Application of Artificial Intelligence to the Diagnosis of Schizophrenia

A thesis submitted towards partial fulfilment of the requirements for the degree of

Master of Engineering in Biomedical Engineering

Submitted by

Moumita Paul

Class Roll No.: 001630201001

Examination Roll No.: M4BMD18003

Registration No.: 71080 of 1998-99

Under the guidance of

Dr. Monisha Chakraborty

Associate Professor

School of Bio-Science and Engineering

Jadavpur University

Kolkata - 700032

Course affiliated to

Faculty of Interdisciplinary Studies, Law and Management

Jadavpur University

Kolkata-700032

India

2018

M.E. (Biomedical Engineering) course affiliated to
Faculty of Interdisciplinary Studies, Law and Management
Jadavpur University
Kolkata, India

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled **“Application of Artificial Intelligence to the Diagnosis of Schizophrenia ”** is a bonafide work carried out by **Moumita Paul** under my supervision and guidance for partial fulfilment of the requirement of Master of Engineering (Biomedical Engineering) in School of Bio-Science and Engineering , during the academic session 2016-2018.

------------------------------------
THESIS ADVISOR
Dr. Monisha Chakraborty
Associate Professor,
School of Bio-Science and Engineering
Jadavpur university,
Kolkata-700 032

------------------------------------
DIRECTOR
School of Bio-Science and Engineering
Jadavpur University,
Kolkata-700 032

------------------------------------
DEAN
Faculty Council of Interdisciplinary Studies, Law and Management
Jadavpur University,
Kolkata-700 032

M.E. (Biomedical Engineering) course affiliated to
Faculty of Interdisciplinary Studies, Law and Management
Jadavpur University
Kolkata, India

CERTIFICATE OF APPROVAL **

This foregoing thesis is hereby approved as a credible study of an engineering subject carried out and presented in a manner satisfactorily to warranty its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not endorse or approve any statement made or opinion expressed or conclusion drawn therein but approve the thesis only for purpose for which it has been submitted.

____

Dr. MONISHA CHAKRABORTY
Thesis Advisor
Associate Professor,
School of Bio-Science and Engineering,
Jadavpur University
Kolkata - 700032

____

Signature of Examiner

** Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of his **Master of Engineering (**Biomedical Engineering) studies during academic session 2016-2018.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by this rules and conduct, I have fully cited and referred all material and results that are not original to this work.

NAME: MOUMITA PAUL

CLASS ROLL NO.: 001630201001

EXAMINATION ROLL NO.: M4BMD18003

REGISTRATION NO.: 71080 OF 1998-99

THESIS TITLE: APPLICATION OF ARITIFICIAL INTELLIGENCE TO THE DIAGNOSIS OF SCHIZOPHRENIA

SIGNATURE: DATE:

Acknowledgements

First and foremost, I would like to thank my thesis advisor, Dr. Monisha Chakraborty, Associate Professor, School of Bio-Science and Engineering, Jadavpur University, without whose support, help, cooperation and encouragement, I could not have completed this work. In particular, I would like to express my deepest gratitude for her role in creating an environment that encourages one to engage in research, and the fact that she always came across as approachable, sincere and helpful.

I would also like to thank Dr. Piyali Basak, the director of the School of Bio-Science and Engineering for her cooperativeness and support.

I wish to thank my friends Pratik, Purbanka, Nilotpal, Nisha, Gouranga, Tathagata and Sumant for their encouragement when things did not go as well as expected.

Next I would like to thank my parents who are the two pillars of support on which I rely most.

And last but not the least, I want to sincerely thank Dr. Asish Mukhopadhyay, M.D., and Dr. Subhadeep Bharati, M.D. for their valuable contribution towards helping me understand what schizophrenia really is, how it affects the patients and how it may be diagnosed.

Moumita Paul

Date:

Contents

List of Figures

List of Tables

1. Introduction

1.1. Schizophrenia

Schizophrenia is the most common serious mental disorder with variable signs and symptoms and includes changes in cognition, emotion, perception, thinking and behaviour [**1**]. The expressions of these symptoms may vary from patient to patient but the effects are always severe and often last for as long as the patient lives. Although schizophrenia is discussed as a single disease, it is really comprised of a group of related disorders with varying aetiologies and schizophrenic patients have varying clinical presentations, treatment response and course of illness.

Written descriptions of symptoms commonly observed today in patients with schizophrenia are found throughout history. Early Greek physicians described symptoms of grandeur, paranoia and deterioration of cognitive function and personality. It was only in the 19th century that schizophrenia was recognized as a medical condition that warranted study and treatment.

1.1.1. Genetic Factors

There is a genetic contribution to some, probably, all forms of schizophrenia. First degree relatives of individuals afflicted with schizophrenia and more likely to develop the condition than second degree relatives. In case of monozygotic twins with identical genetic makeup, the chances of both twins getting the disease is about 50%. However, this data proves that genetic factors alone do not have a role to play in the development of schizophrenia, and environmental and other factors are equally significant. Had it not been the case, then the chances of both twins developing schizophrenia would have been 100%. According to some studies, the age of the father may play a decisive role in the child developing schizophrenia. Children born to fathers who are over 60 years of age are more at risk of developing the condition. Presumably, spermatogenesis in older men is susceptible to greater epigenetic damage than in younger men.

The modes of genetic transmission in schizophrenia is unknown but several genes appear to make a contribution to schizophrenia vulnerability. Linkage and association genetic studies have provided strong evidence for nine linkage sites: 1q, 5q, 6p, 6q, 8p, 10p, 13q, 15q and 22q. Further analyses of these chromosomal sites have led to the identification of specific candidate genes, and the best candidate genes are alpha-7 nicotinic receptor, DISC1, GRM3, COMT, NRG1, RGS 4, G72. Recently, mutations of the gene dystrobrevin (DTNBP1) and neureglin 1 have been associated with negative features of schizophrenia.

1.1.2. Biochemical Factors

Dopamine Hypothesis: According to the simplest formulation of the dopamine hypothesis of schizophrenia, the symptoms arise because of too much dopaminergic activity in the brain. This increased activity may be due to too much dopamine secretion, too many dopamine receptors or hypersensitivity of the dopaminergic receptors in the brain, or a combination of these mechanisms. The mesocortical and mesolimbic tracts of the dopamine receptors are most likely to be involved.

Excessive dopamine release in patients with schizophrenia has been linked to the severity of positive psychotic symptoms. There have been reports of an increase in dopamine receptors in the caudate nucleus of drug-free patients with schizophrenia. There have also been reports of increased dopamine concentration in the amygdala, decreased density of the dopamine transporter and an increase number of dopamine type-4 receptors in the entorhinal cortex.

Serotonin: Current hypothesis posits that serotonin excess is a cause of both positive and negative symptoms of schizophrenia.

Norepinephrine: Anhedonia – the impaired capacity for emotional gratification and decreased ability to experience pleasure – has long been associated with schizophrenia. A selective neuronal generation within the norepinephrine reward neural system could account for this particular symptom of schizophrenia.

GABA: The inhibitory amino acid gamma-aminobutyric acid (GABA) has been implicated in the pathophysiology of schizophrenia based on the finding that some patients with schizophrenia have a loss of GABAergic neurons in the hippocampus.

Neuropeptides: Neuropeptides, such as substance P, and neurotensin are localized with catecholamine and indolamine neurotransmitters and influence the action of these neurotransmitters. Alteration in neuropeptide mechanisms could facilitate, inhibit or otherwise modify the pattern of firing of these neuronal systems.

Glutamate: The hypotheses proposed about glutamate included those of hyperactivity, hypo-activity and those of glutamate induced toxicity.

Acetylcholine and Nicotine: Decreased muscarinic and nicotinic receptors in the caudate-putamen, hippocampus and selected regions of the pre-frontal cortex may lead to schizophrenia because these receptors play a role in the regulation of neurotransmitter systems involved in cognition.

1.1.3. Neuropathology

Researchers have discovered what may be a neuropathological basis for schizophrenia, primarily in the limbic system and basal ganglia, including neuropathological and neurochemical abnormalities in the cerebral cortex, the thalamus and the brain-stem. The loss of brain volumes widely reported in schizophrenic brains appears to result from reduced density of the axons, dendrites and synapses that mediate associative functions of the brain. Synaptic density is highest at age 1 year and is pared down to adult values in early adolescence. One theory holds that schizophrenia results from excessive pruning of synapses during this period.

Cerebral ventricles: Schizophrenia patients have consistently shown lateral and third ventricular enlargement and some reduction in cortical volume. Reduced volumes of cortical gray matter have been demonstrated in the earliest stages of the disease. There is some debate regarding whether the abnormalities are progressive or static. Whether an active pathological process is continuing to evolve in patients with schizophrenia is uncertain.

Reduced Symmetry: There is reduced symmetry in several brain areas of the patient including the temporal, frontal and occipital lobes. This reduced symmetry is believed by some researchers to originate in foetal life and to be indicative of a disruption in brain lateralization during development.

Limbic system: Because of its role in controlling emotions, the limbic system has been hypothesized to be involved in the pathophysiology of schizophrenia. There is a decrease in the size of the region including the amygdala, the hippocampus and the parahippocampal gyrus. The hippocampus is not only smaller in size in schizophrenia but is also functionally abnormal as indicated by disturbances in glutamate transmission. Disorganization of the neurons in hippocampus has also been seen in brain tissue sections of schizophrenia patients.

Prefrontal cortex: There is considerable evidence from post-mortem brain studies that supports anatomical abnormalities in the prefrontal cortex in schizophrenia. Functional deficits in the prefrontal brain-imaging region have also been demonstrated. It has long been noted that several symptoms in schizophrenia mimic those found in persons in prefrontal lobotomies or frontal lobe syndromes.

Thalamus: Some studies show evidence of volume shrinkage or neuronal loss in particular sub-nuclei. The medical dorsal nucleus of the thalamus, which has reciprocal connections with the prefrontal cortex has been reported to contain a reduced number of neurons. The total number of neurons, dendrocytes, oligodendrocytes has been shown to be reduced by 30 to 45 percent in schizophrenia patients.

Basal ganglia and cerebellum: Many patients with schizophrenia show odd movements which include an awkward gait and facial grimacing. Basal ganglia and cerebellum are involved in the control of movements. Neuropathological studies of the basal ganglia have produced reports about cell loss or reduction in volume of the globus pallidus or substantia nigra. Studies have also shown an increase in the number of D2 receptors in the caudate, the putamen and the nucleus accumbens.

1.2. Diagnosis of Schizophrenia

According to the Diagnostic and Statistical Manual of Mental Disorders, 5th edition [**2**], the presence of hallucinations or delusions is not necessary for a diagnosis of schizophrenia. A subject can be diagnosed as schizophrenic when he / she exhibits two of the symptoms in Criteria A, below. Criterion B requires that impaired functioning, although not deterioration be present in the active phase of the illness. Symptoms must persist for at least 6 months and a diagnosis of schizoaffective disorder or mood disorder be absent.

Criterion A

1. Delusions
2. Hallucinations
3. Disorganized Speech
4. Grossly disorganized or catatonic behaviour
5. Negative Symptoms

Criterion B

For a significant portion of time since the onset of the disturbance, level of functioning in one or more major areas of functioning, such as work, interpersonal relations, or self-care, is markedly below the level achieved before the onset (or when the onset is in childhood or adolescence, there is failure to achieve interpersonal, academic or occupational functioning).

1.3. The Positive and Negative Syndrome Scale (PANSS)

One of the most popular and effective way of diagnosing schizophrenia is a psychometric test. The results of the test are captured on the Positive and Negative Syndrome Scale, which is a medical scale used for measuring symptom severity in patients with schizophrenia. The name refers to the two types of symptoms in schizophrenia as defined by the American Psychiatrist Association – the positive symptoms which describe an excess of distortion of normal functions, viz. delusions and hallucinations and negative symptoms which represent a diminution or loss of normal functions. The PANSS is a relatively brief interview and it takes approximately 45 to 50 minutes to administer [**3**]. The interviewer must be trained to a standardized level of reliability. At the end of the interview, the patient is rated on a scale of 0 to 6 on thirty different aspects on the basis of answers given during the interview as well as feedback from family members and hospital primary care workers.

Positive Scale (7 items, minimum score= 0, maximum score = 42):

- Delusions
- Conceptual Disorganization
- Hallucinations
- Excitement
- Grandiosity
- Suspiciousness / persecution
- Hostility

Negative Scale (7 items, minimum score = 0, maximum score = 42)

- Blunted affect
- Emotional withdrawal
- Poor rapport
- Passive / apathetic social withdrawal
- Difficulty in abstract thinking
- Lack of spontaneity / flow of conversation
- Stereotyped thinking

General Psychopathology Scale (16 items, minimum score = 0, maximum score = 96)

- Somatic concerns

- Anxiety
- Guilt feelings
- Tension
- Mannerisms and posturing
- Depression
- Motor retardation
- Uncooperativeness
- Unusual thought content
- Disorientation
- Poor attention
- Lack of judgement and insight
- Disturbance of volition
- Poor impulse control
- Preoccupation
- Active social avoidance

Originally the PANSS was designed to assign a rating from 1 to 7 on the various items. However this meant that patients without any symptoms would have a rating of 30 instead of 0. This led to misrepresentation of the effectiveness of atypical antipsychotic medication when tested on schizophrenics. This is why, nowadays, many psychiatrists prefer PANSS rating from 0 to 6 rather than 1 to 7 [**4**] [**5**].

1.4. High Level Design of the Work

The work is divided into five main phases. In the first phase, a fuzzy expert system is designed that takes as its inputs the PANSS ratings assigned to a subject, and returns a crisp rating on a scale of 0 to 2. According to the recommendations of a qualified psychiatrist, a subject is diagnosed as schizophrenic if the rating is above 1.26 and non-schizophrenic if the rating is below 1.26.

In the second phase of the work, a dataset is synthesized for training an artificial neural network for diagnosing schizophrenia.

In the third and fourth phases of the work, the synthetic dataset is used to train a multilayer perceptron and a support vector machine respectively, and the observations are noted.

In the fifth phase, the synthetic dataset is fuzzy clustered and the result is noted with data from real subjects.

1.5. Motivation Behind the Work

There is a lot of stigma attached to mental illness. Many a times, subjects are reluctant to consult psychiatrists in this matter. Also diagnosis of mental illness is tricky because there is so much overlap of symptoms of various different illnesses. In case of an ailment like cancer, a subject may be completely free of symptoms (pain, rapid weight loss, etc.) but still be diagnosed with the disease on account of physiological changes (presence, shape and growth pattern of tumour). In case of a mental illness like schizophrenia, however, diagnosis is based entirely on symptoms. A subject may have all the biomarkers of schizophrenia but still not be diagnosed as such if he or she has no symptoms. Thus it is useful to have a symptom based tool for diagnosing the condition. The fuzzy expert system may serve to offer a second opinion to the practising psychiatrist as to whether or not a subject is afflicted with schizophrenia.

2. Literature Review

There is a lot of interest in the application of artificial intelligence to the diagnosis of psychiatric disorders including schizophrenia.

In the early 1960's and 1970's interview based screening techniques were employed in order to perform psychiatric data analysis. Wing and Giddens [**6**]proposed a tool called "Present State Examination" for rating psychiatric symptoms reported by patients. The "Minnesota Multiphasic Personality Inventory" – a computer program written by Keinmuntz [**7**]could automatically interpret some psychiatric illnesses. However, the above tools have the drawbacks that they are rigid, time-consuming, have a high chance of human error and involve a rigid and monotonous questionnaire based data feed. Heiser and Brooks proposed a tool named HEADMED [**8**]based on a full-fledged questionnaire regarding the nature, severity and course of the symptoms, data processing algorithms, statistical data manipulation technique and inferential capacity according to the input given to it. Besides the prescription, HEADMED was also able to advise on optimal dosing, best route of administration and possible adverse effects of psychiatric drugs. Mulsant and Servan-Schreiber developed the BLUE BOX [**9**], which could diagnose various kinds of nervous breakdowns like depression and come up with a treatment plan after taking into consideration the possibility that the patient might attempt suicide. Johri and Guha designed an expert tool based on a set covering model using a diagonal search method, elicitation system and abduction knowledge [**10**]. Petrovic developed a tool for clustering the behavioural and psychological symptoms in dementia using Spearman's correlation analysis and Principal Component Analysis [**11**].

The above expert systems were designed primarily for research with little scope of practical application. These systems had several drawbacks in that they required large amounts of valid data and that the self-learning procedure was complex [**12**].

In 1996, Zou et al [**13**]used the back-propagation neural network as well as the Kohonen network to fit psychiatric diagnoses. The networks were trained to classify neurosis, schizophrenia and normal people. In 2005, Aruna et al [**14**]proposed a neuro-fuzzy model for the diagnosis of psychosomatic disorders. The symptoms and signs were obtained by interviewing patients and fuzzy membership values were determined for the inputs. The fuzzy values were fed as input to a feedforward multi-layer neural network which was trained using the back-propagation training algorithm. The trained network was then tested with symptoms and signs from another set of patients. The performance of the model was compared with a medical expert and was also compared with probability model based on Bayesian belief network and statistical model based on Linear Discriminant Analysis. Also, in 2005, Li et al [**15**] proposed a method of classification of schizophrenia and depression by EEG with ANNs. They showed that EEG rhythms can be used to distinguish among individuals suffering from schizophrenia and/or depression and normal healthy subjects. Back-propagation ANN and self-organizing competitive ANNs were used to discriminate among three kinds of subjects. The EEG rhythms were used as feature vectors. They also compared the two different kinds of ANNs and demonstrated that back-propagation ANNs have better performance than self-organizing ANNs. In 2009, Chattopadhyay, Pratihar and De Sarkar developed a fuzzy logic based screening and prediction tool for adult psychoses, a group of similar mental illnesses showing various cause-effect relationships among patients [**12**].

Several works have applied pattern recognition to fMRI data for schizophrenia diagnosis. In 2003, Ford et al [**16**] reduced the dimensionality of fMRI statistical spatial maps using Principal Component Analysis (PCA), and then differentiated between controls and patients with schizophrenia, brain injury and Alzheimer's disease by applying Fisher's linear discriminant. In 2003, Cox and Savoy [**17**] applied linear discriminant analysis and a linear support vector machine (SVM) analysis to classify among 10-class visual patterns. In 2004, Wang et al distinguished between brain cognitive states using a linear SVM. In 2006, Martinez-Ramon et al used SVMs for 4-class interleaved classification. In 2003, LaConte et al used a linear SVM for left and right motor activation. In 2006, Shinakreva et al [**18**] used whole brain fMRI time series and identified voxels which had highly dissimilar time courses among groups employing the RV-coefficient. Once those voxels were detected, their fMRI time series data was used for subject classification.

One of the main difficulties of using pattern recognition in fMRI is that each collected volume contains tens of thousands of voxels. This means that the dimensionality of the fMRI data is very high when compared with the number of data points, ie, the number of images collected from subjects, which may be of the order of tens or hundreds. This great difference between data dimensionality and the number of available observations means that the generalization performance of the estimator (classifier or regressor) goes down, and in some cases, the estimator cannot be used at all. This is known as "the curse of dimensionality". Thus it is desirable to reduce the dimensionality in a way that incurs the least loss of information with an affordable computational burden [**19**].

Two approaches to solve this problem are feature extraction and feature selection. PCA is the most popular method of feature extraction and it extracts the important features by casting high dimensional data into a low dimensional space. In 2005, Mourao-Miranda et al [**20**] used PCA for whole brain classification during fMRI attention experiments. The second approach is feature selection, which determines a subset of features that optimizes the performance of the classifier. The latter approach is useful for fMRI under the assumption that the information in the brain is sparse, that is, the useful information is concentrated in only a few areas of the brain making the rest of the areas irrelevant for classification tasks. In addition, feature selection can improve the prediction performance of the classifier as well as provide a better understanding of the underlying process that generated the data. Feature selection methods can be divided into three categories: filters, wrappers and embedded methods. Filters select a subset of features as a preprocessing step to classification. On the other hand, embedded methods and wrappers use the classifier itself to find the optimal feature set. Wrappers make use of the learning machine to select the feature set that increases its prediction accuracy whereas embedded methods incorporate feature selection as part of the training phase of the learning machine. In 2006, Mourao-Miranda et al [**21**] used the filter approach in their work. In 2005, Haynes and Rees [**22**] also applied filter feature selection by selecting the top 100 voxels that had the strongest activation in two different visual stimuli. In 2008, De Martino et al [**23**] used a hybrid filter / wrapper approach by applying univariate voxel selection strategies prior to using recursive feature elimination SVM (RFE-SVM). RFE-SVM is robust but computationally intensive since it eliminates features one at a time in each iteration – this requires the SVM to be trained M times for M-dimensional data.

While it is possible to remove several features at a time, it comes at the expense of classification performance degradation. In 2010, Ryali et al [**24**] presented an alternate approach that incorporates the use of embedded feature selection methods. The disadvantage of this method is that it achieves only average classification accuracy when applied to real fMRI data. Since multivariate non-linear feature selection is computationally expensive, usually only linear methods are applied for doing feature selection in fMRI. A trade-off is region based discrimination – such an approach assumes that voxels that are close to each other and are part of the same region of brain are non-linearly related, while voxels is different brain regions are linearly related.

In 2011, Castro et al [**25**] proposed a method of detecting schizophrenia in subjects by pattern classification of brain imaging data. Brain imaging data typically has high dimensionality. Their work proposes the application of recursive feature elimination using a machine learning algorithm based on composite kernels to the classification of healthy controls and patients with schizophrenia. The framework analyzes whole brain fMRI data that is segmented into anatomical regions and recursively eliminates the uninformative ones based on their relevance estimates, thus yielding the set of most discriminative brain areas for group classification. The collected data was analyzed using two methods – General Linear Method (GLM) and Independent Component Analysis (ICA).

Neural networks have been extensively used in the diagnosis of diseases like diabetes, chest diseases and urological dysfunction. In the 2003 work of Kayaer and Yildrim, [**26**] the 2005 work by Delen, Walker and Kadam [**27**], and the 2009 work by Temurtas [**28**], multi-layer neural networks have replaced conventional pattern recognition methods of disease diagnostic systems. The back-propagation algorithm is the most popular method of training the neural network, but as demonstrated by Brent in 1991 [**29**] and Gori and Tesi in 1992 [**30**], it suffers from a slow convergence rate and often yields sub-optimal solutions. A number of researchers like Gulbag and Temurtas in 2006 [**31**], Hagan, Demuth and Beale in 1996 [**32**], and Hagan and Menhaj in 1994 [**33**] have carried out comparative studies of the various different training algorithms. According to the 1994 work by Hagan and Mehnaj, Levenberg Marqurdt algorithm provides faster convergence and better estimation results than other algorithms. In 1990, Specht [**34**] developed the probabilistic neural network (PNN), which is very useful for classification problems and disease diagnostic systems.

Carpenter and Markuzon in 1998 [**35**], Deng and Kasabov in 2001 [**36**], Kayaer and Yildrim in 2003 [**26**] and Polat and Gunes in 2007 [**37**] performed studies focussing on diabetes disease diagnosis for Pima Indians.

Neural networks classification has been used for diagnosis of chest diseases too. Aliferis, Hardin and Massion in 2002 [**38**], Ashizawa et al in 2005 [**39**], Copini, Miniati, Paterni, Monti and Ferdeghini in 2007 [**40**], El-Solh, Hsiao, Goodnough, Serghani and Grant in 1999 [**41**], Er, Sertkaya, Temurtas and Tanrikulu in 2009 [**42**], Er and Temurtas in 2008 [**43**], Er and Tanrikulu in 2010 [**44**], Hanif, Lan, Daud and Ahmed in 2009 [**45**], Paul, Ben, Thomas and Robert in 2004 [**46**], dos Santos, Pereira and de Seixas in 2004 [**47**] have all carried out studies on the diagnosis of chest diseases with artificial neural networks. These studies have applied different neural network structures to the various chest diseases diagnosis problem and achieved high classification accuracies using different datasets.

In 2004, Paul et al used the MLNN with one and two hidden layers and used back-propagation with momentum as their training algorithm for predicting community acquired pneumonia for patients with chest problems [**46**]. In 2002, Heckerling reported an accuracy ration of 82.2% for pneumonia diagnosis [**48**]. In 2009, Hanif et al used three different neural networks to classify the severity of asthma and the suitable control measures to overcome it [**45**]. These neural networks were feedforward backpropagation neural networks, Elman's neural network and Radial Basis Function neural network. In 2002, Aliferis et al used KNN decision tree induction, feedforward neural networks and support vector machines to classify lung tumours [**38**].

Human diseases can be detected at an early stage and subsequently treated with the help of expert systems. Fuzzy expert systems have been used to detect several illnesses like prostate cancer, cardiac diseases, jaundice, coronary artery disease, malaria, dengue etc. In 2003, Seritas, Alhaverdi and Sert proposed a method of diagnosing prostate cancer. The system used Prostate Specific Antigen (PSA), age and prostate volume as input parameters and prostate cancer risk (PCR) as output parameter [**49**]. The system determines if there is a need for biopsy and in addition, gives the user a range of the risk of cancer diseases. These factors were fuzzified with the linguistic variables very small, small, middle, high, very high, very low and low. The Mamdani max-min inference was used for the inference system. In 2012, Smita Sikchi et al came up with a study on the use of a fuzzy inference system for diagnosis of cardiac diseases. [**50**]. They defined a set of 700 rules using the disease database as well as expert knowledge on the disease domain. They defined 11 input variables and one output variable for the fuzzy inference system. The input variables are nothing but the results of laboratory tests and manifested symptoms. Laboratory test results are converted into fuzzy compatibility values in the range of zero to unity by consideration of the linguistic medical concepts. Next the fuzzified data is used to draw an inference about the diagnosis with the help of knowledge contained in a knowledge base. Finally, defuzzification was used to obtain crisp values on an arbitrary scale of the fuzzy output variable as the risk of heart disease. The fuzzy inference system was Mamdani type and centroid based defuzzification technique was used. In 2013, Manish Rana et al proposed a fuzzy inference system model for diagnosing maladies of the human brain, viz. haemorrhages and tumours, as well as cardiac diseases and thyroid diseases [**51**]. For brain disease, the fuzzy system takes five inputs, viz, protein, red blood cells, lymphocytes, neutrophils and eosinophils, and gives three outputs, viz. normal, haemorrhage and brain tumour. The FIS for diagnosing heart diseases takes just one input viz. the CPK-MB marker which gives an indication of acute myocardial infarction, and returns an output corresponding to whether or not the subject suffers from heart disease. For diagnosing diseases of the thyroid gland, the expert system takes as input the levels of T3, T4 and TSH hormones and returns an output regarding whether or not the subject suffers from hypothyroidism. In 2014, Nitin Sahai, Deepshikha Shrivastava and Pankaj Shrivastava proposed a fuzzy expert system to diagnose jaundice [**52**]. They used both the Mamdani and Sugeno fuzzy inference systems. The symptoms of jaundice are supplied as inputs to the FES while the grade of the disease constitutes the output. In 2014, Niranjana Devi and Anto proposed an evolutionary fuzzy expert system for the diagnosis of coronary artery disease [**53**]. With the help of a decision tree, the most significant attributes are selected and the output is converted to crisp if-then rules. The crisp set of rules is transformed into fuzzy rules which constitute the fuzzy rule base. Genetic algorithm is used to tune the fuzzy membership functions which leads to better accuracy.

We have not come across any study that deals with fuzzy inference systems used for diagnosing schizophrenia on the basis of PANSS ratings. The same can be said about artificial neural networks. Though there are studies that focus on diagnosing schizophrenia with the help of ANNs that use features of functional MRI as input, these studies are not very practical since it is difficult and expensive to obtain imaging data in sufficiently large quantities. On the other hand the PANSS ratings can be obtained easily and inexpensively. Since the PANSS data that is fed to the ANN has only thirty dimensions which is modest compared with MRI data, the PANSS based diagnostic solution does not suffer from the curse of dimensionality unlike the MRI based solutions.

3. Theory

3.1. Fuzzy Systems

3.1.1. Classical Sets and Fuzzy Sets

Once we are able to define a universe of discourse which contains all possible information on a given problem, we may define certain events on the basis of this information. A set defined on this universe is a mathematical abstraction of these events as well as the universe. A classical set is defined by crisp boundaries whereas a fuzzy set is defined by fuzzy or ambiguous boundaries. This is because a fuzzy set has vague or ambiguous properties, whereas there is no room for vagueness is the world of classical sets. [**54**].

Where classical sets are concerned, a member is either a full member of the set or not a member. On the other hand, in case of fuzzy sets, we have the concept of partial membership, wherein an entity may be members of more than one set defined on that universe.

To describe classical sets, define a universe of discourse X, as a group of objects all having same characteristics. The individual elements in the universe will be denoted as x. The features of the elements in X can be either discrete countable integers, or continuous valued quantities on the real line. A useful attribute of sets and the universe on which they are defined is a parameter known as cardinality, or cardinal number. The total number of elements in a universe X is called its cardinal number, denoted as n_x. Discrete universes have finite cardinal numbers since the number of elements contained in these universes is finite; on the other hand continuous universes have an infinite number of elements. Elements in a universe may be grouped into sets, and sets may be further divided into subsets. We define a null set, Φ, as a set containing no elements, and the whole set X, as the set of all elements in the universe. The null set may be thought of as parallel to an impossible event, whereas the whole set may be thought of as parallel to a certain event. All possible sets of X constitute a special set, called the power set, P(X).

The various operations on classical sets in set-theoretic terms are:

Union: $A \cup B$

Intersection: $A \cap B$

Complement: $\bar{A}$

Difference: $A \mid B$

Properties of classical (crisp) sets - The most appropriate properties of classical sets for demonstrating their similarity to fuzzy sets are:

Commutativity: $A \cup B = B \cup A$

$$A \cap B = B \cap A$$

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

Idempotency: $A \cup A = A$

$$A \cap A = A$$

Identity: $A \cup \emptyset = A$

$$A \cap X = X$$

$$A \cap \emptyset = \emptyset$$

$$A \cup X = X$$

Involution: $\bar{\bar{A}} = A$

Two special principles of set properties are known as “excluded middle axioms” and “De Morgan’s principles”. Of all the axioms described, only the excluded middle axioms are not valid for both classical sets and fuzzy sets.

Axiom of the excluded middle: $A \cup \bar{A} = X$

Axiom of the contradiction: $A \cap \bar{A} = \emptyset$

De Morgan’s principles: $\overline{A \cap B} = \bar{A} \cup \bar{B}$

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

In general, De Morgan’s principle can be extended for n sets, as provided here for events E_i:

$$\overline{E_1 \cap E_2 \cap \dots . \cap E_n} = \overline{E_1} \cup \overline{E_2} \cup \dots \cup \overline{E_n}$$

$$\overline{E_1 \cup E_2 \cup \dots \cup E_n} = \overline{E_1} \cap \overline{E_2} \cap \dots \cap \overline{E_n}$$

In fuzzy sets there is a gradual transition from membership to non-membership, but in classical sets, the transition is abrupt. A fuzzy set, then, is a set containing elements that have varying degrees of membership in the set. Elements of a fuzzy set are mapped to a universe of membership values between 0 and 1. This is where the membership function comes in. A membership function maps elements of a fuzzy set A_f to a real numbered value on the interval 0-1. A notation convention for fuzzy sets, when the universe of discourse, X, is discrete and finite, is as follows for a fuzzy set A_f:

$$A_f = \left\{\frac{\mu_{Af}(x_1)}{x_1} + \frac{\mu_{Af}(x_2)}{x_2} + \dots\right\} = \left\{\sum_i \frac{\mu_{Af}(x_i)}{x_i}\right\}$$

The horizontal bar is a delimiter rather than a division sign. The numerator in each term is the membership value in set A_f associated with the element of the universe indicated in the denominator. The summation symbol denotes that aggregation or collection of each element and has nothing to do with algebraic summation.

Now, let us try to understand fuzzy set operations. Let us define three fuzzy set A_f, B_f, C_f on the universe of X. For a given element x of the universe, the function theoretic operations, union, intersection and complement for the set-theoretic operations of union, intersection and complement are defined for A_f, B_f and C_f: De Morgan's principle for classical sets also holds for fuzzy sets; however, the excluded middle axioms do not hold for fuzzy sets.

Fuzzy sets follow the same properties as crisp sets. Indeed classical sets can be thought of as a subset of fuzzy sets.

3.1.2. Classical Relations and Fuzzy Relations

An ordered sequence of r elements written in the form (a1, a2, a3,…,ar) is called an order-r tuple; an unordered tuple is simply a collection of r elements in which order is not significant. For crisp sets A1, A2,…Ar, the set of all r-tuples (a1, a2, …, ar), where a1 ϵ A1, a2 ϵ A2, ar ϵ Ar, is called Cartesian product of A1, A2,…Ar, and is denoted by A1xA2x…xAr. The Cartesian product of two or more sets is different from the arithmetic product of two or more sets. A subset of Cartesian product A1xA2x…xAr is called a r-ary relation over A1, A2,…Ar.

The Cartesian product of two universes X and Y is determined as:

X x Y = {(x,y) | xϵX, yϵY}, which forms an ordered pair between every x and y forming unconstrained matches between X and Y. This operation establishes a complete relationship between every element of universe X and every element of universe Y. The strength of this relationship between ordered pairs of elements in each universe is measured by the characteristic equation K, where a value of 1 indicates complete relationship and value of 0 denotes no relationship. This strength of relationship may be thought of as a mapping from ordered pairs of the universe or ordered pairs of the sets defined on the universes to the characteristic function. A matrix called relation matrix may be conveniently used to describe this relationship when the elements and sets of the universe are finite.

Cardinality of crisp relations – Suppose n element of universe X are related to m elements of universe Y. The cardinality of X is n_x and the cardinality of Y is n_y, then the cardinality of the relation R between the two universes is $n_{XxY} = n_x * n_y$. Let us define R and S as two separate relations on the Cartesian universe X x Y, and define the null relation and complete relation as the relation matrices O and E respectively. An example of the 4 x 4 form of the O and E matrices are given below:

$$O = \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix} ; E = \begin{bmatrix} 1\,1\,1\,1 \\ 1\,1\,1\,1 \\ 1\,1\,1\,1 \\ 1\,1\,1\,1 \end{bmatrix}$$

The following function theoretic operations for the two crisp relations R and S can be defined as:

Union: $R \cup S \rightarrow K_{R\cup S}(x,y) : K_{R\cup S}(x,y) = \max[\, K_R(x,y), K_S(x,y)]$

Intersection: $R \cap S \rightarrow K_{R\cap S}(x,y) : K_{R\cap S}(x,y) = \min[\, K_R(x,y), K_S(x,y)]$

Complement: $\bar{R} \rightarrow K_{\bar{R}}(x,y) : K_{\bar{R}}(x,y) = 1 - K_R(x,y)$

Identity: $\emptyset \rightarrow 0, and\ X \rightarrow E$

The properties of commutativity, associativity, distributivity, involution and idempotency all hold for crisp relations in the same way they do for classical set operations. Moreover De Morgan's principles and excluded middle axioms hold for crisp relations just as they hold for crisp sets. One may think of the null relation O, and the complete relation E as being analogous to null set Φ and whole set X respectively in the set-theoretic case.

Composition – Let R be a relation that relates or maps elements from universe X to universe Y, and let S be a relation that relates or maps elements from universe Y to universe Z. An operation called composition may be used to find a relation T, that relates the same elements in universe X that R contains to the same elements in universe Y that S contains. There are two common forms of composition operation – one is called the max-min composition and the other is called the max-product composition. The max-min operation is defined as:

T = R o S

$$K_T(x,z) = \max(\min(K_R(x,y), K_S(y,z)))$$

The max-product operation is defined as:

T = R o S

$$\max(K_R(x,y) * \ K_S(y,z))$$

Fuzzy relations use Cartesian product of two universes to map elements of one universe, say X, to those of another universe, say Y. However the strength of the relationship between the ordered pairs of the two universes is measured with membership function and not characteristic function, and various degrees of strength of the relation may be expressed on the unit interval [0,1]. Hence a fuzzy relation R$_f$ is a mapping from the Cartesian space XxY to the interval [0,1], where the strength of the mapping is expressed by the membership function of the relation for ordered pairs from the two universes or $\mu_{Rf}(x,y)$.

The cardinality of fuzzy sets on any universe is infinity, and so, the cardinality of fuzzy relations is also infinity.

Operations on fuzzy relations – Let R$_f$ and S$_f$ be fuzzy relations on the Cartesian space XxY. Then the following operations apply for the membership values of various set operations:

Union: $\mu_{R_f \cup S_f}(x,y) = \max(\mu_{R_f}(x,y), \mu_{S_f}(x,y))$

Intersection: $\mu_{R_f \cap S_f}(x,y) = \min(\mu_{R_f}(x,y), \mu_{S_f}(x,y))$

Complement: $\mu_{\overline{R_f}}(x,y) = 1 - \mu_{R_f}(x,y)$

All properties like commutativity, associativity, distributivity, involution and idempotency hold for fuzzy relations just as they hold for crisp relations. Also, De Morgan's principles hold, and the null relation O and the fuzzy relation E are analogous to the null set and the whole set in set-theoretic form respectively. Similar to fuzzy sets, the excluded middle axioms are not applicable to fuzzy relations.

Equivalence Relations and Tolerance Relations: The three most significant properties of a relation are reflexivity, symmetry and transitivity. A relation R on a universe X may be considered as a relation from X to X. The relation R is an equivalence relation if it has the properties of reflexivity, symmetry and transitivity. For example, for a matrix relation, the following properties will hold:

Reflexivity: $(x_i, x_i) \epsilon R \; or \; K_R(x_i, x_i) = 1$

Symmetry: $(x_i, x_j) \; \epsilon \; R \; \rightarrow (x_j, x_i) \; \epsilon \; R$, or,

$$K_R(x_i, x_j) = K_R(x_j, x_i)$$

Transitivity: $(x_i, x_j) \epsilon \; R \; and \; (x_j, x_k) \epsilon R \; \rightarrow (x_i, x_k) \epsilon \; R$

Or, $K_R(x_i, x_j) \; and \; K_R(x_j, x_k) = 1 \; \rightarrow K_R(x_i, x_k) = 1$

A crisp tolerance relation R, which is also called a proximity relation, on a universe X is a relation that exhibits only the properties of reflexivity and symmetry but not transitivity. A tolerance relation R can be reformed into an equivalence relation by at most (n-1) compositions with itself, where n is the cardinal number of the set defining R, in this case, X.

Fuzzy tolerance and equivalence relations: A fuzzy relation R_f on a single universe X is also a relation from X to X. It is a fuzzy equivalence relation if all three of the following properties of the matrix relations define it:

Reflexivity: $\mu_{R_f}(x_i, x_i) = 1$

Symmetry: $\mu_{R_f}(x_i, x_j) = \mu_{R_f}(x_j, x_i)$

Transitivity:

$\mu_{R_f}(x_i, x_j) = \rho_1 \; and \; \mu_{R_f}(x_j, x_k) = \rho_2 \rightarrow \mu_{R_f}(x_i, x_k) = \rho, \qquad where \; \rho \geq \min(\rho_1, \rho_2)$

3.1.3. Membership Functions, Fuzzification and De-fuzzification

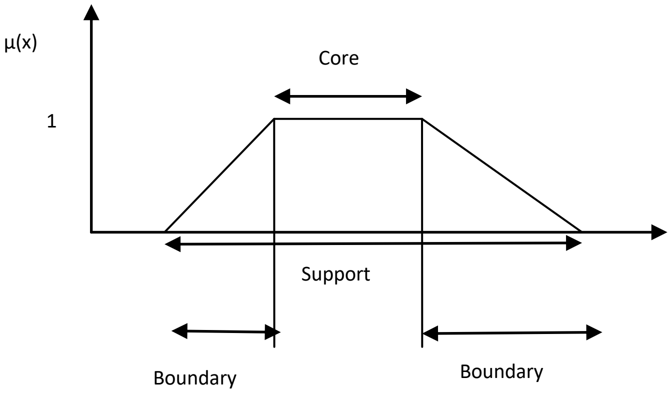

Figure 3.1.3-1: Membership Function Showing Core, Boundary and Support

As shown in the Figure 3.1.3-1, the core of a membership function constitutes those values of x for which the membership value is 1. The support of the membership function for some fuzzy set is defined as that region of the universe for which the membership value is greater than zero. The boundaries of a membership function are constituted by that region of the universe for which the membership value is greater than zero but less than one. A normal fuzzy set is one whose membership function has at least one element x in the universe whose set membership value is unity. In fuzzy sets, where only one element has membership value equal to unity, the element is generally referred to as the prototype of the set or prototypical element.

A convex fuzzy set is described by a membership function whose membership values monotonically increase or monotonically decreases or monotonically increase and then monotonically decrease with increasing values for elements in the universe. A special property of two convex fuzzy sets, A_f and B_f is that the intersection of these sets is also a convex fuzzy set.

The crossover points of a membership function are defined as the elements of the universe for which a particular fuzzy set A_f has membership values greater than 0.5. The height of a fuzzy set A_f is the maximum value of its membership function. The most common forms of a membership function are those that are normal and convex. However subnormal and non-convex membership functions are also possible. Membership functions may also be symmetrical or asymmetrical.

Fuzzification – This is the process of making a crisp quantity fuzzy. Quantities that are thought of as crisp or deterministic may in truth have an element of uncertainty associated with them. For example, hardware such as a digital voltmeter generates crisp data but the data may be subject to experimental / observational errors. A useful though not compulsory step is to represent imprecise data as fuzzy sets when that data is used in fuzzy systems.

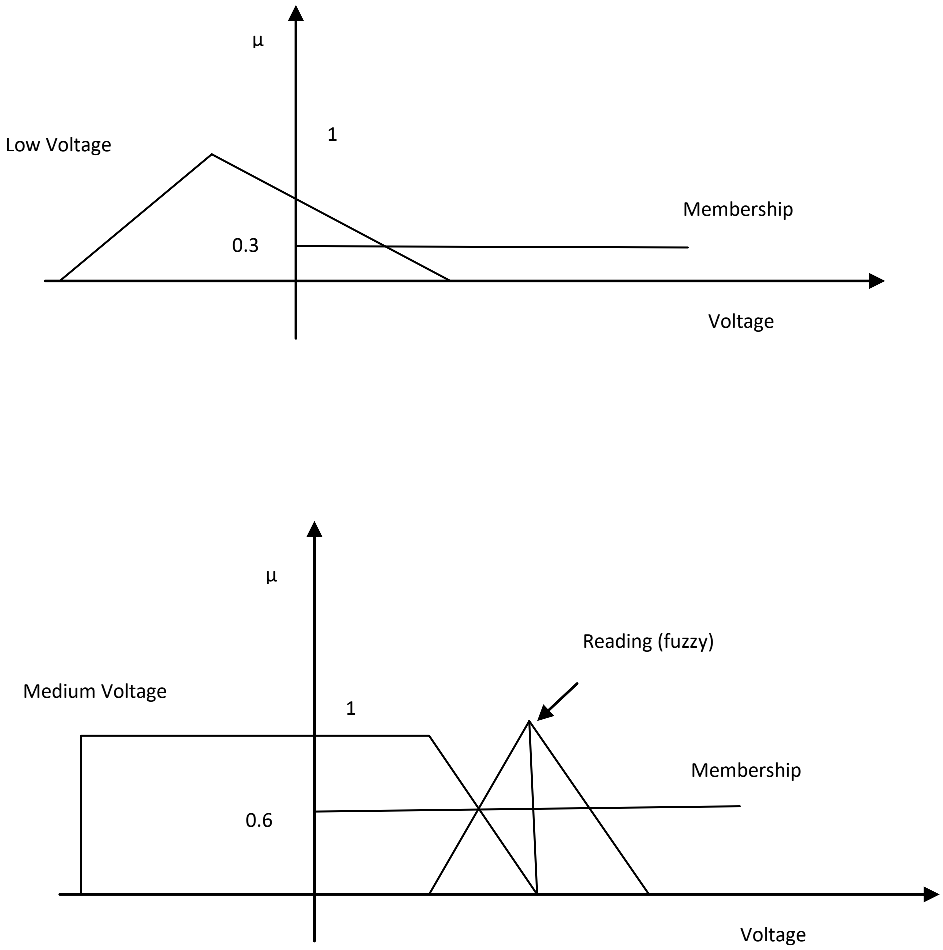

Figure 3.1.3-2: Comparison of Fuzzy Set and Crisp of Fuzzy Readings

Figure 3.1.3-2 shows the comparison of a crisp voltage reading to a fuzzy set, for example, “low voltage”. In the figure we see that the crisp reading intersects the fuzzy set at a membership value of 0.3, that is, the fuzzy set and the reading can be said to concur at a membership value of 0.3. In the lower figure, the intersection of the fuzzy set “medium voltage” and a fuzzified voltage reading occurs at a membership of 0.6. We see that the set intersection of the two fuzzy sets is a small triangle whose largest membership occurs at membership value of 0.6.

Defuzzification: The output of a fuzzy process needs to be a single scalar quantity as opposed to a fuzzy set. Defuzzification is the process of converting a fuzzy quantity to a precise quantity. The output of a fuzzy process can be the logical union of two or more fuzzy membership functions defined on the universe of discourse of the output variable. For example, let us consider a fuzzy output comprising two parts: C_{1f}, a trapezoidal shape and C_{2f}, a triangular membership shape. The unification of these two membership functions involves the max operator, which graphically, is the outer envelope of the two shapes as shown in the figure 3.1.3-3 below:

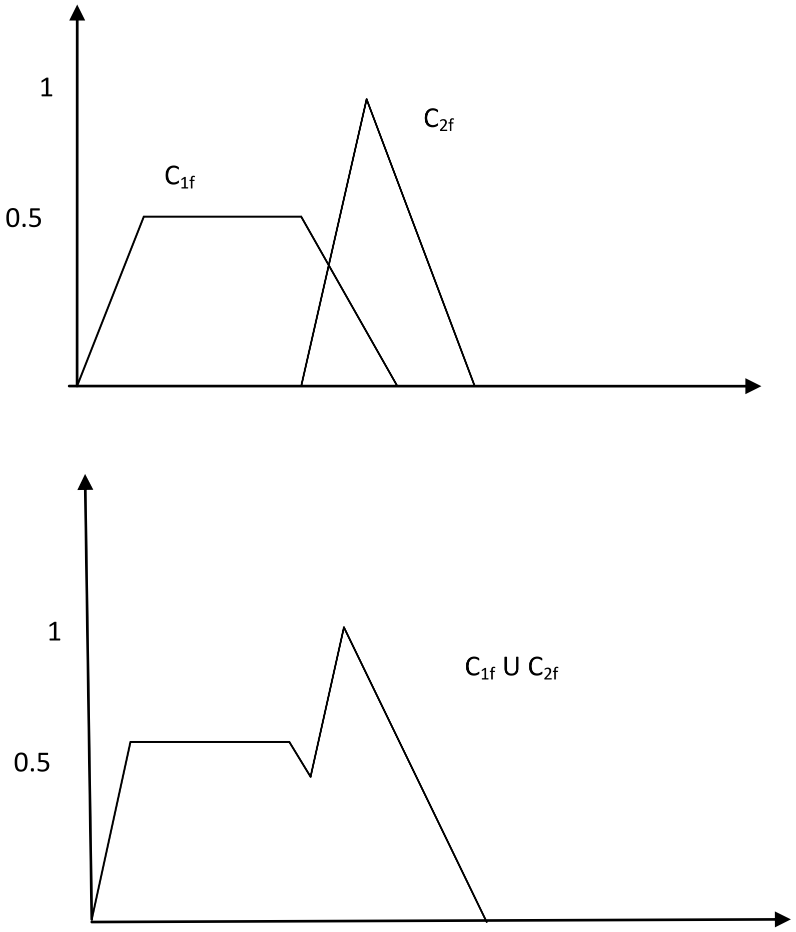

Figure 3.1.3-3: Union of Two Fuzzy Sets

A general fuzzy output process can involve many output parts (more than two), and the shapes of the membership functions representing each part of the output need not be restricted to triangular or trapezoidal. Among the many methods described in literature over the years, the following are the most popular when it comes to defuzzifying fuzzy output functions:

a) Max membership principle: Also known as the height method, this scheme is limited to peaked output functions. This method is given by the algebraic expression:

$$\mu_{C_f}(z*) \geq \mu_{C_f}(z), \qquad for\ all\ z \in Z$$

b) Centroid method: This method, also called centre of area or centre of gravity is the most commonly used and physically appealing of all de-fuzzification methods. It is given by the algebraic equation:

$$z^* = \frac{\int \mu_{C_f}(z).z\,dz}{\int \mu_{C_f}(z)dz}$$

c) Weighted average method: The weighted average method is one of the most frequently used in fuzzy applications since it is computationally efficient. It is usually restricted to symmetrical output membership functions. It is given by the algebraic expression:

$$z^* = \frac{\sum \mu_{C_f}(\bar{z}).\bar{z}}{\sum \mu_{C_f}(\bar{z})},$$

$$where\ \bar{z}\ is\ the\ centroid\ of\ each\ symmetric\ membership\ function$$

d) Mean max membership: This method, also called middle-of-maxima, is closely related to the first method, except that the locations of the maximum membership can be non-unique. This method is given by the expression,

$$z^* = \frac{a+b}{2},\ where\ a\ and\ b\ are\ the\ two\ values\ of\ z\ for\ which\ maxima\ occurs$$

e) Centre of sums: This method is not restricted to symmetric membership functions. This process involves the algebraic sum of individual output fuzzy sets, say, C_{1f} and C_{2f}, instead of their union. The de-fuzzified value is given as follows:

$$z^* = \frac{\sum_{k=1}^{n} \mu_{C_{fk}}(z) \int \bar{z}\,dz}{\sum_{k=1}^{n} \mu_{C_{fk}}(z) \int dz}$$

f) Centre of largest area: If the output set has at least two convex sub-regions, then the centre of gravity of the convex fuzzy sub-region with the largest area is used to obtain the de-fuzzified value of the output.

g) First (or last) of maxima: This method uses the overall output or union of all individual output fuzzy sets to determine the smallest value of the domain with maximized membership degree.

3.1.4. Fuzzy Logic

Consider the story of “The Barber of Seville”. In the town of Seville there is a rule that all and only those men who do not shave themselves are shaved by the barber. Who shaves the barber? This paradox can only work when the statement is both true and false at the same time. This can be shown using set notation. Let S be the proposition that the barber shaves himself and $\bar{S}$ (not S) that he does not. Then since, $S \rightarrow \bar{S}$, and $\bar{S} \rightarrow S$, the two propositions are logically equivalent, that is, $S \leftrightarrow \bar{S}$. Equivalent propositions have the same truth value, that is, $T(S) = T(\bar{S}) = 1 - T(S)$, which yields the expression, T(S) = ½.

As seen, paradoxes reduce to half truths or half-falsities mathematically. Multi-valued logic can address a more subtle kind of paradox. Consider the example of a litre –full glass of water. Consider the situation wherein the water is removed from the glass 1ml at a time. The question is at what point does the glass become empty. No single ml of water represents that decisive value. Rather, the glass transitions from full to empty gradually with the removal of water 1ml at a time.

A fuzzy logic proposition P_f is a statement involving some concept with boundaries that are fuzzy and not clearly defined. Most natural language is fuzzy in that it involves vague and imprecise terms. The truth value assigned to P_f can be any value in the interval [0,1]. Suppose proposition P_f is assigned to fuzzy set A_f, the truth value of a proposition for an element x is the membership grade of x in the fuzzy set A_f. The logical connectives of negation, disjunction, conjunction, implication are also defined in a fuzzy logic. The connectives are given below for two simple propositions – proposition P_f defined on fuzzy set A_f and proposition Q_f defined on fuzzy set B_f.

Negation: $T(\overline{P_f}) = 1 - T(P_f)$

Disjunction: $P_f \vee Q_f$: x is A_f or B_f $T(P_f \vee Q_f) = \max(T(P_f), T(Q_f))$

Conjunction: $P_f \wedge Q_f$: x is A_f and B_f $T(P_f \wedge Q_f) = \min(T(P_f), T(Q_f))$

Implication: $P_f \rightarrow Q_f : x\ is\ A_f, then\ x\ is\ B_f$

$$T\left(P_f \rightarrow Q_f\right) = T\left(\overline{P_f}\ V\ Q_f\right) = \max(T(\overline{P_f}), T(Q_f))$$

Fuzzy (Rule-Based) Systems: In the field of artificial intelligence there are various ways of representing knowledge. The most common way of expressing human knowledge is to form it into natural language expression of the type:

IF premise (antecedent), THEN conclusion (consequent)

This form is called the if-then rule based form, also known as the deductive form. It typically expresses an inference such that if we know a fact (premise or hypothesis or antecedent), then we can infer or derive another fact called the consequent. This type of knowledge representation, characterized as shallow knowledge is quite appropriate in the context of linguistics, because it expresses human expression or heuristic knowledge in our own language of communication. The fuzzy rule based system is most useful in modelling some complex systems observed by humans because they make use of linguistic variables as their antecedents and consequents.

3.1.5. Development of Membership Functions

The process of assigning membership values of function to some fuzzy variables can be intuitive or can be based on some algorithmic or logical operations. The following is a list of six straightforward methods described in literature to assign membership values or functions to fuzzy variables:

Intuition

Inference

Rank ordering

Neural networks

Genetic algorithm

Inductive reasoning

3.2. Artificial Neural Networks

3.2.1. Introduction

[**55**] A neural network is a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1) A learning process is used by the network to gather knowledge from the environment.
2) The acquired knowledge is stored in the form of synaptic weights which are nothing but inter-neuron connection strengths.

The learning process is performed by a procedure called a learning algorithm, which changes the synaptic weights of the network in an orderly manner to attain a desired design objective.

Benefits of neural networks:

- Non-linearity – An artificial neural network may be linear or non-linear. A neural network, made up of an interconnection of non-linear neurons, is itself non-linear, and this non-linearity is distributed throughout the network.
- Input – Output Mapping – Supervised learning or learning from examples is a popular method of updating the synaptic weights. The neural network is fed data which consists of an input pattern and a desired response. The actual output of the network is typically different from the desired response. Now, the training process updates the synaptic weights such that the difference between the actual output and the desired output is minimized. This step is performed for an input/output pair chosen at random and repeated for other pairs. The process is continued till the time there is no significant changes in the synaptic weights between one iteration and the next.
- Adaptivity – Neural networks have the inherent ability to modify their synaptic weights to changes in the surrounding environment. If there are minor changes to the operating environment of a trained neural network, then the network can be easily retrained to take those changes into account and adapt the synaptic weights accordingly.
- Evidential Response – In the context of pattern classification, a neural network may be designed to provide information not only about which particular pattern to select, but also about the confidence in the decision made.
- Contextual Information – The structure and activation state of the neural network provides a means of representing knowledge. The activity of any one neuron in the network can potentially affect all other neurons. As a result, contextual information is dealt with naturally by a neural network.
- Fault Tolerance – A neural network, implemented in hardware form, has the potential to be inherently fault tolerant in the sense that its performance degrades gracefully under adverse operating conditions.
- VLSI Implementability – The massively parallel nature of neural networks makes it potentially fast for the computation of certain tasks. This same feature makes a neural network well-suited for implementation with VLSI technology.
- Uniformity of Analysis and Design – Neurons in one form or another represent an ingredient common to all neural networks. This commonality makes it possible to share theories and learning algorithms in different applications of neural networks. Modular networks can be built by a seamless integration of modules.

- Neurobiological analogy – The design of a neural network is inspired by the structure of the human brain, which is living proof that fault-tolerant parallel processing is not only physically possible but also fast and powerful. Neurobiologists draw from artificial neural networks as a research tool for the interpretation of neurobiological phenomena. On the other hand, engineers draw from neurobiology for new ideas to solve problems that are more complex than those based on conventional hardwired design techniques.

Models of a neuron – As shown in Figure 3.2.1-1, there are three basic elements of the neuron model:

- A set of synapses or connecting links, each of which is characterized by a weight or strength of its own. Specifically, a signal x_j at the input of synapse j connected to neuron k is multiplied by synaptic weight w_{kj}. The synaptic weight of an artificial neuron may lie in a range that includes negative as well as positive values.
- An adder for summing the input signals, weighted by the respective synaptic strengths of the neuron.
- An activation function for limiting the amplitude of the output of a neuron. The activation function is also referred to as a squashing function because it limits or squashes the permissible amplitude range of the output signal to some finite value. Typically, the normalized amplitude range of the output of the neuron is written as the closed unit interval [0, 1] or alternatively, [-1, 1].

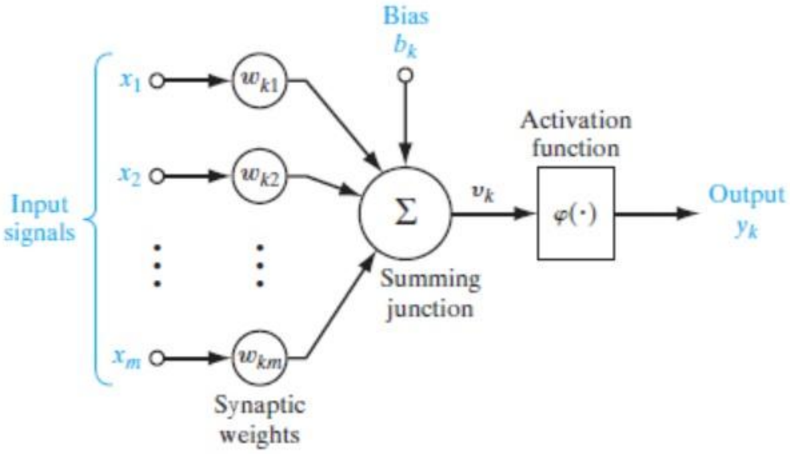

Figure 3.2.1-1: Model of a Neuron

Types of activation function:

The activation function, denoted by φ(v), defines the output of a neuron in terms of an induced local field v. We identify two basic types of activation function:

- Threshold Function – For this type of activation function,

  φ (v) = 1, if v is greater than or equal to 0

  0, if v is less than 0

  Correspondingly, the output of neuron k employing such a threshold function is depicted as:

  y_k = 1, if v_k is greater than or equal to 0

  0, if v_k is less than 0,

  where v_k is the induced local field of the neuron; that is:

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k$$

- Sigmoid Function – The sigmoid function, whose graph is 'S' shaped is by far the most common type of activation function used in the construction of neural networks. It is defined as a monotonically increasing function and provides a good balance between linear and non-linear behaviours. An example of the sigmoid function is the logistic function defined by:

  $\varphi(v) = \frac{1}{1+\exp{(-av)}}$, where a is the slope parameter of the sigmoid function. The sigmoid function assumes a continuous range of values from 0 to 1, whereas the threshold function takes a value of either 0 or 1. Also, the sigmoid function is differentiable, whereas the threshold function is not.

  The above activation functions range from 0 to +1, but it is sometimes desirable to have an activation function that ranges from -1 to +1. Such a function is the tanh function.

Knowledge Representation – Knowledge refers to stored information or models, used by a person or machine to interpret, predict, and appropriately respond to the outside world. Knowledge representation is primarily concerned with two considerations: (1) What information is actually made explicit, (2) How the information is physically encoded for subsequent use. By the very nature of it therefore, knowledge representation is goal oriented.

The design of a neural network is based directly on real-life data with the dataset being allowed to speak for itself. The examples used to train a neural network may consist of both positive and negative examples. The four rules of knowledge representation are:

- Similar inputs from similar classes generally ought to produce similar representations inside the network, and should therefore be grouped as belonging to the same class.
- Items to be classified as separate classes should be given significantly different representations in the network.
- If a particular feature is important then there should be a large number of neurons involved in the representation of that item in the network.
- Prior information and invariances should be built into the design of the neural network, whenever they are available, in order to simplify the network design by its not having to learn them.

Learning Processes – The learning process may be supervised or unsupervised. Under supervised learning, the neural network is trained with examples that include inputs and outputs. Under unsupervised learning process, the examples do not have an expected output.

3.2.2. Rosenblatt's Perceptron

The perceptron is the simplest form of neural network used for classification of patterns said to be linearly separable (that is, lying of opposite sides of a hyperplane). It is nothing but a single neuron with adjustable synaptic weights and bias. A perceptron built around a single neuron can perform pattern classification with only two classes. By expanding the output layer of the perceptron to include more than one neuron, we may correspondingly perform classification with more than two classes. However in order that the perceptron may work properly, the classes have to be linearly separable. Rosenblatt proved that if the patterns used to train the perceptron are drawn from two linearly separable classes, then the perceptron algorithm converges and positions the decision surface in the form of a hyperplane between the two classes. This is the perceptron convergence theorem. The neural model consists of a linear combiner followed by a hard limiter. The summing node of the neural model performs a linear combination of the inputs applied to its synapses as well as incorporates an externally applied bias. The resulting sum, that is, the induced local field is applied to a hard limiter. Accordingly, the neuron produces an output of +1 if the hard limiter input is positive, and -1 if it is negative.

The induced local field is given by:

$$v = \sum_{i=1}^{m} w_i x_i + b$$

Where w_1, w_2, w_m etc are the synaptic weights, x_1, x_2, x_m etc are inputs and b is the externally applied bias. The goal of the perceptron is to correctly classify the set of externally applied stimuli x_1, x_2,…x_m etc into two classes, C_1 and C_2. The decision rule of the classification is to assign the point represented by the inputs x_1, x_2,…x_m to class C_1 if perceptron output y is +1 and class C_2 if y is -1. In the simplest form of the perceptron, there are two decision regions separated by a hyperplane, which is defined by:

$$\sum_{i=1}^{m} w_i x_i + b = 0$$

The effect of the bias b is merely to shift the decision boundary away from the origin.

The synaptic weights, w_1, w_2,…w_m of the perceptron can be adapted on an iteration-by-iteration basis till the time comes when the weights do not change for increasing iterations.

3.2.3. The Least Mean Square Algorithm

The least mean square (LMS) algorithm developed by Widrow and Hoff was the first linear adaptive filtering algorithm developed for solving problems like prediction. Development of the LMS was inspired by the perceptron. Though the LMS and perceptron have different applications, they both involve the use of a linear combiner. Where computational complexity is concerned, the LMS algorithm's complexity is linear with respect to adjustable parameters. This makes the algorithm computationally efficient yet effective in performance. Simple to code and easy to build, importantly, the algorithm is robust with respect to external disturbances.

From an engineering perspective, the above qualities of LMS are highly desirable. Hence, the popularity of the LMS algorithm has remained intact over the years.

Consider an unknown dynamic system that is stimulated by an input vector consisting of the elements, $x_1(i)$, $x_2(i)$,…$x_m(i)$, where "i" denotes the instant of time at which the stimulus is applied to the system. The time index, i = 1,2,…n. In response to the stimulus, the system produces the output y(i). Thus the external behaviour of the system is described by the data set:

D: {x(i), d(i); i = 1,2,…n,…}

Where $x(i) = [x_1(i), x_2(i),\ldots x_m(i)]^T$

The sample pairs constituting D are identically distributed according to an unknown probability law. The dimension M pertaining to the input vector x(i) is referred to as the dimensionality of the input vector space. The stimulus x(i) can may be spatial or temporal.

- The M elements of x(i) originate in different points in space. In this case, we refer to x(i) as a snapshot of data.
- The M elements of x(i) represent the set of present and (M-1) past values of some excitation that are uniformly spaced in time.

We address the problem of how to design a multiple-input-single-output model of an unknown dynamic system by building it around a single linear neuron. The neural model operates under the influence of an algorithm that controls necessary adjustments to the synaptic weights of the neuron, with the following points in mind:

- The algorithm starts from an arbitrary setting of the neuron’s synaptic weights.
- Continuous adjustments to the synaptic weights in response to the statistical variations in the system’s behaviour are made.
- Computations of adjustments to the synaptic weights are completed inside an interval that is one sampling period long.

The neural model just described is referred to as an adaptive filter. Figure 3.2.3-1 below shows the signal flow graph of an adaptive filter:

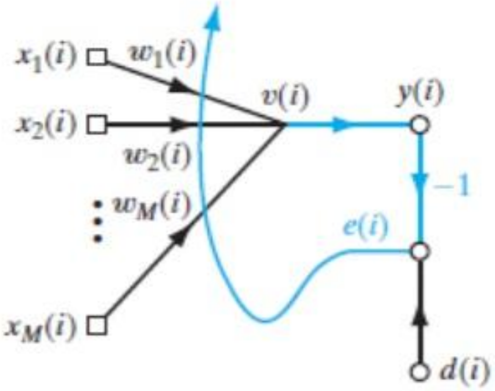

Figure 3.2.3-1: Signal Flow Graph of an Adaptive Filter

Its operation consists of two continuous processes:

1. Filtering process, which involves computation of two signals:

- An output denoted by y(i) that is produced in response to the M elements of the stimulus vector x(i).
- An error signal denoted by e(i) that is obtained by comparing the output y(i) with the corresponding output d(i) produced by the unknown system. In effect, d(i) acts as a desired response.

2. Adaptive process, which involves the automatic adjustment of the synaptic weights of the neuron in accordance with the error signal e(i).

As shown in Figure 3.2.3-1, the combination of the above two processes acting together constitutes a feedback loop acting around the neuron. Since the neuron is linear, the output y(i) is exactly equal to the induced local field v(i); that is,

$$y(i) = v(i) = \sum_{k=1}^{M} w_k(i)x_k(i)$$

Where w_1(i), w_2(i),…w_M(i) are the M synaptic weights of the neuron measured at time i. The neuron’s output y(i) is compared with the corresponding output d(i) which is the desired output of the unknown system at time i. Generally, y(i) is different from d(i), which results in the error signal

$$e(i) = d(i) - y(i)$$

The manner in which the error signal e(i) is used to control the adjustments to the neuron's synaptic weights is determined by the cost function used to derive the adaptive filtering algorithm of interest.

The least means square (LMS) algorithm is designed to minimize the instantaneous value of the cost function

$$\varepsilon(\hat{\boldsymbol{w}}) = \frac{1}{2}\, e^2(n)$$

Where e(n) is the error signal measured at time n. Differentiating the above equation yields:

$$\frac{\delta\varepsilon(\hat{\boldsymbol{w}})}{\delta\hat{\boldsymbol{w}}} = e(n)\frac{\delta e(n)}{\delta\hat{\boldsymbol{w}}}$$

The LMS algorithm works with a linear neuron. So the error signal may be expressed as

$$e(n) = d(n) - \ \boldsymbol{x}^T(n)\hat{\boldsymbol{w}}(n)$$

Hence,

$$\frac{\delta e(n)}{\delta\hat{\boldsymbol{w}}(n)} = \ -x(n)e(n)$$

Using the latter result as the instantaneous estimate of the gradient vector, we may write

$$\hat{\boldsymbol{g}}(n) = \ -\boldsymbol{x}(n)e(n)$$

Finally, we may formulate LMS algorithm as follows:

$$\hat{\boldsymbol{w}}(n+1) = \ \hat{\boldsymbol{w}}(n) + \ \eta\boldsymbol{x}(n)e(n)$$

It is worth noting that the inverse of the learning rate parameter η acts as a measure of the memory of the LMS algorithm: The smaller we make η, the longer the memory span over which LMS algorithm remembers past data will be. Consequently, the LMS algorithm performs accurately when η is small, but the convergence rate of the algorithm is slow.

3.2.4. Virtues and Limitations of the LMS Algorithm

Computational simplicity and efficiency – Two virtues of the LMS algorithm are computational simplicity and efficiency. Coding of the algorithm consists of two or three lines. Computational complexity is linear in the number of adjustable parameters.

Robustness – The algorithm is robust with respect to external disturbances.

The primary limitations of the LMS algorithm are its slow rate of convergence and its sensitivity to variations in the eigenstructure of the input. The LMS algorithm typically requires a number of iterations equal to about 10 times the dimensionality of the input space for it to reach a steady state condition. The slow rate of convergence of the LMS algorithm becomes particularly serious when the dimensionality of the input space becomes high.

3.3. Multilayer Perceptron / Back Propagation Algorithm

3.3.1. Introduction

The multi-layer perceptron is designed to overcome the limitations of Rosenblatt's perceptron and the LMS algorithm. The following points highlight the basic features of multi-layer perceptrons:

- The model of each neuron in the network includes a non-linear activation function that is differentiable.
- The network contains one or more layers that are hidden from the output as well as the input nodes.
- The network exhibits a high degree of connectivity, the extent of which is determined by the synaptic weights of the network.

On the flip side, because of the hidden nodes, our knowledge of the working of the network is incomplete. Firstly, the theoretical analysis of a multilayer perceptron is difficult to undertake owing to the presence of a distributed form of non-linearity and the high connectivity of the network. Secondly, the learning process is harder to visualize owing to the presence of hidden neurons.

A popular method of training the multilayer perceptron is the back-propagation algorithm, which includes the LMS algorithm as a special case. The training proceeds in two phases:

- In the forward phase, the synaptic weights of the network are unchanged and the input signal is propagated through the network, one layer at a time, till it reaches the output. Hence, in this phase, changes are confined to the activation potentials and the outputs of the neurons in the network.

- In the backward phase, an error signal is produced by comparing the output of the network with the desired response. The resulting error signal is propagated through the network, layer by layer, in the backward direction. In the second phase, successive adjustments are made to the synaptic weights of the network. Calculation of the adjustments for the output layer is straightforward, but it is much more complicated for the hidden layer.

The development of the back-propagation algorithm was very significant in neural networks in that in that it provided a computationally efficient method for training multi-layer perceptrons.

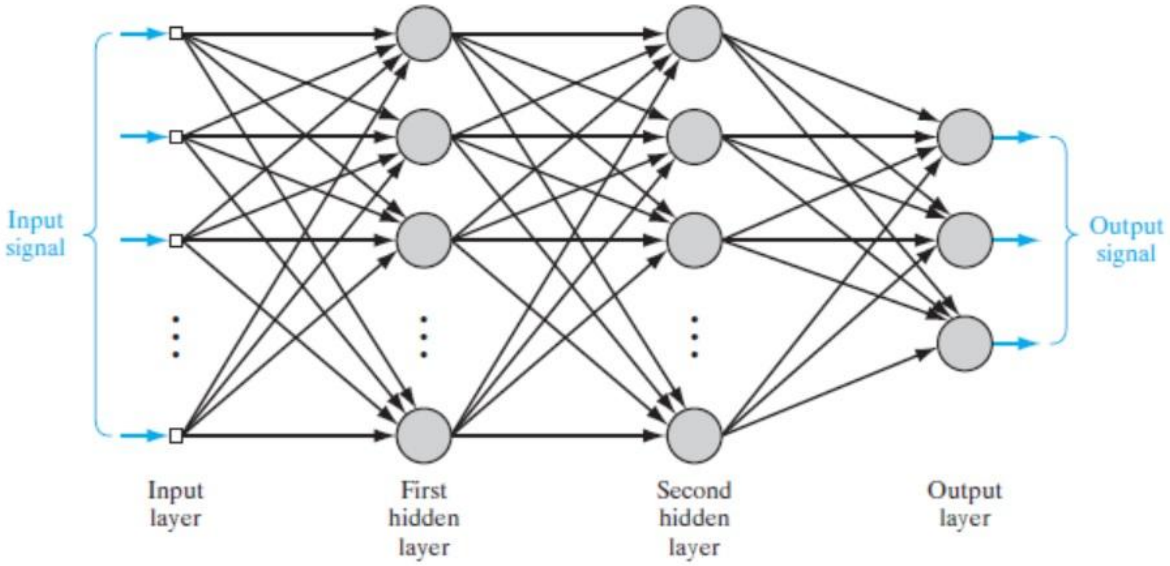

Figure 3.3.1-1: Architecture of a Multi-Layer Perceptron

The Figure 3.3.1-1 shows the architecture of a multilayer perceptron with two hidden layers. The network shown here is fully connected. This means that a neuron in any layer of the network is connected to all the neurons in the previous layer. Signal flow through the network progresses in the forward direction, from left to right, on a layer-by-layer basis. There are two kinds of signals:

- Function signals – A function signal is an input signal or stimulus that comes in at the input end of the network, propagates forward, layer-by-layer and emerges at the output end of the network. At each neuron of the network through which the function signal passes, the signal is calculated as a function of the inputs and the associated weights applied to the neuron. The function signal is also referred to as the input signal.

- Error signals – An error signal originates at the output neuron of the network and propagates backward through the network.

The first hidden layer is fed from the input layer made up of sensory units (source nodes); the resulting outputs of the first hidden layer are applied to the next hidden layer, and so on, for the rest of the network.

Each hidden or output neuron of a multilayer perceptron is meant to perform two computations:

- The calculation of the function signal appearing at the output of each neuron, which is expressed as a continuous non-linear function of the input signal and synaptic weights associated with that neuron.
- The calculation of an estimate of the gradient vector (i.e., the gradients of the error surface with respect to the weights connected to the inputs of a neuron), which is needed for a backward pass through the network.

The hidden neurons which act as feature detectors play a critical role in the operation of multilayer perceptrons. With the progress of the learning process, the hidden neurons gradually begin to discover the important features that characterize the training data. They do so by performing a non-linear transformation of the input data into a new space called the feature space. In this new space, the classes of interest in a pattern classification task may be separated more easily from each other than could be the case in the original input data space. Note the Rosenblatt's perceptron has no concept of feature space.

Credit assignment problem – The credit assignment problem is the problem of assigning credit or blame for overall outcomes to each of the internal decisions made by the hidden computational units of the distributed learning system, recognizing that those decisions are responsible for the overall outcomes in the first place. In order to perform a designated task, the neural network must assign certain kinds of behaviour to all of its neurons through a specification of the error correction learning algorithm. It is possible to supply a desired response to guide the behaviour of an output neuron since each output neuron is visible to the outside world. Thus as far as output neurons are concerned, it is easy and straightforward to adjust the synaptic weights of each output neuron in accordance with the error-correction algorithm. However, doing the same thing for the hidden neurons is more involved.

3.3.2. Batch Learning and Online Learning

Consider a multilayer perceptron with an input layer of source nodes, one or more hidden layers and an output layer consisting of one or more neurons. Let

$$T = \{x(n), d(n)\}^{N}{}_{n=1}$$

denote the training sample used to train the network in a supervised manner. Let $y_j(n)$ denote the function signal produced at the output of neuron j in the output layer by the stimulus x(n) applied to the input layer. Correspondingly, the error signal produced at the output of neuron j is defined by

$$e_j(n) = d_j(n) - y_j(n)$$

where $d_j(n)$ is the jth element of the desired response vector **d**(n). The instantaneous error energy of neuron j is defined as:

$$\varepsilon_j(n) = \frac{1}{2} e_j^2(n)$$

The total instantaneous error energy of the entire network is:

$$\varepsilon(n) = \sum_{j \,\epsilon\, C} \varepsilon_j(n)$$

$$= \frac{1}{2} \sum_{j \epsilon C} e_j^2(n)$$

where the set C includes all the neurons in the output layer. With the training samples consisting of N examples, the error energy averaged over the training samples, or the empirical risk, is defined as:

$$\varepsilon_{av}(n) = \frac{1}{2} \sum_{n=1}^{N} \varepsilon(n)$$

$$= \frac{1}{2N} \sum_{n=1}^{N} \sum_{j \epsilon C} e_j^2(n)$$

Batch Learning – In the batch method of supervised learning, adjustments to the synaptic weights of the multilayer perceptron are performed after all the N examples in the training sample T have been presented - this constitute one epoch of training. The cost function for batch learning is given by the average error energy ε_{av}. The synaptic weights of the multilayer perceptron are adjusted on an epoch-by-epoch basis. The advantages of batch learning are given below:

- Accurate estimation of the gradient vector (i.e., the derivative of the cost function ε_{av} with respect to the weight vector **w**), thereby guaranteeing under simple conditions, convergence of the method of steepest descent to a local minimum;
- Parallelization of the learning process.

However, from a practical perspective, batch learning is rather demanding in terms of storage requirements.

On-line learning – In the on-line method of supervised learning, the synaptic weights of the multi-layer perceptron are adjusted on an example-by-example basis. Therefore, the cost function to be minimized is the total instantaneous error energy ε(n). Since the training examples are presented to the network randomly, online learning is sometimes referred to as a stochastic method. The stochasticity has the desirable effect of making it less likely for the learning process to be trapped in a local minimum, which is a definite advantage of online learning over batch learning. Another advantage of on-line learning is the fact that it requires much less storage than batch learning. When training data is redundant, on-line learning is able to take advantage of this redundancy because examples are presented one at a time. Another useful property of on-line learning is its ability to track small changes in the training data. On-line learning is popular for two important practical reasons:

- On-line learning is simple to implement.
- It provides effective solutions to large-scale and difficult pattern-classification problems.

3.3.3. The Back Propagation Algorithm

The popularity of online learning for the supervised training of multilayer perceptrons has been further enhanced by the development of the back propagation algorithm. To describe this algorithm, consider the Figure 3.3.3-1 below:

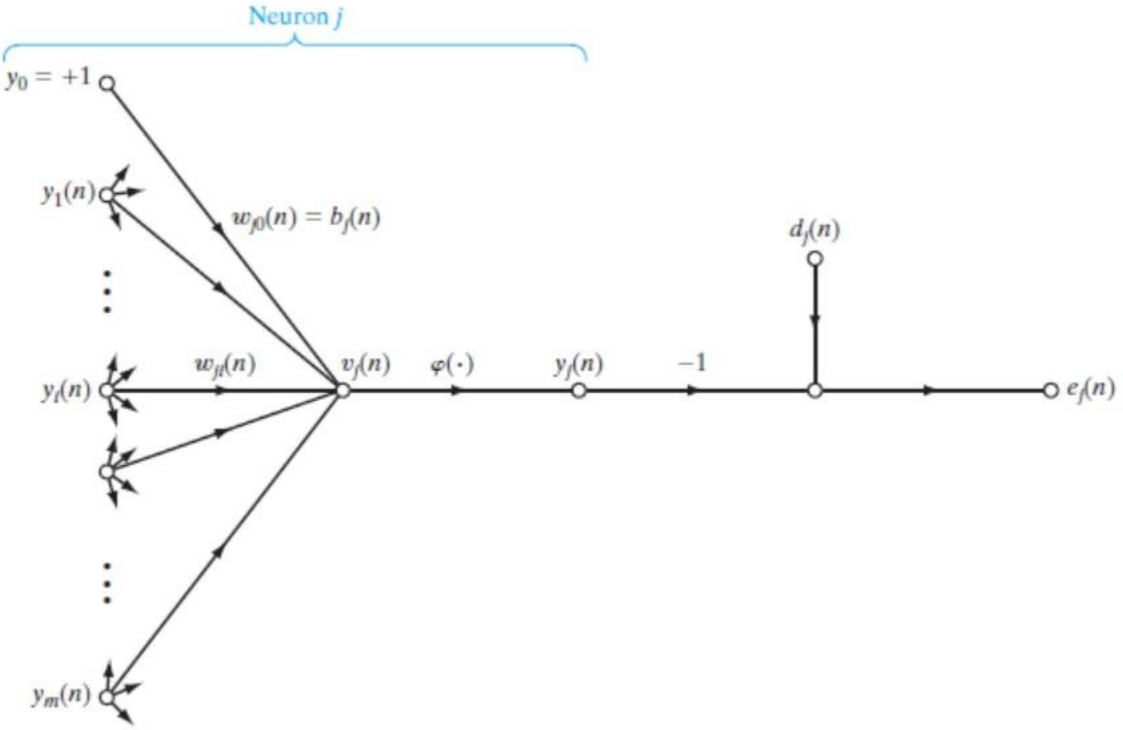

Figure 3.3.3-1: Signal Flow Graph Highlighting the Details of Output Neuron 'j'

It depicts neuron j being fed by a set of function signals produced by a layer of neurons to its left. The induced local field $v_j(n)$ produced at the input of the activation function associated with neuron j is therefore:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n)$$

Where m is the total number of inputs excluding the bias applied to neuron j. The synaptic weight w_{j0} (corresponding to the fixed input $y_0 = +1$) equals the bias b_j applied to neuron j. Hence the function signal $y_j(n)$ appearing at the output of neuron j at iteration n is:

$$y_j(n) = \varphi_j(v_j(n))$$

The back-propagation algorithm applies a correction $\Delta w_{ji}(n)$ to the synaptic weight $w_{ji}(n)$, which is proportional to the partial derivative $\frac{\delta\varepsilon(n)}{\delta w_{ji}(n)}$. According to the chain rule of calculus, we may express the gradient as:

$$\frac{\delta\varepsilon(n)}{\delta w_{ji}(n)} = \frac{\delta\varepsilon(n)}{\delta e_j(n)} \frac{\delta e_j(n)}{\delta y_j(n)} \frac{\delta y_j(n)}{\delta v_j(n)} \frac{\delta v_j(n)}{\delta w_{ji}(n)}$$

The partial derivative $\frac{\delta\varepsilon(n)}{\delta w_{ji}(n)}$ represents a sensitivity factor, determining the direction of search in weight space for the synaptic weight w_{ji}. Now,

$$\frac{\delta\varepsilon(n)}{\delta e_j(n)} = e_j(n); \frac{\delta e_j(n)}{\delta y_j(n)} = -1; \frac{\delta y_j(n)}{\delta v_j(n)} = \varphi_j'(v_j(n)); \frac{\delta v_j(n)}{\delta w_{ji}(n)} = y_i(n)$$

Therefore, $\frac{\delta\varepsilon(n)}{\delta w_{ji}(n)} = -e_j(n)\varphi_j'\left(v_j(n)\right)y_i(n)$

The correction $\Delta w_{ji}(n)$ applied to $w_{ji}(n)$ is defined by the delta rule, or

$$\Delta w_{ji}(n) = -\eta\frac{\delta\varepsilon(n)}{\delta w_{ji}(n)}$$

Where η is the learning rate parameter of the back propagation algorithm. The use of minus sign accounts for gradient descent in weight space (i.e., seeking a direction for weight change that reduces the value of ε(n)).

Stopping criteria – In general, the back-propagation algorithm cannot be shown to converge, and there are no well-defined criteria for stopping its operation. Rather, there are some reasonable criteria, each with its own practical merit, which may be used to terminate the weight adjustments.

The back propagation algorithm is considered to have converged when the Euclidean norm of the gradient vector reaches a sufficiently small gradient threshold.

Another criterion for convergence is:

The back-propagation algorithm is considered to have converged when the absolute rate of change in the average squared error per epoch is sufficiently small.

The rate of change of the average squared error is typically considered to be small enough if it lies in the range of 0.1 to 1 percent per epoch. Unfortunately, this criterion may result in premature termination of the learning process.

3.3.4. Virtues of Back Propagation Learning

The back-propagation algorithm is a computationally efficient technique for computing the gradients (i.e., first order derivatives) of the cost function ε(w), expressed as a function of the adjustable parameters (synaptic weights and bias terms) that characterize the multilayer perceptron.

The computational power of the algorithm is derived from two distinct properties:

- The back-propagation algorithm is simple to compute locally.

- It performs stochastic gradient descent in weight space, when the algorithm is implemented in its on-line (sequential) mode of learning.

3.4. Support Vector Machines

3.4.1. Introduction

The advantage of the multilayer perceptron trained with back-propagation algorithm is its simplicity, but the algorithm is slow to converge and is not optimal. In this section, another class of feedforward network, called support vector machines, is presented. Basically, the support vector machine is an elegant binary learning machine. Given a training sample, the support vector machine constructs a hyperplane as the decision surface in such a way that the margin of separation between positive and negative samples is maximized. This basic idea is extended in a systematic manner to deal with patterns that are difficult to separate non-linearly.

At the heart of the development of the support vector machine is the notion of the inner-product kernel between the support vector x_i and a vector x drawn from the input data space. Most importantly, the support vectors consist of a small subset of data points extracted by the learning algorithm from the training sample itself. Indeed, it is because of this central property, the learning algorithm, involved in the construction of the support vector machine is also referred to as the kernel method.

The support vector machine can be used to solve both pattern classification and non-linear regression problems. However, their impact has been mostly felt in case of classification of non-linearly separable patterns.

3.4.2. Optimal Hyperplane for Linearly Separable Patterns

Consider the training sample $\{x_i, d_i\}^N{}_{i=1}$, where x_i is the input pattern for the i-th example and d_i is the corresponding desired response (target output). In the beginning, let us assume that the pattern (class) represented by the subset $d_i = +1$ and the pattern represented by the subset $d_i = -1$ are linearly separable. The equation of a decision surface in the form of a hyperplane that does the separation is

$$w^T x + b = 0$$

Where x is an input vector, w is an adjustable weight vector and b is the bias. We may thus write

$$w^T x_i + b \geq 0 \quad for\ d_i = +1$$
$$w^T x_i + b < 0 \quad for\ d_i = -1$$

For a given weight vector w and bias b, the separation between the hyperplane and the nearest data point is known as the margin of separation, denoted by ρ. The aim of a support vector machine is to find the particular hyperplane for which the margin of separation, ρ, is maximized. If this condition is met, the decision surface is referred to as the optimal hyperplane. Such an optimal hyperplane is depicted in the Figure 3.4.2-1 below:

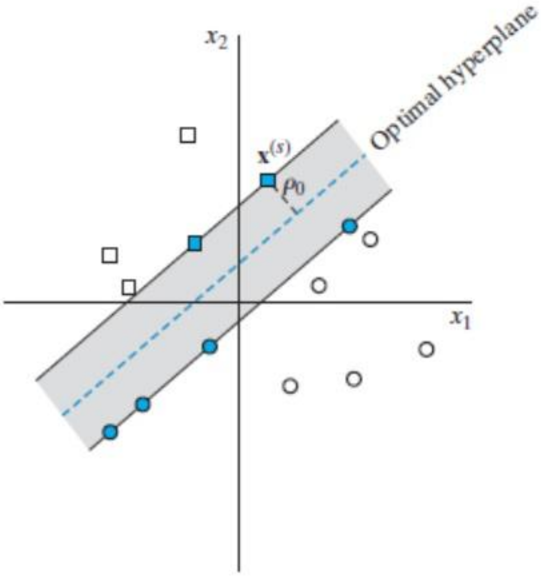

Figure 3.4.2-1: The Optimal Hyperplane of a Support Vector Machine

Let $\mathbf{w}_0$ and b_0 denote the optimum values of the weight vector and bias respectively. Correspondingly, the optimal hyperplane, representing a multidimensional linear decision surface in the input space, is defined by

$$\boldsymbol{w}_0^T \boldsymbol{x} + b_0 = 0$$

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane: $g(\boldsymbol{x}) = \boldsymbol{w}_0^T \boldsymbol{x} + b_0$

The easiest way to see this is to express **x** as:

$$x = x_p + r \frac{w_0}{||w_0||}$$

Where $\mathbf{x}_p$ is the normal projection of **x** on to the optimal hyperplane and r is the desired algebraic distance; r is positive if **x** is on the positive side of the optimal hyperplane and is negative if **x** is on the negative side. Since, by definition, $g(\mathbf{x}_p) = 0$, it follows that

$$g(\boldsymbol{x}) = \boldsymbol{w}_0^T\boldsymbol{x} + b_0 = r||\boldsymbol{w}_0||$$

Or equivalently,

$$\boldsymbol{r} = \frac{g(\boldsymbol{x})}{||\boldsymbol{w}_0||}$$

Particularly, the distance from the origin to the optimal hyperplane is given by $b_0/||w_0||$.
If $b_0 > 0$, the origin is on the positive side of the optimal hyperplane; if $b_0 < 0$, it is on the negative side. If $b_0 = 0$, the optimal hyperplane passes through the origin. Now we need to find the parameters w_0 and b_0 for the optimal hyperplane, given a training set. The particular data points (x_i, d_i) for which

$$\boldsymbol{w}_0^T\boldsymbol{x} + b_0 = 1$$

is satisfied are called support vectors – hence the name “support vector machines”. All the remaining examples in the training sample are completely irrelevant. Generally speaking, the support vectors are those data points that lie closest to the optimal hyperplane and are most difficult to classify. As such, they have a direct bearing on the optimum location of the decision surface.
Consider a support vector $x^{(s)}$ for which $d^{(s)} = +1$. Then, by definition we have,

$$g\left(\boldsymbol{x}^{(s)}\right) = \boldsymbol{w}_0^T\boldsymbol{x} + b_0 = \pm 1, \quad for\ d^{(s)} = \pm 1$$

The algebraic distance of the support vector $x^{(s)}$ to the optimal hyperplane is

$$r = \frac{g(\boldsymbol{x}^{(s)})}{||\boldsymbol{w}_0||}$$

Let ρ denote the optimum value of the margin of separation between the two classes that constitute the training sample. Then, it follows that,

$$\rho = 2r$$

$$= \frac{2}{||w_0||}$$

Maximizing the margin of separation between binary classes is same as minimizing the Euclidean norm of the weight vector **w**.

3.4.3. The Support Vector Machine Viewed As a Kernel Machine

Inner Product Kernel – Let **x** denote a vector drawn from input space of dimension m_0. Let $\{\varphi_j(x)\}$represent a set of non-linear functions that, between them, transform the input space of dimension m_0 to a feature space of infinite dimensionality. Given this transformation, the hyperplane acting as a decision surface may be defined in accordance with the formula

$$\sum_{j=1}^{\infty} w_j \ \varphi_j(\boldsymbol{x}) = 0$$

Where $\{w_j\}$ denotes an infinitely large set of weights that transforms the feature space into the output space. It is in the output space where the decision is made on whether the input vector **x** belongs to one of two possible classes, positive or negative. Using matrix notation,

$$\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}) = \ 0$$

Where $\boldsymbol{\Phi}(\boldsymbol{x})$ is the feature vector and **w** is the corresponding weight vector. We seek linear separability of the transformed patterns in the feature space.

The scalar term $\boldsymbol{\varphi}^{\mathbf{T}}(\mathbf{x_i})\boldsymbol{\varphi}(\mathbf{x})$ is an inner product. Accordingly, let this inner product term be denoted as the scalar

$$\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x_i}) = \ \boldsymbol{\Phi}^{\boldsymbol{T}}(\boldsymbol{x_i})\boldsymbol{\Phi}(\boldsymbol{x})$$

$$= \sum_{j=1}^{\infty} \varphi_j(x_i)\varphi_j(x), \qquad\qquad i = 1,2,3, \dots . N_s$$

$K(x, x_i)$ is the kernel, which is a function that calculates the inner product of the images produced in the feature space under the embedding Φ of two data points in the input space. The kernel $k(x, x_i)$ is a function that has two basic properties:

Property 1: The function $k(x, x_i)$ is symmetric about the centre point x_i, that is,

$$k(x, x_i) = \ k\ (x_i, x) \qquad for\ all\ x$$

The function attains its maximum value at point $x = x_i$.

Property 2: The total volume under the surface of the kernel $k(x, x_i)$ is constant.

The Kernel Trick – We can now make two important observations:

- Where pattern classification in the output space is concerned, specifying the kernel k(x, x_i) is sufficient; in explicit computation the weight vector $\mathbf{w}_0$ is not needed. This is commonly referred to as the kernel trick.
- Though it has been assumed that the feature space could be of infinite dimensionality, the definition of the optimal hyperplane consists of a finite number of terms that is equal to the number of training patterns used in the classifier.

Because of the above observations, the support vector machine is also referred to as the kernel machine. For pattern classification, the machine is parameterized by an N-dimensional vector whose i-th term is specified by the product $\alpha_i d_i$, for i = 1,2,…,N.

We may view k(x_i, x_j) as the ij-th element of the symmetric N-by-N matrix

$$\boldsymbol{K} = \{k(x_i, x_j)\}^{N}_{i,j=1}$$

The matrix **K** is a non-negative definite matrix called the kernel matrix; it is also referred to simply as the Gram. It is non-negative definite or positive semi-definite in that it satisfies the condition

$$\boldsymbol{a^T K a} \geq 0$$

For any real valued vector **a** whose dimension is compatible with that of **K**.

4. Design of Experiments

A set of five experiments have been conducted as part of the work. As part of the first experiment, a fuzzy expert system has been developed for diagnosing schizophrenia. This expert system is harnessed as part of Experiment-2 to generate a synthetic dataset for training artificial neural networks for diagnosing schizophrenia. Experiments 3, 4 and 5 shed light on various points observed when an artificial neural network, a support vector machine and a fuzzy clustering utility are fed this synthetic data.

4.1. Experiment 1 – Design of the Fuzzy Inference System for Diagnosing Schizophrenia

A fuzzy logic based system has been developed that captures the expertise of the psychiatrist in diagnosing the condition. The solution was based on the diagnosis of schizophrenia based on the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) criteria [**2**]. According to these criteria, the subject must experience at least two of the following: delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behaviour and negative symptoms. Out of the two symptoms, if one is either catatonic behaviour or negative symptoms then the other must be delusions or hallucinations or disorganized speech. The solution is based on six fuzzy inference systems (FIS), the outputs of five of which serve as inputs to the sixth. Schizophrenia is a psychotic illness and has a lot in common with other psychotic illnesses and mania. The final diagnosis of schizophrenia and in general the degree of psychosis manifested in the subject's behaviour is evident in the crisp output of the sixth fuzzy inference system. Fuzzy based systems have successfully been deployed to diagnose diseases, in particular heart disease and diseases of the thyroid gland [**56**]. In this work, we have cascaded a set of FIS's for diagnosing schizophrenia.

A fuzzy inference system accepts crisp inputs from the user, fuzzifies it and passes it to the inference engine. The inference engine generates an output depending on the rules of the knowledge base, de-fuzzifies it and gives the user a crisp output. There are several methods of de-fuzzifying the output of which we used the centroid method, which is the most popular. The fuzzification of the input is done with the help of membership functions. In the world of fuzzy logic, it is rarely black and white, but rather it is all various shades of gray. A membership function dictates how black or how white a variable is on a scale of 0 to 1. This is called the degree of membership. Apart from input membership functions, there are output membership functions. Besides the membership functions, there are rules which determine how the inputs affect the output. The schematic of the FIS's is shown below in the Figure 4.1-1:

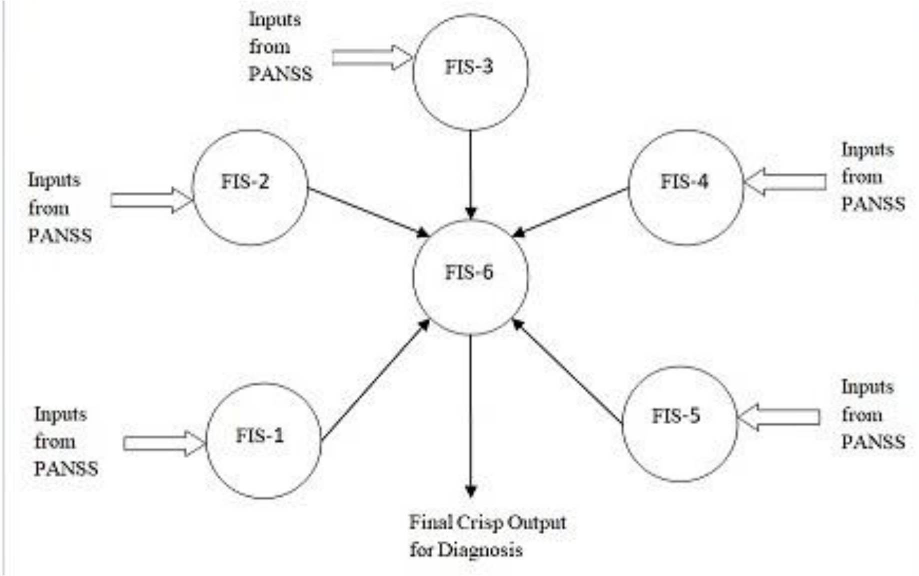

Figure 4.1-1: Cascaded Fuzzy Inference Systems

FIS-1 measures the subject's delusions, FIS-2 measures his hallucinations, FIS-3 measures the degree of disorganized speech, FIS-4 measures the degree of highly disorganized or catatonic behaviour, while FIS-5 measures the level of negative symptoms evident in the subject's behaviour. The outputs of each of FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5 serve as inputs for FIS-6. The PANSS scale is mapped to the following items, as given in Table 4.1-1:

Table 4.1-1: Symptoms of Schizophrenia as Noted on the PANSS Scale

Delusions	Passive / Apathetic Social Withdrawal	Motor Retardation
Conceptual Disorganization	Difficulty in Abstract Thinking	Uncooperativeness
Hallucinatory Behaviour	Lack of Spontaneity and Flow of Conversation	Unusual Thought Content
Excitement	Stereotyped Thinking	Disorientation
Grandiosity	Somatic Concern	Poor Attention
Suspiciousness / Persecution	Anxiety	Lack of Judgement and Insight
Hostility	Guilt Feelings	Disturbance of Volition

Blunted Affect	Tension	Poor Impulse Control
Emotional Withdrawal	Mannerisms and Posturing	Preoccupation
Poor Rapport	Depression	Active Social Avoidance

These PANSS items are selectively mapped to the subject's level of delusions, hallucinations, disorganized speech, highly disorganized or catatonic behaviour and negative symptoms. The mapping is shown below in Table 4.1-2:

Table 4.1-2: Mapping of PANSS Item to DSM-5 Trait

DSM-5 Trait	**PANSS Item**	**Degree of Impact on DSM-5 Trait**
Delusion	Delusions	High
	Suspiciousness / Persecution	
	Hostility	
	Unusual Thought Content	
	Disturbance of Volition	
	Conceptual Disorganization	
	Grandiosity	Medium
	Lack of Judgement	
	Uncooperativeness	
	Anxiety	Low
	Guilt	
Hallucination	Hallucinatory Behaviour	High
	Disturbance of Volition	
	Poor Impulse Control	Medium
	Somatic Concern	Low
Disorganized Speech	Stereotyped Thinking	High
	Lack of Spontaneity	
	Difficulty in Abstract Thinking	Medium
	Disorientation	Low
Grossly Disorganized or Catatonic Behaviour	Mannerisms and Posturing	High
	Motor Retardation	

	Stereotyped Thinking	
	Depression	
	Excitement	
	Anxiety	Medium
	Disorientation	
	Tension	
	Poor Attention	Low
	Preoccupation	
Negative Symptoms	Blunted Affect	High
	Emotional Withdrawal	
	Social Avoidance	
	Passive / Apathetic Social Withdrawal	
	Poor Rapport	Medium
	Lack of Spontaneity and Flow of Conversation	
	Difficulty in Abstract Thinking	Low
	Stereotyped Thinking	

Each of FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5 receives several inputs and each input is associated with two membership functions – “ofconcern” and “notofconcern”. There is a rule base that determines how the inputs – which are fuzzified by the membership functions – determine the fuzzified output. Each FIS has a single output variable that is associated with three membership functions – “low”, “medium” and “high”. A de-fuzzification technique based on the centroid method is chosen that generates a crisp output for the FIS. This output is given as input to FIS-6. As a specific example, the schematic of FIS-2 which evaluates the subject’s hallucinations is given below:

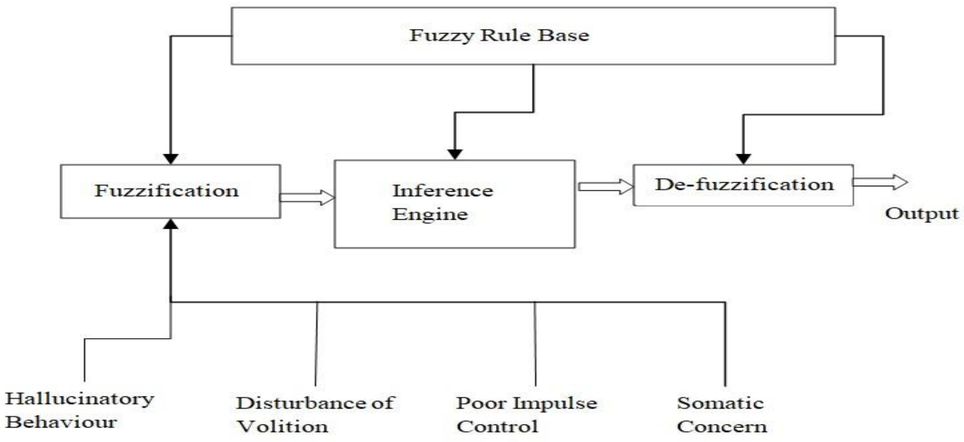

Figure 4.1-2: Structure of the FIS for Measuring Hallucinations

The rule bases:

The rule base for FIS-1 is given as:

Rule1: If "Delusions" is "ofconcern" or "Suspiciousness" is "ofconcern" or "Hostility" is "ofconcern" or "Unusual Thought Content" is "ofconcern" or "Disturbance of volition" is "ofconcern" or "Conceptual Disorganization" is "ofconcern", then "Delusions" is "high"

Rule2: If "Delusions" is "notofconcern" and "Suspiciousness" is "notofconcern" and "Hostility" is "notofconcern" and "Unusual Thought Content" is "notofconcern" and "Disturbance of volition" is "notofconcern" and "Conceptual Disorganization" is "notofconcern" and "Grandiosity" is "ofconcern", then "Delusions" is "medium"

Rule3: If "Delusions" is "notofconcern" and "Suspiciousness" is "notofconcern" and "Hostility" is "notofconcern" and "Unusual Thought Content" is "notofconcern" and "Disturbance of volition" is "notofconcern" and "Conceptual Disorganization" is "notofconcern" and "Lack of judgement" is "ofconcern", then "Delusions" is "medium"

Rule4: If "Delusions" is "notofconcern" and "Suspiciousness" is "notofconcern" and "Hostility" is "notofconcern" and "Unusual Thought Content" is "notofconcern" and "Disturbance of volition" is "notofconcern" and "Conceptual Disorganization" is "notofconcern" and "Uncooperativeness" is "ofconcern", then "Delusions" is "medium"

Rule5: If "Delusions" is "notofconcern" and "Suspiciousness" is "notofconcern" and "Hostility" is "notofconcern" and "Unusual Thought Content" is "notofconcern" and "Disturbance of volition" is "notofconcern" and "Conceptual Disorganization" is "notofconcern" and "Uncooperativeness" is "notofconcern" and "Anxiety" is "ofconcern" , then "Delusions" is "low"

Rule6: If “Delusions” is “notofconcern” and “Suspiciousness” is “notofconcern” and “Hostility” is “notofconcern” and “Unusual Thought Content” is “notofconcern” and “Disturbance of volition” is “notofconcern” and “Conceptual Disorganization” is “notofconcern” and “Uncooperativeness” is “notofconcern” and “Guilt” is “ofconcern” , then “Delusions” is “low”

Rule7: If “Delusions” is “notofconcern” and “Suspiciousness” is “notofconcern” and “Hostility” is “notofconcern” and “Unusual Thought Content” is “notofconcern” and “Disturbance of volition” is “notofconcern” and “Conceptual Disorganization” is “notofconcern” and “Uncooperativeness” is “notofconcern” and “Anxiety” is “notofconcern” and “Guilt” is “notofconcern” , then “Delusions” is “low”

The rule base for FIS-2 is given as:

Rule1: If “Hallucinatory behaviour” is “ofconcern” or “Disturbance of volition” is “ofconcern”, then “Hallucinations” is “high”

Rule2: If “Hallucinatory behaviour” is “notofconcern” and “Disturbance of volition” is “notofconcern” and “Poor impulse control” is “ofconcern”, then “Hallucinations” is “medium”

Rule3: If “Hallucinatory behaviour” is “notofconcern” and “Disturbance of volition” is “notofconcern” and “Poor impulse control” is “notofconcern” and “Somatic concern” is “ofconcern”, then “Hallucinations” is “low”

Rule4: If “Hallucinatory behaviour” is “notofconcern” and “Disturbance of volition” is “notofconcern” and “Poor impulse control” is “notofconcern” and “Somatic concern” is “notofconcern”, then “Hallucinations” is “low”

The rule base for FIS-3 is given as:

Rule1: If “Stereotyped thinking” is “ofconcern” or “Lack of spontaneity” is “ofconcern”, then “DisorganizedSpeech” is “high”

Rule2: If “Stereotyped thinking” is “notofconcern” and “Lack of spontaneity” is “notofconcern” and “Difficulty in abstract thinking” is “ofconcern”, then “DisorganizedSpeech” is “medium”

Rule3: If “Stereotyped thinking” is “notofconcern” and “Lack of spontaneity” is “notofconcern” and “Difficulty in abstract thinking” is “notofconcern” and “Disorientation” is “ofconcern”, then “DisorganizedSpeech” is “low”

Rule4: If “Stereotyped thinking” is “notofconcern” and “Lack of spontaneity” is “notofconcern” and “Difficulty in abstract thinking” is “notofconcern” and “Disorientation” is “notofconcern”, then “DisorganizedSpeech” is “low”

The rule base for FIS-4 is given as:

Rule1: If “MannerismOrPosturing” is “ofconcern” or “MotorRetardation” is “ofconcern” or “StereotypedThinking” is “ofconcern” or “Excitement” is “ofconcern” then “Catatonic” is “high”

Rule2: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Depression” is “ofconcern” then “Catatonic” is “medium”

Rule3: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Anxiety” is “ofconcern” then “Catatonic” is “medium”

Rule4: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Disorientation” is “ofconcern” then “Catatonic” is “medium”

Rule5: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Tension” is “ofconcern” then “Catatonic” is “medium”

Rule6: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Depression” is “notofconcern” and “Anxiety” is “notofconcern” and “Disorientation” is “notofconcern” and “Tension” is “notofconcern” and “PoorAttention” is “ofconcern” then “Catatonic” is “low”

Rule7: If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Depression” is “notofconcern” and “Anxiety” is “notofconcern” and “Disorientation” is “notofconcern” and “Tension” is “notofconcern” and “Preoccupation” is “ofconcern” then “Catatonic” is “low”

Rule8: : If “MannerismOrPosturing” is “notofconcern” and “MotorRetardation” is “notofconcern” and “StereotypedThinking” is “notofconcern” and “Excitement” is “notofconcern” and “Depression” is “notofconcern” and “Anxiety” is “notofconcern” and “Disorientation” is “notofconcern” and “Tension” is “notofconcern” and “PoorAttention” is “notofconcern” and “Preoccupation” is “notofconcern” then “Catatonic” is “low”

The rule base for FIS-5 is given as:

Rule1: If “BluntedAffect” is “ofconcern” or “EmotionalWithdrawal” is “ofconcern” or “ActiveSocialAvoidance” is “ofconcern” or “PassiveSocialWithdrawal” is “ofconcern” then “NegativeSymptoms” is “high”

Rule2: If “BluntedAffect” is “notofconcern” and “EmotionalWithdrawal” is “notofconcern” and “ActiveSocialAvoidance” is “notofconcern” and “PassiveSocialWithdrawal” is “notofconcern” and “PoorRapport” is “ofconcern” then “NegativeSymptoms” is “medium”

Rule3: If “BluntedAffect” is “notofconcern” and “EmotionalWithdrawal” is “notofconcern” and “ActiveSocialAvoidance” is “notofconcern” and “PassiveSocialWithdrawal” is “notofconcern” and “LackOfSpontaneity” is “ofconcern” then “NegativeSymptoms” is “medium”

Rule4: If “BluntedAffect” is “notofconcern” and “EmotionalWithdrawal” is “notofconcern” and “ActiveSocialAvoidance” is “notofconcern” and “PassiveSocialWithdrawal” is “notofconcern” and “PoorRapport” is “notofconcern” and “LackOfSpontaneity” is “notofconcern” and “DifficultyInAbstractThinking” is “ofconcern” then “NegativeSymptoms” is “low”

Rule5: If “BluntedAffect” is “notofconcern” and “EmotionalWithdrawal” is “notofconcern” and “ActiveSocialAvoidance” is “notofconcern” and “PassiveSocialWithdrawal” is “notofconcern” and “PoorRapport” is “notofconcern” and “LackOfSpontaneity” is “notofconcern” and “StereotypedThinking” is “ofconcern” then “NegativeSymptoms” is “low”

Rule6: If “BluntedAffect” is “notofconcern” and “EmotionalWithdrawal” is “notofconcern” and “ActiveSocialAvoidance” is “notofconcern” and “PassiveSocialWithdrawal” is “notofconcern” and “PoorRapport” is “notofconcern” and “LackOfSpontaneity” is “notofconcern” and “DifficultyInAbstractThinking” is “notofconcern” and “StereotypedThinking” is “notofconcern” then “NegativeSymptoms” is “low”

Input and output membership functions:

Delusions (Range: 0-6):

$$\mu_{\text{notofconcern}} = e^{-\frac{x^2}{0.248}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-6.4(x-1)}}$$

Conceptual Disorganization (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{3.395(x-1.83)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-2.27(x-2.85)}}$$

Hallucinatory Behaviour (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{9.79(x-0.518)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-7.091(x-0.598)}}$$

Excitement (Range: 0 – 6):

$$\mu_{\text{no tofconcern}} = \frac{1}{1 + e^{2.263(x-2.89)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-1.65(x-2.85)}}$$

Grandiosity (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{1.53(x-2.44)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-1.65(x-2.85)}}$$

Suspiciousness /Persecution (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1+e^{1.53(x-2.44)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-1.73(x-2.08)}}$$

Hostility (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{1.11(x-3.07)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-1.16(x-3.21)}}$$

Blunted Affect (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{1.125(x-3)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1 + e^{-1.61(x-2.78)}}$$

Emotional Withdrawal (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1 + e^{1.186(x-4.36)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.19(\mathrm{x}-3.04)}}$$

Poor Rapport (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + \mathrm{e}^{1.186(\mathrm{x}-4.36)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.71(\mathrm{x}-3.83)}}$$

Passive / Apathetic Social Withdrawal (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \mathrm{e}^{\frac{-\mathrm{x}^2}{8.6528}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.32(\mathrm{x}-3.35)}}$$

Difficulty in Abstract Thinking (Range: 0 – 6):

$$\mu_{notofconcern} = e^{\frac{-x^2}{8.6528}}$$

$$\mu_{ofconcern} = \frac{1}{1 + |\frac{x-6.99}{2.99}|^{2.04}}$$

Lack of Spontaneity (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + \mathrm{e}^{1.08(\mathrm{x}-3.23)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.32(\mathrm{x}-3.35)}}$$

Stereotyped Thinking (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + \mathrm{e}^{1.1.25(\mathrm{x}-3.23)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.3(\mathrm{x}-3.5)}}$$

Somatic Concern (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + \mathrm{e}^{1.49(\mathrm{x}-2.5)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.49(\mathrm{x}-2.49)}}$$

Anxiety (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \mathrm{e}^{-\frac{\mathrm{x}^2}{9.946}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.14(\mathrm{x}-3.52)}}$$

Guilt Feelings (Range: 0 -6):

$$\mu_{\mathrm{notofconcern}} = \mathrm{e}^{-\frac{\mathrm{x}^2}{9.9458}}$$

$$\mu_{\mathrm{ofconcern}} = \mathrm{e}^{-\frac{(\mathrm{x}-6.1)^2}{9.1592}}$$

Tension (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + \mathrm{e}^{1.2(\mathrm{x}-3.36)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + \mathrm{e}^{-1.07(\mathrm{x}-2.61)}}$$

Mannerisms and Posturing (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = e^{-\frac{x^2}{0.7466}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-3.58(x-1.29)}}$$

Depression (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{1.108(x-2.6)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-1.3(x-2.89)}}$$

Motor Retardation (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{1.152(x-2.6)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-1.55(x-2.5)}}$$

Uncooperativeness (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{1.32(x-2.14)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-1.19(x-3.04)}}$$

Unusual Thought Content (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{6.49(x-0.853)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-4.02(x-1.08)}}$$

Disorientation (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{8.34(x-0.534)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-7.53(x-0.724)}}$$

Poor Attention (Range: 0- 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{1.16(x-3.11)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-1.27(x-3.1)}}$$

Lack of Judgment and Insight (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{(x-3.2)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-1.42(x-2.45)}}$$

Disturbance of Volition (Range: 0 – 6):

$$\mu_{\mathrm{notofconcern}} = \frac{1}{1 + e^{1.129(x-3.01)}}$$

$$\mu_{\mathrm{ofconcern}} = \frac{1}{1 + e^{-3.1(x-1.3)}}$$

Poor Impulse Control (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1+e^{1.79(x-2.26)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1+e^{-1.42(x-2.45)}}$$

Preoccupation (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = \frac{1}{1+e^{1.37(x-2.68)}}$$

$$\mu_{\text{ofconcern}} = \frac{1}{1+e^{-1.39(x-2.91)}}$$

Active Social Avoidance (Range: 0 – 6):

$$\mu_{\text{notofconcern}} = e^{-\frac{x^2}{10}}$$

$$\mu_{\text{ofconcern}} = e^{-\frac{(x-6.12)^2}{9.4178}}$$

The membership functions of the output variables of FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5 are:

FIS-1: Variable Name: Delusions; Range:[0 - 2]:

$$\mu_{\text{low}} = e^{-\frac{x^2}{0.2683}}$$

$$\mu_{\text{medium}} = e^{-\frac{(x-1)^2}{0.2304}}$$

$$\mu_{\text{high}} = e^{-\frac{(x-2)^2}{0.2805}}$$

FIS-2: Variable Name: Hallucinations; Range: [0 – 3]:

$$\mu_{\text{low}} = e^{-\frac{x^2}{0.4564}}$$

$$\mu_{\text{medium}} = e^{-\frac{(x-1.5)^2}{0.51836}}$$

$$\mu_{\text{high}} = e^{\frac{-(x-3)^2}{0.38176}}$$

FIS-3: Variable Name: Disorganized Speech; Range: [0 – 3]

$$\mu_{\text{low}} = e^{\frac{-x^2}{0.43115}}$$

$$\mu_{\text{medium}} = e^{\frac{-(x-1.58)^2}{0.5356}}$$

$$\mu_{\text{high}} = e^{\frac{-(x-3)^2}{0.3748}}$$

FIS-4: Variable Name: Catatonic; Range: [0 – 2]

$$\mu_{\text{low}} = e^{\frac{-x^2}{0.4215}}$$

$$\mu_{\text{medium}} = \frac{1}{1+e^{-5.99(x-0.519)}} - \frac{1}{1+e^{-15.5(x-1.82)}}$$

$$\mu_{\mathrm{high}} = e^{\frac{-(x-2)^2}{0.04205}}$$

FIS-5: Variable Name: Negative Symptoms; Range: [0 - 4]

$$\mu_{\mathrm{low}} = e^{\frac{-x^2}{0.6389}}$$

$$\mu_{\mathrm{medium}} = \frac{1}{1+e^{-3.601(x-1)}} - \frac{1}{1+e^{-3.601(x-3)}}$$

$$\mu_{\mathrm{high}} = e^{\frac{-(x-4)^2}{0.9218}}$$

The Final Diagnosis – FIS-6, the fuzzy inference system that gives the final diagnosis, takes as its inputs the crisp outputs of FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5. The five input variables are delusions, hallucinations, disorganizedspeech, catatonicbehaviour and negativesymptoms. Each input variable has three membership functions, viz. “low”, “medium” and “high”. FIS-6 has a single output variable called finaldiagnosis, which has three membership functions – normal, psychotic and schizophrenic. The membership functions are given as:

FIS-6: Input variable names: delusions, hallucinations, disorganizedspeech, catatonicbehaviour and negativesymptoms; Range [0 – 6]:

$$\mu_{low} = e^{\frac{-x^2}{2}}$$

$$\mu_{medium} = e^{\frac{-(x-3)^2}{1.486}}$$

$$\mu_{high} = e^{\frac{-(x-6)^2}{2}}$$

FIS-6: Output variable name: finaldiagnosis; Range [0 – 2]

$$\mu_{normal} = e^{\frac{-x^2}{0.663}}$$

$$\mu_{psychotic} = \frac{1}{1+e^{-5.68(x-0.717)}} + \frac{1}{1+e^{-21.9(x-1.81)}}$$

$$\mu_{schizophrenic} = e^{\frac{-(x-2)^2}{0.04381}}$$

The rule base for FIS-6 is:

Rule1: If “catatonicbehaviour” is “high” or “negativesymtoms” is” high” then “finaldiagnosis” is “schizophrenic”

Rule2: If “catatonicbehaviour” is “medium” or “negativesymtoms” is “medium” then “finaldiagnosis” is “schizophrenic”

Rule3: If “delusions” is “high” and “hallucinations” is “high” then “finaldiagnosis” is “schizophrenic”

Rule4: If “hallucinations” is “high” and “disorganizedpseech” is “high” then “finaldiagnosis” is “schizophrenic”

Rule5: If “delusions” is “high” and “disorganizedspeech” is “high” then “finaldiagnosis” is “schizophrenic”

Rule6: If “delusions” is “medium” and “hallucinations” is “medium” then “finaldiagnosis” is “psychotic”

Rule7: If “hallucinations” is “medium” and “disorganizedpseech” is “medium” then “finaldiagnosis” is “psychotic”

Rule8: If “delusions” is “medium” and “disorganizedspeech” is “medium” then “finaldiagnosis” is “psychotic”

Rule9: If “delusions” is “medium” and “hallucinations” is “low” and “disorganizedspeech” is “low” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule10: If “delusions” is “low” and “hallucinations” is “medium” and “disorganizedspeech” is “low” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule11: If “delusions” is “low” and “hallucinations” is “low” and “disorganizedspeech” is “medium” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule12: If “delusions” is “high” and “hallucinations” is “low” and “disorganizedspeech” is “low” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule13: If “delusions” is “low” and “hallucinations” is “high” and “disorganizedspeech” is “low” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule14: If “delusions” is “low” and “hallucinations” is “low” and “disorganizedspeech” is “high” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “psychotic”

Rule15: If “delusions” is “low” and “hallucinations” is “low” and “disorganizedspeech” is “low” and “catatonicbehaviour” is “low” and “negativesymptoms” is “low” then “finaldiagnosis” is “normal”

The graphical representations of the various membership functions are depicted below in Figure 4.1-4 to Figure 4.1-17:

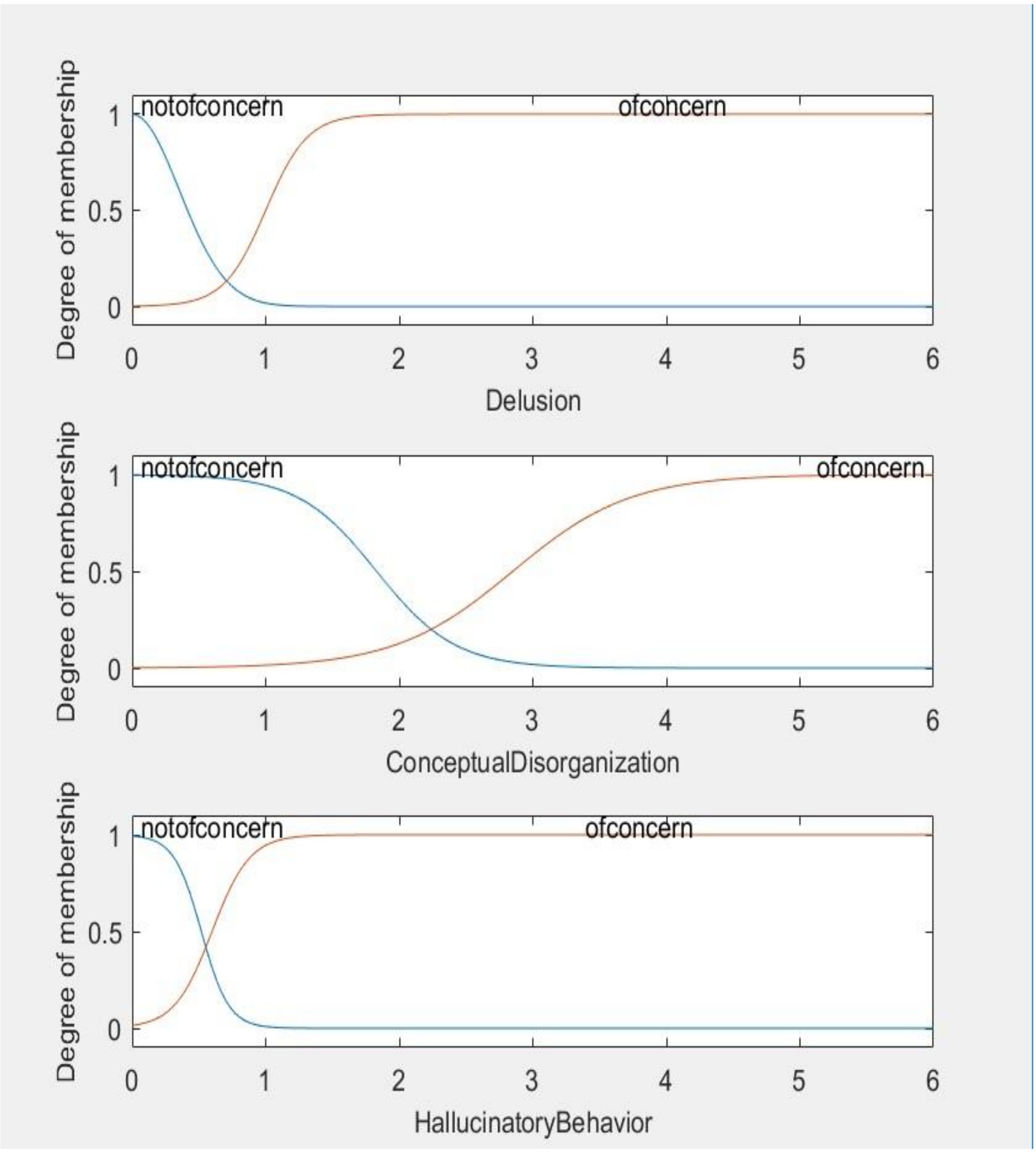

Figure 4.1-3: Input Membership Functions For Delusion, Conceptual Disorganization and Hallucination

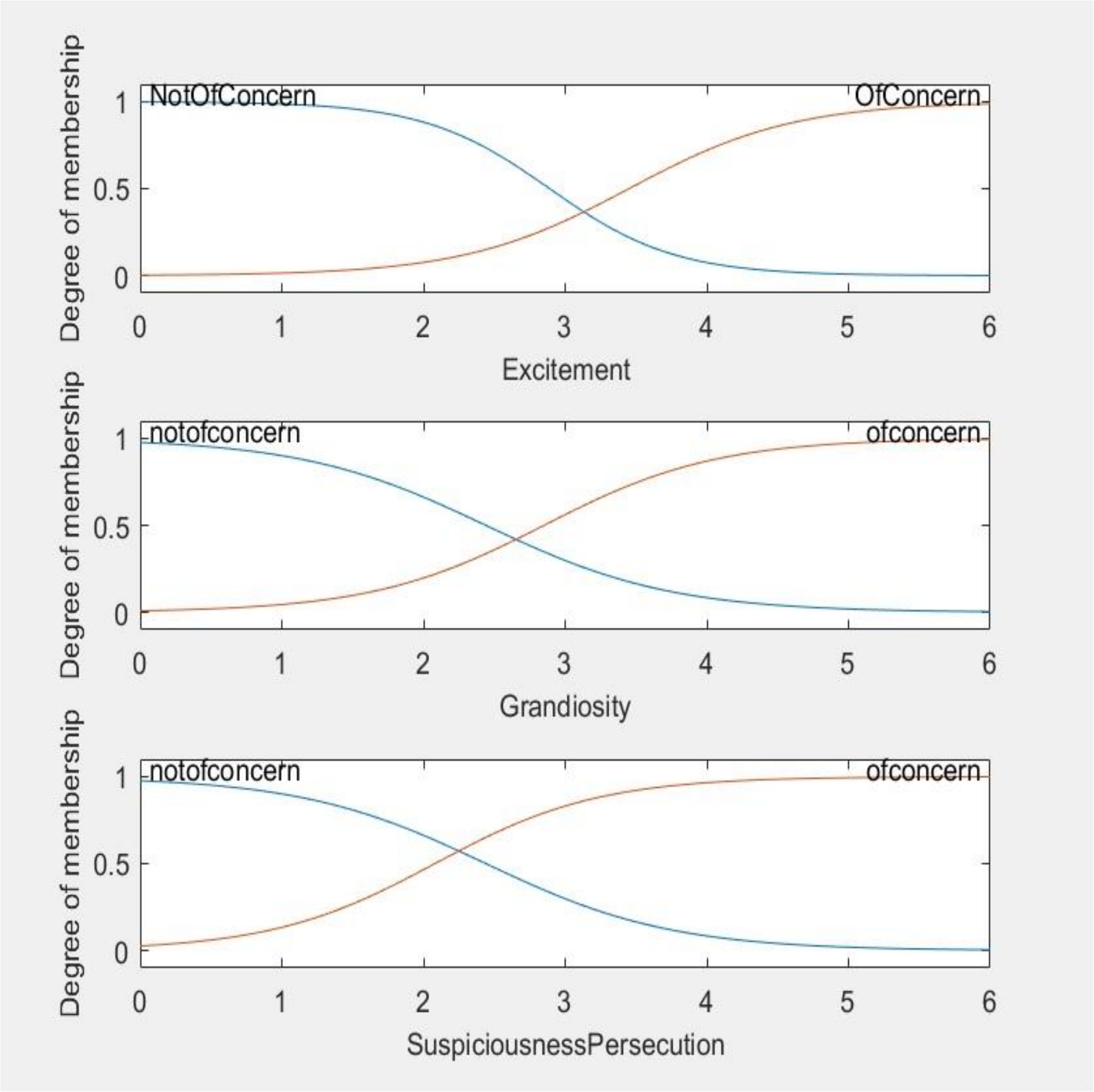

Figure 4.1-4: Input Membership Functions for Excitement, Grandiosity and Suspiciousness / Persecution

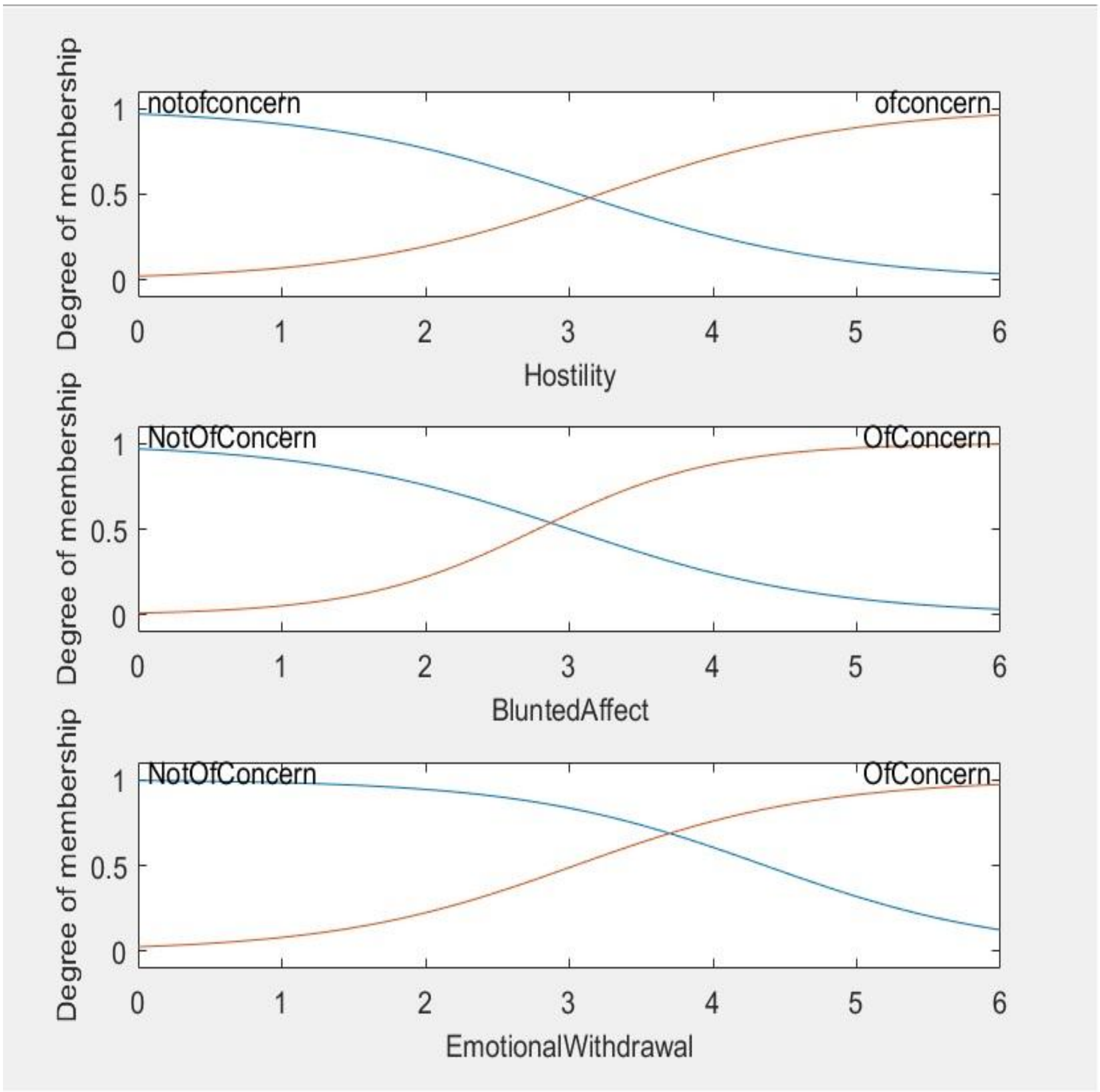

Figure 4.1-5: Input Membership Functions for Hostility, Blunted Affect and Emotional Withdrawal

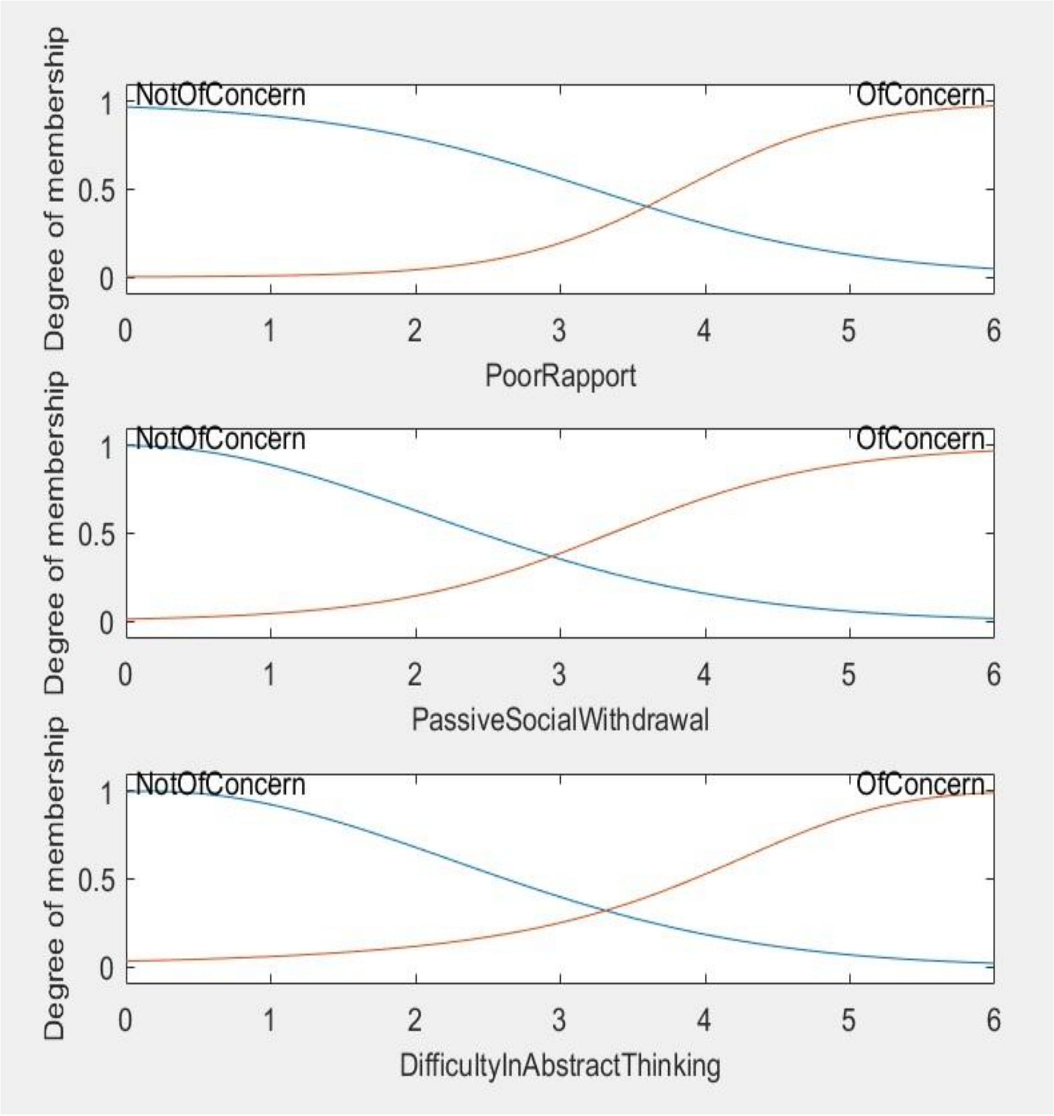

Figure 4.1-6: Input Membership Functions for "Poor Rapport", "Passive Social Withdrawal" and "Difficulty in Abstract Thinking"

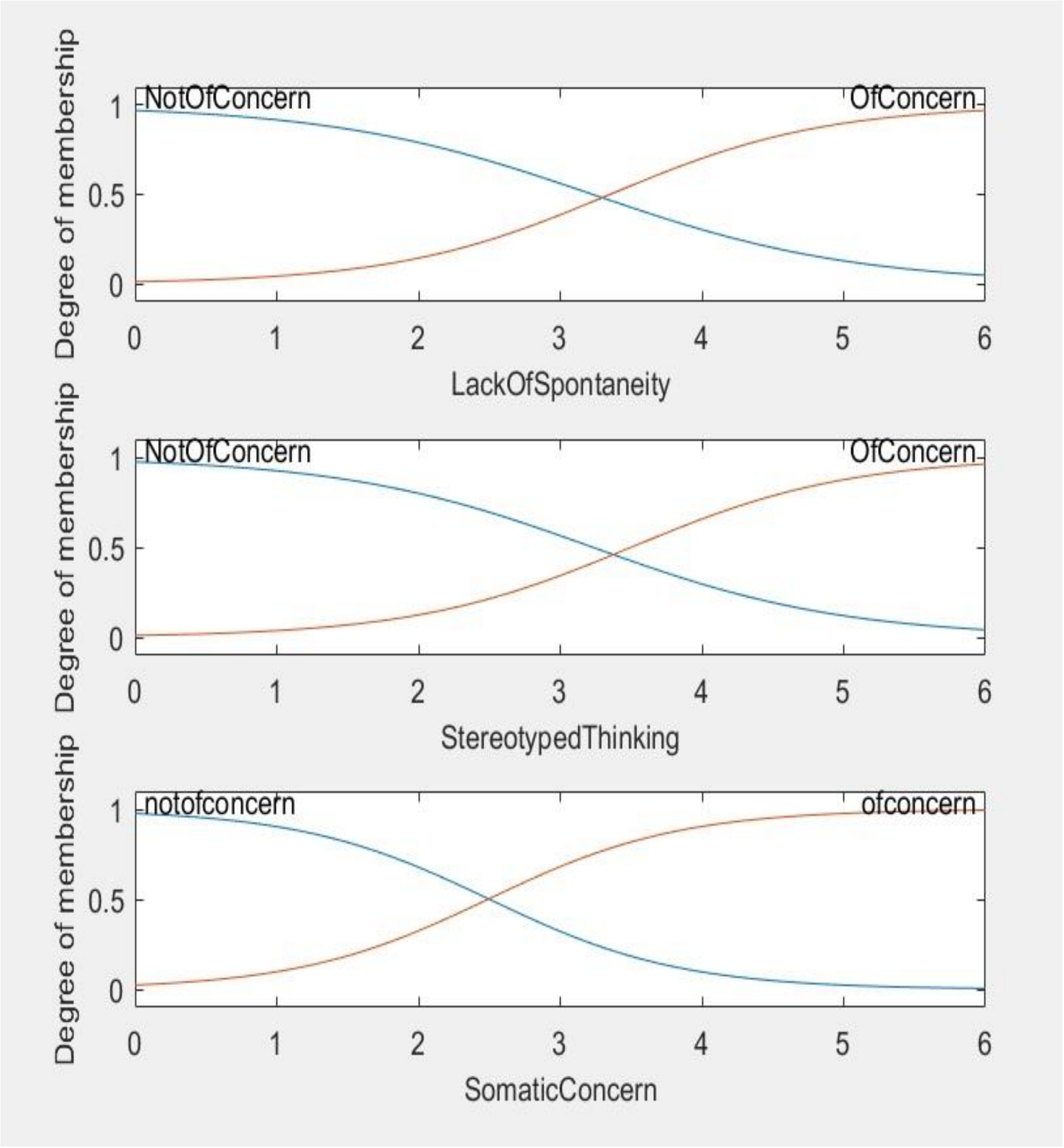

Figure 4.1-7: Input Membership Functions for "Lack of Spontaneity", "Stereotyped Thinking" and "Somatic Concern"

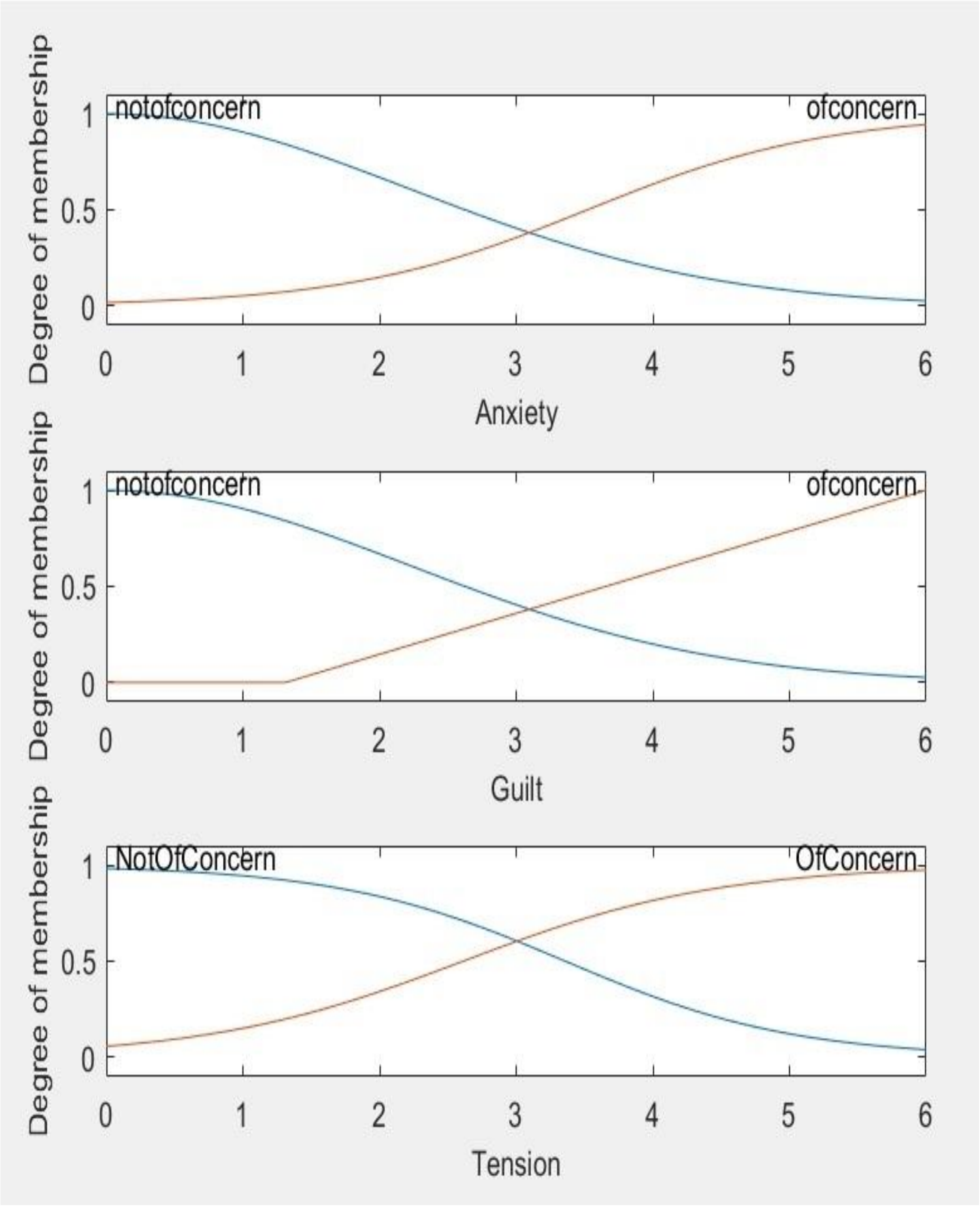

Figure 4.1-8: Input Membership Functions for Anxiety, Guilt and Tension

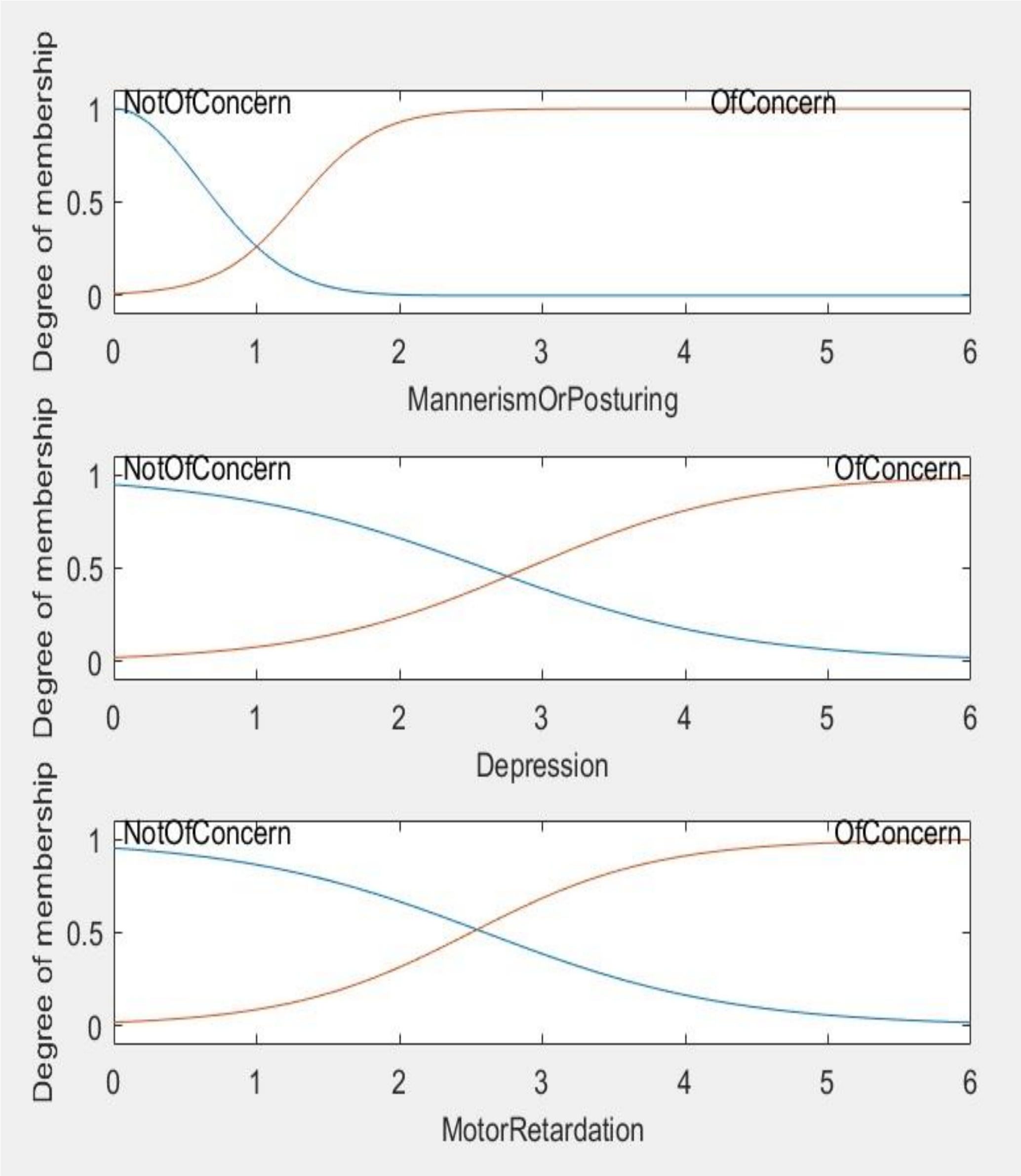

Figure 4.1-9: Input Membership Functions for "Mannerisms and Posturing", "Depression" and "Motor Retardation"

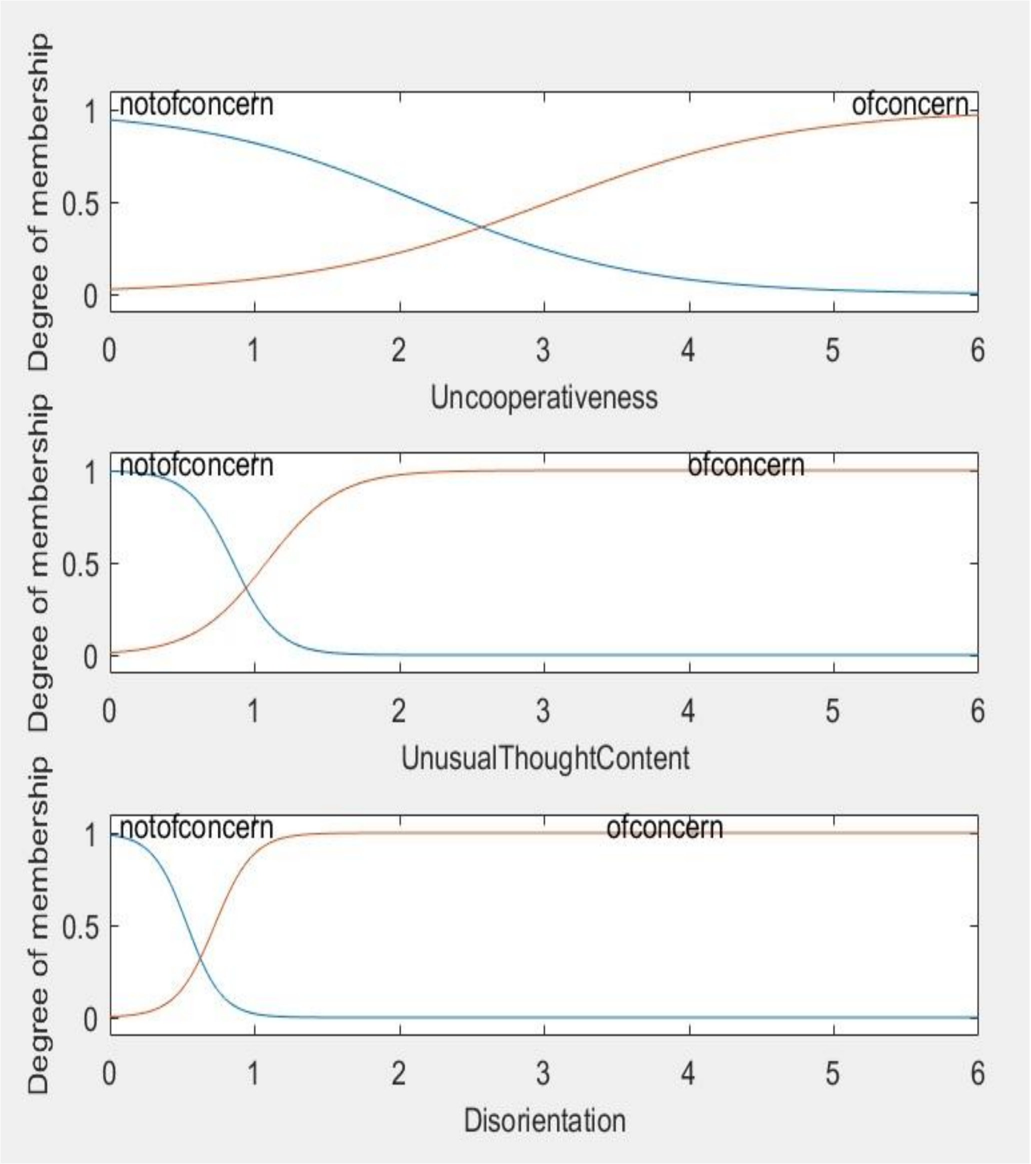

Figure 4.1-10: Input Membership Functions for "Uncooperativeness", "Unusual Thought Content" and Disorientation

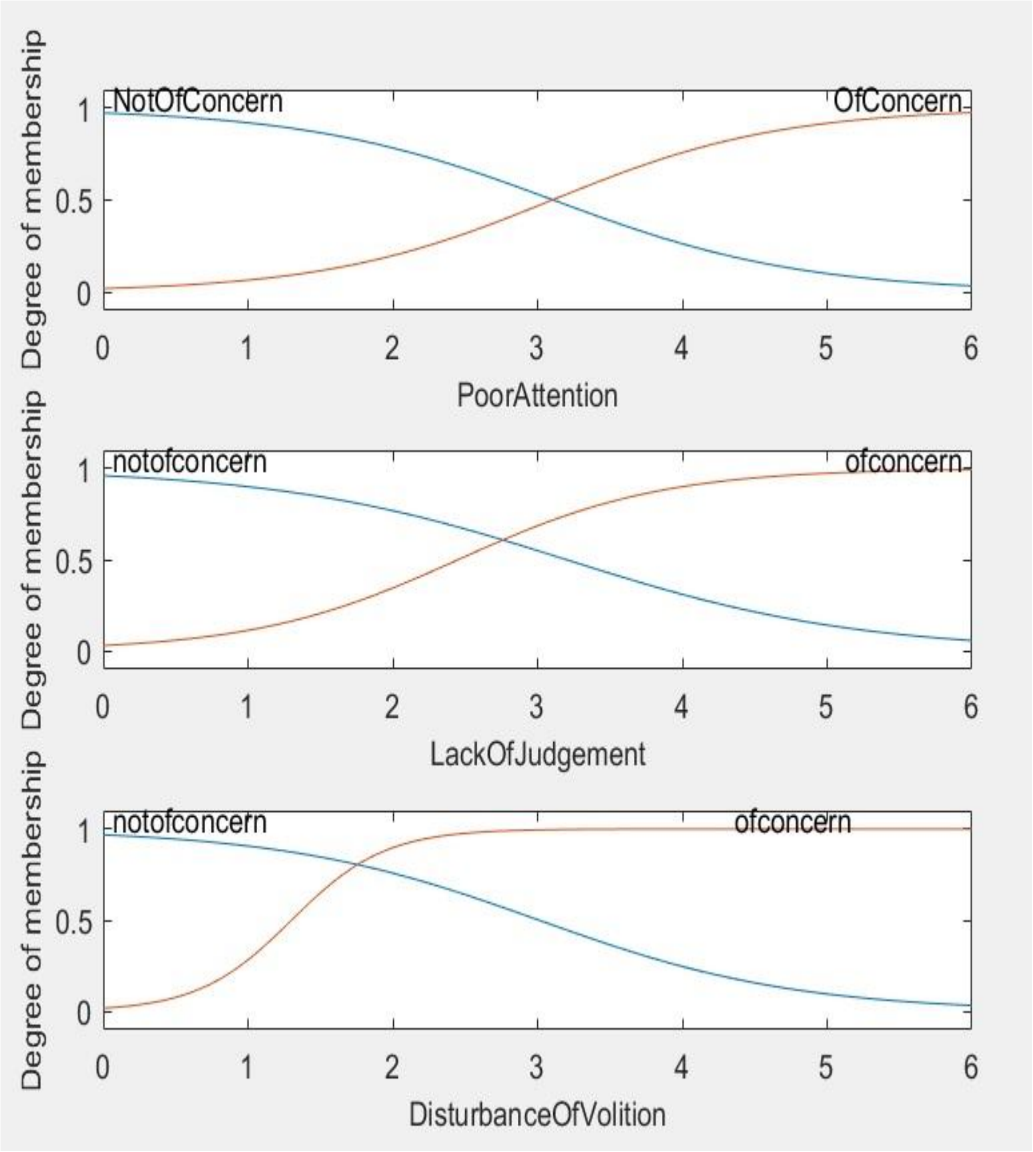

Figure 4.1-11: Input Membership Functions for "Poor Attention", "Lack of Judgment" and "Disturbance of Volition"

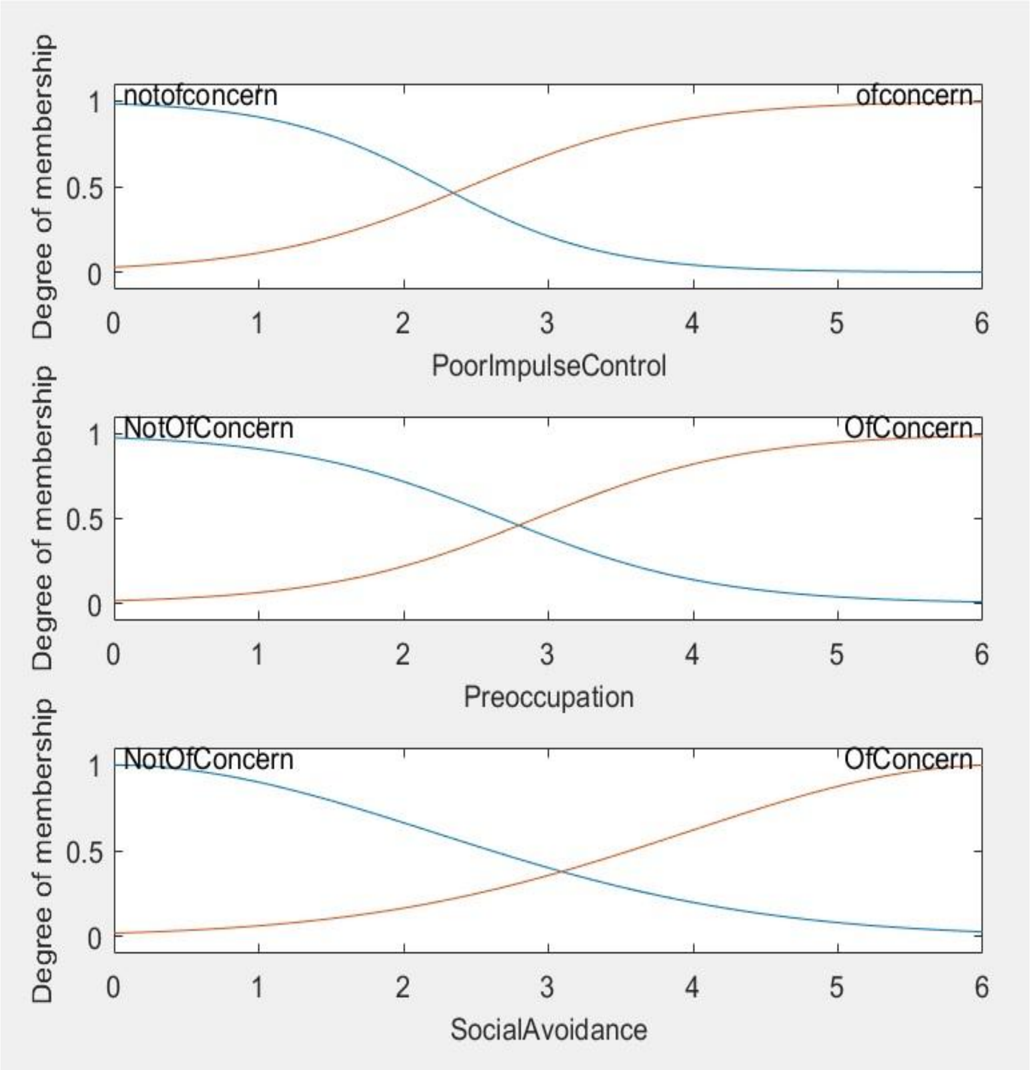

Figure 4.1-12: Input Membership Function for "Poor Impulse Control", Preoccupation and "Social Avoidance"

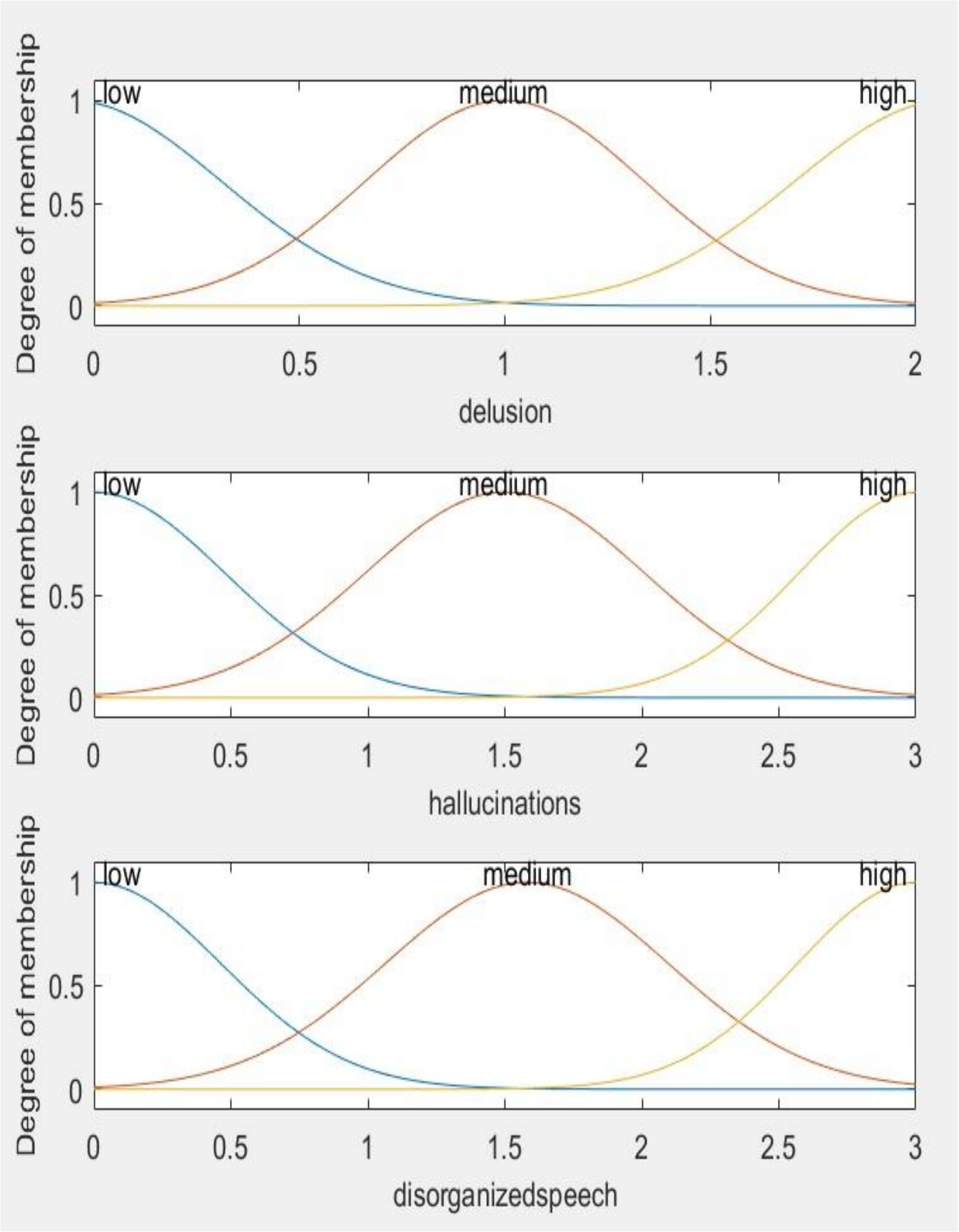

Figure 4.1-13: Output Membership Functions for FIS's Giving Outputs for Delusions, Hallucinations and "Disorganized Speech"

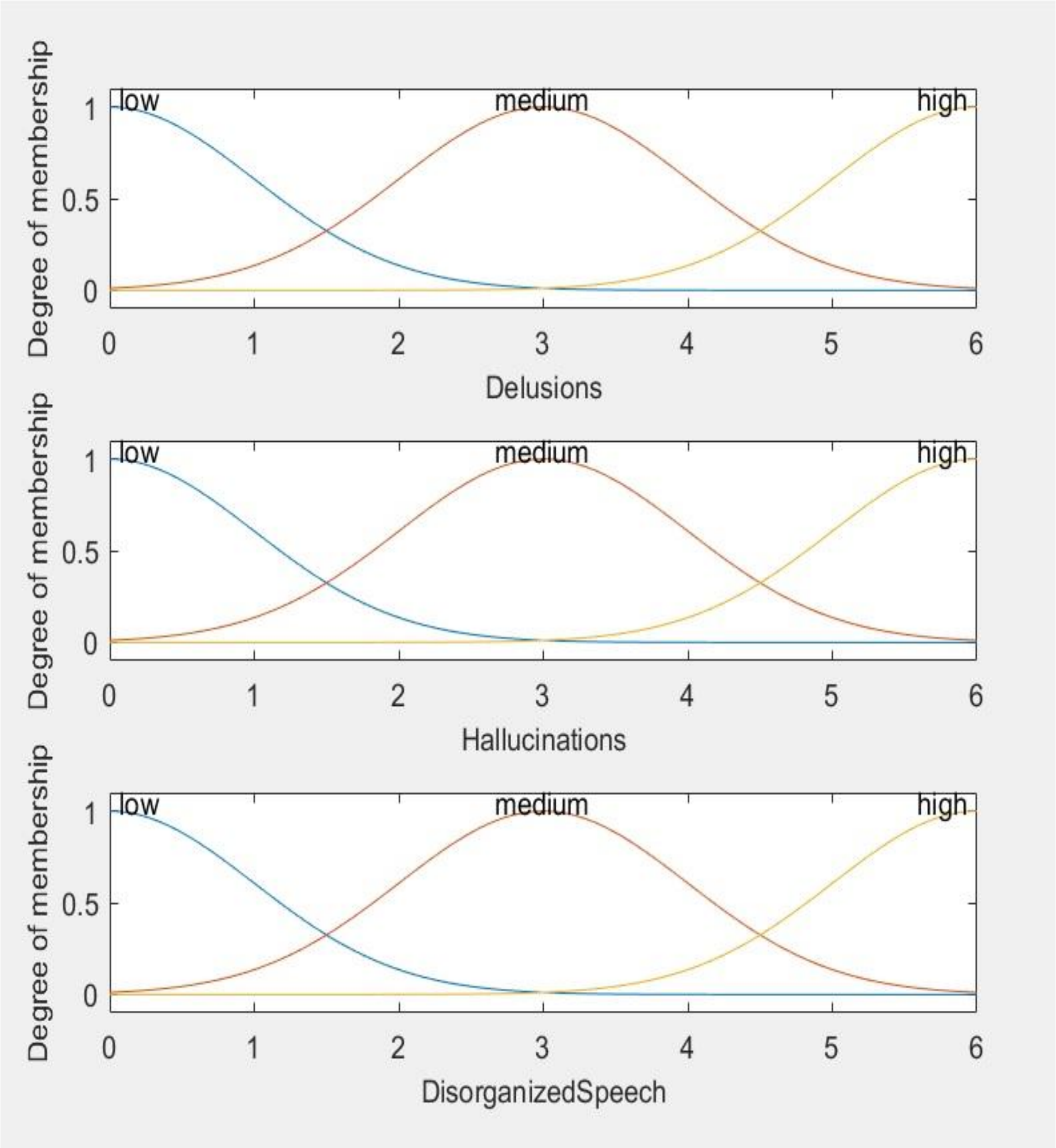

Figure 4.1-14: Input Membership Functions for Membership Variables Delusions, Hallucinations and DisorganizedSpeech of FIS-6

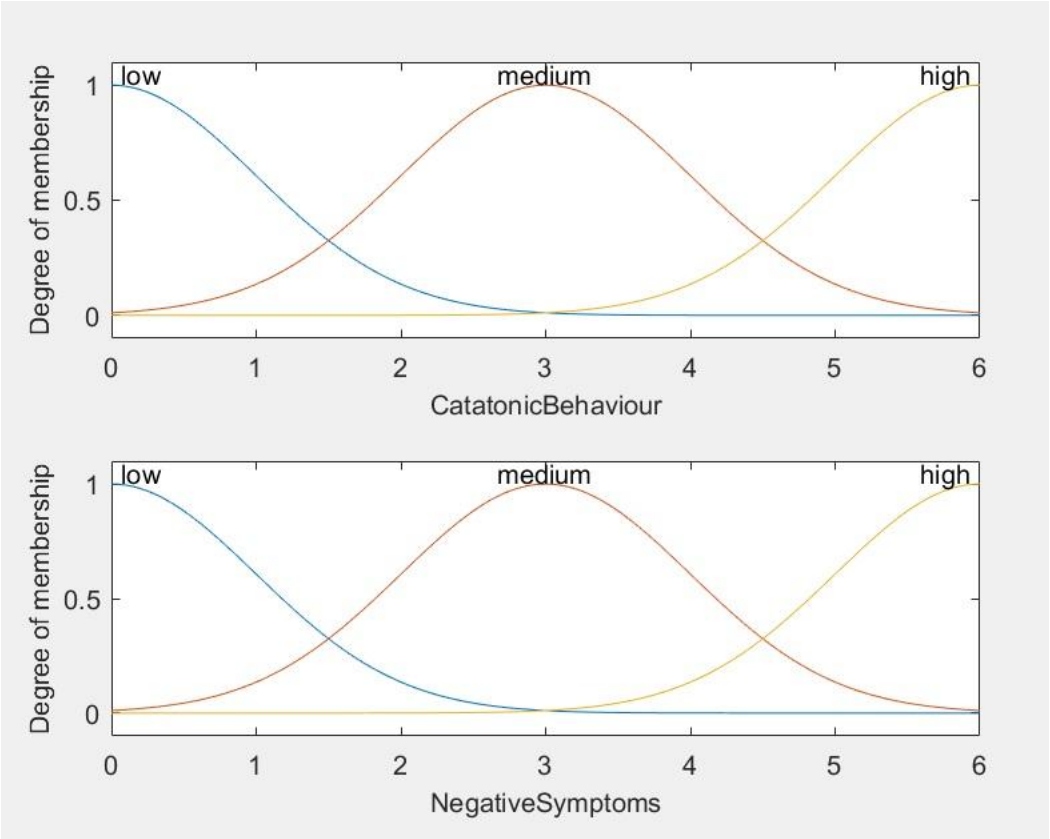

Figure 4.1-15: Input Membership Functions for Membership Variables CatatonicBehaviour and NegativeSymptoms of FIS-6

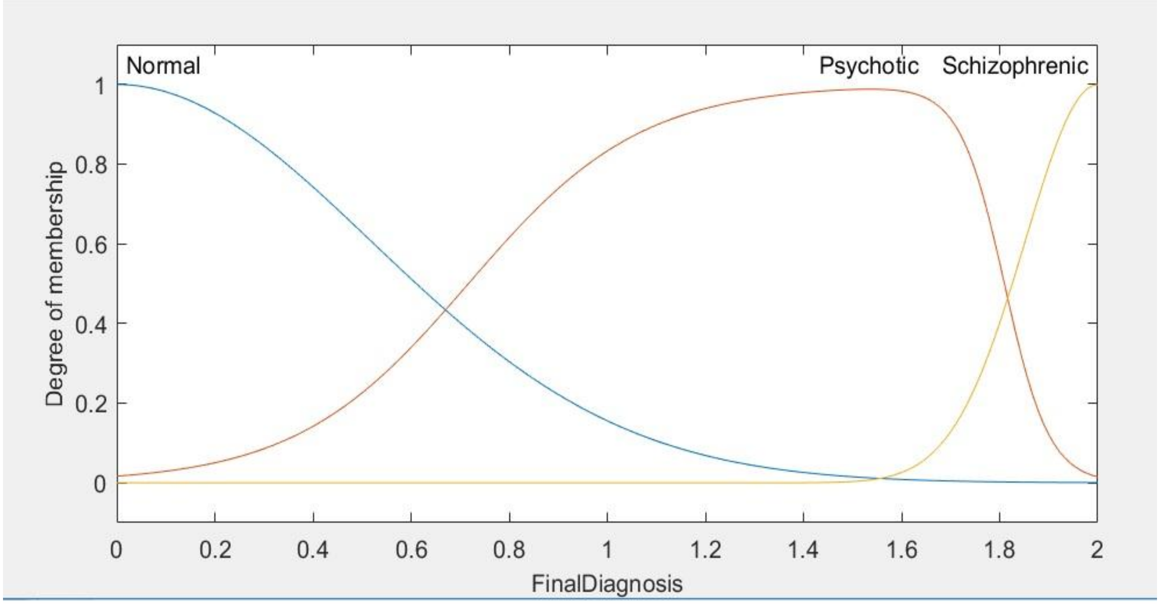

Figure 4.1-17: Output Membership Functions for Membership Variable FinalDiagnosis of FIS-6

4.2. Experiment 2 – Synthesis of Training Dataset for Artificial Neural Network for Diagnosing Schizophrenia

A subject’s PANSS ratings may be fed to a neural network which would then return a diagnosis of schizophrenia or otherwise. If the output of the neural network is closer to 0 than 1 for a particular input, then the subject corresponding to that input is not schizophrenic. If the output is closer to 1 than to 0, then the subject is schizophrenic. However, it is difficult to obtain the necessary number and variety of readings from real subjects, so, the training data may be synthesized with the help of a fuzzy inference system that diagnoses schizophrenia from the psychometric test ratings. The synthetic data obtained in such manner may be used to train neural networks that may be used to diagnose subjects based on the results of psychometric tests.

Artificial neural networks have been used to classify subjects as schizophrenic or otherwise based on resting state functional network connectivity [**57**], blood-based gene expression signatures [**58**], and eye tracking [**59**]. However these methods involve collection of medical data from real subjects– a process that is expensive and inconvenient. On the other hand, PANSS offers a convenient and inexpensive way of assessing a patient’s mental state. Each dimension of the PANSS scale would serve as an input dimension of the neural network. However before such a network can be used, it would be necessary to first train it. Training a neural network requires a large amount of data that must be representative of practical input types. It is difficult to collect so many data points from actual schizophrenics, and so it becomes necessary to synthesize the data. Ulloa *et al* showed that synthetic data may be generated for training deep networks with structural MRI images [**60**]. Castro *et al* showed the synthesis of structural magnetic resonance imaging data that grew the original dataset by a factor of ten [**61**].

In this experiment, synthetic data based on PANSS ratings has been generated. The fuzzy expert system designed as part of Experiment-1 has been deployed to provide data points for training the artificial neural network. Pseudo-random input combinations are fed into the system which returns a crisp output as diagnosis of schizophrenia. For training the neural network, the pseudo-random input combination is taken as input and 0 or 1 as output. The output is assigned a value of 0 or 1 depending on the crisp output of the said fuzzy inference system. As the PANSS is a thirty-item scale and each item can be rated from zero to six, a total of 7^{30} different input combinations are possible. The simplest and easiest thing to do would be to feed each combination into the fuzzy expert system and note the diagnosis. That way, one would end up with 7^{30} data points for training the neural network. However, the above strategy would be unacceptable on a standard PC because of the large amount of data involved. Instead, input combinations must be taken selectively. As described in the previous section, five fuzzy inference systems, FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5 are used to take various items of the PANSS scale as inputs and give outputs of the subject's level of delusions, hallucinations, disorganized speech, catatonic behaviour and negative symptoms. A sixth fuzzy inference system, FIS-6 takes the outputs of FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5 as inputs and gives a final diagnosis. The relationship between the fuzzy inference systems and the PANSS items are tabulated below in Table 4.2-1:

Table 4.2-1: Inputs and Outputs of the Fuzzy Inference Systems

Name of Fuzzy Inference System	Inputs (Items on the PANSS Scale)	Degree of Impact on Fuzzy Inference System Output	Output of Fuzzy Inference System
FIS-1	Delusions	High	Delusions
	Suspiciousness / Persecution		
	Hostility		
	Unusual Thought Content		
	Disturbance of Volition		
	Conceptual Disorganization		
	Grandiosity	Medium	
	Lack of Judgement		
	Uncooperativeness		
	Anxiety	Low	
	Guilt		
FIS-2	Hallucinatory Behaviour	High	Hallucinations
	Disturbance of Volition		
	Poor Impulse Control	Medium	
	Somatic Concern	Low	
FIS-3	Stereotyped Thinking	High	Disorganized_Speech
	Lack of Spontaneity		
	Difficulty in Abstract Thinking	Medium	
	Disorientation	Low	
FIS-4	Mannerisms and Posturing	High	Catatonic_Behaviour
	Motor Retardation		
	Stereotyped Thinking		
	Depression		
	Excitement		
	Depression	Medium	
	Anxiety		
	Disorientation		
	Tension		
	Poor Attention	Low	

Name of Fuzzy Inference System	Inputs (Items on the PANSS Scale)	Degree of Impact on Fuzzy Inference System Output	Output of Fuzzy Inference System
	Preoccupation		
FIS-5	Blunted Affect	High	Negative_Symptoms
	Emotional Withdrawal		
	Social Avoidance		
	Passive / Apathetic Social Withdrawal		
	Poor Rapport	Medium	
	Lack of spontaneity and Flow of Conversation		
	Difficulty in Abstract Thinking	Low	
	Stereotyped Thinking		

The strategy is to take selective permutations of the various input combinations and present them to FIS-1, FIS-2, FIS-3, FIS-4 and FIS-5. A psychiatrist was consulted to divide each input variable range, zero to six, into three regions – high, medium and low, depending on whether the input variable is highly, moderately or only slightly indicative of schizophrenia. Also, from the table, it is seen that all inputs to a given fuzzy inference system may be classed as high-impact, medium-impact and low-impact. The rules of the fuzzy inference system are such that the following permutations of data would give the largest variety of outputs with an optimum number of data points:

High – High – High

Medium – Medium – Medium

Low-Low-Low

High-Medium-Low

Low-Medium-high

Low-Low-Medium

Low-Low-High

To explain this grouping, let's consider the specific example of FIS-2, the output of which is "hallucinations". The input variables for this fuzzy inference system are: "hallucinatory behaviours", "disturbance of volition", "poor impulse control" and "somatic concern". Of these variables, "hallucinatory behaviour" and "disturbance of volition" have a high effect on an outcome of hallucinations, "poor impulse control" has moderate effect while "somatic concern" has a low effect. Again, the range of evaluation for each variable is [0 – 6]. For each variable, this range is divided into low, medium and high. For example, for "hallucinatory behaviour", 0 is considered low, 1 is considered medium and [2 – 6] is considered high. Similarly, for “disturbance of volition”, [0 – 1] is considered low, 2 is considered medium while [3 – 6] is considered high. For “poor impulse control”, [0 – 1] is considered low, [2-3] is considered medium and [4-6] is considered high. For “somatic concern”, [0 – 1] is considered low, [2 – 3] is considered medium and [4 – 6] is considered high. So, for example, “low-medium-high” means that the permutations consist of low readings from the first group input variables that have a high impact on the output of FIS-2, viz., “hallucinatory behaviour” and “disturbance of volition”, medium readings from “poor impulse control” and high readings from “somatic concern”. Therefore, the set “low-medium-high” would contain the following permutations:

[0, 0, 2, 4] ; [0, 0, 2, 5]; [0, 0, 2, 6]; [0, 0, 3, 4]; [0, 0, 3, 5]; [0, 0, 3, 6]; [0, 1, 2, 4]; [0, 1, 2, 5]; [0, 1, 2, 6]; [0, 1, 3, 4]; [0, 1, 3, 5]; [0, 1, 3, 6]

Similar permutations are taken for FIS-1, FIS-3, FIS-4 and FIS-5.

These input combinations are fed to the fuzzy inference system under consideration, (FIS-2, say) and outputs are generated. In order to keep the number of data points manageably modest, only those data points are chosen for which outputs are unique.

In the next step, the data sets obtained from the different fuzzy inference systems are unified. Since FIS-2 has just four contributory PANSS items (Refer to Table 1), it generates a four dimensional data. Likewise, FIS-1 generates eleven dimensional data, and so on. Observe how the data sets for FIS-1 and FIS-2 have a common parameter, viz. “disturbance of volition”. Upon unifying the data sets of FIS-1 and FIS-2, what is obtained is fourteen dimensional data – the common dimension, “disturbance of volition” is considered only once. The unification is done by combining each data point of FIS-2 with each data point of FIS-1 wherever the value of the common parameter (“disturbance of volition”) is same. For example, suppose the FIS-2 data set consists of the following data points:

[0, 1, 0, 0 | 0.4678]; [3, 3, 2, 1| 1.2968]; [2, 1, 1, 3 | 1.0156]

Also, suppose the FIS-1 input data set consists of the following data points:

[1, 2, 2, 1, 0, 3, 2, 1, 1, 0, 4 | 0.9873]; [2, 4, 0, 1, 3, 2, 2, 1, 3, 2, 3 | 1.4438]

In the above, the field after the ‘|’ denotes the output of the respective fuzzy inference system. The result of the unification of FIS-2 and FIS-1 would be:

[3, 3, 2, 1, 2, 4, 0, 1, 2, 2, 1, 3, 2, 3 | 1.2968, 1.4438]

Note the unification happens only where “poor impulse control” (second field in FIS-2 dataset and fifth field in FIS-1 dataset) has the same value in both datasets. The result of unification also gives a two-dimensional output data-point: [1.2968, 1.4438]

The unified dataset is now again unified with the dataset for the next fuzzy inference system (FIS-3, say) and the process is repeated till all five datasets are unified. What is obtained is N number of thirty-dimensional input data points and N number of five-dimensional output data points. The five-dimensional data is fed to the FIS-6 fuzzy inference system to obtain a crisp output. A psychiatrist was consulted and determined that subjects for whom the output of FIS-6 exceeds 1.26 are the ones who need treatment for schizophrenia. Hence the crisp outputs for these subjects are set to 1 and the crisp outputs of the other subjects are set to 0.

Thus a set of N thirty-dimensional input data points and a set N single-dimensional output data points is obtained. These data points can be used to train the artificial neural network.

4.3. Experiment 3 – Observation of Training Neural Network for Diagnosing Schizophrenia

Artificial intelligence is being increasingly used to diagnose disorders including mental illness. At the centre of the diagnostic system is the artificial neural network, which can classify subjects based on their thirty-dimensional PANSS ratings as either schizophrenic or not schizophrenic. A simple perceptron is good enough to classify data that is linearly separable; however in case of non-linearly separable data, it is necessary to project the data into a higher dimension. This is where the multi-layer perceptron (MLP) comes in. The MLP has one input layer, one output layer and one or more hidden layers. The number of nodes in the input layer is equal to the dimension of the input data, which in this case is thirty. The number of nodes in the output layer is one; and the number of nodes in the hidden layer as well as the number of hidden layers may be varied to get optimum classification performance out of the multi-layer perceptron.

Many researchers have tried to establish the ideal number of hidden layers and the ideal number of nodes in each hidden layer. If N_t is the number of training samples, N_i is the number of input nodes, N_h is the number of neurons in the hidden layer, and N_o is the number of output nodes, then the various values of N_h may be arrived at by the following methods:

According to Li, Chow and Yu's method [**62**] :

$$N_h = \frac{\sqrt{1+8N_i}-1}{2}$$

According to Tamura and Tateishi's method [**63**]:

$$N_h = N_i - 1$$

According to Xu and Chen's method [**64**]:

$$N_h = \frac{1}{2} \frac{N_t}{N_i \log N_t}$$

According to Shibata and Ikeda's method [**65**]:

$$N_h = \sqrt{N_i N_o}$$

According to Sheela and Deepa's method [**66**]:

$$N_h = \frac{4\,{N_i}^2 + 3}{{N_i}^2 - 8}$$

According to Trenn [**66**]:

$$N_h = \frac{(N_i + N_o - 1)}{2}$$

Apart from the above, there are several rules of thumb [**67**], viz.

- Size of hidden layer must be between size of input layer and output layer.
- Size of hidden layer must be two-thirds the size of the input layer plus the size of the output layer.
- Size of hidden layer must be less than twice the size of the input layer.

As for the number of hidden layers, the common consensus is that one or two hidden layers are sufficient for most situations [**68**] [**69**].

The MATLAB neural network toolbox was used for creating and training the MLPs. For training the neural network we used 960 training samples that were generated as an outcome of Experiment-3. The samples were randomly distributed into training, validation and testing sets by the MATLAB software. The Levenberg-Marquardt training algorithm was used and the mean square error was considered as the error criterion.

Many different models of neural network with varying number of hidden nodes were created. First, the training was done on a neural network with a certain number of nodes in a single hidden layer. Then the training was repeated for a neural network with two hidden layers and the same number of nodes as above in each hidden layer.

The same training pairs were used to train all models of the neural network.

When deploying an artificial neural network in software, the user has the flexibility of adding as many or as few neurons as he or she wants. However there may be a situation wherein a user must use a hardware implementation of a neural network [**70**] [**71**] [**72**] which does not offer the same flexibility. Under such a circumstance, the user may continue to get good performance out of the neural network even if the number of neurons in the hidden layer is larger than what is recommended, provided the neural network has just one hidden layer. However, the validation performance will be unacceptably poor if multiple hidden layers are present. The conclusion is that in the matter of classifying data for schizophrenia patients, the user must avoid using neural networks with more than one hidden layer.

4.4. Experiment 4 – Choice of Support Vector Machine Kernel for Classifying Schizophrenia Data

A kind of feedforward network called "Support Vector Machines" may be used to classify data on several diseases like heart disease, lymph diseases, cancer, acute coronary syndrome etc. They may also be used to classify schizophrenics and non-schizophrenics based on PANSS data. In order to do so it is first necessary to train the SVM using training data points.

A fuzzy expert system was used to generate the training data points. The expert system takes as input the various different possible PANSS readings, and gives a diagnosis regarding whether or not the PANSS reading corresponds to a schizophrenic or a normal individual. The fuzzy expert system consists of six fuzzy inference systems, five of which deal directly with the PANSS ratings while the sixth takes as input the outputs generated by the five fuzzy inference systems.

Ideally, 7^{30} different PANSS readings are possible. In order to limit the number of PANSS ratings so that the program would run on a standard PC, some custom MATLAB code was designed and executed. The result of the above exercise is a set of 960 training samples containing a mix of data points with positive as well as negative expected outputs.

The above samples were further divided into two groups of 480 samples each, such that each group would contain positive as well as negative data points. One group was used for training the SVM while the other was used for testing the accuracy of classification.

SVMs with various different kernels were created and used with the above data.

4.5. Experiment 5 – Fuzzy Clustering for Diagnosing Schizophrenia

Fuzzy clustering is a kind of soft clustering technique that groups data into two or more clusters. In hard clustering, a data point either belongs completely to a cluster or does not belong to that cluster at all. In soft clustering however, there can be overlap between two clusters, and a data point may belong partially to two or more clusters. However, the sum of membership values for all the clusters for any given data point must be 1. The number of clusters may be specified by the user.

A trivial case would be that just one cluster is formed. The 960 data points generated as an outcome of Experiment-2 is clustered into two clusters using the fuzzy "fcm" utility. The option that signifies cluster overlap is specified as 1.1. The data points generated as an outcome of Experiment-2 are valid PANSS scores for hypothetical subjects. Some of these subjects are schizophrenic while the others are not. Upon clustering the data, the "fcm" utility is able to correctly separate the schizophrenic data points from the non-schizophrenic data points.

Next, the PANSS scores for four actual subjects are added to the pool of data points such that there are now 964 data points. The clustering utility is again executed and the membership values for the real subjects are checked. The real subjects include two schizophrenic and two non-schizophrenic individuals. One expects that "fcm" would be able to correctly include the four subjects into the two clusters. The results of clustering can provide valuable insights into the symptoms and diagnosis of the illness.

5. Results and Discussion

5.1. Experiment1 – Fuzzy Based System for Diagnosing Schizophrenia

The curves for Normal and Psychotic output membership functions of FIS-6 intersect for a value of FinalDiagnosis = 0.7. This is the point at which the psychiatrist's diagnosis tends to favour a diagnosis of psychosis which is not necessarily schizophrenia. Similarly, the curve for Psychotic and Schizophrenic output membership functions indicates that for a FinalDiagnosis of 1.82, the physician would be more certain of a diagnosis of schizophrenia than general psychosis.

The fuzzy based system gives excellent results where diagnosing schizophrenia is concerned. Consider the following PANSS ratings for actual patients given in Table 5.1-1:

Table 5.1-1: PANSS Scores of Four Actual Subjects

PANSS Items	Patient Ratings; Scale: [0 – 6]			
	Patient01	Patient02	Patient03	Patient04
Delusions	1	0	0	4
Conceptual Disorganization	1	0	0	2
Hallucinatory Behaviour	1	1	1	3
Excitement	2	0	1	2
Grandiosity	2	2	1	0
Suspiciousness / Persecution	1	0	0	5
Hostility	3	3	2	4
Blunted Affect	1	0	0	3
Emotional Withdrawal	3	0	2	5
Poor Rapport	3	2	2	2
Passive / Apathetic Social Withdrawal	2	0	1	3
Difficulty in Abstract Thinking	2	0	1	3
Lack of Spontaneity and Flow of Conversation	3	2	2	5
Stereotyped Thinking	1	0	0	5
Somatic Concern	3	0	2	2
Anxiety	4	0	2	3
Guilt Feelings	4	2	2	3

PANSS Items	Patient Ratings; Scale: [0 – 6]			
	Patient01	**Patient02**	**Patient03**	**Patient04**
Tension	4	0	2	4
Mannerisms and Posturing	0	0	0	1
Depression	5	2	4	5
Motor Retardation	2	0	1	3
Uncooperativeness	4	2	2	4
Unusual Thought Content	1	0	0	4
Disorientation	3	0	2	2
Poor Attention	4	1	2	2
Lack of Judgement and Insight	2	0	1	3
Disturbance of Volition	1	0	0	5
Poor Impulse Control	3	3	2	3
Preoccupation	3	0	2	4
Active Social Avoidance	3	2	2	2

Patient01 was diagnosed with Bipolar Personality Disorder as well as schizophrenia. She was put on medication for schizophrenia as well as Bipolar Personality Disorder. In the beginning, there was much confusion among clinicians regarding her diagnosis. The treating psychiatrist could not be certain of a diagnosis of schizophrenia even though patient exhibited marked psychosis. However, treatment for schizophrenia helped to reduce her symptoms. The crisp output of the FIS-6 fuzzy inference system for Patient01 was 1.4040, which is somewhat greater than 1.26. This indicates a diagnosis of schizophrenia which agrees with the ultimate clinical outcome of the treatment for schizophrenia.

As such, Patient02 was not afflicted with any illness, but demonstrated marked psychosis on account of substance addiction. He was uncooperative with his care-givers and initially refused to seek professional help, though he had guilty feelings about his addiction. He also experienced auditory hallucinations which was owing to the addiction. The crisp output of the FIS-6 fuzzy inference system for Patient02 was 1.1787. According to our solution, he is not schizophrenic, which agrees with the clinical diagnosis.

Patient03 was not schizophrenic. He experienced acute depression on account of which he was indulged in substance addiction. Because of his difficult family situation he experienced anxiety and guilt feelings. His substance addiction made him hallucinate and triggered an allergic reaction which caused some somatic concern, viz., crawling sensation on the skin. The patient was treated for depression and afterwards put on therapy for his addiction. The crisp output of the FIS-6 fuzzy inference system for Patient03 was 1.1937 which agrees with his diagnosis.

Patient04 was diagnosed as schizophrenic. His family sought treatment very late – at a stage when he was catatonic. The patient reported hearing voices in his head and possessed far-fetched beliefs that had no link to reality. The crisp output of the FIS-6 fuzzy inference system for Patient04 was 1.6314 which agrees with his diagnosis.

5.2. Experiment2 – Synthesis of Training Dataset for Artificial Neural Network for Diagnosing Schizophrenia

A pattern recognition neural network was trained with the data generated. The neural network was configured with thirty input nodes, one output node and two hidden layers. Each hidden layer had thirty nodes. The results of training the network are shown in Figure 5.2-1:

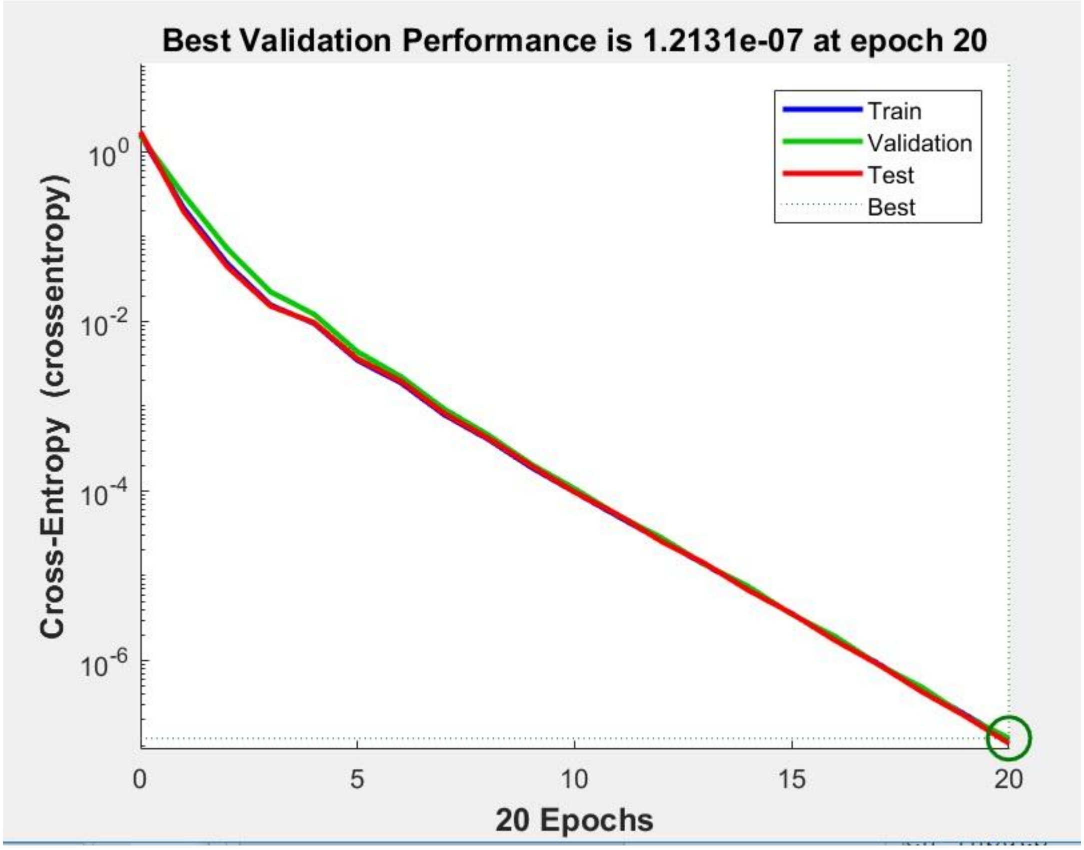

Figure 5.2-1: Validation Performance of 2-layer Neural Network Trained with Synthetic PANSS Data

Now to test the neural network with actual data from subject’s afflicted with mental disorders that may or may not be schizophrenia and let us compare the results with actual diagnosis in Table 5.2-1.

Table 5.2-1: Neural Network Outputs While Classifying Four Real Subjects

PANSS Items	Patient Ratings (Scale: 0 – 6)			
	Patient 1	Patient2	Patient3	Patient4
Delusions	1	0	0	4
Conceptual Disorganization	1	0	0	2
Hallucinatory Behaviour	1	1	1	3
Excitement	2	0	1	2
Grandiosity	2	2	1	0
Suspiciousness / Persecution	1	0	0	5
Hostility	3	3	2	4
Blunted Affect	1	0	0	3
Emotional Withdrawal	3	0	2	5
Poor Rapport	3	2	2	2
Passive / Apathetic Social Withdrawal	2	0	1	3
Difficulty in Abstract Thinking	2	0	1	3
Lack of Spontaneity and Flow of Conversation	3	2	2	5
Stereotyped Thinking	1	0	0	5
Somatic Concern	3	0	2	2
Anxiety	4	0	2	3
Guilt Feelings	4	2	2	3
Tension	4	0	2	4
Mannerisms and Posturing	0	0	0	1
Depression	5	2	4	5
Motor Retardation	2	0	1	3
Uncooperativeness	4	2	2	4
Unusual Thought Content	1	0	0	4
Disorientation	3	0	2	2
Poor Attention	4	1	2	2
Lack of Judgment and Insight	2	0	1	3
Disturbance of Volition	1	0	0	5
Poor Impulse Control	3	3	2	3
Preoccupation	3	0	2	4
Active Social Avoidance	3	2	2	2
Output From Neural Network	0.9944	3.4194e-05	0.0194	1
Actual Diagnosis of Schizophrenia (Yes/No)	Yes	No	No	Yes

If the output of the neural network is closer to one than to zero, it means that that input data vector corresponds to a schizophrenic. If the output is closer to zero than to one, then that corresponding input vector corresponds to a non-schizophrenic.

Hence, it is seen that the neural network can perform well even when trained with synthetic data.

5.3. Experiment3 – Observation on Training Neural Network for Diagnosing Schizophrenia

Neural networks with varying number of nodes were created and tested with the synthetic data. The validation error obtained in each training instance is tabulated below in Table 5.3-1:

Table 5.3-1: Validation Performance with Various Neural Network Configurations

No. of nodes in each hidden layer	Best Validation Performance	
	No. of hidden layers = 1	No. of hidden layers = 2
4	5.3428e-15	4.0688e-15
5	1.2837e-18	5.4858e-15
7	6.7224e-16	8.7213e-16
15	2.3247e-15	3.018e-15
20	2.1164e-15	9.2449e-16
29	4.5492e-16	9.8195e-11
30	8.478e-16	6.7969e-16
35	7.5571e-16	4.4422e-16
40	2.362e-11	1.6613e-05
60	1.7206e-15	2.9723e-04
80	4.8873e-13	4.5297e-04
100	4.4468e-05	2.3377e-03

It is seen that if the number of neurons in the hidden layer is greater than 35, the validation performance deteriorates sharply if a second hidden layer is added. The general consensus is that adding more hidden layers is overkill in the sense it does not improve performance. But our work has demonstrated that adding more hidden layers not only does not improve performance, it will cause the performance to decline sharply. The validation performance curves for different designs of the neural network are given below in Figure5.3-1 to Figure 5.3-24:

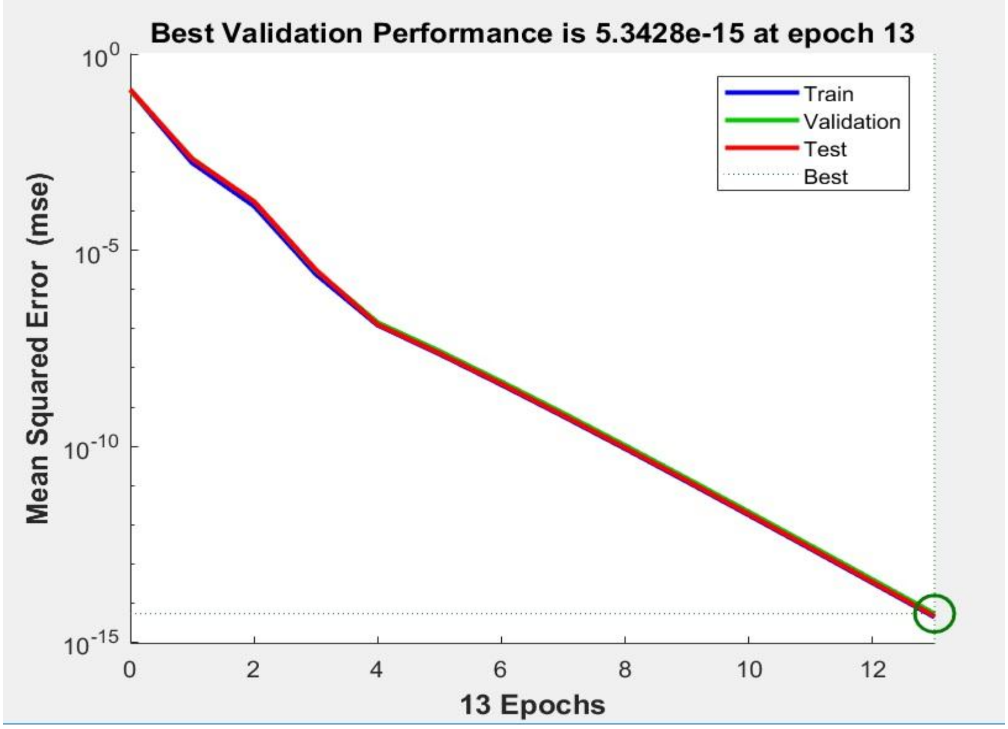

Figure 5.3-1: Validation Performance with One Hidden Layer Having Four Nodes

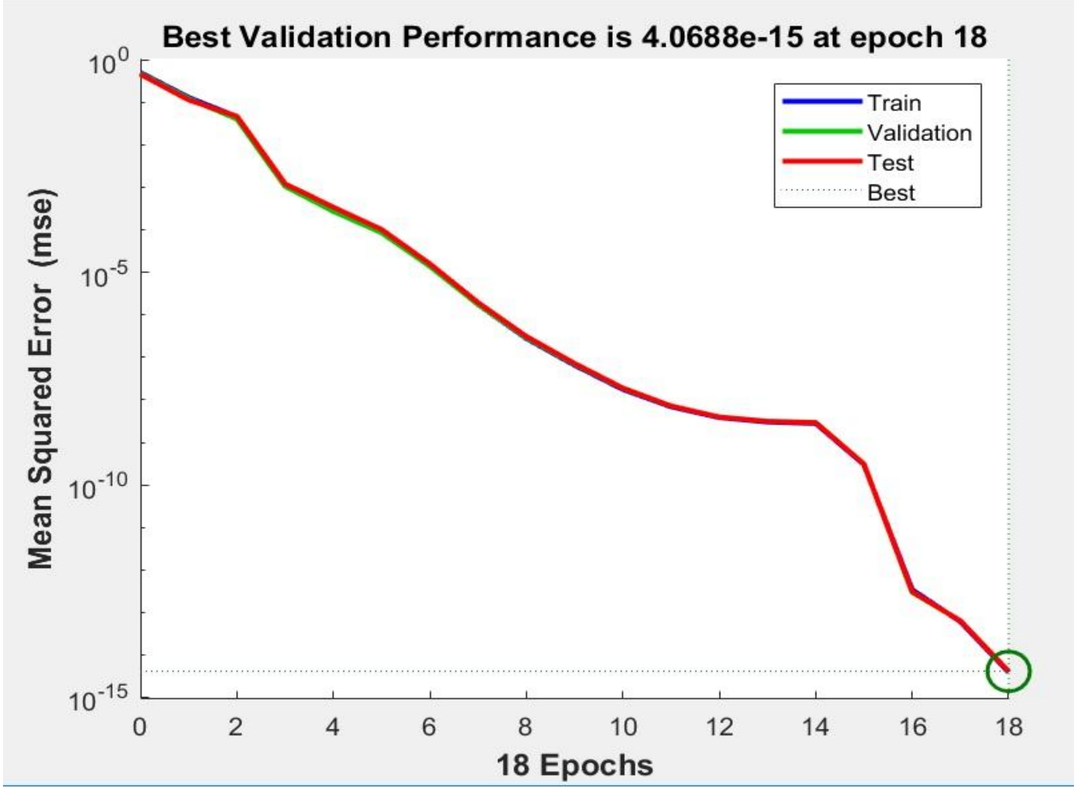

Figure 5.3-2: Validation Performance with Two Hidden Layers Having Four Nodes in Each Layer

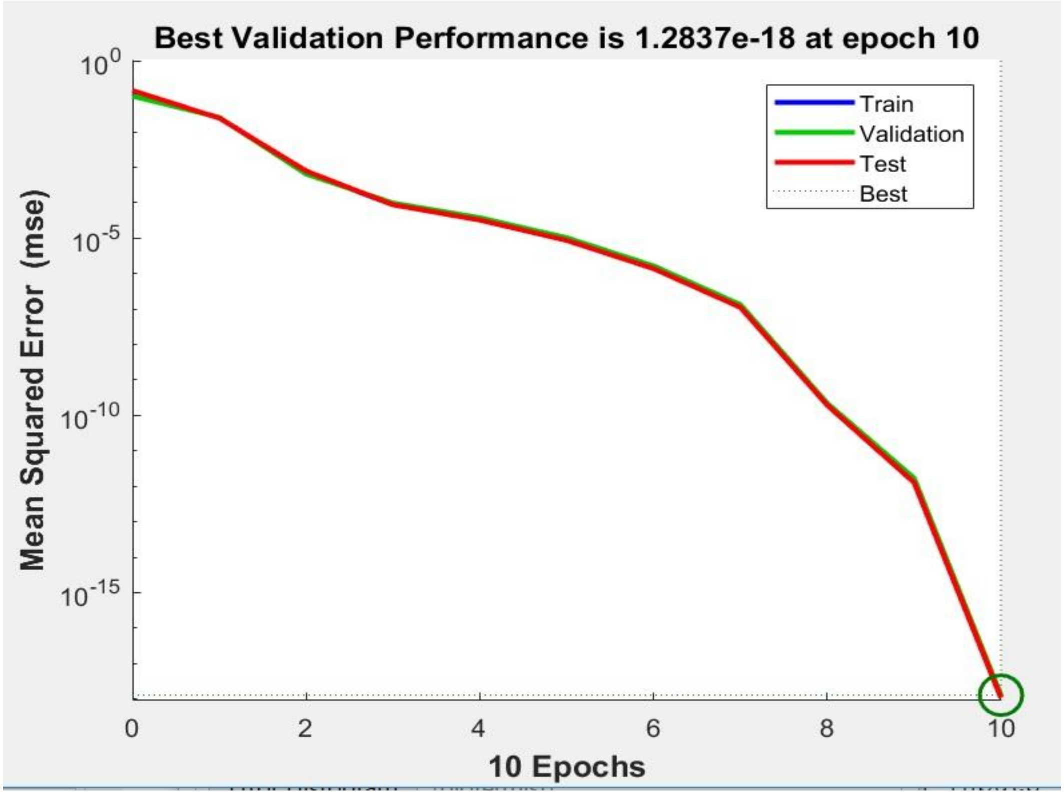

Figure 5.3-3: Validation Performance with One Hidden Layer Having Five Nodes

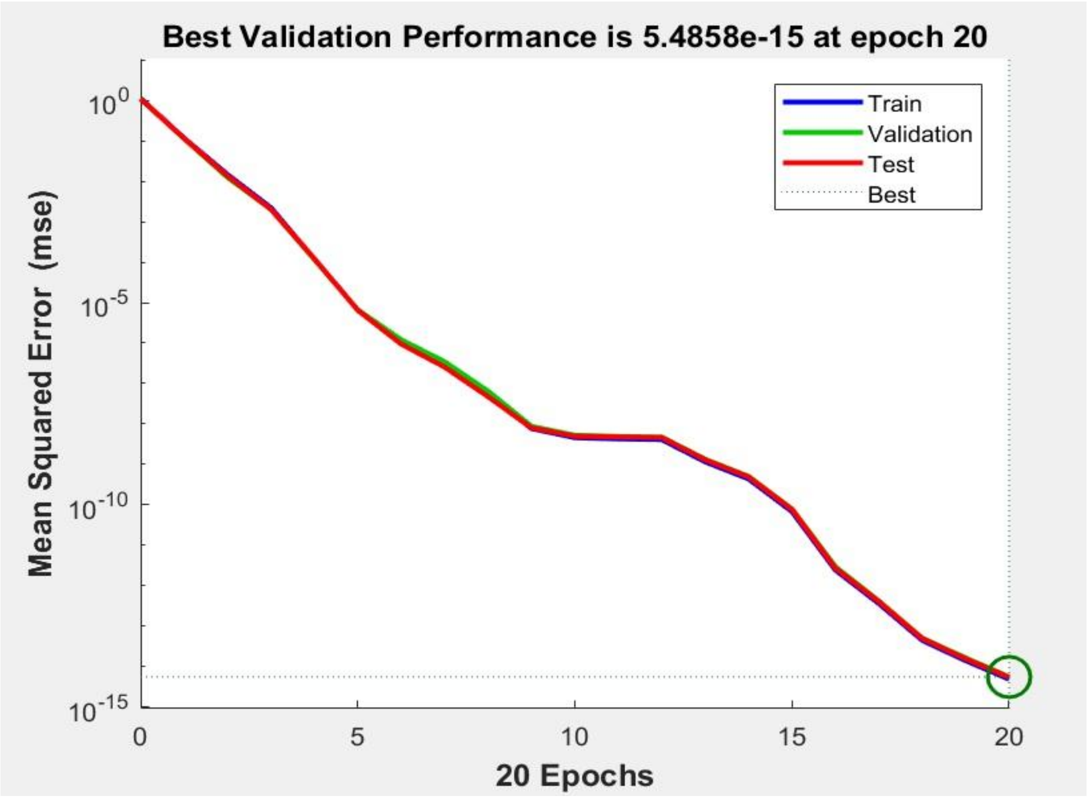

Figure 5.3-4: Validation Performance with Two Hidden Layers Having Five Nodes in Each Layer

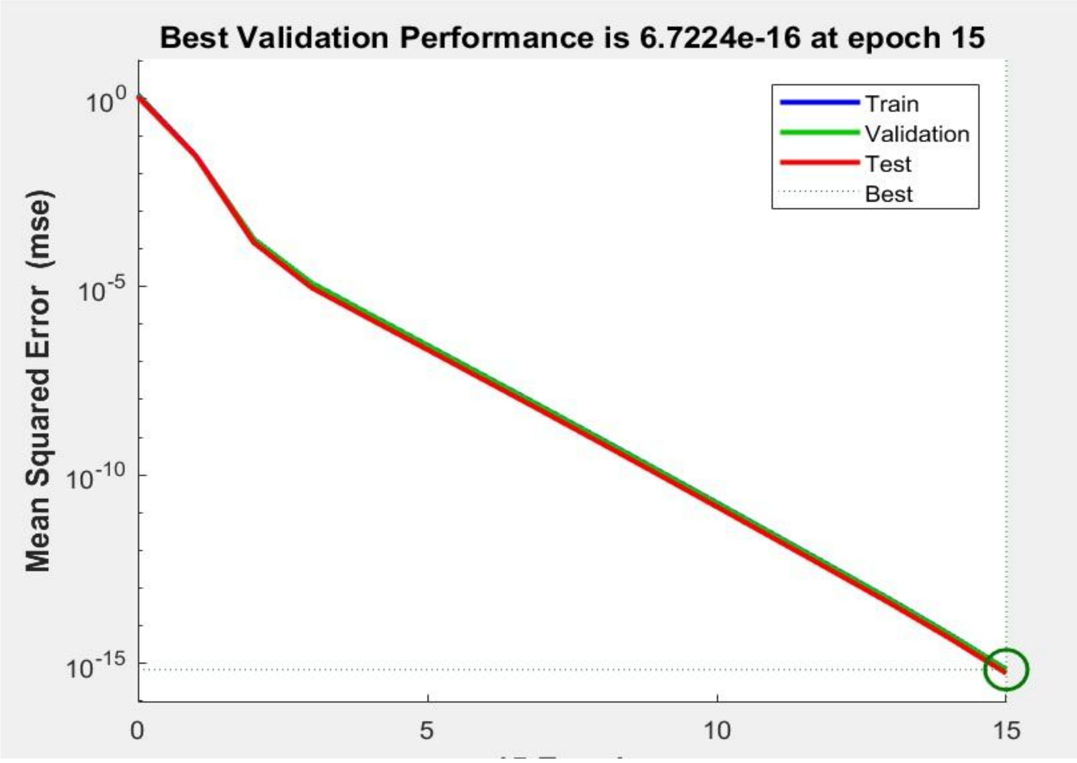

Figure 5.3-5: Validation Performance with One Hidden Layer Having Seven Nodes

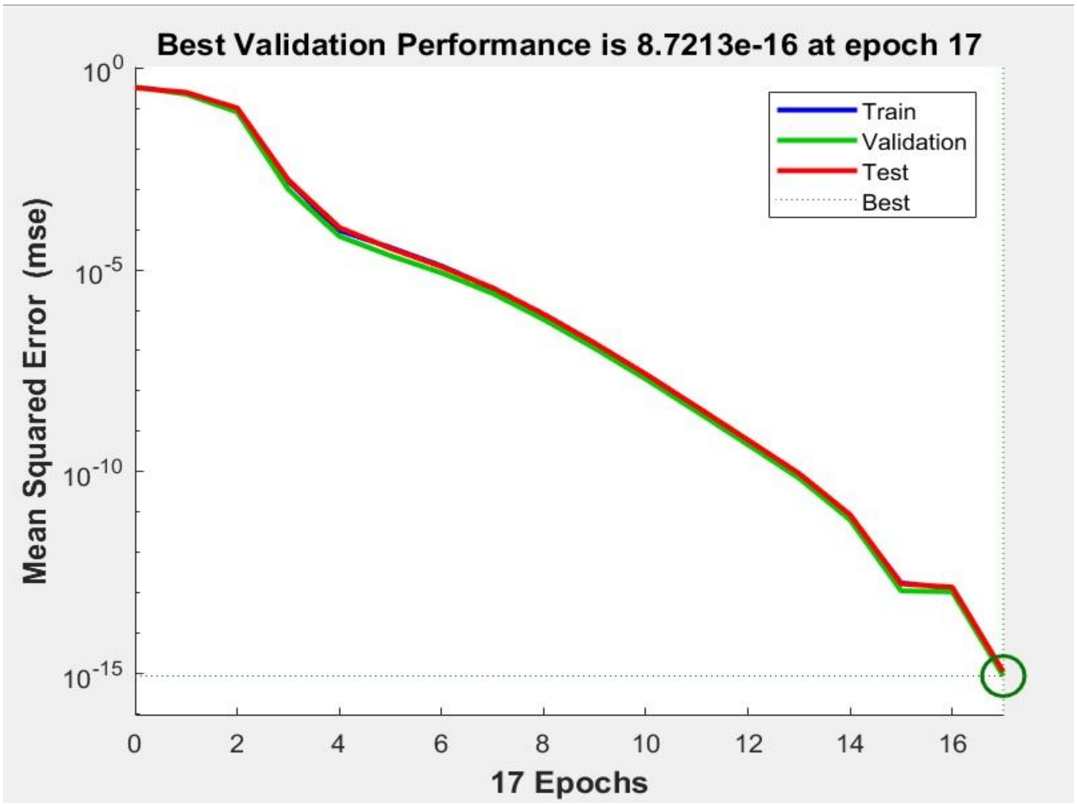

Figure 5.3-6: Validation Performance with Two Hidden Layers Having Seven Nodes in Each Layer

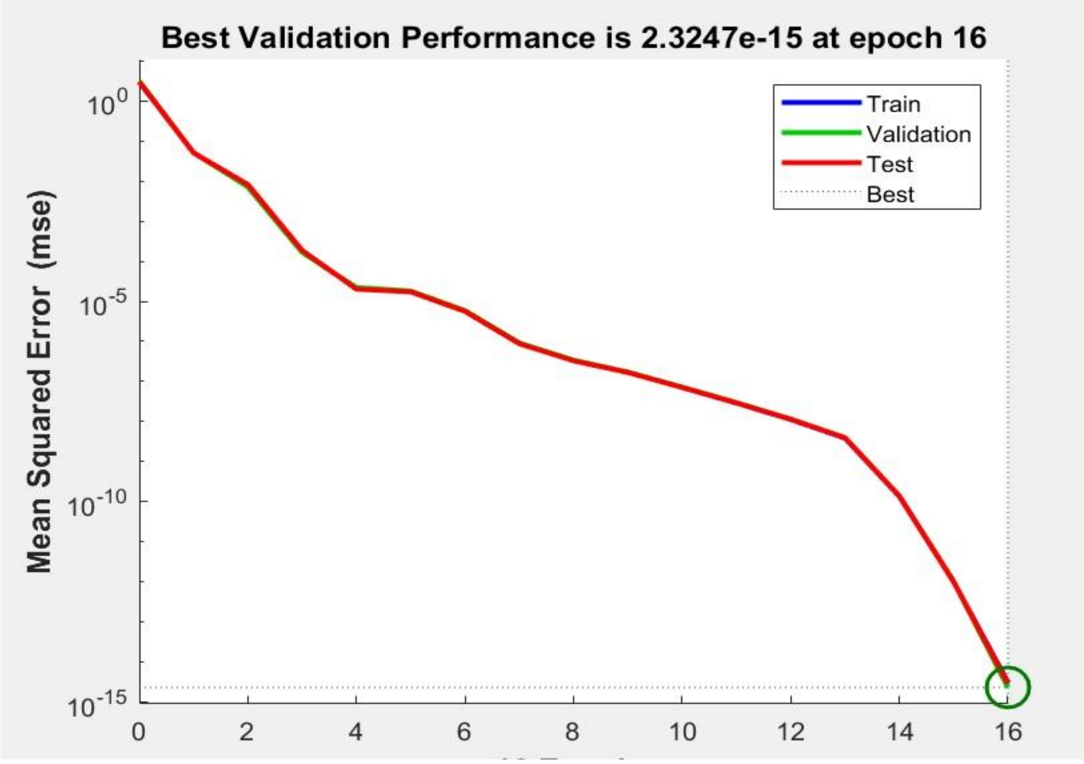

Figure 5.3-7: Validation Performance with One Hidden Layer Having Fifteen Nodes

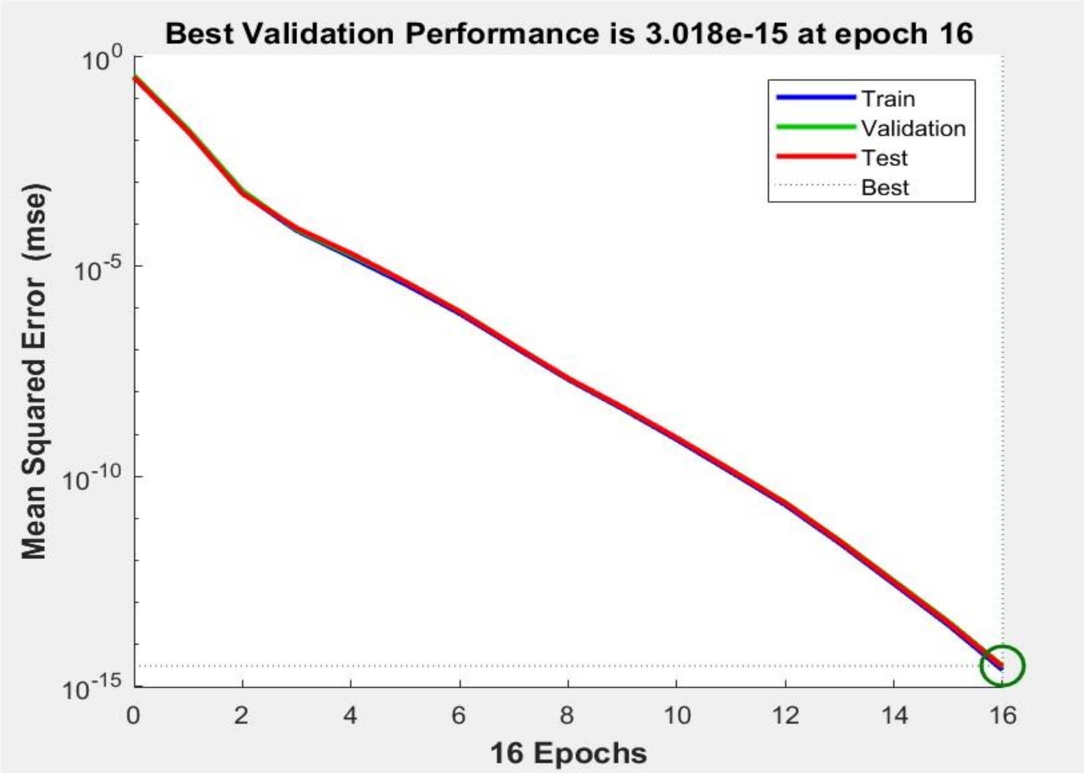

Figure 5.3-8: Validation Performance with Two Hidden Layers Each Having Fifteen Nodes

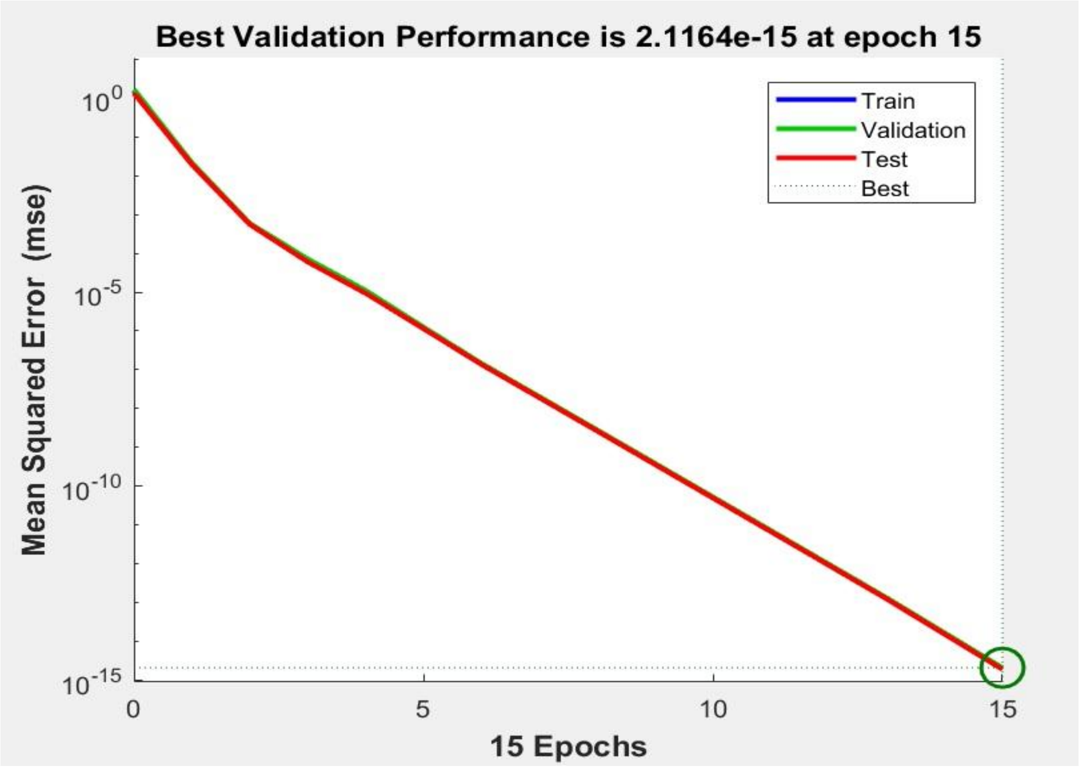

Figure 5.3-9: Validation Performance with One Hidden Layer Having Twenty Nodes

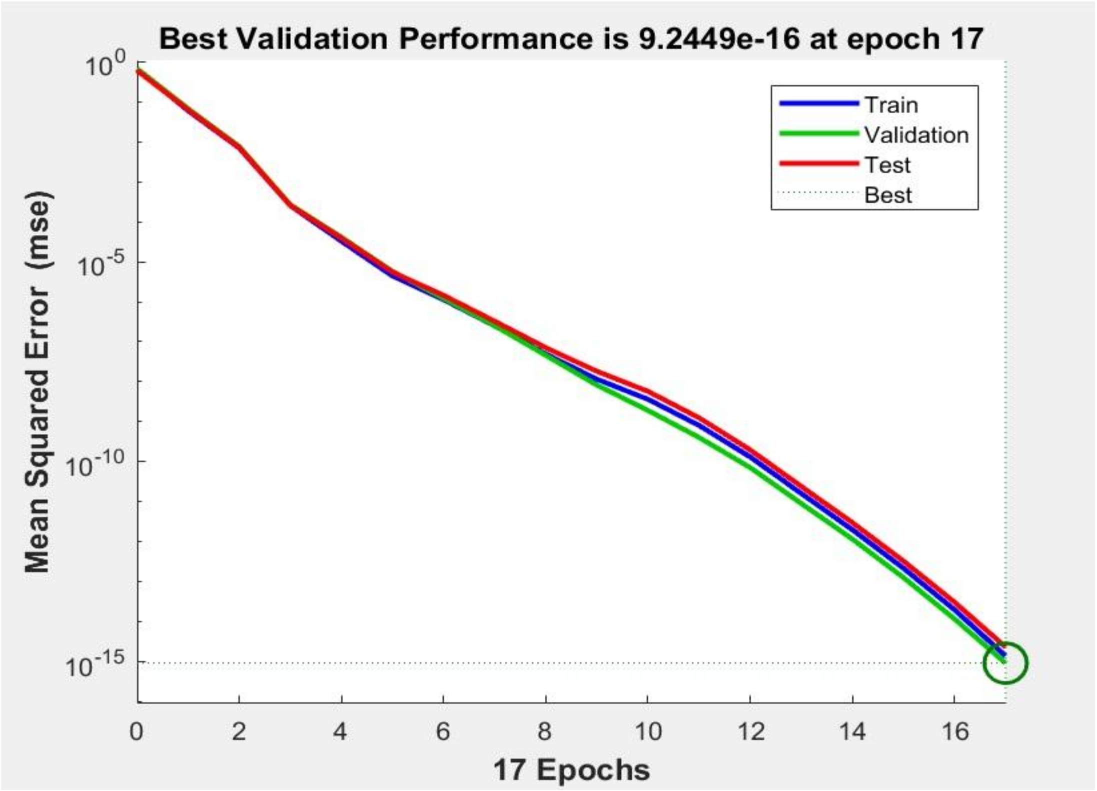

Figure 5.3-10: Validation Performance with Two Hidden Layers Each Having Twenty Nodes

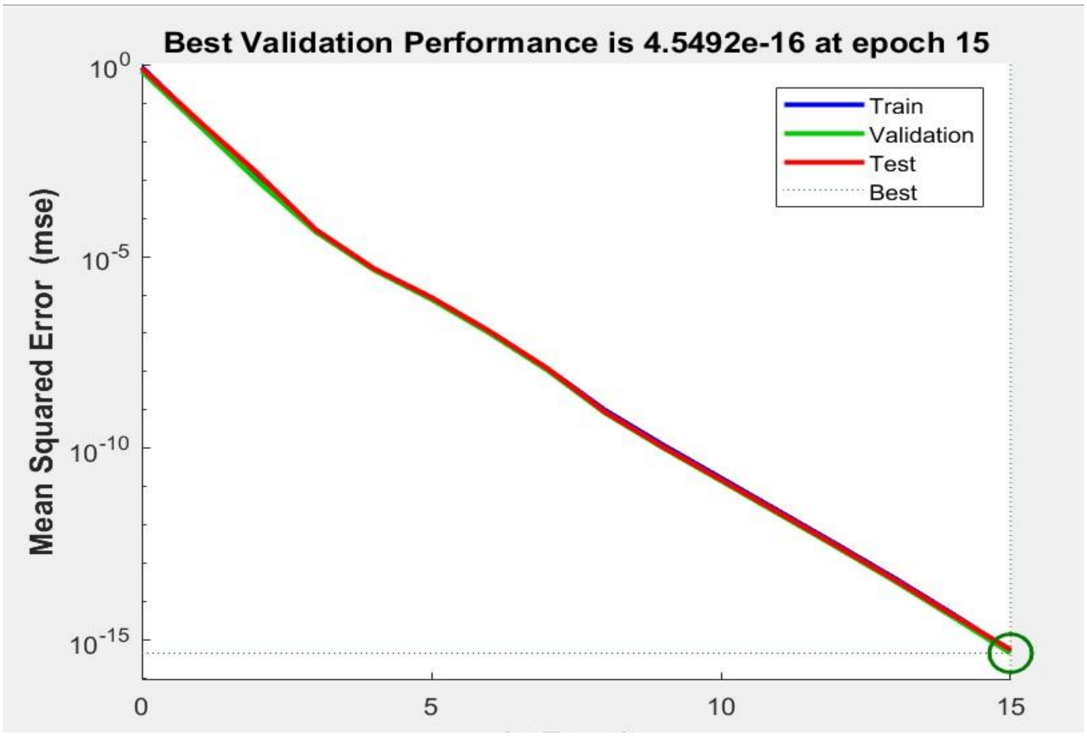

Figure 5.3-11: Validation Performance with One Hidden Layer Having Twenty Nine Nodes

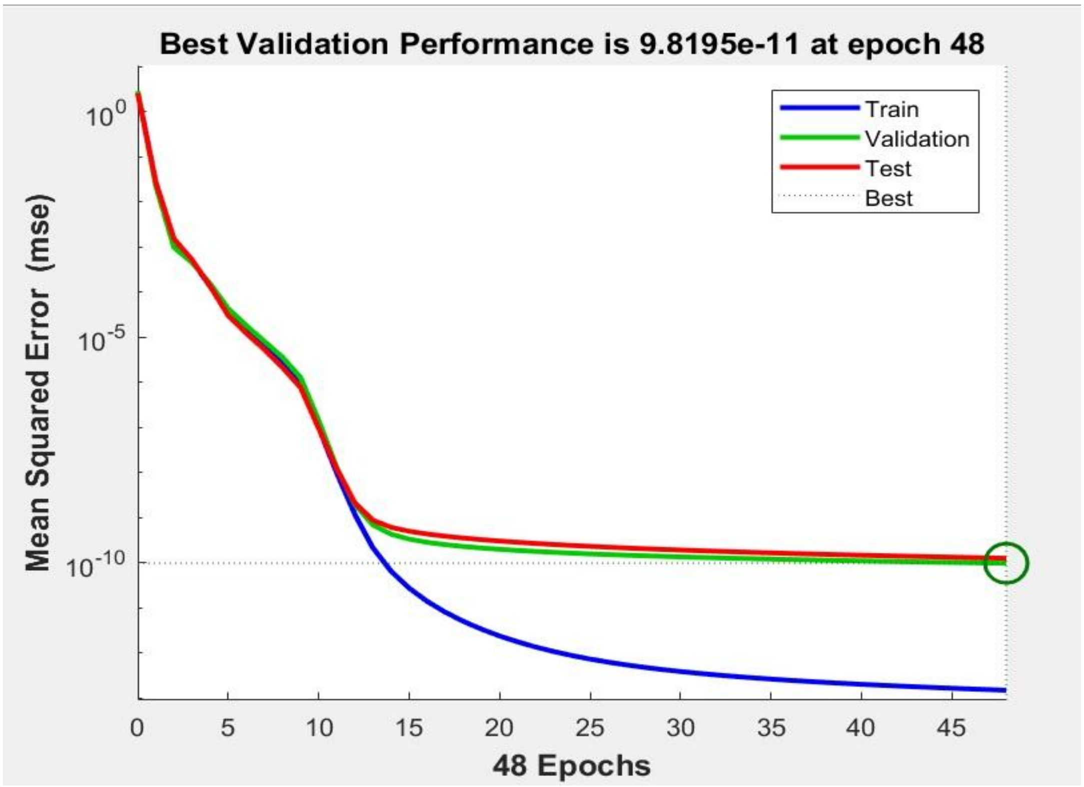

Figure 5.3-12: Validation Performance with Two Hidden Layers Each Having Twenty Nine Nodes

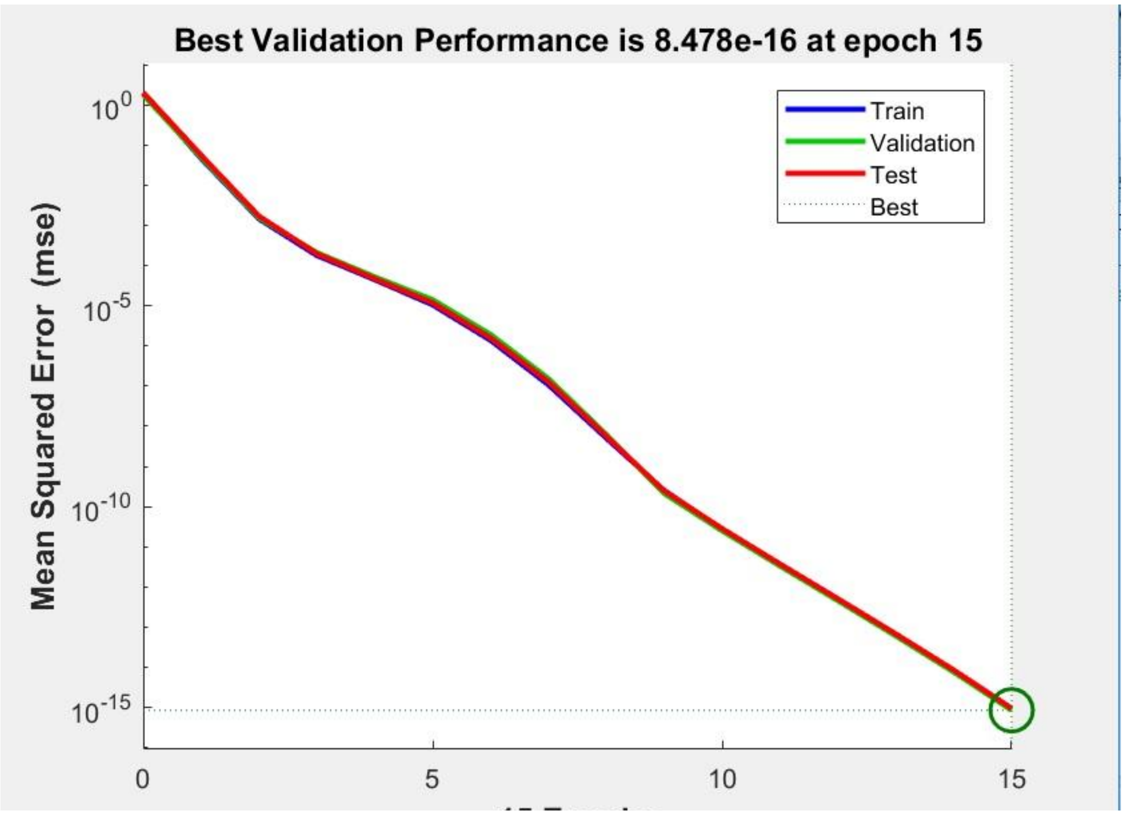

Figure 5.3-13: Validation Performance with One Hidden Layer Having Thirty Nodes

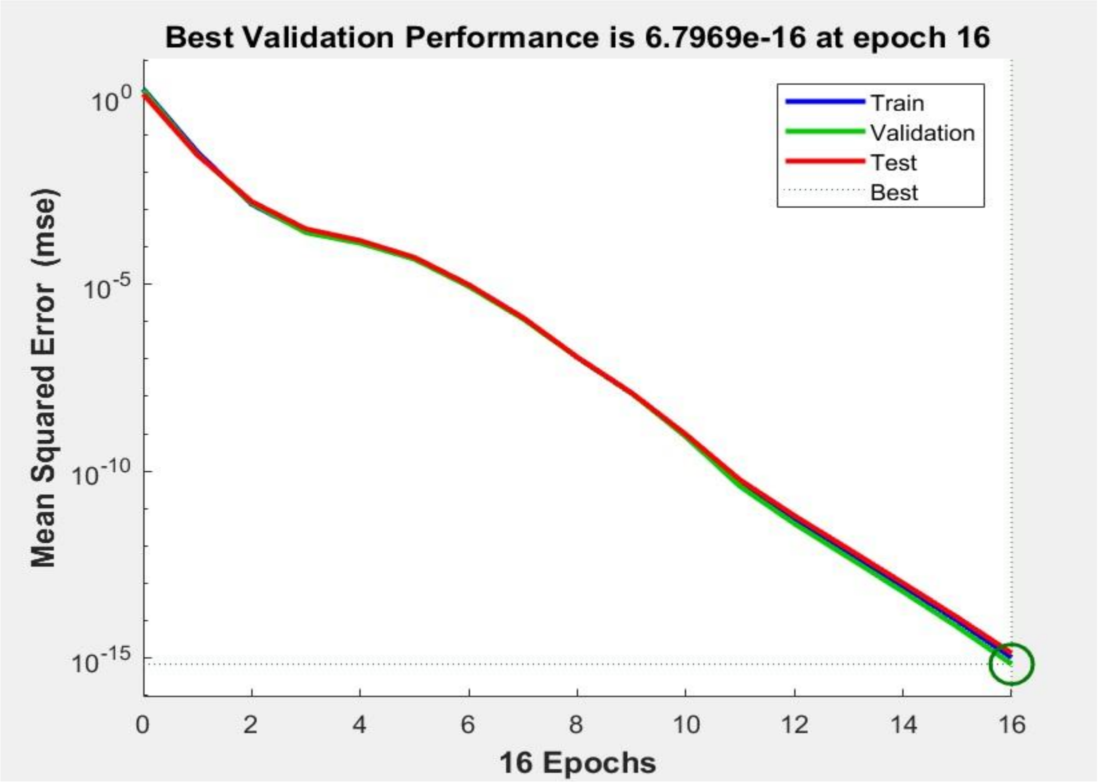

Figure 5.3-14: Validation Performance with Two Hidden Layers Each Having Thirty Nodes

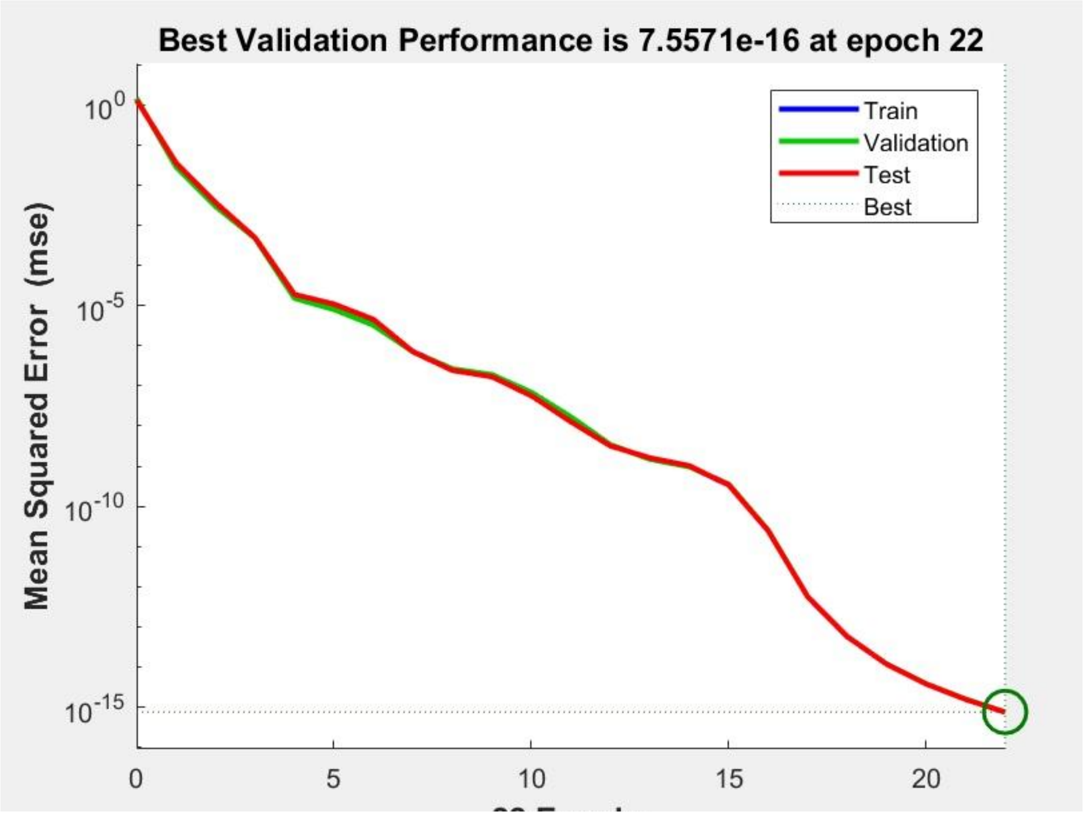

Figure 5.3-15: Validation Performance with One Hidden Layer Having Thirty Five Nodes

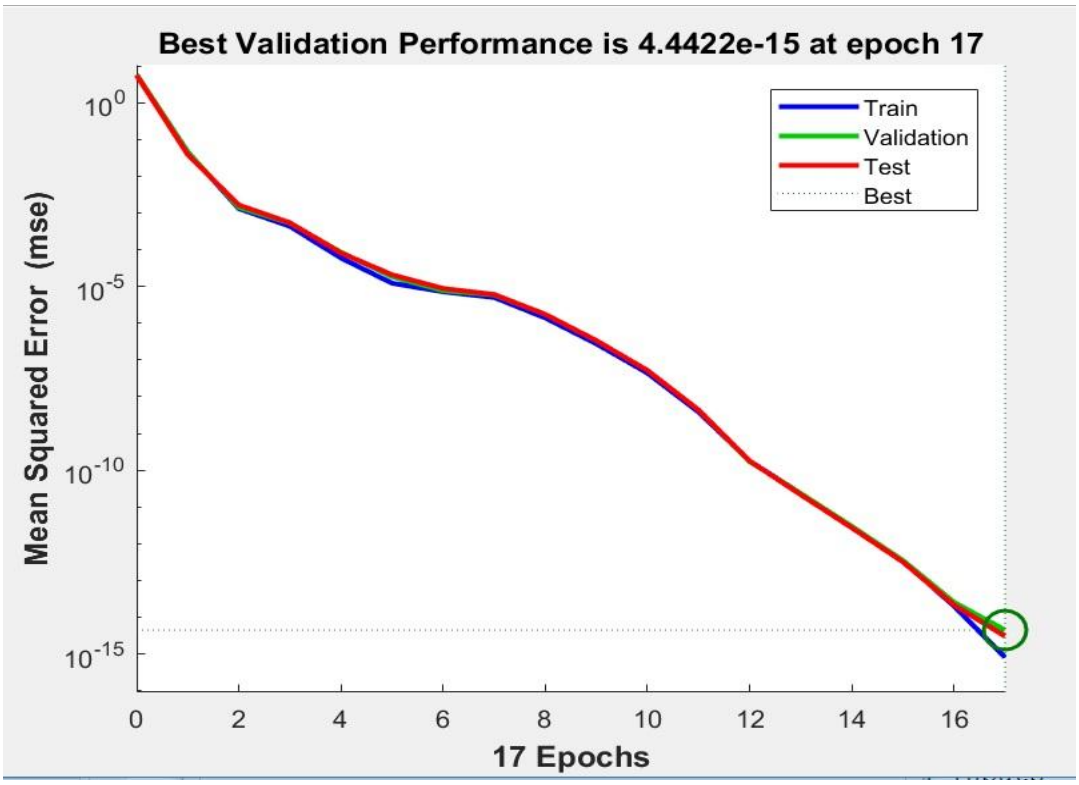

Figure 5.3-16: Validation Performance with Two Hidden Layers Each Having Thirty Five Nodes

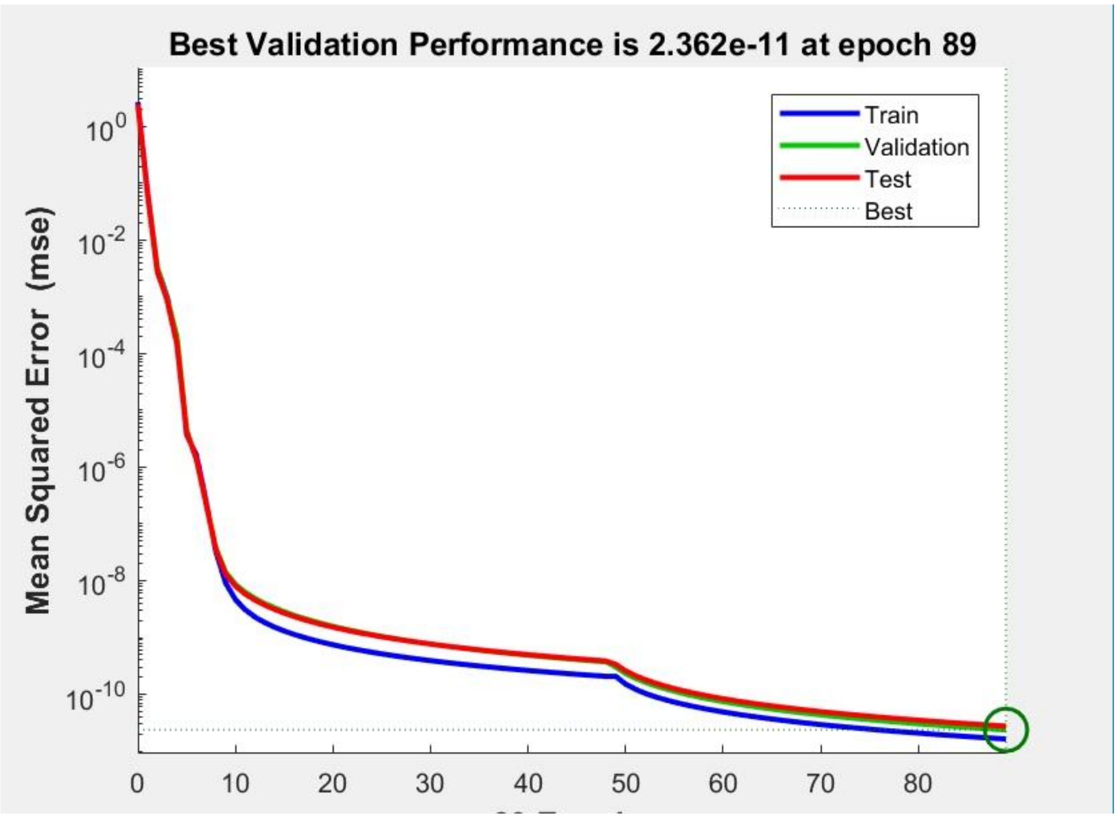

Figure 5.3-17: Validation Performance with One Hidden Layer Having Forty Nodes

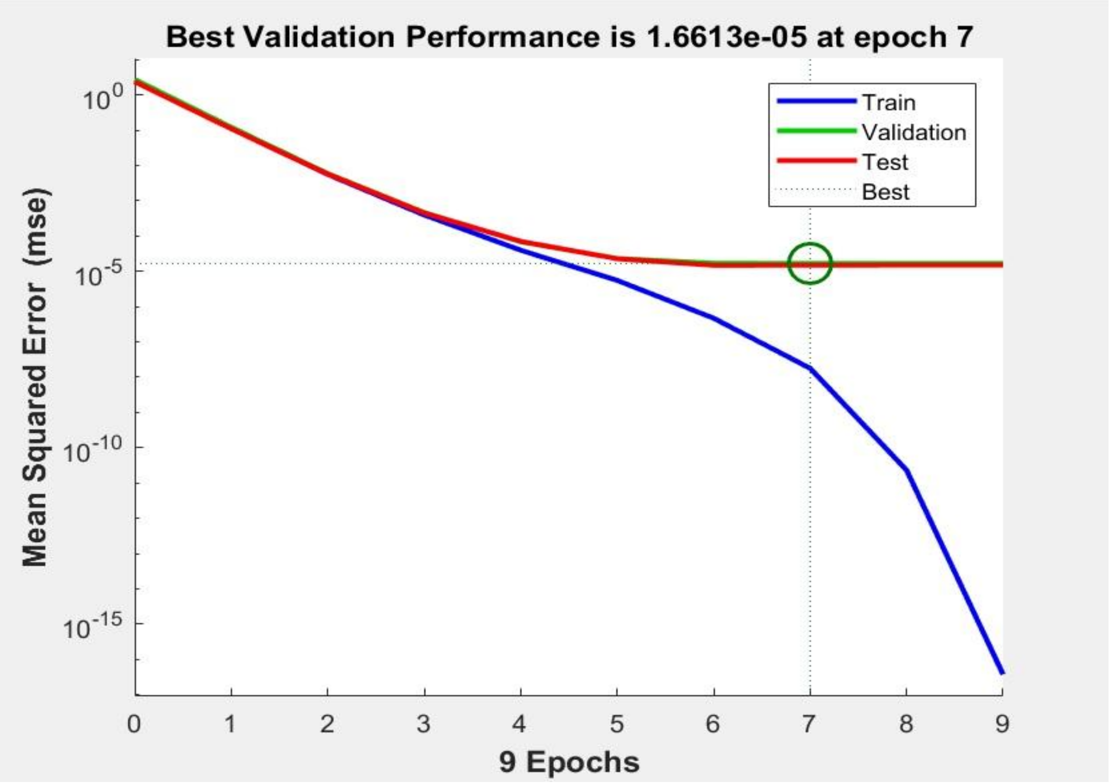

Figure 5.3-18: Validation Performance with Two Hidden Layers Each Having Forty Nodes

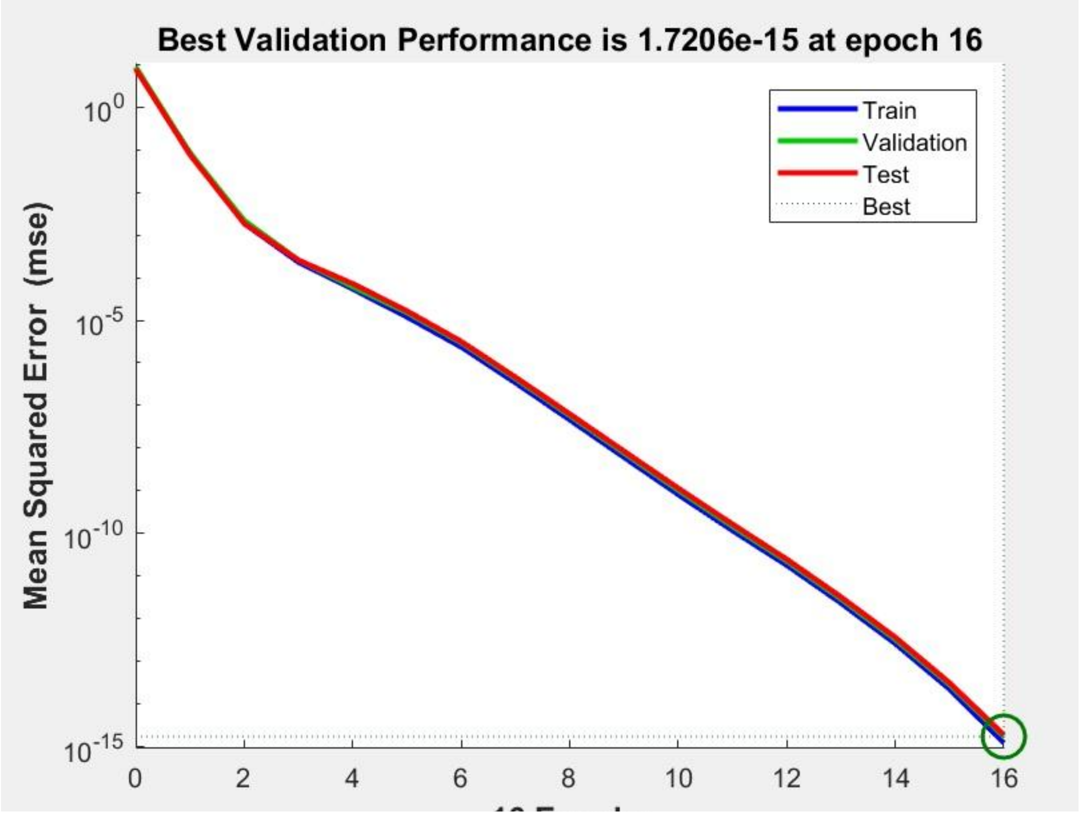

Figure 5.3-19: Validation Performance with One Hidden Layer Having Sixty Nodes

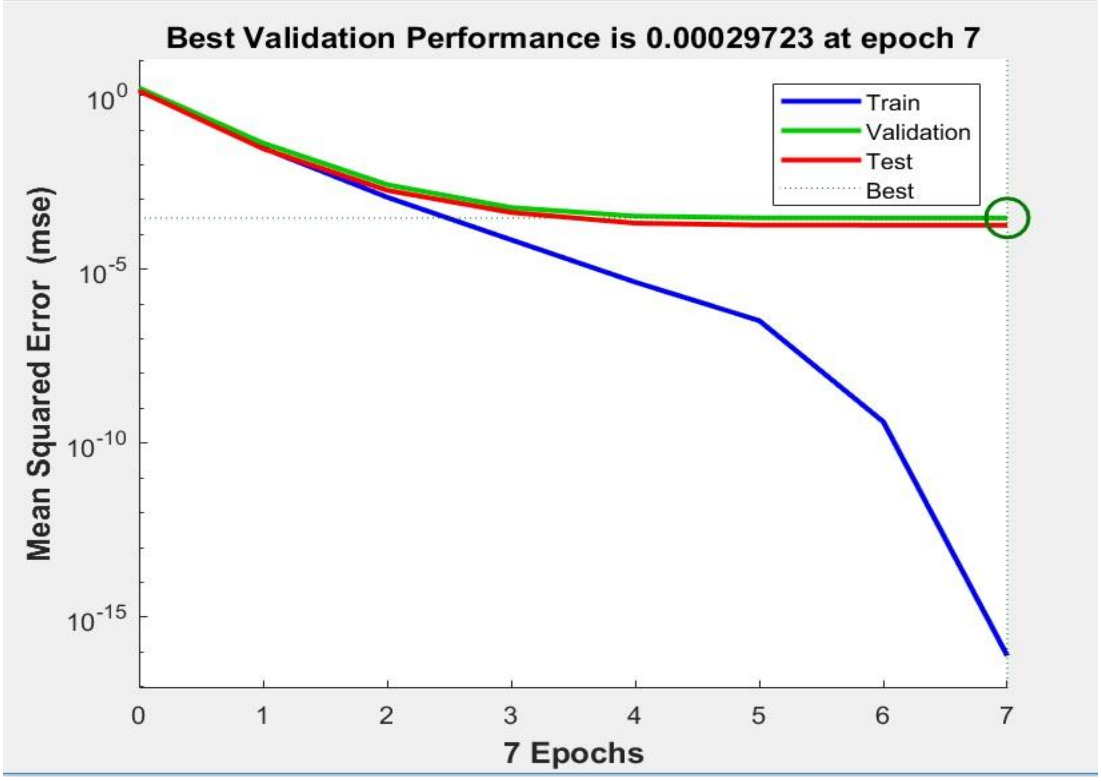

Figure 5.3-20: Validation Performance with Two Hidden Layers Each Having Sixty Nodes

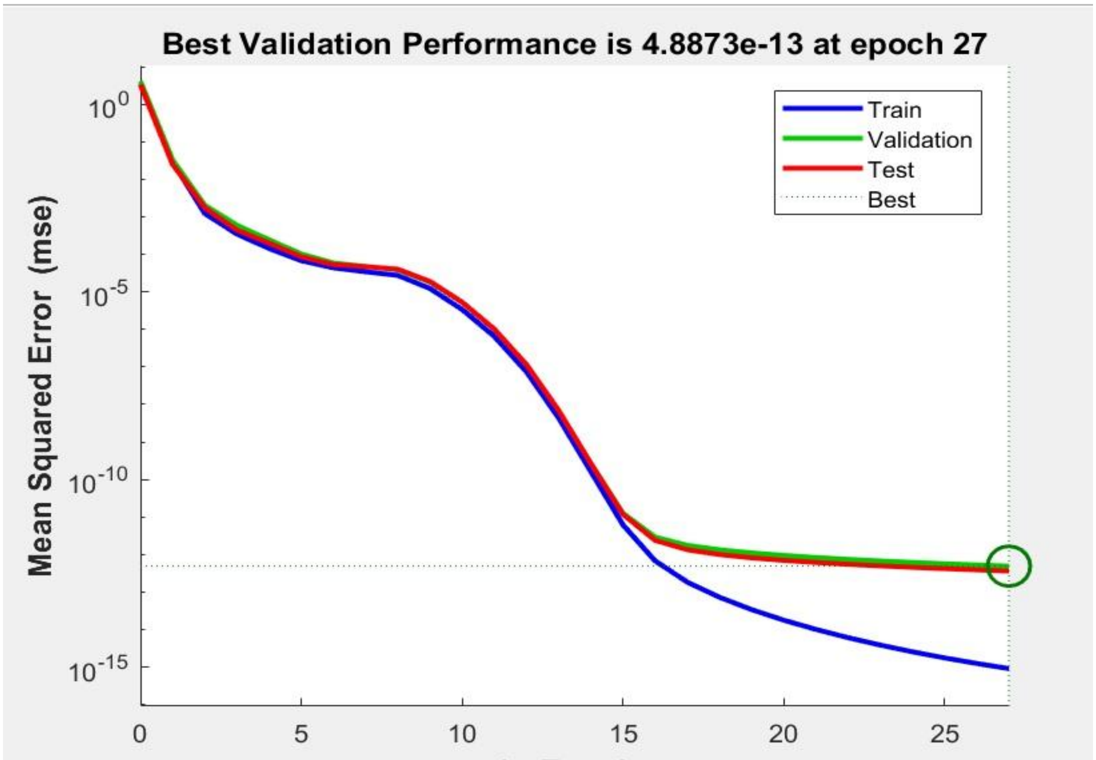

Figure 5.3-21: validation Performance with One Hidden Layer Having Eighty Nodes

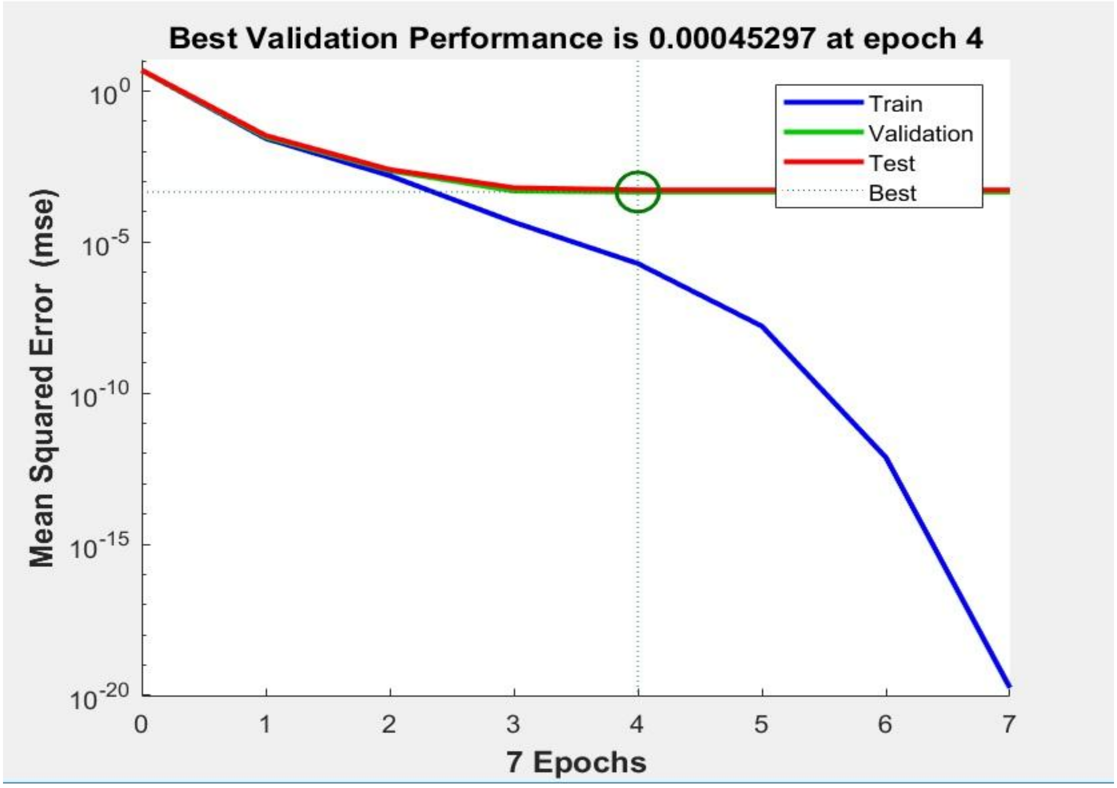

Figure 5.3-22: validation Performance with Two Hidden Layers Each Having Eighty Nodes

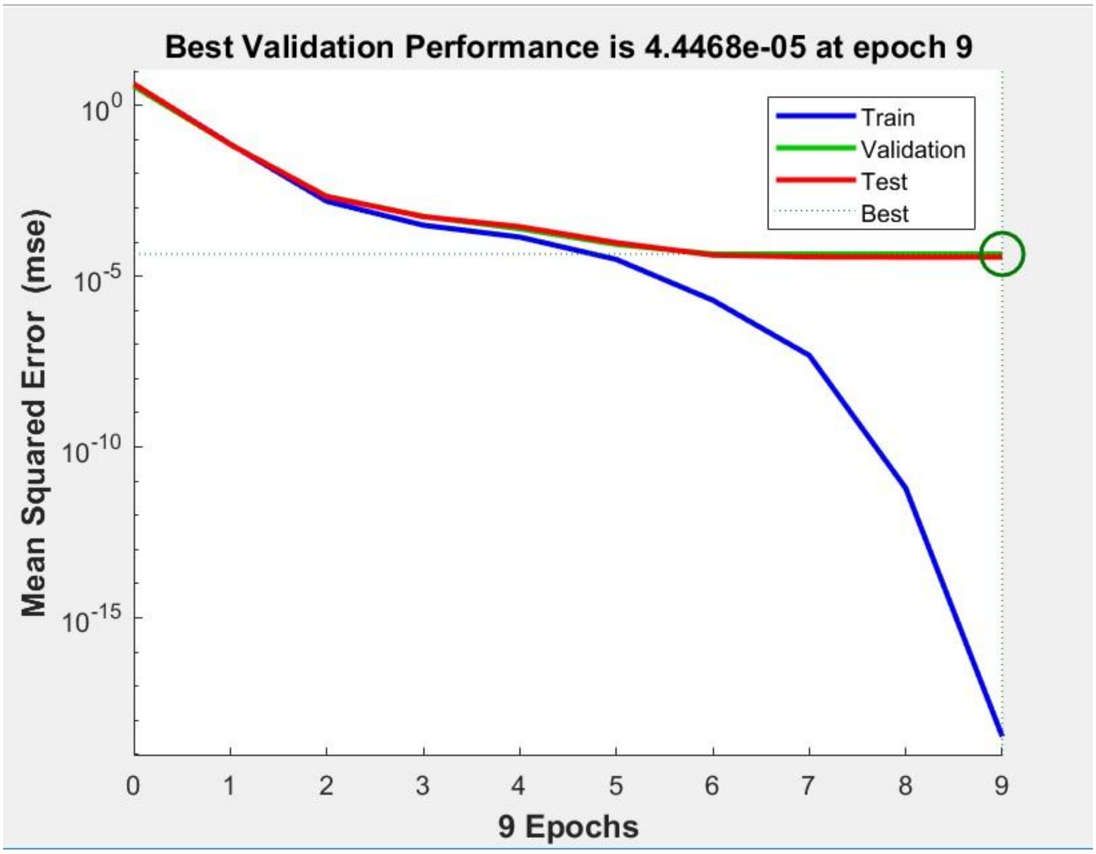

Figure 5.3-23: Validation Performance with One Hidden Layer Having Hundred Nodes

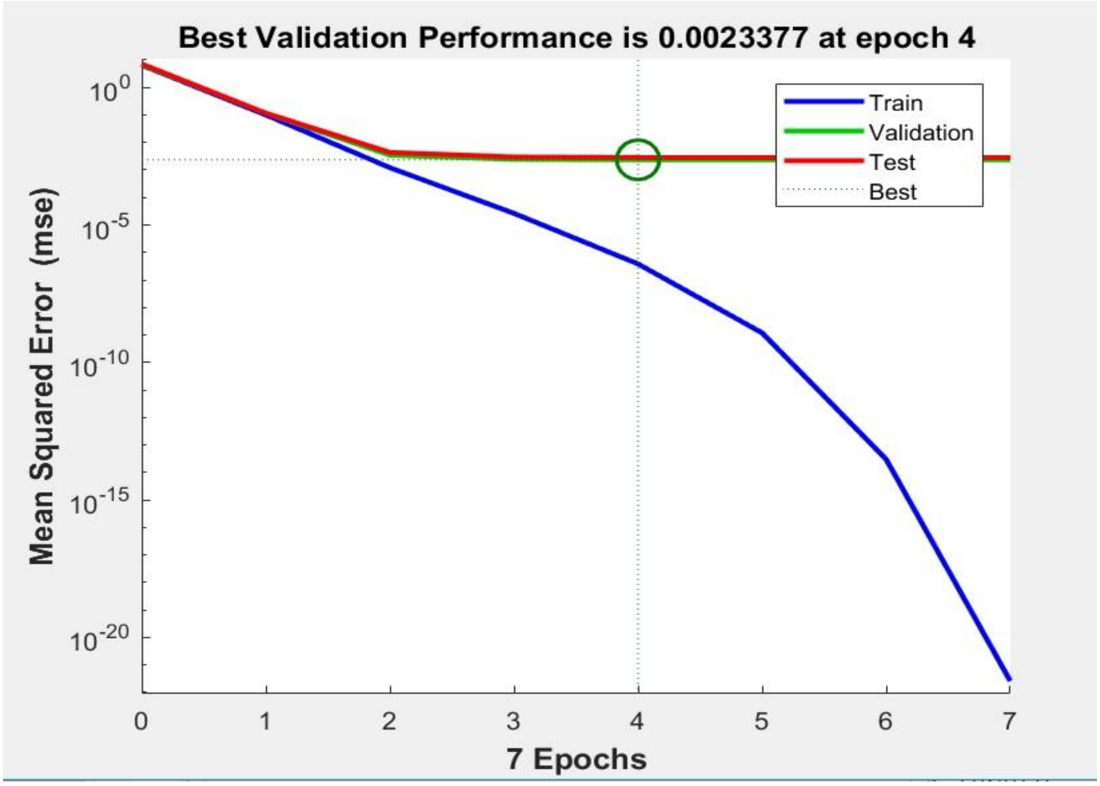

Figure 5.3-24: Validation Performance with Two Hidden Layers Each Having Hundred Nodes

5.4. Experiment4 – Choice of Support Vector Machine Kernel for Classifying Schizophrenia Data

The results obtained upon training the SVM with various different kernels are tabulated below:

Table 5.4-1: Support Vector Machine Classification Performance with Various Kernels

Type of Kernel	Number of Support Vectors	Classification Accuracy
Linear Kernel	19	100%
Gaussian Kernel	480	86.67%
RBF Kernel	480	86.67%
Polynomial kernel (order 1)	19	100%
Polynomial kernel (order 2)	20	100%
Polynomial kernel (order 3)	21	100%
Polynomial kernel (order 4)	1	33.33%
Polynomial kernel (order 5)	5	100%
Polynomial kernel (order 15)	0	100%
Polynomial kernel (order 25)	0	66.67%
Polynomial kernel (order 100)	0	66.67%

Since the number of support vectors cannot be zero, the results show that this data cannot be classified with a polynomial kernel of order higher than three. The Gaussian and RBF kernels have relatively low classification accuracy whereas the linear kernel has a high accuracy (100%) and relatively low number of support vectors. The linear kernel is also the simplest of all kernels. Therefore, taking all things into consideration, it is recommended that a linear kernel be used for classification of this data set.

5.5. Experiment5 – Fuzzy Clustering for Diagnosing Schizophrenia

The synthetic data generated as part of Experiment-2 was clustered using the MATLAB "fcm" utility. The tool was able to separate all the data correctly into two clusters. Further, the PANSS ratings of four actual subjects were added to the synthetic dataset and the "fcm" tool was executed again. The PANSS ratings of the four subjects are given in Table 5.5-1:

Table 5.5-1: PANSS Ratings of Actual Subjects

PANSS Items	Patient Ratings; Scale: [0 – 6]			
	Patient01	**Patient02**	**Patient03**	**Patient04**
Delusions	1	0	0	4
Conceptual Disorganization	1	0	0	2
Hallucinatory Behaviour	1	1	1	3
Excitement	2	0	1	2
Grandiosity	2	2	1	0
Suspiciousness / Persecution	1	0	0	5
Hostility	3	3	2	4
Blunted Affect	1	0	0	3
Emotional Withdrawal	3	0	2	5
Poor Rapport	3	2	2	2
Passive / Apathetic Social Withdrawal	2	0	1	3
Difficulty in Abstract Thinking	2	0	1	3
Lack of Spontaneity and Flow of Conversation	3	2	2	5
Stereotyped Thinking	1	0	0	5
Somatic Concern	3	0	2	2
Anxiety	4	0	2	3
Guilt Feelings	4	2	2	3
Tension	4	0	2	4
Mannerisms and Posturing	0	0	0	1

PANSS Items	Patient Ratings; Scale: [0 – 6]			
	Patient01	Patient02	Patient03	Patient04
Depression	5	2	4	5
Motor Retardation	2	0	1	3
Uncooperativeness	4	2	2	4
Unusual Thought Content	1	0	0	4
Disorientation	3	0	2	2
Poor Attention	4	1	2	2
Lack of Judgement and Insight	2	0	1	3
Disturbance of Volition	1	0	0	5
Poor Impulse Control	3	3	2	3
Preoccupation	3	0	2	4
Active Social Avoidance	3	2	2	2

According to a qualified psychiatrist, Patient01 and Patient 04 are schizophrenic, while Patient02 and Patient03 are not. The membership values of the four subjects in the two clusters are given below in Table 5.5-2:

Table 5.5-2: Membership Values in Fuzzy Clusters of Real Subjects

	Patient01	Patient02	Patient03	Patient04
Cluster1 (Schizophrenic)	0.9727	0.4073	0.8146	0.9827
Cluster2 (Non-schizophrenic)	0.0273	0.5927	0.1854	0.0173

We observe that Patient03 has been wrongly classified as schizophrenic by the clustering tool. Now, let us consider Patient03's medical history. He was given to substance abuse as a result of which he experienced hallucinations. He also suffered from depression, anxiety and guilt. An allergic reaction led to somatic concerns. Thus he exhibited some of the classic symptoms of schizophrenia even though he was not schizophrenic. As a result, he had such a high membership value in Cluster1. Even though Patient03 is not schizophrenic he has more in common with other schizophrenics than non-schizophrenics. This opens up the discussion as to whether an individual who exhibits Patient03’s symptoms without the use of hallucinogens should be diagnosed as schizophrenic even though such a diagnosis is at odds with the guidelines provided in the DSM as it stands today.

6. Conclusion

It is seen that artificial intelligence can go a long way in the diagnosis of a complex mental disease like schizophrenia. This work demonstrates how a fuzzy expert system compatible with the Diagnostic and Statistical Manual, 5th edition (DSM-5) can be built. The output of the expert system is such that a subject may be classified as normal, psychotic or schizophrenic. The expertise of a qualified psychiatrist was leveraged in order to identify the membership functions and fuzzy rules of the expert system.

Next, the fuzzy expert system was leveraged to create a synthetic dataset which may be used to train an artificial neural network for diagnosing the disease. Training neural networks with more than one hidden layers laid bare some interesting observations. The synthetic dataset was also used to train a support vector machine wherein the best SVM kernel for diagnosing schizophrenia was identified.

Finally, the synthetic dataset and some additional data from real subjects was clustered using fuzzy clustering. The result opened up a discussion on the diagnostic criterion for schizophrenia itself.

7. Scope of Future Work

The considerable volume of the work notwithstanding, most of the experiments were performed with synthetic training data and not data from actual subjects. In future, PANSS ratings of actual subjects may be collected and used to refine the membership functions with the help of neuro-fuzzy modelling. Additionally, the membership functions may be modified using genetic algorithms. The membership functions used with the fuzzy expert system captures just one physician's ideas and is subject to human errors. If a good amount and variety of data from human subjects are available, then a feedback loop can be created which would keep on modifying and refining the membership function till there is no significant change. This is something that can be taken up for future research.

8. References

[1] Benjamin J. Sadock, Virginia A. Sadock, and Pedro Ruiz, *Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry*.: Wolters Kluwer, 2014.

[2] American Psychiatric Association. (2013) www.displus.sk. [Online]. http://displus.sk/DSM/subory/dsm5.pdf

[3] Wikipedia. [Online]. https://en.wikipedia.org/wiki/Positive_and_Negative_Syndrome_Scale

[4] M. Obermeier, "Should the PANSS be Rescaled?," *Schizophrenia Bulletin, 36*, pp. 455-460, 2010.

[5] S.R. Kay, A. Fiszbein, and L.A. Opler, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophrenia Bulletin*, pp. 261-276, 1987.

[6] J.K. Wing and R.G.J Giddens, "Industrial Rehabilitation of Male Chronic Shcizophrenic Patients," *Lancet*, vol. 2, no. 7101, pp. 505-507, October 1959.

[7] B. Keinmuntz, "The Computer as Clinician," *American Psychology*, vol. 75, pp. 1269-1274, 1967.

[8] J.F. Heiser and R.E. Brooks, "Design Considerations for a Clinical Psychopharmacology Advisor," *Proceedings of 2nd IEEE Annual Symposium on Computer Application in Medical Care*, vol. 24, pp. 278-285, 1978.

[9] B. Mulsant and D. Servan-Schreiber, "Knowledge engineering: A daily activity on a hospital ward," *Computers and Biomedical Research*, vol. 17, no. 1, pp. 71-81, February 1984.

[10] S.K. Johri and S.K. Guha, "Set-covering diagnostic expert system for psychiatric disorders: The third world context," *Computational Methods and Programs in Biomedicine*, vol. 34, no. 1, pp. 1-7, January 1991.

[11] M. Petrovic, C. Hurt, D. Collins, and A. et al Burns, "Clustering of Behavioural and Psychological Symptoms in Dementia (BPSD): A European Alzheimer's Disease Consortium Study," in *Acta Clin Belg.*, 2007, pp. 426-432.

[12] S. Chattopadhyay, D. Pratihar, and S.C. De Sarkar, "Fuzzy Logic Based Screening and Prediction of Adult Psychoses: A Novel Approach," *IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans*, vol. 39, no. 2, pp. 381-387, March 2009.

[13] Y. Zou, Y. Shen, and L. Shu, "Artificial Neural Network to Assist Psychiatric Diagnosis," *British Journal of Psychiatry*, vol. 169, pp. 64-67, 1996.

[14] P. Aruna, N. Puviarasan, and B. Palaniappan, "An investigation of neurofuzzy systems in psychosomatic disorders," *Expert Systems Applications*, vol. 28, no. 4, pp. 673-679, May 2005.

[15] Y. Li and F. Fan, "Classification of Schizophrenia and Depression by EEG with ANNs," *Proc. 27th Annu. Conf. IEEE Eng. Med. Biol.*, pp. 2679-2682, 2005.

[16] J. Ford and et al., "Patient Classification of FMRI Activation Maps," *Proc. of the 6th Annual International Conference on Medical Image computing and Computer Assisted Intervention*, pp. 58-65.

[17] D.D. Cox and R.L Savoy, "Functional Magnetic Resonance Imaging (fMRI) "brain reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex," *Neuroimage*, vol. 19, no. 2, pp. 261-270.

[18] S.V. et al Shinkareva, "Classificiation of Functional Brain Images with a Spatio-Temporal Dissimilarity Map," *Neuroimage*, vol. 33, no. 1, pp. 63-71.

[19] E. et al. Castro, "Characterization of groups using multiple kernels and multi-source fMRI analysis data: Application to schizophrenia," *Neuroimage*, vol. 58, pp. 526-536, 2011.

[20] J Mourao-Miranda and et al, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data," *Neuroimage*, vol. 28, no. 4, pp. 980-995, 2005.

[21] J. Mourao-Miranda and et al, "The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fmri data," *Neuroimage*, vol. 33, no. 4, pp. 1055-1065, 2006.

[22] J.D. Haynes and G. Rees, "Predicting the orientation of invisible stimuli from activity in human primary visual cortex," *Nature Neuroscience*, vol. 8, no. 5, pp. 686-691, 2005.

[23] F. De Martino and et al, "Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns," *Neuroimage*, vol. 43, no. 1, pp. 44-58, 2008.

[24] S. Ryali and et al, "Sparse logistic regression for whole brain classification of fmri data," *Neuroimage*, 2010.

[25] E. Castro and et al., "Characterization of groups using composite kernels and multi-source fMRI analysis data: Application to schizophrenia," *Neuroimage*, vol. 58, pp. 526-536, 2011.

[26] K. Kayaer and T. Yildrim, "Medical diagnosis on Pima Indian diabetes using general

regression neural networks," in *Proceedings of the international conference on artificial neural networks and neural information processing*, pp. 181-184.

[27] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113-127, 2005.

[28] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Systems with Applications*, vol. 36, pp. 944-949, 2009.

[29] R.P. Brent, "Fast training algorithms for multi-layer neural nets," *IEEE Transactions on Neural Networks*, vol. 2, pp. 346-354, 1991.

[30] M. Gori and A. Tesi, "On the problem of local minima in back-propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 76-85, 1992.

[31] A. Gulbag and F. Temurtas, "A study on quantitative classification of binary gas mixture using neural networks and adaptive neurofuzzy inference systems ," *Sensors and Actuators B*, vol. 115, pp. 252-262.

[32] M.T. Hagan, H.B. Demuth, and M.H. Beale,.: PWS Publishing, 1996.

[33] M.T. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, pp. 989-993.

[34] D.F Specht, "Probabilistic Neural Networks," *Neural Networks*, vol. 3, no. 1, pp. 109-118, 1990.

[35] D.A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Network*, vol. 11, pp. 323-336, 1998.

[36] D. Deng and N. Kasabov, "On-line Pattern Analysis by Evolving Self-Organizing Maps," *Proceedings of the fifth biannual conference on artificial neural networks and expert systems*, pp. 46-51, 2001.

[37] K. Polat and S. Gunes, "An expert systems approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, vol. 17, no. 4, pp. 702-710, 2007.

[38] C.F. Aliferis, D. Hardin, and Massion P.P., "Machine learning models for lung cancer classification using array comparative genomic hybridization," in *Proceedings of the AMIA symposium on biomedical informatics*, Nashville, 2002, pp. 7-11.

[39] Ashizawa K. et al., "Artificial neural networks in chest radiography: Application to the differential diagnosis of interstitial lung disease," *Academic Radiology*, vol. 11, no. 1,

pp. 29-37, 2005.

[40] G. Coppini, M. Miniati, M. Paterni, S. Monti, and E.M. Ferdeghini, "Computer-aided diagnosis of emphysema in COPD patients: Neural network based analysis of lung shape in digital chest radiographs," *Medical Engineering and Physics*, vol. 29, pp. 76-78.

[41] A.A. El-Solh, C.B. Hsiao, S. Goodnough, J. Serghani, and B.J.B. Grant, "Predicting active pulmonary tuberculosis using an artificial neural network," *Chest*, vol. 116, pp. 76-78, 1999.

[42] O. Er, C. Sertkaya, F. Temurtas, and A.C. Tanrikulu, "A comparative study on chronic obstructive pulmonary and pneumonia diseases diagnosis using neural networks and artificial immune system," *Journal of Medical Systems*, vol. 33, no. 6, pp. 485-492, 2009.

[43] O. Er and F Temurtas, "A study on chronic obstructive pulmonary disease diagnosis using multi-layer neural networks.," *Journal of Medical Systems*, vol. 32, no. 5, pp. 429-432, 2008.

[44] O. Er, F. Temurtas, and A.C. Tanrikulu, "Tuberculosis disease diagnosis using artificial neural networks," *Journal of Medical Systems*, vol. 34, no. 3, pp. 299-302, 2010.

[45] N.H.H.M. Hanif, W.H. Lan, H.B. Daud, and J. Ahmad, "Classification of control measures of asthma using artificial neural networks," in *Malaysia: Scientific and technical Publishing Company*.: ACTA Press, 2009.

[46] S.H. Paul, S.G. Ben, G.T. Thomas, and S.W. Robert, "Use of genetic algorithm for neural network to predict community acquired pneumonia," *Artificial Intelligence in Medicine*, vol. 30, pp. 71-84, 2004.

[47] A.M. dos Santos, B.B. Pereira, and J.M. de Seixas, "Neural networks: An application for predicting smear negative pulmonary tuberculosis," *Proceedings for Statistics in Health Sciences*.

[48] P.S. Heckerling, "Parametric receiver operating characteristic (ROC) curve analysis using Mathematica," *Computer Methods and Programs in Biomedicine*, vol. 69, pp. 65-73, 2002.

[49] I. Seritas, N. Allahverdi, and I.U. Sert, "A fuzzy expert system design for diagnosis of prostate cancer," in *International Conference on Computer Systems and Technologies*.

[50] S. S. Sikchi, S. Sikchi, and M.S. Ali, "Design of fuzzy expert system for diagnosis of cardiac diseases," *International Journal of Medical Science and Public Health*, vol. 2, no. 1, pp. 56-61, 2013.

[51] M. Rana and R.R. Sedamkar, "Design of Expert System for Medical Diagnosis Using Fuzzy Logic," *International Journal of Scientific and Engineering Research*, vol. 4, no. 6, pp. 2914-2921, June 2013.

[52] N. Sahai, D. Shrivastava, and P. Shrivastava, "Diagnosis of the Jaundice Using Fuzzy Expert Systems," *International Journal of Biomedical Science and Engineering*, vol. 1, no. 3, pp. 15-19, October 2014.

[53] Y. Niranajan Devi and S. Anto, "An Evolutionary-Fuzzy Expert System for the Diagnosis of Coronary Artery Disease," *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 3, no. 4, pp. 1478-1484, April 2014.

[54] Timothy J. Ross, *Fuzzy Logic With Engineering Applications, 3rd ed..*: Wiley, 2010.

[55] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Marcia J. Horton, Ed. New Jersey, United States of America: Prentice Pearson Hall, 2009.

[56] S.A. Biyouki, I.B. Turksen, and M.H.F. Zarandi, "Fuzzy rule-based expert system for diagnosis of thyroid disease," in *Computational Intelligence in Bioinformatics and Computational Biology*, Ontario, 2015.

[57] M.R. Arbabshirani and et al., "Classification of schizophrenia patients based on resting-state functional network conncectivity," *Frontiers in Neuroscience*, 2013.

[58] M. Takahashi, "Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures," *Schizophrenia Research*, pp. 210-218, 2010.

[59] A. Campana, "An artificial neural network that uses eye-tracking performance to identify patients with schizophrenia," *Schizophrenia Bulletin*, pp. 789-799, 1999.

[60] A. Ulloa, "Synthetic structural magnetic resonance image generator improves deep learning prediction of schzophrenia," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, 2015.

[61] E. Castro, "Generation of synthetic structural magnetic resonance images for deep learning pre-training," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, New York, 2015.

[62] J.Y. Li, T.W.S. Chow, and Y.L. Yu, "The estimation theory and optimization algorithm for the number of hidden units in the higher-order feedforward neural network," in *Proceedings, IEEE Internation Conference on Neural Networks, 1995*, Perth, 1995.

[63] S. Tamura and M. Tateishi, "Capabilities of a Four-Layered Feedforward Neural Network: Four Layers Versus Three," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 251-255, March 1997.

[64] S. Xu and L. Chen, "A Novel Approach for Determining the Optimal Number of Hidden Layer Neurons for FNN's and Its Application in Data Mining," in *5th International Conference on Information Technology and Applications*, 2008, pp. 683-686.

[65] K. Shibata and Y. Ikeda, "Effect of Number of Hidden Neurons on Learning in Large-Scale Layered Neural Networks," in *ICROS-SICE International Joint Conference*, Fukuoka, 2009, pp. 5008-5013.

[66] K.G. Sheela and S.N. Deepa, "Review on Methods to Fix Number of Hidden Neurons in Neural Netwroks," *Mathematical Problems in Engineering*, vol. 2013, May 2013.

[67] S. Karsoliya, "Approximating Number of Hidden Layer Neurons in Multiple Hidden Layer BPNN Architecture," *International Journal of Engineering Trends and Technology*, vol. 3, no. 6, pp. 714-717, 2012.

[68] F. Panchal and M. Panchal, "Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Networks," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 11, pp. 455-464, November 2014.

[69] T. Vujicic, T. Matijevic, J. Ljucovic, A. Balota, and Z Sevarac, "Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network," , Varazdin, Central European Conference on Information and intelligent Systems, pp. 219-223.

[70] N.J. Cotton and B.M Wilamowski, "Compensation of Nonlinearities Using Neural Networks Implemented on Inexpensive Microcontrollers," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 3, pp. 733-740, March 2011.

[71] A. Dinu, M.N. Cirstea, and S.E. Cirstea, "Direct Neural-Network Hardware-Implementation Algorithm," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 5, pp. 1845-1848, May 2010.

[72] A. Gomperts, A. Ukil, and F. Zurfluh, "Development and Implementation of Parameterized FPGA-Based General Purpose Neural Networks for Online Applications," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 78-89, February 2011.

9. List of Publications

1. M. Paul, M. Chakraborty, ‘*Fuzzy Based System for Diagnosing Schizophrenia*’, International Journal of Engineering, Science and Technology, January, 2018.
2. M. Paul, M. Chakraborty, ‘*Syntehsis of Training Dataset for Artificial Neural Network for Diagnosing Schizophrenia’*, International Journal of Computer Engineering and Applications, December, 2017
3. M. Paul, M. Chakraborty, *‘Observation on Training Neural Network for Diagnosing Schizophrenia’,* International Journal of Advanced Research in Computer Science, January-February, 2018
4. M. Paul, M. Chakraborty, *‘Choice of Support Vector Machine Kernel for Classifying Schizophrenia Data’*, International Journal of Computer Science & Information Technology Research Excellence, January-February, 2018
5. M. Paul, M. Chakraborty, *‘Fuzzy Clustering for Diagnosing Schizophrenia’*, International Journal of Computer Engineering and Applications, April, 2018