Research on Water Levels Prediction for Disaster Management Using Machine Learning Models

A Thesis Submitted to the Department of Computer Science and
Communications Engineering, the Graduate School of Fundamental Science
and Engineering of Waseda University
in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering

July 24th , 2018.

Tin Nilar Lin

(5116FG18-1)

Advisor: Prof. Hiroshi Watanabe

Research guidance: Research on Audiovisual Information Processing

DEDICATION

To my beloved parents:

Tin Soe

&

Kyi Kyi Sein

ACKNOWLEDGEMENTS

First and Foremost, I would like to express my deepest gratitude to my research supervisor Professor Hiroshi Watanabe of Department of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University for his invaluable guidance and supervision, timely suggestions, continuous encouragement, kindly supporting and understanding throughout the research project to a successful completion.

I would like to express my special thanks to the lecturers of Department of Computer Science and Communications Engineering, Waseda University whose lectures, knowledge and contributions have assisted me in broadening my research skills.

Moreover, special thanks are extended to Japan International Cooperation Agency (JICA) who supports me the financial aid to study in Waseda University through the JDS scholarship program.

Furthermore, I would like to express my sincere thanks to all lab members for their motivating, encouragement and stimulating discussions about the research during the entire period of the study and for our fun times together.

Last but not least, my heartily grateful to my parents and two sisters for their unconditional supports and encouraging constantly during these two years of study and all those who directly and indirectly involved in this research to be well completed.

ABSTRACT

Nowadays, Flood hazards become the common natural disaster among the natural hazards and the significant challenge for disaster management around the world because of frequently occurrences and leading cause of devastating effect on many countries in every year. It is important to protect the people life and their property from disaster effected area by giving the early warnings using the effective prediction approaches. Therefore, Flood prediction and early warning issues according to the geographical locations become the important issue in natural disaster management field. This study addresses the need of the water level prediction model for flood hazards and warnings depend on the area of interest. Machine learning prediction approach based on the time series data are proposed and making optimization to solve the water level prediction over the three different locations in Myanmar is described. The model focus on the prediction performance based on data driven machine learning approaches using the historical hydro-metrological data for river flood prediction. The proposed prediction approaches that can provide a head of time for estimation of water levels. Three different machine learning models: k-nearest neighbor (KNN), support vector machine (SVR) and multiple linear regression (MLR) are analyzed to predict the five days ahead water level and compare the accuracy and show the predicted and observed result of hydrological observation stations of the study region. As the experimental results, K-NN and SVR have the better prediction performance than MLR on three study locations within the acceptable error rate and also potential for future water levels prediction data driven machine learning approaches of Myanmar.

Keywords: water level prediction, time series analysis, KNN, SVR, MLR

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER 1
INTRODUCTION

For centuries, human have been challenged by natural disasters and the challenges have been greater today. Natural disasters have been increasing around the world because of the globalization, climate changes and deterioration of natural environment of human alterations. Among the natural disasters, flood is the common one because of the leading causes of devastating effects on many countries all over the world.

Devastating caused by flood is extremely harmful to the social economic life of the people and country very often especially in developing countries that cannot apply the advanced technologies and methods to predict for weather forecasting and warning issues of flood management. In most of the developing countries, accurate flood prediction models are not equipped properly. As a result, people from flood prone areas are suffering more than they should. However, natural disasters cannot be prevented from occurring.

Thus, the best way is to prevent the loss of life, people injury and reduce economic, material and environment impact is "effective prediction approach". That is why, implementing more accurate and suitable flood prediction with the consideration of the most important weather data combination of water level and rainfall is crucially important.

Actually, estimation of weather data phenomenon is complex to understand. There are many influences factors, sequential and temporal structure have to be collected from different locations. Changing the locations is another important variation of weather condition and the effect of the weather of one place can extend for a distance of another location and become the weather for that location. Therefore, it is essential to consider the weather data conditions along the path going to the one location to another to get the accurate prediction model.

Thus, prediction of weather data such as water levels, rainfall, temperature, pressure, wind speed, wind direction etc.,) based on time series prediction become important in their corresponding fields. Efficient, flexible and accurate data driven model based on these times data is needed to develop. Our approach in this work is to propose the effective water level prediction using the data driven machine learning models. We investigated the three different machine learning models and trained the models using our real time historical data of our study areas and making the prediction for days ahead water levels for flood disaster management.

1.1. Causes of Flood

In general**,** flood is an overflow of water from rivers, lakes reservoirs and etc., whereby water inundate outside of the water bodies. There are many causes to become flood and major causes are: Monsoon rain and abnormally heavy rain, coastal storm surges, failure of dams, poor drainage system, rapid snow melts, deforestation, global warning, urbanization increase and surface runoff and many others.

1.2. Problem Statement

There are many causes to occur the flood every year all over the world. But we can’t be escaped from the ageing natural disaster. But, if the prediction model and pre-warning is effective, loss may be minimized and even eliminated. In most of the developing countries, flood prediction systems under the metrological and hydrological department are not properly equipped and lack of applying the ICT technologies, insufficient expertise and limited budgets. Thus, people from the affected flood prone area suffering every year as the consequences of the flood hazards. Moreover, efficient and cost-effective flood prediction systems are not very common in many developing countries. Therefore, several damages due to the flood disasters are still occurring and bring loss of life and properties.

In general, flooding occurs when the river water surpasses the normal level and it becomes overflowing. Therefore, determining river water level and notify the pre-warning for flood hazards with the proper water level prediction approaches based on the geographical locations is crucially important.

Thus, water level prediction systems must be accurate according to the locations. If so, people from flood prone area can be notified earlier. Our work based on data driven machine learning approach that can be provided the simple and effective water level prediction for flood and disasters management and will help the people to save their lives from hazard conditions with our projected information.

1.3. Study Region

Myanmar is the agricultural based developing country and its economy is mainly depend on changes of weather and climate conditions and application of water and climate factors play an important role for the rice and other staple food production, transportation, irrigation and water resources and other sectors.

Myanmar is situated between latitudes 9° 32' N and 28° 31' N and longitudes 92° 10' E and 101° 11' E and located in Southeast Asia with the area of 676528sq Km (square kilometres). It has a long sea coast facing the Bay of Bengal in the west, continues south-ward facing the Andaman Sea in the south and south-west. Myanmar enjoys the south west monsoon with three seasons: the rainy or monsoon season, from June to September, the cool season, from October to February and the hot season from March to May.

Most of the area receives 90% of annual rainfall by Monsoon season. At the same time, there is threat of the storms and weather disturbance such as cyclones in the Bay of Bengal, frequency passage of Western disturbances from NE India, frequency of Easterly Waves and Typhoon Remnants from the China Sea towards Myanmar can cause strong wind, abnormal heavy rain and cause the flood especially in Ayeyarwady river and can cause the loss of life and properties.

The evidence of flood disasters due to cyclones in last decades in Myanmar are: Sittwe Cyclone in 1968, Pathein Cyclone in 1975, Gwa Cyclone in 1982, Maungdaw Cyclone in 1994, Cyclone Mala in 2006, Cyclone Nargis in 2008 and Cyclone Komen in 2015. The most striking evidence is that Cyclone Nargis outbreak in 2008, killing over 100,000 people and displacing many others and causing the enormous economic losses and significant human suffering.

1.4. Data Used

In our work, hydro-meteorological data from three stations along the Ayeyarwady river and two stations data of Bago river are used and location of these stations are shown in Table 1.1.

Ayeyarwady river is the largest and vital river of Myanmar and one of the greatest rivers in Mekong region. It flows through the northern part to southern part, Andaman Sea in the Bay of Bengal. It is 2,170 kilo meters long and river’s basin is about 413,674 square kilo meters covering about 60% of total area of Myanmar. There are 16 stations for hydrological forecasting along Ayeyarwady river, the main river of Myanmar and often face the local floods especially in rainy seasons and sometimes even in dry seasons due to the influence of climate changes. That is why, we have selected the three stations Satation-1(Myitkyinar), Station-2 (Bhamo) and Station-3 (Mandalay) of Ayeyarwady river that is the region of interest and expressed in Figure 1.1.

Bago river is the 331 kilo meters long and the catchment area is 5,348 square kilo meters. It starts flowing from central mountainous region called Bago Yoma .The largest portion of river itself is in Bago region and small portion of river outlet is in Yangon region and it joined with Yangon river and from there, flows into the Gulf of Mottama. Bago river basin is also flood prone area and evidence is in 2011 there are two severe floods occurred and in monsoon season and nearly all rivers and creeks are flooded and destroyed thousands of households and adjacent paddy fields. In Bago river, there is only two water levels observation stations. Bago river which is the important flood prone are in lower part of Myanmar and illustrated in Figure 1.1.

Station No	Name	Region	Lat	Long
Station 1	Myitkyinar	Upper	25° 39' N	97° 38' E
Station 2	Bhamo	Upper	24° 26' N	97° 23' E
Station 3	Mandalay	Middle	21° 95' N	96° 08' E
Station 4	Zaungtu	Lower	17° 37' N	96° 14' E
Station 5	Bago	Lower	17° 19' N	96° 28' E

Table 1.1. Detail of hydro/metro stations with latitude and longitude

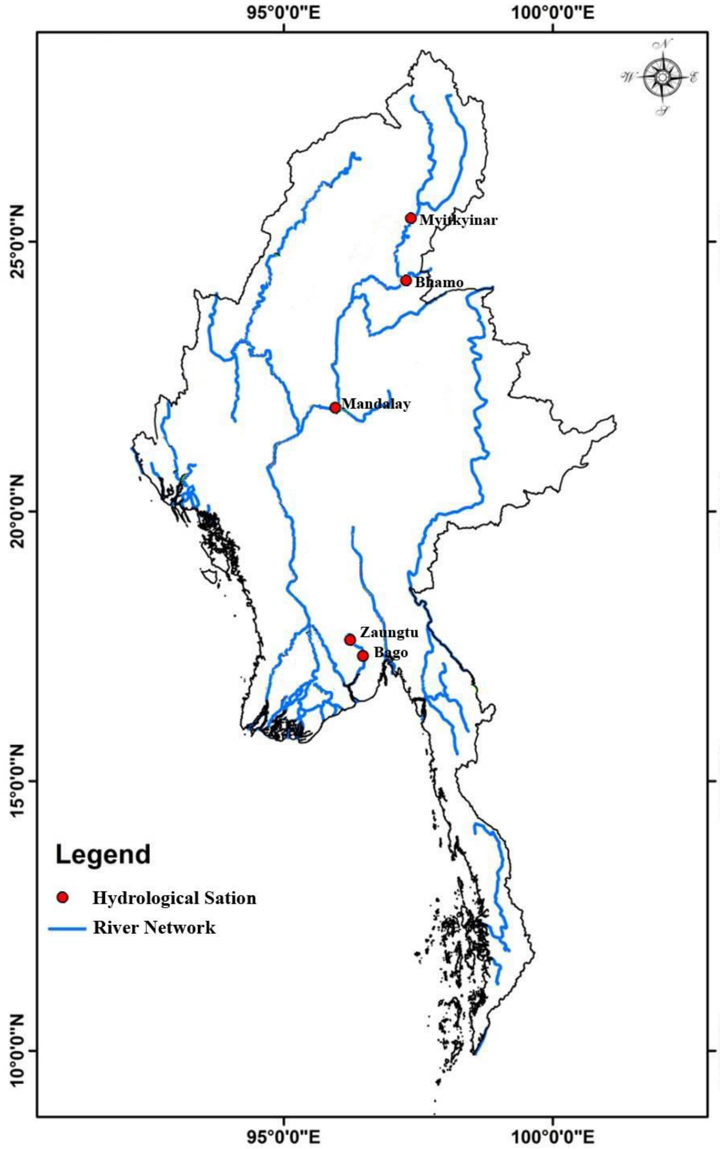

Figure 1.1. Myanmar Map with water level observation stations

1.5. Aims and Objectives

The main aim of this thesis is to investigate the effective, flexible and more reliable machine learning prediction model for the area of the study. The specific objectives of this study are summarized as follows:

- To investigate effective data driven machine learning models based on the historical data.
- To propose the simple and effective machine learning approaches for prediction of days ahead water levels.
- To analysis the capabilities of proposed machine learning approaches using rainfalls and water levels time series data.

- To predict the water level according to geographical locations with the combination of time series hydro-metrological data.
- To generate the predicted water level time series suitable for the assessment of pre-warning and flood management.
- To evaluate the proposed model's accuracy with time series prediction.

1.6. Motivation

There are several different methods for prediction and recently flood prediction has been focused because of frequent occurrences. Generally, flood prediction model can be divided as the two approaches: physically based and data driven approaches [1].

Physically based approaches are fully distributed models and can increase the levels of model complexity. For prediction system, distributed models need much time to develop and appropriate expert is required to remark the result. Data driven machine learning model can find the relationship between the input features and the water levels while the physically based models aims to reproduce the hydrological process in a physically realistic.

Data driven model is useful for real time prediction in various areas with accurate predicted result in their respective fields and it is still open in research area. For real time prediction models, data driven approach is obviously developed and easy to build up. Our approach is to predict days ahead water levels for the flood pre-warning and disaster management of our area of study based on data driven machine learning model.

1.7. System Overview

Our system is used to predict the five days ahead water levels using the three different machine learning models and compared the accuracy of the models and make the prediction. We consider the upper stations and lower station water levels and rainfall data when we make prediction for lower station's days ahead water levels. The water levels and rainfall of upper stations along the distance can greatly effect on the lower station water level condition as the rivers are flowing from the northern part to southern part in our study region.

As shown in Figure 1.2, collected data from study regions are analysed, cleaned and preparation for data pre-processing. And then, the models are trained with the training dataset with our proposed machine learning approaches. After that, the model is tested with testing dataset and evaluated the model's accuracy. Finally, predicted water levels for five days ahead for each day is generated by using our proposed approaches with better accuracy.

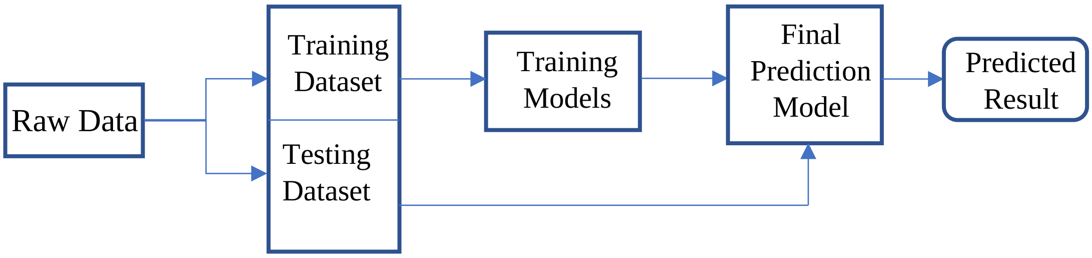

Figure 1.2. General Overview of the System

1.8. Thesis Organization

The whole structure of thesis is organized as follows: Chapter (1) highlights the background, problem statement, research objective, motivation and thesis organization. Chapter (2) provides the background knowledge, time series concepts and relationship between machine learning models and time series data and literature review on various types of machine learning models for weather prediction. Chapter (3) describes the methodology of the simple and efficient data driven machine learning approaches and experimental set up for next five days water level prediction. Chapter (4) implements the proposed approaches for water levels time series data using the proposed approaches. It is mainly stated that the step by step experimental procedure for training and testing over proposed machine learning approaches. Chapter (5) summarizes about the estimated results, progress conclusion from extraction of something from these results and future possible approaches.

CHAPTER 2
LITERATURE REVIEW

2.1.Introduction

In this section, background knowledge and information of the studies were gathered to implement water levels prediction system by reviewing some academic literature, studies from various sources such as books, online articles, journals and papers. All information and knowledge obtained from these studies was applied as the guidance to get the optimal solution. The important theoretical and methodological of the previous studies related to the water levels prediction for flood and disaster management are also reviewed to develop the next five days water level prediction on the study region.

There are several intensive researches and studies for weather data prediction for flood early warning and disaster management depending on the different locations. Various models and parameters are implemented in order to drive the optimal solution based on the underlying concepts and theories. Most of the research associated with water level prediction for flood and disaster management systems in small and large study area have already succeeded in finding the appropriate results. In this chapter, the author reviews some applications' background knowledge and some important concepts used for the research.

2.2. Overview

In today's globalization society, there is abundance of unstructured and structured data. Thus, in order to drive and handle large amount of data manually, several human work forces are needed and time consuming. In the middle of the twentieth centuries, machine learning included as the part of the Artificial Intelligence (AI) as more extensive knowledge. Machine Learning included the development of self-learning algorithms to gain knowledge from data to make optimize prediction [1].

The system was learnt with example data or past experience data to identify the data and apply this knowledge to make the prediction with performance criterion. It avoids knowledge intensive model building and reduces the reliance on expert knowledge as due to its nature solution to automation. Machining learning is concerned with obtaining data from knowledge. It uses the theory of statistic in building mathematical models. Machine Learning plays a key role in the fields of statistics, data mining, computer science and engineering.

In our work, we evaluated the effective of short range water levels prediction using three different machine learning models with three different real time datasets. Water levels predictions play a critical role in weather data prediction, flood warning and disaster management. There are several articles have been published according to three contextual behaviour: white box, black box and gray box models. White box model that tests internal structures or workings of application based on the physical equations and geometry. Black box (classical statistics and artificial intelligence technique) in which internal working of the application are not needed to be known as the closed box and the gray box model is the combination of different kinds of the models in which somewhat internal workings are needed to be known to improve their prediction performance [1][2].

Our water levels prediction approaches are characterized by the black box process in which physical equations and relationships are not required to be known to make day heads water levels prediction. They adjust the models in order to fit the data by automatically extracting the existing information from past experience data. Thus, we applied the data driven machine learning approach that is quickly developed and very useful for accurate water level prediction based on time series data.

2.3. Time series analysis

Time series is an ordered sequence of observations at equally space time intervals. The idea of time series analysis is that the future will continue with the past evolution. Data points over time may have the internal structure such as auto correlation, trend, and seasonal variation that differ by the shape of the line which best fits the observed data. Prediction with time series analysis is based on the following procedures:

- Select the model
- Model parameter optimization
- Estimate the performance and make the prediction.

2.4.Related Works

There are several different machine learning models that have been developed in recent years with different objectives and locations for time series weather data prediction.

Thanh-Tung Nguyen et al [3] predicted water levels of Mekong River by using three machine learning models LASSO, Random Forest and Support Vector Regression (SVR) and compare the prediction performance and choose the least mean absolute error (MAE) result of prediction model is 0.486 meter(m) for area of study on Mekong River.

Felan Carlo C.Garcia et.al [4] has applied Random Forest machine learning algorithm based on data taken from two different stations to predict the water levels on Cagayan River basin in Philippines. The result shows the good prediction accuracy and recommend for the other stations situated on major river basins across the Philippines.

Kitsuchart Pasupa et.al [5] proposed water levels forecasting approaches in Chao Phyra river in Thailand using the different machine learning algorithms (Linear regression, Kernel regression, Support vector regression, K-Nearest neighbor, and Random Forest) and presented the best one and showed the proposed approaches prediction performance are better than the current approach used by the Royal Thai Navy.

Mohammad Sajjad Khan and Paulin Coulibalay [6] investigated that Support Vector Machine has promising for the lake water level prediction with reasonable accuracy by comparing with widely used neutral network model (Multilayer perceptron (MLP) and seasonal autoregressive model (SAR)).

C. Damle [7] has applied the Time series data mining to the area of flood forecasting. High, medium and low flood occurrences from three observation stations were investigated and the calculated the prediction accuracy in terms of positive prediction accuracy (PPA) and Correct Prediction Accuracy (CPP).

Jun-He Yang [8] et.al proposed the time series forecasting model for water level forecasting in reservoir by comparing the five machine learning regression methods (KNN, RBF network, K star, KNN, Random forest) with key variable selection such as reservoir-IN, reservoir-OUT, pressure, rainfall and relative humidity. The result shows that the random forest with /without variable selection has better forecasting performance and feasible for forecasting water level of Shimen reservoir in Taiwan.

W.T.Zaw et.al [9] implemented the prediction model for rainfall in Myanmar using the multiple linear regression model based on 15 predictors. According to the several experiment results, the predicted rainfall amount is close to the observed value.

Zan.C et.al [10] has developed the rainfall forecasting model using Hidden Markov model for the rainy season in Myanmar based on the time series data of 12 rainfall observations stations and shows the close agreement of actual and predicted results.

The rainfall prediction model has been developed by using multiple linear regression (MLR) model and computed the Pearson coefficient for five years by M.Kannan et al.[11]. Result shows the approximate value of observed and predicted value but does not show the enough accuracy.

M.A.kulkarni et.al [12] proposed the efficient approach for wind speed direction using different statistical regression and neural network and calculated the root mean square error (RMSE) in prediction zonal component of wind speed for all months.

Abhishek Agrawal et.al [13] has applied Artificial Neural Network the multilayer perceptron to forecast the maximum and minimum temperature based on the historical time series data of India Metrology department and computed the mean square error (MSE) to confirm the prediction performance and make the temperature prediction.

Afiq Hipni et.al [14] proved that Support Vector Machine (SVM) is excellent model for daily forecasting of dam water level of Klang reservoir in Malaysia compared with Adaptive Neuro Fuzzy Inference System (ANFIS) and show all mean root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) of SVM is a superior model to ANFIS.

Z.Jan et.al [15] utilized data mining algorithm: k-Nearest Neighbour (KNN) to predict the weather data using historical data and proved that how to develop a system that use numeric historical data (instead of geographical location) to forecast the climate of a specific region, city or country months in advance.

The advantages of using the time series machine learning approaches for weather data estimation can be seen in these related works. However, they suffer from their own constraints and have uncertainties in prediction depending on regional and local basic. Most of the existing researches performed process on the one machine learning algorithm which is assumed to be the best algorithm for their system or lack of consideration of valid and key features and small numbers of real workload traces. Therefore, it is impossible to choose the best algorithm for the system unless different machine learning algorithms are trained on different real-time data. It is also important to be sure for creating the better prediction on future workloads demands with appropriate machine learning approaches. In this work, three different machine learning approaches are analyzed with three different real-time series data and valid features that related with ahead of the time water levels mechanism are utilized.

2.5. Summary

In this section, we mainly stated that the essential concepts, background knowledge and analysis to the rest of the thesis. We've studied other literature review of various weather data prediction and water levels prediction approaches that are concerned to our research topic.

CHAPTER 3
METHODOLOGY

The research investigates the simple, flexible and effective prediction methods of accurate water levels for days ahead using machine learning data driven models based on time series data. In this section, types of machine learning, machine learning prediction models in this work, their properties, procedures, the proposed approaches and how we use these approaches to predict the water levels will be described.

3.1. Introduction to Machine Learning System

Machine learning is one of the most popular and evolved as the subfields of artificial intelligence that involved the development of self-learning algorithms to gain knowledge from that data in order to make the predictions. In our age, data are abundance; we can turn this abundance data into knowledge by using self-learning algorithms from the field of machine learning.

Machine learning provides the more efficient alternative for capturing the knowledge in data to gradually increase the prediction models' performance and make data driven decisions. Many famous companies such as Google, Amazon, IBM, Apple, Facebook etc., invest significantly in machine learning research and applications with several objectives. It becomes important not only in computer sciences fields but also in daily lives [16]. Machine learning field is quite vast and developing rapidly because it can apply in various areas such as weather data prediction, email spam filter, voice recognition software , medical data prediction , reliable web search engine etc.,.

In 1997, Mitchel form Carnegie Mellon University, well described the definition of the machine learning that has proven more useful and suitable to engineering types. That is "A computer program is said to learn from Experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." [17].

Thus, in our work, it means that if we want to predict the next five days water level at a location (Task T), we can run it though a machine learning algorithm with our historical data (Experience E) and if it is successfully "learned", then it will do better prediction for future water levels with performance measurement (P).

3.2 Different types of Machine Learning

There are three different types of machine learning: supervised learning, unsupervised learning and reinforcement learning. In our work, we applied the supervised machine learning regression models for prediction water levels for our study region.

3.2.1 Supervised Learning

In supervised learning, the goal is to make the prediction from labelled training data that allows us for making prediction about unseen and unknown data as illustrated in Figure 3.1. There are two subcategories in supervised learning: Classification and regression [18].

Classification is a subcategory of supervised learning in which we can predict the categorical of class labels of new instances based on past observations. These class labels are discrete, unordered values that can be understood as the group memberships of the instances. Regression analysis is another subcategory of supervised machine learning to predict the continuous outcomes. In regression, we provide the number of predictor (explanatory) variables and continuous response variable (outcome) and then find the relationship between those variables and predict the outcome. Thus, in supervised learning, we can predict the output with learning algorithm based on input trained data with labelled.

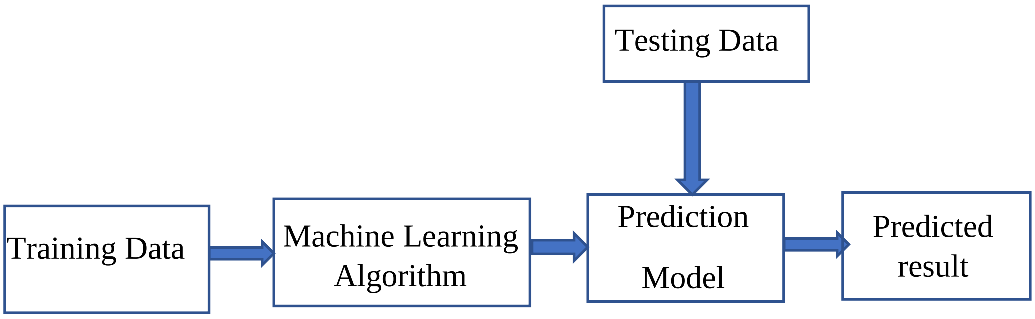

Figure 3.1: Prediction with supervised learning [16]

3.2. 2 Unsupervised Learning

In unsupervised learning, we can learn the algorithm to inherent structure with the unlabelled data. . In supervised learning, we know about the right answer before ahead when we trained the models. But, in unsupervised learning, we have to deal with unlabelled data or unknown structured data. To extract the meaningful information, we can explore the structure of the data without the guidance of known outcome variable or reward function Thus, in the unsupervised learning, the goal is to model the underlying structure or distribution in the data in order to learn about the data [18], [19].

3.2.3 Reinforcement Learning

Reinforcement Learning is another type of machine learning .The goal of reinforcement learning is to develop the system (agent) that improves its performance based on the interaction with the environment. Simple reward feedback is needed to learn its behaviours based on feedback from the environment.

We can think that reinforcement learning as a related field of supervised learning since the current state of the environment information typically includes a so called reward signal. However, this feedback is not the correct ground truth label or value in reinforcement learning but it is a measurement of reward function for how well the action. An agent can use the reinforcement learning through the interaction with the environment to learn a series of actions. We define the measure of reward for particular actions by the agent and reinforcement learning allows the agent or machine to learn its behaviour based on the feedback from the environment [16] [20].

3.3 K-nearest Neighbor (KNN)

In general, the consideration for the methods of data driven modeling is building model with the available training data (learning process) and then were put together to operate when numerical classification and prediction was taking places. These methods are sometimes referred to as eager learning that is they eager to build the model first. In eager learning, the training set is pre-classified and all objects in the training set are clustered with regarding to their neighbors such as artificial neural network [21].

However, there is a group of learning methods when the models are actually constructed, and it is called the instance-based learning. The instance-based learning simply stores the data and postpones the generalization (building the model) until the new instance classification or prediction is made. In instance based learning, when the object is input to the algorithm, the distance is calculated such as K- nearest neighbors.

K-nearest neighbor is (KNN) is the non-parametric instance based method for classification and regression. Machine learning algorithm can be classified as: parametric and non-parametric. The parametric algorithm has a fixed number of parameters and computationally faster. But, it makes stronger assumption on data and work well if the assumption turns out to be correct otherwise, its performance will be bad. But non-parametric algorithm uses the flexible number of parameters that can increase as the learning with more data. An example of non-parametric algorithm is KNN. In KNN, the input of the model consists of the K- closest training samples and the output will depend on the applying of KNN for classification or regression [22].

In KNN classification, the output is the class membership. It is method to classify the objects depend on the closest matching entries (the most common class among its K nearest neighbors) obtained from training data. In KNN regression case, the output will be the property value of the object. This value will be the average value of its K- nearest neighbors.

3.3.1 K-nearest Neighbor (KNN) Algorithm

K-nearest Neighbor (KNN) algorithm is the way to classify the objects based on closest matching neighbor’s entries from learning data. KNN is non- parametric method that can be used for classification regression. KNN is a type of lazy learner not because of its simplicity but because of its doesn’t learn the discriminative function from the training data but memorize the training data instead [16][23]. KNN algorithm can directly search through the all training samples by calculating the distance between testing data and all training sample to identity its nearest neighbors when the classification is performed and then produce the output of classifier. Training samples have various attributes that representing its characteristics and modeled as multi-dimensional space representation [23][24].

3.3.2 K-NN Classification

Classification is the identification of the characteristics of the class that specify to which class each record belongs. In classification algorithm, the dataset is divided into training dataset and testing dataset. Training dataset is used for learning the model and testing dataset is used to evaluate and compute for the model accuracy. In KNN classification, K is user defined and test sample is labeled by the most common class among the K closest training samples as show in Figure 3.2. Particularly, the distance between two data points is defined and calculated by the distance function [16] [25]. KNN algorithm is straight forward and summarized by the following steps.

- Assigning the number of K and calculating the distance matrix with the suitable distance function.
- Find the K neighbors of the test sample that we want to classify.
- Assign the class label by majority vote.

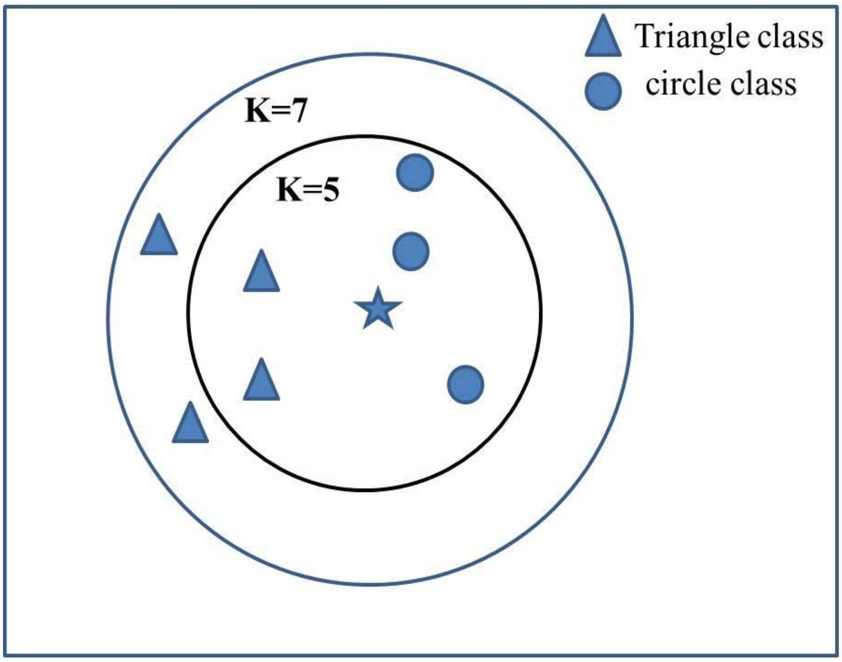

Figure 3.2. K- Nearest Neighbours illustration

In figure, we classify the new test instance "star" in two dimensional spaces and there are two classes – first class of circle and second class of triangle, so the output will be "circle" or "triangle". The new instances also called new query will classify based on the majority voting among its nearest neighbours. If we apply 1-NN method, we will consider just only one training instance and the output will be "circle". If we assign the number of K=7, the classifier output will be "triangle" because there are 3 circle and 4 triangle inside the circle as the 7-NN method. If we use 5-NN method, the output will be "circle". KNN algorithm find the K samples in the training datasets that are closest to the point that we want to classify based on the chosen distance matrix. Then, the class label of new data point is determined by majority of the vote among its K nearest neighbours.

3.3.3 K-NN regression

Regression is the prediction of the continuous value based on value of other variables on the basic of linear and nonlinear dependence model. K- Nearest Neighbours algorithm is easily adopted to predict the continuous value in KNN regression. The algorithm stores all variable cases and predicts the numerical target data based on distance functions that measure the similarity. One such algorithm uses the weight average of the K nearest neighbours, weighted by the inverse of their distance. A widely used distance metric for continuous variable is Euclidean distance function [22] [26] [27]. The algorithm worked as follows:

- Compute the distance function based on suitable distance calculating methods from the testing instance to the labeled training instances.
- Order the labeled training examples by increasing distance.
- Find the optimal number of K nearest neighbors heuristically based on root mean square error (RMSE) that can be done by cross validation.
- Calculate the inverse distance weighted average of K nearest multivariate neighbors.

3.3.4 K-NN Distance Functions

KNN use the same distance functions in classification and regression. There are various distance functions to measure the distance between point X and Y in a feature space. Some of the mostly used distance functions are as follows [28] [26]:

Let consider $X = (x_1 , x_2 , x_3 , \ldots\ldots , x_n)$ and $Y = (y_1 , y, y_3 , \ldots\ldots , y_n)$ where n is the feature space dimensionality. Euclidean distance metric to calculate the distance between X and Y is used by

$$d(X,Y) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{m}} \qquad (1)$$

Other distance functions (Manhattan and Minkowsky) also available in KNN are as follows:

$$Manh.d(X,Y) = |x_i \text{ - } y_i| \qquad (2)$$

$$Min.d(X,Y) = \left(\sum_{i=1}^{m} |x_i - y_i|^r\right)^{1/r} \qquad (3)$$

These above-mentioned distance functions are only useful for the continuous variables. Among the various distance functions, Euclidean distance function is the widely used one. Euclidean distance function performs well over the numerical datasets according to the literature [28]. In our work, we used the rainfalls and water levels of current and previous two days data and make the prediction for five days ahead water levels and all the data are numerical data so that we applied the Euclidean distance function in the proposed K-NN regression approach.

3.3.5 Parameter Selections

In K-NN, parameter selection is the important one and the different number of K can vary the accuracy of the model. The best choice of K can vary on the dataset and it can be done by hyperparameter optimization. In a machine learning models, hyper parameter which is the parameter whose value is set before the learning process begins. Different hyperparameters are required in different model training algorithms.

By tuning the hyperparameters of the training algorithm can measure how much the performance can be improved by tuning it. Hyperparameters optimization is to find tuple of the hyper parameters that generate optimal models which can minimize the predefined loss function on given in independent data. The aim of the function is to take the tuple of hyper parameters and return the associated loss [29] [30]. Cross validation can be used for this generalization process. In KNN, K is the hyper parameter and can be tuned to define the loss function of the training model and the choice of the best value of K can be done by cross validation. For inverse distance weighted value of K nearest neighbors can be calculated as follows [29]:

$$\omega_i = {}^{1}/_{d_i} \tag{4}$$

3.3.6. Pros and Cons of KNN

KNN algorithm is simple to explain and understand and it can be useful for both classification and regression. There is no assumption about the underlying data distribution that is pretty useful because most of the real time data doesn’t follow the typical theoretical assumptions. Thus, accuracy can be pretty high in some real time data classifications and regression.

The computational cost and memory requirement of KNN will be high because algorithm stores all the data training data. Prediction stage can be slow when we use the large amount of data. The accuracy of KNN is sensitive to the irrelevant features and range of data.

3.4 Introduction of Support Vector Regression

Support Vector Regression a valid supervised learning algorithm and extension of the classic Support Vector Machine algorithm that firmly grounded on the framework of statistical learning theories [31][32]. Support Vector Machine algorithm is can be considered as the extended multilayer perceptron algorithm that minimizes the classification errors. However, in SVMs, to maximize the margin that is the distance between separating hyperplane which referred as support vectors as illustrated in Figure 3.3[16].

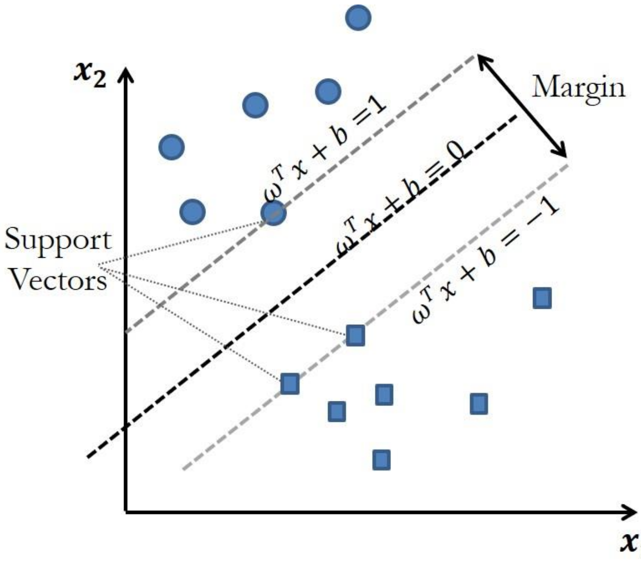

Figure 3.3. Illustration of Support vector machine (SVM)

Support vector machine (SVM) is widely used in many applications and become popular in machine learning [33] [34]. Initially, Support vector machine was designed to provide the classification problems but later their principle could be adopted to solve the other problems such as detection or regression. The major objective of SVM is to improve the performance in terms of generalization errors when testing input are provided. For improving the generalization error in SVM, the structural risk minimization principle is adopted which controls both the error on the training dataset and the penalty term for controlling the model’s complexity. If the functions of the model have higher complexity, the higher risk of overfitting will occur that losing the generalization ability [2] [35] [36].

In our work, we investigated the supervised learning techniques that can generalize well on the unseen data to predict the five days ahead water level based on continuous time series data. Thus, we applied the support vector regression as another prediction model which is the extension of support vector machine and well based on statistical learning theory.

3.4.1 Why Support Vector Regression?

The Support Vector Regression (SVR) is proposed by Vapnik [32] and the basic idea is to find the linear regression function $f(x)$ in a high dimensional feature space. Suppose our training dataset T is composed of m samples and n features.

$$T = \{(x_1, y_2), (x_2, y_2), \ldots\ldots\ldots, (x_m, y_m)\} \qquad (5)$$

where $x_i \in R^n$ is the input vector of m observed samples and $y_i \in R^1$ is the corresponding target values for i=1,2,3,…..,m. The basic support vector machine algorithm only provides the linear function with possible simplest hyperplane that fit to our dataset as close as possible. In that case, loss function has to be defined to measure the performance of the fitting and this measurement can be regarded with noise. Thus, we applied the ε-insentive loss function in our work. The goal ε-SV regression is to perform the linear regression to find the function $f(x)$ that has at most predefined deviation ε from the obtained targets, y_i for all training data T[2] [36]. In other words, we neglect the error smaller than predefined threshold of ε and if the error higher than the threshold, we will penalize it. Our target is to find the function $f(x)$ with the smallest ε and the generic SVR takes the form.

$$f(x) = \omega^T \emptyset(x) + b \qquad (6)$$

where $\omega \epsilon R^n$ and $b \epsilon R$, $\emptyset(x)$ denotes the non-linear transformation function to map the data point into higher dimensional feature space where the linear regression is performed. The idea is to determine the optimal function $f(x)$ that can estimate future values accurately by minimizing the loss function.

$$\text{minimize } \frac{1}{2}\left||\omega|\right|^2$$

$$\text{subject to } |y_i - (\omega^T x_i + b)| \leq \varepsilon \qquad (7)$$

Such assumption in equation (7), function $f(x)$ exists that approximates all pairs $(x_{i,}y_i)$ with ε precision. But sometimes, there is a case to allow some errors. Analogously to loss function, Cortex and Vapnik (1995) [31] proposed the soft margin loss function. According to this function, we can define slack variables ξ and ξ^* to loss function to penalize the points beyond the predefined variable ε through the cost parameters C. The SVR solution of convex quadratic optimization problem [6] with minimizing the loss function is as follows:

$$\text{minimizing} \qquad \frac{1}{2}\left||\omega|\right|^2 + C \sum_{i=1}^{m}(\xi + \xi^*)$$

$$\text{subject to} \qquad y_i - (\omega x_i + b) \leq \varepsilon + \xi \qquad (8)$$

$$\xi,\ \xi^{*}\ \geq 0$$

where the constant C is regularization parameter and C > 0 determines penalties to estimate the error. Thus, optimization problem can be expressed in its optimal object function by dual set of variables [7].

$$L = \frac{1}{2}\left|\left|\omega\right|\right|^{2}\ +\ C\ \sum_{i=1}^{m}(\ \xi + \xi^{*}) - \sum_{i=1}^{m}(\eta_i\xi_i +\ \eta_i^{*}\xi_i^{*})$$
$$-\sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_I - y_i + \omega^{T}x_i + b)$$
$$-\sum_{i=1}^{m}\alpha_i^{*}(\varepsilon + \xi_i^{*} + y_i - \omega^{T}x_i - b) \qquad (9)$$

where L is the Lagrangian and $\eta_i\ ,\eta_i^{*},\xi_i,\xi_i^{*}$ are Lagrange multipliers.

The first partial derivatives of L with respect to the primal variables $(\omega, b, \xi_i,$ $\xi_i^{*}\ )$ can be substituting in equation (9) and we obtain the dual formulation of non-linear SVR solution with ε loss function (13).

$$\partial_b\, L = 0 => \sum_{i=1}^{m}(\ \alpha_i^{*} - \alpha_i) = 0 \qquad (10)$$

$$\partial_\omega\, L = 0 => \omega - \sum_{i=1}^{m}(\ \alpha_i - \alpha_i^{*})x_i = 0 \qquad (11)$$

$$\partial_{\xi}^{(*)}L = 0 => \mathrm{C}\alpha^{(*)} -\ \eta^{(*)}{=}0 \qquad (12)$$

maximizing $-\ \frac{1}{2}\sum_{i,j=1}^{m}\ \ (\alpha_i +\ \alpha_i^{*})\ (\ \alpha_j - \alpha_j^{*})\phi(x_{i,} - x_j)$

$$-\varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^{*}) + \sum_{i=1}^{m}y_i(\alpha_i - \alpha_i^{*}) \qquad (13)$$

subject to $\sum_{i=1}^{m}(\alpha_i + \alpha_i^{*}){=}0$ and $\alpha_i\ ,\alpha_i^{*}\epsilon\ (0, C), i, j\ = 1,2, \ldots m.\ \alpha_i^{(*)}$ refers to α_i^{*} and α_i. Lagrange multiplier pair (α_i,α_i^{*}) is concerned with each point of the dataset. When the optimal hyperplane has been found in the training time of support vector machine, only the point that lies outside the predefined error threshold ε in which one of their Lagrange multiplier is not equal to zero (i.e $\ \alpha_i > 0\ \ or\ \ \alpha_i^{*} > 0)$ are called the support vectors. Support Vector Machine only rely on these points as shown in Figure 3.4 [36].

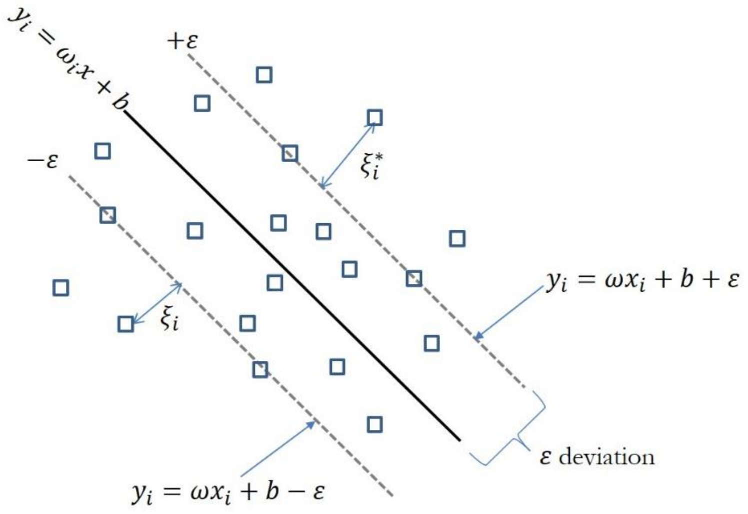

Figure 3.4. Illustration of ε tude with slack variable and data point

In equation (13), the dual variables η_i , η_i^* are already eliminated through condition (12) that can be reformulated as $\eta^{(*)}$= C- α_i^* and rewriting the solution of (11) is

$$\omega = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i \tag{14}$$

Thus $f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)(x_i, x) + b$ (15)

This is called support vector expansion that is ω can be completely described as a linear combination of the training patterns x_i .The complexity of the function representation by support vectors (SVs) is depend only on the numbers of SVs and not depend on the dimensionality of the input space X.

3.4.2 Kernel Functions

Support Vector Regression (SVR) uses the kernel function to map the original data into the higher dimensional space where the linear function exists. There are different kernels functions detailed in [36] and choosing the appropriate kernel type and kernel parameter of SVR is based on the application domain knowledge and distribution of the training data. If the measured points are close enough having the small value of ε, it may be impossible to find the hyperplane in the input space. However, the kernel and its transformation function can be used to arrange the input data in the feature space with a different way. The optimal hyperplane could be found easily in this new feature space than the input space [2].

In our experiment, we applied the Radial Basic Kernel Function (RBF) that most widely used and very flexible in machine learning technique. The RBF kernel can be defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) \tag{16}$$

where, γ is a scaling parameter.

3.4.3 Parameter Selection

Support Vector Regression (SVR) estimation accuracy (generalization performance) mainly depends on varying the parameters of the model. In SVR, the free parameters C, ε and kernel parameter γ are strong influence on the prediction performance. Existing SVR parameters implementation is user defined and the optimal parameters selection and model complexity is depending on all these three hyper parameters. In SVR, these parameters can be tuned to get the better prediction performance [36]. In our experiment, we tuned the hyperparameters of SVR to estimate our prediction model accuracy by using cross validation.

3.4.4 Pros and Cons of SVR

SVR is quite robust and flexible based on well theoretically basic for non-linearly separable input data and can evaluate more relevant data in a suitable way. The judgment of optimal linearization function of nonlinear input data does not need the expertise assessment since SVMs linearize the data by means of kernel transformation. It can be generalized and the choice of the hyperparameters has strong influence on the model performance. These parameters can be tuned and free to choose.

In SVR, the way of estimation of parameters needed to be considered and parameters estimation can be computationally intense. On the other hands, choosing of the kernel functions also have to consider for the model performance improvement.

3.5. Multiple Linear Regressions

Regression is a statistical analysis used to predict the value of dependent variable (the value we wish to predict) based on at least one independent variable (the values we used to predict the dependent variable). Regression model predict one dependent variable by using one independent variable is called linear regression. Linear regression is as statistical function that analyses the relationship between dependent variable and independent variable and makes the prediction of one variable based on another variable. Simple linear regression function is as follows:

$$y = \omega_0 + \omega_1 x + \varepsilon \tag{17}$$

where ω_0 is the estimation of the regression (interception), ω_1 is the estimation of the regression slope of independent variable x and ε is the random error of x as illustrated in Figure 3.5. Linear Regression provides the line that closest to the data by finding the slope and interception that define the line and minimize the regression errors. That best fitting line is called regression line and the vertical lines from the regression line to the sample points are called offsets or residuals (the error of our prediction)[16].

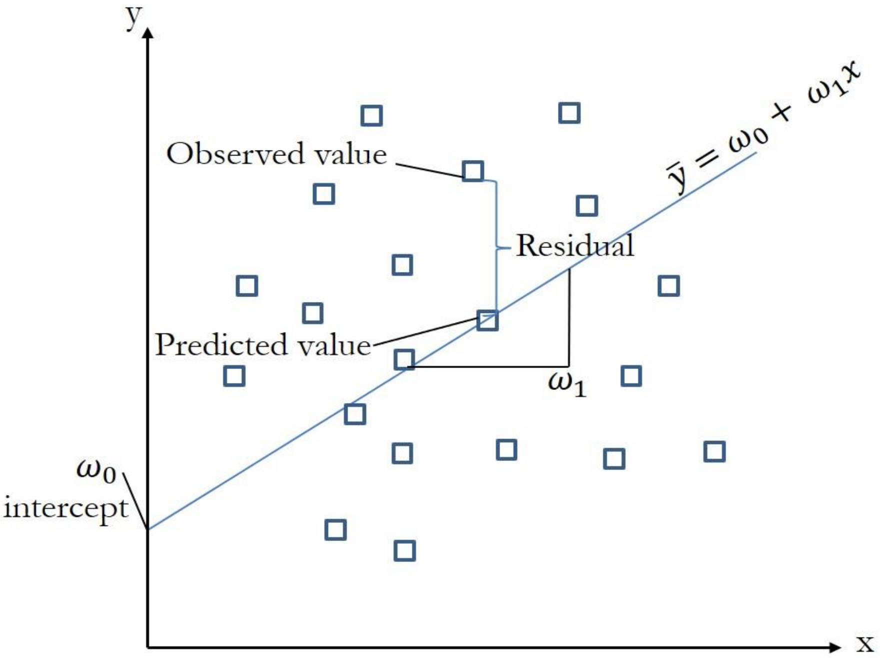

Figure 3.5. Illustration of Linear Regression Function

However, the relationship of many data does not follow the straight line and we can apply the multiple linear regression function in that case. The regression model with one dependent variable and more than one independent variable is known as the multiple linear regression which generalize the linear regression model to multiple independent variables. The estimation of dependent variable y with more than one independent variables $\{x_1, x_2, x_3, \ldots., x_m\}$ and the relationship are as follows:

$$y = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n \qquad (18)$$

Hence, ω_0 is the y axis intercept with x_0 =1. In our work, we applied multiple linear regression to predict the days ahead water levels (dependent variable) using more than one independent variables (historical rainfalls and water levels data of two stations). According to the equation (19), y represents the predicted output and the $\{x_1, x_2, x_3, \ldots., x_m\}$ represents the rainfalls and water levels data of upper and lower stations. The unknown coefficients $\omega_0, \omega_1, \omega_2, \ldots\ldots, \omega_n$ can be solved through the multiple linear regression approach by reducing the sum of the square errors (SSE) as shown in equation (19) [39] [40][41][42].

$$SSE = \sum_{i=1}^{m} (y_i - \overline{y_i})^2 \qquad (19)$$

where y_i is observed value and $\overline{y_\imath}$ is the actual value.

In our experiment, we've applied multiple linear regression approach to calculate our prediction model accuracy and to compare the accuracy with others two proposed approaches. Multiple Linear Regression approach is currently using as the data driven based water level forecasting approach in our study region, Myanmar. The experimental results of five days ahead water levels with the data of water levels observation stations of our study areas in Myanmar by using the K-nearest neighbours, support vector regression and multiple linear regression approaches are described in next chapter.

CHAPTER 4

Experiment Implementation, Evaluation and Result Discussion

The system is managed to predict the five days ahead of the water levels using real time series data depending on our areas of study by using our proposed machine learning approaches. In this chapter, our experiment implementation procedures, evaluation the accuracy each proposed model on the areas of study and prediction result for each day will be described.

4.1. Building Machine Learning Prediction Model

We build our machine learning prediction model using time series data with the following procedures: pre-processing our real time water levels and rainfalls data, learning the prediction models and then evaluation the model accuracy and finally, making the prediction for five days ahead for each day. The overview of our prediction model of each proposed approach is as shown in Figure 4.1.

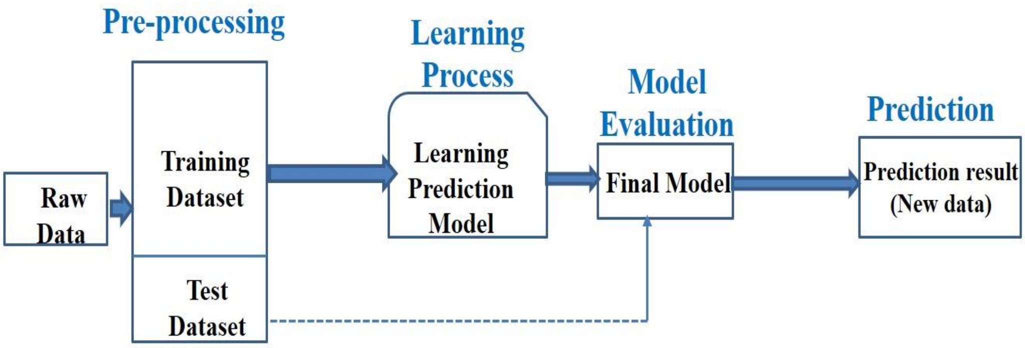

Figure 4.1. Machine learning prediction model

In data pre-processing stage, we collect the time series water levels and rainfalls data from five weather observation stations (three stations on Ayeyarwady river and two stations on Bago river) in Myanmar. We investigated the important features and analyse the data to be easy to understand and useful for preparing the datasets to predict the next five days water levels for our study areas. In learning step, we applied three different machine learning models to train the model with training dataset and making cross validation and parameters optimization to estimate our proposed models accuracy with performance metric. In the evaluation step, we get the final model that we applied to make prediction on the new data. In this step, our prediction models are tested with new examples of input data to predict our expected output. After testing the proposed models with new data and calculate the performance with the evaluation metrics and then produced the predicted results.

4.2. Cross Validation

In machine learning, there are two useful cross validation techniques: holdout cross validation and k-folds cross validation. These can provide to obtain the reliable estimation of the model's generalization error which is how well the model performs on the unseen data (test data).

4.2.1. Holdout Method

In holdout method, the data was separated into training set, validation set and test set to get the better model and hyperparameters tuning to improve the model prediction performance. In this method, after training the model using the different parameters values, the validation set was used in order to repeatedly evaluate the performance of the model. If the tuning of the parameters is acceptable, the generalization error of the model was estimated with test set. The holdout method strategy is as described in Figure 4.2. The disadvantage of the hold out method is the performance estimation of the model is sensitive to the how partition the training set into training and validation subsets where the estimation can vary for different samples of data. [16]

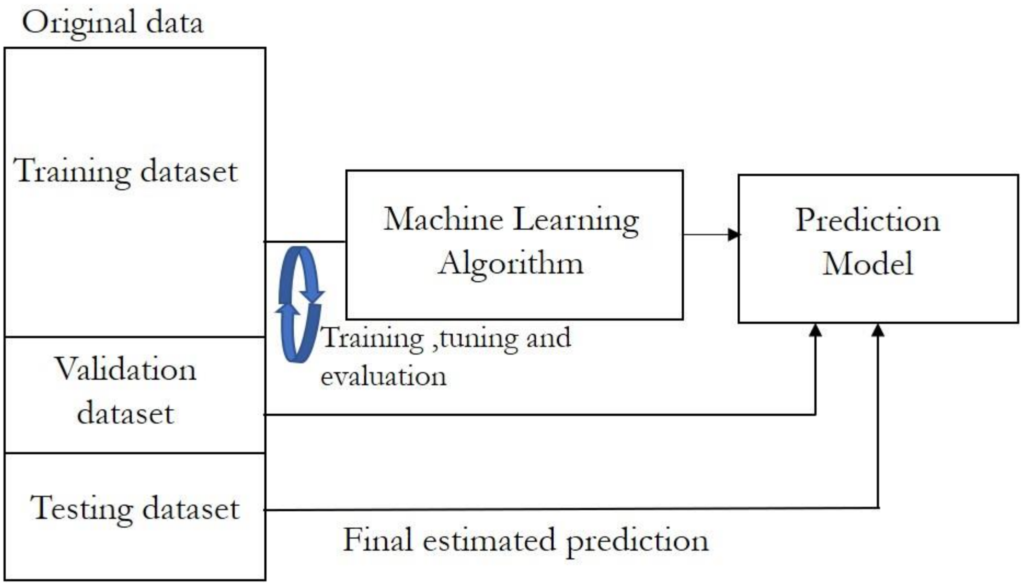
Original data
Training dataset
Validation dataset
Testing dataset
Machine Learning Algorithm
Prediction Model
Training ,tuning and evaluation
Final estimated prediction

Figure 4.2. Illustration of holdout strategy [16]

4.2.2. K-fold Cross Validation Method

K-fold Cross Validation is more robust technique for estimation of performance by repeating the holdout method k-times on k subsets of the training data. In this method, the training set was randomly split into k folds without replacement where k-1 folds are used for training set and one-fold is used for test set and repeated times for performance estimation. Typically, k- folds cross validation can provide the optimal hyperparameters value that satisfying the generalization performance of the model. After obtaining the optimal hyperparameters values, the model will retrain on the complete training set and obtain the final prediction performance using the unseen data (test data).

The advantage of the k-fold cross validation that each sample point will be the part of the training and test data exactly once which yields a lower variance estimate of the model performance than the hold out method. The illustration of k-folds cross validation with k=10 is described in Figure 4.3. Training set is divided into 10 folds and 9 folds are used for the training set and one-fold is used for test set for the model evaluation during the 10 times iterations. The estimated average performance of the model E is calculated by using the estimated performance E_i for each fold. In most applications, the standard k value 10 is generally a reasonable choice [16] [43] [44].

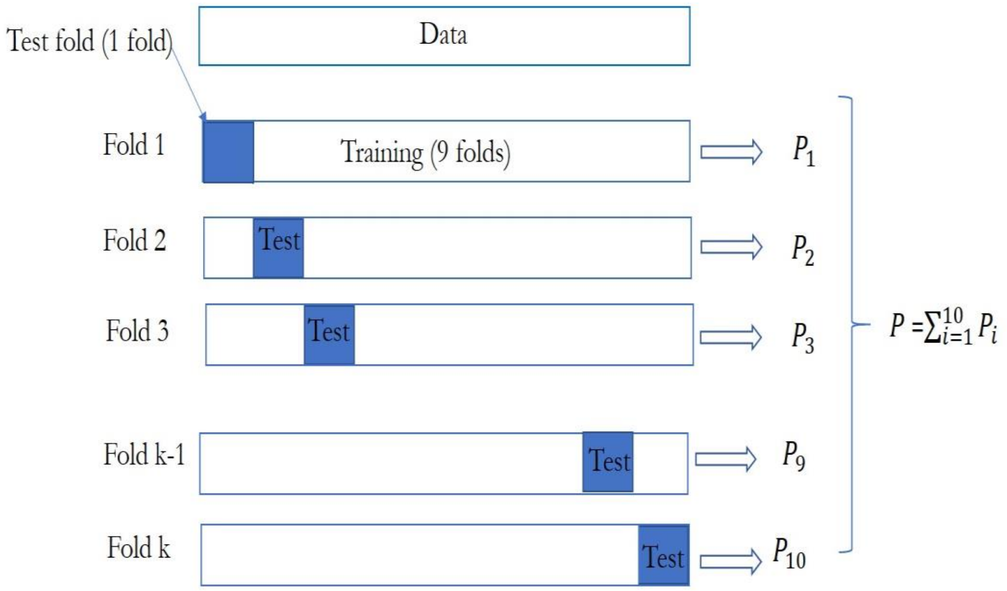

Figure 4.3. Illustration of k-fold strategy

In our work, we investigated the optimal hyperparameters to get the reliable performance of the models and better estimation with k-folds cross validation where k value is 10.

4.3. Model Evaluation

We investigated the datasets by applying different machine learning techniques to evaluate the prediction accuracy in terms of strength and weakness using correlation coefficient, mean absolute error and root mean square error. In this experimental study, three machine learning models KNN (K-nearest Neighbour), Support Vector regression (SVR), multiple linear regression were implemented. In this work, we model water level prediction, so we focus on mean absolute error (MAE). The acceptable error rate of upper Ayeyarwady river is (0.5~0.7) meters. The prediction accuracy can be calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}[|O_i - P_i| \qquad (20)$$

Correlation coefficient (CE) that measures the linear relationship between the actual and predicted values and root mean square error (RMSE) can be used to evaluate the model performance.

$$\text{CE}=1-\frac{\sum_{i=1}^{N}[O_i-P_i]^2}{\sum_{i=1}^{N}[O_i-\bar{O}_i]^2} \qquad (21)$$

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}[O_i - P_i]^2} \qquad (22)$$

In these equations, N is the total number of instances, O_i and P_i illustrates observed water levels and predicted water levels and $\bar{O}_i$ is the mean value of observed water levels at i^{th} time respectively.

4.4. Data Used

We applied three different machine learning models (KNN, SVR, and LR) to predict the five days ahead water levels for our study areas (Ayeyarwady river and Bago river). In Ayeyarwady river, we've collected the weather data (water levels and rainfalls) of three metrology and hydrology observation stations: namely Myitkyinar, Bhamo, Mandalay and two stations (Zaungtu and Bago) of Bago river from Department of Metrology and Hydrology in Myanmar.

4.5. Experimental result of Bhamo station

We used about 3,000 data instances (from year 2009 to 2016) and 9 features of Myitkyinar station (station-1) and Bhamo station (station-2). Myitkyinar station is the upper station and Bhamo is the lower one. Thus, we make the prediction of five days ahead water levels for lower stations Bhamo (station-2) by using the current and previous two days water levels and rainfalls from upper station-1 (Myitkyinar) and lower station-2 (Bhamo). Because the major flow of rivers in our study region is from the northern part to the southern part and the upper station's water levels is strongly influence on the lower station's water levels. The dataset with input and output features are as shown in Table 4.1 and their inputs and output relationship are described in equation (22).

$$S_2\ \text{WL}\ (\text{t}+5) = \int(S_1\ \text{WL (t)}, S_1\text{WL (t-1)}, S_1\ \text{WL (t-2)}, S_1\ \text{RF (t)}, S_2\ \text{WL (t)}, \\ S_2\ \text{WL (t-1)}, S_2\ \text{WL (t-2)}, S_2\ \text{RF (t)}, S_2\ \text{RF (t)}) \qquad (22)$$

FEATURE	DESCRIPTION
S_1 WL (T)	Current water level of myitkyinar (station-1)
S_1 WL (T-1)	Previous water level of Myitkyinar (station-1)
S_1 WL (T-2)	Previous two days water level of Myitkyinar (station-1)
S_1 RF (T)	Current rainfall data of Myitkyinar (station-1)
S_2 WL (T)	Current water level of Bhamo (station-2)
S_2 WL (T-1)	Previous water level of Bhamo (station-2)
S_2 WL (T-2)	Previous two days water level of Bhamo (station-2)
S_2 RF (T)	Current rainfall data of Bhamo (station-2)
S_2 WL (T+5)	Five days ahead water levels of Bhamo (station-2)

Table 4.1. Different features and their description

4.5.1. Model Parameter Estimation and Evaluation with K-NN

K-NN model was applied to predict the water levels for five days ahead for Bhamo station (station-2) in the upper part of Myanmar. The collected data from Myitkyinar station and Bhamo station are randomly separated 2/3 and 1/3 of the data instances for training and test dataset for the water level prediction of Bhamo station. Choosing the optimal parameters, training, validation and testing of the K-NN have been performed as the following steps:

- Determining the dependent variable and independent variables.
- Randomly split the collected data into training, validation and testing set.
- Investigating the estimated optimal parameter for the K-NN model with different value of K.
- The model with optimal value of K (less RMSE) is selected and evaluate the model with 10 folds cross validation.
- Evaluating the performance of the model with testing dataset.
- Making the prediction of five days ahead water levels for each day.

For investigating for the optimal K value, the training process has been analysed several times with different value of K. For Bhamo staion, we tuned the K value from the range (1,2,..,500) and selected the optimal value of K= 12 with less

RMSE (error) =0.7465. The some resulting RMSE for different value of K can be seen in Table 4.2.

Parameter	RMSE
K=1	0.9397
K=2	0.8328
K=3 ⋮	0.8085
K=10	0.7495
K=11	0.7497
K=12	0.7465
K=13 ⋮	0.7508
K=50 ⋮	0.7657
K=100	0.7842
K=200 ⋮	0.8153
K=500	0.8946

Table 4.2.K parameter adjusting of Bhamo station

As mentioned before, K-NN with 10-fold- cross validation has been used to estimate the prediction model performance is mentioned in Table 4.3.

Evaluation Index	K-NN (Cross Validation)
CE	0.9297
MAE	0.4809
RMSE	0.7465

Table 4.3. K-NN Prediction performance of Bhamo station with cross validation

After obtaining the optimal parameter for the K-NN model, the model accuracy was calculated by the testing dataset. The experimental result of the prediction performance is illustrated in Table 4.4.

Evaluation Index	K-NN (Testing Dataset)
CE	0.9508
MAE	0.3961
RMSE	0.6473

Table 4.4.K-NN Prediction performance of Bhamo station with testing dataset

The experimental result shows that K-NN is satisfactory even when it is used to predict the unseen data. Thus, the prediction 5 leads day water levels of each day for Bhamo station is done by K-NN as shown in Figure 4.4.

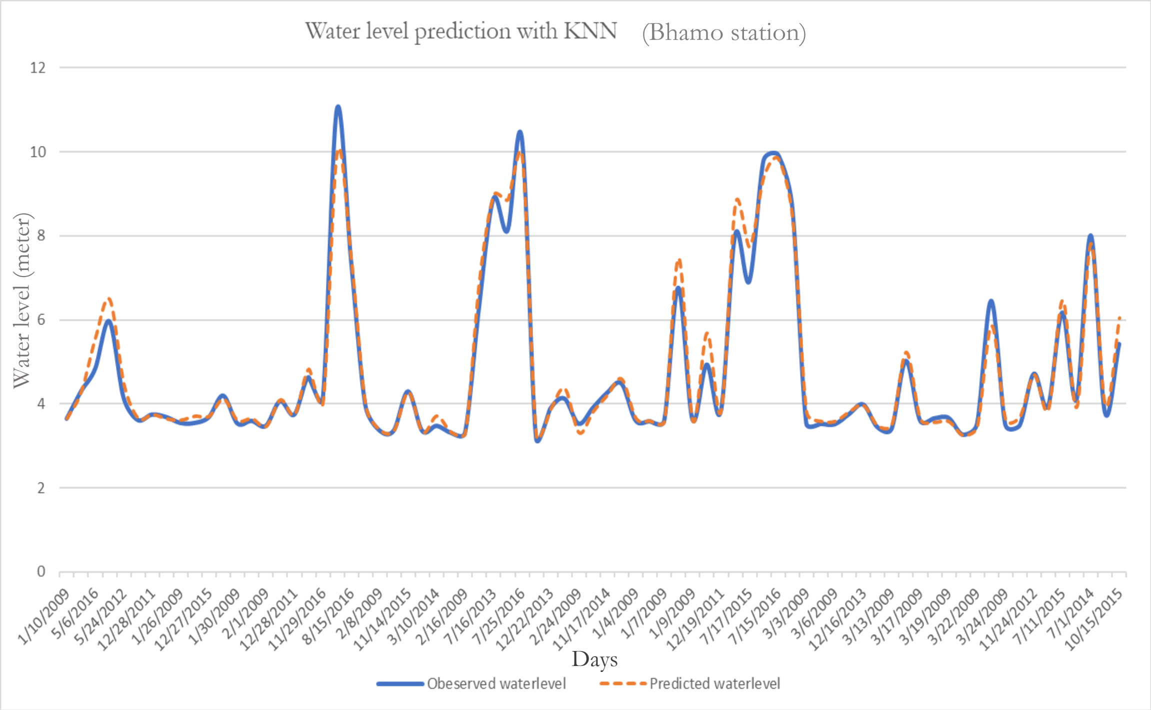

Figure 4.4. Predicted 5 lead day water levels of Bhamo station by KNN

Figure 4.4 shows the prediction of water levels of next five days ahead is closely agreement with the observed value by applying the K-NN.

4.5.2. Model Parameter Estimation and Evaluation with SVR

Another machine learning technique Support Vector Regression is also applied to investigate the five lead day water levels prediction for Bhamo station. The following steps have been performed in SVR for choosing the optimal parameters and evaluating the model performance.

- Investigating the estimated optimal parameters of the algorithm by different combinations free parameters (C & γ).
- The optimal value (C & γ) is selected and evaluate the model with 10 folds cross validation.
- Evaluating the performance of the model with testing dataset.
- Making the prediction of five days ahead water levels for each day.

The estimated optimal parameters of SVR is obtained by training several times with different combination of C & γ (ε has been set to 0.01). Some resulting RMSE (error) of different combinations of (C & γ) are described in Table 4.5. In this work, C $\in (1,2,3,\ldots,5,10,20,\ldots\ldots,100)$ and $\gamma \in (0.01,0.02,0.03,\ldots.,0.1)$ are investigated and using the C= 2 and γ =0.02 with minimum RMSE was achieved.

γ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
C=1	0.7948	0.7761	0.7689	0.7681	0.7715	0.7761	0.7811	0.7873	0.7936	0.8007
C=2⋮	0.7788	**0.7677**	0.7712	0.7785	0.7857	0.7918	0.7994	0.804	0.8093	0.8130
C=3	0.7719	0.7695	0.7775	0.7865	0.7915	0.7971	0.8029	0.8047	0.8105	0.8135
C=5⋮	0.7679	0.7767	0.7858	0.7907	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=10	0.7728	0.7826	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=20	0.7763	0.7825	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=30⋮	0.7765	0.7825	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=50	0.7765	0.7825	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=60⋮	0.7765	0.7825	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135
C=100	0.7765	0.7825	0.7867	0.7906	0.7948	0.7982	0.8018	0.8048	0.8105	0.8135

Table 4.5. RMSE for the different combination of adjustable parameters of Bhamo station

The result of SVR with 10 fold-cross validation can be seen in Table 4.6.

Evaluation Index	SVR(Cross Validation)
CE	0.9257
MAE	0.5189
RMSE	0.7677

Table 4.6. SVR prediction performance of Bhamo station with cross validation

After calculating the prediction performance of SVR with cross validation, the model was tested with testing dataset and the experimental result was mentioned in Table 4.7.

Evaluation Index	SVR (Testing Dataset)
CE	0.9621
MAE	0.4380
RMSE	0.6629

Table 4.7. SVR prediction performance of Bhamo station with testing data

Prediction performance of SVR is convenience for five days water level prediction of Bhamo station. Then, for day by day water level prediction is performed and the observed and predicted value is close as show in Figure 4.5.

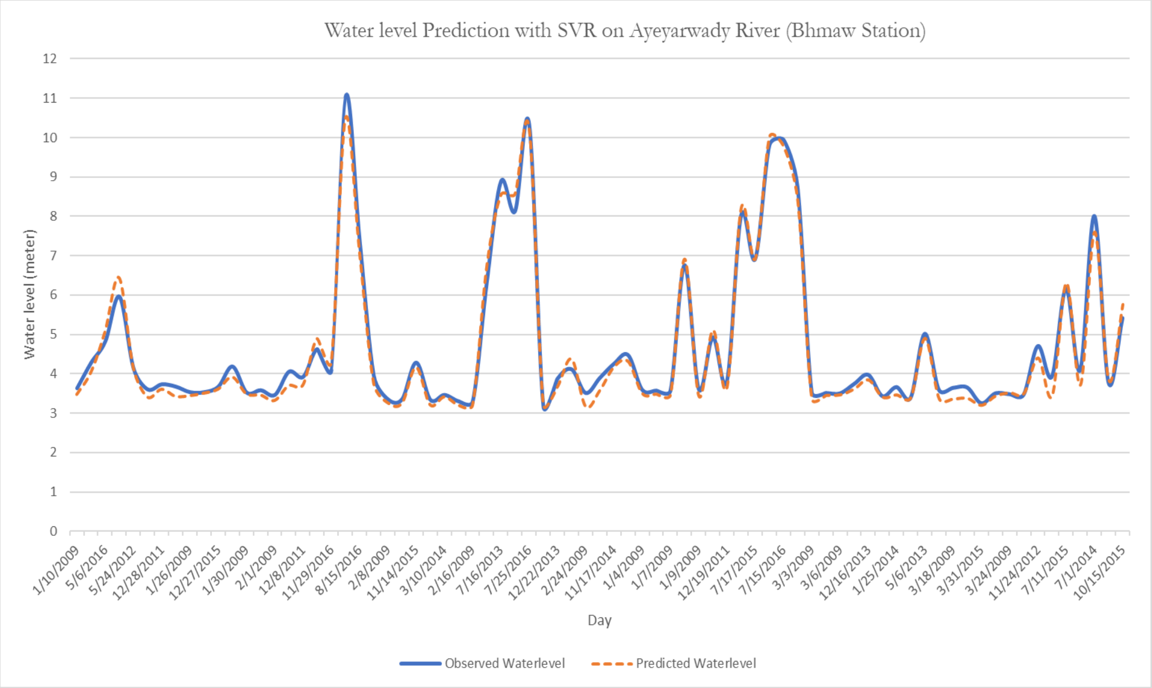

Figure 4.5. Predicted 5 lead day water levels of Bhamo station by SVR

4.5.3. Model Evaluation with Multiple Linear Regressions

As mentioned before, we applied the three different machine learning models to predict the five lead day water levels of our areas of study. Thus, we applied another machine learning approach (multiple linear regressions) to compare only the accuracy of the other approaches (K-NN and SVR). The experimental result of the prediction performance with testing dataset is described in Table 4.8.

Evaluation Index	MLR
CE	0.7684
MAE	2.0011
RMSE	2.2485

Table 4.8. MLR prediction performance of Bhamo Station

Compared with the K-NN and SVR, the error rate of multiple linear regressions is not proper outcomes to predict the five-lead day water level of Bhamo station in the upper part of Myanmar.

4.6. Experimental result of Mandalay station

For the Mandalay station (Station-3) in middle part of Myanmar, the same experimental set up and procedures are performed to prediction 5 leads day water levels as above mentioned in Bhamo station. Three different machine learning models (K-NN, SVR and MLR) was conducted with current and previous two days water levels and rainfalls data from year (2009 to 2016) of Bhamo station (upper station) and Mandalay station (lower station). The data is randomly splitting into 2/3 and 1/3 for training and testing dataset.

4.6.1. Model Parameter Estimation and Evaluation with K-NN

For optimal parameters investigation, K-range from 1 to 500 are tuned and the optimal value of K = 6 with least RMSE (0.4852) is obtained as mentioned in Table 4.9. Prediction accuracy with 10-fold cross validation for Mandalay station is described in Table 4.10.

PARAMETER	RMSE
K=1	0.6003
K=2	0.5217
K=3	0.4958
K=6	0.4852
K=10	0.4889
K=15	0.4952
K=20	0.5017
K=50	0.5330
K=100	0.5698
K=200	0.6151
K=500	0.7631

Table 4.9. K parameter adjusting of Mandalay Station

Evaluation Index	**K-NN (Cross Validation)**
CE	0.9839
MAE	0.31
RMSE	0.4852

Table 4.10. K-NN prediction performance of Mandalay station with cross validation

The prediction performance of K-NN with the optimal K for 5 leads day water levels of Mandalay station is calculated by testing dataset is described in Table 4.11 and day by day prediction of observed and predicted close agreement is illustrated in Figure 4.6.

Evaluation Index	K-NN (Testing Data)
CE	0.9888
MAE	0.2109
RMSE	0.3991

Table 4.11. K-NN prediction performance of Mandalay station with testing data

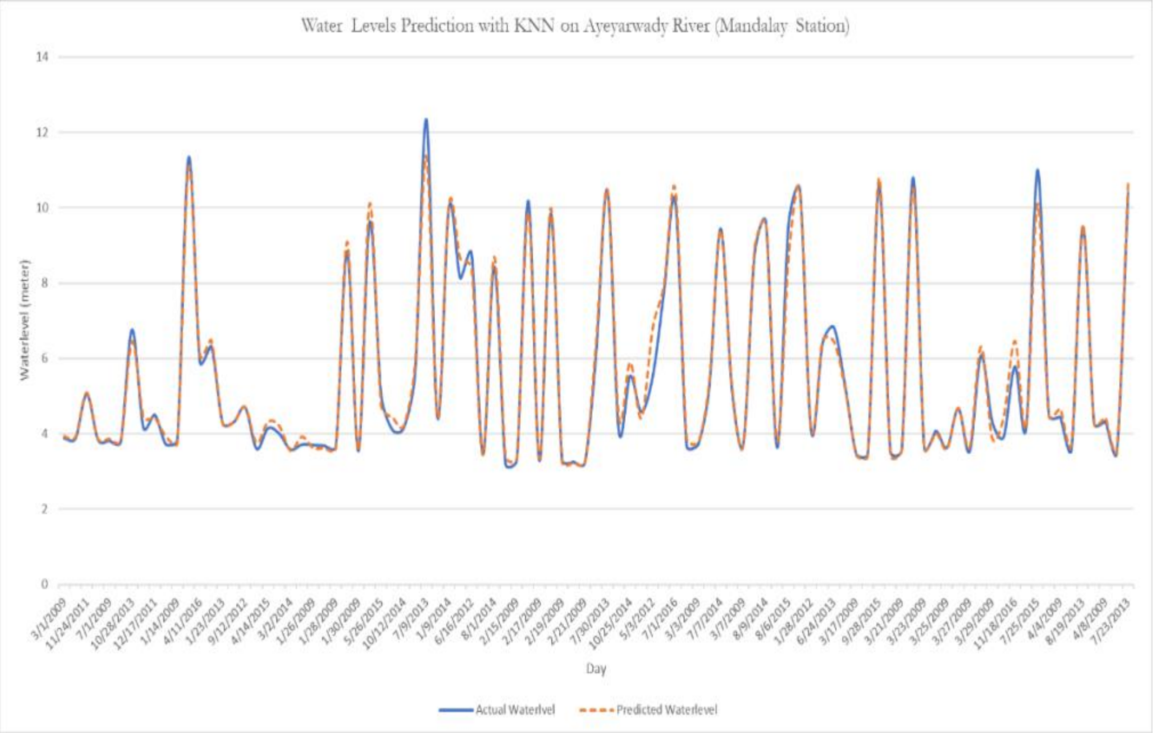

Figure 4.6. Predicted 5 lead day water levels of Mandalay station by KNN

4.6.2. Model Parameter Estimation and Evaluation with SVR

Support Vector Regression (SVR) was applied the prediction water levels of Mandalay station (Staion-3). The procedure to predict the next five days water levels with SVR for Mandalay station is same with Bhamo station that described in previous section. First, different combination of C & γ are tuned to get the estimated optimal parameters of the SVR. $\varepsilon = 0.01$ and C and γ are tuned with the different combination in the range form $(1,2,3, \ldots ,5,10,20, \ldots\ldots ,100)$ and (0.01, 0.02, 0.03,..., 0.1) as mentioned in Table 4.12 . Among these, the less RMSE with C=3 and γ =0.02 is achieved for Mandalay station water levels prediction.

γ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.9	1.0
C=1	0.5348	0.5181	0.5132	0.5131	0.5156	0.5214	0.5277	0.5356	0.5462	0.5564
C=2	0.5187	0.5088	0.5068	0.5092	0.5124	0.5190	0.5258	0.5361	0.5463	0.5578
C=3	0.5107	**0.5059**	0.5073	0.5079	0.5141	0.5184	0.5268	0.5348	0.5455	0.5589
C=5 ⋮	0.5067	0.5061	0.5081	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=10	0.5069	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=20 ⋮	0.5071	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=30	0.5072	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=50 ⋮	0.5072	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=60	0.5072	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589
C=100	0.5072	0.5073	0.5088	0.5110	0.5136	0.5187	0.5267	0.5348	0.5455	0.5589

Table 4.12. RMSE for different combination of adjustable parameters of Mandalay station

Water levels prediction performance of Mandalay station with SVR by cross validation is mentioned in Table 4.13 and the resulting accuracy with testing dataset is shown in Table 4.14.

Evaluation Index	SVR (Cross Validation)
CE	0.9826
MAE	0.3446
RMSE	0.5059

Table 4.13. SVR prediction performance of Mandalay station with cross validation

Evaluation Index	SVR (Testing Dataset)
CE	0.9875
MAE	0.3415
RMSE	0.4948

Table 4.14. SVR prediction performance of Mandalay station with testing dataset

Prediction accuracy of SVR is also acceptable for next five days water level of Mandalay Station in the middle part of Myanmar. The close agreement of day by day prediction is as illustrated in Figure 4.7.

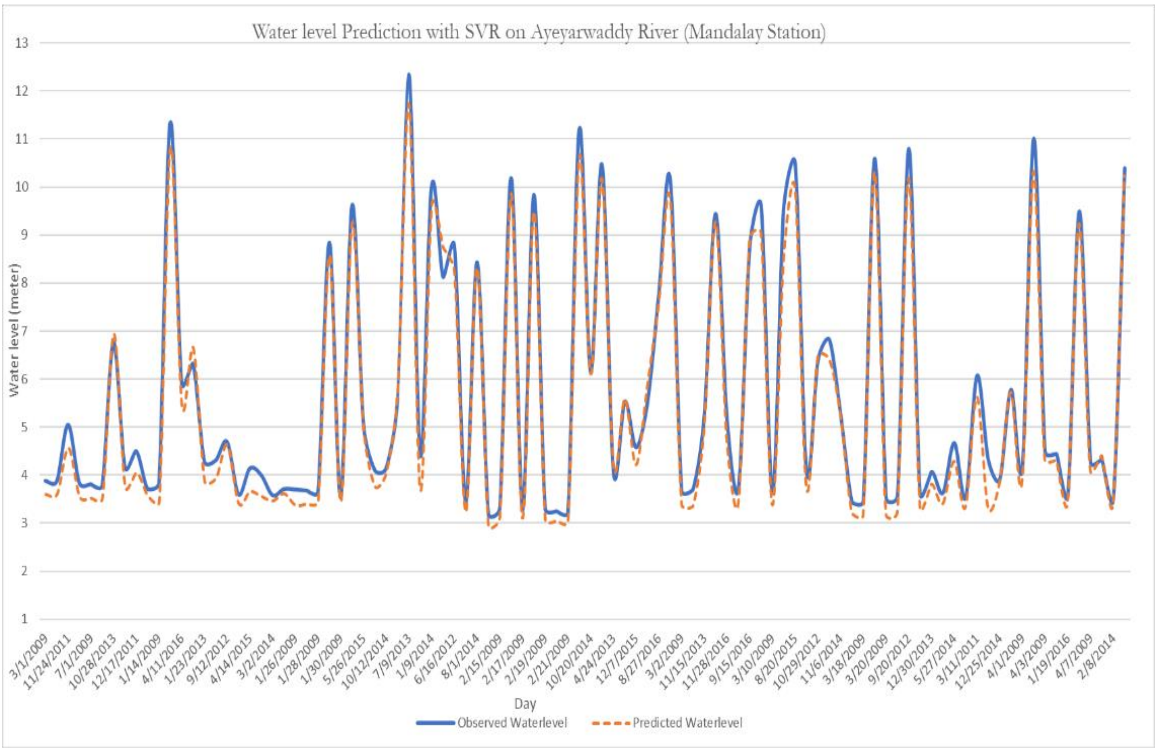

Figure 4.7. Predicted 5 lead day water levels of Mandalay station by SVR

4.6.3. Model Evaluation with Multiple Linear Regressions

Another machine learning model multiple linear regression is also applied on the Mandalay station to compare the prediction accuracy with other proposed prediction approaches. The prediction accuracy with testing dataset is described in Table 4.15.

Evaluation Index	MLR (Testing Dataset)
CE	0.9737
MAE	2.5840
RMSE	2.7868

Table 4.15. MLR prediction performance of Mandalay station

The prediction error with MLR is higher than the KNN and SVR prediction approaches for Mandalay station. That’s why; MLR is not working properly for the prediction of five lead day water levels of Mandalay station in the middle part of Myanmar.

4.7. Experimental result of Bago station

Bago station situated on Bago river at the lower part of Myanmar, three machine learning models (KNN, SVR, and MLR) were utilized for water levels prediction as other two stations (Bhamo and Mandalay) of upper and middle part of Myanmar.

We've collected water levels and rainfall data within 11 years period from 2001 to 2011 from two stations on Bago river. Data collected from the upper station (Zaungtu) and the lower station (Bago) of Bago river are randomly split into 2/3 and 1/3 for training and testing dataset and make the prediction for next 5 days water levels of Bago station on the Bago river that located in lower part of Myanmar.

4.7.1. Model Parameter Estimation and Evaluation with KNN

For estimating of the optimal parameters in KNN model, K value range from 1 to 500 are tuned as shown in Table 4.16 and K= 35 is achieved as the optimal one with less RMSE for Bago station.

Parameter	RMSE
K=1	1.0273
K=2	0.8871
K=3	0.8400
K=10	0.783
K=11	0.7807
K=12	0.7777
K=30	0.7702
K=35	**0.7682**
K=50	0.7700
K=200	0.7898
K=500	0.8269

Table 4.16. K parameter adjusting of Bago station

The KNN model predction performance with estimated optimal value by 10 fold cross validation and prediction evaluation result with unnseen data are described in Table 4.17 and 4.18 respectively.

Evaluation Index	KNN (Cross Validation)
CE	0.9186
MAE	**0.5149**
RMSE	0.7682

Table 4.17. KNN prediction performance of Bago station with cross validation

Evaluation Index	KNN (Testing Dataset)
CE	0.9442
MAE	**0.3720**
RMSE	0.5605

Table 4.18. KNN prediction performance of Bago station with testing data

The prediction outcomes from KNN approach for Bago station is acceptable. Thus, making the prediction for five lead days water levels for each day using KNN and their close agreement is illustrated in Figure 4.8.

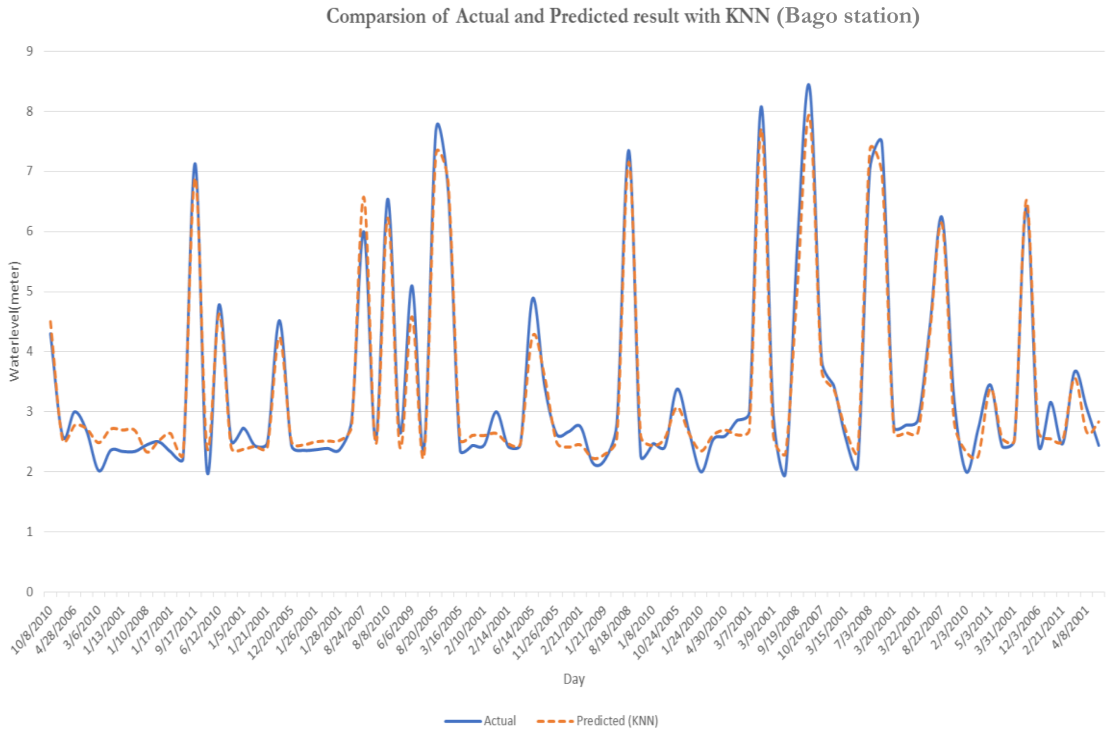

Figure 4.8. Predicted 5 lead day water levels of Bago station by KNN

4.7.2. Model Parameter Estimation and Evaluation with SVR

The adjuastable free parameters C & γ of the SVR algorithm are tuned to get the estimated optimial parameters for Bago station water levels predction with better accuracy. The combination of C within the range (1,2,3…,50,60..,100) and γ (0.01,0.02,0.03,…,0.1) are trained and the SVR with optimal error (RMSE =0.7843) has been selected for predcition . In Bago station, C=1 and γ=0.03 were obtained as the estimated optimal parmaeters of SVR model as mentioned in Table 4.19.

γ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
C=1	0.8031	0.7889	**0.7834**	0.7864	0.7925	0.8032	0.8155	0.8284	0.8407	0.8552
C=2	0.7942	0.7861	0.7965	0.8150	0.8336	0.8524	0.8655	0.8770	0.8844	0.8933
C=3	0.7898	0.7932	0.8158	0.8402	0.8562	0.8691	0.8767	0.8837	0.8890	0.8950
C=5	0.7885	0.8116	0.8380	0.8546	0.8632	0.8713	0.8768	0.8832	0.8893	0.8940
C=10	0.8002	0.8302	0.8442	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940
C=20	0.8115	0.8304	0.8445	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940
C=30	0.8117	0.8304	0.8445	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940
C=50	0.8118	0.8304	0.8445	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940
C=60	0.8118	0.8304	0.8445	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940
C=100	0.8118	0.8304	0.8445	0.8542	0.8640	0.8708	0.8771	0.8832	0.8893	0.8940

Table 4.19.RMSE for the different combination of adjustable parameters of Bago station

Estimation of predciton acccuracy for Bago station (lower part of Myanmar) with cross validation and testing dataset are as shown in Table 4.20 and 4.21 .The predction accuracy of SVR for Bago station is also proper result and their prediction results for each day are mentioned in Table 4.20.

Evaluation Index	SVR (Cross Validation)
CE	0.9152
MAE	**0.5358**
RMSE	0.7843

Table 4.20. SVR Prediction performance of Bago station with cross validation

Evaluation Index	SVR(Cross Validation)
CE	0.9234
MAE	**0.4317**
RMSE	0.6558

Table 4.21. SVR Prediction performance of Bago station with testing data

The SVR's predctiion accuracy with the unseen values of Bago station is also achieved to be acceptable and the predicted value for next five day water levels is as shown in Figure 4.9.

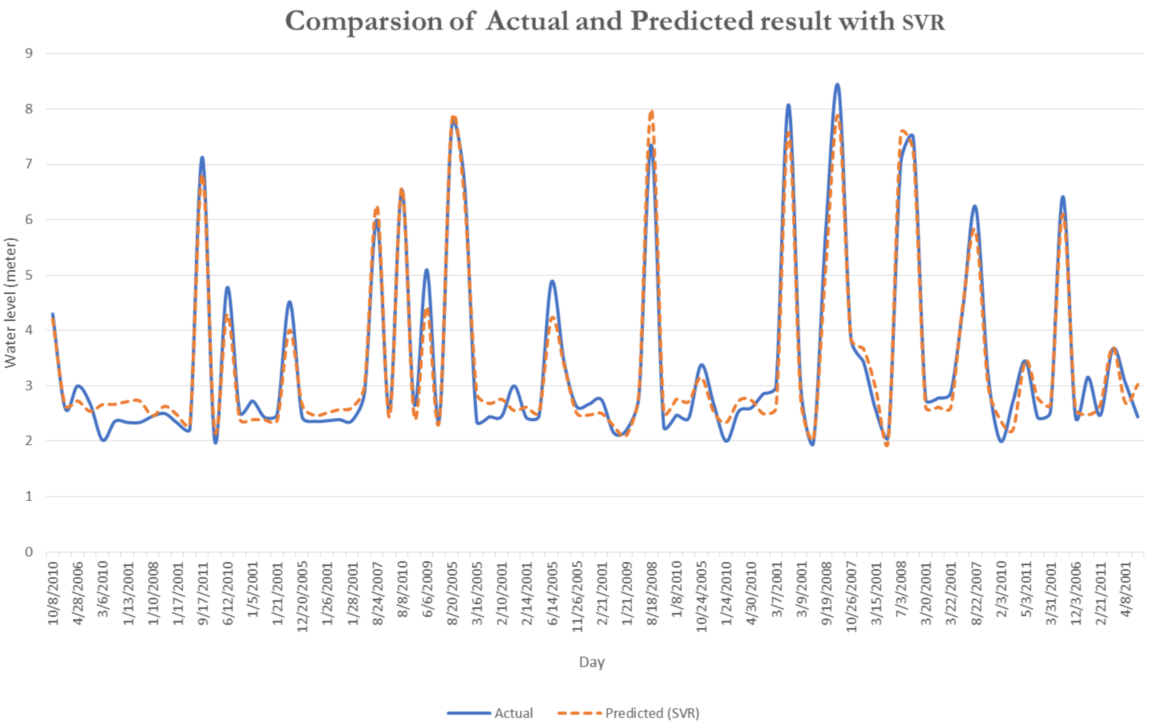

Figure 4.9.Predicted 5 lead day water levels of Bago station by SVR

4.7.3. Model Evaluation with Multiple Linear Regressions

As described early, water levels predction is also investigated with MLR to compare the accuracy of predction with KNN and SVR. The investigating result with testing data is according to the Table 4.22.

Evaluation Index	MLR (Testing Dataset)
CE	0.8500
MAE	1.5972
RMSE	1.7621

Table 4.22.MLR prediction performance of Bago station

Analysing the five lead day water levels of Bago station with MLR approach is not convenience because of the error rate is significantly higher than with other two approaches.

4.8. Comparison of Prediction Performance

Three machine learning approaches (KNN, SVR, and MLR) based on time series data have been applied for the prediction of five lead day water levels of three stations : Bhamo station and Mandalay station on the Ayeyarwady river and Bago stations on Bago river of the upper, middle and lower part of Myanmar. Then, the accuracy of three different machine learning approaches is compared. The comparison of prediction accuracy results for three different station using three machine learning techniques are presented in Table 4.23, 4.24 and 4.25 respectively.

Model	CE	MAE	RMSE
MLR	0.7684	2.0011	2.2485
SVR	0.9621	0.4380	0.6629
KNN	0.9508	0.3961	0.6473

Table 4.23. Comparison of water level prediction performance of Bhamo station

First, Multiple linear regression predictive model obtained MAE 2.0011 and RMSE 2.2485 whereas CE is 0.7684. Secondly, SVR model showed the better result than multiple linear regressions and MAE is 0.4380 and RMSE is 0.6629 whereas CE is 0.9621. Finally, K-NN model achieved the better result and MAE is 0.3961 and RMSE is 0.6473 whereas CE is 0.9508.

MODEL	CE	MAE	RMSE
MLR	0.9737	2.5840	2.7868
SVR	0.9875	0.3415	0.4948
KNN	0.9888	0.2109	0.3991

Table 4.24. Comparison of water level prediction performance of Mandalay station

Table 4.24 shows the prediction accuracy of Mandalay station. MLR model presented MAE= 2.5840 and RMSE=2.7868 whereas CE is 0.9737. SVR obtained MAE =0.3415 and RMSE=0.4948 whereas CE is 0.9875 and the better result of MAE =0.2109 and RMSE=0.3991 were achieved by applying KNN whereas CE is 0.9875.

Water level prediction performance of Bago station on Bago river with three different prediction approaches and their accuracy comparison are described in Table 4.25.

MODEL	CE	MAE	RMSE
MLR	0.8500	1.5972	1.7621
SVR	0.9234	0.4317	0.6558
KNN	0.9442	0.3720	0.5605

Table 4.25. Comparison of water level prediction performance of Bago station

In Bago station, KNN achieved the proper result with MAE=0.3720 and RMSE=0.5605 whereas CE is 0.9442. Prediction error of MAE =0.4317 and RMSE=0.6558 were obtained using SVR algorithm whereas CE is 0.9234. MLR has high error of MAE =1.5972 and RMSE=1.7621 whereas CE is 0.8500.

In our work, the portion of dataset 65 % is analysed via cross validation to find the optimal parameters for the proposed prediction and hyperparameters of the related algorithms are adjusted to avoid over fitting. Then, proposed algorithms were learnt with optimal parameters and used the remaining portion of the data (35%) to test the prediction performance. The training and testing accuracy of existing KNN and SVR is good enough and the prediction error on test dataset is lesser than the training set that is less likely to be overfitting.

To sum up, the present KNN and SVR prediction accuracy is better than MLR and have been worked out for water levels prediction for Bhamo station and Mandalay station on Ayeyarwady river in the upper and middle part of Myanmar and Bago station on Bago river in the lower part of Myanmar according to our experimental results.

4.9. Summary

In this chapter, step by step procedures for implementation of our proposed approaches for water levels prediction are presented with real valued datasets. Model validation and evaluation is also performed with this datasets and the prediction results of different areas are also illustrated with tables and graphical representations.

CHAPTER 5

Conclusion and Recommendations

In this research, different machine learning techniques K-Nearest Neighbours (KNN), Support Vector Regression (SVR), Multiple Linear Regression (MLR) have been applied to predict the short-range water level prediction. Simple, more efficient and reliable time series data driven machine learning prediction approaches are developed to improve the performance of water levels predictions on the study region.

5.1. Conclusion

Water levels prediction is one of the significant factors in flood early warning and disaster management due to its accurate prediction is one of the major challenges facing the meteorologist all over the world. In our work, data driven machine learning approaches that is based on analysis of past historical weather data and their relationship are implemented to predict the short range water levels of study region.

As the goal of the machine learning, there is no perfect prediction that our proposed approaches KNN and SVR are good enough to be useful for short range water levels prediction.

According to the experimental result from Chapter (4), three different machine learning approaches are compared with each other for time series water levels prediction. The objective is to reduce the error rate and increase the level of the performance accuracy using the optimization of the algorithm. The proposed approaches are analysed with different combination of parameters to improve the generalization performance (estimation accuracy). We can adjust the hyperparameters and can generalize well on training and test data to overcome the overfitting and underfitting that is one of the common machine learning problems. Moreover, we got the advantage of prediction approaches to be useful from single site water levels prediction to multiple site water levels prediction according to the experiment results of three different locations.

Prediction accuracy of KNN and SVR significantly outperform than MLR and KNN accuracy is slightly better than SVR. On the other hand, SVR has the larger number of controlling parameters and use the structural risk minimization principle and more generalization capacity compared with simple KNN. Therefore, SVR is also a potential for further improvement of the accuracy in terms of using large number of data, selection of training set and improved generalization.

Thus, the current KNN and SVR are promising alternatives for making five lead days water levels prediction for our study region. Both can handle to predict the low, normal and high water levels within the acceptable error rate. In this work, seriously varied time series data (from low to very high) such as 4 meters to 12 meters are not included. Thus, other pertinent distribution functions are needed to be considered to handle seriously varied water levels for some regions water levels prediction.

5.2. Recommendation

In this work, the proposed machine learning approaches are investigated only with the time series values of our study region, Myanmar. Different prediction patterns can be got with different time series data and key features from various geographical locations. Some key variables of geography, topology, climate and hydro metrology characteristics should be identified in here to promote the reliable multiple site prediction from single site prediction.

For some regions, water levels may vary seriously from low to very high because of the influence on the both human and nature, other reliable functions and kernel functions can be added and applied to the present approaches to perform real time prediction and practically test their performance as a further work. Ensemble learning methods and Artificial Neural Network Topology can be analysed to have better prediction accuracy. Furthermore, other weather data estimation such as rainfall, temperature and pressure prediction can be implemented using the properties of proposed prediction as future attempts.

REFERENCES

[1] A. Soni, "www.quora.com," Quora, 5 October 2017. [Online]. Available: https://www.quora.com/What-is-the-difference-between-white-box-black-box-and-gray-box-testing.

[2] Magoulès, Frédéric, Piliougine, Michel, Elizondo, David, "Support Vector Regression for Electricity Consumption Prediction in a Building in Japan," in 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), Paris, France, pp.24-26, Aug. 2016.

[3] T.T. Nguyen, Q. N. Huu, M. J. Li, "Forecasting Time Series Water Levels on Mekong River Using Machine Learning Models," in 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, pp. 292-297, Oct. 2015.

[4] F. Carlo, C. Garcia, Alvin, E. Retamar, Joven C. Javier, "Development of a predictive model for on-demand remote river level nowcasting: Case study in Cagayan River Basin, Philippines," in 2016 IEEE Region 10 Conference (TENCON), Singapore, pp.22-25, Nov. 2016.

[5] K.Pasupa, S. Jungjareantrat, "Water levels forecast in Thailand: A case study of Chao Phraya river," in 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), Phuket, Thailand, pp.13-15, Nov. 2016.

[6] P. C. Mohammad Sajjad Khan, "Application of Support Vector Machine in Lake Water Level Prediction," Journal of Hydrologic Engineering, vol.11, no.3, pp. 199-205, May. 2006.

[7] C. Damle, A.Yalcin, "Flood prediction using Time Series Data Mining," Journal of Hydrology, vol. 333, no. 2, pp. 305-316, 2007.

[8] J.Yang, C. Cheng, C. Chan, "A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method," Journal of Hydrology, vol. Computational Intelligence and Neuroscience Volume 2017, no.3, pp. 1-11, 2017.

[9] W. T. Zaw, T. T. Naing, "Empirical Statistical Modeling of Rainfall Prediction over Myanmar," International Journal of Computer and Information Engineering, vol. 2, no. 10, 2008.

[10] C. T. Zan, T. T. Naing," Myanmar Rainfall Forecasting Using Hidden Markov Model," in IEEE International Advance Computing Conference, Patiala, India, 2009.

[11] M.Kannan, S.Prabhakaran, P.Ramachandran, "Rainfall Forecasting Using Data Mining," International Journal of Engineering and Technology, vol. 2, no. 6, pp. 397-401, June. 2010.

[12] A. Makarand, Kulkarni, S.Patil, G. V. Rama, P. N. Sen, "Wind speed prediction using statistical regression and neural network," Journal of Earth System Science, vol. 117, no. 4, p. 457- 463, 2008.

[13] A. Agrawal, V. Kumar, A. Pandey, I.Khan, "An Application of Time Series Analysis for Weather Forecasting," International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 2, pp. 974-980, 2012.

[14] A.Hipni, A.El-shafie, A.Najah, O.Abdul Karim, A.Hussain, M.Mukhlisin, "Daily Forecasting of Dam Water Levels: Comparinga Support Vector Machine (SVM) Model With Adaptive Neuro Fuzzy Inference System (ANFIS)," Water Resources Management, vol. 27, no. 10, pp. 3803–3823, August 2013.

[15] Jan Z., Abrar M., Bashir S., Mirza A.M, "Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique, "Wireless Networks, Information Processing and Systems (IMTIC 2008), vol 20, no.1, pp.40-51, April 2008.

[16] Raschka, S. (September 2015). Python Machine Learning. Birmingham, United Kingdom: Packt Publishing Ltd.

[17] http://www.data-machine.com/, data:machine: tools computer intelligence , 2015. [Online]. Available: http://www.data-machine.com/ nmtutorial/ introduction1.htm

[18] J. Brownlee, "https://machinelearningmastery.com," Machine Learning Mastery Pty. Ltd., 16 March 2016. [Online]. Available: https:// machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

[19] N. Mccrea, "Toptal," 2014. [Online]. Available: https:// www.toptal.com /machine-learning/machine-learning-theory-an-introductory-primer.

[20] A.J. Champandard, "http://reinforcementlearning.ai-depot.com/Main.html," Reinforcement Learning Warehouse, [Online]. Available: http://reinforcementlearning.ai-depot.com/.

[21] "data:machine: tools for computational intelligence," 2015. [Online]. Available: http://www.data-machine.com/nmtutorial/introduction1.htm.

[22] "WIKIPEDIA," [Online]. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#The_1-nearest_neighbor_classifier. [Accessed June 2018].

[23] X. Wu, V.Kumar, J.R.Quinlan, J. Ghosh, Q. Yang , H. Motoda "Top 10 algorithms in data mining," in IEEE International Conference on Data Mining, Hong Kong, 2006.

[24] Y. Gu, G.Yu, X. Yu, "The efficient Spatial Index Method for K nearest Neighbours," Journal of Information Science And Engineering, vol. 30, pp. 1569-1583, 2014.

[25] M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm," International Journal of Computer Engineering and Information Technology, vol. 8, no. 6, pp. 90-95, June 2016.

[26] D. Saed, "http://www.saedsayad.com/data_mining_map.htm," [Online]. Available: http://www.saedsayad.com/k_nearest_neighbors_reg.htm.

[27] C. M. Farrelly, "KNN Ensembles for Tweedie Regression: The Power of Multiscale Neighborhoods," in eprint arXiv:1708.02122, 07/2017.

[28] Hu L-Y, Huang M-W, Ke S-W, Tsai C-F. The distance function effect on k-nearest neighbor classification for medical datasets. Springer Plus. 2016;5(1):1304. doi:10.1186/s40064-016-2941-7.

[29] "WIKIPEDIA," [Online]. Available:https://en.wikipedia.org/wiki/ Hyper parameter_optimization. [Accessed June 2018].

[30] B. D. M. Marc Claesen, "Hyperparameter Search in Machine Learning," in The XI Metaheuristics International Conference, Agadir, 2015.

[31] V. N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 2nd edition, 2010.

[32] V. N. Vapnik. Statistical Learning Theory. Wiley-Interscience, 1998.

[33] N. Hasan, N. Chandra Nath, R. Islam Rasel, "A support vector regression model for forecasting rainfall," in 2nd International Conference on Electrical Information and Communication Technologies (EICT), Khulna, Bangladesh, 2015.

[34] P.-S. Yu, S.-T. Chen, I-F. Chang, "Real-Time Flood Stage Forecasting Using Support Vector Regression," Practical Hydro Informatics, vol. 68, pp. 359-373, 2009.

[35] F.Wang , G. Tan, C. Deng, Z. Tian, "Real-time traffic flow forecasting model and parameter selection based on ε-SVR," in 2008 7th World Congress on Intelligent Control and Automation, Chongqing, China, pp.2870-2875, 25-27 June 2008.

[36] Alex J, B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp 199–222, August 2004.

[37] M. N. Bernstein, "Computer Sciences User Pages," University of Wisconsin-Madison, 2015. [Online]. Available: http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf.

[38] S. Ray, "Analytics Vidhya," 13 SEPTEMBER 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[39] Luminto, Harlili, "Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate," in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), Denpasar, Indonesia, 16-18 Aug. 2017.

[40] A. Blokhin, "INVESTOPEDIA," [Online]. Available: https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp.

[41] "ReliaWiki.org," ReliaSoft Corporation, [Online]. Available: http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis.

[42] K. R. M. Supapo, R. V. M. Santiago,M. C. Pacis, "Electric load demand forecasting for Aborlan-Narra-Quezon distribution grid in Palawan using multiple linear regression," in 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Manila, Philippines, 1-3 Dec. 2017.

[43] "https://www.kdnuggets.com/," Kduggets, August 2017. [Online]. Available: https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.htm

[44] Sanjay Yadav, Sanyam Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in 2016 IEEE 6th International Conference on Advanced Computing, Bhimavaram, India, 27-28 Feb. 2016.

LIST OF ACADEMIC ACHIEVEMENT

Conference Publications

1. Tin Nilar Lin and H.Watanabe, “Water Level Prediction for Disaster Management Using Machine Learning Models," in Forum on Information Technology, FIT, H-029, Tokyo, Japan, 2017.
2. Tin Nilar Lin and H.Watanabe, “Weather Data Estimation by Sensitive features Selection,” in ITE Winter Annual Convention, 13B-6, Tokyo, Japan, 2017.
3. Tin Nilar Lin and H.Watanabe, “Prediction of Next Day’s Sea Level Pressure by Support Vector Regression,” IEICE General Conference, D-20-11, Tokyo, Japan, 2018.