**Research Log**

**08/01/2024**

* Completed the playlist - https://www.youtube.com/watch?v=VGUOQV8Eyfg&list=PL6xNS6KCt75ccBh6dB1mkSr3V69OP7p6B

**09/01/2024**

* Completed 18 lectures of playlist - <https://www.youtube.com/watch?v=pid0lUH467o&list=PLE6Wd9FR--Ecf_5nCbnSQMHqORpiChfJf> ,
* Wrote a code for image processing : /home/ibab/SEM\_4/image\_processing1.py

**10/01/2024**

* Revised the above playlist
* Wrote a code for text processing : /home/ibab/SEM\_4/text\_processing.py

**11/01/2024**

* Watched video on neural network - Gradient descent, how neural networks learn | Chapter 2, Deep learning
* Completed- <https://www.youtube.com/watch?v=YtvP5A5OHpU>
* Completed-<https://www.spectrumnews.org/features/multimedia/webinars/webinar-shafali-jeste-discusses-brain-development-high-risk-infants/>

**12/01/2024**

* Completed reading paper - <http://pediatrics.aappublications.org/content/136/Supplement_1/S10>

**13/01/2024**

* Completed reading paper - <http://pediatrics.aappublications.org/content/pediatrics/136/Supplement_1/S60.full.pdf>
* Completed reading review paper - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5915621/>

**14/01/2024**

**HOLIDAY**

**15/01/2024**

* Started with Deep Learning playlist - <https://www.youtube.com/watch?v=CS4cs9xVecg&list=PLpFsSf5Dm-pd5d3rjNtIXUHT-v7bdaEIe>

**16/01/2024**

* Learned about many libraries in python which are to be used in ML

**17/01/2024**

* Wrote a code on classification model and regression model and uploaded it on github.

**18/01/2024**

* A/C alt allele

**19/01/2024**

* Read papers on autism
* Tried working on sample vcf files.

**20/01/2024**

* Read papers on deep learning
* Completed code on sample vcf files.

**21/01/2024**

**SUNDAY**

**22/01/2024**

* Read paper on autism and whole genome sequencing
* Read paper on “ Application of deep learning algorithm on whole genome sequencing data uncovers structural variants”

**23/01/2024**

* Read on data standardisation and K-cross fold validation and wrote the codes as well.

**24/01/2024**

* Continued reading papers on autism and machine learning.

**25/01/2024**

* Tried working on sample vcf files for extracting structural variants.

26/01/2024

**HOLIDAY**

January 29th - February 4th, 2024

•Downloaded a few data files and worked on them.

•Annotated one pvcf file using snpEff.

•Annotated few gcvf file using snpEff

•Comparing the output of gvcf and pvcf.

•Made changes and validated the Annotated output.

•Downloaded snpEff on the server and worked on data on snpEff.

•Read papers on autism and machine learning

•Tried validating the Annotated output.

•Read about VEP and look at its output.

•Worked on previously written logistic regression code.

**05/05/2024**

* Fever and cold

**06/05/2024**

* Copied few gvcf files from server to my system to work on
* Wrote a code to filter out read based on quality and DP.
* Annotated it using snpEff
* Filtered out reads which had variants present in genes related to ASD.
* Made a tsv file after taking features(chr num, chr pos, variant type,gene name)
* Encoded the columns(chr num, variant type,gene name) and made a final tsv

**07/02/2024**

* Read about deep learning and embedding vector

**08/02/2024**

* Studies about vector flattening

**09/02/2024 - 11/02/2024**

* Tried reading about deep learning and CNN
* Read about word2vec and embedding vectors
* Searched papers on machine learning in biology
* Studies about vector flattening
* Worked on gvcf files to convert the output to vector
* Implemented vector flattening on output file

**13/02/2024-**

* Made a final matrix which had all genes name, variant type from 20 files to put in machine learning model
* Wrote a code to extract information from vcf files
* Wrote a code to count the genes and variant type
* Read about RNN, CNN, word2vec
* Took folder number 0 of gvcf and run the pipeline,filtered them,annotate them,extracted info,count the variant type and gene names,ran the final classification model
* Accuracy was 61%
* Read about PyTorch and deep learning

**18/03/2024**

* **Visualised variant distribution** with a barplot.
* From SFARI dataset, I extracted only genes with "high confidence" score.
* **Identified high-confidence genes** from gVCF files and extracted their counts.
* **Processed genes** using previous scripts .
* **Performed K-Fold Random Forest** analysis on the matrix (only high confidence gene) and accuracy was lower than previous results.
* **Extracted sequences** of ASD-related genes for clustering.

**26/03/2024**

* Reduced the number of genes and again ran the entire machine learning pipeline.
* I then incorporated variant count, categorised by chromosome, as a new feature within the data matrix. This aimed to provide the model with additional context.
* I explored using a FeedForward Neural Network (FFN) architecture for the machine learning task. However, the initial implementation led to a decrease in accuracy.
* To address potential data quality issues, I removed features that contained a high proportion of missing values (NaN).
* Tried clustering variant type based on their count using K Means.

**02/04/2024**

* Read about hierarchical clustering to implement on my dataset.
* Processed my dataset through a feedforward neural network.
* Read more about feedforward NN
* Studied about SHAP analyses

**23/04/2024**

* Wrote abstract for IISc Big Data Biology symposium
* Tried SHAP plot for models
* Learned about transformers
* Read about CNN

**29/04/2024**

* Worked on hierarchical clustering
* Did SHAP analysis for other models
* Tried making heatmap for genes
* Continued working on transformers
* Learned about CNN for my model

**12/05/2024**

* Working on writing thesis - introduction, methods and results
* Analysed SHAP results
* Did Pathway Analysis