CAP 5610 Machine Learning Course Project

Human Protein Atlas Image Classification

Yan Xiao

July 28, 2020

Florida International University

Jinglin Lyu

Instructor: Dr. Miguel Alonso

**Table of Contents**

[**ABSTRACT** 2](#_Toc46685669)

[**1.** **INTRODUCTION** 2](#_Toc46685670)

[**2.** **LITERATURE REVIEW** 2](#_Toc46685671)

[**2.1.** **Data Preprocessing and Prediction Methods Used in EHR Challenge** 2](#_Toc46685672)

[***2.1.1.*** ***Data Preprocessing*** 2](#_Toc46685673)

[***2.1.2.*** ***Prediction Methods*** 5](#_Toc46685674)

[*2.1.2.1.* *Gradient Boosting Decision Tree Methods* 5](#_Toc46685675)

[*2.1.2.2.* *Generalized Boosted Regression Method* 7](#_Toc46685676)

[*2.1.2.3.* *Logistical Regression Method* 7](#_Toc46685677)

[*2.1.2.4.* *Neural Network* 7](#_Toc46685678)

[*2.1.2.5.* *Combination of Methods* 8](#_Toc46685679)

[***2.1.3.*** ***Prediction Performance*** 8](#_Toc46685680)

[**2.2.** **Prediction Methods Used in Literature** 9](#_Toc46685681)

[**2.3.** **Summary of Literature Review** 10](#_Toc46685682)

[**3.** **DATA PREPROCESSING** 10](#_Toc46685683)

[**4.** **PREDICTION MODEL DELEVLOPMENT** 17](#_Toc46685684)

[**5.** **RESULTS ANALYSIS** 18](#_Toc46685685)

[**6.** **CONCLUSIONS AND RECOMMENDATIONS** 22](#_Toc46685686)

[**REFERENCE** 23](#_Toc46685687)

**List of Figures**

[Figure 2‑1 Modeling Flow Chart Using by Banda et al. (2020) 8](#_Toc46685701)

[Figure 3‑1 An Example of Person and Death Data Preprocessing 11](#_Toc46685702)

[Figure 3‑2 An Example of Person and Death Data Preprocessing (Source: Reich et al. 2016) 13](#_Toc46685703)

[Figure 3‑3 Age Distribution in Training Data Set 15](#_Toc46685704)

[Figure 3‑4 Gender Distribution in Training Data Set 15](#_Toc46685705)

[Figure 3‑5 Race Distribution in Training Data Set 16](#_Toc46685706)

[Figure 3‑6 Ethnicity Distribution in Training Data Set 16](#_Toc46685707)

**List of Tables**

[Table 2‑1 Feature Extracted from Data (Brygere et al., 2020) 5](#_Toc46685745)

[Table 2‑2 Model Parameters Used by Choi et al., 2020 8](#_Toc46685746)

[Table 2‑3 Feature Extracted from Data (Brygere et al., 2020) 8](#_Toc46685747)

[Table 2‑4 XGBoost Modeling Parameters (Togias and Pandi., 2020) 9](#_Toc46685748)

[Table 2‑5 Modeling Parameters (Gaudelet, 2020) 10](#_Toc46685749)

[Table 4‑1 LGBMClassifier Parameters 19](#_Toc46685750)

[Table 4‑2 Values of LGBMClassifier Used 20](#_Toc46685751)

[Table 5‑1 Impacts of Boosting Types 21](#_Toc46685752)

[Table 5‑2 Impacts of Resampling Strategies 21](#_Toc46685753)

[Table 5‑3 Impacts of Modeling Strategies 21](#_Toc46685754)

[Table 5‑4 Limit Data within Six Months 22](#_Toc46685755)

[Table 5‑5 Parameter Sensitivity Analyses 22](#_Toc46685756)

[Table 5‑6 Performance Ranking 24](#_Toc46685757)

# **ABSTRACT**

High-resolution microscope images of protein provide useful information regarding diseases. The Human Protein Atlas hosted an image classification challenge at the Kaggle website. A convolutional neural network model was developed in this project to identify mixed protein pattern in human cells using images provided by the challenge.

# **INTRODUCTION**

Protein pattern in human cell contains important information regarding diseases. Correct classification of protein patterns is very helpful for better understanding diseases and production of medicine. Traditional studies only classify protein into a single pattern or very limited cell types, which may not be able to capture the variation in human cells.

With the availability of high-resolution microscope images, a further examination of protein complex patterns become feasible. The Human Protein Atlas hosted an image classification challenge at the Kaggle website. It would be interesting to develop a machine learning model based on the image data provided by this challenge to identify mixed protein patterns in human cells. Various machine learning techniques have been used in practice for image classification, for example, support vector machine, decision trees, k nearest neighbor, convolutional neural network, and so on (Jain, 2020). Among these methods, convolutional neural network seems to be the most commonly used deep learning method for analyzing images. Therefore, the goal of this project is to develop a convolutional neural network model to classify protein types that exist or coexist in human cell images.

# **LITERATURE REVIEW**

This section provides a brief review of modeling techniques, structures, and performance that have been reported in the literature for protein image classification.

Yang et al. (2019) provided a summary of Human Protein Atlas Image Classification competition, which covered different aspects of competition including participation and performance, and strategies used by the top-ranking solutions. According to the survey conducted by the competition organizer, 44 out of 56 teams used variation of neural network architectures such as Inception, Densenet, and Resnet. 34 out of 56 teams used binary cross entropy as loss function. Other strategies for handling unbalanced classes and improving performance include applying class weights, data augmentation, and adding supplementary images.

Pärnamaa and Parts (2017) trained 11-layer neural network model based on high-throughput microscopy data to classify 12 fluorescent protein subcellular localization patterns in yeast cells. The first 8 were convolutional layers with rectified linear units. The number of units used in the first two convolutional layers was 64, 128 for the layers 3 and 4, and 256 for the remaining 4 convolutional layers. The last 3 layers were fully connected layers with a number of units of 512, 512, and 12. The output class was determined by the softmax function.

Rumetshofer et al. (2019) proposed a new CNN architecture (GapNet-PL) to determine the existence or coexistence of 13 major organelles in Human Protein Atlas project protein images. The GapNet-PL consisted of two-steps. The first step was to use an encoder that was made of several convolutional layers to learn features on different spatial resolutions. The second step was to reduce features to vectors through global average pooling and paas them to a fully connected network with two hidden layers to make final prediction. SELU was used as activation function. The proposed architecture was compared to DenseNet, Multi-scale Convolutional Neural Network (M-CNN), DeepLoc, FCN-Seg, and Convolutional Multiple Instance Learning (Convolutional MIL). The proposed GapNet-PL showed better performance in terms of the metrics of F1 score and AUC.

Ezat et al. (2019) applied convolutional neural network that was pre-trained on CNN CAFFE Image-Net based on large dataset of ILSVRC to classify the PASCAL VOC 2007 image dataset into 4 classes. The performance of the developed model was compared to support vector model and showed better performance.

The Kaggle website posted a leaderboard for Human Protein Atlas Image Classification Challenge (Kaggle, 2020). The performance is quantified by macro F-1 score using 29% of the test data. There is a total number of 2160 teams listed on the leaderboard. The highest score is 0.65602 while the lowest score is 0.0.

# **DATA EXPLORATION AND VISUALIZATION**

## **Human Protein Atlas Image Classification Challenge Dataset**

The dataset used in this project was downloaded from the Kaggle website for Human Protein Atlas Image Classification Challenge. It consists of two csv files (tain.csv and sample\_submission.csv) and two image folders (one for training and another one for testing). Figure 3-1 shows the first 5 rows of train.csv. As shown in this figure, there are two columns. The first column lists the image id, which can be used to find the corresponding images under the train image folder. The second target column lists the types of protein found in that image. A total number of 28 protein types have been included in the challenge dataset. In order to use those target protein types in the development of multi-class classification model, the target column was converted into multiple binary columns that correspond to 28 protein types in this study, as shown in Figure 3-2. A total number of 62,144 samples is included in the training dataset and 23,404 samples for the testing dataset.



**Figure 3‑1 Top 5 Rows in Train.csv File**

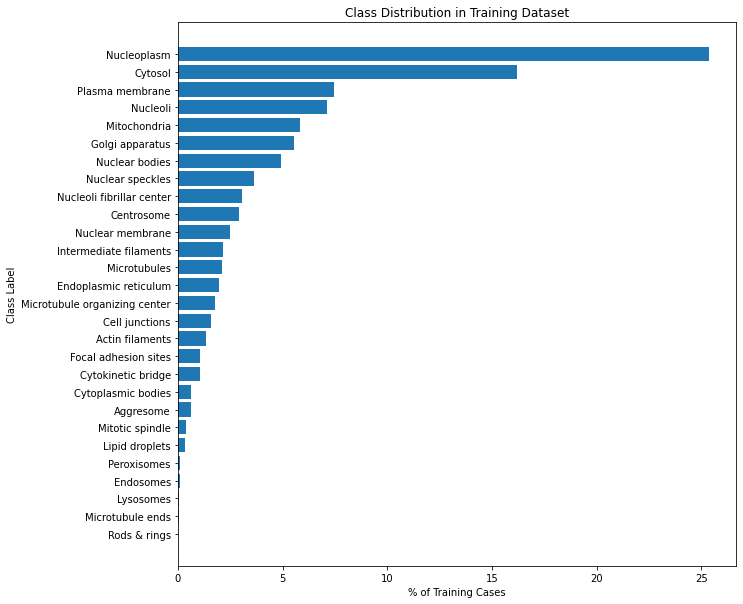


**Figure 3‑2 Example of Binary Target Columns after Conversion**

## **CSV File Exploration and Visualization**

* **Examination of Class Distribution in Training Dataset**

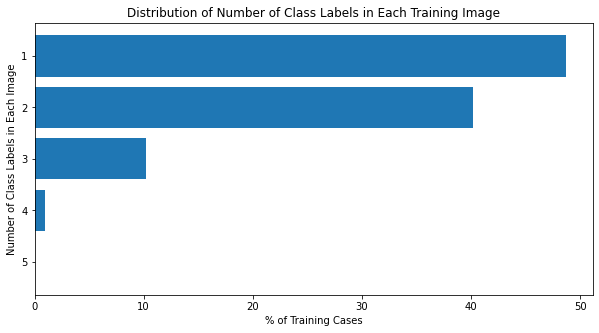
As the first step of data exploration and visualization, the occurrence of each class in the training data were counted. Note that one sample image may have multiple classes. The corresponding percentage of each class were calculated and shown below. It is seen from this figure that the most common protein is Nucleoplasm (Class 0) having a percentage of 25.4%, followed by Cytosol (Class 25) with a percentage of 16.2%, and 7.43% of Plasma Membrane (Class 21).



**Figure 3‑3 Class Distribution in Training Dataset**

* **Number of Coexisting Classes in Training Images**

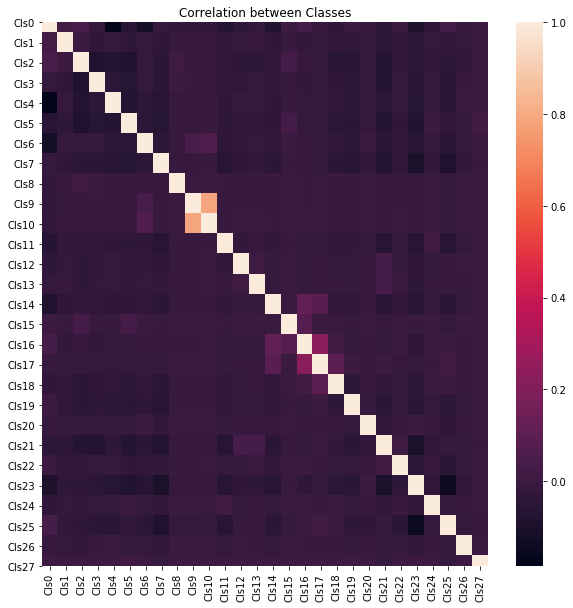
The number of classes that exist or coexist in a training image were counted and the corresponding percentage distribution is shown in Figure 3-4. This figure shows that about 50% of images only have one type of protein. About 40% of training images have two class labels. The images having 3 or more class labels are relatively rare.



**Figure 3‑4 Distribution of Number of Classes in One Image**

* **Correlation between Class Labels**

The purpose of this exploration is to examine which pair of classes more likely coexist in a Human cell image. The correlation between two class columns were found and visualized in the heat map below (Figure 3-5). The correlation matrix shows that Classes 9 and 10 (that is, Endosomes and Lysosomes) have a higher correlation, which means that they are usually located in the same image. Classes 16 and 17 (i.e., Cytokinetic bridge and Mitotic spindle) have certain correlation.



**Figure 3‑5 Correlation between Protein Types**

## **Image File Exploration and Visualization**

* **Check If Each Image ID Have 4 Images**

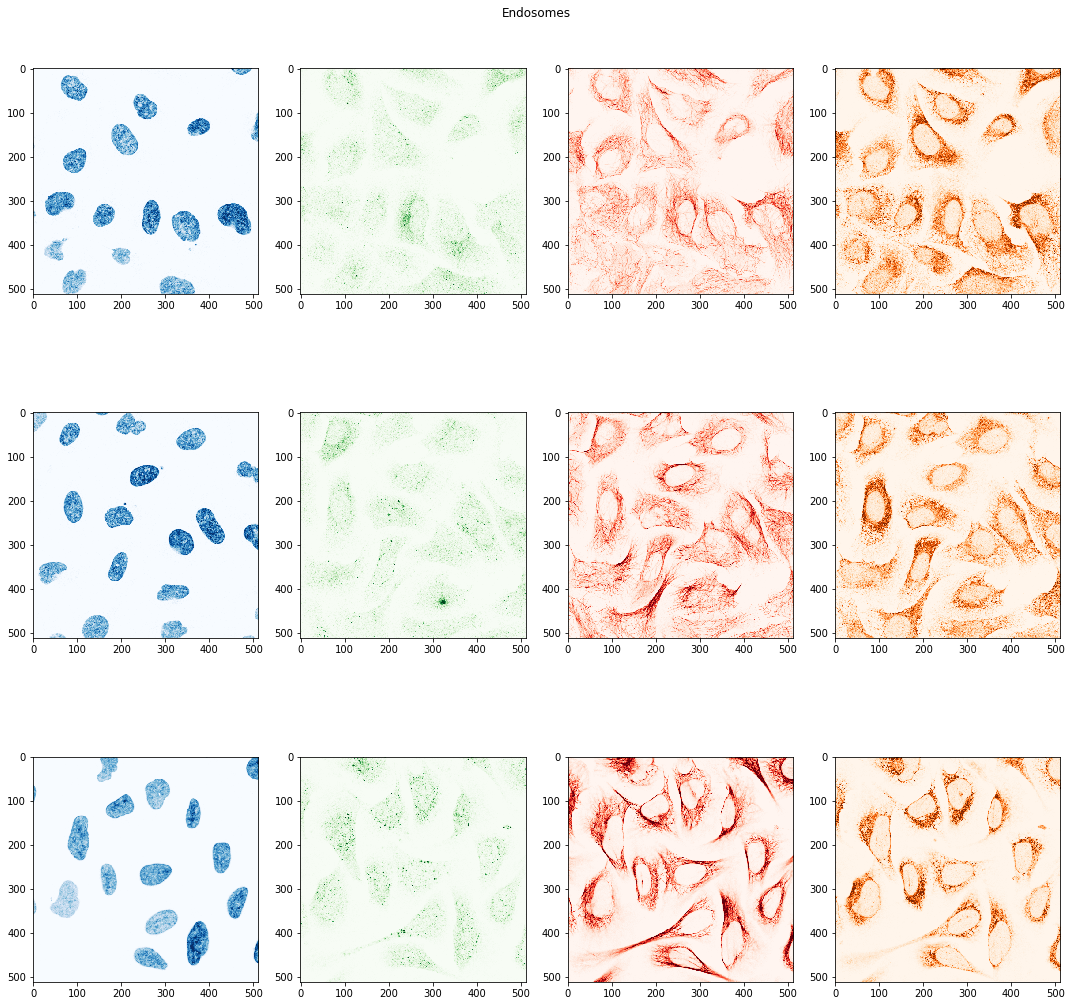
Each image id in Human Atlas Image dataset include 4 images that are corresponding to 4 color channels, red, green, blue, and yellow, respectively. A total number of 62,144 image ids is listed in train.csv file and there are 124,288 (that is, 62,144x4) images in the training image folder. It means that no images are missing. Similarly, it is found that all the 23,404 image ids listed in the submission csv file have 4 corresponding single-channel images.

* **Check If Any Image File is Corrupted**

Codes were written to open each image file and check if any image file is corrupted. The check results indicate that every image file is good.

* **Image Visualization**

Example of images were loaded to program and visualized. Three samples of Class Endosomes are visualized in Figure 3-6. As shown in this figure, these three samples have some similar patterns but some differences can also be observed



**Figure 3‑6 Three Samples of Protein Cells with a Class of Endosomes**

# **METHODOLOGY**

The methods have been used in the Human Protein Image Classification challenge as well as the multi-class classification methods reported in the literature will be reviewed. Based on literature review, a convolutional neural network will be developed to classify mixed protein patterns using the image datasets from Human Protein Image Classification challenge. The model performance will be quantified in terms of macro average of F1 score. Strategies that may improve model performance will be explored.

## **Data Preprocessing**

## **Model Development**

## **Results Visualization**

# **RESULTS AND DISCUSSION**

This section lists and analyzes the results that have been obtained in the study.

<https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/73246> threshold 0.15

<https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77269> threshold 0.2-0.3

# **CONCLUSIONS AND RECOMMENDATIONS**

This study

# **REFERENCE**

Ezat, W. A., M. M. Dessonuky, and N.A. Ismail. (2020). “Multi-class Image Classification Using Deep Learning Algorithm.” Journal of Physics: Conference Series. pp. 1447.

Jain, T. (2020). “Basics of Image Classification Techniques in Machine Learning.” <https://iq.opengenus.org/basics-of-machine-learning-image-classification-techniques/>. Accessed on July 28, 2020.

Kaggle. (2020). “Human Protein Atlas Image Classification.” <https://www.kaggle.com/c/human-protein-atlas-image-classification/overview>. Accessed on July 28, 2020.

Pärnamaa, T. and L. Parts (2017). “Accurate Classification of Protein SSubcellular Localization from High-Throughput Microscopy Images Using Deep Learning.” *G3 Genes | Genomes | Genetics*. Volume 7, pp. 1385-1392.

Rumetshofer, E., M. Hofmarcher, C. Rohrl, and S. Hochreiter. (2019). “Human-Level Protein Location with Convolutional Neural Networks.” Conference Paper at ICLR 2019.

Yang, W., C. F. Winsnes, M. Hjelmare, and et al. (2019). “Analysis of the Human Protein Atlas Image Classification Competition.” *Nature Methods.* Vol. 6, pp. 1254-1261.