Automatic Identification of Algae using Low-cost Multispectral Fluorescence Digital Microscopy, Hierarchical Classification & Deep Learning

by

Jason Deglint

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Stefan C. Kremer Professor School of Computer Science, University of Guelph
Supervisor:	Alexander Wong Associate Professor Systems Design Engineering, University of Waterloo
Internal Member:	Katharine Scott Assistant Professor Systems Design Engineering, University of Waterloo
Internal Member:	John Zelek Associate Professor Systems Design Engineering, University of Waterloo
Internal-External Member:	Zhou Wang Professor Electrical and Computer Engineering, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

There are three journal papers that make up parts of this thesis. Each paper will outline the contributions of each author that are relevant to the work presented in this thesis. I am the primary author and major contributor of all these papers.

The first journal paper was published in the Journal of Computational Vision and Imaging Systems in 2018 and was titled *SAMSON: Spectral Absorption-fluorescence Microscopy System for ON-site-imaging of algae*. The authors on this paper were Jason L. Deglint (JLD), Lyndon Tang (LT), Yitian Wang (YW), Chao Jin (CJ) and Alexander Wong (AW). This paper published the implementation of the imaging system designed by JLD, CJ and AW, as discussed in Section 4.4. The printed circuit board (PCB) and the 3D printed frame from this paper are presented in Section 4.4.2 of this thesis. The PCB and 3D print was designed by JLD and LT while LT implemented these designs. The graphical user interface (GUI) in this paper is presented in Section 4.4.3 of this thesis. The GUI was designed by JLD and YW while YW implemented the design.

The second journal paper was published in IEEE Access in 2018 and was titled *The Feasibility of Automated Identification of Six Algae Types using Feed-forward Neural Networks and Fluorescence-based Spectral-morphological Features*. The authors on this paper were JLD, Angela Chao (AC), CJ, and AW. This paper demonstrated that features could be extracted from multispectral fluorescence data and used as input into a feedforward neural network for classification of six algae. JLD and AC implemented the imaging apparatus in this paper, which inspired the work presented in Section 4.4.2. The feedforward model in this paper inspired the work in Section 5.2.1, and the data preprocessing in this paper was similar to Section 6.3. All the data collection and analysis in this paper are independent of this thesis.

The third paper was published in Springer: Lecture Notes in Computer Science in 2019 and was titled *Investigating the Automatic Classification of Algae via Deep Residual Learning*. The authors of this paper were JLD, CJ, and AW. The residual network developed for this paper is the same model presented in Section 5.2.2 and was designed by JLD and AW. All the data collection and analysis in this paper are independent of this thesis.

Abstract

Harmful algae blooms (HABs) can produce lethal toxins and are a rising global concern. In response to this threat, many organizations are monitoring algae populations to determine if a water body might be contaminated. However, identifying algae types in a water sample requires a human expert, a taxonomist, to manually identify organisms using an optical microscope. This is a tedious, time-consuming process that is prone to human error and bias. Since many facilities lack on-site taxonomists, they must ship their water samples off site, further adding to the analysis time.

Given the urgency of this problem, this thesis hypothesizes that multispectral fluorescence microscopy with a deep learning hierarchical classification structure is the optimal method to automatically identify algae in water on-site. To test this hypothesis, a low-cost system was designed and built which was able generate one brightfield image and four fluorescence images. Each of the four fluorescence images was designed to target a different pigment in algae, resulting in a unique autofluorescence spectral fingerprint for different phyla groups.

To complement this hardware system, a software framework was designed and developed. This framework used the prior taxonomic structure of algae to create a hierarchical classification structure. This hierarchical classifier divided the classification task into three steps which were phylum, genus, and species level classification. Deep learning models were used at each branch of this hierarchical classifier allowing the optimal set of features to be implicitly learned from the input data.

In order to test the efficacy of the proposed hardware system and corresponding software framework, a dataset of nine algae from 4 different phyla groups was created. A number of preprocessing steps were required to prepare the data for analysis. These steps were flat field correction, thresholding and cropping. With this multispectral imaging data, a number of spatial and spectral features were extracted for use in the feature-extraction-based models.

This dataset was used to determine the relative performance of 12 different model architectures, and the proposed multispectral hierarchical deep learning approach achieved the top classification accuracy of 97% to the species level. Further inspection revealed that a traditional feature extraction method was able to achieve 95% to the phyla level when only using the multispectral fluorescence data. These observations strongly support that: (1) the proposed low-cost multispectral fluorescence imaging system, and (2) the proposed hierarchical structure based on the taxonomy prior, in combination with (3) deep learning methods for feature learning, is an effective method to automatically classify algae.

Acknowledgements

Completing a PhD is a substantial commitment and as with any journey, has many ups and down as well as twists and turns. There is no way I could have completed my dissertation without the guidance, love and support of many individuals and organizations whom I am honoured to acknowledge now. All the people mentioned below helped me get to where I am today.

First I want to thank Alexander Wong, my supervisor and research dad. Without Alex I wouldn't have begun a PhD as he encouraged me to apply for NSERC and supported me in preparing my application which in turn allowed me to pursue a full-time PhD without financial burden. Thank you Alex for taking me on as your student and for allowing me the freedom to select my research topic and pursue my passion for entrepreneurship. I greatly value your honesty and guidance to help me think critically and objectively about my research.

Secondly I want to thank Chao Jin, my co-supervisor and friend. When I started my PhD in September 2016, I intended to pursue research in multispectral imaging of art paintings. After chatting extensively with Chao I decided to switch my PhD topic to focus on water. Thank you Chao for your support and conversations about developing imaging technology to ensure clean drinking water, as this has a direct impact on people's lives.

Next I want to thank my dear wife, Taylor Jade. Thank you Taylor for your unwavering support in my studies and encouraging me in my entrepreneurial ambitions. I couldn't have completed this PhD without you.

Thank you to de Gaspé Beaubien Foundation for hosting AquaHacking and for generously awarding our team (Alex, Chao and myself) the first place prize in 2017. This was the beginning of my entrepreneurial journey in the water industry and the origin of Blue Lion Labs. Winning AquaHacking accelerated my path as an entrepreneur and was a springboard that opened doors into Toronto, Montreal, Vancouver, China and the USA.

Thank you Angela Chao, Lyndon Tang, and Yitian Wang (co-ops and URAs) who assisted with my research. I value the work you contributed to my thesis and am grateful for your time. Also, thank you Heather Roshon at the CPCC for all your hard work preparing samples and answering all my questions.

Thank you NSERC for awarding me a full scholarship and thank you University of Waterloo for hosting me as a student. Thank you Velocity for providing free workspace and mentorship for Blue Lion Labs. Thank you to my committee for your time and energy.

Finally, thank you to all my family and friends who supported me in my PhD studies in one way or another. Specifically, I want thank my mom (Jane Deglint) for proofreading this thesis for spelling and grammar.

Dedication

ded · i · cate: devote (time, effort, or oneself) to a particular task or purpose.

I have been given a gift.
- The gift of freedom
  - to explore my interests and to pursue my passions.
- The gift of opportunity
  - to study and research this universe.
- The gift of time
  - to ponder and travel.

With this gift comes responsibility.
- Responsibility to work
  - on meaningful solutions that build and create.
- Responsibility to share
  - my knowledge and discoveries.
- Responsibility to serve
  - those in need and my fellow man.

This responsibility calls me to devote myself.
- Devote myself to constant gradual improvement
  - based on ruthlessly seeking reality and taking ownership of my actions.
- Devote myself to build up others
  - by encouraging them and treating each person with respect and dignity.
- Devote myself to creating a better world
  - not out of payment for the gift, but out of thankfulness for it.

First off, I dedicate this work to my amazing wife Taylor Jade. You have encouraged me to pursue reality and held me accountable to a higher standard. I admire your love for life and internal joy that you share with others. You can do anything.

Secondly, I dedicate this work to the spirit of innovation and exploration that each person has. May it take us further than we can ever imagine. May it allow us to steward our home and care for each other.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Access to water is access to education, work, and the kind of future we want for all the members of our human family."

– Co-founder of water.org, Matt Damon (1970 - present)

This introductory chapter provides the reader with a high level overview of the entire thesis. First, in Section 1.1, an overview of harmful algae blooms (HABs) will be given. Then, in Section 1.2, the problem of manual identification of algae will be discussed. Finally, in Section 1.3, the main scientific contributions and outline of the thesis will be presented.

1.1 Motivation

We live in an amazing time. In general, people are living healthier and longer lives due to advances in medicine and health. Thanks to modern transportation one can now travel anywhere in the world. And thanks to the digital age, we can communicate with anyone from nearly anywhere on the planet. Despite these optimistic observations, the world still has significant concerns. These include, but are not limited to: poverty, lack of access to clean drinking water and basic education, human sex trafficking, gender inequality, our rapid energy consumption, and climate change. In fact, in 2015 the United Nations (UN) created the Sustainable Development Goals outlining 17 major goals which they hope to reach in the near future [1]. The 150 targets included in these goals were proposed by sector experts.

Figure 1.1: The Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua satellite showing a harmful algae bloom (HAB) on Lake Erie on October 9, 2011 [2].

However, this overwhelming amount of targets created its own problems. How are the targets prioritized? Which of these problems are actually solvable given our current technologies, resources and time frame? Bjorn Lomborg, who led a team of economists, set out to answer these questions and took on the task of prioritizing and ranking the order of the UN targets. Lomborg explains that while these targets are commendable, many are also unrealistic and unpractical. He argues that by pursuing unachievable goals, we will actually hinder our progress instead of advancing it. For these reasons Lomborg and his team proposed a set of 19 achievable targets, resulting in the greatest return on every dollar spent on a given problem [3].

In the spirit of tackling a major world problem as proposed by the UN, while honouring the focus on urgent problems as identified by Bjorn Lomborg, this thesis will make a contribution to the challenge of providing clean water to all regions of the earth. The specific focus of this thesis is the identification of algae in water with the goal of more efficient management of harmful algae blooms (HABs). Recently, harmful algae blooms (HABs) have become a common experience around the globe. For example, as seen in Figure 1.1, in the summer of 2011 Lake Erie experienced a severe harmful algae bloom [4]. According to the Great Lakes Environmental Research Laboratory this bloom was primarily *Microcystis aeruginosa*, a type of lethal cyanobacteria [2].

Cyanobacteria can be extremely dangerous for humans and animals. For example,

swallowing *Microcystis* can lead to serious reactions, such as abdominal pain, diarrhea, vomiting, blistered mouths, dry coughs, and headaches. In addition, *Anabaena*, another common cyanobacteria, can produce lethal neurotoxins called anatoxin-a which are shown to have caused death by progressive respiratory paralysis [5]. Therefore it is essential for the proper management of any water that the water quality is monitored for different cyanobacteria and other algae [6]. The preservation and maintenance of our water directly affects marine wildlife, as well as the recreational, fishing and tourism industries. Moreover, water maintenance is crucial for water treatments plants that are in charge of distributing clean drinking water to the population.

However, as we will see in Section 1.2, this poses a significant problem as samples of water must be taken from the source of water and shipped to a certified laboratory for inspection. Then these samples must be manually inspected under traditional microscopy techniques and each algae type is manually identified and enumerated.

1.2 Overview of Problem

The current method to monitor algae requires a human expert, a taxonomist to use a microscope in order to manually identify and enumerate each algae. Due to the low number of algae taxonomists available and the extensive experience required to classify different algae, the process to identify and enumerate algae is time-consuming [6].

In addition, since the sample must travel to the taxonomist, there is additional transit time, which further increases the turn-around time and the cost of analysis. Due to this long process it is nearly impossible to maintain ongoing active monitoring of a given water body for different algae and cyanobacteria [7, 8]. To understand the environmental precursors to algae blooms, and to study the behaviour of algae under different environmental factors, a in-situ method to quantitatively identify and measure different algae species is essential and much needed [8].

The problem that will be addressed and solved in this thesis is the need for human experts to manually identify different types of algae by looking through a microscope. This problem will be solved by creating a novel imaging system that using machine learning methods to automatically identify different types of algae in water samples. Solving the problem of manual identification removes the need to transport the sample away from the source water, removing the burden of analyzing back-logged samples by trained taxonomists. Therefore this thesis provides a method for on-site analysis of water samples which can be analyzed in real-time using machine learning models.

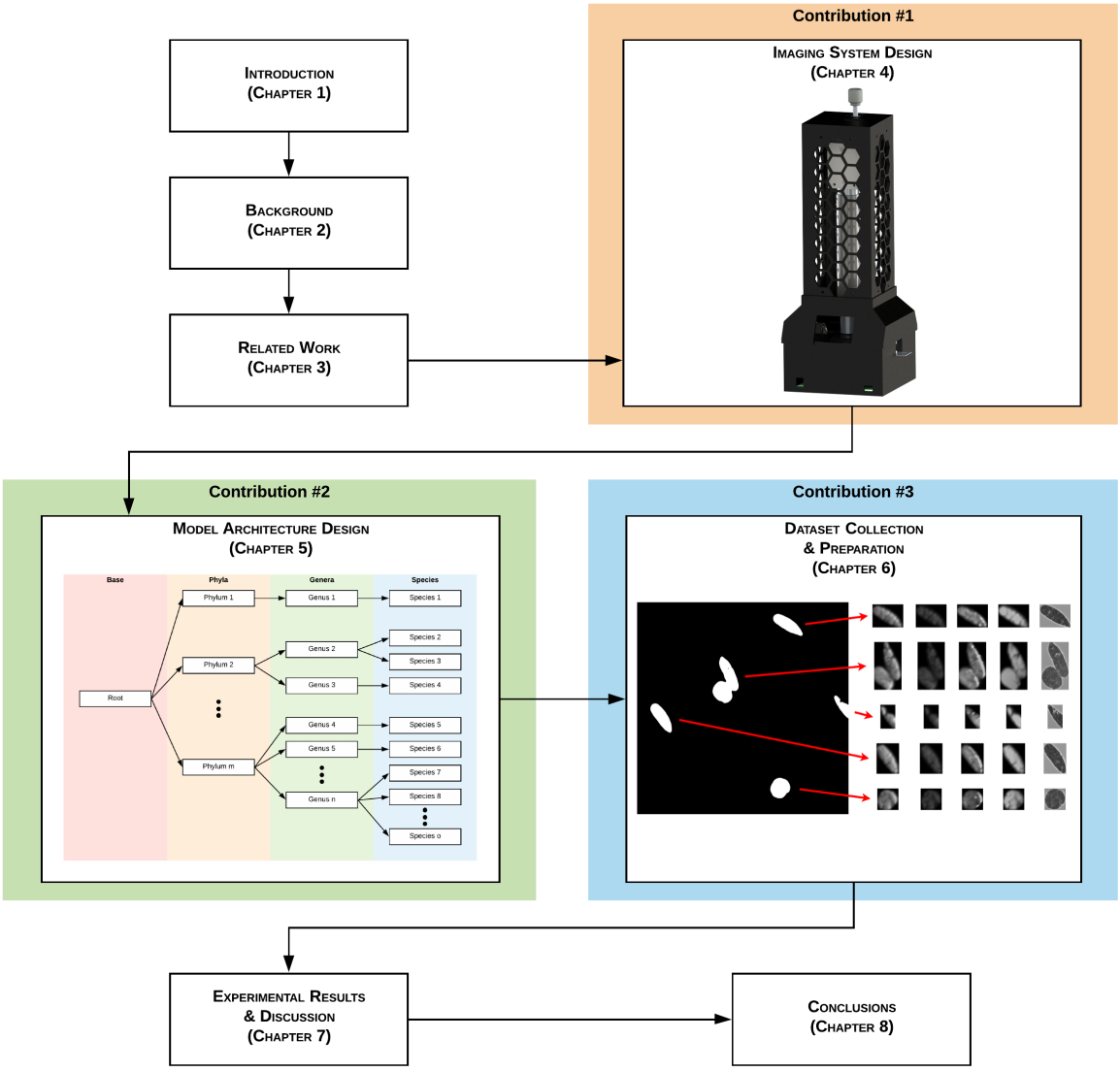


Figure 1.2: This thesis is composed of 8 main chapters. After this introductory Chapter 1, Chapter 2 goes into detail of the background of HABs and algae, while Chapter 3 present the current methods of automatic identification of algae. Chapter 4 presents the hardware contribution of this thesis and Chapter 5 describes the software framework. Then, in Chapter 6 a dataset is created in order to evaluate the efficacy of the Contribution 1 and Contribution 2. In Chapter 7 the results are presented and in Chapter 8 the final conclusions are discussed.

1.3 Thesis Contributions & Outline

Given the importance of identifying algae and the current limitations of the manual process, this thesis proposes a new method to automatically identify algae in water. This work is broken in eight chapters, as seen in Figure 1.2. After this introductory chapter, Chapter 2 will go into further details of HABs and the current practices of manual identification. Chapter 2 extensively covers the problem being solved in this thesis. We will see how manual identification and enumeration using a microscope is both tedious and time-consuming, and is very prone to human bias and error.

As discussed by Walker *et al.*, developing a method to automatically identify algae down the genus or species level using modern imaging systems in combination with pattern recognition techniques would be extremely valuable [8]. Sieracki *et al.* add that these systems can potentially be used as an early warning sign for harmful algae blooms in water bodies since these imaging systems can record the contents of predetermined water volumes at high rates [9]. Therefore, in Chapter 3, the various methods to automate the identification of algae will be presented. These methods include digital microscopy, imaging flow cytometry, spectral analysis, fluorescence probes, and genomics.

To date, each method discussed in Chapter 3 has been fairly independent from the other. Therefore, the proposed research combines two of these methods, the spatial element from digital microscopy with the spectral element of fluorescent spectroscopy. The proposed solution has three main scientific contributions, which are presented in the next three chapters:

1. Contribution 1: Creation of a novel imaging device for low-cost acquisition of absorption and multispectral fluorescence images of algae (Chapter 4).

2. Contribution 2: Creation of a software framework to automatically identify algae using multispectral-based classification & taxonomy-based-hierarchical classification (Chapter 5).

3. Contribution 3: Creation of a novel dataset using Contribution 1 in order to test Contribution 2 (Chapter 6).

Given this dataset, Chapter 7 will present the relative performance of all these models and highlight the use-case for each one in order to determine the efficacy of Contribution 1 and Contribution 2. Finally in Chapter 8, the main conclusions and future work will be presented.

Chapter 2

Problem

"In the year 1657 I discovered very small living creatures in rain water."

– 'the Father of Microbiology', Antonie van Leeuwenhoek (1632 – 1723)

After the brief description of the problem in Chapter 1, this chapter presents further details of HABs in Section 2.1 and the process of manual identification of algae in Section 2.2. We will see that it is imperative that drinking water treatment plants closely monitor algae as this can be an indicator to assess which toxins are present in their source water. Furthermore, we will examine why the current method of manual identification is time-consuming and tedious.

2.1 Harmful Algae Blooms (HABs)

Harmful algae blooms (HABs) develop when different types of algae grow out of control in a water body causing them to produce lethal toxins. HABs are happening all around the world, from the North American Great Lakes, to the African Great Lakes and from Lake Taihu in China to the Baltic Sea [12]. These HABs are having severe global economic and social impact as they are affecting drinking water quality, recreational use of water and tourism, as well as the fishing industry [13]. For example, during the 2014 Toledo water crisis over 400,000 people lost access to clean drinking water for nearly three days [14].

Cyanobacterial Genera	Hepatotoxins		Neurotoxin	
	Cylindrospermopsin	Microcystin	Anatoxin	Saxitoxin
Anabaena / Dolichospernum	x	x	x	x
Anabaenopsis		x		
Aphanizomenon	x		x	x
Aphanocapsa		x		
Cylindrospermopsis	x			x
Haplosiphon		x		
Lyngbya (Plectonema)	x			x
Microcystis		x		
Nostoc		x		
Oscillatoria (Planktothrix)		x	x	x
Phormidium			x	
Pseudanabaena		x		
Raphidiopsis	x		x	
Umezakia	x			
Synechococcus		x		
Synechocystis		x		

Table 2.1: The State of Ohio presents the common algae which are known to produce different toxins [10]. Knowing whether toxic producing algae are in a water body is critical information for a drinking water treatment plant [11].

Furthermore, the number of blooms being observed has been increasing rapidly [15], and due to phosphorus and nitrogen runoff in combination with climate change these blooms are only expected to increase in severity. [12, 16].

One common toxin produced by a number of algae is known as microcystin, as seen in Table 2.1. This toxin has a threshold guideline set by the World Health Organization (WHO) since it is lethal for humans [17]. The maximum acceptable concentration (MAC) for the cyanobacteria toxin microcystin-LR in drinking water is 1.5 μg/L, according to the Government of Canada [18, 11]. In addition, in 2014 the U.S.A. released the Harmful Algal Bloom and Hypoxia Research and Control Amendments Act (HABHRCA), which requires the National Oceanic and Atmospheric Administration (NOAA) and United States Environmental Protection Agency (USEPA) to advance the scientific understanding and ability to detect, monitor, assess, and predict HABs and hypoxia events in marine and freshwater in the United States [19].

Many different algae are the source of different toxins in our drinking water, as seen in Table 2.1, which is presented by the State of Ohio [10]. Therefore knowing which algae are

present in a water sample provides insight into whether an algae bloom will have toxins or not. To quote the Ohio State EPA:

> "Phytoplankton samples can be collected to determine the cause of the bloom. If cyanobacteria are present, the manager should use [Table 2.1] ... to determine if the bloom is capable of producing cyanotoxins, and which cyanotoxins should be analyzed."[10]

Health Canada also voices the need for accurate counts which are provided by a highly trained professional:

> "The use of a trained microscopist with experience in identifying cyanobacteria is favourable when performing cell counts."[11]

2.2 Manual Algae Identification

Given the importance of having a trained professional with the ability to identify different algae, it is worthwhile to understand this history and current practices of trained taxonomists. As explained by He *et al.*, the WHO also suggests to use cyanobacteria and algae identification and enumeration by a human expert as a screening tool to asses the severity of an algae bloom. However, this process requires a high level of expertise as well as time and for that reason becomes unsuitable as an early warning sign [20].

As seen in Figure 2.1, the standard method of identifying and enumerating microalgae consists of three main steps which are: (1) sample preparation, (2) classification, and (3) enumerating. The purpose of the sample preparing is to condense the algae down to a higher concentration that can then be observed. The three main methods to condense the sample are filtration, centrifugation and use of sedimentation. Identification of different genera and species is done manually by the human taxonomist. Lastly, a counting chamber is used to aid the taxonomist in enumerating the sample. Before analysis of the sample can occur the sample must first be concentrated. The two methods to concentrate live samples are filtration and centrifugation, while sedimentation is used for preserved samples. In general, centrifugation and sedimentation are the most common depending on whether the cells are alive or are being preserved, however filtration is an effective means as well [21].

When identifying unicell or small colonies by way of filtration or centrifugation, counting chambers such as the haemacytomter, the Thoma chamber, the Fuchs-Rosenthal or the Burker chambers can be utilized to estimate the density of a variety of cultures. All these

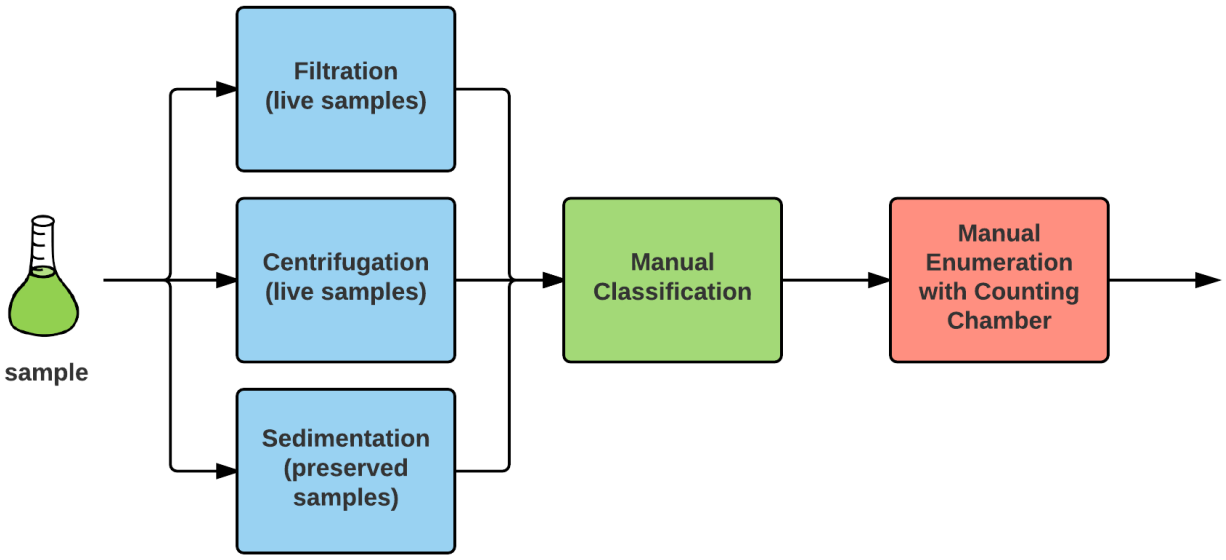


Figure 2.1: The standard method of identifying and enumerating microalgae consists of three main steps which are: (1) sample preparation, (2) classification, and (3) enumerating. The three main methods to condense the sample are filtration, using centrifugation and sedimentation [21].

chambers have a grid with known spacing in order to determine the different microalgae counts. In addition, most of these cell counters have very delicate surfaces that cannot be scratched and must be handled very carefully [21]. When using sedimentation a special type of counting chamber known as an Utermöhl chamber can be used to identify and count different microalage. These Utermöhl chambers are usually very expensive and also require time (approximately one day) to settle to the bottom of a solution. Furthermore, a solution is added to the sample that kills and preserves the sample, which allows them to sink to the bottom of the solution on account of gravity. The advantage is that now the sample can be directly placed on an inverted microscope which can be used to observe random sections of the condensed algae which facilitates the identification and enumeration of microalgae [21].

Throughout the history of phycology, the study of algae, considerable effort was put forth to classify and organize the diversity of algae into distinct groups. However little effort was made to provide "identification (ID) keys" in order to facilitate its usefulness to others, until 1931 when Lily Newton published her handbook of algae in the British Isles [22]. 31 years later, in 1962, Eifion William Jones published his key to the genera

Figure 2.2: Manual identification of algae requires a human expert, a taxonomist, to manually look through a microscope and classify all the organisms present in a water sample. This method requires years of experience and is both time-consuming and tedious.

of British seaweed [23]. Throughout the last 50 years many others have released keys in form of books [24, 25]. Since the advent of the internet numerous credible sources have released online ID keys providing images of algae under a microscope to assist both novice and professional taxonomists [26, 27, 28].

Regardless of what key taxonomists use, they must observe the algae under a microscope, and compare their observations against the ID key, as seen in Figure 2.2. Constantly switching between between these two tasks is tedious, time consuming [29], and may cause straining on the eyes. A study by Culverhouse *et al.* estimate that human taxonomists have a self-consistency identification accuracy between 67% to 83% and a 43% consensus when comparing between taxonomists [30]. Furthermore, as identification can take hours the taxonomist will likely fatigue, increasing the likelihood of a miss-classification. Due to this time-consuming process the average time it takes for a given organization to get the taxa breakdown of a water sample can be anywhere from a few days to a week. It is for these reasons that many people have explored the automation of this task using digital imaging and pattern recognition, as will be discussed in Chapter 3.

Chapter 3

State of the Art

"If I have seen further it is by standing on the shoulders of giants."

– Sir Isaac Newton (1643 - 1727)

Over the years there have been many different methods to automatically identify algae. First off, in order to get an understanding of the biological classification of algae, an overview of algae taxonomy and pigmentation will be presented in Section 3.1. We will see that different algae groups have different associated pigments, and that these pigments also have different spectral absorption curves.

Next the automated methods will be presented. In this thesis these methods will be grouped into five main categories: microscopy based methods (Section 3.2), imaging flow cytometry based methods (Section 3.3), spectral based methods (Section 3.4), fluorescence probe based methods (Section 3.5), and finally genomics based methods (Section 3.6).

Microscopy methods can be further broken down into brightfield microscopy (Section 3.2.1), fluorescence microscopy (Section 3.2.2), and a combination of brightfield and fluorescence microscopy (Section 3.2.3). Spectral methods remove any spatial components and consist of absorption spectroscopy (Section 3.4.1) and fluorescence spectroscopy (Section 3.4.2). After exploring these methods a summary of the main findings for each method will be presented in Section 3.7.

3.1 Algae Taxonomy & Pigmentation

The term algae has no clear definition since many organisms that are referred to as algae come from significantly different branches in the tree of life. For example, cyanobacteria (blue-green algae) belong to the bacteria domain, while green algae belong to the eukarya domain. The domain division is the lowest base rank split in the taxonomic structure, illustrating that the term "algae" can include significantly different organisms. The one characteristic of all algae is that they have chlorophyll and occasionally other pigments to carry out photosynthesis. Using pigmentation to classify different algae was first done by William Henry Harvey in 1836 when he divided algae into four major divisions: Rho-

Table 3.1: Different phlya of algae contain different pigments within their structure. The most noticeably pigments are the Chlorophylls, C-Phycoerythrin (C-PE), C-Phycocyanin (C-PC), and allophycocyanin (APC) [21].

Phylum	**Common Name**	**Pigments**			
		Chlorophylls	**Phycobilins**	**Carotenoids**	**Xanthophylls**
Cerozoa (Chlorarachniophyta)	n/a	***a, b***	Absent	Absent	Lutein, Neoxanthin, Violaxanthin
Charophyta	n/a	***a, b***	Absent	α-, **β-**, & γ-Carotene	Lutein, Neoxanthin, Violaxanthin
Chlorophyta	Green algae	***a, b***	Absent	α-, **β-**, & γ-Carotene	Lutein Prasinoxanthin
Cryptophyta	Cryptomonads	***a, c***	Phycoerythrin-545 R-Phycocyanin	α-, **β-**, & ε-Carotene	Alloxanthin
Cyanophyta	Cyanobacteria Blue-green algae	***a, b***	**C-Phycoerythrin C-Phycocyanin Allophycocyanin**	**β-Carotene**	Myxoxanthin Zeaxanthin
Euglenzoa	Euglenoids	***a, b***	Absent	**β-**, & γ-Carotene	Diadinoxanthin
Glaucophyta	n/a	***a***	**C-Phycocyanin Allophycocyanin**	**β-Carotene**	Zeaxanthin
Haptophyta	Coccolithophorids	***a, c***	Absent	α- & **β-Carotene**	Fucoxanthin
Myzozoa (Dinophyta)	Dinoflagellates	***a, c***	Absent	**β-Carotene**	Peridinin, **Fucoxanthin,** Diadinoxanthin Dinoxanthin Gyroxanthin
Ochrophyta	Golden algae Yellow-green algae Diatoms Brown algae	***a, c***	Absent	α-, **β-**, & ε-Carotene	**Fucoxanthin,** Violaxanthin
Rhodophyta	Red algae	***a***	B-Phycoerythrin R-Phycoerythrin R-Phycocyanin **Allophycocyanin**	α- & **β-Carotene**	Lutein

dospermae (red algae), Melanospermae (brown algae), Chlorospermae (green algae) and Diatomaceae (diatoms) [31].

As provided by Barsanti *et al.* [21], a common and modern classification of algae into different groups can be seen in Table 3.1. This classification scheme groups algae into different phylum groups, as seen on the left in Table 3.1. Cyanophyta, as previously mentioned in Chapter 1 and Chapter 2, are commonly known as cyanobacteria or blue-green algae, and is the group of algae that is most associated with harmful algae blooms. Other common phyla groups are Chlorophyta (commonly known as green algae), Euglenzoa, dinoflagellates, as well as diatoms and red algae.

Table 3.1 also highlights the major pigments present in each phyla group. While all algae contain chlorophylls, certain algae groups contain specific pigments that other groups do not have. For instance, Cyanophyta have three common phycobilin pigments, C-Phycoerythrin (C-PE), C-Phycocyanin (C-PC), and allophycocyanin (APC), while these phycobilin pigments are completely absent from Chlorophyta, Euglenzoa and other phyla groups. A similar pattern can be observed when inspecting the carotenoid pigments and the xanthophyll pigments. For example, Cyanophyta only contains β-Carotene, while other groups contain additional carotenoid pigments. The other important thing to consider is that even if two phyla groups have the same pigments present, they likely have different concentrations of that pigment. For example, while both Cyanophyta and Chlorophyta contain chlorophyll-*a*, one phyla may have more of that specific pigment than the other. As we will see later in Chapter 6, this in fact can be observed.

As one may expect, each pigment has a different associated spectral curve that occupies a unique part of the electromagnetic spectrum. This is illustrated in Figure 3.1, where we see seven common pigments that were reported by Coltelli *et al.* [32]. Inspecting the spectral absorption curves reveals that chlorophyll-*a* and chlorophyll-*b* have peaks in the blue part of the spectrum (400 nm - 475 nm) as well as in the orange and red part of the spectrum (600 nm - 700 nm). This is opposed to B-Phycoerythrin (B-PE) that occupies the green part of the spectrum (500 nm - 600 nm) and C-Phycocyanin (C-PC) which occupies the orange part of the spectrum (550 nm - 650 nm).

Since different pigments have unique spectra, and since different algae phylum groups have different combinations of pigments, one can deduce that a given algae group will likely have a unique spectrum relative to another algae group. As we will see in Chapter 3, this reality is commonly used in spectral methods (Section 3.4), however, is it rarely used in combination with digital microscopy (Section 3.2). Only a few researchers have explored leveraging both the spectral uniqueness of algae due to their pigments and the spatial uniqueness of algae due to their different morphologies (Section 3.2.3).

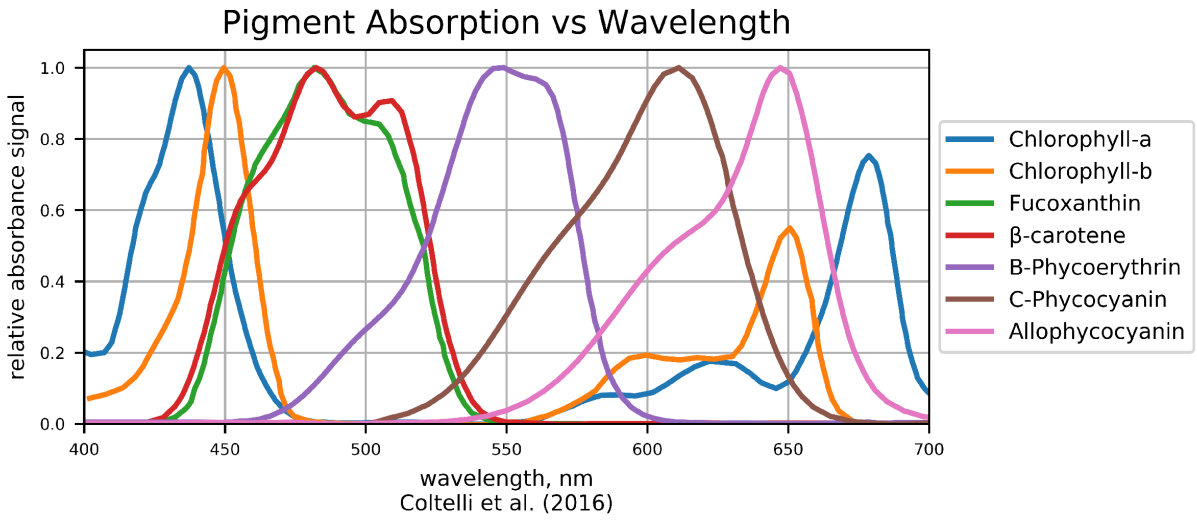


Figure 3.1: As observed in Table 3.1, different phyla of algae contain different antenna pigments [21]. Here the major pigments are shown on a single plot [32]. Note how different pigment are spread across different parts of the visible spectrum.

3.2 Microscopy Methods

A microscope is an optical device that magnifies a given sample by orders of magnitude in order to viewed by human eyes or a digital sensor. Two common forms of microscopy are known as brightfield microscopy (Section 3.2.1) and epi-fluorescence microscopy (Section 3.2.2). Newer methods have been combining both brightfield and fluorescence microscopy (Section 3.2.3) as research has shown that a combination of this data allows for higher performance when classifying organisms. By attaching a digital camera to either of these microscope systems, one can capture one or more digital images which can be used as input into a pattern recognition algorithm. The majority of the related work surrounding automatic identification of algae uses feature extraction methods where distinct descriptors are measured from a given image and then inputted into a machine learning algorithm such as a support vector machine or decision tree.

3.2.1 Brightfield Digital Microscopy

Brightfield microscopy is an imaging modality where a broad band light source is placed below the sample in order to illuminate it. Then a set of objective lenses are placed above

the same which focus the light into an eyepiece for viewing by a human, or onto a camera sensor for capturing a digital image. Brightfield microscopy is the most elementary and low-cost forms of microscopy [33]. Early work on automatic algae classification began in 1995 when Thiel *et al.* collected brightfield images of nine genera of blue-green algae and two genera of green algae, used 14 Fourier descriptors, 6 cell features, 7 moment invariants and 20 statistical features for texture, and used discriminant analysis to build a classifier [7]. They achieved 98.10% accuracy when classifying from these 11 different genera when training and testing on 158 samples. However, due to the small dataset and the omission of evaluating their discriminant classifier with a test set, the results likely showed an overfitted model to the training data.

Major work was done by Walker *et al.* in 2000 when they used a multiclass hierarchical classifier structure to properly classify four different species of Anabaena and two different species of Microcystis [8]. They measured 123 object features, including morphometric properties, object boundary shape properties, frequency domain properties, and second-order statistical properties and then used stepwise regression to find discriminatory features. Using a general Bayes decision function they achieved 97% accuracy when classifying. Although they only used cultured cyanobacteria, their work has shown that when enough data is present classifying multiple species is relatively accurate. In 2014, Promdaen *et al.* [34] developed a method to classify 12 different genera of microalgae with 97.22% classification accuracy, using data collected from a variety of brightfield microscopes. These genera included three types of toxic blue-green algae *(Anabaena, Oscillatoria, Microcystis)* as well as 7 genera of green algae, and one genus of Euglenoids. Feature extraction involved using Fourier descriptors, moment invariants, shape measures and texture features, while their classifier was a support vector machine.

Next, Coltelli *et al.* released a paper in 2014 in which they describe how they used a self-organizing map (SOM), an unsupervised learning method, to achieve 98.6% accuracy from a set of 53,869 images of 23 different microalgae representing the major algal phyla [6]. After acquiring the images and segmenting the algae, the RGB images were converted to the L*c*h* color space (lightness, chroma, and hue) and the morphological features were extracted. To recognize the different algae, these features were grouped into classes using clustering. Very recently, in 2019 Iamsiri *et al.* [35] used three geometric shape features (solidity, eccentricity, and convexity) as well as 13 features derived from a gray level co-occurrence matrix (GLCM) to train a support vector machine and achieve a classification accuracy of 91.30%. This was achieved while classifying five filamentous types of algae: *Anabaena*, *Oscillatoria*, *Spirogyra*, *Spirulina*, and *Anabaenopsis*.

Overall, supervised learning methods such as support vector machines (SVMs), naive Bayes, decision trees, and k-nearest neighbour have all been utilized to learn the optimal

classifiers for different genera of microalgae when imaged under brightfield microscopy. As inputs to these models the features used are primarily morphometric properties (diameter, area, convex perimeter, elongation, etc.), but also Fourier descriptors, moment invariants, and statistical texture features. It is important to note that all these methods are traditional feature extraction methods and do not leverage feature learning capabilities. In addition, active learning and unsupervised learning methods have also been utilized and have shown to be effective means of classification of different types of microalgae. Therefore brightfield microscopy coupled with different classifiers is a viable method to classify different types of cyanobacteria both to the genus level [7, 34, 6, 35], as well as to the species level [8].

3.2.2 Fluorescence Digital Microscopy

The most widely used form of fluorescence microscopy is known as epi-fluorescence microscopy, which was invented by Johan Sebastiaan Ploem (1927 - present) [36]. In this image modality a broadband light source, usually a mercury arc lamp, emits high energy light into a filter cube. The light entering the filter cube passes through an excitation filter which selects a narrow band of light which will reflect off a dichroic mirror and then focused onto to the sample using the objective lens. This high energy light causes the sample to fluorescence, resulting in lower energy light being emitted from the sample. This lower energy light passes through the objective lens and into the filter cube. Given the properties of the dichroic mirror, the low energy light passes through the dichroic mirror, and then is filtered by an emission highpass filter. Finally, this emission signal is then focused into an eyepiece for viewing by a human or onto a camera sensor in order to capture an image.

In 2006 Ernst *et al.* [37] developed an automated system to count filamentous *Planktothrix rubescens* using image processing. By using a single band epi-fluorescence setup they were able to identify and estimate the cell density for three different environmental samples and one cultured sample of *Planktothrix rubescens*. They found that the *Planktothrix rubescens* could be easily separated from algae when imaging under a fluorescence setup.

In 2016 Jin *et al.* used fluorescence microscopy to separate *Microcystis aeruginosa* [38] and *Anabaena flos-aquae* [39] from the background by exciting the respective samples at 546 nm and leveraging a Maximum Likelihood (ML) classifier. In both papers, Jin *et al.* then enumerated the samples and calculated size statistics. When comparing their results to the manual enumeration data using an hemacytometer they found their method achieved higher accuracy using much less time and resources. These papers [38, 39] illustrate the power of

using the auto-fluorescence of cyanobacteria to separate the cyanobacteria samples from the background. However, these methods only inspect a single species of cyanobacteria at a single fluorescence wavelength and do not explore using multi-band fluorescence microscopy to classify between different genus or species of cyanobacteria and other microalgae.

3.2.3 Brightfield & Fluorescence Digital Microscopy

As we have seen, both brightfield microscopy and fluorescence microscopy can be used to classify different genera of microalgae. However, when combining these two modalities we have the potential to improve our classification capabilities.

In 2006 Rodenacker *et al.* [40] developed a system called PLASA (Plankton Structure Analysis) that captured four fluorescence images and a brightfield image by using two fluorescent filters in tandem with a RGB camera. The first fluorescent filter cube had an excitation of 450 nm - 490 nm with a emission long pass filter of 515 nm. The second fluorescent filter cube had an excitation wavelength of 546 nm with a bandpass filter of 600 nm $\pm$ 40 nm. They also developed a segmentation system which fed into a feature extraction system in order to automatically identify different algae types. Hense *et al.* published a paper in 2006 using the PLASA system [40] which allowed for both bright-field and fluorescence images to be captured. They showed that autofluorescence information improves the discrimination between algae and non-algal objects and also distinguished between phycoerythrin (PE) containing algae and other algae [41]. One of the few important insights learned from Hense *et al.* is that thresholds based on fluorescent ratios opposed to fluorescent intensities proved more effective for discriminating between algae vs non-algae objects. In addition, they observed that under repeated excitation, chlorophyll-*a* and phycoerythrin emission intensity show an exponential decay, which they call fluorescent fading. In order to combat this fluorescent fading, appropriate image acquisition is required. Overall they found that the main benefit of using fluorescent information was the separation of algae species from non-algae species and large amounts of detritus; therefore Hense *et al.* recommend using multiple fluorescent features.

Walker *et al.* came to the similar conclusion that if accurate species level classification is required, it is necessary to capture both fluorescence and brightfield images [42]. Therefore, in 2002, Walker *et al.* illustrated that by capturing a single fluorescence image and a single brightfield image over 97% classification accuracy is possible when looking for *Anabaena sp.* and *Microcystis sp.* in natural populations found in Lake Biwa, Japan [42]. Without the use of fluorescence imaging the automated analysis of microalgae in the sediment saturated samples is nearly impossible. In order to achieve this high accuracy they

accomplished image registration using template matching and region growing techniques. As a result 120 different features (morphometric, boundary shape, frequency domain, and second-order statistics) were found and used to build a general Bayes decision function with Gaussian distributions. However, they acknowledge that due to current hardware limitations this is not viable, and therefore they limited their research to a single fluorescence image and a single brightfield image. Walker *et al.* believe that future improvements in fluorescence imaging will enable a low-cost, automated, species-level analysis and classification of microalgae. In addition, they only explored a traditional feature extraction based method and they did not leverage modern feature learning methods such a deep neural networks.

3.3 Imaging Flow Cytometry

Imaging flow cytometry takes an existing microscope system and incorporates a flow system in order to increase the throughput of the system. This imaging flow cytometry is a hybrid of the speed and statistical capabilities of flow systems, combined with the imaging feature of digital microscopy [43]. This allows for more samples to be collected, resulting in a more representative distribution of a given algae population. Some common, commercially available imaging flow cytometry systems built for automated algae analysis are the Cytobuoy, the Flow Cytometer And Microscope (FlowCam) by Fluid Imaging Technologies, and the Imaging FlowCytobot (IFCB) by McLane Labs. As we will see, these flow systems are often acquired by research labs to investigate the efficacy of such a system to collect unique data which can be used in a machine learning model. Several important studies related to the automatic identification of algae using imaging flow cytometry are highlighted below.

Blaschko *et al.* conducted a thorough investigation using 982 images of 13 different classes of plankton collected from a FlowCAM device in 2005 [44]. They used five different groups of features which were simple shape, moments, contour representations, differential and texture features. They then evaluated the classification performance using the following algorithms: decision trees, naive Bayes, ridge regression, k-nearest neighbour, and support vector machines. In the end the best classifier, a SVM, was able to achieve only 71.08% accuracy.

In 2007 Heidi M. Sosik created the Imaging FlowCytobot (IFCB) to explore automatic classification of phytoplankton [45]. In this study the submersible device collected images which were used to extract features (size, shape, symmetry, texture, invariant moments, and co-occurrence matrix statistics) to be entered into a machine learning algorithm. They

trained and tested a 22 category classifier which achieved 88% overall accuracy. This revolutionary technology allowed an unbiased approach to classify large amounts of phytoplankton data.

In 2013, Colares *et al.* used active learning in order to boost the performance of a classifier from microalgae data which were captured using a FlowCAM system [46]. The FlowCAM system provided 26 different features to represent the data, however Colares *et al.* only used seven of those features from two datasets. The first dataset contained 1,526 images consisting of four classes: Flagellates (1,003 images), others (500 images), pennate diatoms (14 images) and mesopores (9 images). The second dataset is composed by 923 images consisting of four classes: Pennate Diatoms (112 images), Flagellates (669 images), gymnodinium (65 images) and prorocentrales (77 images). Two metrics, accuracy and maxF1, were used to evaluate the performance of their approach and their final Gaussian mixture model with expectation-maximization classifier had an accuracy of 92%.

In 2016, Corrêa *et al.* used supervised learning on FlowCAM data which consisted on an imbalanced dataset of 24 types of microalgae divided in 19 classes [47]. They used the Synthetic Minority Oversampling Technique (SMOTE) to achieve oversampling strategy to compensate for the imbalances of their data. They evaluated five different supervised classification schemes which included multilayer perceptrons, the naive Bayes classifier, decision trees, and k-nearest neighbour (kNN). The input to these classifiers were ten manually-chosen features selected from the FlowCAM data. Three metrics were evaluated to produce the results which where Kappa, MaxF1 and MinF1. The best performance classifier was a kNN and has the scores of 98.1%, 98.2%, and 98.2% for Kappa, MaxF1 and MinF1 respectively.

Newer methods of imaging flow cytometry are beginning to use deep learning models to automatically identify algae. In 2018, Görögcs *et al.* [48] used a deep learning approach with a portable imaging flow cytometer in natural water samples. They captured diffraction patterns of flowing microorganisms and then use object detection and deep learning-based hologram reconstruction. They used this device for the detection of microplankton and nanoplankton ocean in samples along the Los Angeles coastline. By combining the ability of computational algorithms with deep learning Görögcs *et al.* were able to create a cost-effective, high-throughput flow cytometer to monitor algae.

3.4 Spectral Methods

Having covered the microscopy based approaches we will now shift to systems that have no spatial component, but only a spectral component in the data. Spectral based methods

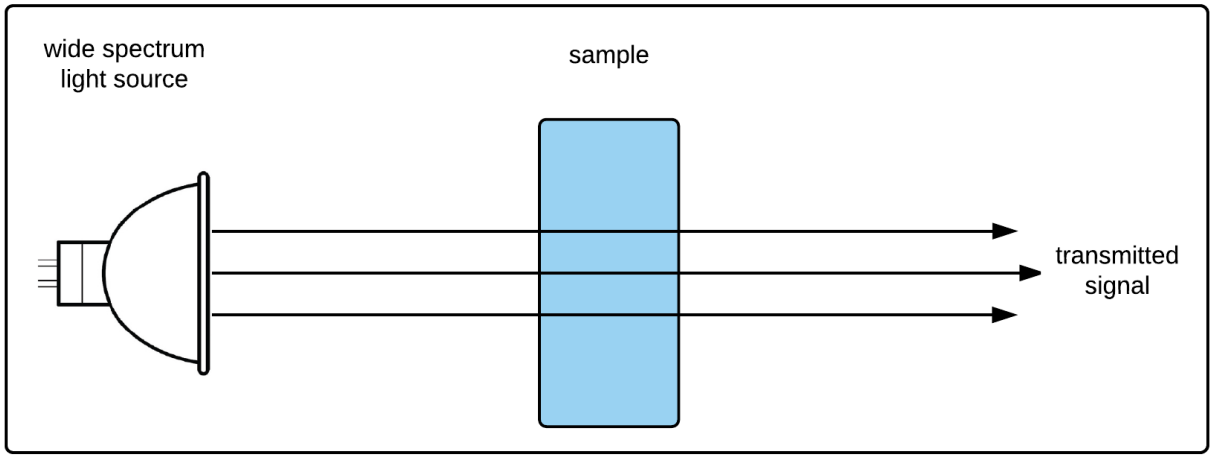


Figure 3.2: The absorption of light of a solution can be measured by passing a known broadband light source through the solution and measuring the transmitted signal with a spectrometer.

can be further broken down into absorption spectroscopy methods (Section 3.4.1) and fluorescent spectroscopy methods (Section 3.4.2). A spectrometer is a device that has the ability to measure the relative intensity of different wavelengths across the electromagnetic spectrum. Since the light-matter interaction is unique for different substances, light will reflect, absorb, fluoresce and transmit at different intensities for different wavelengths. Spectral methods are most accurate when the sample is assumed to be homogeneous: care is taken that only one type of algae is present. Due to spectral mixing, which occurs when different algae are present in the same sample, the spectrum of each is combined into a single signal, making it difficult to determine the relative amount of each organism.

3.4.1 Absorption Spectroscopy

As seen in figure 3.2, the absorption spectra of a solution can be measured by passing a known light source through a sample, and then measuring the transmitted light by a spectrometer. The absorbance of a material, denoted by A is defined as:

$$A = log_{10}\frac{\Phi_i}{\Phi_t} = -log_{10}T, \tag{3.1}$$

where, Φ_i is the radiant flux incident on the material, Φ_t is the radiant flux transmitted by the material, and T is the transmittance of the material.

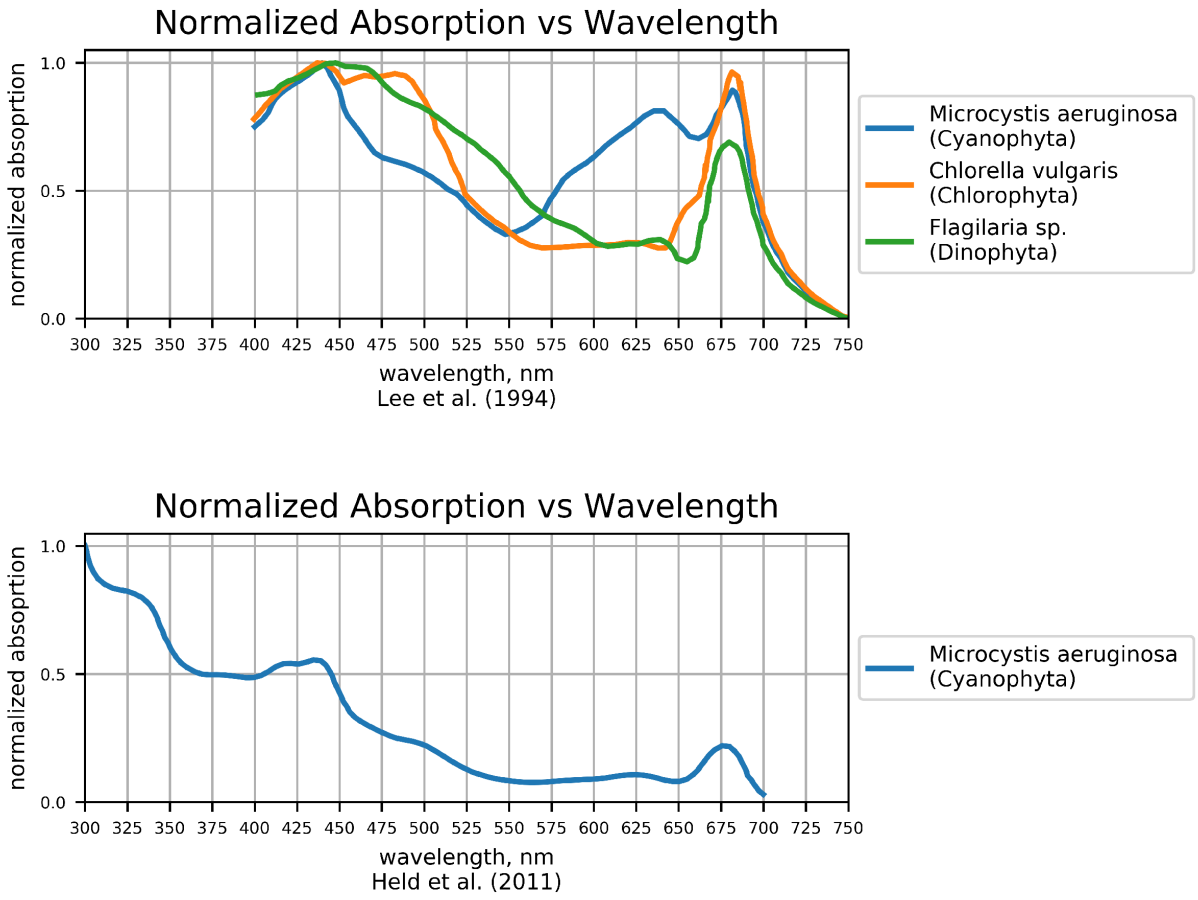


Figure 3.3: The absorption of algae for different phyla groups reported by Lee *et al.* (top) [49] and Held *et al.* (bottom) [50].

The absorption of different algae, as seen in Figure 3.3, has been described in two papers. The first paper is by Lee *et al.*, who in 1994 measured three algae, each from a different phyla group [49] from 400 nm - 750 nm. These algae were from the Cyanophyta phylum, Chlorophyta phylum and the Dinophyta phylum. As seen from Figure 3.3 all three algae peaked in the 400 nm - 475 nm range as well as in the 650 nm - 725 nm range. The Cyanophyta also had a peak in the 575 nm - 650 nm range These spectral responses all match the pigmentation data presented earlier in Table 3.1 as only Cyanophyta are reported to have phycobilin pigments.

In addition, Held *et al.* reported the absorption spectra of *Microcystis aeruginosa* in

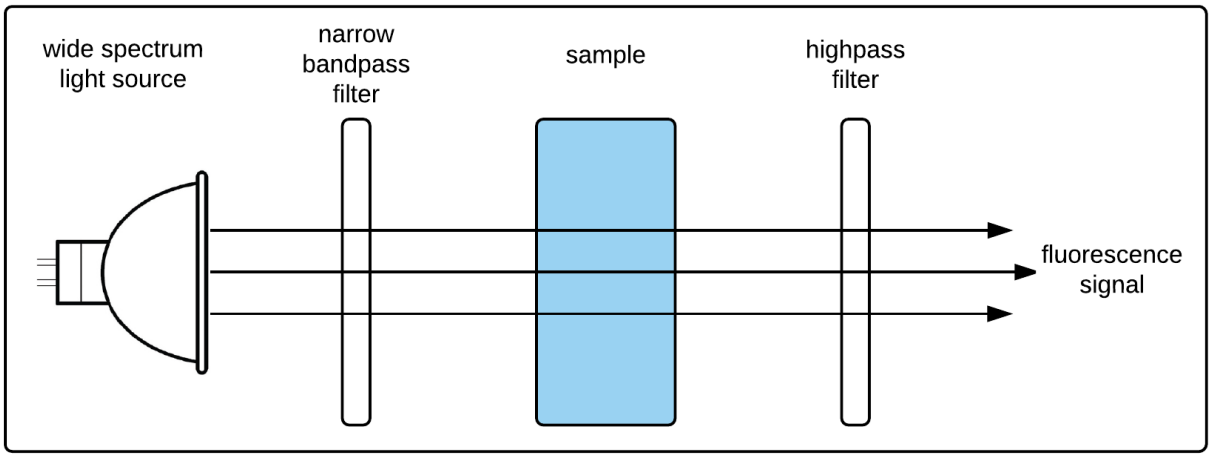


Figure 3.4: A fluorescence spectra can be measured by using a broadband light source which passes through a narrow bandpass filter to isolate the excitation wavelengths. This excitation light enters the sample, causes the sample to fluoresce, and emits a lower energy light signal. This lower energy light gets filtered by an additional highpass filter and then measured by a spectrometer.

2011 [50]. This spectrum was measured from 300 nm - 700 nm and has similar trends as the data reported by Lee *et al.* [49], but with a slightly different ratio of peaks. It is also interesting to observe the strong absorption of this algae in the ultraviolet part of the electromagnetic spectrum.

These spectra show that there is the potential to discriminate between different algae phyla groups when using absorption spectra methods. However, as seen in Section 3.4.2, most methods of spectroscopy use fluorescence based methods, where the excitation and emission spectra of different algae tend to range much more than the absorption spectra of different algae.

3.4.2 Fluorescence Spectroscopy

As seen in Figure 3.4, fluorescence spectroscopy involves using a light source and a narrow bandpass filter to excite a sample at a high energy, corresponding to a lower wavelength and then measuring the emission spectra at a lower energy, using a highpass filter, as seen in Figure 3.4. The excitation wavelength excites the sample by absorbing a photon of light and causes an electron to jump in energy state. When this electron then falls from this higher

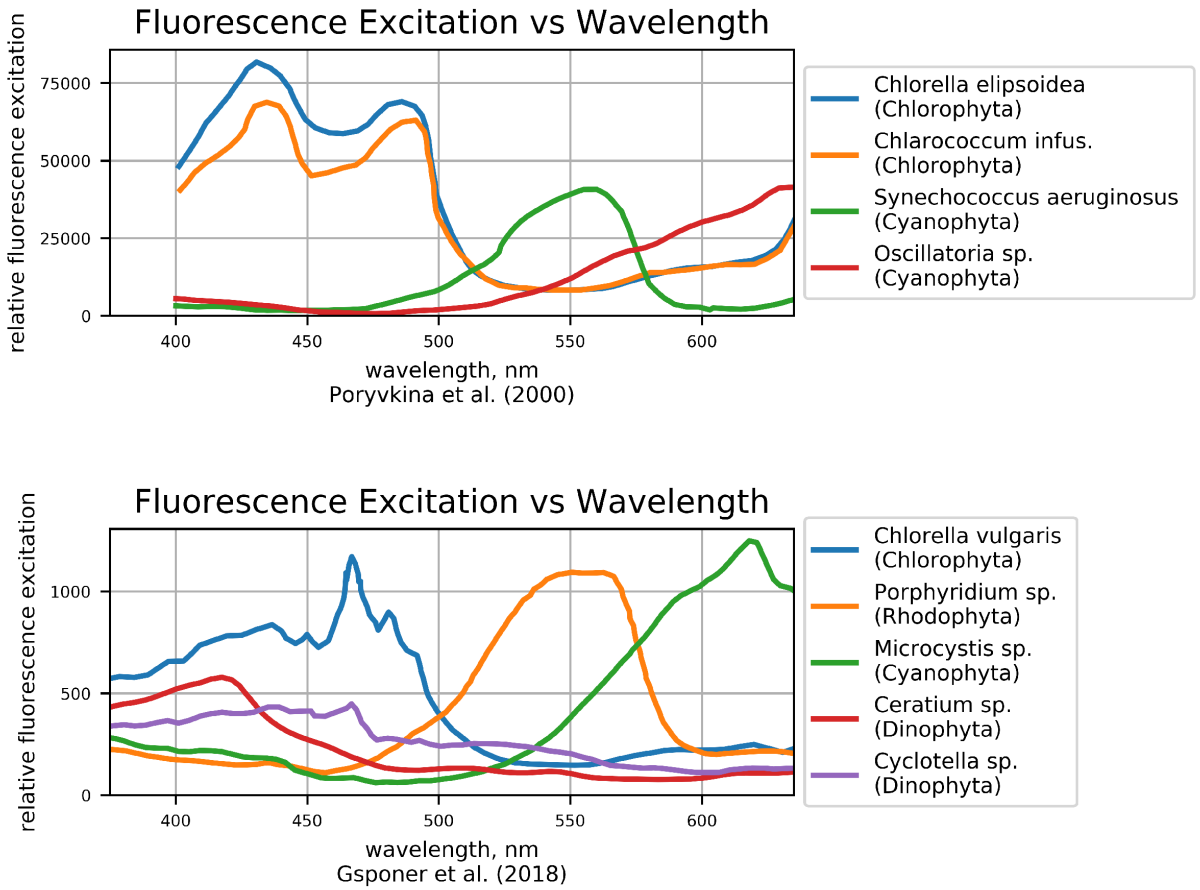


Figure 3.5: The excitation spectra from 400 nm - 650 nm was reported by Poryvkina *et al.* [51] (top) and Gsponer *et al.* [52] (bottom). These spectra reveal significant differences in the excitation wavelengths for different phyla of algae.

state, the energy is emitted as light at a higher wavelength. This is, in fact, the same technique that epi-fluorescence methods (Section 3.2.2) use. Since fluorescence spectra have both an excitation and corresponding emission spectra, both will be discussed. Some papers only present either the excitation or emission spectra while others present both.

Excitation Spectra

Poryvkina *et al.* [51] measured the excitation spectra of 31 algae species from seven phyla. In Figure 3.5 (top), four of the 31 excitation spectra can be seen from two of the phyla (Chlorophyta, Cyanophyta). In addition, Gsponer *et al.* [52] measured the norm excitation spectra of five algae species from four phyla (Chlorophyta, Rhodophyta, Cyanophyta, and Dinophyta) as seen in Figure 3.5 (bottom). Both of these studies [51, 52] demonstrate that different algae types have unique fluorescence excitation spectra from each other caused by the difference in pigments in a given algae type, as well the relative concentration of a given pigment, once again matching the different pigments from Table 3.1.

Another very interesting observation from Figure 3.5 is that excitation spectra differ significantly compared to the absorption spectra seen in Figure 3.3. This reveals the potential use of fluorescence data as a method to achieve high separability as well as a method of classifying different algae types. In fact, we will leverage this information when designing our own hardware system in Chapter 4 and then report the corresponding classification accuracy of using only fluorescence spectra in Chapter 7.

Emission Spectra

To determine this we must first look at work previously published by French *et al.* [53] and Millie *et al.* [54] as seen in Figure 3.6.

French *et al.* [53], as seen in Figure 3.6 (top), measured the emission spectra of *Porphyridium cruentum*, a type of red algae, from 570 nm - 750 nm when exciting the suspended sample at 11 different narrow-band wavelengths from 405 nm - 546 nm by using a high pressure mercury lamp and additional optics. Figure 3.6 (top) illustrates that when plotting three of the eleven emission curves the fluorescence emission increases when going from 490 nm to 515 nm to 546 nm. In fact, this matches the work presented by Gsponer *et al.* [52], whose data can be seen in Figure 3.5 (bottom), since the only species from the Rhodophyta phyla peaks around 550 nm.

Millie *et al.* [54]. excited five different samples of microalgae from four different phyla at either 440 nm or 490 nm. In Figure 3.6 (bottom), four samples, from three different phyla, are presented, where the spectra were normalized to a range of zero and one. Each of these fluorescence emission spectra from the presented three phyla (Cyanophyta, Dinophyta, and Chlorophyta) are relatively unique as they each have a distinct peak wavelength. Therefore, as expected, the emission spectra for different phyla are relatively unique which is caused by the distinct pigments in each phyla as previously shown in Table 3.1. Furthermore, this information can be leveraged when designing and building an imaging system.

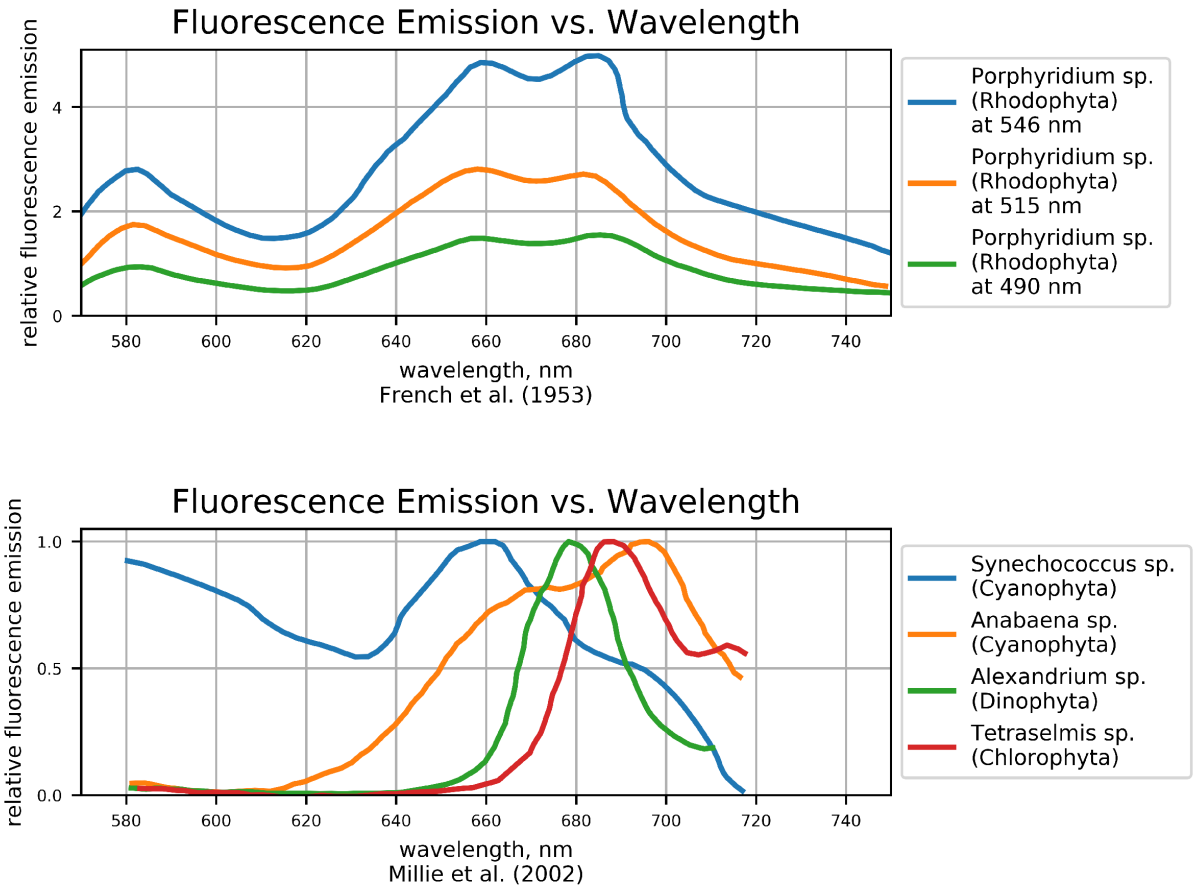


Figure 3.6: The emission spectra of *Porphyridium sp.* by French *et al.* [53] (top) and the emission spectral of four common algae from three phyla groups by Millie *et al.* [54]. These spectra show that different algae have different emission spectra which is the result of these algae having different different pigments. Notice the wavelength range in this figure compared to the wavelength range in Figure 3.5. Since the emission spectra is at a lower energy, the spectra will be at a higher wavelength.

In addition to measuring the excitation spectra of five different types of algae, Gsponer *et al.* [52] took one sample (*Microcystis sp.*) and also measured the emission spectra. As seen in Figure 3.5 (bottom) and Figure 3.7, *Microcystis sp.* has a peak excitation wavelength between 600 nm and 625 nm. As seen in Figure 3.7, the corresponding peak emission wavelength is approximately 680 nm.

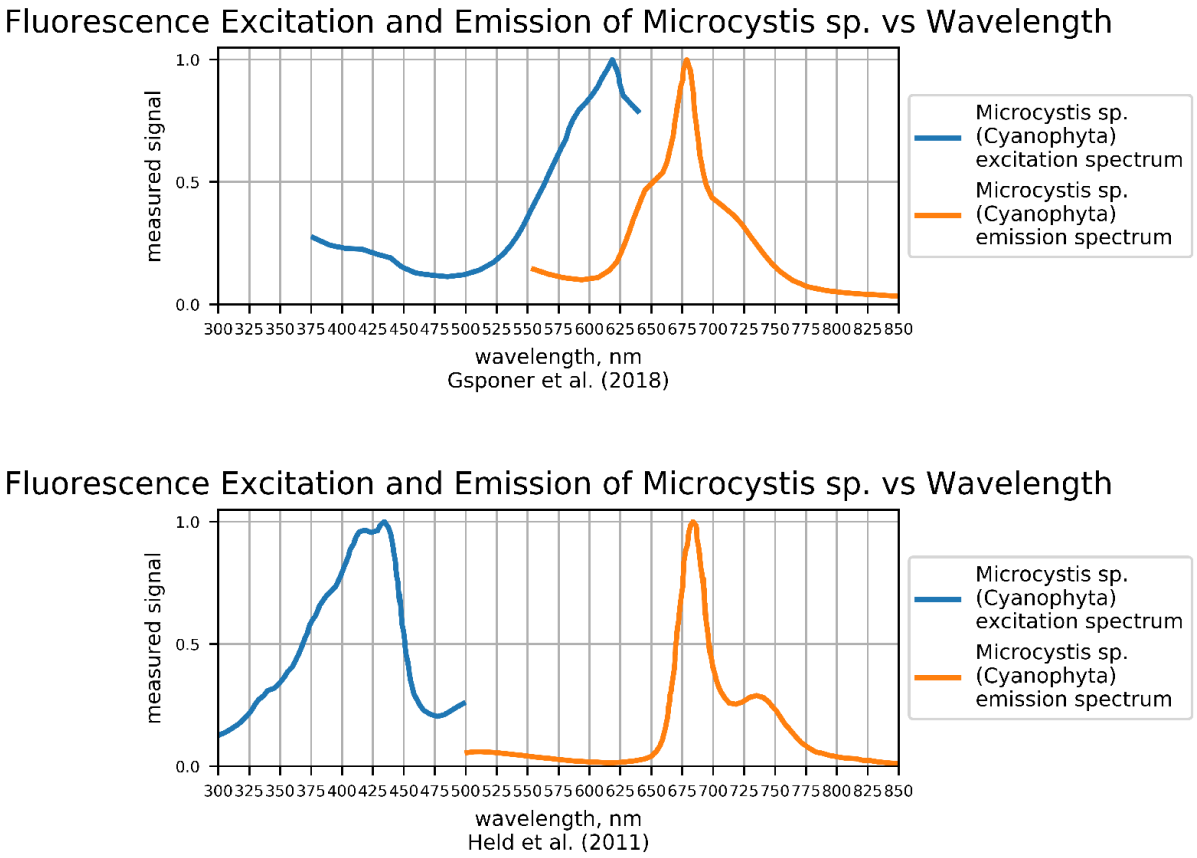


Figure 3.7: The excitation and emission spectra of *Microcystis sp.* measured by Gsponer *et al.* (top) [52] and by Held *et al.* (bottom) [50]. While the emission spectra is very similar, the excitation spectra varies between these two authors.

In addition to these reports about excitation and emission spectra, work has been done in exploring a fluorescence-based approach to identifying algae. In 2002, Beutler *et al.* built a custom device that used five distinct wavelength LEDs (450 nm, 525 nm, 570 nm, 590 nm, and 610 nm) to excite different pigments in five different algae spectral groups, which were: green algae (Chlorophyta), glue-green algae (Cyanobacteria), brown algae (Bacillariophyceae and Dinophyceae) and mixed algae (Cryptophyceae). These LEDs excited the Chlorophyll *a* pigments as well as other antenna pigments such as Chlorophyll *c*, phycocyanobilin, phycoerythrobilin, fucoxanthin and peridinin [55]. By placing a bandpass filter between the algae sample and the sensor they were able to take five measurements

of the emission signal from the microalgae between 650 nm - 750 nm by turning on each consecutive LED. Beutler *et al.* fitted a norm spectrum to each of the spectral classes and found that each spectral group was differential from each other. They then proceeded to use this new information to determine the chlorophyll concentrations of different diluted samples of each spectral group and compared this to the known chlorophyll concentration and found very high correlation.

Gregor *et al.* also used in vivo fluorescent methods as a tool to quantify phytoplankton organisms in water in 2007 [56]. They found that phycocyanin fluorescence was a sensitive indicator of the presence of cyanobacteria in water since the phycocyanin pigment in the cyanobacteria is excited around 590 nm to 630 nm, and has a max emission around 650 nm. This is unique behaviour compared to eukaryotic algae which are excited around 430 nm to 530 nm and has their peak emission 685 nm, which is also the peak of Chlorophyll *a*. When taking the ratio of the two fluorescent measurements in all samples they found a very strong correlation to the true ratio of cyanobacteria to eukaryotic algae.

Hu *et al.* utilized the fact that different algae species have different ratios of antenna pigments, which results in different fluorescence emission spectra [57]. Hu *et al.* imaged twenty different algae from six algae divisions (Dinophyta, Bacillariophyta, Chrysophyta, Cyanophyta, Cryptophyta, and Chlorophyta) by illuminating the samples at four different wavelengths (440 nm, 470 nm, 530 nm, and 580 nm), and then measure the emission spectra from 600 nm - 750 nm with a 5 nm resolution. In order to create a norm vector of each algae division, each of these four emission spectra were concatenated to form a single feature vector. Then, using a discrimination method established by multivariate linear regression and weighted least-squares, it was shown that each of the norm vectors from each phylum were independent from each other. This insight now allowed Hu *et al.* to measure the spectra of a mixed sample and predict the relative ratios of different algae phyla when comparing samples with two different species from different phyla.

In summary, spectrometers that measure multiple fluorescent wavelengths have the ability to differentiate between different phyla of algae, as seen in Beutler *et al.* [55] and Hu *et al.* [57]. Therefore spectral based approaches are suitable when looking at pure algae types to determine which phylum the sample is from.

3.5 Fluorescent Probes

Just like imaging flow cytometry added a flow element to microscopy based approaches, fluorescence probes takes the spectrometer approach and applies the theory to build an

in-situ monitoring device. Based on these underlying fluorescent properties, a number of probes have been developed and are currently on the market.

For example, McQuaid *et al.* used a YSI 660 V2-4 water quality multi-probe in their experiments. This YSI probe was designed to measure the cyanobacterias phycocyanin pigment at 590 nm (with a passband of 565 nm - 605 nm) and measures the pigments emission at 660 nm ± 20 nm [58]. They found that the phycocyanin probe can be used to monitor cyanobacterial biovolume in surface water when the cyanobacterial blooms were dominated by *Microcystis sp.* and microcystin. Furthermore, they found that a taxonomist analysis was found to underestimate the risk of a microcystin contamination, due to less frequent sampling.

However, in more recent years, additional studies have been conducted on these probes. In 2012, Zamyadi *et al.* took five different probes, some of which were YSI probes, and found that there was no correlation between a given probe's reading and the true cell count in a given sample [59]. However, the authors did find that the correlation between the probe's readings and the total biovolume in the sample could be trusted.

Zamyadi *et al.* continued their work and in 2016 tested six different in-situ fluorometric probes from major brands such as bbe, TriOS, Tuner Designs, and YSI [60]. One conclusion from their research is that a major disadvantage of all these probes is the fact that they cannot distinguish between species of cyanobacteria and that the use of a microscope would be required to accomplish this task. Furthermore, it was highlighted that access to real-time data and automatic frequent sampling (at least every 60 minutes) is a major advantage of the probes and a highly desired functionality of a monitoring device.

Finally, Bowling *et al.* also used a YSI EXO2 fluorometric probe to measure the chlorophyll *a* and phycocyanin and found that a good correlation between phycocyanin and total cyanobacterial biovolume in two of the three ponds they investigated [61]. They also found that phycocyanin did not correlate well with cell counts, and Chl-a was a poor measure of cyanobacterial presence.

Therefore, probes that measure the Chl-a and phycocyanin can estimate the total biovolume in a cyanobacteria bloom, but are very poor at estimating actual cell counts of different species of cyanobacteria in the samples [56, 58, 59, 60, 61]. In order to determine the actual cell count of different organisms in a water sample, microscopy methods must be utilized.

In addition, these probes are known to underperform when not in ideal conditions, as stated by the state of Ohio:

> "Phycocyanin is a pigment unique to cyanobacteria. Sensors are available which

measure the presence of this pigment and report in either relative fluorescence units (RFUs) or cyanobacteria concentrations in cells/mL. The cell concentration data, however, should be used with caution because sensors are typically calibrated to a pure *Microcystis* culture, and *Microcystis* may not be the dominant cyanobacteria in the water source. Also, other factors such as turbidity and overall light availability can impact the amount of phycocyanin that is produced per cyanobacterial cell. It is often best for a water system to review the general changes in RFUs over time as an indication of an increase in bloom severity instead of a particular cell/mL reading."[62]

3.6 Genomics

Recently the major increase in genetic data has propelled the birth of DNA taxonomy for algae where species are classified by sequencing the precise order of nucleotides within a DNA molecule [63]. For DNA taxonomy to be successful a database of prior DNA sequences must first be built, and then DNA barcoding methods can be utilized to match the newly sequenced DNA to the known database. However, one disadvantage of DNA sequencing is that it can be difficult to match the new DNA sequence with the standard Linnaean taxonomical names.

The most recent methods are genomics based methods, which primarily consist of quantitative polymerase chain reaction (qPCR). Water utilities are beginning to use commercial molecular assays, however, since these molecular techniques require specialized equipment and training to run, only a select few drinking water treatment plants can leverage this new technology [64]. In summary, as explained by Clerck *et al.*, while molecular based technique are promising, the technology needs to advance in both the speed and success rate before it can be a viable option for algae taxonomy [63].

3.7 Summary of Methods

As we have seen in this chapter, there are five main methods that are currently being used for automated identification of algae:

1. Digital Microscopy (Section 3.2)
2. Imaging Flow Cytometry (Section 3.3)
3. Spectral Analysis (Section 3.4)
4. Fluorescent Probes (Section 3.5)
5. Molecular Methods (Section 3.6)

Each of these methods have advantages and disadvantages. Digital microscopy approaches and imaging flow cytometry have the advantage of capturing spatial information but lose the spectral data. Furthermore, the majority of these microscopy based approaches only capture a single brightfield image. Only a few researchers have begun to look at multispectral imaging and combining brightfield with fluorescence imaging. Spectral based methods and fluorescence probes measure the spectral fluorescence properties of algae, however, they don't capture any spatial information. While this can achieve phylum level classification when inspecting a simple water sample, fluorescent probes are not dependable when inspecting more complex samples. Finally, genomics based methods show promise as an emerging technology, with the restriction that it requires specialized equipment and training in order for the method to be effective.

Given all these trade-offs, in Chapter 4, we will present a novel cost-effective system that captures a single brightfield images as well as multiple fluorescence images. To complement this new instrument, a software framework will be presented in Chapter 5 in order to process the data captured by the proposed imaging system. Then, in Chapter 6, a dataset will be created in order to test these different contributions. Finally, the proposed imaging system is used to test efficacy of using multiple fluorescence images when automatically identifying algae. These results will be presented in Chapter 7.

Chapter 4

Imaging System Design

"I begin with an idea and then it becomes something else."

– Pablo Picasso (1881 - 1973)

All five approaches presented in Chapter 3 have shown to be effective methods to automatically identify algae in a water sample. However, to date each method has been fairly independent from the other. However, the research presented here proposes to combine two of these methods, the spatial element from digital microscopy with the spectral element of fluorescent spectroscopy. By fusing these, the spectral and spatial information, our system is able to capture salient information from both modalities.

In Section 4.1 the optical design requirements will be presented. These include the design requirements to capture a single brightfield image, as well to excite antenna pigments of algae to generate four fluorescence images. Additional requirements are to generate data on-site and analyze this data in real-time. Next in Section 4.2, three different fluorescence setups will be compared and the orthogonal fluorescence microscopy approach will be selected. Then in Section 4.3, the optical design configuration will be discussed. In this section the specific LED wavelengths and filters will be selected to allow the imaging system to capture both brightfield and multiple fluorescence images. In Section 4.4, the design will be implemented by creating a 3D printed frame to house the optics, camera and printed circuit board (PCB). In addition, a graphical user interface (GUI) will be created in order to control the imaging system. Finally, in Section 4.5, a summary of the final system will be presented.

4.1 Imaging System Design Requirements

When designing a system to test the proposed method against existing methods careful consideration must be taken to determine what must be included and excluded from the system. In order to determine this, the hardware design requirements will based on the related work as discussed in Chapter 3. The requirements to build such a device can be broken down into the following sections:

1. Brightfield Requirements (Section 4.1.1)
2. Spatial Resolution Requirements (Section 4.1.2)
3. Multispectral Fluorescence Requirements (Section 4.1.3)
4. On-site Requirements (Section 4.1.4)
5. Real-time Analysis Requirements (Section 4.1.5)

4.1.1 Brightfield Requirements

As discussed in Chapter 2 and Chapter 3 a number of methods rely on the information in a single brightfield image. As discussed in Section 2.2, nearly all manual identification is done with a brightfield microscope. In addition, both brightfield methods (Section 3.2.1) as well as imaging flow cytometry (Section 3.3) regularly capture a single brightfield image for automated algae identification. In order to match current methods, the brightfield modality must be incorporated into the design of the system. This will allow a direct comparison between the baseline method of brightfield imaging against the proposed method of using multiple fluorescence images. Therefore it is a design requirement of the hardware system that it must be able to capture a single brightfield image.

4.1.2 Spatial Resolution Requirements

Algae can range in size from smaller organisms like single celled *microcystis aeruginosa* (3-7 um in diameter) to larger organisms such as *ceratium* (150 um in length) [21, 24, 25, 65]. As discussed in Section 2.1, *microcystis* is known to produce lethal toxins. In order for our system to be effective at monitoring HABs, it must be able to resolve smaller microalgae such as *microcystis*. Therefore, the first priority is to have a spatial resolution of at least

0.75 um/pixel, as in this situation smaller microcystis cells would be four pixels in width. However, given that there are many larger algae that are known to produce toxins, such as filamentous *Anabaena*, it is also desired to achieve the largest field of view possible.

4.1.3 Multispectral Fluorescence Requirements

As seen in in Section 3.2.2, early work using fluorescence digital microscopy only captures a single fluorescence image [38, 39]. However, as seen in Section 3.2.3, if accurate classification of algae is to be achieved, it is necessary to capture both fluorescence and brightfield images [42]. Work by Hense *et al.*, who used two fluorescence images, showed conclusively that using these two fluorescent bands improves the discrimination of algae from non-algal objects as well as phycoerythrin (PE) containing algae from others. Their final recommendations was to use multiple fluorescent features [41]. In addition, as discussed in Section 3.4, spectral based methods commonly collect more than two fluorescent spectral features but lose the spatial information provided by digital microscopy. Section 3.4 showed that different fluorescent signals yield separability between different types of algae. Beutler *et al.* collected five fluorescent bands to separate five algae spectral groups [55], and Hu *et al.* collected four fluorescent bands from six algae divisions [57]. Therefore a design requirement for the proposed system is that it must capture more than two fluorescent bands, where each band selects targets a unique pigment.

4.1.4 On-site Requirements

Another design consideration is the cost of the hardware of the system. Currently, on-site data acquisition is possible from systems such as the FlowCam Cyano [44, 46, 47] as well as the IFCB system [45]. However, these systems cost well over $100,000 USD and provide limited automated image analysis. As both the FlowCam Cyano and the IFCB are flow systems, removing the flow component dramatically reduces the costs. Even when removing the flow components, a standard epi-fluorescence microscope with camera system starts at $20,000 USD for lower end models such as AmScope and can easily cost $60,000 - $80,000 USD for an Olymupus, Zeiss or Nikon setup. To reduce the costs of the proposed system, this research aims to build a system for less than $10,000 USD. Therefore the goal of this research is to build a low-cost device which is affordable for rural communities and third world countries.

4.1.5 Real-time Analysis Requirements

The final design requirement is to process and automatically analyze the data generated from the system in near real-time. Having a near real-time analysis allows individuals and communities who were previously limited by cost and distance the opportunity to analyze their water bodies. For example, creating a low-cost system with real-time analysis allows individuals and communities in rural areas or third world countries to determine their water quality which is something they are currently unable to do due limited capital. Therefore, the final design requirement of the hardware system is that must generate and process data in real-time.

4.1.6 Summary of Requirements

In conclusion, a summary of design requirements for the system are as follows:

1. Capture a single brightfield image with a spatial resolution of at least 0.75 um/pixel.
2. Capture at least three fluorescence images where each image targets a different pigment.
3. Be low-cost (less than $10,000) as to remove any financial limitations of rural communities and third world countries to generate data on-site.
4. Automatically analyze the generated data in real-time to allow for real-time inspection of water samples.

Given these design requirements, Section 4.2 will present the fluorescence design considerations. These considerations must weigh the benefits of the different modalities of fluorescence imaging. These modalities are diascopic fluorescence microscopy (Section 4.2.1), epifluorescence microscopy (Section 4.2.2), and orthogonal fluorescence microscopy (Section 4.2.3).

4.2 Optical Design Considerations

The goal of the imaging system is to capture one brightfield image and multiple fluorescent wavelength images. There is only one method to capture a brightfield image, that is with the illumination source underneath the sample and with the camera sensor and objective lens above the sample. However, there are three main methods to collect fluorescent images:

1. diascopic fluorescence microscopy
2. episcopic fluorescence / epi-fluorescence microscopy
3. orthogonal fluorescence microscopy

The advantages and disadvantages of each method will be discussed for the specific application of building a low-cost device that can generate data in real-time.

4.2.1 Diascopic Fluorescence Microscopy

Diascopic fluorescence microscopy, also known as transmitted light fluorescence microscopy, is a method where the illumination source is placed beneath the sample, and then a excitation filter allows a narrow set of wavelengths of light to excite the sample. In the standard brightfield setup, both this excitation light as well as the emission light continue to an emission filter, which ideally only allows the emission light to pass onto a camera sensor. Theoretically this setup works but practically it is not a viable means to collect fluorescent data. Very high quality excitation and emission filters are required to ensure that no stray light at unwanted wavelengths enter into the optical path. If the excitation filter allows light to pass at higher wavelengths, any emission light from the sample will be lost since the emission light from the sample can be several orders of magnitude weaker than the intensity of the excited light [36].

Therefore the solution is to purchase high quality filters and switch to a darkfield transmission setup to block out additional stray light. However both these improvements drastically increase the price of the system. Using a darkfield setup with high quality filters introduces additional problems since only a fraction of the excitation light will excite the sample due to the darkfield setup. This results in exposure times of tens of seconds for a single capture and quickly becomes impractical given the design requirement to capture multiple fluorescence images. Smearing effects would occur in the image since the algae in the water sample are alive and can move within the sample. Different wavelength images would have organisms in different pixel locations due to the motion of algae over time. For these reasons the diascopic fluorescence setup does not achieve our design requirements.

4.2.2 Epifluorescence Microscopy

Episcopic fluorescence or epifluorescence microscopy is the most common method to capture fluorescence data and was invented by Johan Sebastiaan Ploem in 1967 [66]. As described by Ploem himself:

> "A fluorescence microscope is designed to provide an optimal collection of the fluorescence signal from the specimen, while minimizing the background illumination consisting of unwanted excitation light and autofluorescence. This requires rather sophisticated technology, since the specimen can be several orders of magnitude weaker than the the intensity of the excited light."[36]

This setup uses a filter cube to direct the excitation light through the objective lens onto the sample. The main advantage is that all the excitation light that passes through the filter cube gets concentrated directly onto the field of view of the imaging system. Therefore the capture times for a given fluorescent image are in the normal range of milliseconds, making it a viable setup for our application. However the two main drawbacks of this setup are: (1) it requires additional optical components such as a dichroic mirror as well as an infinite corrected objective lens which increases the cost of the system, and (2) it requires moving parts to switch between different filter cubes when collecting multiple fluorescence images at different excitation wavelengths. Once again, given that algae move within a water sample, this creates blurring and motion artifacts in the multispectral images. For these reasons the epifluorescence setup is also unfit to achieve our design requirements.

4.2.3 Orthogonal Fluorescence Microscopy

The third method may be less common, yet it is valid as it is able to produce a fluorescence signal. This method called orthogonal fluorescence microscopy places the excitation source orthogonal to the sample. This removes the issue of stray light entering the optical path as in the diascopic fluorescence setup, and it removes the need for the dichroic mirror in the epifluorescence setup. The first reported use of this setup was in 1902 when Richard Adolf Zsigmondy and Henry Siedentopf created a "ultramicroscope" in order to observe single gold molecules [67]. More recently side-on illumination is used in light sheet fluorescence microscopy (LSFM) where planar illumination techniques are used to send a sheet of light, usually a few hundred nanometers to a few micrometers, through a sample [68]. LSFM is an advanced and expensive method of fluorescence imaging, but the concept of side-on illumination is a valid one. It is a suitable setup to achieve our design requirements as it doesn't have the issue of stray light (as in the diascopic fluorescence setup), and it doesn't require a dichroic mirror (as in the epifluorescence setup).

For these reasons it is the most suitable optical configuration considering the design constraints. However, if we were to use a standard broadband light source for fluorescence excitation, such as a mercury arc lamp, then a filter wheel would be required to select the desired excitation band. Even worse, a second filter wheel would be needed in the

orthogonal fluorescence setup in order to filter out the desired emission spectra. The main advantage of the epi-fluorescence setup is that it combines the excitation filter, dichroic mirror, and emission filter into one unit, called the filter cube. This results in only one filter wheel in the optical system. So the design choice to use orthogonal fluorescence does no longer seem to be the optimal setup. However, as seen in Section 4.3, we area able to completely eliminate both these filter wheels, removing any moving components from the optical system.

4.3 Optical Design Configuration

Having chosen the optimal fluorescence setup, that is, orthogonal fluorescence microscopy, Section 4.3.1 will discuss a more detailed design of this setup. Specifically, the LEDs and specific filters will be decided based on the spectral information presented in Chapter 3. In Section 4.3.2, the brightfield LED and filter setup will be discussed. Finally, in Section 4.3.3 the main benefits of this design will be highlighted.

4.3.1 Fluorescence Optical Configuration

As previously discussed, the orthogonal fluorescence setup is the optimal choice since it allows for fast acquisition times and it removes the need for a dichroic mirror. The problem to be discussed is the removal of moving parts while still capturing a fluorescence signal.

In order to remove the excitation filter wheel, the broad band light source will be replaced with high-power narrow-band LEDs. To determine the optimal LED wavelengths, an understanding of what antenna pigments in each algae group is required. SooHoo *et al.* [69] showed that for different marine phytoplankton the highest fluorescence yields occur in the blue-green region of the spectrum, corresponding to bands of peak absorption by the accessory pigments. They confirmed this for five different species: one diatom, two dinoflagellates, one chrysophyte (golden algae), and one chlorophyte (green algae).

As discussed in Chapter 3, different algae groups are known to have different excitation peaks based off of the photosynthetic pigments in their structures. These excitation peaks range from approximately 400 nm to 650 nm as seen in Figure 4.1. Figure 4.1 incorporates the excitation spectra presented by Poryvkina *et al* [51] and Gsponer *et al.* [52] with the emission data presented by French *et al* [53] and Millie *et al.* [54] which was discussed in Chapter 3.

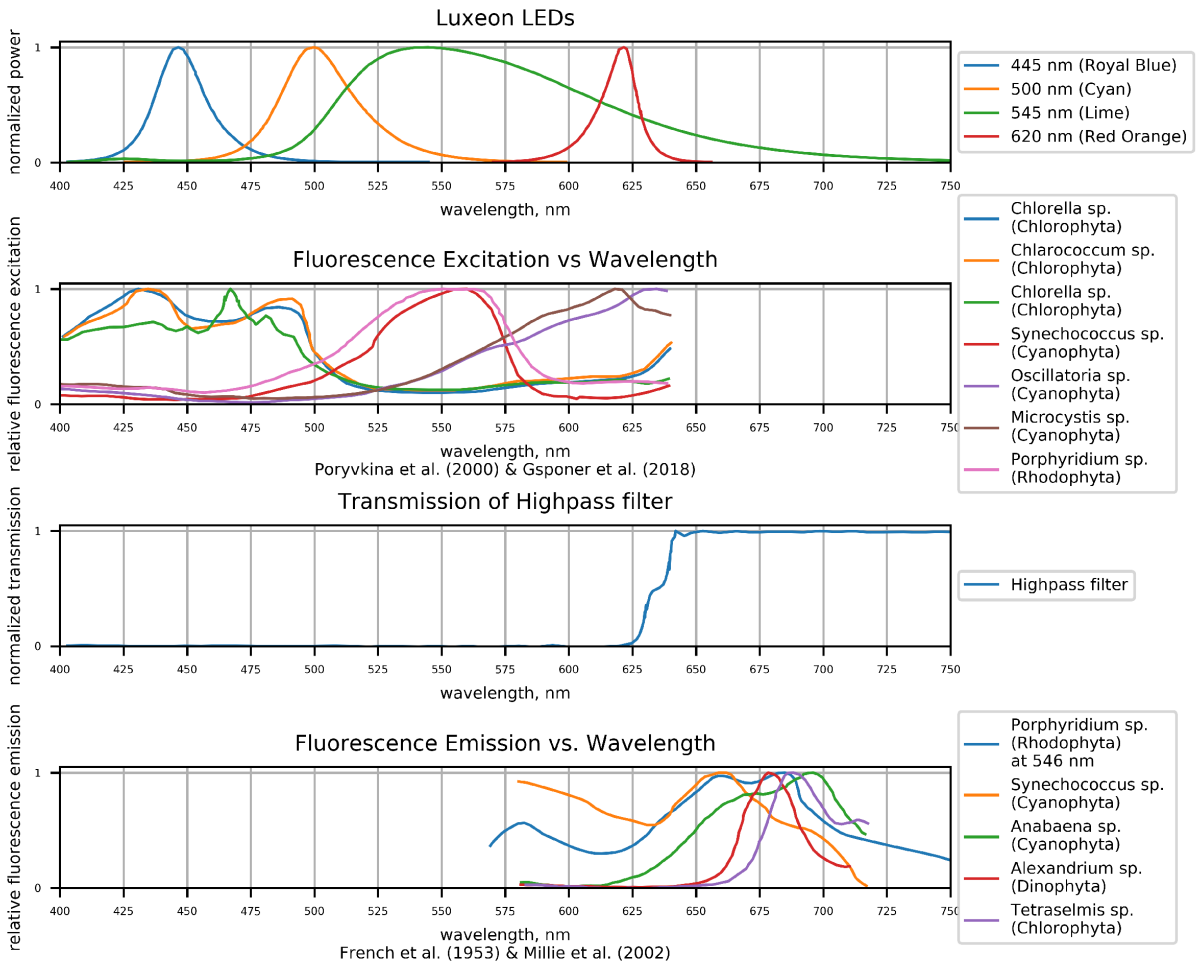


Figure 4.1: The optimal LEDs (top) where chosen based from the excitation spectra presented by Poryvkina *et al* [51] and Gsponer *et al.* [52] (second from top). Furthermore, the transmission of the high-pass filter (third from top) was selected based off the emission data presented by French *et al* [53] and Millie *et al.* [54] (bottom).

Therefore, as seen in Figure 4.1, four Luxeon high-power LEDs where chosen to excite the major pigment peaks in this range. The wavelength peaks of these LEDs are 445 nm (Royal Blue), 500 nm (Cyan), 545 nm (Lime), and 620 nm (Red Orange). There are two main advantage of using narrowband LEDs, opposed to a traditional fluorescent lamp or mercury arc lamp. First since the LEDs are already at a specific wavelength, and therefore the excitation filter is now longer required to isolate the excitation light from the broadband

light source. Secondly, LEDs can rapidly switch on and off, something broad band light sources are unable to do. In addition, the narrow wavelength of these LEDs remove the need of any mechanical filter wheel, allowing for faster acquisition rates. Compared to broad band light sources LEDs also waste less heat and have longer lifetimes. Therefore narrow-band LEDs are far superior compared to using a broadband light source with a filter wheel.

The removal of the emission filter wheel resulted from gaining insight into the excitation and emission spectra of algae. As seen in Figure 4.1, the majority of the emission spectra of algae range from 575 nm to 750 nm, and therefore only a single highpass filter at 635 nm is needed between sample and the sensor. From data presented in Chapter 3, we know there is often fluorescence emission below the 635 nm cutoff, however, Figure 4.1 demonstrates that for different phyla groups there is still adequate amounts of signal above the 635 nm cutoff. It is also important to note that a highpass filter is still required in the system. If this highpass filter were not placed in the optical path of the fluorescence emission signal, it would get mixed with any diffuse or specular reflections caused by the interaction of the LED light and the sample.

4.3.2 Brightfield Optical Configuration

Since the design of the fluorescence optics must be incorporated into the same design of the brightfield optics, only the range of 635 nm or greater can be observed in brightfield mode. This is because the highpass filter will block any light less than 635 nm, and transmit any light that is greater than 635 nm. However, the optical design uses the fact that the absorption spectrum of algae, as seen in Figure 4.2, has two major peaks as reported by Lee *et al.* [49]. The first peak, from 400 nm to 525 nm will not be observed, as that signal is blocked by the highpass filter. However, the second peak, from 650 nm to 725 nm will be observed, as that signal will pass right through the highpass filter. By placing a 2700 K LED directly underneath the sample, a broadband LED can interact with the pigments above 635 nm and a brightfield image can be captured.

It is important to note that the three absorption spectra shown in Figure 4.2 (middle) are from three different phyla groups: Cyanophya, Chlorophyta, and Dinophyta [49]. This same information was also shown in Figure 3.3 where it was discussed that the absorption spectra for different phyla groups are relatively similar in nature. This observation allows a highpass filter to be used at 635 nm as different phyla groups have a significant absorption response above 635 nm.

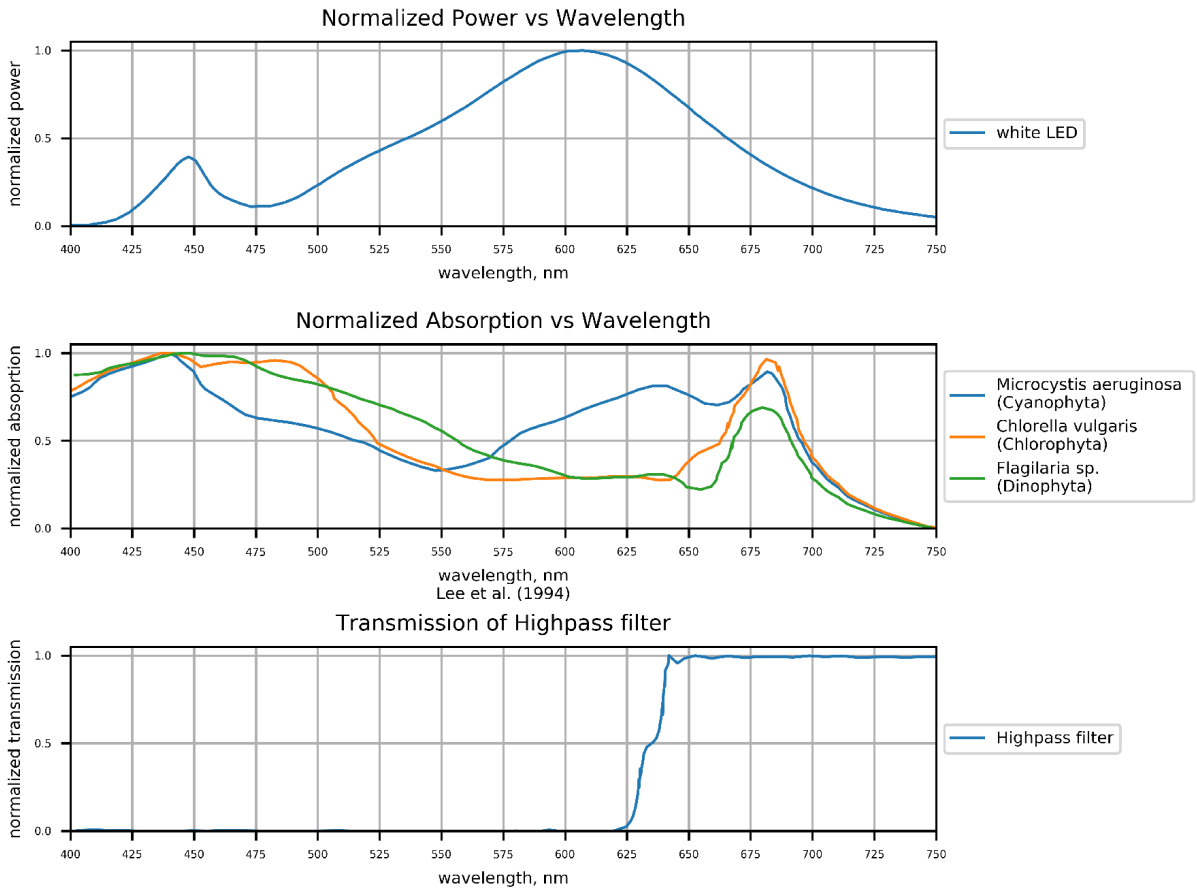


Figure 4.2: A 2700 K LED (top) is used as a light source for the brightfield imaging modality. This LED will interact with the absorption spectra of different algae, a few of which are presented by Lee *et al.* [49] (middle). The highpass filter (bottom) then blocks any light lower than 635 nm.

4.3.3 Optical Configuration Benefits

In this manner the identical setup, as seen in Figure 4.3, can be used to rapidly capture both multispectral fluorescence images and a single brightfield image. The orthogonal fluorescence setup with high-powered LEDs and a single highpass filter dramatically reduces the costs of the overall system in comparison to standard epi-fluorescence setup. This system allows for four spectral images at four different excitation wavelengths to be captured in a rapid manner without any moving parts. Each of these four spectral images measure

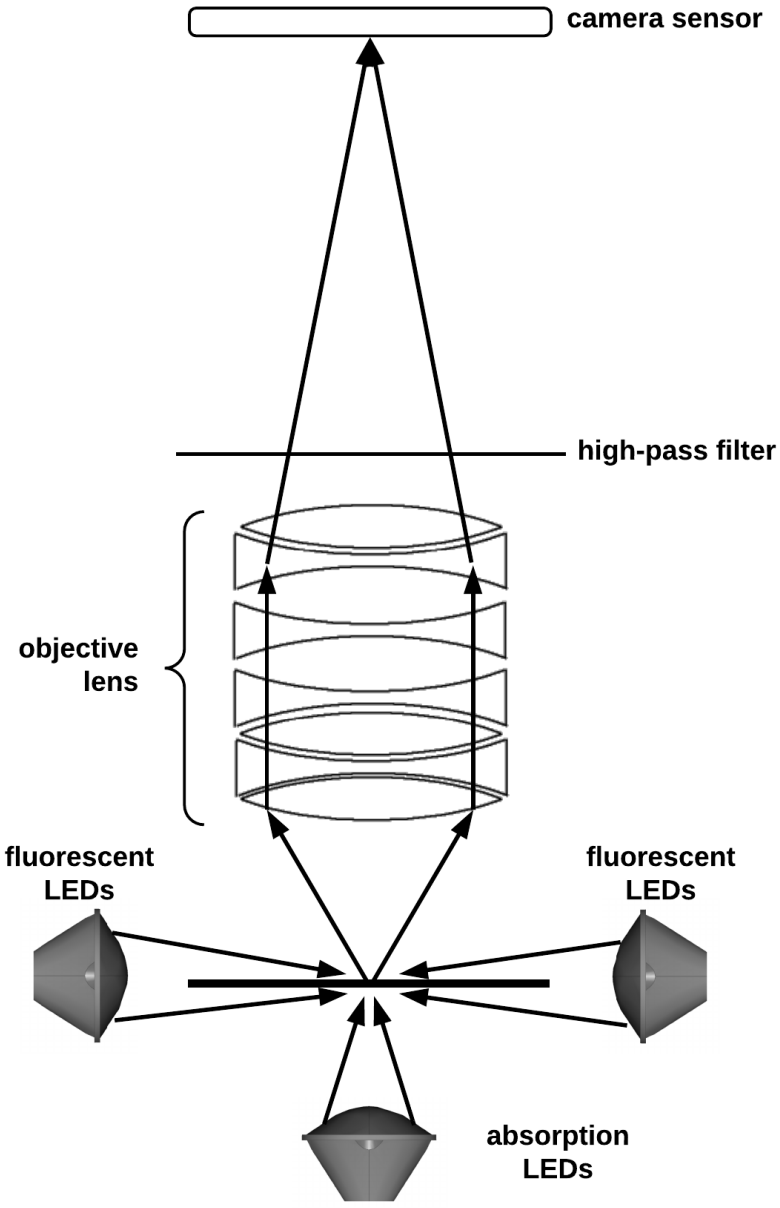


Figure 4.3: The optical system of the proposed imaging device allows the camera sensor to capture a single brightfield image as well as four fluorescence images at different excitation wavelengths. It is this configuration that will be used to generate a dataset which can be fed into a software framework for automatic identification of algae.

the relative concentration of different pigments in the algae, leveraging the fact that different algae groups have different pigments, and that those who have the same pigments likely have them in different concentrations. Since different algae contain different pigment concentrations, the relative fluorescence for different algae types will be different. It is this information that can be used while building and testing a machine learning model.

4.4 Imaging System Implementation

Having completed the optical design in Section 4.3 the final step is to build the imaging system. In section 4.4.1, the optimal microscope objective lens and camera sensor will be chosen. The combination of the lens with the sensor is shown to be able to resolve single-celled *microcystis aeruginosa*. In Section 4.4.2, the 3D printed hardware chassis will be presented. This chassis houses all the optics, electronics, and other components. Finally, Section 4.4.3, presents a graphical user interface that can control the camera and LEDs, allowing on-site data collection to be possible.

4.4.1 Spatial Resolution

The theoretical spatial resolution limit, d, of a microscope system is determined by the Abbe diffraction limit:

$$d = \frac{\lambda}{2n \sin\theta} = \frac{\lambda}{2\text{NA}} \tag{4.1}$$

where n is the refractive index, θ is the half-angle, NA is the numerical aperture of an objective lens, and λ is the wavelength of light passing through the objective lens [70]. Given that the average wavelength of light passing through our system in 700 nm, and that a standard 40x objective lens has a NA of 0.65, the theoretical value of d is 0.539 um.

Given that this distance is magnified by the lens and given that we want to meet the Nyquist criterion (N) in order to avoid any aliasing, results in the following equation:

$$dM = pN \tag{4.2}$$

Here d is spatial resolution limit, M is the magnification of the lens, p is the pixel size, and N is the Nyquist criterion (a minimum value of 2). Substituting the d value of 0.539 μm, the M value of 40x, and the minimum value of $N = 2$, results in a required pixel size 13.18 um. However, no low-cost sensor exists with such a large pixel size. Furthermore, a value of $N = 2$ is on the threshold of the Nyquist criterion and therefore is susceptible to

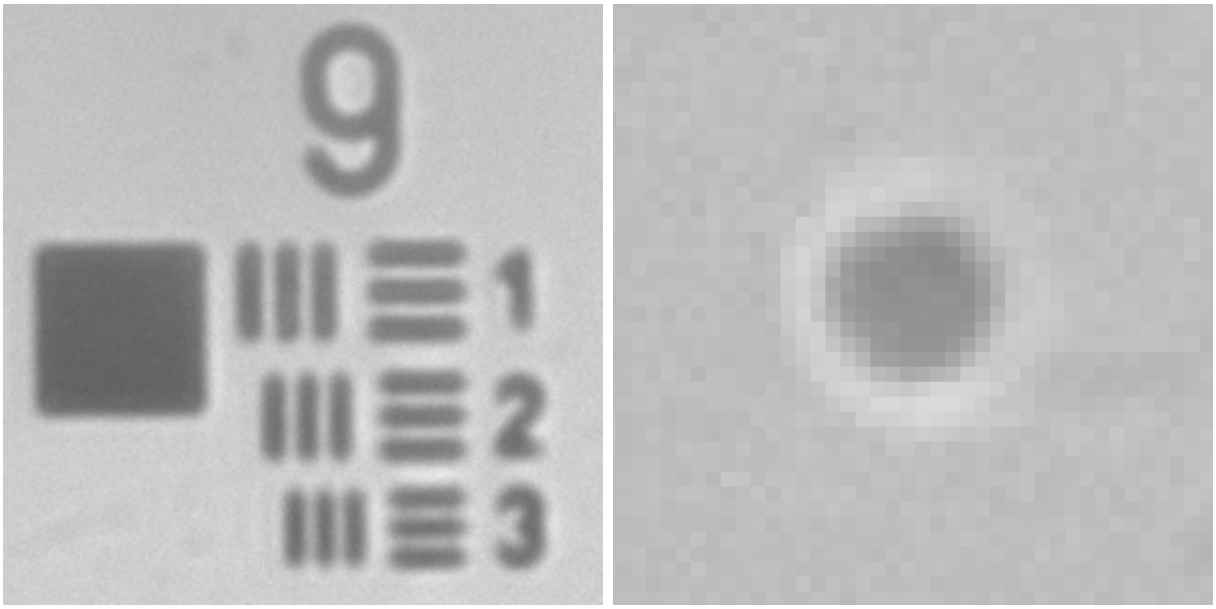


Figure 4.4: Left: Using the USAF 1951 chart a spatial resolution of 0.65 um / pixel was able to be achieved. Right: This is sufficient to measure single-celled *microcystis aeruginosa* which is known to be 3-7 um [65]. This *microcystis aeruginosa* from the dataset in Chapter 6 is 10 pixels in diameter, resulting in a diameter of 6.5 um.

aliasing. For these reasons, N was chosen to be 4, resulting in a desired pixel size of 5.4 um, which is a readily available pixel pitch. A FLIR Grasshopper 3 NIR monochromatic camera was selected as it had a pixel size of 5.5 um, and allowed full software control of the sensor parameters (white balance, exposure time, frame rate, etc.). This sensor was specifically suitable as it has a strong response in the NIR part of the electromagnetic spectrum, matching the emission response of the data presented in Figure 4.1.

To calculate the the actual spatial resolution of the imaging system the 1951 U.S. Air Force (USAF) chart was used, as seen in Figure 4.4 (left). Given this chart the spatial resolution was calculated to be 0.65 um / pixel. Given the sensor dimensions of 2048 $\times$ 2048 results in a field of view of 1.33 mm $\times$ 1.33 mm.

To further validate this spatial resolution of 0.65 um / pixel, a single-celled *microcystis aeruginosa* was imaged, as seen in Figure 4.4 (right). As reported by Xiao *et al.* *microcystis aeruginosa* is expected to be anywhere between 3-7 um in diameter [65]. The observed *microcystis aeruginosa* from the dataset in Chapter 6 is 10 pixels in diameter and therefore is 6.5 um in diameter. This check validates that the desired spatial resolution was

achieved as the initial design requirement was to be able to resolve a single-celled *microcystis* organism. Therefore the 40x microscope objective with a FLIR Grasshopper 3 NIR monochromatic camera achieves a high enough spatial resolution to resolve single-celled *microcystis* while simultaneously maximizing the field of view.

4.4.2 Hardware Chassis

By using an orthogonal fluorescence setup with high-powered LEDs and a single highpass filter the cost of the system can be dramatically reduced while simultaneously allowing both florescence and brightfield images to be captured in a rapid manner. The final consideration in the design of the imaging system is to decide how to package all the components into a single system that can be used by an individual. To achieve this design requirement, this optical setup was encapsulated into a 3D printed frame, as seen in Figure 4.5. By housing all the different optical and electrical components in a single chassis, this allows the imaging system to be portable. Furthermore, the 3D printed structure also blocks any ambient light from interacting with the sample, and thus preventing any additional errors into the system.

The system works as follows. First, light is emitted from either the fluorescent illumination sources or from the brightfield illumination sources onto the slide that contains a water sample. If the fluorescent illumination source is active, it will cause the algae to autofluorescence. If the absorption illumination source is active, the algae will either absorb or transmit the light. The emitted or transmitted light then passes through a series of lenses (acting together as an objective magnification lens) as well as a high-pass-filter, and finally onto an imaging sensor. A vertical translation stage which controls the fine-tune focusing of the image, is connected to the sensor and a focusing knob extends from the top of the system, allowing a user to focus the image. In order to control the different LEDs, a custom printed circuit board (PCB) was designed and manufactured.

The material cost of all the components and the 3D printed frame was approximately $4,000 USD. Therefore the design requirement of building a low-cost system for under $10,000 was met.

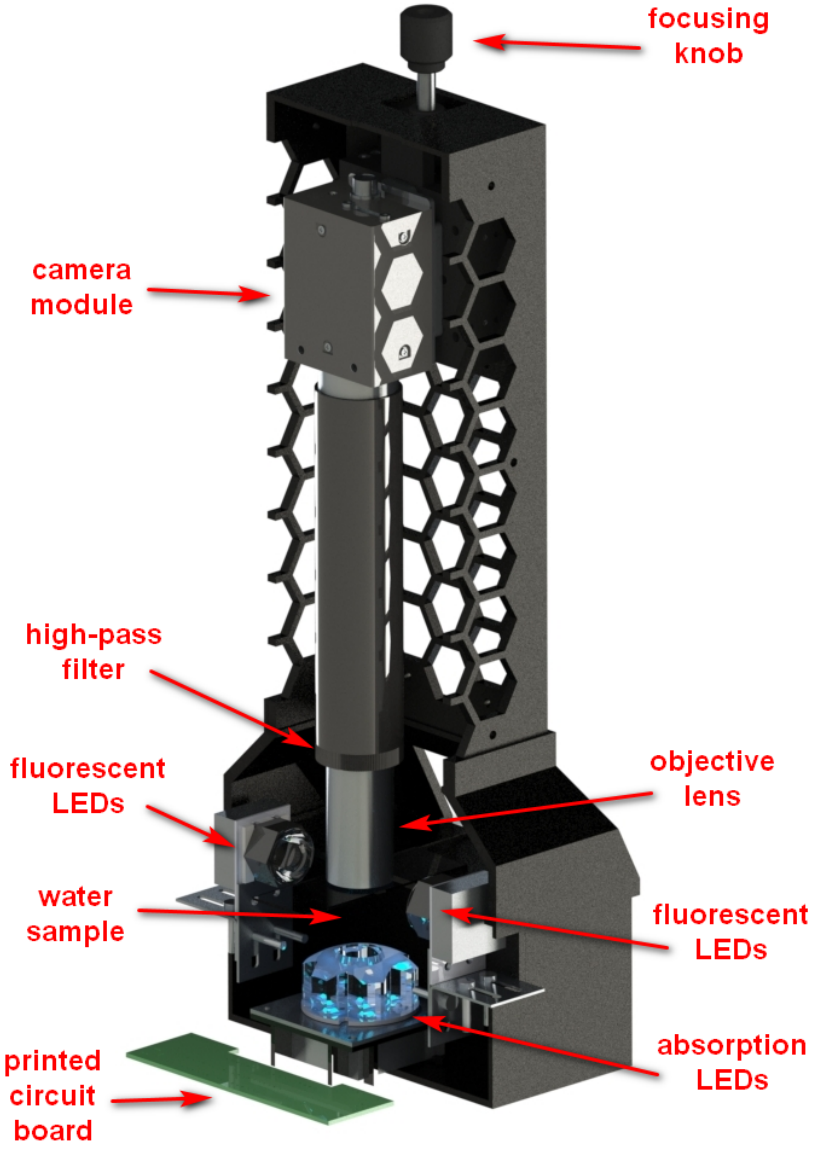


Figure 4.5: The proposed imaging system was built using off the shelf components and housed by a 3D printed frame. The user places the water sample in the imaging path and then adjust the focus with the focusing knob. All control over the illumination sources and sensor of SAMSON is done through the graphical user interface (Fig. 4.6).

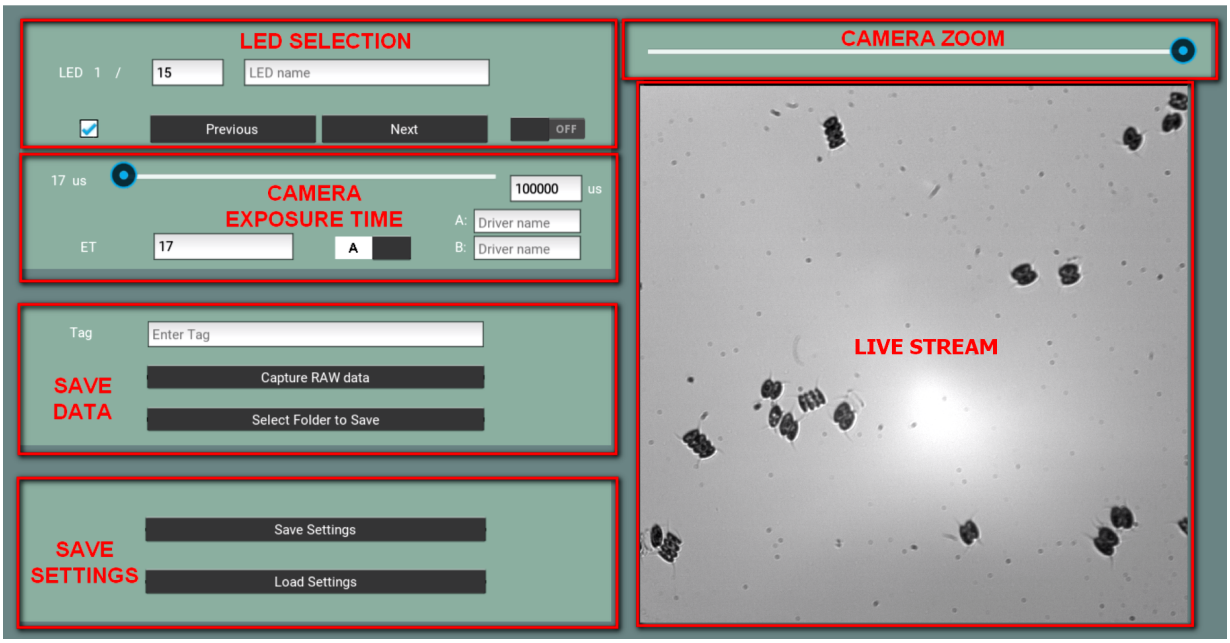


Figure 4.6: The graphical user interface (GUI) enables the flexible selection of different illumination sources and changes in exposure time of the sensor, during the process of viewing the water sample in real-time.

4.4.3 Software Control Tool

In order to control the LEDs and camera of this system, a graphical user interface (GUI) was built, as seen in Figure 4.6. In a standard scientific environment, multiple software programs are necessary for the purpose of image acquisition to control the various components (e.g., sensor, individual illumination sources, etc.). As such, this typically requires the user to switch between programs during the image acquisition process, which is time-consuming and prone to error. Therefore, this user interface was designed to provide all required functions in one concise and user-friendly interface, thus drastically improving the usability of the image acquisition process. The three main functions in the GUI are: i) live-stream visualization, ii) sensor exposure time control, and ii) illumination source selection and control. All operations can be performed in the GUI sub-system through sliders. Therefore, by using this software system on a local laptop, a user can use the imaging system on-site to generate data in real time, meeting the requirement to generate data on-site.

4.5 Summary of Imaging System

At the beginning of this chapter there were four design requirements for the imaging system. As review, they were:

1. Capture a single brightfield image with a spatial resolution of at least 0.75 um/pixel.
2. Capture at least three fluorescence images where each image targets a different pigment.
3. Be low-cost (less than $10,000) as to remove any financial limitations of rural communities and third world countries to generate data on-site.
4. Automatically analyze the generated data in real-time to allow for real-time inspection of water samples.

By carefully reviewing existing literature and by designing our system from first principles, a low-cost system with on-site data generation was created. The data created by this instrument is a single brightfield image and four fluorescence images. This imaging system has a spatial resolution of 0.65 um/pixel, which is sufficient to observe a single-celled *microcystis aeruginosa*. Furthermore, this entire imaging system costs approximately $4,000 USD to build, meeting the cost requirement to be under $10,000 USD. Having realized this system, the next step is to design and implement the complementary software framework in order to achieve the last design requirement of real-time data analysis. To fulfill this final design requirement a software framework will be developed, as presented in Chapter 5.

Chapter 5

Model Architecture Design

"All models are wrong, but some are useful."

– George E. P. Box (1919 - 2013)

In the previous chapter we discussed the motivation and rationale behind building a low-cost imaging system that can capture multiple fluorescence images and a single brightfield image. Having designed and built this system, the doors now opens to exploring how multispectral fluorescent data can be used when classifying algae. To explore this a number of different model architectures must be designed and compared against each other in order to determine the optimal configuration. These different architecture design considerations can be summarized into three fundamental questions. These questions are:

1. Which data do we use? (Section 5.1)
2. How is the data represented? (Section 5.2)
3. How does the data propagate through the model? (Section 5.3)

These questions will guide the different architecture-design decisions in the following sections.

5.1 Data Modality

Question 1 (which data do we use?) requires a choice of which imaging modality is fed into a given machine learning model. Given the data generated from the imaging system proposed in Chapter 4, the choices consist of three options:

1. Brightfield Based Classification (Section 5.1.1)
2. Multispectral Fluorescence Based Classification (Section 5.1.2)
3. Combined Brightfield & Multispectral Fluorescence Based Classification (Section 5.1.3)

To determine which data modality is best, machine learning models must be trained and tested where the only difference is the input data that is being used. By determining the relative performance difference between these different models, we can determine whether it is necessary to include multispectral fluorescence images as input to the model.

5.1.1 Brightfield Classification

Given that the majority of methods from Chapter 3 use a single brightfield image, that configuration must be tested as a baseline. For instance, both brightfield imaging methods (Section 3.2.1) and imaging flow cytometry methods (Section 3.3) capture a single brightfield image which is used for automatic identification. By building machine learning models based only on the brightfield data, we can determine a baseline performance. If the brightfield image contains enough information to adequately identify different algae, then the fluorescence data is not required, reducing the number of images needed to be acquired by the imaging system. This is important, as acquiring more images adds to the complexity of the imaging system and also takes additional time.

5.1.2 Multispectral Fluorescence Classification

An alternative option is not to use the brightfield data, but to input only the multispectral fluorescence data into a given machine learning model. This will allow us to test the classification performance when just capturing the fluorescence imaging. If this performance is much higher than the brightfield classification method, then one must decide whether the increase in performance is required. This will depend on how much better the classification accuracy is and the specific use-case of the the system. Furthermore, this is

an important setup to test as spectral methods in Section 3.4 have been able to achieve phyla-level classification of algae, by just using spectral measurements.

5.1.3 Combined Brightfield & Multispectral-Fluorescence Classification

The final option is to use both the brightfield data and the multispectral fluorescence data when training and testing a machine learning model. This would leverage all the data being acquired by the imaging system that was designed and built in Chapter 4. Once again, if there is a performance increase compared to the other two options, once must decide whether the performance increase is worth the additional data collection.

5.2 Data Representation

Question 2 (How is the data represented?) determines whether the data should be transformed into some new space, or if the data should be left in its original form. The first option is called feature extraction and the second option is called feature learning. This results in two different classification methods:

1. Feature Extraction Based Classification (Section 5.2.1)
2. Feature Learning Based Classification (Section 5.2.2)

Since we are dealing with images, the feature learning paradigm will leverage convolutional neural networks (ConvNets or CNNs). To stay in the domain of neural networks, the feature extraction method will leverage feedforward neural networks. By simultaneously exploring these two independent paradigms we can determine the relative performance changes when using different image modalities (Section 5.1), and when using different methods to propagate the data through the network (Section 5.3).

5.2.1 Feature Extraction Based Classification

Feature extraction requires preprocessing steps to transform the original data into a feature space. These features are usually hand-crafted and often based off domain expertise. Low-level feature extraction includes edge detection and corner detection as well as Scale

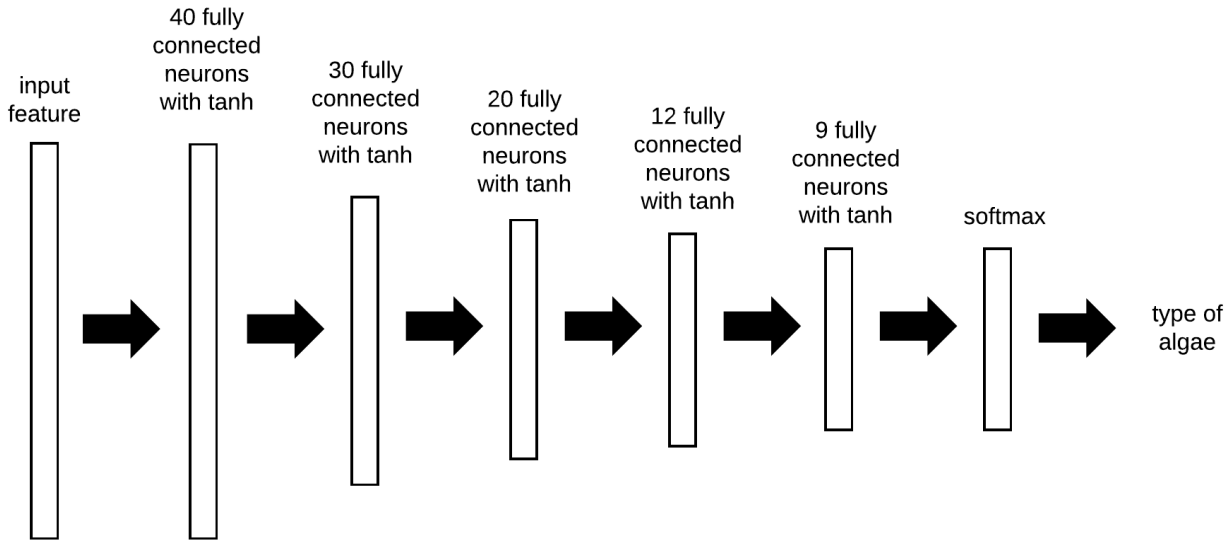


Figure 5.1: The feedforward neural network architecture used in feature extraction based models. The extracted features are input into the model. The data propagates through the network in order to classify a given type of algae.

Invariant Feature Transform (SIFT) methods. High-level feature extraction includes methods utilizing fixed shape matching (e.g. Fourier descriptors and the Hough transform), as well as deformable shape analysis (e.g. active contours). There are many other common object descriptors such as chain codes and moment analysis. Finally there are methods to describe the structure within an image which include structural and statistical approaches [71]. These features are then passed into a machine learning approach such as a decision tree, support vector machine or simple neural network in order to learn a mapping from the extracted feature to the output class [72].

Given that the majority of the methods in Section 3.2 use a feature extraction in their models, this configuration must be tested as a baseline. Therefore, to mimic these methods, a set of features will be extracted and used to train and test a machine learning model. As previously mentioned, convolutional neural networks are an appropriate choice for dealing with images, however, we have now extracted a set of features from that given image. A number of different machine learning methods can be leveraged here such as support vector machines, decision trees, and neural networks. Neural networks were chosen as they can theoretically approximate any non-linear function [73], and they are shown to outperform other machine learning paradigms when the dataset size increases [74].

The architecture of the feedforward model can be seen in Figure 5.1, and has an input feature, five fully connected layers with a tanh activation function, followed by a softmax. In a feedforward network the data propagates through the network from left to right. Each layer of the network consists of multiple neurons that take a weighted sum of the inputs, x_k and bias, b and transform them with a non-linear activation function, which in this case is a the tanh function. This process is repeated until the softmax layer normalizes the output data.

The architecture presented in Figure 5.1 was determined through empirical testing over many iterations. Initially the model consistently underfit the data and was unable to learn any mapping from input features to the output classes. Care was taken to not exceed a network with five layers to ensure that the network would not experience an unstable gradient during backpropagation. Different number of layers, the width of each layer, as well as the activation function was varied to find a model with an appropriate capacity to learn the task at hand.

5.2.2 Feature Learning Based Classification

The advancements of feature learning, where neural networks automatically extract the optimal features while simultaneously classifying, allows the raw image data to be passed directly into the neural network. This completely removes the need to extract any features, removing an often complicated step from the data pipeline.

The most common approach to feature learning for images uses convolutional neural networks (CNNs) [75], and therefore it is the most appropriate method for dealing with the images data generated from our imaging system. Over the last decade many CNN architectures have been publicly released, each architecture superior to its predecessors. Krizhevsky *et al.* released AlexNet (2012) [76], Szegedy *et al.* released GoogLeNet (2014) [77], and Simonyan *et al.* released VGG (2014) [78]. Finally, He *et al.* released ResNet (2015) [79] which has quickly become a standard convolutional neural network architecture in computer vision. Li *et al.* demonstrate that the loss function for a network with skip connections is easier to traverse when learning [80]. For these reasons a convolutional neural network with a residual framework was chosen as the architecture for the feature learning model.

ResNets are built off the idea of residual layers, which skip connections between different layers of the network. As explained by He *et al.*, the ResNet has a basic building block called the residual learning block. This residual block is defined as

$$\mathbf{y} = \mathcal{F}\left(\mathbf{x}, \{W_i\}\right) + \mathbf{x}. \text{ [79]} \tag{5.1}$$

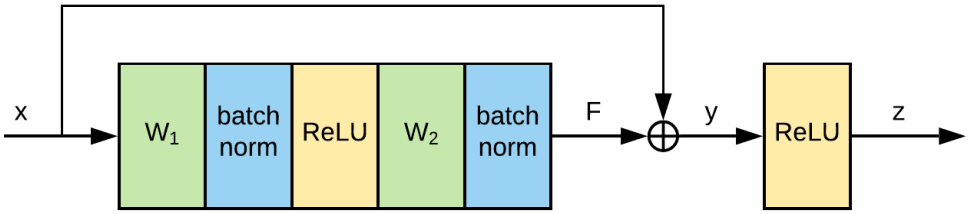


Figure 5.2: A two layer residual block as proposed by He *et al.* [79] where the convolutional layers are in green, the batch normalization layers are in blue, and the ReLU layers are in yellow. This basic building block is repeated in the proposed model as seen in Figure 5.3.

More specifically, this residual block learns the function $\mathcal{F}$ which consists of multiple convolutional layers and ReLu activation functions. A common implementation of this, as seen in Figure 5.2, consists of two convolutional layers, W_1 and W_2, which are weight matrices. The function $\mathcal{F}$ is then added element-wise with the input layer $\mathbf{x}$ to create $\mathbf{y}$. He *et al.* also explain that this term is then passed through another ReLU layer, which we will call $\mathbf{z}$. Therefore, for the two layer case:

$$\mathcal{F} = W_2\ \sigma(W_1\mathbf{x}). \tag{5.2}$$

The output of the residual block is then:

$$\mathbf{z} = \sigma(\mathbf{y}) = \sigma(\mathcal{F} + \mathbf{x}), \tag{5.3}$$

where $\sigma()$ is the ReLU function. He *et al.* explain the counter-intuitive idea that adding more layers to a standard network without the residual layers actually decreases the performance. The main advantage of skip connections is that it avoids the problem of a vanishing gradient which can occur during back-propagation, allowing the network to improve in performance as more layers are added on.

Due to these advantages, the proposed deep convolutional neural network architecture for algae detection will be based off the residual network architecture, as seen in Figure 5.3. More specifically, the base architecture that will be used is the ResNet18 architecture, which has which has 18 layers of the base residual block. The input layer in Figure 5.3 consists of a 2D convolutional layer followed by a batch normalization layer and then followed by a max pooling layer. This input layer takes one image for the brightfield data, four images for the fluorescence data, and five images for the combined brightfield and fluorescence data. The residual block layers (green) follow the structure presented in Figure 5.2. The downsampling layers (red) consist of a 2D convolutional layer followed by

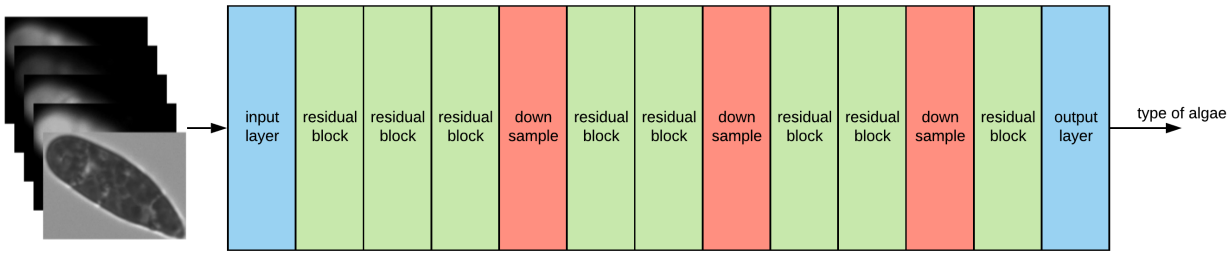


Figure 5.3: The proposed deep convolutional neural network architecture for algae classification from multispectral imaging data. The architecture is based on a ResNet-18 architecture [79] with the input layer designed to take in a stack of multispectral images, and the output layer designed to predict the type of algae in the imaging stack. The feature extraction component of the architecture was pretrained using the ImageNet dataset [81, 82].

a batch normalization. Finally, the output layer is an average pooling layer, followed by a fully connected layer, which feeds into a softmax function.

In addition to using convolutional neural networks for feature learning in an end-to-end manner, many large datasets exist which can be used to pretrain a given CNN architecture in order to leverage transfer learning. Transfer learning is the ability to leverage an existing dataset from another domain for a different use-case. Opposed to training with random weight initialization, the network comes with specific initialized weights as determined by training on another dataset. This is of significant importance in our situation because a deep learning model requires a large dataset of images in order to adequately learn a mapping from input to output. As we will see in Chapter 6, the dataset collected is a total of 6330 images, which is not sufficient to train a deep neural network.

For this reason we pretrained our proposed model using the ImageNet dataset [81, 82]. ImageNet consists of over 14 million images with more than 20,000 categories of common every day items (e.g. appliance, plant, tool, vehicle, etc.) [81, 82]. However, ImageNet and our dataset in Chapter 6 are likely to have different data distributions, and therefore the higher level features in the pretrained model will have little similarity to our dataset. Often in transfer learning it is desirable to pretrain a network on a large dataset that follows a similar data distribution of the smaller dataset. Given that the ImageNet dataset is so diverse, the lower-level and mid-level features learned by the deep neural network can be leveraged in our dataset. For this reason transfer learning using the ImageNet dataset will still be beneficial in this application since we can allow the model to fine-tune its weights to our application.

In summary, the proposed deep learning architecture for algae detection was built off the ResNet18 architecture and was pretrained on the ImageNet dataset. This will allow us to gather a relatively small dataset while still leveraging the power of feature learning using deep neural networks.

5.3 Data Propagation

Question 3 (How does the data propagate through the model?) requires a choice between two methods of how the data will travel through the model. This decision has two choices:

1. Flat Structure Based Classification (Section 5.3.1)
2. Hierarchical Structure Based Classification (Section 5.3.2)

5.3.1 Flat Structure Based Classification

The first method of propagating the data through the model is assuming a flat structure, where all output classes have equal rank. This is the default method when training a machine learning classifier as all output classes are independent of each other. Since this is the standard approach, this classification method will be compared to the hierarchical structure.

5.3.2 Hierarchical Structure Based Classification

The branch science called taxonomy deals with how we categorize and classify living organisms based on shared characteristics. Organisms with similar features, homologous structures, traits and behaviours are grouped into a single taxon (plural: taxa). The fact that there are course and fine similarities across living organisms brings rise to a hierarchy of groups, that is, a hierarchy of taxa, where different taxa are nested inside other taxa. Inherent to any hierarchy is a rank system. In biological taxonomy, a higher rank refers to more general and broad characteristics and a lower rank refers to a grouping with very high similarity in characteristics. From highest to lowest, the ranks are as follows: domain, kingdom, phylum, class, order, family, genus, and species.

Over the last 200 years many taxonomy systems have been proposed, each evolving from the previous. Currently the most widely accepted biological taxonomy structure is

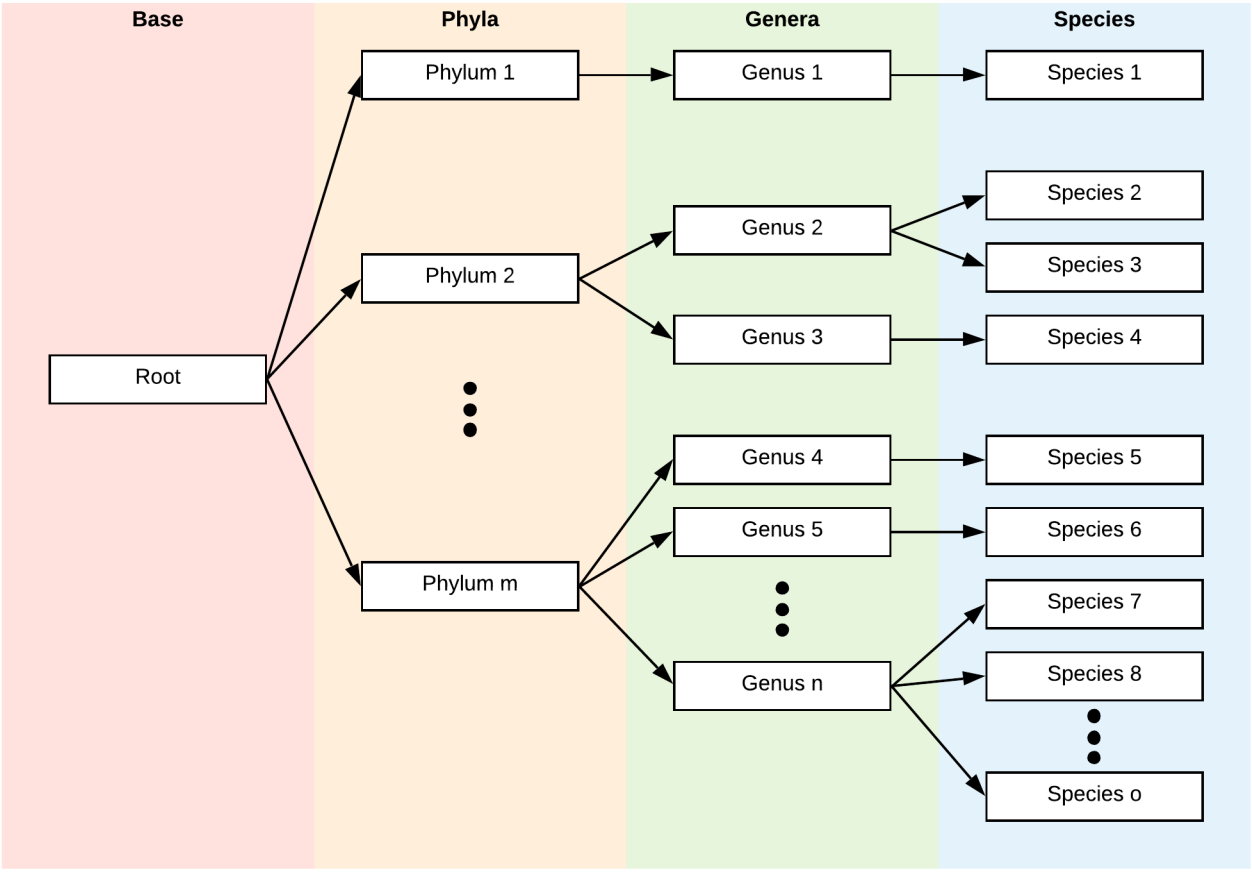


Figure 5.4: The proposed hierarchical structure is broken into four levels: the base level, the phyla, level, the genera level, and the species level. Building a machine learning model in a hierarchical manner allows online learning and explainability, both of which are important in a regulated industry such as drinking water treatment plants.

the three-domain system as proposed by Carl Woese *et al.* in 1977, which breaks all life into three domains: bacteria, archaea, and eukarya [83]. All organisms have a unique location in this hierarchical structure, which forms the entire phylogenetic tree, which is commonly knows as the tree of life.

Since the biological taxonomy is divided into a hierarchical system, we can use this prior knowledge when constructing our classifier. Opposed to building a single model and attempting to classify down to the species level, it is much more advantageous to build specific models for the different classification levels. For example, opposed to classifying

30 different species from 6 different genera of cyanobacteria, is more logical to first build a classifier to separate the different genera, and then build sub-classifiers for each genus in order to determine the different species with that genus. This will allow each sub-model to be much simpler when compared to the total complexity of a single model, ultimately resulting in significantly less data needed to train each sub-model.

Such an approach was taken by Walker *et al.* when they used a hierarchical classifier to achieve species level classification of four different species of *Anabaena* and two different species of *Microcystis* [8]. Inspired by this, we also propose to use a hierarchical classification scheme divided into four levels: the base level, the phyla level, the genus level, and the species level, as seen in Figure 5.4. The entire model will be referred to as the global model, while a given node model will be referred to a local model. Any node that has more than one child node must make a decision as to what the output class should be. Therefore, the nodes with more than one child node will have a unique machine learning model, that is a node / local model. In this manner the global model is composed of a hierarchy of local models.

The main advantage of this this approach is that it allows more explainability of what the global model is learning, which in turn makes online learning simpler. Explainability is an important feature of a model, as algae identification for drinking water treatment plants is a regulated space. Having a hierarchy of local models allows one to determine at which taxa level the global model is struggling. Not only does this allow one to understand the global model, this allows one to focus on improving a single local model resulting in the performance improving of a global model. For instance, in Figure 5.4, let's assume that the Genus 2 node is achieving low classification performance between Species 2 and Species 3. This allows focused attention on improving that specific local model by either additional data collection of just those two species, or experimenting with different machine learning paradigms. Opposed to having to retrain an entire end-to-end flat model, we can now train a specific branch within the hierarchy.

One potential disadvantage of this hierarchical approach becomes manifest when the taxa identification of an algae changes. As biologists learn more about certain organisms, their taxa identification can change. As we will see in Chapter 6, *Anabaena flos-aquae* changed its name twice during the course of this PhD research. First it changed species type from *Anabaena flos-aquae* to *Anabaena variabilis*, and then it changed genus type from *Anabaena variabilis* to *Trichormus variabilis*. In a flat structure approach, the model does not need to be retrained as just the class label will change. However, in a hierarchical approach multiple local models will need to be updated. Specifically, the local model that the algae left and the local model that it achieved will both need to be retrained.

model #	data modality	data representation	hierarchical structure
1	Brightfield	Feature Extraction	no
2			yes
3		Feature Learning	no
4			yes
5	Fluorescence	Feature Extraction	no
6			yes
7		Feature Learning	no
8			yes
9	Brightfield + Fluorescence	Feature Extraction	no
10			yes
11		Feature Learning	no
12			yes

Table 5.1: The 12 models that will be evaluated side-by-side result from three set of options when building a model. The first option is: which data do we use? The second options is: how do we represent that data? And the third option is: how does that data propagate through the network?

Even with this disadvantage, a hierarchical approach is likely to perform better in a real world scenario since the number of output classes is constantly increasing as new algae are being observed. In a use case such a drinking water treatment plant, new organisms will regularly be detected, and therefore the machine learning model will need to be updated to include these new organisms. In a flat structure, the entire model will need to be retrained to accommodate the single new class, while in a hierarchical approach only a few local model needs to updated, which will not affect all the other local models.

In summary, a hierarchical classification scheme allows for online learning and better explainability of the classification as opposed to a flat structure. Ultimately, both these structures need to be compared side by side using the same data and learning parameters to see their relative performance. These results will be presented and discussed in Chapter 7.

5.4 Summary of Model Architectures

We have seen that we have three independent sets of options with multiple outputs, therefore we must test each combination to determine the optimal classifier. To recap, these three questions where:

1. Which data do we use? (Section 5.1)
2. How is the data represented? (Section 5.2)
3. How does the data propagate through the model? (Section 5.3)

These three questions result in a total of 12 unique models that must be tested, as seen in Table 5.1. The first decision (which data do we use?) has three options which consist of using the brightfield data, the multispectral fluorescence data, and the combination of brightfield and multispectral fluorescence data. The second decision (how is the data represented?) has two options which consist of feature extraction based methods and feature learning based methods. For the feature extraction based method the model of choice is a standard feedforward neural network. For the feature learning based method the model choice is a modified ResNet18. The final decision (how does the data propagate through the model?) has two options which consist of using a flat structure or a hierarchical structure. To test the 12 models presented in Table 5.1, a dataset must be created using the proposed imaging system from Chapter 4. The dataset collection and preparation will be discussed in Chapter 6.

Chapter 6

Dataset Collection & Preparation

"Though perfectly transparent and colourless when held between the eye and the light, or a white object, it yet exhibits in certain aspects, and under certain incidences of the light, an extremely vivid and beautiful celestial blue colour..."[84]

– the first reported observation of fluorescence by Sir F.W. Herschel (1738 - 1822)

It is important to develop an intuition of what data are being created by the proposed hardware system as proposed in Chapter 4. In this chapter we will discuss the steps which are needed to prepare the data for further analysis by the software framework presented in Chapter 5.

First, in Section 6.1, we will discuss the nine types of algae chosen for our experiments. Then in Section 6.2 the image acquisition of the data collection will be discussed. Given that the raw data needs to be preprocessed, Section 6.3 will discuss how we clean and crop regions of interest. This process consists of flat field correction (Section 6.3.1), thresholding (Section 6.3.2), and cropping (Section 6.3.3).

Given that a region of interest has been located, Section 6.4 discusses the feature extraction process. The first set of features are brightfield features (Section 6.4.1) which consist of Fourier descriptors, Hu's invariant moments, geometric shape features, and texture features. The second feature set consists of multispectral fluorescence features (Section 6.4.2). Finally, in Section 6.5, an overview of the entire dataset will be presented.

6.1 Algae Selection

In order to build a machine learning model one must first have a dataset to train and test the model. However, to the best knowledge of the author, currently there is no multispectral fluorescence dataset of algae that can be used to evaluate a given machine learning model. Given that no dataset exists, we were motivated to build our own instrument (Chapter 4) to image water samples containing algae.

Taking images of algae in water samples solves the problem of creating a dataset, but it poses the problem of having unlabelled images. Solving this problem requires a human to manually label the images, or it requires an unsupervised learning approach. Having a human manually label all the images requires hiring a highly trained taxonomist, someone with decades of experience. But even if such a person could be located, it would still take a significant amount of time to do the labelling, which leaves this option unsuitable for this research. It is possible to use unsupervised learning methods to separate the data into the respective number of classes, however, this increases the likelihood of having incorrectly labelled images. For this reason also unsupervised learning is not desirable when building a dataset of labelled images.

In order to solve the problem of creating a labelled dataset we purchased pure algae cultures and imaged each of these under our custom microscope. By imaging pure algae cultures, it can be assumed that every organisms in the image belongs to the same class. This results in a multispectral image that has one or more regions of the same organism. This multispectral image can be further processed to locate regions of interest (ROIs) that can be assigned a known class label, as discussed in Section 6.3.

Therefore, nine pure algae cultures were purchased from the Canadian Phycological Culture Centre (CPCC). The taxonomic breakdown of these algae can be seen in Figure 6.1. These algae cultures, as seen in Figure 6.2, are as follows:

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*; AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*; AKA *Scenedesmus obliquus* (CPCC 005)

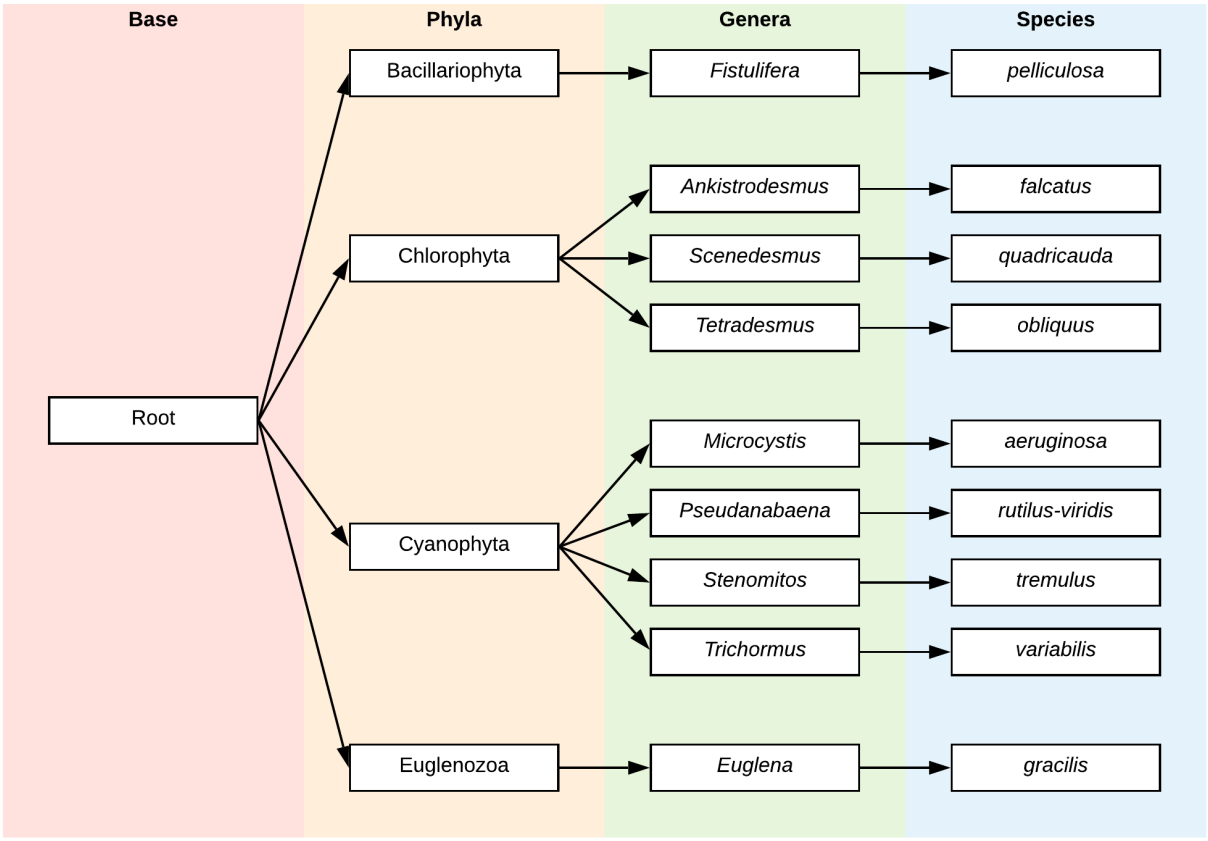


Figure 6.1: The taxonomic breakdown of the nine types of algae used to test the proposed hardware system and software framework. These nine algae come from four different phyla groups and were purchased from the Canadian Phycological Culture Centre (CPCC). The physical appearance of these nine algae can be seen in Figure 6.2.

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*; formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*; formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)

As discussed in Section 2.1, there are algae species that are known to produce toxins. Two of the highest toxin producers that come from the Cyanophyta phylum are *Microcystis sp.* and *Anabaena sp.*. For this reason CPCC 300 (*Microcystis aeruginosa*) and CPCC 067 (*Trichormus variabilis*; AKA *Anabaena variabilis*; formerly *Anabaena flos-aquae*) were chosen. CPCC 300 was specifically chosen as it is a small single-celled organism, allowing us to evaluate the spatial resolution of the system. It is also known to be one the major producers of microcystin, the toxin regulated by Health Canada, the USEPA and the World Health Organization. Moreover, it is important to note that CPCC 300 was observed as single-celled organisms, which is not its naturally occurring form as these single cells commonly aggregate into large colonies.

As observed, the specific name of a given algae type can change over a period of time based on what experts believe it to be. During the course of this PhD research CPCC 067 changed its name twice. When it was initially selected it was labelled as *Anabaena flos-aquae*, but then its species type changed to *Anabaena variabilis*. At the time of this writing, its genus and species name had recently changed to *Trichormus variabilis*. The other selected Cyanophyta algae were CPCC 697 (*Pseudanabaena rutilus-viridis*) and CPCC 471 (*Stenomitos tremulus*; formerly *Pseudanabaena tremula*). These two were chosen as they are similar in appearance to CPCC 067 and for that reason are commonly mistaken for CPCC 067. Also in this case, the biological name of CPCC 471 changed at both the genus and species levels.

One goal of this research is to explore whether a hierarchical approach would improve classification performance and for that reason algae types from three other phyla groups were chosen. From the Bacillariophyta phylum CPCC 552 (*Fistulifera pelliculosa*; AKA *Navicula pelliculosa*) was chosen because it is similar in appearance to CPCC 300 (*Microcystis aeruginosa*), as it is a small single-celled organism. From the Euglenozoa phylum CPCC 095 (*Euglena gracilis*) was chosen because they are a common single-celled organism and are known to bloom in both freshwater and saltwater [85].

Finally, three algae type were selected from the Chlorophyta phylum which were CPCC 366 (*Ankistrodesmus falcatus*), CPCC 158 (*Scenedesmus quadricauda*), and CPCC 005 (*Tetradesmus obliquus*; AKA *Scenedesmus obliquus*). CPCC 366 (*Ankistrodesmus falcatus*) was chosen as it was a readily available Chlorophyta algae that was easy to maintain. CPCC 158 and CPCC 005 were chosen as they were both part of the *Scenedesmus* genera, however, CPCC 158 changed names from *Scenedesmus obliquus* to *Tetradesmus obliquus*. Even with the name change the underlying research question can be explored as the basic branching structure still exists within the hierarchy at different levels.

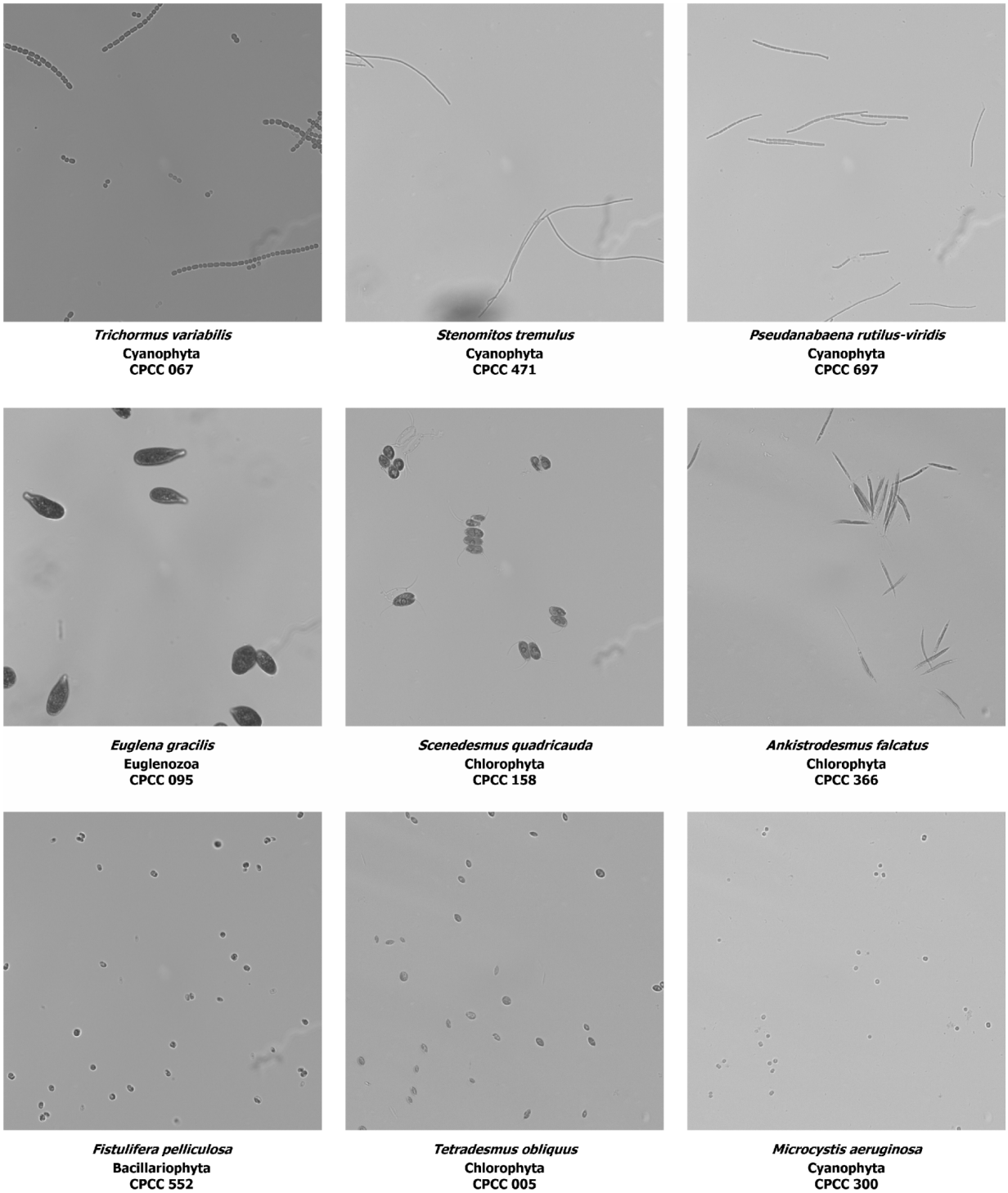


Figure 6.2: The nine types of algae used to test the proposed hardware system and software framework. Note that some algae types are very similar in appearance, as observed in the filamentous algae (top row) as well as the single celled algae (bottom row).

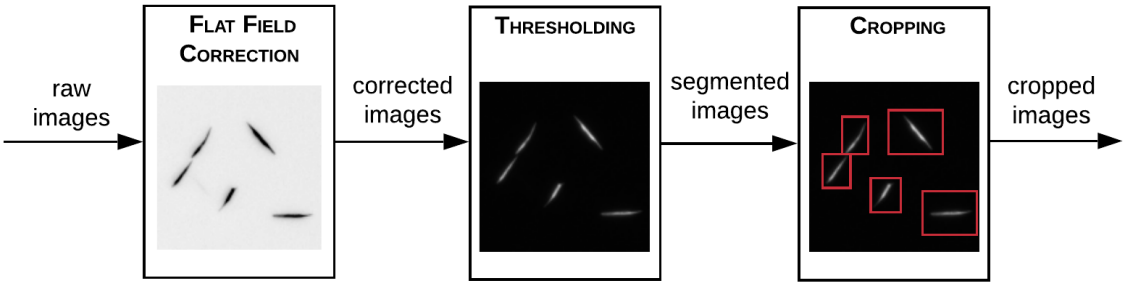


Figure 6.3: The raw images go through a three step process in order to create cropped images ready to be used by an image classifier. These steps are flat field correction, thresholding and cropping.

6.2 Image Acquisition

Given the nine pure algae cultures in Section 6.1, 7 uL of a given algae type was pipetted onto a blank 3" × 1" slide, and was then covered with a standard cover slip. This prepared slide was then placed into the imaging system from Chapter 4. Multiple images using the graphical user interface from Section 4.4 were taken from a single slide under a 40x finite conjugate objective lens. This process was repeated numerous time to build up a set of multispectral raw images for all nine algae types. Each multispectral image consisted of a single brightfield image and four fluorescence images.

Given these raw images, a number of steps need to occur before the data can be used by the software framework presented in Chapter 5. They first need to be cleaned and cropped (Section 6.3), after which additional features need to be extracted (Section 6.4).

6.3 Region of Interest Detection

Given a captured multispectral raw image, additional cleaning and preprocessing must occur before it can be used by a image classification model. This cleaning and preprocessing can be broken down into three main steps, as seen in Figure 6.3. Firstly, in Section 6.3.1 flat field correction will remove any illumination inhomogeneities. Next, in Section 6.3.2 a binary image will be created separating the foreground object from the background objects. Finally, in Section 6.3.3 the different foreground objects will be cropped.

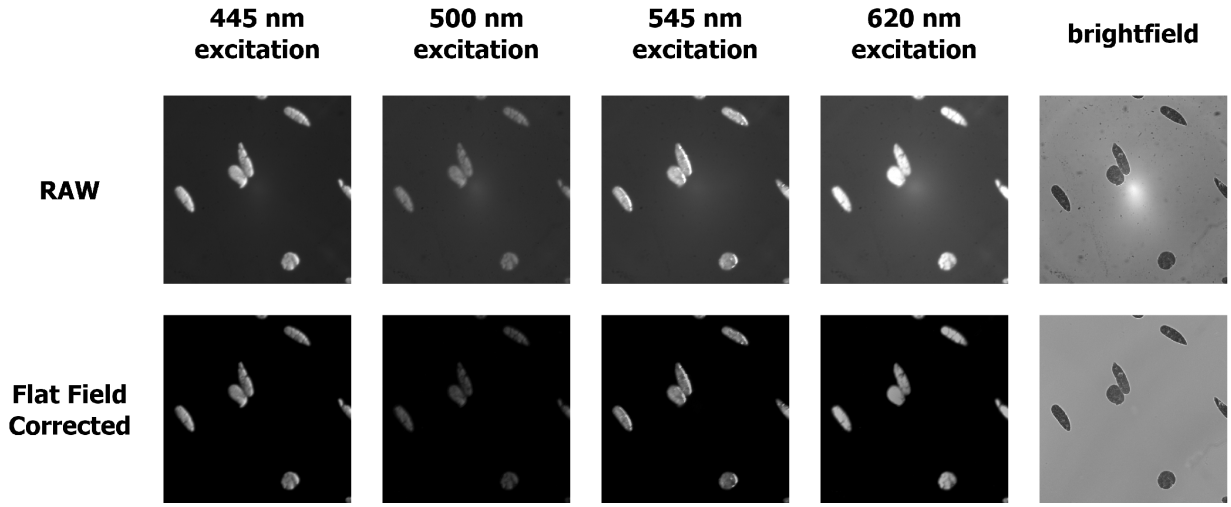


Figure 6.4: Flat field correction takes the raw camera image and corrects for noise, illuminations and optical distortions [33].

6.3.1 Flat Field Correction

A captured multispectral raw image contains two significant types of noise that render it unsuitable for scientific analysis. First, the image contains a bias signal and the resulting noise increases the pixel values compared to the true photometric values. And second, a raw image will contain a number of illumination and optical distortions inherent in the imaging system. In order to restore the photometric accuracy and remove imaging defects, a process known as flat field correction can be used [33]. Flat field correction can be mathematically described as

$$I_C = \frac{I_R - I_D}{I_F - I_D} \tag{6.1}$$

where I_R is the raw image, I_D is an image captured with no light source, that is a dark image, I_F is a image with no sample and only the light source and I_C is the corrected image. In Figure 6.4, the raw images I_R can be seen on the top and the corrected image I_C can be seen on the bottom. We can observe the non-uniformity of the light as there is a noticeable bright spot in the raw images. After flat-field correction, the corrected image has a complete uniform background. The other major benefit of flat-field correction is that it removes any other background artifacts, such as dust or impurities on the optical elements or camera sensor.

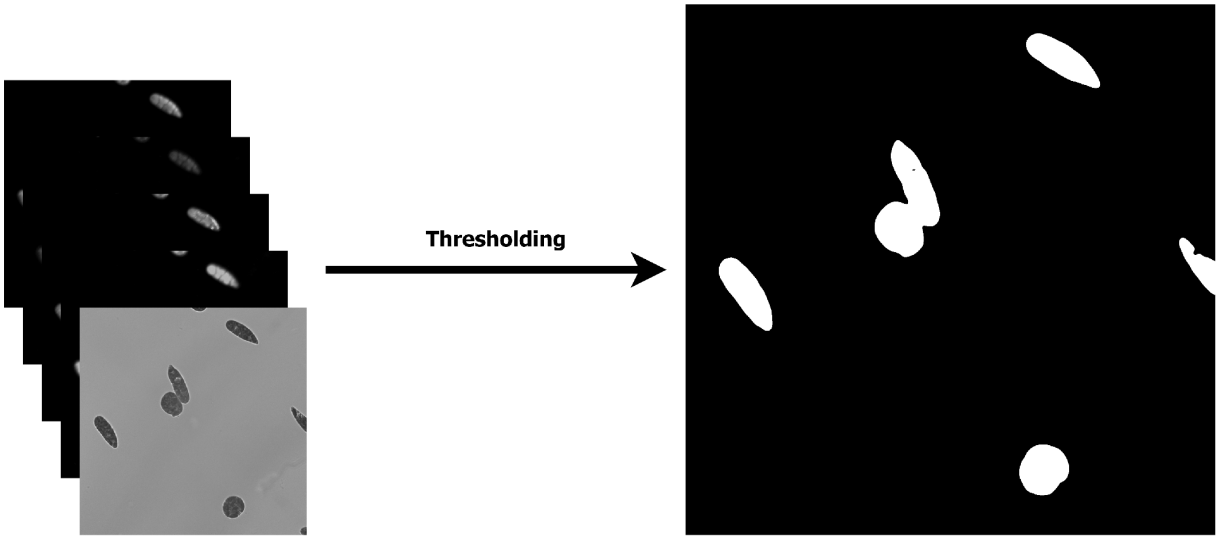


Figure 6.5: The highest contrast image of the flat field corrected images was manually chosen to be used in the thresholding task. Thresholding results in a binary mask which distinguishes the foreground and background from each other.

6.3.2 Thresholding

Given a corrected multispectral image, the next challenge is to separate the background from the foreground, as the algae samples are considered to be foreground objects as seen in Figure 6.5. Since only one segmentation mask is needed for all five images, the highest contrast image of the five spectral images was manually chosen for the thresholding task. In the case where the brightfield image was chosen, the inverse of the image was taken before the thresholding was applied.

To achieve this task a binary classifier was defined to classify each pixel into either the foreground class, C_f or the background class, C_b. The decision boundary of this classifier, θ, was learned by implementing Otsu's method [86], where the inter-class variability of the image is maximized, which simultaneously minimizes the intra-class variability. For any given pixel $\underline{x}$ the class, $C(\underline{x})$, was determined by:

$$C(\underline{x}) = \begin{cases} C_f & \text{if } f(\underline{x}) > \theta \\ C_b & \text{otherwise} \end{cases} \tag{6.2}$$

where $f(\underline{x})$ is the pixel intensity at pixel $\underline{x}$.

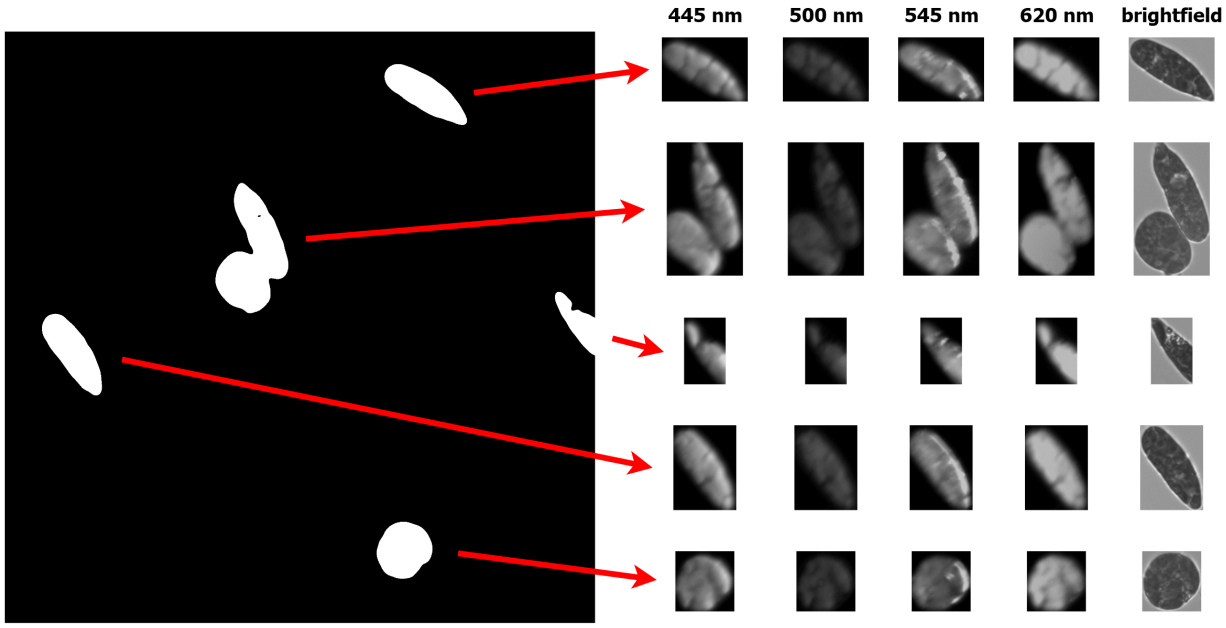


Figure 6.6: Given the binary mask which separates the foreground and the background, each foreground object can be cropped, resulting in a multispectral cropped image.

6.3.3 Cropping

Once all the organisms in a given multispectral image are segmented, each foreground group of pixels in the image were extracted and cropped, as seen in Figure 6.6. The brightfield cropped region of interest for each of the nine species can be seen in Figure 6.7. It is important to note that each cropped image will need to be resized to a fixed dimension as required for input to the deep convolutional neural network. For example, *Microcystis aeruginosa* will appear larger than in the original image and the *Anabaena flos-aquae* will appear smaller. This resizing results in the images losing their relative scale information, potentially discarding useful information when classifying these different organisms. Therefore this a potential limitation of the existing method, but is required for the current neural network architecture.

These cropped binary masks and multispectral images are now ready for the feature extraction process which will be discussed in Section 6.4.

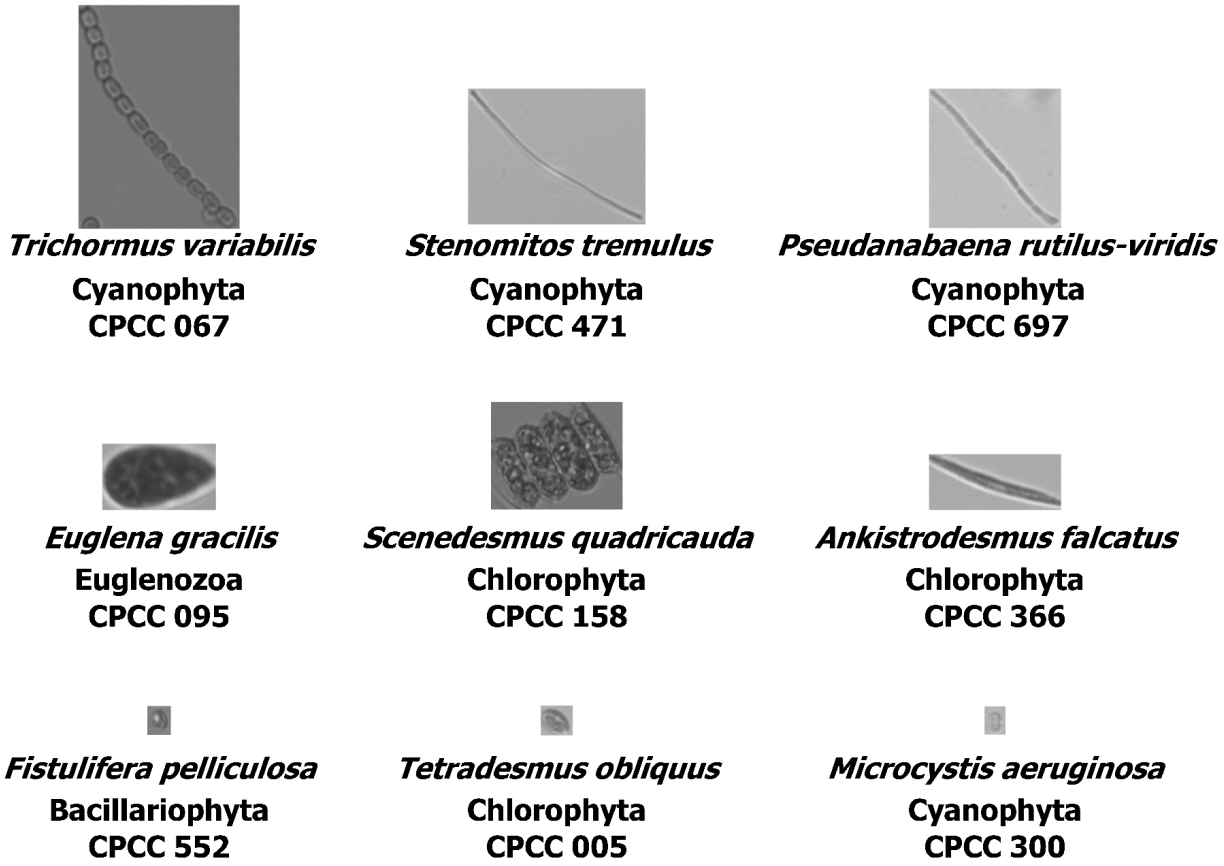


Figure 6.7: An example brightfield image crop for each of the nine algae types. Note how certain algae are significantly smaller in scale compared to other algae.

6.4 Feature Extraction

The feature extraction can be broken into two main parts, as seen in Figure 6.8. First, in Section 6.4.1, the brightfield features will be presented. These brightfield features are created using Fourier descriptors, Hu's invariant moments, geometric shape features, as well as texture descriptors. Next, in Section 6.4.2, the fluorescence features will be presented. These features take the mean of the pixels within a foreground object. These features are extracted for use in the feature extraction based models as discussed in Section 5.2.1.

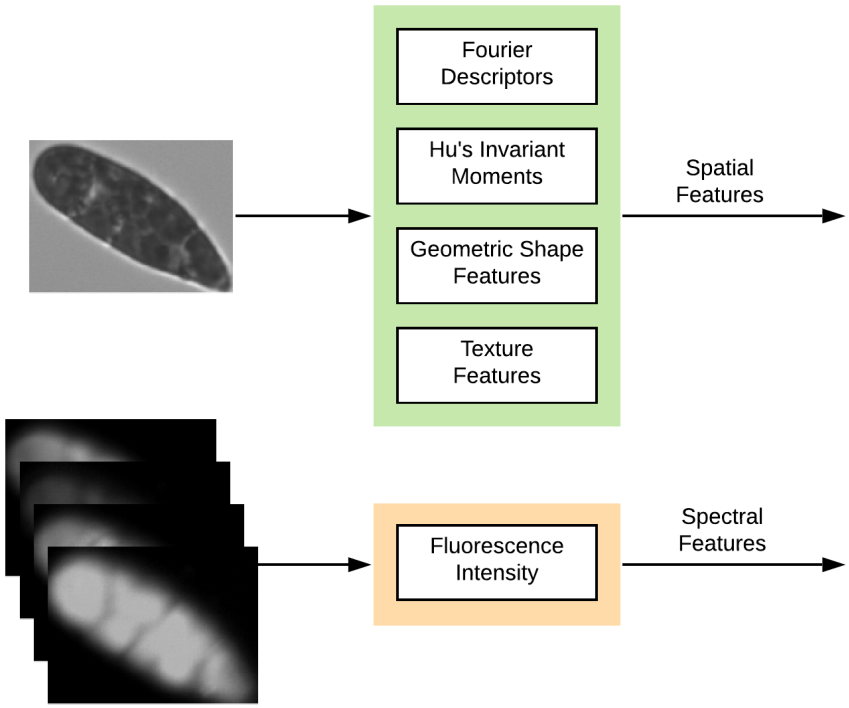


Figure 6.8: The brightfield cropped images were used to generate a set of spatial features using Fourier descriptors, Hu's invariant moments, geometric shape features, and texture features. The four band fluorescence spectral image was used to generate spectral features.

6.4.1 Brightfield Spatial Features

The set of spatial features extracted from the segmented organisms were generated using Fourier descriptors, Hu's invariant moments, geometric shape features, as well as texture features. These features will each be discussed in more detail in the follow sections. These spatial features were chosen to reproduce previous work [7, 34, 87, 35], which was discussed in detail in Chapter 3.

Fourier Descriptors

The Fourier transform is a method to decompose a signal into its different frequency components. Fourier descriptors use the Fourier transform to generate a feature vector that describes the shape and relative size of a given 2D object in an image [88].

Fourier descriptors are an appropriate way to describe algae as they are translation invariant, rotation invariant as well as scale variant [89]. First, Fourier descriptors are

translation invariant, which means they are independent of where the 2D objects are in an image. Secondly, only the phase of the frequency signal is altered when a image is rotated or the starting point of the boundary is changed. Since the feature vector is the magnitude of the complex frequency signal, this makes the feature vector also rotation invariant. Finally, Fourier descriptors preserve scale in the transformation, that is, for two identically shaped objects where object A is twice the size of object B, the Fourier descriptors for object A will be twice as large as object B. This is an important property of this transform as different algae types can be distinct sizes, and therefore scale can be used as a discriminating feature.

By only keeping N Fourier descriptors, a large amount of extraneous information is removed. For example, given M points of the boundary of an object, this can be reduced to N points in the Fourier space, where N << M. When these N points are transformed back into the image space by taking the inverse Fourier transform, the same basic shape is still preserved. In this thesis, 30 Fourier points were preserved (N=30) for different different organisms, each with significantly different M values.

Hu's Invariant Moments

Hu's invariant moments [90] are a set of seven statistics calculated from the grey scale pixel intensities in a segmented area from an image. These seven values are a set of commonly extracted features as they are translation, scale, and rotation invariant and are derived from a weighted combination of different central moments of a region. A central moment μ_{pq} is defined as

$$\mu_{pq} = \sum_{x} \sum_{y} (x - \overline{x})^p (y - \overline{y})^q I(x, y) \tag{6.3}$$

where $(\overline{x}, \overline{y})$ is the location of the centroid for a region, and $I(x, y)$ is the pixel intensity at pixel coordinate (x, y) within a region. By combining the 30 Fourier descriptors with these seven moment invariant features brings the current total of features to 37.

Geometric Shape Features

Geometric shape features are commonly used as they are intuitive to understand and easy to calculate [91]. The following six geometric shape parameters where extracted from each region with the goal of having features to describe the spatial characteristics of a given region:

1. **Area**: The number of pixels within a given region.
2. **Convex Area**: The number of pixel withing the convex hull of a given region.

3. **Circularity**: Measures how round a region is and is computed as $\frac{4\pi A}{P^2}$, where A is the area and P is the perimeter of the region.

4. **Major Axis Length**: The length of the major axis of an ellipse that is fitted to the region.

5. **Minor Axis Length**: The length of the minor axis of an ellipse that is fitted to the region.

6. **Eccentricity**: Measures the ratio of the major axis length and minor axis length.

Combining the 30 Fourier features, the seven moment invariant features and these six geometric shape features brings the current total number of features to 43.

Texture Features

In 1973, Haralick *et al.* introduced the Gray-Level Co-Occurrence Matrix (GLCM) to describe the texture of a given image which was then used for in a classification task [92, 93]. The co-occurance matrix tends to be very sparse, and therefore common metrics are measured from the GLCM. Commonly four GLCM features are measured [94]:

1. **Contrast**: measures the local variation within the GLCM.

2. **Correlation**: measures the joint probability of specific pixel pairs.

3. **Energy**: computes the sum of squared elements in the GLCM.

4. **Homogeneity**: measures the similarity of the distribution of the values in the GLCM to that of the diagonal values in the GLCM.

Combining these 4 texture features with the previous 43 features results in a total of 47 features extracted from the brightfield image. All of these 47 features describe the spatial element of a given algae in a water sample.

6.4.2 Fluorescence Multispectral Features

Whereas the spatial features are calculated from the brightfield image, the spectral features are calculated from the multispectral fluorescence images. To extract fluorescent spectral features, the mean emission intensity at each of the four wavelengths was measured for a given foreground region. More specifically, the four fluorescent spectral features are as follows:

1. **445 nm fluorescence**: Average fluorescence intensity values within a given region when excited by the 445 nm LED.

2. **500 nm fluorescence**: Average fluorescence intensity values within a given region when excited by the 500 nm LED.

3. **545 nm fluorescence**: Average fluorescence intensity values within a given region when excited by the 545 nm LED.

4. **620 nm fluorescence**: Average fluorescence intensity values within a given region when excited by the 620 nm LED.

For a given wavelength, the same LED, optics, camera sensor, and exposure time were used, allowing a comparison to be done across different algae types and within the data collected under these experimental settings. However, any spectral curves generated from this data cannot be seen as true excitation fluorescence curves since they have not been corrected. These corrections are required when the true excitation curve is needed as the signal is altered due to the amplifications and attenuation at different wavelengths, which are caused by the radiometric power differences of each LEDs (even though the LEDs were all run at the same current, their power outputs can vary), the spectral response of the optics, as well as the quantum efficiency of the sensor and the exposure time of the camera. While the exposure time of the camera at each wavelengths was known, the previously remaining parameters are unknown, and while they could be measured, they were not measured in this experiment since the required equipment was not available. However, since all the data was captured in the same manner, a comparison of different algae types can be accomplished. This is because they only variable changing within the system is the type of algae being imaged.

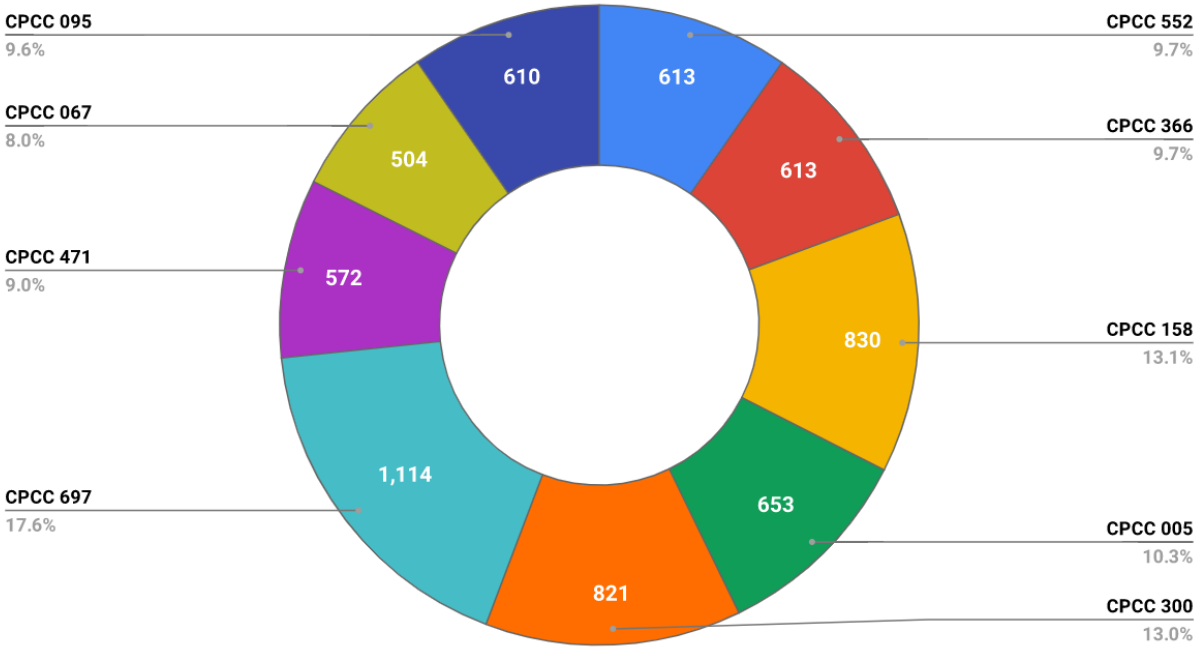


Figure 6.9: A total of 6330 segmented and cropped multispectral images were generated from the raw image collected from the imaging system. The class distribution of the nine types of algae can be seen above.

6.5 Overview of Dataset

The distribution of the number of samples for each algae class, as chosen in Section 6.1, can be seen in Figure 6.9. The total number of multispectral images was 6330, that is, each of these 6330 images are composed of five sub-images, four of which are fluorescence based, and one which is an absorption image. These images were created by first capturing raw multispectral data (Section 6.2) from which the Region of Interest detection was extracted (Section 6.3). Given these cropped images, both brightfield and fluorescence features were extracted for each multispectral image (Section 6.4). For example, for CPCC 158, there are 830 multispectral images (four band fluorescence with one band brightfield). For each of those 830 multispectral images the brightfield spatial features were extracted and the fluorescence multispectral features were extracted. The total number of extracted features for every image was 51, where 47 come from the brightfield image and 4 come from the fluorescence images. This set of images and features makes up the available data to now train and test different classification schemes, which will be discussed in Chapter 7.

Chapter 7

Experimental Results & Discussion

"The way to get good ideas is to get lots of ideas and throw the bad ones away."

– Linus Pauling (1901 - 1994)

In Chapter 4, a novel low-cost imaging system was proposed that could capture multispectral fluorescence images and a single brightfield image. In Chapter 5, it was proposed to use a hierarchical classification approach based on the prior information of the taxonomic structure. Finally, in Chapter 6, a dataset was created using the proposed imaging system in order to explore whether this hierarchical approach would be beneficial.

This chapter will demonstrate that using multispectral fluorescence data has a higher classification accuracy compared to the standard brightfield imaging modality. Furthermore, this chapter demonstrates that there is no change in performance between a flat structure compared to that of a hierarchical structure. Therefore a hierarchical structure is preferred as it is more suitable for online learning and for explainability.

The evidence for these conclusions will be presented in two sections. First, in Section 7.1, a qualitative analysis will be conducted on the data in order to build intuition and gain understanding of the spatial and spectral components of the data. Then, in Section 7.2, a quantitative analysis will be done on the 12 different model architectures presented in Chapter 5.

7.1 Qualitative Analysis

Before a quantitative analysis of the results can be done, we will present a qualitative analysis. This will provide context to the difficulty of the classification task and allow us a better look at the data. This qualitative analysis will be broken into image analysis (Section 7.1.1) and spectral analysis (Section 7.1.2). From the spatial analysis we will see that different groups of algae have similar morphological characteristics and fluorescence spectra while other groups have differences. From the spectral analysis we will quantify the relative intensity of the fluorescence images allowing a deeper understanding into the spectral nature of our data.

7.1.1 Image Analysis

To better understand the task of classifying nine types of algae, a sample algae image from each class is shown in Figure 7.1. In this figure, we can see both the four band fluorescence images as well as the brightfield image for each of the nine types of algae. To better understand the spatial features these images have been resized to approximately the same dimensions, and therefore these images are not to scale. To compare the relative scale of each of the nine algae classes, please refer to Figure 6.7.

The first observation is that the three filamentous types of algae (CPCC 067, CPCC 471, CPCC 697) are all similar in appearance. Furthermore, these three algae types, in addition to CPCC 300, belong to the Cyanophyta phylum and therefore have similar fluorescence responses. Since Cyanophyta are known to contain phycobilin pigments (e.g. C-Phycoerythrin (C-PE), C-Phycocyanin (C-PC), and allophycocyanin (APC)), as discussed in Section 3.1, it is understandable that the highest fluorescence signal is at 620 nm, as this matched the peak excitation wavelength of the phycobilin pigments.

Another observation is that the three single-celled organisms CPCC 552, CPCC 005 and CPCC 300 are very similar in appearance as seen in the brightfield images. However, all three of these algae types come from three different phyla, resulting in their fluorescence spectrum to be quite different. CPCC 552 (Bacillariophyta) has the highest fluorescence response at 445 nm with a progressively lower response at 500 nm, 545 nm and 620 nm respectively. CPCC 005 (Chlorophyta) follows a similar pattern, however, the intensity at 445 is much larger compared to CPCC 552. In addition, the response at 620 nm is much higher for CPCC 005 compared to that of CPCC 552. Also, CPCC 300 (Cyanophyta) has the highest and a similar response to CPCC 005 at 620 nm. Relative to the CPCC 552 and CPCC 005 the fluorescence spectrum is extremely unique.

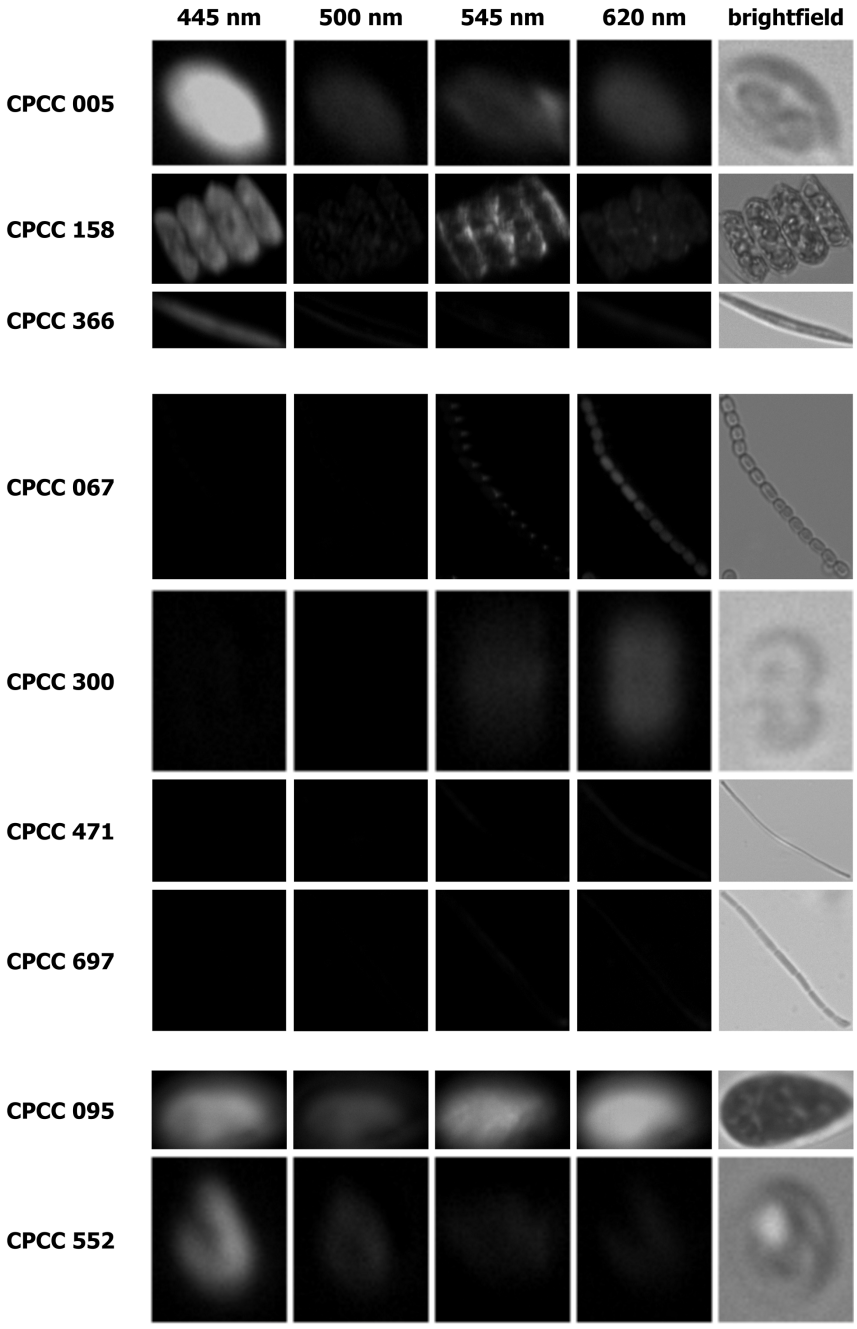


Figure 7.1: The nine algae types from four phyla groups at four excitation wavelengths (445 nm, 500 nm, 545 nm, and 620 nm) as well as the single brightfield image. This data was collected with the imaging device from Chapter 4 and preprocessed using the methods in Chapter 6. The average numerical values of the entire dataset can be seen in Figure 7.2.

The other three algae types (CPCC 095, CPCC 158, CPCC 366) are seen to have significantly different fluorescent spectra. CPCC 095 belongs to the Euglenozoa phyla and is seen to have strong fluorescence across the entire spectrum, indicating a strong presence of many different pigments. CPCC 158 and CPCC 366 both belong in the Cyanophyta phylum and follow a similar fluorescence spectrum. However, CPCC 158 has a much stronger signal at 545 nm compared to CPCC 366, indicating a higher concentration of phycobilin pigments. This observation shows that there can be significant pigmentation variations within a given phyla.

7.1.2 Spectral Analysis

To gain further insights into the fluorescence data the spectral features from Section 6.4.2 were averaged over all the different classes. A plot of these features for the nine different algae types can be seen in Figure 7.2. The excitation spectra of the Chlorophyta algae, the Cyanophyta, and the Euglenozoa with the Bacillariophyta algae can be seen in the top, middle and bottom plots respectfully. It is important to note that the y-axis scale is different for the three plots in this figure. These plots also validate the speculation from Section 3.1 that different algae types have different concentration of pigments.

One initial observation is that all three of the Chlorophyta algae species have a much larger fluorescence signal at 385 nm and 405 nm compared to the Cyanophyta algae, which matches results presented by Poryvkina *et al.* [51]. This difference in fluorescent intensity is due to the difference in pigmentation between each phylum, as previously discussed in Section 6.1. Another observation is that *Euglena* is consistently higher in excitation intensity compared to the other samples as seen in Figure 7.2 (bottom). In fact, the lowest *Euglena* signal, 0.2 at 500 nm, is approximately the same value of the highest value of excitation intensity of the other organisms. Furthermore, at 620 nm the range of Chlorophyta algae is the same as the Cyanophyta algae (from 0.0 to 0.1). This observation indicates that given the presence of certain algae, the single 620 nm excitation wavelength is likely not sufficient to discriminate between Cyanophyta and other organisms that have pigments which fluoresce at at 620 nm. For example, if *Euglena* was present in the same water sample as other Chlorophyta or Cyanophyta, the *Euglena* signal would overpower the other algae emission.

From this analysis in Section 7.1.1 and Section 7.1.2, we can conclude that the spectral fluorescence images are likely to improve the classification performance compared to when we classify these algae with just the brightfield images. We have observed that different phyla have significant different fluorescence spectra, and within one phylum there

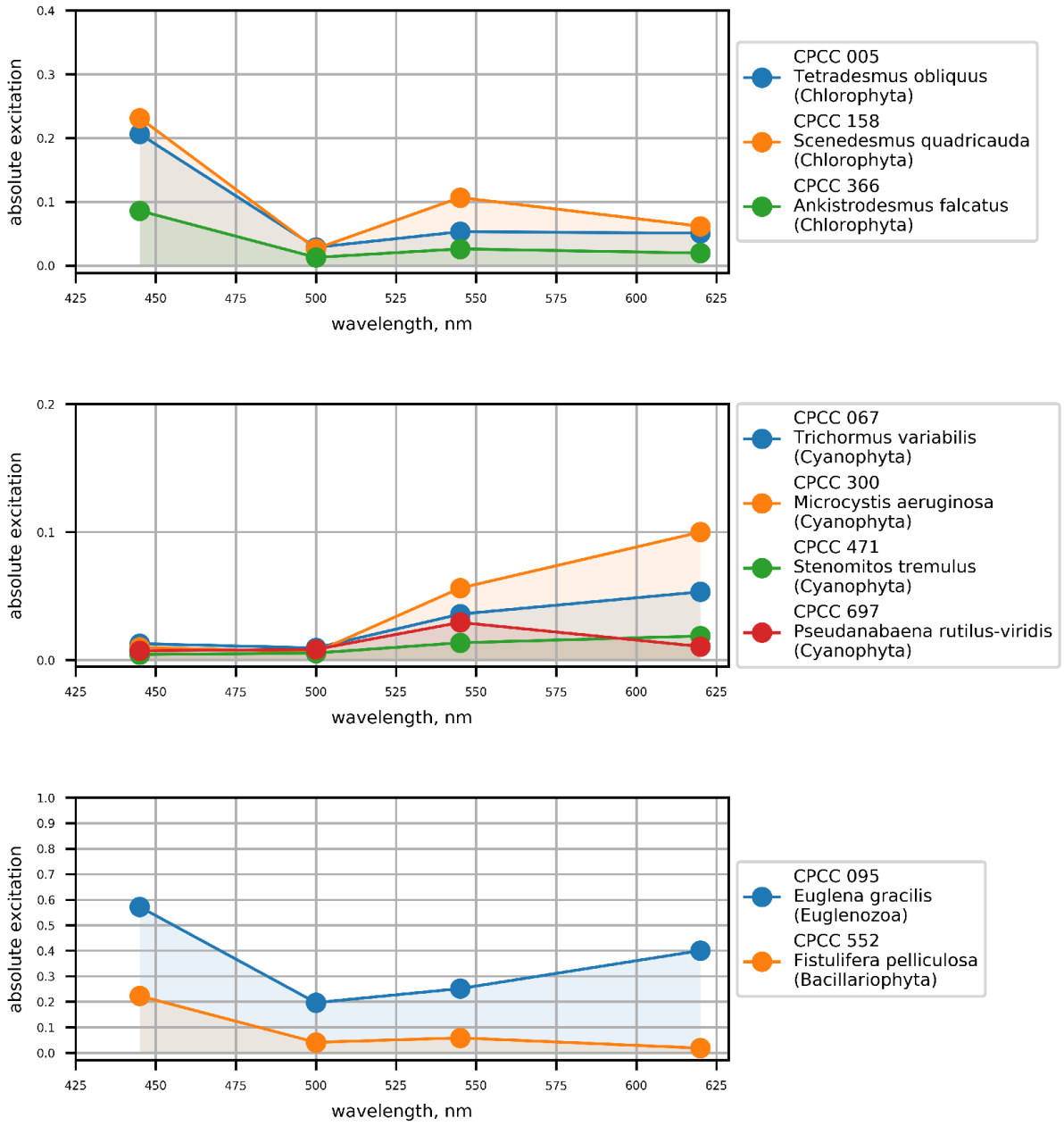


Figure 7.2: The multispectral emission spectra from nine types of algae when excited at 445 nm, 500 nm, 545 nm and 620 nm. Note the changes in the y-axis scale for each subplot. These spectra show that different phyla groups have similar emission spectra. The images for these nine algae are presented in the same order as in Figure 7.1.

can even be some variation. Given that many of these algae have similar morphological characteristics, this additional fluorescence information will likely improve the classification performance.

7.2 Quantitative Analysis

In Section 7.1 we visually inspected the spatial and spectral aspects of this data in order to build an intuition of the data and the classification task. Having completed this qualitative analysis the current section will conduct a quantitative analysis. In order to perform a quantitative analysis on the relative performance of the 12 proposed models, a dataset was created in Chapter 6. In Chapter 5 three main questions where presented, which resulted in 12 unique models. To recap, those questions were:

1. Which data do we use? (Section 5.1)
2. How is the data represented? (Section 5.2)
3. How does the data propagate through the model? (Section 5.3)

Each of the 12 proposed models from Chapter 5 were trained and tested using a seven-fold monte-carlo cross validation scheme. In each run, the training set consisted of 70% of the data, while the validation set and test set each consisted of 15% of the data. The learning rate for all models was kept at 0.1 and the batch size for all runs was 64.

The final mean accuracy with standard deviation of each of the 12 models, averaged over the seven runs, can be seen in Table 7.1. In addition, the confusion matrices of the first run of these 12 models can be found in Appendix A. Having these results allows us to now answer the questions from Chapter 5. Each question will be explored in detail; however, they will be answered out of order since, as seen in Table 7.1, the data representation (Question 2) has the largest impact on model performance, followed by the data modality (Question 1), followed by whether a hierarchical structure was used (Question 3). Therefore:

1. Question 2 will be explored in Section 7.2.1: Data Representation Analysis
2. Question 1 will be explored in Section 7.2.2: Data Modality Analysis
3. Question 3 will be explored in Section 7.2.3: Data Propagation Analysis

data representation	data modality	hierarchical structure	train accuracy	validation accuracy	test accuracy
Feature Extraction	Brightfield	no	78.19% ± 1.78%	77.22% ± 1.71%	76.38% ± 1.21%
		yes	77.75% ± 0.71%	77.54% ± 0.97%	76.77% ± 1.02%
	Fluorescence	no	80.93% ± 1.39%	80.64% ± 1.42%	80.52% ± 1.48%
		yes	79.62% ± 0.72%	80.92% ± 1.32%	79.12% ± 1.45%
	Brightfield + Fluorescence	no	84.41% ± 0.97%	83.50% ± 0.98%	83.18% ± 0.64%
		yes	83.85% ± 0.93%	84.40% ± 0.95%	82.63% ± 0.87%
Feature Learning	Brightfield	no	99.99% ± 0.01%	94.40% ± 0.95%	93.02% ± 0.97%
		yes	99.02% ± 0.95%	91.75% ± 1.15%	90.72% ± 1.92%
	Fluorescence	no	99.98% ± 0.02%	97.97% ± 0.27%	97.39% ± 0.28%
		yes	99.80% ± 0.20%	98.11% ± 0.25%	96.91% ± 0.72%
	Brightfield + Fluorescence	no	99.98% ± 0.03%	98.23% ± 0.33%	97.81% ± 0.54%
		yes	99.78% ± 0.15%	98.27% ± 0.18%	97.63% ± 0.44%

Table 7.1: The classification accuracy is reported for the train, validation, and test sets when evaluating the 12 proposed software frameworks.

7.2.1 Data Representation Analysis

In order have a fair and direct comparison between feature extraction and feature learning based models no data augmentation was used. Data augmentation is a common practice when using images, however, it is more challenging to augment a dataset from extracted features. If the feature learning model used data augmentation it would automatically see a performance increase compared to the feature extraction based model. This would bias the results as it would favour the feature learning model.

In addition, the feature extraction model (feedforward neural network) ran for 300 epochs, while the feature learning models (modified ResNet18) ran for 100 epochs. The difference in epochs was required to ensure that different model types each converged and then plateaued at their optimal solutions. The feature learning model (modified ResNet18) only trained the new input and output modified layers for the first 15 epochs while keeping the original weights of the modified ResNet18 model frozen. After 15 epochs, all the layers were unfrozen to allow the model to fine-tune its weights in order to learn the optimal feature set across the entire convolutional neural network.

As seen in Table 7.1, all the feature learning models achieved 99% accuracy on the training data compared to the feature extraction based models, where the highest training accuracy was 84.41% for the brightfield and fluorescence classifier that did not use a

hierarchical structure. This immediately demonstrates that feature-learning methods are preferred, compared to feature-extraction methods as they have higher performance and require less preprocessing steps. Although the feature learning based models have very high training accuracy, not all these corresponding models have high validation and test accuracies. Specifically, the brightfield models have significantly lower validation and test accuracies compared to the training accuracy, indicating that the brightfield models that used feature learning are overfit to the training data.

One potential drawback of the feature learning approach with a fixed architecture is that the input image size is fixed to a certain dimension. For example, in our case, the feature learning model expects an image with the dimensions of 224×224. However, as seen in Figure 6.7, a given crop ranges widely in dimensions. Smaller organisms have a very small region of interest (ROI), while larger organisms have a dramatically larger ROI. Furthermore, the filamentous organisms, such as CPCC 067, CPCC 471, and CPCC 697, can be very elongated, resulting in one dimension being significantly larger than the other dimension. Since the feature-learning model requires a 224×224 image input size, all the cropped ROI images must be resized before entering the network. Therefore all scale information is lost when resizing these images, which results in the ResNet18 model being scale invariant. In this situation, where the classification accuracy is still very high, losing the scale information has no impact on the model performance and the feature learning models still outperforms the feature extraction models. However, given that the different sizes and scales of these different microorganisms can be used to classify them, scale variance might a useful characteristic to retain, especially when dealing with more output classes. That is, as new organisms are added into the database scale variance may be a useful feature to maintain.

In summary, even with a one third the number of epochs, the feature learning approach consistently outperformed the feature extraction approach. In addition, the feature extraction approach requires prepossessing the data before it can enter into the neural network, while the feature learning approach directly takes the image as input into the network. Therefore feature learning is a superior method over feature extraction as feature learning requires less work up front, and it achieves higher classification accuracy compared to feature extraction.

7.2.2 Data Modality Analysis

Inspecting Table 7.1 one can observe that in both the feature extraction and feature learning models, the test accuracy is the lowest for the brightfield models, slightly higher for the

fluorescence models, and the highest for the combined brightfield and fluorescence models. For the feature extraction models, this is a nearly linear increase as the brightfield model has approximately 76% test accuracy, the fluorescence model has approximately 80% test accuracy, and the brightfield and fluorescence model has approximately 83% test accuracy. For the feature learning models, the relationship is more asymptotic as the brightfield model has approximately 90% - 93% test accuracy, the fluorescence model has approximately 97% test accuracy, and the brightfield and fluorescence model has approximately 97% test accuracy.

These results indicate that the fluorescence data modality has more of an impact on model performance than the brightfield data modality. Since we have four times more data for the multispectral fluorescence data (4 images) compared to the bright field data (1 image), this increase is to be expected for the feature learning. However, in the case of feature extraction, there are 47 features extracted from the brightfield image (see Section 6.4.1) and only four features extracted from the multispectral fluorescence images (see Section 6.4.2). Therefore the four fluorescence spectral features contain more useful information than all of the 47 brightfield spatial features. This shows the value of capturing the natural autofluorescence of algae, as fluorescence data is able to achieve approximately 80% test accuracy with only the extracted fluorescence spectral features, and approximately 97% test accuracy when using the images as input to a feature learning model.

These results also show that there is value in simultaneously using both the brightfield data and the fluorescence data as inputs to the model when using feature extraction, but not when using feature learning. For the feature extraction models the combined brightfield and multispectral fluorescence data achieves the highest accuracy of approximately 83%, which is 3% higher than just the fluorescence model. However, for the feature learning model from a strict observable improvement when using just the fluorescence data compared to using the combined brightfield and fluorescence data. Therefore, one could leverage just the fluorescence data in a feature learning paradigm and achieve the same performance when learning a model from both the brightfield and fluorescence data. As concluded in Section 7.2.1, this reinforces the value of using a feature learning paradigm over a feature extraction paradigm.

In summary, in both the feature extraction and feature learning paradigms the fluorescence data had a higher accuracy compared to the brightfield data. This shows the value of capturing the auto-fluorescence of algae and demonstrates the efficacy of the proposed imaging system in Chapter 4.

data representation	data modality	hierarchical structure	test accuracy	T-Test null hypothesis (p value = 0.05)	T-Test null hypothesis (p value = 0.01)
Feature Extraction	Brightfield	no	76.38% ± 1.21%	TRUE	TRUE
		yes	76.77% ± 1.02%		
	Fluorescence	no	80.52% ± 1.48%	TRUE	TRUE
		yes	79.12% ± 1.45%		
	Brightfield + Fluorescence	no	83.18% ± 0.64%	TRUE	TRUE
		yes	82.63% ± 0.87%		
Feature Learning	Brightfield	no	**93.02% ± 0.97%**	**FALSE**	TRUE
		yes	**90.72% ± 1.92%**		
	Fluorescence	no	97.39% ± 0.28%	TRUE	TRUE
		yes	96.91% ± 0.72%		
	Brightfield + Fluorescence	no	97.81% ± 0.54%	TRUE	TRUE
		yes	97.63% ± 0.44%		

Table 7.2: A T-Test between two consecutive models was run where the only difference in the models is the use of a flat structure as opposed to a hierarchical structure. For a p-value of 0.01 there was no noticeable improvement between any of the models. For a p-value of 0.05 only one model showed a statically significant decrease in performance when using a hierarchical approach.

7.2.3 Data Propagation Analysis

The final aspect when inspecting Table 7.1 is whether a flat structure or a hierarchical structure was used while learning a model. For a given model, irrespective of which data modality was used and irrespective of how that data was represented, the flat structure consistently outperformed the hierarchical structure by a marginal amount.

T-Test Analysis

Since the increase in accuracy of the flat structure is so small compared to that of the hierarchical structure, Welch's t-test was performed between each of these two models types. The null hypothesis of the t-test is that the two sets of accuracy averages are not statistically significant from each other. In order words, the null hypothesis assumes that the two sets of accuracy averages are very similar to each other. Therefore the alternative hypothesis is that the mean of two sets of accuracies are statistically significant from each other, and therefore there is a noticeable difference from one model to the other. Thus any two models that reject the null hypothesis indicate a statically significant difference.

As seen in Table 7.2, when using a p-value of 0.01 the null hypothesis is true in each case, and therefore there is no observable improvement between the flat structure and the hierarchy structure. However, Table 7.2 also shows that when a p-value of 0.05 is used the feature learning brightfield model does reject the null hypothesis, and therefore there is a noticeable difference between the the flat structure and the hierarchy structure. In this case, the flat structure model had an accuracy of 93.02% while the hierarchy structure had an accuracy of 90.72%. Therefore, from a strictly accuracy standpoint, in this situation it is better to allow the network to learn an implicit hierarchy when classifying in an flat structure opposed to reinforcing a given hierarchy based off the taxonomic structure of the biological organisms. However, this is not strongly supported as it depends on whether a p-value of 0.05 or 0.01 is used.

Overall, since the flat structure and the hierarchy structure have similar performance, the hierarchy model is preferred. The main advantages of using a hierarchical approach is that it allows for online learning and it increases the explainability of the model compared to using a flat structure model.

Hierarchy Performance

As previously mentioned, one advantage of having a hierarchical model is that the tree structure can be decomposed, which allows for increased understanding of the sources of error. Table 7.3, breaks down the test accuracy from Table 7.1 into the three taxonomy levels of phyla, genus and species. Inspecting Table 7.3 reveals the astonishing fact that the feature extraction fluorescence based hierarchical classifier was able to achieve 95.78% accuracy to the phylum level. This is noticeably higher than feature extraction brightfield and fluorescence based hierarchical classifier which achieved 91.90% accuracy to phylum level. This indicates that adding the brightfield features with the fluorescence features actually reduces the performance of the model classification.

In addition to outperforming the feature extraction brightfield and fluorescence hierarchical classifier, the feature extraction fluorescence based hierarchical classifier had similar performance to that of the feature learning brightfield classifier, which achieved 94.45% to the phylum level. This shows that the four average values from the fluorescence spectra features provide the same amount of useful information to that of the single brightfield image in order to classify a microorganism to phyla level. This provides strong evidence that the multispectral fluorescence data is critical to the success of identifying different algae type.

data representation	data modality	hierarchical level	test accuracy
Feature Extraction	Brightfield	Phyla	84.79% ± 1.15%
		Genus	76.77% ± 1.02%
		Species	76.77% ± 1.02%
	Fluorescence	Phyla	**95.78% ± 0.99%**
		Genus	79.12% ± 1.45%
		Species	79.12% ± 1.45%
	Brightfield + Fluorescence	Phyla	**91.90% ± 0.71%**
		Genus	82.63% ± 0.87%
		Species	82.63% ± 0.87%
Feature Learning	Brightfield	Phyla	**94.45% ± 1.71%**
		Genus	90.72% ± 1.92%
		Species	90.72% ± 1.92%
	Fluorescence	Phyla	99.03% ± 0.45%
		Genus	96.91% ± 0.72%
		Species	96.91% ± 0.72%
	Brightfield + Fluorescence	Phyla	99.25% ± 0.29%
		Genus	97.63% ± 0.44%
		Species	97.63% ± 0.44%

Table 7.3: A further breakdown of the 12 model architectures reveals how the accuracy of the model throughout the three levels of the tree.

7.3 Summary of Performance

This chapter started by qualitatively analyzing the data generating from Chapter 6. Inspecting this data resulted in the conclusion that the multispectral fluorescence data was likely to improve the classification accuracy as there were noticeable differences in the spectra between the four phyla. This was then confirmed to be true by conducting a quantitative analysis of the 12 proposed model architectures in Chapter 5.

In summary, we observed that (1) feature learning drastically outperformed feature extraction, (2) using fluorescence data results in a higher classification accuracy compared to brightfield data, and (3) a flat structure and a hierarchical structure do not have statistically significant different performance.

Chapter 8

Conclusions & Future Work

"The time to work on a problem is after you've solved it."

– R. H. Bing (1914 - 1986)

In this final chapter we will discuss the main conclusions of this work in Section 8.1 as well as the potential future research directions in Section 8.2.

8.1 Conclusions

This thesis introduced a low-cost system that can generate data on-site as well as analyze this data in real-time for the use case of algae identification. As discussed in Chapter 4, the on-site data generation was accomplished by building a low-cost imaging system capable of capturing a single brightfield image and four fluorescence images. This imaging system was built for approximately $4,000 USD and was able to achieve a spatial resolution of 0.65 um/pixel. The imaging system is composed of magnification optics, a camera sensor, a printed circuit board (PCB), and LEDs. All these elements are housed in a 3D printed frame and are controlled by a customer graphical user interface (GUI).

To complement this hardware system a software framework was developed to analyze the data in real-time, as discussed in Chapter 5. This software system uses a feature learning approach along with a hierarchical classification scheme which was inspired by

the taxonomic breakdown of algae into the phylum, genus and species levels. Twelve model architectures were discussed and created by asking three main questions:

1. Which data do we use? (Section 5.1)
2. How is the data represented? (Section 5.2)
3. How does the data propagate through the model? (Section 5.3)

To answer these questions a dataset was created in Chapter 6, imaging nine types of algae from four different phyla groups. In order prepare the raw images for data analysis the region of interests (ROIs) had to be located. These ROIs were located by first conducting flat field correction, then thresholding and finally cropping. Next a number of spatial and spectral features were extracted from these images in order answer question 2 from Chapter 5. After the raw images had been preprocessed there were a total of 6330 images representing all nine algae classes.

Having a dataset made it possible to determine the efficacy of the proposed imaging system and software framework, as discussed in Chapter 7. We observed that (1) feature learning outperformed feature extraction, (2) using fluorescence data resulted in a higher classification accuracy compared to brightfield data, and (3) a flat structure and a hierarchical structure had statistically significant similar performance. From these three observations we can conclude that the proposed multispectral fluorescence imaging system (Chapter 4), along with the proposed hierarchical structure that is based on a taxonomic prior (Chapter 5), is an effective method to automatically classify algae.

8.2 Future Work

The future direction of this research can be broken into three main sections. In Section 8.2.1 will we will discuss working with natural water samples. In Section 8.2.2 future work around semantic segmentation will be presented. In Section 8.2.3 a quick analysis of separating live and dead cells will be discussed. Finally, Section 8.2.4 a brief discussion of future work regarding the hierarchical structure will be presented.

8.2.1 Natural Samples

Future work involves collecting natural samples and evaluating the classification performance on this new data. The dataset described in Chapter 6 was created by imaging nine

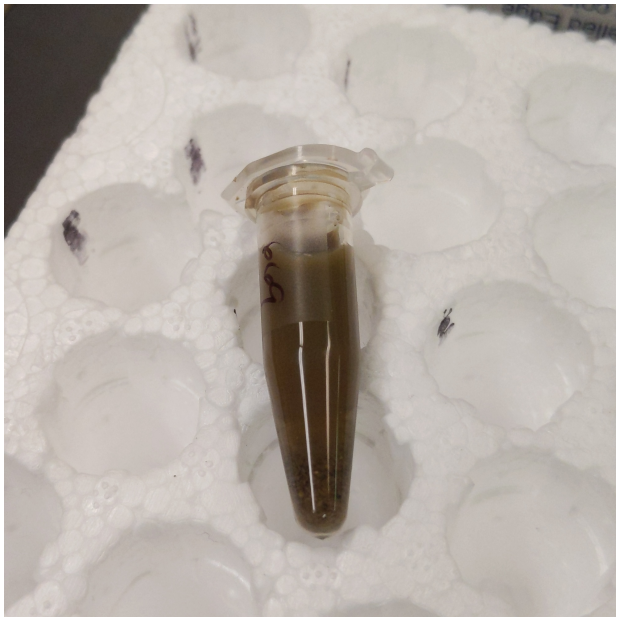


Figure 8.1: A sample was prepared by artificially mixing pure algae samples and then adding dirty water from an outdoor puddle. The corresponding microscope images can be seen in Figure 8.2.

sets of pure algae cultures. However real-world samples, as seen in Figure 8.1, can be extremely dirty due to particulates in the sample. The sample in Figure 8.1 was placed under a brightfield and epifluorescence microscope, as see in Figure 8.2. As seen from this figure, it is very unlikely that the algae can be properly identified by only using the brightfield image. However, the fluorescence signal causes the algae to fluoresce, resulting in the algae being easily identified from the background.

In order to conduct a study with natural samples, a professional taxonomist must be employed who is qualified to label regions within a given image. This adds significant time to exploring this path, but nonetheless it is still a worthwhile direction to pursue. The purpose of the proposed imaging system in this thesis is its usefulness to an individual who doesn't have the time or budget to send a sample away for analysis. In order to reach this goal, the water samples must graduate from lab samples to real-world samples.

One path to aid in the creation of a labelled dataset from natural samples is the use of unsupervised learning approaches [95]. Given the human taxonomist can determine how many different types of organisms are in a water sample, an unsupervised learning approach can be taken where only the images near the class boundaries are given to a

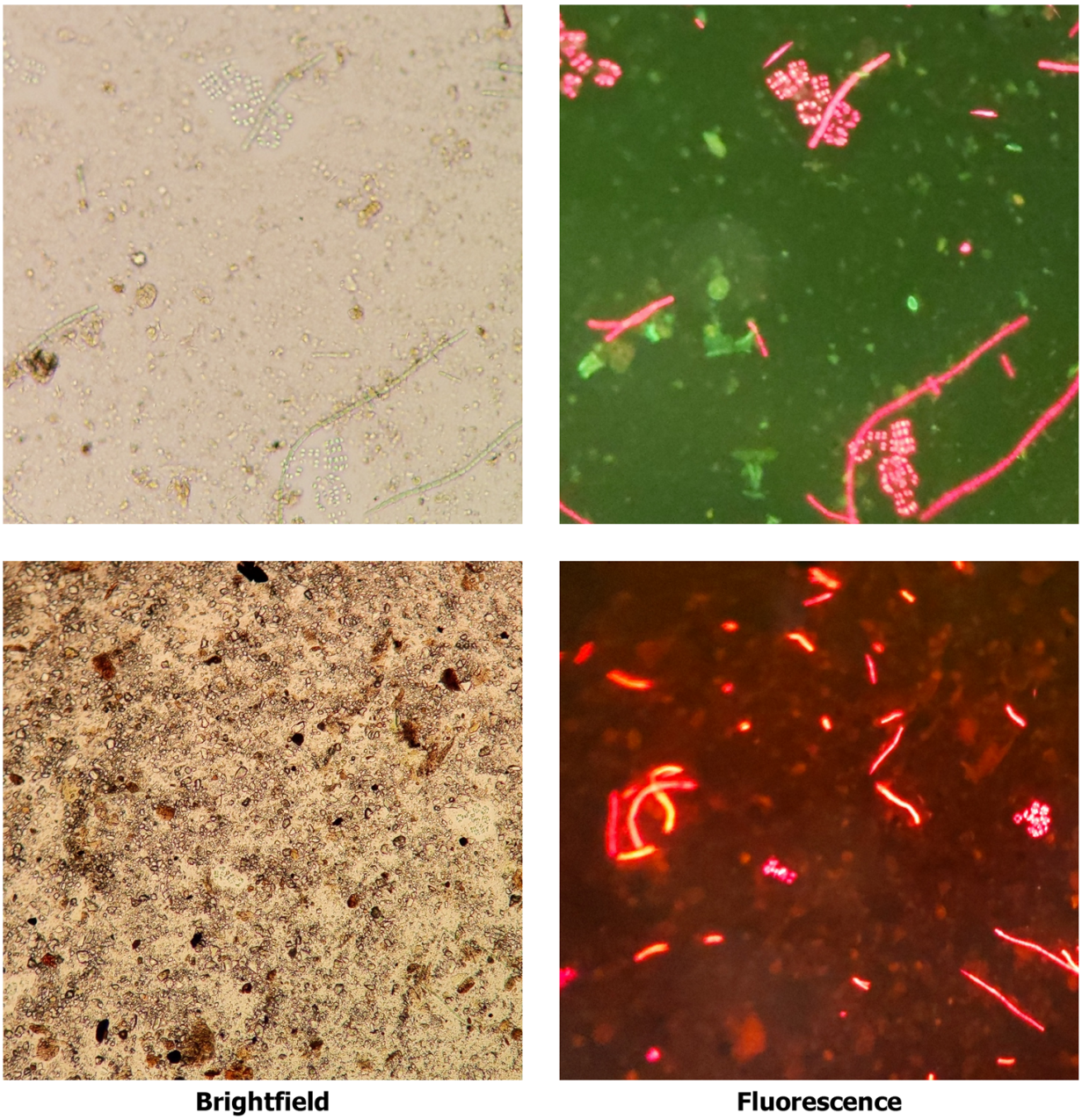


Figure 8.2: A sample of water from Figure 8.1 was placed under a microscope to capture a brightfield image and fluorescence image. These images illustrate the benefit of using fluorescence for identification of algae in a contaminated sample.

human for labelling. This process can be iterative and thus large amounts of data can be rapidly labelled and then verified by a human. This avoids the human expert having to comb through all the data and manually labelling each one.

8.2.2 Semantic Segmentation

Another future research direction is to use semantic segmentation approaches to solve the problem of identifying algae types in an image. For instance, inspecting Figure 8.3, we can see that there are multiple algae types in a single image and that some of these organisms overlap each other. Specifically, in Figure 8.3, many filamentous algae cover a large area of the image and intersect with other algae. In this situation the current approach of selecting a region of interest is no longer suitable, since multiple algae occupy a single region. Given that a binary mask was created in Section 6.3.2, this mask can be used to learn a semantic segmentation model to automatically segment and classify algae in images. This also removes the issue of resizing the images to a fixed dimension when passing into a image classifier (such as a ResNet18), which removes all scale information.

The scale of an organism is strongly related to its identity, and thus retaining this scale information by using semantic segmentation approaches could improve the ability to automatically identify algae in a water sample through a digital microscope. Many standard semantic segmentation models exists which offer a great starting point to determine how effective this approach could be. These models include U-Net by Ronneberger *et al.* [96], DeepLab by Chen *et al.* [97], and RefineNet by Lin *et al.* [98].

8.2.3 Separating Live & Dead Cells

To date a number of researchers have explored staining certain phytoplankton to separate live and dead organisms [99, 100, 101]. Figure 8.4 illustrates two unstained cells beside each other under brightfield and fluorescence illumination. In the brightfield configuration the cells look similar in nature. However in the fluorescence setup the dead cell is no longer visible, while the living cell has a very strong fluorescence response. Therefore a potential avenue to explore with the proposed setup is to determine if live and dead cells can be separated using the autofluorescence signal.

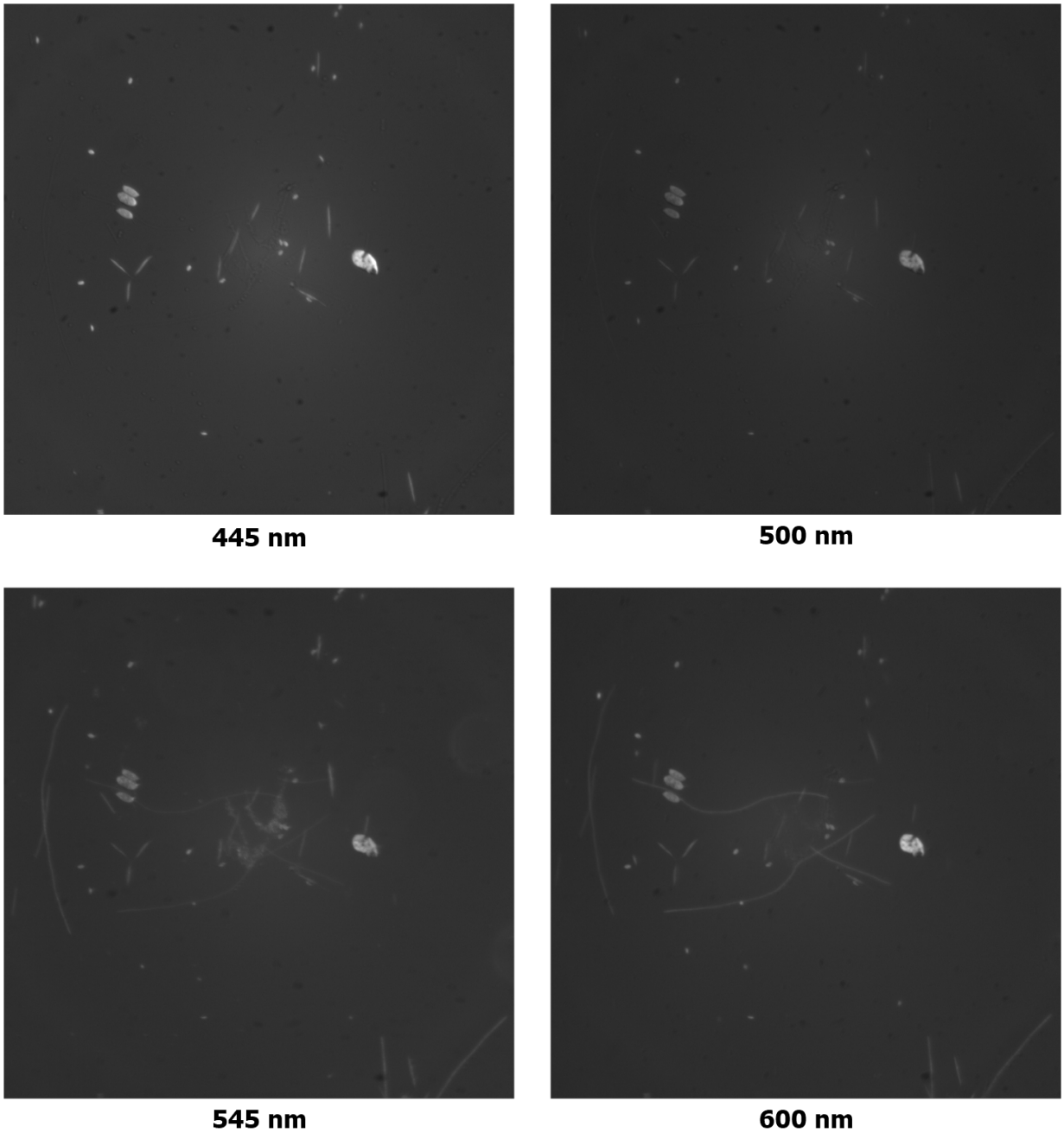


Figure 8.3: A mixed sample under 445 nm, 500 nm, 545 nm and 600 nm. Note that multiple algae types intersect with each other, making the current approach of cropping a region of interest no longer suitable. Therefore it is recommended that future work explores using semantic segmentation to achieve a per pixel classification.

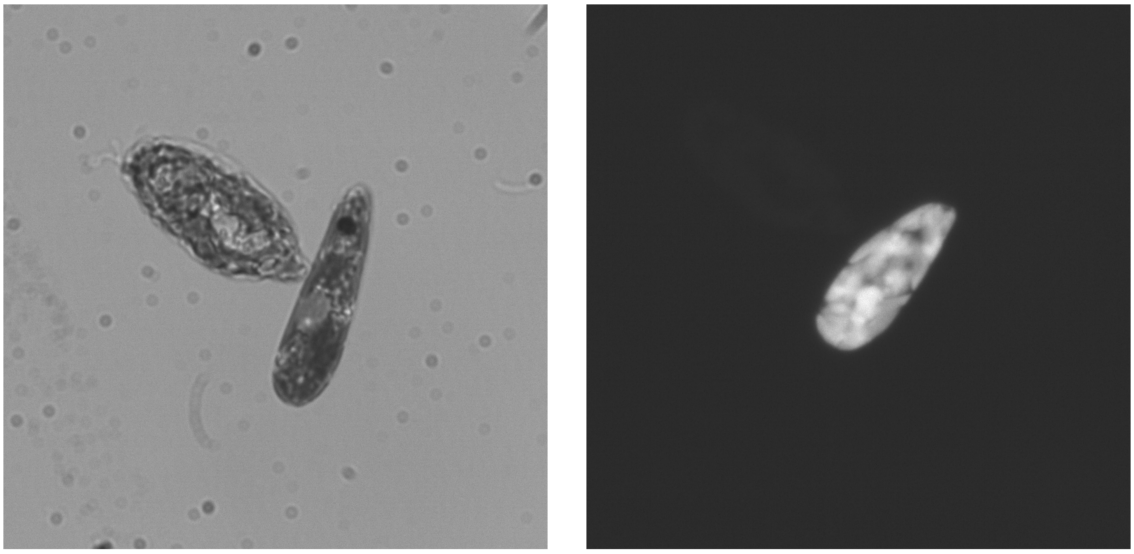


Figure 8.4: A dead cell and a live cell under brightfield and fluorescence illumination. Since only the living cell is visible under fluorescence light, an interesting research direction is to explore using the autofluorescence of algae to separate between live and dead cells.

8.2.4 Exploring the Hierarchical Structure

A final research direction to explore concerns the hierarchical nature of the model architecture. To further explore the benefits of this architecture it would be interesting to determine how effective it is at online learning. This could be done by adding a single new class to the structure and only retraining the node that makes the final decision. Assuming parent nodes in the tree structure are not updated, how well would the overall performance change? This is an important aspect to explore, because when the parent node misclassifies an image, the error will propagate throughout the entire tree.

This approach of only training the last child node may perform even better when we don't use the existing taxonomy structure as a prior, but to learn a taxonomy structure based off the data. This research field is known as phenetics or taximetrics, and strictly classifies organisms based on similar traits by deliberately ignoring the phylogentic tree [102]. The advantage of using phenetics with the generated dataset is that we have both fluorescence spectra information and the standard spatial imaging data. Having this spectral-spatial data may offer new insights into the origins of certain organisms and the relationships between them.

8.3 Closing Words

This final chapter concludes that the proposed system is able to achieve fine-grained classification when using multispectral fluorescence data with a hierarchical framework. Some initial ideas of future work were also discussed. In addition to these future research directions there are many other interesting possibilities to explore as this thesis is only the tip of the iceberg of research that could be conducted with the proposed imaging system. It is the author's hope that this research will inspire others to join in solving this problem as well as work on solving other important problems that stand in the way of people living healthy and productive lives.

References

[1] General Assembly. sustainable development goals. *Transforming our world*, 2015.

[2] Toxic algae bloom in lake erie. *NASA*, Oct 2011.

[3] Bjorn Lomborg. *Prioritizing development: a cost benefit analysis of the United Nations' sustainable development goals*. Cambridge University Press, 2018.

[4] Anna M Michalak, Eric J Anderson, Dmitry Beletsky, Steven Boland, Nathan S Bosch, Thomas B Bridgeman, Justin D Chaffin, Kyunghwa Cho, Rem Confesor, Irem Daloğlu, et al. Record-setting algal bloom in lake erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences*, 110(16):6448–6452, 2013.

[5] Ian R Falconer. Potential impact on human health of toxic cyanobacteria. *Phycologia*, 35(6S):6–11, 1996.

[6] Primo Coltelli, Laura Barsanti, Valtere Evangelista, Anna Maria Frassanito, and Paolo Gualtieri. Water monitoring: automated and real time identification and classification of algae using digital microscopy. *Environmental Science: Processes & Impacts*, 16(11):2656–2665, 2014.

[7] Stefan U Thiel, Ron J Wiltshire, and Lance J Davies. Automated object recognition of blue-green algae for measuring water qualitya preliminary study. *Water Research*, 29(10):2398–2404, 1995.

[8] Ross F Walker and Michio Kumagai. Image analysis as a tool for quantitative phycology: a computational approach to cyanobacterial taxa identification. *Limnology*, 1(2):107–115, 2000.

[9] Michael E Sieracki, Mark Benfield, Allen Hanson, Cabell Davis, Cynthia H Pilskaln, David Checkley, Heidi M Sosik, Carin Ashjian, Phil Culverhouse, Robert Cowen, et al. Optical plankton imaging and analysis systems for ocean observation. *Proceedings of ocean Obs*, 9:21–25, 2010.

[10] J Kasich, C Butler, J Zehringer, and L Himes. State of ohio harmful algal bloom response strategy for recreational waters. *Department of Health, Environmental Protection Agency and Department of Natural Resources*, 2012.

[11] Minister of Health. *Guidelines for Canadian recreational water quality: Third Edition*. Health Canada, Ottawa, 2012.

[12] Hans W Paerl and Jef Huisman. Blooms like it hot. *Science*, 320(5872):57–58, 2008.

[13] Isabella Sanseverino, Diana Conduto, Luca Pozzoli, Srdan Dobricic, and Teresa Lettieri. Algal bloom and its economic impact. *European Commission, Joint Research Centre Institute for Environment and Sustainability*, 2016.

[14] Morgan M Steffen, Timothy W Davis, R Michael L McKay, George S Bullerjahn, Lauren E Krausfeldt, Joshua MA Stough, Michelle L Neitzey, Naomi E Gilbert, Gregory L Boyer, Thomas H Johengen, et al. Ecophysiological examination of the lake erie microcystis bloom in 2014: linkages between biology and the water supply shutdown of toledo, oh. *Environmental science & technology*, 51(12):6745–6755, 2017.

[15] JC Ho, AM Michalak, and N Pahlevan. Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, pages 1–1, 2019.

[16] JM Oneil, TW Davis, MA Burford, and CJ Gobler. The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. *Harmful algae*, 14:313–334, 2012.

[17] World Health Organization et al. Cyanobacterial toxins: microcystin-lr. *Guidelines for drinking water quality*, 2, 1998.

[18] Health Canada. Canadian drinking water guidelines. *Cyanobacterial Toxins - Microcystin-LR*, July 2002.

[19] US Congress. Harmful algal bloom and hypoxia research and control amendments act of 2014. *Pub. S*, 1254, 2014.

[20] Xuexiang He, Yen-Ling Liu, Amanda Conklin, Judy Westrick, Linda K Weavers, Dionysios D Dionysiou, John J Lenhart, Paula J Mouser, David Szlag, and Harold W Walker. Toxic cyanobacteria and drinking water: impacts, detection, and treatment. *Harmful algae*, 54:174–193, 2016.

[21] Laura Barsanti and Paolo Gualtieri. *Algae: anatomy, biochemistry, and biotechnology*. CRC press, 2014.

[22] Lily Newton. Handbook of the British seaweeds. 1931.

[23] William Eifion Jones. A key to the genera of the british seaweeds. 1964.

[24] Edward G Bellinger and David C Sigee. *Freshwater algae: identification and use as bioindicators*. John Wiley & Sons, 2015.

[25] John D Wehr, Robert G Sheath, and J Patrick Kociolek. *Freshwater algae of North America: ecology and classification*. Elsevier, 2015.

[26] M Huynh and N Serediak. Algae identification field guide agriculture and agric food canada. In *Report number: Cat. No. A125-8/2-2011E-PDF*. 2006.

[27] AL Baker et al. Phycokey–an image based key to algae (ps protista), cyanobacteria, and other aquatic objects. *University of New Hampshire Center for Freshwater Biology*, 2012.

[28] Chris Carter, Joanna Wilbraham, and David John. Algaevision: Virtual collection of freshwater algae from the british isles. version ii. *Natural History Museum*, 2016.

[29] Chen Li, Kimiaki Shirahama, and Marcin Grzegorzek. Application of content-based image analysis to environmental microorganism classification. *Biocybernetics and Biomedical Engineering*, 35(1):10–21, 2015.

[30] Phil F Culverhouse, Robert Williams, Beatriz Reguera, Vincent Herry, and Sonsoles González-Gil. Do experts make mistakes? a comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247:17–25, 2003.

[31] Peter Stanley Dixon. *Biology of the Rhodophyta*, volume 4. Oliver & Boyd Edinburgh, 1973.

[32] Primo Coltelli, Laura Barsanti, Valter Evangelista, Anna Frassanito, and Paolo Gualtieri. Reconstruction of the absorption spectrum of an object spot from the colour values of the corresponding pixel (s) in its digital image: the challenge of algal colours. *Journal of Microscopy*, 264(3):311–320, 2016.

[33] Douglas B Murphy. *Fundamentals of light microscopy and electronic imaging*. John Wiley & Sons, 2002.

[34] Sansoen Promdaen, Pakaket Wattuya, and Nuttha Sanevas. Automated microalgae image classification. *Procedia Computer Science*, 29:1981–1992, 2014.

[35] Saowanee Iamsiri, Nuttha Sanevas, Chakrit Watcharopas, and Pakaket Wattuya. A new shape descriptor and segmentation algorithm for automated classifying of multiple-morphological filamentous algae. In *International Conference on Computational Science*, pages 149–163. Springer, 2019.

[36] William T Mason. *Fluorescent and luminescent probes for biological activity: a practical guide to technology for quantitative real-time analysis*. Elsevier, 1999.

[37] Bernhard Ernst, Stephan Neser, Evelyn OBrien, Stefan J Hoeger, and Daniel R Dietrich. Determination of the filamentous cyanobacteria planktothrix rubescens in environmental water samples using an image processing system. *Harmful algae*, 5(3):281–289, 2006.

[38] Chao Jin, Maria Mesqutia, Monica Emelko, and Alexander Wong. Automated enumeration and size distribution analysis of microcystis aeruginosa via fluorescence imaging. *Journal of Computational Vision and Imaging Systems*, 2(1), 2016.

[39] Chao Jin, Maria Mesqutia, Monica Emelko, and Alexander Wong. Computerized enumeration and bio-volume estimation of the cyanobacteria anabaena flos-aquae. *Journal of Computational Vision and Imaging Systems*, 2(1), 2016.

[40] Karsten Rodenacker, Burkhard Hense, Uta Jütting, and Peter Gais. Automatic analysis of aqueous specimens for phytoplankton structure recognition and population estimation. *Microscopy research and technique*, 69(9):708–720, 2006.

[41] Burkhard A Hense, Peter Gais, Uta Jütting, Hagen Scherb, and Karsten Rodenacker. Use of fluorescence information for automated phytoplankton investigation by image analysis. *Journal of Plankton Research*, 30(5):587–606, 2008.

[42] Ross F Walker, Kanako Ishikawa, and Michio Kumagai. Fluorescence-assisted image analysis of freshwater microalgae. *Journal of microbiological methods*, 51(2):149–162, 2002.

[43] Veronika Dashkova, Dmitry Malashenkov, Nicole Poulton, Ivan Vorobjev, and Natasha S Barteneva. Imaging flow cytometry for phytoplankton analysis. *Methods*, 112:188–200, 2017.

[44] Matthew B Blaschko, Gary Holness, Marwan A Mattar, Dimitri Lisin, Paul E Utgoff, Allen R Hanson, Howard Schultz, and Edward M Riseman. Automatic in situ identification of plankton. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 79–86. IEEE, 2005.

[45] Heidi M Sosik and Robert J Olson. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5(6):204–216, 2007.

[46] Rafael G Colares, Pablo Machado, Matheus de Faria, Amália Detoni, Virgínia Tavano, et al. Microalgae classification using semi-supervised and active learning based on gaussian mixture models. *Journal of the Brazilian Computer Society*, 19(4):411–422, 2013.

[47] Iago Corrêa, Paulo Drews, Márcio Silva de Souza, and Virginia Maria Tavano. Supervised microalgae classification in imbalanced dataset. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pages 49–54. IEEE, 2016.

[48] Zoltán Grcs, Miu Tamamitsu, Vittorio Bianco, Patrick Wolf, Shounak Roy, Koyoshi Shindo, Kyrollos Yanny, Yichen Wu, Hatice Ceylan Koydemir, Yair Rivenson, et al. A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples. *Light: Science & Applications*, 7(1):66, 2018.

[49] Tsaiyun Lee, Mikio Tsuzuki, Toshifumi Takeuchi, Kenji Yokoyama, and Isao Karube. In vivo fluorometric method for early detection of cyanobacterial waterblooms. *Journal of Applied Phycology*, 6(5):489–495, 1994.

[50] Paul Held. Monitoring of algal growth using their intrinsic properties. *Biofuel Research, Application note*, pages 1–5, 2011.

[51] Larisa Poryvkina, Sergey Babichenko, and Aina Leeben. Analysis of phytoplankton pigments by excitation spectra of fluorescence. In *EARSeL-SIG-Workshop LIDAR. Institute of Ecology/LDI, Tallinn, Estonia*, pages 224–232, 2000.

[52] Natalia S Gsponer, M Claudia Rodríguez, Rodrigo E Palacios, and Carlos A Chesta. On the simultaneous identification and quantification of microalgae populations based on fluorometric techniques. *Photochemistry and photobiology*, 2018.

[53] C Stacy French and Violet K Young. The fluorescence spectra of red algae and the transfer of energy from phycoerythrin to phycocyanin and chlorophyll. *The Journal of general physiology*, 35(6):873–890, 1952.

[54] David F Millie, Oscar ME Schofield, Gary J Kirkpatrick, Geir Johnsen, and Terence J Evens. Using absorbance and fluorescence spectra to discriminate microalgae. *European Journal of Phycology*, 37(3):313–322, 2002.

[55] M Beutler, Karen Helen Wiltshire, Bettina Meyer, C Moldaenke, C Lüring, M Meyerhöfer, U-P Hansen, and H Dau. A fluorometric method for the differentiation of algal populations in vivo and in situ. *Photosynthesis research*, 72(1):39–53, 2002.

[56] Jakub Gregor, Blahoslav Maršálek, and Helena Šípková. Detection and estimation of potentially toxic cyanobacteria in raw water at the drinking water treatment plant by in vivo fluorescence method. *Water Research*, 41(1):228–234, 2007.

[57] Xupeng Hu, Rongguo Su, Fang Zhang, Xiulin Wang, Hongtao Wang, and Zhixi Zheng. Multiple excitation wavelength fluorescence emission spectra technique for discrimination of phytoplankton. *Journal of Ocean University of China*, 9(1):16–24, 2010.

[58] N McQuaid, A Zamyadi, M Prévost, DF Bird, and S Dorner. Use of in vivo phycocyanin fluorescence to monitor potential microcystin-producing cyanobacterial biovolume in a drinking water source. *Journal of Environmental Monitoring*, 13(2):455–463, 2011.

[59] Arash Zamyadi, Natasha McQuaid, Sarah Dorner, David F Bird, Mike Burch, Peter Baker, Peter Hobson, Michele Prevost, et al. Cyanobacterial detection using in vivo fluorescence probes: managing interferences for improved decision-making. *Journal-American Water Works Association*, 104(8):E466–E479, 2012.

[60] Arash Zamyadi, Florence Choo, Gayle Newcombe, Richard Stuetz, and Rita K Henderson. A review of monitoring technologies for real-time management of cyanobacteria: Recent advances and future direction. *TrAC Trends in Analytical Chemistry*, 85:83–96, 2016.

[61] Lee C Bowling, Arash Zamyadi, and Rita K Henderson. Assessment of in situ fluorometry to measure cyanobacterial presence in water bodies with diverse cyanobacterial populations. *Water research*, 105:22–33, 2016.

[62] John R Kasich, M Taylor, and Craig W Butler. Public water system harmful algal bloom response strategy. *Ohio Environmental Protection Agency*, 2019.

[63] Olivier Clerck, Michael D Guiry, Frederik Leliaert, Yves Samyn, and Heroen Verbruggen. Algal taxonomy: a road to nowhere? *Journal of Phycology*, 49(2), 2013.

[64] United States Environmental Protection Agency. Water treatment optimization for cyanotoxins. *Office of Water*, 2016.

[65] Man Xiao, Ming Li, and Colin S Reynolds. Colony formation in the cyanobacterium microcystis. *Biological Reviews*, 93(3):1399–1420, 2018.

[66] JS Ploem. The use of a vertical illuminator with interchangeable dichroic mirrors for fluorescence microscopy with incidental light. *Zeitschrift fur wissenschaftliche Mikroskopie und mikroskopische Technik*, 68(3):129–142, 1967.

[67] Richard Zsigmondy. *Colloids and the ultramicroscope: a manual of colloid chemistry and ultramicroscopy*. J. Wiley & sons, 1909.

[68] Peter A Santi. Light sheet fluorescence microscopy: a review. *Journal of Histochemistry & Cytochemistry*, 59(2):129–138, 2011.

[69] Janice Beeler SooHoo, Dale A Kiefer, Donald J Collins, and I Stuart McDermid. In vivo fluorescence excitation and absorption spectra of marine phytoplankton: I. taxonomic characteristics and responses to photoadaptation. *Journal of plankton research*, 8(1):197–214, 1986.

[70] Ariel Lipson, Stephen G Lipson, and Henry Lipson. *Optical physics*. Cambridge University Press, 2010.

[71] Mark S Nixon and Alberto S Aguado. *Feature extraction & image processing for computer vision*. Academic Press, 2012.

[72] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[73] Balázs Csanád Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[74] A Ng. Machine learning yearning: Technical strategy for ai engineers in the era of deep learning, 2019.

[75] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[77] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[80] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399, 2018.

[81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[82] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[83] Carl R Woese, Otto Kandler, and Mark L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[84] John Frederick William Herschel. IV.Aμόρφω a, no. I.on a case of superficial colour presented by a homogeneous liquid internally colourless. *Philosophical Transactions of the Royal Society of London*, (135):143–145, 1845.

[85] Peter V York and Leslie R Johnson. *The freshwater algal flora of the British Isles*. Cambridge University Press, 2002.

[86] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[87] E Yousef Kalafi, Christopher Town, and S Kaur Dhillon. How automated image analysis techniques help scientists in species identification and classification? *Folia morphologica*, 77(2):179–193, 2018.

[88] Pete E Lestrel. *Fourier descriptors and their applications in biology*. Cambridge University Press, 1997.

[89] Eric Persoon and King-Sun Fu. Shape discrimination using fourier descriptors. *IEEE Transactions on systems, man, and cybernetics*, 7(3):170–179, 1977.

[90] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962.

[91] Rafael C Gonzalez and Richard E Woods. Digital image processing, 2012.

[92] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[93] V Sebastian, A Unnikrishnan, Kannan Balakrishnan, et al. Gray level co-occurrence matrices: generalisation and some new features. *arXiv:1205.4831*, 2012.

[94] MATLAB. Texture analysis using the gray-level co-occurrence matrix (glcm). *The MathWorks, Inc.*, 2019.

[95] Richard O Duda, Peter E Hart, and David G Stork. Unsupervised learning and clustering. *Pattern classification*, pages 517–601, 2001.

[96] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[97] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[98] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[99] Eva-Maria Zetsche and Filip JR Meysman. Dead or alive? viability assessment of micro-and mesoplankton. *Journal of Plankton Research*, 34(6):493–509, 2012.

[100] Kam W Tang, Michail I Gladyshev, Olgo P Dubovskaya, Georgiy Kirillin, and Hans-Peter Grossart. Zooplankton carcasses and non-predatory mortality in freshwater and inland sea environments. *Journal of Plankton Research*, 36(3):597–612, 2014.

[101] Hugh L MacIntyre and John J Cullen. Classification of phytoplankton cells as live or dead using the vital stains fluorescein diacetate and 5-chloromethylfluorescein diacetate. *Journal of phycology*, 52(4):572–589, 2016.

[102] Ernst Mayr. Numerical phenetics and taxonomic theory. *Systematic Zoology*, 14(2):73–97, 1965.

APPENDICES

Appendix A

Confusion Matrices

This appendix presents the confusion matrices for the 12 models presented in Chapter 5 and Chapter 7. The original data in chapter 6 was split into 80% train data, 15% validation data, and 15% test data. These confusion matrices were created using the test data from the first of seven monte-carlo cross validation runs.

A.1 Model 1

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

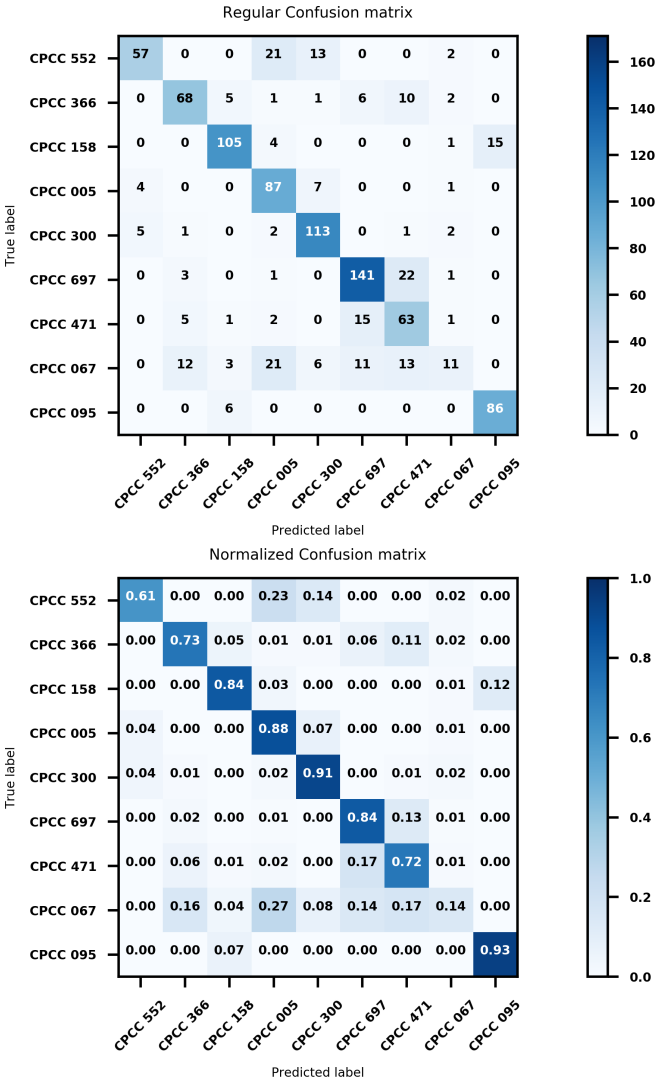
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)



A.2 Model 2

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

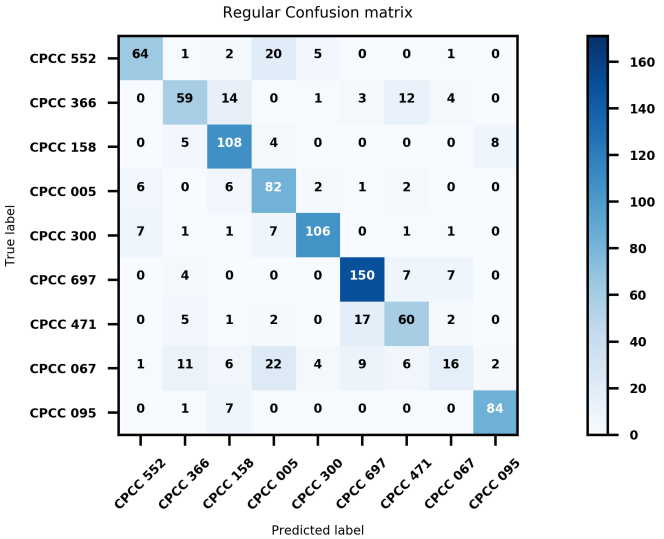
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

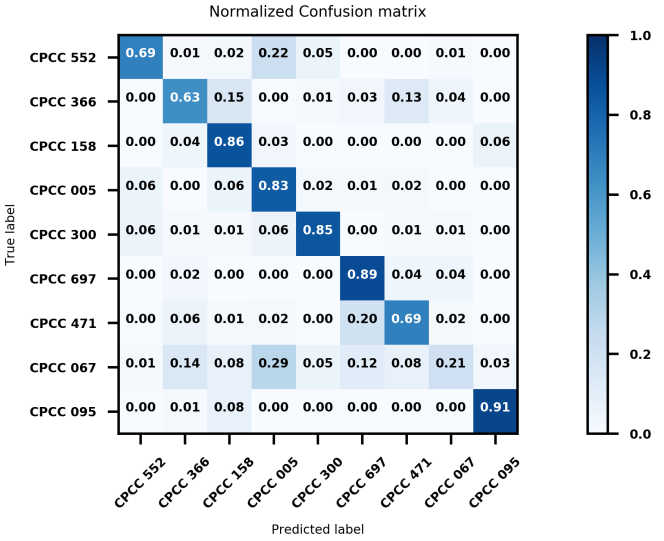
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.3 Model 3

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

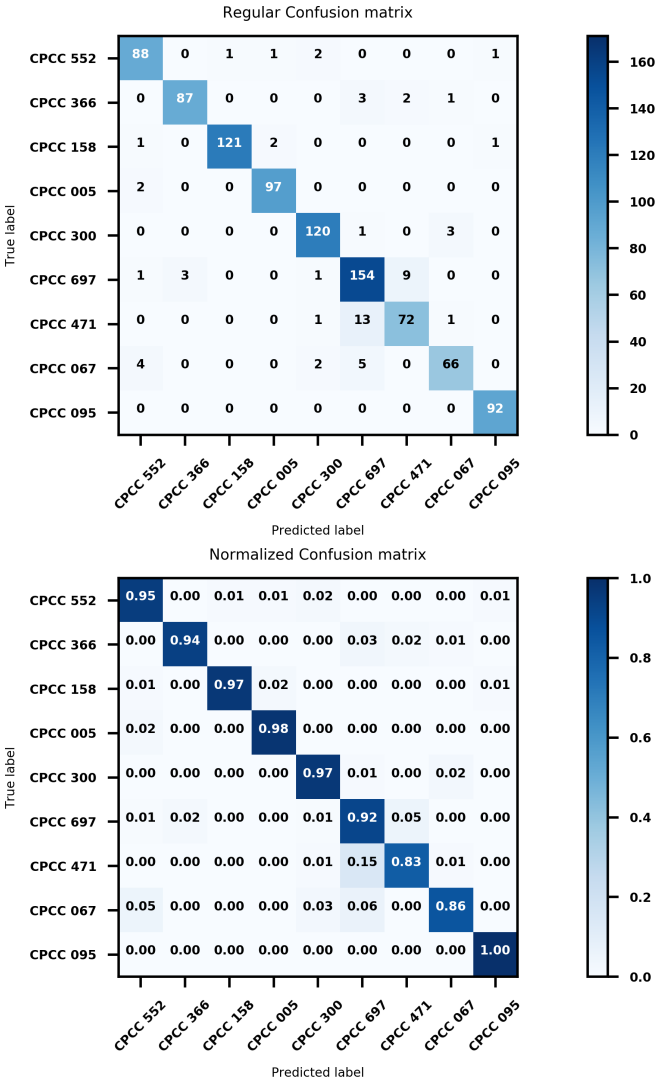
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)



A.4 Model 4

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

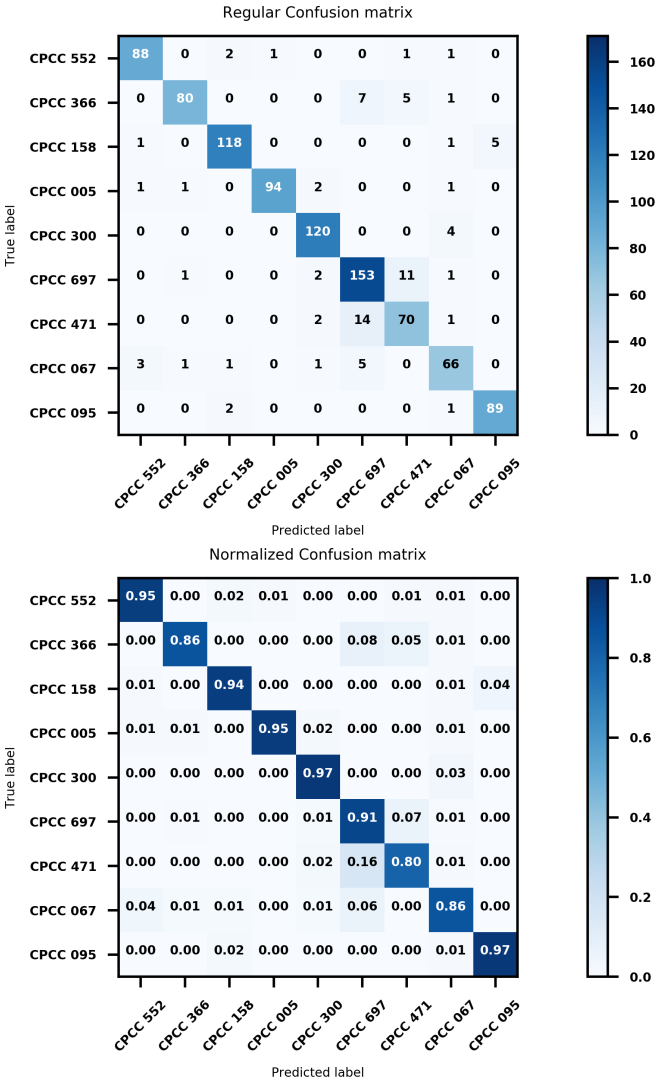
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)



A.5 Model 5

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

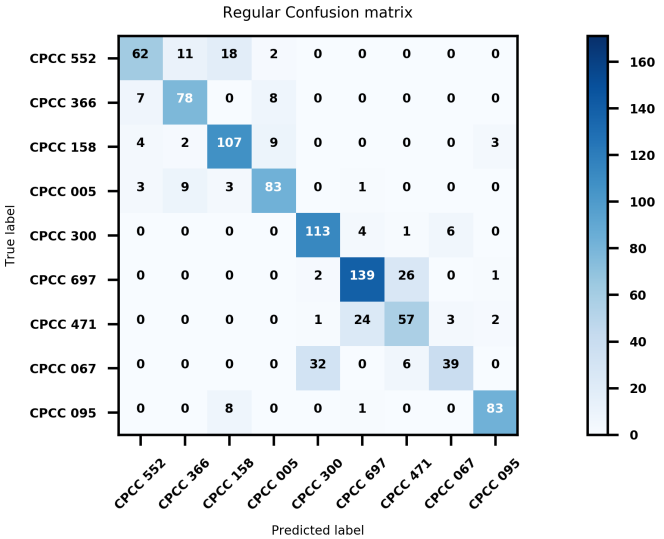
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

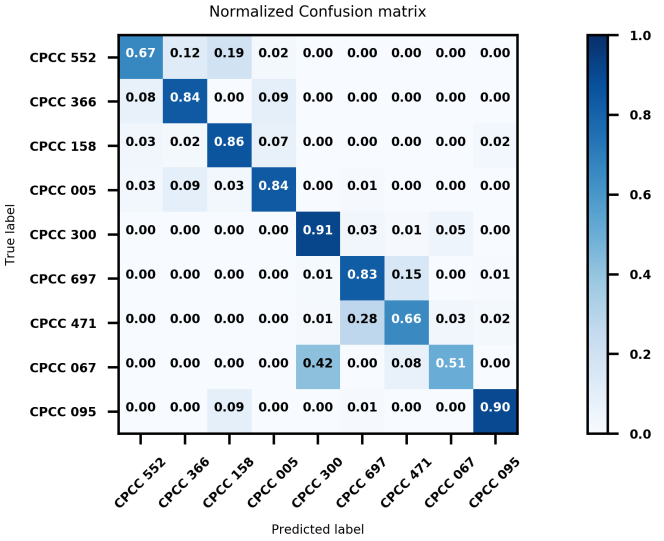
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.6 Model 6

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

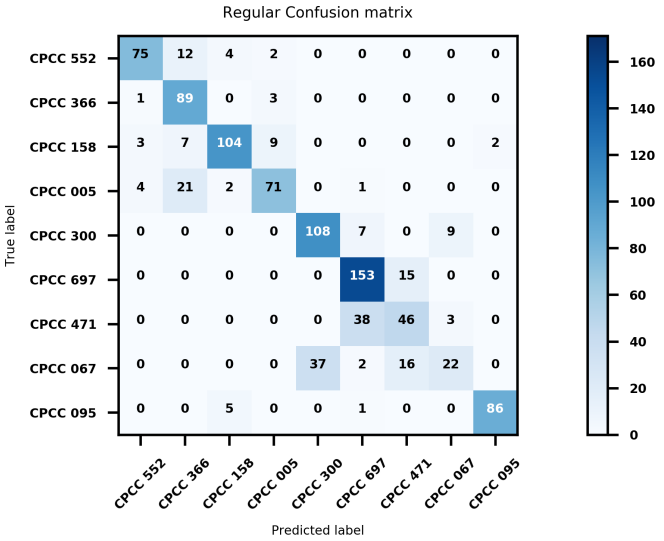
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

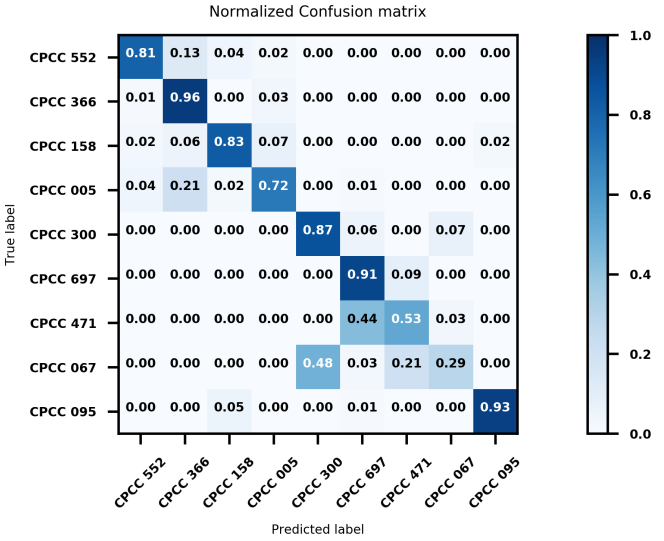
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.7 Model 7

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

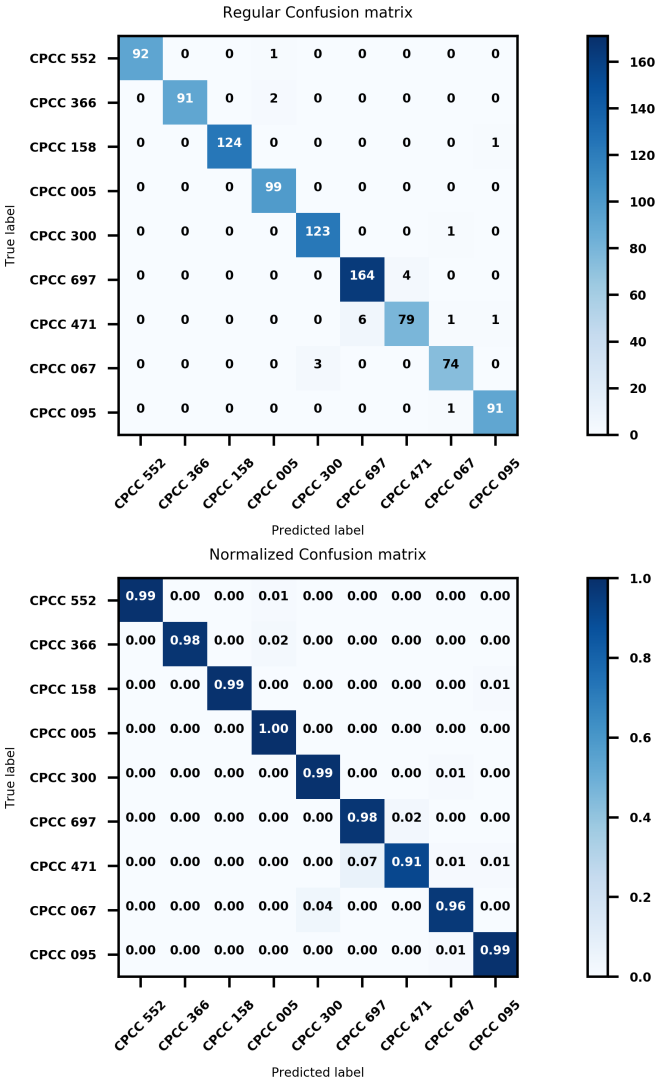
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)



A.8 Model 8

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

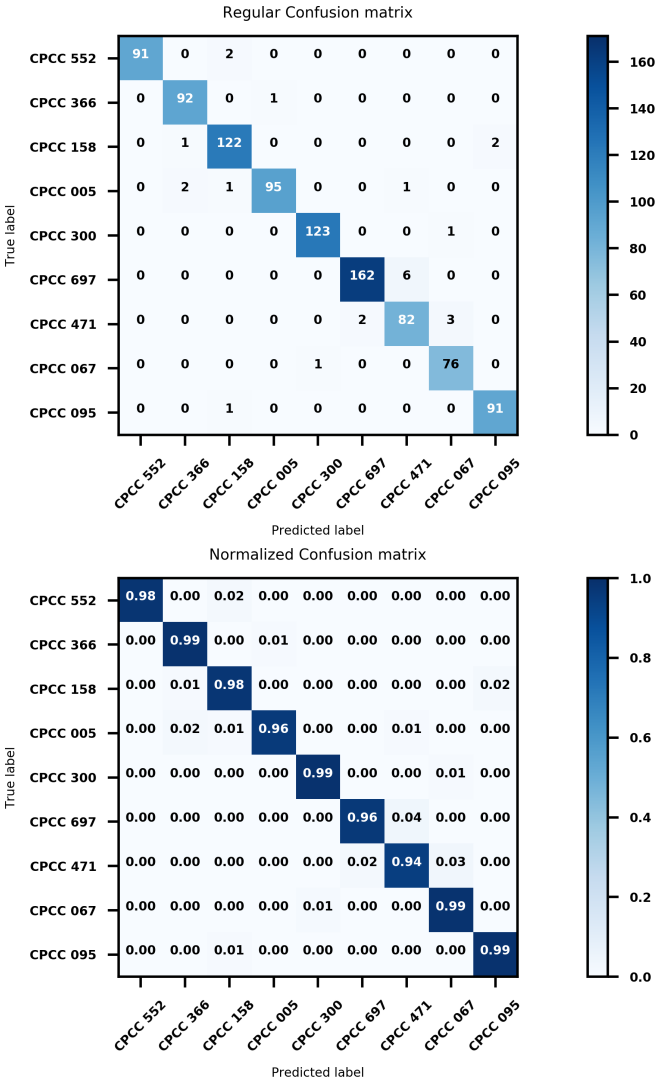
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)



A.9 Model 9

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

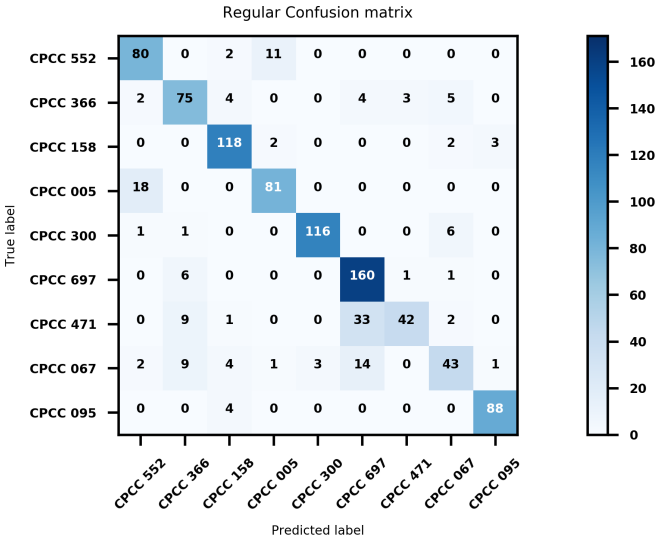
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

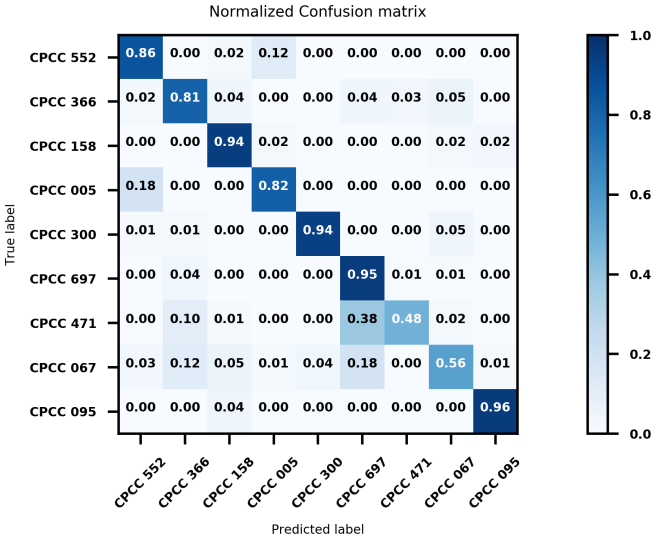
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.10 Model 10

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

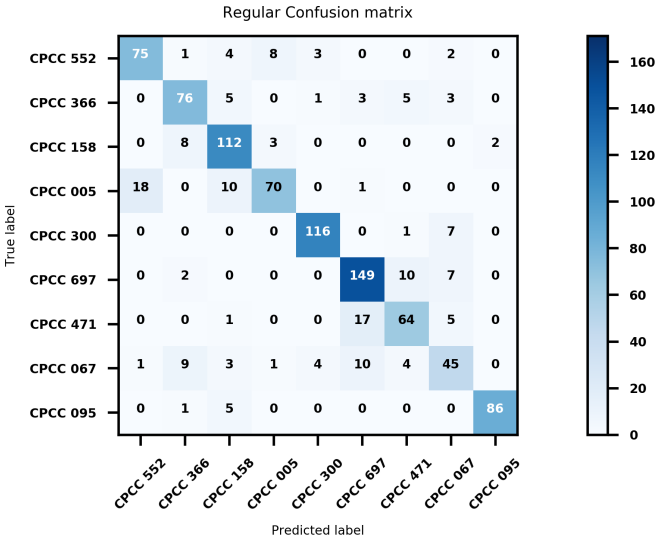
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

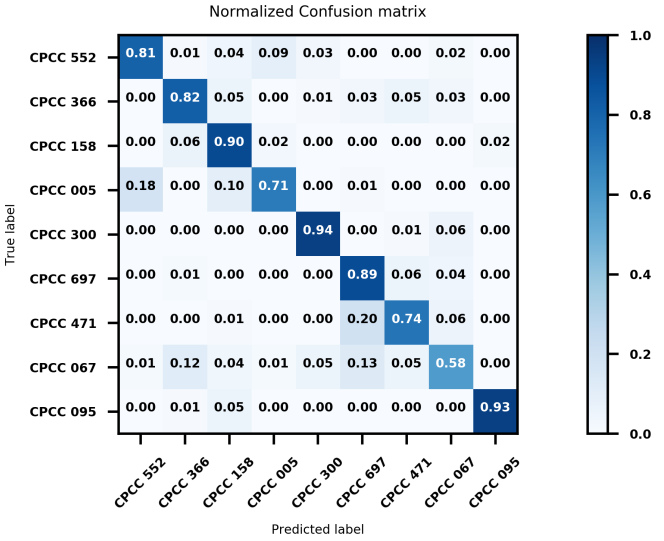
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.11 Model 11

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

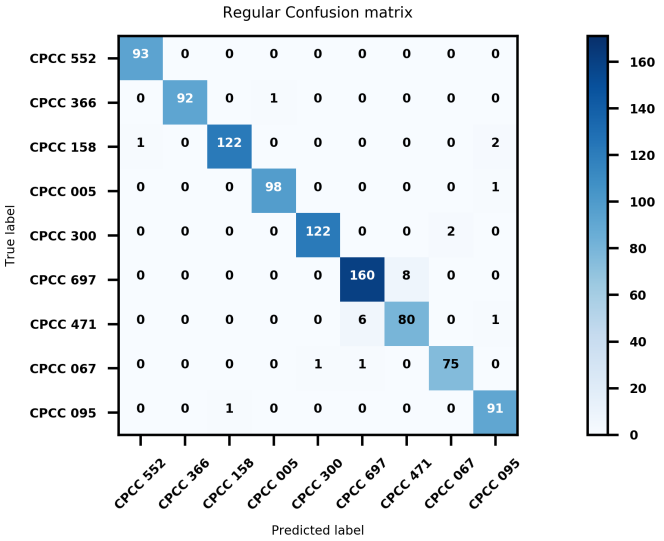
2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

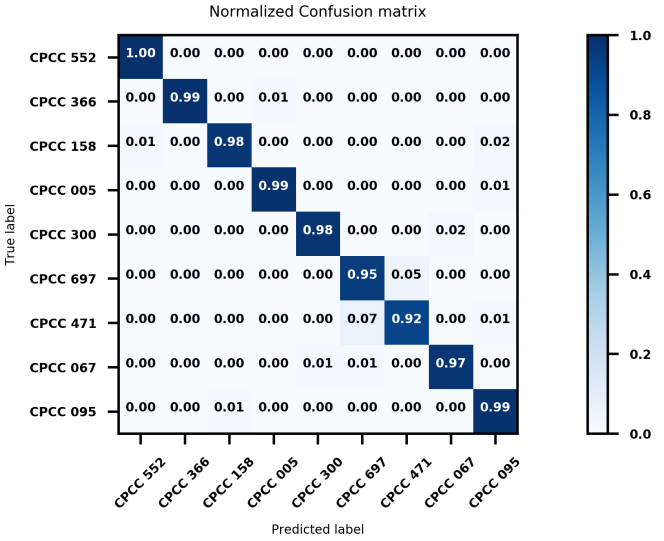
III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)





A.12 Model 12

data modality	data representation	hierarchical structure	model #
Brightfield	Feature Extraction	no	1
		yes	2
	Feature Learning	no	3
		yes	4
Fluorescence	Feature Extraction	no	5
		yes	6
	Feature Learning	no	7
		yes	8
Brightfield + Fluorescence	Feature Extraction	no	9
		yes	10
	Feature Learning	no	11
		yes	12

I. **Bacillariophyta**

1. *Fistulifera pelliculosa*;
   AKA *Navicula pelliculosa* (CPCC 552)

II. **Chlorophyta (green algae)**

2. *Ankistrodesmus falcatus* (CPCC 366)
3. *Scenedesmus quadricauda* (CPCC 158)
4. *Tetradesmus obliquus*;
   AKA *Scenedesmus obliquus* (CPCC 005)

III. **Cyanophyta (blue-green algae or cyanobacteria)**

5. *Microcystis aeruginosa* (CPCC 300)
6. *Pseudanabaena rutilus-viridis* (CPCC 697)
7. *Stenomitos tremulus*;
   formerly *Pseudanabaena tremula* (CPCC 471)
8. *Trichormus variabilis*; AKA *Anabaena variabilis*;
   formerly *Anabaena flos-aquae* (CPCC 067)

IV. **Euglenozoa**

9. *Euglena gracilis* (CPCC 095)