Machine Learning Algorithm for the Analysis of Cardiac Tissue Cross Sections

Technical Report
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Biomedical Engineering

By
Godwin Oluwafemi

April 12, 2024

Technical Team Members: Marcus Elward

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Timothy Allen, Professor, Department of Biomedical Engineering

Jeffrey Saucerman, Professor, Department of Biomedical Engineering

Machine Learning Algorithm for the Analysis of Cardiac Tissue Cross Sections

Marcus S. Elward[a,b], Godwin Oluwafemi[a,b], Alice Luanpaisanon[a], Jeffrey Saucerman[a,c,1]

[a] Department of Biomedical Engineering, University of Virginia
[b] These authors contributed equally to this work
[c] Senior Advisor for the project
[1] Correspondence: jsaucerman@virginia.edu, MR5 Room 2111, (434) 924-5095

Abstract

Hypertrophic cardiomyopathy (HCM) is the hypertrophy, or cell growth, of cardiomyocytes (CMCs), or heart muscle cells, in response to damage and cell death within the heart. The hypertrophy of CMCs, while strengthening the heart contractions to compensate for the lower number of present CMCs, alters the heart's architecture and leads to other health complications, such as diastolic dysfunction, obstructions, and arrhythmias. The Wolf Lab at the University of Virginia is currently developing drugs to treat HCM by both ameliorating CMC hypertrophy and promoting CMC proliferation to replenish the cells that died over the course of the injury. To test the efficacy of drug candidates, the Wolf Lab performs in vivo testing on mice which have been genetically modified to have HCM. They then take tissue slices from the mouse hearts, stain them, and examine the cardiac tissue under a microscope. The current method for quantifying hypertrophy employed by the Wolf Lab is to select 30 cells at random and measure them by hand. This process is time consuming and is vulnerable to bias and human error. Our capstone project developed a machine learning algorithm approach to objectively segment and quantify hypertrophy and nuclear content of CMCs in the tissue slice images. The algorithm will save researchers time by allowing for high throughput testing of HCM drugs while also providing more data points and greater objectivity to the measurements taken.

Keywords: Machine learning, hypertrophic cardiomyopathy, hypertrophy, DNA content, drug screening

Introduction

Hypertrophic Cardiomyopathy (HCM) is a cardiovascular condition characterized by abnormal thickening (hypertrophy) of the heart muscle, particularly the left ventricle. This thickening makes it harder for the heart to pump blood effectively, leading to symptoms such as shortness of breath, chest pain, fatigue, and in some cases, fainting or sudden cardiac arrest. HCM is often inherited, although it can also develop due to other factors. Diagnosis typically involves a combination of medical history, physical examination, imaging tests such as echocardiography, and genetic testing. Management of HCM aims to alleviate symptoms, prevent complications, and reduce the risk of sudden cardiac death through medications, lifestyle modifications, and in some cases, surgical interventions like septal myectomy or alcohol septal ablation. Long-term monitoring and regular follow-ups are essential for individuals with HCM to manage their condition effectively and minimize its impact on their quality of life[1].

Quantifying Hypertrophic Cardiomyopathy (HCM) in cells involves measuring and assessing various cellular characteristics, such as structural changes, gene expression profiles, or protein levels, that are associated with the presence or progression of HCM[2]. This process aims to systematically analyze and quantify the cellular features relevant to HCM, providing insights into the underlying molecular mechanisms of the disease. Quantification may involve techniques such as microscopy, genetic analysis, or proteomic profiling to examine cellular morphology, gene expression patterns, or protein abundance in cardiac cells. By quantifying these cellular aspects, researchers can gain a deeper understanding of HCM pathophysiology, identify potential biomarkers for diagnosis or prognosis, and develop targeted therapies to treat the condition effectively.

Various machine learning algorithms are employed for cell identification tasks across biology, medicine, and microscopy. Support Vector Machines (SVM) offer robust performance in classifying cells by learning to distinguish between different types based on extracted features. Random Forest, an ensemble learning method, is adept at handling large datasets and mitigating overfitting, making it a suitable choice for cell identification. Convolutional Neural Networks (CNN) excel in processing image data, automatically learning hierarchical features from raw pixels, thus proving effective for cell classification in microscopy images. K-Nearest Neighbors (KNN) assigns cells to classes based on the majority class among its nearest neighbors, while Decision Trees provide interpretable results by learning simple decision rules from data. Deep learning architectures like Recurrent Neural Networks (RNNs) and Transformers are also applicable, particularly in handling sequential or time-series cell data.

These algorithms can be utilized individually or in combination, depending on factors such as data complexity, dataset size, computational resources, and interpretability requirements. The choice of algorithm is crucial and often tailored to suit specific characteristics and demands of the cell identification task at hand, contributing to advancements in various fields through accurate and efficient cell analysis. The softwares we used in order to quantify HCM are CellProfiler and Ilastik[3].

In the Wolf lab, researchers embarked on a crucial investigation aimed at evaluating the impact of Galunisertib, a potential therapeutic agent, on animals afflicted with cardiac hypertrophy. To conduct this study, they employed a meticulously designed experimental protocol utilizing mice as the model organism. The mice were subjected to Transverse Aortic Constriction (TAC), a widely recognized method for inducing cardiac hypertrophy, thereby simulating the pathological conditions observed in human patients. Following TAC surgery, the mice were divided into distinct groups: one group received treatment with Galunisertib, while the other served as the control group, remaining untreated throughout the study period.

This experimental setup enabled the Wolf lab to systematically investigate the efficacy of Galunisertib in ameliorating cardiac hypertrophy. By comparing the physiological and molecular characteristics of the treated mice with those of the untreated controls, the researchers sought to elucidate the drug's potential therapeutic effects on hypertrophic cardiac remodeling. Parameters such as cardiac function, hypertrophic marker expression, tissue morphology, and overall survival were meticulously assessed to comprehensively evaluate the therapeutic efficacy and safety profile of Galunisertib in mitigating the pathological consequences of cardiac hypertrophy[4].

Results

Hypertrophy Measurements

Cross sectional area of the CMCs was used to quantify the hypertrophy within a tissue sample. Using 12 images from control group mice and 16 images from the drug treated mice, the average cross sectional area was 497 ± 44 px^2 and 547 ± 44 px^2 for control and treatment groups respectively. The average number of cells counted per image was 650 ± 85 cells. A p-value of 0.0082 was determined from conducting a two-tailed t-test. An experiment was conducted which drew from six images (three control and three treatment), with 10 cells being manually measured from each image and compared to the automated measurements. The manual method produced an average cell area of 499 ± 59 px^2 and 654 ± 188 px^2 for control and treatment groups respectively, while the automated method generated area values of 483 ± 32 px^2 and 546 ± 44 px^2 for the control and treatment groups for those same images (Figure 1).

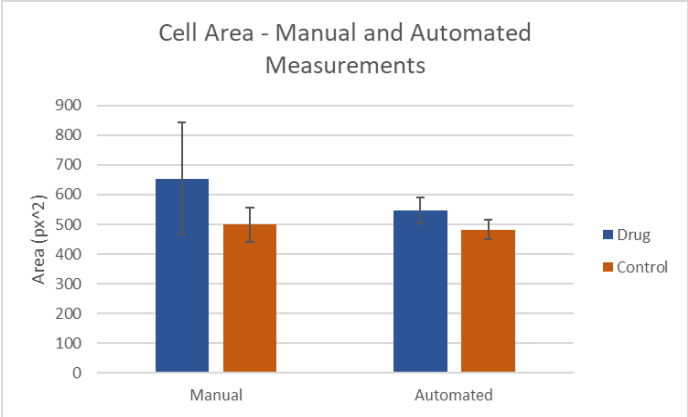

Fig. 1. Manual and Automated Cell Area. Comparison of automated cell segmentation area with manually segmented drug and control sample averages. Manual segmentation provides a control to compare the automated results to. Automated measurements were acquired using the hypertrophy pipeline. It was determined that the pipeline has minimal discrepancies when calculating area.

To further validate the results generated by the automated pipeline, 40 cells were manually segmented to be compared to the automated area values for those same cells using an indexing system. The manual measurements from both researchers and the automated measurements from the algorithm were plotted as a scatter plot in Figure 2. The algorithm was able to accurately recreate the manual measurements taken, giving R2 values of 0.9707 and 0.9651 when plotted with each of the researcher's measurements.

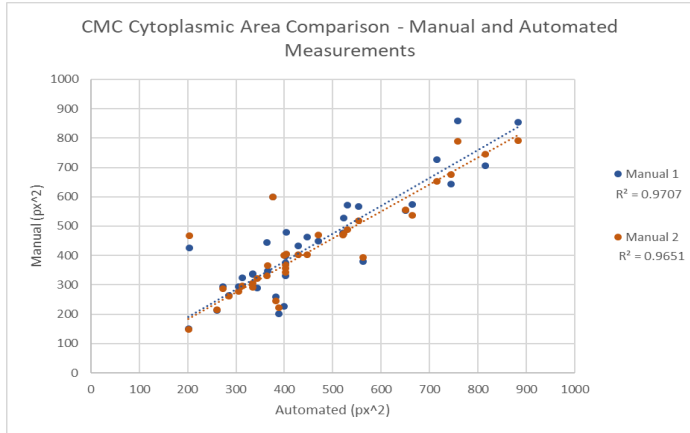

Fig. 2. Scatterplot comparing area measurement methods. Automated and manual measurements of cell area are plotted on the x and y axes, respectively. Two researchers conducted manual analysis, which validated the automated results.

After filtering cells which had eccentricity values greater than 0.85, which allowed for the isolation of cells aligned directly with the field of view to obtain a more accurate measurement of cross sectional area, the average number of cells counted per image was 365 ± 59 cells (Figure 3). The average cell area for the control group cells was 440 ± 46 px^2, and the average cell area for the drug treatment group was 487 ± 34 px^2. Figure 4 shows a comparison of the filtered and unfiltered measurements for CMC cross sectional area. A two-tailed t-test gave a p-value of 0.0087. The difference between the treatment groups was preserved after filtering while providing greater reliability and consistency in the measurements, and the average number of cells counter per image continued to give an adequate sample size.

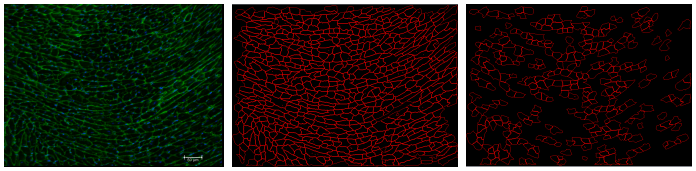

Fig. 3. Demonstration of the filtering process to exclude cells with >0.85 eccentricity. From left to right: (a) the raw tissue slice image, (b) cell segmentation without a filter for eccentricity, (c) cell segmentation after filtering.

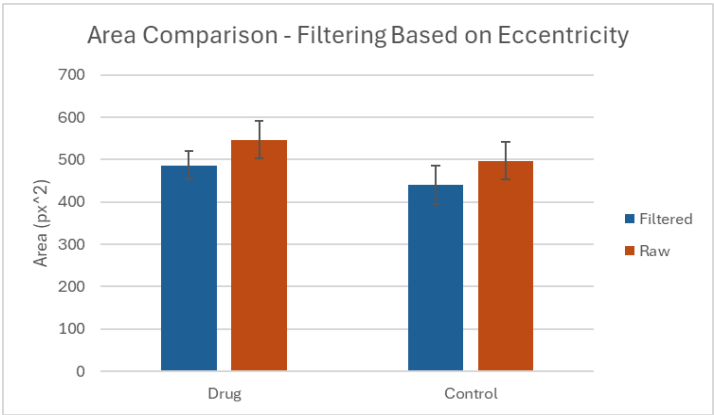

Fig. 4. Comparison of automated nuclei segmentation area with manually segmented drug and control sample averages. Manual segmentation provides a control to compare the automated results to. Automated measurements were acquired using the nuclear segmentation pipeline. It was determined that the pipeline has minimal discrepancies when calculating area.

Nuclear Segmentation Measurements

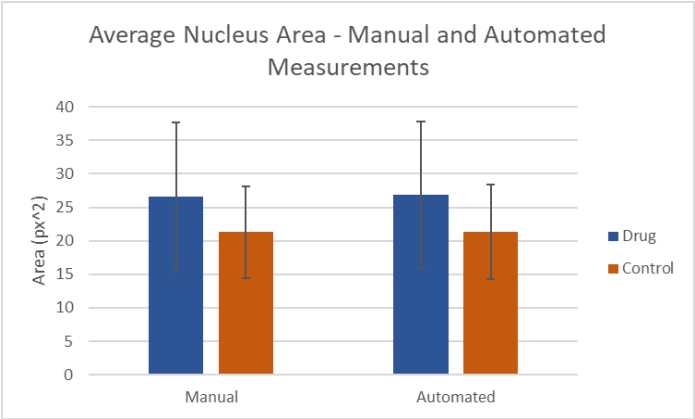

Fig. 5. Comparison of automated nuclei segmentation area with manually segmented drug and control sample averages. Manual segmentation provides a control to compare the automated results to. Automated measurements were acquired using the nuclear segmentation pipeline. It was determined that the pipeline has minimal discrepancies when calculating area.

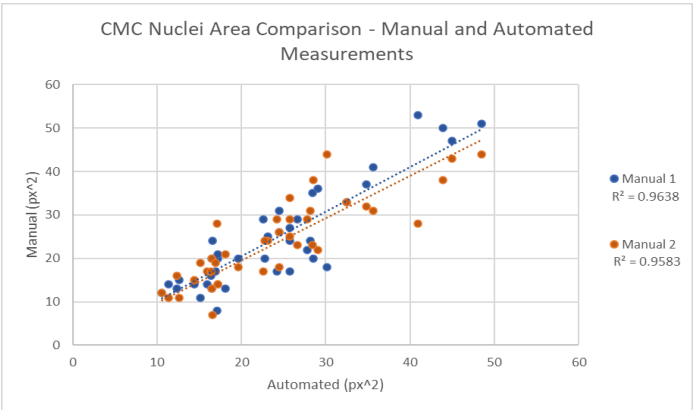

Fig. 6. Scatterplot comparing manual and automated nuclei area measurements. This figure illustrates the correlation between points plotted with the x-axis representing the automated counts retrieved by Cell Profiler and the y-axis representing the manual counts. Two separate researchers conducted manual analysis to validate the automated results. A linear regression of the data is plotted to demonstrate the discrepancies, accuracy, and the trend line. Determined that the pipeline has a high level of accuracy when generating automated quantification.

Nuclear segmentation was analyzed by quantifying hypertrophy of the CMCs. Twelve images of control mice and 16 of drug treated mice were quantified. The average area of nuclear areas were 26.7 ± 5.37 px^2 and 21.65 ± 1.90 px^2 for control and treatment groups respectively. In this study, we examined six images, dividing them evenly between the control and treatment groups, with three images in each. We conducted manual measurements on 10 cells from each image and contrasted them with automated measurements. The manual technique revealed an average nuclear area of 26.7 ± 5.37 px^2 for the control group and 21.65 ± 1.90 px^2 for the treatment group. Conversely, the automated method provided area values of 24.3 ± 4.47 px^2 for the control group and 19.75 ± 1.79 px^2 for the treatment group (Figure 5). A p-value of 0.0063 was determined from a two-sided t-test. To further validate the results obtained from the automated pipeline, we manually segmented 40 cell nuclei and compared them with the automated area values using an indexing system. The scatter plot in Figure 6 illustrates both the manual measurements conducted by researchers and the automated measurements generated by the algorithm. The algorithm accurately mirrored the manual measurements, achieving R^2 values of 0.9638 and 0.9538 when compared against the measurements of each researcher.

DNA Content Analysis

DNA Content analysis was performed on segmented internal nuclei that were embodied in a cell membrane. Simultaneously, this analysis was performed on cell membranes nuclei. Using 10 images from the control group of mice and 10 from the drug treated mice, intensity of DAPI stains and channels could be determined. Though images were hyper stacked, there was inaccuracy in determining the level of intensity of individual channels in the cells. As a result, nuclear area was measured which quantifies the size of nuclei and the amount of DNA there could be by approximation. The average nuclear area of internal nuclei was 37 ± 12 px^2 and the average nuclear area of cell membrane nuclei was 20.5 ± 5 px^2. Though the mean intensity, which is a representation of the amount of pixels per area, was 6.5 ± 0.5 px^2 for the internal nuclei and for the cell membrane nuclei it was 7.5 ± 0.4 px^2 (Figure 7).

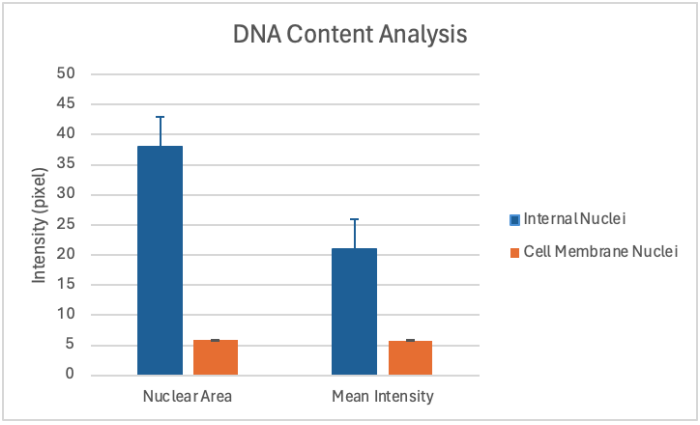

Fig. 7. Graph of nuclear area and mean intensity. Intensity refers to the intensity of fluorescence signals within defined regions of interest (ROIs). These metrics provide insight into signal distribution and consistency across analyzed images. This quantifies the amount of DNA content in the cell membrane and the internal nuclei. As expected, internal nuclei have a higher nuclear area and mean intensity.

Discussion

Through machine learning algorithms and image analysis, variations in cell cross-sectional area and DNA content can be effectively discerned. This technological approach enables the precise filtering of cell populations based on their distinctive characteristics, facilitating an in-depth analysis of cardiomyocyte hypertrophy. With batch processing capabilities, researchers can swiftly analyze vast datasets, accelerating the pace at which results are obtained and enhancing the iterative nature of experiments. DNA content analysis can be quantified by determining proliferation and growth of cells through treatment of drugs. Moreover, the shift from manual segmentation to machine learning methods significantly diminishes the risk of bias and variation stemming from human error and sampling bias, ensuring more robust and reliable findings in cellular research.

Our algorithm not only accelerates the quantification of images and pixels, thereby saving time, but it also ensures a consistent level of objectivity across all quantification metrics. By streamlining this process, it minimizes variability and enhances the reliability of

results. Uniformity is achieved through our algorithm which translates into greater precision and reproducibility, crucial for scientific research and analysis. Researchers can rely on consistent measurements, reducing the likelihood of erroneous conclusions and facilitating robust data interpretation.

As a result, the application of our algorithm holds significant promise in drug screening endeavors. By expediting the assessment process, it empowers drug developers to swiftly evaluate the efficacy of their compounds. This rapid feedback loop enables timely decision-making, potentially accelerating the pace of drug discovery and development. Ultimately, our algorithm serves as a valuable tool in the pharmaceutical industry, facilitating the identification of promising candidates and expediting the journey from bench to bedside.

Future Directions

All of the images used in this project were from healthy mice, meaning none of the samples had HCM when they were being treated. Testing the algorithm on a study that includes both healthy and HCM mice, control and treatment groups, would provide a better understanding of the effectiveness and applicability of the pipeline. Analyzing HCM mice would verify the algorithm's ability to detect and quantify changes in hypertrophy, and this would also allow for an analysis of the effectiveness of galunisertib and other drugs being used to ameliorate hypertrophy in vivo. Additionally, understanding the algorithm's ability to measure hypertrophy would allow for the refinement of the pipeline to make it more efficient. Knowing whether more or less training is required to generate the desired results will ensure researchers are only spending the necessary time and resources in training the program. Lastly, the pipeline currently utilizes multiple open source softwares in sequence to produce the results for hypertrophy, nuclear segmentation, and DNA content analysis. A potential advancement to be made would be to connect these separate programs into a single software to eliminate unnecessary complexity and expedite the analysis process.

Materials and Methods

Fig. 8. Workflow Schematic. Schematic of the workflow and outputs for each of the pipelines developed.

The materials used in this project were images of murine heart tissue cross sections in tiff file format from the Wolf Lab at UVA. The tissue samples were stained with DAPI and WGA to identify the tissue and cellular structure of the samples. All mice used in the study were healthy (i.e., none had HCM), but half of the mice were treated with galunisertib, a drug used to ameliorate hypertrophy in HCM, while the other half served as a control. The softwares used to create the protocol were Ilastik for the training and machine learning aspect of the project, FIJI to convert the file format to be compatible with CellProfiler, and CellProfiler for size and eccentricity thresholding, segmentation, and measurements. Excel was used for downstream analysis and plotting figures. Figure 8 shows a schematic of the analysis pipeline for each aspect of the project.

Hypertrophy

Six images were selected (three from the control group and three from the drug treated group) were inputted into Ilastik to serve as the training data. Color/intensity, edge, and texture were selected at various sigma values during feature selection, as shown in Figure 9. For label one, which are the objects to be identified for the analysis, five cell cytoplasms were selected along with five cell nuclei (so the algorithm would not identify the nuclei as a separate structure for the purpose of this pipeline). For label two, five cell membranes were selected to contrast with the cell cytoplasms. Also in this label, vacant space or scale bars were marked to help prevent them from being falsely identified as cells. This process was repeated for each of the six images, and an example of the training is shown in Figure 10. After the initial training, with the probability map live update activated, adjustments to the training were made until the predictions from Ilastik appeared satisfactory. The rest of the images from the data set were loaded into Ilastik, and the probability maps were batch processed and exported as tiff files for further analysis.

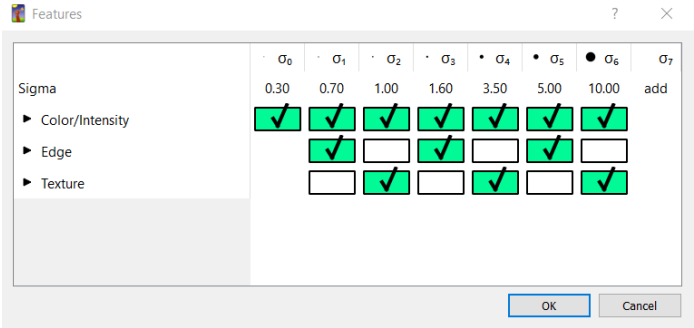

Fig. 9. Feature Selection. This figure shows the sigma value options during feature selection in Ilastik. These options tell the software how to discriminate between different types of pixels in the image to classify them properly.

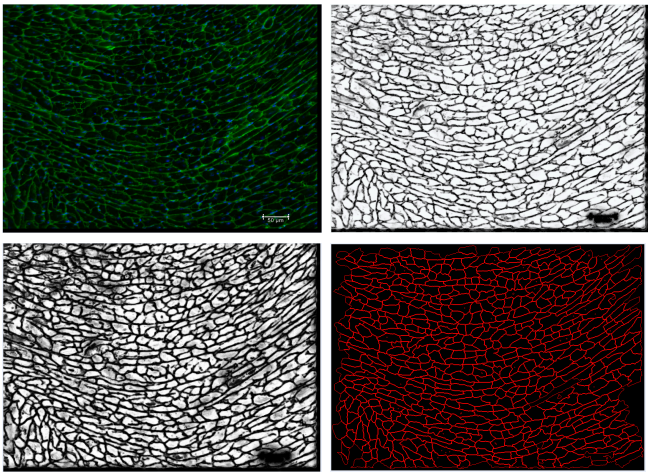

Fig. 10. Demonstration of the image analysis pipeline. Moving from right to left and top to bottom: (a) raw image, (b) Ilastik probability map, (c) Fiji filtering output, and (d) CellProfiler cell outlines.

The probability maps from Ilastik were not initially in a proper format to be compatible with CellProfiler, so an intermediate step to convert the file format in FIJI was added. The images from Ilastik were converted to a 16-bit grayscale format tiff file using the batch processing feature in FIJI, which allowed them to be effectively processed by CellProfiler.

After conversion, the grayscale probability maps were loaded into CellProfiler. An IdentifyPrimaryObjects module allowed for the size thresholding of cells to assist the algorithm in eliminating small particles or large vacant areas as being cells. It also incorporated smoothing scales, correction factors, and was set to declump groups of cells based on intensity. The MeasureObjectSizeShape module gave measurements for the objects identified in the previous module. A FilterObjects module allowed for thresholding based on eccentricity to eliminate elongated and unaligned cells whose true cross sectional diameters could not be determined. The eccentricity cutoff for this module was set to 0.85 as the maximum value, which was determined based on manual sampling of both circular and elongated cells from the training images. A second MeasureObjectShapeSize module was included to measure only the remaining cells after the eccentricity filtering. Another module was added to display cell numbers onto each cell within each image, which allowed manual comparisons to the automated values generated retroactively. CellProfiler outputted csv files of measures for both filtered and unfiltered images as well as images of the segmentation performed by CellProfiler. The CellProfiler pipeline is available within the Supplementary Materials.

The data from CellProfiler was exported as a csv, which was then converted into an Excel file for the final steps of analysis. All graphs, plots, and data analysis were performed in Excel.

Nuclear Segmentation and DNA Content Analysis

Six images were selected (three from the control group and three from the drug treated group) were inputted into Ilastik to serve as the training data. In the context of label one, representing the objects slated for identification in the analysis, five cell nuclei were specifically chosen, accompanied by five cell membranes. This selection ensured that the algorithm refrained from recognizing the nuclei as distinct structures within the pipeline's framework. In label two, the focus shifted to selecting five cell membranes to serve as a point of contrast with the cell cytoplasms. This procedure was replicated for each of the six images. Following the initial training phase, adjustments were iteratively made while utilizing the live update feature of the probability map in Ilastik, ensuring the predictions reached a satisfactory level. Subsequently, the remaining images from the dataset were imported into Ilastik, and the probability maps were subjected to batch processing, yielding exported tiff files for subsequent analysis (Figure 11).

Initially, the probability maps generated by Ilastik were incompatible with CellProfiler. To address this, an intermediary step involving the conversion of file formats using FIJI was introduced[5,6]. Through batch processing in FIJI, the images from Ilastik were converted into 16-bit grayscale format tiff files, facilitating seamless integration with CellProfiler for efficient processing[7,8]. After conversion, the grayscale probability maps were loaded into CellProfiler (Figure 12).

Fig. 11. Nuclear Segmentation Probability Mask. The figure above depicts a probability mask that is rendered after segmentation is completed from the cross-sectioned cell images. The mask indicates the likelihood of where a nuclei can be based on the intensity of the nuclei and the training given.

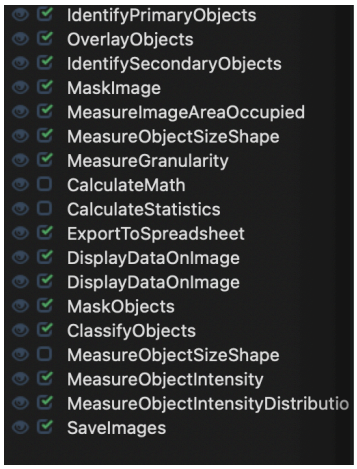

Fig. 12. CellProfiler Pipeline for Nuclear Segmentation and DNA Content Analysis. This figure shows the total selection for the process of analysis. It details the steps taken to analyze a cell image into the quantification of internal nuclei, cell membrane nuclei, and DNA nuclear content.

The CellProfiler analysis process began with the utilization of Primary Objects identification to effectively sift through nuclei based on their pixel sizes, ensuring the elimination of those that were either too large or too small. Additionally, a size filter was integrated to guarantee the selection of nuclei with an optimal eccentricity, maximizing accuracy. These meticulously chosen nuclei represented internal components within cells, characterized by distinct membrane boundaries. The Identify Secondary Objects function was employed to generate a comprehensive map of cell membranes. This step specifically targeted cells lacking a nucleus within their confines. Leveraging the primary image input alongside the probability map of primary objects, this phase meticulously outlined the boundaries of cell membranes. Quantitative analysis ensued through the Measure Image Area Occupied tool, facilitating the precise determination of the areas occupied by both nuclei and cell membranes. Furthermore, employing the Measure Object Size Shape module allowed for the calculation of diameter and perimeter metrics for both nuclei and cell membranes, utilizing Zernike features for enhanced accuracy. Masks representing these objects were then generated, visually depicting the extent of mapping achieved and color-coded to indicate size variations (Figure 13, 14). Moreover, the Measure Object Intensity function was instrumental in quantifying DNA content within both nuclei and cell membranes. Utilizing merged DAPI channels, this method provided a robust measurement, with a preference for Nuclear Area measurements due to their superior representation of nuclear content9. The pipeline is within supplementary objects. The data from CellProfiler was exported as a csv, which was then converted into an Excel file for the final steps of analysis. All graphs, plots, and data analysis were performed in Excel.

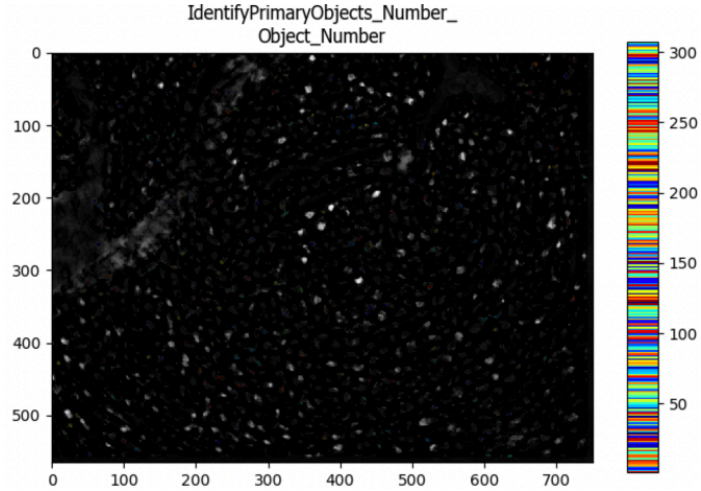

Fig. 13. Cardiomyocyte Nuclear Segmentation in CellProfiler. This figure above depicts a mask created of the internal nuclei based on size and color. The size and color are assorted on the right of the map and the nuclei are highlighted across the mask.

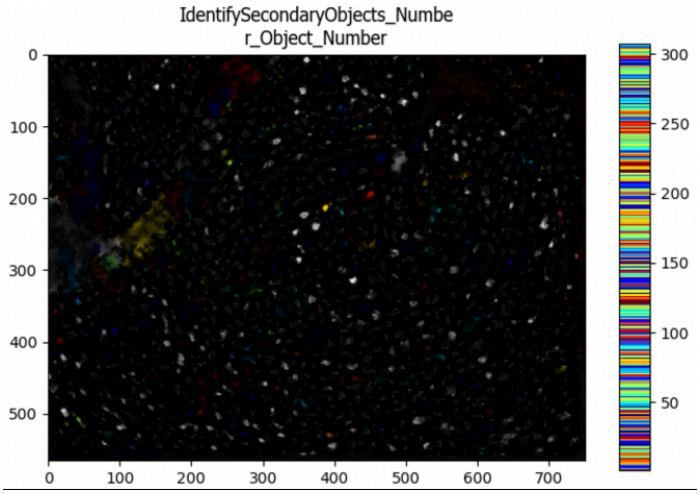

Fig. 14. Membrane Nuclear Segmentation in CellProfiler. This figure above depicts a mask created of the cell membrane nuclei based on size and color. The size and color are assorted on the right of the map and the nuclei are highlighted across the mask.

Conclusion

The development of a machine learning algorithm for the automated analysis of tissue sections will permit the more rapid screening of drugs during in vivo testing. Manual quantification methods are slow and do not offer the same level of consistency and objectivity as can be offered by our pipeline. In the development of drugs to treat HCM, the ability to quantify changes in hypertrophy and DNA content enable high throughput testing and the processing of larger datasets. The continued application and development of our algorithm within the Saucerman and Wolf Labs will expand our understanding of HCM and accelerate the discovery of new treatment methods to improve the quality of life for patients.

End Matter

Author Contributions and Notes

Elward and Oluwafemi designed research, Elward and Oluwafemi performed research, Elward constructed the hypertrophy pipeline, Elward and Oluwafemi constructed the nuclear segmentation pipeline, Oluwafemi performed the DNA content analysis, Elward and Oluwafemi analyzed data; and Elward and Oluwafemi wrote the paper. Luanpaisanon and Saucerman were advisors for the project.

The authors declare no conflict of interest.

This article contains supporting information online. See Supplementary Materials.

Acknowledgments

Wolf, Matthew J - Associate Professor, Medicine: Cardiovascular Medicine