Chapter 1: Introduction

1.1 Introduction

Occupying a wide range of terrestrial and aquatic environments all over the world, cyanobacteria or blue-green algae constitute one of the largest groups of photosynthetic prokaryotes. Cyanobacteria, representing some of the most ancient life forms on earth (Schopf et al., 1965), are one of the most studied organisms worldwide. Margulis (1975) characterized them as unusual prokaryotic microorganisms that can perform oxygenic photosynthesis. They can also synthesize chlorophyll a. Similar to eukaryotic algae and plants, cyanobacteria use H_2O as an electron donor for the production of oxygen. The oldest fossil findings of cyanobacteria can be dated back to approximately 3500 million years ago. On the last decades cyanobacteria blooms had given special consideration over worldwide because of it is highly effects on water environments by increasing anthropogenic input of nitrogen and phosphorus (Paerl et al., 2001; Khan and Ansari, 2005). To date, cyanobacteria are still largely present in oceans, freshwaters, and soils. Most cyanobacteria live in water as phytoplankton. Cyanobacteria are often called “blue-green algae” and are named after the blue pigment phycocyanin, which, together with chlorophyll a and other pigments, is used to capture light for photosynthesis. Cyanobacteria is important factor used for life evaluation such as phototrophic organisms, they can use the energy of light to produce organic matter (CH_2O) and oxygen (O_2) out of water (H_2O) and carbon dioxide (CO_2) through photosynthesis: $H_2O + CO_2 = CH_2O + O_2$. Cyanobacteria were the first organisms capable of oxygenic photosynthesis; thus, they are considered largely responsible for the rise in atmospheric O_2 over two billion years ago (Van den Hoek et al., 1995). The rise in atmospheric oxygen made the evolution of life that is dependent on oxygenic respiration possible. Furthermore, Raven and Allen (2003) compared ribosomal RNA from cyanobacteria with DNA inside the chloroplasts of eukaryotes and revealed that all photosynthetic eukaryotes derived their photosynthetic capabilities from cyanobacteria through

endosymbiosis. Water quality in freshwater lakes and reservoirs is mostly conducted on a regular basis; it can be determined by measuring the existence organism in water which provides the necessary information about long term of water quality. Moreover, some organisms are producing toxins which also effect water quality. Cyanobacteria used sometimes as indicators of water quality because they are light sensitive and can regulate their buoyancy; it is essential property to detect it over other organisms. Cyanobacteria are considered one of the essential factors in studying water pollution. Thus, the cyanobacteria genus *Oscillatoria* was selected for the current study conducted in the Putrajaya Lake.

1.2 Problem of the Study

Few or limited studies are available on the various aspects of tropical freshwater algae. In addition, few automated systems can recognize and identify tropical freshwater algae, especially in Malaysia. Expertise in both the computational and physiological field is lacking. Consequently, manual procedures for algae identification are time consuming because different parameters need to be considered, and every characteristic must be confirmed to conduct a predefined classification of algae.

1.3 Research objectives

The main objective of this research is to classify cyanobacteria found in tropical freshwater in the Putrajaya Lake by using image processing and an artificial neural network (ANN) approaches. Specific research objectives are listed below:

1. To capture high-quality images of freshwater cyanobacteria;

2. To analyze images through image processing techniques, and isolate each objects separately.

3. To automate the process of algae image classification for the identification of cyanobacteria (*Oscillatoria* sp.).

1.4 Scope of the Study

The main aim of this project is to use a supervised ANN to classify the images of algae found in Putrajaya Lake. Water samples were collected from different sampling sites. Images of cyanobacteria were acquired using plankton nets, and the water samples were analyzed and examined using an electronic microscope (MTC#B1-220ASA). A microscope eyepiece camera (AM432X) was attached to the microscope lens and connected to a computer via a USB port for image acquisition. It was used to capture, load, and store images directly into the computer. The images were processed using MATLAB software and classified using a neural network. ANN was used as a platform to automate a neural network to recognize the selected algae. The neural model selected for this attempt is the multilayer perceptron (MLP) and the generalized feed-forward neural network.

1.5 Constraints and Limitation

Lack of knowledge and experience in using an ANN is a huge problem in this project. Moreover, images captured digitally contain some problems such as the build-up of noise and signal distortion during the capturing process. Nevertheless, these obstacles can be overcome by referring to websites and journals. The lack of knowledge in the field of algae also poses a problem when identifying the algae to train the network. In addition, images are usually defined over two dimensions and perhaps more, thus can be used as digital image processing to correct the captured images. High-quality images need to undergo training and testing through an unsupervised ANN. Larger sets of data would produce better results once they are trained and tested by the neural network. Misidentification of the algae in the training set will affect the outcome of the trained

network by decreasing the general accuracy of the network, possibly rendering it insufficient for real commercial use. Only one computer can be used at a time to operate the NeuroSolution software, causing occasional inconvenience. This research is limited to algae found in tropical freshwater in the Putrajaya Lake in Malaysia. The weather affects the collection of algal samples because rain causes water levels to rise and stir up sediments.

1.6 Outline of the Study

Chapter One: This chapter contains the General Research Framework, which consists of the introduction, problem, objectives, scope, constraints and limitation, and outline of the study.

Chapter Two: This chapter contains the Literature Review, which includes an overview about algae, automated recognition system for algae, identification and classification of algae, introduction to cyanobacteria, discussion of cyanobacteria blooms (occurrence, non-toxicity, undesirability, toxicity in lakes, courses of action), discussion of *Oscillatoria* sp., and advantages of MATLAB, Image Processing Toolbox, ANNs, and MLP.

Chapter Three: This chapter contains the Materials and Methods, which consists of an overview about the Putrajaya Lake and the tools (plankton net, bottles, dropper, glass slides and cover slips, light microscope, and DinoLite-DinoCapture 2.0 camera) used for sampling and analyses, including image processing and ANNs.

Chapter Four: This chapter presents and discusses the Results.

Chapter Five: This chapter contains the Conclusion and future research.

.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview about Algae

Algae are a wide range of plants heterogeneous in all shapes, sizes, and physiological functions, except that they contain chlorophyll pigments and others.

Algae include members of the nucleus of a primitive prokaryote such as blue-green algae. Some algae are microscopic, whereas others reach up to several meters in length. Algae are abundant in salt water and fresh and stagnant pools of water, lakes, and in humid places or on the rocks. Large numbers of algae are also found on soil. The presence of algae in surface water has been a long-standing issue all over the world because of their adverse effects on the treatment process and quality of drinking water. The removal and control of algae in the water treatment industry are important global issues, especially in tropical and semi-tropical zones.

The presence of algae in surface water causes many problems regarding color, odor, taste, and toxic compounds (Gao et al., 2009), posing potential hazards to human and animal health. Furthermore, algae widely affect the drinking water which mostly required a treatment process (Gilbert, 1996). Traditionally, some treatment process involved such as pre-oxidation by chlorine dioxide, ozone, chlorine, or permanganate is usually used to improve algae removal in the coagulation process (Plummer and Edzwald, 2001; Chen and Yeh, 2009; Henderson et al., 2008).

Research found that algae are important factors in most of aquatic ecosystems; they reported that algae are classified as important components of biological monitoring programs. Algae have very short life cycles with rapid reproduction rates which make them an ideal component for water quality assessment of short terms effects. Algal showed wider distribution among geographical regions and ecosystems. Algal divisions are very sensitive to some poisons, and they accumulate pollutants readily. Algae as

primary producers are most organisms that directly affected by physical and chemical factors because their metabolism is also sensitive to natural disturbances and differences in environmental. In laboratory, standard methods exist for evaluation of functional and non-taxonomic structural characteristics of algae communities showed that algae are easily cultured, easy and inexpensive in sampling process, and have a minimal effect on local biota (Van Dam et al., 1994; Stevenson and Pan, 1999; Stevenson and Lowe, 1986; Rott, 1991; Round, 1991; McCormick and Cairns, 1994). Furthermore, as biological indicators algae showed many attributes of spatial and temporal environmental changes, especially as parameter for algae community in structural and functional variables which used as biological monitoring programs. Recently Malaysia governments had been included algae as biological indicator for water quality assessments in some area. Using algae as parameters in identifying the different types of water degradation have been becomes essential and complementary with other environmental indicators over worlds.

Actually, more than 10,000 living diatom species are become well known, and also same number approximately of named fossil forms, about over than 90% of the biosphere is employed by plant life, where diatoms make up approximately a quarter by weight. They are hugely abundant in the upper layers of oceans, where they providing a high-grade of nutrition to a variety of creatures from protozoans to baleen whales.

2.2 Automated recognition system for algae

Many attempts have been made to automate the identification system for organisms. However, such an attempt has not been done for algal species, especially in Malaysia. Furthermore, image analysis of phycological images has been conducted for cyanobacterial classification. Work has been conducted on many cyanobacterial genera, such as *Microcystis*, *Anabaena*, *Planktothrix*, *Aphanocapsa*, *Aphanizomenon*,

Coelosphaerium, *Gloeotrichia*, *Merismopedia*, *Nostoc*, and *Oscillatoria*. A localization method based on image-fusion techniques, high-resolution microscope images, color interest point extraction, self-organizing map network, and aid of textural features was among the methods used to analyze algal images. Most processes collect various algal samples to be observed and automated. Success rate is at most up to 90%. This project attempts to determine the methods that can automate the image recognition process to more than 90% for future use.

The aim in pattern recognition is to use a set of example solutions to solve problems to infer an underlying regularity and to serve as reference for possible solutions in subsequent cases. Among the many models proposed over the years, neural network models have exhibited the most optimal nonlinear boundary for classification problems. The capabilities of nonlinear learning in ANN is supported a powerful tool to be used widely for solving many complex applications such as nonlinear system identification, unsupervised classification and optimization, and functional approximation. The multidisciplinary nature of the neural network classification research is believed to generate more research activities and produce more fruitful outcomes in the future.

2.3 Identification and classification of algae

Image recognition and identification of fresh water algae has been manually executed, and very few automated systems have been implemented. Manual identification and classification are tedious and time consuming. A classification problem for algae exists because they needs to be categorized into predefined groups or division based on a number of observed attributes related to the object shape and measurements. Traditional statistical classification procedures are usually built on Bayesian decision theory, which details the concept of probability by reasoning with uncertain statements. In specific procedures, the probability model is assumed to calculate the posterior probability to

guide the classification process. However, one major problem of probability models is effective only if the underlying assumptions are satisfied. The efficiency of this technique is highly demands on the level of conditions and assumptions made during development process of this model. That is, one must be knowledgeable and experienced in the data properties and model capabilities to successfully apply or create the model (Zhang G.P, 2000).

2.4 Algae taxonomy and classification methods

The classification of algae into interested groups or division was performed on based of the same rules which used for land plants the classification however organization of groups of algae above the order level has been changed tis methods substantially since 1960. Research begins to use the differences of algae feature in classification process such as cell division, organelle structure and function, and flagellar apparatus. Microscopic images showed a similarities and differences among algal, fungal, and protozoan groups which led scientists to propose new major taxonomic changes that are currently under process. Division-level in classification algae is used as well as kingdom-level classification but itis tenuous for algae because classes are mostly distinguished by the structure of flagellate cells such as (scales, microtubular roots, angle of flagellar insertion, and striated roots), the cytoplasmic division process (cytokinesis), the cell covering, and the nuclear division process (mitosis).

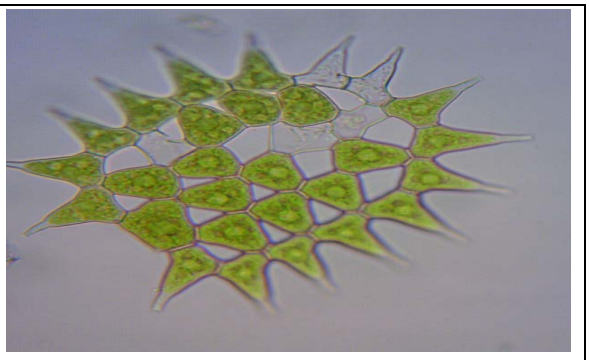
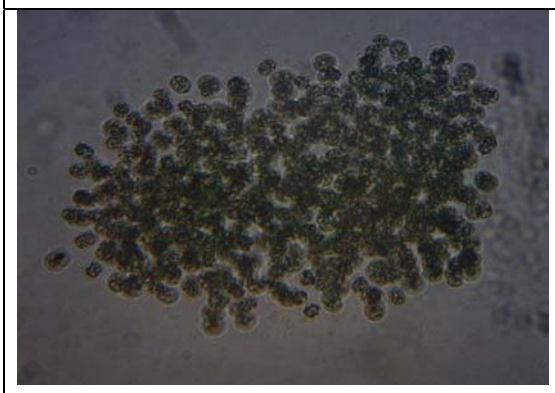
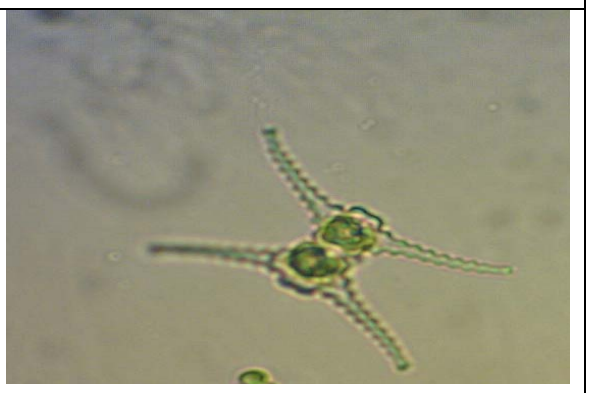
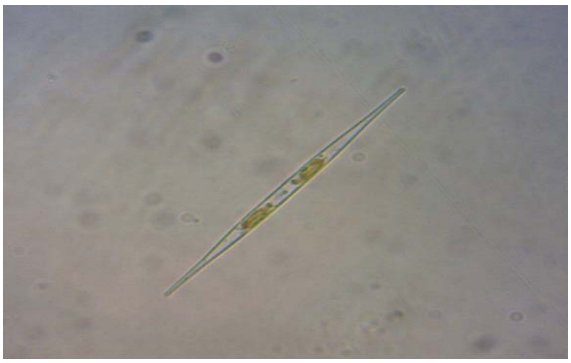
	
Closterium.sp	**Pediastrum duplex**
	
Microcystis aeruginosa	**Staurastrum paradoxum**
	
Anabaena circinalis	**Nietzsche ps.**

Figure 1: Example of images algae collected

Table1: classification and compare some kinds of algae, which collected by images.

Kingdom:	**Plantae**	**Protista**	**Plantae**	**Bacteria**	**Bacteria**
Phylum:	Charophyta	Chlorophyta	Charophyta	Cyanobacteria	Cyanobacteria
Class:	Zygnemophycea	Chlorophyceae	Zygnemophyceae	Cyanophyceae	Cyanophyceae
Order:	Desmidiales	Chlorococcales	Desmidiales	Nostocales	Chroococcales
Family:	Closteriaceae	Hydrodictyaceae	Desmidiaceae	Nostocaceae	Microcystaceae
Genus:	Closterium	Pediastrum	Staurastrum	Anabaena	Microcystis
Species:	Closterium.sp	duplex	paradoxum	circinalis	Microcystis aeruginosa

2.5 Cyanobacteria

Cyanobacteria (blue-green algae) are common members of freshwater algae, the plankton of marine, and brackish throughout the world. It mostly found on rocks, soils, and in symbioses with plants and fungi. Cyanobacteria have a simple structure at the subcellular level and lack a nucleus, also as prokaryotes a characteristic feature defining them along with bacteria (Fogg et al., 1973).

Research found that cyanobacteria have a photosynthetic appearance that supports them to perform photosynthesis as in algae and higher plants, but they lack choloroplasts in which these reactions occur in the latter organisms. Unicellular and filamentous forms are commonly found among cyanobacteria. Both morphotypes are capable of producing structures visible to the naked eye, such as pinhead or larger, spherical, or irregular colonies (e.g., of *Microcystis*), and bundles of filaments (e.g., of *Aphanizomenon*) like sawdust in shape and size. Further differentiation among cyanobacteria includes the ability of certain filamentous genera, such as *Anabaena*, *Aphanizomenon*, *Gloeotrichia*,

Nostoc, and *Nodularia* to fix atmospheric nitrogen enzymically in specialized cells termed heterocysts. Several of the filamentous genera also produce other differentiated cells termed akinetes (spore stages), which permit them to survive periods of adverse conditions such as cold and drought. In addition, research found that cyanobacteria produce high mass populations in natural and controlled water resources, which lead to the common cyanobacteria blooms, scums, and mats, but not invariable consequences of eutrophication with rich nutrients for waters and plant (FWR, 2000). Cyanobacteria large growths and accumulations are often considered undesirable because of it is effect on water color, quality, and odors, it also cause some turbidity in recreational and amenity facilities, and have potentially synthesize over many low molecular weight compounds that cause taste and odor problems. These substances often result in complaints regarding recreational and amenity water bodies as well as the quality of raw and treated drinking water. Low molecular weight compounds such as cyanobacteria toxins or cyanotoxins produced by cyanobacteria are highly concern because of their high toxicity to vertebrates including mammals. These compounds are ignored during associated the problems with taste and odor compounds because they are colorless and odorless (Codd G.A, 1995; Sivonen and Jones, 1999). Cyanobacteria have higher toxicity if compared with other biological toxins such as plant, fungal, and shellfish toxins. Saxitoxins produced by cyanobacteria are considered as chemical weapons by the international Chemical Weapons Convention (Organization for the Prohibition of Chemical Weapons, 2000). Also, Saxitoxin and microcystin are listed in the core list of toxins preventes to export and control by the Australia Group (The Scientific Response to Terrorism, 2003). Among the cyanobacterial genera that include toxin-forming species, the ones of particular concern when mass populations occur include *Microcystis*, *Anabaena*, *Planktothrix* (formerly known as *Oscillatoria*), *Aphanizomenon*, *Cylindrospermopsis*, *Phormidium*, *Nostoc*, *Anabaenopsis*, and

Nodularia). Together, these genera can produce a wide range of cyanobacterial toxins of varying structures, toxicities, and modes of action.

Cyanobacteria are prokaryotes have a simple cell structure with a real nucleus Prokaryotic. Their body is made of Cyanobacteria from a single cell, often clustered cells as colonies of different shapes. The optimum temperatures for the growth of blue-green algae, ranging from 35-40 C. Cyanobacteria are typically much larger than bacteria in size, it contain many types of pigments such as carotenoids and phycocyanin. A characteristic of water soluble pigment in cyanobacteria gives the group of cyanobacteria their blue green coloration. In addition, cyanobacteria living in individuals places in fresh and salt water, and some other types live in moist soil. The water distinctive bluish color is results for cyanobacteria blooms when it dies (Crayton M.G., 1993).

2.6 Cyanobacterial Blooms

The blue-green algae include many groups, widespread, and lives in fresh and salt water and most of them live submerged in fresh water. Most marine species living on the beach, whereas some species live in soil and on the rocks, and when there are favorable conditions can be seen and clear the murky waters became Growth with a very green, blue, greenish, or reddish brown, within a few days only. Many types of organization of buoyancy float have the surface to form a thin scum of green and blue colors. Also, maintain these cyanobacteria population does not abnormally consider as high for a long time, and will die quickly within one to two weeks. If conditions remain favorable, we can replace the last bloom quickly than its predecessor. Interfere successive blooms are making the flowers look like continuing growth for several months (Crayton MG, 1993). Phenomenon of fast growth for cyanobacteria called bloom which effects water color to make lake water appears like pea soup. Cyanobacteria are classified as algae

commonly and called blue-green algae which a distinct group of bacteria capable of photosynthesis. The sunlight and nutrients are converted by cyanobacteria into energy for growth and reproduction.

2.7 Occurrence of Cyanobacteria Blooms

Cyanobacteria have different types of adaptations which allow them for persistence, optimal growth, and support them by the ability to outcompete algae during good conditions. In general, many species produce hidden cells or components that remain inactive until the condition is become suitable then they arise again. Some spices have specific cells which able to convert nitrogen gas into nitrogen fixation forms, dissimilar to algae which produce photosynthesis from carbon dioxide gas. Most cyanobacteria types can able to utilize different carbon sources.

2.8 Toxicity of Cyanobacteria Blooms

It is not truth that all cyanobacteria blooms are toxic, even some blooms caused by identified toxin producers aren't produce toxins or may produce toxins at low levels. When bloom occurs in specific area, warning should be given to the peoples using their water source, even if the toxins detected in relatively small amounts. It is still ambiguous until now what is the triggers that cause to produce toxin by cyanobacteria, at least scientist have not been reported the main reasons about that. Recently, researchers found that only about 10% of all blooms types are considered toxins producer. Recent studies in developed countries such as USA, and CANADA reported that probability of producing toxins for individual bloom including *Anabaena*, *Microcystis*, and *Aphanizomenon* between 45% to 75% which is greater than previous expectation.

Many evident showed that cyanobacteria bloom is toxic and can cause dead to large number of animals within or around water such as fish, waterfowl, and terrestrial animals. Researcher found many symptoms from sub-lethal poisonings with different kind of animals, quantity of toxin consumed, and nature of toxin itself. Mostly, any unexpected animal illness, and unexplained death existing near water, it should be suspected that water containing toxic bloom (Crayton M.G., 1993).

2.9 Toxicity of Cyanobacteria in Lakes

Since more than 100 years ago, numerous accidents existing which cases of animal poisonings over worldwide. Some reports showed that also human illness and death have been documented in many countries lastly. Numerous bloom forming common species reported they produce potent toxins, and cyanobacteria is become known well for long time it produce toxic. Research showed that cyanobacteria produce different types of toxins such as different type of nerve toxins, common liver toxins called microcystins, and shellfish poison named saxitoxin which considered the less common types of toxins.

2.10 Treatment of Cyanobacteria Blooms

To treatments cyanobacteria in water chemical compound must be used. Using chemical in natural lakes is note allowed because it can cause other different problems. Thus, using chemicals can effect organisms and human too because it is also toxic to other form of life. In near future, research may be found suitable solution by reducing the nutrients amount lakes. User activities and entertaining of lakes help to reduce the nutrient amount input into lakes through individual and group action.

2.11 *Oscillatoria* sp.

Oscillatoria sp is one of the simplest filamentous blue-green algae, where each thread is surrounded by gelatinous scabbard, and do not contain different vesicles, and one of the initial signs resulting from the presence of moderate pollution where it appears in large numbers, leading to change the color of the water. *Oscillatoria* uses photosynthesis to survive and reproduce. Each filament of *Oscillatoria* consists of trichomes made up of rows of cells (Delpeuch F et al., 1975). Filaments commonly exhibit oscillating, sporadic flexing, or gliding movements under the microscope, especially near the anterior ends. The filaments are composed of disk-shaped cells that are wider than they are long. Their end cells are usually rounded or tapered. *Oscillatoria,* with an average size of 7 µm, is an organism that reproduces by fragmentation and forms long filaments of cells that can break into fragments called hormogonia, which can grow into new and longer filaments. Breaks in the filament usually occur where dead cells (necridia) are present. *Oscillatoria* reproduces through fragmentation. This genus is categorized under the domain of bacteria because it does not possess membrane-bound organelles, a nucleus, or an endoplasmic reticulum. In addition, metabolically active bacteria include those that are heterotrophic, chemoautotrophic, and photosynthetic. *Oscillatoria* is a photosynthetic bacteria grouped under the kingdom of cyanobacteria. Some *Oscillatoria* species, such as *Oscillatoria plantesis*, are consumed for their nutritional contribution (Delpeuch F et al., 1975) and sold as food, whereas some are useful for nitrogen fixation. Some *Oscillatoria* species are known to produce hepatotoxins, which damage the liver, and neurotoxins, which affect the nerve cells. However, no other useful function has been found for *Oscillatoria curviceps.*

They are considered the primarily affect of liver in animals by producing the *hepatotoxins* toxic which lead to disconnect liver cells with additional blood accumulation within liver that causing death via *hypovolumic* shock.

Recently, experimental results found many evidences that prove at least one of molecular mechanisms action consistent certain known carcinogens, which led researchers to suspect that these toxins are liver poisons, which could prove significant to humans after continuous low-level exposure (Crayton M.G., 1993).

Given the advantages and disadvantages to the environment and to living organisms, algae would be easier to manage if identification and classification can be done in a more automated and convenient manner.

Table 2: Classification of *Oscillatoria*

Kingdom:	Bacteria
Phylum:	**Cyanobacteria**
Class:	**Cyanophyceae**
Order:	**Oscillatoriales**
Family:	**Oscillatoriaceae**
Genus:	**Oscillatoria**
Type Species:	**Oscillatoria SP**

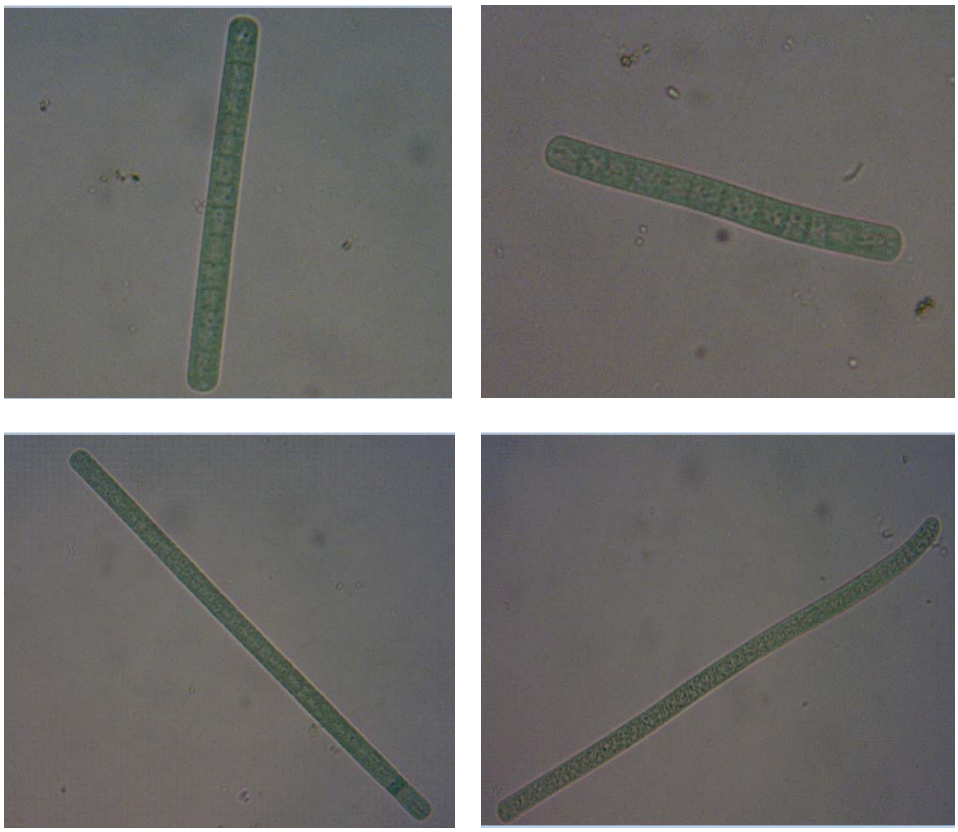


Figure 2: Example of images Oscillatoria collected

2.12 MATLAB Software

MATLAB is considering as an object oriented graphical language, MATLAB has an interactive environment for designing interface and development algorithm to solve complex engineering problems such as data analysis, visualization, and numerical computation. Research reported that using MATLAB to solve the technical computing problems is faster than using the traditional programing language such as FORTRAN, and C. Recently MATLAB included a variety of built in function and procedure that can be used in for different applications, including control design, computational biology , measurements and evaluation, signal and image processing, communications, and financial modeling and analysis. The add-on toolboxes which are a collection of scientific purpose functions was improve MATLAB environments for solving technical problem in wide variety application tasks. Recently MATLAB have been enhanced with

several advantages such as features of documents and sharing projects and research, and it is ability to integrate the code with other programming language which facilitates the distribution process for MATLAB algorithms and applications. In the following table a list of most of includes toolboxes available in the last version of MATLAB:

Table 3: MATLAB toolboxes

Communications	**image Processing**	**System Identification**
Control System	Instrument Control	Wavelet
Data Acquisition	Mapping	MATLAB Compiler
Database	Neural Network	MATLAB C/C++Graphics
Data feed	Optimization	MATLAB C/C++ Math Library
Filter Design	Differential Equation	MATLAB Report Generator
Financial	Robust Control	MATLAB Runtime Server
Frequency Domain	Signal Processing	MATLAB Web Server
Fuzzy Logic	Statistics	Simulink

2.13 Advantages of MATLAB

Using MATLAB can offers significant advantages over the existence traditional programming languages as listed below:

• Using MATLAB for a student or engineer can facilitate solving the diffuclts problems in less time and with less effort which considered the most powerful advantages of MATLAB.

• Researcher found that MATLAB is considered easier to use than other a high-level programming language.

• MATLAB also helps research to emphasis on problem solving directly rather than involving in understanding other language issues. It assists researcher on concentrating more about the problems that need to be solved.

• High level function and built in MATLAB tools have been decreased the runtime errors, and the mean time for error correction.

These advantages provide a glimpse of the power and flexibility of the MATLAB system (Tahir and Pareja, 2010).

2.14 MATLAB image Processing Toolbox

MATLAB is included a set of Image Processing Tools and function to provide engineers and scientists with a widespread suite of reliable functions for digital image processing analysis. Image processing tools integrated with MATLAB is developed by scientific researcher and professional technical to reduce the time consuming of coding and debugging for the essential image processing tasks. Using built in function of image processing tools is more reliable and faster than building application from scratch, it is illustrated a significant results in saving time and operation cost which enable users to spend less time in design and coding algorithms. MATLAB also help user by allowing them to learn and apply specialized technology in varying area of image processing approaches (Tahir and Pareja, 2010).

2.15 Artificial Neural Networks (ANNs)

ANN is abbreviations for artificial neural networks which is the scientific field that model information-processing structures as massively parallel structure similar to human brain. ANN is usually consisting of several nodes connected together through adjusted functions to construct the neural networks (Gaston and O'Neill, 2004). ANN is

constructed from three different layers often which are the input layer, hidden layer, and output layer. ANN is considering as a powerful data modeling tool that create a representation form for complex input/output relationships.

ANN has become increasingly significantly in solving complex problems which required special consideration of thinking and decision in many disciplines. ANN is designed to be used for solving particularly complex problems related to highly non-linear relationships. Particularly, ANN has been widely used as scientific tools in many disciplines to solve different types of problems, such as identification and control, forecasting, classification, and optimization because the heterogeneous systems are extremely difficult to model in mathematic models. Flexible structure of ANN is providing simple and reasonable solutions to various types of problems (Misra and Dehuri, 2007). In addition, ANNs were commonly employed in computerized pattern recognition tool for automated taxonomic identification based on several types of features such as morphological features, shape features, etc. ANN is simple has two phases training phase and decision phase, in training phase learning process of specific information is performed, and in decision phase the classification or recognition results is produced based on the learning phase. In training phase the internal structure of ANN is adjusted in response to a representative sample of data patterns for each taxon to be identified with related information identification by using the training data.

ANN used to generate special mapping form between input and output variables and produced a complex arbitrarily for nonlinear decision boundaries. Mostly neural network complexity increases with the number of nodes and layers of net. Neural networks usually use the sequential training process to observe data, and distinguish between patterns of interest that makes them an efficient pattern recognizer (Jayanta K.B. et al., 2010).

ANNs is developed to perform specific information computation which totally different from conventional methods. Neural networks construct from nodes with numerous simple elements that process individually several aspects of a large problem. A processing element (PE) is used to multiply the input elements by a set of weights, to transfer the results into nonlinear outputs values. Neural computation is represented a massive interconnection among PEs to share the load of overall processing task, and adaptive nature of the parameters (weights) with PEs. ANN usually has several different layers of PEs. Figure (3) below showed simple diagram for MLP neural networks. The circles are represents PEs which arranged in different layers. The input is the left node, hidden layer are the two columns in the middle, and outputs is represented by the right nodes in last column. The lines that connected between PEs is represented the weighted or sometimes known as a scaling factor.

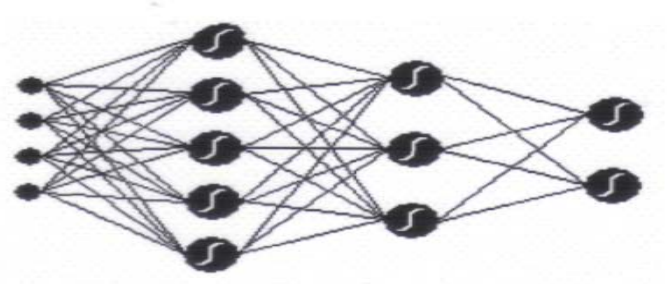


Figure 3: A simple multilayer perceptron

Neural network works to find optimal solution based on a measurement of its performance by adapting its weights during training process. ANN performance for a supervised learning is measured in terms of error criteria, a desired signal, and iteration cycles or epoch, whereas the performance for unsupervised ANN is measured based on learning law, and topology constraints.

In this research we employed an ANN to automate the recognition system of selected algae. in general, ANNs is a mathematical models attempts to mimic the functioning of human brain, which consists of billions of neurons and some algorithms assumed to encode knowledge neutrally (Noriega. L, 2005).

2.16 MLP Neural networks

MLP is one type of feed-forward ANN model which used to map the input sets of data into a set of appropriate output or desired results. MLP designed with multiple node layers in a directed graph where each layer fully connected to the next layer. Each node in MPL ANN is represented a neuron, or processing element with a nonlinear activation function that control the process of weight adjusted except input nodes. Layers of MPL ANN are important in solution, and optimization results, where one or two hidden layers are nearly sufficient to achieve classification tasks with a great accuracy. Increasing number of hidden layers is considered a disadvantageous because the training process will need extra times which can effects network performance. The input layer in MPL ANN are denoted the set of information or data need to be processed by multiplied it with the weight of hidden layers of network. This procedure is repeated many times until the desired results obtain which considered the learning process for the network (L. Noriega, 2005). A few learning algorithms are exist recently which available to obtain the correct values for the weights, the supervised and unsupervised learning paradigm is the common methods employed to adjust the networks weight values.

MLP is used a back propagation technique as supervised learning for training the ANN network. MLP is considered as one of the most neural network topologies used widely for general application purpose, for example MLP with two hidden layers can be used

as general purpose for statistic pattern classifier. In other words, discriminant functions of MPL are used to cluster a set of input data which can be taken any shape.

Furthermore, researcher found that MLP is usually achieves better performance in computing the maximum posteriori receiver which optimize the classification tasks especially when weights are normalized properly including output classes which usually normalize into 0 or 1 values. MLP has the ability also of approximating arbitrary functions, this capability is important on design nonlinear dynamics system. MLPs are trained normally by using the back propagation algorithm which depends on propagation the rules to optimize errors through the network by automated for weight of hidden PEs. MLP neural network mostly trains by using error correction of learning procedure, where perceptron elements of a given layer connected with all other elements of next layer, and the desired values must be initializing at early stages. Each input data sets need to be associated with desired outputs data set which considers the heart of pattern recognition process because the ANN is usually used to transfer the input data sets into appropriate values during training phase

MPL ANN is first start off by defining the initial estimates of the correct output, then the process of weight randomize for the input pattern presented. The main difference between the actual network output and desired output is evaluated during training process. The average squared error between actual output and target value is minimized as much as possible by network. This minimization process of networks is used a gradient descent in MLP ANN, which obtains the common back propagation algorithm for training neural networks. Gradient descent is used mostly to find the minimum values of a function by taking steps to the negative function of the gradient at current point. Back propagation method is a common method used for learning the MPL ANN. This technique of calculating the derivatives in a computationally efficient manner has a

graphical interpretation and operates by propagating error backwards from output nodes to inner nodes. Momentum learning is commonly applied in this process. This type of learning is used to prevent the network and the system from converging to a local minimum or saddle point. It also allows changes to persist for a number of adjustment cycles and takes into account the previous weight change.

Chapter Three: Materials and Methods

3.1 Study Site: Putrajaya Lake

The Putrajaya lake system occupies more than 600 hectares, including man-made wetlands and small-perched lakes located on the major incoming tributaries immediately above the main body of the lake. The lake system has a total upstream catchment of 51 square kilometers. The Putrajaya Lake occupying 390 hectares was created by inundating the valleys of Sungai Chuau and Sungai Bisa. It is characterized by a deep main basin at the south (approximately 9 meters to 13 meters deep) and a narrow arm approximately north of the dam (2 meters to 7 meters deep). The full supply level of the lake is EL 21.0 meters, corresponding to a water volume of 26.5 million cubic meters. The lake is in its second phase of construction. The first phase (Phase 1A) of 110 hectares has been completed and filled behind a temporary dam, and the second phase of the lake (Phase 1B) covers an additional area of approximately 310 hectares. The western arm of the primary lake between the northern weir and the dam in the south generally varies in width between 140 and 800 meters, whereas the narrower eastern arm of the lake following the eastern shoreline of the core island is approximately 100 meters in width. The distance between the northern most part was below the Central Wetland waterbody, and the dam wall in the far south is approximately 6 kilometers in a straight line or approximately 9 kilometers by way of a midstream navigation path. The total shoreline length of the main lake is 38 kilometers (Hijjas Kasturi Associates Sdn, 2001).

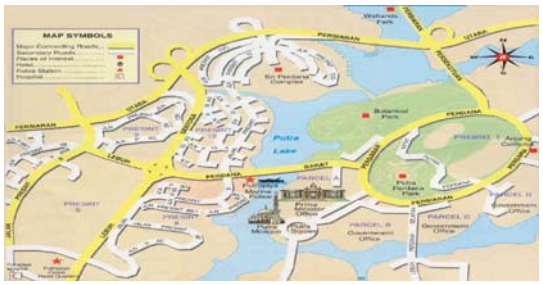


Figure 4: location of Putrajaya Lake

3.2 Tools used for Sample Collection

3.2.1 Plankton net

Plankton net is a shaft shaped with smooth mesh through the water as shows in (Figure 2). Phytoplankton were collected by throwing the plankton net into the lake and pulling it inwards continuously for approximately five minutes. The phytoplankton net was washed with lake water after withdrawing it from the lake to allow trapped algae to flow down. The samples were then collected and placed in bottles.

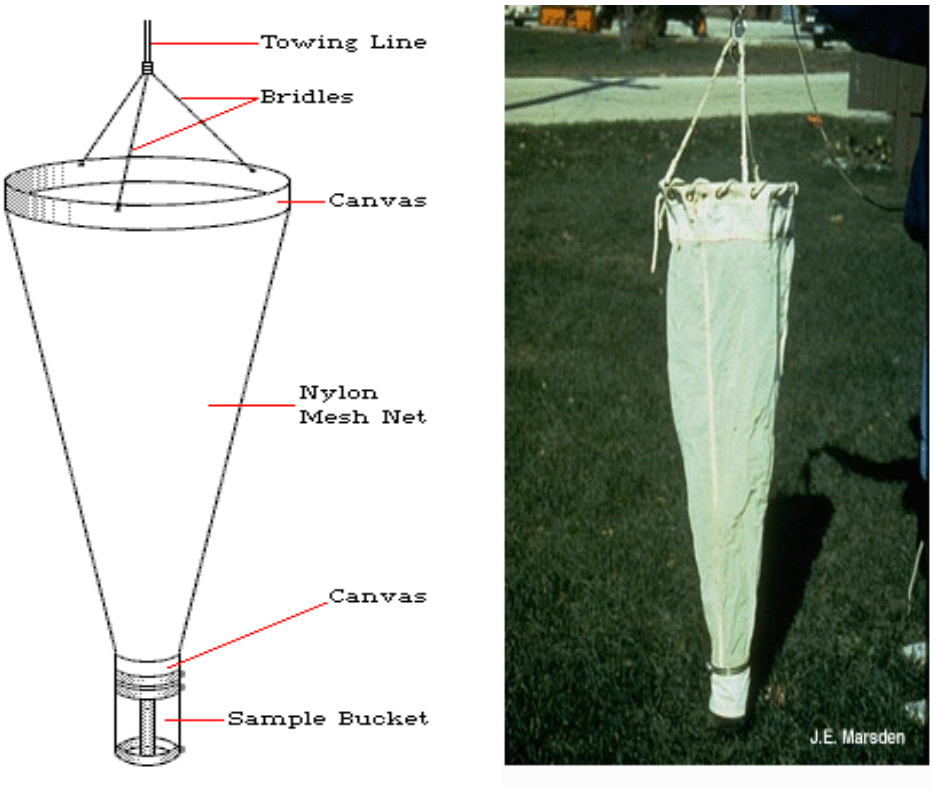


Figure 5: Plankton net

3.2.2 Bottles: Plastic bottles were used to store the water samples.

3.2.3 Dropper: A dropper was used to take water samples from the bottles and place them on glass slides.

Figure 7: Dropper

3.2.4 Glass slide and Cover slip: Water samples were placed dropwise on glass slides. The specimen was then covered with a cover slip as shown in figure (8).

Figure 8: -Glass slide and Coverslip

3.2.5 Light microscope: The water samples were analyzed and examined under a light microscope (MTC#B1-220ASA) as shown in figure (9) with a starting magnification of 10× and 40×.

Figure 9: Light microscope

3.2.6 DinoLite-DinoCapture 2.0 camera:

One of the lenses on the microscope was replaced with a DinoLite - DinoCapture 2.0 camera and connected with a PC via a USB port for images. The microscope was then used to capture images with DinoLite-DinoCapture 2.0. The images were stored into a computer as digital images in JPEG format.

Figure 10: DinoLite - DinoCapture 2.0 camera

3.3 Image Processing

The image was scanned for algae using the automatic detection algorithm. The genus *Oscillatoria* was selected from among several algae. Images were processed to remove noise and are cropped into a 300×300px template for further image processing. The 300 × 300px template was preferred after trials with ANN-NeuroSolutions. The 300 × 300px image size was verified as the optimum size for efficient processing during network building, which will be conducted later for image classification. The format of the images used for all the processes were in bitmap for the convenience of the subsequent methods and processes.

One of the lenses on the microscope has been replaced with a DinoLite - DinoCapture 2.0 camera. The microscope images are then captured using DinoLite - DinoCapture 2.0 and stored as digital images in JPEG format. However, only selected of the images were deemed ideal for image processing and image classification in the subsequent process. The number of images for other species was randomly selected and was less than the main selected species because of the comparison percentage suitable to train the network.

In this project, two selected species of algae were used to train the network because of their easy retrieval and availability. The number of images for each algae sample was taken into consideration to allow more variability as possible. Thus, the species selected for comparison and training of the network include *Oscillatoria*, and *Diatoma.*

Diatoma elongatum **Oscillatoria sp.**

Figure 11: Images of algae

Table 4: Species used for comparison and image classification

Kingdom	**Bacteria**	**Chromobiota**	**Chromalveolata**	**Chromobiota**
Phylum:	**Cyanobacteria**	**Bacillariophyta**	**Heterokontophyta**	**Bacillariophyta**
Class:	**Cyanophyceae**	**Coscinodiscophycidae**	**Bacillariophyceae**	**Fragilariophycidae**
Order:	**Oscillatoriales**	**Aulacoseirales**	**Bacillariales**	**Fragilariales**
Family:	**Oscillatoriaceae**	**Aulacoseiraceae**	**Bacillariaceae**	**Bacillariaceae**
Genus:	**Oscillatoria**	**Aulacoseira**	**Nitzschia**	**Diatoma**
Species:	**curviceps c.**	**Aulacoseira granulata**	**Diatom nitzchia palea**	**elongatum**

The images were consecutively processed using MATLAB R2011. The images were converted into grayscale after transferring and cropping them into 300 × 300px image size to allow better edge detection in the following operation.

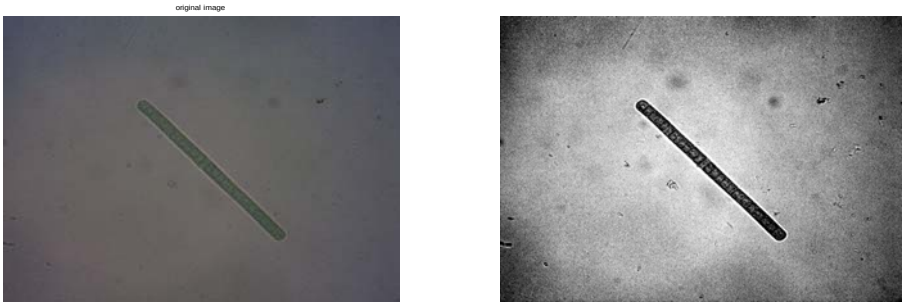


Figure 12: Conversion of original image into grayscale using MATLAB.

The next operation is edge detention by binary gradient mask or “Sobel” operator. Binary gradient mask was used to ensure maximum contrast of the boundary of the object of interest with the background. “Sobel” operator is commonly used in edge detection algorithms by computing an approximation of the image gradient intensity function. “Sobel” operator results in better edge detection because of its sensitivity to the image intensity gradient at each point, enabling it to identify part of the image that represents the edge. in this study we used specific sensitivity threshold called Sobel method. The syntax will ignore all edges that are not stronger than the threshold value stated. Nevertheless, the threshold value empty may also be left empty, allowing the program to select the value automatically.

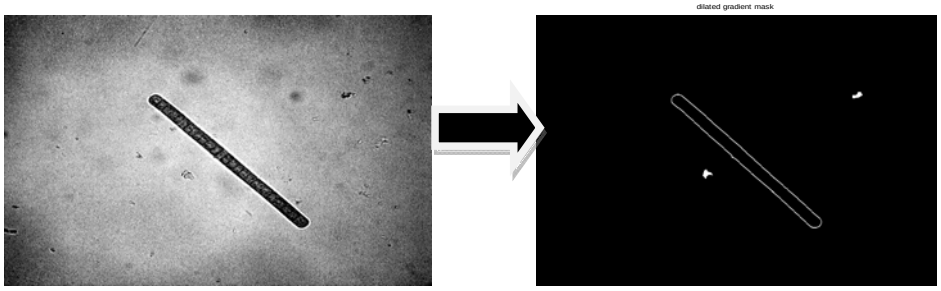


Figure 13: 'Sobel' operation on a grayscale image.

Imaging after the "Sobel" operation was performed to produce images with edge-detected object. However, gaps and holes still exist within the boundary of the object of interest. Hence, the dilation operator was subsequently performed on the current image. Dilation is done to enhance and enlarge the lines within the boundary of the object of interest, reducing holes and gaps. Binary dilation was performed on the images using the "imdilate" command. Images were still maintained in binary form.

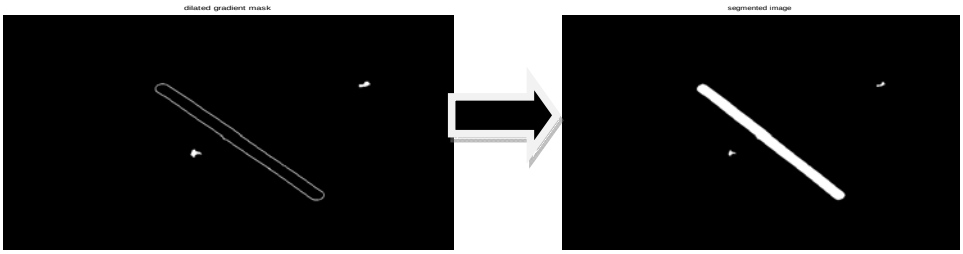


Figure 14: Conversion of dilated gradient mask image into segmented image

The images were then operated using the "fill" operator. Area is represented as a whole object illustrated by dark pixels and surrounded by lighter pixels. Hence, the "imfill" command will fully fill the remaining holes in the boundary of the object of interest. The detected holes will be converted into lighter pixels, in contrast to its original dark pixel value, making the pixel values of the holes equal to the area surrounding them.

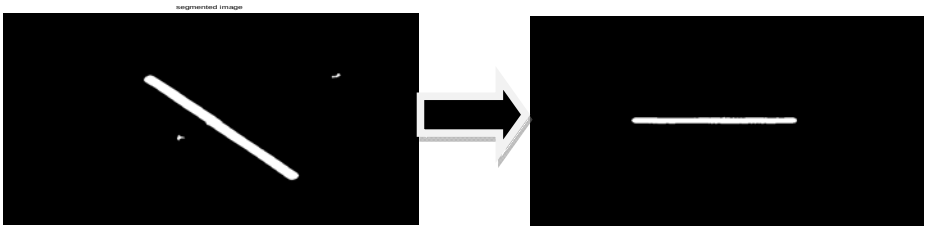
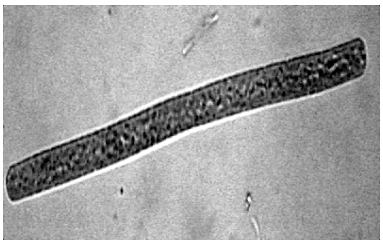


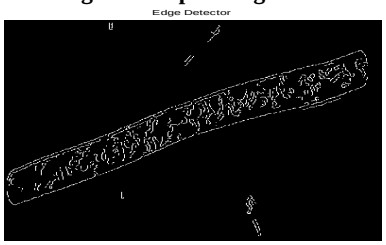
Figure 15: Image fill operation conducted on image after image segmented image.

O

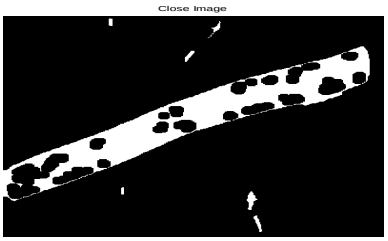
riginal sample image



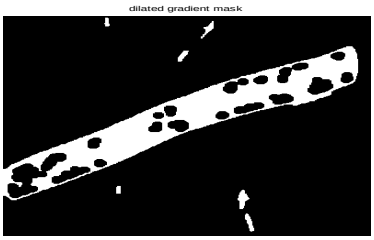
Grayscale image



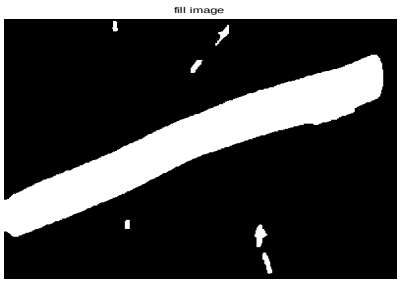
Edge detection and segmentation



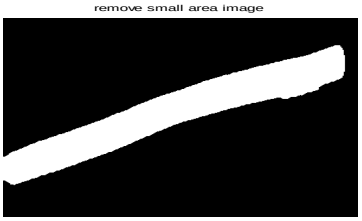
close image



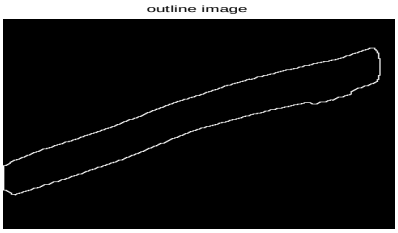
Dilated gradient mask



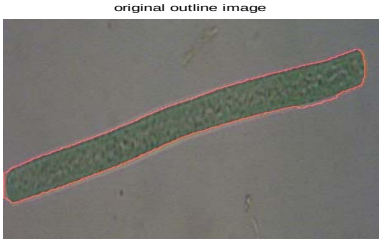
Fill image



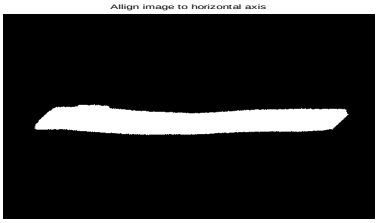
Remove small area image



outline image



Original outline image



Align image to horizontal axis

Figure 16: The image processing.

3.4 Method of system development

MATLAB 7.9 was used for the system development process because it can integrate technical computing environment suitable for algorithm design and development. In addition, MATLAB 7.9 is a high-level programming language that includes several functions supporting the image-processing field together with ANN. The system architecture is shown in Figure 17.

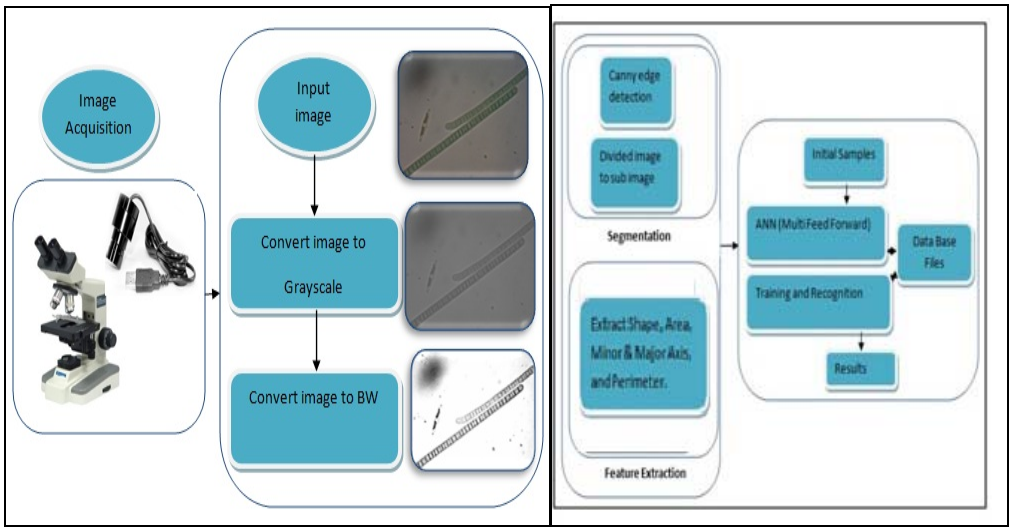


Figure 17: Developed System Architecture.

3.5 Image Preprocessing Model

Image preprocessing was used to prepare the microscope image to be utilized efficiently by the developed system. Most acquitted images from microscope or any other sources suffer from low contrast quality. They may contain some noise and unwanted areas or may be blurry. Preprocessing for captured images is a preparation and treatment process for image feature enhancement to produce clearer details, remove noise, remove intelligibility of images, and improve the overall appearance. The following is a simplified list of the basic steps used for preprocessing in our system:

(1) All image samples extracted from the water samples were captured using a microscope with an attached camera and then transferred into one folder into our computer.

(2) The acquired images were classified into different groups. Each single algal genus had it is own folder, which contains two categories: one for training images and the other for testing images.

(3) Preprocessing was performed to prepare the training set for each genus, whereas preprocessing was performed automatically for the testing set.

(4) The images were uploaded into the system using a GUI interface designed to facilitate user interaction.

(5) Contrast enhancement was performed to enhance the uploaded images, remove dark areas, enhance image brightness, and make the images clearer. Histogram equalization built in MATLAB functions was used to enhance the color and contrast intensities of the image. The images were transferred to a gray scale, and the image adjust functions were used to increase the image contrast.

(6) The median filter was used to reduce image noise and preserve edges. Some unwanted areas and small objects were removed when the median filter was applied.

(7) The images were converted from gray scale to binary images. Image complement was performed to produce the image background in black and image objects in white.

(8) Essential steps for binary images were performed, such as image border removal, filling of the boundary area, and exclusion of any small region < 50px. These three steps are important for the segmentation process.

3.6 Segmentation Model

Images of the selected algal genera rarely exist with only one object. They also contain other objects, such as other microorganisms. Image segmentation was used to split the input images into several sub-images based on the number of detected objects, where each sub-image contains only one object only. Particularly, it was used to identify the location of feasible objects and their boundaries. Therefore, image segmentation was used to consolidate the objects included inside the images to process every sub-image separately. Based on the previous section of image pre-processing, the input image was processed to become a binary image with some regions. A canny edge detector considered as the most powerful edge detector for image segmentation was used for the segmentation process (Canny, 1986). The region of binary image was detected using the canny edge approach, and each region was represented on a sub-image. Each sub-image was used as a mask to obtain the same region of the original image (color image). Both regions of color and binary images were associated with the corresponding index numbers, and the image resolution was resized to 300 × 300px image.

3.7 Feature Extraction Model

A feature extraction model was used to transform binary and color image features into a set of parameters used to describe the object features under consideration. The extraction techniques used for the object feature extraction were independent features, such as color, texture, and shape. In our system, the feature extraction method from both

binary and color image features mainly includes shape, area, minor and major axes, and perimeter. Different feature extraction techniques were applied in this study for both binary image and corresponding color images. These techniques are described in detail as follows:

1 Centroid: A 1-by-Q vector that specifies the mass center of the region. The first and second elements of Centroid are the horizontal (or x-coordinate) and vertical (or y-coordinate) coordinates of the mass center, respectively. All other elements of the Centroid are in order of dimension.

2 EquivDiameter: scalar number that specifies the size of diameter for a circle that cover the same area of objects as region. This property is computed as following equation:

sqrt(4*Area/pi), which supported only for 2-D input label matrices.

3 Extract Object Length and Width: Major and minor axes extracted using another routine that identifies two points automatically by calculating the maximum distance between given points in an object vector. The major axis represents the line segment connecting between the base points in the x-axis, and the minor axis represents the maximum width perpendicular to the major axis (Figure 18a). The object width factor was calculated by slicing across the major axis and parallel to the minor axis. Feature points were then normalized into a number of vertical strips. For each strip, the ratio of strip length to object width was calculated using the following equation:

$$Rc = Wc/L$$

where Rc is the ratio at column c, Wc is the object width at column c, and L represents the object length (Figure 18b). In Addition, other morphological features such as area and perimeter were extracted.

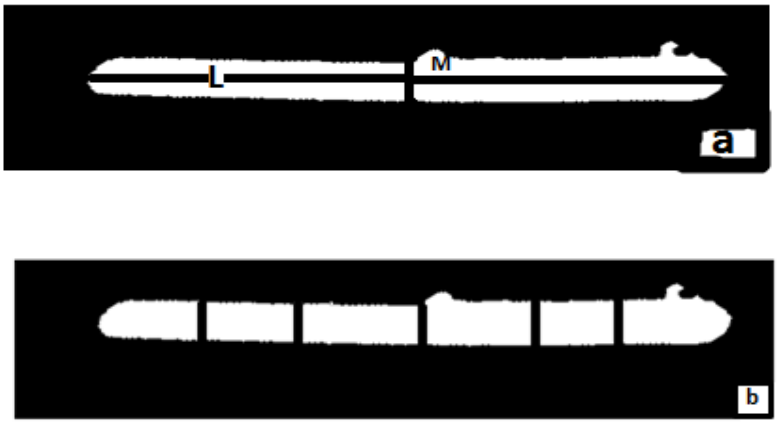


Figure 18: (a) Major Axis L, and Minor Axis m. (b) Object Width Factor strips

4 **Extract Area**: The area that represents the actual number of white pixels in the selected region. The object area was calculated using the number of white or "1" pixels (Figure 19a). The area was included in this study as a parameter for the classifying process.

5 **Extract Perimeter:** The perimeter of an object, shown in red pixel in Figure 19b, is the summation of the distance between each adjoining pair of pixels around the object border.

Figure 19: Examples for Extracted (a) Area, and. (b) Perimeter.

Table 5 : Extracted Feature used by classifier in this study.

Feature No.	Feature Descriptions
F1,F2,F3	Length, width, extent area
F4,F5	Area, Perimeter
F6,F7,F8	Equ. Diameter, per/area, per/length
F9-F10	Perm/width, width/length
F11-F12	Area/length, area/width

3.8 Classification and Identification Model

The classifying model was applied to categorize the detected objects based on the extracted feature vectors. Multilayer perception and feed-forward ANN were applied to perform the identification process for the selected blue-green algal genera. The vector feature dimension was equal to 12 features extracted from the input image. R represents the input vector which identified how many number of features is used, and Q is represents the total number of training phase between transition from inputs and outputs pairs. The ANN architecture consists of 12 inputs, 1 output, and 4 neurons in the hidden layer. The training image samples were used to extract vector features for each image and then store in comma-separated files. The element of vector features is shown in Table 1. For the training phase, a total of 50 images of oscillatoria were used to train the feed-forward neural networks.

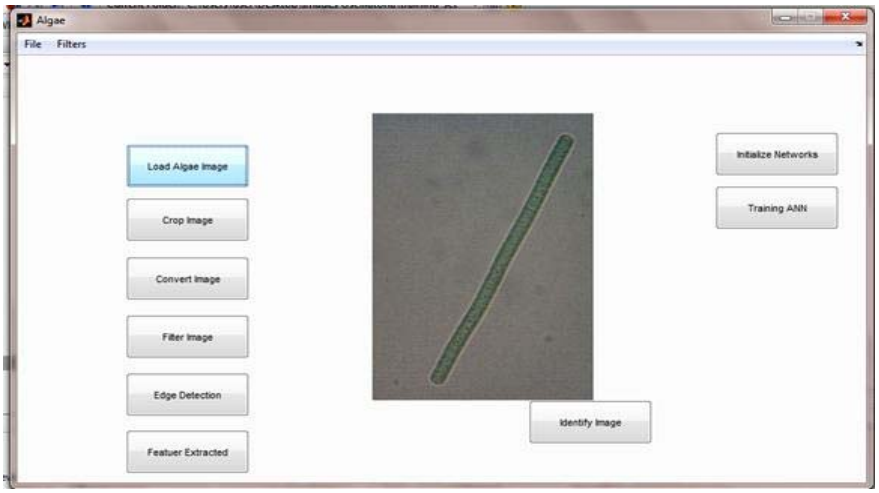


Figure 20: System Interface

3.9-Artificial Neural Networks

NeuroSolution 5.0 is used in this project for design the ANN layeres and peromed classification and testing data as shown in figure 21.

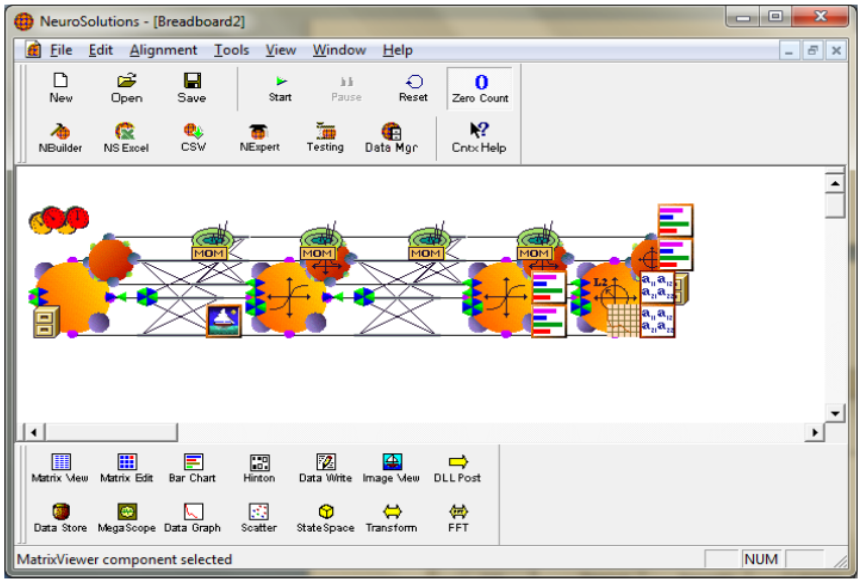


Figure 21: Screenshot of MLP ANN

The neural network is built by selecting the method, processing elements, learning rule, input and desired data as shown in table (6) below:

Table 6: Elements used to build the MLP ANN automated image recognition.

Method:	Multilayer Perceptron
Learning Parameter:	Momentum: 0.750 Supervised Learning} Step size: 1.000
Processing elements (PEs)/ artificial neurons:	2 in input layer 4 } 2 in output layer
Transfer function:	Tanh (Axon)
Hidden layer	1
Performance Measure:	Confusion Matrix
Training Epoch (cycle):	500
Testing Epoch:	1

We used default set testing for most of selected options during implemented the ANN. MLP topologies usually employ the function Tanh(Axon) transfer function in hidden and output layers for adjusting the neuron weight. We applies Tanh(Axon) function to each neuron in each layer to enforce the output result to be between 1 and -1 for each single neuron. The Neural Builder comes with a side display which gives a brief explanation on the options in the section. Most of the default settings are sufficient enough to carry out the training of the network and there seem to be no added improvement in altering the settings.

After completing the building of the network, the training, testing and cross validation set of images will be inserted into the neural model to be used as a desired output to be compared with the actual outcome of the neural model. This input process takes time as currently, there is no shortcut to select and input all the images at once. Thus, it has to be done one-by-one.

Training, cross validation and testing set of images will all be input in this file inspector.

Table7: Division of images used in the classification process.

Files	Percentage (%)	Number of objects
Train	50	Oscillatoria= 93 128} Others = 35
Test	30	Oscillatoria= 37 46 } Others = 9
Cross Validation	20	Oscillatoria= 26 31 } Others = 5
Total	100	205

Chapter four: Results and Discussions

4.1 Results

Developed system was tested for evaluation purpose. Then system performance was calculated by comparing the system outputs results for tested images with the manual identified results. The system interface allowed the user to load the store dataset files yield from training phase, and also allowed him to initialization his own training image samples. The automatic training procedure is taking approximately 5 minutes; all extracted vector features were included in training and testing mode of the feed forward ANN.

After training phase, the final classification system was tested by using image test samples; direct assessment was performed manually with automated classification system. Total objects founded inside sample images was 205 objects, unidentified object was 49 objects in manual process while the identified object was 156 objects. In addition, unidentified object was 54 in automatic process while the identified object was 148 objects as illustrated in more details in table 8.

Table 8: Comparison Results between Automatic and manual process

	Automatic	**Manual**
Total objects	194	205
Identified object(Oscillatoria)	148	156
Unidentified object	54	49

To evaluation system accuracy and performance another test has been performed for total of 50 test images. In this test, we calculate the time requires for training phase which was about 5 minutes, and the time required for identifying and classifying process for the input images which was about 1.5 minute. The results showed that our system identified 45 input images from the 50 input images, and average recognition

accuracy of our system is calculated which equals to 90% when included all extracted features as shown in confusion matrix below figure(22) and table (9). The overall system accuracy for our proposed system depends essentially on the ability of system to isolate and segmented image objects within input image, and also the accuracy of classification system to identify the detected object based on selected feature. Also, the training volume samples are playing an important role in system accuracy and performance results.

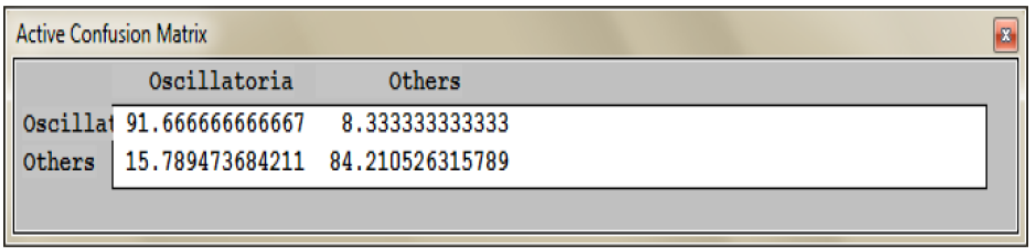


Figure 22: Confusion matrix of testing data from 'Test' file for GFNN.

Table 9 : Confusion matrix of testing data from 'Test' file for GFNN.

	Predicted outcome (%)	**Predicted outcome (%)**
Actual value (%)	91.67	8.33
Actual Value (%)	15.79	84.21

4.2 Discussion

This project specifically targets the shape or boundary of the object of interest as a parameter to detect and identify the algae. The project is aims to model an ideal automated system for *Oscillatoria sp.* which are freshly obtained with no broken or damaged samples.

From the result, it can be seen that the training of an artificial neural network to automatically recognize, identify and classify the selected algae is plausible with an accuracy percentage of about 91%. Also, when tested with a new set of images, the trained network is able to correctly recognize about 88% of the images as the selected algae species.

Images used to train the network is in binary form means there is distinct contrast and difference of the edge and the background, thus allowing the network to accurately learn to detect the boundary and shape of the objects and train itself to recognize similar variable in the images.

In addition, the image analysis and processing to detect its edge and boundary and convert them into binary form takes up time. Although it is more convenient to process the image using computational methods with the current technology and software available, the objective of making image classification a faster procedure seems to be quite disappointing.

Also, the images captured were freshly taken from the water. Being in different depth in a drop of water means the capturing of the image will be partly blur due to the water effect of refraction. Since the selected algae *Oscillatoria* is a rod shape species, the position of the algae while being observed under the microscope may not be planar. The shape of the algae will be affected in the captured images. Hence, the captured image of the species might not be of high quality.

Recognition procedures done manually in the ecological field to evaluate many aspects and characteristics of an living organism before classifying it, that too with a chance of misidentification. Hence, allowing an automated system to recognize an algae based on one parameter only is insufficient to 100% confirm the accuracy of the classification. But the good news is that this automated system does allow higher speed rate in eliminating non-selected species and narrowing down the scope for identification of the

selected species. Plus, comparing this system to a manual classification, this computational technique still has a chance of improving and increasing its efficiency, while the former does not.

Chapter Five: Conclusion and future work

5.1 Conclusion

In this research we presented an image processing techniques with ANN approach to identify and classify *Oscillatorias*. This study illustrated that computational recognition approach is important for freshwater algae, and prove that the classifying process is feasible. The developed system results showed better results in both accuracy and performance for automatic identification of the selected algae. The better accuracy resulted was obtained due to the well pre-processing used techniques, and due to the specific features selected during extract feature process. The main limitation of our system its inability to work well with images that include a huge number of objects. We would like to solve these limitations in our future work and make the system even more robust. Experiment showed that results was near optimal probably which indicates that the combined of ANN with rule-based classification is efficient to use with for classification and additional improvements for image preprocessing and extracted extra features can be used to obtain more accuracy results.

5.2 Future work

This project can be considered as a preliminary study towards the development of computer software that can identify and recognize all different types of species and/or algae. The new technique developed here as classifier can be extended to categorize all existing types of algae based on their boundary for optimization ANN processing time.