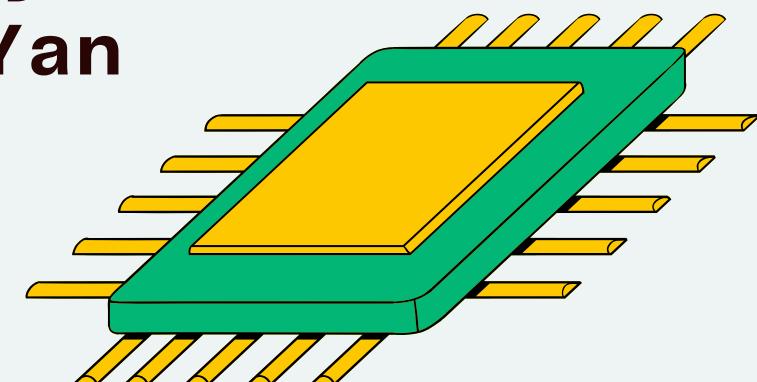


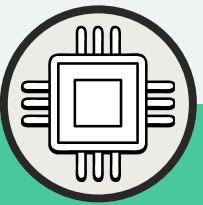
# **PREDICTION OF AML SUBTYPES THROUGH SUPERVISED MACHINE LEARNING**

**PRESENTED BY: GROUP H**

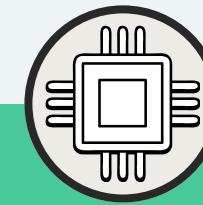
**Group Member: Tao Shengmin, Keerthana Kathavarayan,  
Cui Minghang, Liu Lihui, Li Zhuofei, Elizabeth Ng Jia Yan**



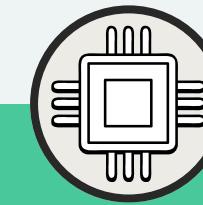
# TABLE OF CONTENTS



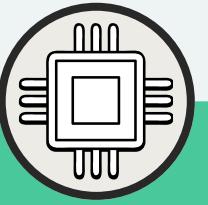
INTRODUCTION



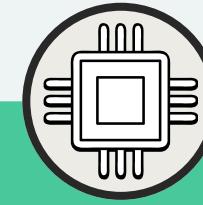
DATA OVERVIEW



METHODOLOGY



RESULTS & DISCUSSION



CONCLUSION

# INTRODUCTION

## ACUTE MYELOID LEUKAEMIA (AML)

- Cancer of **blood** and **bone marrow**.
- Overproduction of immature white blood cells (**myeloblasts**).
- Various subtypes-> critical for treatment plans.

## FAB CLASSIFICATION

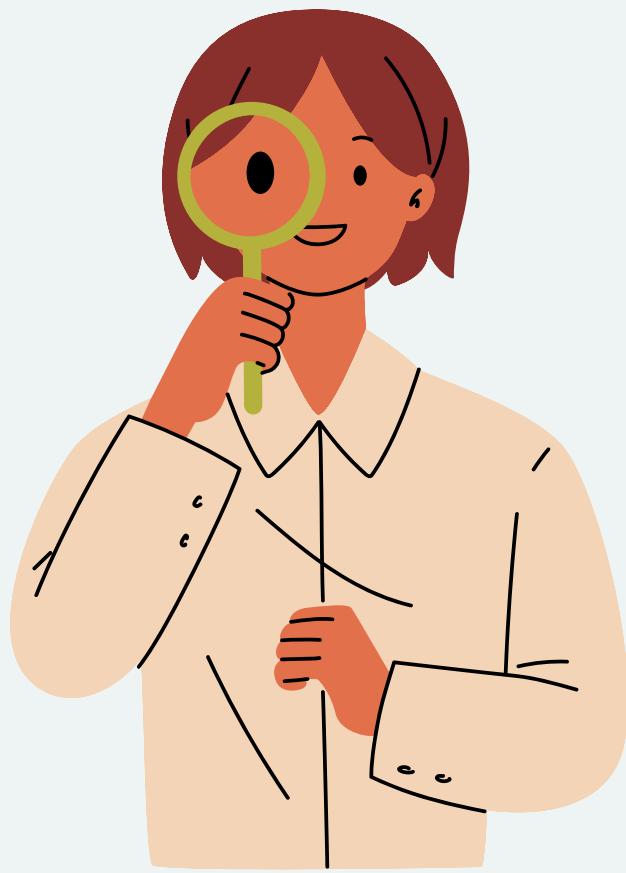
- **French-American-British Classification:**  
8 subtypes (**M0-M7**) based on cell morphology.
- M2, M3, and M4 most common.

FAB subtype	Name
M0	Undifferentiated acute myeloblastic leukemia
M1	Acute myeloblastic leukemia with minimal maturation
M2	Acute myeloblastic leukemia with maturation
M3	Acute promyelocytic leukemia (APL)
M4	Acute myelomonocytic leukemia
M4	Acute myelomonocytic leukemia with eosinophilia
M5	Acute monocytic leukemia
M6	Acute erythroid leukemia
M7	Acute megakaryoblastic leukemia

# GENETIC COMPLEXITY

- **Known Mutations:** Includes fusion genes RUNX1/CBFA2T1, CEBPA mutations, etc.
- **Research Challenges:** High heterogeneity complicates marker identification.
- **Importance of Genetics:** Essential for targeted therapies and understanding disease mechanisms.

## DIAGNOSTIC METHODS



- **Current Approaches:** Imaging (X-rays), Microscopy, Cytogenetics, and PCR-based genetic tests.
- **Limitations:** Labour intensive, insufficient genetic markers, for precise subtype classification and developing treatment plans.

# PROJECT BASIS



## Objective:

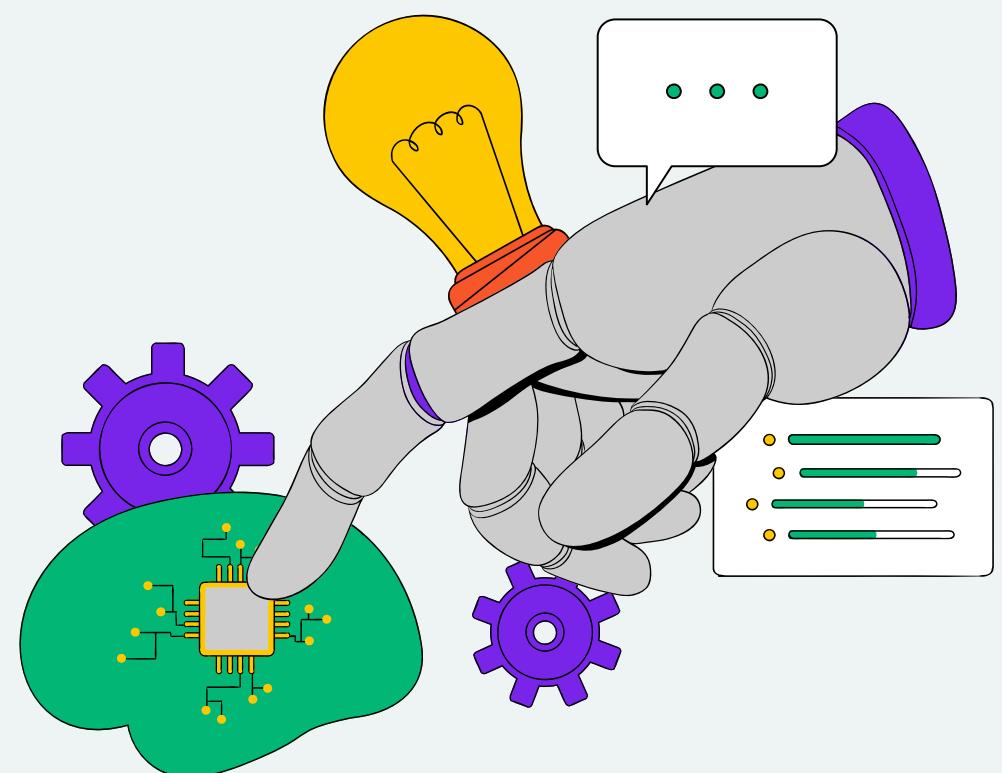
To enhance FAB classification with gene expression data for precise subtype identification.

## Method:

Gene expression patterns analyzed as features for prediction of AML subtypes through Machine Learning models.

## Future Impact:

- Improve diagnostic accuracy
- Enable efficient prognosis & treatment
- Advance precision oncology



# DATASET OVERVIEW

## Database:

- Gene Expression Omnibus (GEO)

## Microarray data:

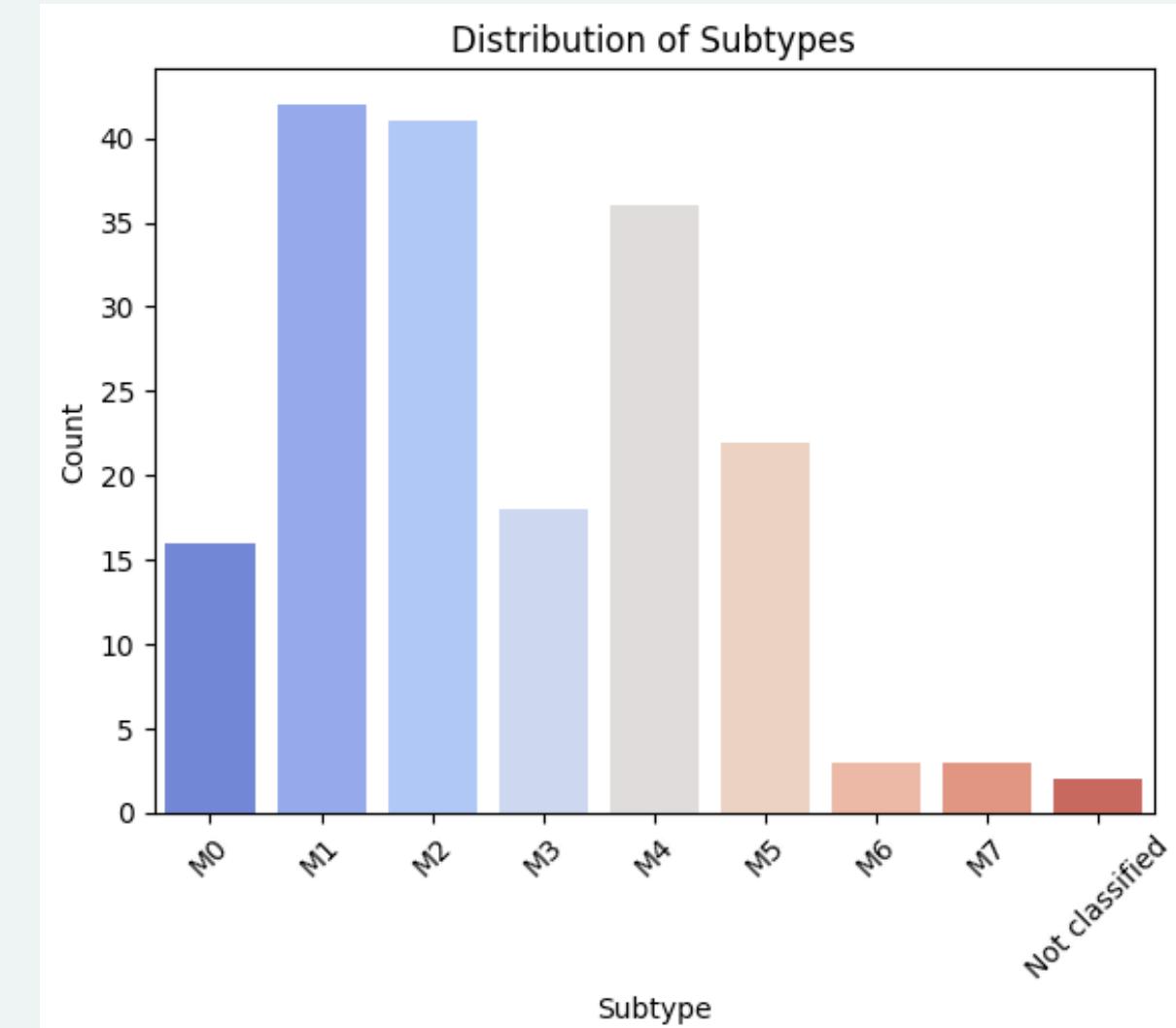
- Affymetrix HG-U133 Plus 2.0
- 16384 Probe IDs
- M0 – M7 AML Subtypes + Not Classified

## Samples:

- Whole Blood Tissue
- 183 AML Patients



**183 x 16384**



# METHODOLOGY

## DATA PRE-PROCESSING

- Cleaning
- Annotation
- Normalization
- Feature selection

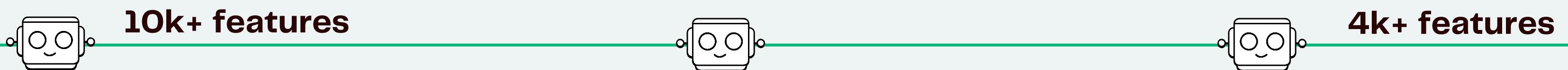
## MODEL DEVELOPMENT

- Classification Model
- Best-K hyper-parameter fine-tuning
  - Gridsearch
  - Cross Validation
  - Kfold
- CNNs model architecture to further explore

## MODEL EVALUATION

- Metrics
  - Accuracy
  - Weighted-Average F1 score
- Visualisation
  - Confusion Matrix
  - Accuracy Barchart
  - Validation lineplot

# DATA PRE-PROCESSING



## TRANSFORMATION

- Annotated Gene symbols
- Removed Probe IDs with no match
- Dropped Classes outside of FAB framework.

## NORMALISATION

- Log-Transformations

## STATS

- Anova threshold  
 $p\text{-value} < 0.05$

# FEATURE SELECTION

## Fold change

- Threshold → optimal features for downstream

## Random-Forest

- n\_estimators → no. of trees in the forest

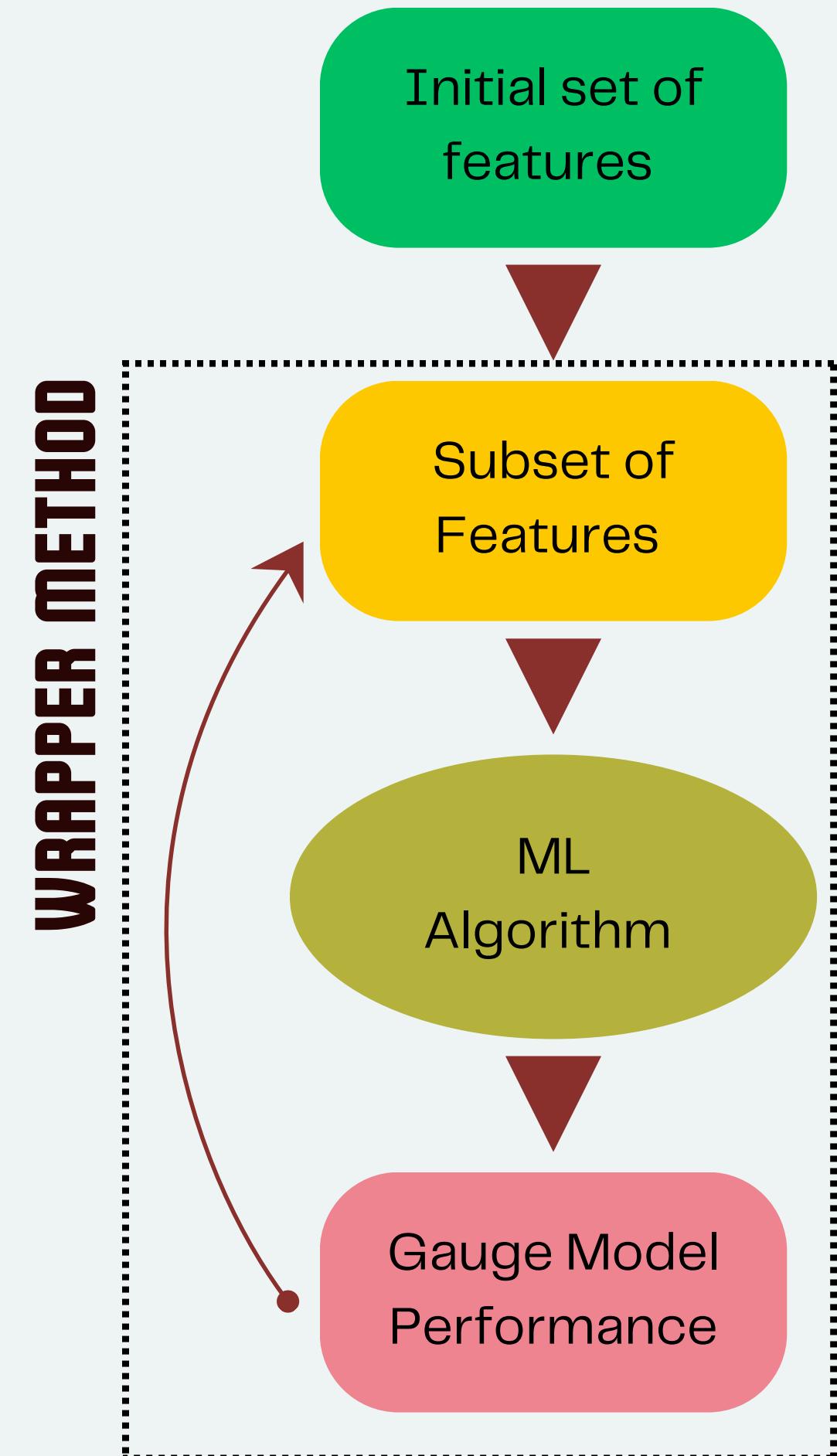
## Ridge Regularisation (L2)

- 'penalty' → penalise the  $\lambda$  coefficient

## Recursive Feature Elimination with Cross-Validation

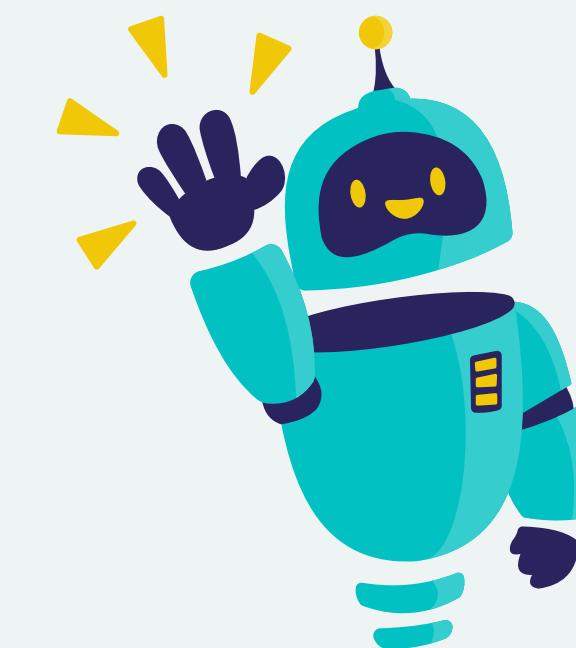
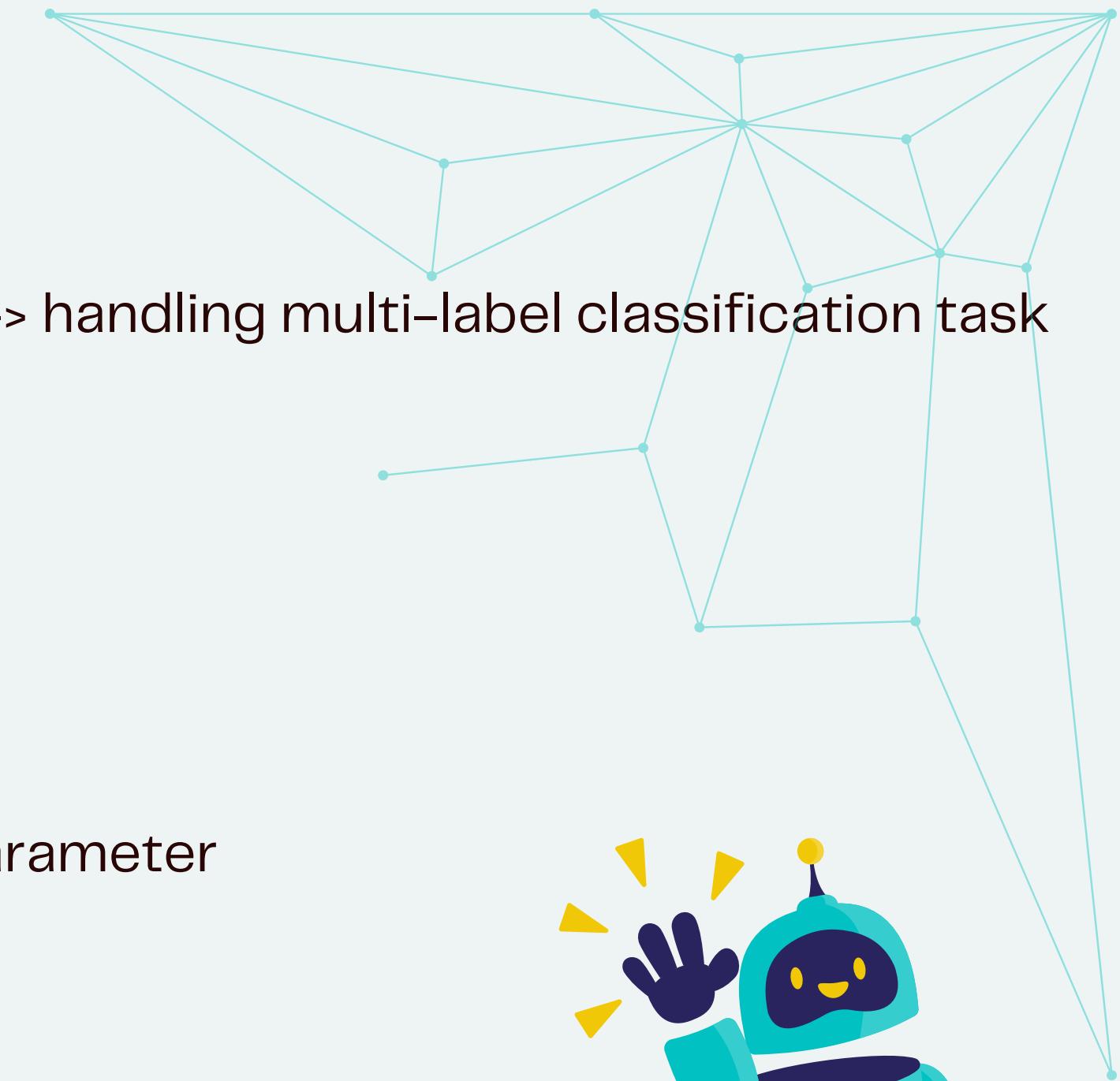
- 'step' → no. of features to remove at each iteration till min. features is achieved.

330 gene features



# MODELING: OVERVIEW

- **Methods defining:** Multi-class (9 classes) prediction → handling multi-label classification task
- **Model selection:**
  - Random Forest
  - Supportive Vector Machine
  - Ensemble Model
  - Convolutional Neural Networks
- **Best model fine-tuning:** Finding the Best-K Hyper-parameter
  - GridSearch
  - KFold + Cross Validation
- **Evaluation metrics:**
  - Accuracy, ROC score, F1 recall score
  - confusion matrix
- **Visualisation**

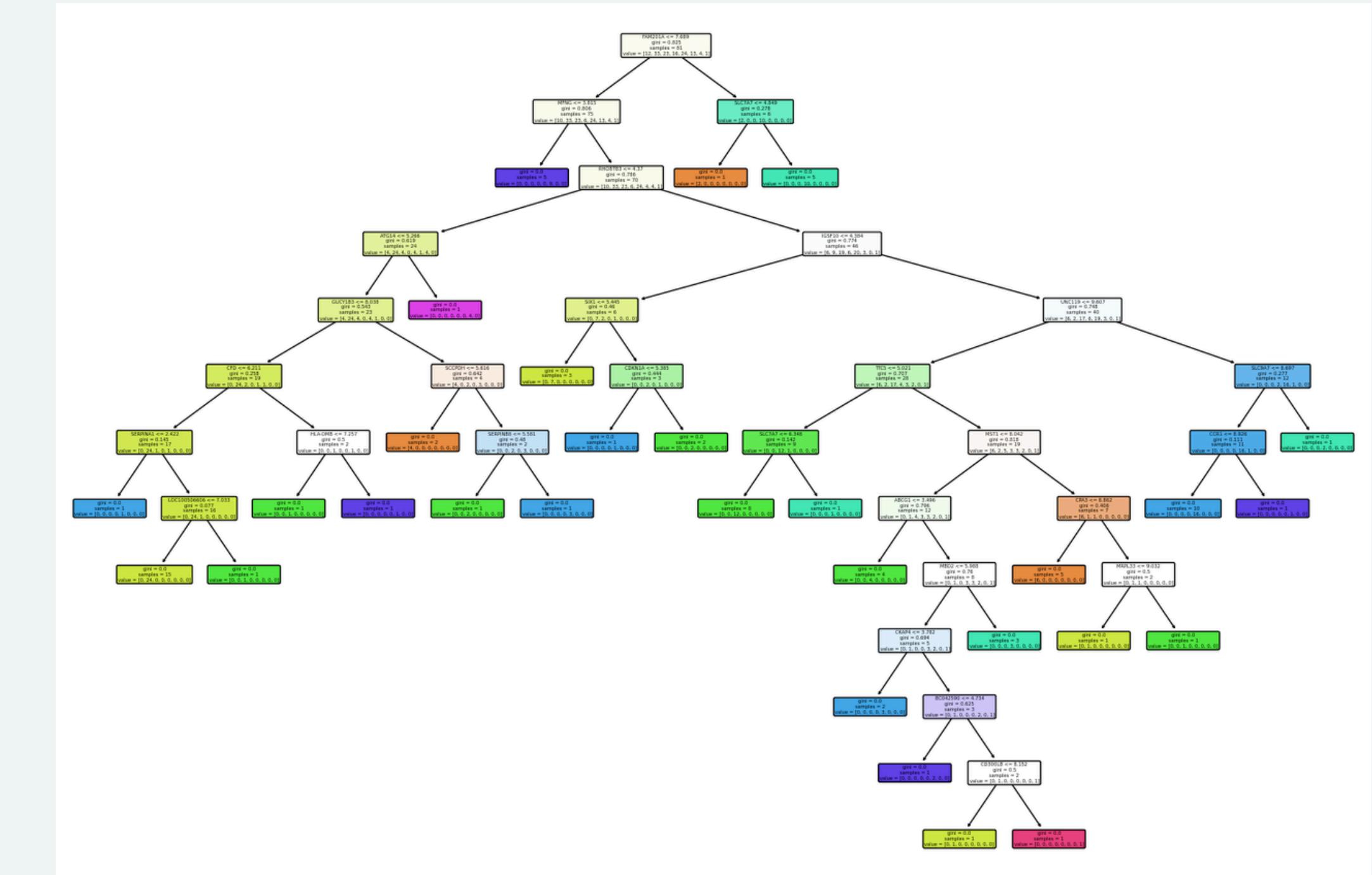


# MODELING: RANDOM FOREST

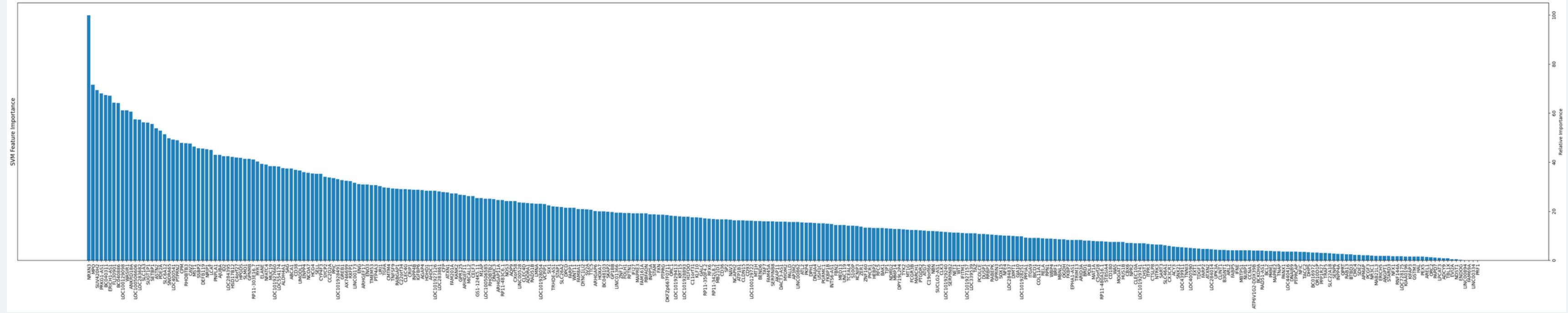


## RESULT

- Accuracy: 0.61
- Weighted-Average F1 Score: 0.57



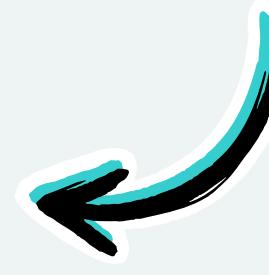
# MODELING: SVM



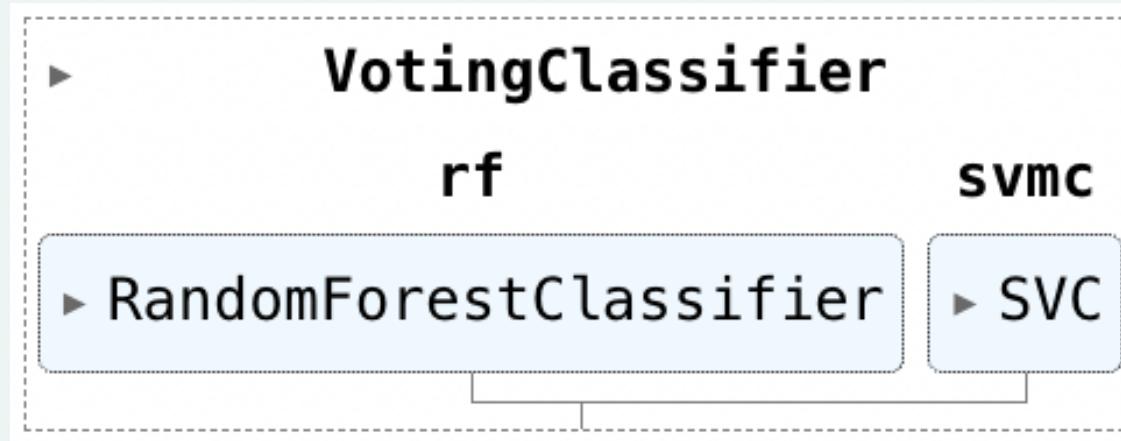
## RESULT

- Accuracy: 0.75
- Weighted-Average F1 Score: 0.73

- Can make further improvement on feature cutting based on the `Feature Importance Bar chart`



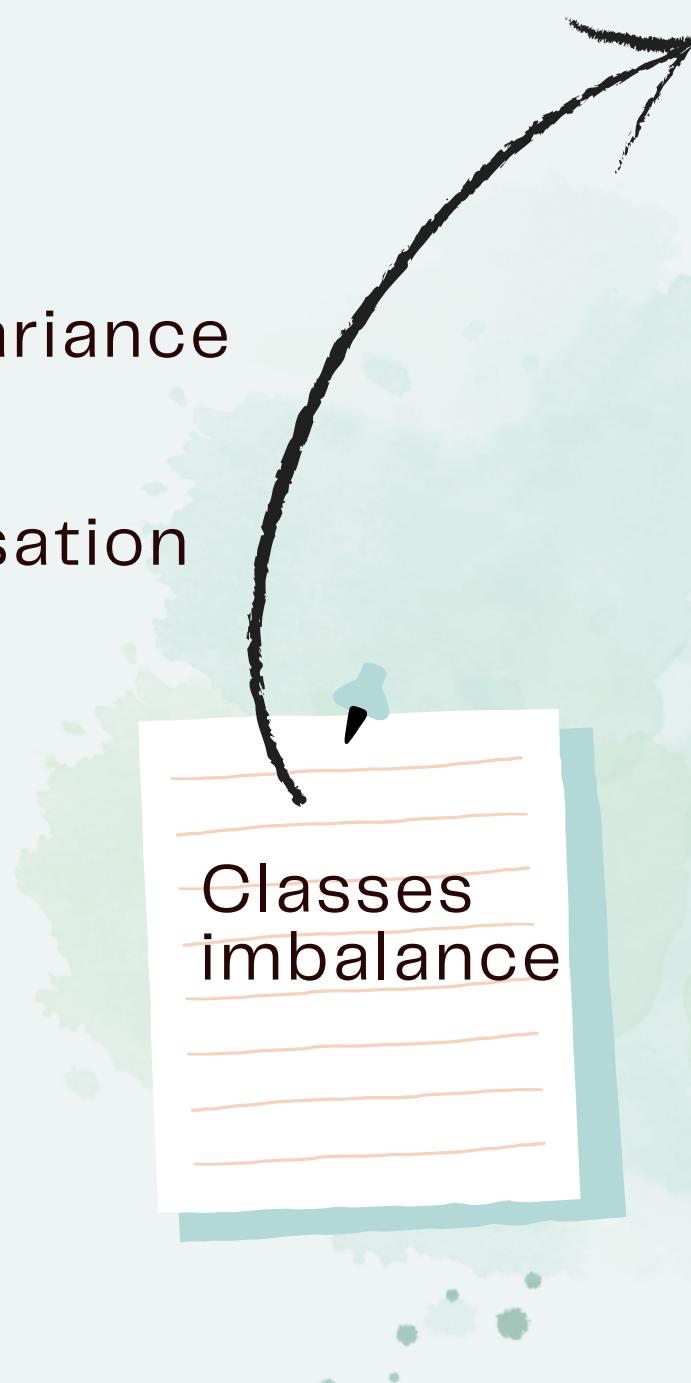
# MODELING: ENSEMBLE SOFT-VOTING



- Address overfitting, bias and variance
- Improve accuracy: > +4%
- Improve single model generalisation

## RESULT

- Accuracy: 0.76



Confusion Matrix of AML Patient Case/Control Predictions						
	0	1	2	3	4	5
0	2	0	1	0	0	0
1	2	8	1	0	0	1
2	0	3	10	0	0	0
3	0	0	0	3	0	0
4	2	1	1	0	11	0
5	0	0	0	0	0	8
6	0	0	0	0	0	0

# MACHINE LEARNING ACCURACY COMPARISON

## Strength

- **Random Forest:**

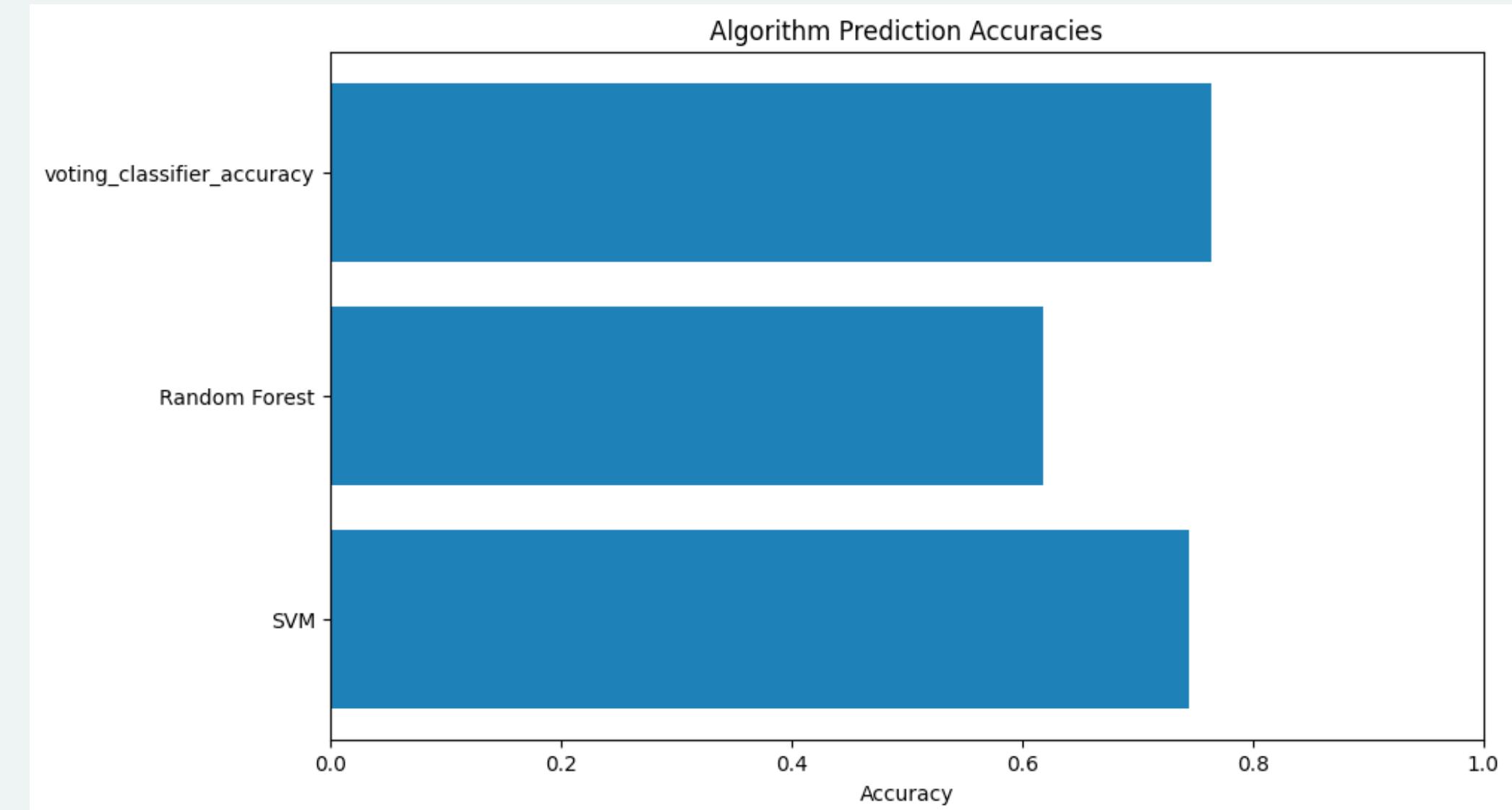
- Feature identification
- Dimension reduction

- **SVM:**

- Nonlinear data
- Optimal hyperplane

## RESULT

- Relatively stable accuracy
- Overall: +16%

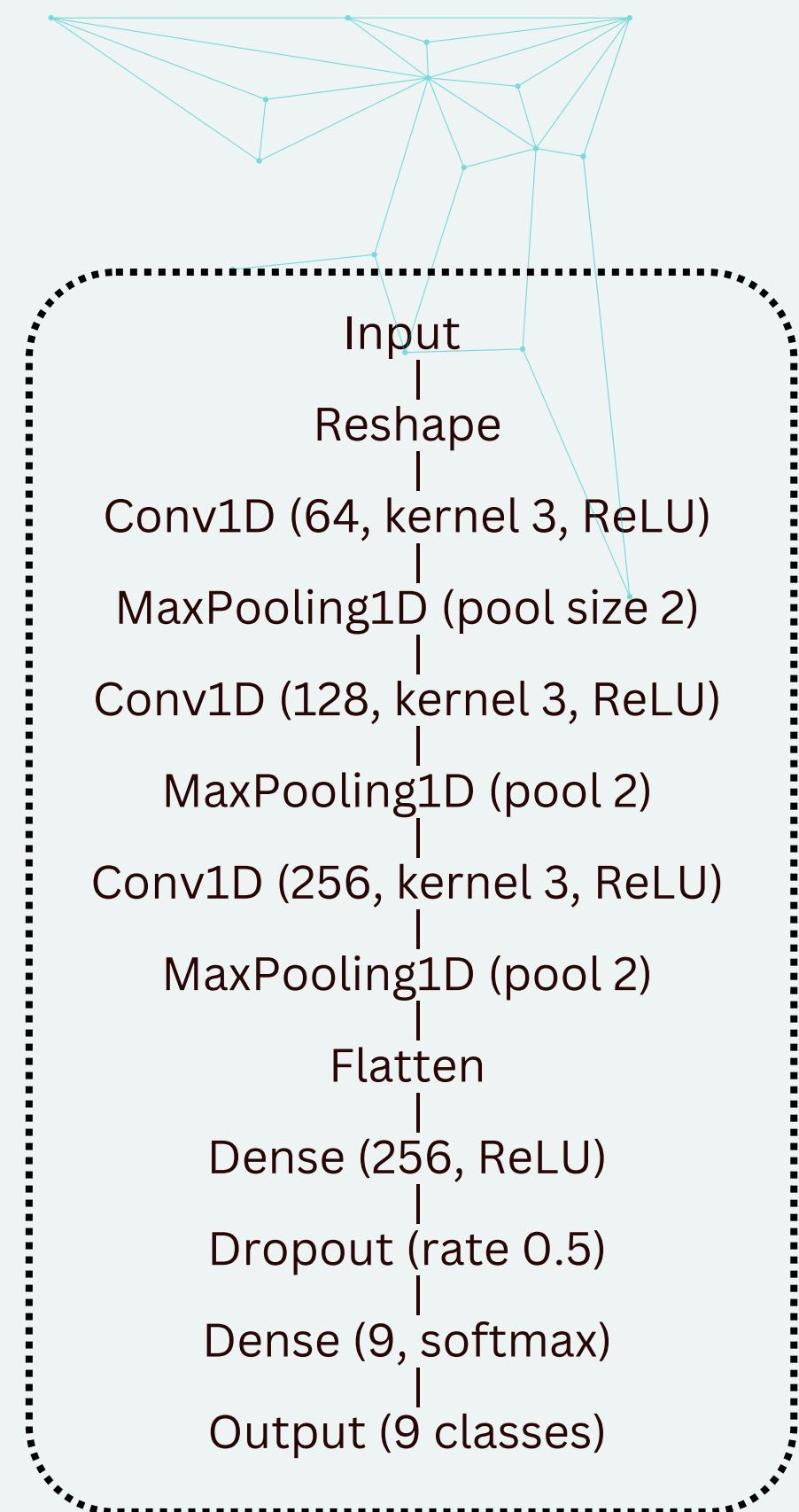
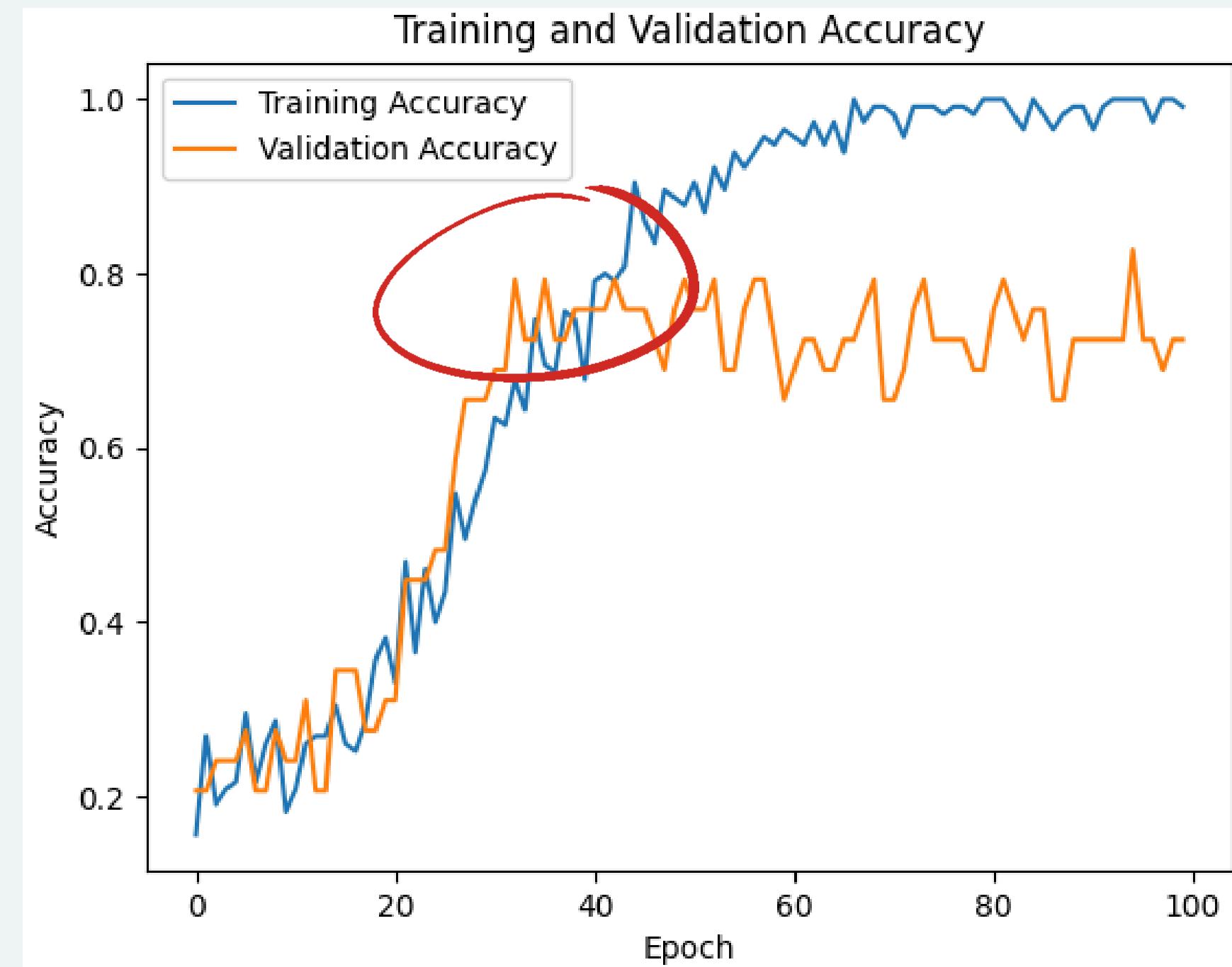


# MODELING APPENDIX: CNN

- CNN converges at 40ish epochs.
- Over-fitting: Validation not always increasing & aligned with training.

## RESULT

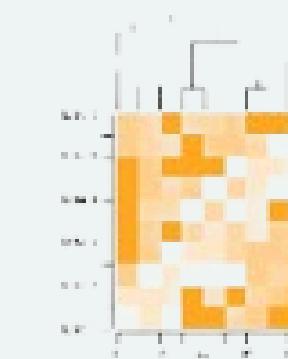
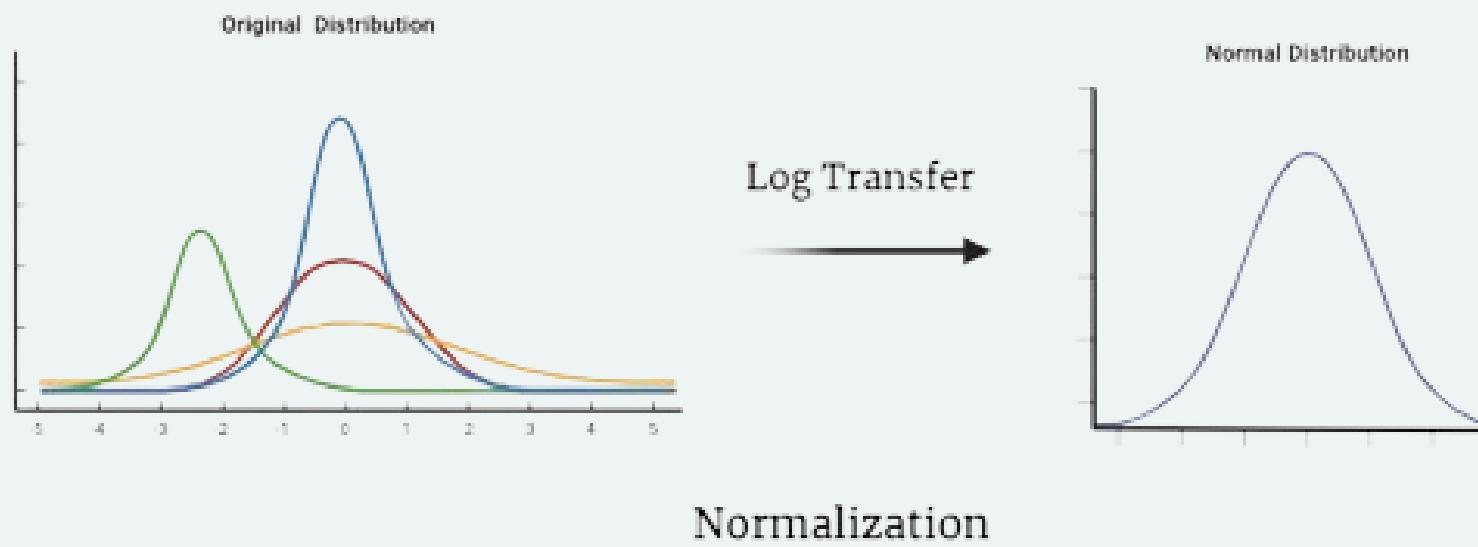
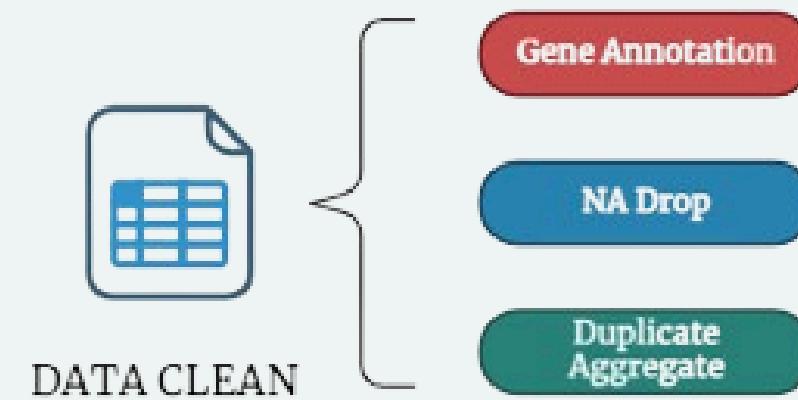
- Accuracy differs each time, but within range of (0.65, 0.70)



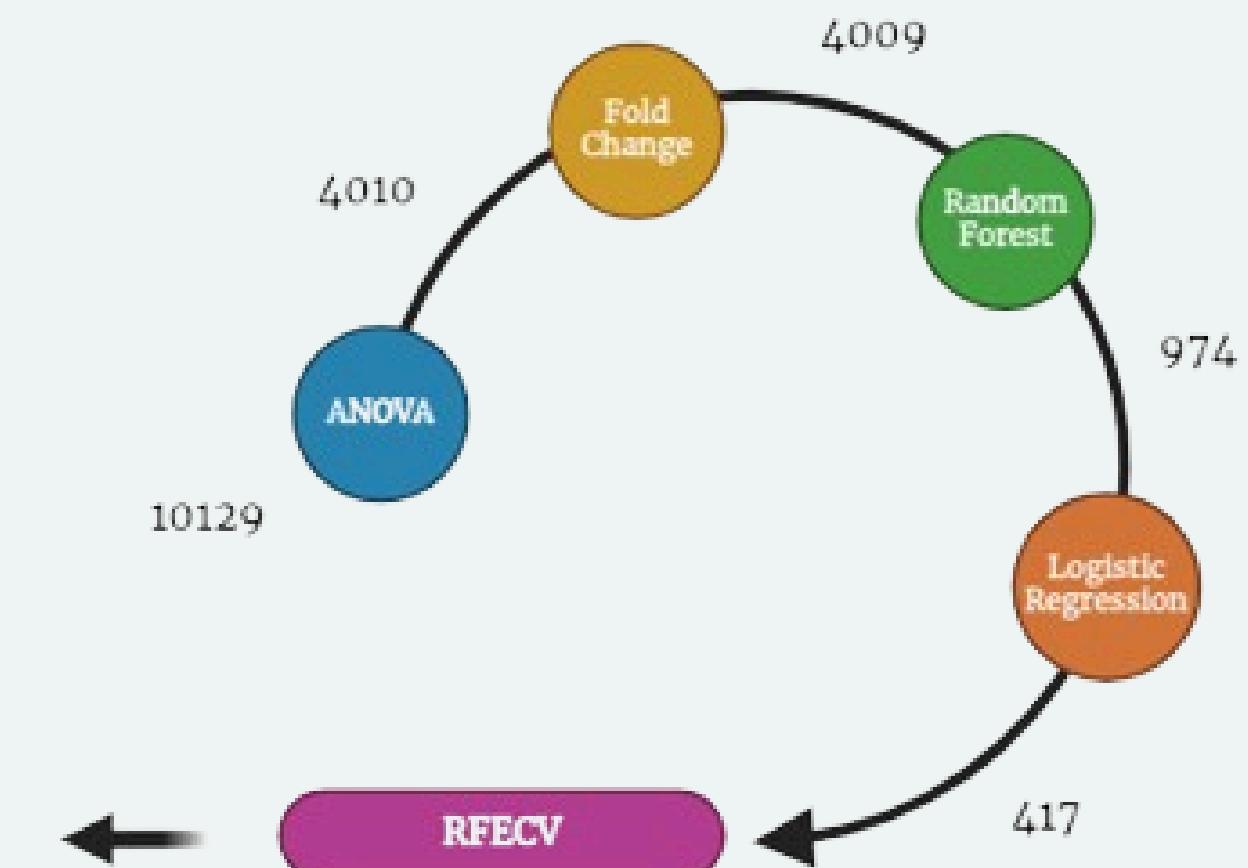
# STRENGTHS



## Comprehensive Preprocessing and Feature Selection



330 BioMarker



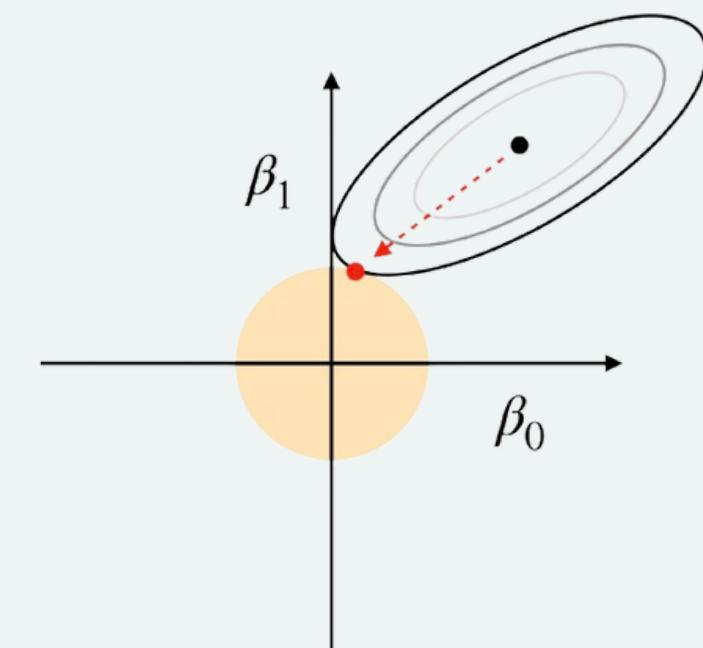
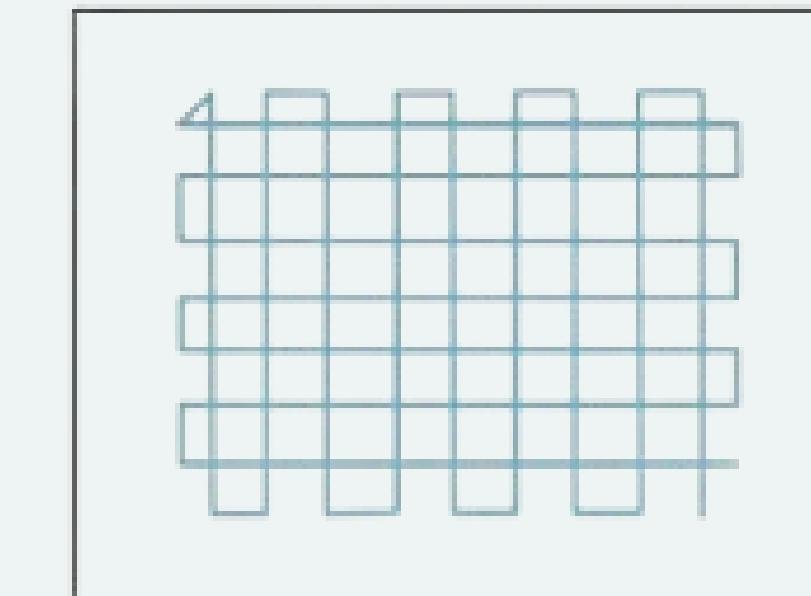
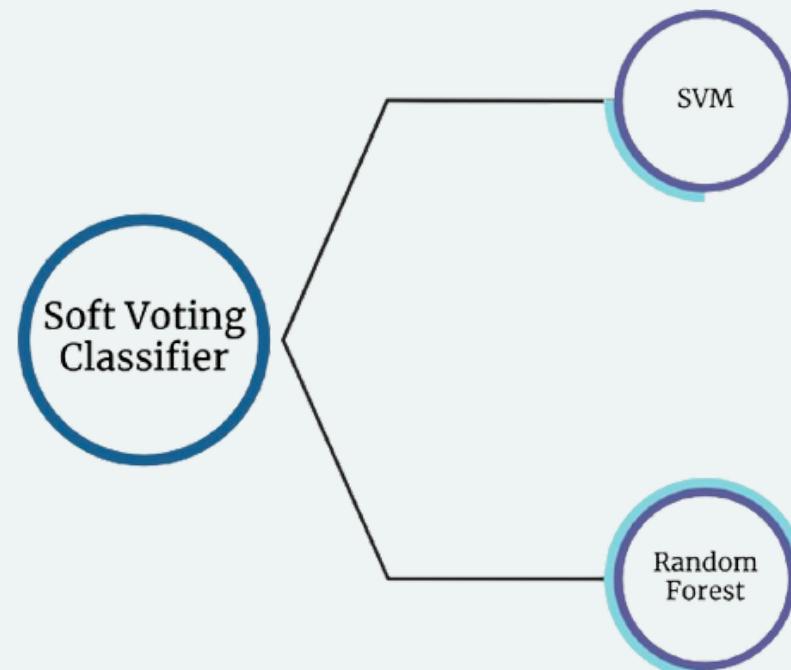
# STRENGTHS



- Use of Multiple Models and Ensemble Method
- Hyperparameter Tuning and Regularization

- Comparative Analysis
- Integrate results from different models
- Crucial for optimizing model performance
- L2 regularization help in feature selection and combating overfitting

Random Forest  
SVM  
Soft Voting Classifier  
CNN



Multiple Models

Ensemble Method

Hyperparameter Space Search

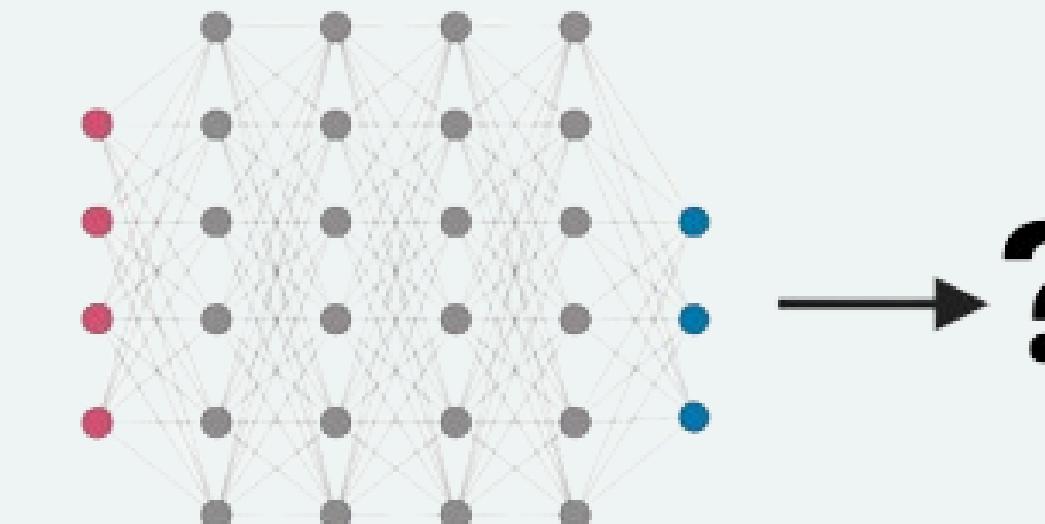
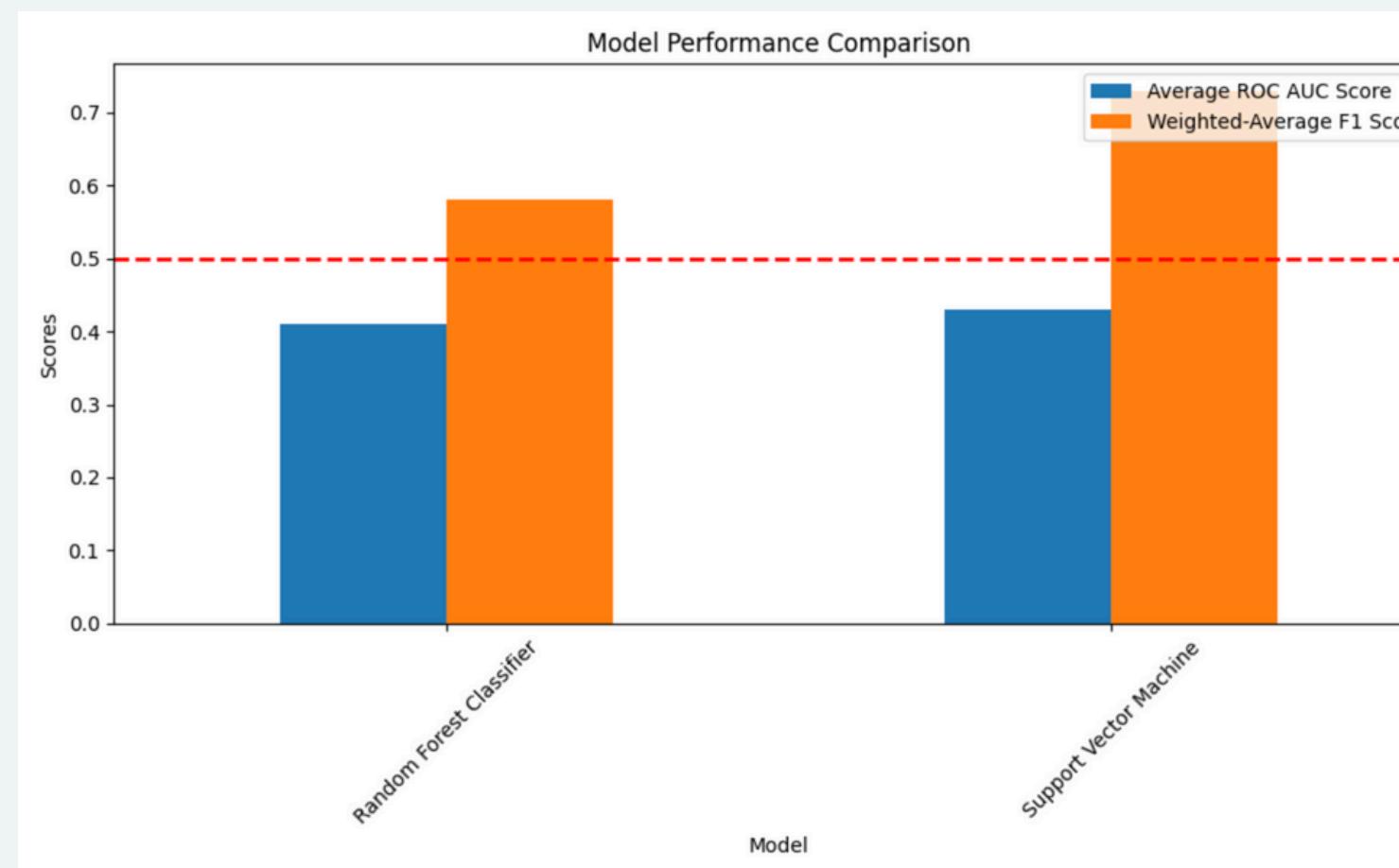
L2 Regularization

# LIMITATIONS

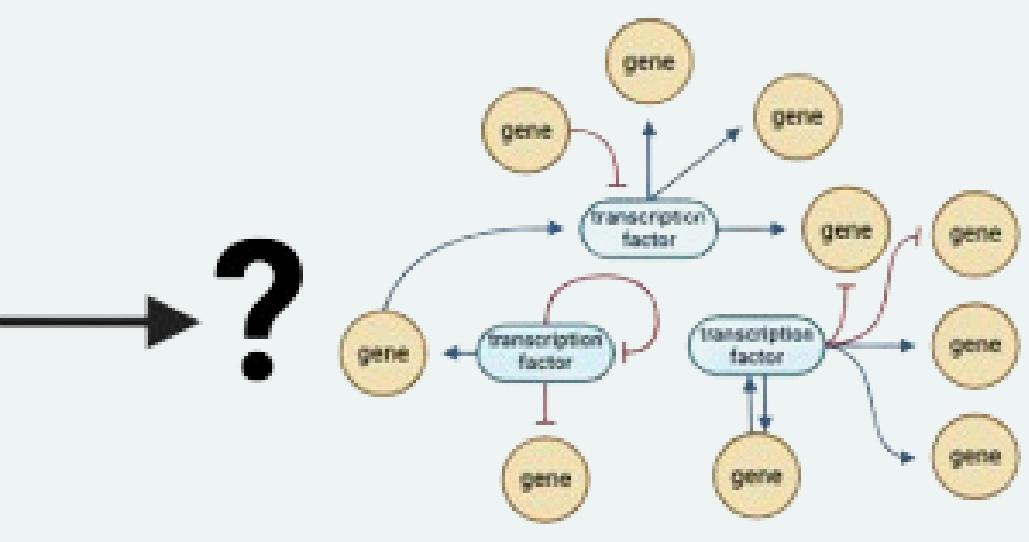


- **Moderate Model Performance**
- **Model Complexity and Interpretability**

- **The accuracy and average ROC AUC across the model indicate a significant room for improvement**
- **The use of ensemble methods and multiple layers of feature selection increase model complexity**



Machine Learning Model

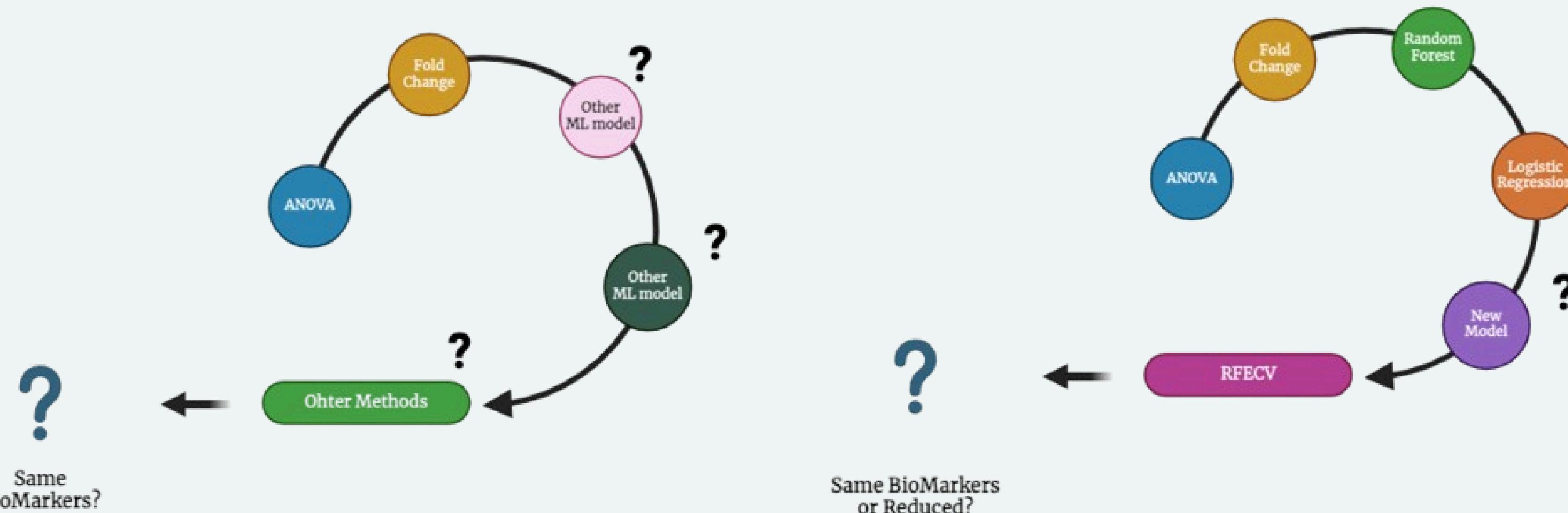


Biological Meaning

# LIMITATIONS

## Feature Selection

- The threshold set for Fold Change is empirical, may result in the exclusion of potentially informative genes
- The convergency of features selected by our workflow do not evaluated

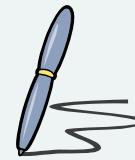


# FUTURE DIRECTIONS



- **Combine other biological information for predicting**
  - Blood routine examination
  - Microscope image data
- **Use Biological field knowledge to help feature selection**
  - Conduct GO and KEGG pathway enrichment analysis
  - Conduct a GSEA analysis to focus on the feature selection in a significant gene set
- **Find a convergence feature set using different methods**
- **Use a larger dataset to do feature selection and model training**
- **Apply other Machine Learning Methods to improve the prediction**
  - LSTM, Neuron Network, Transformer

# REFLECTIONS



## Technical Learning:

- Struggled with coding and dataset comprehension, but improved through hands-on experience and learning data-preprocessing and feature selection techniques.



## Coding Insights:

- Found joy in coding, seeing it as a puzzle-solving activity despite no background.



## Methodology Focus:

- Emphasize the importance of data cleaning, preprocessing, and feature engineering over hyper-parameter tuning.
- Realized traditional machine learning is often more practical than deep learning for real-world scenarios.



## Project Management:

- Faced challenges with back-to-back deadlines, highlighting the need for better time management.



# REFERENCES

- [1]GEO Accession viewer. (n.d.). <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68833>
- [2]Bystrykh, L. (2022, June 23). Python for gene expression. *F1000Research*, 10, 870. <https://doi.org/10.12688/f1000research.53842.2>
- [3]Acute Myeloid Leukemia (AML) Subtypes and Prognostic Factors. (n.d.). American Cancer Society. <https://www.cancer.org/cancer/types/acute-myeloid-leukemia/detection-diagnosis-staging/how-classified.html>
- [4]heatmaps of RNA-seq data: what are accepted normalizations? (n.d.). <https://www.biostars.org/p/356041/>
- [5]GeneChip™ Human Genome U133 Plus 2.0 Array. (n.d.). <https://www.thermofisher.com/order/catalog/product/900466>
- [6]Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., .. Campbell, P. J. (2016, June 9). Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine*, 374(23), 2209–2221. <https://doi.org/10.1056/nejmoa1516192>
- [7]De Kouchkovsky, I., & Abdul-Hay, M. (2016, July 1). 'Acute myeloid leukemia: a comprehensive review and 2016 update.' *Blood Cancer Journal*, 6(7), e441–e441. <https://doi.org/10.1038/bcj.2016.50>
- [8]Weinberg, O. K., Porwit, A., Orazi, A., Hasserjian, R. P., Foucar, K., Duncavage, E. J., & Arber, D. A. (2022, October 20). The International Consensus Classification of acute myeloid leukemia. *Virchows Archiv*, 482(1), 27–37. <https://doi.org/10.1007/s00428-022-03430-4>

# CHATGPT HISTORY

- <https://copilot.microsoft.com/sl/ibWvCoqJy5Q>
- <https://copilot.microsoft.com/sl/ekvQMb4vVmK>
- <https://copilot.microsoft.com/sl/cJJiGYEtrWu>
- <https://chat.openai.com/share/c2c3f278-5b28-4d7b-971b-3743a04ddb31>



Thaw  
you!