Cosmic Learning: Classifying Stars, Quasars, and Galaxies

...

Team Members:

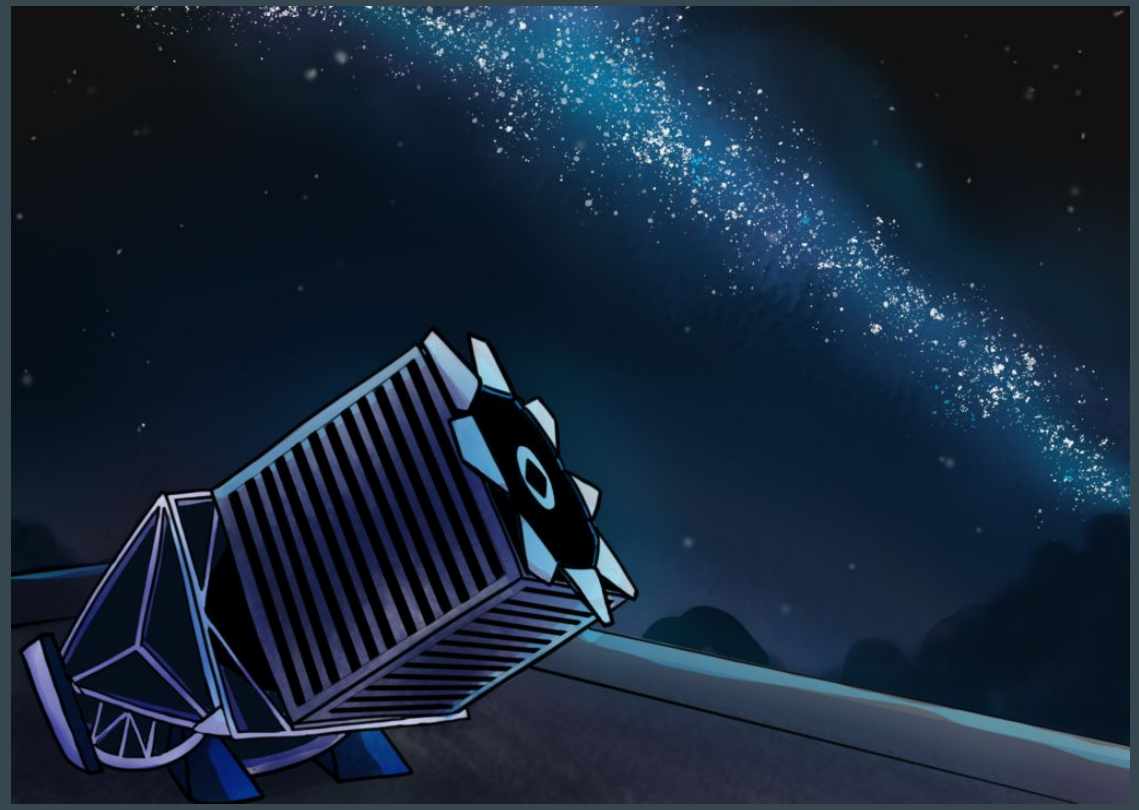
Ahmad Allaou, Daelyn Bergsman, Spencer Morris

DATASCI 207 Section 007

14 December 2023

Motivation & Background

- Astronomical surveys are huge
  - Vera C. Rubin Observatory will generate 5 PB of post-processed data per year
- Next-gen observatories require automated object classification
  - Quasars, Galaxies, Nebulae, Stars, Asteroids, Satellites, ... and many others!
- ML models can be trained from curated labeled datasets
  - Sloan Digital Sky Survey (SDSS)
  - Galaxy Zoo 2 (GZ2)

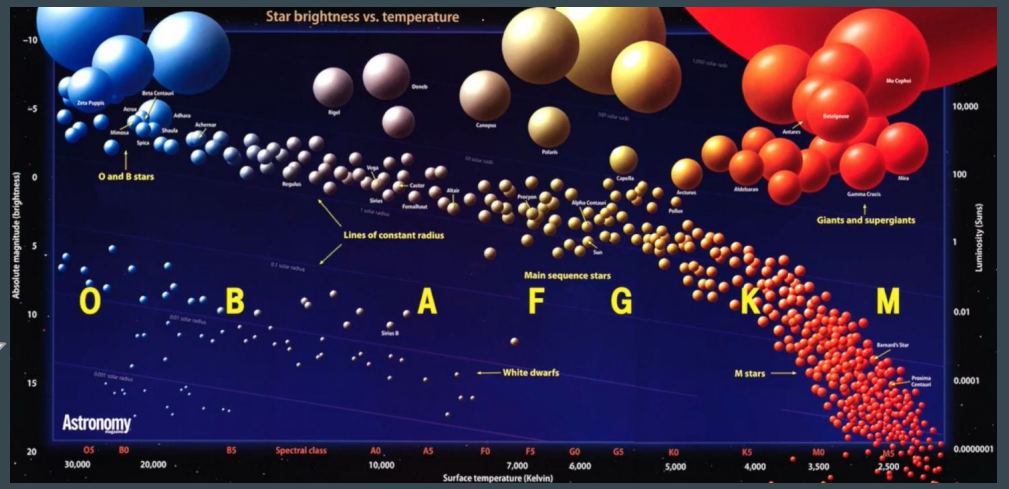


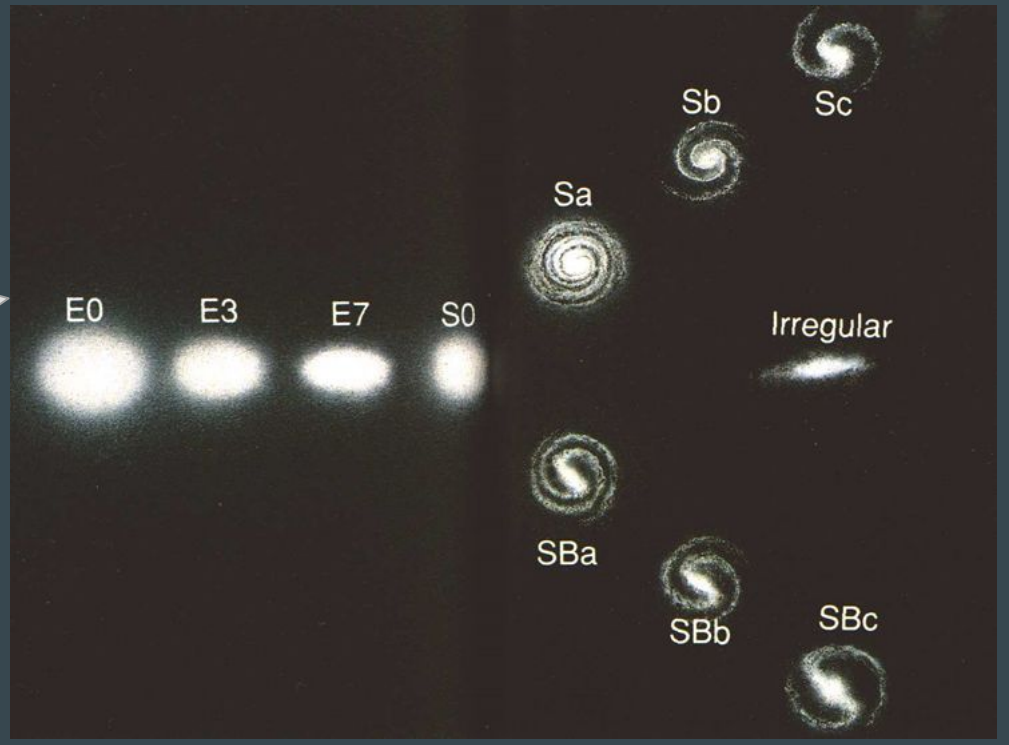
Classification Schema

1. Superclass
   - Stars
   - Galaxies
   - Quasars

2. Stars
   - Subclassified according to MK system (based on spectroscopic color)

3. Galaxies
   - Subclassified according to Hubble system (based on geometric morphology)





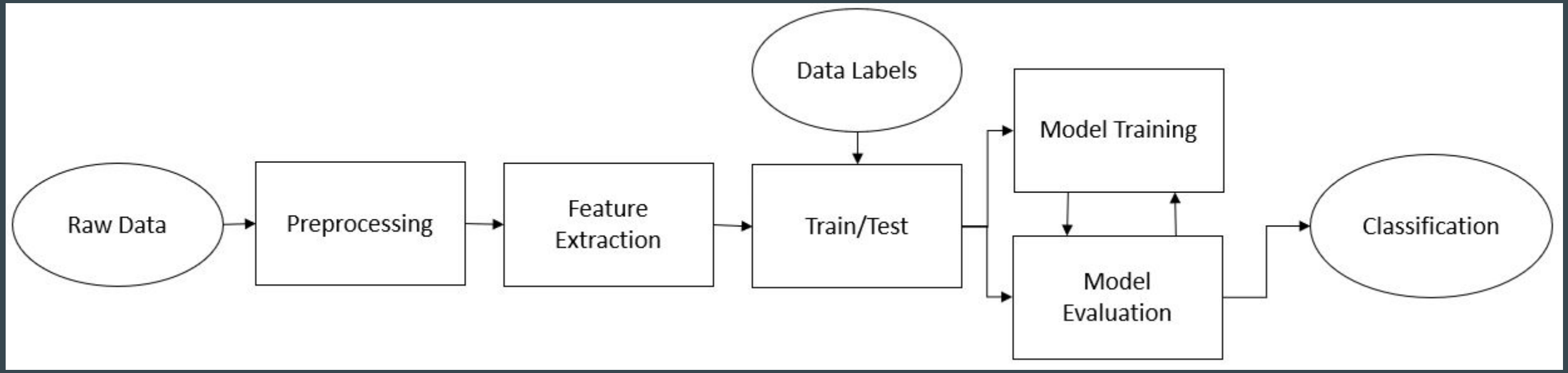
Datasets

- Provenance ensured by directly accessing/querying from original repositories

- SDSS
  - Data coming from many sensors
  - Included ~2 million unique objects
  - Tabular (0-d shape)
  - Labels:
    - Superclass
    - Stellar subclass
    - Galaxy subclass
  - Features
    - Photometric (eg. 5 color channels)
    - Spectroscopic (eg. redshift)

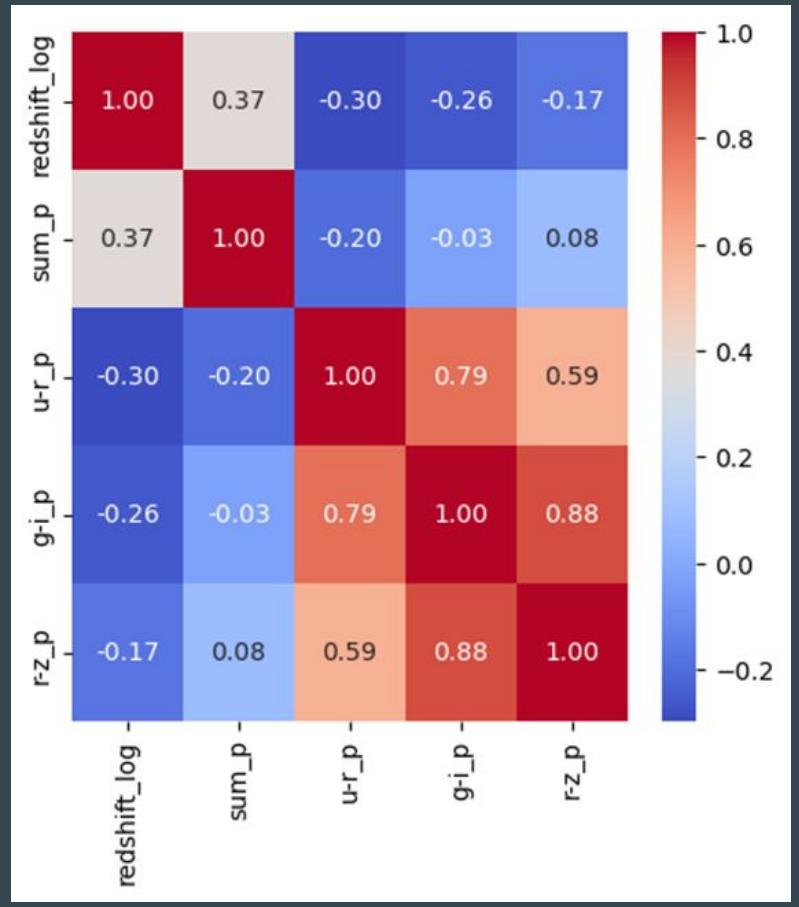
- GZ2
  - Labeling citizen-science project
  - Included ~few hundred thousand unique objects
  - Images (3-d shape, 424x424 pixels)
  - Labels:
    - Galaxy subclass
  - Features:
    - Photometric (3 color channels)
    - Geometric

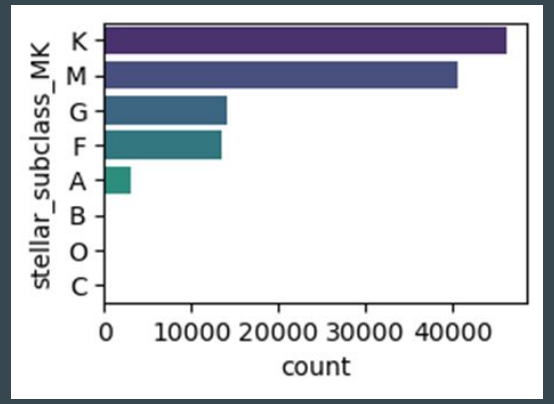
Modeling Block Diagram



Tabular preprocessing

- Label engineering
  - Stars: 171 → 10
  - Galaxies: 818 → 14
- Feature engineering
  - Derived features
  - Transformations
  - 2 feature-sets (n=5, n=8)
- Filtering
  - Brightness (p_r < 17 maggies)
  - Angular size (R_r > 17 arcsec)





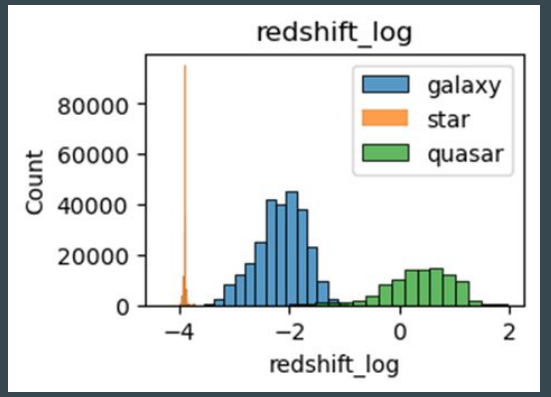
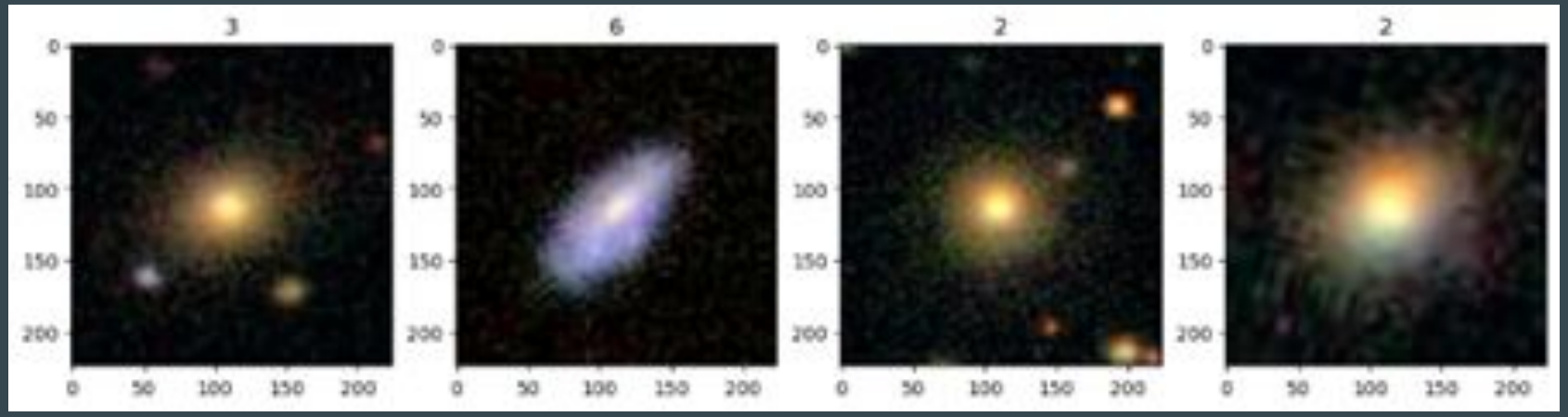


Image Preprocessing

- Images loaded, and resized from 424x424 to 224x224 to maintain consistency and for ResNet requirements
- Adjustments were made to the brightness and contrast of the images
- Randomly shuffled, split into training, validation, and test sets (60-20-20)
- Scaled RGB values from 0-255 to [0,1]

Example of images after preprocessing:



Models

Model	Superclass	Stars	Galaxies	Dataset
PCA	x	x	x	tabular
Linear Regression		x		tabular
K-means	x	x	x	tabular
KNN	x	x	x	tabular
Logistic Regression	x	x	x	tabular
SVM	x	x	x	tabular
Decision Tree	x	x	x	tabular
Random Forest	x	x	x	tabular
Feedforward NN	x	x	x	both
Res-Net			x	imagery
CNN			x	imagery
LSTM			x	imagery

Some notable findings for superclass and stars...

Decision Tree (superclass)

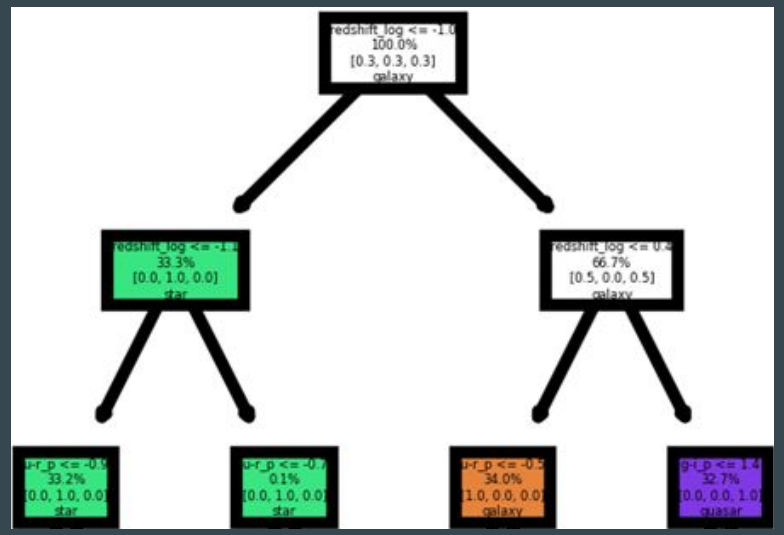
- Redshift is largely explanatory

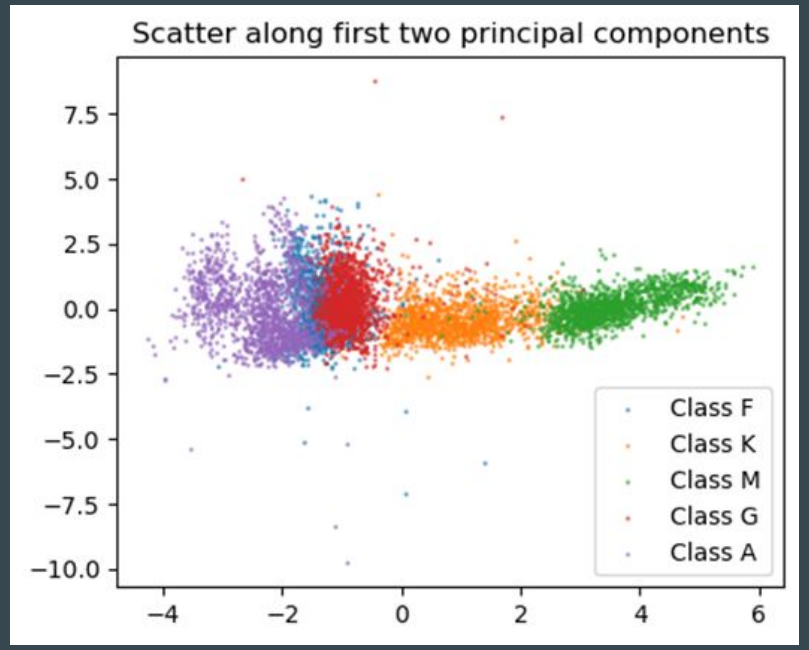
PCA (stellar)

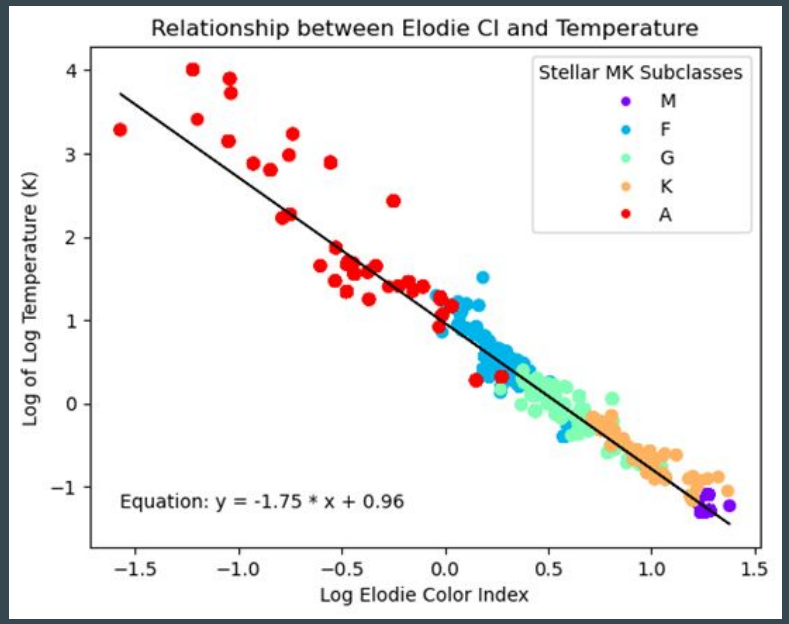
- 2 PC's explain ~85% of variation

Linear regression (stellar)

- Temperature is predictive of Elodie Color Index (R2~0.98)







Classification Experiments

							Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	N Neighbors	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
					5	140	0.9997	0.9997	0.9997	0.9997	0.9994	0.9994	0.9994	0.9994
					20	200	0.9994	0.9994	0.9994	0.9994	0.9992	0.9992	0.9992	0.9992
				Smote	100	308	0.9991	0.9991	0.9991	0.9991	0.9987	0.9987	0.9987	0.9987
					5	50	0.9995	0.9995	0.9995	0.9995	0.9992	0.9992	0.9992	0.9992
					20	63	0.9992	0.9992	0.9992	0.9992	0.999	0.999	0.999	0.999
			Superclass	Rarest	100	98	0.9987	0.9987	0.9987	0.9987	0.9984	0.9984	0.9984	0.9984
					5	16	0.991	0.991	0.991	0.991	0.9809	0.9809	0.9809	0.9809
					20	21	0.9797	0.9798	0.9897	0.9897	0.9751	0.9754	0.9751	0.9751
				Smote	100	33	0.9628	0.9634	0.9628	0.9625	0.9643	0.9643	0.9657	0.9643
					5	4	0.9668	0.9667	0.9668	0.9666	0.9614	0.9622	0.9614	0.9614
					20	4	0.9445	0.9449	0.9447	0.9442	0.9517	0.9536	0.9517	0.9517
			Stars	Rarest	100	4	0.9147	0.9171	0.9147	0.9133	0.9303	0.9343	0.9303	0.9398
					5	31	0.7005	0.6917	0.7005	0.6905	0.2376	0.2728	0.2376	0.2435
					20	40	0.535	0.5157	0.535	0.5176	0.2501	0.3055	0.2501	0.2567
				Smote	100	65	0.4258	0.4065	0.4258	0.4064	0.2641	0.33	0.2651	0.2665
					5	6	0.4993	0.5043	0.4993	0.4962	0.2409	0.2718	0.2409	0.24429
					20	7	0.4047	0.398	0.4047	0.3985	0.2666	0.3076	0.2666	0.2711
KNN	X0	1000	Galaxies	Rarest	100	9	0.3667	0.3556	0.3667	0.351	0.2716	0.3223	0.2716	0.2728

- Number of hyperparameters tuned ranged from 2 to 6
- Studied the impact of data augmentation by using a ‘rebalancing mode’ hyperparameter
- Measured run-time performance as well as usual predictive performance metrics
- Focused on F1 score for non-balanced validation and test sets

Image Classification: Baseline

FFNN used as baseline for subclassification

- Comprised of 4 dense layers that feed into each other, followed by batch normalization
- At the end is a single dense layer with SoftMax activation.
- Hidden layers used Relu activation
- This model had the lowest predictive performance

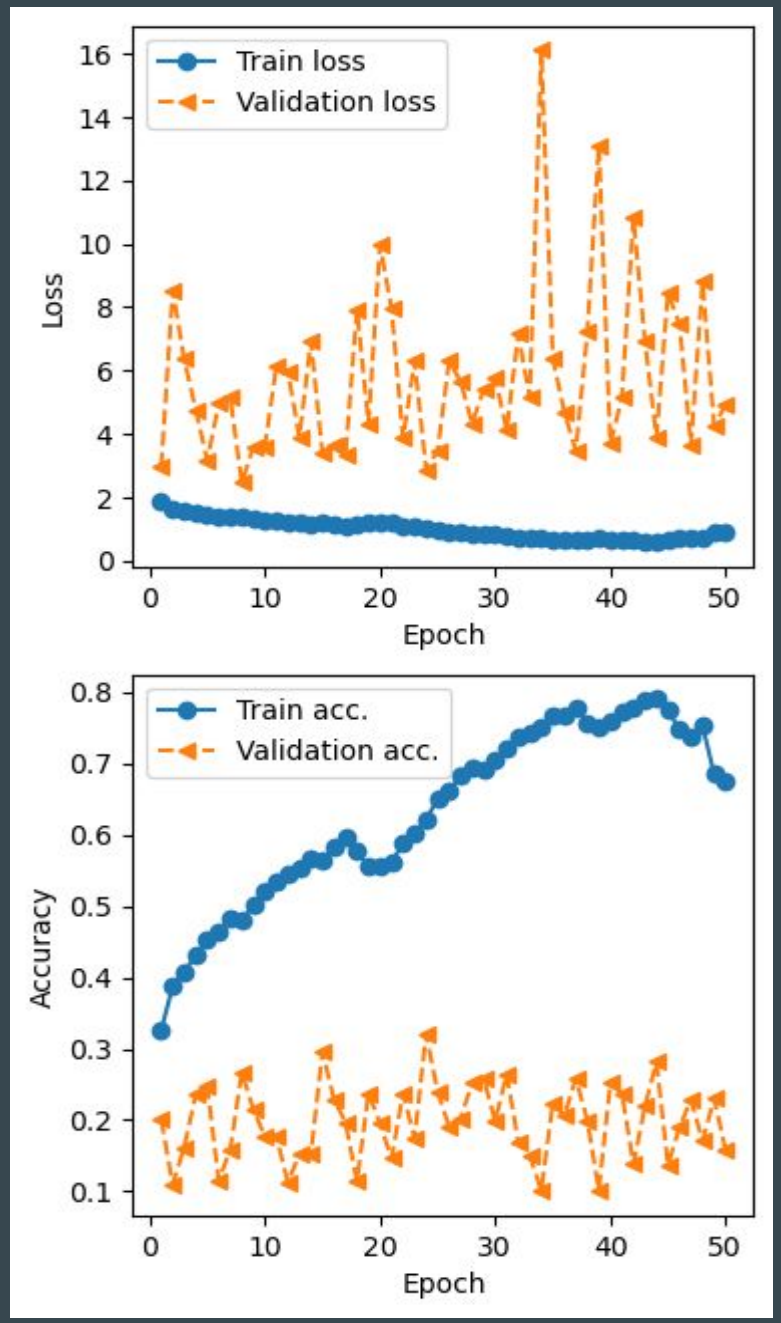
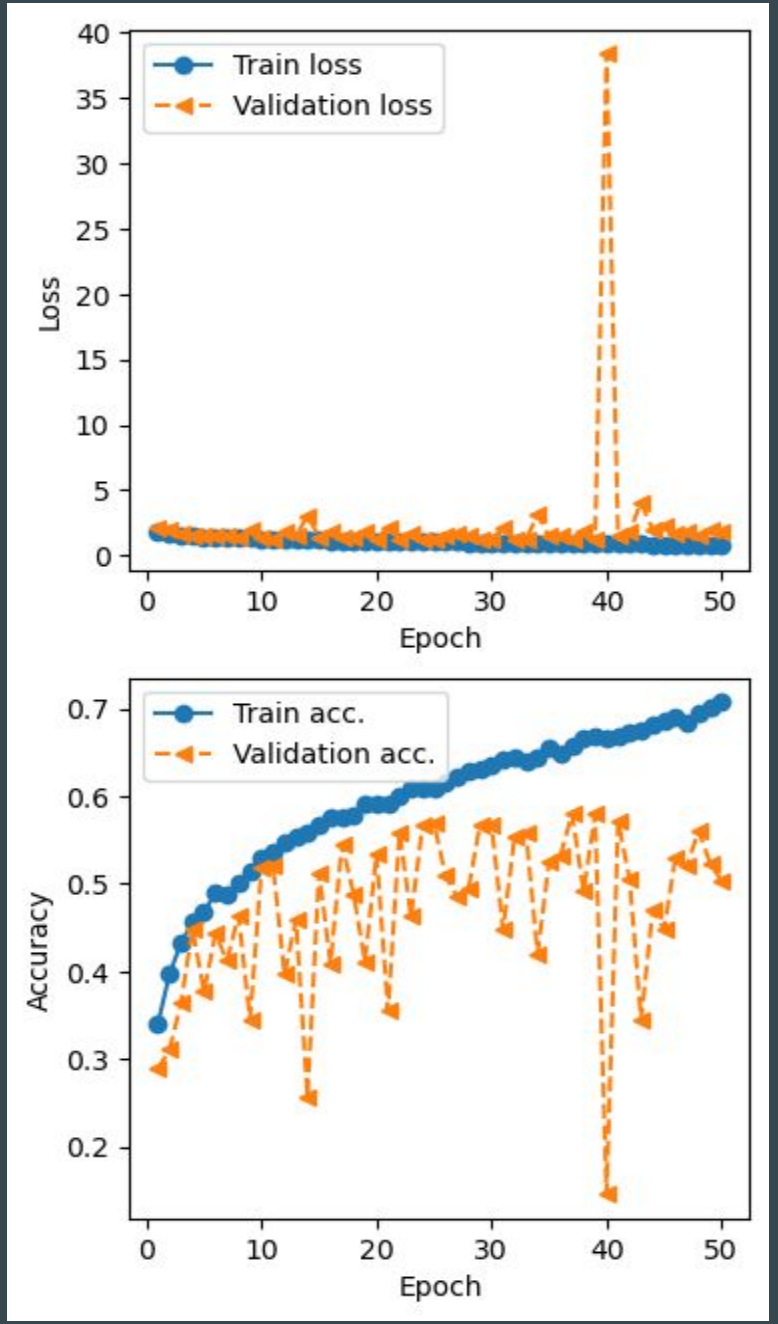


Image Classification: CNN

Best galaxy subclassification model:

- Begins with a layer that randomly rotates the image
- Followed by 3 series of convolutional layers.
- Each series is comprised of a conv2D, a MaxPool2D, and a batch normalization
- Fed directly into the next series, until they are flattened after the third convolution.
- After flattening, they go through 2 dense layers, each followed by batch normalization. Finally, feed into the final SoftMax layer

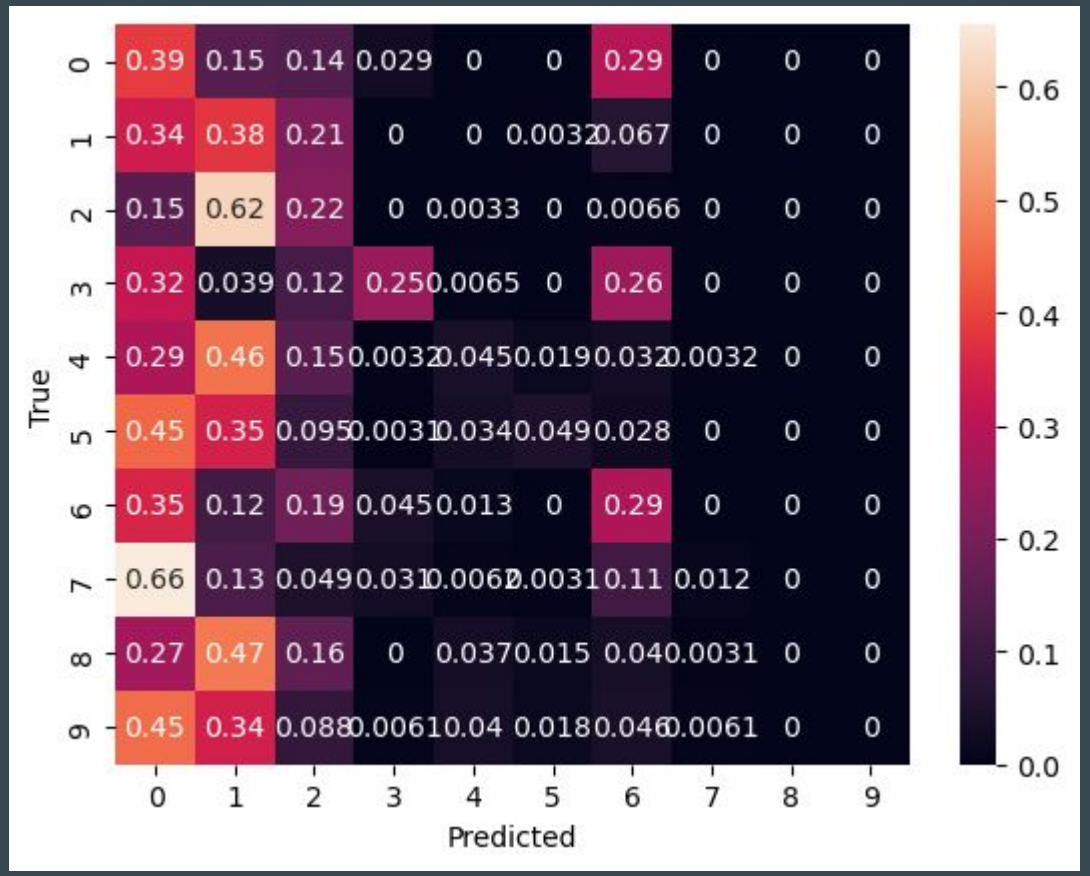


Model Summaries - FFNN

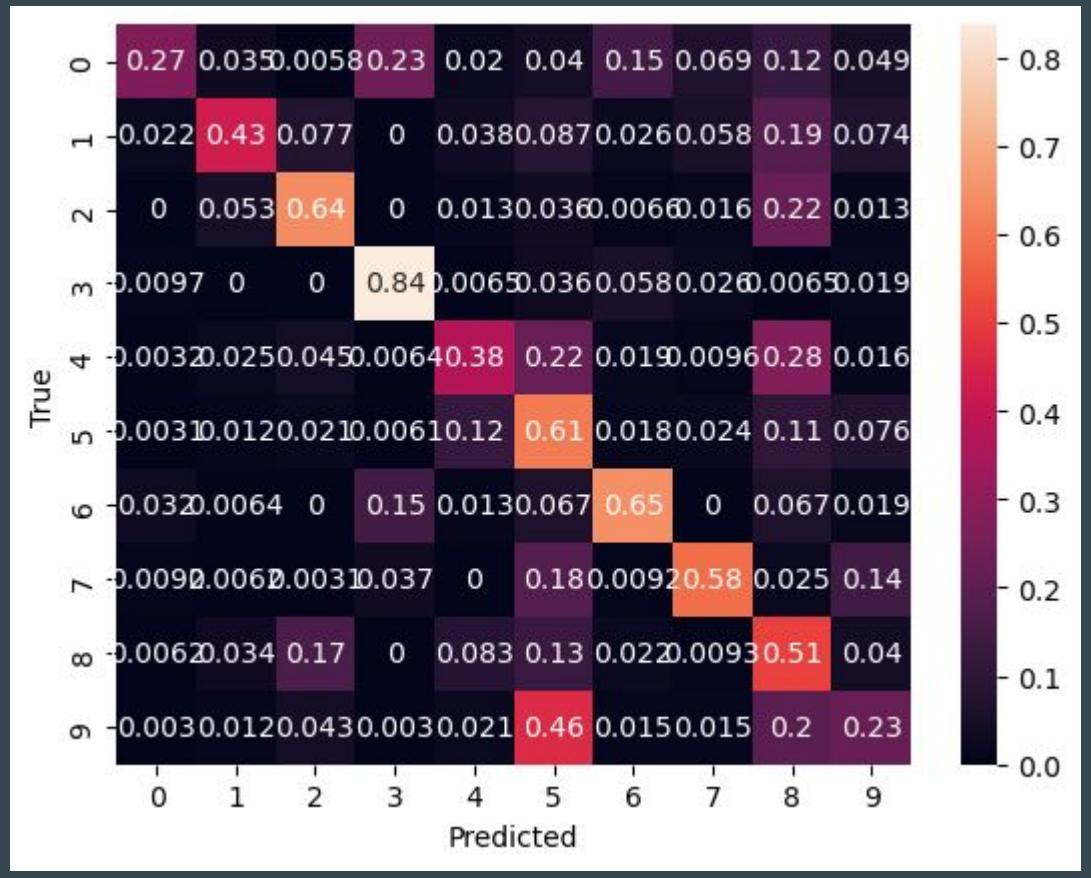
Feed Forward Neural Network

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 150528)	0
fc_1 (Dense)	(None, 256)	38535424
batch_normalization (BatchNormalization)	(None, 256)	1024
fc_2 (Dense)	(None, 512)	131584
batch_normalization_1 (BatchNormalization)	(None, 512)	2048
fc_3 (Dense)	(None, 256)	131328
batch_normalization_2 (BatchNormalization)	(None, 256)	1024
fc_4 (Dense)	(None, 128)	32896
batch_normalization_3 (BatchNormalization)	(None, 128)	512
fc_5 (Dense)	(None, 10)	1290

Total params: 38837130 (148.15 MB)
Trainable params: 38834826 (148.14 MB)
Non-trainable params: 2304 (9.00 KB)



Model Summaries - CNN

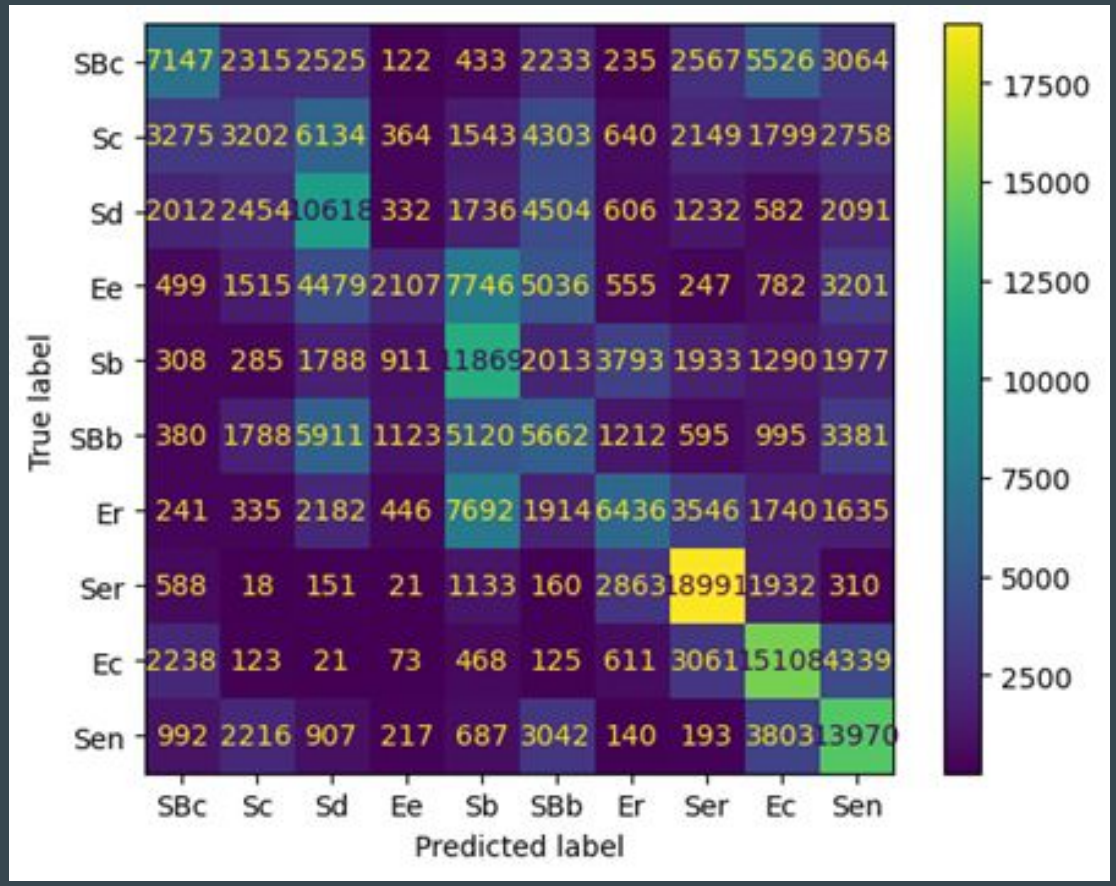



```
Convolutional Neural Network
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 rot (RandomRotation)        (None, 224, 224, 3)       0

 conv_1 (Conv2D)             (None, 112, 112, 32)      896

 pool_1 (MaxPooling2D)       (None, 56, 56, 32)        0

 batch_normalization_4 (Bat  (None, 56, 56, 32)        128
 chNormalization)

 conv_2 (Conv2D)             (None, 56, 56, 64)        51264

 pool_2 (MaxPooling2D)       (None, 28, 28, 64)        0

 batch_normalization_5 (Bat  (None, 28, 28, 64)        256
 chNormalization)

 conv_3 (Conv2D)             (None, 28, 28, 128)       204928

 pool_3 (MaxPooling2D)       (None, 14, 14, 128)       0

 batch_normalization_6 (Bat  (None, 14, 14, 128)       512
 chNormalization)

 flatten_1 (Flatten)         (None, 25088)             0

 fc_1 (Dense)                (None, 2048)              51382272

 batch_normalization_7 (Bat  (None, 2048)              8192
 chNormalization)

 fc_2 (Dense)                (None, 1024)              2098176

 batch_normalization_8 (Bat  (None, 1024)              4096
 chNormalization)

 fc_3 (Dense)                (None, 10)                10250

=================================================================
Total params: 53760970 (205.08 MB)
Trainable params: 53754378 (205.06 MB)
Non-trainable params: 6592 (25.75 KB)
_________________________________________________________________
```

Tabular Model Summary - FFNN

Best tabular galaxy subclassification model:

- Training data augmented using SMOTE
- 2 hidden dense layers
- Relu activations
- Learning Rate of 0.1
- Overall Accuracy 0.37



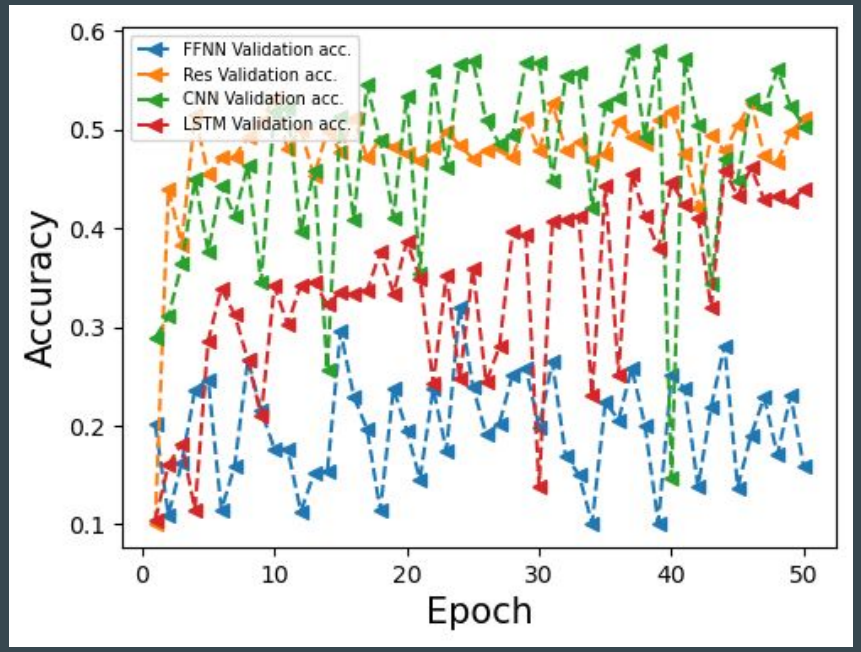
Conclusion

- Tabular data for superclass and star sub classes performed incredibly well
- Tabular data did not perform well on the geometric properties of the galaxies

- Of all models for image classification, CNN performed the best, but inconsistently.
- Resnet was incredibly performance heavy compared to all other models
- Image recognition may not be the best application of FFNN and LSTM models

- We were not able to reach the target of 80% accuracy we set earlier in the semester

			Test				
Objects	Data Shape	MODEL	Run-Time (s)	Accuracy	Precision	Recall	F1
		KNN	111	0.2693	0.3343	0.2693	0.2702
		Logistic	7	0.2242	0.3091	0.2242	0.2107
		SVM	31	0.1697	0.2031	0.17	0.1607
		DT	17	0.2639	0.3260	0.2639	0.2649
		RF	91	0.2844	0.3438	0.2844	0.2840
	Tabular	FFNN	144	0.3635	0.3565	0.3635	0.3412
		FFNN	3240	0.1548			
		ResNet50	34920	**0.5109**			
		CNN	7200	0.5034			
Galaxies	Imagery*	LSTM	5040	0.4397			



Questions?

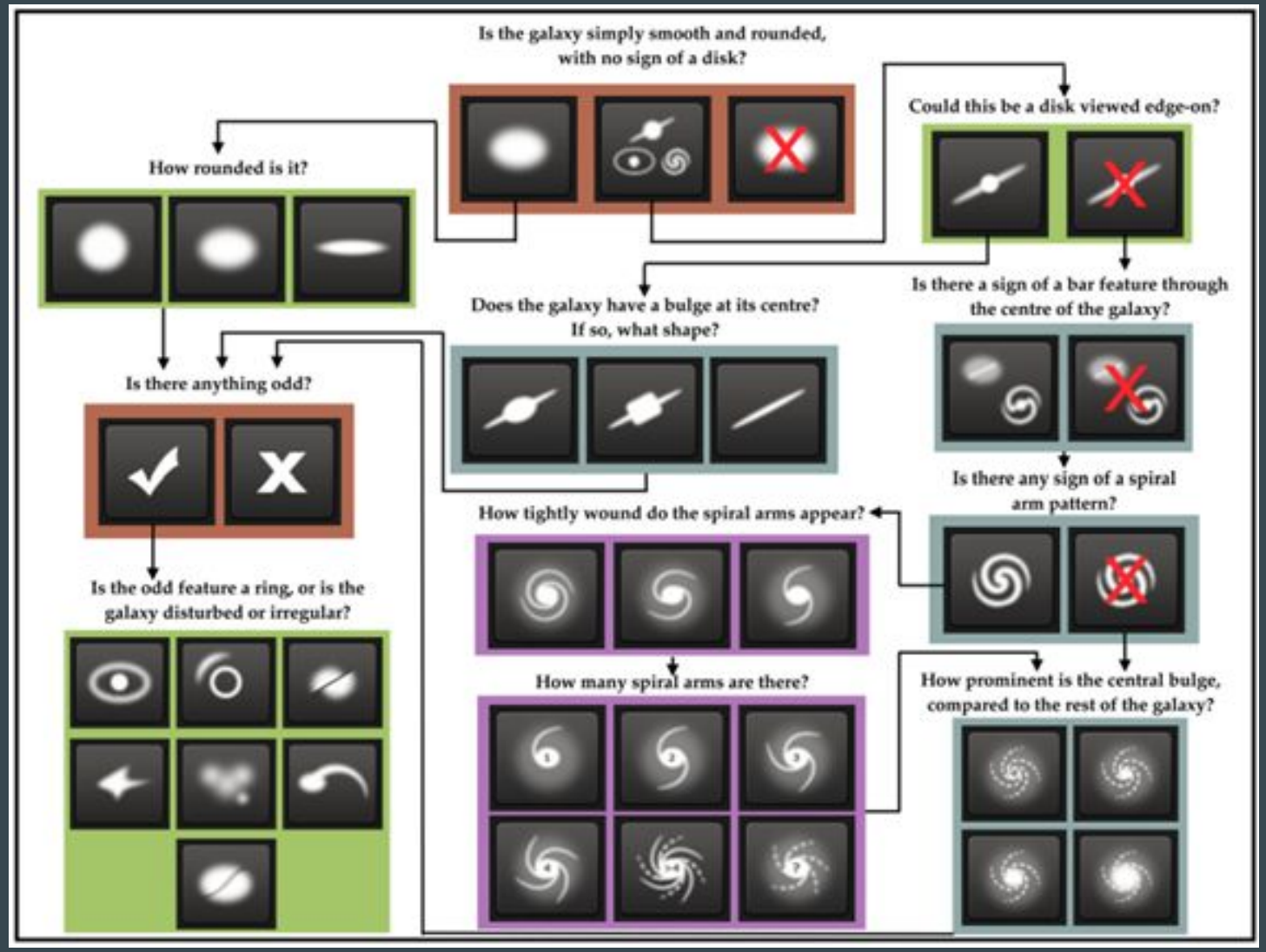
Backup Slides

Computing Environments

- PC 1
  - Intel Core i7-7660U CPU
  - 16GB DDR3 RAM
- PC 2
  - AMD Ryzen 7 7800X3D CPU
  - 32GB DDR5 RAM

- Virtual Environment
  - Python 3.10
  - Conda Package Manager
  - Prototyping in Jupyter Notebooks
  - Key libraries: pandas, numpy, matplotlib, seaborn, sklearn, tensorflow, keras

Galaxy Zoo 2 Decision Tree



Simplified Labels for each Classification Task

Superclass

Label
galaxy
quasar
star

Stars

Label	Description
O	blue
B	bluish white
A	white
F	yellowish white
G	yellowish white
K	light orange
M	orangeish red
D	white dwarf
C	carbon star
d	cool (red or brown) dwarf

Galaxies

Label	Description
Er	elliptical with low eccentricity (round)
Ee	elliptical with intermediate eccentricity
Ec	elliptical with high eccentricity (cigar-shaped)
Sa	spiral with large bulge
Sb	spiral with medium bulge
Sc	spiral with small bulge
Sd	spiral with no bulge
SBa	barred-spiral with large bulge
SBb	barred-spiral with medium bulge
SBc	barred-spiral with small bulge
SBd	barred-spiral with no bulge
Ser	edge-on spiral with round bulge
Seb	edge-on spiral with boxy bulge
Sen	edge-on spiral with no bulge

Test metrics of tuned models for superclass & star

			Test				
Objects	Data Shape	MODEL	Run-Time (s)	Accuracy	Precision	Recall	F1
		KNN	156	0.9992	0.9992	0.9992	0.9992
		Logistic	15	0.9991	0.9991	0.9991	0.9991
		SVM	44	0.9568	0.9627	0.9568	0.9578
		DT	25	0.9992	0.9992	0.9992	0.9992
		RF	26	0.9994	0.9994	0.9994	**0.9994**
Superclass	Tabular	FFNN	252	0.9993	0.9993	0.9993	0.9993

			Test					
Objects	Data Shape	MODEL	Run-Time (s)	R^2	Accuracy	Precision	Recall	F1
		KNN	21		0.9819	0.9820	0.9819	0.9819
		Logistic	2		0.9346	0.9358	0.9346	0.9351
		SVM	3		0.9107	0.9164	0.9107	0.9078
		DT	2		0.9972	0.9972	0.9972	**0.9972**
		RF	2		0.9956	0.9956	0.9956	0.9956
Stars	Tabular	FFNN	62		0.9832	0.9836	0.9832	0.9833

Superclass Classification: Baseline

RFC used as baseline for superclass classification:

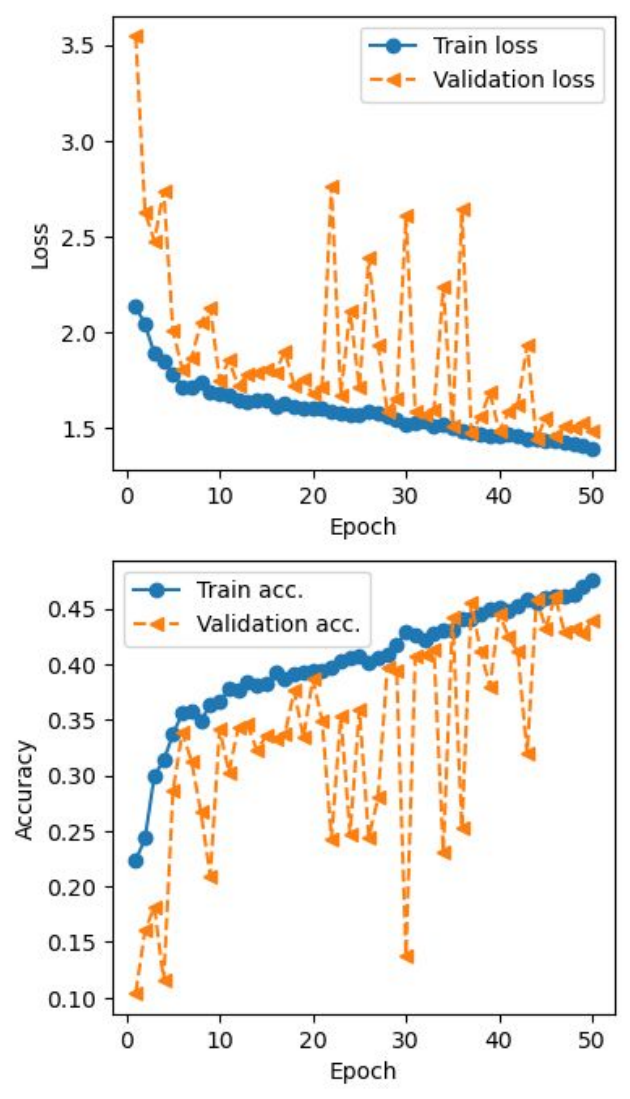
- Hyperparameters used during training:
  - Rebalancing Mode
  - Bootstrap sampling
  - Number of estimators
- Oversampling led to better results at the cost of 10x run time
- 10 estimators proved to be sufficient for optimal performance

Superclass Classification: Best Model

FFNN was best model for superclass classification

- Used Adam optimizer with 2 hidden layers
- First hidden layer had 512 neurons, second with 128 neurons
- 69,123 trainable parameters for 5-feature dataset, 70,917 parameters for 8 feature dataset
- 2048 batch size, 5 epochs
- Results with best hyperparameters (Validation F-1):
  - Superclass: ~100%
  - Stellar subclass: ~98%
  - Galaxy subclass: ~28%.

Model Summaries - LSTM



Long Short Term Memory Model

Layer (type)	Output Shape	Param #
reshape (Reshape)	(None, 1, 224, 224, 3)	0
rot (TimeDistributed)	(None, 1, 224, 224, 3)	0
c1 (TimeDistributed)	(None, 1, 112, 112, 32)	896
p1 (TimeDistributed)	(None, 1, 56, 56, 32)	0
bn1 (TimeDistributed)	(None, 1, 56, 56, 32)	128
c2 (TimeDistributed)	(None, 1, 56, 56, 64)	51264
p2 (TimeDistributed)	(None, 1, 28, 28, 64)	0
bn2 (TimeDistributed)	(None, 1, 28, 28, 64)	256
c3 (TimeDistributed)	(None, 1, 28, 28, 128)	204928
p3 (TimeDistributed)	(None, 1, 14, 14, 128)	0
bn3 (TimeDistributed)	(None, 1, 14, 14, 128)	512
flat (TimeDistributed)	(None, 1, 25088)	0
lstm (LSTM)	(None, 256)	25953280
fc_2 (Dense)	(None, 128)	32896
batch_normalization_12 (BatchNormalization)	(None, 128)	512
dense (Dense)	(None, 10)	1290

Total params: 26245962 (100.12 MB)
Trainable params: 26245258 (100.12 MB)
Non-trainable params: 704 (2.75 KB)