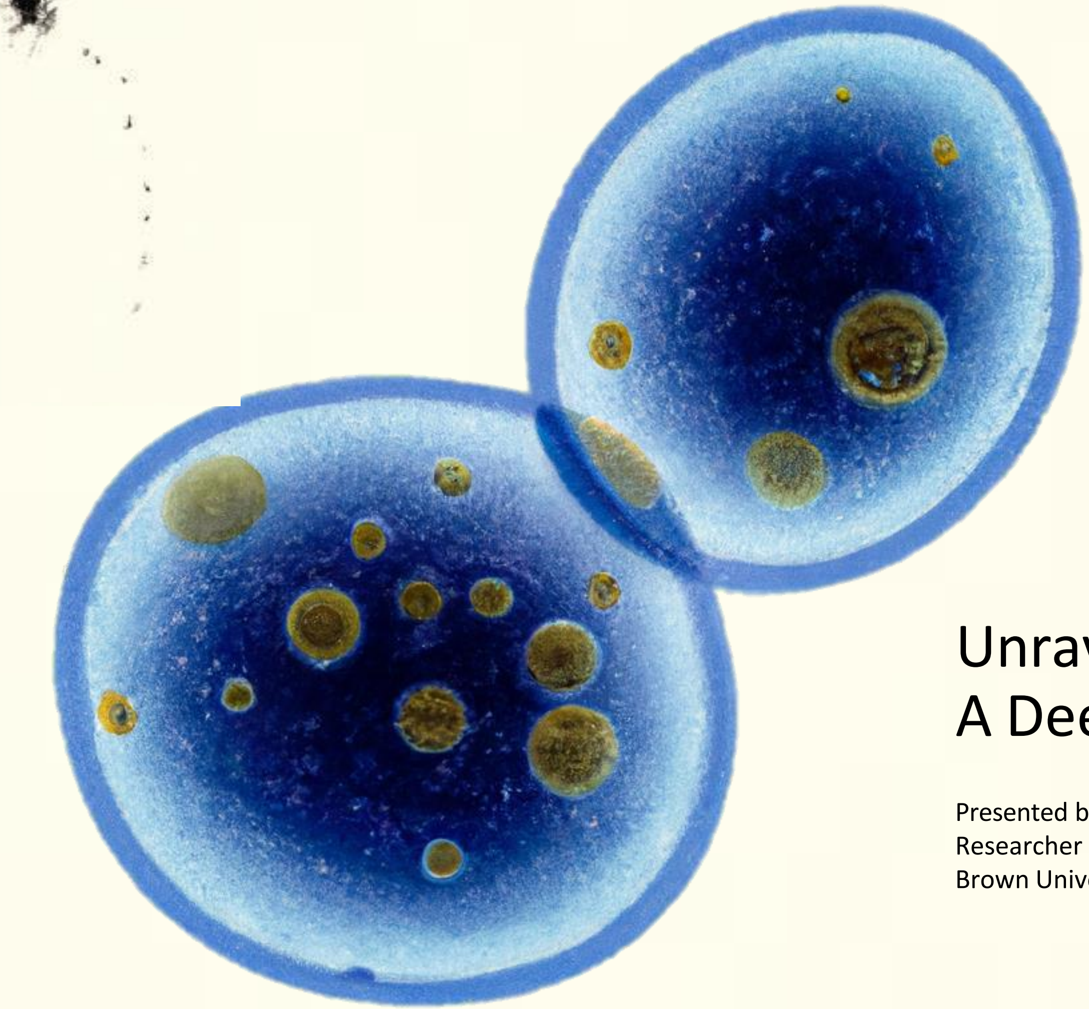
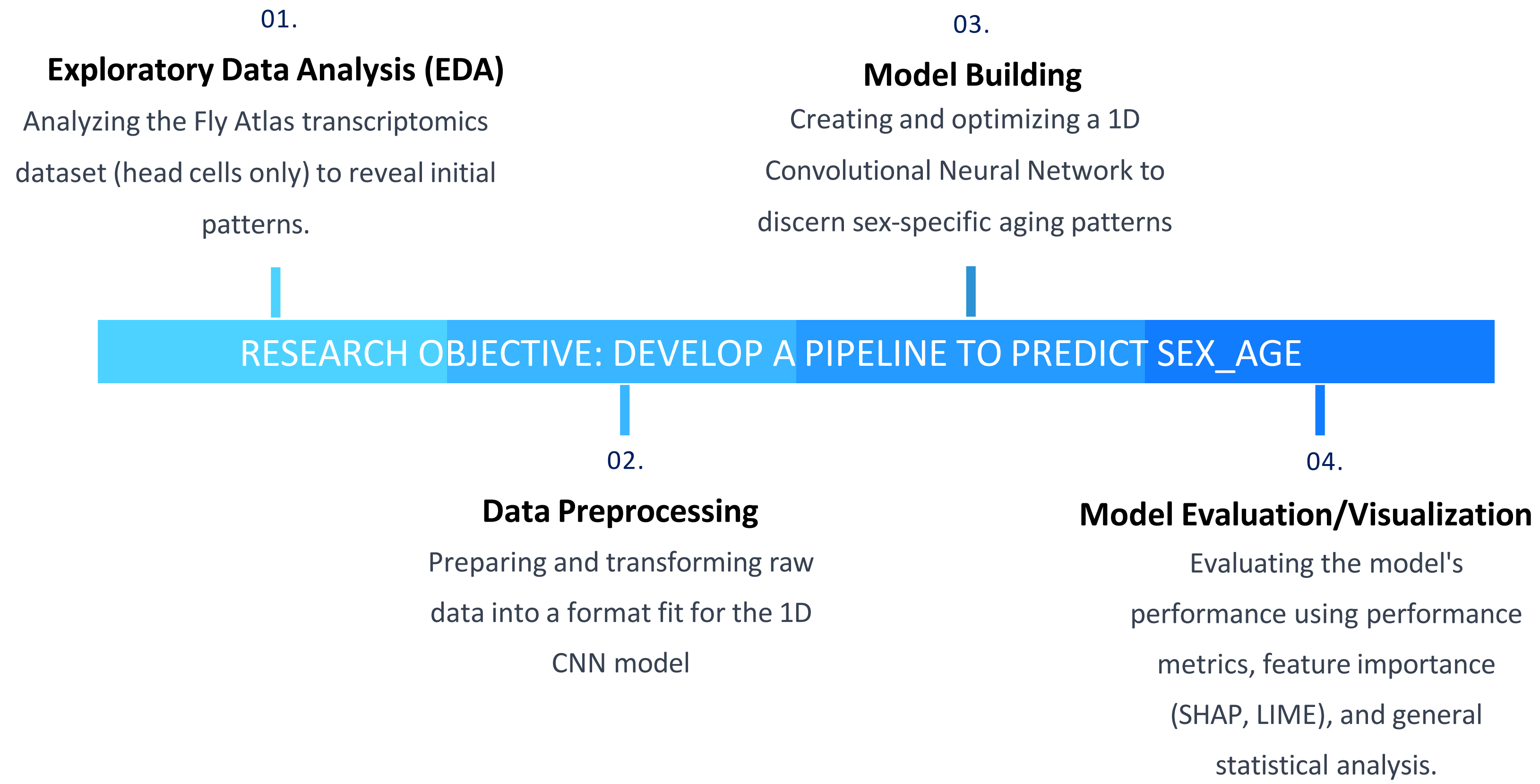


Unravelling Sex-Specific Aging on Flies: A Deep Learning Approach

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

METHODOLOGY



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.

RESEARCH OBJECTIVE: DEVELOP A PIPELINE TO PREDICT SEX_AGE

EDA

Dataset

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

- Obtained raw (UMI-based data) samples from Hongjie Li Lab

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.

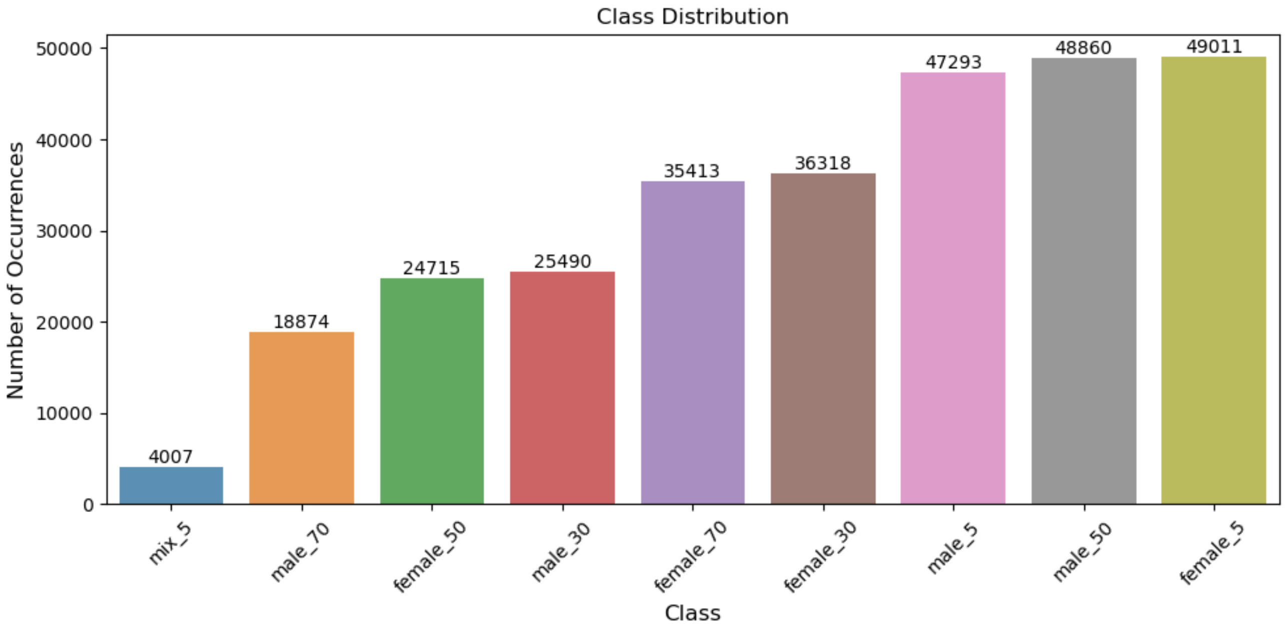


Figure 1: Bar Plot Showing Class Distributions for sex_age

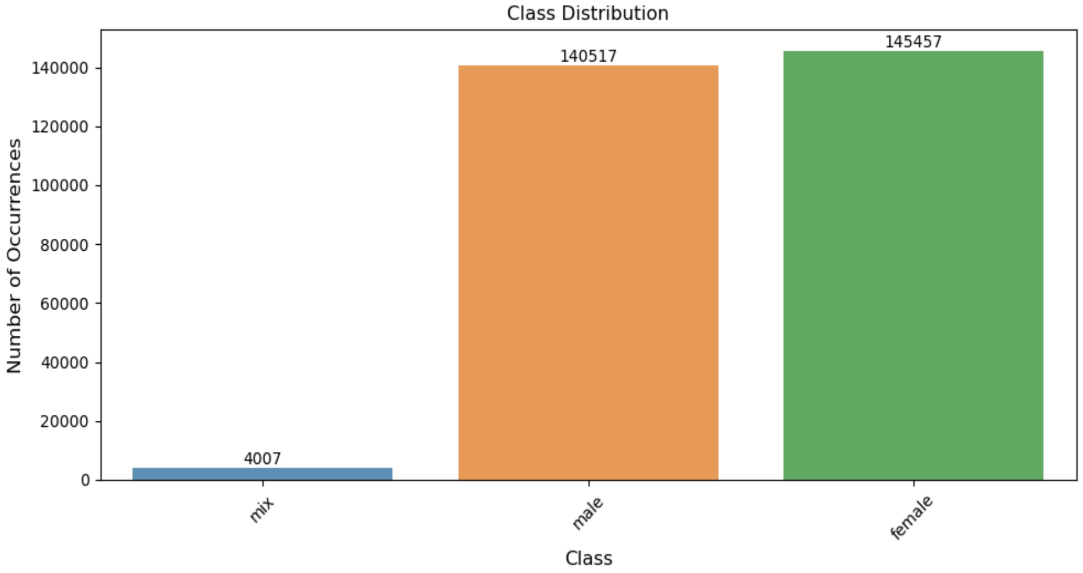


Figure 2: Bar Plot Showing Class Distributions for Sex

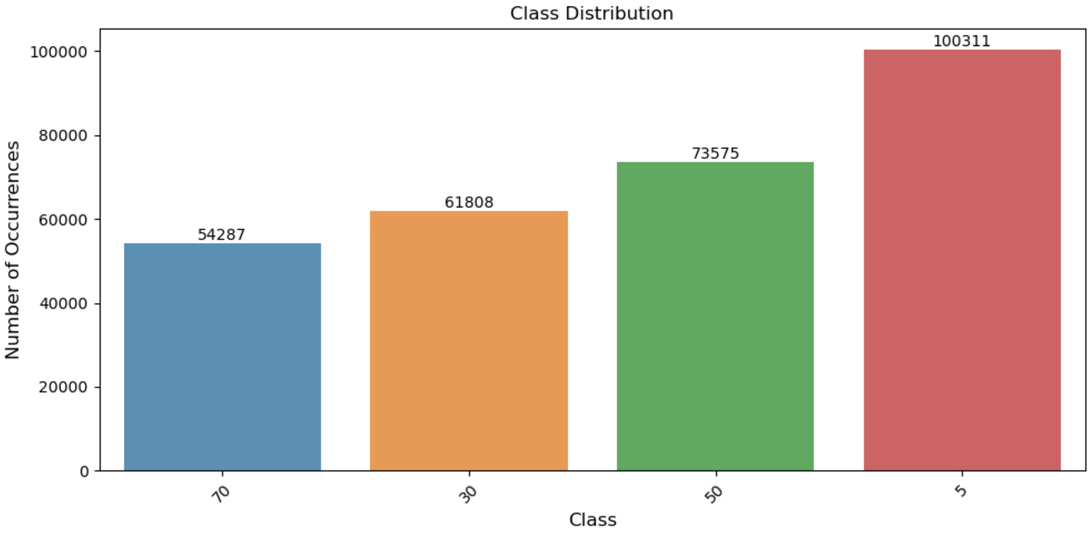
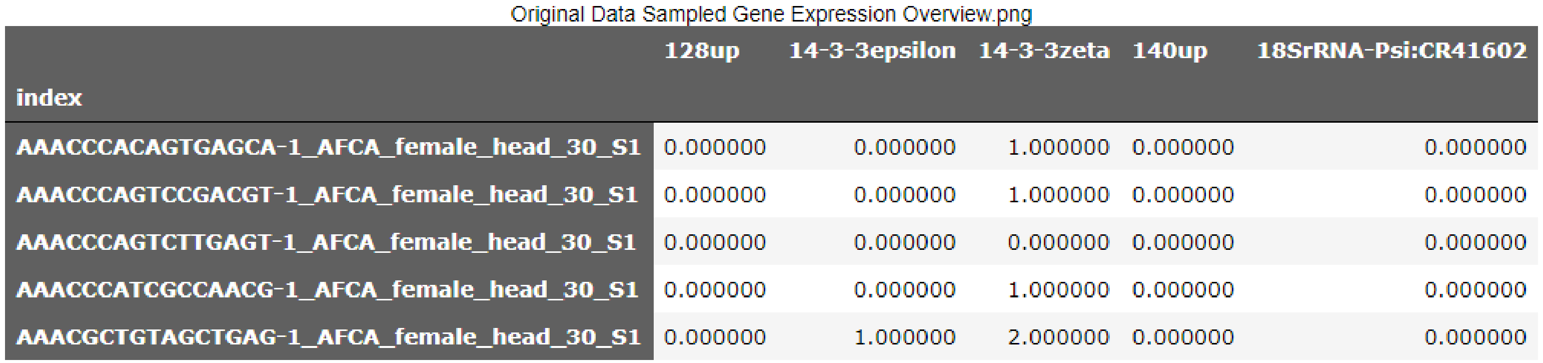


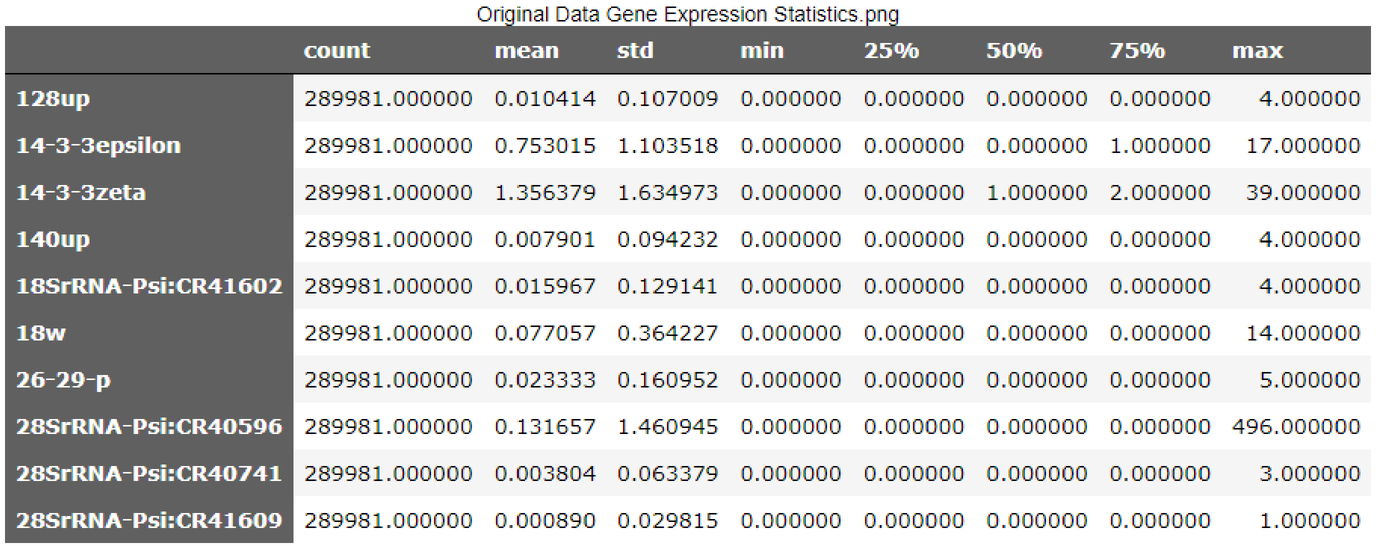
Figure 3: Bar Plot Showing Class Distributions for Age



Original Data Sampled Gene Expression Overview.png

	128up	14-3-3epsilon	14-3-3zeta	140up	18SrRNA-Psi:CR41602
index					
AAACCCACAGTGAGCA-1_AFCA_female_head_30_S1	0.000000	0.000000	1.000000	0.000000	0.000000
AAACCCAGTCCGACGT-1_AFCA_female_head_30_S1	0.000000	0.000000	1.000000	0.000000	0.000000
AAACCCAGTCTTGAGT-1_AFCA_female_head_30_S1	0.000000	0.000000	0.000000	0.000000	0.000000
AAACCCATCGCCAACG-1_AFCA_female_head_30_S1	0.000000	0.000000	1.000000	0.000000	0.000000
AAACGCTGTAGCTGAG-1_AFCA_female_head_30_S1	0.000000	1.000000	2.000000	0.000000	0.000000

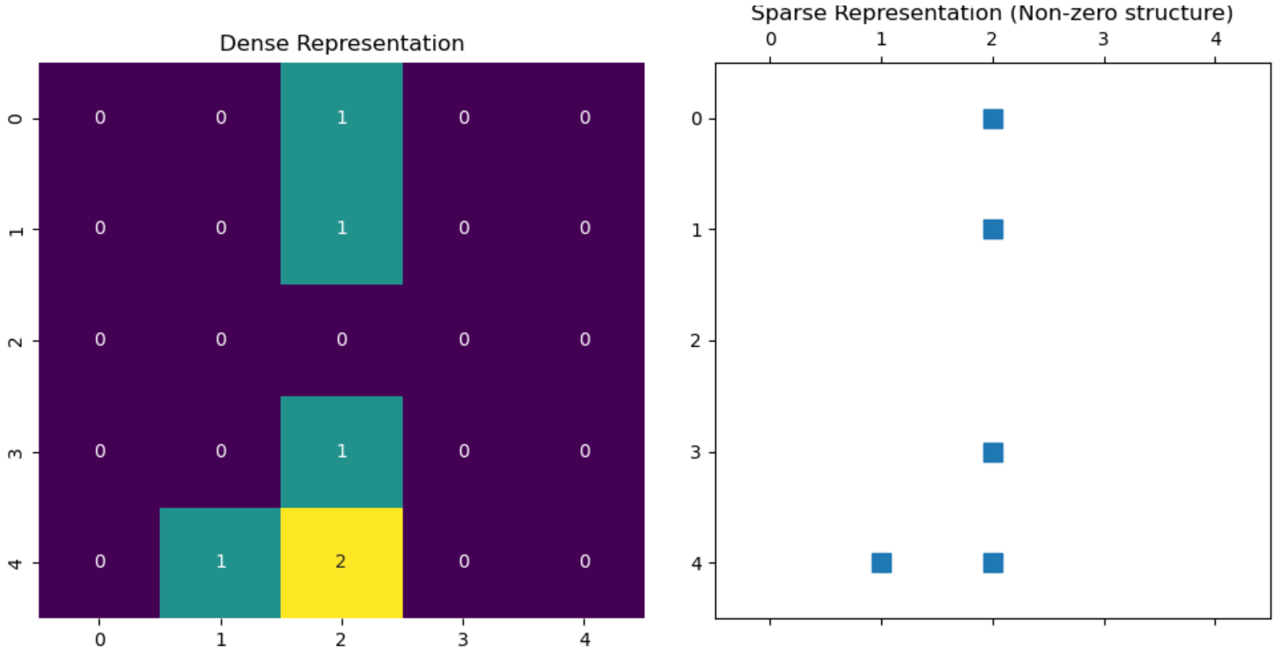
Figure 4: Gene Expression Matrix for First Five Cells



Original Data Gene Expression Statistics.png

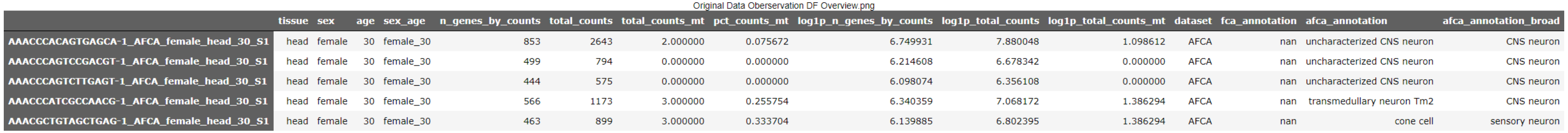
	count	mean	std	min	25%	50%	75%	max
128up	289981.000000	0.010414	0.107009	0.000000	0.000000	0.000000	0.000000	4.000000
14-3-3epsilon	289981.000000	0.753015	1.103518	0.000000	0.000000	0.000000	1.000000	17.000000
14-3-3zeta	289981.000000	1.356379	1.634973	0.000000	0.000000	1.000000	2.000000	39.000000
140up	289981.000000	0.007901	0.094232	0.000000	0.000000	0.000000	0.000000	4.000000
18SrRNA-Psi:CR41602	289981.000000	0.015967	0.129141	0.000000	0.000000	0.000000	0.000000	4.000000
18w	289981.000000	0.077057	0.364227	0.000000	0.000000	0.000000	0.000000	14.000000
26-29-p	289981.000000	0.023333	0.160952	0.000000	0.000000	0.000000	0.000000	5.000000
28SrRNA-Psi:CR40596	289981.000000	0.131657	1.460945	0.000000	0.000000	0.000000	0.000000	496.000000
28SrRNA-Psi:CR40741	289981.000000	0.003804	0.063379	0.000000	0.000000	0.000000	0.000000	3.000000
28SrRNA-Psi:CR41609	289981.000000	0.000890	0.029815	0.000000	0.000000	0.000000	0.000000	1.000000

Figure 5: Statistical Summary of First 10 Genes


Dense Representation

Sparse Representation (Non-zero structure)

Figure 6: Dense vs Sparse Matrix Transformation Comparison


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	2.000000	0.075672	6.749931	7.880048	1.098612	AFCA	nan	uncharacterized CNS neuron	CNS neuron
AAACCCAGTCCGACGT-1_AFCA_female_head_30_S1	head	female	30	female_30	499	794	0.000000	0.000000	6.214608	6.678342	0.000000	AFCA	nan	uncharacterized CNS neuron	CNS neuron
AAACCCAGTCTTGAGT-1_AFCA_female_head_30_S1	head	female	30	female_30	444	575	0.000000	0.000000	6.098074	6.356108	0.000000	AFCA	nan	uncharacterized CNS neuron	CNS neuron
AAACCCATCGCCAACG-1_AFCA_female_head_30_S1	head	female	30	female_30	566	1173	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	3.000000	0.333704	6.139885	6.802395	1.386294	AFCA	nan	cone cell	sensory neuron

Figure 7: Observation DF Overview

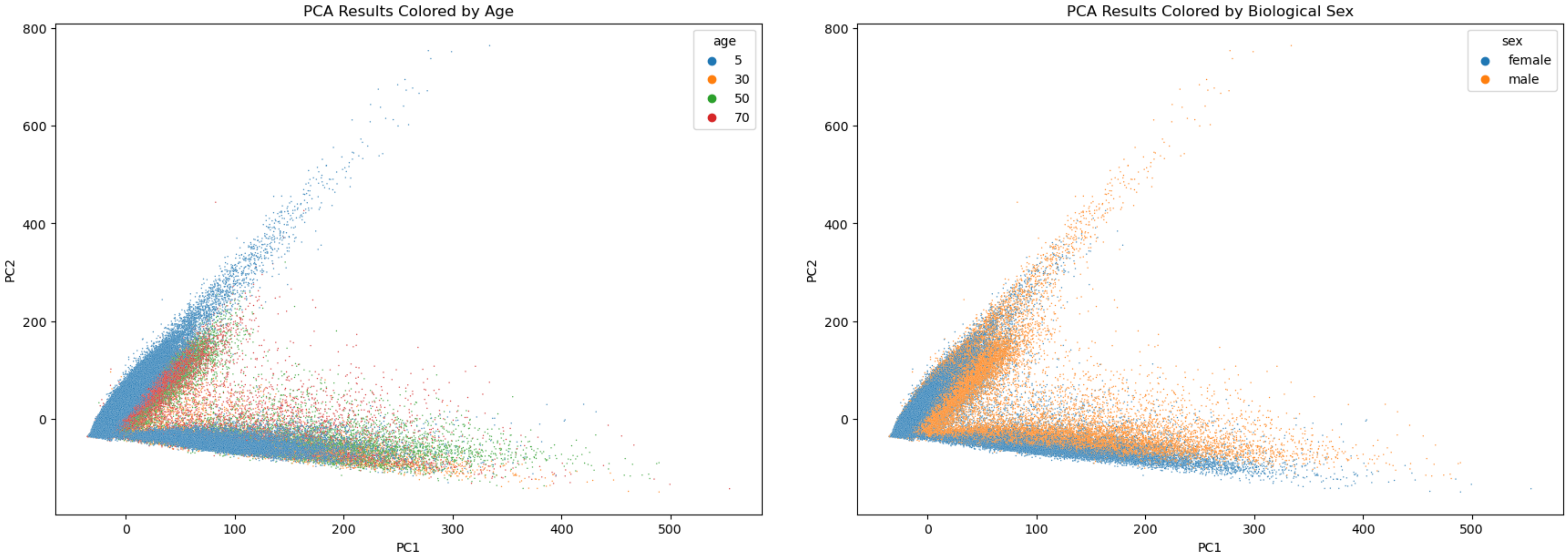


Figure 8: PCA scatter plots assessing the impact of biological sex and age on dataset variability, aimed at distinguishing significant biological patterns from noise.

PREPROCESSING

Initial Data Split

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

Data Handling

- Selective handling of categories and properties, such as 'mix' category exclusion and gene shuffling.
- Further customizable by removing certain genes (for example, autosomal, non-autosomal, lnc, or any combination therfore)

Stratified Subset Selection (Skipped)

- Selected a manageable stratified subset of 200,000 cells from the training dataset.

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 performance reduction.

Data Saving and Loading

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

MODEL BUILDING AND TRAINING

Configurable Model Building

- Model construction guided by a flexible configuration file.

Model Construction

- Three base models built: 'sex', 'age', and 'sex_age'.
- Further variations built by removing certain genes (for example, autosomal, non-autosomal, lnc, or any combination)
- Type: 1D CNN
- Architecture: Convolutional blocks, flattening layer, fully connected layers, and output layer.

Optimization and Compilation

- 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.

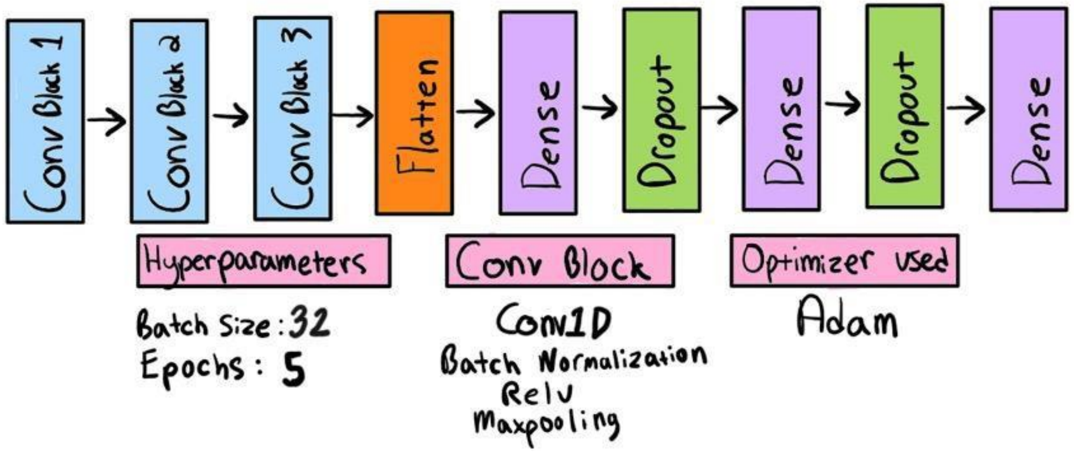
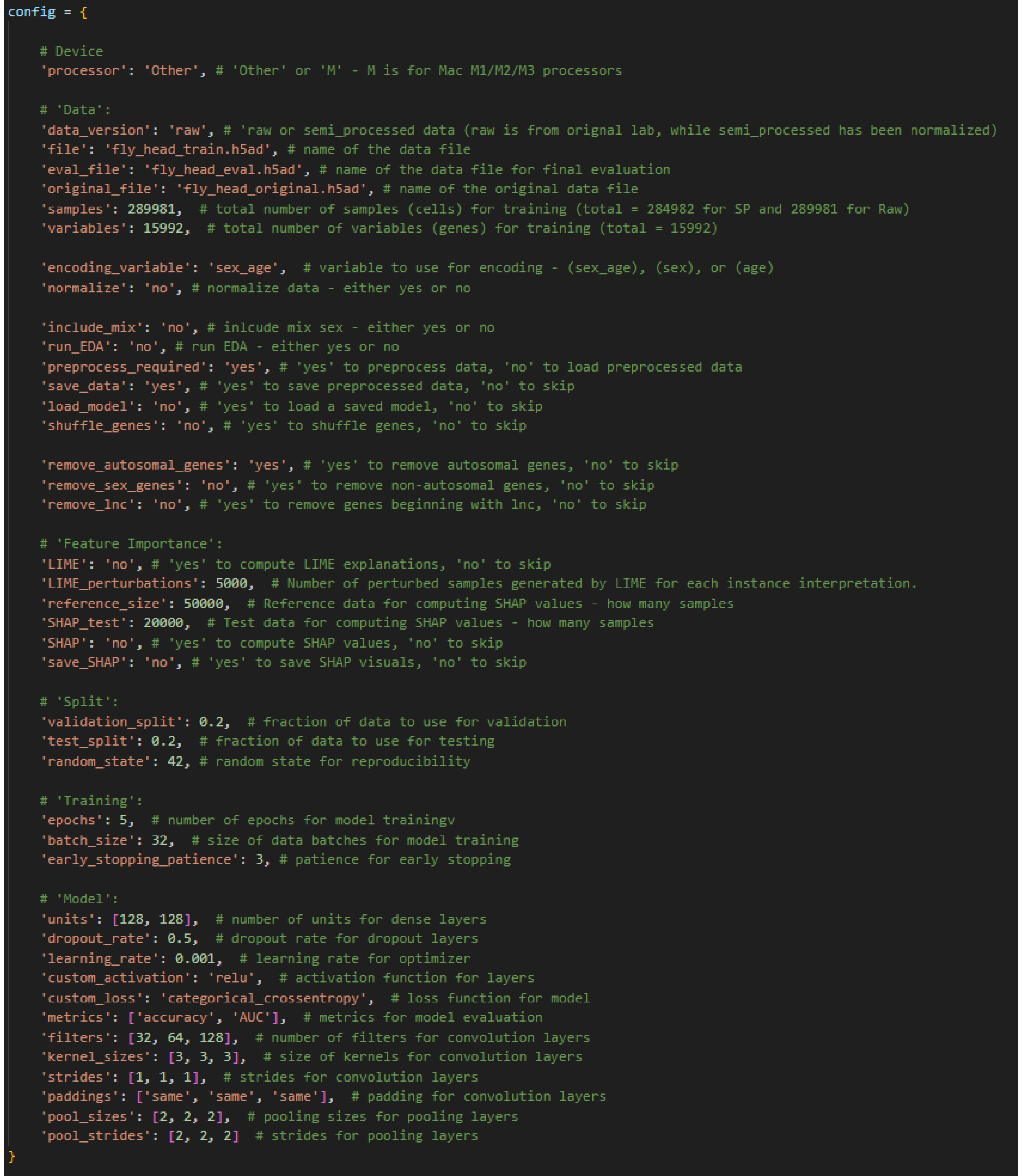


Figure 9: Convolution Neural Network Architecture



```
config = {

    # Device
    'processor': 'Other', # 'Other' or 'M' - M is for Mac M1/M2/M3 processors

    # 'Data':
    'data_version': 'raw', # 'raw or semi_processed data (raw is from orignal lab, while semi_processed has been normalized)
    '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': 289981,  # total number of samples (cells) for training (total = 284982 for SP and 289981 for Raw)
    '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)
    'normalize': 'no', # normalize data - either yes or no

    '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': 'yes', # 'yes' to save preprocessed data, 'no' to skip
    'load_model': 'no', # 'yes' to load a saved model, 'no' to skip
    'shuffle_genes': 'no', # 'yes' to shuffle genes, 'no' to skip

    'remove_autosomal_genes': 'yes', # 'yes' to remove autosomal genes, 'no' to skip
    'remove_sex_genes': 'no', # 'yes' to remove non-autosomal genes, 'no' to skip
    'remove_lnc': 'no', # 'yes' to remove genes beginning with lnc, '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': 50000,  # Reference data for computing SHAP values - how many samples
    'SHAP_test': 20000,  # 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

    # 'Training':
    'epochs': 5,  # number of epochs for model trainingv
    '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
}
```

Figure 10 : Config Dictionary

RESULTS

Metrics

SEX AGE

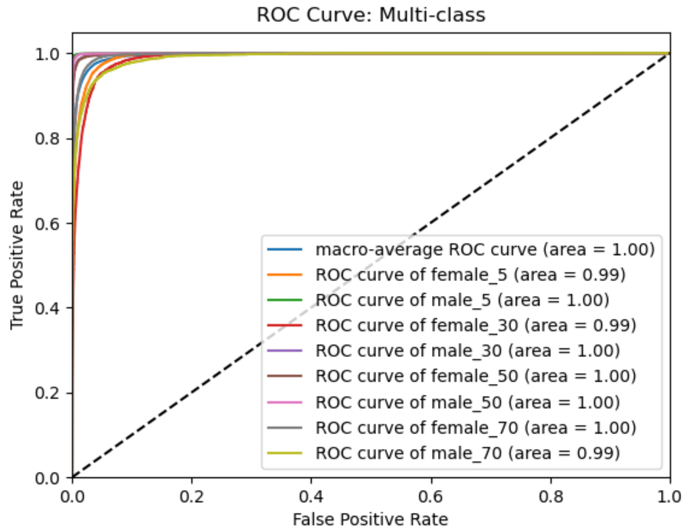


Figure 11: sex_age ROC Curve

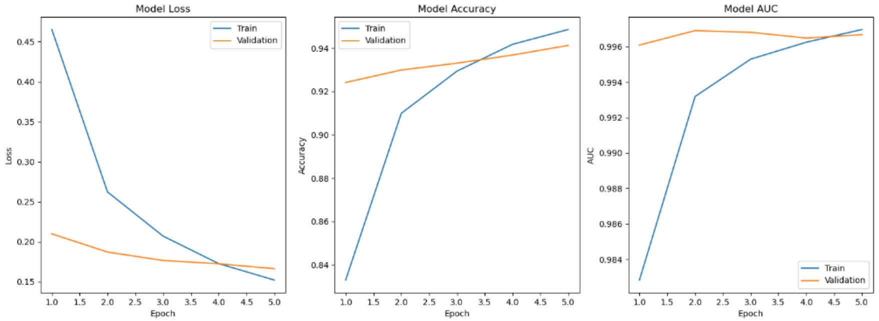
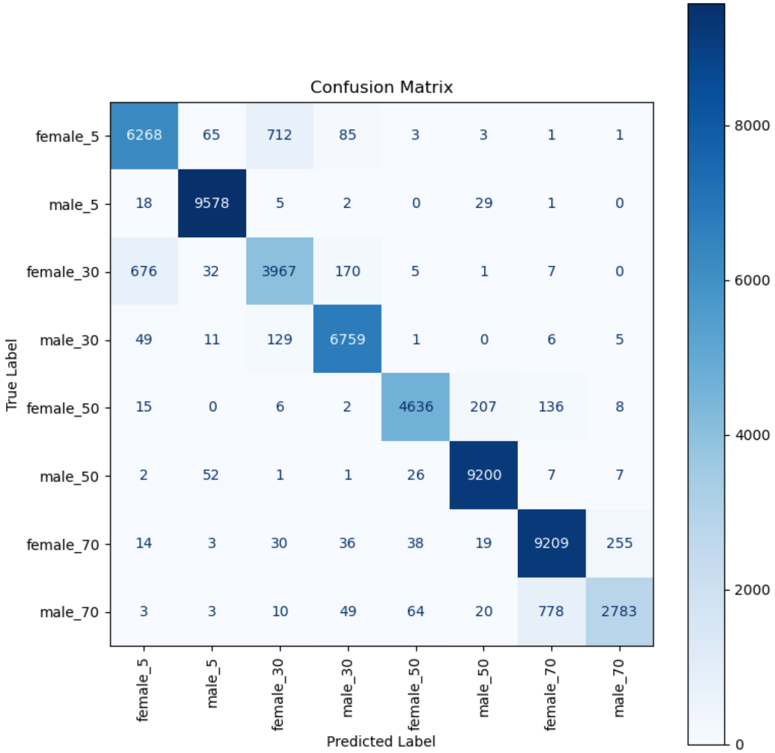


Figure 12: sex_age Training History Plot



	female_5	male_5	female_30	male_30	female_50	male_50	female_70	male_70
female_5	6268	65	712	85	3	3	1	1
male_5	18	9578	5	2	0	29	1	0
female_30	676	32	3967	170	5	1	7	0
male_30	49	11	129	6759	1	0	6	5
female_50	15	0	6	2	4636	207	136	8
male_50	2	52	1	1	26	9200	7	7
female_70	14	3	30	36	38	19	9209	255
male_70	3	3	10	49	64	20	778	2783

Figure 13: sex_age Confusion Matrix

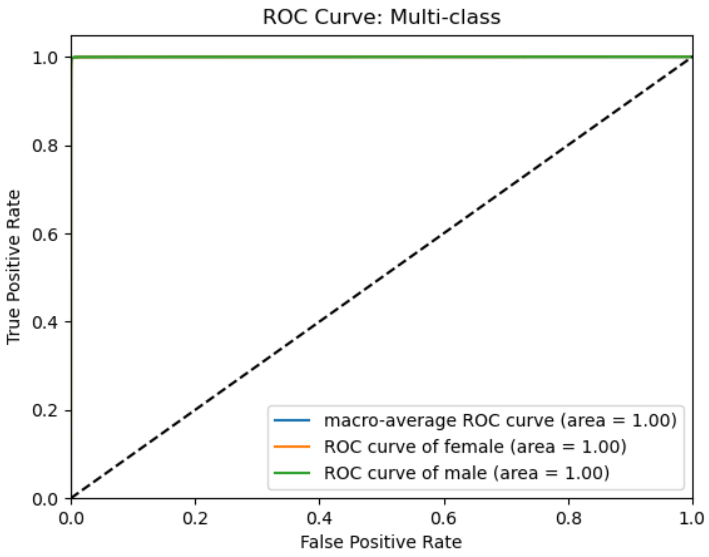


Figure 14: Sex ROC Curve

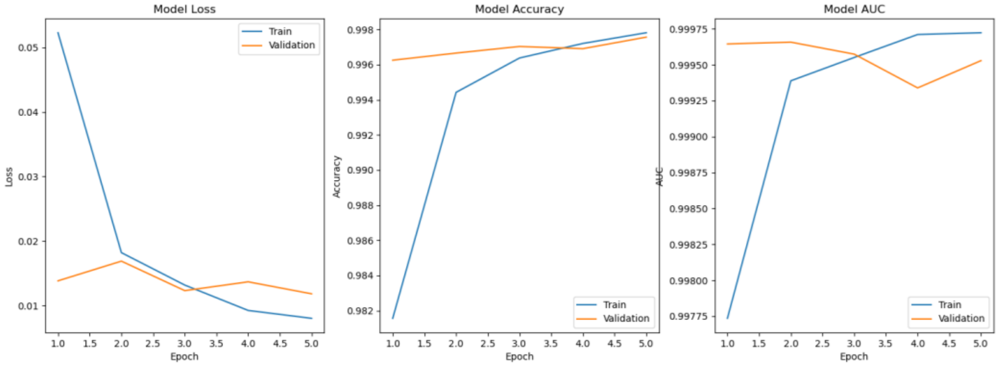


Figure 15: Sex Training History Plot

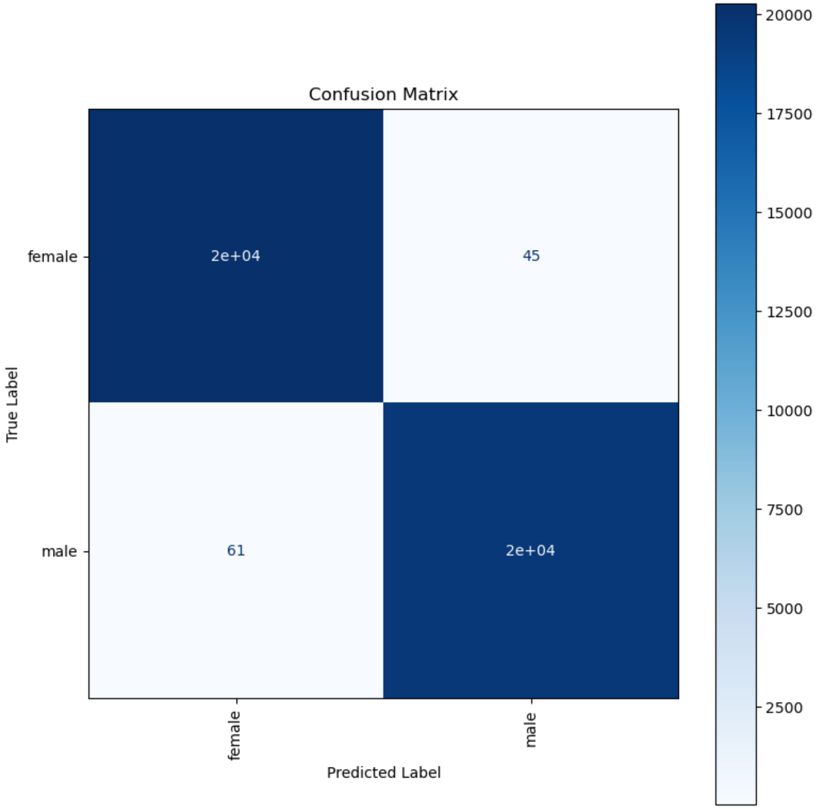


Figure 16: Sex Confusion Matrix

SEX

AGE

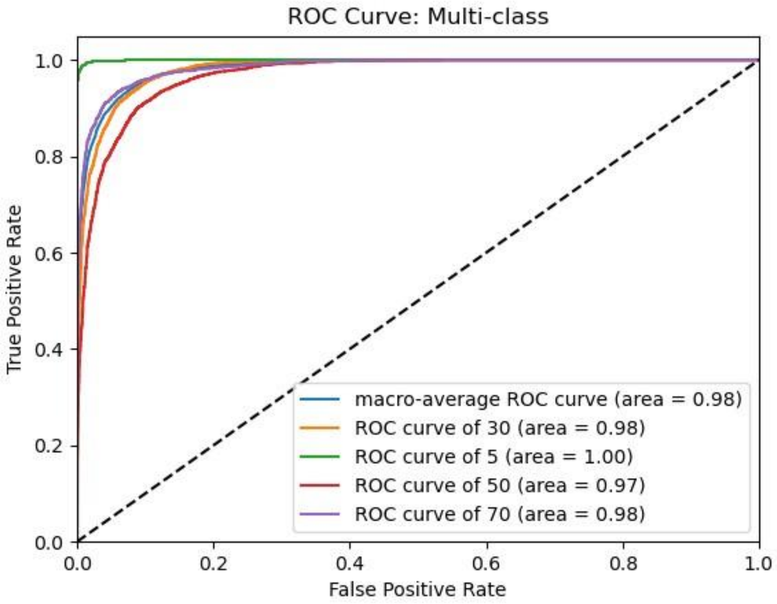


Figure 17: Age ROC Curve

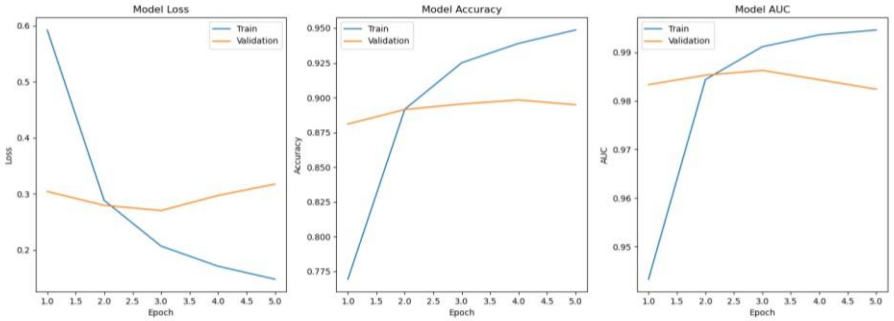


Figure 18: Age Training History Plot

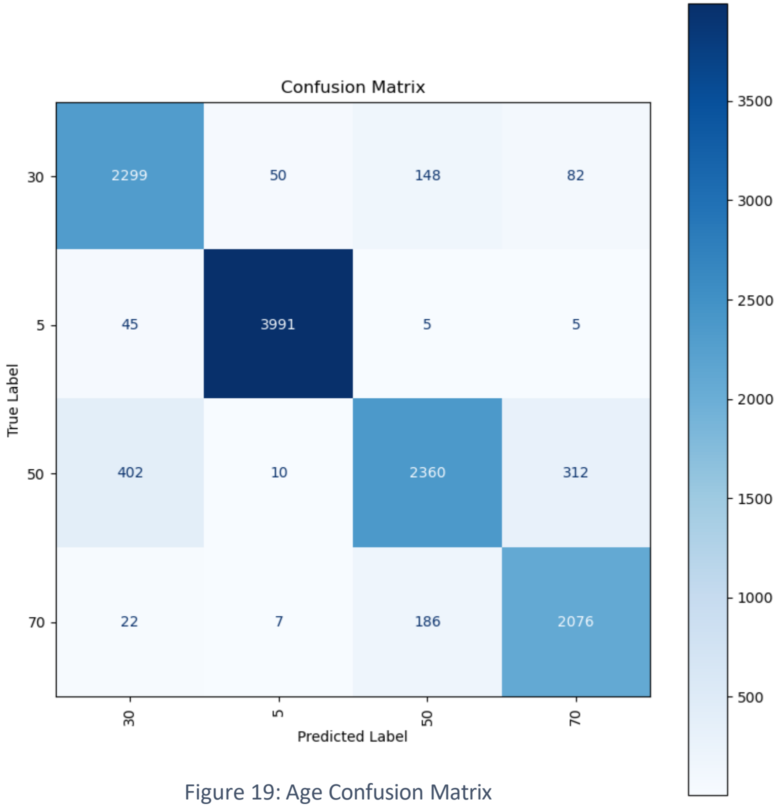


Figure 19: Age Confusion Matrix

TRAINED MODELS (1D CNN)

	ACCURACY	PRECISION	RECALL	F1	AUC
SEX _AGE	94%	93%	93%	93%	99.65%
AGE	94%	94%	94%	94%	99.44%
SEX	99.7%	99.7%	99.7%	99.7%	99.97%

BASELINE

(Majority Classifier)

ACCURACY	PRECISION	RECALL	F1	
17.14%	2.14%	12.50%	3.66%	SEX_AGE
33.67%	8.42%	25%	12.5%	AGE
50.87%	25.44%	50%	33.72%	SEX

BASELINE

(Comparable Models)

	ACCURACY	PRECISION	RECALL	F1	AUC
XGBOOST	87%	86%	83%	84%	98.6%
Random Forest	51%	53%	39%	31%	90%
MLP	94%	93%	92%	93%	99.58%

RESULTS

Feature Importance

SEX AGE

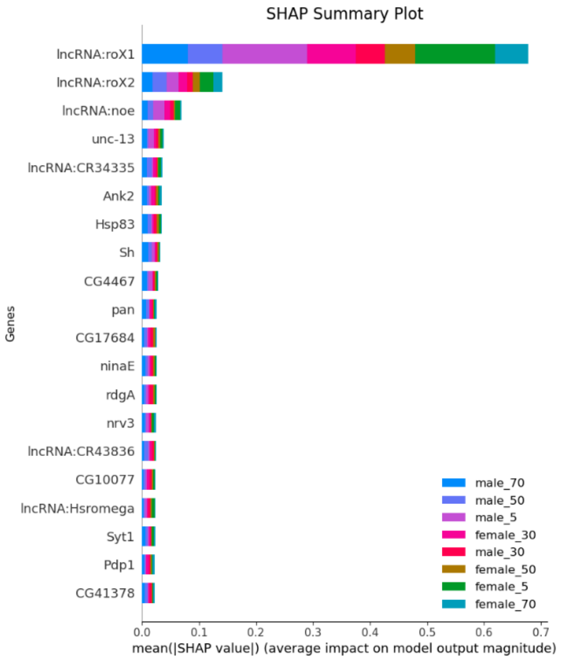


Figure 20 : Sex_Age SHAP Summary Plot – All Genes (ACC: 94%)

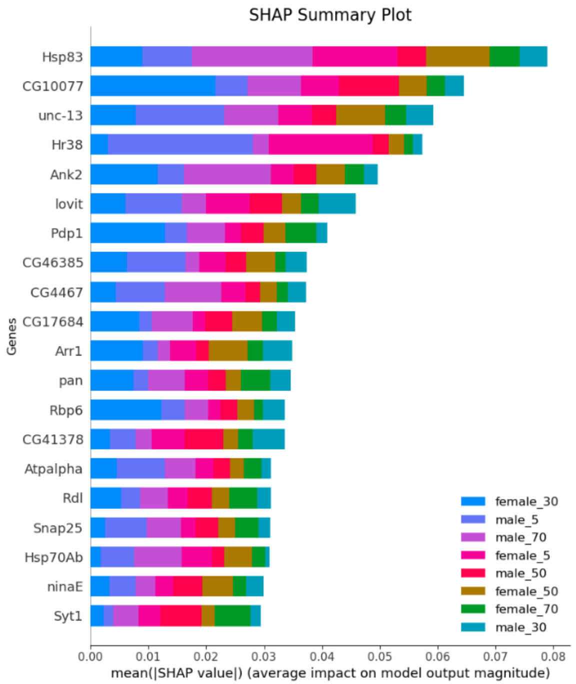


Figure 21 : Sex_Age SHAP Summary Plot – Only Autosomal Genes (ACC: 85%)

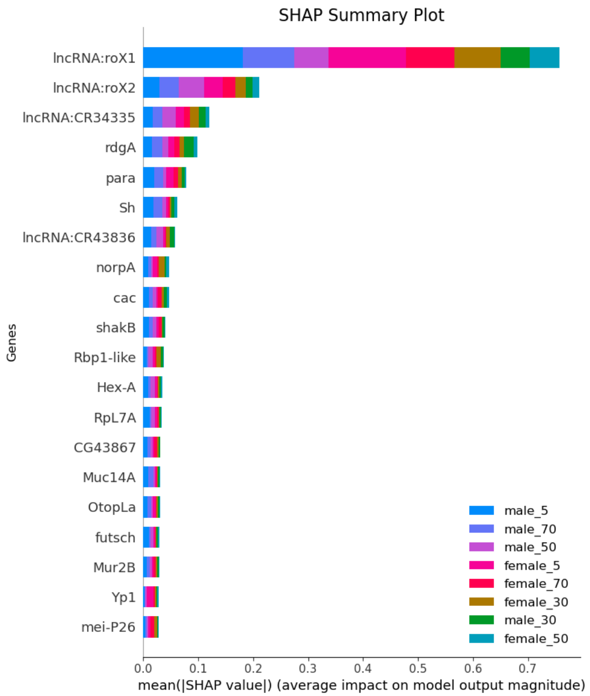
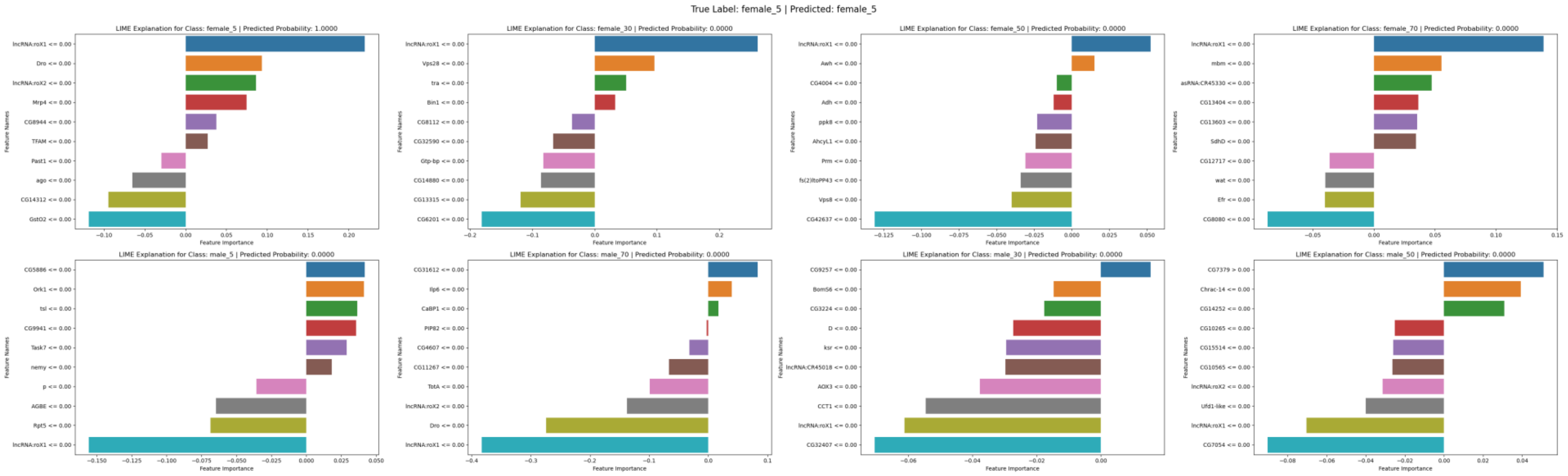
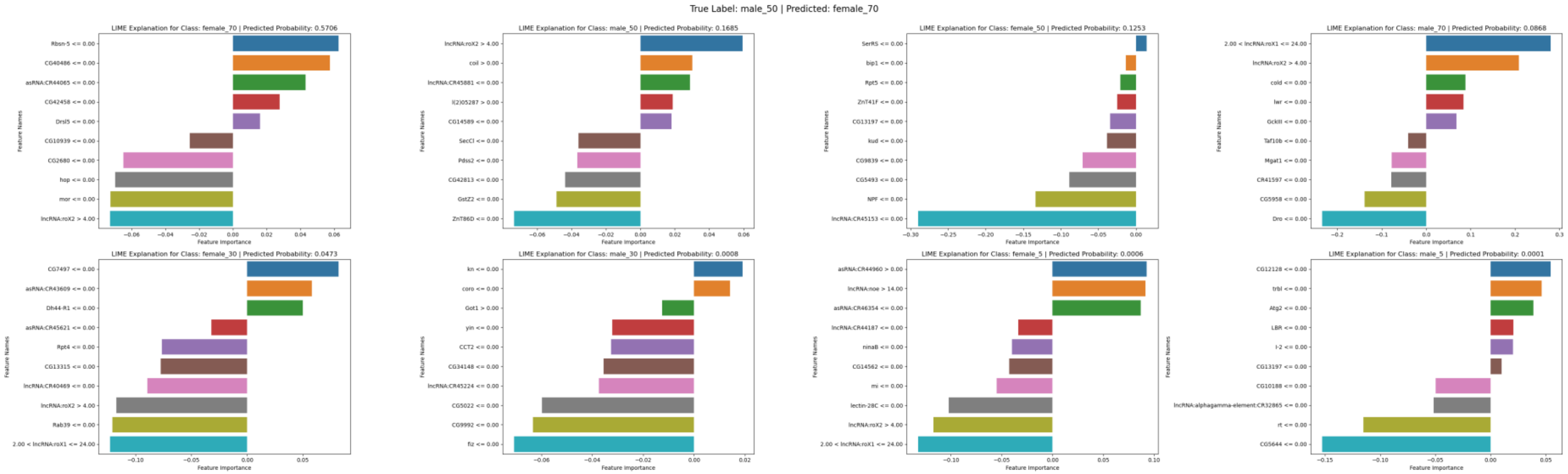


Figure 22 : Sex_Age SHAP Summary Plot – Only lnc and non-austosomal Genes (ACC: 77%)





Figures 23 & 24: LIME Explanations (T: correct, B: incorrect)

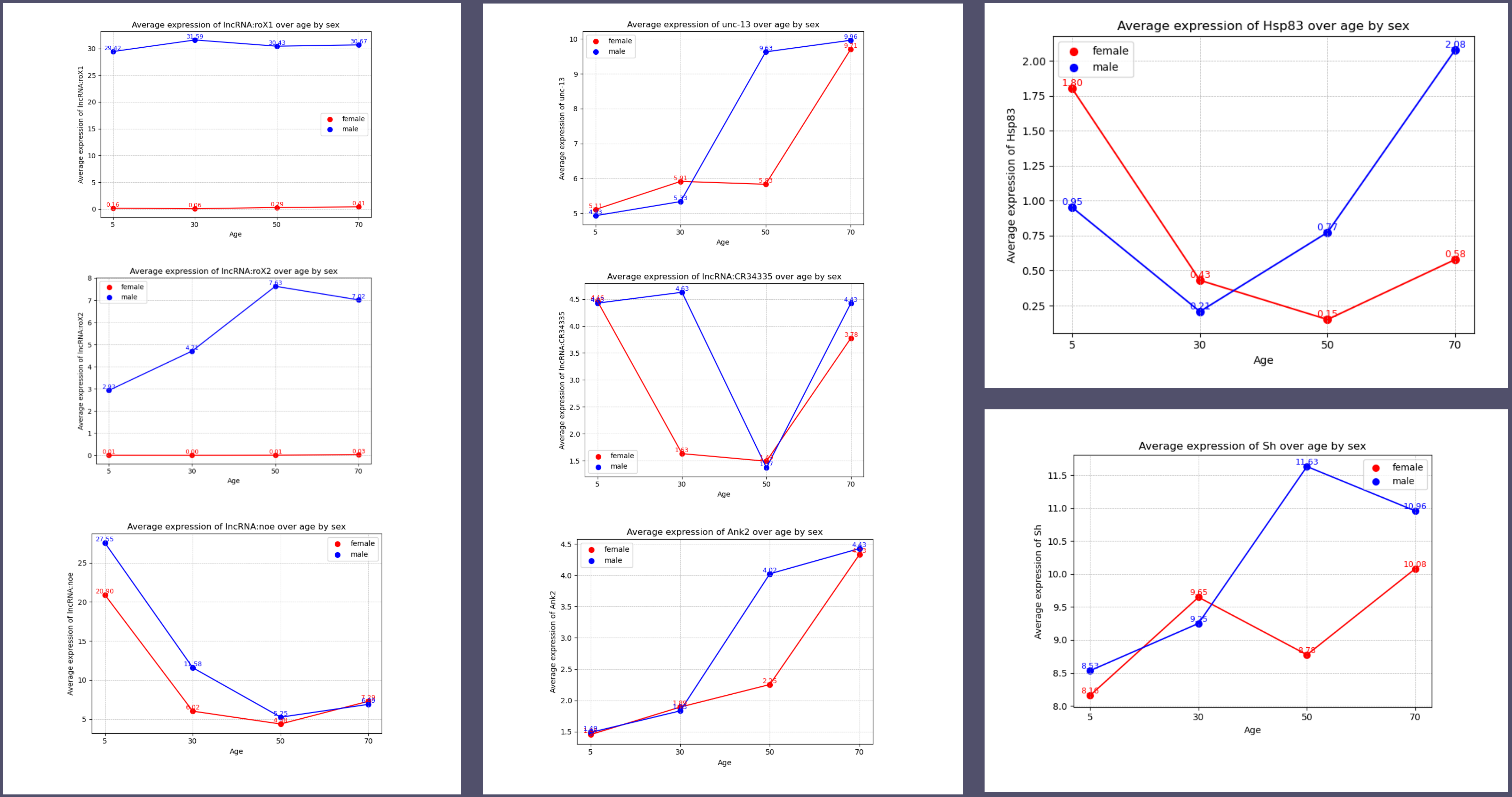
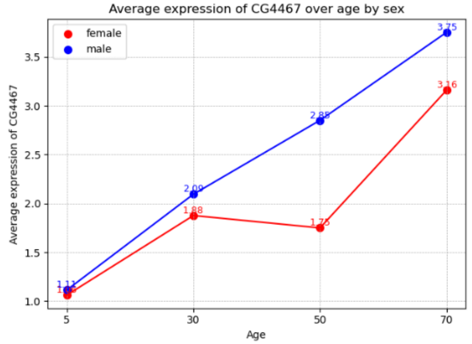
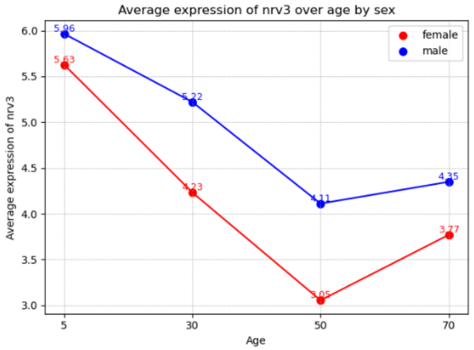
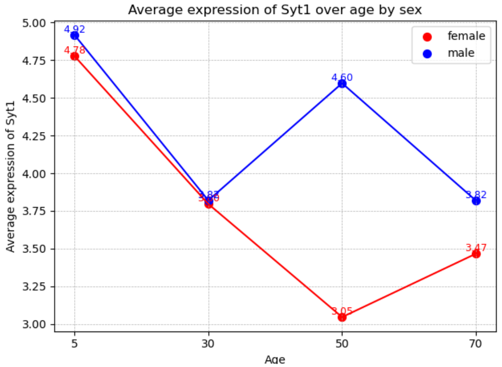
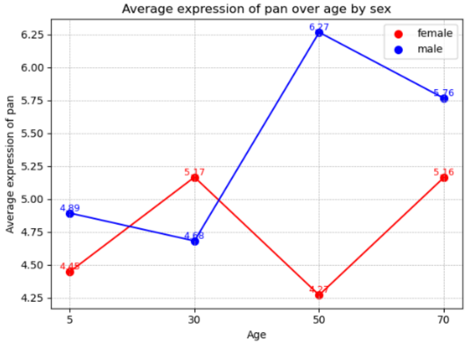


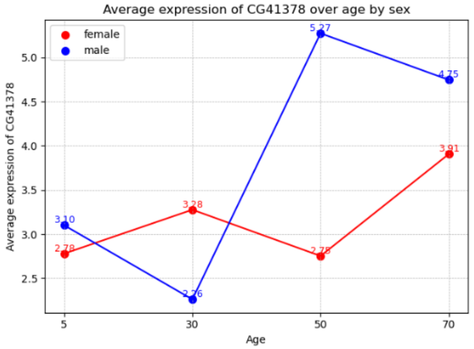
Figure 25: Differential expression of the top 20 SHAP-identified features across age groups by sex.

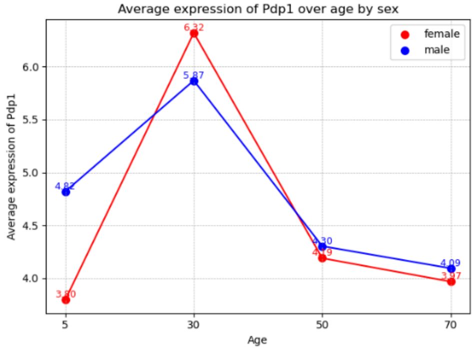

Average expression of CG4467 over age by sex
female
male
Average expression of CG4467
Age
1.0
1.5
2.0
2.5
3.0
3.5
5
30
50
70
1.11
1.06
2.09
1.88
2.85
1.75
3.75
3.16

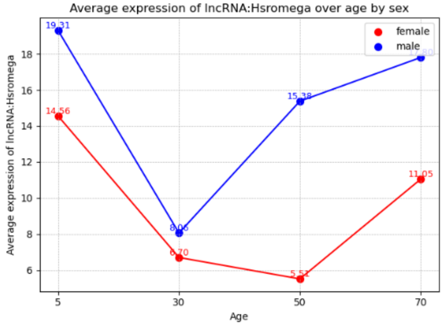

Average expression of nrv3 over age by sex
female
male
Average expression of nrv3
Age
3.0
3.5
4.0
4.5
5.0
5.5
6.0
5
30
50
70
5.96
5.63
5.22
4.23
4.11
3.05
4.35
3.77

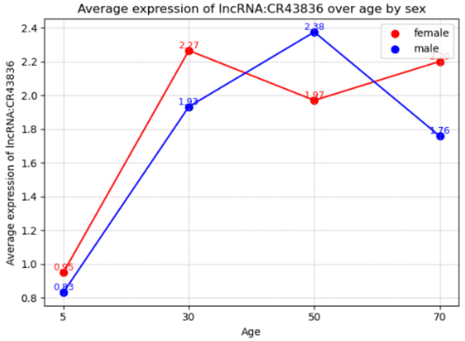

Average expression of Syt1 over age by sex
female
male
Average expression of Syt1
Age
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00
5
30
50
70
4.92
4.78
3.82
3.80
4.60
3.05
3.82
3.47

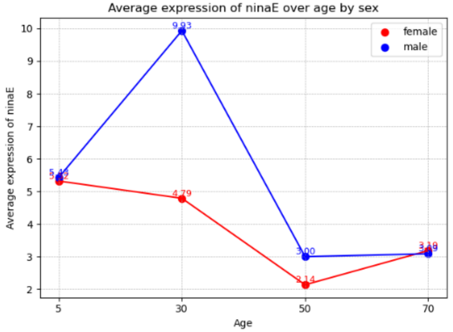

Average expression of pan over age by sex
female
male
Average expression of pan
Age
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
5
30
50
70
4.89
4.45
5.17
4.68
6.27
4.27
5.76
5.16

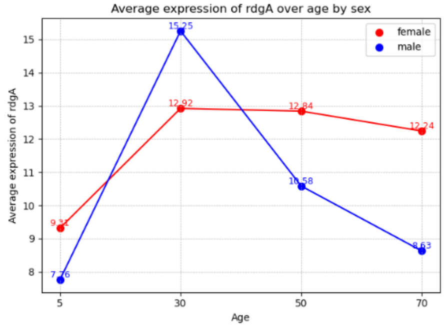

Average expression of CG41378 over age by sex
female
male
Average expression of CG41378
Age
2.5
3.0
3.5
4.0
4.5
5.0
5
30
50
70
3.10
2.78
3.28
2.26
5.27
2.75
4.75
3.91

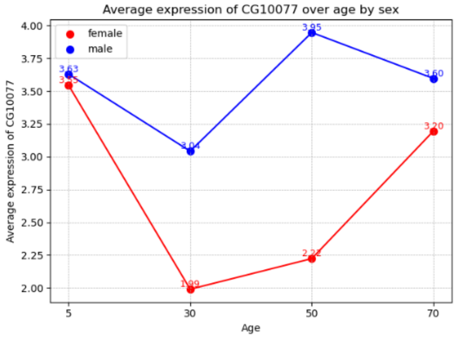

Average expression of Pdp1 over age by sex
female
male
Average expression of Pdp1
Age
4.0
4.5
5.0
5.5
6.0
5
30
50
70
4.82
3.80
6.32
5.87
4.30
4.19
4.09
3.97

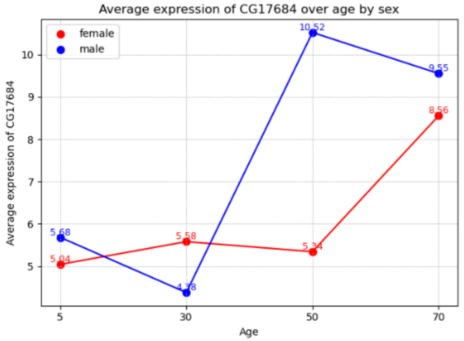

Average expression of lncRNA:Hsromega over age by sex
female
male
Average expression of lncRNA:Hsromega
Age
6
8
10
12
14
16
18
5
30
50
70
19.31
14.56
8.06
6.70
15.38
5.51
17.80
11.05


Average expression of lncRNA:CR43836 over age by sex
female
male
Average expression of lncRNA:CR43836
Age
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
5
30
50
70
0.95
0.83
2.27
1.93
2.38
1.97
2.20
1.76


Average expression of ninaE over age by sex
female
male
Average expression of ninaE
Age
2
3
4
5
6
7
8
9
10
5
30
50
70
5.43
5.32
9.93
4.79
3.00
2.14
3.19
3.09


Average expression of rdgA over age by sex
female
male
Average expression of rdgA
Age
8
9
10
11
12
13
14
15
5
30
50
70
9.31
7.76
15.25
12.92
12.84
10.58
12.24
8.63


Average expression of CG10077 over age by sex
female
male
Average expression of CG10077
Age
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
5
30
50
70
3.63
3.55
3.04
1.99
3.95
2.22
3.60
3.20


Average expression of CG17684 over age by sex
female
male
Average expression of CG17684
Age
5
6
7
8
9
10
5
30
50
70
5.68
5.04
5.58
4.38
10.52
5.34
9.55
8.56

NEXT STEPS

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.

Scientific Publication

Prepare a paper detailing the results and analysis for peer-reviewed publication.

Diversified Datasets

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

PRIOR WORKS AND METHODS

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

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

Tzu-Chiao Lu et al.: Provided the raw UMI data.

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).