ACKNOWLEDGEMENT

First and foremost I thank the God Almighty for giving me the strength, knowledge, ability, opportunity to undertake this research work and to persevere and complete it successfully.

I extend my thanks to **Dr. (Mrs.) R. Nandini, Chairperson,** PSGR Krishnammal College for Women, Coimbatore for having given me the opportunity to undertake this research work in this esteemed institution..

I express my whole hearted thanks to **Dr. (Mrs.) N. Yesodha Devi, M.Com, M.Phil, Ph.D., Secretary,** PSGR Krishnammal College for Women, Coimbatore for her continuous motivation and encouragement.

My heartfelt thanks to **Dr. (Mrs.) S. Nirmala, MBA, M.Phil., Ph.D., Principal,** PSGR Krishnammal College for Women, Coimbatore for her kind support and for all the resources provided.

I place on record my deep sense of gratitude to my research supervisor **Dr. (Mrs.) M.S. Vijaya, M.Sc., M.Phil., Ph.D.,** Associate Professor and Head, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore for her sustained interest and advice that have contributed to a great extent to the completion of this work. I am thankful for her appropriate guidance, insightful suggestions and support in the completion of this research work. Her patience and immense knowledge has helped me during the tenure of research and writing of thesis.

I wish to thank my friend and research colleague **Mrs. K. Sathiyakumari** for her timely help and words of encouragement during the course of study. My sincere thanks to the faculty members of Department of Computer Science (PG) for their co-operation rendered.

Finally, I place on record my deep sense of gratitude to my beloved parents, husband, son, my sister and in-laws for supporting me in successfully completing this work. This would not have been possible without their unwavering and unselfish love and support given to me at all times.

V.Pream Sudha

ABSTRACT

Globally an increasing number of children are being diagnosed with Autism Spectrum Disorder (ASD) and still more persons remain unidentified in the society. ASD is characterized by genetic heterogeneity and is defined by a range of conditions that affect persons to varied degrees. High risk factors for ASD include genes and genetic mutations, chromosomal conditions, family factors, prenatal influences and birth complications. Prediction of pretentious genes that underlie this disease is a significant challenge in biomedical research. This research titled "Deep Learning Framework For Efficient Prediction Of Causative Mutations, Genes And Their Susceptibility To Autism Spectrum Disorder" aims at predicting ASD causing genes, their susceptibility and contributing mutations by building models through user defined and self learned features based on conventional machine learning and contemporary deep learning approaches.

The core objectives of this research work are as follows.

1. To create a synthetic gene sequence database that mimics the causative ASD gene sequences
2. To identify and capture distinctive features from the diseased gene sequences that contribute to the classification of genes, their susceptibility to the disorder and the underlying mutations
3. To develop a framework based on conventional machine learning techniques for prediction of causative ASD genes and the underlying mutations
4. To develop a model to predict the susceptibility of the ASD genes to the disorder using supervised machine learning techniques
5. To build a deep learning framework for predicting the causative ASD genes, their vulnerability to the disorder and the driving mutations through user defined features
6. To employ Recurrent Neural Network (RNN) variants namely Bidirectional Recurrent Neural Network (BRNN), Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) with user defined features for creating computational models to predict ASD causing genes, their susceptibility and the driving mutations
7. To develop two kinds of encoding schemes namely codon encoding and one hot encoding to build DNN, BRNN, LSTM and GRU models through self-learnt features for predicting ASD causative genes

The thesis explains a novel and an unprecedented approach wherein the problem of predicting ASD genes, their vulnerability to the disorder and the driving mutations are

formulated as pattern classification tasks and solved through conventional supervised learning and contemporary deep learning methods. These approaches simplify the prediction problem in generating reliable solution based on intelligent hints collected from simulated gene sequences.

Accurate prediction of ASD causing genes and mutations is a complicated task as the pattern of the gene sequence varies for every individual. As diseased gene sequences are not readily available, they are simulated with the gene mutational information collected from the Human Gene Mutational Database (HGMD). The genes associated with two kinds of ASD are examined and ten genes namely FMR1, MECP2, TSC1, CACNA1C, SHANK3, CHD8, FOXP2, CNTNAP2, GABRB3 and HOXA1 that are key players for syndromic and asyndromic ASD are considered for the study. The reference genes are identified from OMIM (Online Mendelian Inheritance in Man) database and its corresponding reference gene sequences are downloaded from NCBI. The raw sequence obtained from HGMD is processed to form cDNA sequence and the nucleotide base variation is done based on the mutational information obtained from HGMD database for four types of mutations namely Nonsense, Missense, Frameshift and Silent Mutations. The synthetic mutated gene sequences are generated and stored as fasta files. In each gene category 100 gene sequences are generated and a corpus comprising of 1000 synthetic gene sequences is developed.

The research work is carried out in three stages using conventional machine learning and the contemporary deep learning methods for building the predictive models.

In the first stage, the traditional learning approach is employed for building the prediction models and the key idea here is to identify and extract distinctive features from synthetic disease gene sequences. Various descriptors such as gene features, codon features, alignment features accounting to 43 attributes have been captured from 1000 gene sequences of 10 gene types and the dataset called Codon Measures Dataset (CMDS) is developed for learning the gene classification model. Mutational discriminators including gene mutation features, amino acid change features and published matrix features are identified and extracted. The dataset comprising 1000 instances with a dimension of 15 is created and named as Mutation Dataset (MDS) which is exploited for constructing the mutation recognition model. The cumulative strength of evidence for each ASD associated gene with attributes pertaining to gene, mutation, conserved protein domains, gene expression profiles and pathway interactions are integrated for all 1000 instances. A consolidated score is calculated by summing the various features for each individual variant of an ASD implicated gene leading to a clear understanding of their relevance to the disorder. Finally one of the

three class labels namely low(score < 0.5), medium (score >=0.5 and < 0.8), high(score >=0.8) is assigned to the gene depending on the range. Around 25 features have been captured from 1000 gene sequences of 10 gene types and the dataset called Gene Susceptibility Dataset (GSDS) is developed for the gene susceptibility identification model.

To predict the ASD candidate genes and mutations concurrently by classifying them based on the contributing features, a multi-dimensional machine learning approach is proposed. The profound association between genes and mutations is modeled as a multi-label problem by capturing the dependencies between them. The pooled mutation dataset (PMDS) with 58 features including codon measures and mutation features from 1000 instances is created by pooling the gene, mutations, amino acid substitution attributes and is used for multidimensional modeling of gene - mutation prediction problem. Min - max normalization is done to standardize the feature values of all the four datasets.

Various experiments are carried out by implementing supervised classification algorithms such as Decision tree, Multilayer Perceptron, Support Vector Machines using the above three dataset CMDS, MDS, GSDS in Scikit Learn environment and different independent data driven models are built correspondingly to identify 10 types of genes, to recognize 4 forms of mutations and to predict gene susceptibility. Multi-dimensional classifiers such as Bayesian classifier chains, Nearest Set Replacement, Class relevance, Ensemble of classifier chains with base classifiers are implemented using PMDS in MEKA environment and multi-dimensional models are built to predict the ASD candidate gene-mutation simultaneously.

In the second stage, the deep learning approach is employed for building the prediction models and the key idea here is to explore the competence of deep models. As deep architectures are able to discover the high-level features, detect complex interactions among them, increase interpretability and support variable-size data, they can be potentially powerful in discriminating ASD genes, their susceptibility and mutations. Initially, Deep Neural Network (DNN), architecture is attempted for learning feature representations, modeling their sequential dependencies and prediction. In the subsequent works, the variants of Recurrent Neural Network (RNN) namely Bidirectional Recurrent Neural Network (BRNN), Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) are employed to build the classifiers with user defined features and representation learning for the above mentioned tasks.

The deep learning experiments are carried out in Keras environment with Tensorflow as backend. The same three user defined feature sets with one hot encoded class labels are

used for training the above deep networks. Independent deep models are built correspondingly to identify 10 types of genes, to recognize 4 forms of mutations and to predict gene susceptibility.

In the last stage, two types of encoding schemes are proposed for deep learning and the key idea here is to utilize the self-learned features of deep learning models. The gene sequences are utilized as raw input data without any domain expertise and thereby the time consuming task of feature engineering is avoided. The first encoding scheme adopts codon encoding of diseased gene sequences and the Codon Encoded Dataset (CEDS) with 1000 instances of dimension 2582 is prepared. The next scheme uses one hot encoding technique wherein each input sequence is transformed into a 7746 x 4 one-hot encoded vectors and the One Hot Encoded Dataset (OHEDS) of 1000 samples is developed. DNN, BRNN, LSTM and GRU architectures are employed to build the ASD causative gene type identification model.

Various experiments have been carried out by implementing the above deep learning algorithms using Keras environment with Tensorflow as backend. The two encoded datasets with one hot encoded class labels are used for training the deep networks and independent deep classifiers are built to identify 10 types of genes.

The performances of all the predictive classifiers are evaluated using 10 fold cross validation and their effectiveness in recognizing the genes, mutations and in predicting the gene susceptibility with respect to various metrics like precision, recall, accuracy, F-measure is analyzed. The experimental results demonstrate that deep learning based GRU model is efficient in predicting the genes and mutations whereas LSTM model is competent in detecting the gene susceptibility. Also the experimental results proved that the GRU model utilizing the one hot encoded gene sequences is efficient in discriminating the ASD causing genes.

Identification of pretentious genes and mutations that underlie ASD is a significant challenge in biomedical research and hence an attempt is made to carry out research in this domain.

CONTENTS

Acknowledgement i

Abstract ii

List of Figures viii

List of Tables xii

List of Abbreviations xv

1. INTRODUCTION 1

1.1 Data Mining 2

1.2 Machine Learning 10

1.2.1 Supervised Learning 11

1.2.2 Deep Learning 19

1.3 Overview of Autism Spectrum Disorder 26

1.4 Review of Literature 40

1.5 Motivation 46

1.6 Objectives of the Research 47

1.7 Organisation of the Thesis 48

2. DEEP LEARNING ARCHITECTURES 50

2.1 Deep Neural Networks 50

2.2 Convolutional Neural Networks 58

2.3 Restricted Boltzmann Machines 60

2.4 Deep Belief Networks 61

2.5 Recurrent Neural Networks 63

2.6 Long Short Term Memory Units 69

2.7 Gated Recurrent Units 72

3. PROBLEM MODELLING 76

3.1 Corpus Development 77

3.2 Design of Features and Datasets 82

3.3 Training and Testing 88

4. SUPERVISED LEARNING MODELS TO PREDICT ASD CAUSING GENES AND MUTATIONS 92

4.1 Model to Predict the ASD Causative Genes 92

4.2 Model to Predict ASD Causing Mutations 101

4.3 Multidimensional Model to Predict ASD Genes and Mutations 110

5. SUPERVISED LEARNING MODEL TO PREDICT GENE SUSCEPTIBILITY TO ASD **123**

6. DEEP NEURAL NETWORK MODEL TO PREDICT ASD GENES, THEIR SUSCEPTIBILITY AND MUTATIONS 138

7. RECURRENT NEURAL NETWORK MODELS TO PREDICT ASD GENES, THEIR SUSCEPTIBILITY AND MUTATIONS 149

7.1. BRNN Based Model to Predict ASD Genes, their Susceptibility and Mutations 149

7.2. LSTM Based Model to Predict ASD Genes, their Susceptibility and Mutations 158

7.3. GRU Based Model to Predict ASD Genes, their Susceptibility and Mutations 167

8. DEEP LEARNING MODELS TO PREDICT ASD GENES USING GENE SEQUENCE ENCODING 181

8.1 Codon Encoding of Gene Sequences for Deep Learning Based Prediction of ASD Causing Genes 181

8.2 One Hot Encoding of Gene Sequences for Deep Learning Based Prediction of ASD Causing Genes 195

9. CONCLUSION 207

References 211

List of Publications 219

Appendix

A. Datasets 220

B. Coding 227

C. Screenshots 237

LIST OF FIGURES

Figure No.	Title	Page No.
Fig. 1.1	Data Mining Tasks	2
Fig. 1.2	Linear Support Vector Machine	16
Fig. 1.3	A Typical ANN Architecture	19
Fig. 1.4	Comparison Between Artificial Neural Network and Deep Learning Model	21
Fig. 2.1	Basic Architecture of Deep Neural Networks	52
Fig. 2.2	Architecture of CNN	59
Fig. 2.3	RBM Architecture	60
Fig. 2.4	DBN Architecture	62
Fig. 2.5	Architecure of RNN	64
Fig. 2.6	RNN Unfolded	65
Fig. 2.7	RNN Types Based on Input / Output	67
Fig. 2.8	LSTM Cell	69
Fig. 2.9	Repeating Module in LSTM	70
Fig. 2.10	Gated Recurrent Unit	73
Fig. 3.1	Architecture of the Proposed System	76
Fig. 3.2	Mutation information retrieved for SHANK3 gene from HGMD	80
Fig. 3.3	cDNA sequence of SHANK3 Gene	81
Fig. 3.4	Mutated SHANK3 Gene	82
Fig. 4.1	Proposed Framework to Identify ASD Causing Genes	93
Fig. 4.2	Precision, Recall, F-Measure, Accuracy of ASD Gene Classifiers	98
Fig. 4.3	Precision of ASD Gene Classifiers	99
Fig. 4.4	Recall of ASD Gene Classifiers	99
Fig. 4.5	Mean Absolute Error of ASD Gene Classifiers	99
Fig. 4.6	ROC of the Decision Tree Classifier	100
Fig. 4.7	Architecture of ASD Causing Mutation Prediction Model	102
Fig. 4.8	Precision of Mutation Classifiers	108
Fig. 4.9	Recall of Mutation Classifiers	108

Fig. 4.10	Performance Comparison of Mutation Classifiers	109
Fig. 4.11	ROC Curve of Decision Tree Classifier	109
Fig. 4.12	Proposed Framework for Multi Dimensional Classification of Genes and Mutations	114
Fig. 4.13	Hamming Loss and Zero-One Loss of Gene - Mutation Classifiers	119
Fig. 4.14	Exact Match of Multidimensional Gene - Mutation Classifiers	120
Fig. 4.15	Accuracy and Average Training Time of Multi-Dimensional Classifiers	120
Fig. 4.16	Accuracy and Average Training Time of Base Classifiers	120
Fig. 5.1	Architecture of ASD Gene Susceptibility Prediction Model	124
Fig. 5.2	Overlap Matrix of Genes and Biological Processes	128
Fig. 5.3	Overlap Matrix of Gene and Cellular Components	130
Fig. 5.4	Precision, Recall, F-Measure, Accuracy of Gene Susceptibility Classifiers	133
Fig. 5.5	Precision of Gene Susceptibility Classifiers	133
Fig. 5.6	Recall of Gene Susceptibility Classifiers	133
Fig. 5.7	Performance Metrics of Gene Susceptibility Classifiers	134
Fig. 5.8	Accuracy of Gene Susceptibility Classifiers	134
Fig. 5.9	ROC Curve of SVM Classifier	134
Fig. 5.10	Pie Chart Showing the Gene Susceptibility	135
Fig. 5.11	Genes with Susceptibility Score	136
Fig. 6.1	Architecture of DNN Models For Identifying Genes, Susceptibility and Mutations	139
Fig. 6.2	Epochwise Accuracy of DNN Models	144
Fig. 6.3	Epochwise Log Loss of DNN Based Gene Identification Model	144
Fig. 6.4	Epochwise Log Loss of DNN Based Mutation Prediction Model	145
Fig. 6.5	Epochwise Log Loss of DNN Based Gene Susceptibility Prediction Model	145
Fig. 6.6	Accuracy of DNN Models	145
Fig. 6.7	Comparative Performance Measures of DNN with Shallow Models	147
Fig. 7.1	Architecture of the BRNN Based Prediction Model	150

Fig. 7.2	Accuracy of BRNN Models	154
Fig. 7.3	Epochwise Log loss of BRNN Models	154
Fig. 7.4	Performance Results of BRNN Based Gene Prediction Model	155
Fig. 7.5	Performance of BRNN Based Mutation Prediction Model	155
Fig. 7.6	Performance of BRNN Based Gene Susceptibility Prediction Model	155
Fig. 7.7	Specificity of BRNN Models	156
Fig. 7.8	Comparative Performance of BRNN and DNN Based Models	157
Fig. 7.9	Architecture of the LSTM Based Prediction Model	159
Fig. 7.10	Epochwise Accuracy of LSTM Models	163
Fig. 7.11	Epochwise Log Loss of LSTM Models	163
Fig. 7.12	Performance of LSTM Models	164
Fig. 7.13	Precision of LSTM Models	164
Fig. 7.14	Comparative Performance of DNN, BRNN and LSTM Based Models	166
Fig. 7.15	Architecture of GRU Based Prediction Model	168
Fig. 7.16	Epochwise Accuracy of GRU Based Models	173
Fig. 7.17	Epochwise Logloss of GRU Based Models	173
Fig. 7.18	Performance of GRU Based Gene Prediction Model	173
Fig. 7.19	Performance of GRU Based Mutation Prediction Model	174
Fig. 7.20	Performance of GRU Based Gene Susceptibility Prediction Model	174
Fig. 7.21	Comparative Performance of GRU Based ASD Gene Prediction Model with Other Deep Models	177
Fig. 7.22	Comparative Performance of GRU Based ASD Mutation Model with Other Deep Models	177
Fig. 7.23	Comparative Performance of GRU Based ASD Gene Susceptibility Prediction Model With Other Deep Models	178
Fig. 7.24	Log Loss Comparison of GRU Based Prediction Models With Other Deep Models	178
Fig. 8.1	Architecture of Codon Encoding Based Deep Models for Identifying ASD Genes	185
Fig. 8.2	Epochwise Accuracy of Deep Models with Codon Encoding	189

Fig. 8.3	Epochwise Log Loss of Deep Models with Codon Encoding	189
Fig. 8.4	Performance of DNN Model with Codon Encoding	190
Fig. 8.5	Performance of BRNN Model with Codon Encoding	190
Fig. 8.6	Performance of LSTM Model with Codon Encoding	190
Fig. 8.7	Performance of GRU Model with Codon Encoding	191
Fig. 8.8	Performance Comparison of Deep Models with Codon Encoding	191
Fig. 8.9	Comparative Performance of Deep Models - Self Learned Features (CEDS) Vs User Defined Features (CMDS)	193
Fig. 8.10	Log Loss Comparison of Deep Models - Self Learned Features (CEDS) Vs User Defined Features (CMDS)	194
Fig. 8.11	Architecture of One Hot Encoding Based Deep Models for Identifying ASD Genes	196
Fig. 8.12	Epochwise Accuracy of Deep Models with One Hot Encoding	200
Fig. 8.13	Epochwise Log Loss of Deep Models with One Hot Encoding	200
Fig. 8.14	Performance of DNN Model with One Hot Encoding	201
Fig. 8.15	Performance of BRNN Model with One Hot Encoding	201
Fig. 8.16	Performance of LSTM Model with One Hot Encoding	202
Fig. 8.17	Performance of GRU Model with One Hot Encoding	202
Fig. 8.18	Comparative Performance of Deep Models - Self Learned Features (OHEDS) Vs User Defined Features (CMDS)	204
Fig. 8.19	Log Loss Comparison of Deep Models - Self Learned Features (OHEDS) Vs User Defined Features (CMDS)	205

LIST OF TABLES

Table No.	Title	Page No.
Table I	Genes and Gene Functions Associated with Idiopathic ASD	35
Table II	Common Genetic Syndromes and Genes Associated	36
Table III	Clinical Features of the Syndromes and their Prevalence in ASD	38
Table IV	Summary of Literature Survey	43
Table V	ASD Causative Genes Taken for Study and their Associated Behavior	78
Table VI	Summary of Training Datasets	87
Table VII	Performance Results of ASD Gene Classifiers	97
Table VIII	Performance Results of Mutation Classifiers	107
Table IX	Classwise Performance Results of Decision Tree Classifier	107
Table X	Performance of Multi Dimensional Gene – Mutation Classifiers Based On Loss Measures	117
Table XI	Training Time, Test Time, Accuracy for Each Configuration	118
Table XII	Average Accuracy, Training and Testing Time of The Gene – Mutation Classifiers	119
Table XIII	Protein - Protein Interaction Pathways Associated with Genes and their P-Values	126
Table XIV	Biological Components Associated with Genes and Their P-Values	127
Table XV	Cellular Components Associated with Genes and Their P-Values	129
Table XVI	Performance of Gene Susceptibility Classifiers	132
Table XVII	Hyperparameter Setting for DNN Training	142
Table XVIII	Accuracy of DNN for Varying Epochs	142
Table XIX	Epochwise Log loss of DNN Models	143
Table XX	Performance Comparison of DNN Models	143

Table XXI	Comparative Performance of DNN and Shallow Methods	146
Table XXII	Epochwise Accuracy of BRNN Models	152
Table XXIII	Epochwise Log Loss of BRNN Models	153
Table XXIV	Performance Results of BRNN Models	153
Table XXV	Comparative Performance of BRNN and DNN Based Models	157
Table XXVI	Epochwise Accuracy of LSTM Models	161
Table XXVII	Epochwise Log loss of LSTM for Three Datasets	162
Table XXVIII	Performance Results of LSTM Models	162
Table XXIX	Comparative Performance of DNN, BRNN and LSTM Based Models	165
Table XXX	Epochwise Accuracy of GRU Models	171
Table XXXI	Epochwise Log loss of GRU Models	171
Table XXXII	Performance Results of GRU Based Models	172
Table XXXIII	Comparative Results of GRU Based ASD Gene Prediction Model with Other Deep Models	175
Table XXXIV	Comparative Results of GRU Based ASD Mutation Prediction Model with Other Deep Models	175
Table XXXV	Comparative Results of GRU Based ASD Gene Susceptibility Prediction Model with Other Deep Models	176
Table XXXVI	Comparative Results of Log Loss of GRU Based Prediction Model with Other Deep Models	176
Table XXXVII	Genetic Code Showing Relation Between Codons and Amino Acid	183
Table XXXVIII	Epochwise Accuracy of Deep Models with Codon Encoding	187
Table XXXIX	Epochwise Log loss of Deep Models with Codon Encoding	188
Table XL	Performance Results of Deep Models with Codon Encoding	188
Table XLI	Comparative Performance of Deep Models - Self Learned Features (CEDS) Vs User Defined Features (CMDS)	192
Table XLII	Log Loss of Deep Models - Self Learned Features (CEDS) Vs User Defined Features (CMDS)	193

Table XLIII	Epochwise Accuracy of Deep Models with One Hot Encoding	198
Table XLIV	Epochwise Logloss of Deep Models with One Hot Encoding	199
Table XLV	Performance Results of Deep Models with One Hot Encodings	199
Table XLVI	Comparative Performance of Deep Models - Self Learned Features(OHEDS) Vs User Defined Features (CMDS)	203
Table XLVII	Log Loss of Deep Models - Self Learned Features(OHEDS) Vs User Defined Features (CMDS)	204

LIST OF ABBREVIATIONS

A	Adenine
AARC	Amino Acid Residue Changes
ANN	Artificial Neural Network
ASD	Autism Spectrum Disorder
BLOSUM	Block Substitution Matrix
C	Cytosine
CEDS	Codon Encoded Dataset
CMDS	Codon Measures Dataset
CNN	Convolutional Neural Network
CNV	Copy Number Variants
DAWN	Detecting Association with Networks
DBN	Deep Belief Networks
DNN	Deep Neural Networks
DSM- 5	Diagnostic and Statistical Manual of Mental Disorders version 5
FT	Functional Trees
FXS	Fragile X Syndrome
G	Guanine
GRU	Gated Recurrent Units
GS	Gene Specific Features
HGMD	Human Gene Mutational Database
LMT	Logistic Model Trees
LSTM	Long Short Term Memory
MDS	Mutation Dataset
MLP	Multi Layer Perceptron
NB	Naive Bayes
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
OHEDS	One Hot Encoded Dataset
OMIM	Online Mendelian Inheritance in Man
PMDS	Pooled Mutation Dataset
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines

RNN	Recurrent Neural Networks
RTT	Rett syndrome
SGD	Stochastic Gradient Descent
SM	Substitution Matrix
SVM	Support Vector Machine
T	Thymine