Deep learning model in analysing the Genetic variation associated with the occurrence and progression of Neurodevelopmental disorders

Dr. S. Pitchumani Angayarkanni, Associate Professor, Department of Computer Science, Lady Doak College, Madurai, TN, India
Dr. S. Kalaivani Priyadarshini, Assistant Professor, Department of Biotechnology, Lady Doak College, Madurai, TN, India
Dr. Sofia, Associate Professor, Department of Computer Science, Lady Doak College, Madurai, TN, India
Ms. Raga Priya, UG Student, Bioinformatics, TN Agricultural University, Coimbatore, TN, India

Summary:

Neuro developmental disorders are group of childhood onset disorders. The most severe NDD affects the multiple domains of cognitive development are intellectual disability (ID), pervasive disorders of social communication like (Autism Spectrum Disorder (ASD)), motor functioning and cognition (epilepsy encephalopathies) and behavioural regulations (Attention Deficit Hyperactive Disorder, ADHD). Under this category some of them are single gene disorders. ASD and ADHD are common and they result in major functional impairment related to high co-morbidity rates. Identification of the disorder-gene association is mainly used to understand the pathogenies and therapeutic targets discovery. Relationship between the disease/disorder and gene can be determined by analysing the genomic sequences. One of the challenges in predicting the complex human disease status is using genomic data. The curse of dimensionality results in unsatisfied performance of many algorithms. Recent advancements in machine learning is the deep learning which can be used to extract meaningful features from high-dimensional and complex datasets through stacked and hierarchical learning process. Deep Learning algorithms shows promising predictive potential by applying learning strategies based on pattern classification of the input gene sequence to the type of possible disorders(Mohammed et. al., 2019).

Objectives:

Design and optimize the pathway for diagnosis, therapeutic intervention, and prognosis by using large multidimensional biological datasets that capture individual variability in genes, function and environment to identify neuro developmental disorders.

- Duchenne muscular dystrophy
- Cerebral palsy
- Autism
- ADHD

Scope:

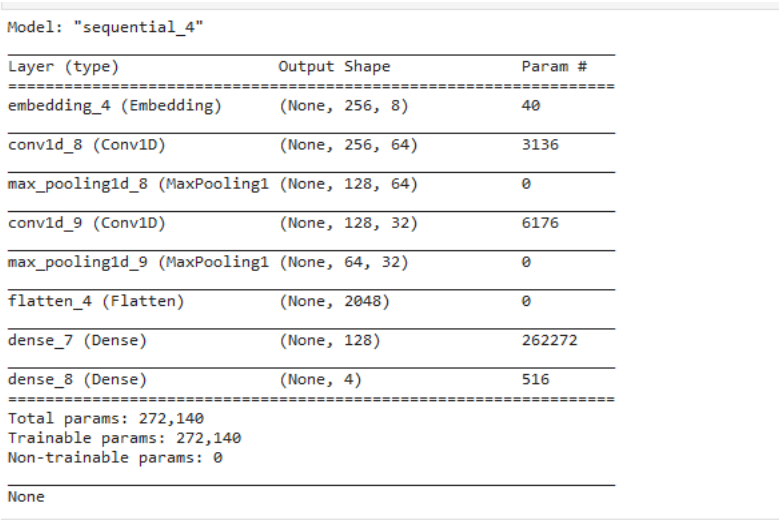
To identify and predict the genomic variations among children in the following neuro developmental disorders using deep learning model

□ Duchenne muscular dystrophy
□ Cerebral palsy
□ Autism
□ ADHD

The effective development of deep learning model helps to the early detection of embryonic neurodevelopmental disorders (ENDs) based on its prognostic values could render quality diagnosis and health management.

Methodology and Outcome:

In this paper we propose methodologies to formulate the Neuro Developmental Disorder dataset which comprises of the fasta sequence corresponding to ADHD, ASD, Duchene Muscular Disorder (DMD) and Cerebral Palsy (CP) using web scrapping approach and natural language processing. The formulated dataset is validated by splitting the gene id from the sequence using natural language processing technique and matching with the dataset provided by NCBI related to developmental brain disorder https://www.dbdb.urmc.rochester.edu and dbGAP for ADHD through web scrapping technique. The dataset is fed as input to the convolution neural network to classify the gene sequence based on the class label which corresponds to ADHD, ASD, DMD and CP. The proposed CNN provides an accuracy of 95%. Proposed CNN architecture is shown in figure 1.



```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 256, 8)            40
_________________________________________________________________
conv1d_8 (Conv1D)            (None, 256, 64)           3136
_________________________________________________________________
max_pooling1d_8 (MaxPooling1 (None, 128, 64)           0
_________________________________________________________________
conv1d_9 (Conv1D)            (None, 128, 32)           6176
_________________________________________________________________
max_pooling1d_9 (MaxPooling1 (None, 64, 32)            0
_________________________________________________________________
flatten_4 (Flatten)          (None, 2048)              0
_________________________________________________________________
dense_7 (Dense)              (None, 128)               262272
_________________________________________________________________
dense_8 (Dense)              (None, 4)                 516
=================================================================
Total params: 272,140
Trainable params: 272,140
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 1: Proposed CNN for classification of NDD gene sequence

Hyperparameter	Range
Kernel size for convolution	3
Number of kernels (in two convolution layers)	256X64 and 128X32
Pooling method	Max pooling
Pooling in second layer	Max Pooling
Number of units in hidden layer (ratio to input layer)	1/3, 1/2, 2/3, 3/4, 1
Learning algorithm	Adam

Table 1: Hyperparameters used for CNN

This was followed by the statistical approach to find the correlation between the genes which plays a vital role in diagnosing the disorder and which has least correlation in the diagnosis and which type of gene overlap between the disorders. To perform this process we used the bioinformatics tools like metascape for enrichment gene analysis, Malacards for correlation analysis and VLAD: Gene List Analysis and Visualization. Further the predicted genes which play a less significant role in the identification of the disorders were identified and the results are compared with the literature review to justify the resultant output. This research work has clearly revealed considerable overlap of genes involved in more than one NDD. The proposed outcome is validated with the WES approach which clearly demonstrated in a recent study based in consanguineous families with NDDs, in which 14 new candidate genes not previously associated with NDD disorders were identified (*GRM7*, *STX1A*, *CCAR2*, *EEF1D*, *GALNT2*, *SLC44A1*, *LRRIQ3*, *AMZ2*, *CLMN*, *SEC23IP*, *INIP*, *NARG2*, *FAM234B*, and *TRAP1*) all in patients who were homozygous for truncating mutations in each of the genes and with SFARI Gene bioinformatics tool. The phylogenetic tree generated for the formulated dataset to identify the similar and dissimilar gene sequences. The phylogenetic tree plotted between the gene sequences clearly depicts that Each major clusters has sub-clusters. DMD disease sequences are clustered in the first and third major clusters. They are, NM 001365584.1 Homo sapiens neuroligin 4 Y-linked (NLGN4Y) transcript variant 6 mRNA DMD and NR 028319.1 Homo sapiens neuroligin 4 Y-linked (NLGN4Y) transcript variant 4 non-coding RNA DMD , NM 001365591.1 Homo sapiens neuroligin 4 Y-linked (NLGN4Y) transcript variant 10 mRNA DMD and NM 001365586.1 Homo sapiens neuroligin 4 Y-linked (NLGN4Y) transcript variant 7 mRNA DMD , NM 001282145.2 Homo sapiens neuroligin 4 X-linked (NLGN4X) transcript variant 3 mRNA DMD and NM 181332.3 Homo sapiens neuroligin 4 X-linked (NLGN4X) transcript variant 2 mRNA DMD were closely related. CP and DMD disease sequence comes under the second and third major clusters respectively.

The CNN algorithm was implemented for classification of the gene sequence resulted in an accuracy of 95% with Area under ROC curve=0.90. The Statistical Interpretation between the gene sequences using metascape.org enrichment analysis was done. The genes with negative correlation was analysed and validated using gene analytics tool.

	Gene	**GO:0071625 vocalization behavior**	**GO:0042391 regulation of membrane potential**	**GO:2000146 negative regulation of cell mo**	**GO:0071560 cellular response to transform**
Gene	**1**				
GO:0071625 vocalization behavior	**-0.33386**	**1**			
GO:0042391 regulation of membrane potential	**-0.25214**	**0.527101636**	**1**		
GO:2000146 negative regulation of cell mo	**-0.2014**	**0.527101636**	**0.206349206**	**1**	
GO:0071560 cellular response to transform	**-0.10325**	**0.168408267**	**0.213844343**	**-0.02916**	**1**

Table 1: Negative Gene Correlation

Genes with negative correlation related to Vocalization behaviour GO:0071625 are CNTNAP2,NLGN3,NLGN4X,NLGN4Y, Regulation of membrane potential GO:0042391 are DMD,HTR3A,MEF2C,NLGN3,NLGN4X, Negative regulation of cell motility GO:2000146 are DAG1,KANK1,MEF2C,SPOCK3 and Cellular response to transforming growth factor beta stimulus GO:0071560 are DUSP15,LTBP4,MEF2C.

Justification:

Positive correlation of the finding with review of literature & Gene ontology study

Pathogenic mutations in the X-linked Neuroligin 4 gene (NLGN4X) in autism spectrum disorders (ASDs) and/or mental retardation (MR) are rare (Daoud , 2009).

According to gene antology annotation DMD and NLGN4X has not been associated with Regulation of membrane potential while MEF2C the gene associated with AUTISM, DMD, ADHD and NLGN3,NLGN4X which is associated with autism is based on positive regulation of excitatory postsynaptic potential and it is unclear according to the literature of how mutations in *NLGN4X* result in neurodevelopmental defects is associated with autism (Lingling, 2013). According to gene ontology study SPCOK3 is not associated with Negative regulation of cell motility because it is associated with Hemostatic Risk Factors and Arterial Thrombotic Disease (Reiner,2001) and MFC2C negative regulation of blood vessel endothelial cell migration (Schechter DS et. al., 2017).

Cellular response to transforming growth factor beta stimulus DUSP15 which is associated with ADHD is identified as a key regulator gene for oligodendrocytes differentiation which is associated with autism(Tian Y et. al.,2017). HTR3A gene involved in Autism is associated with regulation of membrane potential according to gene ontology annotation but it is associated with suicidal behaviour(Souza et. al., 2011). LTBP4 is associated with transforming growth factor beta receptor signalling pathway and leads to kidney disease (https://maayanlab.cloud/Harmonizome/gene_set/Kidney+Diseases/CTD+Gene-Disease+Associations)

Negative correlation of the finding:

Neurobiological, genetic, and imaging data provide strong evidence for the CNTNAP2 gene as a risk factor for ASD and related neurodevelopmental disorders (Peñagarikano et. al.,2012). Negative regulation of cell mobility DAG1 gene responsible for DMD is associated based on gene ontology study, Negative correlation of MEF2C gene responsible for Autism is a Gene to cellular response to transforming growth factor beta stimulus based on gene ontology study online tool mismatches with the findings.

Code Repository:

- Github Repository of the Project: angayarkannipitchumani/**DeepLearning-for-NDD-Classification**

Recommendations:

Electronic health record pertaining to the on medical profiles and diagnostic testing like patient’s profile, vital signs, systems review, clinical impression and diagnosis, medical orders and disposition, if made available in the public repository for NDD it will help in identifying the major cause.

Due to the very complex nature of NDDs, interdisciplinary approaches combining genetics, functional genomics, robust biological models and objective measures of response, such as biomarkers, as well as the capability of researchers and clinicians to work side by side, will be essential.

Acknowledgement:

We are grateful to the initiative and the support rendered by experteze.org research team headed by **Dr. MOHAN VENKATARAMANA**, President/CEO and **Mr. SARAVANAN DHANDAPANI**, Senior Vice President for their motivation and systematic planning in helping us in shaping our project and achieve the result within the time frame.

References:

1. Daoud, Hussein & Bonnet-Brilhault, Frédérique & Marouillat Vedrine, Sylviane & Demattéi, Marie-Véronique & Vourc'h, Patrick & Bayou, Nadia & Andres, Christian & Barthélémy, Catherine & Laumonnier, Frédéric & Briault, Sylvain. (2009). Autism and Nonsyndromic Mental Retardation Associated with a De Novo Mutation in the NLGN4X Gene Promoter Causing an Increased Expression Level. Biological psychiatry. 66. 906-10. 10.1016/j.biopsych.2009.05.008.
2. Lingling Shi, Xiao Chang, Peilin Zhang, Marcelo P. Coba, Wange Lu, Kai Wang, The functional genetic link of *NLGN4X* knockdown and neurodevelopment in neural stem cells, *Human Molecular Genetics*, Volume 22, Issue 18, 15 September 2013, Pages 3749–3760, https://doi.org/10.1093/hmg/ddt226
3. Peñagarikano, Olga & Geschwind, Daniel. (2012). What does CNTNAP2 reveal about autism spectrum disorder?. Trends in molecular medicine. 18. 156-63. 10.1016/j.molmed.2012.01.003.
4. Reiner, Alex & Siscovick, David & Rosendaal, Frits. (2001). Hemostatic Risk Factors and Arterial Thrombotic Disease. Thrombosis and haemostasis. 85. 584-95. 10.1055/s-0037-1615638.
5. Sampath S, Bhat S, Gupta S, et al. Defining the contribution of CNTNAP2 to autism susceptibility. *PLoS One*. 2013;8(10):e77906. Published 2013 Oct 17. doi:10.1371/journal.pone.0077906
6. Schechter DS, Moser DA, Pointet VC, Aue T, Stenz L, Paoloni-Giacobino A, Adouan W, Manini A, Suardi F, Vital M, Sancho Rossignol A, Cordero MI, Rothenberg M, Ansermet F, Rusconi Serpa S, Dayer AG. The association of serotonin receptor 3A methylation with maternal violence exposure, neural activity, and child aggression. Behav Brain Res. 2017 May 15;325(Pt B):268-277. doi: 10.1016/j.bbr.2016.10.009. Epub 2016 Oct 5. PMID: 27720744.
7. Souza, Renan & de Luca, Vincenzo & Manchia, Mirko & Kennedy, James. (2011). Are serotonin 3A and 3B receptor genes associated with suicidal behavior in schizophrenia subjects?. Neuroscience letters. 489. 137-41. 10.1016/j.neulet.2010.11.079.

8. Tărlungeanu, D.C., Novarino, G. Genomics in neurodevelopmental disorders: an avenue to personalized medicine. *Exp Mol Med* **50,** 100 (2018). https://doi.org/10.1038/s12276-018-0129-7
9. Tian Y, Wang L, Jia M, Lu T, Ruan Y, Wu Z, Wang L, Liu J, Zhang D. Association of oligodendrocytes differentiation regulator gene DUSP15 with autism. World J Biol Psychiatry. 2017 Mar;18(2):143-150. doi: 10.1080/15622975.2016.1178395. Epub 2016 May 25. PMID: 27223645.
10. Uddin, M., Wang, Y. & Woodbury-Smith, M. Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digit. Med.* **2,** 112 (2019). https://doi.org/10.1038/s41746-019-0191-0