**IST 512: Thesis Proposal   
The use of artificial neural networks for phytoplankton species identification**

**Section 1: Introduction**

    There are many pressing issues facing our world including fossil fuels and energy generation, climate change, population growth and preserving important ecosystems. This research will be utilizing an interdisciplinary focus combining Biology, Computer Science and Statistics to help expand our understanding in these fields. Biology, like many other fields of science, has already been transformed by the rapid advancement of computers; from the sequencing of the human genome to artificial intelligence and neural networks these advances have happened fairly quickly given the span of overall scientific progress.

The goals of this research is to utilize principles of machine learning to enable better identification of phytoplankton species. It will involve some new research in developing the algorithms for the artificial neural networks and open up ways to better understand large scale climate changes through more accurate and quicker sampling methodology.

**Hypothesis:** Utilizing artificial neural networks can accurately and quickly identify phytoplankton species during flow cytometric analysis

Hopefully, it will expand our capabilities in a small area of both machine learning and biological oceanography. It also will open up doors for future research I am interested involving the use of artificial intelligence and machine learning to process and handle much larger data sets from current and past research in a way that a team of researchers would not be able to. If these methods worked it could open the way to develop and test previously taxing research questions or even find correlations that were unnoticed by human scientists.

**Section 2: Literature Review**

**Section 2.1: Ocean Ecology and Phytoplankton**

    Phytoplankton are foundational to ocean ecosystems and also for life on earth. They are responsible for nearly fifty percent of all global carbon fixation; a process converting inorganic carbon to organic carbon (Vaulot, 2008). They also are massive contributors to other biogeochemical cycles which are important to the sustaining of life on earth (Toseland**,** 2013). Phytoplankton are autotrophs utilizing PAR irradiance in the euphotic zone where they function as a primary producer in the pelagic food web. Besides their important conversation of carbon they are often grazed upon by larger organisms like microzooplankton (Miller, 2012).

Phytoplankton are comprised of a number of groups, classes and a large number of species. In spite of the large number of species it was determined that they generally follow the Redfield Ratio which is that their elemental stoichiometry makeup consists of fairly similar ratio of carbon, nitrogen and phosphorous. They also tend to have a number of trace elements which is important to researching changes they are undergoing. Because of the importance of phytoplankton to ecosystems our understanding of them, the effects they have on the ecosystem and how climate change is affecting the environment they inhabit is incredibly important to understand (Quigg, 2003).

One example of this change is the role that climate induced temperature fluctuations has on their metabolism and elemental stichometry. Research has found that at higher temperatures eukaryotic phytoplankton appear to require a lower density of ribosomes to produce a required amount of cellular protein. This could lead to a higher demand for nitrogen which could be problematic in areas with shifts towards N-limitation. This research shows temperature may play a previously unrecognized role that could have implications for larger biogeochemical cycles they are so integral to (Toseland**,** 2013).

Our understanding and ability to classify and look at large scale data sets of phytoplankton is thus important to looking at large scale climate changes and shifts in the environment. Changes and shifts occurring over time could indicate changes to the oceans redux state, temperature fluctuations, nutrient depletion or many other important factors.

**Section 2.2: Machine Learning and Neural Networks**

To look at and classify this scientific data we hope to utilize principles of artificial intelligence and machine learning. Machine learning is primarily about the ability to access data and utilize this to discover patterns and learn for themselves. It is useful in research like this where there is a large data set with small fluctuations. Artificial neural networks are non-parametric and computationally efficient ways of machine learning and were designed to mimic neural networks found in the brain (Beck, 2018). The two main types of machine learning are supervised and unsupervised learning. With supervised training data patterns with known identity are presented to the algorithms in order to “train” it. With unsupervised learning patterns are presented to the network and it derives its own groupings of the data. In our research we will utilize supervised learning.

Biologists and ecologists in both pelagic and tertial ecosystems have already been utilizing computer science in almost every aspect of their research. This research will hopefully build upon this in a small but useful way.

**Section 3: Methodology  
Section 3.1: Phytoplankton Cultures**

Three sets of phytoplankton data will be collected that contains 50 species of phytoplankton per data set. The large range of data will cover a range of size and morphological shapes. Currently, the research will plan for over 80% of species being common to all three sets and the remaining potentially being an outlier. The location of where the phytoplankton are harvested is important but has not been determined. Ideally work with Biological Oceanography faculty will allow for sample collection during a research cruise the Summer of 2020.

These batch cultures will most likely be analyzed utilizing 300 mL conical flasks or polycarbonate bottles. If culturing is utilized they will be sub-cultured approximately every 4 days to maintain exponential growth rates (Guillard, 1962).

**Section 3.2: Machine Learning and Computational Methodology**

Algorithms will be constructed utilizing Python and a number of open source machine learning python modules. The processing will take place on an AWS Linux server for a number of reasons.

* While early on processing power and speed will not be a factor in the future with large data sets we may wish to greatly increase our computational power and this can be done quickly utilizing AWS
* These Linux servers will also allow for the development of a RESTful API in order to allow HTTP requests and respond with standardized JSON output

One of the reasons Python was chosen is that it is one of the leading languages that both machine learning and artificial intelligence algorithms are being developed. It is fairly user friendly and also provides a range of statistical packages that will be used for statistical analysis.

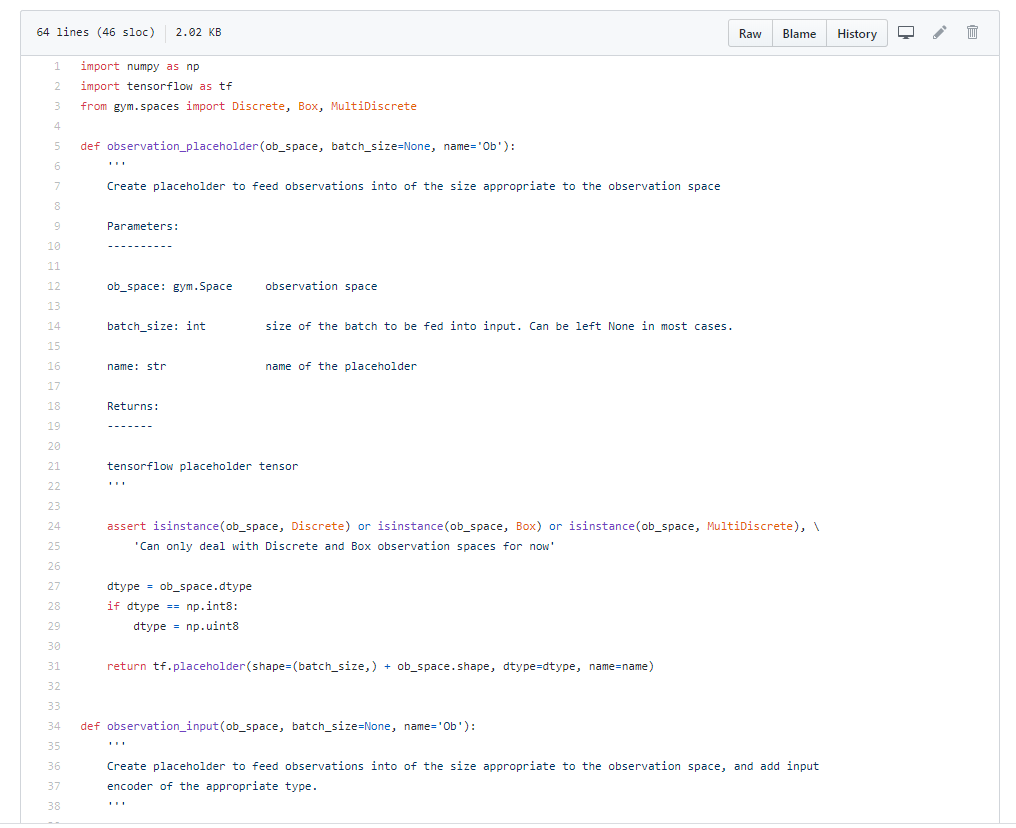


Figure 3.1: Machine Learning Code utilizing Tensor Flow

**Section 3.3: Flow Cytometric Analysis**

Flow cytometry analysis will be done utilizing the Flow Cytometry and Cell Sorting (FACS) on the Oregon State University Campus. The analysis will be done utilizing either the CytoFLEX S flow cytometer or the MoFlo XDP high speed cell sorter. Most likely the MoFlow XDP will be used since it can sort up to 4 populations simultaneously. Depending on how the research is developed the samples will either be sorted into tubes or if needed into culture plates.

**Section 3.4: Processing Cytometric Data**

Before analysis of the data it will be gated to insure that clusters of events caused by noise particles will be removed. These could consist of inorganic particles, bacteria, contaminants or cellular debris. Currently, this will most likely be done by omitting events that have low red fluorescence signals since these should not contain photosynthetic pigments.

**Section 4: Analysis**

**Section 4.1: Early Analysis**

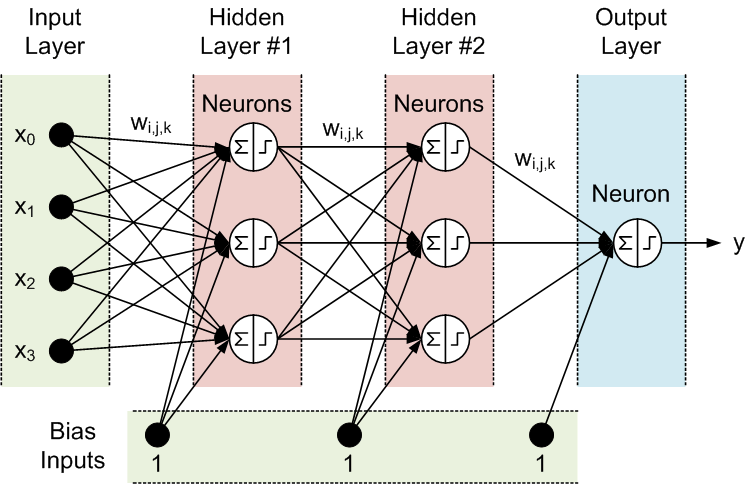


Figure 4.1: Artificial Neural Network Logic

**Section 4.2: Final Analysis**

Early analysis will look at a simple neural network with four nodes; the input layer, hidden layers and output layer. Every hidden layer will represent a separate basis function. The shape of the basis function will be determined by a Euclidian distance and the more computationally intensive Mahalanobis distance. These results will be analyzed with Complexity Theory and Computability Theory to look at tradeoffs between running time, algorithm complexity of the system and accuracy of species identification on a limited scale (Sipser, 2019).

**Section 4.3: Creating a Dendogram**

Ideally the data will be placed into a Dendogram similar to Figure XXX. This will be useful for a number of reasons. The first is that it will be an easy to use tool for researchers to look at and analyze the success of the algorithms. Checking the data by hand will also be easier with it classified in an easy to understand, scientific way. The next is that this type of data is easy to store and sort in a number of different types of data structures including arrays, specially designed classes or a relational or NoSQL database. Access to this data could then be readily available to other scientists through a RESTful API allowing for standardized JSON output.

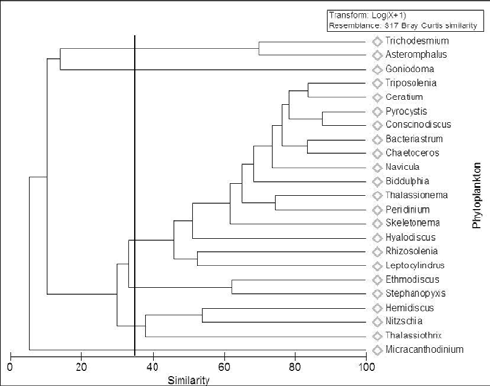


Figure 4.2: Taxonomic Dendrogram of Phytoplankton

**Section 4.4: Statistical Analysis**

Statistical analysis will also be run on this data to determine how accurate the machine learning algorithms worked. This can be then compared to data analyzed by hand and through genomic analysis. Work will be done with other faculty to determine how accurate the outcomes should be for future research questions.

**Section 5: Timeline:**

A tentative timeline has been put together for this research. One potential hurdle is that I am still taking the graduate level pre-requisites and will not be able to get to the graduate Machine Learning course tell late in year two. This problem could pose a challenge but will still provide a solid foundation for future research once the needed curriculum and thesis has been completed.

Table 5.1: Proposed Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Year 1** | | | |
| **Fall 2020** | **Winter 2020** | **Spring 2020** | **Summer 2020** |
| Coursework | Coursework | Coursework | Begin AFC processing |
| Early Research Preparation | Develop Thesis Proposal | Work with Phytoplankton Samples | Begin construction of simple shape recognition algorithms |
|  | Begin to Gather Phytoplankton Samples | Literature Review | Literature Review |
| **Year 2** | | | |
| **Fall 2020** | **Winter 2021** | **Spring 2021** | **Summer 2021** |
| Coursework | Coursework | Coursework | Write and Defend Thesis |
| Develop Algorithms | Develop Algorithms | Write Thesis |
| Conduct Research | Conduct Research | Statistical Analysis of Data |

**Section 6: Conclusions**

Ideally this research will provide a novel way for quick, large scale species identification. While the initial research is aimed at phytoplankton it could be potentially used for classification of other species, bacteria, viruses and organisms. Another outcome of the research is to hopefully enable future research that utilizes advanced methods in computer science and artificial intelligence to look at and combine data collected from years of previous research. This data could be made so that it is easily accessible through a standardized API and algorithms could be developed to answer fascinating, previously untouchable questions from this massive data set.

**Citations:**

A. Toseland, S. J. Daines, J. R. Clark, A. Kirkham, J. Strauss, C. Uhlig, T. M. Lenton, K.

Valentin, G. A. Pearson, V. Moulton, and T. Mock. "The Impact of Temperature on Marine Phytoplankton Resource Allocation and Metabolism." Nature Climate Change 3.11 (2013): 979-984. Web.

Beck, Marcus W. “NeuralNetTools: *Visualization and Analysis Tools for Neural Networks.”*

*Journal of statistical software* vol. 85,11 (2018): 1-20. doi:10.18637

Behrenfeld MJ. 2010. *Abandoning Sverdrup's Critical Depth Hypothesis on phytoplankton*

*blooms*. Ecology 91:977–89

Miller, C. B., & Wheeler, P. (2012). *Biological Oceanography*. Hoboken, NJ: John Wiley and

Sons.

Pershing AJ, Mills KE, Record NR, Stamieszkin K, Wurtzell KV, Byron CJ, Fitzpatrick D,

Golet WJ, Koob E. 2015 Evaluating trophic cascades as drivers of regime shifts in

different ocean ecosystems. *Phil. Trans. R. Soc.* B 370: 20130265.

Quigg, A., Finkel, Z., Irwin, A. et al. *The evolutionary inheritance of elemental stoichiometry in*

*marine phytoplankton*. Nature 425, 291–294 (2003)

Sipser, M. 2019. *Introduction to the Theory of Computation* (3rd. ed.). International Thomson

Publishing.

T. Platt, S. Sathyendranath *Ecological indicators for the pelagic zone of the ocean from*

*remote sensing***.** Remote Sensing of Environment, 112 (8) (2008), pp. 3426-3436

Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. *The diversity of small eukaryotic*

*phytoplankton in marine ecosystems.* FEMS Microbiol. Rev. 32, 795820 (2008).