Classification of Mice on Protein Expression Level

-Team E

Team Leads

Vaibhavi	Lead
Nandita	Co-Lead
Akshar	Co-Lead

Extended Members

Aachal Gupta

Aarjav Jain

Pabitra Goswami

Saili Ashok Dhuri

Pratham Jindal

Ashmit Zanzote

Shubham Arvind Rangari

Abeed Mohammed

Ketan Kumar Sandre

Juweriya Shayhmeeh

Kunal Khumbhar

Sudarshan Sopane

Table of contents

Introduction

Our project, "Classifying Mice Based on Protein Expression Levels," explores the relationship between protein expression, genetics, behavior, and treatments in mice. Using machine learning, we developed a model to classify mice into eight groups based on 77 proteins. This research provides insights into how Down syndrome affects learning and memory and evaluates the potential benefits of the drug memantine for trisomic mice.

Steps Involved

- A Data Preprocessing
- B Exploratory Data Analysis
- C Feature Selection
- D Model Training
- E Model Evaluation
- F Interpretation & Analysis

Data Preprocessing

Handling Missing Values

Encoding Categorical Variables

Data Normalisation

EDA

Data Distribution

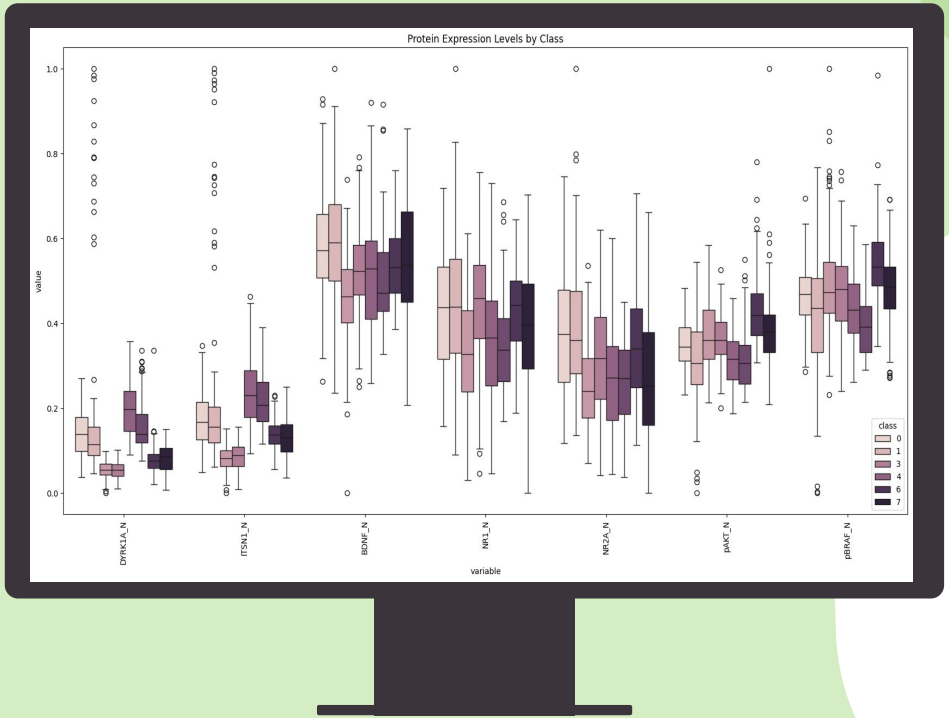
Distribution of protein expression levels across different classes.

Class Balance

Examination of the distribution of samples across different classes to check for class imbalance.

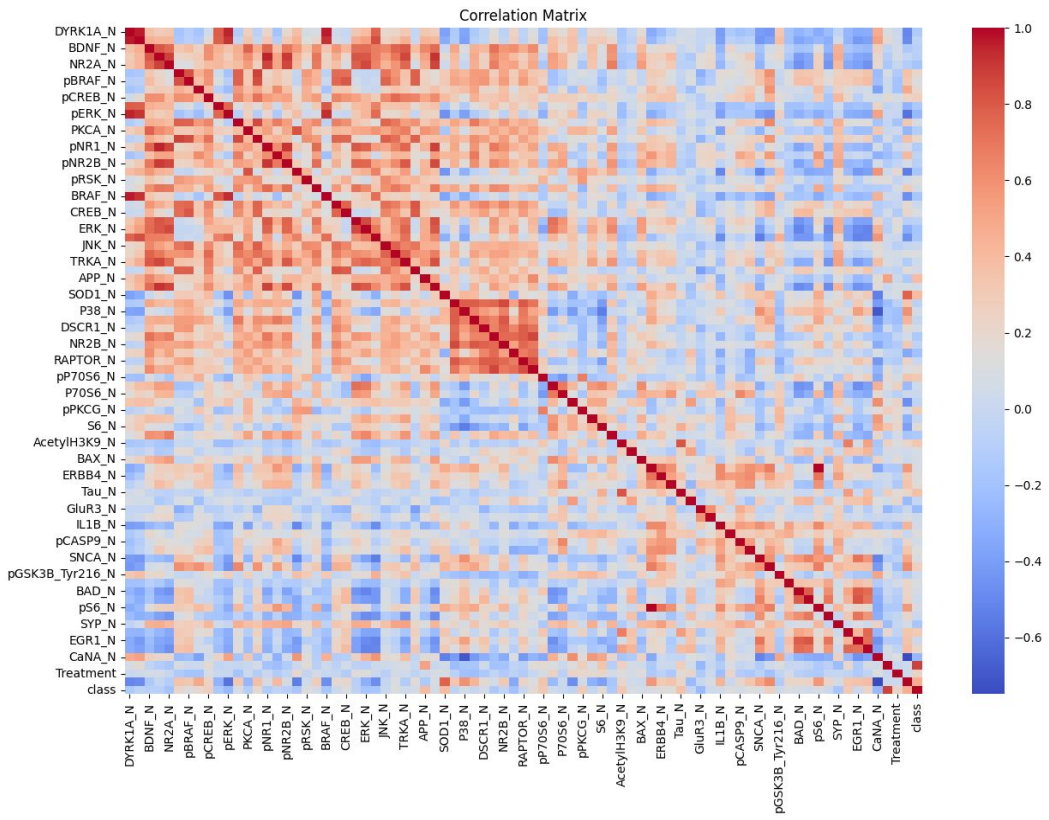
Missing Values

Identification and proportion of missing values in the dataset.



Box Plots

Created to visualise the distribution and variance of protein expression levels across different classes.

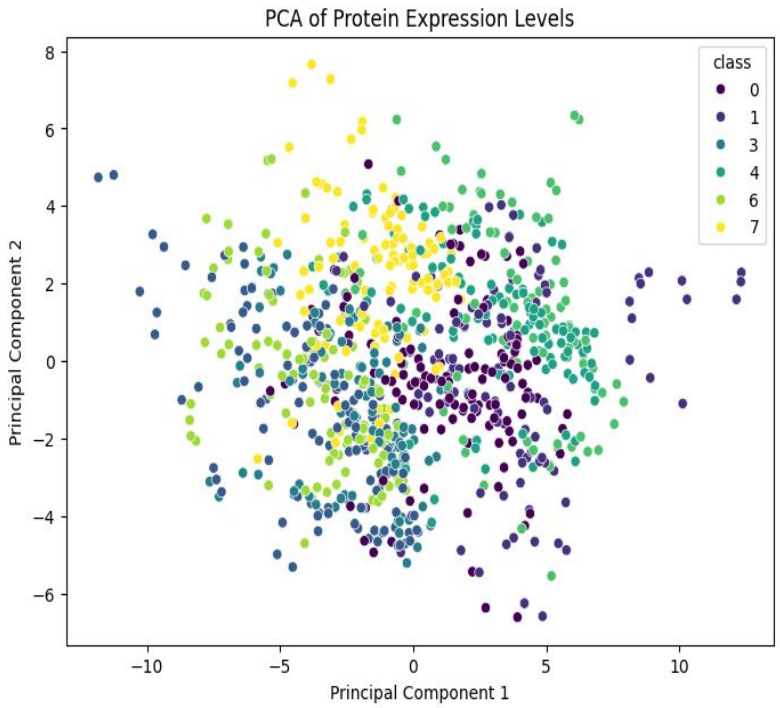

Correlation Matrix
DYRK1A_N
BDNF_N
NR2A_N
pBRAF_N
pCREB_N
pERK_N
PKCA_N
pNR1_N
pNR2B_N
pRSK_N
BRAF_N
CREB_N
ERK_N
JNK_N
TRKA_N
APP_N
SOD1_N
P38_N
DSCR1_N
NR2B_N
RAPTOR_N
pP70S6_N
P70S6_N
pPKCG_N
S6_N
AcetylH3K9_N
BAX_N
ERBB4_N
Tau_N
GluR3_N
IL1B_N
pCASP9_N
SNCA_N
pGSK3B_Tyr216_N
BAD_N
pS6_N
SYP_N
EGR1_N
CaNA_N
Treatment
class
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Feature Selection

Techniques Used

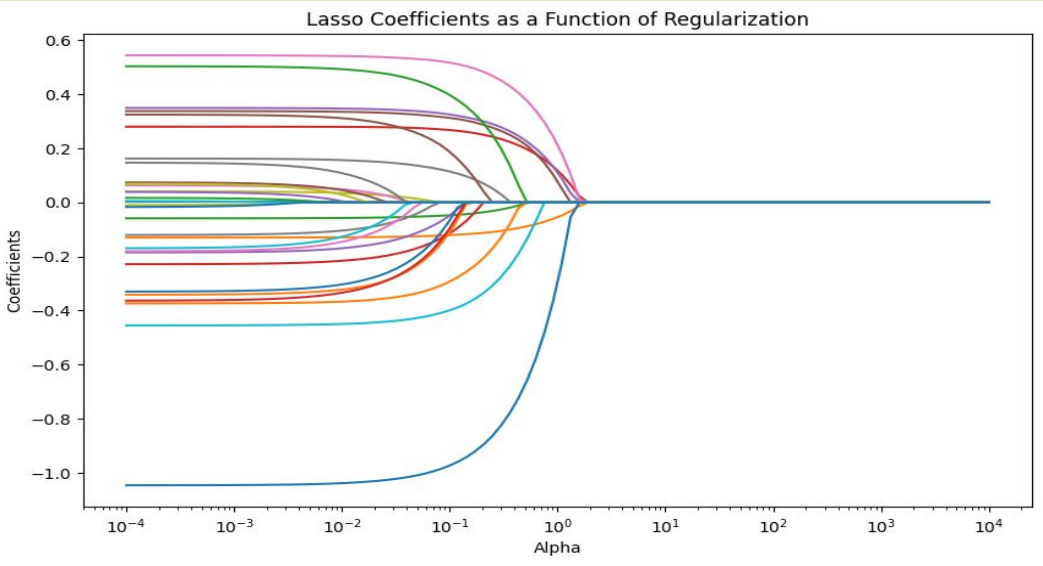
- **A** Correlation Analysis
- **B** Feature Importance from Models (Random Forest)
- **C** Principal Component Analysis

Feature Selection



PCA reduces the complexity of the dataset while retaining most of the variance.
Tight, distinct clusters indicate good separability, while overlapping clusters suggest that the classes are not well separated by the chosen features.

Feature Selection



Use of Lasso

The Lasso coefficients plot shows how different feature coefficients shrink to zero as the regularisation parameter (alpha) increases, many coefficients shrink to zero, indicating that Lasso regression effectively reduces the number of features by retaining only the most significant ones.

A subset of key proteins was identified as highly discriminative between control and trisomic samples.

Model Training

Techniques Used

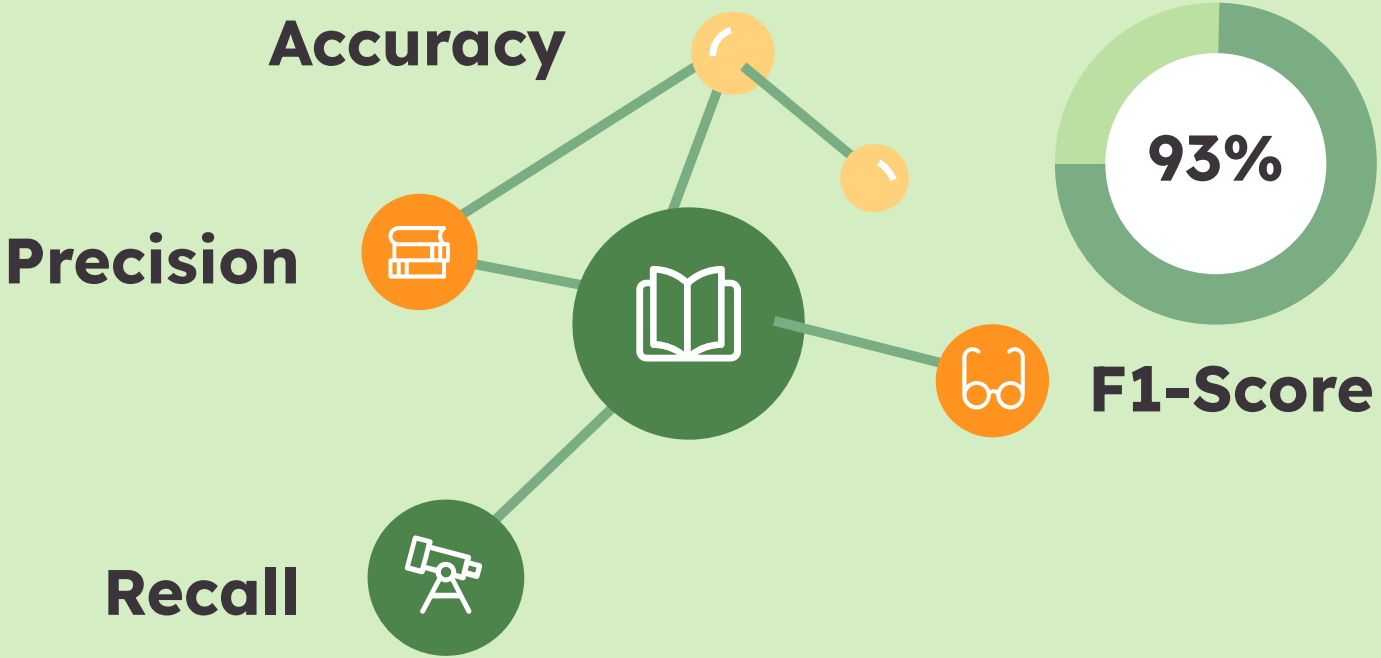
- A Grid Search
- B Parameters Tuned
- C Validation

What we used??

Random Forest: An ensemble method that constructs multiple decision trees and outputs the mode of the classes for classification.

HyperParameter Tuning: An exhaustive search method that tests all possible combinations of predefined hyperparameters to find the best-performing model.

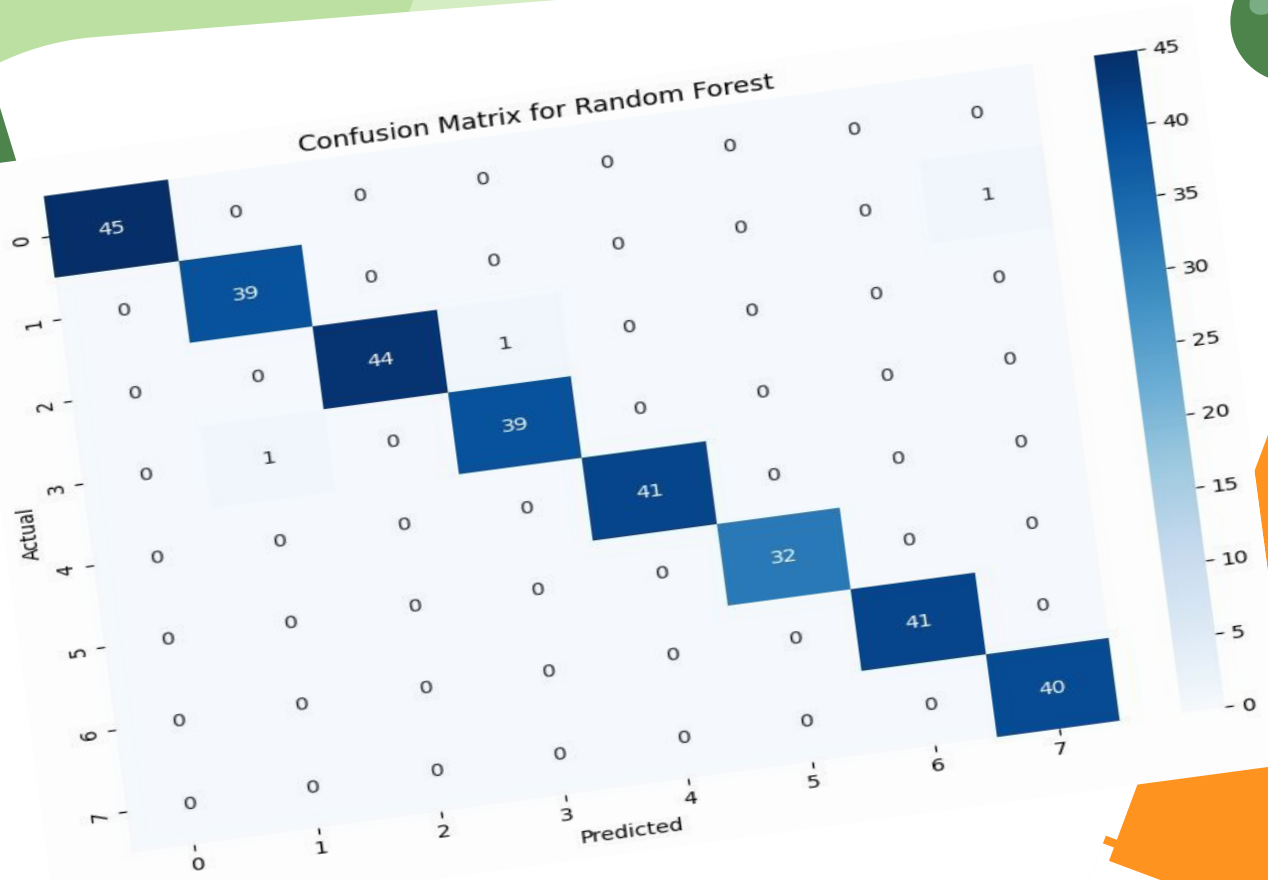
Model Evaluation

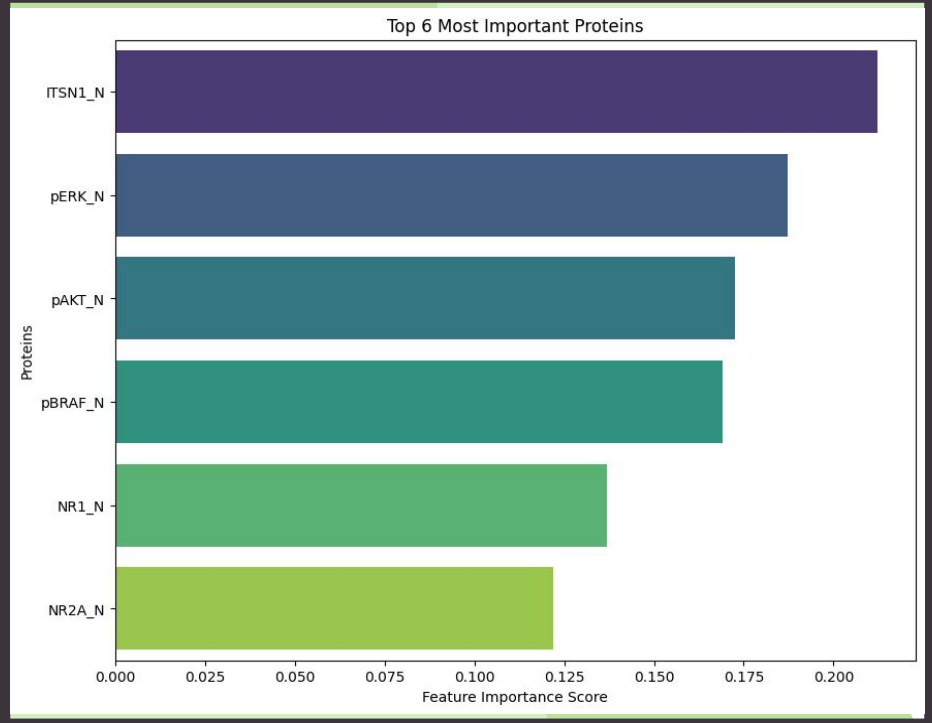


Here’s a table

Classification Report for Reduced Random Forest Model:

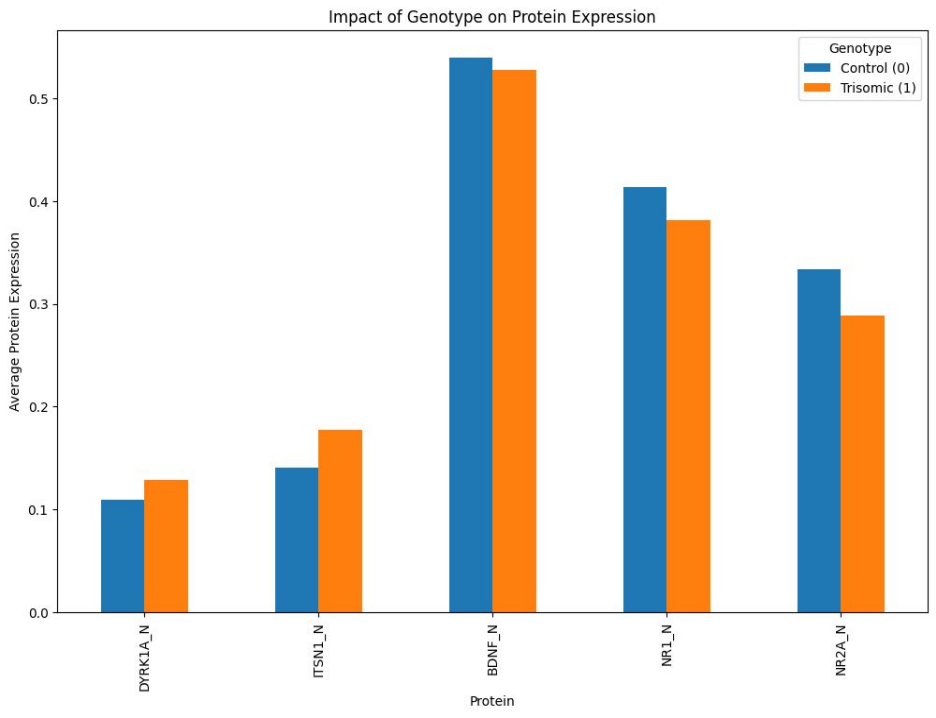
	precision	recall	f1-score	support
0	0.95	1.00	0.98	21
1	0.96	0.87	0.92	31
2	0.97	0.97	0.97	32
3	0.81	0.96	0.88	27
4	0.96	0.88	0.92	25
5	1.00	0.85	0.92	20
6	0.90	0.88	0.89	32
7	0.90	1.00	0.95	28
accuracy			0.93	216
macro avg	0.93	0.93	0.93	216
weighted avg	0.93	0.93	0.93	216


Confusion Matrix for Random Forest
Actual
Predicted
0 1 2 3 4 5 6 7
45 0 0 0 0 0 0 0
0 39 0 0 0 0 0 1
0 0 44 1 0 0 0 0
0 1 0 39 0 0 0 0
0 0 0 0 41 0 0 0
0 0 0 0 0 32 0 0
0 0 0 0 0 0 41 0
0 0 0 0 0 0 0 40
45 40 35 30 25 20 15 10 5 0



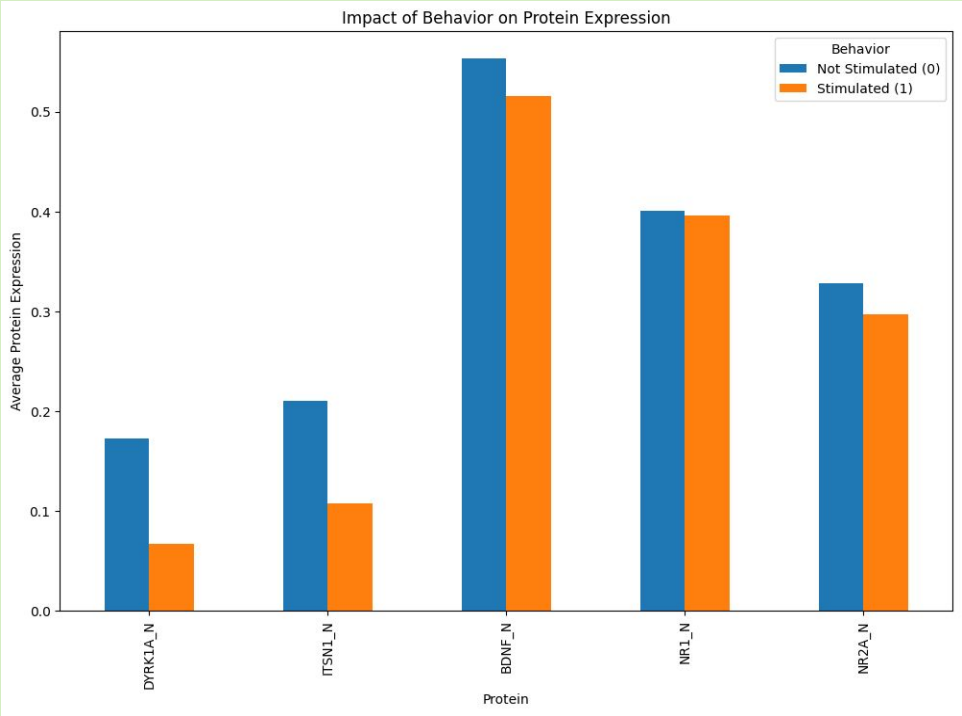
Top 6 Most Important Proteins in Mice Classification

Interpretation and Analysis



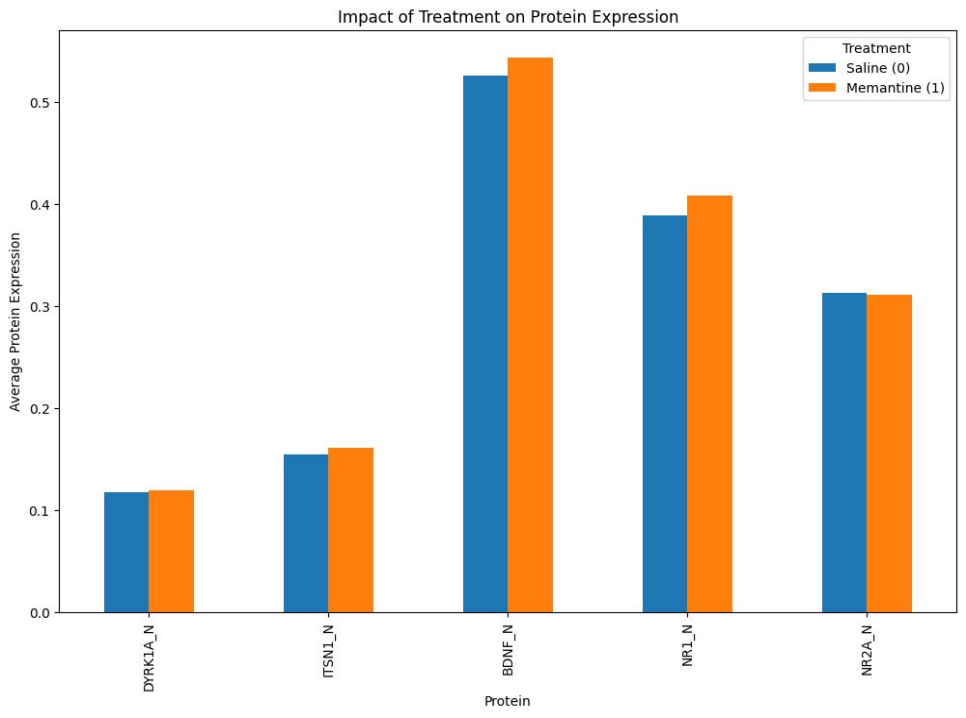
- Trisomic mice show higher levels of ITSN1_N and DYRK1A_N, while control mice have higher BDNF_N, NR1_N, and NR2A_N. These protein differences may be linked to the cognitive impairments in Down syndrome.

Interpretation and Analysis



- Unstimulated mice show higher levels of DYRK1A_N, ITSN1_N, BDNF_N, and NR2A_N compared to stimulated mice. This suggests that stimulation affects these proteins, potentially influencing associative learning mechanisms.

Interpretation and Analysis



Memantine treatment does not drastically alter most protein expression levels compared to saline. However, slight increases in BDNF_N and NR1_N in the memantine group suggest subtle effects on these proteins, possibly related to its therapeutic effects.

Challenges we faced....

Overfitting

High-dimensional data risks overfitting; regularization and cross-validation are essential.

Missing Data

Technical issues may lead to missing protein values, requiring effective handling techniques.

Biological Variability

Natural variations in protein expression among mice can complicate the identification of subtle class differences.

Conclusion

The study identified key proteins, such as DYRK1A_N, ITSN1_N, and BDNF_N, that differentiate between control and trisomic mice. However, the findings are limited by sample size and biological complexity, suggesting the need for further validation with larger datasets and additional experiments. Future research should focus on confirming these results in broader cohorts, exploring underlying biological mechanisms, and integrating other omics data for a more comprehensive understanding of Down syndrome.

Thank you!

Does anyone have any questions?