IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 11, 2021, accepted July 30, 2021, date of publication August 3, 2021, date of current version August 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102044

A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model

NUR AQILAH PASKHAL ROSTAM[1], NURUL HASHIMAH AHAMED HASSAIN MALIM[1], ROSNI ABDULLAH[1], ABDUL LATIF AHMAD[2], BOON SENG OOI[2], AND DEREK JUINN CHIEH CHAN[2]

[1]School of Computer Sciences, Universiti Sains Malaysia, USM, Penang 11800, Malaysia
[2]School of Chemical Engineering, Universiti Sains Malaysia, Nibong Tebal, Penang 14300, Malaysia

Corresponding author: Nurul Hashimah Ahamed Hassain Malim (nurulhashimah@usm.my)

This work was supported by the Malaysian Ministry of Higher Education through the Transdisciplinary Research Grant under Grant TRGS/1/2018/USM/01/5/4-203.PKOMP.67612.

ABSTRACT An end-to-end process to achieve a complete framework methodology for Harmful Algal Bloom (HAB) growth prediction is crucial for water management, especially in implementing robust predictive modelling of HAB to prevent water pollution. Previous works have separately focused on the prediction part or the implementation of the water monitoring system that involves the integration of sensors through the Internet of Things (IoT). These studies lack in terms of discussion of both IoT with the algae ecological domain and prediction method. Therefore, this paper takes the initiative to provide a wider coverage on the end-to-end process including the assembly and integration of sensors, data acquisition and predictive modelling using data-driven approaches, for example, machine learning, deep learning and deep time series forecasting algorithm for future algal bloom outbreak mitigation. This paper believes that discussion in a complete framework perspective based on the execution of each phase is important besides providing a true understanding of the algae growth factors and prediction problems to achieve a robust prediction algorithm for algal growth. In the end, this paper presents proof that selecting the right features and utilising time series with deep learning are much better for tackling the issues of highly non-linear and dynamic algae ecological data that are briefly introduced in this paper. Among all the algorithms selected, Long Short-term Memory (LSTM) is the best fit for the prediction method and has outperformed other basic machine learning methods in accurately predicting algal growth through the prediction of chlorophyll-a (Chl-a) as a strong indicator of algal presence for coastal studies.

INDEX TERMS Deep learning, harmful algal bloom (HAB), IoT, long short-term memory (LSTM), machine learning, time series forecasting.

I. INTRODUCTION

Recently, water resources have been reported to be polluted by the increase of nutrients and minerals that consequently promote excessive algal growth [1]. Harmful algal bloom (HAB) has long been a threat to water sources due to the rapid increase and accumulation of algae population that can cause harm. HAB toxin negatively affects human health, the environment, and the economy whilst non-toxic ones can damage fisheries resources and equipment [2], [3]. Harmful toxins such as neurotoxins released by the algae encourage decomposers' growth that increases the biochemical oxygen demand (BOD) of the water. As the algae decompose, oxygen is removed from the water, which then starves the fish and plants of oxygen and damages the local ecology. BOD acts as a measure of the amount of dissolved oxygen (DO) that has been consumed. When BOD is high, less DO will be available for other organisms. This promotes competition for oxygen and finally, causes water pollution known as eutrophication [4].

With the current advancement of the Internet of Things (IoT), the use of various sensors has also increased, which

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno Garcia.

further facilitates the process of monitoring and profiling water quality and eutrophication mitigation. The IoT is a network of physical objects that have evolved into a network of devices such as smartphones, cameras, etc. for homes and vehicles that all are connected, communicating and sharing information [5]. Due to the development of sensors as part of IoT for monitoring, fields such as ecological informatics [6] and bioinformatics have gained various benefits where manual profiling is automated.

A sensor is a device that receives a signal (physical, chemical or biological) and converts it into an electric signal output such as current or voltage [7]. Since profiling processes are arduous, time-consuming and lack real-time outcomes to stimulate proactive response to water pollution, the use of sensors is considered a promising alternative for water quality control.

To date, the key challenges in the study and management of HABs are species variety, life histories, ecosystems, and the impacts involved. For example, algae communities such as phytoplankton or cyanobacteria that are categorised as potentially harmful do not fit a sole, evolutionarily distinct group [8]. Since algae communities comprise various species and differ in nonlinear ways, they are complex and hard to analyse and are not well understood, resulting in unreliable predictive models [9]. The dynamic growth of algae, which can vary on short timescales (e.g., hours to days) has made identifying the condition that favours HABs a major research effort.

In algae or HAB prediction, algal count and chlorophyll concentration, especially chlorophyll-a (Chl-a), have been widely used to indicate the presence or growth of algae. Algae concentration can change abruptly where the current chlorophyll content can sometimes increase or decrease up to 5 times than before, causing great difficulties in predicting accurately [10].

Therefore, the prediction of algae remains difficult and unreliable due to the dynamic nature of the time series algae ecological data [11], [12]. Besides, this dynamic nature creates highly nonlinear data which results in randomness issues in model fitting [10], [12]–[14], [15]. Randomness issues are rooted in anomalies that have made algal bloom predictions extremely complicated and not well understood. Various research works randomly selected factors for algal growth and depended only on the domain knowledge for feature selection by including all the factors that seemed to be important. This led to model fitting issues and caused fluctuating performance. This paper believes that if the dynamic issues can be tackled accurately, which considers from the data or features level until the algorithm level, all these mentioned strategies might improve the overall prediction performance more.

Based on past literature on the prediction method, due to the success of the data-driven prediction method either with or without considering temporal behaviour [16], researchers used historical data to predict algal blooms by incorporating machine learning techniques [11], [17]–[19]. Machine learning provides a principled set of mathematical methods for extracting meaningful features from data into distinct and meaningful patterns that can be exploited for decision making, estimation and forecasting. The most applied machine learning methods include Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), and regressions. Even though only the input and output of the model are needed for data-driven models, the prevailing data-driven models, especially those using basic machine learning techniques as mentioned above, are unable to effectively extract features of multi-factor timing data and solve the dynamic issues.

Another issue concerns the implementation of the monitoring system. Inspired by the high cost of the commercialised sensor and the dynamic nature of algae that has complicated the prediction process, this article also discusses the development of a solar-powered and low-cost real-time monitoring system to profile the quality of water. Water quality data collected in the data acquisition phase will be used for the development of a predictive model.

Our previous work [3] has managed to implement a solar-powered and low-cost water quality monitoring system (WQMS) for coastal studies. However, it was only a preliminary study focusing on simple data analysis of the parameter readings. This paper extends from that previous work and will discuss in detail the modelling phase for the predictive modelling, especially in tackling the dynamic problem of algae ecological data and will briefly mention the enhancement progress of the previous system. Later, the chosen predictive modelling will be applied to the data collected in the previous study [3]. Hence, a complete framework is presented in this paper, which covers the end-to-end process of developing a water monitoring system, deployment, installation, and prediction model development, which was missing in our previous work in this domain.

Algae can damage fisheries equipment. This can further affect the algae ecological data with missing values and consequently, reduce the quality of the data. Not limited to algae ecology, in general, time series data are vaguely defined expert knowledge due to the existence of random variables, incomplete and inaccurate data, and approximate estimations rather than measurements, which rendered the understanding of data to remain elusive [20], [21]. Algae ecological data that are exposed to high missing values are due to the dependence on monitoring sensors or systems that need frequent maintenance.

Coralline algae would attach to the equipment and damage all the installed sensors [3] which would later lead to missing data collection for the day. This is one of the main reasons why early algae prediction remains crucial as more time is provided for facilities that use coastal water to shut down before their equipment is damaged [22], [23].

Existing ecological studies, especially those on the algae population are lacking in several aspects. To achieve robust predictive modelling of algal growth, several issues must be highlighted and addressed, for example, (i) the features must be mapped to the dynamic issues of algae ecology and (ii) a suitable algal growth predictive modelling must be found, particularly to tackle dynamic algae for coastal studies. This is because, previously, prediction has been mostly done in rivers [24]–[26] and lakes [27]–[30]. Thus, more research is needed for coastal areas and other types of water sources such as estuary. An algorithm that worked well for one type of water source might not work well for different water sources as they have different characteristics in terms of their hydrologic, geographic, climatic, morphologic, physical, chemical, geochemical, and biological features. Addressing the problems through the features (water parameter) level and algorithm level might help to achieve the main aim of this paper of solving the dynamic issues by discussing the development of coastal WQMS with predictive modelling. Hence, this paper aims to answer this research question:

1) How to map the features and the right prediction method to concurrently tackle the dynamic issues of algae to provide a wider coverage from the end-to-end process based on the IoT perspective until the predictive modelling?

This paper's primary objective is to address and discuss the research question which will later complete the discussion on the overall framework of coastal WQMS. The two contributions of this paper are first, a low-cost solution to the arduous manual sample collection process. Second, is the development of the most suitable predictive modelling which can be used to tackle the dynamic issues of algal growth prediction in coastal studies. In the end, the predictive modelling will be embedded into the IoT architecture, indirectly implementing a smart IoT system. Besides, additional experiments using suitable data-driven approaches, especially deep time series, are performed to capture the non-linearity and temporal-dynamic of coastal algae ecological study as a proof of concept. Finally, a complete framework that consists of the end-to-end process in achieving predictive modelling for coastal algal growth is presented in this paper. The paper believes that implementing the framework would benefit ecological informatics and other domains in artificial intelligence in terms of observing the importance of tackling both the features until the algorithm level.

The paper is organised as follows. A short review of past studies on identifying algal growth factors, development of WQMS and algal growth prediction is presented in Section II. Then, the research framework extended from our previous study is introduced in Section III. This framework that focuses more on predictive modelling using machine learning and time series with deep learning methods is discussed in detail. Section IV discusses the results and analysis while Section V concludes the paper.

II. RELATED WORK

There are numerous continuous studies on algae ecology including algal classification, algal detection, water quality profiling, algal analysis, and algal prediction [31]–[35]. Nevertheless, this paper focuses on both WQMS and algal prediction model development.

Prediction of algae in terms of Chl-a concentration is a strong indicator that can contribute to the eutrophication issue and is considered an important water quality (WQ) parameter for the management of water resources. Nonetheless, since conventional profiling processes are arduous, time-consuming and lack real-time findings to encourage proactive response to water pollution, the use of sensors is considered a promising alternative. Profiling water quality for the eutrophication mitigation process will help to provide a solution for a cost-effective and high-quality profiling technique to solve the environmental concern due to highly polluted water discharge. It will also help provide insights from the data for robust predictive modelling at the later stage. However, to the best of the researchers' knowledge, discussion on both monitoring work that also discussed the implementation of the monitoring system with the predictive modelling part has never been done together in the past, especially for coastal studies that are still lacking in predictive modelling. As such, this paper will review several past studies on the factors that contribute to algal growth in general, the monitoring progress, and the algal growth prediction methods.

A. IDENTIFYING IMPORTANT FACTORS OF ALGAL GROWTH

Phytoplankton comprises several groups of algae and cyanobacteria. In general, phytoplankton is described as free-floating and reliant on water movement for maintenance and transport [36]. Various factors affect their population dynamics and these differ based on the type of phytoplankton being scrutinised. Nevertheless, all algae species depend on light as a basic input for photosynthesis and need nutrients, for example, nitrogen and phosphorus for growth and reproduction. Factors such as water temperature, turbidity, mixing, competition, and grazing are also pertinent to the population dynamics of algae.

External pollution loading coupled with hydrodynamic force influence the concentrations of nutrients, which, subsequently, together with the underwater light intensity, affect phytoplankton evolution [37]. The four main factors that are crucial for algal growth are [38], (i) the growth coefficient, (ii) the influence of solar radiation which under conditions of unlimited nutrient availability is the main driving force for algal growth, (iii) the influence of turbulence that is inclined to increase with rising flow leading to resuspension of sediment material and decreasing light penetration, and (iv) the self-shading factor where algal populations continue to grow until light penetration is decreased by the algae themselves.

Twenty environmental parameters that facilitate cyanobacteria bloom in freshwater are listed in the latest study [39]. Amongst them are water temperature (WT), ambient temperature, Secchi disk depth (SD), transparency, turbidity, solar radiation, total phosphorus (TP), total nitrogen (TN), NH4-N, NO3-N, ammonium ion concentration, DO, conductivity, alkalinity, calcium concentration, total suspended solids (TSS), silica, pH, salinity, and chlorophyll-a.

Chlorophyll is the green pigment in leaves that allows plants to generate energy light via photosynthesis. The amount of photosynthesising is indirectly determined by measuring chlorophyll. In a water sample, such plants are algae or phytoplankton. Chlorophyll is the measure of all the green pigments regardless if they are alive or dead. Meanwhile, Chl-a is the measure of the portion of the pigment that is still alive. Both algae number and Chl-a concentration are affected by factors such as sunlight, temperature, nutrients, and wind. During spring, when water starts to warm, the days are sunnier, and nutrients are abundant, the first outbreak or ''bloom'' of algae might happen. As the days become progressively warmer and sunnier, algae will continue growing. Predicting algae concentration that can be measured in total chlorophyll form in raw water has previously been carried out as a strong algal growth indicator [19], [40].

Other factors that are commonly used as a strong indicator are turbidity, water depth, nutrient loading, total dissolved solids (TDS), light intensity, temperature, climate change, and water quality parameter. A liquid's measure of relative clarity is turbidity. It is water's optical characteristic and signifies the amount of light dispersed by material in the water when light is shone through the water sample. Turbidity will be higher when the light scattering intensity is high. Among the materials that make the water turbid are clay, silt, finely divided inorganic and organic matter, algae, soluble coloured organic compounds and other microscopic organisms. Water turns cloudy or opaque due to turbidity. Algae growth increases water turbidity. This is because algae block the light from passing through the water, hence, narrowing the light spectrum below the water surface [41], [42]. Thus, other than Chl-a, turbidity also plays a huge role in indicating the growth rate of algae.

Numerous researchers have related essential parameters such as water temperature, turbidity, pH, and Chl-a and other standard parameters to algae growth. Water temperature, pH, and DO were reported to be positively associated with cyanobacterial community dynamics and concentrations of microcystins [17]. Other than that, nutrient level, phosphate, and nitrogen concentration were determined to be the vital factors for cyanobacterial proliferation. Water depth and nutrient loading are the two major factors causing HABs [43]. Shallow waters can raise the water temperature, encouraging the growth of HABs. This is linked with seasonal changes such as during spring and summer that generate more blooms since there is more sunlight. Another factor that encourages bloom growth is nutrient loading from tributaries and phosphorus and nitrogen are the main nutrients for microcystins, a type of toxins produced during HABs.

TDS is a measurement of inorganic salts, organic matter and other dissolved materials in water [44]. This measurement does not distinguish among ions. TDS results in toxicity via increases in salinity, changes in water's ionic composition, and toxicity of individual ions. Surges in salinity have been demonstrated to result in shifts in biotic communities, reduce biodiversity, eliminate less-tolerant species and cause acute or chronic effects at specific life stages. In some studies, a significant and positive link between Chl-a concentration (an estimate of primary production) and concentrations of Na+, Mg2+, SO42−, HCO3− and CO32− has been shown [45]. Light is crucial for autotrophic growth and photosynthetic activity because algae contain chlorophyll a and b, which are important light-harvesting pigments sensitive to blue and red light.

Other than light, temperature affects the cellular chemical composition, uptake of nutrients, CO2 and the growth rates of each algae species. A thorough study on the impact of temperature and light intensity on various species of algae (green, blue-green algae, red algae, brown algae, phytoplankton, and seaweed) in terms of algal growth was conducted [46]. This study revealed the optimal temperature range of 20°C to 30°C for different algae species growth and found that temperature and light were vital growth factors in the form of photon flux density. The study indicated that algae growth was restrained by shading light. After eliminating the shading light materials, algae can continue its rapid growth [46].

A previous paper has provided extensive coverage of the past and current situation of algal blooms, emphasising how climate change affects the marine planktonic system globally [42]. The authors described the linkage among several environmental factors that undergo alterations when pressured by climate change. These factors are temperature, stratification, light, ocean acidification, precipitation-induced nutrient inputs, and grazing. Besides, an analysis by the United States Environmental Protection Agency (EPA) in 2013 has summarised the effect of climate change on the occurrence of HABs via an assortment of mechanisms such as warmer water temperature, changes in salinity and rainfall pattern, rise in carbon dioxide concentration, coastal upwelling and increase in sea level.

Lastly, research on water quality determines the chemical and physical characteristics of water bodies and identifies the likely pollution sources that reduce water quality. This can be an indicator to show the growth of algae since Chl-a belongs under the biological factor (BF) in a water quality study. Table 1 summarises the most measured qualitative parameters in a water quality study [47].

B. MONITORING AND PROFILING WATER QUALITY FOR GROWTH OF ALGAE

Automated WQMS for fields such as aquaculture enable the industry to decrease catastrophic losses, lower production cost and enhance product quality. A promising alternative by using sensors to ensure the successful integration of obtaining the needed data and predictive modelling development, and water quality control will be the key aspects in both water and fisheries management success. Sensors are generally inexpensive and allow remote measuring in real-time and with little human intervention. Aquatic ecosystems have a crucial role in preserving water quality and are a key indicator of the

TABLE 1. Commonly measured qualitative parameters [47].

Water Quality Parameter	Abbreviation	Unit
Chlorophyll-a	Chl-a	mg/L
Secchi Disk Depth	SDD	m
Temperature	T	°C
Coloured Dissolved Organic Matters	CDOM	mg/L
Total Organic Carbon	TOC	mg/L
Dissolved Organic Carbon	DOC	mg/L
Total Suspended Matters	TSM	mg/L
Turbidity	TUR	NTU
Sea Surface Salinity	SSS	PSU
Total Phosphorus	TP	mg/L
Total Nitrogen	TN	mg/L
Orthophosphate	PO_4	mg/L
Chemical Oxygen Demand	COD	mg/L
Biochemical Oxygen Demand	BOD	mg/L
Electrical Conductivity	EC	Ms/cm
Ammonia Nitrogen	$NH_{3\text{-}}N$	mg/L

suitability of the water for other usages and the reasons why the water authority has implemented various measures to profile the water quality. Moreover, since discoloured water can be a sign of polluted water caused by HAB, measures such as image processing, real-time monitoring via satellites or drones have been carried out on image data of polluted water areas. Due to the limitation of manual profiling, the task of modelling algal bloom has long been initiated. Based on the literature, algal bloom predictions are reliable and can be performed by monitoring (involving detection, prediction, and tracking) the mechanism of algal growth. Several techniques were developed for remote detection and identification of algal blooms. For example, in situ sensors and low-cost sensors were developed for algae detection [48]–[51], [47], [52]. These sensors replace human effort for water sampling as well as human observation (judgment).

A review of the development of in situ sensors to measure chlorophyll concentration was performed [51] to quantify and analyse freshwater and seawater phytoplankton in situ. The review outlined the enhancement of probe design, excitation light sources, detectors, and calibrations of in situ fluorometers. Numerous optical designs to increase the effectiveness of fluorescence measurement and the development of electronic technology to fulfill and enhance in situ measurement were discussed. A smart sensors network for in situ and continuous space-time monitoring of surface seawater bodies to evaluate water quality was presented [50] and this formed strong support to strategic decisions regarding serious environmental problems. Since internal distances are in two scenarios (less than 1 km and the maximum is 1.852 km (1 mile)), the authors have proposed two different probes i.e. A and B that can be integrated solutions for field coverage (surface etc.).

Besides, a low-cost sensor to solve the problem of high-priced commercially produced in situ fluorometers has been proposed [49] to perform the depth-resolved measurement of phytoplankton biomass by measuring in situ phytoplankton fluorescence. The sensor managed to log over 500 Chl-a readings where 10 measurements were taken every 10 minutes. The authors mentioned that the same approach could be used in an array to help detect HABs. Finally, a concept paper was proposed [52] on the Smart River Monitoring System for river sustainability. Inspired by these studies, additional modifications to the hardware to suit the low-cost concept for seawater monitoring system was studied and proposed in our previous work [3].

This current paper will extend the previous work [3] on the predictive modelling phase to explain both IoT for WQMS development and predictive modelling using data-driven methods, especially deep learning for coastal studies. Furthermore, this paper also discusses the factors that contribute to algal growth in coastal areas to address the gap and improve the conventional method of selecting salient features for algal growth.

C. DATA-DRIVEN ALGAL GROWTH PREDICTION METHODS

Currently, the algorithms employed for algal bloom prediction are mostly separated into data-driven and process-driven models. The process-driven models usually need several parameters, such as initial conditions and ecological variables. Even though the process-driven models are very precise in their predictions, they need comprehensive knowledge of the system [18] and are known to suffer from the uncertainty of kinetic coefficients utilised in such models. The complexity to get all the data during the simulation has limited the application of process-driven methods. However, various investigations have reported the effective usage of data-driven artificial intelligence-based methods. This is because data-driven models are usually based on computational intelligence and machine-learning techniques [53].

A machine-learning algorithm is employed to discover the connection between a system's inputs and outputs using a training dataset that represents all the behaviours found in the system. When the model is trained, an independent dataset can be used to test it to determine how well it can be generalised to unseen data. Nevertheless, since this involves the physical, chemical, and biological processes and the interaction among them, to properly model and predict algae blooms in such a complex system is quite challenging. The occurrence of water pollution or the eutrophication phenomenon with the algae mechanism itself is a complex function of all the possible influencing factors [54]. These limitations, nonetheless, may be addressed through machine learning or artificial intelligence to gain insight into the algal community. The most common machine learning methods applied in algal prediction are ANN, SVM, DT, RF, and regression method.

A study in 1997 was the first to conduct modelling of algal blooms in freshwaters using ANN [55]. Back-propagation was used during training where the inputs were observable water quality parameters whereas the outputs were the biomass quantities of specific algal groups [55]. Next, another study forecasted algal growth by increasing the sample data size by merging in-situ and remote sensing data [56]. An ANN was employed to build the empirical data-driven models using water quality data, meteorological data, and bloom grade data gathered via remote sensing and in-situ monitoring [56]. The same work was conducted in another study [57] that predicted algal bloom using Extreme Learning Machine (ELM) models at artificial weirs [57]. An SVM-based prediction was proposed [15] to understand and predict a dynamic algae population transformation in freshwater reservoirs and to address the high complex nonlinearity that required only a small number of samples but produced a high degree of prediction accuracy [16]. Throughout the study, SVM was shown to be better in the prediction of algal growth and could cater to the non-linear aspect. SVM was also better than ANN in some cases if observed from the performance aspect. This could be because SVM is recognised as a strong predictor since it has a higher chance to attain the globally optimum solution in comparison to a weak predictor such as ANN that is often trapped in a local minimum [58]. Based on the observation, the performance still depends on the variable or features that are used. Furthermore, a fluctuating result was observed, which highlighted the highly non-linear and dynamic algal ecology data despite the water sources. Regardless of its success, SVM still has problems in handling noisy datasets and it does not work very well for a large dataset [59], which could be very challenging for a multivariate time series predictive problem. Several other works have been conducted in the past which utilised other machine learning methods [14], [60]–[62].

Even though for data-driven models only the input and output of the model must be determined, the prevailing basic machine learning models cannot efficiently extract features of multi-factor timing data and most of the models do not reflect the temporal characteristics of the data. An alternative approach to capture the temporal aspect for forecasting algal dynamics is therefore necessary. The use of time series forecasting might be able to address the problems.

The key difference between the time series problem and traditional prediction problem is that the data points in traditional prediction such as classification are assumed to be independent of one another. In contrast, in time series, the data points possess a temporal nature. The time dimension gives an explicit ordering to the data points that must be maintained since they can offer extra or crucial information to the learning algorithms [63] and this cannot be efficiently learnt by basic machine learning. Classical time series statistical forecasting models such as Auto-regressive Integrated Moving Average (ARIMA) and its variants (autoregressive models (AR), moving average (MA) and autoregressive and moving average (ARMA)) can be identified as frequently employed for forecasting methodologies. Although these models can capture temporal behaviour and produce acceptable forecasts for linear time series data, they are not appropriate for analysing non-linear data. These methods generally assume certain distribution or function form of time series, which renders them incapable of capturing complex underlying non-linear relationships and reflecting reality. Moreover, most of them disregard the reliance between variables when addressing multivariate time series, which lowers the forecasting accuracy.

Fortunately, recent studies have shown that deep learning with time series models (e.g., Recurrent Neural Networks (RNN) and Long Short-term Memory (LSTM)) can provide better accuracy in predictions than traditional machine learning models and traditional time series models because of their capability to persist information, tackle non-linearity and recognise temporal relationships. RNN and its variants, especially LSTM, show a good performance in exploiting long-term dependencies and managing non-linear dynamics [38]. Only very recently have time series been applied with deep learning for algae prediction and very few studies have discussed RNN or LSTM [29] and its improvement, especially for the coastal dataset, in algal growth prediction. Therefore, this concept paper takes the initiative to prove and observe the capability of deep time series, especially LSTM, in capturing temporal-algal dynamic behaviour for coastal studies.

Deep learning is a subfield of machine learning regarding algorithms inspired by the structure and function of the brain termed ANN. Besides, several kinds of deep learning models are usually utilised in time-series forecasting, for instance, RNN and its variant LSTM. RNN has been suggested to elucidate the dynamics [64], [65]. It is a network with feedback connections from the hidden and output layers to the preceding ones, through which the dynamics of sequential data can be recorded, and the memories of the prior patterns are kept via cycles in the network.

LSTM is an RNN architecture created to model temporal sequences and its long-range dependencies make predictions according to past data, hence, it is more precise compared to conventional RNNs. LSTM does not employ the activation function within its recurrent components, the stored values are not altered, and the gradient is not inclined to disappear during training like RNN.

LSTM has been evolving and is applied in many fields, especially in time series forecasting. Nonetheless, for algae prediction, only a few studies have adopted the LSTM algorithm. For example, the LSTM model was employed for algal bloom prediction on a newly constructed WQMS on 16 rivers [29]. Besides, a study using other time series non-linear models has been conducted [66] to improve the error caused by time series. LSTM has showcased excellent performance in other water sources, mainly river [24]–[26]. In a coastal study, deep time series was utilised via enhanced RNN [67]. Hence, this paper proposes LSTM as predictive modelling and examines the performance using the designed dataset and proposed approach in selecting features for this study. The in-depth explanation of LSTM as the proposed method used to accomplish the research objectives will be provided in Section III.

III. RESEARCH FRAMEWORK AND METHODS

As a revision to our previous work [3], this paper presents an enhanced and more detailed predictive modelling stage that has not been discussed yet. The complete framework is presented in Fig. 1.

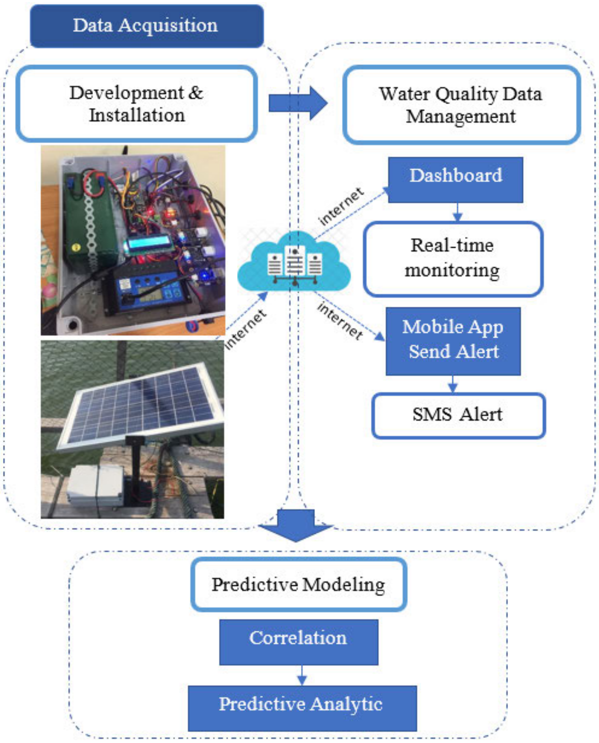

FIGURE 1. Complete framework of coastal WQMS [3].

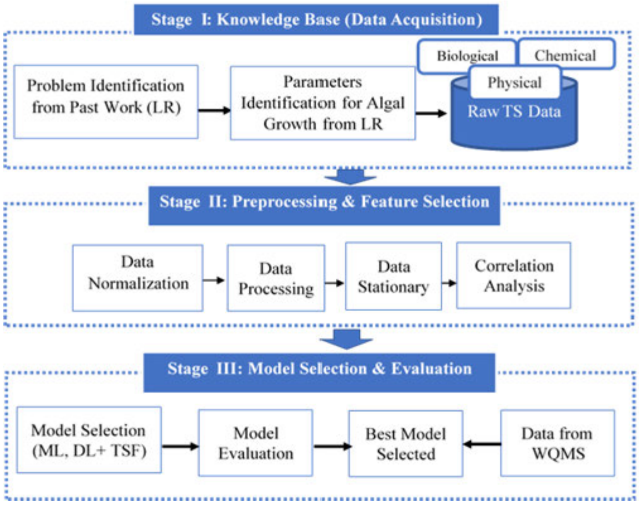
Stage I: Knowledge Base (Data Acquisition)
Problem Identification from Past Work (LR)
Parameters Identification for Algal Growth from LR
Biological
Chemical
Physical
Raw TS Data
Stage II: Preprocessing & Feature Selection
Data Normalization
Data Processing
Data Stationary
Correlation Analysis
Stage III: Model Selection & Evaluation
Model Selection (ML, DL+ TSF)
Model Evaluation
Best Model Selected
Data from WQMS

FIGURE 2. Predictive modelling framework.

A. STAGE 1: KNOWLEDGE BASE (DATA ACQUISITION)

In this stage, the problems of this research topic must be identified first to find and collect the right data and determine the right method to tackle the problems. Previous studies have revealed a research gap in tackling the issues of highly non-linear, uncertainty and complexity due to the dynamic behaviour of algae aquatic ecosystems [68]. Another gap concerns the way features or the parameters (factors of algal growth) are chosen.

This paper opines that the method of selecting the features and the features themselves are important in improving the prediction performance as proven in past research [56]–[59]. Furthermore, the dynamic problem itself originates from the features level. Despite only focusing on the algorithm level in tackling the dynamic issues, this paper will further investigate the preparation and designing of the dataset at the features level. This is because features selected in past works were mostly based on domain knowledge or were random [60] where most researchers considered all the features to be important and had no specific feature selection method. To address the issues, this paper has proposed a combination of knowledge based on the literature, and the features are then inspected using the feature selection method at the features level, as shown in Fig. 2.

Next, at the algorithm level, there is a gap for the coastal dataset where the use of deep learning with time series has the least investigation performed, especially using LSTM. For coastal studies, only one study [67] utilised deep time series using enhanced RNN. To the best of the authors' knowledge, LSTM alone with an improved feature selection method in one study has never been applied for coastal studies. The question of whether LSTM can still outperform other algorithms using coastal datasets remains a grey area to be investigated.

Based on these gaps, the proposed method was compared and studied from the literature. After problem formulation, suitable and important parameters were identified using knowledge extracted from the literature and feature selection method to further inspect and validate the importance of the predictor or features. Several parameters (features) that were reviewed could be categorised as biological factor (BF), physical factor (PF), chemical factor (CF), and meteorological factor (MF), as listed in Table 2.

The summary of analysis and categories extracted based on the domain knowledge discussed in past literature is provided in Table 2. After further analysis, features from the dataset were chosen based on these categories. For dataset, the monthly/biweekly water quality monitoring data gathered by the Hong Kong Environmental Protection Department were utilised for modelling where the data were set up and designed according to certain guidelines [13]. The data collection in this paper was based on individual indicators (water parameters) from the most weakly flushed monitoring station, TM3.

Nine water parameters or indicators were considered [13] as the target variables. The parameters were Chl-a (μg/l), total inorganic nitrogen, TIN (mg/l), orthophosphate, PO4 (mg/l), total phosphorus, TP (mg/l), water temperature, temperature (°C), DO (mg/l), Secchi disc depth, SD (m), daily solar radiation, SR (MJ/m2), and daily average wind speed, WS (m/s). Five years and 5 months (Jan 1997 to June 2002) of data for training and 2 years and 5 months (July 2002 to 7 Dec 2004) of data for testing were used.

To improve the features, this research differed from a previous work [13] whereby the features in this study had an

TABLE 2. Categorical variables.

Abbreviation	Variable	Factor Category
Chl-a	Chlorophyll-a	Biological Factor **(BF)**
BC	Bloom Cases (Incident)	
SGR	Specific Growth Rate	
WT	Water Temperature	Physical Factor **(PF)**
Salin	Salinity	
DO	Dissolved Oxygen	
Turb	Turbidity	
pH	pH	
SD	Secchi Disk Depth	
SS	Suspended Solid	
DC	Depth Code	
FI	Freshwater Inflow	
EV	Estuarine Velocity	
SRT	Salinity Recovery Time	
TIN	Total Inorganic Nitrogen	Chemical Factor **(CF)**
PO_4	Orthophosphate	
TP	Total Phosphorus	
TN	Total Nitrogen	
AN	Ammonia Nitrogen	
NO_2-N	Nitrite Nitrogen	
NO_3-N	Nitrate Nitrogen	
COD	Chemical Oxygen Demand	
Si	Silica	
Hg	Mercury	
Pb	Lead	
Zn	Zinc	
Al	Aluminium	
Rf	Rainfall	Meteorological Factor **(MF)**
T_{min}	Minimum Temperature	
T_{avg}	Average Temperature	
T_{max}	Maximum Temperature	
Hum	Humidity	
SR	Daily Solar Radiation	
WS	Daily Average Wind Speed	

additional length of data until 2018 (according to availability). Besides, via knowledge-based extraction, several other parameters were determined and correlation analysis for feature importance inspection was performed. The additional length of data and exclusion of some variables following the domain knowledge in the literature were to strengthen each factor and for comparison purposes. Based on past research on identifying important algal growth factors (Section II. A), investigation on PF (refer to Table 2) such as turbidity, DO, and other vital factors is still lacking [13]. Hence, more features under the PF were included. Additional physical features common in water quality studies, such as salinity, turbidity, pH, suspended solids, and total nitrogen along with some CF variables were also included, as listed in Table 3, as the final features used for modelling.

As explained previously, water quality study parameters, especially PF and CF, are usually used to indicate algal growth. The identification of these parameters was also based on WQMS's available probe. From here, knowledge-based extraction or selection of features was conducted, and the chosen features were further inspected using correlation analysis in the next stage, as illustrated in Fig. 2.

TABLE 3. Additional dataset design and description.

	Variable	Category
Chl-a	Chlorophyll-a	Biological (BF)
Salin	Salinity	Physical (PF)
DO	Dissolved Oxygen	
Turb	Turbidity	
pH	pH	
SD	Secchi Disk Depth	
SS	Suspended Solid	
Wtemp	Water Temperature	
TIN	Total Inorganic Nitrogen	Chemical (CF)
PO_4	Orthophosphate	
TP	Total Phosphorus	
TN	Total Nitrogen	
AN	Ammonia Nitrogen	
NO_2-N	Nitrite Nitrogen	
NO_3-N	Nitrate Nitrogen	
Si	Silica	

Next, experiments were designed to investigate the effect and relation between all the factors with the amount of Chl-a as the indicator of total algal biomass. For comparison purposes, the MF was purposely excluded in the initial experiment. This was to observe the results with and without MF since recent scientific research related climate change with the outbreak of algal growth. This paper wanted to investigate if the exclusion of MF could still improve the current performance using the benchmark dataset from the literature. Such an arrangement gave an observable comparison from the dataset used. The dataset was divided using a 70:30 ratio of training and testing datasets as recommended in previous work [13] but with several modifications as explained earlier.

The guideline of sampling rate was followed exactly as stated in the previous study [13] to conduct a fair comparison in terms of the division of data. Moreover, after the study, it was found that not all the methods such as cross-validation could be directly applied to time series data since there was a temporal dependency between the observations, and this relation must be preserved during testing [69]. Hence, this study had 1,556 observations with 16 attributes or features from the biological, physical, and chemical categories as the predictors. Meanwhile, date was included as an attribute to preserve the temporal order. The target variable in the dataset was Chl-a as the strong indicator to predict the presence of algal growth. The Tolo Harbour dataset description is provided in Table 4 while Table 5 presents a brief description of the sampling rate.

TABLE 4. Tolo harbour dataset description.

Date Range	Factor	Update	Instances	Attributes
1986–2018	BF, PF, CF	Sporadic	1556	17

B. STAGE 2: PRE-PROCESSING AND FEATURE SELECTION

This stage was divided into two major steps i.e. (a) data pre-processing from data collection and design and (b) feature

TABLE 5. Sampling rate description.

Dataset	Date Range	70:30
Train Model	Start	Jan 1986
Train Model	End	April 2008
Test Model	Start	May 2008
Test Model	End	December 2008

selection using correlation analysis. The details of each step are discussed as follows:

1) DATA PRE-PROCESSING

Data cleaning for time-series data is different from other data as there are special ways to efficiently handle time-series data. Data pre-processing was divided into:

a: MIN MAX NORMALISATION

Min Max Normalisation is a rescaling of data from the original range so that all the values are within the range of 0 and 1. Equation (1) is used to normalise the dataset, which scales all the data in the range of [0 1]:

$$X_{new} = \frac{X}{X_{max}} \tag{1}$$

where X is daily observation of time series data obtained as described in dataset design that comprises 17 features from the range date, X_{max} is the highest value of observation of a particular feature, and X_{new} is obtained after normalisation.

b: DATA PROCESSING

The data usually comes in a very untidy form, contain a lots of noise and missing data. To impute the missing values, linear interpolation is used. Incomplete data set usually caused bias due to differences between observed and unobserved data. A direct but unreliable approaches to deal with this problem is to ignore the missing data and to discard those incomplete cases from the data set. This approach is generally not valid for time-series prediction, in which the value of a system typically depends on the historical time point. A method was adopted based on reference [70] that successfully compared and proved that linear interpolation method provides a very good fit to the time series data if compared with other method. Besides that, time-series data was framed as a supervised learning problem in this stage. The detail is discussed in Model Selection section.

c: DATA STATIONARY

The next step was to find out whether a given series was stationary. A stationary time series' statistical properties (e.g., mean, variance, and autocorrelation) are constant over time. There are two ways to check data stationary, namely via (i) visual test and (ii) statistical test. This research employed unit root statistical tests. Unit root denotes that a given series' statistical properties are not constant with time, which is the condition for stationary time series. An explanation using equations 2–4 is provided as follows:

Suppose there is a time series:

$$yt = a \times y_{t-1} + \varepsilon_t \tag{2}$$

where y_t is the value at the time instant t and ε_t is the error term. To calculate yt the value of y_{t-1} is needed, which is:

$$y_{t-1} = a \times y_{t-2} + \varepsilon_{t-1} \tag{3}$$

If (3) is applied for all the observations, the value of y_t will be:

$$y_t = a^n \times y_{t-n} + \sum \varepsilon_{t-1} \times a^i \tag{4}$$

If the value of a is 1 (unit), as in the above equation, then, the predictions will be equal to y_{t-n} and the sum of all the errors from $t - n$ to t, which implies that the variance will increase with time. This is known as unit root in a time series. For a stationary time series, the variance should not be a function of time. The unit root tests determine the existence of unit root in the series by checking if the value of a $= 1$. Augmented Dickey-Fuller (ADF) is the most utilised unit root stationary test.

d: AUGMENTED DICKEY-FULLER (ADF)

This test is a popular statistical test used to find out the existence of unit root in the series and it helps to determine if the series is stationary. The null hypothesis of this test is the series has a unit root (value of a $= 1$) while the alternate hypothesis is the series has no unit root. If the null hypothesis is rejected, the series is non-stationary. Hence, the series can be linear or difference stationary. To test for stationarity, if the test statistic is less than the critical value, the null hypothesis (the series is stationary) can be rejected. On the other hand, if the test statistic is higher than the critical value, the null hypothesis is rejected (i.e., the series is not stationary). If the ADF statistic employed in the test is a negative number, the more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

e: FEATURE SELECTION USING CORRELATION ANALYSIS

Correlation analysis [71] is a technique for investigating the relationship and measuring the strength between two quantitative, continuous variables to represent their interdependencies. Pearson's correlation is the most common one where the coefficient scales range from -1 to 1, where 1 represents the strongest positive correlation while -1 represents negative correlation and 0 indicates no correlation at all. For instance, a positive correlation means that if feature A increases, feature B will also increase, or if feature A decreases, feature B also decreases. A and B have a linear relationship. Meanwhile, a negative correlation implies that if feature A increases, feature B will decrease and vice versa. Highly positive correlation features can range from 0.5 to 0.7 while a strong and perfect positive correlation is signified by a correlation score value of 0.9 or 1.0. Data and feature correlation is a key step in the feature selection phase of the data pre-processing stage, especially if the data type for the features is continuous. An experiment using the algae ecological time-series dataset was performed using this algorithm to observe the correlation between the 16 features which were modelled based on the PF, CF, and BF that were determined from the literature via knowledge-based extraction. Hence, using Pearson's correlation, this research could validate the significance of the features selected at Stage 1.

C. STAGE 3: MODEL SELECTION AND EVALUATION

This stage was divided into three major steps, namely (a) model selection, (b) evaluation, and (c) evaluation on a real case study using WQMS data. The details of each step are discussed as follows:

1) MODEL SELECTION–LSTM

Based on past literature, SVM, DT, RF, ANN, and Multiple Linear Regression (MLR) were usually used and the performance could be improved further. However, basic machine learning has been pointed out as inefficient in handling the dynamics of ecological data. Recently, deep learning with time series capabilities such as RNN and LSTM has shown outstanding performance to capture non-linear and temporal behaviour but it has not been applied on coastal datasets.

The LSTM neural network, a variant of RNN, was initially introduced to solve the vanishing/exploding gradient issue, which causes training divergence in RNN. Like RNN, LSTM is very capable of capturing the dynamic features via cycles in the graph [73]. Furthermore, LSTM shares the same parameters (i.e., network weights) across all time steps that significantly lowers the number of unknowns [74]. LSTM was introduced in 1997 [74]. The key components of the LSTM network are its memory cells, which differentiates it from the traditional RNN. The input gate, the output gate and the forget gate in the memory cells are the three types of multiplicative units existing in the LSTM model structure. These gates alter the state of the memory cells based on several steps [74]. First, by activating the input gate, as the latest data enter, the input message can be accumulated to the cell. Second, by activating the forget gate, the former cell states are to be abandoned during the procedure. Finally, the output gate is in charge of determining if the latest cell output is propagated to the final state, as illustrated in Fig 3.

The architecture of LSTM has made LSTM networks suitable to classify, process, and make predictions based on time series data since lags of unknown duration can exist between crucial events in a time series. LSTM was developed to handle the vanishing gradient issue that could occur when training traditional RNNs. The prediction of the algal growth can be made using LSTM and is computed as follows [74]:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{5}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{7}$$

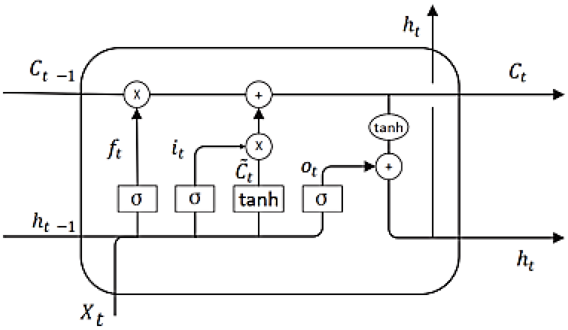

FIGURE 3. **LSTM internal architecture.**

$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{8}$$

$$c_t = f_t.c_{t-1} + i_t.\tilde{c}_t \tag{9}$$

$$\mathrm{h_t} = \mathrm{o_t}.\sigma_\mathrm{h}(\mathrm{c_t}) \tag{10}$$

where the initial values are $\mathrm{c_{0=}0}$ and $\mathrm{h_{0=}0}$ and the operator "." denotes the element-wise product. The subscript *t* indexes the time step.

For each variable, the definition is as follows:

- $x_t \in R^d$: input vector to the LSTM unit
- $f_t \in R^h$: forget gate's activation vector
- $i_t \in R^h$: input/update gate's activation vector
- $o_t \in R^h$: output gate's activation vector
- $h_t \in R^h$: hidden state vector also known as output vector of the LSTM unit
- $\tilde{c}_t \in R^h$: cell input activation vector
- c_t: cell state vector
- $W \in R^{h\times d}$, $U \in R^{h\times h}$ and $b \in R^h$: weight matrices and bias vector parameters which need to be learned during training

where the superscripts *d* and *h* denote the number of input features and the number of hidden units, respectively. The activation functions definitions are as follows:

- σ_g: sigmoid function
- σ_c: hyperbolic tangent function

To summarise, the gates contain sigmoid activations. A sigmoid activation is like the tanh activation. Rather than squishing values between −1 and 1, it squishes values between 0 and 1. This helps to update or forget data because any number multiplied by 0 is 0, causing values to disappear or be "forgotten."

Information from both the previous hidden state and the current input is passed through the sigmoid function. The values are between 0 and 1. A value closer to 0 means to forget, whereas closer to 1 indicates to keep. Input gate is to update the cell state.

Therefore, based on past literature, this paper proposed an LSTM model to be used in this research because LSTM continuously outperformed the other methods in terms of the performance measure root-mean-square-error (RMSE), mean-absolute-error (MAE), etc.) in rivers and other fields (compared in the results section). Since our framework is designed for seawater or coastal studies, finding a suitable algorithm for coastal predictive modelling is crucial as the performance might not be the same because the characteristics of each water source are varied as mentioned before. Moreover, in machine learning itself, the No Free Lunch Theorems [75] indicated that a general-purpose, common strategy is unfeasible. The only way one strategy can outperform another is if it is specialised to the structure of the specific issue under scrutiny [75]. In this study, to cater for all types of water sources, the indicator used must be considered. Hence, LSTM methods were compared to other algorithm performances using only the coastal dataset designed in this study. The experiment results would be the baseline and benchmark of this research work.

Choosing a good hyperparameter for deep learning models needs numerous experiments, which is a laborious and time-consuming task. Most scholars depend on their experience in selecting appropriate parameters for a deep neural network. Based on the dataset size, up to several days are needed to train a single model. Hence, a common way is a meticulous selection of limited values of the hyperparameters to train several models and then select the model that performs the best on a validation set. The description of the hyperparameter used in the LSTM [76] is provided in Table 6.

TABLE 6. LSTM hyperparameter description.

Parameter	Description
LSTM Cells	Number of LSTM memory cells that store the temporal information
Batch Size	Number of samples per gradient update
Activation Function	Type of nonlinear activation function
Optimiser	Type of optimiser to update weights during training

For LSTM initial benchmark and preliminary study, the parameter setup for LSTM involved the sequential model, epoch of 50, batch size of 72, only one hidden or dense layer and Adam as the optimiser. The batch size was set up according to the discussion in past works. In general, a batch size of 32 is a good starting point, together with 64, 72, 128, and 256. The optimal batch size was 72 after experimenting. Like the epoch number, the program implemented early stopping, as a callback was used to check at the end of every epoch whether the validation loss was no longer decreasing. Once it was found that there was strictly no decrease for 3 epochs consecutively, the training process was terminated. The maximum epoch was set up as 50. This is because it is a good practice to start with 50 epoch and it can be increased from time to time. Other methods' setups are listed in Table 7.

2) MODEL EVALUATION

The "loss" values are usually reported by deep learning algorithms. Technically, loss is a type of penalty for a poor prediction. More precisely, the loss value will be zero, if the model's prediction is perfect. As such, the goal is to minimise the loss values by gaining a set of weights and biases that minimises the loss. Besides loss, which is utilised by the deep learning algorithms, researchers frequently employ the RMSE, mean-square-error (MSE), and MAE to evaluate the prediction performances [29]. The goodness-of-fit measures used in this study were RMSE and MAE.

TABLE 7. Experimental setup.

Method	Parameter
MLR	–
SVM	Kernel: linear, poly, rbf, sigmoid C=1, gamma=0.5
DT	Depth=from 2 until 9
RF	–
ANN	Simple MLP Hidden Layer=1

DNN	Sequential Model
	Dense Layer=3
	activation='relu'
	optimiser='adam'
	verbose=2
	epochs=50

RNN	Sequential Model
	Simple RNN Model
	Dense Layer=2
	optimiser='adam'
	activation='relu'
	epochs=50
	batch_size=72,
	verbose=2

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{11}$$

RMSE is commonly utilised to measure the difference between the values predicted by a model and the actual values in the data or the square root of the mean/average of the square of all the error. The n is the total number of data, Y_i and $\hat{Y}_i$ are the actual and simulated data, respectively, and the average value of the related variable is signified by the 'bar' above the variable. Meanwhile, MAE measures the average magnitude of the errors in a set of predictions, regardless of their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all the individual differences have equal weight.

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|^2 \tag{12}$$

3) EVALUATION OF A REAL CASE STUDY USING A WQMS STUDY

This stage overlapped with Stage 1 where the development of the water monitoring system was conducted simultaneously with the study from the literature. However, the development of the water station (WQMS) is discussed in this stage. Data from the data acquisition process served as the dataset for the real case study and were applied using the proposed model. The architecture of the system during data collection could be separated into two major sections, i.e. hardware and software. These sections existed in three phases, namely kit development, installation, and transmission of data to the server. In the case of hardware, the kit development phase consisted of all the necessary sensors, whereas the software section encompassed the development of a water monitoring programme using the C/C++ language in Arduino Bluno. The sensors to build the water station are summarised in Table 8.

TABLE 8. Sensor specification.

Sensor/Equipment	Sensor/Equipment
Arduino Bluno	Temperature
Expansion Shield	Turbidity
pH	Solar Panel and Controller
DO	Battery
TDS	Jumper Wire
EC	Waterproof Box

There is an antenna outside the box to send or receive data via the 3G mobile network. After kit development was completed, kit installation was remotely carried out at the fish-farm site. Since sufficient power must be supplied to the sensors and Arduino, a rechargeable battery with a solar panel was also fixed at the site. After confirming that the solar panel was linked to the battery via the solar panel controller, the panel box was closed and firmly tied beside the solar panel. Four units of panel box were developed and fixed at different sides of the fish farm to monitor the algal population via profiling of the water quality parameters. Based on our previous work [3], the sensors were in the process of being improved to use a water pump instead of direct contact with water due to coralline algae [3]. Fig. 4(b) shows the broken probe. Hence, the implementation of the proposed enhancement architecture from the previous WQMS is still ongoing, as displayed in Fig. 4(c). After we have completed the enhancement of the water station development, the real case study will be ready for data collection and at least 6 months of data must be integrated with the proposed predictive model discussed in this paper. Overall, the complete implementation of the WQMS framework [3] can be illustrated as in Fig. 4.

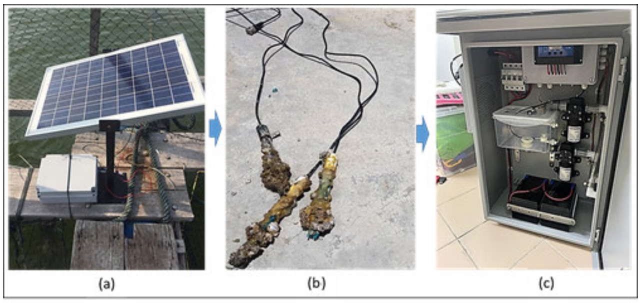

FIGURE 4. The complete implementation of the previous WMQS [3] from (a) Initial development, (b) After installation and deployment until (c) Enhancement.

IV. ANALYSIS OF RESULTS AND DISCUSSION

This section discusses in detail the development of the predictive modelling in the last stage. The discussion comprises the exploratory data analysis (EDA) and the results of comparing different algorithms such as SVM [77], DT, RF [78], ANN, Deep Neural Network [79], MLR [80], RNN, and LSTM [74] to the dataset. The chosen method would be used as the final prediction method and as a proof of concept. The results of some of the preliminary EDA are discussed in Section A and the comparison model results are discussed in Section B.

A. PRELIMINARY RESULTS OF EDA

It is vital to make sense of the dataset and clean it to attain success and improve the prediction algorithm. First, the data must be understood via the EDA method [82] because it assists in creating the logical approach to solve the research problem. Besides, it helps to determine issues such as the presence of outliers in the dataset through visual observation. However, it will be complicated when dealing with datasets that have hidden properties, for example, time series datasets. The time series datasets are a kind of data that are ordered chronologically and require special attention for managing intrinsic elements such as trend and seasonality. Therefore, this paper took the initiative to also discuss the results of EDA and the pre-processing part.

Since the interest of our prediction was Chl-a as the strong indicator of algae presence, Chl-a concentration was the target variable in the prediction and must be analysed in detail from the beginning. Initially, for the EDA, the time series was visualised using the average amount of Chl-a over time, as presented in Fig. 5.

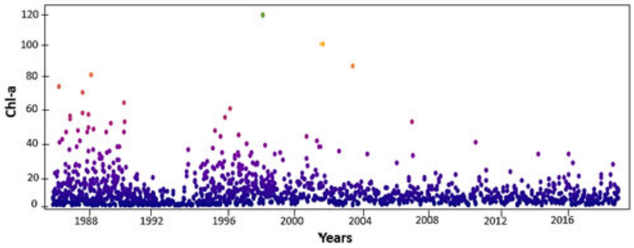

FIGURE 5. Average amount of Chl-a over time.

The higher concentration of Chl-a observed between the years 1996 to 2005 can be indicated as outliers (Fig. 5). An outlier is an observation that substantially varies from other observations of the same feature. If a time series is plotted, in general, outliers are the unexpected spikes or dips of observations at given points in time. A temporal dataset with outliers possesses several characteristics and has either a systematic pattern (deterministic) or some variations (stochastic). Only a small number of data points are outliers. The outliers considerably differ from the rest of the data [81]. The outliers of Chl-a concentration (Fig. 5) have a value of more than 60 until greater than 120. This could be further investigated using the Statistical Summary, as shown in Table 9.

There are various methods in statistics to make the detection of outliers easy such as interquartile range and standard deviation. From the statistical summary, a huge difference

TABLE 9. Statistical summary of the variables.

Variable	Min	Mean	Std	25%	50%	75%	Max
Chl-a	0.20	9.33	11.09	2.70	6.00	12.0	130
Salinity	14.5	30.84	2.22	30.3	31.3	32.1	34.6
DO	0.10	6.87	2.50	5.40	6.90	8.40	17.0
Turb	0.10	3.98	3.58	1.60	3.00	5.40	51.0
pH	7.20	8.23	0.29	8.00	8.20	8.40	9.30
SD	0.30	2.02	0.78	2.00	2.50	5.00	5.00
SS	0.50	3.32	4.46	1.50	2.20	3.50	93.0
W/Temp	11.6	23.6	4.44	19.7	24.3	27.6	32.0
TIN	0.01	0.14	0.14	0.04	0.09	0.19	1.17
PO_4	0.00	0.02	0.03	0.00	0.01	0.03	0.22
TP	0.02	0.07	0.09	0.02	0.05	0.09	2.40
TN	0.11	0.56	0.38	0.28	0.45	0.73	3.97
AN	0.00	0.10	0.10	0.03	0.06	0.13	0.99
NO_2-N	0.00	0.01	0.02	0.00	0.00	0.01	0.35
NO_3-N	0.00	0.03	0.05	0.00	0.00	0.01	0.60
Si	0.05	0.79	0.63	0.32	0.68	1.10	7.10

was noted between the 75th percentile and the max values of certain fields such as "Chl-a", "Turbidity", "Suspended Solid", etc. Thus, this observation implies the presence of extreme values or outliers in the dataset. The same conclusion was made looking at the time series graph of the observed Chl-a distribution from the year 1986 until 2016. The daily, monthly, and yearly Chl-a amount is shown in Fig. 6. Once the data were cleaned by removing the outliers, the graph of the Chl-a series prediction was produced, as illustrated in Fig. 7.

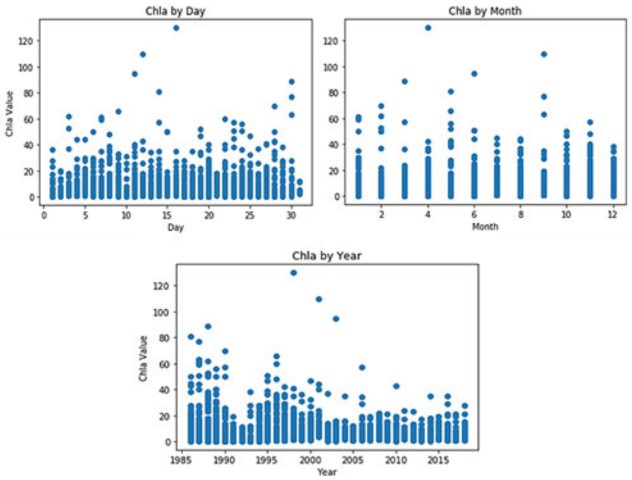

FIGURE 6. Chl-a concentration distribution by day, month, and year.

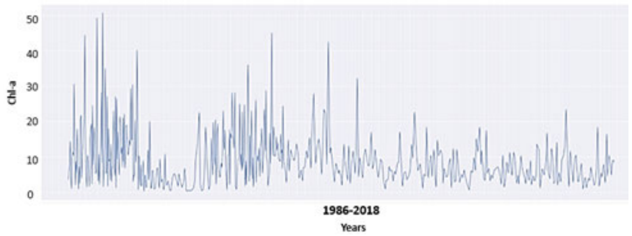

FIGURE 7. Chl-a concentration distribution from 1986–2018 after outliers removal.

As shown in Fig. 7, Chl-a was highly uncertain and non-linear. It was observed that the data were hardly analysed in terms of trend or seasonality. The data were quite erratic, having neither an upward nor downward trend. Hence, this time series data showed an irregular trend, whereby the components were mostly unpredictable. However, for prediction, the key aim is to model all the components to the point that the only component that is still unexplained is the random component. This is where decomposition techniques help to extract trend, seasonality, and error/irregular components of a time series dataset. The next step was to find out if a given series was stationary. The results for the stationary test using ADF is listed in Table 10.

TABLE 10. Augmented dickey-fuller test.

Results of ADF for Chl-a	
Test Statistic	–5.497272
p-value	0.000002
Critical Value (1%)	–3.434560
Critical Value (5%)	–2.863399
Critical Value (10%)	–2.567760

In Table 10, the test statistic value < the critical value, implying that the series for Chl-a is already stationary. The test was also applied to other variables that yielded the same results, whereby the test statistic value < the critical value and was negative. A time series can be easily modelled if it is stationary. Statistical modelling techniques assume or need the time series to be stationary to be effective.

B. FEATURE SELECTION USING CORRELATION ANALYSIS

Next, the correlation between the features was checked. A quick way to check is by visualising the correlation matrix as a heatmap. Fig. 8 shows the correlation analysis results using Pearson's correlation matrix heatmap plots.

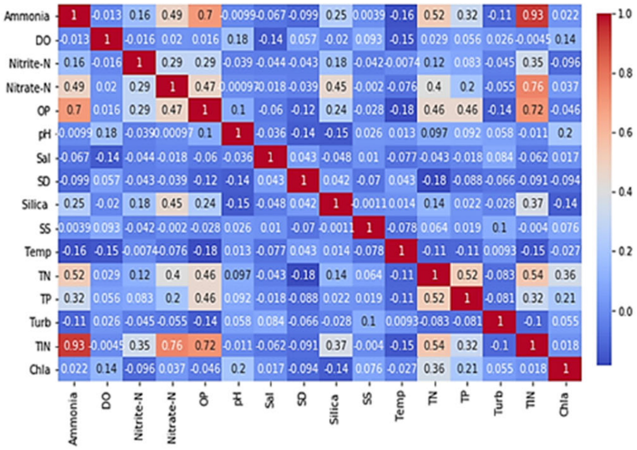

FIGURE 8. Correlation analysis using pearson's correlation heatmap.

According to Fig. 8, some variables are highly positively and negatively correlated while some are not. A high positive correlation was seen for ammonia and total inorganic nitrogen (TIN) with 0.93. Besides, TIN was repeatedly shown to be highly positively correlated with other variables. Some highly negative correlations were between the variable temperature and ammonia, DO, and orthophosphate. Negative correlations were seen when temperature increased while suspended solid decreased, ammonia decreased, and DO increased.

Despite only some variables were considered to have a higher correlation with Chl-a, the correlation matrix did not exclude any of the selected variables. Thus, this is considered a list of good features for our predictive model and the features can be used as the input to our model.

C. COMPARISON OF PREDICTION MODEL

This section presents the results, analysis and discussion arranged according to the experimental design in Section III for the initial experiment. The results are shown in Table 11 while Table 12 presents the comparison of LSTM performance from the literature review. Table 11 lists the experimental findings using the designed dataset with various algorithms reviewed in the literature. LSTM outperformed the other basic algorithms and efficiently addressed the challenge of the dataset.

TABLE 11. Comparison model performance evaluation for testing data.

Method	MAE	RMSE	MSE
SVM	0.4772	0.5923	0.3508
DT	0.4840	0.5940	0.3528
RF	0.4453	0.5686	0.3233
MLR	0.4477	0.5632	0.3171
ANN	0.5607	0.6359	0.4044
TSP	0.4772	0.5923	0.3508
RNN	**0.0594**	**0.0696**	**0.0048**
DNN	**0.0319**	**0.0440**	**0.0019**
LSTM	**0.0256**	**0.0360**	**0.0013**

TABLE 12. Comparison model performance of our approach and LR.

Author(s)	Method	Source	MAE	RMSE	MSE
[23]	LSTM	River	NP	0.0486	NP
[28]	LSTM	River	NP	7.67	NP
[25]	Merge-LSTM	River	NP	0.0459	NP
[67]	DA-RNN	Coastal	0.790	1.269	NP
[12]	SVM	Coastal	0.926	1.583	NP
Ours	**SVM**	**Coastal**	**0.477**	**0.592**	**0.351**
Ours	**RNN**	**Coastal**	**0.091**	**0.083**	**0.008**
Ours	**DNN**	**Coastal**	**0.032**	**0.044**	**0.002**
Ours	**LSTM**	**Coastal**	**0.026**	**0.036**	**0.001**

*NP=Not provided

In Table 12, the RNN or LSTM model performance from the literature is not only limited to coastal datasets. This is to show that LSTM has continuously displayed impressive performance despite using various types of water sources with a high reduction of error. This is also true when compared with the results of the past study [13]. With the changes in the selection of features and by using longer data, our approach of using SVM has produced much better results than the previous work [13]. This indicates that the right features are important with and without the temporality aspect. Even though SVM was not as efficient as the deep learning method, the results are better using the improved and selected features proposed in this paper. It can also be concluded that factors such as PF, BF, and CF are sufficient for predictive modelling using this dataset.

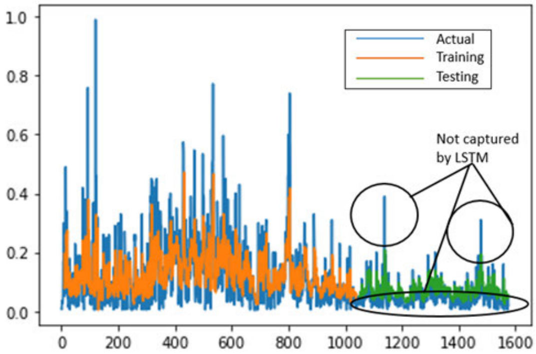

FIGURE 9. Model fitting of LSTM.

As shown in Fig. 9, LSTM can fit very well, which concurs with a previous study [21] which mentioned that LSTM is a type of neural network with powerful nonlinear fitting ability. This also shows that LSTM is better in predicting the dynamics of Chl-a concentration and continues improving during the testing phase and this includes the observation of the overfitting issue during training. The problems are addressed during testing. Nevertheless, LSTM has room for improvement as more attention is needed, especially in predicting the sudden high peak that is not captured by LSTM (refer to the circles in Fig. 9). This has motivated us to proceed with the improvement part and to further reduce the errors.

It can be concluded that PF, BF, and CF factors are enough as inputs to the development of predictive modelling. This research has further strengthened the justification that our proposed approach is better where LSTM with improvement through the selection of significant features is the best in handling non-linear, uncertain and dynamic data. This research has further proven that the improvement must be tackled from data level until algorithm level. Another insight is even though the features were tested to be significant using correlation analysis alone, the results somehow only showed the correlation between two variables of the 16 features. As mentioned in the review, one drawback of the correlation method is the missing of the sense of direction or cause and effect relationship. Hence, investigation on the relation of each feature could further improve the explanations.

V. CONCLUSION

Due to the issue of dynamics of algae that are highly non-linear and uncertain, robust predictive modelling that tackles from the end-to-end process is necessary. Selecting the right features are crucial in tackling the dynamic issues, and from the results, the algae ecology is dependent on the number and types of the features. Based on the discussion and analysis, it was observed that LSTM with the right features outperformed the other methods and grasped the temporal behaviour and tackled the dynamic issues. Besides, even though during this study the MF was excluded, and more CF and PF were included, this study outperformed the other studies. This indicates that the factors are dependent on the characteristics of the data to improve the prediction further.

Additionally, this paper has concisely presented a complete framework that discusses in detail both IoT and predictive modelling that consists of the main phases such as data acquisition, data management and lastly, predictive modelling. Later, the predictive model can be integrated into our system for future HABs prevention. Hence, with the suitable method that has been chosen during the predictive modelling stage, each phase has now been completed, which comprises all the phases in the framework, and overall has achieved all the objectives mentioned. For future work, the LSTM method can be improved further using the hybrid method with other suitable learning methods. Besides, the MF that might further improve the performance can be incorporated. To include MF, the discussion will be big in scope as it will include data segmentation, processing, and feature engineering. Finally, future research should investigate the relation of each feature that could enrich further the explanations.

ACKNOWLEDGMENT

This work is a collaborative research between the Universiti Sains Malaysia and Sg. Udang Fish Farm Community.

REFERENCES

[1] Y. F. Huang, S. Y. Ang, K. M. Lee, and T. S. Lee, "Quality of water resources in Malaysia," *Res. Practices Water Qual.*, T. S. Lee, Ed. Rijeka, Croatia: Books on Demand, 2015, pp. 65–94, doi: 10.5772/58969.

[2] D. M. Anderson, "Approaches to monitoring, control and management of harmful algal Blooms (HABs)," *Ocean Coast Manag.*, vol. 52, no. 7, p. 342, Jul. 2009, doi: 10.1016/j.ocecoaman.2009.04.006.

[3] N. A. P. Rostam, N. H. Ahamed, H. Malim, and R. Abdullah, "Development of a low-cost solar powered & real-time water quality monitoring system for Malaysia seawater aquaculture: Application & challenges," in *Proc. 4th Int. Conf. Cloud Big Data Comput.*, Virtual, U.K., Aug. 2020, pp. 86–91, doi: 10.1145/3416921.3416928.

[4] U.S. Geological Survey. (2020). *Biological Oxygen Demand (BOD) and Water*. [Online]. Available: https://www.usgs.gov/special-topic/water-science-school/science/biological-oxygen-demand-bod-and-water?qt-science_center_objects=0#qt-science_center_objects

[5] R. Ande, B. Adebisi, M. Hammoudeh, and J. Saleem, "Internet of Things: Evolution and technologies from a security perspective," *Sustain. Cities Soc.*, vol. 54, Mar. 2020, Art. no. 101728, doi: 10.1016/j.scs.2019.101728.

[6] F. Recknagel and W. K. Michener, *Ecological Informatics: Data Management and Knowledge Discovery*, 3rd ed. Cham, Switzerland: Springer Verlag, 2017, doi: 10.1007/978-3-319-59928-1.

[7] B. C. Patel, G. R. Sinha, and N. Goel, "Introduction to sensors," in *Advances in Modern Sensors*, G. R. Sinha, Ed. Bristol, U.K.: IOP, Nov. 2020, ch. 1, p. 367, doi: 10.1201/9781315218274.

[8] K. Davidson, D. M. Anderson, M. Mateus, B. Reguera, J. Silke, M. Sourisseau, and J. Maguire, "Forecasting the risk of harmful algal Blooms," *Harmful Algae*, vol. 53, no. 6, pp. 1–7, Mar. 2016, doi: 10.1016/j.hal.2015.11.005.

[9] J. H. W. Lee, Y. Huang, M. Dickman, and A. W. Jayawardena, "Neural network modelling of coastal algal Blooms," *Ecol. Modell.*, vol. 159, nos. 2–3, pp. 179–201, Jan. 2003, doi: 10.1016/S0304-3800(02)00281-8.

[10] J. Lu, T. Huang, and R. Hu, "Data mining on algae concentrations (Chlorophyll) time series in source water based on wavelet," in *Proc. 5th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Oct. 2008, pp. 611–616, doi: 10.1109/FSKD.2008.540.

[11] Y. Wang, Z. Xie, I. Lou, W. K. Ung, and K. M. Mok, "Algal Bloom prediction by support vector machine and relevance vector machine with genetic algorithm optimization in freshwater reservoirs," *Eng. Comput.*, vol. 34, no. 2, pp. 664–679, Apr. 2017, doi: 10.1108/EC-11-2015-0356.

[12] L. Wang, T. Zhang, X. Jin, J. Xua, X. Wang, H. Zhang, J. Yu, Q. Sun, Z. Zhao, and L. Zheng, "Multi-factor nonlinear time-series ecological modelling for algae Bloom forecasting," *Desalination Water Treat.*, vol. 122, pp. 91–99, Feb. 2018, doi: 10.5004/dwt.2018.22661.

[13] X. Li, J. Yu, Z. Jia, and J. Song, "Harmful algal Blooms prediction with machine learning models in Tolo Harbour," in *Proc. Int. Conf. Smart Comput.*, Nov. 2014, pp. 245–250, doi: 10.1109/SMARTCOMP.2014.7043865.

[14] L. Wang, T. Zhang, J. Xu, J. Yu, X. Wang, H. Zhang, and Z. Zhao, "An approach of improved dynamic deep belief nets modeling for algae Bloom prediction," *Cluster Comput.*, vol. 22, no. S5, pp. 11713–11721, Dec. 2017, doi: 10.1007/s10586-017-1460-9.

[15] J. A. McGowan, E. R. Deyle, H. Ye, M. L. Carter, C. T. Perretti, K. D. Seger, A. Verneil, and G. Sugihara, "Predicting coastal algal Blooms in southern California," *Ecology*, vol. 98, no. 5, pp. 1419–1433, May 2017, doi: 10.1002/ecy.1804.

[16] Z. Xie, I. Lou, W. K. Ung, and K. M. Mok, "Freshwater algal Bloom prediction by support vector machine in Macau storage reservoirs," *Math. Problems Eng.*, vol. 2012, pp. 1–12, Nov. 2012, doi: 10.1155/2012/397473.

[17] L. Wang, X. Wang, X. Jin, J. Xu, H. Zhang, J. Yu, Q. Sun, C. Gao, and L. Wang, "Analysis of algae growth mechanism and water Bloom prediction under the effect of multi-affecting factor," *Saudi J. Biol. Sci.*, vol. 24, no. 3, pp. 556–562, Mar. 2017, doi: 10.1016/j.sjbs.2017.01.026.

[18] M.-H. Bui, T.-L. Pham, and T.-S. Dao, "Prediction of cyanobacterial Blooms in the Dau Tieng Reservoir using an artificial neural network," *J. Mar. Freshwater Res.*, vol. 68, no. 11, pp. 2070–2080, May 2017, doi: 10.1071/MF16327.

[19] W. Tian, Z. Liao, and J. Zhang, "An optimization of artificial neural network model for predicting chlorophyll dynamics," *Ecol. Model.*, vol. 364, pp. 42–52, Nov. 2017, doi: 10.1016/j.ecolmodel.2017.09.013.

[20] T. Egerton, R. Morse, H. Marshall, and M. Mulholland, "Emergence of algal Blooms: The effects of short-term variability in water quality on phytoplankton abundance, diversity, and community composition in a tidal estuary," *Microorganisms*, vol. 2, no. 1, pp. 33–57, Jan. 2014, doi: 10.3390/microorganisms2010033.

[21] L. Zhang, Y. Cheng, Y. Niu, and J. Jiang, "Analysis and prediction of eutrophication for advanced warning of the water quality concerns in Gaoyou Lake," *Water Supply*, vol. 20, no. 1, pp. 186–196, Oct. 2019, doi: 10.2166/ws.2019.148.

[22] D. A. Caron, M.-È. Garneau, E. Seubert, M. D. A. Howard, L. Darjany, A. Schnetzer, I. Cetinić, G. Filteau, P. Lauri, and B. Jones, "Harmful algae and their potential impacts on desalination operations off southern California," *Water Res.*, vol. 44, no. 2, pp. 385–416, Jan. 2010, doi: 10.1016/j.watres.2009.06.051.

[23] A. J. Lewitus, R. A. Horner, D. A. Caron, E. Garcia-Mendoza, B. M. Hickey, M. Hunter, D. D. Huppert, R. M. Kudela, G. W. Langlois, J. L. Largier, E. J. Lessard, R. RaLonde, J. E. J. Rensel, P. G. Strutton, V. L. Trainer, and J. F. Tweddle, "Harmful algal Blooms along the North American west coast region: History, trends, causes, and impacts," *Harmful Algae*, vol. 19, pp. 133–159, Sep. 2012, doi: 10.1016/j.hal.2012.06.009.

[24] H. Cho, U. J. Choi, and H. Park, "Deep learning application to time-series prediction of daily chlorophyll—A concentration," *WIT Trans. Ecol. Environ.*, vol. 215, pp. 157–163, Oct. 2018, doi: 10.2495/EID180141.

[25] S. Lee and D. Lee, "Improved prediction of harmful algal Blooms in four major South Korea's rivers using deep learning models," *Int. J. Environ. Res. Public Health*, vol. 15, no. 7, p. 1322, Jun. 2018, doi: 10.3390/ijerph15071322.

[26] H. Cho and H. Park, "Merged-LSTM and multistep prediction of daily chlorophyll—A concentration for algal Bloom forecast," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 351, no. 1, Jul. 2019, Art. no. 012020, doi: 10.1088/1755-1315/351/1/012020.

[27] S. Malek, S. M. S. Ahmad, S. K. K. Singh, P. Milow, and A. Salleh, "Assessment of predictive models for chlorophyll—A concentration of a tropical lake," *BMC Bioinf.*, vol. 12, no. S13, pp. 1–11, Nov. 2011, doi: 10.1186/1471-2105-12-S13-S12.

[28] Q. Chen, H. Rui, W. Li, and Y. Zhang, "Analysis of algal Bloom risk with uncertainties in lakes by integrating self-organizing map and fuzzy information theory," *Sci. Total Environ.*, vols. 482–483, pp. 318–324, Jun. 2014, doi: 10.1016/j.scitotenv.2014.02.096.

[29] X. Li, J. Sha, and Z.-L. Wang, "Chlorophyll—A prediction of lakes with different water quality patterns in China based on hybrid neural networks," *Water*, vol. 9, no. 7, p. 524, Jul. 2017, doi: 10.3390/w9070524.

[30] A. Voutilainen and L. Arvola, "SOM clustering of 21-year data of a small pristine boreal lake," *Knowl. Manage. Aquatic Ecosyst.*, vol. 418, no. 36, pp. 1–16, 2017, doi: 10.1051/kmae/2017027.

[31] J. Tao, W. Chen, B. Wang, X. Jiezhen, J. Nianzhi, and T. Luo, "Real-time red tide algae classification using Naive Bayes classifier and SVM," in *Proc. 2nd Int. Conf. Bioinform. Biomed. Eng.*, Shanghai, China, 2008, pp. 2888–2891, doi: 10.1109/ICBBE.2008.1054.

[32] J. Tao, W. Cheng, W. Boliang, X. Jiezhen, J. Nianzhi, and L. Tingwei, ''Real-time red tide algae recognition using SVM and SVDD,'' in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, Xiamen, China, Oct. 2010, pp. 602–606, doi: 10.1109/ICICISYS.2010.5658453.

[33] M. A. A. Mosleh, H. Manssor, S. Malek, P. Milow, and A. Salleh, ''A preliminary study on automated freshwater algae recognition and classification system,'' *BMC Bioinf.*, vol. 13, no. 17, Dec. 2012, Art. no. S25, doi: 10.1186/1471-2105-13-s17-s25.

[34] P. J. G. Nieto, E. García-Gonzalo, J. R. A. Fernández, and C. D. Muñiz, ''Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain),'' *Ecol. Model.*, vol. 404, pp. 91–102, Jul. 2019, doi: 10.1016/j.ecolmodel.2019.03.009.

[35] A. M. Hussein, M. A. Elaziz, M. S. M. A. Wahed, and M. Sillanpää, ''A new approach to predict the missing values of algae during water quality monitoring programs based on a hybrid moth search algorithm and the random vector functional link network,'' *J. Hydrol.*, vol. 575, pp. 852–863, Aug. 2019, doi: 10.1016/j.jhydrol.2019.05.073.

[36] C. S. Reynolds, ''What factors influence the species composition of phytoplankton in lakes of different trophic status?'' *Hydrobiologia*, vol. 369, pp. 11–26, May 1998, doi: 10.1023/A:1017062213207.

[37] J. P. Heiden, K. Bischof, and S. Trimborn, ''Light intensity modulates the response of two Antarctic diatom species to ocean acidification,'' *Frontiers Mar. Sci.*, vol. 3, pp. 1–17, Dec. 2016, doi: 10.3389/fmars.2016.00260.

[38] P. G. Whitehead and G. M. Hornberger, ''Modelling algal behaviour in the River Thames,'' *Water Res.*, vol. 18, no. 8, pp. 945–953, Dec. 1984, doi: 10.1016/0043-1354(84)90244-6.

[39] S. Huang, D. Wang, X. Wu, and A. Tang, ''DSANet: Dual self-attention network for multivariate time series forecasting,'' in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, Beijing, China, 2019, pp. 2129–2132, doi: 10.1145/3357384.3358132.

[40] Y. Shin, ''Prediction of chlorophyll—A concentrations in the Nakdong River using machine learning methods,'' *Water*, vol. 12, no. 6, p. 1822, Jun. 2020, doi: 10.3390/w12061822.

[41] M. Järvenpää and K. Lindström, ''Water turbidity by algal Blooms causes mating system breakdown in a shallow-water fish, the sand goby *Pomatoschistus minutus*,'' *Proc. Roy. Soc. London B, Biol. Sci.*, vol. 271, no. 1555, pp. 2361–2365, Nov. 2004, doi: 10.1098/rspb.2004.2870.

[42] M. L. Wells, V. L. Trainer, T. J. Smayda, B. S. O. Karlson, C. G. Trick, R. M. Kudela, A. Ishikawa, S. Bernard, A. Wulff, D. M. Anderson, and W. P. Cochlan, ''Harmful algal Blooms and climate change: Learning from the past and present to forecast the future,'' *Harmful Algae*, vol. 49, pp. 68–93, Nov. 2015, doi: 10.1016/j.hal.2015.07.009.

[43] K. Kim, M. Park, J.-H. Min, I. Ryu, M.-R. Kang, and L. J. Park, ''Simulation of algal Bloom dynamics in a river with the ensemble Kalman filter,'' *J. Hydrol.*, vol. 519, pp. 2810–2821, Nov. 2014, doi: 10.1016/j.jhydrol.2014.09.073.

[44] P. K. Weber-Scan and L. K. Duffy, ''Effects of total dissolved solids on aquatic organisms: A review of literature and recommendation for salmonid species,'' *Amer. J. Environ. Sci.*, vol. 3, no. 1, pp. 1–6, Jan. 2007, doi: 10.3844/ajessp.2007.1.6.

[45] J. F. H. Bierhuizen and E. E. Prepas, ''Relationship between nutrients, dominant ions, and phytoplankton standing crop in Prairie saline lakes,'' *Can. J. Fisheries Aquatic Sci.*, vol. 42, no. 10, pp. 1588–1594, Oct. 1985, doi: 10.1139/f85-199.

[46] S. P. Singh and P. Singh, ''Effect of temperature and light on the growth of algae species: A review,'' *Renew. Sustain. Energy Rev.*, vol. 50, pp. 431–444, Oct. 2015, doi: 10.1016/j.rser.2015.05.024.

[47] M. H. Gholizadeh, A. M. Melesse, and L. Reddi, ''A comprehensive review on water quality parameters estimation using remote sensing techniques,'' *Sensors*, vol. 16, no. 8, p. 1298, 2016, doi: 10.3390/s16081298.

[48] J. S. Erickson, N. Hashemi, J. M. Sullivan, A. D. Weidemann, and F. S. Ligler, ''*In situ* phytoplankton analysis: There's plenty of room at the bottom,'' *Anal. Chem.*, vol. 84, no. 2, pp. 839–850, Jan. 2012, doi: 10.1021/ac201623k.

[49] T. Leeuw, E. Boss, and D. Wright, ''*In situ* measurements of phytoplankton fluorescence using low cost electronics,'' *Sensors*, vol. 13, no. 6, pp. 7872–7883, Jun. 2013, doi: 10.3390/s130607872.

[50] F. Adamo, F. Attivissimo, C. G. C. Carducci, and A. M. L. Lanzolla, ''A smart sensor network for sea water quality monitoring,'' *IEEE Sensors J.*, vol. 15, no. 5, pp. 2514–2522, May 2015, doi: 10.1109/JSEN.2014.2360816.

[51] L. Zeng and D. Li, ''Development of *in situ* sensors for chlorophyll concentration measurement,'' *J. Sensors*, vol. 2015, pp. 1–16, Apr. 2015, doi: 10.1155/2015/903509.

[52] O. Elijah, T. A. Rahman, C. Y. Leow, H. C. Yeen, M. A. Sarijari, A. Aris, J. Salleh, and T. H. Chua, ''A concept paper on smart river monitoring system for sustainability in river,'' *Int. J. Integr. Eng.*, vol. 10, no. 7, pp. 130–139, Nov. 2018, doi: 10.30880/ijie.2018.10.07.012.

[53] N.-C. Jung, I. Popescu, P. Kelderman, D. P. Solomatine, and R. K. Price, ''Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea,'' *J. Hydroinformatics*, vol. 12, no. 3, pp. 262–274, Jul. 2010, doi: 10.2166/hydro.2009.004.

[54] X.-E. Yang, X. Wu, H.-L. Hao, and Z.-L. He, ''Mechanisms and assessment of water eutrophication,'' *J. Zhejiang Univ. Sci. B*, vol. 9, no. 3, pp. 197–209, Mar. 2008, doi: 10.1631/jzus.B0710626.

[55] F. Recknagel, M. French, P. Harkonen, and K. I. Yabunaka, ''Artificial neural network approach for modelling and prediction of algal Blooms,'' *Ecol. Model.*, vol. 96, nos. 1–3, pp. 11–28, Mar. 1997.

[56] H.-S. Yi, S. Park, K.-G. An, and K.-C. Kwak, ''Algal Bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea,'' *Int. J. Environ. Res. Public Health*, vol. 15, no. 10, p. 2078, Sep. 2018, doi: 10.3390/ijerph15102078.

[57] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, ''Ensemble deep learning for regression and time series forecasting,'' in *Proc. IEEE Symp. (CIEL SSCI)*, Orlando, FL, USA, Dec. 2014, pp. 1–6, doi: 10.1109/CIEL.2014.7015739.

[58] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, ''A comprehensive survey on support vector machine classification: Applications, challenges and trends,'' *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[59] A. Rahman and M. S. Shahriar, ''Algae growth prediction through identification of influential environmental variables: A machine learning approach,'' *Int. J. Comp. Intel. Appl.*, vol. 12, no. 2, pp. 1–19, Jul. 2013, doi: 10.1142/S1469026813500089.

[60] L. Knoll, E. Hagenbuch, M. Stevens, M. Vanni, W. Renwick, J. Denlinger, R. S. Hale, and M. Gonzalez, ''Predicting eutrophication status in reservoirs at large spatial scales using landscape and morphometric variables,'' *Inland Waters*, vol. 5, no. 3, pp. 203–214, Jul. 2015, doi: 10.5268/IW-5.3.812.

[61] H. Serry, A. E. Hassanien, S. Zaghlou, and H. A. Hefn, ''Predicting algae growth in the Nile River using meta-learning techniques,'' in *Proc. Int. Conf. Adv. Intell. Syst. Comput. (AISI)*, vol. 639, A. Hassanien, Eds. Cham, Switzerland: Springer, 2018, pp. 745–754, doi: 10.1007/978-3-319-64861-3_70.

[62] R. Adhikari, R. K. Agrawal, and L. Kant, ''PSO based neural networks vs. traditional statistical models for seasonal time series forecasting,'' in *Proc. 3rd IEEE Int. Advance Comput. Conf. (IACC)*, Ghaziabad, India, Feb. 2013, pp. 719–725, doi: 10.1109/IAdCC.2013.6514315.

[63] N. H. An and D. T. Anh, ''Comparison of strategies for multi-step-ahead prediction of time series using neural network,'' in *Proc. Int. Conf. Adv. Comput. Appl. (ACOMP)*, Ho Chi Minh City, Vietnam, Nov. 2015, pp. 142–149, doi: 10.1109/ACOMP.2015.24.

[64] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ''Empirical evaluation of gated recurrent neural networks on sequence modeling,'' in *Proc. Workshop Deep Learn. (NIPS)*, Dec. 2014, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1412.3555

[65] H. Wang, R. Zhu, J. Zhang, L. Ni, H. Shen, and P. Xie, ''A novel and convenient method for early warning of algal cell density by chlorophyll fluorescence parameters and its application in a highland lake,'' *Frontiers Plant Sci.*, vol. 9, p. 869, Jun. 2018, doi: 10.3389/fpls.2018.00869.

[66] X. Wang and L. Xu, ''Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion,'' *Future Internet*, vol. 12, no. 2, p. 34, Feb. 2020, doi: 10.3390/fi12020034.

[67] P. A. Whigham and F. Recknagel, ''An inductive approach to ecological time series modelling by evolutionary computation,'' *Ecol. Modell.*, vol. 146, nos. 1–3, pp. 275–287, Dec. 2001, doi: 10.1016/S0304-3800(01)00313-1.

[68] R. Adhikari and R. K. Agrawal, ''An introductory study on time series modeling and forecasting,'' 2013, *arXiv:1302.6613*. [Online]. Available: http://arxiv.org/abs/1302.6613

[69] M. N. Noor, A. S. Yahaya, N. A. Ramli, and A. M. M. Al Bakri, ''Filling missing data using interpolation methods: Study on the effect of fitting distribution,'' *Key Eng. Mater.*, vols. 594–595, pp. 889–895, Dec. 2013, doi: 10.4028/www.scientific.net/KEM.594-595.889.

[70] S. Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *Int. J. Environ. Res. Public Health*, vol. 15, no. 12, pp. 1–24, 2018, doi: 10.3390/ijerph15122907.

[71] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," Oct. 2015, *arXiv:1506.00019*. [Online]. Available: https://arxiv.org/abs/1506.00019

[72] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, p. 6085, Apr. 2018, doi: 10.1038/s41598-018-24271-9.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Dec. 1997, doi: 10.1162/neco.1997.9.8.1735.

[74] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–81, Apr. 1997, doi: 10.1109/4235.585893.

[75] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting," *Energies*, vol. 13, no. 2, pp. 1–23, 2020, doi: 10.3390/en13020391.

[76] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull, "HABNet: Machine learning, remote sensing-based detection of harmful algal Blooms," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3229–3239, 2020, doi: 10.1109/JSTARS.2020.3001445.

[77] N. Mellios, S. Moe, and C. Laspidou, "Machine learning approaches for predicting health risk of cyanobacterial Blooms in Northern European lakes," *Water*, vol. 12, no. 4, p. 1191, Apr. 2020, doi: 10.3390/W12041191.

[78] X. Bo, Z. Yanchuang, W. Xinyuan, and Z. Xin, "Research on algae Blooms forecasting based on the multivariate data driven method: A case study of the Chaohu Lake," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 46, no. 1, 2016, Art. no. 012044, doi: 10.1088/1755-1315/46/1/012044.

[79] M. Mamun, J.-J. Kim, M. A. Alam, and K.-G. An, "Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches," *Water*, vol. 12, no. 1, p. 30, Dec. 2019, doi: 10.3390/w12010030.

[80] Y. Yu, Y. Zhu, S. Li, and D. Wan, "Time series outlier detection based on sliding window prediction," *Math. Probl. Eng.*, vol. 2014, Oct. 2014, Art. no. 879736. https://arxiv.org/pdf/1506.00019, doi: 10.1155/2014/879736.

[81] S. Ismail, M. Zulkifli, R. Mansor, M. M. Yusof, and M. I. Ismail, "The role of exploratory data analysis (EDA) in electricity forecasting," *Pertanika J. Soc. Sci. Humanit.*, vol. 24, pp. 93–100, Oct. 2016.

NUR AQILAH PASKHAL ROSTAM received the B.Sc. degree (Hons.) in computer science and the M.Sc. degree in software engineering from the Universiti Sains Malaysia, Penang, Malaysia, in 2012 and 2017, respectively, where she is currently pursuing the Ph.D. degree in artificial intelligence. She was a Former System Analyst with the Universiti Sains Malaysia. Since 2018, she has been working as a Graduate Research Assistant with the Department of Computer Science covering the M.Sc. and Ph.D. degrees. Her research interests include the development of IoT and ecological informatics modeling using deep learning and time series forecasting algorithm.

NURUL HASHIMAH AHAMED HASSAIN MALIM received the B.Sc. (Hons.) and M.Sc. degrees in computer science from the Universiti Sains Malaysia, Malaysia, and the Ph.D. degree from The University of Sheffield, U.K. She is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. Her current research interests include high performance computing, chemoinformatics, bioinformatics, data analytics, and sentiment analysis.

ROSNI ABDULLAH received the B.Sc. (Hons.) and M.Sc. degrees in computer science from Western Michigan University, Kalamazoo, MI, USA, and the Ph.D. degree from Loughborough University, U.K., in 2011, specializing in the area of parallel numerical algorithms and is one of the national pioneers in this field. She is currently a Professor with the School of Computer Sciences, Universiti Sains Malaysia. She has been the Head of the Parallel and Distributed Processing Research Group at the School, since its inception in 1994. Her research interests include parallel and distributed computing, parallel numerical algorithms, and parallel algorithms for bioinformatics.

ABDUL LATIF AHMAD received the B.Eng. and M.Sc. degrees from the University of Wales, Swansea, U.K., and the Ph.D. degree from the University of Wales, specializing in membrane technology, in 1995, and is one of the national pioneers in the field. He is currently a Professor with the School of Chemical Engineering, Universiti Sains Malaysia. His current research interests include hollow fiber membranes for environmental protection, such as gas and liquid, liquid membrane for wastewater treatment, and membrane technology for energy application, such as fuel cell and lithium air battery. He is also a Professional Engineer, a Chartered Engineer, and a fellow of IChemE, U.K. and an internationally renowned and acclaimed award-winning researcher in membrane science and technology, receiving a total of 59 personal achievements, including Prince Sultan Bin Abdulaziz International Prize for water by the Kingdom of Saudi Arabia and the Merdeka Award.

BOON SENG OOI received the B.Sc. degree (Hons.) in bioprocess engineering from the Universiti Teknologi Malaysia, in 1998, and the Ph.D. degree in chemical engineering from the Universiti Sains Malaysia, in 2005. He is specializing in the area of membrane science and technology. He is currently a Professor with the School of Chemical Engineering, Universiti Sains Malaysia. His research interests include water treatment and resource recovery using advanced polymer and membrane technology.

DEREK JUINN CHIEH CHAN received the bachelor's degree (Hons.) in chemical engineering from the Universiti Teknologi Malaysia, Malaysia, and the Ph.D. degree specializing in plant cell cultivation technology and proteomic purifications from the Universiti Sains Malaysia, Malaysia, in 2008. He is currently an Associate Professor with the School of Chemical Engineering, Universiti Sains Malaysia, where he has been teaching, since 2008. His research goal is to work toward a sustainable production and consumption strategies in any production activities with the integration of plants. His research interests include microalgae cultivation technology and reserved special interest in phytoremediation.