Scientific Image Classifier

#Track 1 Sponsored by Cactus

Team Dataverse  
Members:  
Harsh Ratna

Shivangi Singh

Prateek Jha

# Introduction

Image classification is crucial for computer vision, allowing machines to interpret visual data effectively. By categorizing images accurately, it enables diverse applications across different fields, from autonomous vehicles to medical diagnostics, revolutionizing machine perception.

**Proposed Classification Project in Biomedical Imaging:**

Develop a robust computer vision tool capable of accurately classifying various types of life science images, including microscopies, histology slides, pathology slides, western blots, gel electrophoresis, flow cytometry, fluorescence activated cell sorting (FACS), and etc. The tool should provide precise classification outputs for individual images, aiding researchers and professionals in efficiently analyzing vast amounts of scientific data

**Model used for classification.**

* We first built a CNN Model based on VGG16 architecture and trained it from scratch. Hyper-parameter tuning was done and it achieved a best validation accuracy of 86%.
* Next, to improve the accuracy, we used the VGG16 pretrained model by Google, and achieved an accuracy of 96%.
* VGG16 is used because of its simplicity and it works well with moderately sized datasets as in our case (2.5k to 3k images per class.

**Challenges and Trends:**

Generalization Limitation: While the code performs well on provided datasets, its capability to generalize to unseen data or diverse domains is unexplored. Further validation across varied datasets is necessary to assess its robustness and limitations beyond the current scope.

Hyperparameter Sensitivity: The performance of the code could be sensitive to hyperparameter settings, requiring extensive tuning to achieve optimal results, which can be time-consuming and resource-intensive.

Interpretable Output : While our classification model may achieve high accuracy, ensuring interpretability of its outputs is essential for researchers and professionals in the life sciences field to understand and trust the model's classifications.

.

# Dataset Exploration

Combining the BioFors Dataset with Graphs Dataset, COVID-19 Radiography Database, and SARS-COV-2 Ct-Scan Dataset resulted in a diverse collection for our classifier, encompassing seven distinct classes. However, the class distribution exhibited significant imbalance, with 27k images for blot/gel class and only 1061 for FACS. To address this, extensive data augmentation was applied to the FACS class to enhance its representation, while moderate augmentation was applied to other classes to maintain balance. Ultimately, each class comprised around 2500 images, ensuring robust training and classification performance.

A screenshot of a medical scan

Description automatically generated

# Screenshots

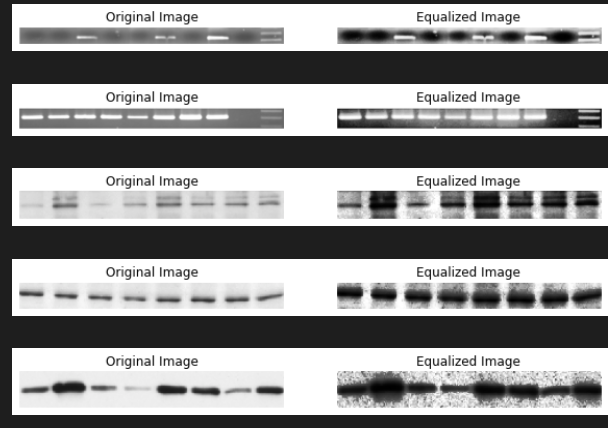
* A screenshot of a computer screen

  Description automatically generated watershed thresholding:

A black and white image of a graph

Description automatically generated

* Histogram Eqaulization



* Sharpening filters.

A screenshot of a computer

Description automatically generated

* Gamma correction on Laplacian filter

A screenshot of a graph

Description automatically generated

**Proposed Approach:**

* Preparing metadata by combining multiple datasets including BioFors, Graphs Dataset, COVID-19 Radiography Database, and Ct-Scan Dataset to ensure diversity and richness of data.
* Dataset Exploration through Graphs and Plots
* Augmentation of classes with limited images to address class imbalance, focusing particularly on classes such as FACS.
* Filtering out the dataset to retain approximately 2500 best images per class for improved model training and performance.
* Application of image enhancement techniques such as watershed thresholding, histogram equalization, and denoising to preprocess the images, enhancing their quality and clarity.
* Utilization of a 48-layer convolutional neural network (CNN) architecture for classification purposes, leveraging the processed images to train the model effectively and achieve accurate classification results.

**Conclusion:**

In conclusion, our developed classification tool exhibits promising capabilities in accurately categorizing diverse life science images, contributing to streamlined data analysis in research and diagnostics. Continual refinement and validation will further enhance its utility and reliability in real-world applications, fostering advancements in the field of biomedical imaging.