SMU Data Science Review

Volume 7 | Number 3 Article 7

2023

# Deep Learning Image Analysis to Isolate and Characterize Different Stages of S-phase in Human Cells

Kevin A. Boyd
*SMU*, kevinboyd76@gmail.com

Rudranil Mitra
*SMU*, rmitra@smu.edu

John Santerre
*SMU*, jsanterre@smu.edu

Christopher L. Sansam
*University of Oklahoma & Oklahoma Medical Research Foundation*, Chris-Sansam@omrf.org

# Deep Learning Image Analysis to Isolate and Characterize Different Stages of S-phase in Human Cells

Kevin A. Boyd[1], Rudranil Mitra[1], John Santerre[1], Christopher L. Sansam[23]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Oklahoma Medical Research Foundation, 825 NE 13th St.,
Oklahoma City, OK 73104 USA
[3]University of Oklahoma, 660 Parrington Oval, Norman, OK 73019
{kevinboyd, rmitra, jsanterre}@smu.edu
Chris-Sansam@omrf.org

**Abstract.** This research used deep learning for image analysis by isolating and characterizing distinct DNA replication patterns in human cells. By leveraging high-resolution microscopy images of multiple cells stained with 5-Ethynyl-2′-deoxyuridine (EdU), a replication marker, this analysis utilized Convolutional Neural Networks (CNNs) to perform image segmentation and to provide robust and reliable classification results. First multiple cells in a field of focus were identified using a pretrained CNN called Cellpose. After identifying the location of each cell in the image a python script was created to crop out each cell into individual .tif files. After careful annotation, a CNN was created from scratch using the TensorFlow Keras package and trained on those images to categorize them into five distinct replication patterns. Using a holdout test set our model was able to achieve an accuracy of 86.5%. This analysis method for segmentation and classification enhances the efficiency and reproducibility of DNA replication analysis, allowing for high-throughput processing and analysis of replication foci. This research can enhance image analysis in cell biology by providing a time-efficient and accurate tool to investigate replication dynamics, advance cancer research, and contribute to scientific discovery in various domains.

## 1 Introduction

For cells to divide properly they must first copy their entire genome so that each of its daughter cells inherits a complete copy of the genome. This process is called DNA replication, and it plays a crucial role in maintaining genomic integrity. Most metazoan cells replicate their genome following a spatiotemporal pattern that becomes apparent when replication sites in S phase nuclei are labeled with thymidine analogs like BrdU or EdU. Quantifying cells with spatiotemporal replication patterns characteristic of early, mid, or late S phase has become a widely used method among

scientists to assess and understand the progression of cells through the S phase of the cell cycle. Identifying and characterizing DNA replication patterns in human cells is essential for understanding the mechanisms underlying cell division and its dysregulation in various diseases such as cancer. Current methods for analyzing replication foci in images of cells rely heavily on manual assessment, leading to subjectivity, time inefficiency, and limited sample sizes. To address these limitations, this research focuses on developing a deep learning image analysis pipeline that can automate the characterization and identification of different patterns during S-phase. S-phase takes place during interphase, before mitosis or meiosis, and is the time that DNA replication occurs.

During S-phase of the cell cycle, the genome must be accurately copied to make sure proper cellular function is maintained. If the inherited genetic information is not copied correctly and contains errors, it can lead to various diseases. One of the key aspects of DNA replication is the initiation of replication forks, which are responsible for unwinding and copying the DNA strands. Replication forks initiate at specific sites on the DNA where the replication machinery that copies the DNA takes place. These dynamic structures are formed by the unwinding of the DNA double helix and the assembly of proteins and enzymes necessary for DNA synthesis. The initiation of replication forks is a highly regulated process involving multiple factors. The regions on chromosomes where these forks can begin replication are known as replication origins (ORI). In the G1 phase of the cell cycle, prior to S-phase where DNA replication occurs, these replication origins are "licensed" or marked for potential initiation (Chong et al. 1995). This licensing process ensures each ORI is used only once during each cell cycle, preventing overactive or incomplete DNA replication. The mechanisms that control the progression of S-phase, like replication fork initiation, and the factors that limit it have implications both developmentally and in disease progression.

Replication foci are specific subnuclear structures that are observed within the nucleus of a cell during S phase. These foci are formed by a combination of various proteins and enzymes at active sites of replication forks. Replication foci represent the localized concentration of replication-associated factors and newly synthesized DNA at clusters of replication forks in the nucleus. They are considered indicators of ongoing DNA replication in a cell. The number and distribution of replication foci within a cell's nucleus can provide insights into the replication status of the genome and can be used to compare specific treatments that affect DNA replication.

The importance of investigating this topic lies in the need for a robust and efficient tool to improve the analysis of replication foci in human cells. Such a tool would enhance the accuracy and reproducibility of DNA replication studies and enable researchers to process and analyze images time-efficiently. By automating the identification of replication patterns, the subsequent pipeline would contribute to the advancement of cancer research and other fields exploring DNA replication dynamics. It is important to note that these automated processes should be maintained and checked often to ensure the validity of the results. It is possible for slight experimental variation to affect the expected accuracy.

Previous studies have identified distinct patterns associated with different stages of DNA replication using specific markers of replication foci such as 5-ethynyl-2′-deoxyuridine (EdU), 5-bromo-2′-deoxyuridine (BrdU), and the replication

processivity factor PCNA. Specifically, we will be focusing on the use of EdU, which is a Thymidine analog. When applied to cells in S-phase, this analog is incorporated instead of Thymidine into newly replicated DNA. Once EdU becomes incorporated into genomic DNA, it can be efficiently labeled by covalently attaching fluorescent azides through a Cu(I)-catalyzed [3 + 2] cycloaddition reaction, commonly referred to as "click" chemistry (Salic et al. 2008). Sites of replication in the nucleus fluorescently labeled with EdU can be observed using fluorescence microscopy. Replication sites have distinct spatial patterns associated with different times in S-phase (O'Keefe, 1992). In the early stages of S-phase, genomic DNA distributed across the nuclear interior undergoes replication, followed by replication of DNA located at the nuclear periphery and surrounding nucleoli in mid-S phase, with DNA within large heterochromatin patches being replicated towards the end of S phase. Hence, a cell's specific stage in the S phase can be identified by categorizing it based on one of three to five distinct spatiotemporal replication patterns. Quantifying these replication factors has evolved into a standard approach for evaluating S phase progression. However, traditionally, scientists have relied on manual categorization of these patterns, which is labor-intensive, subjective, and susceptible to variability. Furthermore, the manual assessment often involves analyzing only a subset of cells within an image, potentially leading to biased conclusions. There is a clear gap in the field for a tool that can increase both the throughput and accuracy of DNA replication analysis while reducing the inherent biases associated with manual categorization.

This research aims to bridge this gap by leveraging deep learning techniques to develop an automated image analysis pipeline. By training a deep learning model on a large dataset of human cell images stained with replication markers, the pipeline will learn to identify and categorize replication patterns with high precision and efficiency. This approach will not only reduce the time and effort required for replication analysis but also enable researchers to analyze larger sample sizes, thereby improving the statistical power of their experiments.

Image segmentation is a kind of computer vision that partitions an image into multiple segments or regions. Each of these segments represents an area of significance or a unique object. The goal of image segmentation is to divide an image into meaningful and semantically consistent regions, making it easier for computers to analyze and understand the contents of the image. The research will discuss architectural variations, modifications, and strengths in handling image segmentation. In image segmentation, each pixel or group of pixels in an image is assigned a label or a unique identifier based on its visual attributes, such as color, intensity, texture, or spatial location. The segmentation process aims to separate objects or regions of interest from the background or separate different objects from each other. There are different approaches to image segmentation, including thresholding, region-based segmentation, edge detection, clustering, and deep learning-based methods. For this research, our focus will be on semantic segmentation.

Semantic segmentation is another computer vision method that labels each pixel in an image with a class or category, aiming to assign a meaningful semantic label to every individual pixel. Unlike other forms of image segmentation that only differentiate regions or boundaries, semantic segmentation provides a more detailed understanding of the image by associating semantics with each pixel. In semantic segmentation, the goal is to partition an image into multiple regions and assign each

pixel a label that represents the category or class it belongs to. The resulting segmented image provides a pixel-wise representation of the scene, highlighting the specific objects or regions of interest.

Semantic segmentation is typically performed using deep learning techniques, specifically convolutional neural networks (CNNs). CNNs are trained on large, annotated datasets, where pixel-level labels are provided for training images. The network learns to recognize and classify different visual patterns and features in the images, allowing it to segment new, unseen images accurately. These networks can generalize well to new, unseen data and provide accurate and real-time segmentation results for various computer vision applications. Deep learning-based image segmentation has significantly advanced the field, enabling breakthroughs in medical imaging, autonomous vehicles, robotics, and many other areas where accurate scene understanding is crucial.

The architecture commonly used for semantic segmentation is the Fully Convolutional Network (FCN). FCNs are used to replace the fully connected layers in a traditional CNN with convolutional layers, enabling the network to produce spatially dense predictions for each pixel. Semantic segmentation has numerous applications in various domains. In medical imaging, it aids in segmenting organs, tumors, or lesions, enabling accurate diagnosis and treatment planning.

## 2 Literature Review

High-resolution microscopy and deep learning algorithms have both shown exciting potential in advancing our understanding of the cell cycle. Medical researchers have leveraged these approaches to classify different cell cycle phases, analyze chromosome replication, and even predict disease progression. Developing and improving deep learning models and refining image analysis techniques, valuable insights can be gained regarding different cellular processes and disease mechanisms. The literature presented here gives some relevant information about visualizing replication foci, explores the background of deep learning image analysis in the field of medicine, and shows some examples of relevant studies that have demonstrated its effectiveness in cell cycle research.

### 2.1 DNA Staining Techniques

An efficient way to label replication foci is with BrdU, a thymidine analog. This technique has been widely used in cell cycle studies to label replication foci. Vogel et al. (1989) employed this technique to analyze human chromosome replication in lymphocyte and amniotic cells. This study highlighted the sensitivity of the technique in detecting replication foci, even in regions where only a very small number of nucleotides were replaced by BrdU. This demonstrates the effectiveness of BrdU labeling for studying replication foci. EdU is another thymidine analog used to label active sites of DNA replication by incorporating into the newly synthesized DNA. While both EdU and BrdU have similar purposes, there are slight differences in the chemical structure and detection method. Another group of researchers used a similar

technique to study the length of S-phase by using a dual EdU-BrdU pulse-labeling technique with incremental thymidine chases (Bialic et al., 2022). With this method, they measured the length of unperturbed S-phase without genome engineering or cell cycle synchronizing the cells. Therefore, S phase progression could be tracked in unmodified adherent or suspended cultured cells or even cells in animals. These approaches reduce the risk of off target effects or artifacts that could undermine the specific goal of any study and are why they are commonly used in medical research today. Another example of the EdU-pulse labeling technique being used to study DNA replication is when a group of researchers found that CDKs (cyclin-dependent kinases) play a crucial role in regulating the length of S phase and replication initiation (Sansam et al., 2015). They investigated the role of CDKs in this process through TICRR/TRESLIN phosphorylation using high resolution microscopy, BrdU/EdU-pulse double labeling, and other techniques. They found that phosphorylated TICRR limits S-phase progression, and the overexpression of a mutant form of TICRR with phosphomimetic mutations resulted in an enhanced replication initiation and a shorter S-phase. This study provides insights into the regulatory mechanisms governing S-phase progression and replication timing using these established labeling techniques.

**2.2 Image Segmentation**

Image segmentation has a long history and has evolved through various techniques and methods over the years. The earliest (1960-1970) image segmentation methods were based on simple thresholding and region-growing techniques. Researchers used basic intensity or color thresholding to separate objects from the background. In the 1980s, edge-based segmentation techniques gained popularity. These methods aimed to identify boundaries or edges between different regions in an image using gradient-based operators or filters. Active contour models, also known as snakes, were introduced in the 1990s. These methods used deformable curves or contours to detect object boundaries by minimizing an energy function. In the 1990s, region-based segmentation techniques gained popularity. These methods grouped pixels based on their similarity in color, texture, or other feature spaces. Graph-based methods emerged as powerful tools for image segmentation in the early 2000s. These methods represented the image as a graph, where pixels were nodes, and edges represented the relationships between neighboring pixels. Machine learning techniques, particularly clustering algorithms like k-means, were applied to image segmentation tasks in the 2000s. Additionally, support vector machines (SVMs) and random forests were utilized to classify pixels into different regions based on feature representations. In the 2010s, the advent of deep learning revolutionized image segmentation. Convolutional neural networks (CNNs) became the dominant approach, enabling accurate image segmentation. Fully Convolutional Networks and architectures like U-Net, DeepLab, and Mask R-CNN propelled the field of image segmentation to new heights, achieving state-of-the-art performance in various applications.

**2.3 Deep Learning Models in Medicine**

Deep learning models have also been applied to medical image analysis studies to predict disease progression. Voets et al. (2019) attempted to replicate the results of an earlier study that developed a deep learning algorithm for detecting diabetic retinopathy. The original study did not use publicly available data for training

and source code was not available. To deal with this the researchers re-implemented the main method using publicly available data sets. In contrast to the original study which had multiple grades per image, this study had only one grade per image. Unfortunately, Voets replicated algorithm achieved lower performance compared to the original study. The researchers suggested these discrepancies may be due to the data having only a single grade per image or differences in the training data they used. This is an example of the challenges of reproducibility in deep learning results and the importance of validation and replication studies, particularly in medical image analysis.

Modern deep learning methods have been useful in performing image segmentation. A survey by Minaee et al. (2022), explains how deep learning methods can be used to perform image segmentation. The paper begins by introducing the historical background of image segmentation and its evolution over the years. It highlights the limitations of traditional segmentation methods and the need for more robust and accurate approaches, leading to the emergence of deep learning-based techniques. The paper provides a comprehensive overview of deep learning architectures commonly used in image segmentation, such as Fully Convolutional Networks (FCNs), U-Net, DeepLab, and Mask R-CNN. It explains the principles behind each architecture, including encoder-decoder structures, dilated convolutions, and feature fusion mechanisms. The survey delves into the role of annotated datasets and transfer learning in training deep learning models for image segmentation. Moreover, the paper discusses the trade-offs between accuracy and computational efficiency in deep learning-based segmentation models, considering the resource constraints in real-time applications. The paper states that fully convolutional networks and encoder-decoder networks were initially developed for medical & biomedical image segmentation and that residual networks (ResNet) can be used as feature extractors in images. Semantic segmentation using deep learning techniques have been used in the field of medical imaging. Araújo, F. H. D. et al. (2019) explores the use of deep learning for cell image segmentation and ranking tasks. Cell image segmentation is the process of accurately identifying and delineating individual cells within an image, which is crucial for various biological and medical research applications. Deep learning techniques like convolutional neural networks (CNNs) have shown remarkable success in automating this process, enabling efficient and accurate cell segmentation even from complex and noisy images. Similar deep learning techniques have also been used to identify Covid-19 lung infections as addressed in Chen, Y et al. (2022). Asgari Taghanaki, S et al. (2021), provides a comprehensive overview of deep learning techniques for semantic segmentation in medical image domains. The paper then delves into the principles and architectures of deep learning models commonly used for semantic segmentation, paying special attention to convolutional neural networks. It presents a thorough overview of popular CNN architectures, such as U-Net, DeepLab, and Mask R-CNN, and explains their design characteristics and strengths. The study concludes with how semantic segmentation can be used in medical image domains. To add further support, Hesamian, M. H. et al. (2019) extensively discusses deep learning techniques and architectures for medical image segmentation. The study begins by

highlighting the critical role of medical image segmentation in various clinical applications, such as disease diagnosis, treatment planning, and monitoring. It emphasizes the significance of accurate and precise segmentation to aid medical professionals in making informed decisions. The paper provides an in-depth explanation of multiple deep learning techniques with examples. It focused on convolutional neural networks (CNNs), which have become some of the most effective models in addressing medical image segmentation. It discusses the architectural components of CNNs and the benefits of using deep learning models over traditional approaches. A sizable portion of the review is dedicated to discussing the achievements of deep learning-based medical image segmentation. The authors present case studies and examples of successful applications, including organ and tumor segmentation, brain lesion detection, and cardiac image analysis. They highlight the improved accuracy and efficiency of deep learning methods in these tasks compared to conventional methods. Our study intends to use residual networks to improve the accuracy of the deep learning model. Cheng, J et al. (2022), proposes residual networks and effective deep learning architecture that address challenges related to variations in appearance and limited training data in medical image segmentation. The authors of the paper emphasize the significance of accurate medical image analysis in diagnosis and treatment planning. They highlight the challenges posed by the complexity and variability of medical images, motivating the need for advanced techniques to improve classification and segmentation accuracy. The paper introduces the ResGANet architecture, which combines the residual connections and group attention mechanisms. Residual connections enable the network to learn residual mappings, facilitating the training of deeper architectures and reducing vanishing gradient issues. Group attention mechanisms enhance the model's ability to capture relevant features by selectively attending to informative regions of the input.

Another example of using machine learning on image analysis comes is when Jaeger et al. (2010) developed an algorithm for classifying cell cycle phases using 3D spinning disk confocal microscopy images. Their method leveraged a 3D k-means approach for image segmentation and was able to extract the shape and curvature features associated with different cell cycle phases. After training a support vector machine (SVM) classifier they achieved high recognition rates for both the chromocenter and PCNA channels, demonstrating the reliability of their automated algorithm. A study using deep learning to predict disease progression was conducted by Yoo et al. (2019). They applied deep learning to predict the risk of conversion to multiple sclerosis (MS) from clinically isolated syndrome. By combining deep learning with user-defined clinical and MRI features, they improved the accuracy of MS conversion prediction compared to traditional imaging biomarkers. This study highlights the potential of deep learning in extracting latent lesion patterns from MR images for enhanced disease progression prediction.

**2.4 Cellpose Image Segmentation**

Stringer et al. (2021) introduced Cellpose, a deep learning-based segmentation method for cellular analysis. Cellpose allows precise segmentation of cell bodies, membranes, and nuclei in high resolution microscopy images. This approach eliminates

the need for model retraining with hundreds or thousands of images, allowing for minimal parameter adjustments. This makes it suitable for various image types without needing much training data. Cellpose provides an efficient and accurate initial segmentation step in cell cycle studies 'out of the box' or with very little training, enabling researchers to analyze individual cells. Saad et al. (2023) used Cellpose and Fiji to create a novel automated protocol for ice crystal segmentation analysis. They were able to improve the throughput and accuracy of their measurements using this automated approach. Yang et al., 2023 tested Cellpose on fluorescent images of HeLa cells. After determining the algorithm was performing well on their data, they developed a workflow that increased their throughput without lowering cell identification accuracy. They later tested the workflow on images of fluorescent labelled cells exposed to polystyrene nanoparticles. This allowed them to investigate the connection between the size of each cell and how many nanoparticles they could absorb. This is a study that was greatly helped by the increased throughput. Another group of researchers decided to investigate deep learning architectures for nuclear image segmentation by comparing U-Net, U-Net ResNet, Cellpose, Mask R-CNN, KG instance segmentation, iterative h-min based water shedding, and attribute relational graphs (Kromp et al., 2021). Their results showed that both Cellpose and instance aware segmentation architectures outperformed the U-Net architectures and conventional methods. This research demonstrates the different methods that use annotated images to train instance aware segmentation architectures that have the ability to accurately segment fluorescent nuclear images.

**2.5 Challenges Using Deep Learning**

Some familiar challenges using deep learning for image segmentation include limited annotated data because deep learning models often need a large amount of annotated data to train, and it can be difficult to obtain high-quality annotations. A sizable portion of the review by Asgari Taghanaki, S et al. (2021), is dedicated to exploring the challenges and datasets related to deep semantic segmentation, both in the context of natural images and medical images. It highlights the importance of annotated datasets and the complexities involved in acquiring high-quality annotations for training deep learning models; Class Imbalance and Multi-Modal Fusion in that some medical image segmentation tasks involve imbalanced classes, where certain structures or pathologies are rare compared to the background or normal regions. Dealing with class imbalance is crucial to prevent the model from biasing towards the majority class. Medical imaging often involves multiple modalities, and fusion of information from different modalities has gained attention. Techniques like early fusion, late fusion, and attention-based fusion are used to combine multi-modal information for more accurate and robust segmentation. A survey by Hesamian, M. et al (2019) addresses the challenges faced in the domain of medical image segmentation using deep learning. It explores issues related to limited annotated data, class imbalance, handling multi-modality images, and ensuring robustness and generalization across different patient cohorts. The paper presents insights into various strategies to address these challenges, including data augmentation techniques, transfer learning, and domain adaptation. It also discusses the significance of model

interpretability in medical applications, as well as the need for uncertainty estimation to assess the reliability of segmentation results; Interpretability in that deep learning models, especially complex architectures, can lack interpretability, which is crucial in medical imaging applications. Understanding the reasoning behind model predictions is essential for clinical acceptance and trust. Cheng, J. et al (2022) emphasizes the interpretability of ResGANet, enabling clinicians to understand the reasoning behind the network's decisions. This interpretability aspect is crucial for building trust in the model's outputs in clinical applications and handling Small Structures in that some medical structures, such as tiny lesions or cellular structures, pose difficulties in segmentation due to their size and limited contrast in the image. Addressing this challenge is essential for accurate and reliable segmentation. Tajbakhsh, N. et al (2020), dedicates their study to examining methods to address the challenges of imperfect medical image datasets. Techniques such as data augmentation, domain adaptation, transfer learning, and semi-supervised learning are explored, which help to mitigate the effects of limited and noisy training data.

### 2.6 Summary

In conclusion, deep learning-based image analysis techniques offer investigators the possibility for significant advancements in many different areas of research including cell cycle studies. From classifying cell cycle phases to analyzing replication patterns and predicting disease progression, these approaches provide automated and accurate analysis, reducing manual efforts and enabling comprehensive investigations. Continued research and development in this area will enhance our understanding of cellular processes and contribute to the advancement of disease diagnosis and treatment.

## 3 Methods

### 3.1 Data Acquisition

The data for this research will be obtained from Dr. Chris Sansam at the Oklahoma Medical Research Foundation. Dr. Sansam has graciously provided a collection of high-resolution microscopy images of human cells (HCT116) stained with the replication marker 5-Ethynyl-2′-deoxyuridine (EdU) at different times. These images capture the dynamic process of DNA replication in the cell's nucleus at various stages.

### 3.2 Image Segmentation

The first step in the analysis pipeline involves segmenting the images to differentiate individual cells. Our initial input data were cell images that had two channels with between 4 and 6 layers each. Using ImageJ, each image was transformed to show a maximum intensity projection. Only one channel is needed for this process, so the channels were split, and we only keep the one. After this all the images were

concatenated so we could use a single file to process as many images as we wanted. For the segmentation process, we utilized a program called Cellpose, which has been demonstrated to be effective in segmenting cells in fluorescence microscopy images. Using the Cellpose convolutional neural network we automatically identified and drew masks around individual cells within the images. The Skimage Regionprops package in python we developed a script to automatically crop out each of the cells in the image based on the masks created by Cellpose. This allowed us to efficiently isolate and crop out all of the cells in our images in order to use them downstream for training our classification model. Each of the output files were saved as individual .tif files with a sequential naming order based on the input name. By segmenting the images this way we can isolate and analyze many more cells than we could using other methods, enabling more accurate characterization of replication patterns.

**3.3 Annotation and Training Data Preparation:**

After the image segmentation, we need to proceed with annotating the segmented cells. Manual annotations were performed to classify the cells into 1 of 5 states. These annotations were then checked by Dr. Courtney Sansam, an experienced researcher in the DNA replication field. The quality of the annotations is particularly important to classify unseen data in the future.

**3.4 Deep Learning Model Development and Training**

To develop a robust and accurate deep learning model, we employed a convolutional neural network (CNN) architecture using TensorFlow and Keras. The segmented cell images were classified into one of the five replication patterns. The training dataset consisted of over 400 annotated cell images, encompassing a diverse range of replication patterns. Data augmentation techniques, including rotation, scaling, and flipping, were applied to increase the diversity and generalizability of the training dataset. The annotated dataset was split into training, validation, and testing sets to evaluate the model's performance accurately. The CNN we used was trained using 798 images with 171 validation images and 171 final test images. The CNN model uses 5 convolution layers with max pooling. Convolutions are needed for learning local features in an image. Small filters called kernels are applied to local regions of the image, allowing the model to capture patterns like edges, corners and textures present in the image. Multiple layers of convolutional are needed for the model. The initial layers capture low-level features such as edges and then the other layers capture complex features and object representations. The model uses max pooling to reduce feature map dimensionality and overfitting. The deep learning model was trained using the well-defined loss function categorical cross-entropy and optimized through the Adam gradient descent algorithm. The model's hyperparameters, such as learning rate, batch size, and network architecture, were tuned through iterative experimentation to achieve optimal performance. During the training process, we regularly monitored the model's performance on the validation set to avoid overfitting. The trained model was then evaluated on the independent testing set to obtain an unbiased estimate of its performance and generalizability.

# 4 Results

## 4.1 Image Segmentation and Annotation

The initial segmentation process was performed on 87 high-resolution fluorescence microscopy images showing groups of cells stained with EdU to produce over 2000 individual images of cells. An example of the output from Cellpose to identify and draw masks around each of the cells in the is shown in the image below.

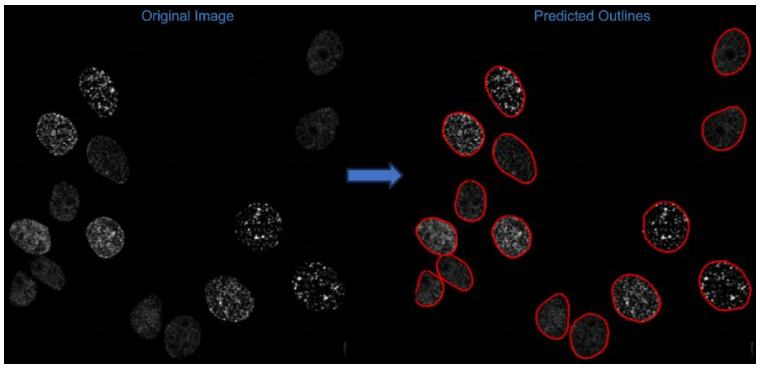


Figure 1: Identification of individual cells with Cellpose

Using those masks, we identified the coordinates of each cell to crop and output as an individual .tif file. Using some of images output from the previous step, 285 images of cells were annotated by eye and augmented creating a final dataset of 1140 images. The annotations classified each of the cells into one of five stages with one being the earliest time point and 5 being the latest. Examples of the cropped images and their associated annotations are shown in the figure below. Each of the stages are defined by specific characteristics, but there can be some ambiguity between some of the stages. These images were imported in greyscale with a single channel and field of focus showing a maximum intensity projection.

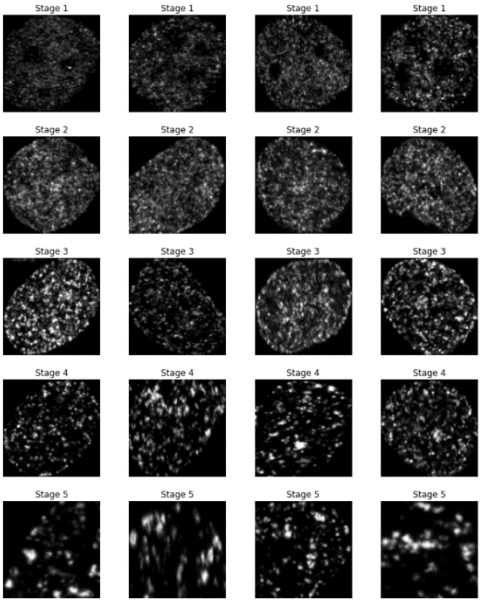


Figure 2: Example of cropped and annotated cells

### 4.2 Classification

We found that a convolutional neural network using categorical cross entropy as its loss function and accuracy as our performance metric was most useful to our task and would be most interpretable. While training this model we achieved an accuracy of 99% on the training set with a loss of 0.0643, but when performing on our unseen test dataset our accuracy dropped to 86.5% with a loss of 0.470.

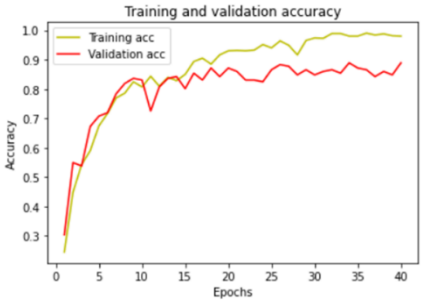


Figure 3: Training and validation accuracy

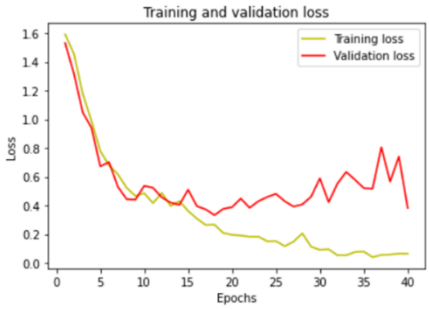


Figure 4: Training and validation loss

When trying to predict the first stage we were able to classify it correctly 56 times out of 62 total times, giving that specific stage a ~90% prediction accuracy. The accuracies on the test data for the distinct stages are seen in the table below. There is a

trend that the least represented stages have lower accuracy which is specifically seen in stages 4 and 5.

| Cell Stage | Test Accuracy |
|---|---|
| Stage 1 | 90.32% |
| Stage 2 | 94.44% |
| Stage 3 | 88.24% |
| Stage 4 | 83.33% |
| Stage 5 | 80.95% |

Table 1: Table of individual stage accuracies

Predicting any of the 5 distinct stages we can see that misclassification happens most often in the adjacent stages, although some stages were easier to predict than others. There is also an imbalance in the number of images representing the last two stages, but the predictions remain accurate as seen in the heat plot below.

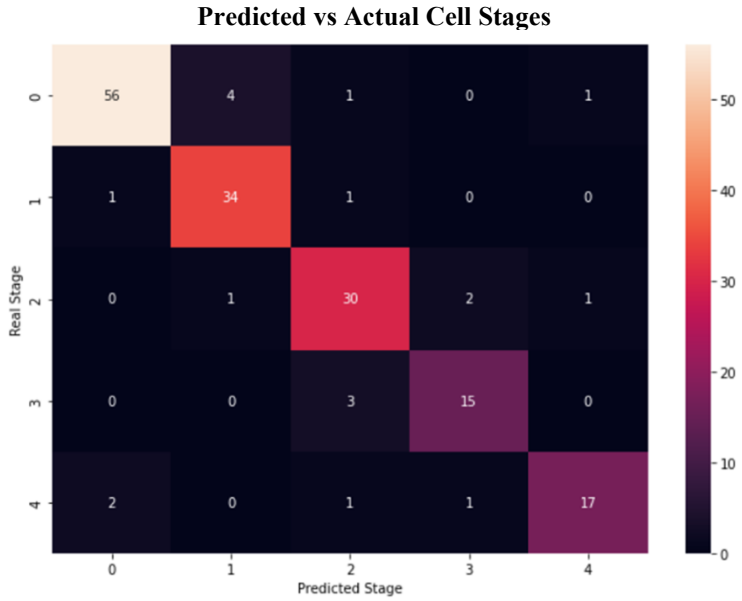


Figure 5: Heat plot of classification of each stage

The images below are examples of each of the states that had a correct prediction. Each of the 5 patterns were predicted accurately even though the first stage was most common in our dataset. While we do see an imbalance in the data, we still have >80% accuracy for all the stages.

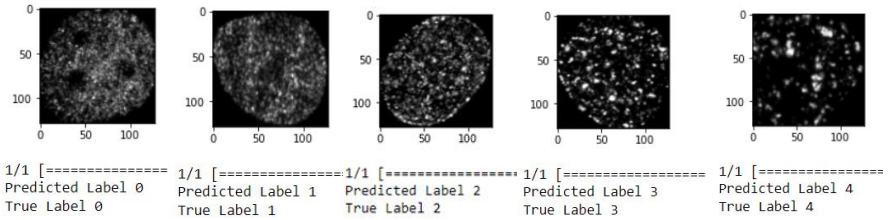


Figure 6: Example of predictions for each stage

## 5 Discussion

Automating the segmentation process was a crucial aspect of this project in order to obtain enough data to train our model. While it is possible to use the base nuclear segmentation model trained by Cellpose, we trained the base model on 4-5 images to ensure the masks were as accurate as possible. This process does not require many images for accurate masks to be drawn. This was the first step before we could crop the images and separate them into individual cells. The script to crop the images in python used the location of each mask that was drawn to define the boundaries of each cell. The masks were extremely important to ensure the entire cell was cropped without cutting off the edges or including multiple cells. Occasionally when cells were layered on top of each other they were cropped out together, but this did not happen very often. After we removed any images with multiple cells, we could annotate them.

The predictions from our model appear to be accurate, which can be a challenging task when observing biological images. There can be a lot of variation between cells in the same stage with only a few key differences differentiating the stages. These key features need to be picked up by the model during the training process for accurate predictions, but we still need to take care not to overfit. The extremely high accuracy of the training dataset suggests that we should be careful of overfitting in our case. Something that can always be considered when looking at a problem like this is getting even more training data to train the model so it can pick up on as many differences as possible that are found naturally in each of the stages.

Now that we have shown this model's ability to perform accurate predictions on cells in various stages of S-phase, it could be applied to biological problems. Knowing how long these cells have been stained before they were imaged and employing sequential, double labeling with EdU and BrdU at multiple time points can allow us to calculate the total length of S-phase. We could then compare cells treated with different drugs or genetic mutations to see if we could alter the time it takes to progress through S-phase. By increasing the number of cells, we can study will allow for a more granular perspective to understand the length of each individual stage in S-phase.

Future possibilities include using unsupervised techniques to see how many cell stages are found. These could possibly produce three stages; an early, mid, and late stages based on the misclassifications seen in the model above. It is also possible that an unsupervised technique could find more than the five we were trying to identify and

that is a possible reason for some of the misclassifications we are seeing. It is not surprising that we are seeing some overlap between the stages considering many of the cells could be transitioning from one stage to another at the time of the image being taken. But this also leads to the exciting possibility that an entirely new stage could be identified if there is a consistent classification using an unsupervised approach.

This model was trained and tested using HCT116 cells. It is possible to also apply this model to other cell types with minor adjustments. The initial segmentation process would be the same. It would be possible to use base nuclear segmentation model from Cellpose or train on a couple of images beforehand. The cropping step uses the masks drawn by Cellpose so this process would not need to be changed. Finally, when using the CNN it is important to check the output to make sure it looks as expected. Checking the accuracy with train and test splits would be most appropriate, but this model can likely be used on other cell lines with the same DNA replication marker. While this is an exciting possibility it is not the only way to apply this model in a slightly different manner. Using transfer learning we could apply our model to other nuclear stains to answer different questions posed by researchers. Specific proteins of interest could be identified at different time points and our model could be used to classify them into distinct groups.

Another exciting possibility for future work would be to make this into a pipeline using Nextflow or Snakemake to completely automate the processes. It would be important to output model performance each time, but it would be possible to start with only an input image with multiple cells in the frame and output all the cells with a predicted cell stage label. These could be checked by eye, or a previously annotated dataset could be used to ensure the model accurately characterizes the cells.

A final important aspect to consider is making sure to put this deep learning model into context. Often deep learning models tend to be overfitted to the data they are trained on. We want to be sure that readers know that every experiment and recording technique could be subject to aberrations that may make this model less accurate for their data ‘out of the box’. Also, all deep learning models will need to be maintained and updated as imaging techniques change over time.

## 6 Conclusion

In conclusion, this research harnessed deep learning convolutional neural networks to efficiently isolate and categorize distinct DNA replication patterns in human cells. The study addressed the limitation of current manual assessment by using an automated method for the segmentation and classification process. This robust tool can also be used for many different applications in medical image analysis as it can be applied to different cell lines, fluorescent stains, and even bright field images of cells with minimal training.

However, it is important to keep in mind that the model's performance may vary with different experimental designs and imaging techniques, and regular maintenance and updates will be necessary to ensure accurate segmentation and classification. This research offers a valuable contribution to the field of cell biology by providing a powerful tool for image segmentation and replication dynamics

investigation with the potential to accelerate the pace of scientific discovery in various scientific domains, particularly cancer research.

## Code Availability

A general workflow, code, examples of output, and any future updates are available at https://github.com/kevinboyd76/Automated_S-phase_Image_Classification.

## References


1. Chong JP, Mahbubani HM, Khoo CY, Blow JJ. 1995. Purification of an MCM-containing complex as a component of the DNA replication licensing system. Nature 375: 418–421

2. Stefan Jaeger, Kannappan Palaniappan, Corella S. Casas-Delucchi, and M. Cristina Cardoso. (2010). "Classification of cell cycle phases in 3D confocal microscopy using PCNA and chromocenter features. In Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '10)." Association for Computing Machinery, New York, NY, USA, 412–418. https://doi.org/10.1145/1924559.1924614

3. Vogel, W., Autenrieth, M. & Mehnert, K. (1989). "Analysis of chromosome replication by a BrdU antibody technique. Chromosoma 98, 335–341"). https://doi.org/10.1007/BF00292386

4. Sansam, C. G., Goins, D., Siefert, J. C., Clowdus, E. A., & Sansam, C. L. (2015). "Cyclin-dependent kinase regulates the length of S phase through TICRR/TRESLIN phosphorylation." Genes & development, 29(5), 555–566. https://doi.org/10.1101/gad.246827.114

5. Voets M, Møllersen K, Bongo LA. (2019). "Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." PLOS ONE 14(6): e0217541. https://doi.org/10.1371/journal.pone.0217541

6. Stringer, C., Wang, T., Michaelos, M. et al (2021). "Cellpose: a generalist algorithm for cellular segmentation.", Nat Methods 18, 100–106 https://doi.org/10.1038/s41592-020-01018-x

7. Youngjin Yoo, Lisa Y. W. Tang, David K. B. Li, Luanne Metz, Shannon Kolind, Anthony L. Traboulsee & Roger C. Tam (2019) Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 7:3, 250-259, DOI: 10.1080/21681163.2017.1356750

8. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

9. Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. The Artificial Intelligence Review, 54(1), 137–178. https://doi.org/10.1007/s10462-020-09854-1

10. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., & Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis, 63, 101693–101693. https://doi.org/10.1016/j.media.2020.101693

11. Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging, 32(4), 582–596. https://doi.org/10.1007/s10278-019-00227-x

12. Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., Wu, W., Liu, S., & Lu, H. (2022). ResGANet: Residual group attention network for medical image classification and segmentation. Medical Image Analysis, 76, 102313–102313. https://doi.org/10.1016/j.media.2021.102313

13. Bialic M, Al Ahmad Nachar B, Koźlak M, Coulon V, Schwob E. Measuring S-Phase Duration from Asynchronous Cells Using Dual EdU-BrdU Pulse-Chase Labeling Flow Cytometry. Genes. 2022; 13(3):408. https://doi.org/10.3390/genes13030408

14. Saad J, Fomich M, Día VP, Wang T. A novel automated protocol for ice crystal segmentation analysis using Cellpose and Fiji. Cryobiology. 2023;111:1-8. doi:10.1016/j.cryobiol.2023.02.002.

15. Kromp F, Fischer L, Bozsaky E, Ambros IM, Dörr W, Beiske K, Ambros PF, Hanbury A, Taschner-Mandl S. Evaluation of Deep Learning Architectures for Complex Immunofluorescence Nuclear Image Segmentation. IEEE Transactions on Medical Imaging. 2021;40(7)

16. Araújo, F. H. D., Silva, R. R. V., Ushizima, D. M., Rezende, M. T., Carneiro, C. M., Campos Bianchi, A. G., & Medeiros, F. N. S. (2019). Deep learning for cell image segmentation and ranking. Computerized Medical Imaging and Graphics, 72, 13–21. https://doi.org/10.1016/j.compmedimag.2019.01.003

17. Yang B, Richards CJ, Gandek TB, de Boer I, Aguirre-Zuazo I, Niemeijer E, Åberg C. Following nanoparticle uptake by cells using high-throughput microscopy and the deep-learning based cell identification algorithm Cellpose. Front Nanotechnol. 2023; 5:1181362. https://doi.org/10.3389/fnano.2023.1181362

18. Chen, Y., Lin, Y., Xu, X., Ding, J., Li, C., Zeng, Y., Liu, W., Xie, W., & Huang, J. (2022). Classification of lungs infected COVID-19 images based on inception-ResNet. Computer Methods and Programs in Biomedicine, 225, 107053–107053. https://doi.org/10.1016/j.cmpb.2022.107053

19. Guo, Z., Li, X., Huang, H., Guo, N., & Li, Q. (2019). Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. IEEE Transactions on Radiation and Plasma Medical Sciences, 3(2), 162–169. https://doi.org/10.1109/TRPMS.2018.2890359
20. Aljabri, M., & AlGhamdi, M. (2022). A review on the use of deep learning for medical images segmentation. Neurocomputing (Amsterdam), 506, 311–335. https://doi.org/10.1016/j.neucom.2022.07.070