Universidad Autonóma de Yucatán

Facultad de Matemáticas

Segmentation of HeLa Cells Using Persistent Homology

THESIS

Daniel Antonio Brito Pacheco

A thesis presented for the degree of

MSc. in Computer Science

Mérida, Yucatán, México
2024

Acknowledgements

First and foremost, I would like to give a huge thanks to my parents Carlos and Elda, as well as my brother David. If it were not for them venturing into the scientific field before me, I would not have been inspired to pursue the academic path of life. They are the people who have encouraged me to go down this road filled with satisfaction, accomplishment and wonder.

Secondly, I would like to thank Dr. Reyes-Aldasoro for giving me such a warm welcome to London and being a guiding light when otherwise surrounded by fear and unknowing. It is with his company that I managed to accomplish focusing on research when home was very far away in time and space and he, together with Dr. Giannopolous have pushed me to achieve a higher quality of work than I thought I could manage.

A special acknowledgement goes towards all the friends and people around me who appreciate me and who value my curiosity, allowing me to embrace it as an important part of who I am. Friends I've made along the way, friends that I've recently met. Friends who have helped me through hard times and who've been with me in the good times. Those friends who live freely, unburdened and are always up for a deep chat or a fun music jam.

Finally, I would like to thank the institutions. Universidad Autonoma de Yucatan for being a space where computer science and mathematics are very much alive and thriving. Where a collection of curious minds and professional scientists exist encouraging the students to keep learning and exploring. City, University of London for being a place that welcomed me for a few months while I worked with Dr. Reyes-Aldasoro and gave me a place where I could focus and develop ideas in a tranquil environment. Lastly, the Francis-Crick institute for providing the very crucial imageset of HeLa cells that is used throughout the project.

Publications Related to this Work

D. Brito-Pacheco, C. Karabag, C. Brito-Loeza, P. Giannopoulos, C.C. Reyes-Aldasoro. Relationship Between Irregularities of the Nuclear Envelope and Mitochondria in HeLa Cells Observed with Electron Microscopy. *International Symposium on Biomedical Imaging* (2024)

Contents

Chapter 1

Introduction

The field of Topological Data Analysis (or Computational Topology) has received a lot of attention in the scientific community during recent years. This is due mainly to its robustness to noise and its versatility in being applied to different contexts. Specifically the concept of Persistent Homology has been applied to many areas, from ecology [1] to image processing [2]. In combination with Machine Learning, TDA can be a powerful tool. The broad objective of this thesis is to continue venturing into the research of image segmentation performed with Persistent Homology such as [3] [4] and [5].

Additionally, in the field of biomedicine, investigations into mitochondrial function and damage to the organelle have been found to be correlated to the presence of disease [6], [7]. In cancer, the role of mitochondria has been highlighted to be manifold [8]. Mitochondria are organelles than can change shape and distribution inside the cell. It has been shown that there exists a considerable amount of communication between mitochondria and the nucleus of a cell [9]. This communication may lead to an alteration on one organelle when the other also becomes altered [10]. Damage in both structures can contribute to the progression and metastasis of cancer [11]. A first step towards studying the communication and deformation between these organelles is performing a segmentation of the two. HeLa Cells are an immortalized line [12] of cells derived from a culture of a now-diseased cervical cancer patient called Henrietta Lacks. These cells have been extensively used for research purposes in cancer [12] [13].

It is with the knowledge presented in the previous paragraphs in mind that the segmentation tests of the developed PH algorithms will be applied to a set of

images of HeLa cells obtained with an Electron Microscope.

Electron Microscopy presents a unique challenge when performing segmentation. While it is able to produce images at a much higher resolution than conventional fluorescence-based imaging techniques - thus being able to capture details in the cells and its organelles; it sacrifices contrast, making segmentation with traditional techniques much harder. Additionally, the complex nature of the organelles inside the cell, as captured by the Electron Microscope, make it hard to capture the shape of the structures.

The main goal of the thesis is to develop an algorithm based on Persistent Homology in combination with traditional image processing methods to segment the mitochondria found in the HeLa cell image-set. This test presents the process through which three different algorithms were created to tackle this goal. These algorithms (or models) are the following:

1. Based on Persistent Homology and Machine Learning.
2. Uses only traditional Image Processing techniques.
3. Combines the two previous ones into a Hybrid Algorithm.

Chapter 2 presents an overview of knowledge and theory that will permit the reader understand the following chapters. This is not a comprehensive review of theory as it is assumed the reader is at least familiar with basic concepts of Topology, Machine Learning, and Image Processing.

Chapter 3 presents some studies which are similar to the research performed in Chapter 4. In this Chapter, the gaps in research are also presented and the reasons behind moving forward with the main body of work in the thesis are highlighted.

Chapter 4 contains the main research. Section 4.1 presents the dataset used to fit the models presented in Section 4.2. In this Section, the three previously mentioned algorithms are presented in detail, and an additional Deep Learning model known as MitoNet is also presented. This model represents the current state of the art in electron microscopy image segmentation.

Chapter 5 compares the performance of the segmentation made with all four segmentation algorithms against each other and against a human-made Inter-Observer segmentation, thus providing a fair comparison point.

Chapter 6 shows an application of the best-performing segmentation algorithm

- in this case, MitoNet. The model is used to segment mitochondria from the cells in the stack of images, and a separate Image Processing procedure is used to extract deformations known as invaginations from the nuclear envelope. The relationship between the shape and size of the nucleus, cytoplasm, and mitochondria are compared by using a statistical metric known as the Pearson Correlation Coefficient.

Finally, Chapter 7 discusses the results shown in Chapters 5 and 6, diving into greater detail about the findings, consequences, and future research directions.

Chapter 2

Theoretical Framework

In this chapter, the concepts and theory behind the research is presented. The chapter is divided into three sections:

- The first section focuses on biomedical images, and provides some background on the biology of cells, this is done so the reader can understand the full picture behind the research that was performed.
- The second section presents the types of image segmentation and a few traditional and recent methods for performing the segmentation task.
- The last section is all about TDA. The way this section is written is targeted towards readers with a slight mathematical background, since the concepts are not explained in full detail.

2.1 Imaging of HeLa Cells

Imaging is widely used in medicine for aiding in diagnoses and prognoses. CT scans, MRI's, microscopes and other machines are employed to take images of bone and organ tissue, cancerous and healthy cells, etc. This thesis focuses on images of HeLa Cells obtained through EM. HeLa Cells are an immortalized line [12] derived from a culture (extracted without permission) of a now-diseased cervical cancer patient called Henrietta Lacks. These cells have been used widely for research purposes [12] [13] as they are extremely durable.

EM produces grayscale images, with low contrast but high resolution. Herein lies

the problem that this project aims to tackle. Current methods rely on brightness of pixels to perform the segmentation, thus, higher contrast on images leads to an easier or more successful algorithm. Recent advances in CNN's have yielded algorithms which are better at segmenting all organelles of cells in EM images [14]. These recent works will be presented in the following chapter.

In this work, the main focus of segmentation is the mitochondria of the HeLa cells. A mitochondrion is an organelle whose main function is producing adenosine triphosphate (ATP). This is the primary source of energy for cells and is used for most biochemical and physiological processes, such as growth, movement and homeostasis [15].

2.2 Machine Learning

Machine Learning is a very large field of study that is concerned with using statistical and optimization methods to study datasets and find patterns that generalize to previously unseen datasets. The basic idea behind machine learning is to get an algorithm to learn to perform tasks without explicitly programming the instructions to do so [16]. In this section two machine learning algorithms are presented in detail (Classification Trees and Random Forest Classifiers) as they form an integral part of the final algorithm developed in this work. Meanwhile, a few more are introduced with a less thorough explanation.

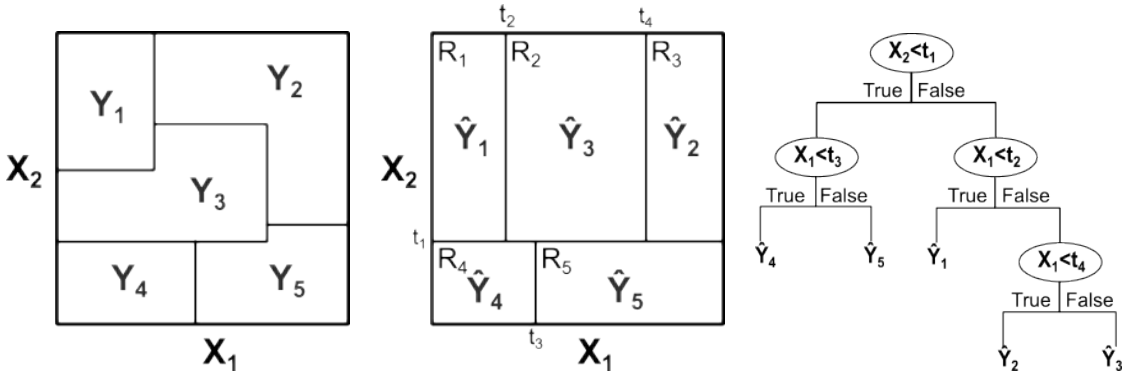
2.2.1 Classification Trees

Classification trees refer to a set of machine learning methods which partition the feature space into a set of rectangles to fit a simple model [17]. Consider a problem with response variable Y which can take any of 5 values Y_1, Y_2, Y_3, Y_4, Y_5 (5 classes) and has inputs X_1 and X_2, each taking values in $[0, 1]$. Visually this is represented in Figure 2.2.1 (a). It is clear to see that the target partition cannot easily be described by simple conditions like $X_i > c$. Tree-based methods aim to obtain partitions by recursively split regions in two like shown in Figure 2.2.1 (b). In the example, the feature space is first partitioned at $X_2 = t_1$; then the resulting region above the threshold is partitioned at $X_1 = t_2$, and finally at $X_1 = t_4$; the region below t_1 is partitioned at $X_1 = t_3$. This process results in the feature space partitioned into corresponding regions called R_1, R_2, R_3, R_4, and R_5. The

regression model is given as the following predictor function:

$$\hat{f}(X) = \hat{Y}_i, \quad \text{if } X \in R_i$$

The more common binary tree representation of these methods is shown in Figure 2.2.1 (c). This way of showing the partition is much more intuitive.



(a) (b) (c)

Figure 2.1: (a) The target partition of the feature space. (b) An example of a partition obtained by the classification tree method. (c) The partition represented as a decision tree.

The question that remains is how to grow the tree by choosing the splitting variable and splitting point. Let m be a node representing a region R_m, with N_m observations. To grow a tree, the algorithm seeks to minimize the Gini index given by:

$$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where K is the total number of classes and $\hat{p}_{mk}$ is the proportion of observations of class k in node m:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k).$$

The algorithm is greedy in the sense that to choose the variable j and split point s to create the region R_m, all possible pairs are tried and the one that minimizes the Gini index shown above is chosen to create the node. The process is repeated with the two newly created regions.

Pruning

Once a tree is fully grown, it is usually a good idea to "prune" the tree. This means to remove nodes in order to avoid overfitting to the training data. This is done by minimizing a cost-complexity measure - some function which quantifies the balance between the simplicity of the model and how good it fit to the training data. For a given terminal node m, representing a region R_m:

$$\frac{1}{N_m} \sum_{i \in R_m} I\big(y_i \neq k(m)\big),$$

which is simply the ratio of incorrect classifications of the training samples for the given node m.

2.2.2 Random Forest

A Random Forest Classifier is part of a large number of methods called Bagging (Bootstrap Aggregating) that involve training multiple models on random subsets of data and aggregating their predictions (usually through averaging or voting). The basic idea is simple:

1. Choose a random subset S of the training data.
2. Use S to grow a decision tree using a subset of the variables.
3. Repeat from step 1 an appropriate amount of times.

A new data point is run through all the created decision trees that were grown, and each assigns a label to the data point as if 'casting a vote'. The label with the most votes is the label that is assigned to the data point.

2.3 Image Segmentation

Image segmentation is one of the most important tasks in image processing. The goal of image segmentation is to simplify a given image in order to analyze it. This is done by dividing the image into different regions [18]. The task of image segmentation can be divided into two main types: Semantic Segmentation and Instance Segmentation. The main difference lies in whether segmented objects are grouped with others or not. This is explained in further detail below:

- Semantic Segmentation: each pixel of the image is simply assigned a class out of a given number of possibilities. For example, in Figure 2.2, taken from [19], the people in the image are all grouped into one class, the table and everything on it gets assigned a class of its own.
- Instance Segmentation: identifies individual instances of the same class. For example, in Figure 2.2, each person has been identified as a separate object.

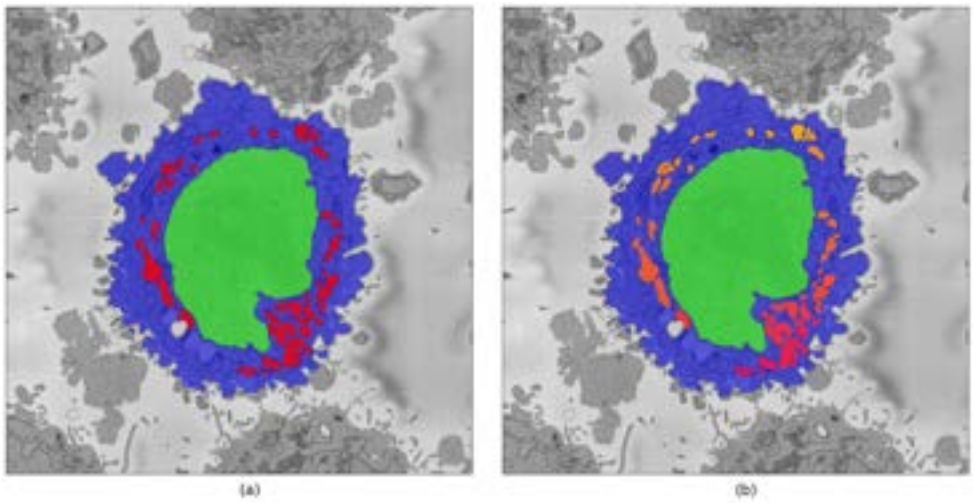


Figure 2.2: Left: Semantic Segmentation. Each different part of the cell is given a certain class (represented by the colors). Right: Instance Segmentation. Each different instance of the parts is given a different color. Notice the difference in the multiple colors given to the mitochondria.

In [19], image segmentation techniques are classified into the following categories:

- Thresholding methods: these methods simply consist in defining one or multiple thresholds for pixel-values such that those pixels that fall within a threshold can be considered as belonging to a given class. Thresholds can be variable over the image. A well-known thresholding method is Otsu's method [20]. This example is used to essentially separate the foreground from the background in an image.
- Edge-based methods: edges are detected by taking advantage of the quick change in intensity in neighboring pixels, they are then connected, this

process yields boundaries of objects in an image. The boundaries segment the image into various regions.

- Region-based methods: regions are segmented by the properties they present. There are two main ways to perform this type of segmentation:

    1. Region growing takes seeds (single pixels) in an image and begins "growing" the region around the pixel by determining if neighboring pixels are similar enough.

    2. Splitting and merging first divides an image into many regions, then, adjacent regions are combined if they are similar enough.

- Clustering-based methods: there is a large area of research on clustering data. If the image is viewed as a point-cloud directly or features of each pixel are extracted to map onto a point cloud, all the theory behind data clustering can be applied. The two main branches are Hard Clustering, where pixels are assigned to only one class; and Soft Clustering, where pixels can be in multiple classes at once with a certain degree of belonging.

  A popular clustering method is k-means clustering [21], in which k centers are initially chosen (manually or at random) and will then move around the point cloud at each iteration - always looking to minimize distances between the points and the centers, and the variances within clusters.

- Watershed-based methods: these methods draw parallels between the image and the physical world. They are easily explained by an analogy: the intensity of pixels are seen as "height above the ground", thus, if one were to place water over a pixel, the water would flow down to a local minimum (a valley) - all pixels whose flow of water ends up at the same valley are considered as the same region.

- PDE-based / Variational methods: they take advantage of mathematical theory developed for solving Differential Equations (either numerically or analitically). They generally consist in defining a PDE over the domain of the image and solving it. A lot of differences exist in the way the initial PDE is defined, as they can involve the Total Variation, a diffusion filter, and more.

- ANN-based methods: by far the most popular segmentation algorithms used recently to segment medical images, ANNs and CNNs can be trained with enough training examples to extract the features of an image and to

perform the segmentation. The invention of the U-Net [22] was a milestone in medical image segmentation and resulted in a quick advancement of precision and accuracy for performing this task.

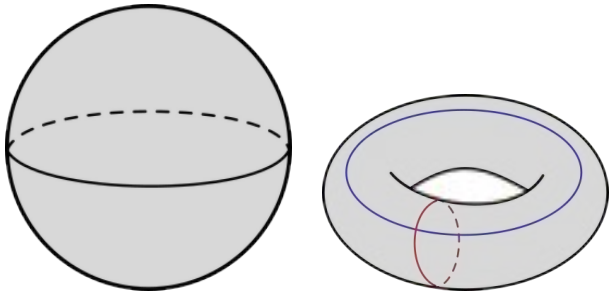
2.4 Topological Data Analysis

Topological Data Analysis is the collection of statistical methods that are used to find structure in data. TDA is a large field that encompasses many concepts but this section focuses mainly on the concept of Persistent Homology. Additionally, some concepts related to PH will not be defined here as this section does not aim to be a rigorous or comprehensive review of PH, the concepts defined and explained in this section are enough for the reader to understand the mathematical theory behind the research performed in the following chapters. All definitions are taken from [23], [24] and [25].

2.4.1 Homology

The homology of a topological space is a set of topological invariants. These invariants essentially form a characterization of the topological space through the k-dimensional Betti numbers β_k. Each number β_k counts the number of k-dimensional holes. Each k-dimensional hole is understood, informally, as cycles which are not the boundary of a $k+1$-dimensional subset of the topological space. In simple terms this means that β_0 counts the number of connected components, β_1 counts the number of 'loops', β_2 counts the number of 'cavities', etc. This is better illustrated with an example. Consider the hollow sphere in Figure 2.3 (a). This set is exactly one connected component, so $\beta_0 = 1$; it has no 1-dimensional holes, so $\beta_1 = 0$ [1]; it contains exactly one cavity, so $\beta_2 = 1$. Similarly, if we consider the torus in Figure 2.3 (b), the Betti numbers in this case yield: $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$. The difference between the two shapes being the number of different 1-dimensional holes on each (the two 1-D holes on the torus are shown in red and blue).

[1] The cycle drawn in black is the boundary of a hemisphere; thus it is not a hole. Any cycle on the sphere forms a boundary of a section of the sphere, thus cannot be considered holes.

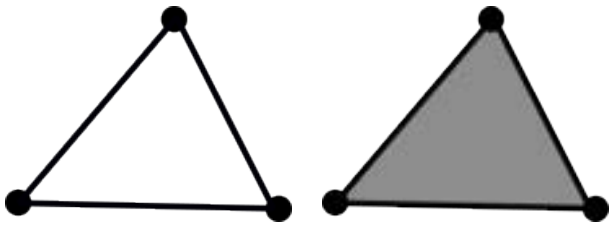


(a) (b)

Figure 2.3: (a) A sphere's Betti numbers are: $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$. (b) A torus' Betti numbers are: $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$.

2.4.2 Simplex and Simplicial Complex

A k-simplex can easily be understood as the k-dimensional generalization of a triangle: a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, etc. More specifically, a k-simplex is the convex hull of its $k+1$ vertices, this essentially means that a simplex has to be "filled in". To illustrate this, consider Figure 2.4 (a) and (b): the first can be thought of as a set of three 0-simplices (vertices), three 1-simplices (edges) and zero 2-simplices (no triangles); whereas the second can has exactly the same 0- and 1-simplices but with the difference that there is a 2-simplex present in the set. Moreover, we say that a simplex F is a face of another simplex S if F is the convex hull of a nonempty set of the points that define S. In Figure 2.4 (b) the 2-simplex (triangle) has three 1-faces.



(a) (b)

Figure 2.4: (a) The diagram shows a set that includes three 0-simplices (points), three 1-simplices (edges) and no 2-simplex. (b) In contrast to (a), this set includes the same 0- and 1-simplices but additionally includes one 2-simplex; shown by filling in the void inside the triangle.

A simplicial complex Σ is a set of simplices such that for every simplex S in Σ, the faces of S are also in Σ. Figure 2.5 shows an arbitrary example of a simplicial complex.

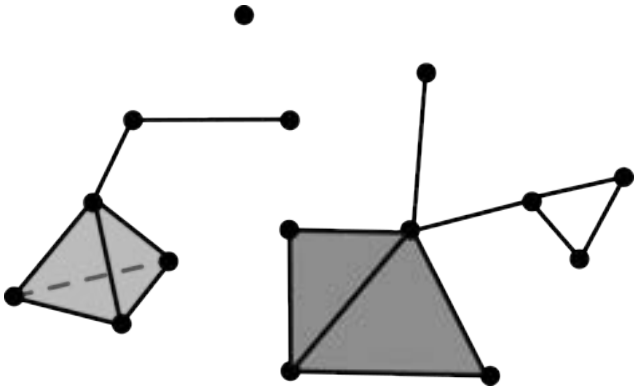


Figure 2.5: The diagram shows a simplicial complex.

2.4.3 Filtration

A filtration F of a space S is a sequence of subspaces (S_k) such that $S_i \subset S_j$ for all $i \leq j$. As an example, take the Simplicial Complex Σ shown in Figure 2.5; a filtration of Σ can be created by adding points from left to right, adding edges when the two points between them exist and adding triangle faces when the three edges that define its boundary exist. This is shown in Figure 2.6.

2.4.4 Persistent Homology

Persistent Homology can be simply thought of as keeping track of how a the topology of a space S changes through a filtration of it. This general definition applies to any filtration of any topological space. However, in the context of TDA it is mostly applied to filtrations of simplicial complexes. In practice, a simplicial complex is formed through connecting points by varying a distance parameter (like the Vietoris-Rips Complex); on images, a filtration can be built by thresholding at different brightness values, etc.

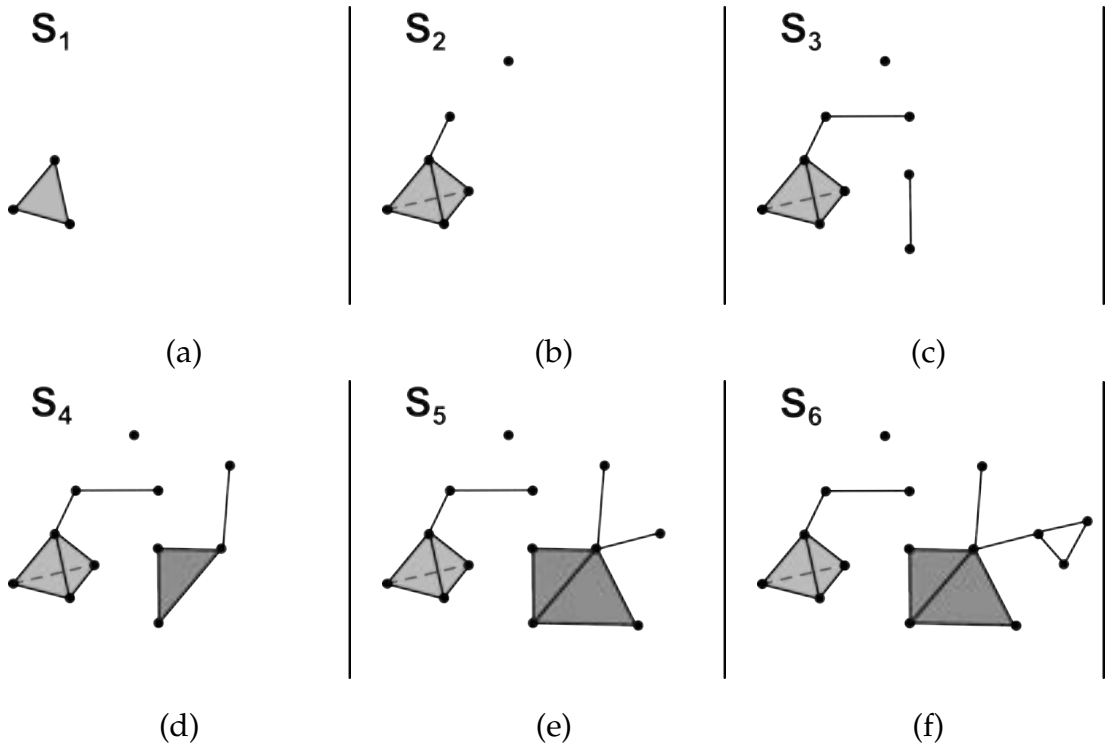


Figure 2.6: (a)-(f) show an example of how a filtration of $\Sigma = S_6$ is made by adding vertices from left to right. Whenever two vertices of an edge (which exists in Σ are present in the subspace S_k, the edge is added. Similarly, for triangles that are present in Σ, if their three edges are present in a given subspace S_k, the triangle face is added.

Chapter 3

Previous Works

Understanding the context of existing research is crucial for performing new research and situating it within the academic landscape. In this chapter, a comprehensive review of primary studies similar to the work done as part of this thesis are presented. By examining these studies, the main reason for performing the research in upcoming chapters is highlighted as gaps in the current state of the art are found and discussed. The compilation of the previous studies demonstrates the relevance and contribution of this study to the field.

3.1 Systematic Review Process

The studies presented in Section 3.2 were found following a Systematic Review process. The guidelines in [26] were followed in order to perform the review. Before identifying the research that will be included in the review, it is first necessary to have a clear goal or question that must be answered from this process. The main research question is the following:

"How well have TDA tools (mainly Persistent Homology) aided in image processing tasks, what are the main tasks it has been applied to (segmentation, classification, etc), and what are some areas still left to explore with these tools?"

This overarching question can be split into three "sub-questions":

1. *How succesful have TDA tools been when applied to image segmentation / classification / denoising / registration? (Have they improved upon the state of the art?)*

2. *What types of imaging data are predominantly processed using TDA?*
3. *What are some gaps that are still to be explored / researched?*

Now that the research questions have been decided upon, the review must be conducted. Broadly, the steps followed for conducting the review are as follows:

1. **Identification of research** - in this step all the studies that are related to the research question are identified. It is during this step that an online search for papers is usually performed using a well-crafted search string. In this study, the database that was used to find such papers was PubMed where the following search string was used:

   ("Topological Data Analysis" OR "Persistent Homology" OR
   "Computational Topology")
   AND
   ("microscop*" OR "medical imag*" OR "biomedical imag*")
   NOT
   Review

   The first parenthesis is used to find all studies that talk about TDA. This field is sometimes called Computational Topology and many times the words Persistent Homology are used without mentioning the field of TDA. Other concepts pertaining to TDA are sometimes used without mentioning the field, such as Reeb Graph, and Morse Theory. However, there are many such concept that covering them all in a single search string would be an impossible task, and more importantly, they are not very relevant to this thesis as PH is the most important part of TDA used in this work. The second parenthesis, together with the first, aims at finding studies that used TDA to perform some kind of medical/biomedical image analysis. Finally, the exclusion of the word "Review" is done in order to find only primary studies.

2. **Selection of studies** - not all the scientific papers found in the initial search will be relevant to the review. Thus, inclusion and exclusion criteria must be used to select those which are. The following criteria were used:

   Inclusion:

   - Primary research articles published between 2010 and 2024 (inclusive) in specialized journals.

Exclusion:

- Articles not focusing on image analysis.
- Articles not including some aspect of TDA or PH for a key step in the analysis.
- Articles not using medical images.
- Duplicate articles.

3. **Assessment of study quality** - this step is presented by Kitchenham et al. in the previously mentioned guidelines [26] but in this process is not explicitly followed, as the initial number of studies found is low. All peer-reviewed papers were assumed to be of sufficient quality to include in the final list of articles.
4. **Data extraction** - when reading through each article, the following quantitative data was extracted in order to have statistical information on each field for the papers:
    - Type of Image Processing - studies are categorized by the task of image processing that is performed; classification, segmentation, or more broadly, some type of analysis.
    - Type of Image Data - the types of the images or acquisition techniques are also registered for each article. Some examples include but are not limited to Bright Field Images (BFI), CT Scans, Microscopy, MRIs.
    - Requirement of Staining - a large part of the problem for the segmentation of EM images is that minimal staining is required to obtain the images. Thus, in order to compare the research performed here with the current state of research, these data were also extracted from the articles found.

    Qualitative data were also extracted from each paper. The following points were qualitatively summarized:

    - Successful or not - when comparing against more traditional state-of-the-art methods, did the proposed methodology perform better according to standard metrics? A synthesis of these results was created for each study.

- Summary of methodology - a brief summary of the methods shown in each paper was created.

3.2 Studies

In this section, the most relevant studies that closely resemble the research performed in this project are presented together with a brief summary of their work.

- In [4], tumors are segmented from large histological Whole Slide Images (WSI). The results are obtained by splitting the WSI into patches, then each patch is classified as tumor or not. Classification is done by first extracting a feature known as Persistent Homology Profile (PHP). By first thresholding the patches with different t values, from 0 to 255 and calculating the first two Betti numbers (β_0, β_1) at each t. The function of Betti numbers versus t is known as the PHP. Finally, a k-NN procedure is done against the PHP of previously found exemplar images to perform the classification.

  This article is presented first as the algorithm developed in it acts as the base for the methodology that is shown in Chapter 4.

- In [27] Persistence Homoology is used to draw a persistence diagram from an image by drawing a parallel between voxels at a given brightness value and simplicial complexes. As a brightness threshold changes, the filtration is built. Using the persistence diagrams for a given segmentation, a topological loss function is defined as the distance from the predicted persistence diagram to the target persistence diagram. This loss function is summed with the Dice score and a customized cross-entropy function known as DeepVess to yield the complete loss function used to train the Neural Network.

  This article is very relevant to the methodology in this project because it is the only article that was found to use PH to aid in performing segmentation on microscopy images.

- [28] and [5] showcase techniques to perform segmentation on histological images. The process is done by 'dismantling' each individual image according to the brightness value of the pixels into a sequence of images. An inclusion tree, and, subsequently, a filtration is built by focusing on the connected components of each resulting image. Then, persistence is calculated,

noise is removed and a region growing procedure is applied to finalize the segmentation process.

- The authors in [29] developed two feature-extraction modules based on persistent homology that are used to guarantee both topological correctness and pixel-wise accuracy are kept when performing segmentation of echocardiography images. In the first module a loss function is defined based on the topology of the segmentation which is used to train the model. The second module ensures myocardial integrity by considering the connectivity of the heart tissue.
- In [30], CT Scans are segmented in order to obtain a 3D model of the small bowel. This is done by defining a Loss function which is then used to fine-tune a Deep CNN. For each 2D slice X, a CNN yields a probability image Y (with values in [0,1]) of the segmentation. By going through the interval [0,1], a filtration is constructed on which Betti numbers are calculated at each step, this way a Persistence Diagram is drawn. A loss function (insert loss functions here) is then used to fine tune a pre-trained U-Net. This loss function is designed to be at a minimum when the topology of the desired object is reached (encoded through the persistence diagrams).

Upon reviewing the above papers, the main differences between the research presented here and the current state of knowledge in the field is two-fold:

1. Most research that involves PH to aid in segmentation of medical images is done by adding a topological constraint to or modifying the loss function used to train a deep learning model. This means that not much research has been performed in combining traditional image processing techniques with PH, as done in this work.
2. When PH is applied to perform the segmentation without using Deep Learning, it has not been applied to EM images, mainly being used on Whole Slide Images. This indicates there is a gap in applying these techniques to an imageset such as the one used here.

Chapter 4

Materials and Methods

4.1 Dataset

4.1.1 Cell Preparation and Image Acquisition

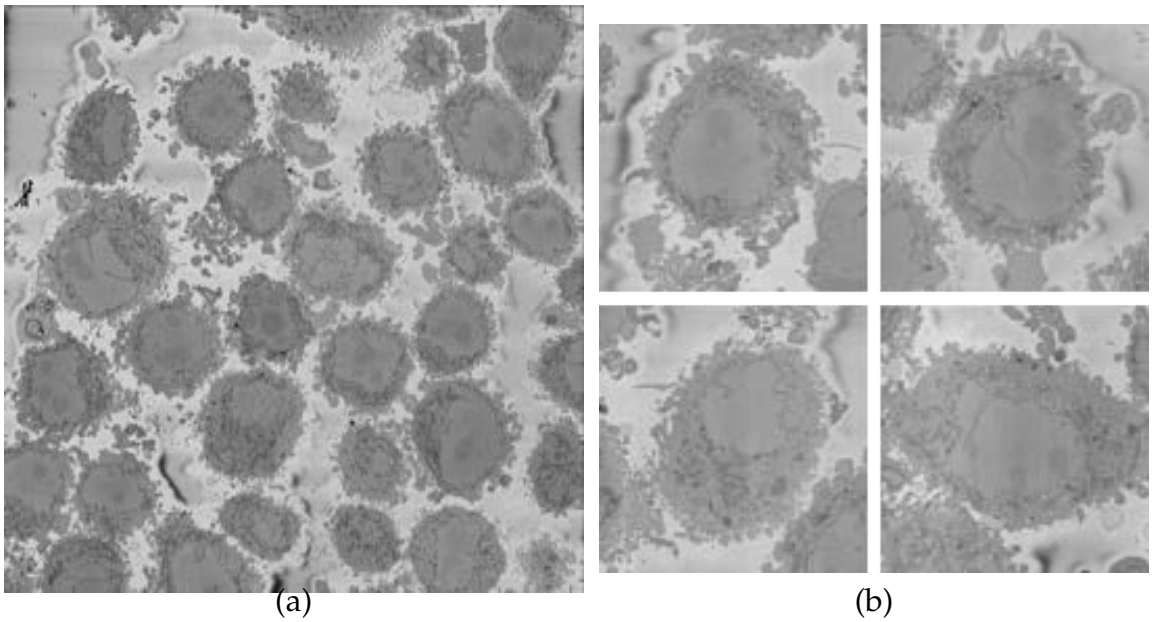
The HeLa cells were prepared through the process explained in [31]. The full details of this process are beyond the scope of this thesis.

Once prepared, the cells were embedded in Durcupan resin. The actual image data were collected using a 3View2XP (Gatan, Pleasanton, CA) attached to a Sigma VP SEM (Zeiss, Cambridge). Each time an image was taken from above, the sample was sliced. In total 517 2D slices were captured with a resolution of 10nm and a slice height of 50nm, yielding a final voxel size of 10nm $\times$ 10nm $\times$ 50nm. The final volume of data is 8-bit, meaning that voxel intensity is in the range of $[0, 255]$.

Finally, Regions of Interest (ROIs) were created by manually selecting the center of the ROI on the centroid of a cell. The final size of these ROIs are 2000 $\times$ 2000 $\times$ 300 voxels. 25 different ROIs were created this way, it is on these volumes that the work presented in this thesis is performed.

The resulting images have a high resolution, making visible the cellular and sub-cellular structure. However, a there is a trade-off for the high resolution - the contrast of the images is low and the capture process only yields one channel (grayscale images). Herein lies the challenge of segmenting EM images. The low contrast coupled with the high complexity and size of the images means that

traditional image processing methods generally fail when performing segmentation. Figure 4.1 shows a representative slice of the original 8000×8000 volume (a) together with a representative slice from four different ROIs (b).



(a) (b)

Figure 4.1: (a) A representative slice of the original 8000×8000 volume of cells. (b) A representative slice of each of four different ROIs.

4.1.2 Ground Truth of Benchmark Set

Five slices were chosen to form a benchmark set of images. It is these images that the algorithms presented in the following chapter will be trained on and tested. The Ground Truth (GT) of the pixels that correspond to the mitochondria was not readily-available. Thus, the GT the main author of this document (D.B.-P) performed the segmentation manually by delineating the perimeter of the mitochondria in each slice. For comparison, one of the supervising researchers (C.C.R.-A) performed a second segmentation blind to the first. This second segmentation will also be useful when presenting the Inter-Observer Jaccard Index. Figure 4.2 shows the five slices manually segmented by the researcher.

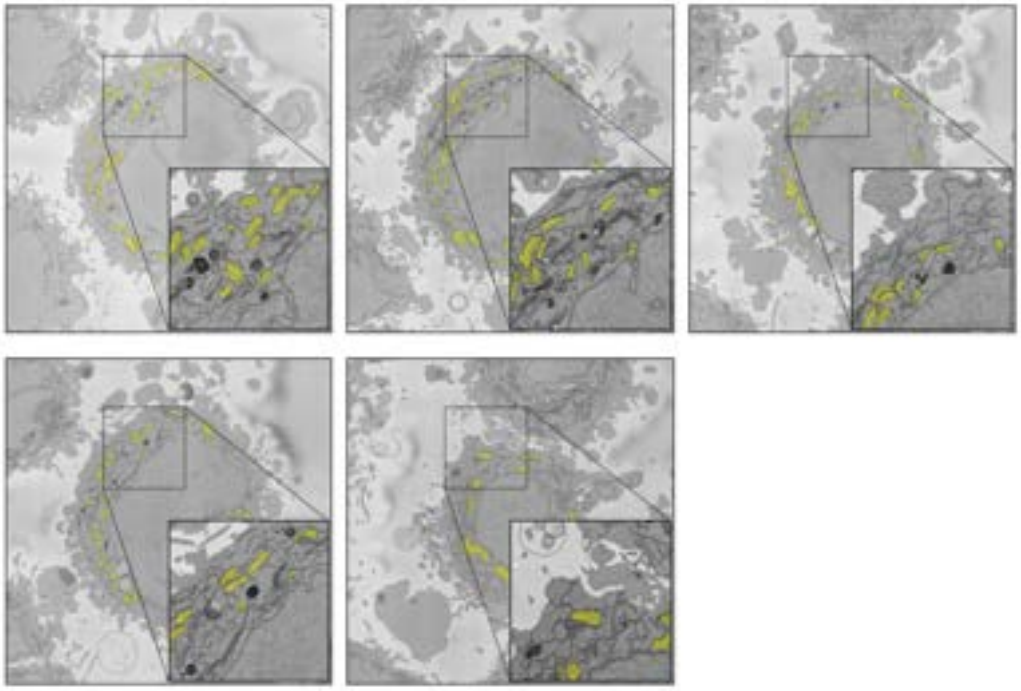


Figure 4.2: Mitochondria (yellow) were manually segmented by delineating their perimeter to create a Ground Truth for the five benchmark images. The Regions of interest (ROI)'s have been darkened to improve contrast.

4.2 Methods

As mentioned previously, the general objective of the thesis is to develop a segmentation algorithm based on Persistent Homology for the afformentioned HeLa cell dataset. For this, many previous works were analysed and explained in Chapter 3. The main contribution of this thesis builds on the work of Qaiser et al. in [4]. In this section the methodology to build this algorithm is presented. It is obvious that the algorithm that is built needs to be tested and compared against the state of the art. The current state of the art is represented by a Deep Learning model known as MitoNet. The details on MitoNet are also explained in this section.

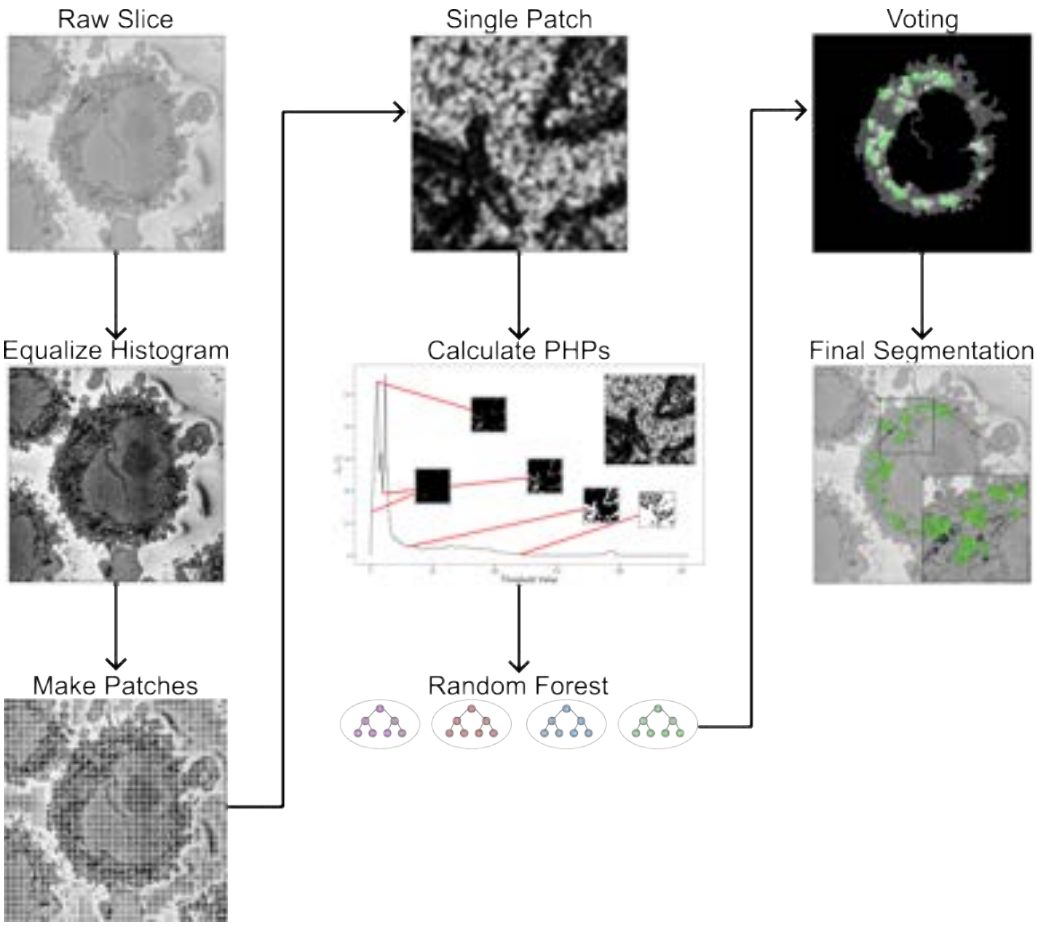


Figure 4.3: A step-by-step overview of the Topological Algorithm used for the segmentation.

4.2.1 Persistent Homology Algorithm

The work in this section is based on [4]. This algorithm was originally created to perform segmentation of stained histology Whole Slide Images. Specifically, the images show colorectal tissue where a tumor is present. The algorithm segments the part that shows the tumor from the rest of the image. This algorithm was taken and slightly modified to perform an initial segmentation of the mitochondria in the HeLa dataset. An overview of the algorithm is shown in Figure 4.3. The algorithm is described in detail following subsections.

Histogram Equalization

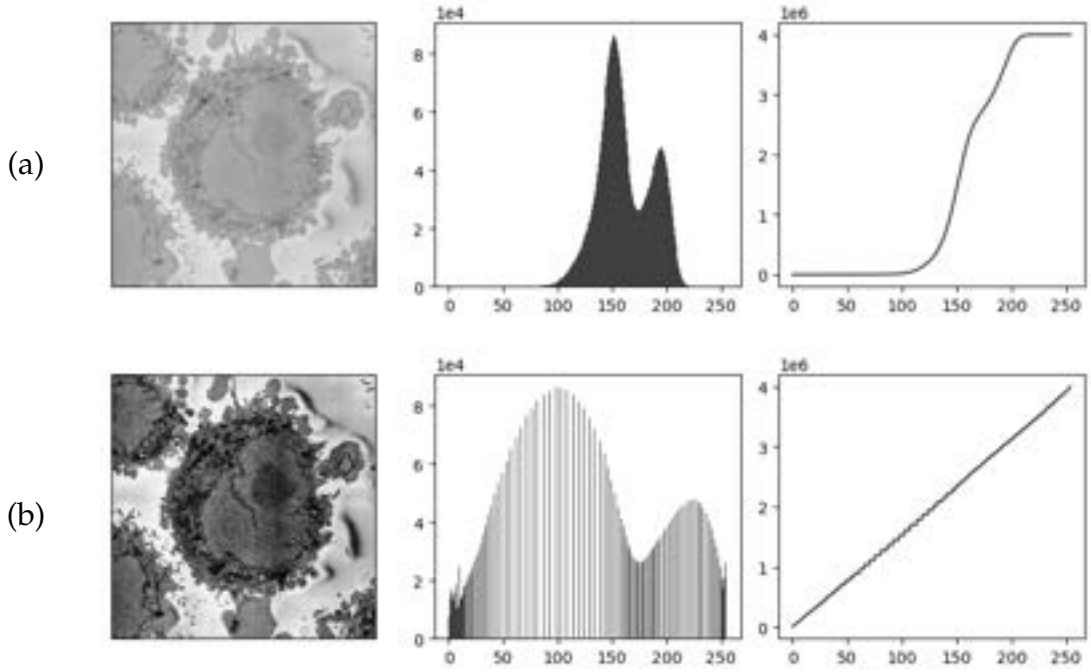


Figure 4.4: (a) Left to right: original example slice of the image volume, histogram of the original sample slice, cumulative sum through the bins of the histogram. (b) Example slice after applying histogram equalization, histogram of the sample slice after applying histogram equalization, cumulative sum through the bins of the histogram. Notice that the histogram is more spread out after applying equalization and the cumulative sum approaches a straight line.

The first step in the Persistent Homology algorithm is to apply Histogram Equalization. Histogram Equalization is a technique employed to improve the global contrast in images [32]. The basic idea is that, generally, images with low contrast have a brightness histogram that is concentrated towards a small sub-interval of the whole possible brightness interval - in 8-bit images this interval is $[0, 255]$. Thus, to increase contrast in an 8-bit image, it is necessary to spread out the histogram over the whole $[0, 255]$ interval.

In order to perform histogram equalization, a transformation is needed of the

form:

$$T : [0, 255] \to [0, 255]$$
$$T(r_k) = s_k,$$

where r_k denotes a brightness value in the interval $[0, 255]$ for the original image and s_k denotes the resulting value after applying the transformation to obtain the new image. Let $p(r_k)$ denote the normalized histogram of the original image evaluated at a given brightness value r_k. This is given by:

$$p(r_k) = \frac{n_k}{MN},$$

where n_k is the total number of pixels with value r_k in the original image, and M, N are the dimensions of the image. The transformation T that is applied for histogram equalization is:

$$s_k = T(r_k) = (255) \sum_{j=0}^{k} p(r_j). \tag{4.1}$$

Essentially, this transformation simply is the product of the number of bins in the initial histogram by the cumulative probability function up to the bin to be transformed (r_k). The transformation makes the cumulative sum through the bins approximate a straight line, as can be seen in Figure 4.2.1. Figure 4.2.1 (a) shows the original image and its histogram previous to equalization, while Figure 4.2.1 (b) shows the image after applying the transformation.

It is important to note that the transformation in 4.1 can be expanded to work on images of a different bit size B by changing replacing the 255 factor to $(2^B - 1)$. For example, for 4-bit images, the transformation would look like

$$T(r_k) = (15) \sum_{j} p(r_j).$$

Patch Generation

A single 2D slice of the image is split into patches of small size. This can easily be done through the patchify package in Python. The slice is split into patches of side-lengths 30, 40, 50, 60, 75, and 90 pixels with a 30% overlap. The patches are kept as different sets (one per each side length). The following steps of the algorithm are performed on these patches.

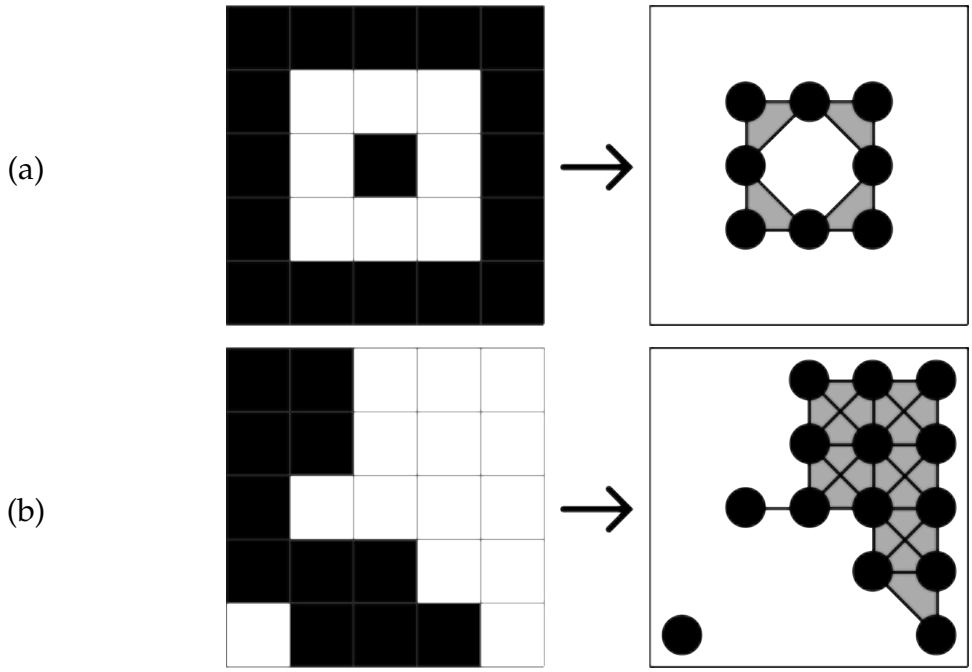


Figure 4.5: Left: sample binary images. Right: Simplicial complex obtained from the corresponding binary image on the left. (a) A white region with a black hole in the middle will yield a simplicial complex with a similar 1-dimensional hole in the middle. (b) Two separate white regions yield two different (disjoint) connected components in the simplicial complex.

Filtration

In order to calculate PH, a filtration must first be built. This concept is explained in detail in Chapter 2, but briefly remember that a filtration of a topological space S is a sequence of subspaces S_k such that:

$$S_1 \subseteq S_2 \subseteq ... \subseteq S_n = S.$$

To make a simplicial complex from a binary image, the following procedure should be followed: place a vertex at each white pixel, an edge between two neighboring pixels (with 2-connectivity), and a triangular face when three vertices a pairwise connected (see Figure 4.2.1). Let S be the simplicial complex made by following the above from a completely white image. It is clear that such a simplicial complex will have no 1-dimensional holes, and be composed of a

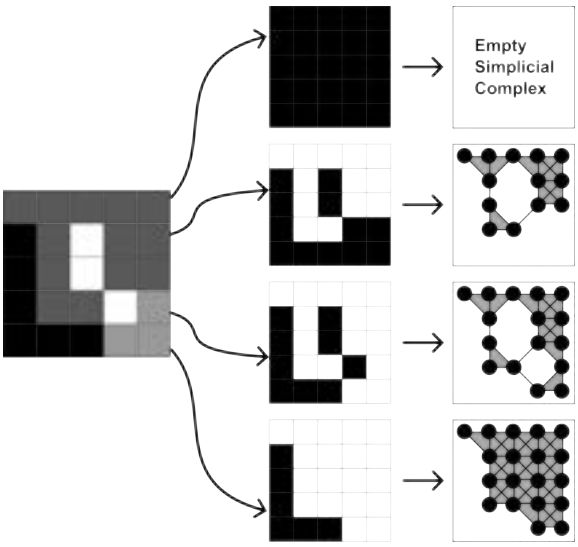


Figure 4.6: From top to bottom, different thresholds are applied to the synthetic grayscale image on the left, yielding the four different binary images (middle column). Simplicial complexes (right column) are obtained from the binary images. It is clear that these simplicial complexes satisfy the condition $S_k \subseteq S_{k+1}$.

single connected component (i.e $\beta_1 = 0, \beta_0 = 1$). A filtration of S is created by applying a lower threshold to a single patch for each brightness value t in the range $[0, 255]$, this way pixels that have a value $p \leq t$ are made white, they are made black otherwise. This thresholding procedure yields 256 binary images per patch.

On each of the binary images, a simplicial complex S_t is defined in the same way as S - a vertex is placed at each white pixel, neighboring white pixels are connected by an edge and when three vertices are connected, a triangular face is placed. This way, as the threshold value increases, black pixels transition to white, creating new vertices, adding new edges, and filling in holes. This ensures that for each threshold value t, the resulting simplicial complexes S_t adhere to the condition of the filtration $S_t \subseteq S_{t+1}$ (see Figure 4.6).

Persistent Homology Profiles

Given a filtration S_t like the ones described above, let $\beta_0(t)$, $\beta_1(t)$ denote the 0-th and 1st Betti numbers calculated at the t-th step of a filtration S_t. Consider the

following functions:

$$\mathcal{B}_0 : [0,255] \to \mathbb{R}^{\geq 0} \qquad t \mapsto \frac{\beta_0(t)}{S_0(t)} \qquad\qquad \mathcal{B}_1 : [0,255] \to \mathbb{R}^{\geq 0} \qquad t \mapsto \frac{\beta_1(t)}{S_1(t)}$$

where S_0 and S_1 denote $\sum_{t\in[0,255]} \beta_0(t)$ and $\sum_{t\in[0,255]} \beta_1(t)$, respectively. These functions essentially map each t value to a probability distribution given by the Betti numbers on the filtration in question. Consider a new function $R(t) = \mathcal{B}_0/\mathcal{B}_1$. The image of this function can be thought of as a 255-dimensional vector where each entry corresponds to a value of t. The functions $\mathcal{B}_0(t), \mathcal{B}_1(t), R(t)$ are known as a Persistent Homology Profiles (PHPs). The algorithm used in this project mainly uses R. Figure 4.7 shows a visualization of profile R for a single patch.

It is important to note that to calculate β_0 and β_1 given a binary image, one must simply count white connected components and black connected components, respectively. β_0 is simply the number of connected components in the simplicial complex, which directly results from the white connected components in the image. β_1 is the number of 1-dimensional holes in the simplicial complex, which directly corresponds to the number of black connected components in the binary image.

The PHPs for patches that contain a part of at least one mitochondrion look visually different from the PHPs of patches that contain no part of a mitochondrion. These two classes will be used to train a Machine Learning model as explained in the following subsections. This can be seen in Figure 4.8

Random Forest

The PHPs can be thought of as features extracted from the patches. There is a clear reduction in dimensionality - for example, a 40×40 patch has 1600 pixels, while the PHP is simply stored as a vector with 256 entries.

Using the Ground Truth of the Benchmark set, patches containing whole mitochondrias were manually created. PHPs were extracted from these patches and were used to train a Random Forest classifier. The Random Forest classifier has been previously explained in Chapter 2. It is important to note that while the patches containing whole mitochondrias may vary in size and shape, their PHPs will always be a 255 dimensional vector, thus, the input to the model remains

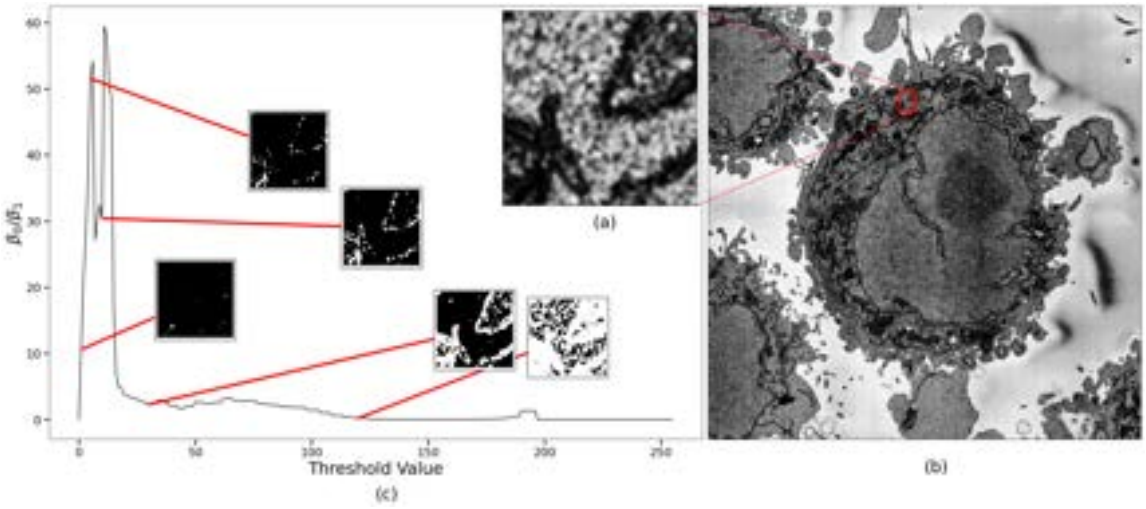


Figure 4.7: (a) Shows an example grayscale patch of the slice shown in (b). (c) Shows the Persistent Homology Profile calculated from a patch. Each binary image shown corresponds to an example threshold value.

constant. The reasoning behind using patches that contain a whole mitochondria is that preliminary tests showed that the PHPs of these patches are much more distinct from the patches which ontain no mitochondria, and more closely resemble patches which contain a part of a mitochondria.

Voting

Preliminary results revealed a problem with the algorithm. The shape of the patches (squares) meant that the segmentation could not capture the complexity and irregular shape of the mitochondria. In order to capture the curvilinear shape of mitochondria the final segmentation is done through what can be informally described as a voting mechanism. Given the possible side-lengths of the patches that the image is split up into and the overlap, it follows that each pixel could be found at most in 20 different patches. So each pixel is assigned a confidence value (or "votes") from 0 - 20. Any pixel with a value above 5 is classified as being part of a mitochondria. The goal of this process is to reduce the patches classified as containing a part of a mitochondria down to the individual pixels that form that mitochondria by finding those pixels that appear more frequently in patches classified as containing a part of a mitochondria.

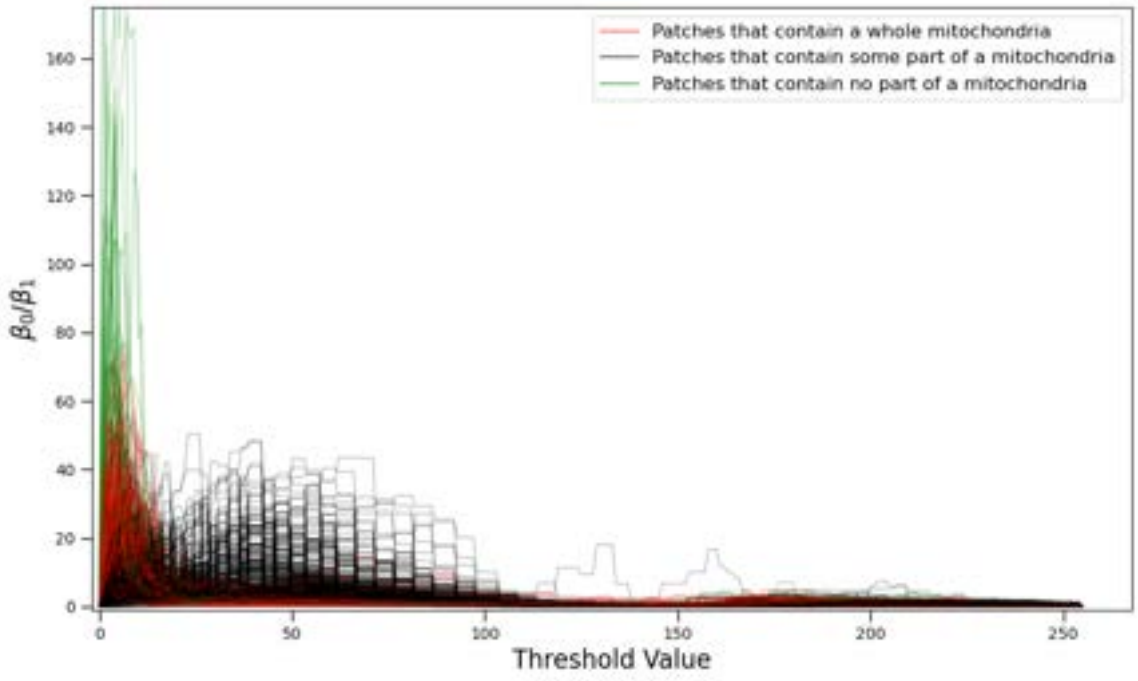


Figure 4.8: The Persistent Homology Profiles of patches that contain some part of at least one mitochondria have a visually different PHP to those patches that contain no part of a mitochondria. The red lines (mitochondria patches) tend to have their peak more concentrated towards dark threshold values (left of the graph). Note that the y axis has been cut off at 175 to make all lines more visible, as some green PHPs stretch much further above the red and black PHPs.

4.2.2 Image Processing Algorithm

An independent algorithm was developed using traditional Image Processing methods. This was done to compare results with the Persistent Homology Algorithm. As can be observed in Figure 4.1, the boundary of the mitochondria in the imageset are darker than the cytoplasm around it. This boundary also forms a closed curve (around the body of the mitochondria). The algorithm presented in this section takes advantage of these two facts to perform the segmentation as explained in the subsections below. It is important to note that mitochondria segmentation is done on a per-slice basis; this means a 2D slice is processed and regions are classified as mitochondria or not on the slice.

Finding Dark Regions

Given a 2D slice of the imageset, the first step in the Image Processing Algorithm is to find the darkest regions inside the cytoplasm. This is not simply done by thresholding as the goal is to obtain whole regions that are visibly darker than their surrounding and thresholding will yield isolated pixels which are of no use. First, define the following values:

$$I_{nuclei} := \text{average intensity of all pixels belonging to the nuclei}$$
$$I_n := n^{\text{th}} \text{ percentile of the intensity values in the cell}$$

Then, in order to find the darkest regions in the cytoplasm, the following steps describe the process to obtain the dark regions in the image. Figure 4.9 shows the step-by-step transformation of the image and highlights the dark regions.

1. Erode the cytoplasm region using a 9×9 structuring element (Fig 4.9 (a)). This is done in order to avoid darker regions caused by the cell membrane.
2. Then, the remaining region is thresholded using I_1; i.e. the darkest 1% of pixels are set to a value of 1, while the rest are set to 0. (Fig 4.9 (b))
3. A morphology operator known in MATLAB® (Mathworks™, USA) as "majority" is applied to the binary image (Fig 4.9 (c)). This operator resembles the dilation operator in that small white regions ("salt noise") are eliminated in the process, but it also has the advantage of removing small black holes inside white regions while smoothing the overall shape of the white regions. The latter is the reason why this morphology operation is applied.
4. After labelling the remaining white regions, their area is calculated and only those with area above 100 pixels are kept (Fig 4.9 (d)).
5. Finally, holes are filled (Fig 4.9 (e)) and the resulting binary image is closed using a 5×5 structuring element (Fig 4.9 (f)). The final binary image yields holeless regions without much intricate boundaries.

The regions segmented in this subsection are the darkest parts found in the cytoplasm. These segmented regions are **not** the mitochondria, but regions darker than them. Usually, the darkest areas are places where the metallic pigment used to obtain the images under EM has over-accumulated. These dark regions will aid in finding the mitochondria as, generally, the regions of over-accumulation of pigment do not touch the mitochondria.

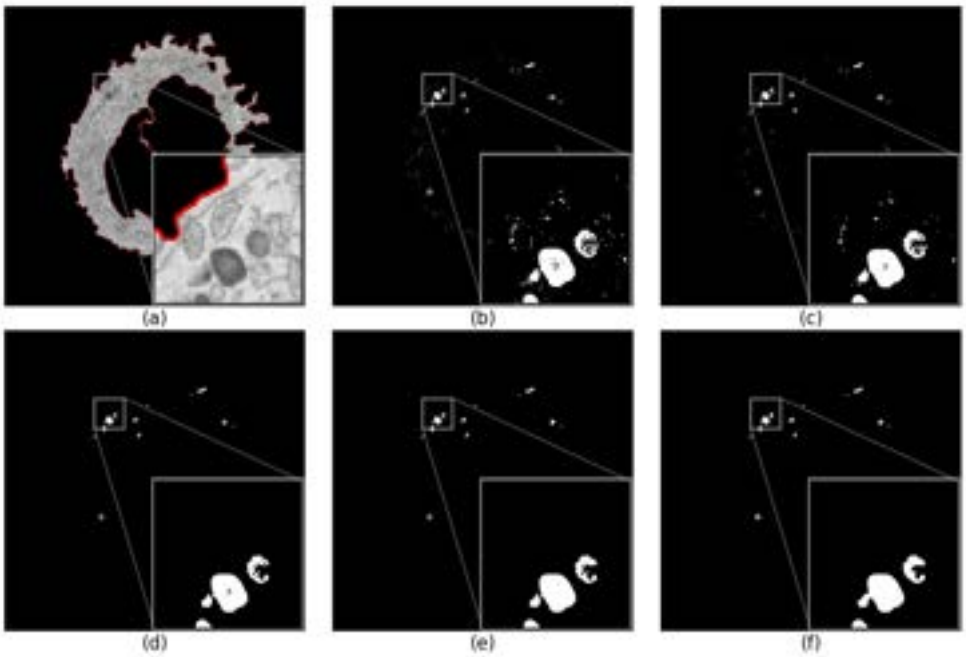


Figure 4.9: (a) Remaining cytoplasm region after erosion, and the pixels that were eroded (red). (b) Binary image after thresholding the cytoplasm region (darkest 1% pixels inside the cytoplasm). (c) Image after applying the majority morphological operator. (d) White regions larger than 100 pixels in area. (e) Binary regions after filling holes. (f) Final image after closing.

4.2.3 Finding Mitochondria

The dark regions found in the previous subsection will aid when thresholding for the mitochondria.

Finding Intermediate Regions

First, it must be shown how regions with intermediate intensity are found according to a parameter α_P:

1. The dark regions are first dilated using a 5×5 structuring element.
2. The region inside the cytoplasm is thresholded to include all intensities p such that:
$$p < \alpha_P \, I_5 + (1 - \alpha_P) \, I_{nuclei}$$
Figure 4.10 shows the effect of different α_P's.
3. Pixels that are also inside the dilated regions created in Step 1. are discarded.
4. The regions are labeled to calculate their area and, again, regions that have an area less than or equal to 100 are discarded.

For a given value of α_P, after the regions with intermediate brightness are found. The following procedure is followed to obtain a segmentation of the mitochondria:

1. The intermediate regions are thinned until they are 1-pixel wide (Figure 4.11 (a)).
2. The holes of the binary image created from step 1. are filled (Figure 4.11 (b)). The first two steps together ensure that for the next step, only closed regions will 'survive'.
3. The binary image is opened with a 3×3 structuring element (Figure 4.11 (c)). This step eliminates any region that was still 1-pixel wide after step 2., that is, all regions that didn't have a large enough hole to be filled in step 2..
4. Finally, regions with areas smaller than 500 pixels are discarded (Figure 4.11 (d)).

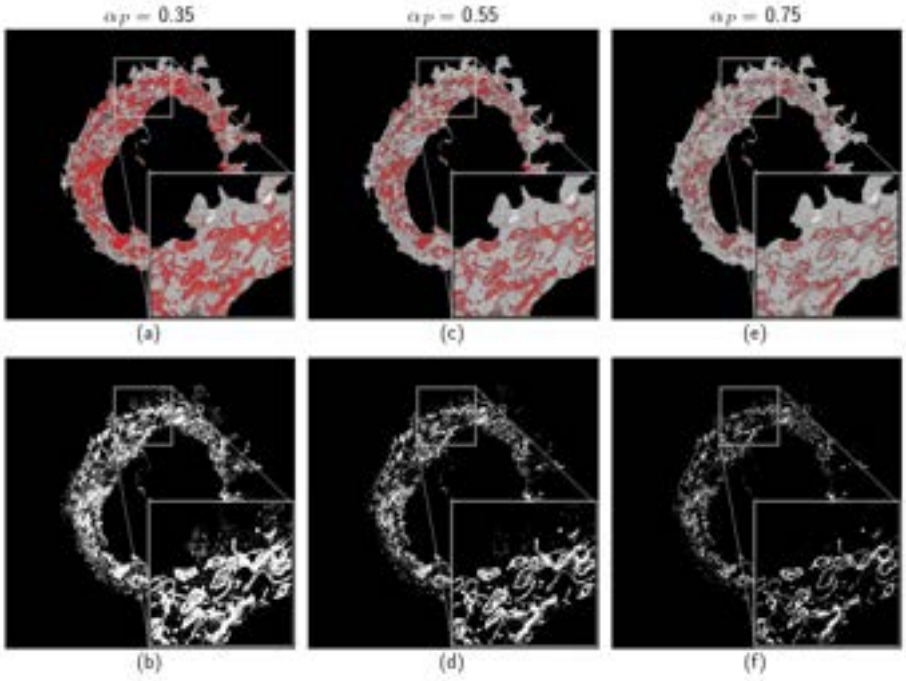


Figure 4.10: Top row: shows cytoplasm region (grayscale) and the thresholded pixels (red) overlaid on top for a given value of α_P. Bottom row: the binary images resulting from thresholding for a given α_P. (a) and (b) correspond to $\alpha_P = 0.35$, (c) and (d) correspond to $\alpha_P = 0.55$, (e) and (f) correspond to $\alpha_P = 0.75$.

For a given α_P, the above procedure yields a segmentation of the mitochondria. However, a confidence parameter is used to perform the actual final segmentation.

Final Segmentation

The confidence parameter is very similar to the voting mechanism described in Section 4.2.1. The parameter α_P takes values in $\{0.35, 0.4, 0.45, ..., 0.75, 0.8\}$. For each of these values a segmentation is generated and if a pixel shows up in two of these segmentations, then it forms part of the final segmentation.

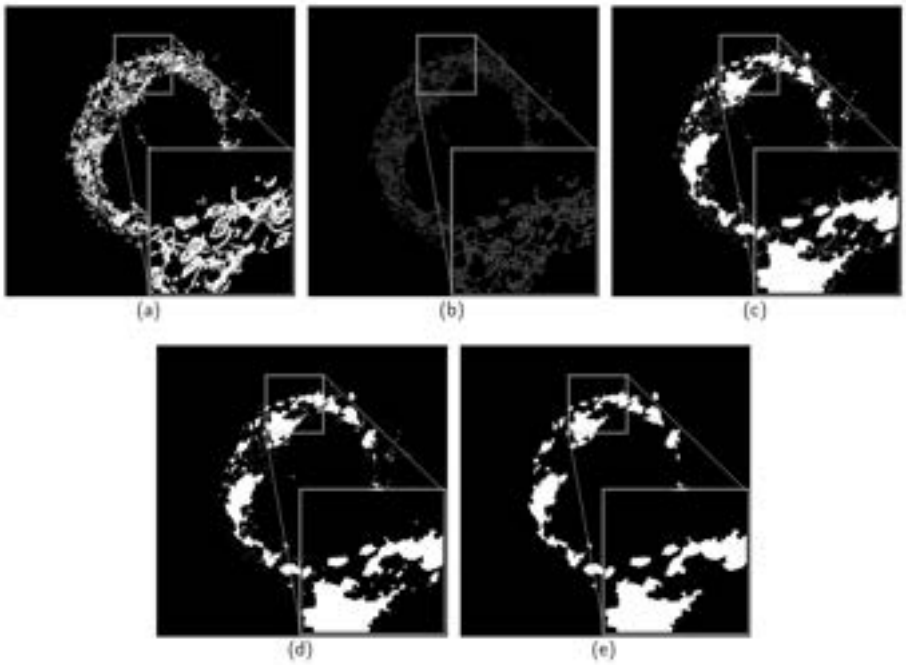


Figure 4.11: For the segmentation shown in this figure a value of $\alpha_P = 0.35$ was chosen to illustrate the segmentation procedure. (a) Thresholded cytoplasm region after discarding all regions present in the dilated dark regions and all regions with area less than 100 pixels. (b) Binary image after thinning until all regions are 1-pixel wide. (c) Image obtained after filling holes. (d) Image after opening with 3×3 structuring element. (e) Final segmentation obtained after discarding all regions with areas smaller than 500 pixels.

4.2.4 Hybrid Algorithm

The last algorithm developed in this research is a hybrid of the previous two (Topological and Image Processing). This algorithm combines both algorithms developed previously to complement each other's performance. The algorithm is described as follows:

1. For a given slice, a segmentation T is obtained with the topological algorithm and a segmentation I is obtained with the image processing algorithm. Think of T and I as sets of pixels, such that if a pixel coordinate (x, y) is in any of these two segmentations, it means that the corresponding algorithm has classified it as being part of a mitochondria.

2. The intersection $N = T \cap I$ is obtained. The final segmentation will be given by a set H which is initialized as N.

3. Take $I' = I - T$ (pixels segmented exclusively by the image processing algorithm). For every region $X \subset I'$, compare their area against adjacent regions $Y \subset H$. If the area of X is larger than 50% the area of Y, then add X to the final segmentation H.

4.2.5 MitoNet

MitoNet is a Deep Learning model based on the Panoptic-DeepLab architecture [33]. The latter model is itself based on a U-Net [22]. The MitoNet model represents the current state of the art in segmentation of EM image. It has been specifically trained to perform segmentation of organelles, in particular, mitochondria found in cells and is presented here simply because it is the model against which the previous three are compared in the following chapter, in order to evaluate their performance.

MitoNet was developed by R. Conrad and K. Narayan in [14]. As mentioned previously, it takes the architecture of Panoptic-DeepLab but trains the model on two datasets:

- CEM1.5M: contains 1,592,753 unlabeled 2D image patches. It is used to pre-train the encoder part, f_θ of the neural network using the methodology described in [34]. Essentially what is done is as follows: an image X is ran through two different augmentation steps, yielding images X_a and X_b. Each of these images is fed forward through f_θ. The output of this is two features z_a and z_b, respectively, that lay on the unit sphere. Afterwards, codes z_a and z_b are mapped to a set of trainable vectors $\{c_1, c_2, ..., c_K\}$, yielding "codes" q_a and q_b. Finally, a swapped prediction problem with an appropriate loss function is used to train the weights in f_θ. The loss function has the following shape:

  $$L(z_a, z_b) = l(z_a, q_b) + l(z_b, q_a)$$

  The idea is that if the two features z_a and z_b manage to capture the same information in the image, then q_a should be able to be predicted from z_b and vice versa.

- CEM-MitoLab: contains 21,860 annotated images. This dataset is used to fully train the network, after pre-training on CEM1.5M.

Chapter 5

Results

The general objective of this project as mentioned in Chapter 1 is to develop an algorithm based on Persistent Homology in combination with traditional image processing methods to segment the mitochondria found in the HeLa cell image-set. A more specific objective is to evaluate the performance of this algorithm in comparison with methodologies that have similar purposes, and the current state of the art in EM segmentation.

In order to achieve the goal, three segmentation algorithms have been developed, as presented in Chapter 5. These are a Persistent Homology Algorithm (4.2.1), an Image Processing Algorithm (4.2.2), and a Hybrid Algorithm (4.2.4). In Chapter 4, the most successful algorithm at the time of writing (MitoNet), was also presented.

In this Chapter, results of the segmentations of the developed algorithms and MitoNet are presented visually, by showing the segmented benchmark dataset and standard metrics are calculated and presented for each of the algorithms so that they may be compared against each other.

For each segmentation, pixel-wise accuracy, F_1-scores, and Jaccard Index were calculated. Table 5.1 shows a summary of all the metrics for the four algorithms described in Chapter 4.

5.1 Segmentation Results

5.1.1 Persistent Homology Algorithm

In order to train the Persistent Homology algorithm, the five benchmark images were split using a procedure similar to k-folds. A single image is left out of the training set, and then the algorithm is trained using the patches containing complete mitochondria from the remaining four. This is done five times (one for each image). The segmentation of each image is done by the model trained on the other four. The final segmentations of all five benchmark images using this algorithm are shown in Figure 5.1.

The pixel-wise scores are calculated for each of the five segmentations and then the average is taken. This is shown in Table 5.1.

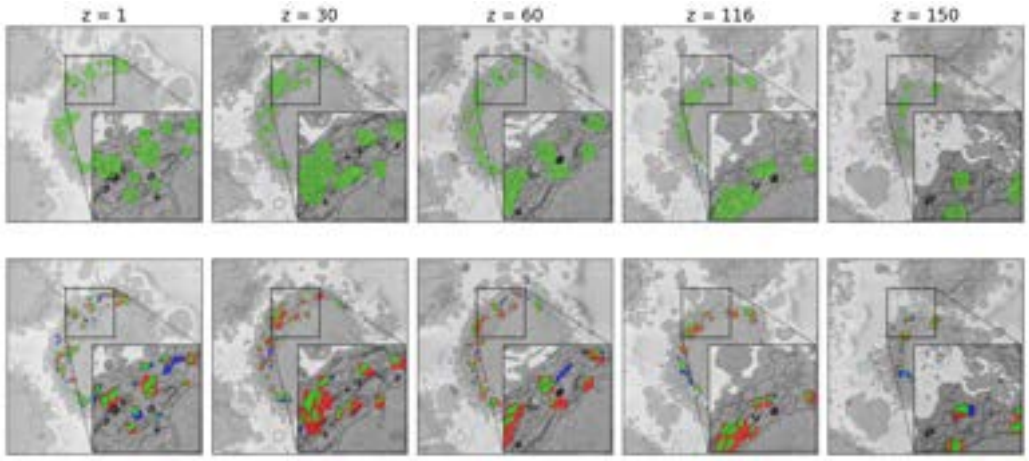


Figure 5.1: Visualization of the performance by the Persistent Homology algorithm. Top: the segmentations on the benchmark set are shown. The green regions show the pixels classified as being part of a mitochondria. A brighter green surrounds such regions, these borders do not form part of the segmentation, only the interior of the green areas. Bottom: green regions show true positives (pixels that form part of a mitochondria correctly classed as such), red regions show false positives (pixels classed as mitochondria which are not), blue regions show false negatives (pixels inside mitochondrias which were not classified as such).

5.1.2 Image Processing Algorithm

The Image Processing algorithm as described in 4 was applied to the benchmark images. The visualizations (Figure 5.2) for each image were generated and the average pixel-wise metrics were also calculated for this algorithm (Table 5.1).

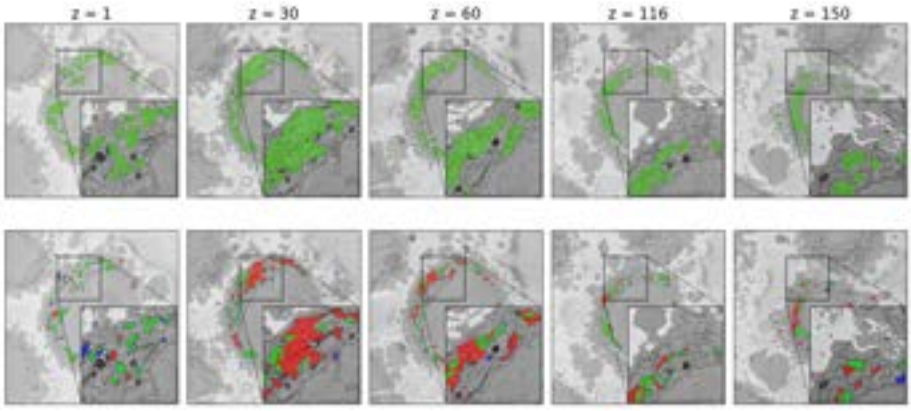


Figure 5.2: Visualization of the performance by the Image Processing algorithm. Top: the segmentations on the benchmark set are shown. The green regions show the pixels classified as being part of a mitochondria. A brighter green surrounds such regions, these borders do not form part of the segmentation, only the interior of the green areas. Bottom: green regions show true positives (pixels that form part of a mitochondria correctly classed as such), red regions show false positives (pixels classed as mitochondria which are not), blue regions show false negatives (pixels inside mitochondrias which were not classified as such).

5.1.3 Hybrid Algorithm

The Hybrid Algorithm presented in 4 was applied to the five images. Effectively, since the inputs of this algorithm are basically the segmentations of the Persistent Homology and the Image Processing algorithms, no additional training needs to be done (after training the PH algorithm). After the first two algorithms have produced a segmentation, these were used to create the visualizations for the Hybrid algorithm shown in Figure 5.3 and to calculate the average metrics shown in Table 5.1.

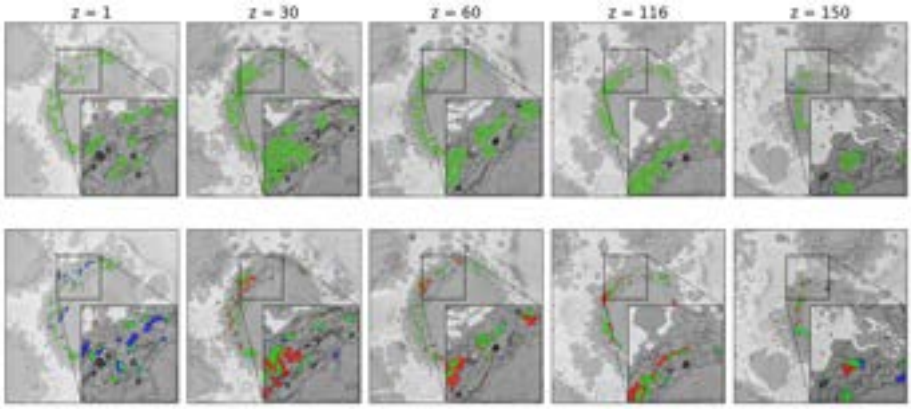


Figure 5.3: Visualization of the performance by the Hybrid algorithm. Top: the segmentations on the benchmark set are shown. The green regions show the pixels classified as being part of a mitochondria. A brighter green surrounds such regions, these borders do not form part of the segmentation, only the interior of the green areas. Bottom: green regions show true positives (pixels that form part of a mitochondria correctly classed as such), red regions show false positives (pixels classed as mitochondria which are not), blue regions show false negatives (pixels inside mitochondrias which were not classified as such).

5.1.4 MitoNet

The MitoNet model was applied individually to the 2D benchmark images. Using these segmentations, the visualizations were generated as shown in Figure 5.4 and the pixel-wise metrics shown in Table 5.1 were also calculated. These metrics, again, are calculated on each slice and the average over all five are registered.

5.1.5 Inter-Observer Score

A separate segmentation was performed by hand by one of the supervisors of the author of this thesis. The same pixel-wise metrics (accuracy, F_1-score, and Jaccard Index) were calculated so that all algorithms could be compared more fairly. The idea is that if the pixel-level detail in the segmentation varies heavily from human to human, then it is expected to vary as heavily between algorithms. That is to say that even a 'close-to-perfect' algorithm, will still present high variance when compared to any human-made ground truth segmentation. The inter-observer scores are also shown in Table 5.1.

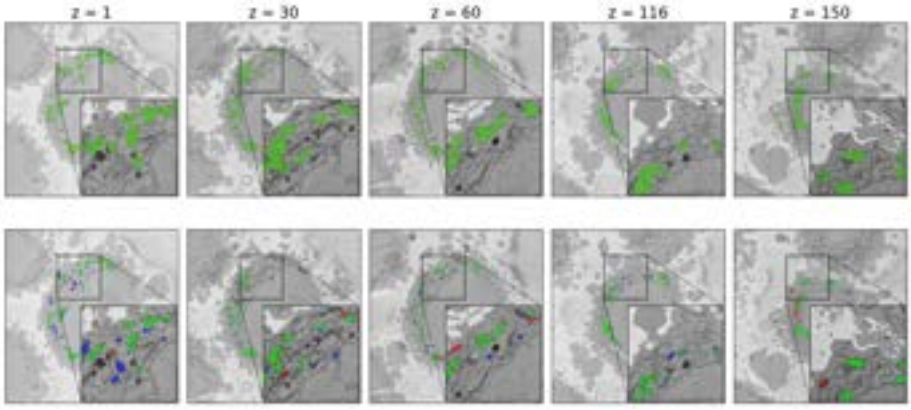


Figure 5.4: Visualization of the performance by MitoNet. Top: the segmentations on the benchmark set are shown. The green regions show the pixels classified as being part of a mitochondria. A brighter green surrounds such regions, these borders do not form part of the segmentation, only the interior of the green areas. Bottom: green regions show true positives (pixels that form part of a mitochondria correctly classed as such), red regions show false positives (pixels classed as mitochondria which are not), blue regions show false negatives (pixels inside mitochondrias which were not classified as such).

5.1.6 Comparative Analysis

Table 5.1 summarizes the pixel-wise scores obtained by each algorithm and the Inter-Observer scores. While accuracy and F_1-score are shown, the main focus of comparison should be done with the Jaccard Index (JI). This score is the standard for evaluating the performance of segmentation algorithms due to the focus on the overlap of the predicted regions with the ground truth, the balance it achieves between false positives and false negatives, and its robustness to class imbalance - in the present case, this last benefit is important since the mitochondria are small areas compared to the whole images.

The results shown on Table 5.1 indicate that the overall best performer is definitely MitoNet (Jaccard Index = 0.6595). While the Persistent Homology and Image Processing algorithms have the lowest performance (JI = 0.4109 and JI = 0.4622, respectively). However, when combined, the hybrid algorithm reaches a JI of 0.5699. The Inter-Observer JI was 0.6961.

The F_1-score of the Persistent Homology annd Image Processing algorithms were similar (0.5816 and 0.6261, respectively) while the Hybrid algorithm and MitoNet

obtained higher and similar results (0.7243, 0.7090, respectively).

These results an overview of the capabilities of the proposed algorithms, as well as the MitoNet model. The next chapter will discuss the implications of these findings and will provide a much more specific comparison between the models.

Algorithm	**Accuracy**	F_1**-score**	**Jaccard Index**
Persistent Homology	0.8945	0.5816	0.4109
Image Processing	0.9039	0.6261	0.4622
Hybrid	0.9413	0.7243	0.5699
MitoNet	0.9329	0.7090	0.6595
Inter-Observer	0.8975	0.5815	0.6961

Table 5.1: Average pixel-wise metrics obtained by each algorithm on the benchmark set of five images and the Inter-Observer scores are summarized.

Once the segmentations of the mitochondria were made, an investigation into relationships with the shape of the cell and its nucleus was performed. The methodology and results of this subsequent research are presented in Chapter 6.

Chapter 6

Morphology of HeLa Cells (nuclei and mitochondria)

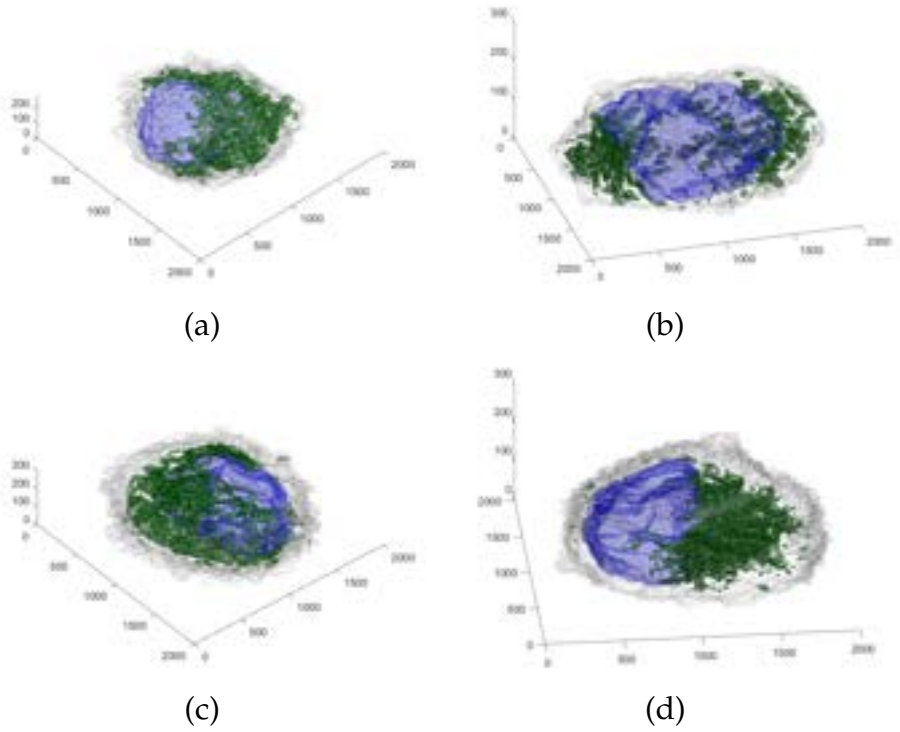
As mentioned in the previous chapter, once the segmentations of mitochondria were obtained, research into their relationship with the shape of the cell and its nucleus was done. Particularly, relationships between the size and shape of mitochondria, and the size and shape of the nucleus and cytoplasm were investigated. This chapter details the methodology behind performing this research and the results obtained from it.

6.0.1 Segmentation of Mitochondria

In order to compare the morphologies of mitochondrias with the cells they live in, the mitochondria must first be segmented. That is where the work shown in previous chapters comes in. One can think of the previous research as a selection process for the presented segmentation models. As MitoNet was the best performer, it was selected to make the final segmentation of all mitochondria in the entire HeLa imageset.

Once MitoNet yielded a segmentation, no further modification was made. It was taken as an accurate segmentation of the mitochondria. In total, 12324 mitochondria were segmented from the volume. Figure 6.1 shows samples of the final segmentation of the mitochondria. It must be highlighted that the mitochondria inside the cells are not always distributed in the same manner. With the segmentation, four different types of distributions were found. Namely, mitochondria

inside the cell can be distributed uniformly 6.1 (a), polarised towards the left and right 6.1 (b), uniformly except for a small region 6.1 (c), and concentrated towards a single side 6.1 (d). The implications or causes of the differences in distributions are beyond the scope of this thesis.



(a) (b)

(c) (d)

Figure 6.1: Four segmentations illustrate the different ways mitochondria can be distributed inside the cell. The green regions represent the segmented mitochondria while the blue region shows the nuclear envelope of the cell. The plasma membrane is shown in light gray. It can be noticed how mitochondria distribute in the cells: (a) uniform, (b) polarised towards left and right, (c) uniform except for a small region, (c) concentrated towards a single side.

6.0.2 Segmentation of Nuclear Invaginations

Invaginations of the nuclear envelope are deformations in the nuclear envelope, which usually is a smooth continuous shape. These deformations can become like large canyons on the otherwise ellipsoidal shape of the nucleus. Invagina-

tions have previously been linked with cancer and can also be used for diagnosis and prognosis in some cases [35].

The invaginations in the nucleus were segmented slice-by-slice using the following steps:

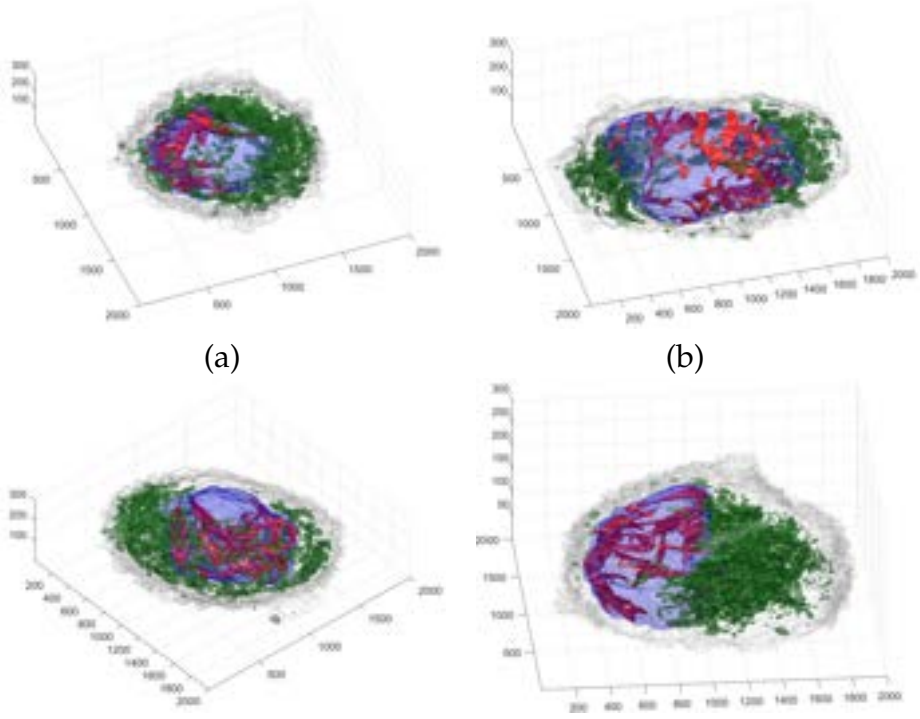
1. First, a slice containing a binary segmentation of the nucleus was filled for holes, in case the segmentation presented a hole in the given slice. This yields a binary image called N_1.
2. The image N_1 is closed using a rather large structuring element (55×55 disk) in order to find any invagination protruding in from the edge of the nucleus, leaving an image named N_2.
3. By eroding N_2 with a 9×9 square structuring element, the region is made smaller, yielding N_3.
4. The regions where N_3 has pixel values greater than N_1 (i.e. regions where N_3 is 1 and N_1 is 0), are large invaginations not close to the surface.
5. Small noise is removed by using a 2×2 disk structural element.

Once all slices are processed by following the above algorithm, all regions below a volume of 15,000 voxels are discarded. Leaving a final segmentation of the invaginations as shown in Figure 6.2 taken from [36].

6.0.3 Morphology Metrics

All segmentations obtained as described previously are saved as 2D images and assembled into 3D volumes. Using MATLAB, metrics about the morphology of each cell's nucleus, cytoplasm and their mitochondria were computed. The following is a list of all the metrics that were calculated for each cell, all in voxels:

- Volume of the cytoplasm: computed as the total volume contained within the plasma membrane minus the total volume of the cell's nucleus.
- Total volume of the invaginations: calculated as the sum of the volumes of all invaginations found using the procedure in 6.0.2.
- Total number of mitochondria: total count of uniquely labeled regions in the stack of segmentations produced by MitoNet.



(a) (b) (c) (d)

Figure 6.2: The same four Regions of Interest as in Figure 6.1 are shown. The nuclear invaginations are displayed in red, while the mitochondria, nuclear envelope, and plasma membrane are shown in green, blue, and light gray, respectively. Figure taken from [36].

- Total volume of mitochondria: it is the total sum of the volumes of all uniquely labeled regions produced by MitoNet.
- Average volume of mitochondria: given by the ratio of total volume of mitochondria and the total number of mitochondria.
- Average aspect ratio of mitochondria: the aspect ratio of a single mitochondrion is given by the ratio of its major axis and its minor axis. The average aspect ratio of all mitochondria in a cell is simply the average of all aspect ratios calculated for each mitochondrion.

The morphological metrics were taken in pairs and a Pearson correlation coefficient [37] was calculated for each pair. The most important findings in these correlations are the following:

1. A positive ($r = 0.5067$) and significant ($p = 0.0097$) correlation between the total volume of invaginations and the total volume of mitochondria. Whilst this could suggest that larger and more complex invaginations were associated with more mitochondria, the positive correlation may just be an indication of the size of the cell.
2. A negative ($r = -0.4466$) and significant ($p = 0.0252$) correlation between the number of mitochondria and the average volume of mitochondria was found. This suggests that the more mitochondria are present in a cell, the smaller they are. This correlation would not be affected by the size of the cell.
3. A negative ($r = -0.4407$) and significant ($p = 0.0275$) correlation between the volume of the cytoplasm and the aspect ratio of the mitochondria, suggesting that the larger the cytoplasm, the thinner and more elongated the mitochondria.

Chapter 7

Discussion

Recall that the main objective of this study was to develop an algorithm based on Persistent Homology in combination with traditional image processing methods to segment the mitochondria found in the HeLa cell imageset and evaluate its performance. It is with this goal in mind that in the previous chapters, namely Chapter 4 and Chapter 5, three different models for performing segmentation of EM images of HeLa Cells were designed and tested for performance. The MitoNet model was also presented as the state of the art and Inter-Observer segmentation metrics were also calculated for comparison against the developed algorithms.

Before continuing it is important to note what the Inter-Observer (IO) scores showed. The Jaccard Index obtained for the IO segmentations was of 0.6961. It is a score quite far away from the "perfect" 1. This implies that there is significant differences between segmentations made even by human researchers. Thus, the segmentations obtained via automatic algorithms shouldn't be judged on the usual 0-1 Jaccard Index scale, and instead be compared against the more lenient IO score of 0.6961.

With the previous point in mind, the results showed, confidently, that MitoNet is the best performer when segmenting mitochondria. This is shown by a considerably higher Jaccard Index than the other three algorithms of 0.6595. This score is remarkably close to the IO score of 0.6961.

The second highest performer is the Hybrid algorithm which combines the Persistent Homology and Image Processing algorithms. With a Jaccard Index of

0.5699, it is not extremely far away from the one calculated with the IO segmentations. Moreover, the average F_1-score achieved by the Hybrid algorithm was higher than the one achieved by the MitoNet model. This could be due to the F_1-score generally putting more weight on the True Positive pixels (mitochondria pixels correctly classified as such) than on False Positive or False Negative pixels. This means that the Jaccard Index generally penalizes incorrect pixels more. Given that mitochondria are small areas of the cell, a higher penalization of falsely predicted pixels yields a lower score.

The previous point means that, while MitoNet might have obtained a higher Jaccard Index, and this is a more commonly used metric to gauge performance, the hybrid algorithm is more accurate at predicting True Positive pixels (covering the whole mitochondria), but ends up over-predicting pixels that are not meant to be Positive. It is also important to note that the sample size used to gauge the performance of these segmentations is relatively small as only five images were used.

A notable point to be taken from Table 5.1 is that the PH and IP algorithms complement each other well. The Hybrid algorithm achieved considerably higher scores than either of the previous two on their own.

It must now be acknowledged that the Persistent Homology algorithm required very minimal training (only four slices were used for each segmentation) and the Image Processing algorithm uses traditional techniques and thus requires no machine learning. With the Hybrid algorithm being essentially an extension of the combination of both these methods, the Hybrid algorithm shows promise in the computational complexity front when compared to MitoNet. The reader is reminded that MitoNet required over 1.5 million images for pre-training and training together. The fact that MitoNet is based on a U-net architecture, also means that prediction on the images is generally slower than any of the three algorithms developed in this study.

7.0.1 Future Work

With the previous points having been mentioned, the following are suggestions to expand upon the research in the future:

1. **Improve the Hybrid Algorithm:** the most obvious way to continue this line of research is to improve upon the hybrid algorithm. The facts that the average F_1-score was higher than MitoNet's and that its Jaccard is not

far from the one obtained from the IO segmentations is encouraging to keep working towards a better way to combine Persistent Homology with traditional Image Processing techniques.

2. **Test on larger image set:** as mentioned previously, the models were tested on a benchmark set of five images. A natural idea to extend the research is to test on more 2D slices. This would yield a clearer picture on the performance of each model.

Even though it was not the main goal of the project, the findings obtained in Chapter 6 must also be discussed. The first result (positive correlation between total volume of invaginations and total volume of mitochondria) may suggest that larger invaginations are associated with more (or larger) mitochondria, but, as mentioned, this correlation may just be an indication of the size of the cell. With the given sample size, research that accounts for the size of the cell must be performed in order to discard this possibility. A simple calculation for average sizes did not yield a significant correlation. This implies that the correlation between total volumes might not be due to the size of the cell.

The second result (negative correlation between number of mitochondria and average volume of them) is more interesting. The initial conclusion is that the more mitochondria are present in a cell, the less space there is for each - thus, meaning that mitochondria cannot grow as freely and are restricted to stay small. However, mitochondria can merge between them as time passes [38], so the cells with more mitochondria may just be younger cells (before they have time enough to merge).

Finally, the last result is the most interesting to the author of this thesis (negative correlation between the volume of cytoplasm and aspect ratio of the mitochondria) as this implies a direct correlation between the size of the cytoplasm and the shape of the mitochondria. The larger the cytoplasm, the more elongated the mitochondria. Further research accounting for the shape and size of the cell is needed.

7.0.2 Conclusion

To conclude, the study presented in this thesis aimed to create and evaluate an image segmentation model based on Persistent Homology and Image Processing techniques tailored towards segmenting mitochondria from images of HeLa Cells taken with an Electron Microscope. Three different algorithms were created for

this task: an algorithm based on Persistent Homology and Machine Learning, an algorithm using solely well-known Image Processing techniques, and a final hybrid algorithm which combines the segmentations of the previous two.

Segmentations of five slices of the image set were generated using the previously mentioned algorithms. Two additional segmentations were created - one by the MitoNet deep learning model, and another by a separate human researcher. The former, represents the current state of the art in medical image segmentation, the latter is used to establish a baseline for comparing all the automatic segmentations.

The models created in this study (PH, IP, and Hybrid) performed more poorly than the MitoNet segmentation model according to the average Jaccard Index. However, the combination of PH and IP using the Hybrid algorithm achieved significantly better results than the first two algorithms by themselves. This, combined with the fact that the PH model requires minimal resources to train, is encouraging to keep working on this topic and advance the project further.

A further investigation using the best-performer, MitoNet into the morphology of the cells, the nucleus, and the mitochondria yielded interesting results which also open the door for further research.

Bibliography

[1] G. Migdałek and M. Żelawski, "Measuring population-level plant gene flow with topological data analysis," *Ecological Informatics*, vol. 70, p. 101 740, 2022.

[2] T. Teramoto *et al.*, "Computer-aided classification of hepatocellular ballooning in liver biopsies from patients with nash using persistent homology," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105 614, 2020.

[3] P. Cheng *et al.*, "Automatically recognize and segment morphological features of the 3d vertebra based on topological data analysis," *Computers in Biology and Medicine*, vol. 149, p. 106 031, 2022.

[4] T. Qaiser *et al.*, "Persistent homology for fast tumor segmentation in whole slide histology images," *Procedia Computer Science*, vol. 90, pp. 119–124, 2016.

[5] R. Rojas-Moraleda *et al.*, "Robust detection and segmentation of cell nuclei in biomedical images based on a computational topology framework," *Medical image analysis*, vol. 38, pp. 90–103, 2017.

[6] R. W. Taylor and T. D. W., "Mitochondrial DNA mutations in human disease," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 389–402, 2005.

[7] B. Van Houten *et al.*, "Mitochondrial DNA damage induced autophagy, cell death, and disease," *Frontiers in bioscience (Landmark edition)*, vol. 21, p. 42, 2016.

[8] S. Vyas *et al.*, "Mitochondria and cancer," *Cell*, vol. 166, no. 3, pp. 555–566, 2016.

[9] R. B. Soledad *et al.*, "The secret messages between mitochondria and nucleus in muscle cell biology," *Archives of biochemistry and biophysics*, vol. 666, pp. 52–62, 2019.

[10] R. Delsite *et al.*, "Nuclear genes involved in mitochondria-to-nucleus communication in breast cancer cells," *Molecular cancer*, vol. 1, pp. 1–10, 2002.

[11] G. Amuthan *et al.*, "Mitochondria-to-nucleus stress signaling induces phenotypic changes, tumor progression and cell invasion," *The EMBO journal*, vol. 20, no. 8, pp. 1910–1920, 2001.

[12] R. Rahbari *et al.*, "A novel l1 retrotransposon marker for hela cell line identification," *Biotechniques*, vol. 46, no. 4, pp. 277–284, 2009.

[13] J. R. Masters, "Hela cells 50 years on: The good, the bad and the ugly," *Nature Reviews Cancer*, vol. 2, no. 4, pp. 315–319, 2002.

[14] R. Conrad and K. Narayan, "Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset," *Cell Systems*, vol. 14, no. 1, pp. 58–71, 2023.

[15] M. D. Brand *et al.*, "The role of mitochondrial function and cellular bioenergetics in ageing and disease," vol. 169, no. s2, pp. 1–8, 2013.

[16] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 44, no. 1.2, pp. 206–226, 2000.

[17] T. Hastie *et al.*, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[18] D. Kaur and Y. Kaur, "Various image segmentation techniques: A review," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809–814, 2014.

[19] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 441–450.

[20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] D. Steinley, "K-means clustering: A half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[22] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[23] U. Fugacci *et al.*, "Persistent homology: A step-by-step introduction for newcomers.," in *STAG*, 2016, pp. 1–10.

[24] H. Edelsbrunner *et al.*, "Persistent homology-a survey," *Contemporary mathematics*, vol. 453, no. 26, pp. 257–282, 2008.

[25] L. Wasserman, "Topological data analysis," *Annual Review of Statistics and Its Application*, vol. 5, pp. 501–532, 2018.

[26] S. Keele *et al.*, *Guidelines for performing systematic literature reviews in software engineering*, 2007.

[27] M. Haft-Javaherian *et al.*, "A topological encoding convolutional neural network for segmentation of 3d multiphoton images of brain vasculature using persistent homology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 990–991.

[28] R. Rojas-Moraleda *et al.*, "Segmentation of biomedical images based on a computational topology framework," *Seminars in Immunology*, vol. 48, p. 101 432, 2020, The Tumor Microenvironment: prognostic and theranostic impact. Recent advances and trends, ISSN: 1044-5323. DOI: https://doi.org/10.1016/j.smim.2020.101432. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1044532320300488.

[29] S. Sun *et al.*, "Topology-sensitive weighting model for myocardial segmentation," *Computers in Biology and Medicine*, vol. 165, p. 107 286, 2023.

[30] S. Y. Shin *et al.*, "Deep small bowel segmentation with cylindrical topological constraints," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, Springer, 2020, pp. 207–215.

[31] M. R. G. Russell *et al.*, "3d correlative light and electron microscopy of cultured cells using serial blockface scanning electron microscopy," *Journal of Cell Science*, vol. 130, no. 1, pp. 278–291, 2017.

[32] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009, pp. 133–140.

[33] B. Cheng *et al.*, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 475–12 485.

[34] M. Caron *et al.*, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[35] A. N. Malhas and D. J. Vaux, "Nuclear envelope invaginations and cancer," *Cancer Biology and the Nuclear Envelope: Recent Advances May Elucidate Past Paradoxes*, pp. 523–535, 2014.

[36] D. A. Brito-Pacheco *et al.*, "Relationship between irregularities of the nuclear envelope and mitochondria in hela cells observed with electron microscopy," *bioRxiv*, pp. 2023–11, 2023.

[37] W. M. Mendenhall and T. L. Sincich, "Statistics for engineering and the sciences," in CRC Press, Taylor & Francis Group, 2016, pp. 513–516.

[38] B. Westermann, "Merging mitochondria matters," *EMBO reports*, vol. 3, no. 6, pp. 527–531, 2002.