Product design / development / engineering | Pharma / biopharma | Improved process performance | Auto-encoders

Computer Vision for Cell Line Development

BUSINESS PROBLEM

One of the most crucial interfaces in the drug development process is Cell Line Development (CLD). The general process consists of screening thousands of single cell clone candidates to determine a high quality Master Cell Bank (MCB). All future drug substance derive from this one step and the MCB. However, some candidates for the MCB are anomalies, requiring operators to spend hundreds of hours manually checking. Using the various forms of data generated in this process, Amgen looks to expedite this down-selection process and ensure anomalies are caught through the use of Machine Learning models.

APPROACH

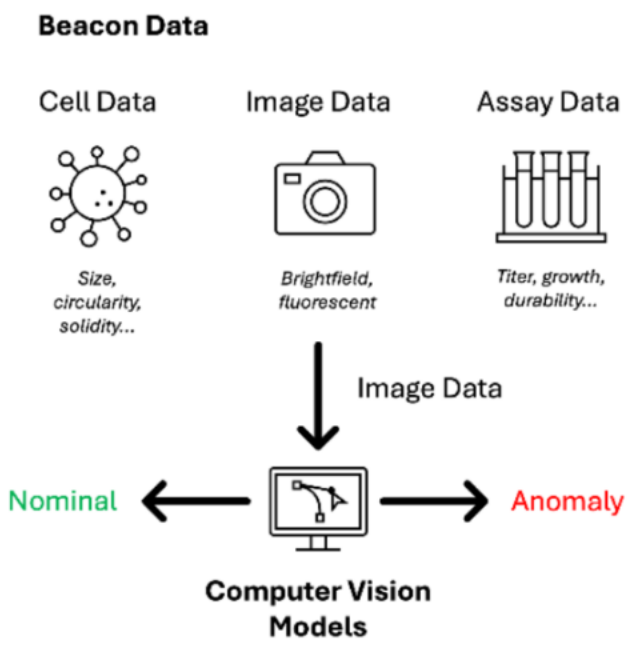
Three computer vision techniques were applied to identify anomalies without human intervention: a Fluorescent Image Autoencoder, an Edge Identification Convolutional Neural Network, and an RGB Channel Support Vector Machine (SVM). These techniques are evaluated for their effectiveness in anomaly detection within a historical dataset and the necessary resources.

DATA SOURCES

The data primarily came from thee Beacon platform. Three types of data are produced inthe cell process: Cell Data, Assay Data, and Image Data. For a computer vision model, Image Data is the primary focus. Image Data consists of brightfield and fluorescent images, but models were only developed for fluorescent image anomaly detection.

Data Types and Format

Image Data is generated as 22 brightfield and fluorescent png images. These images are spliced and split to produce 1,758 individual png images for model input and training.



Author: Jack Albright

IMPACT

Implementing a computer vision model for anomaly detection has clear benefits for the CLD Process. To start, anomaly identification is a highly subjective process. Scientists in general have shared criteria that htey look for when making decisions on what is an anomaly versus a nominal image. However, scientists will focus on different small details that may lead them to separate conclusions. By introducing a model to supplement this process, the key factors that determine an anomalous image are incorporated to better standardize decision making. Reducing the subjective nature of the process means that the model reinforces selection of quality candidates for downstream development, while also providing an interface for scientists to exercise their expertise if they do not agree with the classification prediction. Concurrently, the amount of time scientists need to spend manually assessing whether cell clones are anomalies dramatically reduces. Originally, scientists could spend hundreds of hours going through this process, but by implementing a model to standardize this process, they now would only need to look over what is asserted as an anomaly. Further improving the model could result in scientists not needing to even dedicate time towards checking the validity of predicted anomalies, instead focusing on which clones are truly the top candidates for the MCB.

DRIVERS

Computer vision is widely used in biopharmaceuticals especially for detecting faults or quality issues. Leveraging the application of computer vision for challenges related to the biopharma manufacturing environment allowed focus on techniques most applicable for the CLD use case. Additionally, the data science experience within Amgen's TDC organization provided support.

BARRIERS

Fluorescent images coming from CLD have in-class variation that makes it difficult from a traditional autoencoder to learn what distinguishes a positive from negative class. Additionally, anomalies are outnumbered by nominal images. The class imbalance results in computer vision techniques readily skewing in favor of a nominal prediction as there is more data.

ENABLERS

The Transformative Digital Capabilities (TDC) organization at Amgen, the interface between data science experts and Process Development teams, was pivotal in the completion of this project. Working as part of TDC meant I had quick access to a range of expertise and support in establishing data pipelines to feed the models.

ACTIONS

I collaborated with the CLD team consistently to ensure the models developed would be useful for the scientists that would operate them. I also collaborated with another TDC team to ensure outputs from the computer vision models would serve as inputs to their more comprehensive models used for expedited clone selection.

INNOVATION

Per research on computer vision in biopharmaceuticals, there has not been an application of an autoencoder specifically used for fluorescent image anomaly detection. The innovative aspect comes from the ability of the autoencoder to be continuously trained on newer data after a batch run as been completed and labels for images are generated. The approach is self-reinforcing once implemented.

IMPROVEMENT

In its first iteration, the best performing autoencoder has an accuracy of 94% for anomalies and nominal images. Its precision for anomalies is also over 94%, while its recall remains suboptimal at 79%. However, its curve and composite metrics are also all significantly high, and the model correctly identifies 414 or 523 anomalies present in the evaluation set.

BEST PRACTICES

Establishing a strong data pipeline is critical so each model uses the same data as input. Additionally, it is important to establish a comparison baseline, especially in anomaly detection settings where class imbalances can impact accuracy and precision metrics.

OTHER APPLICATIONS

These solutions can be applied to other image types that may come from the rest of the process, such as brightfield images. The data pipeline for brightfield images has been established, so utilization of the models would merely require some architectural tweaking to determine whether this image type is suitable for making predictions.