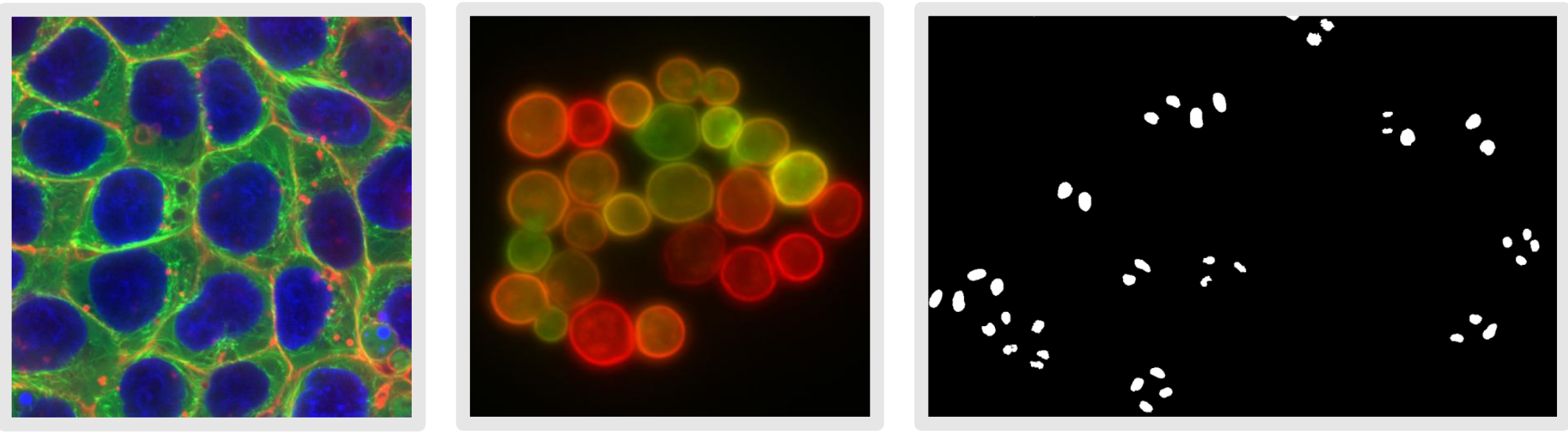
Biocluster
A K-Means Approach

by David Dulkies, David Lehmann, Jonas Schenker, David Schroth

Introduction

Original Datasets:



Cell Nuclei

Yeast Cells

N2DL-HeLa

Motivation:

For quantitative analysis in biological research, precise segmentation of cellular structures from microscopy images—such as nuclei—is essential.

However, there are many challenges due to variations in image quality, noise, and structural complexity. To improve the accuracy of cellular boundary identification, our project focuses on segmentation techniques. We aim to find out which methods work best by comparing different preprocessing steps with the dice score and analyzing the results.

This will help us better understand which approaches are most effective for future projects in cell biology

Methods

Data Preparation & Color Models

- Image Filtering with Gaussian-, Bilateral- and Median-Filter
- Data normalization with division by 255 and z-transformation
- Channel Extraction and combining best filters
- Yeast-Cell segmentation with Watershed
- Convert RGB image to HSV: H (Hue), S (Saturation), V (Value)

Otsu Thresholding

Automatic threshold determination (segmentation method), which separates an image into foreground and background

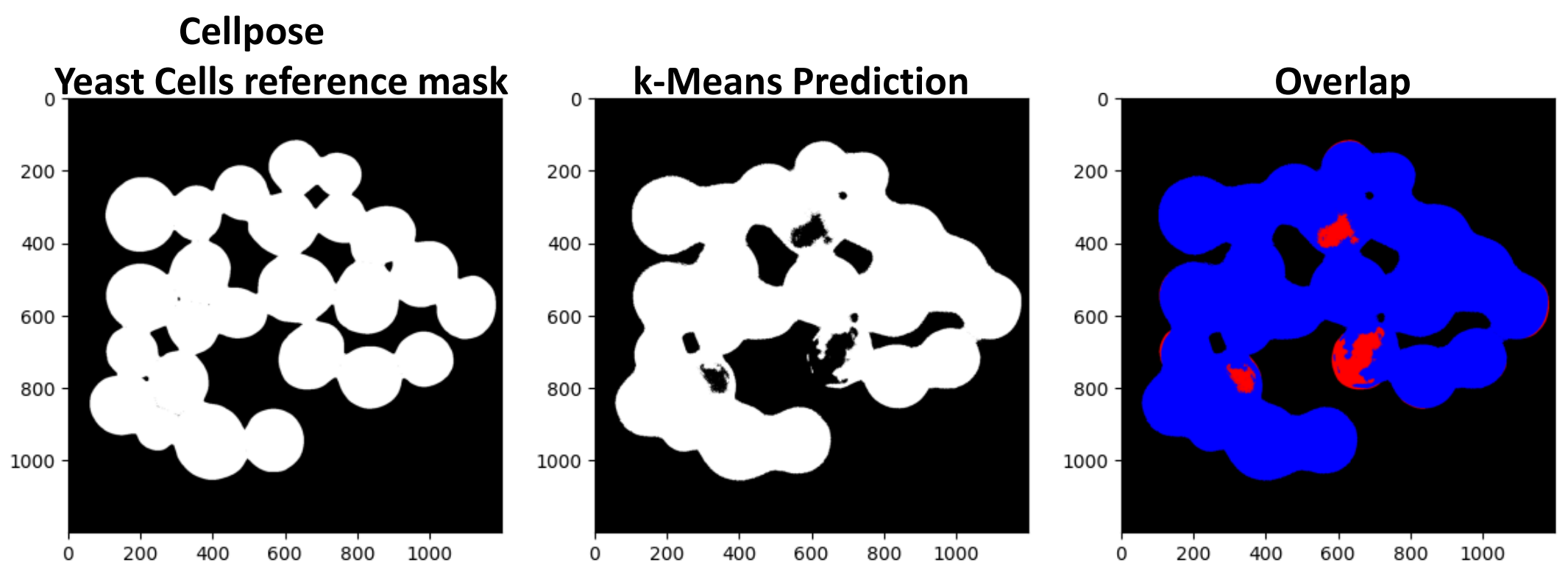
Reference Mask

Cellpose (deep learning-based algorithm), Scikit-learn (Python package)

Dice Score

- Measure of Overlap between two segmentations
- Dice = 2 × (intersection) / (sum of the two areas)
- 1 = perfect match ; 0 = no overlap
- make images binary, standardized image sizes,
- inverted masks if needed, visualized overlays

$$\text{Dice} = \frac{2\times|A\cap B|}{|A|+|B|}$$



k-Means Clustering

- k-Means clustering algorithm for image segmentation in RGB, HSV, and Grayscale → adapted image preprocessing
- Modular implementation with init_centroids, assign_to_centroids,
- update_centroids, reconstruct_segmented_image
- clustering

Different colormodels for segmentation in for- and background:

- Compare colormodels using an ai-generated reference mask via Dice Score
- Segment N2DL-HeLa images (grayscale) with k-Means and Otsu → compare both segmentation methods via Dice Score

Coordinates As Additional Feature

- Create new 3D feature vector by stacking → 3D (intensity, X, Y)
- Run modified k-Means algorithm with t13 from N2DL-HeLa
- Use mask to cluster cells without background

Elbow Method

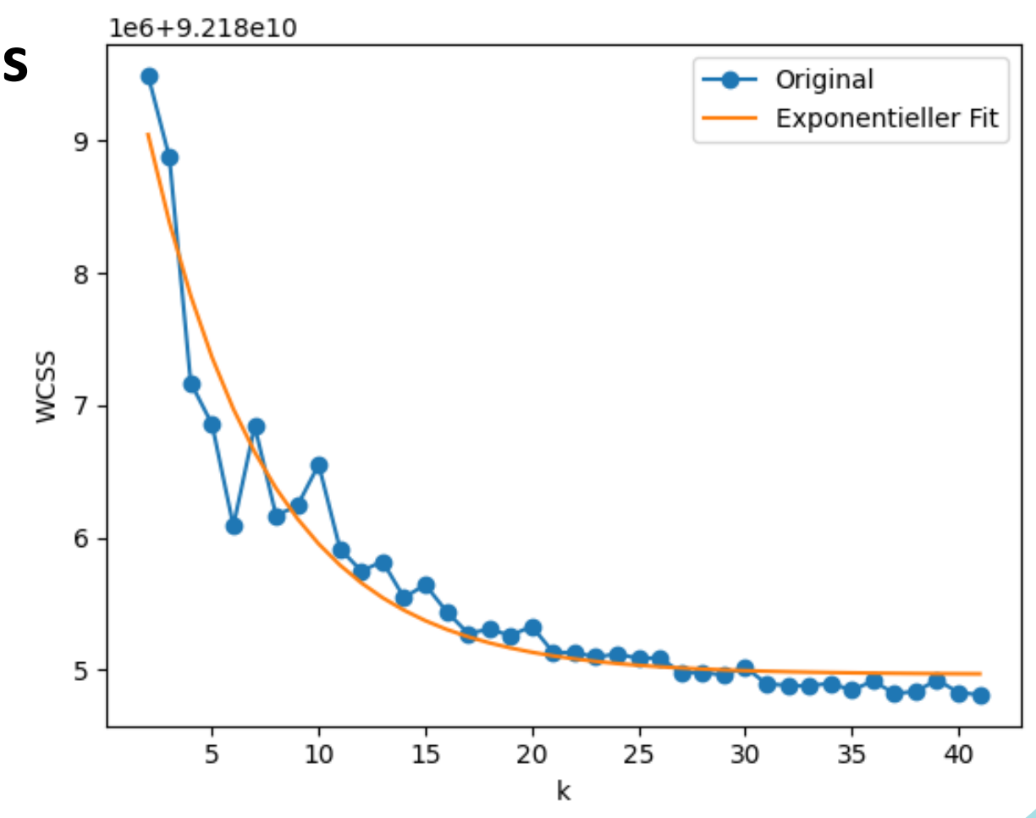
- Elbow identification by maximal distance to line method (DSPA2)

Identify number of dominant fluorescence colors:

- Apply Elbow Method to HSV with all channels, with only hue channel and to RGB Yeast Cells image → which works best?

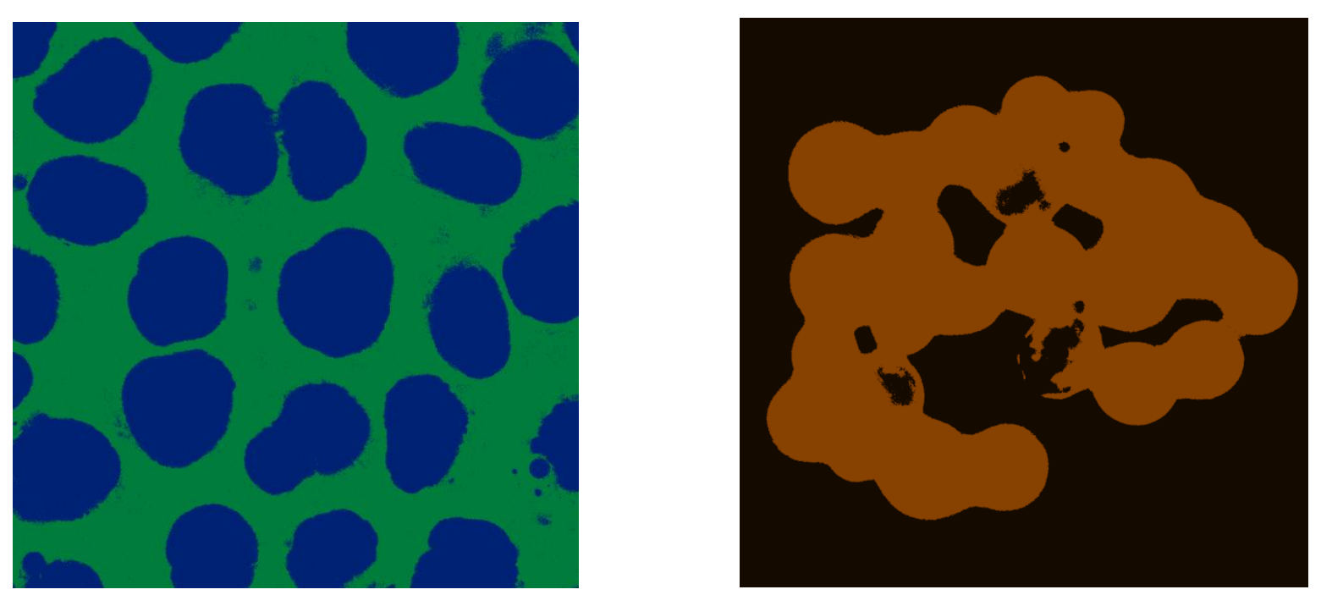
Identify different cell groups based on intensity and position:

- Apply Elbow Method to k-Means with 3D feature vector
- Use Curve Fit to optimize Elbow-Plot for Elbow identification



Results & Discussion

Background Clustering

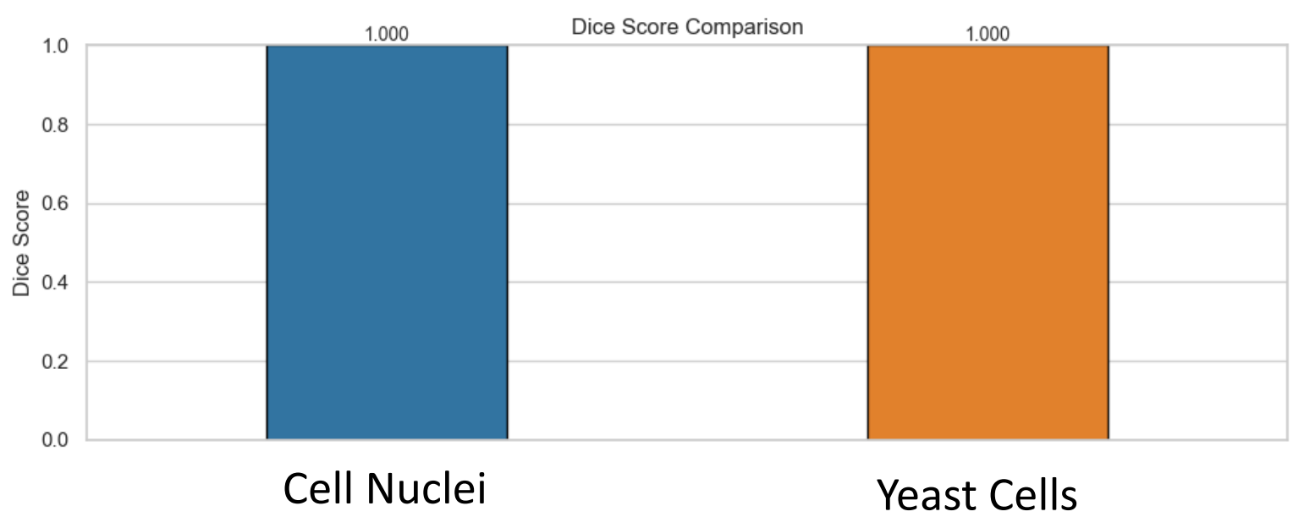


Cell Nuclei

Yeast Cells

k-Means vs. sklearn

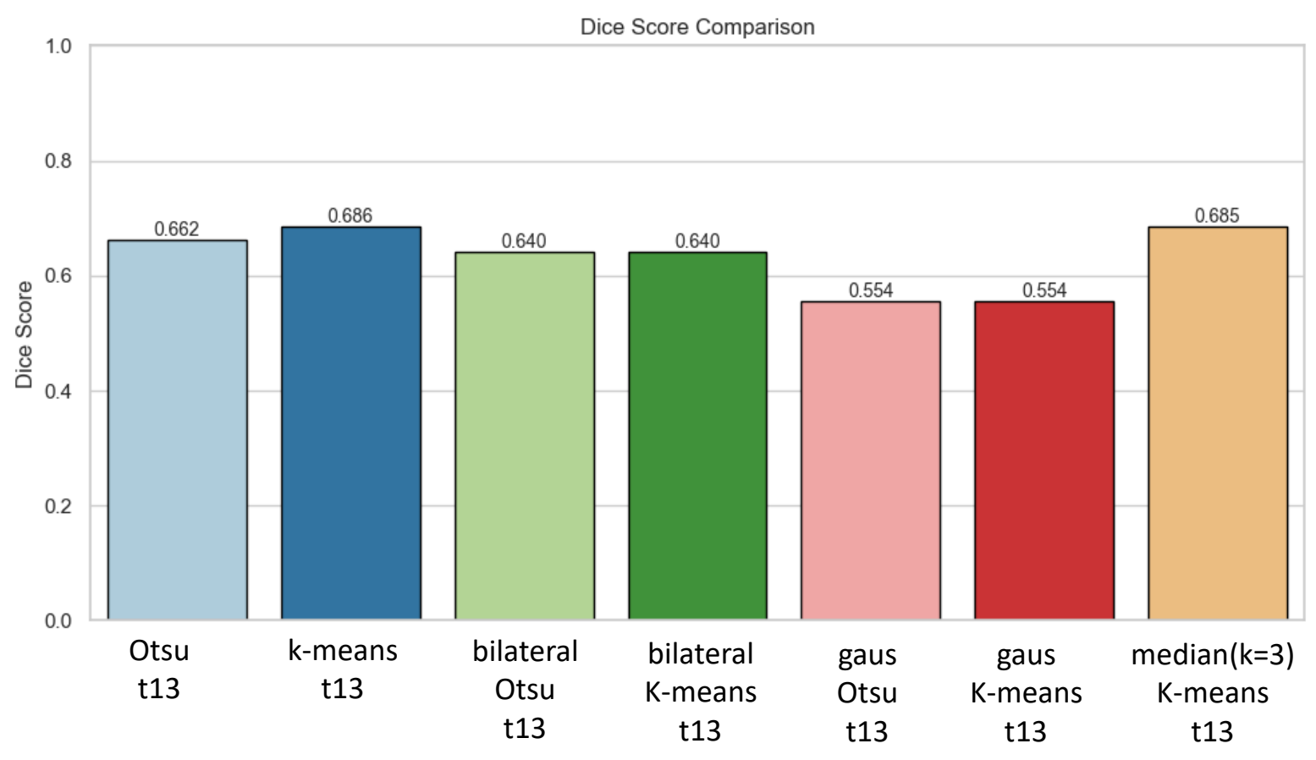
Goal: Prove that the self-implemented k-means works like scikit-learn's k-means.



Findings:
- Self-implemented k-means generates exactly the same results as scikit-learn

→ We can be sure that our k-means implementation works.

Evaluation of different filters on HeLa Dataset



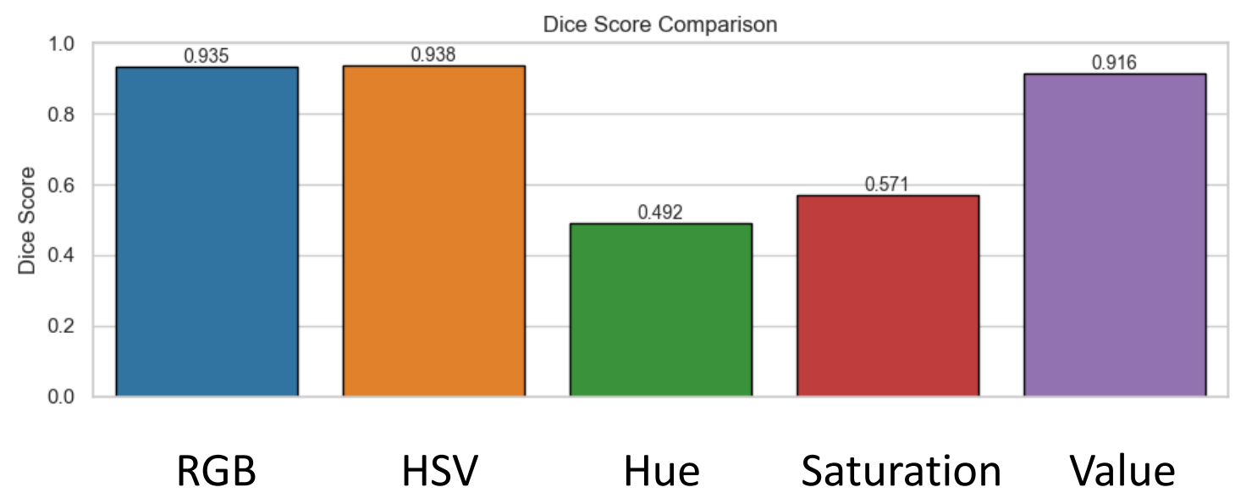
- k-Means has better segmentation than otsu
- bilateral-filter (preserves edges) better than gaus-filter
- Median filter better than other filters
- No filter -> best dice score

- bilateral filter takes additional intensity information --> sharper segmentation
- median filter only smoothens where there is salt & pepper noise. If there is little to no noise, there is no significant influence on dice score

Evaluation of k-Means with different channels

Goal: segment the pictures into foreground (cells or cell nuclei) and background

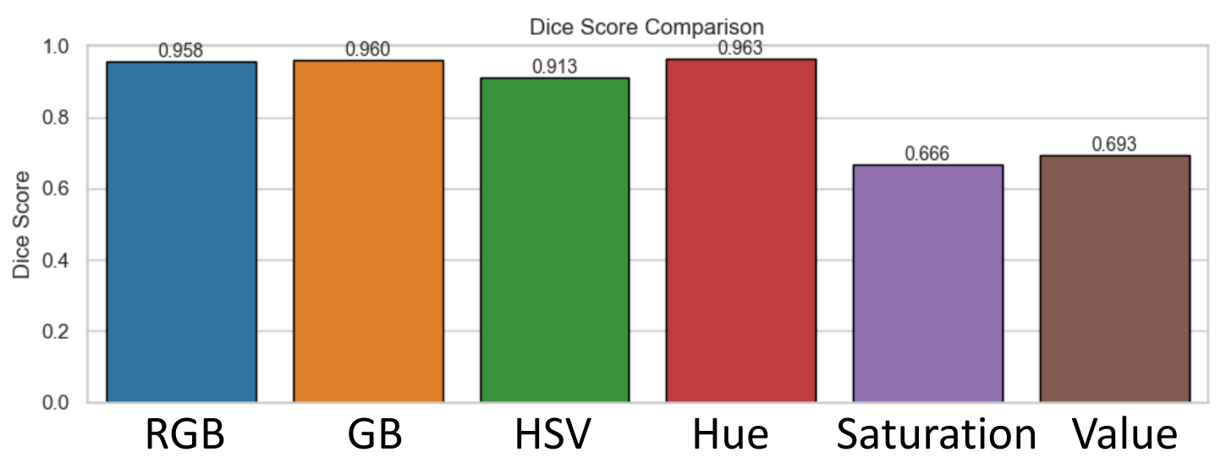
Yeast Cells



Findings:
- HSV is not better than RGB
- Intensity is the dominant feature -> almost as good as all channels combined
- Hue and Saturation alone -> no good segmentation

Factors influencing segmentation:
- Halos, color overlap

Cell Nuclei

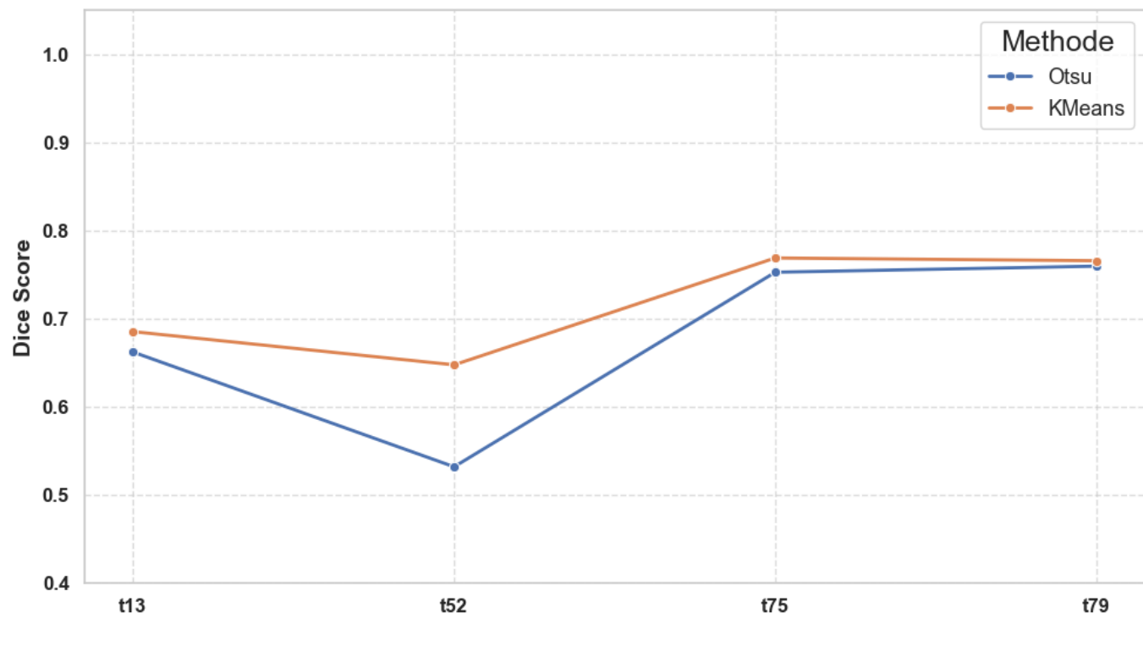


Findings:
- HSV is not better than RGB
- Hue is the dominant feature -> as good as all channels combined
- Saturation and Value alone -> no good segmentation
- Removing the R channel has no significant effect

→ The quality of the segmentation depends more on the dominant channel than on the color model, but the segmentation was very effective

Otsu vs. k-Means

Goal: comparison of the Otsu thresholding and k-means segmentation method



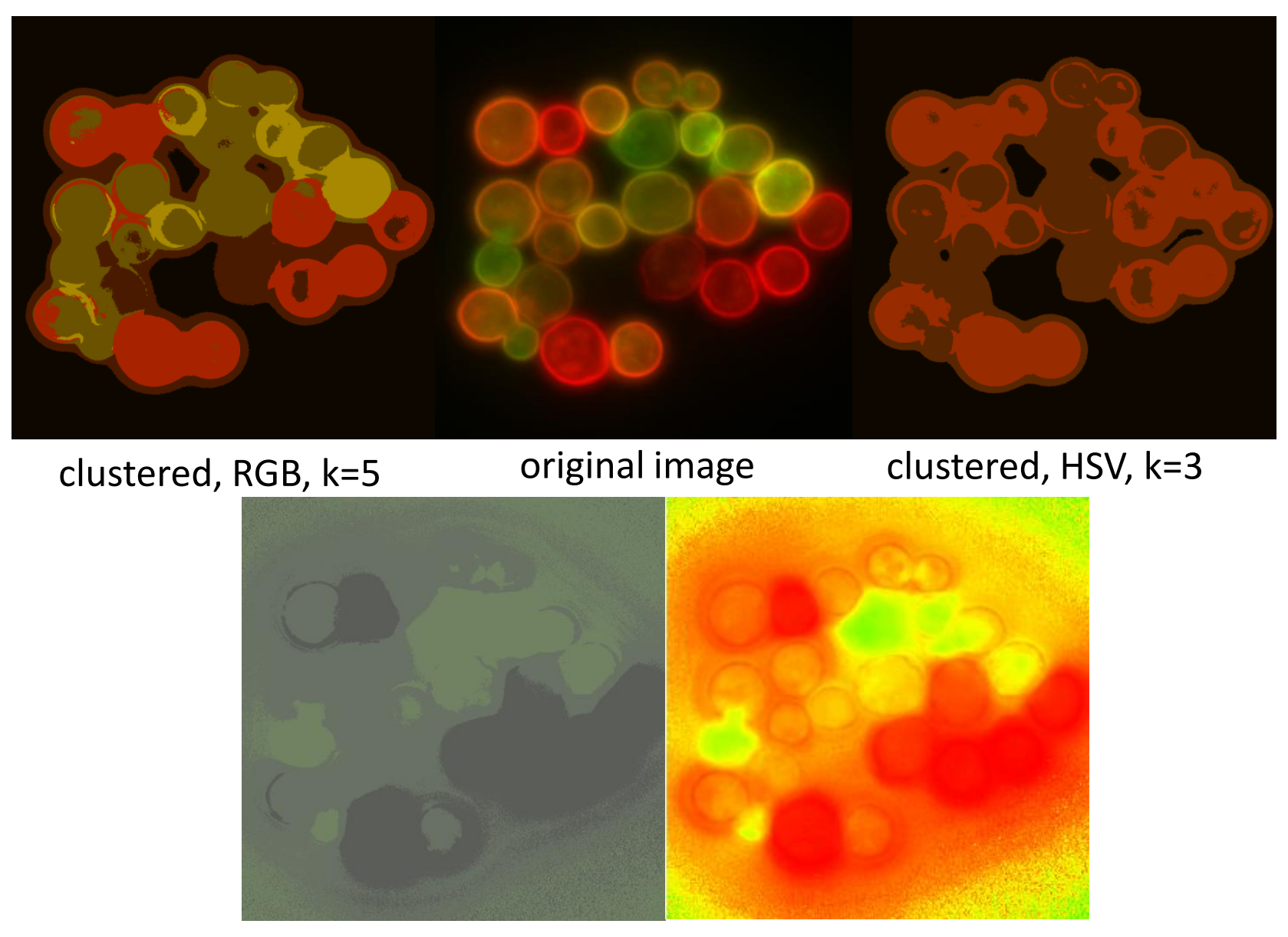
N2DL-HeLa (from left to right: t13, t52, t75, t79)

Findings:
- K-means delivers higher dice-score values than Otsu for all images

Factors influencing segmentation
- Otsu forms a threshold value -> inconsistent lighting, noise, and local variations can affect the threshold

-> k-Means is the better option

Color and cell distinction using the Elbow Method



clustered, RGB, k=5

original image

clustered, HSV, k=3

clustered, Hue, k=3

Original image (Hue)

Optimal clusters with different colormodels:
- RGB: k = 5
- HSV all channels and single Hue channel: k = 3

Emerging problems with different color models:
- **HSV all channels:** Cells with different colors but similar intensities get clustered in the same cluster
- → cells are not clustered correctly by color
- **RGB**: Halos get clustered in one cluster with darker cells → intensity not separated from color
- **HSV Hue:** Huge halos prevent specific segmentation of cells of different colors (cells are poorly visible)

Which colormodel is the best?
- To determine how many different colors exist
- → HSV Hue
- To determine which cell has which dominant color and therefore potentially a specific gene expression
- → RGB **OR** potentially HSV if Halos could be reduced by for example thresholding

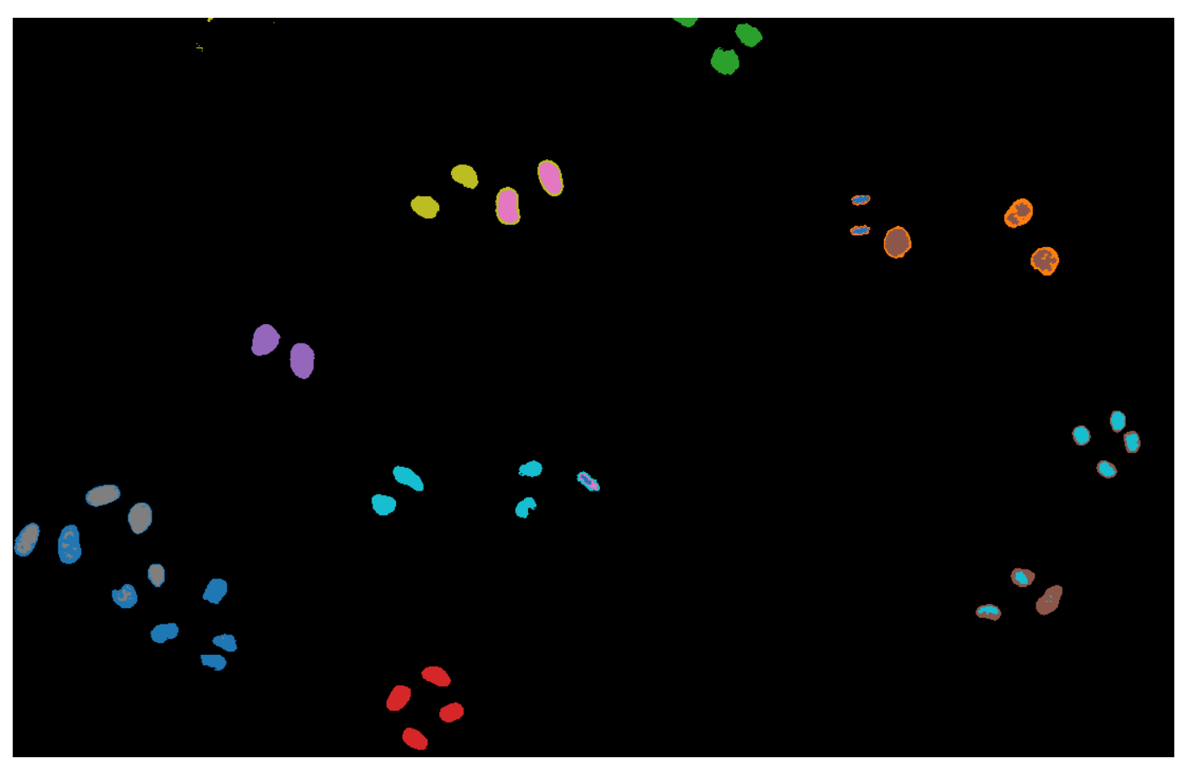


Image t13, clustered by intensity and coordinates, same clusters have same color

- Intensity weighted 2x in comparison to coordinates → achieve clustering where intensity has same weight as coordinates
- Optimal k identified at 12/13 clusters → not quite consistent results due to difficult elbow identification

Conclusion

Filter Evaluation:
- Using no filter is the best because important features are preserved
- Influence of halos is only minimal on Dice Score
- Although advanced segmentation and noise-targeted filters improve clarity, unfiltered data ultimately provides the most accurate results

Colormodels:
- The dominance of individual channels is a key factor for the segmentation
- The colormodel (RGB or HSV) has no significant impact on the result -> but the segmentation was very effective
- The AI-generated ground truth makes the dice score less reliable

Otsu thresholding vs k-Means:
- Otsu is easy and simple to perform, but is more sensitive to interference
- k-Means is more complex, but also less sensitive to interference

-> k-Means is the better segmentation method for our images

Elbow Method:
- Elbow Method is inconsistent in general, Elbow difficult to identify → Shilouette Score should be used next time instead → minimize wrong results
- In our case regarding data preparation, wrong clustering and unprecise segmentation, RGB is the best color model to distinguish cells from each other by their fluorescence color.
- k-Means can be applied with additinonal features like coordinates

Dice Score:
- Challenges: Convert images into a comparable format; missing ground truth for two images (used Cellpose).
  Strengths: Simple and robust when comparing areas.
  Weaknesses: No information about shape or location.
  For future projects: Since fine details are hardly noticeable in the Dice Score due to filtering, crop images for better focus.

References:

Vassilvitskii, Sergei, and David Arthur. "k-means++: The advantages of careful seeding." *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007.

DSPA2: Data Science and Predictive Analytics (UMich HS650), VIII. Unsupervised Clustering.

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nature Methods, 18(1), 100-106.