# DEEP LEARNING FRAMEWORK FOR EFFICIENT PREDICTION OF CAUSATIVE MUTATIONS, GENES AND THEIR SUSCEPTIBILITY TO AUTISM SPECTRUM DISORDER

Thesis submitted to
**BHARATHIAR UNIVERSITY**

for the award of the degree of
**DOCTOR OF PHILOSOPHY**

in
**COMPUTER SCIENCE**

By
**V. PREAM SUDHA M.C.A, M.Phil.,**

Under the guidance of
**Dr. (Mrs.) M. S. VIJAYA M.Sc., M.Phil., Ph.D**
Associate Professor and Head
Department of Computer Science (PG)


**PSGR KRISHNAMMAL COLLEGE FOR WOMEN**
College of Excellence – nirf 22nd Rank
(An Autonomous Institution – Affiliated to Bharathiar University)
Reaccredited with "A" grade by NAAC
An ISO 9001:2015 Certified Institution
**Coimbatore – 641 004, Tamilnadu, India**

**AUGUST 2019**

# 9. CONCLUSION

The thesis titled “Deep Learning Framework for Efficient Prediction of Causative Mutations, Genes and Their Susceptibility to Autism Spectrum Disorder” portrays the research work carried out to predict ASD causing genes, their susceptibility to the disorder and triggering mutations through conventional machine learning and contemporary deep learning approaches.

The core objective of this work is to propose an efficient and appropriate solution for prediction tasks based on mutated gene sequences using two learning approaches. Traditional machine learning approach is implemented by identifying and extracting hand crafted features from mutated gene sequences. Deep learning approach aids in building prediction models through self learned features and high level abstractions.

The process of corpus creation has been carried out by generating 1000 synthetic gene sequences using gene mutational information collected from the Human Gene Mutational Database. Ten types of ASD causative genes and four types of mutations have been taken for study. Various descriptors related to codon measures, mutation features, amino acid change features and published matrix features have been extracted from mutation induced gene sequences and four different datasets have been created.

The discriminative models have been built using traditional supervised pattern classification algorithms such as Decision Tree, Multilayer Perceptron and Support Vector Machines for the prediction of ASD causing genes, their susceptibility to the disorder and the underlying mutations. Also multi-dimensional classifiers have been built to predict the ASD gene - mutation by classifying them concurrently based on the pooled features using multi-label learning.

Deep learning based predictive models have been developed using DNN, BRNN, LSTM and GRU by learning representations from the user defined features to classify ASD genes, their predisposition to ASD and driving mutations.

Further, two kinds of encoding schemes namely codon encoding and one hot encoding have been proposed to transform the gene sequence as direct input to deep architectures DNN, BRNN, LSTM and GRU and ASD gene identification models have been built through self-learning features.

Performances of the classifiers built through traditional learning and the modern deep learning techniques are evaluated using various measures like precision, recall, F-measure, accuracy etc., and the empirical result analysis is carried out and the several valuable outcomes are derived. The observations made and the interpretations drawn from this research work are summarized below.

- Diseased gene sequences can be computationally generated for predicting ASD genes, their susceptibility to the disorder and the underlying mutations
- Traditional machine learning approaches implemented to achieve objectives 3 and 4 attained desirable accuracy which instigated to broaden the research to advanced level using deep learning
- The prediction model is effective when the pooled features are used in multi label learning
- Deep learning enables the prediction task through representation learning by self extraction of intelligent hints from the synthetic gene sequences and hence the predictive accuracy of the deep models is found to be comparatively higher than that of traditional learning methods when manually extracted features are used
- Long Short Term Memory units demonstrates promising results for predicting the gene susceptibility to the disorder when feature engineering is used
- Gated Recurrent Unit model which was built with manual features and shared layers has the advantage of less parameters, easier training and high accuracy for predicting ASD causing genes and mutations.
- The prediction models built using deep learning architectures DNN, BRNN, LSTM and GRU exhibits enhanced performance with encoded datasets in discriminating genes by extracting gene characteristics automatically rather than with user defined datasets
- The proposed encoding schemes considerably reduces the task of feature engineering in building deep models
- The deep architectures have shown promising results with encoded schemes and specifically the GRU model outperforms the other models in predicting ASD causing genes with one hot encoded dataset

The research contributions of the work are listed below.

- Autism Spectrum disorder which is increasing among children has been taken for study
- Gene sequences which have not been investigated for ASD so far have been considered
- Developed a corpus of synthetic mutated ASD gene sequences using mutational information from HGMD
- Identified and captured differentiating attributes from gene sequences related to different types of genes, susceptibility to the disorder and driving mutations and four datasets have been created
- Built prediction models through traditional pattern classification and multi-dimensional modeling
- Built an efficient deep learning framework for predicting ASD causative genes, their susceptibility to the disorder and the driving mutations
- Designed an effective solution to characterize the variable length gene sequence data using RNN variants
- Performance enhancement of deep models through encoding schemes
- Development of a generalized predictive model which is appropriate for any genetic disorder

It is concluded that state of the art deep learning based models are more efficient in predicting the ASD causing genes, their susceptibility to the disorder and driving mutations than conventional pattern recognition methods. Exhaustive experiments carried out ascertain that the deep learning framework has the clear advantage over traditional machine learning as it does not depend on domain expertise for manually extracting the features or any specific preprocessing. It is evident that deep models show enhanced performance as feature engineering is done on its own by discovering new, high - level features in an incremental manner.

The proposed approaches are significantly useful to clinicians in making precise diagnoses and for pursuing targeted genetic testing of individuals with ASD. In a clinical environment, when a diseased gene sequence is provided, the proposed models will facilitate faster targeted treatments as the ASD causing gene, its susceptibility level and the underlying mutation can be efficiently identified.

In future, the work can be extended by proposing new encoding schemes for deep learning to address the problem of predicting the ASD causing mutations and the susceptibility

of genes to the disorder. Also multi-dimensional approach of classifying genes and co-occurring mutations can be attempted using deep learning techniques which will be a novel initiative in this domain.