Classifying Human Cell Proteins in Microscope Images

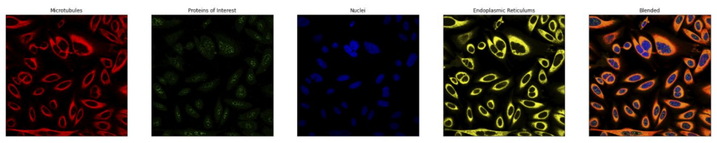
Jacob Lee[1], Scott Bamford[1]
[1] The Graduate Center, City University of New York

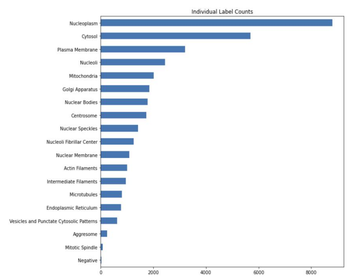
Abstract

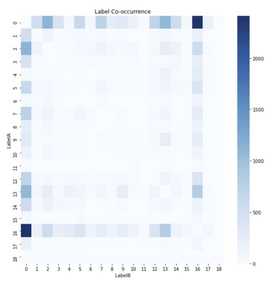
Developing an accurate and reliable model for automatic human cell protein classification can address a major bottleneck in biology research and accelerate the efforts to understand human diseases and find treatments. With a large amount of microscope images manually annotated by subject matter experts, we train a convolutional neural network based on EfficientNetB7 that can classify the cell proteins at the image level with an accuracy of .82 in F1. Furthermore, we provide empirical results of the variations in model performance based on different backbone models and hyperparameter settings. After segmenting individual cells in each image for cell-level classification, the task can be categorized as weak supervision because there are no cell-level gold-standard labels. In order to address this issue, we propose various methods of data analysis for detecting and denoising the anomalies in the dataset. Lastly, we address the label imbalance with class weighting. Code available at: https://github.com/hgjlee/cell_classification

Dataset

- Total image count: 21,806
- 19 Labels
  - 'Nucleoplasm', 'Nuclear Membrane', 'Nucleoli', 'Nucleoli Fibrillar Center', 'Nuclear Speckles', 'Nuclear Bodies', 'Endoplasmic Reticulum', 'Golgi Apparatus', 'Intermediate Filaments', 'Actin Filaments', 'Microtubules', 'Mitotic Spindle', 'Centrosome', 'Plasma Membrane', 'Mitochondria', 'Aggresome', 'Cytosol', 'Vesicles', 'Negative'
- Green channel contains the protein of interest







Backbone Selection and Hyperparameter Tuning

Table 1: Average F1, Recall and Precision Scores for Each Model

Model	Average F1 Score	Average Recall	Average Precision
ResNet152	0.77	0.77	0.77
ResNet1152V2	0.77	0.75	0.70
InceptionNetV2	0.80	**0.81**	**0.81**
EfficentNetB7	**0.81**	**0.81**	**0.81**

Note: The Average Used was Micro Average, in the case that Micro Average was un able to be calculated Macro average was used.

Table 2: Average F1, Recall and Precision Scores for Each Learning Rate

Learning Rate	Average F1 Score	Average Recall	Average Precision	Labeled Correctly
0.1	0.19	0.19	0.19	No
0.01	0.71	0.62	0.72	No
0.001	**0.81**	**0.81**	**0.81**	Yes
0.0001	**0.81**	0.75	0.77	No

Note: In the Case that Micro Average was un able to be calculate the macro average was used.

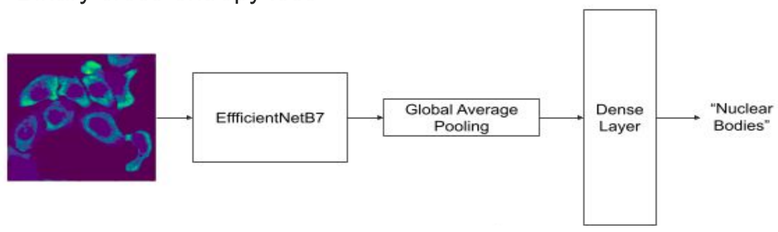
Table 3: Average F1, Recall, and Precision Scores for different Activation Functions

Activation Function	Micro Average F1-Score	Average Recall	Average Precision
ReLu	0.15	0.06	0.01
Tanh	0.15	0.06	0.01
Sigmoid	**0.81**	**0.81**	**0.81**

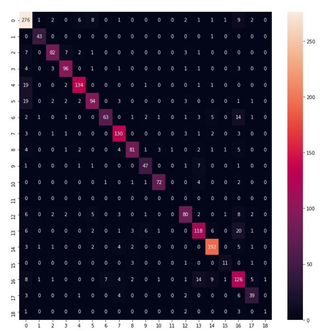
Note: In the case that Micro Average was un able to be calculate the macro average was used

Selected Model

- Layers
  - EfficientNetB7, Global average pooling, Sigmoid activation
- Adam optimizer
- Binary cross-entropy loss



	precision	recall	f1-score	support
0	0.76	0.89	0.82	310
1	0.91	0.98	0.95	44
2	0.87	0.80	0.83	103
3	0.86	0.87	0.86	110
4	0.91	0.85	0.88	158
5	0.82	0.74	0.78	127
6	0.89	0.66	0.76	95
7	0.84	0.90	0.87	144
8	0.87	0.76	0.81	106
9	0.80	0.80	0.80	59
10	0.92	0.89	0.91	81
12	0.80	0.71	0.75	112
13	0.76	0.72	0.74	164
14	0.88	0.91	0.90	211
15	0.73	0.69	0.71	16
16	0.61	0.70	0.65	181
17	0.74	0.71	0.72	55
18	0.50	0.14	0.22	7
micro avg	0.81	0.81	0.81	2083
macro avg	0.80	0.76	0.78	2083
weighted avg	0.81	0.81	0.81	2083

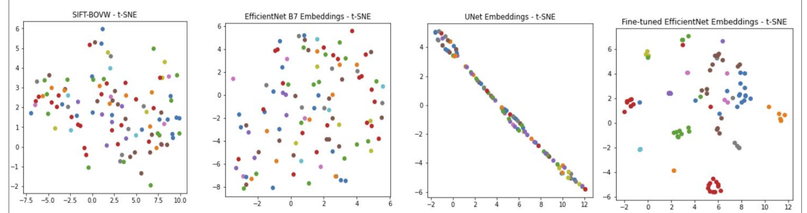


Intensity-based Anomaly Detection

- Mean Intensity Values
- Filtered the images with mean intensities that are more than three standard deviations from the mean value for each class.

	Precision	Recall	Macro F1	F1
No Filtering	0.79	0.72	0.74	0.81
Filtering	0.79	0.76	0.77	0.81

Embedding-based Anomaly Detection



Class Weight

$$1 - \frac{class\ count}{total\ image\ count}$$

	Precision	Recall	Macro F1	F1
Zero Weight	0.79	0.76	0.77	0.81
Class Weight	0.82	0.78	0.79	0.82

Conclusions

- Machine learning models are prone to extensive variations in performance caused by many different factors in hyperparameters.
- Fine-tuning CNN models that have been pre-trained on larger datasets is an effective in solving niche tasks with smaller datasets such as cell classification.
- It is necessary to enforce consistency in the ground truth data, and developing more robust methods of evaluating the data quality can lead to larger performance improvements.
- Label imbalance is a major source of inaccuracy. There are different methods of sampling the imbalanced data and hyperparameter setups during training.