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

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). In this paper we propose methodologies to formulate the Neuro Developmental Disorder dataset which comprises of the fasta sequence corresponding to ADHD, ASD, Duchne Muscular Disorder(DMD) and Cerebral Palsy(CP) using webscrapping 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%. 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 polygenetic tree plotted between the gene sequence 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 sequence with negative correlation was analysed and validated using gene analytics tool and metascape.org. 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.

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

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:

I. Formation of Dataset

The dataset with the gene sequence is formulated using the web scrapping approach through python pipeline library called entrez. Entrez is an online search tool by NCBI. It is a Molecular biology databases with an integrated global query supporting Boolean operators and field search. It returns results from all the databases with information like the number of hits from each databases, records with links to the originating database, etc. Biopython provides an Entrez specific module, Bio.Entrez to access Entrez database. The Bio.Enterz library is used to retrieve the fasta sequence from NCBI based on the keyword search related to the four different neuro developmental disorders in human which is indicated in Figure 1.

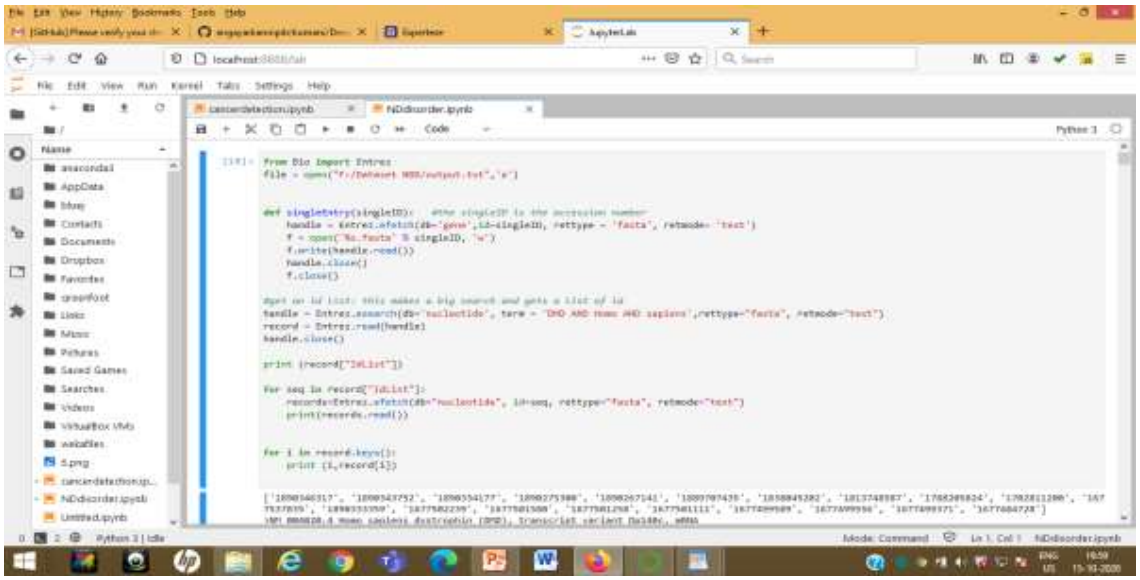


Figure 1: Enterz bio python Library to access the gene sequence from NCBI

The derived sequences are stored in a csv file using pandas library with Sequence Number, Sequence and Sequence Description which is represented in Figure 2.

II Dataset Validation

The collected dataset is validated by retrieving the gene details from the description field of the Genbank nucleotide sequence data files using Natural Language Processing technique. Tokenization is essentially **splitting** a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual **words** or terms. Each of these smaller units are called tokens. The tokens could be **words**, numbers or punctuation marks. We use this concept to split the description words into tokens and collect only the gene details from the description.

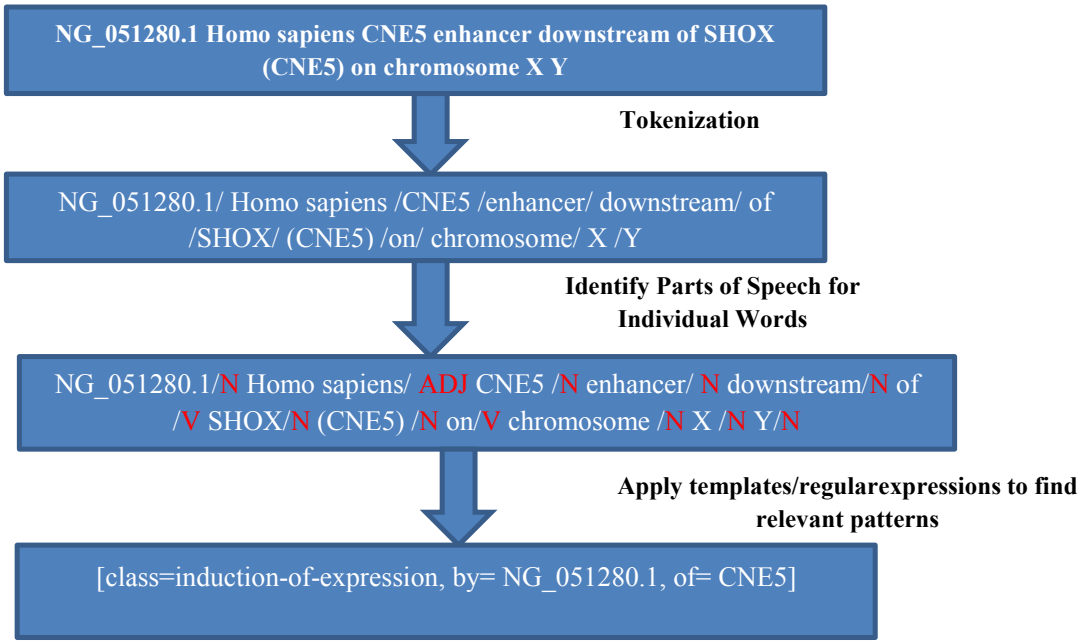


Figure 3: NLP technique implemented to identify the Gene from the description

The resultant gene retrieved from the sequence is cross verified by copying and pasting the gene to https://www.ncbi.nlm.nih.gov/gene/?term=NLGN4Y or by using the following query a new dataset with Gene ID and other relevant information was created

- Autism
- https://ghr.nlm.nih.gov/search?query=autism&tab=gene
- ADHD
- https://ghr.nlm.nih.gov/search?query=adhd&tab=gene
- Duchenne muscular dystrophy
- https://ghr.nlm.nih.gov/search?query=duchenne+muscular+dystrophy&tab=gene&rows=10
- cerebral palsy
- https://ghr.nlm.nih.gov/search?query=cerebral+palsy&tab=gene

The retrived gene id is validated with the dataset formulated with tax_id, Org_name, GeneID, CurrentID, Status, Symbol, Aliases, description, other_designations, map_location, chromosome, genomic_nucleotide_accession.version, start_position_on_the_genomic_accession, end_position_on_the_genomic_accession, orientation, exon_count , OMIM and Class

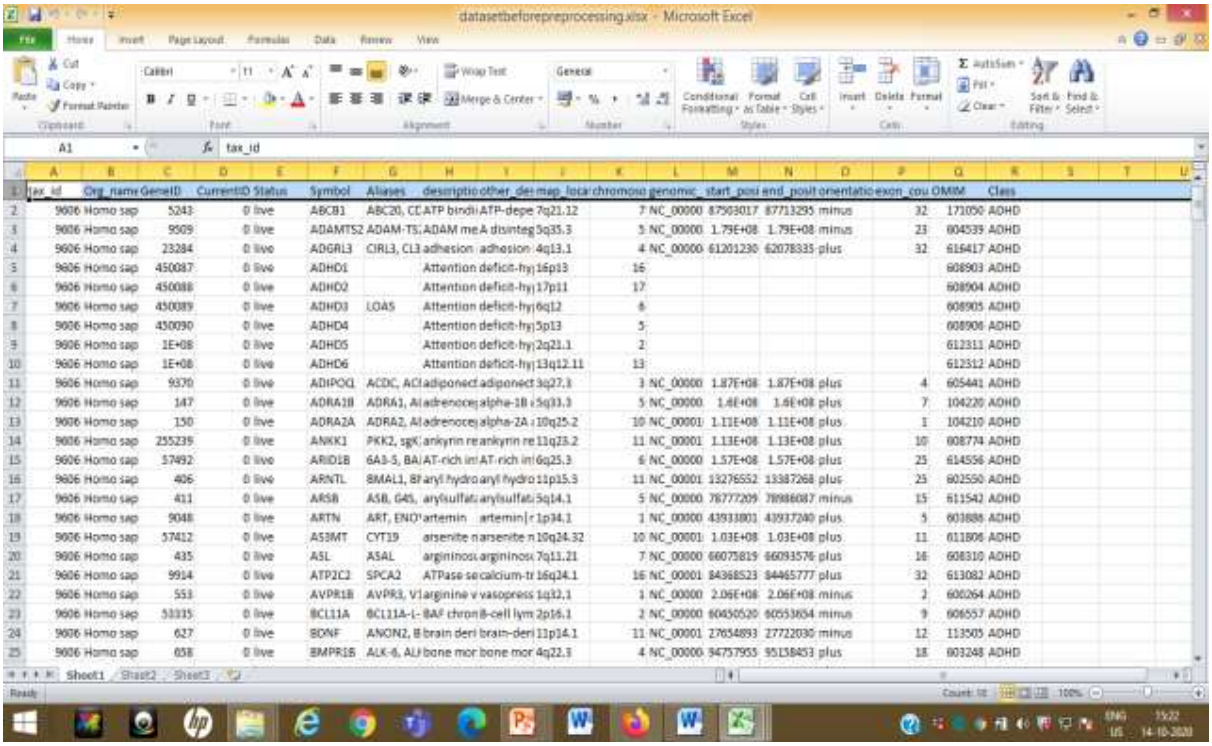


Figure 4: GeneID and other parameters collected from NIH

Gene Sequence Classification using Deep learning model:

Accurate gene prediction in metagenomics fragments is a computationally challenging task due to the short-read length, incomplete, and fragmented nature of the data. Most gene-prediction programs are based on extracting a large number of features and then applying

statistical approaches or supervised classification approaches to predict genes(Al-Ajlan, A., El Allali, A., 2019). We use deep learning techniques to automatically extract significant features from raw data, such as image intensities or DNA sequences. In this research work we had implemented the convolutional neural network (CNN) to the classification problem of DNA sequences based on the four types of neuro developmental disorders. The training of CNNs with distributed representations of four nucleotides has successfully derived position weight matrices on the learned kernels . The proposed architecture is shown in figure 5. The sequence coloumn alone is read from the csv file and given as input to CNN by performing one-hot encoding technique in which the gene sequences are encoded as a binary value using One-hot encoding technique.

We have implemented One-hot encoding to represent the DNA sequence using binary values. This is widely used in dep learning methods and lends itself well to algorithms like convolutional neural networks. In this example, “ATGC” would become [0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0,0,0]. And these one-hot encoded vectors can either be concatenated or turned into 2 dimensional arrays .

Sequence letter	Binary value
A	0001
G	0100
T	0010
C	1000

The proposed model includes four steps in total regarding to different layer embedded. The model contains one embedding layer which will encoded the sequences and one convolutional layer followed by a max-pooling layer which extracts features from representation matrixes of sequences. Then, all the extracted features is merged into one big feature vector using fully connected layer. Finally, the accuracy of the tested model is calculate and will be analysed as a performance result.

```
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 5: 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

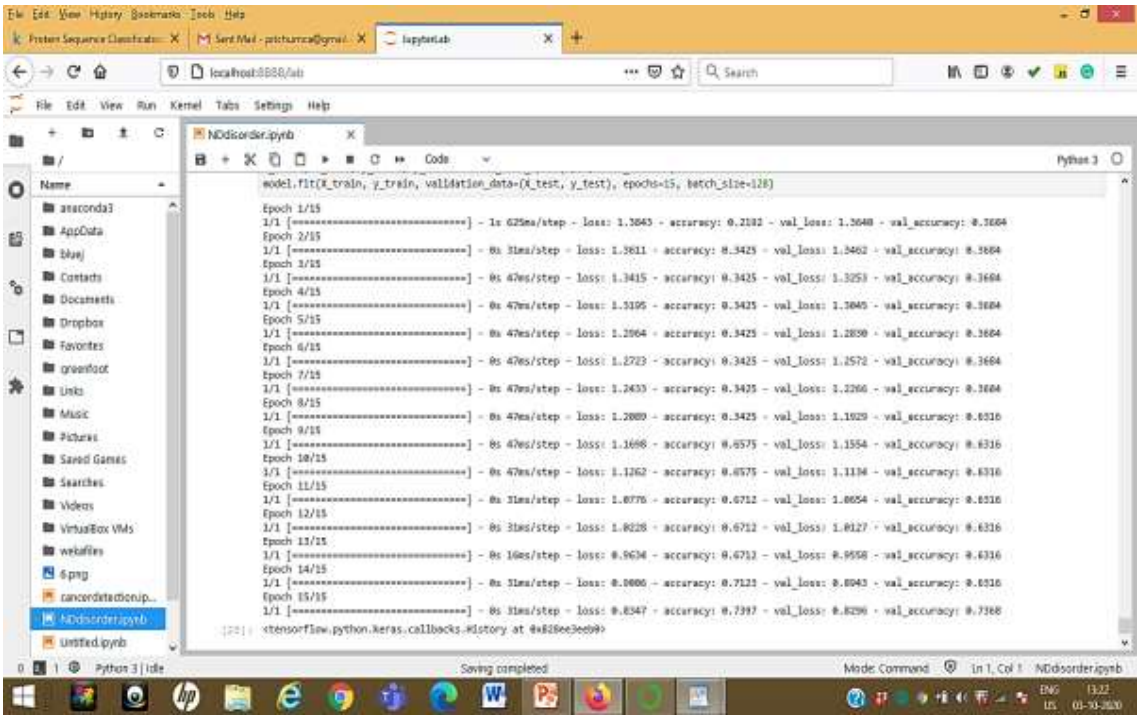


Figure 7: Number of Iterations in CNN is 25

```
train-acc = 1.0
test-acc = 0.9565217391304348
```

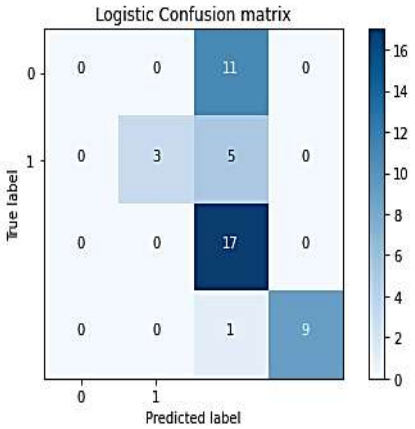


Figure 8: Confusion Matrix of CNN Classification technique

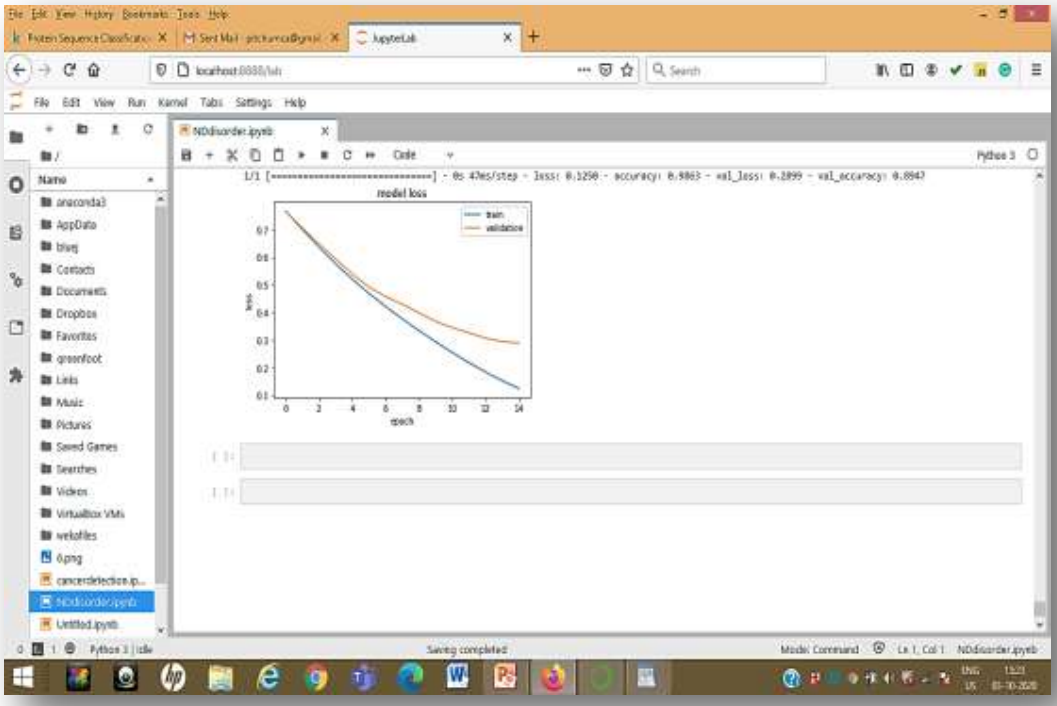


Figure 9: Error rate for CNN model

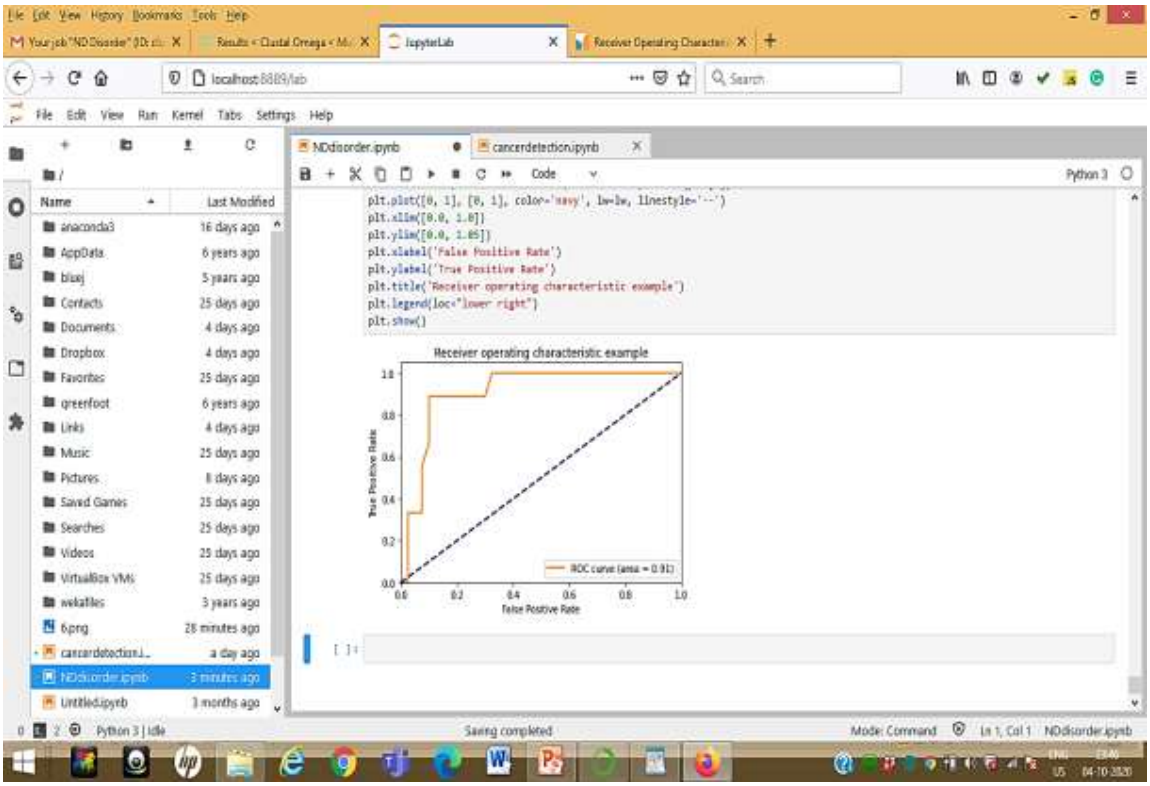


Figure 10: ROC Curve for the proposed CNN model

Confusion matrix is a technique to determine the performance of the classification algorithm. The confusion matrix clearly indicates that the classification of the class labels is very accurate with an overall accuracy of 95%.

Sequence Similarity analysis using Phylogenetic tree:

The evolutionary history was inferred using the UPGMA method. The optimal tree with the sum of branch length = 11.88472227 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of base differences per site. This analysis involved 95 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 39745 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

In the phylogenetic tree four major clusters were found. 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 .

Figure 11: Phylogenetic tree between ADHD Autism DMD and CP

Gene List Analysis and Visualization has been done using the following tools to find the similarity between the sequences

Bioinformatics Tools used for Gene correlation study:

- http://metascape.org for enrich
- VLAD: Gene List Analysis and Visualization
- MALACARD: Human Disease Database

Gene Correlation Analysis using Statistical Approach:

Vocalization Behaviour:

- -0.33386 : Negative Correlation. The relationship between vocalization behaviour and gene is very week. This gene is related to CNTNAP2,NLGN3,NLGN4X,NLGN4Y-Autism

Regulation of Membrane Potential:

- -0.25214: Negative Correlation. The relationship between regulation of membrane potential and gene is very week. The genes which has negative correlation are CD99L2(AUTISM),DAG1(DMD),KANK1(CP),MEF2C(AUTISM),PAX6(AUTISM),SPARC(DMD)

GO:0042391 & GO:0071625 :

- 0.5271801636: Positive Correlation with vocalization behaviour. The relationship between vocalization behaviour and regulation of membrane potential is moderate. This shows perfect positive correlation with regulation of membrane potential

Negative regulation of cell mobility:

- -0.2014: Negative correlation with gene. The relationship between cell mo and gene are very weak. DAG1(Muscular Dystrophy-),KANK1(CP),MEF2C(AUTISM,DMD,ADHD),SPOCK3(ADHD).

Cellular response to transform:

- -0.10325: Negative correlation with gene. The relationship between cell mo and gene are very weak. DUSP15(Autism),LTBP4(DMD),MEF2C(AUTISM,DMD,ADHD)

Total number of genes taken for study is 99 and out of 99 , 15 genes shows negative correlation. The above statistical interpretation was validated using enrichment analysis through metascape.org.

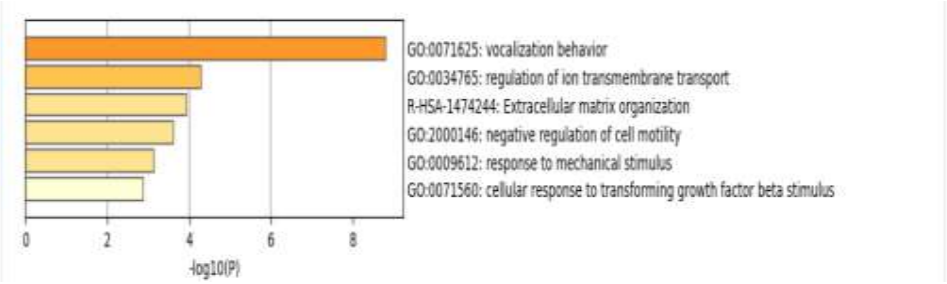


Figure 12(a): Heatmap for Statistical Interpretation

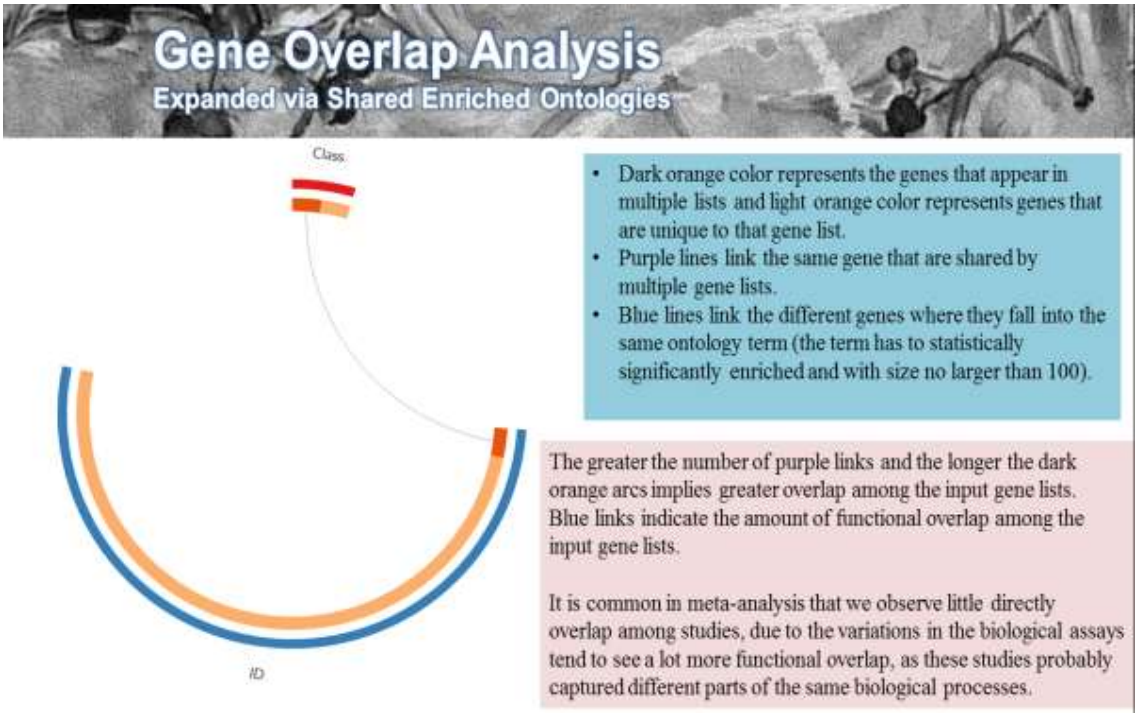
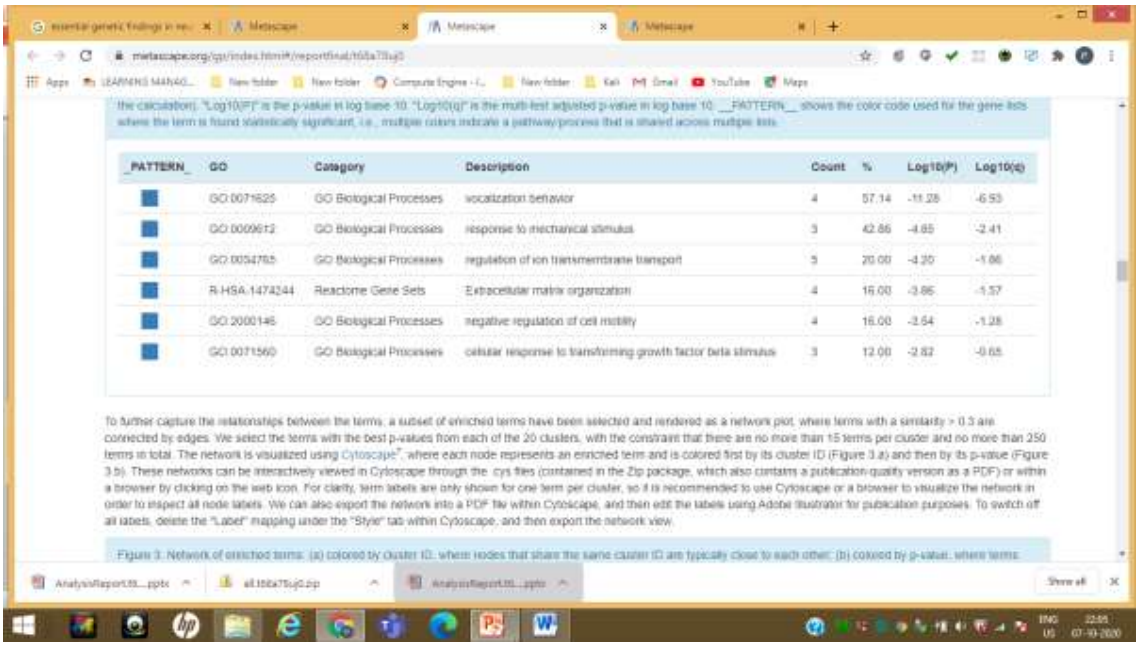


Figure 12(b) Gene Cluster analysis

The Heatmap obtained coincides with the statistical interpretation results. The heatmap cells are colored by their log p-values, white cells indicate the lack of enrichment for that term in the corresponding gene list.



PATTERN	GO	Category	Description	Count	%	Log10(P)	Log10(q)
	GO:0071625	GO Biological Processes	vocalization behavior	4	57.14	-11.28	-6.93
	GO:0009612	GO Biological Processes	response to mechanical stimulus	3	42.86	-4.65	-2.41
	GO:0034765	GO Biological Processes	regulation of ion transmembrane transport	5	20.00	-4.20	-1.86
	R-HSA-1474244	Reactome Gene Sets	Extracellular matrix organization	4	16.00	-3.86	-1.57
	GO:2000146	GO Biological Processes	negative regulation of cell motility	4	16.00	-2.54	-1.28
	GO:0071560	GO Biological Processes	cellular response to transforming growth factor beta stimulus	3	12.00	-2.82	-0.65

Figure 12(b): Enrichment Analysis output from metascape.org

Clustering of the genes is indicated in the figure below

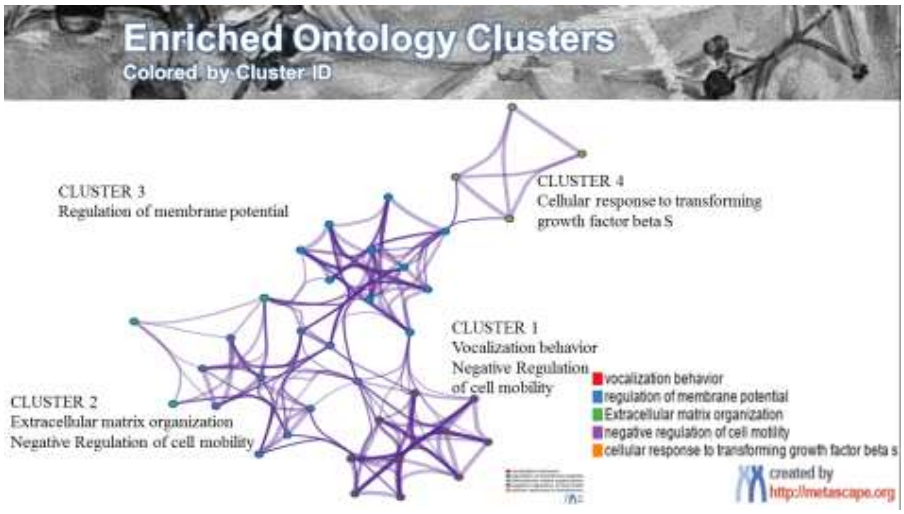


Figure 13: Clustering of the genes sequences under ADHD,DMD,Autism and CP through metascape.org

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 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 for the statistical interpretation:

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

Analysis of Amyloid Tau Immune related and Haemoglobin binding genes in association with Alzheimer disease

1. Introduction

Alzheimer's Dementia is a neuropsychiatric disorder prevalent over many decades among the elderly population. Off late it is reported in above middle age population also. Policy decisions are made in United Nations about the early diagnosis and treatment to prevent death due to Alzheimer's. More than nine hundred million people living worldwide are affected with Dementia (Prince et al. 2014). The number of people affected with Dementia is increase exponentially and expected to be above 80 million by 2030 and to 131 million by 2050 (World Alzheimer Report, 2015)

Many molecular and cellular changes take place in the brain of a person with Alzheimer's disease. These changes can be observed in brain tissue under the microscope after death. Investigations are underway to determine which changes may cause Alzheimer's and which may be a result of the disease. The beta-amyloid protein involved in Alzheimer's comes in several different molecular forms that collect between neurons. It is formed from the breakdown of a larger protein, called amyloid precursor protein. One form, beta-amyloid, is thought to be especially toxic. In the Alzheimer's disease brain, abnormal levels of this naturally occurring protein clump together to form plaques that collect between neurons and disrupt the cell function. Alzheimer's-related brain changes may result from a complex interplay among abnormal tau and beta-amyloid proteins and several other factors. It appears that abnormal tau accumulates in specific brain regions involved in memory.

From the literature it is evident that plaques are commonly present in the brain in the absence of any cognitive pathology. Hence it is necessary to analyse the variation of Aβ in Alzheimer patients and the healthy people. There are several receptors for Aβ on different cells but mainly on microglia, which are brain macrophages that can engulf Aβ and destroy it.

Objective of the study is to identify upregulated genes which involve in the MCI to AD conversion pathway.

2. Methodology

2.1 Dataset

Data is retrieved from Alzheimer's Disease Neuroimaging initiative. Data is classified under three categories as patients who are Cognitively Normal, patients with Mild Cognitive Impairment and Alzheimer's Dementia as given in Table 1. Number of MCI patients reported are higher than individuals with AD. Data from ADNI involves 5 classes of data namely

Early MCI, Late MCI, SMCI, AD and Cognitive Normal. The dataset has records with missing values and diagnosis for multiple visits. ADNI dataset belonging to Cohort1 contains multiple visit diagnosis with some missing data. Records are reviewed to include instances with the complete diagnosis.

Table 1. Details of control group under study

CLASS	MALE	FEMALE
AD	26	17
CN	115	135
MCI	260	184

2.2. Review of Genes analysed in the MCI to AD conversion pathway

Genes involved in the AD pathway is reviewed as given in Table 2, to analyse the varying expression levels in AD, MCI and CN. The variation in expression level among different classes is set as the basic differentiating threshold to analyse genes that contribute higher than the genes given in the literature.

Table 2. Gene review in AD pathway

S.No.	Analysis	Genes referred for Analysis from Existing literature
1.	Amyloid pathway	APOE, APP,PSEN1,PSEN2
2.	Complement system - body immune system	APCA7, CR1, CLU,CD2AP
3	Endocytosis Pathway	BIN1,PICALM,ADRB1,ADRB2,CTSS,PSK1,PAP2, MAPT,GSK3B,APOC1,CELF2,PVRL2,RFC3,TOM M40,TISC1,TTLL7,FAM113B,GAB2,PAX2
4	Genes associated with abnormal levels of haemoglobin	AHSP, HBG2, HBD, SPTB, FECH

Gene expression values are grouped into the three classes AD, MCI and CN. The average expression value of each gene for each class of AD, MCI and CN is found by implementing algorithm in R. Values lying above the average and that contribute significantly to the classification of the records into three classes were examined.

Figure 1 provides an overview of the pre-processing steps and identification of upregulated genes.

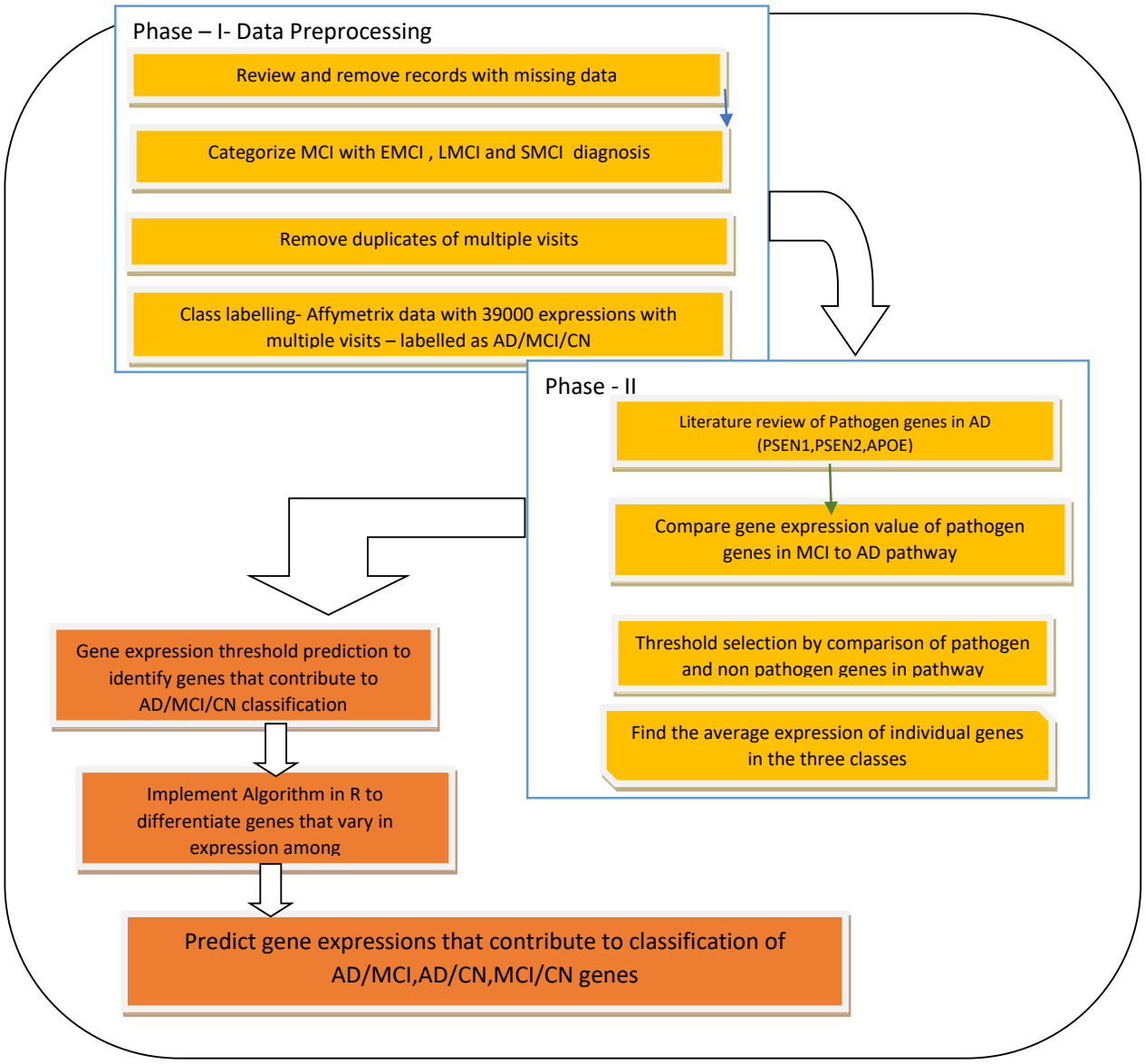


Figure 1. Overview of Methodology

More than thirty-nine thousand gene expression values are examined to identify the outstanding upregulated and down regulated gene expressions. Gene expressions are imported in R to calculate the average expression value under each class. Values that are specifically above threshold values for pathogen related genes are filtered for further analysis. Filtered genes are further analysed by calculating k-fold cross validation and classification to delineate the strongly differentiating the expression presented for AD/CN, AD/MCI and MCI/CN.

3. **Results and Discussion**

Genes that differentiated between AD/CN, AD/MCI and CN/MCI are tabulated in Table 3. The correlation between genetic data and neuropsychological data would throw more light in the early detection and treatment. Affymetrix gene expression data is analysed based on the average threshold value reported from literature. Existing dataset with CN, AD and MCI classes is run through the algorithm implemented in R to identify the expression values that lie well above the given threshold as tabulated in Table 3.

Table 3. Prediction of Genes with expression beyond the threshold for differentiating AD, MCI and CN

Threshold values and corresponding genes that exceed the threshold							
0.5 difference genes- mci - ad	0.5 diff - CN-AD	0.6 diff MCI-AD	0.6 diff - CN-AD	0.7 diff MCI-AD	0.7diff - CN-AD	0.8 diff MCI-AD	0.8diff - CN-AD
HBG2	HBG2	TUBB2A	IFI27	TUBB2A	TUBB2A	TUBB2A	TUBB2A
IFI27 Not yet associated with AD	IFI27	SNCA	TNS1	SNCA	SNCA	RPS4Y1	
SELENBP1	SELENBP1		AHSP	DDX3Y			
TNS1	TNS1		TUBB2A	KDM5D			
TNS1	TNS1		CA1				
TNS1	TNS1		SNCA				
SLC6A8	SLC6A8						
SLC6A8	KANK2						
KANK2	AHSP						
AHSP	TUBB2A						
TUBB2A	TUBB2A						

TUBB2A	MRC2						
GMPR	HBD						
PLEK2	SOX6						
ITLN1	C19ORF77						
KRT1	FECH						
HBD	FECH						
SOX6	XK						
C19ORF77	CA1						
FECH	CA1						
FECH	SPTB						
XK	SLC6A10P || SLC6A10PB*						
CA1	SNCA						
CA1	SNCA						
SPTB							
SPTB							
PAGE2 || PAGE2B							
SLC6A10P || SLC6A8 || SLC6A10PB*							
SLC6A8							
SNCA							
SNCA							

Genes that differentiated between AD/CN, AD/MCI and CN/MCI can be classified into the genes involved in the pre-processing pathway of beta amyloid precursor proteins, microtubule associated protein encoding genes, Proteins encoding genes which are involved in host immune defence against pathogen such as part of complement system and clathrin mediated endocytosis. It is shown from the differential expression pattern of the genes between AD, MCI and CN, there were genes which as highly upregulated and are moderately upregulated. These genes were characterized based on the differential expression levels which is above 0.8, 0.7, 0.6 and 0.5 between AD and CN as shown in Figures 2,3 and 4.

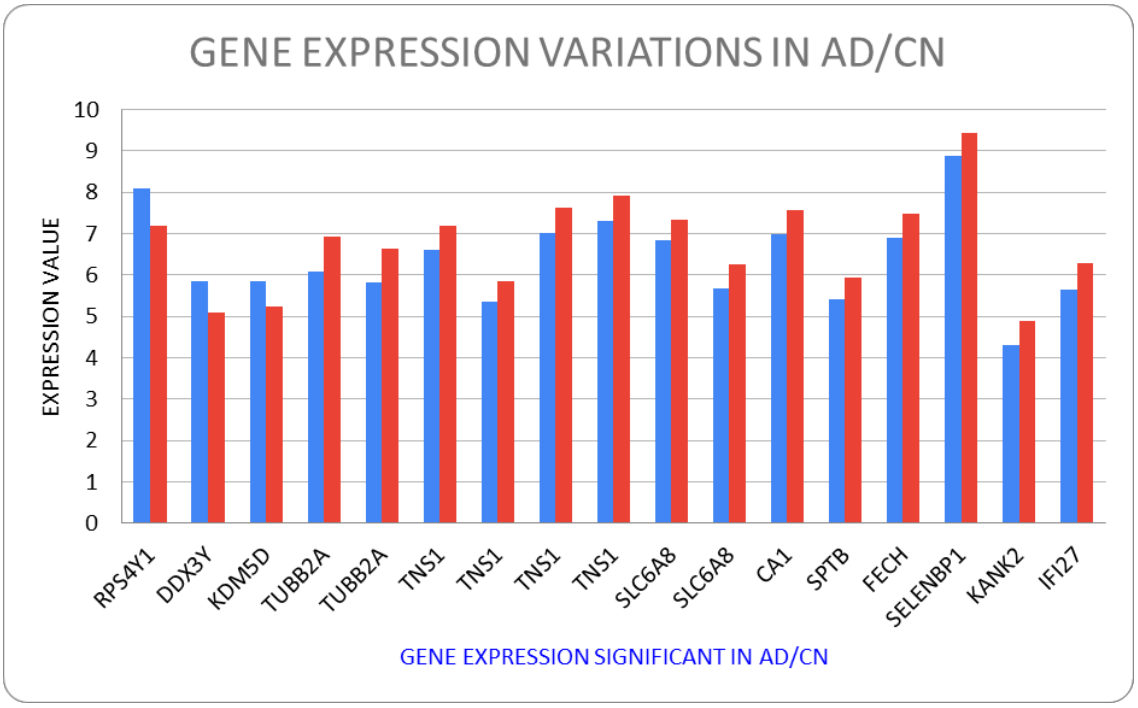


Figure 2. Comparison of Gene expression values beyond threshold among AD

Among the thirty nine thousand expressions with slight variations, the most prominent genes whose expression clearly demarcates the three classes are outlined. The findings add to the existing literature reporting on most common genes found in the pathogenic pathway of AD.

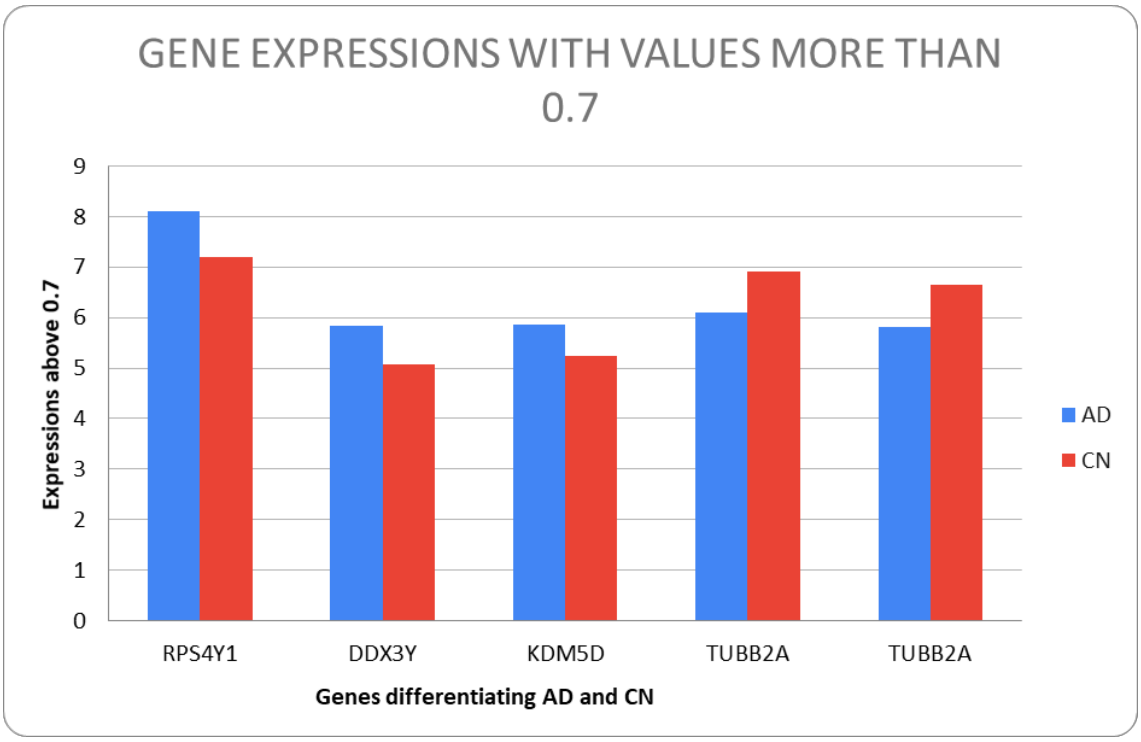


Figure 3. Comparison of Gene expression values beyond threshold among AD and CN

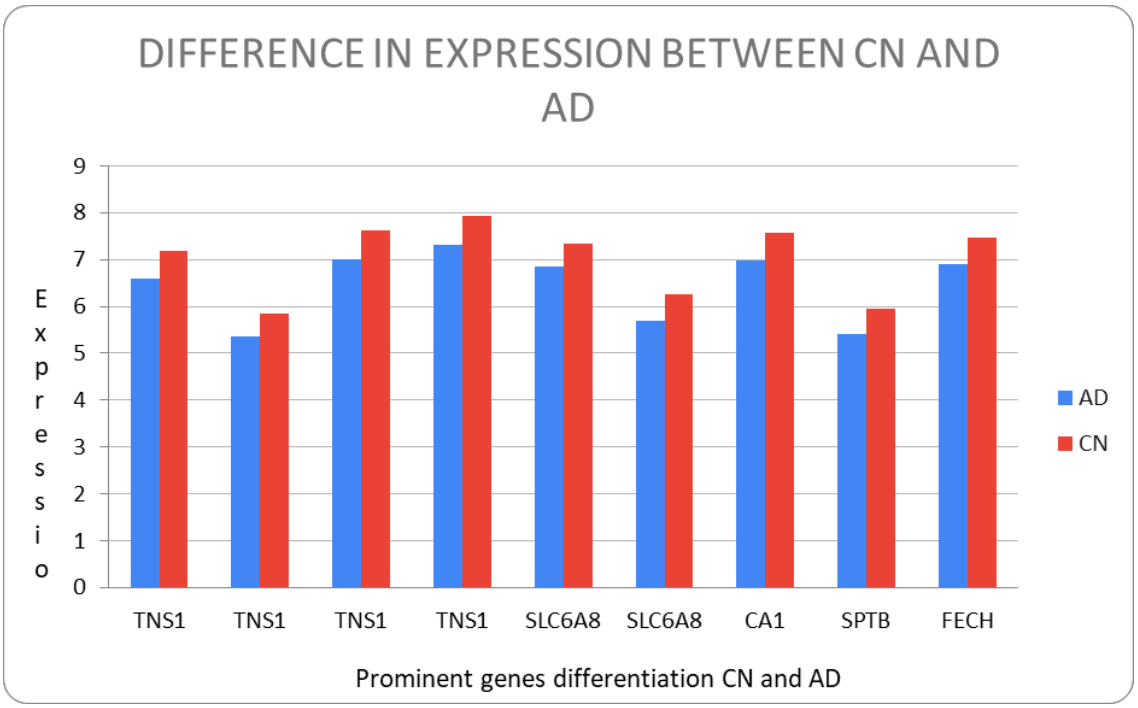


Figure 4. Comparison of Prominent Gene expression values beyond threshold among AD and CN

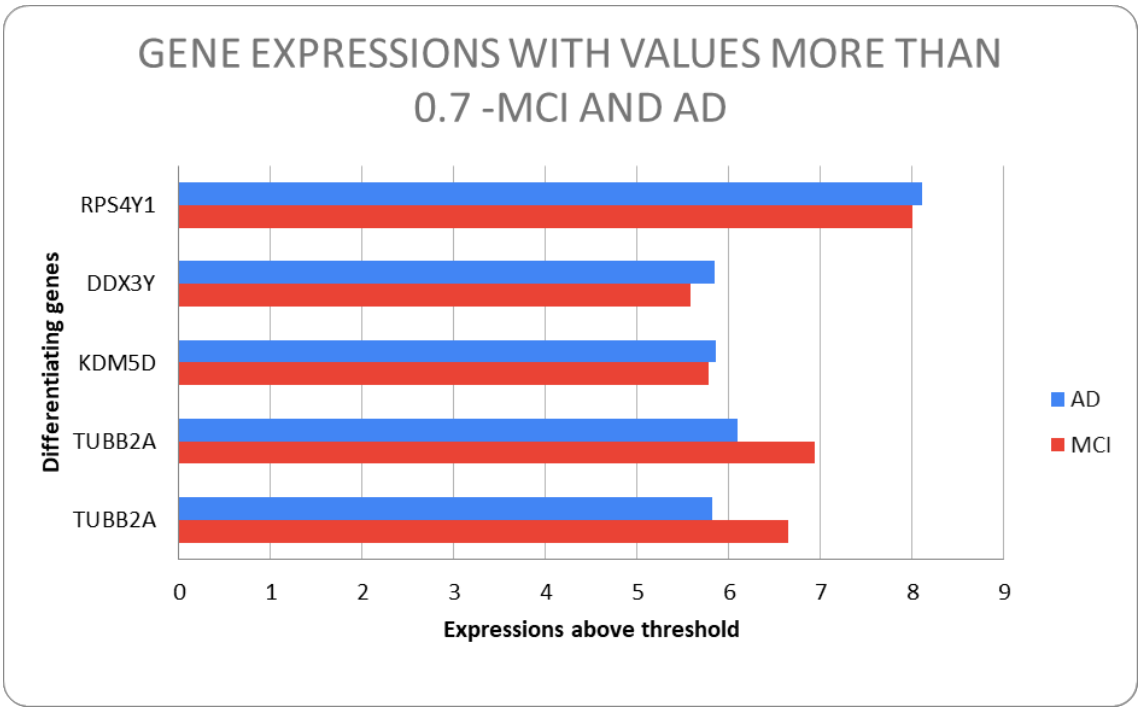


Figure 5. Comparison of Prominent Gene expression values beyond threshold among MCI and AD

The genes which are highly upregulated but more than 0.8 were regulated are directly related to the genes involved in the amyloid plaques formation. TUBB2A genes are directly involved in the amyloid plaque formation as reflected in Figure 5. The genes involved in the heme binding or hemoglobin binding are also upregulated significantly (0.5 fold).

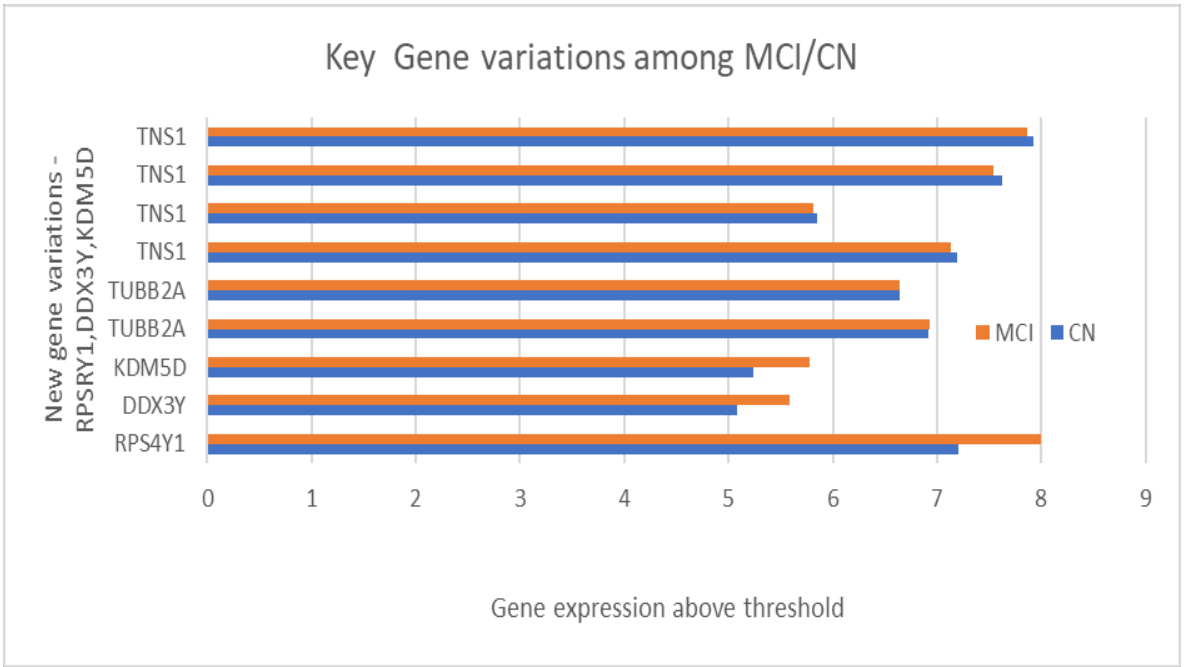


Figure 6. Gene expression values beyond threshold among AD and CN

In addition to that, the following genes were upregulated: RPS4Y1, KDM5D and DDX3Y which encodes for transcriptional activation factors to regulate the RNA level which are necessary during the deterioration of AD as shown in Figure 6. In addition to this the genes DDX3Y, RPS4Y1 and KDMD5 are linked with Y chromosomes. It may not have direct relations with the fact that AD is more prevalent in Female rather than male.

4. Conclusion

In the Alzheimer's brain, functioning of cell is disrupted by abnormal levels of this naturally occurring protein clumps. “APP, APOE, SNCA, PSEN1 and PSEN2, MAPT are directly related to Amyloid plaque pre-processing pathway or its degradation. The increase in the expression level of these genes confirmed its direct role in the development of AD and also the progression of the MCI to AD. AHSP, HBG2, HBD, SPTB, FECH are also shown an increase in the expression levels in the AD (reference Table 2), but not much difference between MCI to AD. This finding shows that Anaemia is closely related to the onset of AD.

This study reports that TUBB2A clearly demarcates the expression value between AD, CN and MCI apart from the genes reported in the literature. KDM5D, DDX3Y, RPS4Y1 also effectively segregate MCI and CN. The study could be further extended to identify gene that could identify conversion from MCI to AD.

5. References

World Alzheimer's Report (2020) Design, Dignity, Dementia: Dementia-related design and the built environment, https://www.alz.co.uk/u/WorldAlzheimerReport2020Vol1.pdf

World Alzheimer Report 2015: The Global Impact of Dementia,https://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf.

Prince. M., Albanese. E., Guerchet. M., and Prina. M. "Alzheimer Report. Dementia and Risk Reduction. An Analysis of Protective and Modifiable Factors", Alzheimer's Disease International, London, UK, 2014, http://www.alz.co.uk/research/WorldAlzheimerReport2014.pdf.

Lee, T., & Lee, H. (2020). prediction of Alzheimer's disease using blood gene expression data. *Scientific reports*, *10*(1), 1-13.

DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular neurodegeneration*, *14*(1), 1-18.

Ma, G., Liu, M., Du, K., Zhong, X., Gong, S., Jiao, L., & Wei, M. (2019). Differential Expression of mRNAs in the Brain Tissues of Patients with Alzheimer's Disease Based on GEO Expression Profile and Its Clinical Significance. *BioMed research international*, *2019*.

Fenoglio, C., Scarpini, E., Serpente, M., & Galimberti, D. (2018). Role of genetics and epigenetics in the pathogenesis of Alzheimer's disease and frontotemporal dementia. *Journal of Alzheimer's Disease*, *62*(3), 913-932.

Carter, C. (2011). Alzheimer's disease: APP, gamma secretase, APOE, CLU, CR1, PICALM, ABCA7, BIN1, CD2AP, CD33, EPHA1, and MS4A2, and their relationships with herpes simplex, C. pneumoniae, other suspect pathogens, and the immune system. *International Journal of Alzheimer's Disease*, *2011*.