Clearer Views for Cancer Research: Upgrading Chromosome Analysis

Sophia Lugo slugo@ucsd.edu

Niha Malhotra n1malhot@ucsd.edu

William Trang wtrang@ucsd.edu

Vineet Bafna vbafna@ucsd.edu

Gino Prasad giprasad@ucsd.edu

Abstract

Cancer research is pivotal for understanding and combating the disease, with cell research playing a crucial role due to the genetic mutations leading to tumors. However, manual analysis of cell images is time-consuming and error-prone. Our project proposes to develop an advanced system for understanding variations in chromosome numbers and structures. Our project targets the development of an imaging system to accurately segment individual chromosomes within cell images, which addresses the key issue of past systems that weren't able to do so. It's especially important because cells with abnormal chromosome counts contribute to cancer growth by changing the number of oncogenes. To create our imaging system, we decided to transition from semantic segmentation to instance segmentation in order to delineate individual chromosomes. This approach contrasts with existing methods, such as ecSeg, which, while useful for identifying chromosome clusters, falls short in accurately quantifying individual chromosomes. Our project aims to bridge this gap by leveraging advanced computational methods to develop optimal bounding boxes for each chromosome, facilitating precise identification of the chromosome centers. The anticipated outcome is a robust tool capable of providing detailed chromosomal profiles in cancer cells during metaphase, thereby contributing to a deeper understanding of cancer genetics and potentially unveiling new avenues for diagnosis and treatment.

Website: https://sophialugo.github.io/Capstone/
Code: https://github.com/williamtrang/VaCe

1 Introduction

Cancer cells have been shown to have genetic mutations that lead to uncontrolled cell division that can often lead to the creation of tumors (ChemoMetec 2199). This happens through oncogenes, a mutated gene that has the potential to turn cancerous, that is carried in extrachromosomal DNA, also known as ecDNA (Turner 2017). The analysis of cells, particularly through the examination of chromosomes, is a cornerstone in understanding cancer's genetic basis. However, this analysis is fraught with challenges, including the labor-intensive and error-prone nature of manual cell image analysis. Our project seeks to address these challenges by developing an imaging system designed to detect individual chromosomes in metaphase cell images with high precision. This is crucial as the replication of ecDNA, which often accompanies the formation of new chromosomes, results in the amplification of oncogenes, thereby increasing the malignancy potential of cancer cells.

EcDNA has been shown to amplify certain oncogenes that cause cancer. EcDNA refers to the DNA molecules that exist outside of the cell's chromosomal DNA. These ecDNA are known to segregate unevenly, leading to high copy numbers in cells and evident across many types of cancer (Turner 2017). But why do we care about ecDNA? It's been discovered that oncogene amplification, an increase in copy number of a specific gene that can give cancer a growth advantage, can occur in ecDNA. The ratio of ecDNA to chromosomes can give us an idea of whether a cell is cancerous or not, with higher ratios being more probable.

We will start by leveraging a tool that already exists, ecSeg. EcSeg is an extrachromosomal segmentation system that discerns chromosomes in fluorescence microscopy images using semantic segmentation (Rajkumar et al. 2019). We will use this system to identify chromosomes and then create our own system that uses the idea of instance segmentation for identifying and characterizing individual chromosomes in the images. EcSeg will be able to tell us what is and is not a chromosome, while our segmentation system will find centers of individual instances, allowing us to count how many are in each image.

This is made possible due to the FISH imaging used to obtain the images. FISH imaging, meaning fluorescence in situ hybridization, is a technique that allows us to identify and locate specific sequences of DNA in a cell (Rajkumar et al. 2019). To do this, probes of varying wavelengths are used to bind to specific sequences of DNA. When these probes attach to their targets in cells and absorb a certain frequency of light, they appear to be fluorescent, which can be easily detected when taking photos or when observed under a microscope. In general, FISH imaging is useful for scientists to help understand a cell's structure and gene expression, while in the context of ecDNA, it will allow us to locate and begin counting and sectioning chromosomes within cells.

DAPI staining solution is one of the most commonly used fluorescent probes in FISH imaging. While all fluorescent probes share the purpose of staining certain parts of a cell or se-

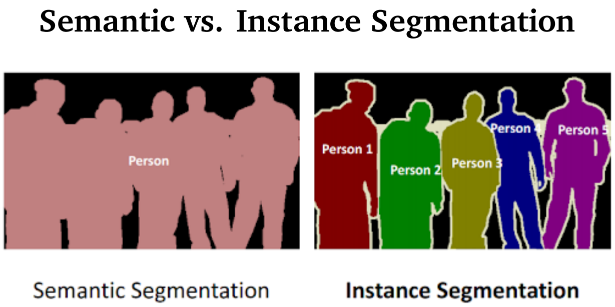

Figure 1: Semantic segmentation identifies the class of a person, where instance segmentation creates boundaries and identifies each person individually, allowing us to count how many are in the group (Team 2023). This is the concept we're using but instead, we're making it a semantic segmentation problem, where we find the centers of the chromosomes and individually identify them that way.

quences of DNA, DAPI only stains DNA making it useful for staining cell nuclei (B Tarnowski 1991). This allows us to look at and segment cells.

Individually identifying chromosomes and ecDNA in cancerous cells is an integral part of cellular research and developing a system to do this relieves lots of time from researchers. There exist lots of challenges with this task, including the overlapping of chromosomes, along with different sizes, shapes, and image qualities that will require unique techniques to be used. Similar research has been conducted to tackle this problem each with its own unique approaches. Our approach utilizes a tool we already have - ecSeg - to identify ecDNA and clusters of chromosomes that haven't been counted as individual chromosomes and further develop a model to do so.

1.1 Relevant Papers

One of the most relevant papers to our project is the paper of Faster R-CNN (Shaoqing Ren 2016). It highlights the challenges of region-based CNNs and the importance of efficient region proposal methods. The introduction of novel Region Proposal Networks (RPNs) is proposed as an effective solution. RPNs predict region proposals, unifying with Fast R-CNN object detection networks. The Faster R-CNN object detection system consists of a deep fully convolutional network for proposing regions and the Fast R-CNN detector for using these proposed regions. The RPN module guides the Fast R-CNN where to focus. It generates rectangular object proposals from an input image using a fully convolutional network. To generate region proposals, a small network slides over the convolutional feature map and produces lower dimensional features, which are then processed by fully connected layers for box regression and box classification.

Another relevant paper to our project is the paper on NuSeT, or Nuclear Segmentation Tool. NuSeT is a tool used to segment normal cells from their nucleus. It creates a binary segmentation of cell images, meaning that it can detect what is and is not a cell, but does not individually identify them (Yang et al. 2020). This paper is highly relevant, as the U-Net (Olaf Ronneberger 2015) model that we are using is derived from the work done with NuSeT.

Another relevant paper to our chromosome segmentation is about segmenting FISH images (Cao, Deng and Wang 2012). Our project contains a very similar idea in that we are both segmenting chromosomes in FISH images, but theirs focuses on the segmentation and classification of chromosomes in different classes, while our tool will be specialized to distinguish individual chromosomes. In addition, the methods used in this paper to segment, Fuzzy C-means Clustering, will not be our algorithm of choice (Cao, Deng and Wang 2012).

The research done on automatic nuclei segmentation for cancer detection, is work towards a very similar goal to ours: early and accurate cancer detection (Kaustav Nandy 2199a). Similarly, they use FISH images to attempt to segment cell nuclei. However, while the concepts are fairly similar, the primary difference is in the datasets. The dataset used to train their model is a mix of cancerous and non-cancerous tissue, while our model will primarily focus on cancerous cells in metaphase, with most of our data coming from a cell line. Also, the images in our dataset have been treated using various cytogenetics solutions, while the tissue used in the paper has not.

Finally, as we're using the tool ecSeg to identify the chromosomes in our images, the paper on ecSeg (Rajkumar et al. 2019) is highly relevant. It primarily describes the tool ecSeg and what its primary use cases are, as well as gets into specifics about its implementation and reliability.

1.2 Description of Relevant Data

Our training data comes from the (Turner 2017) paper, this is the data from ecSeg, which focuses on extrachromosomal oncogene amplification research. A majority of the cells were produced on a cell line, where cells were cultured and propagated in a petri dish, meaning that they were grown and replicated in a controlled environment. The rest of the data came from neurospheres and direct tissue samples. Tissue samples were obtained from the Moores Cancer Center Biorepository Tissue Shared Resource. Patient consent was obtained and all samples were de-identified (**?**). All obtained Institutional Review Board approval. DAPI was applied twice to metaphase cells in slides to visualize DNA and capture the images. We passed these cell images into a trained U-Net model from ecSeg to generate feature maps to train our model.

2 Methods

The core task of our model is to transition from semantic to instance segmentation, focusing on identifying and counting individual chromosomes by their centers. To aid in this task, we followed the Faster R-CNN architecture (Shaoqing Ren 2016) to create a model suitable for the task of segmentation. We settled on the architecture of Faster R-CNN due to its high accuracy, something that is crucial for medical imaging tasks, and its use of feature maps, something we had pre-trained through the work of our mentor.

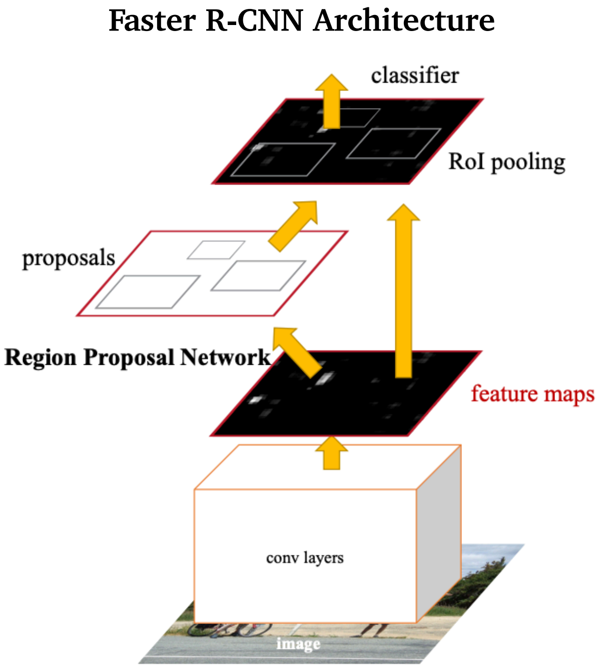

Figure 2: Faster R-CNN Structure that we implemented for our code. From (Shaoqing Ren 2016)

We created a model to identify valid chromosome centers from chromosomes in cell images. Our model takes cancer cell images as input, which are then processed into feature maps using a pre-trained convolutional neural network from ecSeg (Rajkumar et al. 2019). This convolutional neural network is built on an architecture called U-Net, and is specialized for biomedical imaging (Olaf Ronneberger 2015). It provides high dimensional representation of the input images crucial for the rest of our tasks, and for the purpose of our model, helps separate chromosomal and non-chromosomal regions within a cell image. Our model outputs the predicted centers of chromosomes.

Before we began implementing our model, we created a dataset for training our model. This dataset used the Turner et al. images described in the above section. We passed these images into the U-Net to create feature maps for the images. Next, we filtered these feature maps to only include chromosome components, as ecSeg is also able to identify cell nuclei and extrachromosomal DNA. Using these filtered feature maps, we generated bounding boxes around connected components using the scikit-image library (ski 2199b). From there, we moved on to finding valid chromosome centers in the images. To classify a pixel as a valid center, we assessed the Intersection Over Union (IoU) score between anchor boxes, drawn around each pixel, and the bounding boxes from our connected components. If an anchor box had an IoU score that exceeded 0.6, it would be considered. We used anchor boxes of various width and height ratios including various combinations of 15, 20, 25, and 30, which we derived after plotting the distribution of bounding box areas.

Chromosome Component Sizes

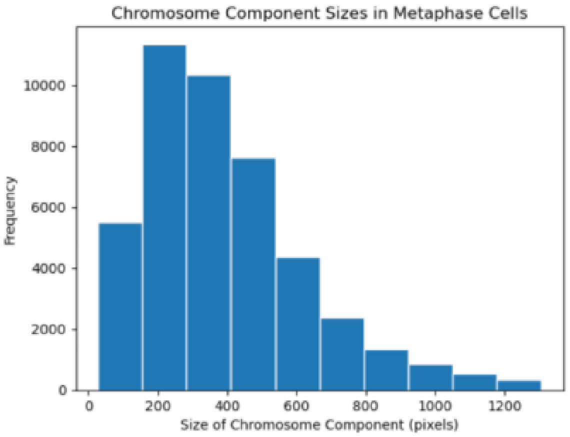

Figure 3: Chart shows a distribution of the chromosome component sizes in our cell images. While the data is skewed left, the component sizes are shown to be highly variable, making it difficult for us to find suitable anchor sizes.

The anchor sizes we chose had to be specific, as our anchors being too large or small would make it impossible for them to meet our IoU threshold. The presence of multiple valid centers around a single chromosome was expected due to variability in chromosome shapes and sizes, so our goal was to identify regions sufficient for individual chromosome segmentation rather than pinpointing exact, single pixel centers. We used the chromosome centers generated by our anchor boxes as the ground truth for the cell images.

To train our model, we first filtered our dataset to only include images with valid centers above a certain threshold. We arbitrarily chose 100, as we felt like 100 pixels as valid centers would be enough to accurately represent that chromosomes are within the data. From there, we further separated the data into 90% training, 5% validation, and 5% test. We randomly sampled 100 true centers and 100 false centers, and used the Adam optimization algorithm (Diederik Kingma 2017) to tune our model. The Adam optimization algorithm was chosen due to its ability to handle high-dimensional data and low overhead. Our model uses two convolutional layers, a 3x3 convolution layer with 512 output channels, and a 1x1 convolution layer with 64 out channels. These numbers were chosen with respect to the original Faster R-CNN paper. The U-Net model that is used to create our input convolves many times over the same data, allowing us to use these smaller kernel sizes, as the individual pixels all contain information about the surrounding pixels. To validate model performance, we used binary cross entropy loss (BCE) and our validation set. The loss function is specific for binary classification tasks, and is effective as it not only penalizes incorrect classifications but also punishes classifications with low confidence. As we saw that our validation loss was fairly similar to our training loss, that is, around .104 against .11, we concluded that our model was tuned fairly.

3 Results

Using a binary cross-entropy loss while training our Region Proposal Network (RPN), we observed the loss converging from approximately 0.18 to 0.11. We visualized the training loss over steps. This visualization demonstrates an initial sharp decline followed by a gradual convergence, suggesting effective model learning over iterations.

Training Loss per 500 Steps

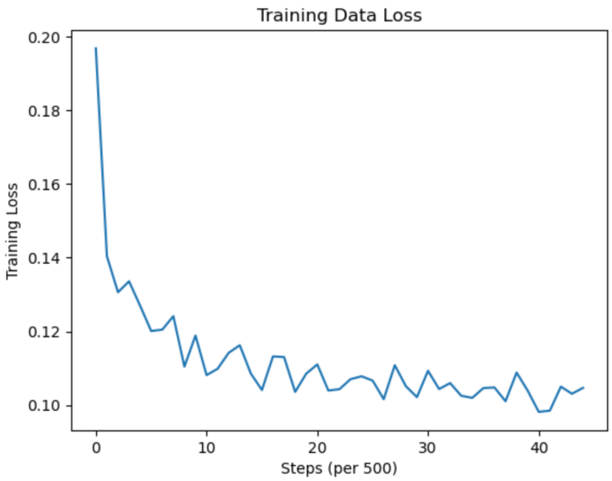

Figure 4: Chart shows training loss per 500 steps.

When assessing loss over epoch, the model exhibited a consistent decrease from 0.14 to 0.11 over the course of 10 epochs. The decline shows the model's progressive refinement in accurate chromosome center identification with each epoch.

Training Loss per Epoch

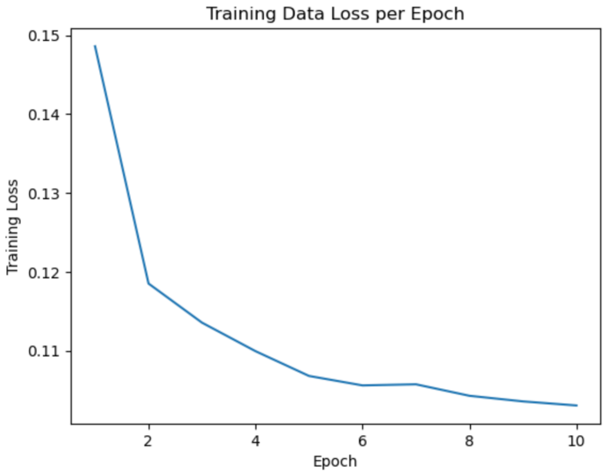

Figure 5: Chart shows the training loss per epoch.

On our test set, we were able to achieve a similar binary cross-entropy loss value of about 0.11, indicating that our model was not just fit well to the training data. Our model achieved a high true positive rate of 97% and a true negative rate of 95%, indicating a strong predictive performance for both chromosome center presence and absence. False positives and false negatives were contained at 4.9% and 2.5% respectively, which reinforces the model's precision.

Table 1: Performance Metrics

Metric	**Value**
Precision	0.952
Recall	0.975
Accuracy	0.963

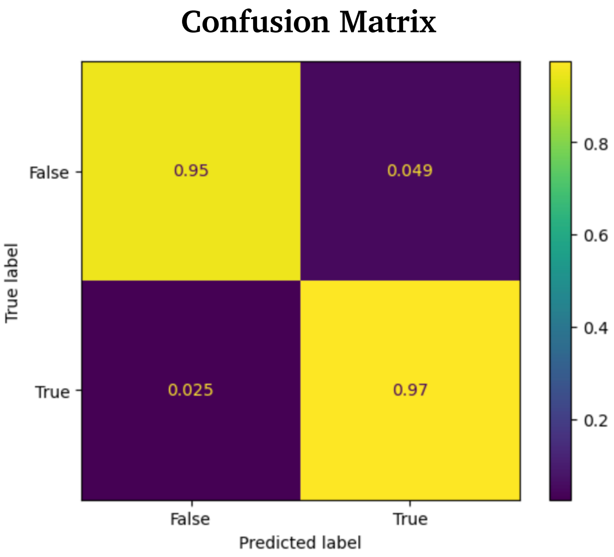

Figure 6: Chart shows the confusion matrix for predictions generated by the model

It was noted that the model's performance was less robust when identifying larger connected components, which were occasionally missed. The large bounding boxes created by overlapping or otherwise connected chromosomes resulted in low IoU scores. As such, our initial creation of the dataset was unable to identify valid centers within those bounding boxes, suggesting a need to adjust the anchor box sizing or the IoU threshold to accommodate a broader range of component sizes.

4 Conclusion

Though we achieved high accuracy rates, there are possibilities for future improvements. Hand labeling the images to show more accurate bounding boxes which represent our true labels, would allow for more accurate IoU scores, and thus more accurate and confident results. This would possibly capture more results, as some of our bounding boxes were too large to capture any IoU greater than 0.6. In our implementation, we've only tried using linear convolutional layers. While we believe that our model, as it stands, is accurate, there are possibilities that linear layers, or such few linear layers, are not able to capture chromosome centers as well as non-linear layers or a model with linear activation could. Enhancements to the model could include an extension where the output from the RPN is transformed into an instance segmentation output. This modification would allow for the precise segmentation of each pixel, distinguishing individual chromosomes with greater accuracy. This could provide detailed outlines of each chromosome, extracting more data and providing more aid to researchers. The count of chromosomes could also be an addition to the end of the model for statistical analysis.

Finally, we could evaluate the relationship between chromosome counts and oncogenes in the cell images. By identifying if a cell has more than the normal chromosome count of

46, we could quickly and efficiently identify cancerous cells. This ability would allow researchers and doctors to find the best cancer treatment for the patient according to their chromosome count and cancer diagnosis.

Though there is still more work to be done, this is a step in the right direction and provides yet another foundation for researchers to build on. The exploration of chromosomes in cancerous cells is extremely important and this model provides an easy and faster way to do so.

References

B Tarnowski, J Nicholson, F Spinale. 1991. "DAPI as a useful stain for nuclear quantitation." *PubMed*. [Link]

Cao, Hongbao, Hong-Wen Deng, and Yu-Ping Wang. 2012. "Segmentation of M-FISH Images for Improved Classification of Chromosomes With an Adaptive Fuzzy C-means Clustering Algorithm." *IEEE Transactions on Fuzzy Systems* 20 (1): 1–8. [Link]

ChemoMetec., "Oncology: Cancer cell counting analysis assays." [Link]

Diederik Kingma, Jimmy Lei Ba. 2017. "Adam: A Method for Stochastic Optimization." [Link]

Kaustav Nandy, Karen J. Meaburn Tom Misteli Stephen J. Lockett, Prabhakar R. Gudla., "Automatic Nuclei Segmentation And Spatial FISH Analysis For Cancer Detection." [Link]

Olaf Ronneberger, Thomas Brox, Philipp Fischer. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." [Link]

Rajkumar, Utkrisht, Kristen Turner, Jens Luebeck, Viraj Deshpande, Manmohan Chandraker, Paul Mischel, and Vineet Bafna. 2019. "EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA." *iScience* 21: 428–435. [Link]

Shaoqing Ren, Ross Girshick Jian Sun, Kaiming He. 2016. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." [Link]

skimage.measure. [Link]

Team, Folio3 AI Editorial. 2023. "Semantic Segmentation vs. Instance Segmentation: Explained." *Folio3AI Blog*. [Link]

Turner, Deshpande V. Beyter D. et al., K. 2017. "Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity." [Link]

Yang, Linfeng, Rajarshi P. Ghosh, J. Matthew Franklin, Simon Chen, Chenyu You, Raja R. Narayan, Marc L. Melcher, and Jan T. Liphardt. 2020. "NuSeT: A deep learning tool for reliably separating and analyzing crowded cells." *PLOS Computational Biology* 16 (9): 1–20. [Link]