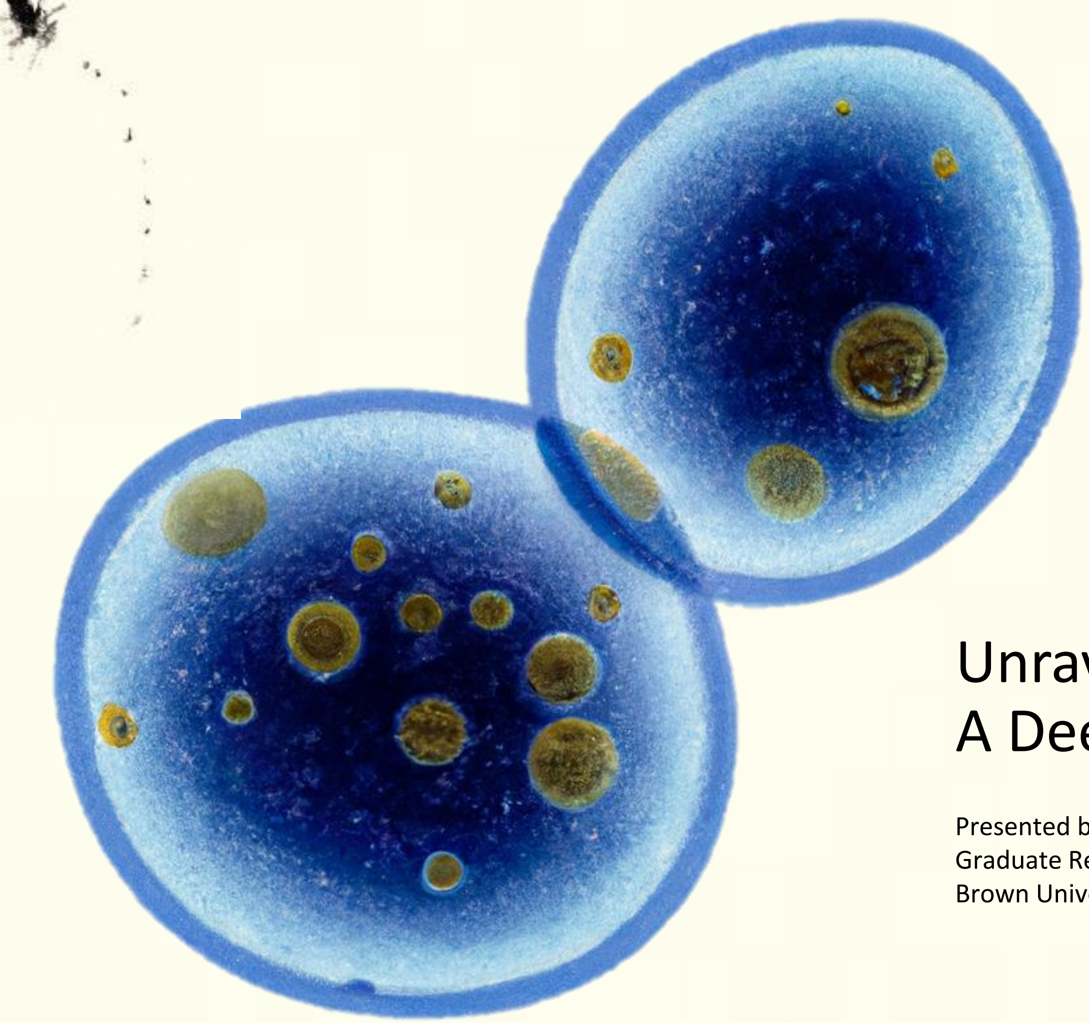
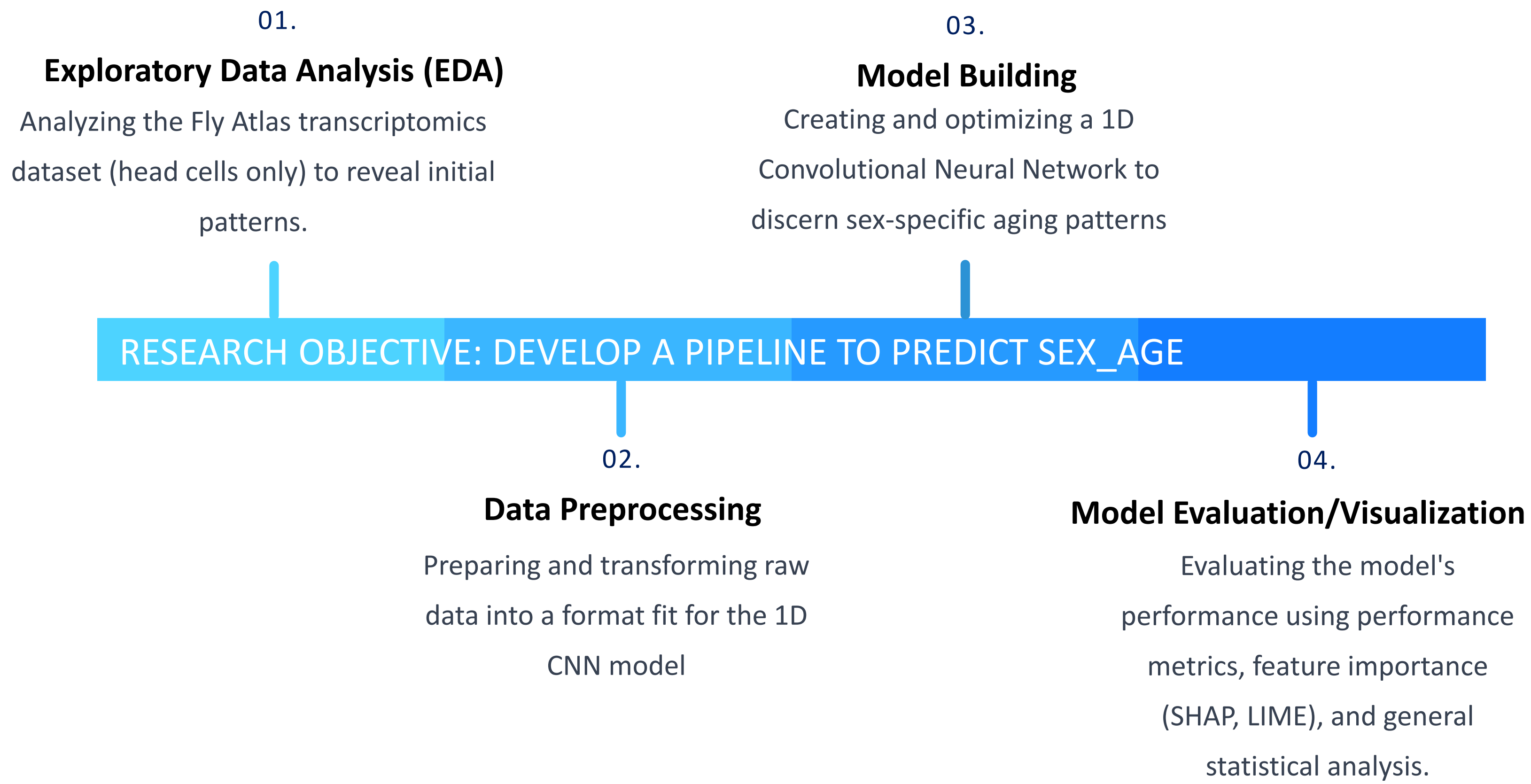


Unravelling Sex-Specific Aging on Flies: A Deep Learning Approach

Presented by **Nikolai Tennant**,
Graduate Researcher at Ritambhara Singh's Lab
Brown University & IISAGE

METHODOLOGY



RESEARCH OBJECTIVE: DEVELOP A PIPELINE TO PREDICT SEX_AGE

01.

Exploratory Data Analysis (EDA)

Analyzing the Fly Atlas transcriptomics dataset (head cells only) to reveal initial patterns.

02.

Data Preprocessing

Preparing and transforming raw data into a format fit for the 1D CNN model

03.

Model Building

Creating and optimizing a 1D Convolutional Neural Network to discern sex-specific aging patterns

04.

Model Evaluation/Visualization

Evaluating the model's performance using performance metrics, feature importance (SHAP, LIME), and general statistical analysis.

EDA

Dataset
Initiated the methodology with the Fly Atlas transcriptomics dataset (head cells only).

Data Integrity
Checked for data integrity, ensured no duplicate rows or missing values in the gene expression data.

Data Format
Shifted data from a sparse to a dense format for easy manipulation and analysis.

Target Variable

- Analysed 'sex_age' as the primary target variable, 'female_5' being the most populous category.
- Also looked at ‘sex’ and ‘age’ variables

Gene Expression Data
Examined gene expression data, decided to keep identified outlier cells due to potential biological significance.

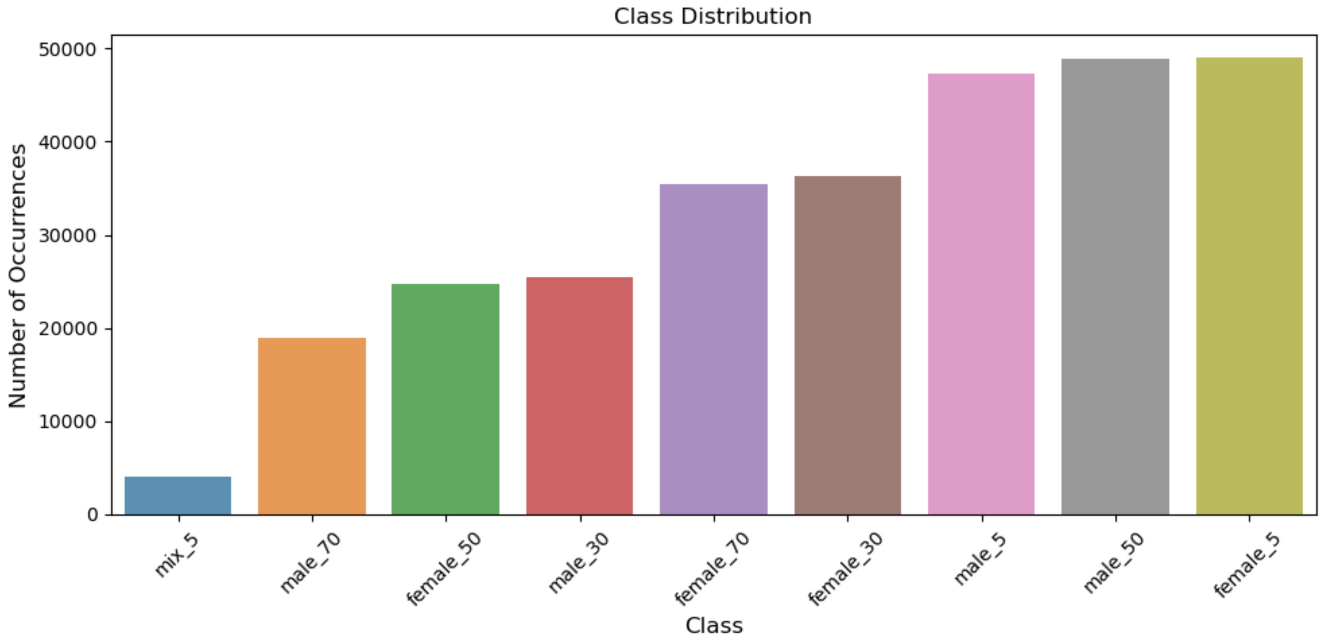


Figure 1: Bar Plot Showing Class Distributions for sex_age

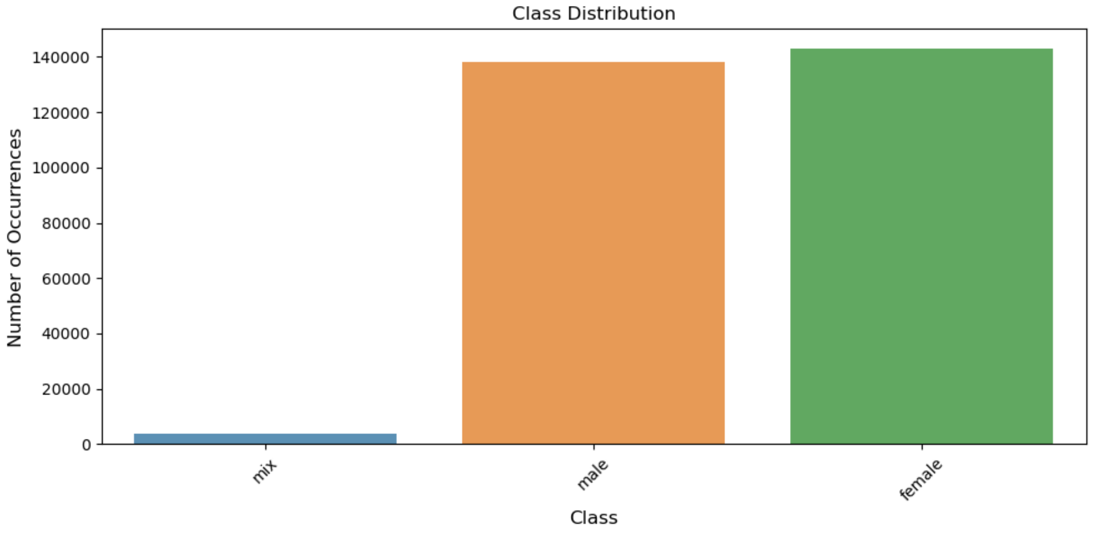


Figure 2: Bar Plot Showing Class Distributions for Sex

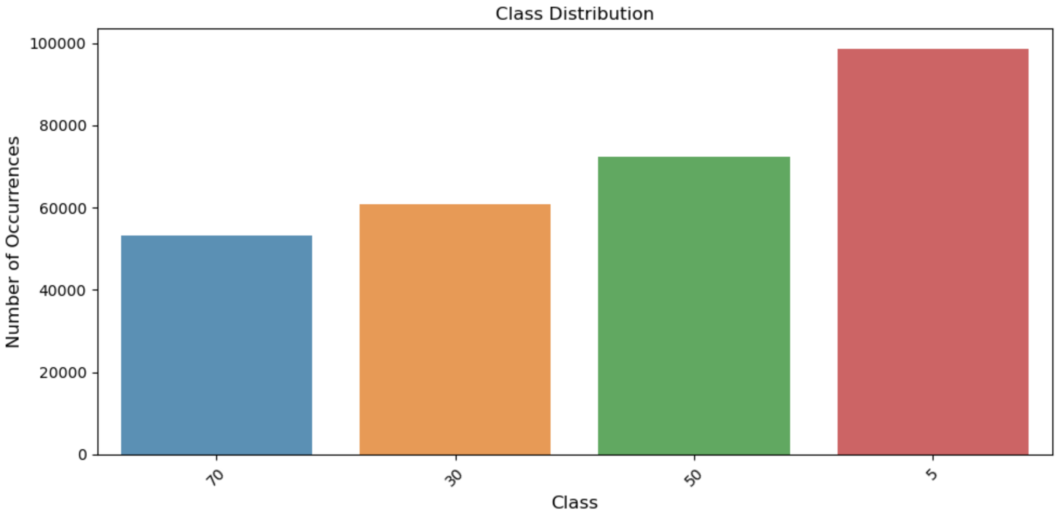
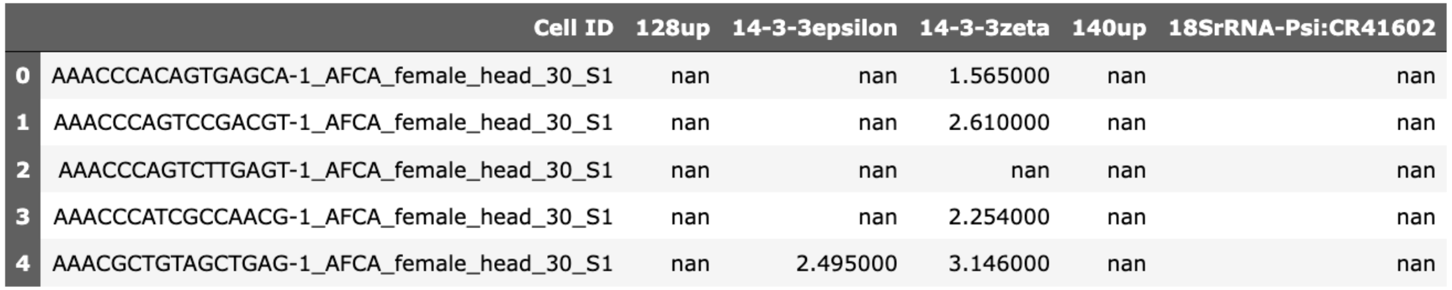
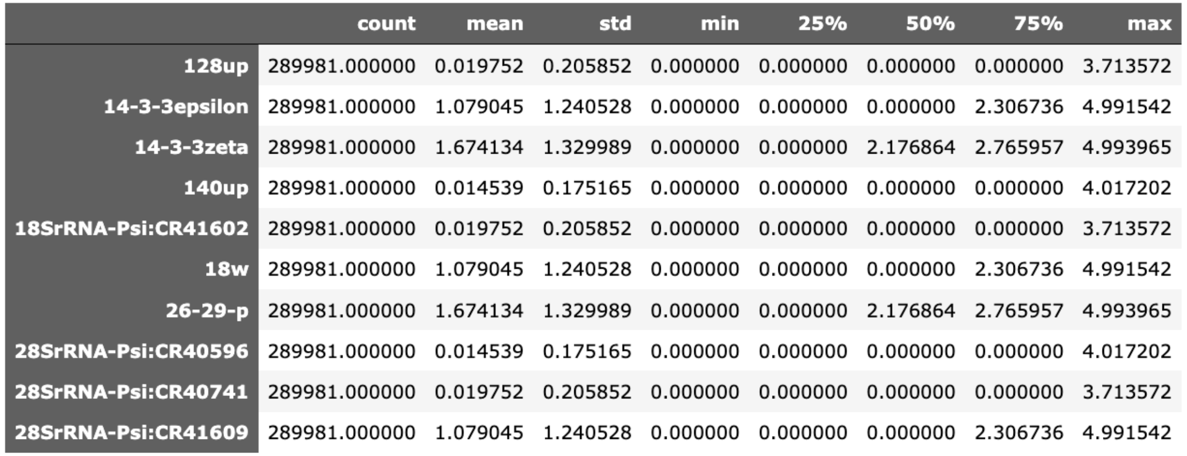


Figure 3: Bar Plot Showing Class Distributions for Age



	Cell ID	128up	14-3-3epsilon	14-3-3zeta	140up	18SrRNA-Psi:CR41602
0	AAACCCACAGTGAGCA-1_AFCA_female_head_30_S1	nan	nan	1.565000	nan	nan
1	AAACCCAGTCCGACGT-1_AFCA_female_head_30_S1	nan	nan	2.610000	nan	nan
2	AAACCCAGTCTTGAGT-1_AFCA_female_head_30_S1	nan	nan	nan	nan	nan
3	AAACCCATCGCCAACG-1_AFCA_female_head_30_S1	nan	nan	2.254000	nan	nan
4	AAACGCTGTAGCTGAG-1_AFCA_female_head_30_S1	nan	2.495000	3.146000	nan	nan

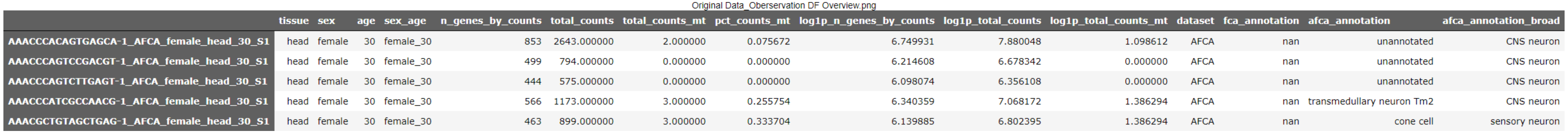
Figure 4: Gene Expression Matrix for First Five Cells



	count	mean	std	min	25%	50%	75%	max
128up	289981.000000	0.019752	0.205852	0.000000	0.000000	0.000000	0.000000	3.713572
14-3-3epsilon	289981.000000	1.079045	1.240528	0.000000	0.000000	0.000000	2.306736	4.991542
14-3-3zeta	289981.000000	1.674134	1.329989	0.000000	0.000000	2.176864	2.765957	4.993965
140up	289981.000000	0.014539	0.175165	0.000000	0.000000	0.000000	0.000000	4.017202
18SrRNA-Psi:CR41602	289981.000000	0.019752	0.205852	0.000000	0.000000	0.000000	0.000000	3.713572
18w	289981.000000	1.079045	1.240528	0.000000	0.000000	0.000000	2.306736	4.991542
26-29-p	289981.000000	1.674134	1.329989	0.000000	0.000000	2.176864	2.765957	4.993965
28SrRNA-Psi:CR40596	289981.000000	0.014539	0.175165	0.000000	0.000000	0.000000	0.000000	4.017202
28SrRNA-Psi:CR40741	289981.000000	0.019752	0.205852	0.000000	0.000000	0.000000	0.000000	3.713572
28SrRNA-Psi:CR41609	289981.000000	1.079045	1.240528	0.000000	0.000000	0.000000	2.306736	4.991542

Figure 5: Statistical Summary of First 10 Genes

Original Data_Oberservation DF Overview.png



	tissue	sex	age	sex_age	n_genes_by_counts	total_counts	total_counts_mt	pct_counts_mt	log1p_n_genes_by_counts	log1p_total_counts	log1p_total_counts_mt	dataset	fca_annotation	afca_annotation	afca_annotation_broad
AAACCCACAGTGAGCA-1_AFCA_female_head_30_S1	head	female	30	female_30	853	2643.000000	2.000000	0.075672	6.749931	7.880048	1.098612	AFCA	nan	unannotated	CNS neuron
AAACCCAGTCCGACGT-1_AFCA_female_head_30_S1	head	female	30	female_30	499	794.000000	0.000000	0.000000	6.214608	6.678342	0.000000	AFCA	nan	unannotated	CNS neuron
AAACCCAGTCTTGAGT-1_AFCA_female_head_30_S1	head	female	30	female_30	444	575.000000	0.000000	0.000000	6.098074	6.356108	0.000000	AFCA	nan	unannotated	CNS neuron
AAACCCATCGCCAACG-1_AFCA_female_head_30_S1	head	female	30	female_30	566	1173.000000	3.000000	0.255754	6.340359	7.068172	1.386294	AFCA	nan	transmedullary neuron Tm2	CNS neuron
AAACGCTGTAGCTGAG-1_AFCA_female_head_30_S1	head	female	30	female_30	463	899.000000	3.000000	0.333704	6.139885	6.802395	1.386294	AFCA	nan	cone cell	sensory neuron

Figure 6: Observation DF Overview

PREPROCESSING

Initial Data Split

- Initial stratified split of dataset into evaluation (for pre-trained models) and larger training subsets (everything else)

Category Handling

- Selective handling of categories, such as 'mix' category exclusion

Stratified Subset Selection

- Selected a manageable stratified subset of 200,000 cells from the training dataset (15,992 genes – all).

Further Split and Transformation

- Further stratified split of selected data into training and testing subsets.
- Data labels transformed using a label encoder and one-hot encoding

Data Reshaping

- Convert gene expression data to dense matrix and reshape to fit CNN input
- Chose to bypass standardization/normalization due to computational demands and performance reduction.

Data Saving and Loading

- Preprocessed data, label encoder, lime explainer and reference data saved for potential future use.
- Feature to load saved preprocessed data and label encoder for subsequent pipeline executions.

Uniform Label Encoding

- Preserved uniformity in label encoding during preprocessing of evaluation data.

MODEL BUILDING AND TRAINING

Configurable Model Building

- Model construction guided by a flexible configuration file.

Model Construction

- Three models built: 'sex', 'age', and 'sex_age'.
- Type: 1D CNN
- Architecture: Convolutional blocks, flattening layer, fully connected layers, and output layer.

Optimization and Compilation

- Optimized performance with a stratified dataset subset size of (200000, 15992) and used a 20% split.
- Models compiled using Adam optimizer and categorical cross-entropy loss function.
- Evaluation metrics: accuracy and AUC.

Training Process

- Incorporated validation split and custom early stopping and model checkpointing.
- Checkpointing ensures retention of the most optimal model parameters and aids visualization.

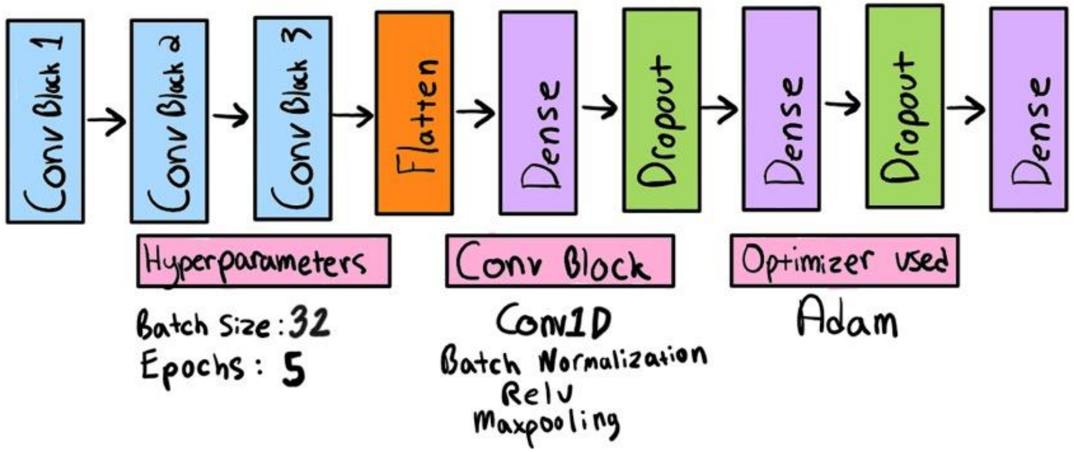
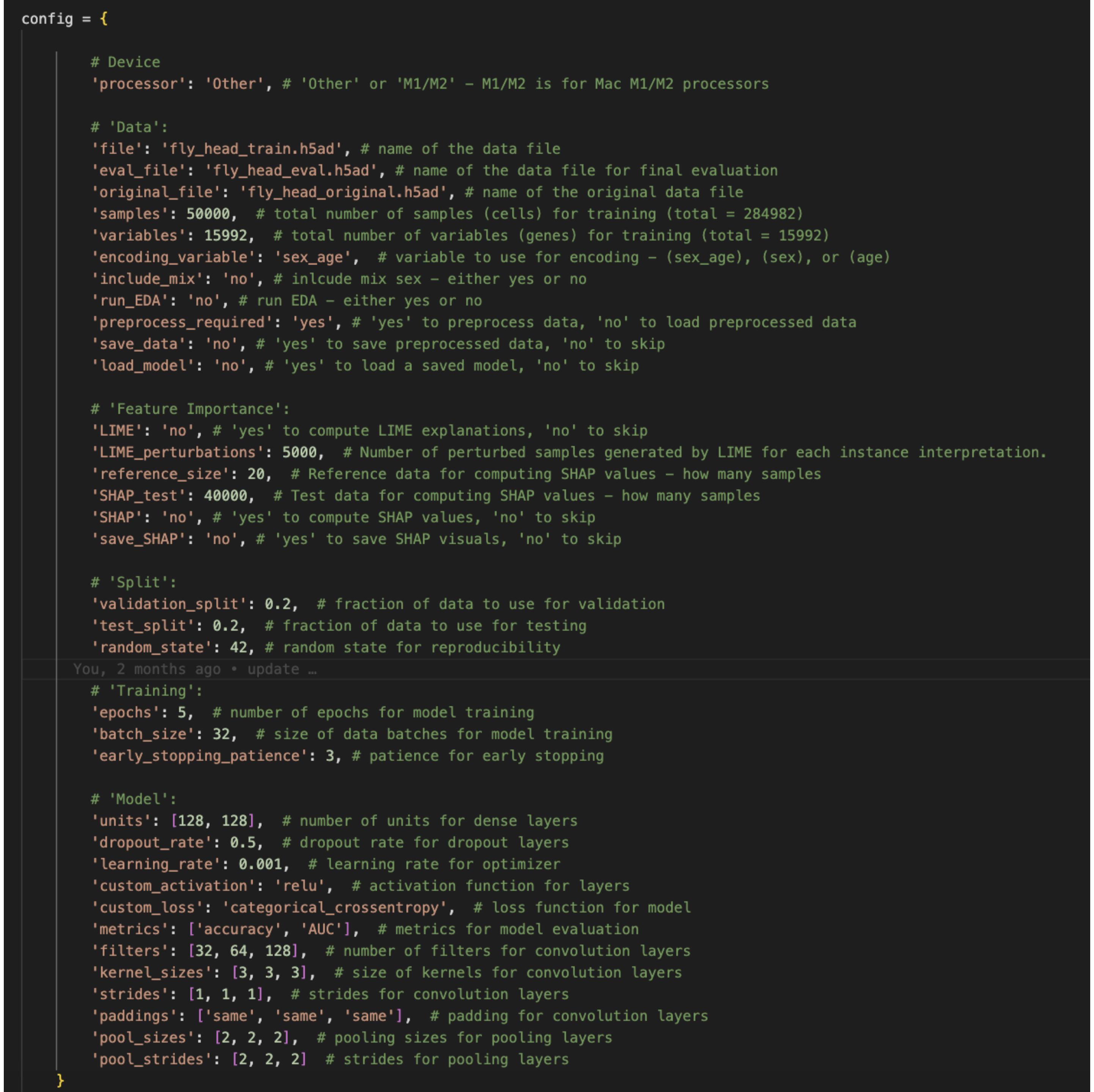


Figure 7: Convolution Neural Network Architecture

Figure 8: Config Dictionary



```
config = {

        # Device
        'processor': 'Other', # 'Other' or 'M1/M2' - M1/M2 is for Mac M1/M2 processors

        # 'Data':
        'file': 'fly_head_train.h5ad', # name of the data file
        'eval_file': 'fly_head_eval.h5ad', # name of the data file for final evaluation
        'original_file': 'fly_head_original.h5ad', # name of the original data file
        'samples': 50000,  # total number of samples (cells) for training (total = 284982)
        'variables': 15992,  # total number of variables (genes) for training (total = 15992)
        'encoding_variable': 'sex_age',  # variable to use for encoding - (sex_age), (sex), or (age)
        'include_mix': 'no', # inlcude mix sex - either yes or no
        'run_EDA': 'no', # run EDA - either yes or no
        'preprocess_required': 'yes', # 'yes' to preprocess data, 'no' to load preprocessed data
        'save_data': 'no', # 'yes' to save preprocessed data, 'no' to skip
        'load_model': 'no', # 'yes' to load a saved model, 'no' to skip

        # 'Feature Importance':
        'LIME': 'no', # 'yes' to compute LIME explanations, 'no' to skip
        'LIME_perturbations': 5000,  # Number of perturbed samples generated by LIME for each instance interpretation.
        'reference_size': 20,  # Reference data for computing SHAP values - how many samples
        'SHAP_test': 40000,  # Test data for computing SHAP values - how many samples
        'SHAP': 'no', # 'yes' to compute SHAP values, 'no' to skip
        'save_SHAP': 'no', # 'yes' to save SHAP visuals, 'no' to skip

        # 'Split':
        'validation_split': 0.2,  # fraction of data to use for validation
        'test_split': 0.2,  # fraction of data to use for testing
        'random_state': 42, # random state for reproducibility
      You, 2 months ago • update …
        # 'Training':
        'epochs': 5,  # number of epochs for model training
        'batch_size': 32,  # size of data batches for model training
        'early_stopping_patience': 3, # patience for early stopping

        # 'Model':
        'units': [128, 128],  # number of units for dense layers
        'dropout_rate': 0.5,  # dropout rate for dropout layers
        'learning_rate': 0.001,  # learning rate for optimizer
        'custom_activation': 'relu',  # activation function for layers
        'custom_loss': 'categorical_crossentropy',  # loss function for model
        'metrics': ['accuracy', 'AUC'],  # metrics for model evaluation
        'filters': [32, 64, 128],  # number of filters for convolution layers
        'kernel_sizes': [3, 3, 3],  # size of kernels for convolution layers
        'strides': [1, 1, 1],  # strides for convolution layers
        'paddings': ['same', 'same', 'same'],  # padding for convolution layers
        'pool_sizes': [2, 2, 2],  # pooling sizes for pooling layers
        'pool_strides': [2, 2, 2]  # strides for pooling layers
    }
```

RESULTS

Metrics

SEX_AGE

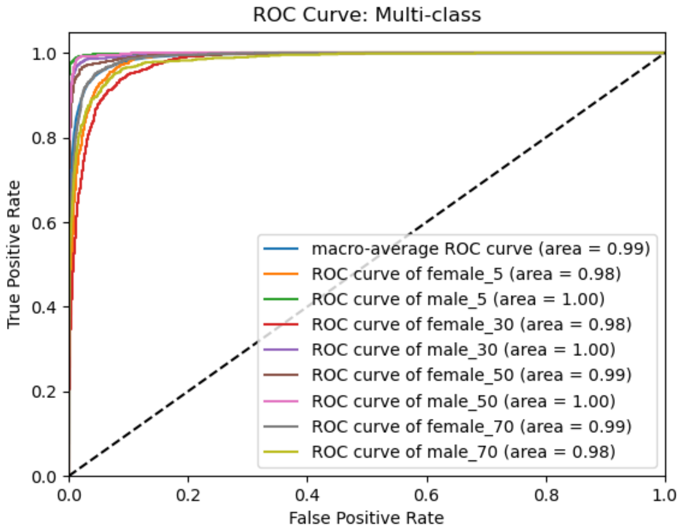


Figure 9: sex_age ROC Curve

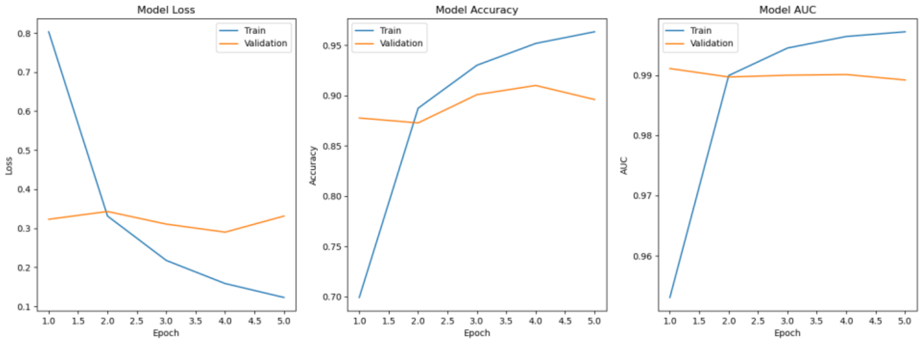
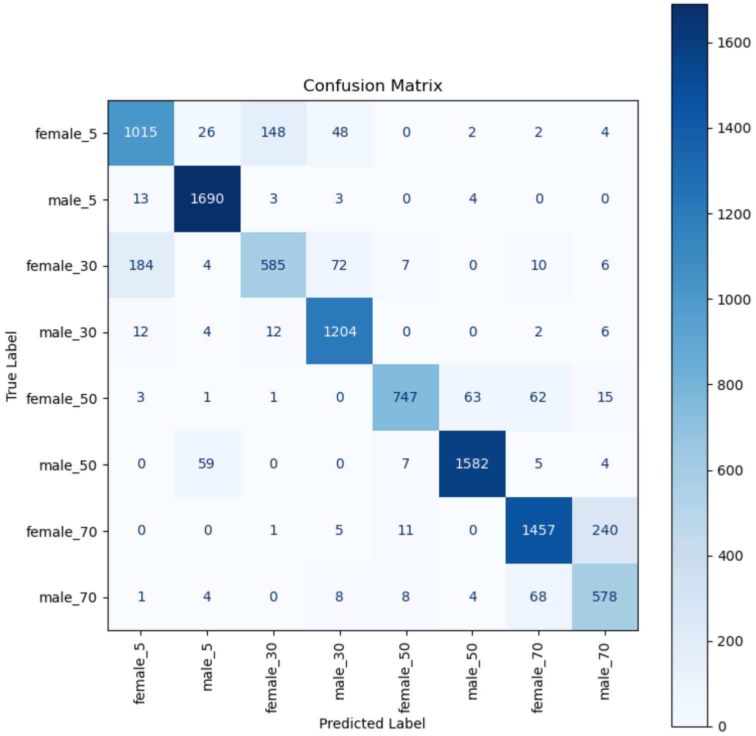


Figure 10: sex_age Training History Plot



True Label \ Predicted Label	female_5	male_5	female_30	male_30	female_50	male_50	female_70	male_70
female_5	1015	26	148	48	0	2	2	4
male_5	13	1690	3	3	0	4	0	0
female_30	184	4	585	72	7	0	10	6
male_30	12	4	12	1204	0	0	2	6
female_50	3	1	1	0	747	63	62	15
male_50	0	59	0	0	7	1582	5	4
female_70	0	0	1	5	11	0	1457	240
male_70	1	4	0	8	8	4	68	578

Figure 11: sex_age Confusion Matrix

SEX

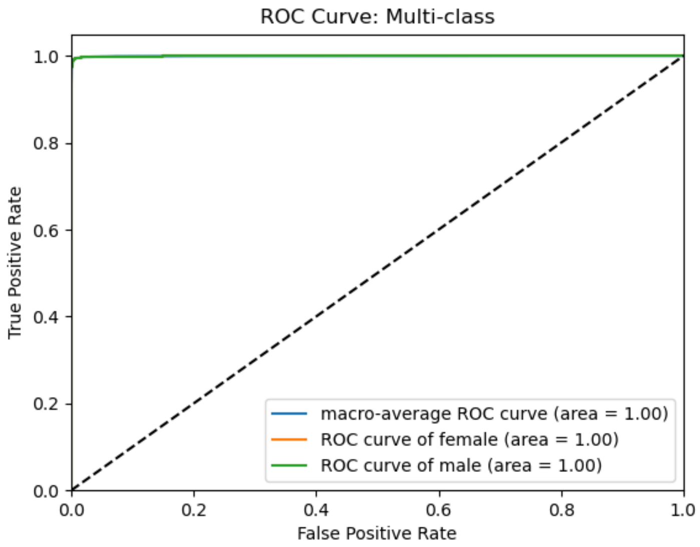


Figure 12: Sex ROC Curve

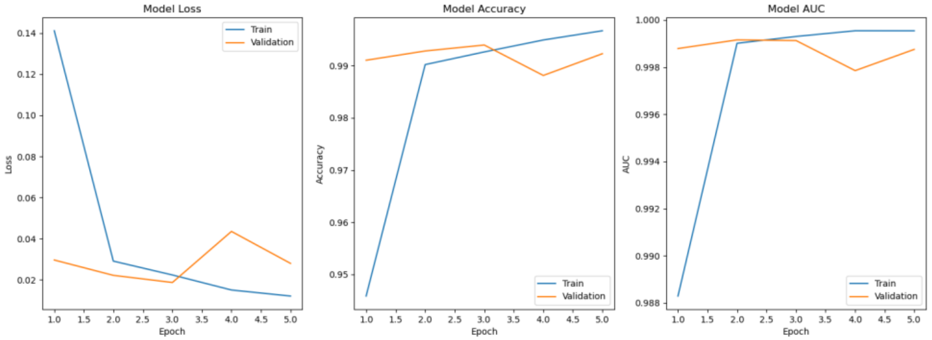


Figure 13: Sex Training History Plot

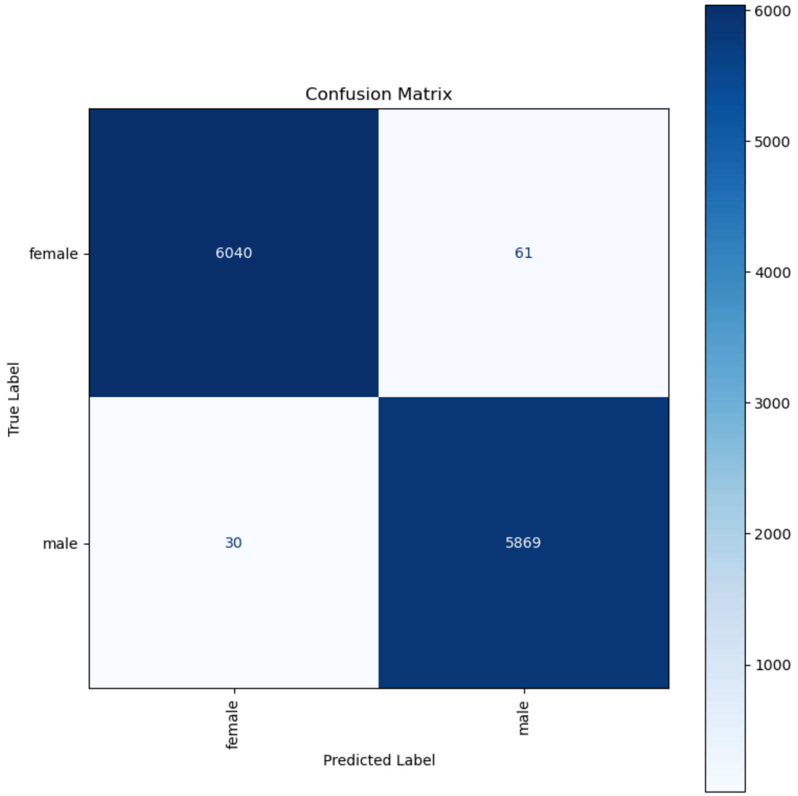


Figure 14: Sex Confusion Matrix

AGE

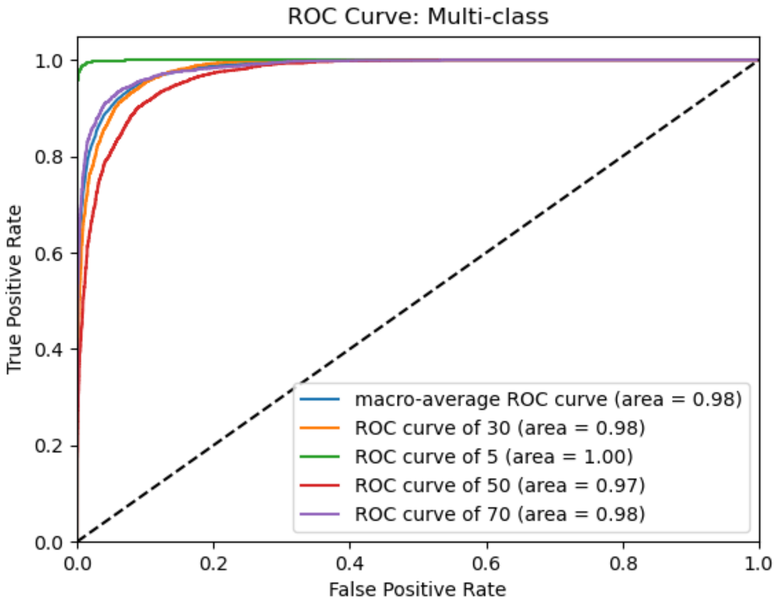


Figure 15: Age ROC Curve

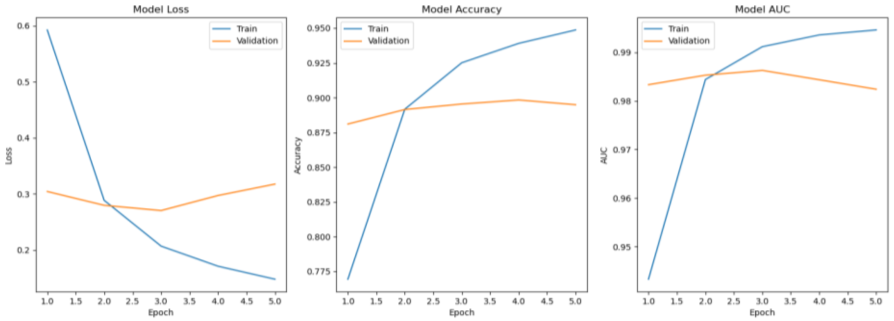
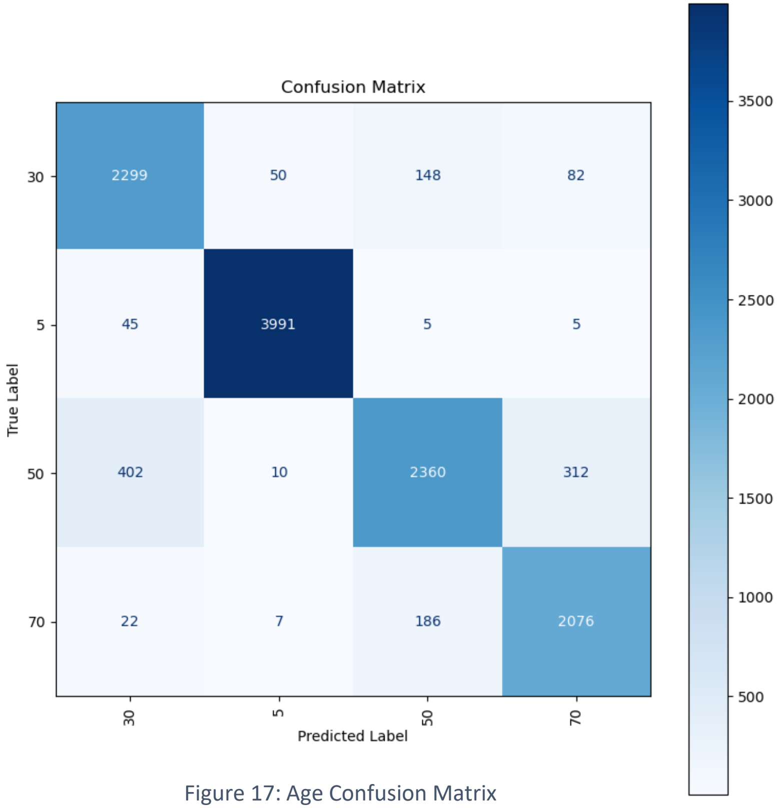


Figure 16: Age Training History Plot



	30	5	50	70
30	2299	50	148	82
5	45	3991	5	5
50	402	10	2360	312
70	22	7	186	2076

Figure 17: Age Confusion Matrix

TRAINED MODELS (1D CNN)

	ACCURACY	PRECISION	RECALL	F1	AUC
SEX_AGE	94%	93%	93%	93%	99.6%
AGE	94%	93%	93%	93%	99.3%
SEX	99.6%	99.6%	99.6%	99.6%	99.98%

BASELINE

(Majority Classifier)

ACCURACY	PRECISION	RECALL	F1	
12.1253%	1.3473%	11.1111%	2.4031%	SEX_AGE
33.7167%	8.292%	25.0%	12.6075%	AGE
50.8417%	25.4208%	50%	33.7053%	SEX

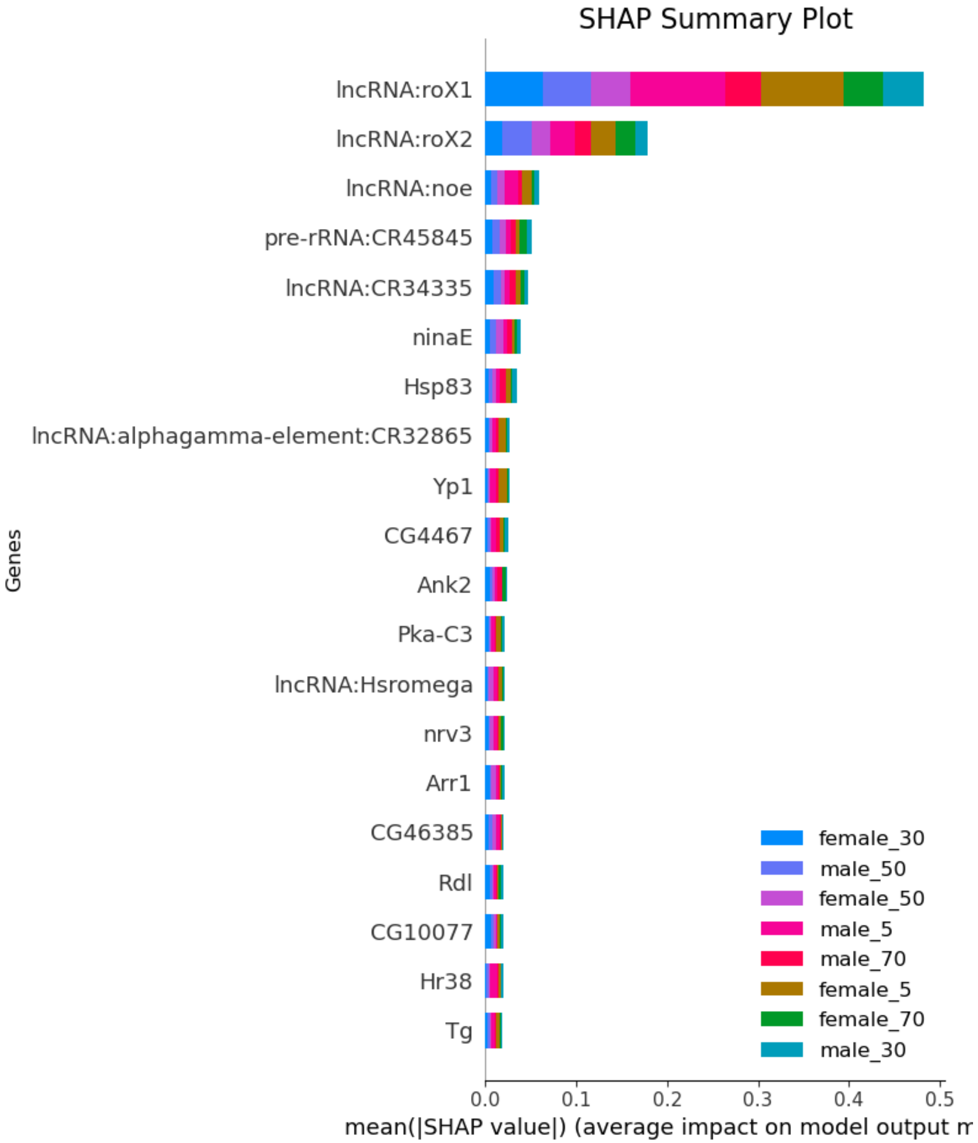
BASELINE

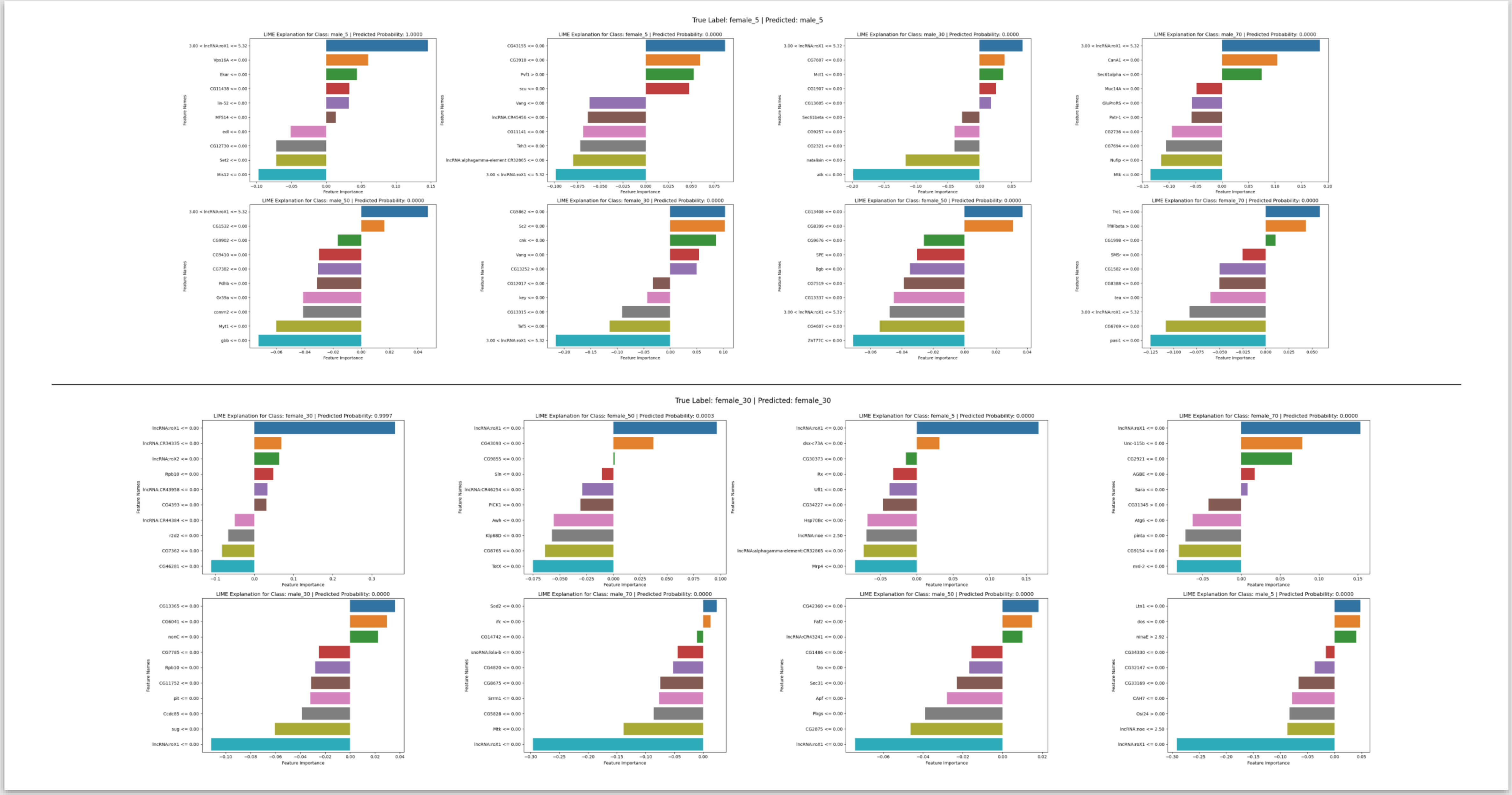
(Comparable Models)

	ACCURACY	PRECISION	RECALL	F1	AUC
XGBOOST	86%	95%	92%	93%	98.5%
Random Forest	51%	54%	39%	32%	92%
MLP	95%	94%	93%	94%	99.7%

RESULTS

Feature Importance


SHAP Summary Plot
lncRNA:roX1
lncRNA:roX2
lncRNA:noe
pre-rRNA:CR45845
lncRNA:CR34335
ninaE
Hsp83
lncRNA:alphagamma-element:CR32865
Yp1
CG4467
Ank2
Pka-C3
lncRNA:Hsromega
nrv3
Arr1
CG46385
Rdl
CG10077
Hr38
Tg
Genes
female_30
male_50
female_50
male_5
male_70
female_5
female_70
male_30
0.0
0.1
0.2
0.3
0.4
0.5
mean(|SHAP value|) (average impact on model output m



True Label: female_5 | Predicted: male_5

LIME Explanation for Class: male_5 | Predicted Probability: 1.0000

LIME Explanation for Class: female_5 | Predicted Probability: 0.0000

LIME Explanation for Class: male_30 | Predicted Probability: 0.0000

LIME Explanation for Class: male_70 | Predicted Probability: 0.0000

LIME Explanation for Class: male_50 | Predicted Probability: 0.0000

LIME Explanation for Class: female_30 | Predicted Probability: 0.0000

LIME Explanation for Class: female_50 | Predicted Probability: 0.0000

LIME Explanation for Class: female_70 | Predicted Probability: 0.0000

True Label: female_30 | Predicted: female_30

LIME Explanation for Class: female_30 | Predicted Probability: 0.9997

LIME Explanation for Class: female_50 | Predicted Probability: 0.0003

LIME Explanation for Class: female_5 | Predicted Probability: 0.0000

LIME Explanation for Class: female_70 | Predicted Probability: 0.0000

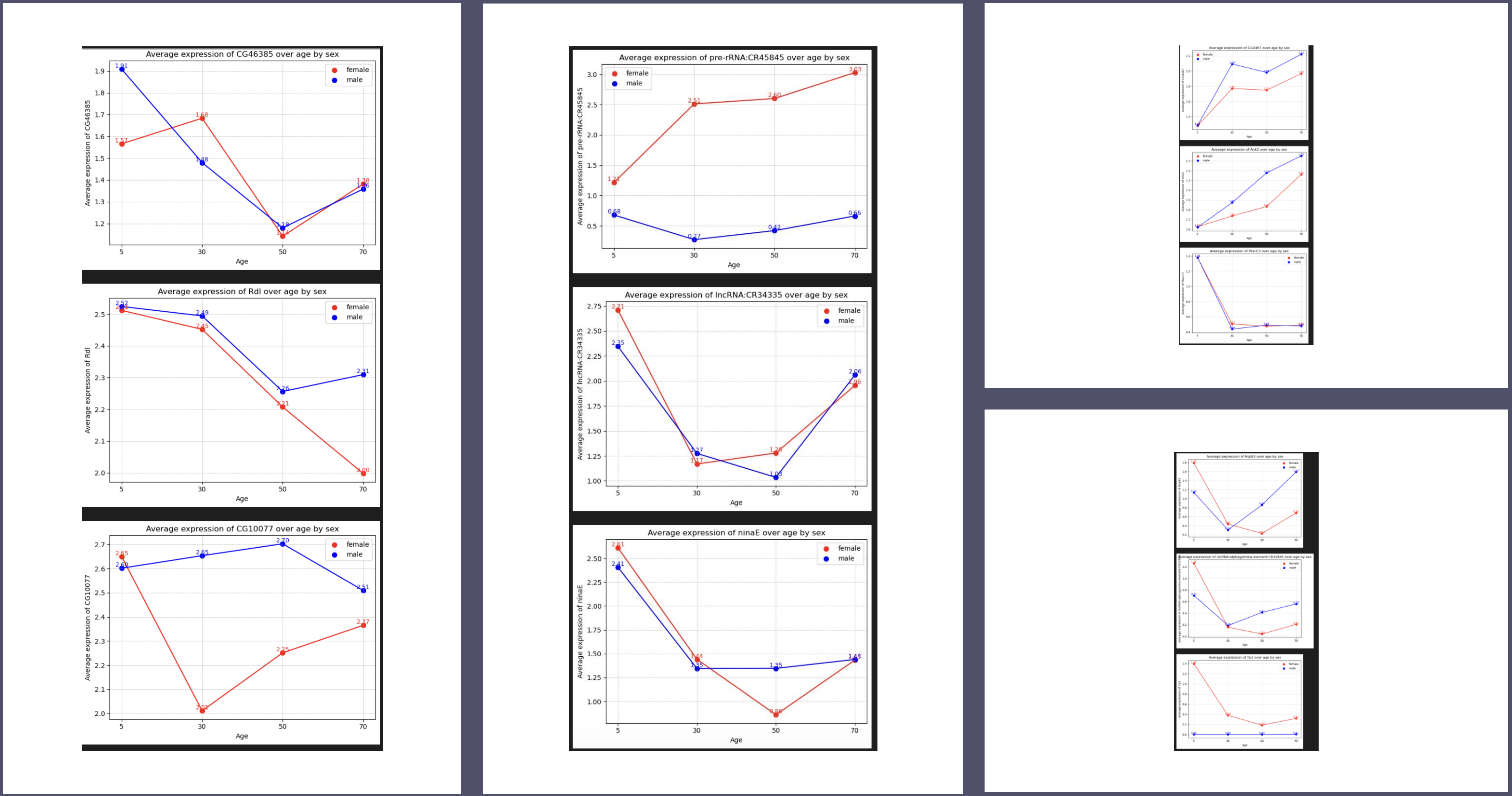
LIME Explanation for Class: male_30 | Predicted Probability: 0.0000

LIME Explanation for Class: male_70 | Predicted Probability: 0.0000

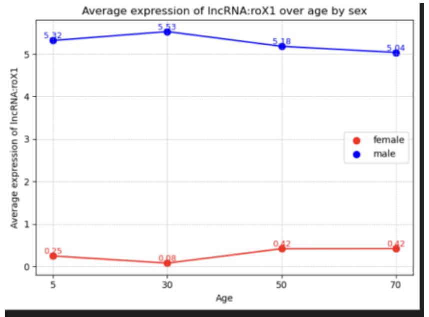
LIME Explanation for Class: male_50 | Predicted Probability: 0.0000

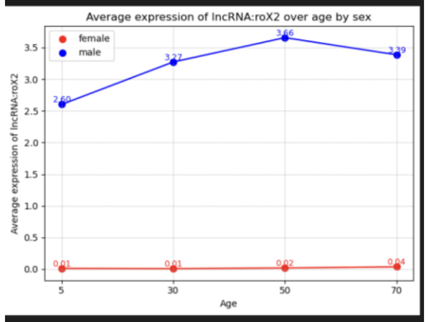
LIME Explanation for Class: male_5 | Predicted Probability: 0.0000

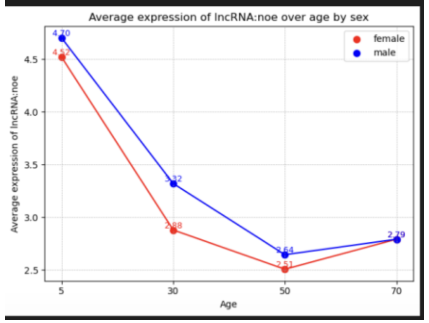
LIME Explanations (One correct and one incorrect)

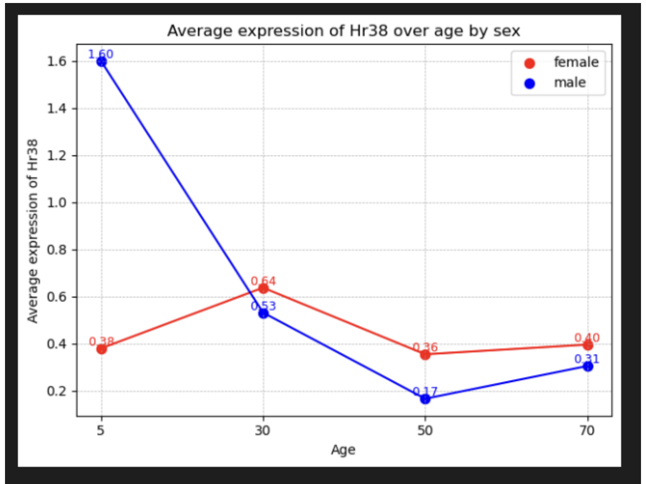


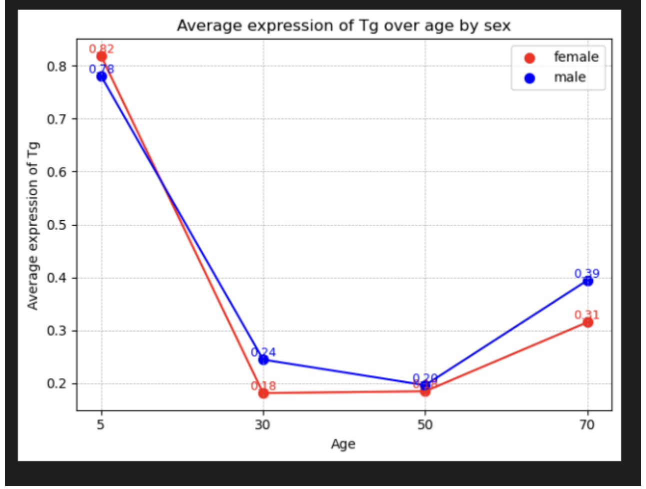
Individual plots showcasing the feature importance of the top 20 genes (sex_age), compared across all ages for both male and female expression

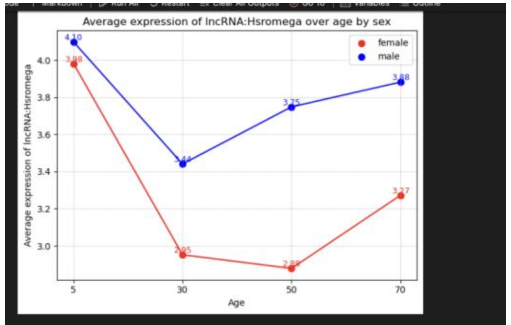

Average expression of lncRNA:roX1 over age by sex

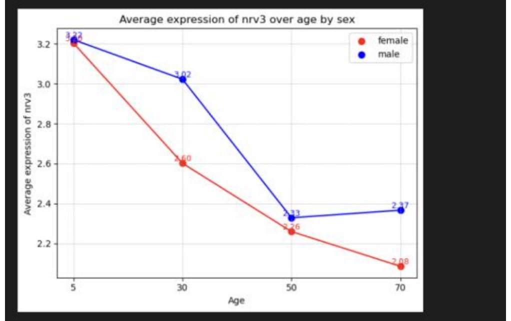

Average expression of lncRNA:roX2 over age by sex

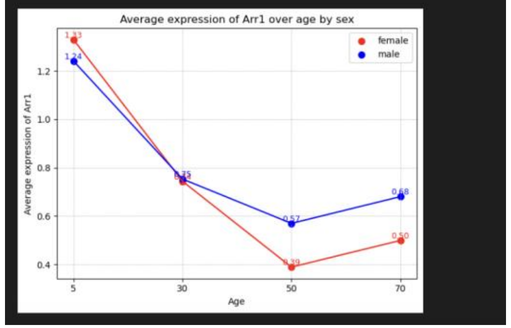

Average expression of lncRNA:noe over age by sex


Average expression of Hr38 over age by sex


Average expression of Tg over age by sex


Average expression of lncRNA:Hsromega over age by sex


Average expression of nrv3 over age by sex


Average expression of Arr1 over age by sex

NEXT STEPS

Hyperparameter Tuning

Plans to introduce automated hyperparameter tuning for better model performance despite its computational heaviness and time consumption.

Diversified Datasets

Aim to apply the model to a wider range of datasets collected by IISAGE.

Further Genomic Analysis

To discern the roles of key genes in sexual dimorphisms and ageing in flies, it's essential to delve into existing literature, analyse biological functions using tools like GO or KEGG, validate findings through lab experiments, and collaborate with genetics and fly biology experts.

PRIOR WORKS AND METHODS

- **López-Otín et al.:** Detailed the biological processes associated with aging.

- **Zhou and Troyanskaya:** DeepSEA utilizes a CNN to predict functional effects of non-coding genomic variants.

- **Ji et al.:** DNABERT, an adaptation of BERT, captures patterns in DNA sequences for genomics tasks.

- ***Avsec et al.***: Introduces Enformer, a new deep learning model, that integrates long-range genomic data to enhance gene expression prediction from DNA, aiding in disease association mapping and cis-regulatory evolution insights.

- **Tan et al.:** Investigated the use of denoising autoencoders for gene expression data.

- **Eraslan et al.:** Delved into the application of deep learning in genomics, focusing on 1D CNNs.

- **Alipanahi et al.:** Presented the DeepBind method for genomic sequence analysis.

- ***Bronikowski et al.***: Provided a comprehensive perspective on sex-specific ageing.

CONCLUSION

Key Points

- Developed a powerful pipeline to analyse high-dimensional genomics data.
- Deep learning-based 1D CNN model effectively recognized intricate data patterns, enabling precise predictions about sex and age from gene expression profiles.
- Feature Importance gives insight into which genes are most important.
- Insights generated will enhance understanding of sex-specific aging biology and underscore the efficacy of deep learning in genomics.

Project Materials

- All project materials including code, onboarding, lab materials and deliverables can be found on the Singh Lab GitHub (Branch: Nikolai’s-CNN).