1. INTRODUCTION

Lake Chini and Lake Bera are the largest natural lake situated in Pahang state of Malaysia. These lakes play role as wetland which help preventing flood and erosion also as habitat for various flora and fauna which some of them are endangered species. It is important to keep this source to stay in good condition to balance the ecosystem and indirectly the lakes also one of the daily water source for the local people. Water quality monitoring is important to help in managing the water quality so that it will be safe whether as drink water, daily chores uses or for the fish to live.

1.1. Lake Bera and Lake Chini

Listed as a RAMSAR in 1994 for its importance of nature conservation Lake Bera has been preserved from development because it can disturb its nature. It is an example of blackwater ecosystem which consists of swamp area and swamp forest with grassland on the periphery. It lies in the basin of Peninsula's Malaysia largest river of Pahang river. Lake Bera eventually discharging into the South China Sea from as it flow through Bera River and then Pahang River. Lake Bera is a known habitat for many types creature such as animals and plants which some are rare and endangered species .This situation give a great opportunity for researchers as it is an important area containing a great combination of research to conduct.

Lake Chini is ranked as after Lake Bera as the largest natural lake where it is in second place. Lake Chini also plays role which Lake Bera has because they both served as wetlands such as preventing the floods and erosion of its riverbank. Then it can help the surrounding area save from the damage made by the flood. Lake Chini also provided a very important source especially to the local people as they can get the fishes in its lake since the lake also create a very good nature for fishes to breed and a place for fishes that migrate. Lake Chini flows by Chini River and then flows into Pahang River the same as Lake Bera.

However somehow the lakes encounter threat from its basin or within the lake itself and this has made some changing with their ecosystem. With development happen around the lake it affected to the damages of the plant around the area, the aquatic habitat and their system and made the lake to be polluted. Their role as wetlands are lost due to these problems mentioned and thus decreases the water quality. Therefore a study is needed to assess the quality of water because it is important for managing its quality for a sustainable management.

1.2. Water Quality Monitoring

Water quality have different requirement based on what purpose it will be used for such as to supply clean water to household or to cool down generator. Water quality as it has different requirement can be defined as chemical, physical, and biological characteristics of the water.

Water-quality monitoring is important because we can monitor and use it to control pollution that happen with water so that it can stay clean and safe to be used. There are several factors that can influenced water quality as it comes from physical factor, human activities such as development project, meteorology, chemical effects and many other factor.

In Malaysia Water Quality Monitoring has been established through National Monitoring Network Established in which was started in 1978. Several aims been created where the water quality monitoring is created as a platform to monitor the water quality status of river water in Malaysia and also to check the development activities as if it is affecting the water quality. They check the water quality based on continuous basis. DOE and ASMA has been collaborating to do the water quality monitoring as a result from high demand of monitoring.

As establish by Interim Water Quality Standards for Malaysia which is Interim National River Water Quality Standards, there are five classes of water quality standards table as

below and several important parameter used in monitoring. As in this project we use DO as a parameter to be predict based on the class of high, medium or low because it is an important parameter which can easily show the water quality whether it is in good condition or it is not in good condition.

Table1. 1 Interim National River Water Quality Standards

Class	**Description**
Class I	Conservation of natural environment, Water Supply I – practically no treatment necessary, Fishery I – very sensitive aquatic species.
Class IIA	Water supply II – conventional treatment required, Fishery II – sensitive aquatic species.
Class IIB	Recreational use with body contact
Class III	Water supply III – extensive treatment required, Fishery III – common, of economic value, and tolerant species livestock drinking
Class IV	Irrigation
Class V	None of the above

Source: Department of Environment Malaysia

Table 1.2 Class and parameters standard

Class **Parameter**	**I**	**II**	**III**	**IV**	**V**
BOD	***<1***	***1-3***	***3-6***	***6-12***	***>12***
COD	***<10***	***10-25***	***25-50***	***50-100***	***>100***
NH3N	***<0.1***	***0.1-0.3***	***0.3-0.9***	***0.9-2.7***	***>2.7***
DO	***>7***	***5-7***	***3-5***	***1-3***	***<1***
pH	***>7***	***6-7***	***5-6***	***<5***	***>5***
SS	***<25***	***25-50***	***50-150***	***150-300***	***>300***
WQI	**>92.7**	**76.5-92.7**	**51.9-76.5**	**31.0-51.9**	**<31.0**

Source: Department of Environment Malaysia

1.3. Support Vector Machine (SVM)

SVM is a tools created to solve the classification problem. It works by separating data into training and testing data and are gaining popularity due to many attractive features. Furthermore SVM now can also solve the regression problems. SVM and neural network share a quite similar process and system however SVM is has more advantages in solving a complex and nonlinear data since it use kernel function that can provide more solving method for the problems.

SVM modeling works by finding the best line to separate training data according to the classification chosen and place the data in the plane made by the line that has been created. There are vector that will be place near the line and it is the support vector.

1.4. Scope

Water Quality Monitoring requires a lot of effort, time and also cost. Besides, calculating water quality parameter is hard to get the accurate results since it is a nonlinear system. Some model has been recognized to solve this problem such as the Artificial Neural Network and SVM. SVM has been applied in many research fields and successfully solve the nonlinear system with its regression algorithm.

This project will use data from 15 stations from Lake Chini and Lake Bera which were sampled every two month from February 2005 until October 2009. 4 parameters were analyzed such as pH, temperature, turbidity, dissolved oxygen (DO) as the output to be predicted. The SVM also will be implemented using R.

1.5. Objectives

This project serves two objectives as below:

1. To predict DO classification of low medium and high at Lake Chini and Lake Beraby using SVM model.
2. To develop the SVM model based on the DO classification.
3. To test the accuracy of SVM model for its accuracy in the predicted DO.

2. LITERATURE REVIEW

Water is very important in our daily life in fact all living creature needs water to survive however the importance of this source also means that it is very possible to be polluted which will reduce its quality hence it will become a bigger problems in the end. Xiang and Jiang (2009), Xul, Wang, Guan and Huang (2007) in their articles agreed that monitoring the water quality and the forecast for it is a very important task to do. They mentioned that development such as economical activities is contributing to the pollution of water and that make it is a crucial task to monitor the water quality as to quickly solve the water pollution problems. As for the arisen problem there have been many researches that conducted the water quality forecast model. They mentioned that when we can determine the water quality parameter it can show at which level the evaluation of the water is and thus we can prevent the water pollution. In addition to the monitoring the traditional method used before offer a lot range of the monitoring parameters used. However water quality unfortunately can be directly polluted by so many factors such as limited manpower, materials, climate, landform, hydrologic conditions therefore the traditional way to monitor water quality is not really efficient to solve these problems because to do water quality forecast it involves a complex and nonlinear data. Moreover there are many problems with traditional way that may be from the human source and limitation and some other technical problems arise during inspection and make the monitoring become not productive. Therefore a new method which can increase the forecast of water quality is necessary to maintain the water quality.

Regarding the methods used in water quality monitoring, currently there are two main methods for monitoring and evaluating water quality as first by physical and chemical analysis, and second, biological monitoring methods as mentioned by Liao, Xu and Wang (2012). Water quality is evaluated by determining the existence and content of hazardous substances within the water directly using a variety of instruments. These physical and chemical analysis methods are accurate and sensitive, but they are time-consuming and cannot be used continuously in situ. While biological monitoring is to detect if there is any changes whether it involves in water quality itself or if there is presence of pollution by identifying changes in the health status, physiological characteristics, and behavioral responses of individuals or populations of aquatic organisms, providing a basis for environmental quality monitoring and evaluation from a biological point of view. Biological methods are once a system is established it can provide automatic alarms and can be used for long-time online monitoring of water quality. Furthermore the response of aquatic organisms to water quality is more sensitive and reliable and biological methods are also useful for detecting mixed pollution. Lastly they have a low cost and can easily be incorporated into a digital system.

Liao, Xu and Wang (2012) also agreed that traditional method do not solve the complex nonlinear relationships between assessment factors and water quality, and the assessment result is greatly affected by subjective factors of the assessing person.

Bouamar and Ladjal (2007), Xul, Wang, Guan and Huang (2007), Liu, Chang and Ma (2009) in their articles said that the automation tool of artificial intelligence techniques can provide a better result from the data that will be used which get directly from the monitoring station or the raw data. These data is known to be complex and is a nonlinear data which is hard to deal so the tool mention can be used to do the decision making aid. One of the tools that gain attention is SVM which has been successfully applied in many areas to do forecasting such as in biological area. SVM is a statistical learning theory (SLT) where it uses structural risk minimization principle with good generalization ability. It can solve the problem that conventional methods face in assessing water quality and can overcome the defects of slow training speed, poor network generalization, and low learning accuracy in artificial neural networks (ANNs). It also can fully utilizes the distribution feature of training samples to construct discriminant function based on part of the training samples, describing such nonlinear relationship. They also said that calculation result of their data shows that SVM has favorable classification performance and can be applied in water quality assessment. Besides, Liu, Chang and Ma (2009) mentioned when using SVM there are three main issues need to be considered such as feature selection, kernel function selection, and the penalty and inner parameters of kernel function selection.

In Bouamar and Ladjal (2007) they use the SVMs technique to solve pattern recognition and clearly satisfied with the result produce from the SVM technique but there also error produce from it and they conclude that with the increasing of training data and new sensor the precision can be improved.

In Liu, Chang and Ma (2009) in their journal mentioned that as water quality assessment is a complex data (SVM) can transform the learning process into a convex quadratic planning problem to get a global optimization by using the rule of minimum structure risk, which is appropriate to solve small-sample, nonlinear classification and regression issues. They apply SVM in water quality assessment for karst groundwater sample at the Niangziguan fountain region of Haihe River basin to obtain the grade of water quality assessment. The result shows that such a method solves the complex nonlinear relationship between assessment factor and water quality grade. It offers high prediction accuracy and is a reasonable and feasible assessment method.

When we mentioned about SVM there is another tool we cannot forget which is really close to SVM. Researchers often do the evaluation between these two tools. Bouamar and Ladjal (2007), Xiang and Jiang (2009) evaluate both tools in their paper which other tool is the Artificial neural network (ANN). ANN can also process nonlinear data and it also can gives result with high accuracy however it is itself a complex structure and made it poor in its performance.

Bouamar and Ladjal (2007), in their finding found SVM can deal with complex and highly nonlinear data with good result even the training sample is only few. They did an evaluation of ANN and SVM techniques which both of the tools shows a highly good results about 86%. They found that the corresponding time of ANN to the training data is better but it is not very sensitive to the noise produce while SVM is good when dealing

with this noise and it shows that as for water quality monitoring the SVM is a better tool to do the forecasting of water quality.

SVM prediction method used to do DO prediction however has been done in many areas such as in Najah et al. (2011), they use different kind of machine learning method to do prediction of DO such as the ANN, ensemble and also SVM. The research was done for the river water of Johor state which is for Johor River. Based on the research SVM give the best performance among all the methods that have been used in doing the DO prediction. Other machine learning method used to do DO prediction such as ANN method which was establish earlier then SVM and also by using mathematical method. In Palani et al, (2008) they have agreed that DO prediction is successfully done using ANN method. Based ontheir finding the ANN method used gives a good result when they do the prediction for seawater for Singapore which they obtain acceptable accuracy. In Junsawang, Pomsathit and Areerachakul (2011), Naik and Manjapp (2010) they agreed that DO is the best parameter to indicate the water quality. The prediction done in their research for river in Thailand and India is by using regression method where they predicted the DO value and not the classification of DO. The results they get show a good production of predicted value of DO. It is shown that the prediction is almost as the actual value in their data. They are satisfied with the result obtain by using their methods.

3. METHODOLOGY

3.1. Support Vector Classification

SVM can solve classification or the regression problem and in this project SVM is used to solve classification problem. As the early development of SVM it was create to solve classification problems. The classification problem involves separating data into training and testing sets where it contains the class label and variables input of data. SVM then from the training data will create a model which going to be used to predict the class label on the testing data.

SVM works by predicting the labels of training data as $D = \{(\overrightarrow{x_i}, y_i), i = 1..N\}$ with $y_i \in \{-1, +1\}$ is separable by a hyperplane. Where if the data is positive it belongs to class 1 and if the data is negative it belongs to class -1. When data is linearly separable and support vector existed it can be derived as:

$$f(\vec{x}) = \vec{w}^T \vec{x} + b (= w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n + b)$$

W and b are determined during training process where it locates the support vector. W is defined as weight vector while b is term for bias and is a scalar. T stands for transpose operator.

Data in SVM can be a linear or nonlinear data. For a linear and separable dataw and b is minimize to $^1/_2 \|w\|^2$

Subject to constraints: $y_i(w^T x_i + b) \geq 1, \ \forall i$

To solve the optimization Lagrangian function defined as below is used:

$$L(w, b, a) = \frac{1}{2} w^T w - \sum_{i=1}^{N} a_i [y_i(w^T x_i + b) - 1, \qquad a_i \geq 0, \forall i]$$

α_i is required to express w. When $\alpha_i > 0$ it is called the support vectors and it resulted in

$$w = \sum_{i=1}^{N} a_i y_i x_i$$

$$\sum_{i=1}^{N} a_i y_i = 0$$

The Lagrangian then derived into the dual optimization problem as follow

maximizationα

$$\sum_i a_i - {}^1/_2 \sum_i \sum_j a_i a_j y_i \, y_j x_i^T x_j$$

then

$$\sum_{i=1}^{N} a_i y_i = 0 \, a_i \geq 0, \forall i$$

SVM for linear predictor can be expressed from as below

$$f(x) = w^T x_i + b = \sum_{i=1}^{N} a_i y_i x_i^T x + b$$

where

$$b = \frac{1}{\left|I_{support}\right|} \sum_{i \in I_{support}} \left(y_i - \sum_j a_j y_j x_j^T x_i \right)$$

andI_{support} is the set of support vectors.

If f(x) is positive the test data will belong to class of y_i =1 and for negative it will be classified into the other class.

The nonlinear data equation on the other hand can be derived as

$$^1/_2 \|w\|^2 + C \sum_i \varepsilon_i^2$$

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \qquad \varepsilon_i \geq 0, \forall i$$

Here $C > 0$ is a constant of positive number on a selected parameter. $\varepsilon_i$shows distances of data lying on false class side and its margin of predicted class.

As in the linear problem where it can be optimized into dual problem it also can be used in nonlinear case. Optimization problem can be converted into dual problem as:

$$\sum_i a_i - {}^1\!/_2 \sum_i \sum_j a_i a_j y_i \, y_j x_i^T x_j$$

subject to

$$\sum_{i=1}^{N} a_i y_i = 0,$$
$$0 \le a_i \le C, \ \ \forall i$$

By this it is dependable on the data to choose the appropriate C.

In SVM kernel function is introduced as it offer a better performance when dealing with a nonlinear data. Here data is mapped and derived as below where K is the kernel function

$$\Phi(x)^T \Phi(y) = K(x, y)$$

By introducing the kernel it can be derived as

$$\sum_i a_i - {}^1\!/_2 \sum_i \sum_j a_i a_j y_i \, y_j K(x_i x_j)$$

subject to

$$\sum_{i=1}^{N} a_i y_i = 0,$$
$$0 \le a_i \le C, \ \ \forall i$$

The equation produce is as below

$$f(x) = w^T \Phi(x_i) + b = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b$$

The kernel selection depends on the data distribution but kernel selection also generally done through trial and error.

3.2. Kernel-based Machine Learning Lab (Kernlab)

The SVM was implemented using R software. Kernel-based Machine Learning Lab (Kernlab) package was used to do the SVM. Along kernlab there are a few more packages in R that can be used to do SVM. Such as R package e1071 which is very efficient SVM implementation. Another SVM related R package available is klaR.

In kernlab R user is provided with basic kernel functionality and other functions. Besides user can also create own function to the kernel based on the kernel is in the package. It is applied to class system of S4 where declaration is needed of every step taken by the user and it is stricter however it is still consistent.

There are prepared data included in this package such as spam data set which classifies spam or non-spam. Promotergene, ticdata data set, spirals data set and lastly the income data set are more of the data included in this package.

In kernlab it supports about seven kernels such as:

The linear kernel

$$k(x, \acute{x}) = \langle x, \acute{x} \rangle$$

Gaussian radial basis kernel

$$k(x, \acute{x}) = exp(-\sigma \|x - \acute{x}\|^2)$$

The polynomial kernel

$$k(x, \acute{x}) = (scale. \langle x, \acute{x} \rangle + offset)^{degree}$$

The Hyperbolic tangent kernel

$$k(x, \acute{x}) = tanh(scale. \langle x, \acute{x} \rangle + offset)$$

The Bessel kernel

$$k(x, \acute{x}) = \frac{Bessel^{n}_{(v+1)(\sigma \|x - \acute{x}\|)}}{(\|x - \acute{x}\|)^{-v(v+1)}}$$

The Laplace radial basis kernel

$$k(x, \acute{x}) = exp(-\sigma \|x - \acute{x}\|)$$

The ANOVA radial basis kernel

$$k(x, \acute{x}) = \left(\sum_{k=1}^{n} exp(-\sigma(x^k - \acute{x}^k)^2) \right)^d$$

The linear kernel is known as vanilladot is used when we predict the linear data. The Gaussian radial basis kernel is known as the rbfdot which is usually used for doing the classification prediction and for general purpose and in the rbfdot the parameter introduced is the sigma. This is the same as Laplacian radial basis kernel or the laplacedotwhich has the sigma parameter. Lapelacedot is usually used for general purpose also. The Polynomial kernel used three different parameter which known as scale, offset and degree in its function and usually the polynomial kernel that is known as polydot is used for the image classification. In Hyperbolic tangent kernel which is known as tanhdot is usually used in neural network purpose and the parameter we need to determine are scale and offset. The Bessel kernel is used for general purpose and it is known as besseldot. The parameters used in this kernel are sigma, order and degree. Lastly the Anova kernel is known as anovadot usually deal with the multidimensional regression problems and the parameters used are sigma and degree.

Ksvm as an implementation used in kernlab is a very efficient method because of it functions and also because it applied the C-SVM classification which can also predict for multiclass classification problems where some method that it uses are like one-against-one method and the other method is pairwise classification method. These methods used

voting method to do prediction and have been shown to produce good results when used with SVM.

Another method that can be used is by solving the problem by including the data from all classes such as derived:

$$t(w_{n,}\varepsilon) = \frac{1}{2}\sum_{n=1}^{k}\|w_n\|^2 + \frac{C}{m}\sum_{i=1}^{m}\varepsilon_i$$

subject to

$$\langle x_i, w_{yi}\rangle - \langle x_i, w_n\rangle \geq b_i^n - \varepsilon_i$$

where

$$b_i^n = 1 - \delta_{\mathrm{yi,n}}$$

where the decision function is

$$argmax_{m=1,\ldots\ldots,\mathrm{k}}\langle x_i, w_n\rangle$$

For R package kernlab it is adapted to R new and modern functions where it allows the used to explore using its package by bravely constructing new kernel function of the algorithm existing in it. As for ksvm it helps improving the prediction by allowing multiclass problem classification.

3.3. Data

Lake Bera consist of 26 000 hectares of its core zone and 27 500 hectares the buffer zone all has been preserved as RAMSAR sites and it is coordinate at 3°49′00″N102°25′00″E. Lake Chini consists of about 5026 hectares and is situated in coordinate 3°26′N102°55′E. Both Lakes have the climate of equator in Peninsular Malaysia which having the humidity, temperature and rain fall at an average characteristic.

Figure 3.1 Lake Bera Map, Source: go2travelmalaysia.com

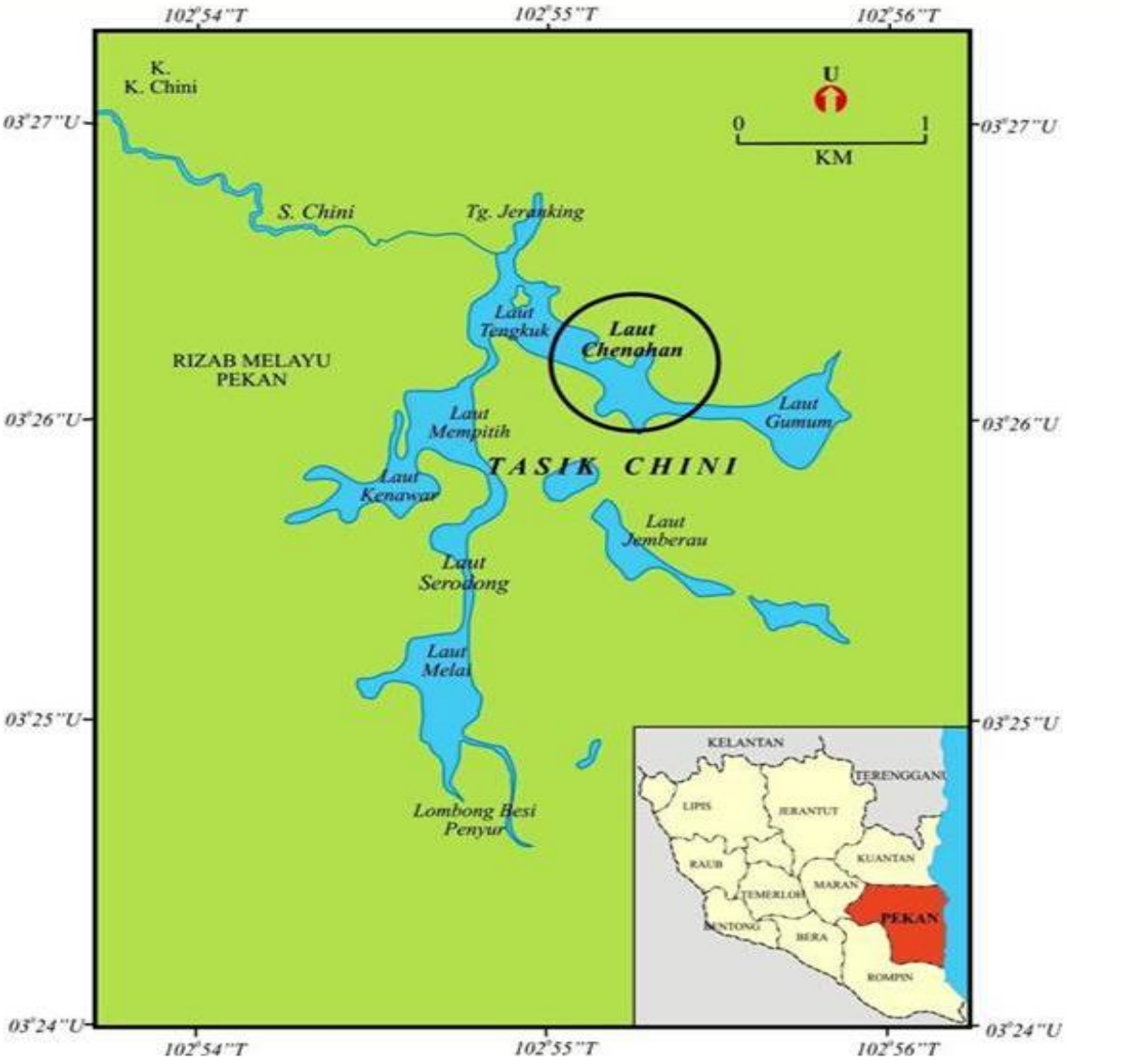

Figure 3.2 Lake Chini Map. Source: http://www.ukm.my/ahmad/tesispelajar/fitochenahan_files/image319.jpg

In Shuhaimi Othman, Lim and Mushrifah (2007) based on the research conduct on Lake Chini the water quality is decreasing due to the pollution that happen because of development.

Lake Bera and Lake Chini water quality data were collected starting from February 2005 until October 2009. The lakes were monitored regularly for every two month during mentioned years above which is from 2005 until 2009. It was monitored from six stations for Lake Bera and from Lake Chini it was monitored by nine stations. The stations monitored are as followed below:

Table3. 1 Monitored station of the lakes

Lake	Stations
Lake Bera	4PH03, 4PH07, 4PH66, 4PH67, 4PH71, 4PH72
Lake Chini	4PH75, 4PH76, 4PH77, 4PH78, 4PH79, 4PH80, 4PH81, 4PH82, 4PH83

In this project we used 4 parameters of the lake data. The parameters that were measured and put in the data such named as pH (pH), temperature, turbidity and dissolved oxygen (DO).

The pH is used as the indicator for the water to be determined whether it is acidic or alkaline. In the standard state by the Department of Environment the pH suitable and safe for Malaysian rivers range between 5.00 to 9.00 and the lakes results are within the range which is from the lowest are 4.97 until 7.94. The pH was slightly fall from the range when February 2005. Temperature varying from 26.51 until 33.63 degree Celsius and turbidity are from 1 and 282.2 *nephelometric turbidity units (*NTU). All of these variables are the factor that has the most affecting factor to the DO level whether it is good or not. As for the DO chosen to be the output because DO is a good indicator to know whether

the river is clean and save or not because it shows how many oxygen can be dissolved in water and if the creature in water can survive with the DO condition at certain level.

3.4. Data Prediction Using SVM

The water quality data has about 147 samples and 11 variables. 10 variables were the input used to do the SVM and DO is used as the output in this project. DO variables were labeled in classification as high, medium and low based on the classification standard by the Interim National Water Quality Standard, Malaysia (INWQS) and Department of Environment as in table 1.2. In this classification the high class of DO range for value 7 above while the medium class ranges between 5 until 7 and lastly for the DO that had value below 5 is in class low. Data selection then is done by eliminating the data that was empty or was not available data. Then data was divided into training and testing data set to do the SVM. It was divided by 80% for training data and 20% for testing data. 80% data for training data consist of 80% of class high, medium and low also for testing data is vice versa. Next data was converted into comma delimited (CSV) format before it can be used to run in the R software.

To run SVM in R we need to call the kernlab package. Then data that was in the CSV format was imported into R by using function read csv. Data was reviewed in R by using summary function where we can see the information of data such as the mean, median, the min and others. The parameter was tested one by one to see which data has the least error where here we can see which parameter is relevant to use to do the prediction using SVM. Parameter was tested by using linear kernel which known as vanilladot to

determine the cross validation error. Result from using the linear kernel function then was ranked from the least error result until the parameter that has the big error.

After determined which parameter was ranked in ascending ranking then we do the selection of the best parameter that can compute highest accuracy by doing forward selection. Parameter was added one by one and after that kernels function was tested on them to see which kernel suits best by seeing the error that produced when running the program. The kernel that has least error was then chosen to be used as kernel for the SVM model. In this step default value was used for each kernel.

Kernels have different parameter used in their function and depend on their parameter we need to choose the best parameter value which can generate least error and finally when we do prediction function it gives high accuracy. To determine the least error for parameter in kernel we did the loop function in the kernel chosen then from the resulted calculation we chose the value with least error. From the best value of each parameter we then inserted it to the model and run it. After that we did the prediction function to know the prediction that has been done by the SVM. To see the cross tabulation table of prediction and actual data table function is used. Lastly the calculation for accuracy of prediction is done automatically using R. The model evaluation is done by using the sampling of cross validation error and using the sensitivity, specificity and accuracy method.

4. RESULTS

The data of Lake Bera and Lake Chini shows that DO levels of these lakes mostly is at the level medium and from this level the lake is classified as the water need to be treated to be as a supply and it involves the sensitive aquatic species.

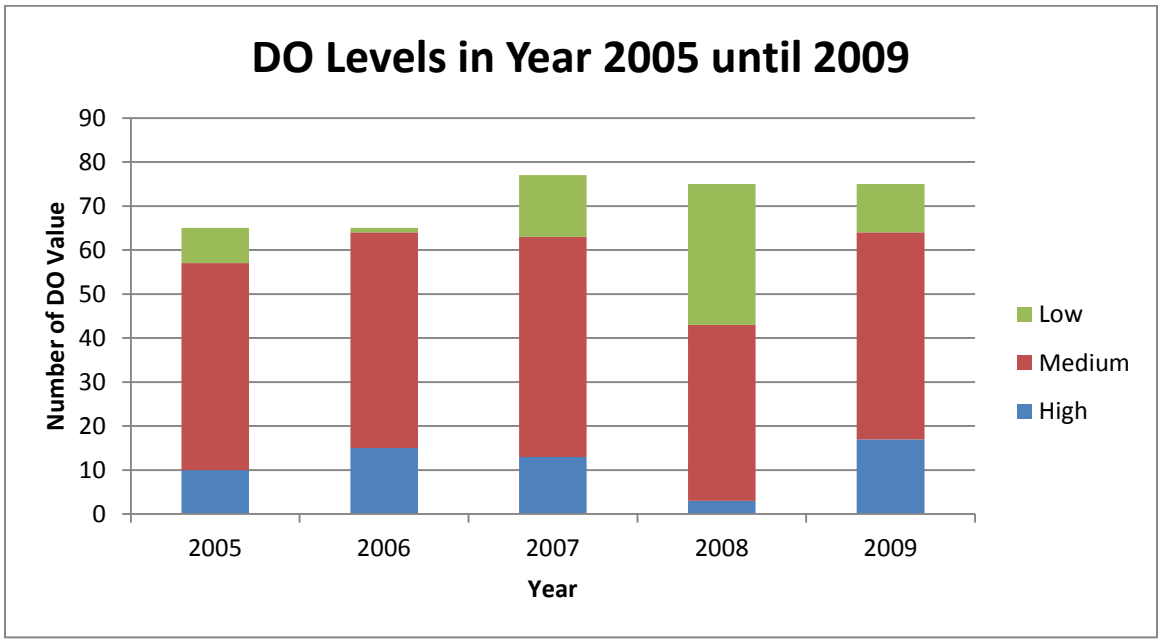

Figure 4. 1 DO Levels at Lake Bera and Lake Chini

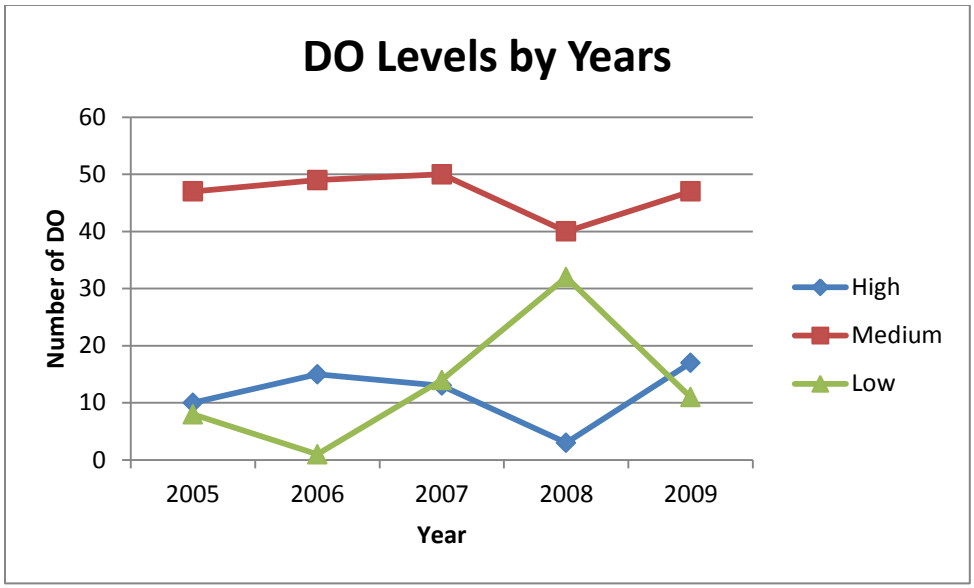

Figure 4.2 Levels of DO by year

As seen in the graph year 2008 shows that the lakes decreases in high level and low level of DO is increasing.

In the SVM method used to do prediction based on DO classification firstly each parameter is tested to see whether it is suitable to use as an input and from the result that generate by R shown below on its ranking with the least error until parameter that has big error:

Table 4.1 Parameter ranking

Ranking	Parameter	Error
1	Temperature	0.458333
2	pH	0.516667

Table 4.1 continue…

3	Condition	0.575
4	Coliform	0.608333
5	Turbidity	0.625
6	Natrium	0.675
7	Salinity	0.691667
8	Phosphate	0.7
9	*E.coli*	0.708333
10	Nitrate	0.8

When doing forward selection the data was added up one by one according to the ranking above. In each step of determining which parameter is good, the determination of which kernel was best also been done. When 2 and 10 inputs were used the best kernel was the radial basis and 3, 4, 5, 7, 8 and 9 inputs show that Anova is the best kernel. Lastly with 6 inputs of parameter it was suitable using Laplacian kernel. With the best kernel then SVM model are run and the accuracy produce from these kernels when run with certain inputs parameter is shown in table below.

Table 4.2 Accuracy for kernels

Inputs	Parameter	Kernel	Accuracy (%)
2	Temperature, pH	Radial Basis	44.44
3	Temperature, pH, Condition	Anova	74.07
4	Temperature, pH, Condition, Coliform	Anova	62.96
5	Temperature, pH, Condition, Coliform, Turbidity	Anova	70.37
6	Temperature, pH, Condition, Coliform, Turbidity, Natrium	Laplacian	66.67
7	Temperature, pH, Condition, Coliform, Turbidity, Natrium, Salinity	Anova	42
8	Temperature, pH, Condition, Coliform, Turbidity, Natrium, Salinity, Phosphate	Anova	48. 14
9	Temperature, pH, Condition, Coliform, Turbidity, Natrium, Salinity, Phosphate, *E.coli*	Anova	44.44
10	Temperature, pH, Condition, Coliform, Turbidity, Natrium, Salinity, Phosphate, *E.coli*, Nitrate	Radial Basis	51.85

From the table shown when using 3 inputs parameter using kernel Anova the highest accuracy that was obtain about 74%. This means that the best 3 parameter is relevant to use in doing the SVM prediction. Furthermore it is shown that when parameter is added

up the accuracy is decreasing. The best five parameters which shown good accuracy result is actually the factor that gives effect to the DO. With the result shown it is also prove that DO is affected by these five parameters. However in this project 3 inputs were used as it provides highest accuracy among all.

In kernel Anova the parameters need to be determined in this function are degree, sigma and cost C. By doing the loop function the degree obtain is 1.5, sigma is 1 and the cost C used for this model is 24. This model computes the cross validation error about 0.575. The accuracy as mentioned above is 74.07%. The prediction made by this model can be seen in the figure below.

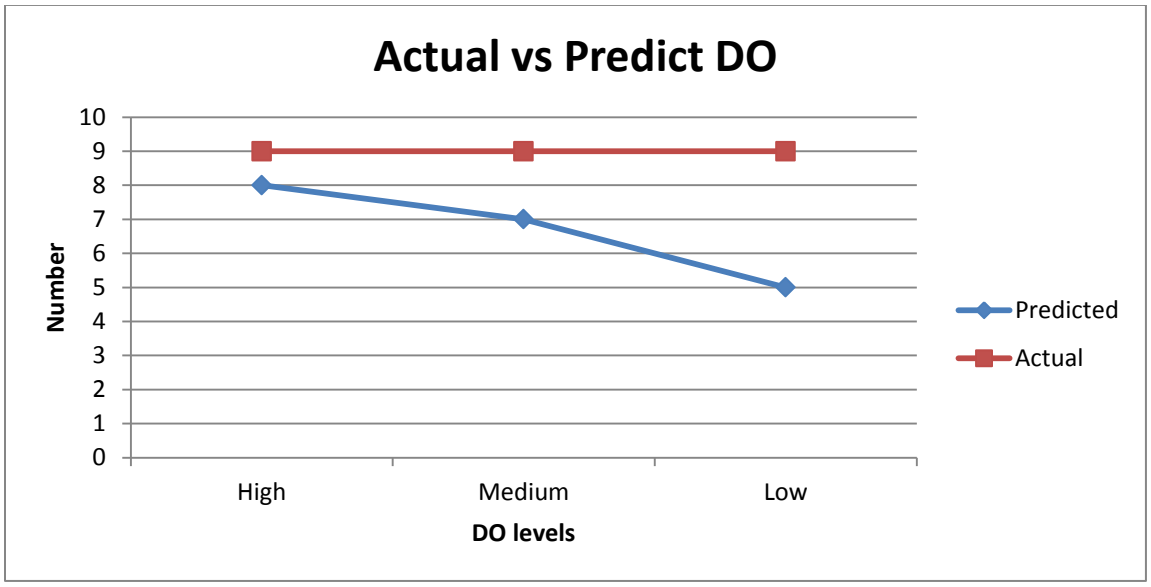

Figure 4.3 Predictions and actual value of testing data

Table 4.3 Prediction result on testing data

DO	pH	TEMP	COND	Prediction	Prediction result
Medium	6.18	30.13	28	Medium	TRUE
Medium	6.45	31.17	28	Medium	TRUE
Medium	6.34	30.97	29	Medium	TRUE
Medium	6.77	31.065	28	Medium	TRUE
Medium	6.24	30.67	32	Medium	TRUE
Medium	6.3	30.6	29	Medium	TRUE
Medium	6.07	29.93	28	Medium	TRUE
Medium	7.59	26.12	30	High	FALSE
Medium	7.77	27.7	31	Low	FALSE
Low	5.79	28.46	26	Low	TRUE
Low	5.81	28.12	22	Low	TRUE
Low	6.41	29.88	22	Low	TRUE
Low	7.03	29.67	78	Low	TRUE
Low	6.66	28.37	33	High	FALSE
Low	6.24	32.64	37	Medium	FALSE
Low	6.15	30.95	25	Medium	FALSE
Low	5.98	32.62	26	Medium	FALSE
Low	6.9	29.82	69	Low	TRUE
High	6.67	30.72	26	High	TRUE
High	6.68	30.47	25	High	TRUE
High	7.01	33.47	23	High	TRUE
High	7.26	30.71	125	Medium	FALSE
High	6.64	30.91	43	High	TRUE
High	6.71	30.28	26	High	TRUE
High	6.42	32.86	27	High	TRUE
High	7.89	31.54	21	High	TRUE
High	7.42	32.54	23	High	TRUE

Results shown that this model predicted the high classification as the most correctly predicted, and after that medium class and after that is the low class. From the prediction that has been made sensitivity, specificity and accuracy for each class can be calculate as this shown figure below.

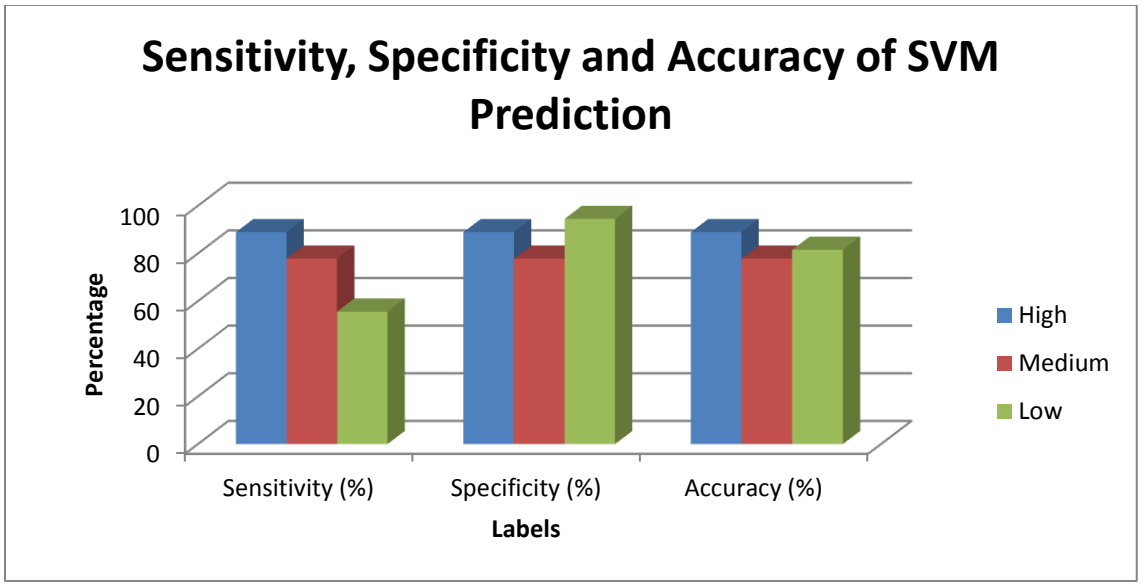

Figure 4.4 Sensitivity, Specificity and Accuracy Graph

Table 4.4 Calculated sensitivity, specificity and accuracy

	Sensitivity (%)	Specificity (%)	Accuracy (%)
High	88.89	88.89	88.89
Medium	77.78	77.78	77.78
Low	55.56	94.44	81.48

Overall accuracy is 74.07 % and as shown in the table the sensitivity for high class is 88.89% while specificity is 88.89% and for medium class the sensitivity is 77.78% while specificity is also 77.78%. For low class of DO sensitivity is 55.56% and specificity is 94.44%. The accuracy for high class calculated about 88.89% and medium class accuracy is 77.78%. Lastly class low accuracy is 81.48%.

5. DISCUSSIONS

Based on the result present in the previous chapter, the prediction made using SVM model produce error which is about 0.5 from the cross validation error method and the accuracy about 74% when done automatically using R. The accuracy is calculated for the prediction that matches the actual data over overall data. The research done by other researcher such as Najah et al. (2011), Bouamar and Ladjal (2007), Xul, Wang, Guan and Huang (2007), Liu, Chang and Ma (2009)when using SVM to predict data produce small value of error and the accuracy of their prediction is above 70%. The training was set to 80% containing each levels of DO also for 80% because we want the data to be sampled at fairly value. So SVM is predicted to give the best prediction when the entire sample is divided fairly.

This is proven in the prediction when data is divided properly and the sample use for training and testing data is chosen carefully the result can show a good result. Such as in the parameter selection when the parameter with high error was added it produced bad accuracy and it reduced the performances of SVM to do prediction correctly. Since in SVM the term bias is introduced it is affecting the sample selection where if the data has more high class then the prediction will be bias towards high class label and so on.

The introduction of empty sample also is problem if we do not remove it because the SVM will not include it in the calculation and we will have problems with the data in the future step when doing SVM. The forward selection is used because it is easy to

recognize which parameter actually works or not because we add the parameter one by one and if it gives bad result we can eliminated it and substitute with another parameter.

The sensitivity for each class is quite high except for the low class however it is acceptable because the high and medium class sensitivity is higher. This is important because the prediction can predict the DO high class and medium class very effectively as high class of DO is clean water and medium class require treatment and low class of DO need an extra treatment for the water. The sensitivity is high because the true negative result in this prediction is high because it is not wrongly predict the data that was not supposed to be predicted in wrong class but the accuracy shows that the prediction is accurate for all classes. However the error that obtains from this prediction is quite high and it is because the value used in the kernel parameter is affecting the error result. Such as the cost C is about 24 and it shows that this model tolerates more error.

Water quality since it is not easy to be predicted is same with these data because several factors which arise in these lakes such as climate and the development that happen around them. Such as in 2008 where the DO levels drop mainly from Lake Chini is because some factor of development such as the land activities and from the river flow that drained into Lake Chini especially during wet season. Prediction on water quality also expected to falls below predicted model because of random error cause by nature factor.

6. CONCLUSION

Water quality monitoring is hard to be forecast. However with the introduction of artificial intelligence techniques it helps in predicting the water quality successfully. SVM is nowadays recognized to do the forecast of this complex data. In this project prediction made by the SVM for the Lake Bera and Lake Chini produces accuracy about 74.07%.It accurately predicts the class of DO. This prediction can be used to predict new data for future data. Besides in ecology the precision is not as important as the range because we want to know whether the water quality is acceptable to be use or not. In the objective mentioned in chapter 1 show that these two lakes are in safe condition as based on the standard given for DO quality. All objective produces for this project is achieved.