Faculty of
Computer Science

Data Science and
Business Analytics

Moscow
2025

Deep learning Models for Galaxy Clusters Search in Sky Surveys

Presented by

Nadezhda Vladimirovna Fomicheva,
2nd year student,
group 232, DSBA

Project supervisor

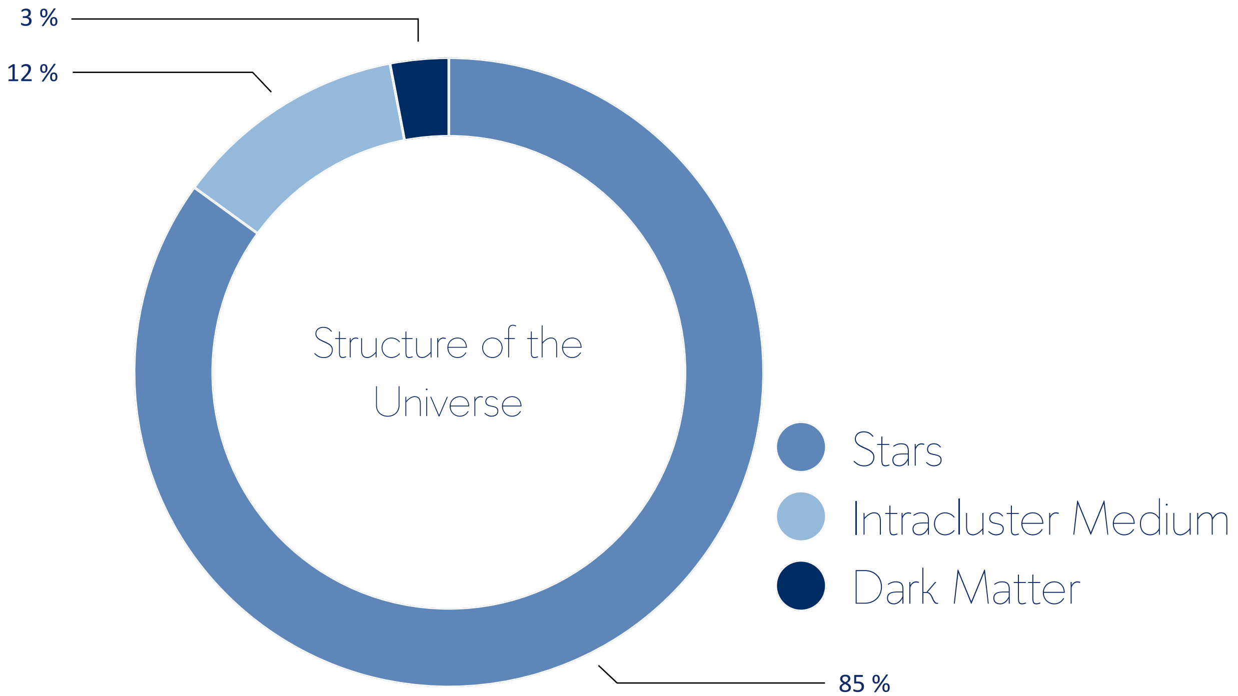
Svetlana Alekseevna Voskresenskaia,
Research Fellow,
Faculty of Physics, HSE University

Consultant

Matvey Sergeevich Zekhov,
Teacher,
Big Data and Information Retrieval School,
HSE University

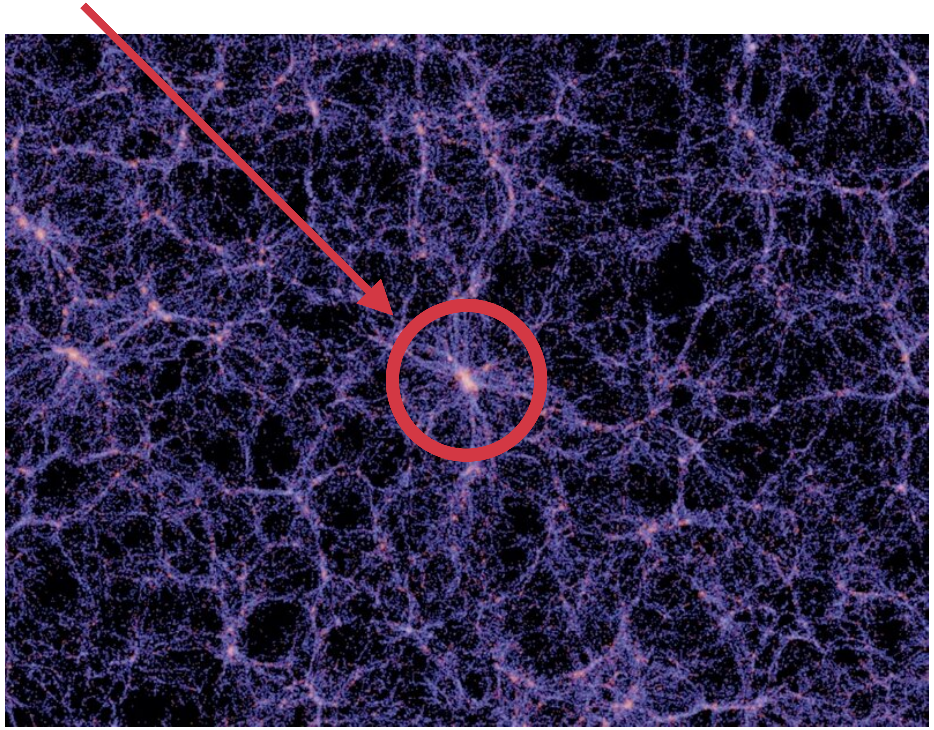
Galaxy clusters

Galaxy clusters are the largest gravitationally bound systems in the Universe.



ΛCDM model of the Universe

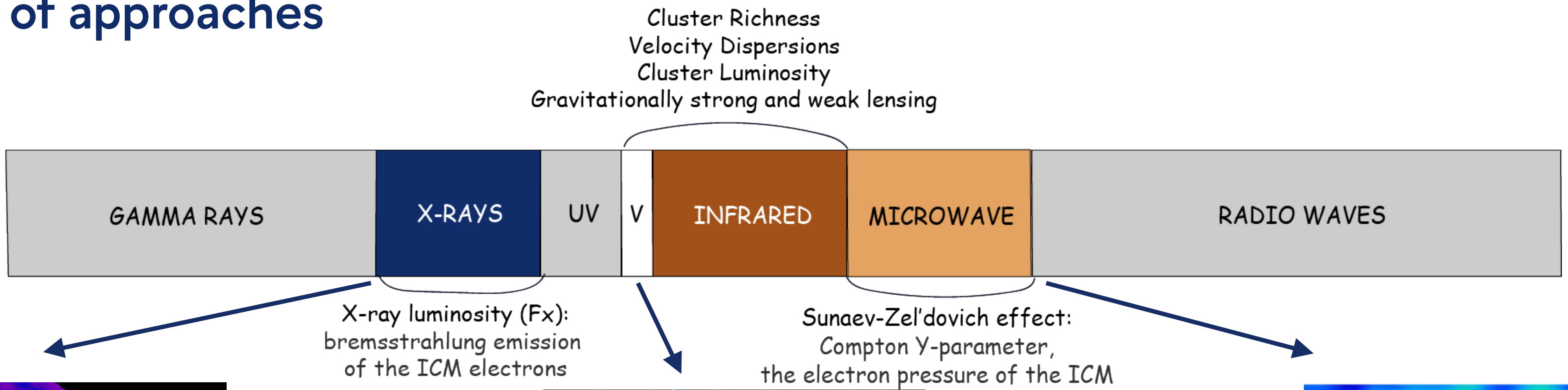
A node of a large-scale structure

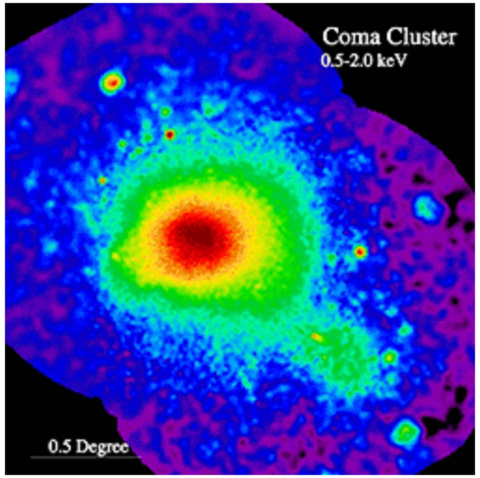


Computer simulation of the evolution of the large-scale structure of the Universe of good agreement with observations.

Credit Springel et al. (2005)

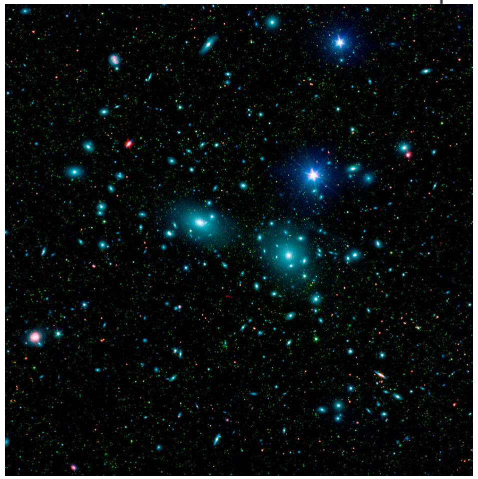
Variety of approaches





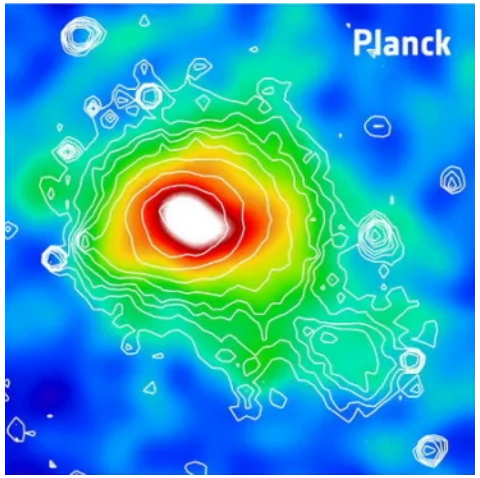
The Coma Cluster in the soft X-Ray

Credit: ROSAT/MPE/S. L. Snowden



The Coma Cluster in the optics

Credit: NASA/JPL-Caltech/L. Jenkins (GSFC)



The Coma Cluster in the microwave range

Credit: Planck

Goal

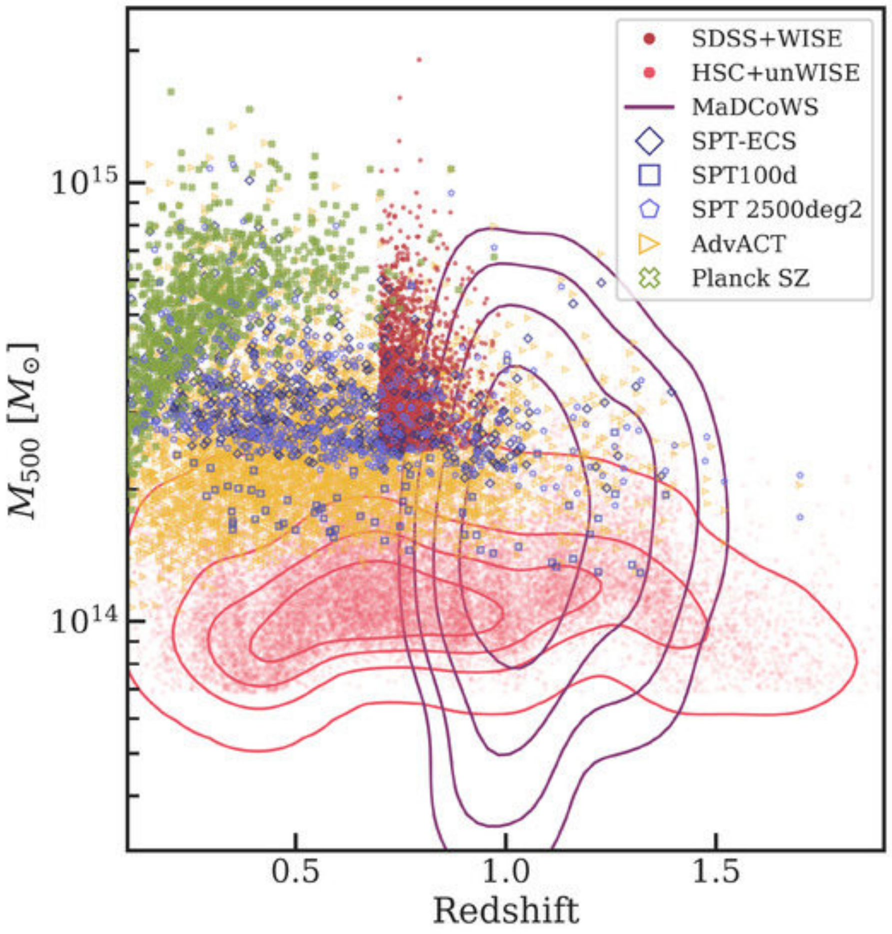
The project aimed to develop a neural network-based algorithm for classifying galaxy clusters in the infrared and microwave lengths to **expand existing cluster catalogues.**

The infrared:

- allows to detect high redshift clusters with no optical counterparts
- possible false detection in the overdensity regions

Microwave:

- provides less noisy data, independent of redshift
- intensity of signal depends on the mass an object



Clusters identified via various sky surveys.

Dataset 1/2

01

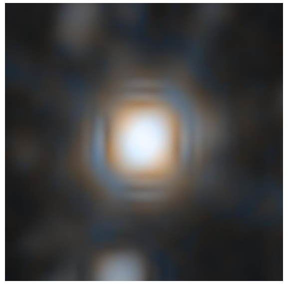
Galaxy clusters

Targeted objects of the survey

02

Sources

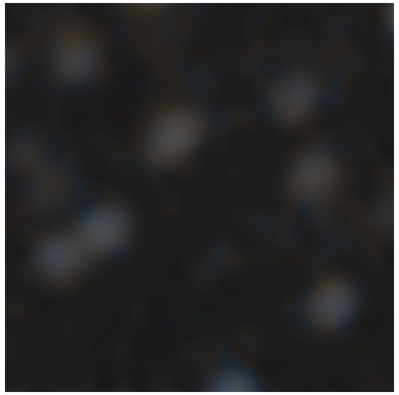
Galaxies, stars, active galactic nuclei



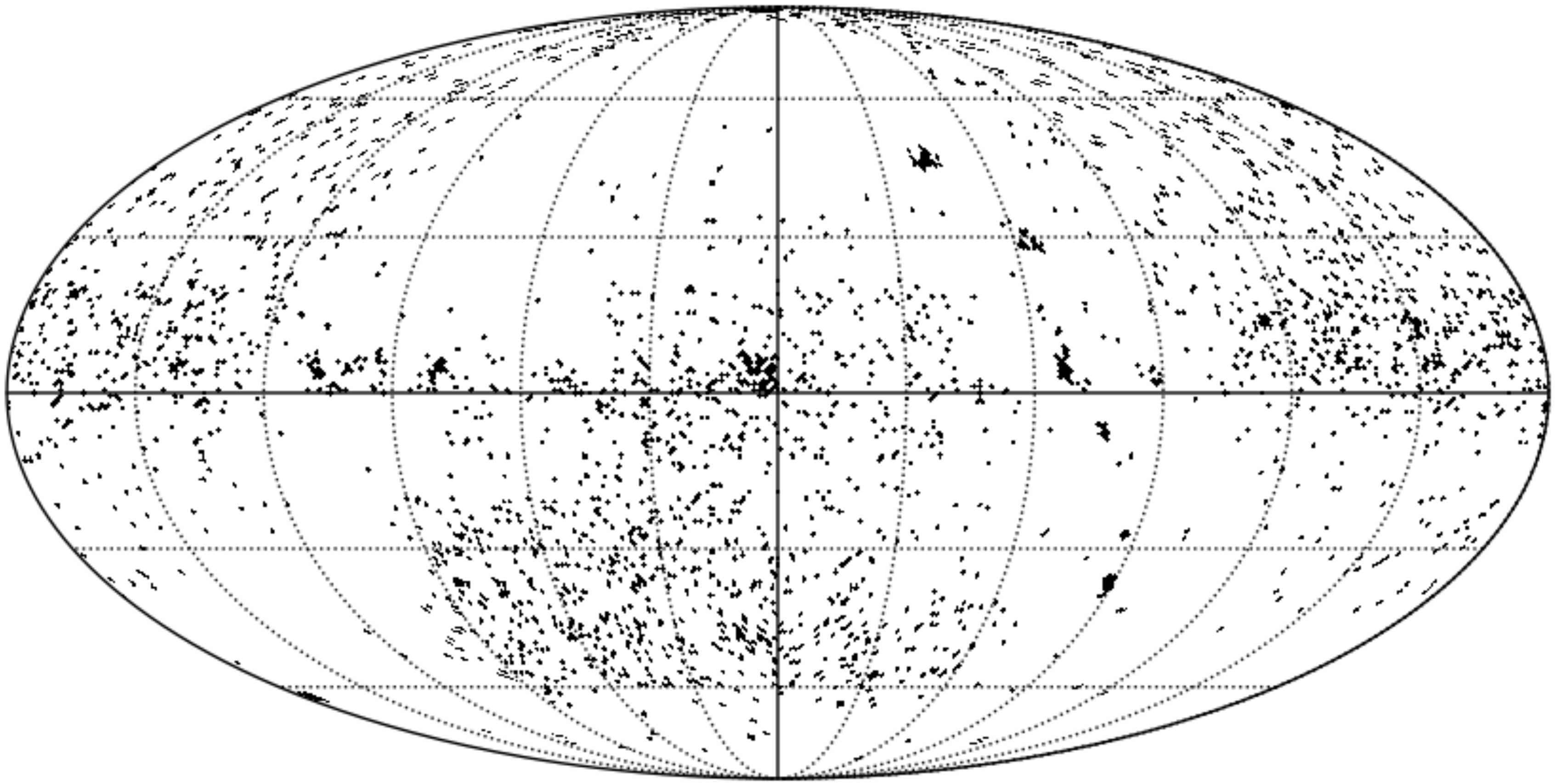
03

Random fields

«Background» images



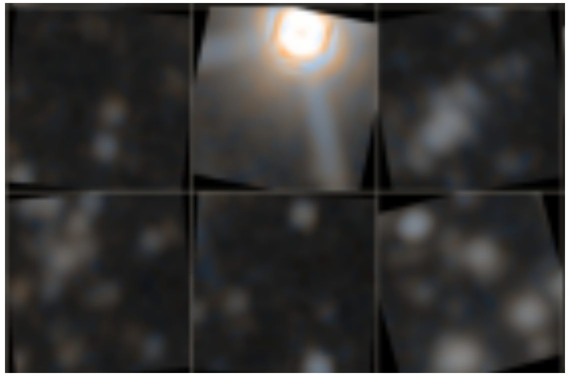
Dataset 2/2



Sky coverage by objects from the test sample from NEOWISE survey

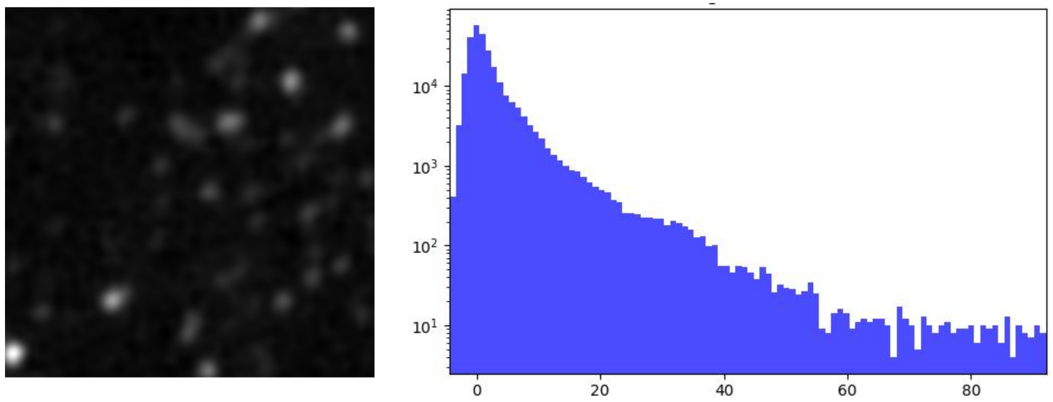
Data preprocessing

1. **Resizing images** to match input dimensions of the networks
2. **Normalization** of right-skewed data to enhance model performance
3. **Data augmentation** to detect objects with similar features
   - random rotations up to 15°
   - horizontal flips
   - random shifts

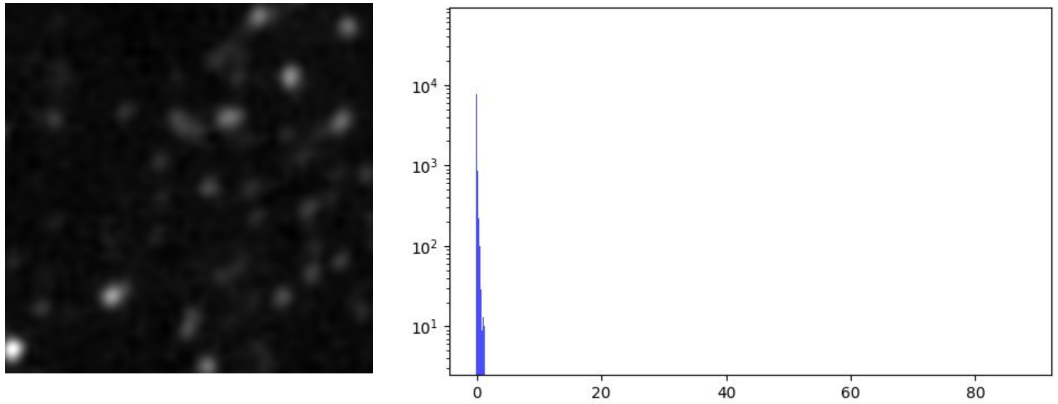


Examples of data augmentation applied to RGB images

Normalization of an object in the infrared band W1



(a) the initial image and the distribution of the signal



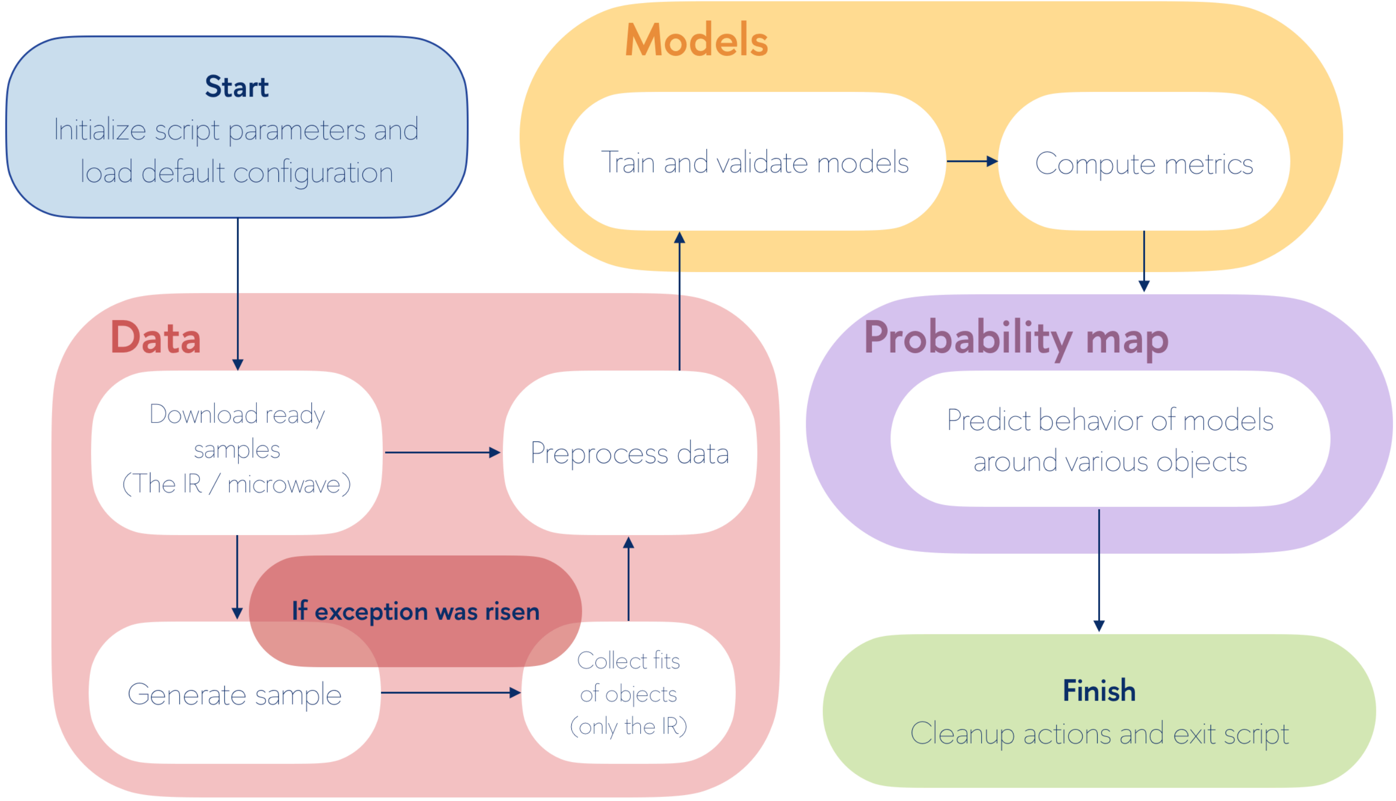
(b) the image and the distribution of the signal after normalization

Plan of experiments

1. **Choose dataset structure** based on the models' performance on the IR data
2. **Apply the networks to the infrared data.**
3. **Apply the networks to the microwave data.**

All networks were tested using **Adam optimizer.**

Learning rate for each model was **calculated automatically** via Learning Rate Range Test (LRRT).



Script structure

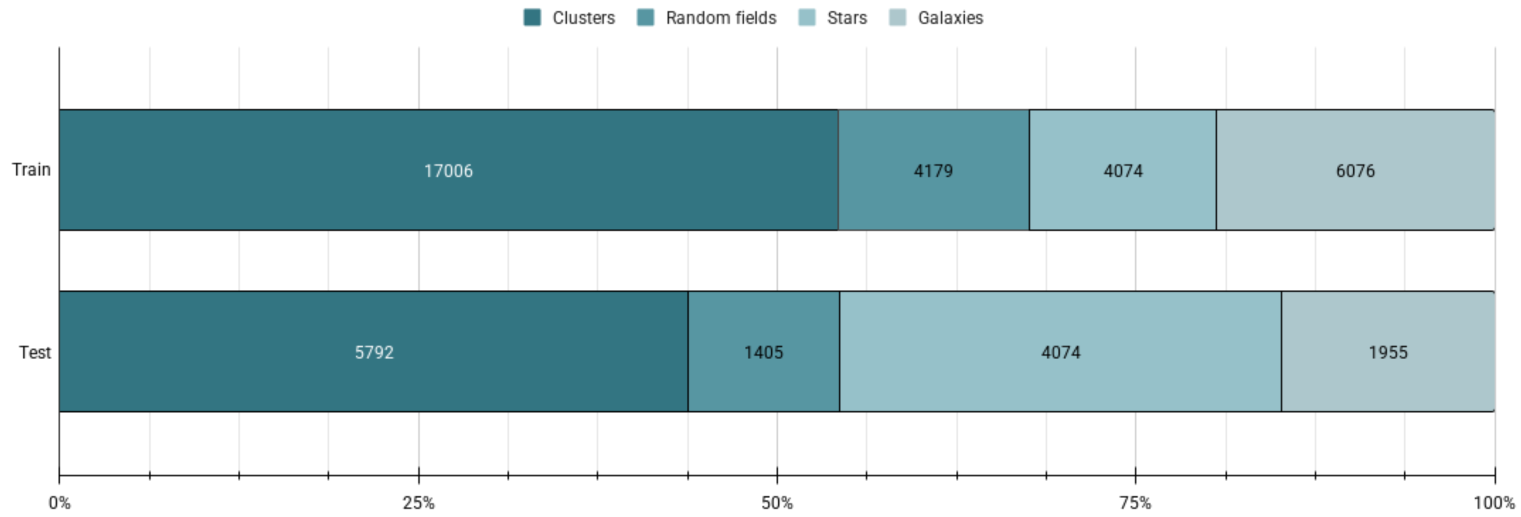
Instruments used for implementing script: **PyTorch**, **PyTorch Image Module**.

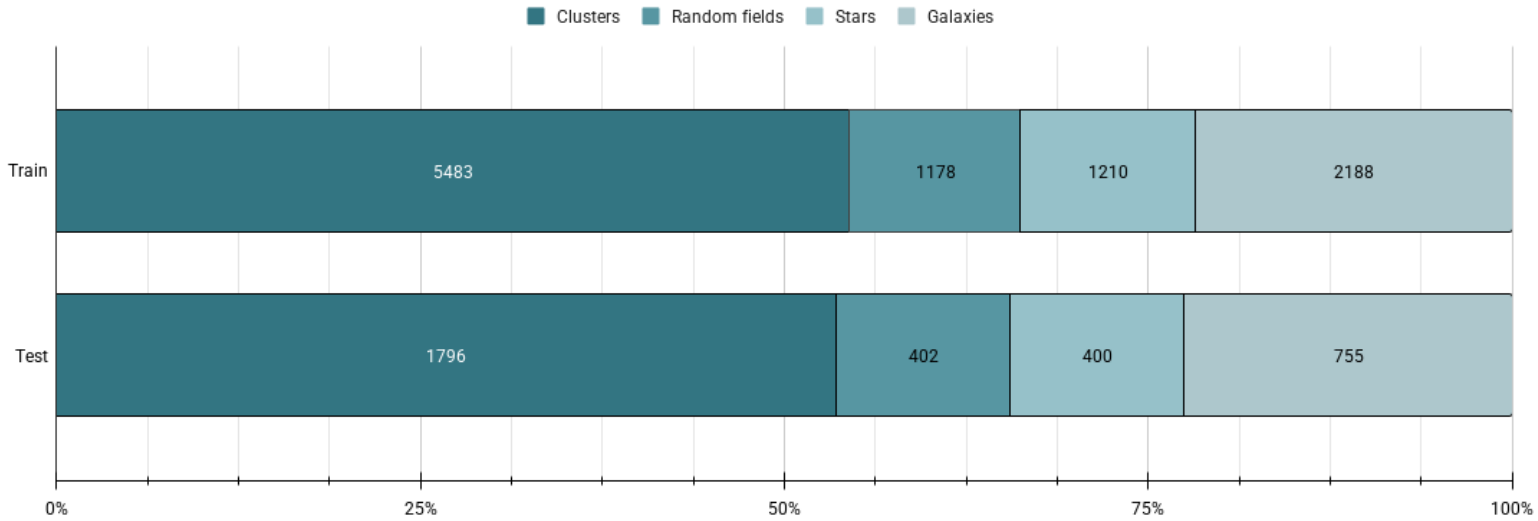
Github: https://github.com/pelancha/galaxyHackers

Documentation: https://pelancha.github.io/galaxyHackers/

Choosing dataset structure

Initial dataset structure that resulted in the overfitting of the models

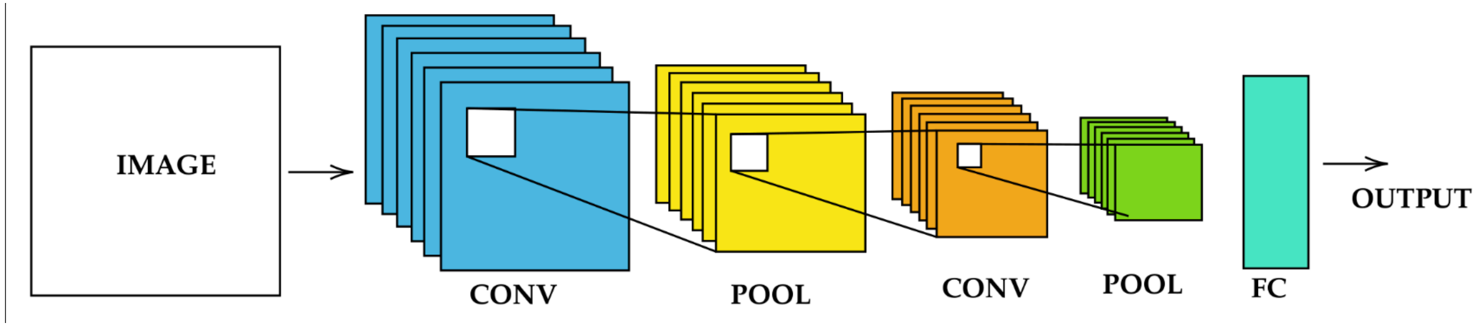




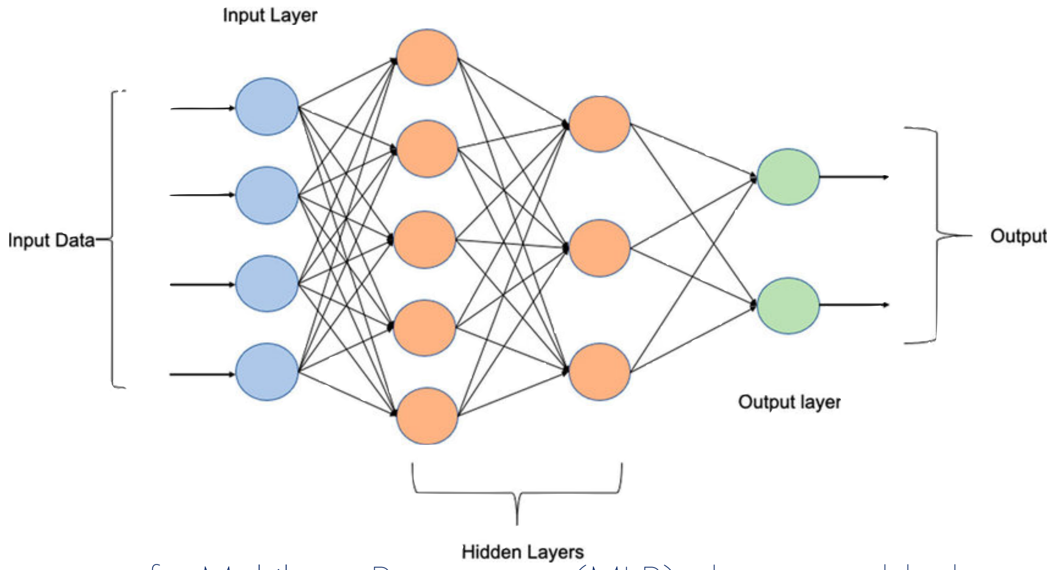
Final dataset structure that was used in tests

Studied networks 1/2

CNN models dominate in image classification tasks.

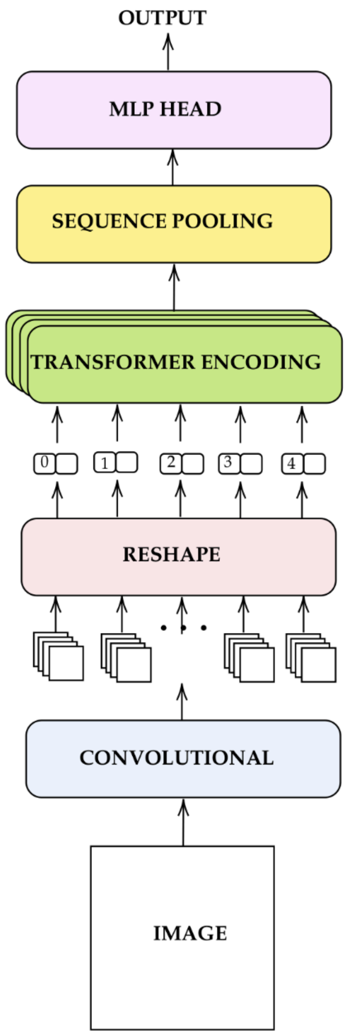


Architecture of a Convolutional Neural Network (CNN)
Credit: Kamilya Smagulova et al.



Architecture of a Multilayer Perceptron (MLP), that was added to one of the CNNs at the end
Credit: Haitham Abdulmohsin Afan et al.

Studied networks 2/2



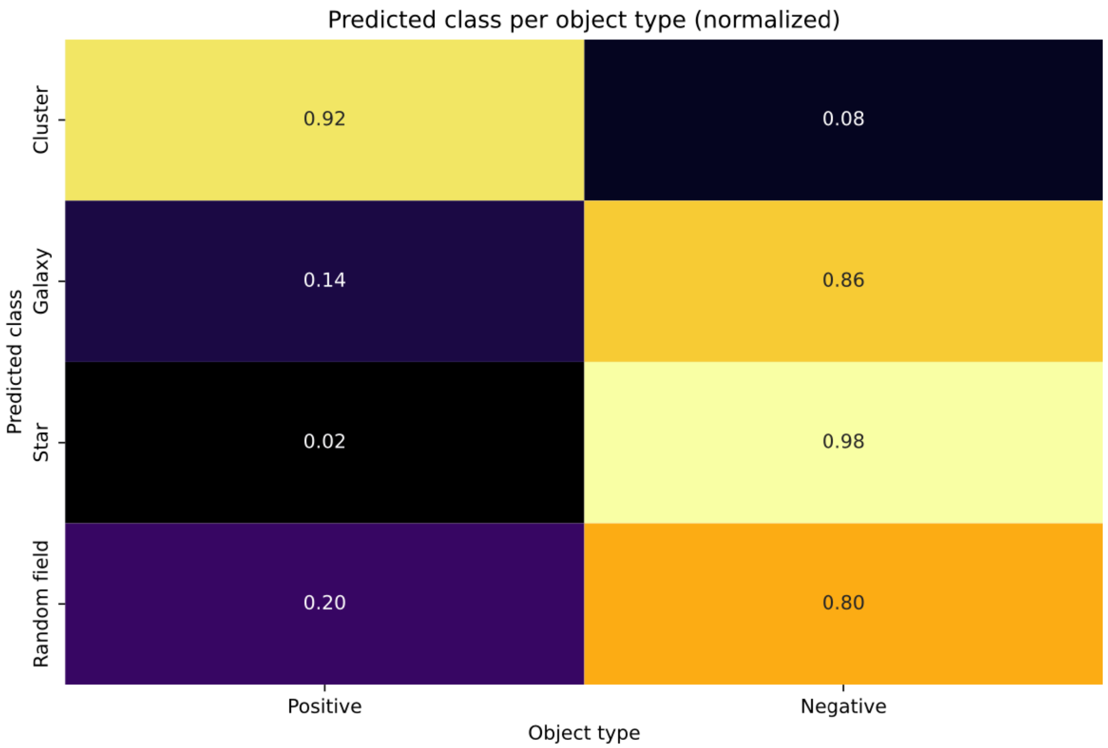
Architecture of SpinalNet
Credit: Kamilya Smagulova et al.

Models' performance

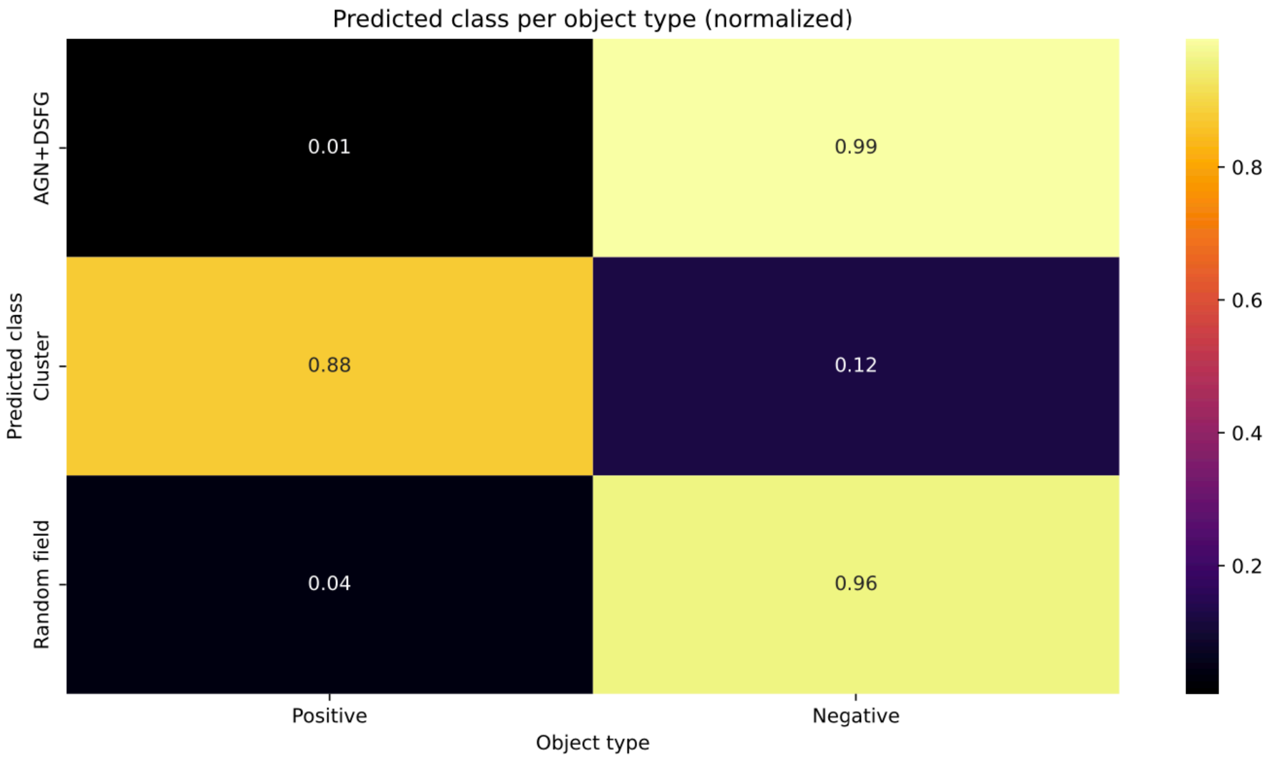
	ViTL16	SpinalNetVGG	SpinalNet+ResNet	ResNet18	EfficientNet	DenseNet
Accuracy (IR data)	0.882	0.857	0.542	0.568	0.398	0.681
Accuracy (microwave data)	0.727	0.915	0.994	0.892	0.765	0.899

	CNN+MLP	AlexNetVGG	Baseline
Accuracy (IR data)	0.807	0.905	0.668
Accuracy (microwave data)	0.836	0.937	0.648

Best performance: AlexNetVGG



(a) performance on the IR data



(b) performance on the microwave data

Results 1/2

Architectures that showed **the best performance** on the:

- the **IR data**: SpinalNets, Transformer, CNNs
- the **microwave data**: SpinalNets, CNNs

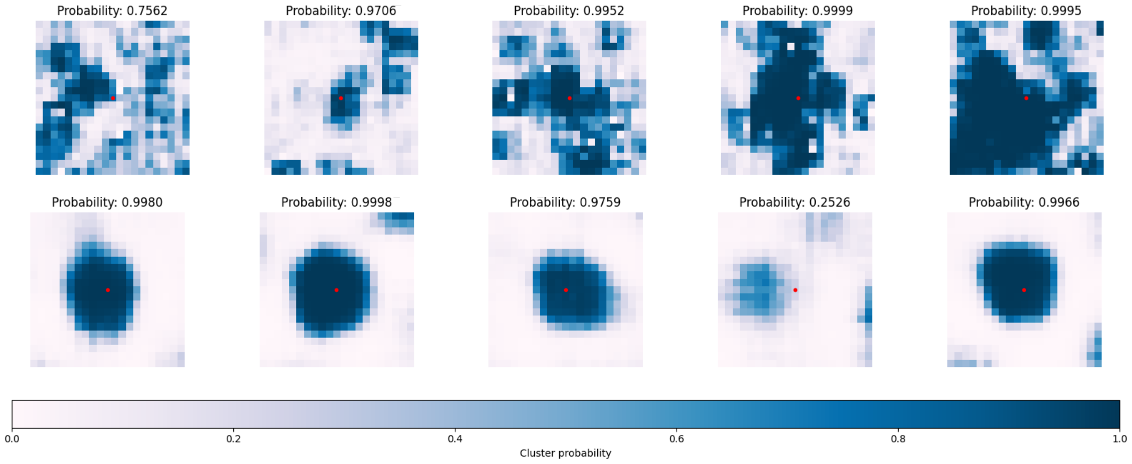
The networks with **good performance** on **both types of data**:

SpinalNetVGG, AlexNetVGG

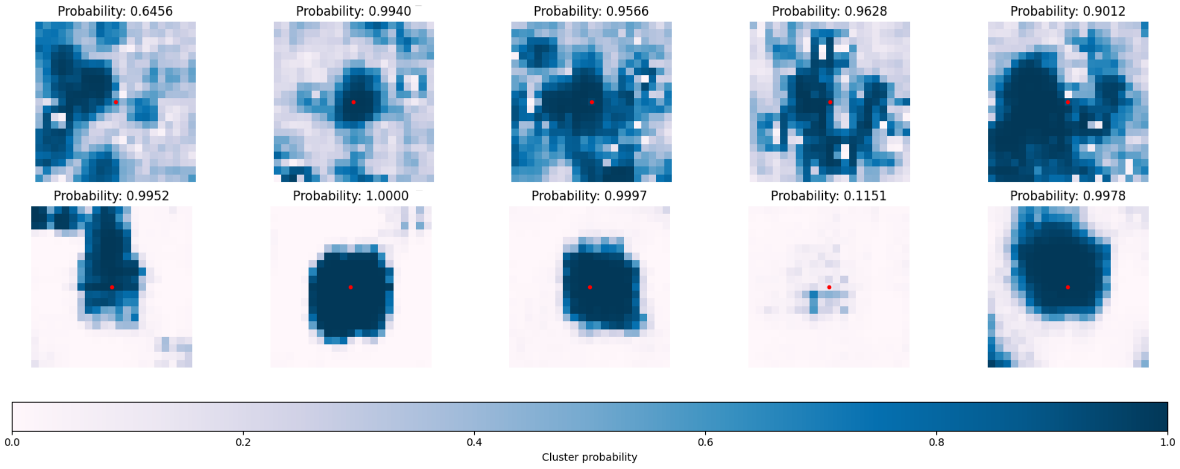
The networks that **show potential:**

DenseNet, CNN+MLP

Probability maps of networks on the IR (top line) and on the microwave (bottom line) data



(a) SpinalNetVGG



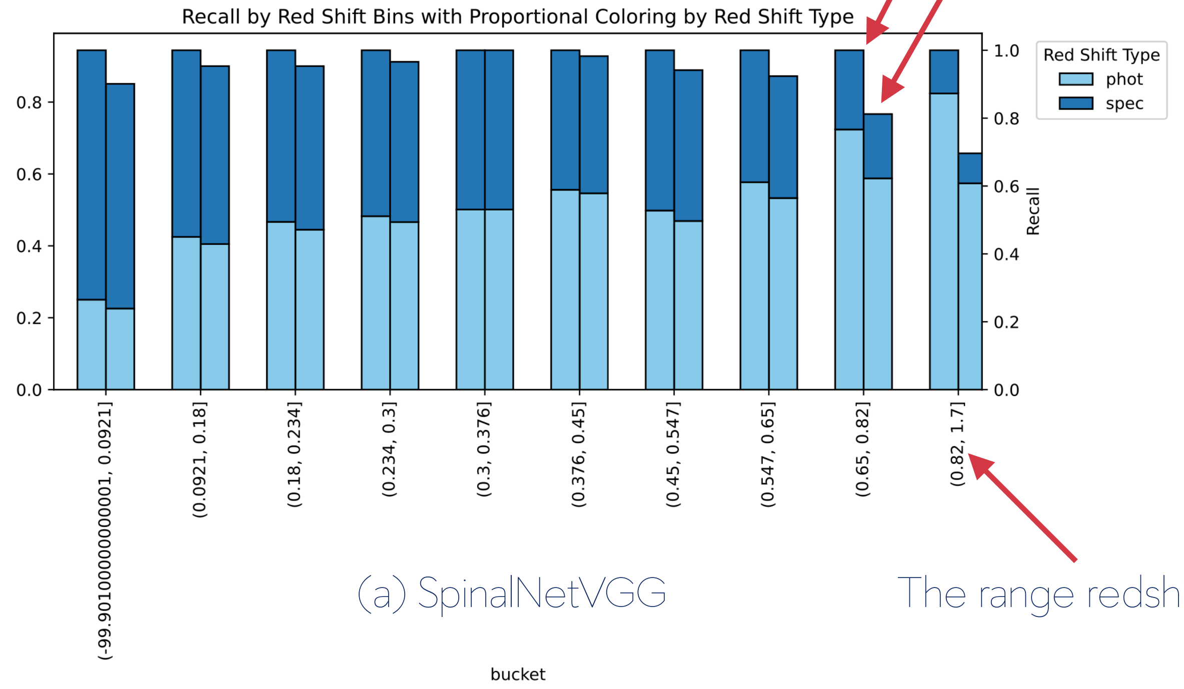
(a) AlexNetVGG

Results 2/2

Recall by redshift achieved on the IR data

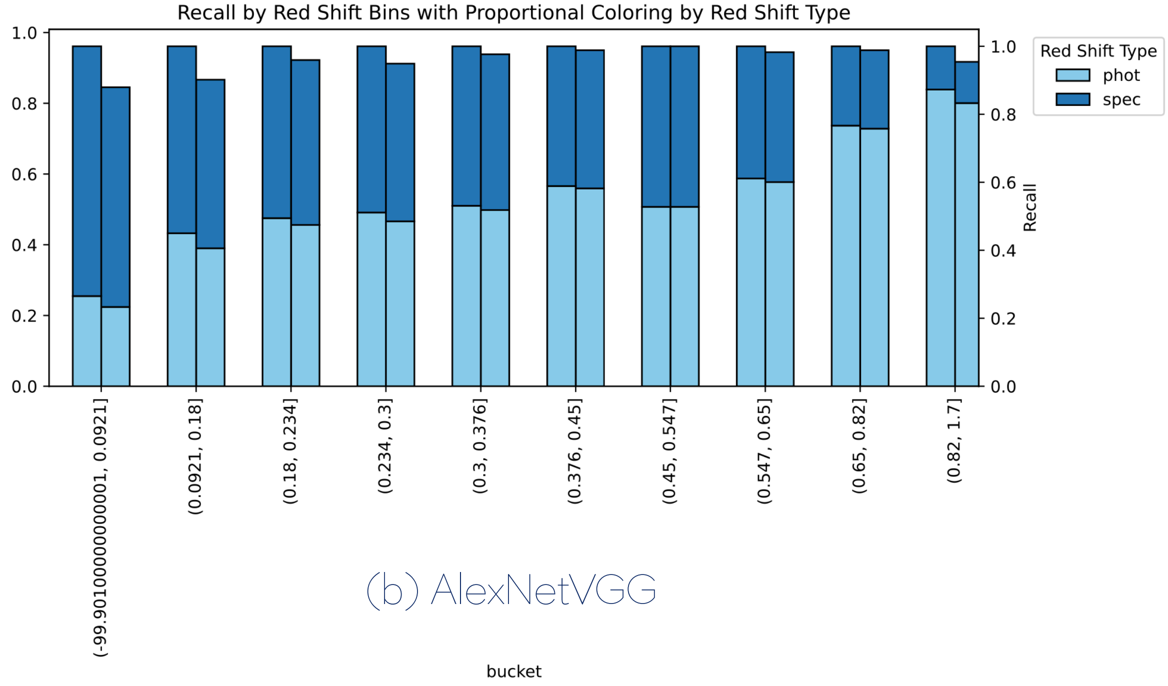
All clusters of the test sample with photometric and spectroscopic redshifts

Clusters identified by the model



(a) SpinalNetVGG

The range redshifts



(b) AlexNetVGG

AlexNetVGG is the most applicable model
for classifying clusters **at high redshifts.**

Future goals

1

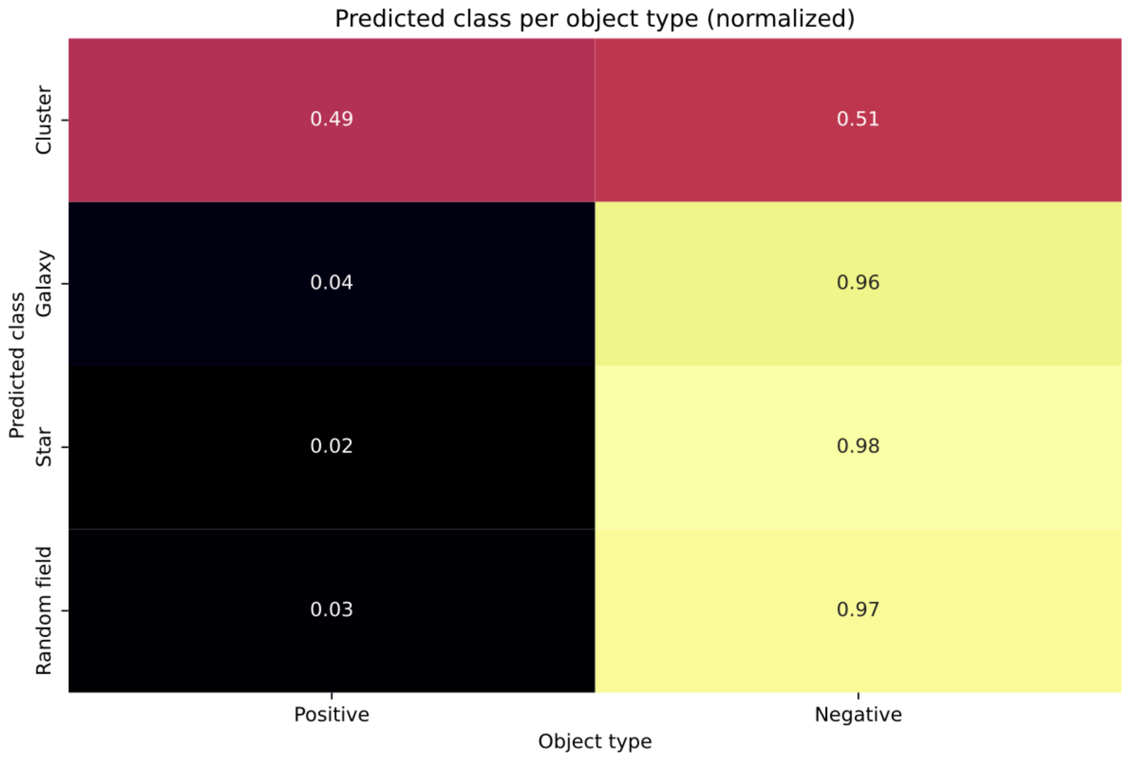
Improve the networks' architectures that show potential in the task of classifying galaxy clusters.

2

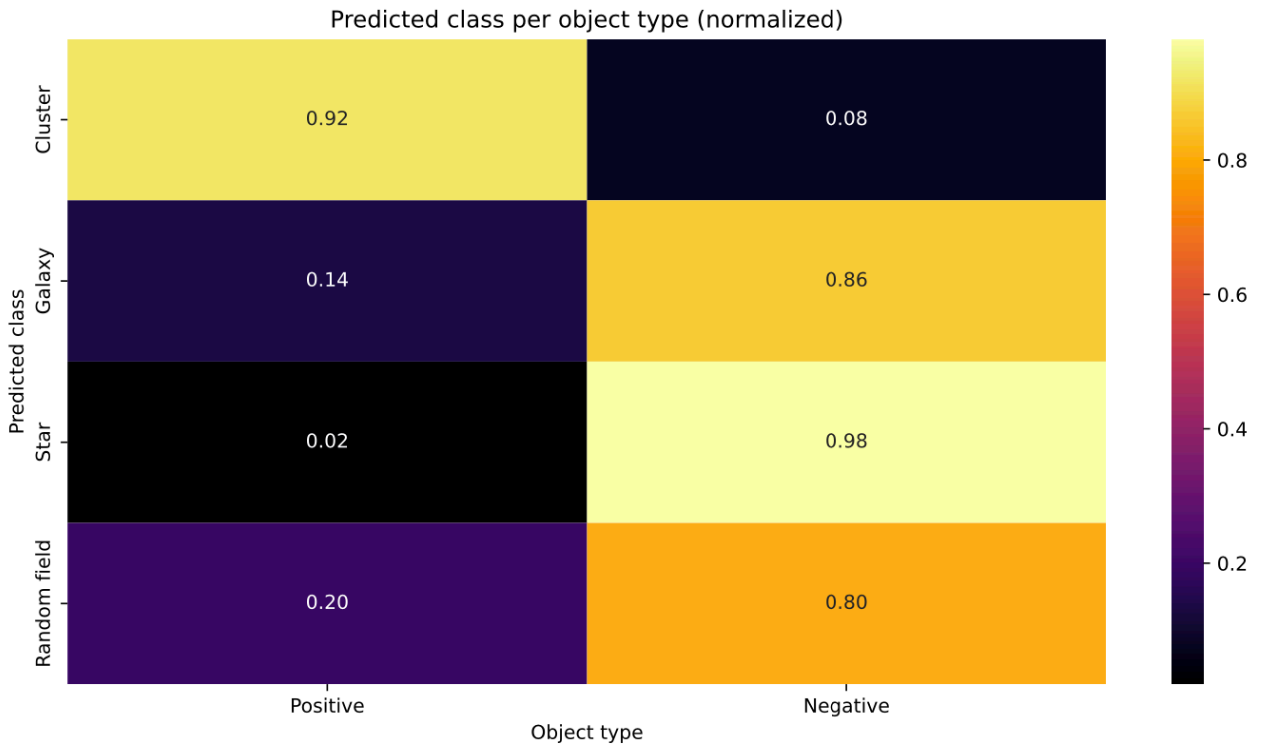
Improve combination of data used in training to overcome limitations and enhance models' performance.

HSE UNIVERSITY • EST 1992

Could be improved, IR data

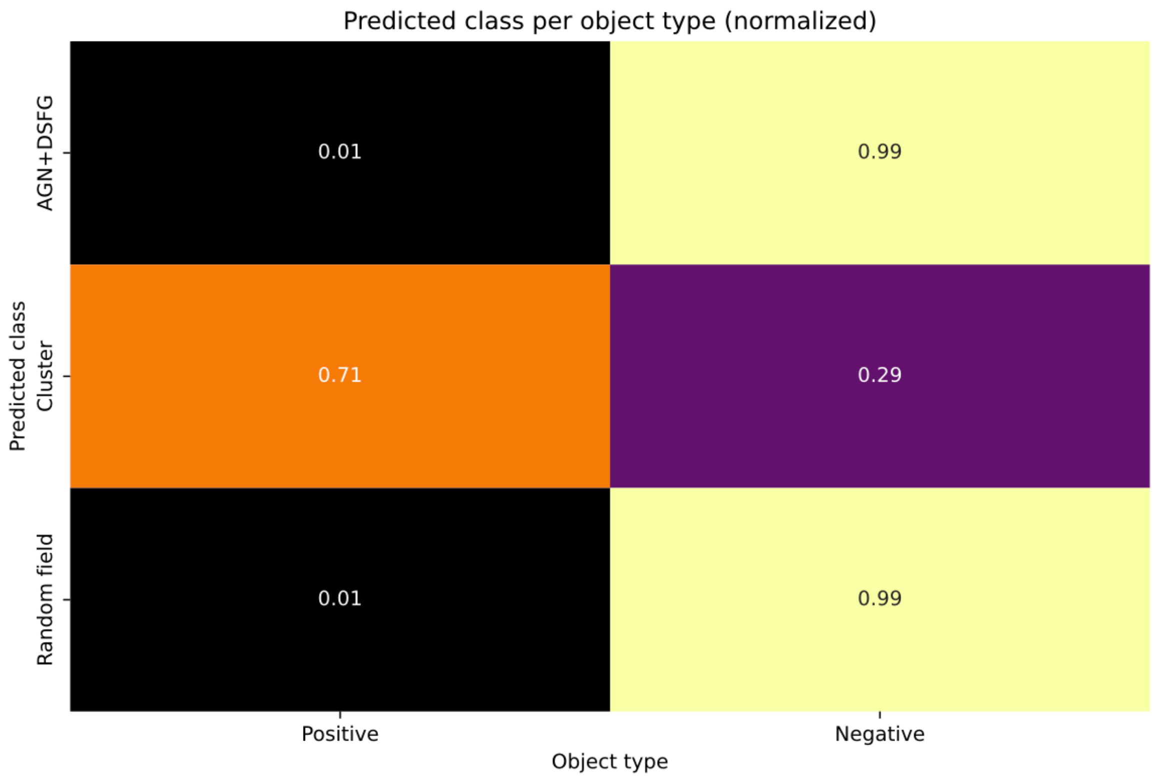


DenseNet

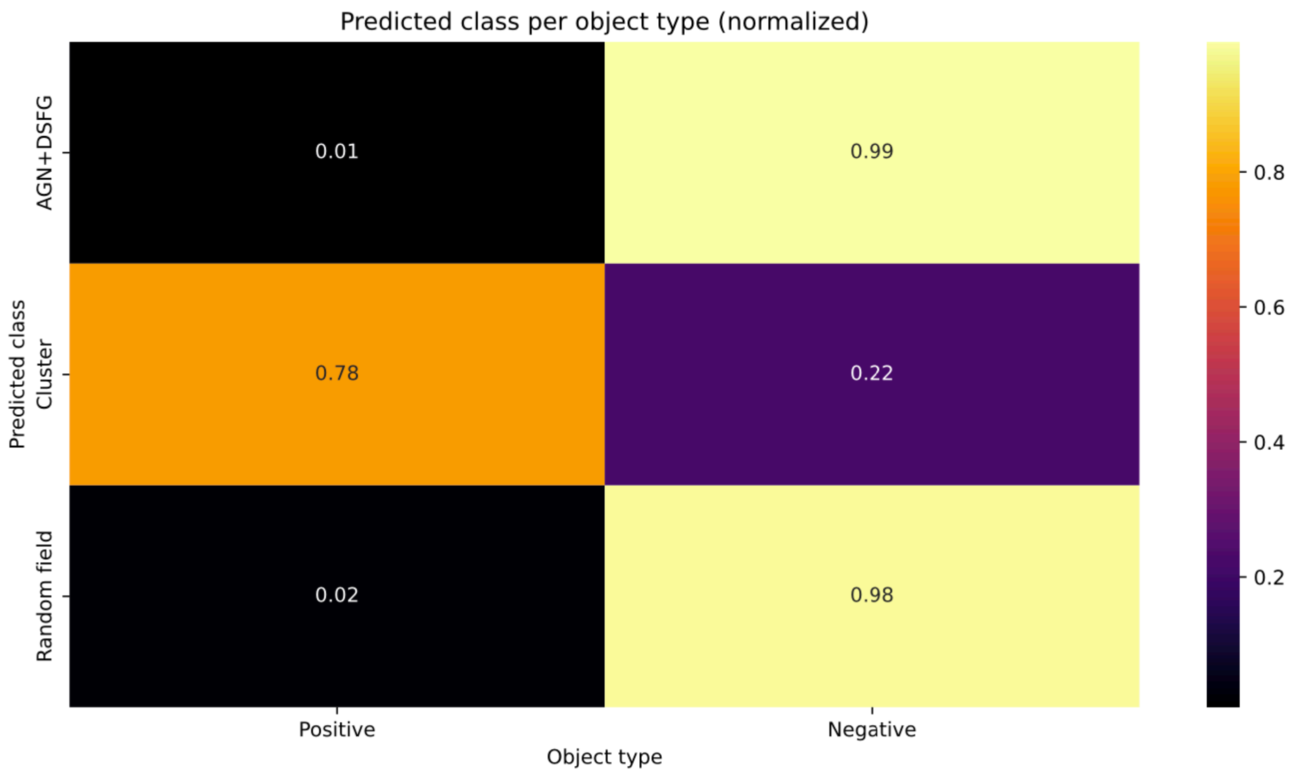


AlexNetVGG

Could be improved, microwave data

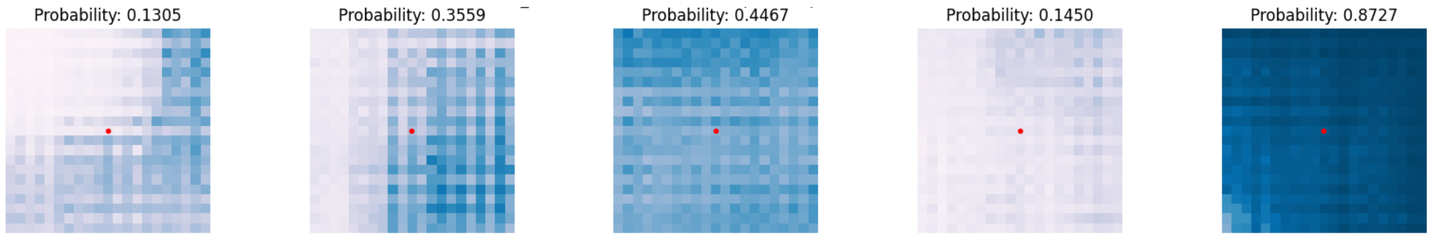


DenseNet

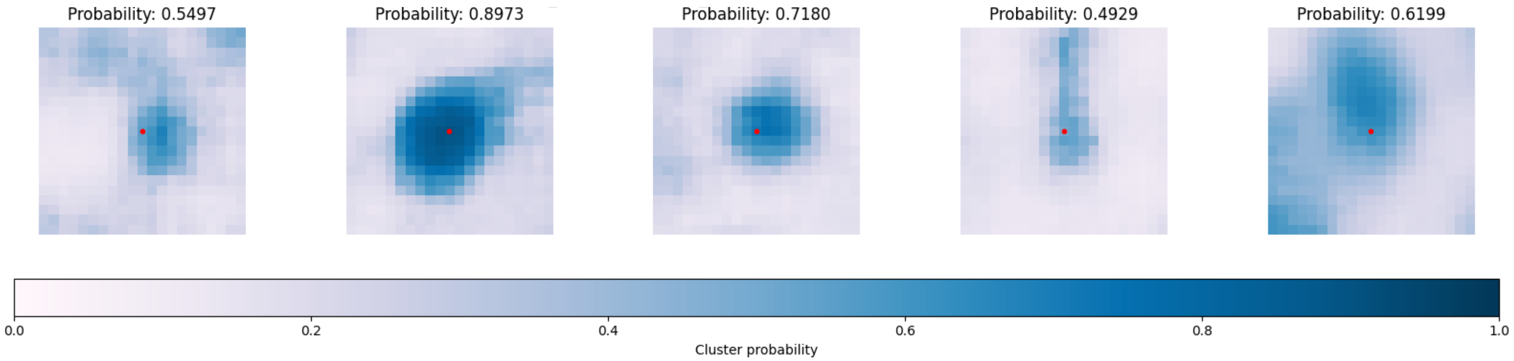


SpinalNetVGG

Examples of poor performance on probability maps

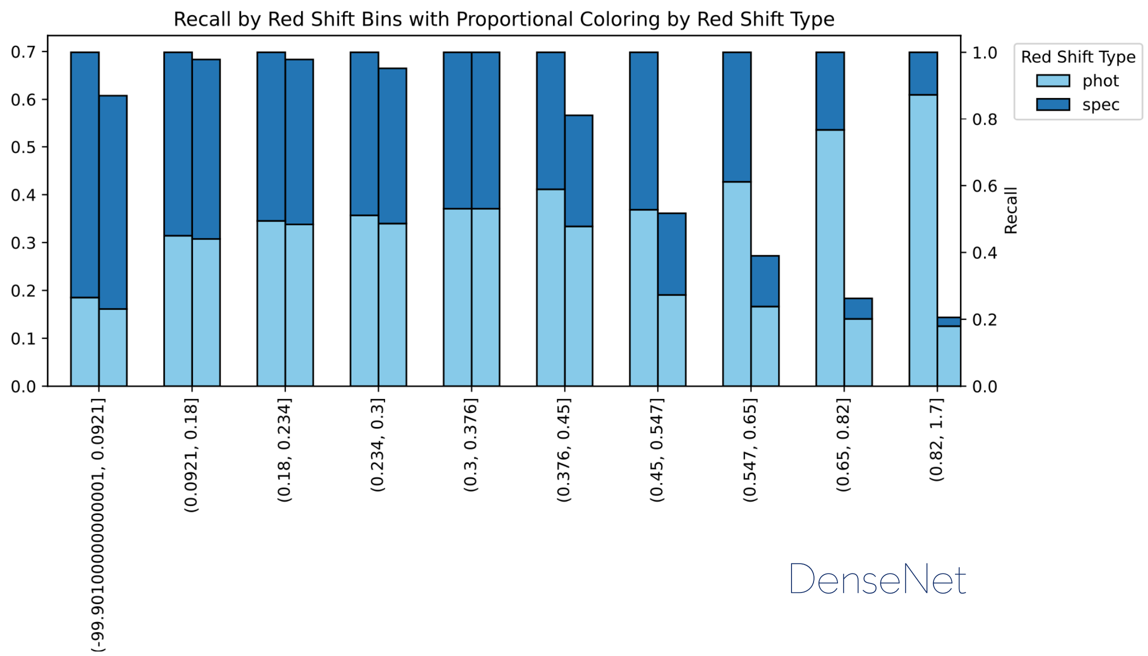


Clusters in microwave bands, ViTL16

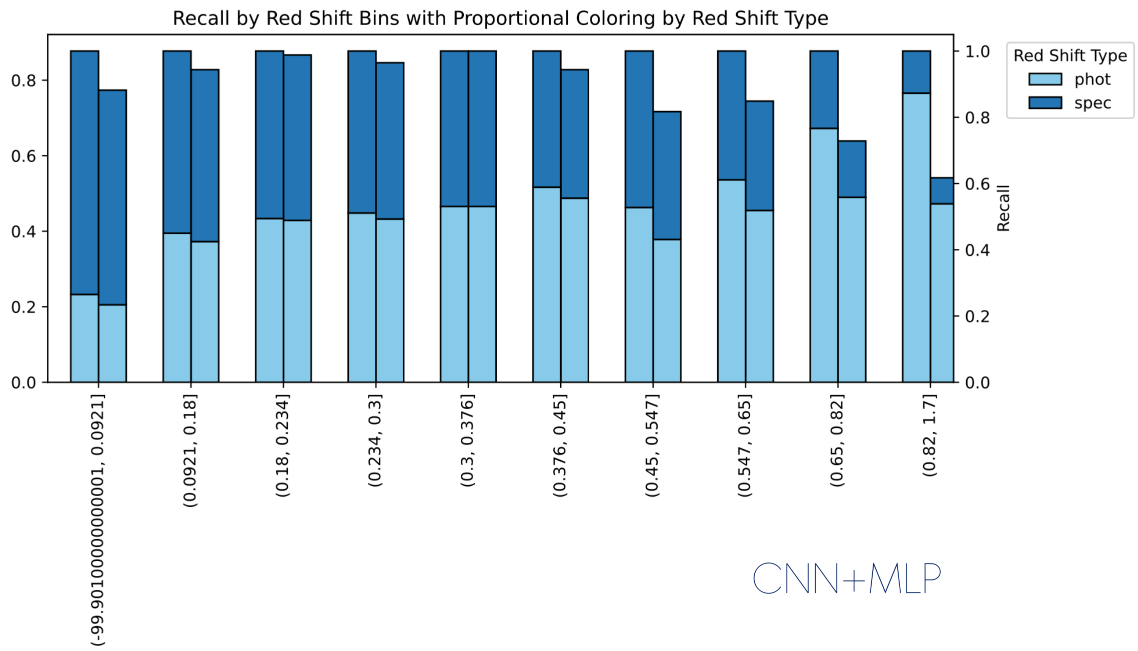


Clusters in microwave bands, CNN+MLP

Recall by redshift on the IR data for the networks that show potential



DenseNet



CNN+MLP