0.02054665

On the basis of mid value obtained form Table 10. The number of parameters excreted form Table 4 was based on the value obtained for each method and shown in Table 11. From the Table 11 we can observed that Quick method will be based on 12 parameters. Dynamic method will be based on eight parameters. RBFN method will be based on 30 parameters. Table 10 was also showing the common parameters in more than one method. Maximum common parameters were obtaining in RBFN and Quick method i.e. 12 parameters.

The accuracy at mid-level for all the reduced parameter based models was shown in Table 12. The first column of Table 12 showed the ANN method that performs sensitivity analysis. The column second and third contains accuracy for decision tree artificial neural network models. The values in the brackets were number of parameters.

Table 12: Mid-level accuracy of different models

Method	Decision Tree Accuracy	ANN Model Accuracy
Quick	A4: 94.59% (12)	A1: 94.59% (12)
Dynamic	A5: 96.4% (8)	A2: 94.59% (8)
RBFN	A6: 100% (30)	A3: 100% (30)

In mid-level analysis, model A3 and A6 produced 100% accuracy. Model A1, A2 and A4 produced lowest accuracy. From the Table 12, we can observe that number of parameters in RBFN was significant for the diagnosis that's why its accuracy was 100% for both the models.

9. DECISION TREE AND ARTIFICIAL NEURAL NETWORK COMBINATION

In this analysis the parameters obtained from the decision tree analysis used to train and test three ANN methods. Table 13 showed the result obtained from different ANN methods.

Table 13: Test result of decision tree combination with ANN

Decision Tree Combined With	Test Result (%)
Quick	100
Dynamic	100
RBFN	100

We can obtain diagnosis from decision tree as shown in Figure 1. The decision tree construction depends upon four parameters. These four parameters used for construction of three types of ANN method. The combined decision tree and ANN analysis result shown in Table 13. It was observed from Table 13 that all ANN produced 100% test data accuracy because these four parameters were the important parameters among 38 parameters for deciding the design of decision tree. Same dependency of ANN method on these four parameters was showing 100% accuracy. Decision tree can also be used as filtration techniques other than sensitivity analysis for parameters filtration and these

Table 11: Number of parameters at mid-level

Quick	Dynamic	RBFN	Quick-Dynamic	Dynamic-RBFN	RBFN-Quick	Quick-Dynamic-RBFN
12	8	30	2	8	12	2

filtered parameters further used by different ANN methods for analysis.

10. CONCLUSION

Data mining analysis showed the importance of parameters that actually involve in the diagnosis process. The efficiency of the diagnosis process depends upon the number of parameters contributed for the detection of diseases. Three ANN method analysis showed that RBFN was the best method to produce efficiency on the reduced and non-reduced parameters. Sensitive analysis combination with ANN as well as decision tree result analysis shows that level D3 and above produced unacceptable results. Decision tree combination with different ANN methods only used four parameters that showed the test cases accuracy 100% but not medically reliable for diagnosis. These methods showed the efficiency in terms of number of parameters involved in the classification of five neuropsychiatric diseases but unable to produce reasoning for the diagnosis. In medical computing reasoning can increased confidence level for diagnosis. So to overcome with this problem there was a need of another method that use combination of rule based reasoning and case based reasoning for the diagnosis of neuropsychiatric diseases.

REFERENCES

[1] Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: An overview. AI magazine, 1992;13(3):57.

[2] Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artificial intelligence in medicine, 2007;41(3):251-262.

[3] Herskovits E, Gerring JP. Application of a data-mining method based on Bayesian networks to lesion-deficit analysis. Neuroimage, 2003;19(4):1664-1673.

[4] Petriè I, Urbanèiè T, Cestnik B. Discovering hidden knowledge from biomedical literature. Informatica, 2007;31(1):15-20.

[5] Herzog H. Multiple View Framework for Highly Structured Data. VDM Verlag, 2008.

[6] Angell RL, Friedlander RR, Kraemer JR. Differential diagnosis of neuropsychiatric conditions. U.S. Publication US8388529 B2, 2013.

[7] Roy M, Kushwaha A. Mining on Medicine Data. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2012;1(7):16-22.

[8] Quinlan J. Ross. C4. 5: programs for machine learning. Morgan kaufmann, 1993.

[9] Haykin S, Network N. A comprehensive foundation. Neural Networks, 2, 2004.

[10] Prince J, Euliano N, Lefebre W. Neural and Adaptive System. John Wiley & Sons Inc., New York, 2000.

[11] Chang Chun-Lang, Chen Chih-Hao. Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Systems with Applications, 2009;36(2):4035-4041.

[12] Miyoshi K, Morimura Y. Clinical Manifestations of Neuropsychiatric Disorders. In: Neuropsychiatric Disorders. Springer Japan, 2010. pp. 1-14.

[13] Taniguchi E, Kawaguchi T, Sakata M, Itou M, Oriishi T, Sata M. Lipid profile is associated with the incidence of cognitive dysfunction in viral cirrhotic patients: A data mining analysis. Hepatology Research, 2013;43(4):418-424. doi: 10.1111/j.1872-034X.2012.01076.x

[14] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes, 2011;4(1):299.

[15] Sigurðsson HM. Data mining brain regions with neuroimaging databases. 2012.. http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6268/pdf/imm6268.pdf (accessed on 10 November 2013)

[16] Maia TV, Mcclelland JL. A neurocomputational approach to obsessive-compulsive disorder. Trends in cognitive sciences, 2012;16(1):14-15.

[17] American Psychiatric Association, *et al*. The Diagnostic and Statistical Manual of Mental Disorders: DSM 5. Bookpoint US, 2013.

[18] http://www.mayoclinic.com (accessed on 10 November 2013)

[19] World Health Organization *et al*. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. 1992.

[20] http://synapse.org.au/get-the-facts/an-overview-of-mood-disorders-fact-sheet.asp. (accessed on 10 November 2013)