*Computing and Artificial Intelligence 2023; 1(1): 100.*
Review Article

# Harmful algal blooms (HAB) open issues: A review of ecological data challenges, factor analysis and prediction approaches using data-driven method

**Nur Aqilah Paskhal Rostam[1], Nurul Hashimah Ahamed Hassain Malim[1,*], Nur Afzalina Azmee[2], Renato J. Figueiredo[3], Mohd Azam Osman[1], Rosni Abdullah[1]**

[1] *School of Computer Sciences Universiti Sains Malaysia, Penang 11800, Malaysia*

[2] *Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Perak, Malaysia*

[3] *Department of Electrical and Computer Engineering of the University of Florida, Gainesville, FL 32611, United States*

*** Corresponding author:** Nurul Hashimah Ahamed Hassain Malim, nurulhashimah@usm.my

**ARTICLE INFO**

Received: 27 June 2023
Accepted: 3 August 2023
Available online: 12 August 2023

doi: 10.59400/cai.v1i1.100



**ABSTRACT:** Ongoing research on the temporal and spatial distribution of algae ecological data has caused intricacies entailing incomprehensible data, model overfit, and inaccurate algal bloom prediction. Relevant scholars have integrated past historical data with machine learning (ML) and deep learning (DL) approaches to forecast the advent of harmful algal blooms (HAB) following successful data-driven techniques. As potential HAB outbreaks could be predicted through time-series forecasting (TSF) to gauge future events of interest, this research aimed to holistically review field-based complexities, influencing factors, and algal growth prediction trends and analyses with or without the time-series approach. It is deemed pivotal to examine algal growth factors for useful insights into the growth of algal blooms. Multiple open issues concerning indicator types and numbers, feature selection (FS) methods, ML and DL forms, and the time series-DL integration were duly highlighted. This algal growth prediction review corresponded to various (chronologically-sequenced) past studies with the algal ecology domain established as a reference directory. As a valuable resource for beginners to internalize the algae ecological informatics research patterns and scholars to optimize current prediction techniques, this study outlined the (i) aforementioned open issues with an end-to-end (E2E) evaluation process ranging from FS to predictive model performance and (ii) potential alternatives to bridge the literature gaps.

***KEYWORDS:*** data-driven prediction method; harmful algal bloom; time series forecasting; machine learning; deep learning

## 1. Introduction

The escalation of HAB trends over the past years has caused much global concern. The HABs entail various bloom types that cause harm without exception as HAB toxins adversely impact human, environmental, and economic health while their non-toxic counterparts could prove detrimental to fishery-based resources and tools[1]. Rapid algae generation and growth, which depict its susceptibility to shifts in environmental conditions[2], have been thoroughly examined in ecological informatics as typical ecological data issues. As an emerging field that integrates computational methods for ecological evaluation, ecological informatics regards the intensive nature of ecological data with its valuable content and the necessity to convey empirical outcomes and make research-, conservation-, and resource

management-oriented decisions[3]. The conceptual framework associates ecological components (genomes, organisms, populations, communities, ecosystems, and landscapes) with data management, analysis, and synthesis.

Early prevention and algae growth control were executed as various places were impacted by this natural phenomenon until 2015 and continue to be affected to date. Nevertheless, several prediction-oriented issues resulting from spatial and temporal algae distribution remain unaddressed despite the exertion of time and effort to forecast HAB growth. The rapid generation and growth of algae under favorable environmental conditions could differ on short timescales ranging from several days to weeks or hours[4]. The concentration could shift abruptly as current chlorophyll content could occasionally be five times more concentrated than before and vice versa. This complexity further instigates the non-linearities and ambiguities in ascertaining HAB-favouring conditions, thus complicating the prediction process and causing forecasting errors due to fluctuations and ambiguities[5].

Apart from the nature of algal blooms, highly non-linear data and time-varying processes that remain vague[6] hinder current physical prediction models from establishing a clear coefficient, thus highlighting the correlation between every factor in algal bloom prediction and the multiple variable data sources required for analysis[7]. Such shortcomings are further hampered by prolonged durations and high financial costs and undermine prediction accuracy. Spatial and temporal distributions are impacted by multiple climatic, geographical, and ecological elements. Temporal distribution, which catalyzes the interconnection of indicators, inevitably increases prediction-related intricacies[8]. Summarily, all the aforementioned concerns have rendered prediction to become more complex and imprecise.

Effective algae bloom modeling and prediction in such an intricate system is significantly challenging given the presence of physical, chemical, and biological processes and their subsequent interrelations. Water pollution or eutrophication with algae implies an intricate operation of all potentially impacting factors[9]. The drawbacks could be resolved using ML or artificial intelligence (AI) for insights into the algal community. In line with previous studies, multiple scholars have employed past historical data to forecast algal bloom by integrating ML methods following the successful data-driven approach in algal growth prediction. Essentially, ML methods constituting artificial neural network (ANN), support vector machine (SVM), decision trees (DT), random forest (RF), and regression offer a principled set of mathematical techniques to elicit meaningful features from the data in generating distinctive patterns that could be manipulated for decision-making, estimating, and forecasting purposes.

The necessity to predict future HAB events has resulted in the incorporation of time series, which considers the temporal aspect. Multiple time-series approaches were implemented through the traditional method or integration with ML techniques following past research. Regarding the primary variation between the time series and traditional statistical forecasting, data points in the traditional statistical prediction (classification) could be independent of one another while the counterparts in time series denote a temporal nature that induces interdependence. This time dimension adds an explicit ordering to data points that should be conserved to offer vital information to learning algorithms[10].

Traditional time-series statistical forecasting models resembling auto-regressive integrated moving average (ARIMA) and its variants, such as autoregressive models (AR), moving average (MA), and autoregressive and moving average (ARMA) is extensively employed to make predictions. The models prove inappropriate for non-linear data evaluation despite their capacity to determine temporal behavior and generate satisfactory forecasts for linear time-series data[10] given their function in assuming specific distribution or function types of time series, thus deterring them from ascertaining the intricate

underpinnings of non-linear associations and depicting reality. As most of the models disregarding variable interdependencies in terms of multivariate time series induced low prediction accuracy, initiatives to incorporate DL in environmental study problems have intensified as the DL model reflects optimal performance for time-series data forecasting. Research on DL, such as recurrent neural network (RNN) and its variant, long short-term memory (LSTM), remains scarce given the recent integration of time series with DL in algal growth prediction studies[7].

The input-output variable relationship denotes a crucial study component. Concerning aquatic systems, information involving the impacts of physical, chemical, and biological water quality parameters on algal dynamics is necessary for optimal system internalization and management. Other elements (meteorological) also induce algae growth parallel to the growing data. On practical grounds, an efficient prediction approach denotes intricate algorithms, a holistic understanding of the blooms mechanism, and a reliable dataset with essential components. Determining this relationship by assessing the aforementioned set of information would generate insightful data to unveil the interconnection between factors. It is crucial to review past literature on algae prediction techniques and present complexities in determining an optimal algal growth prediction method from various aspects as precise decision-making arise from a sound interpretation of forecasting in various prediction tasks. The same framework could be generalized across other domains if the dynamics and comprehension of algae are addressed factor-wise. In this vein, the worldwide trend on water resource conservation and protection could be sustained.

The remaining sections are organized as follows: Section 2 provides an overview of the algae ecology, algal growth factor analysis, and research concerns; Section 3 highlights current literature on the data-driven method through ML, traditional time series, and time series with DL; Section 4 elaborates on open issues and future study directions; Section 5 summarizes the research.

## 2. Background study of algae ecology

Radmer[11] identified two distinct algae types that commonly live in water or damp environments: macroalgae and microalgae. Macroalgae imply bigger algae (seaweed) while microalgae denote the smaller counterpart (phytoplankton or cyanobacteria and green and red algae, red algae). Algae require (i) the sun as its primary source of energy, (ii) water, (iii) conducive temperature, and (iv) nutrients for growth: elements that could be easily found in water. Algae also require carbon dioxide, which is generated from pollutants (smoke, fumes from cars, and a little carbon dioxide) when plants breathe at night in the absence of sunlight, to produce sugar. The HAB implies an excess of harmful algae. Water conditions with nutrient-rich water columns (specifically phosphorus and nitrogen) following regular fish-feeding and fertilizer (or sewage) discharge into an aquatic system catalyze HAB growth. Toxic HAB could prove detrimental to human health and aquatic life (including fishes) in the form of ailments and demise, respectively, whereas non-toxic counterparts could damage fishery resources[1]. Harmful algal blooms could resemble foam, scum, or mats on the water surface in different colors and release life-threatening toxins (microcystin or MC, anatoxin, cylindropermopsin, and saxitoxin). In this regard, HAB has become a serious environmental concern on a global scale[12]. Discolored water could also be a HAB indicator. **Table 1** presents common HABs and their implications on human health following Bui et al.[12].

**Table 1.** Common HABs and health effect.

| Organism | Water type | Color | Toxin | Health effects |
|---|---|---|---|---|
| Alexandrium sp. | Salt | Red or brown | Saxitoxins | Paralytic shellfish poisoning, paralysis, death |
| Karenia brevis | Salt | Red | Brevetoxins | Gastrointestinal illness, muscle cramps, seizures, paralysis, respiratory problems |
| Pseudo-nitzschia | Salt | Red or brown | Domoic acid | Amnesiac shellfish poisoning, vomiting, diarrhea, confusion, seizures, permanent short-term memory loss, death |
| Microcystis | Fresh | Blue-green | MC | Gastrointestinal illness, liver damage |

## 2.1. Factor analysis affecting algae growth

Past literature demonstrated the severity and essentiality of algae prediction as a means of early prevention. Based on Whigham and Recknagel[13], recent advancements in algal bloom modeling encounter two drawbacks: insufficient ecological information for deductive modeling and inefficient data analysis for inductive modeling. The HAB heterogeneity must be evaluated to derive a comprehensive understanding and control their formation[14]. In other words, algal growth analysis would provide an optimal understanding of the aquatic system. The analysis method is an initial step pre-predictive modeling to gain useful insights and internalize algal ecology factors and interconnections in managing two primary elements: algae interrelationships and dynamics.

Many efforts have been exerted to evaluate algae ecology to comprehend the dynamics, ambiguity, and non-linear nature of algal growth prediction. Several studies solely emphasized this evaluation sans predictive work. The analysis techniques were performed by categorizing the algae. As some algorithms could also be utilized in pre-processing or FS, the algal bloom method analysis requires a thorough examination to internalize the pertinent factors and interrelations in algal ecology.

All algae species depend on light as a fundamental input for photosynthesis and nutrients for growth and reproduction: nitrogen and phosphorus. Factors encompassing water temperature, turbidity, mixing, competition, and grazing also hold relevance to the algae population dynamics. For example, Huang and Zheng[15] listed 20 environmental parameters entailing water temperature (WT), ambient temperature, secchi disk depth (SDD), transparency, turbidity, solar radiation, total phosphorus (TP), total nitrogen (TN), ammonia-nitrogen ($NH_3$–N), ammonium-nitrogen ($NH_4$–N), nitrate nitrogen ($NO_3$–N), ammonium ion concentration, dissolved oxygen (DO), conductivity, alkalinity, calcium concentration, total suspended solids (TSS), silica, pH, salinity, and chlorophyll-a (Chl-a) that catalyze cyanobacteria bloom.

Chlorophyll, the green pigment in leaves, enables plants to create energy from light using photosynthesis. The amount of photosynthesizing plants are implicitly assessed by measuring chlorophyll. Such plants would be algae or phytoplankton in a water sample. Overall, chlorophyll denotes the measure of all (dead or living) green pigments while Chl-a implies the measure of a pigment portion that remains alive. Sunlight, temperature, nutrients, and wind collectively impact both the prevalence of algae and Chl-a concentration. The first algae outbreak or 'bloom' may occur and grow in spring when nutrients are in rich supply and the water temperature and days turn warm. Algae concentration prediction, which could be measured in total chlorophyll form with raw water, was previously executed as a robust algal growth indicator[16].

Tian et al.[17] highlighted water quality, hydrology, and climate conditions as the key determinants of chlorophyll dynamics while Bui et al.[12] indicated that WT, pH, and DO were positively related to cynobacterial community dynamics and MC concentration. Meanwhile, nutrient level, phosphate, and

nitrogen concentration were also identified as fundamental elements. Water quality-oriented research ascertains the chemical and physical attributes of water bodies and potential contaminants that reduce the water quality and forecast algae growth. Notably, Chl-a also constitutes the biological elements in water quality research. **Table 2** lists the most extensively-assessed qualitative water quality parameters following Gholizadeh et al.[18].

**Table 2.** Common qualitative measures.

| Water quality parameter | Abbreviation | Units |
|---|---|---|
| chlorophyll-a | Chl-a | mg/L |
| Secchi disk depth | SDD | m |
| Temperature | T | ˚C |
| Coloured dissolved organic ma | 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 |
| Ortho-phosphate | $PO_4$ | mg/L |
| Chemical oxygen demand | COD | mg/L |
| Biochemical oxygen demand | BOD | mg/L |
| Electrical conductivity | EC | Ms/cm |
| Ammonia nitrogen | $NH_3$-N | mg/L |

Wells et al.'s[14] study extensively covered past and present algal blooms with emphasis on how climate shifts globally impact the marine planktonic system with elaborations on the connection of specific environmental aspects (temperature, stratification, light, ocean acidification, precipitation-induced nutrient inputs, and grazing) that undergo alterations amidst climate change. The 2013 United States Environmental Protection Agency (EPA) analysis summarised the implications of climate change on HABs with multiple mechanisms: warmer WT, altered salinity and rainfall patterns, high carbon dioxide concentration, coastal upwelling, and rising sea levels.

Other elements including microclimates and the thermal, hydric, and radioactive conditions in the first meter above and below the Earth surface may optimize prediction performance despite the paucity of scholarly attention. Microclimates are frequently disregarded in ecology and evolution despite empirical proof of their essentiality in ecosystem dynamics and processes, such as the organism responses to climate change[18] although Kearney and Porter[19] emphasized the importance of understanding microclimates in ecology given its representation of the physical conditions experienced by organisms. Amsler et al.'s[20] initial work that examined algal abundance with microclimate in 1992 aimed to morphologically, physiologically, and behaviourally connote the early germling life-history stages of algal for survival. Resultantly, the planktonic environment is chemically heterogeneous on a macroalgal propagule scale.

Biological oceanographers have come to acknowledge the high variability of nutrient concentration in water columns, which could then be assessed with classical approaches. Following Shi et al.[21], the

cyanobacterial dynamics sensitivity to climate conditions differed across regions based on hydrodynamics, morphology, and specified chemical parameters. This phenomenon exemplifies one of the microclimates examined in this study. Some of the crucial variables were elaborated on throughout the current review. All the aforementioned and additional variables were classified into several categories (see **Table 3**).

**Table 3.** Categorical variables.

| Abbreviation | Variables | Factor category |
|---|---|---|
| Chl-a | Chlorophyl-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 | |
| SDD | 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 | Aluminum | |
| 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 | |

Algal bloom-oriented research could also include factors constituting population density and algal bloom cases. The elements are explicitly inspired by a distinctive domain, such as the dengue outbreak where population density and dengue cases are key determinants of dengue prediction[21,22]. Shi et al.[21] indicated that population density impacts the dengue outbreak following economic and income variances. Essentially, 'dengue cases' denote one of the key variables recognized by relevant scholars[23–26]. It could prove advantageous to thoroughly examine human factors for algae prediction as high nutrient loading and carbon dioxide concentration is induced by human activities. Careful consideration or adoption of additional determinants with data fusion could provide high-impact outcomes for the prediction process. Extensive research on whether data fusion optimizes prediction performance proves necessary to date.

## 2.2. Ecological data issue and challenges

Notably, HAB research and management struggle to ascertain species variety, life histories, ecosystems, and subsequent implications. The potentially-harmful algae ecology that does not fall under one distinctive group[1] leads to dynamic growth and complex prediction. Issues concerning dynamics and interconnections frequently appear in ecological modeling parallel to the aforementioned factors catalyzing algal growth.

Frequently correlated algae ecological variables have caused redundant information as high algae data interrelations primarily result from the indicator connections that are highly associated with one another. For example, multiple variables in aquatic environments implicitly or explicitly rely on the amount of oxygen available. Algae prediction needs to consider abiotic factors and their association with WT and nutrient concentration[27]. This occurrence has instigated low data quality, ambiguities, and variabilities. The formation of algal blooms depicts high uncertainty in addition to spatial variation following complex mechanisms.

Kim[28] thoroughly outlined dynamic-related issues. Specifically, dynamic algae growth, which could differ based on short timescales (hours to days), has rendered the identification of favorable HAB conditions a significant empirical effort among scholars. Algae concentration could alter abruptly when the present chlorophyll concentration is five times higher than before and vice versa, thus hampering accurate forecasting. Based on scholarly perspectives, natural factors undeniably induce impromptu changes in algae content. To date, ecological data are ambiguously connoted as expert knowledge following the presence of random variables, incomplete and inaccurate data, and approximate predictions (rather than measurements) that lead to data incomprehensibility[29].

Algae ecological data experience high missing values following their reliance on frequently-maintained monitoring sensors or systems, which could be damaged by the presence of algae. As coral-like algae would attach itself to the utilized equipment and damage all the installed sensors in line with Rostam et al.[30], early algae prediction remain essential given the extended timeframe offered for coastal water facilities to shut down before the equipment is damaged[31,32].

Highly non-linear behavior and dynamics emphasize conventional approaches resembling model-fitting[33]. Regarding aquatic systems, holistic knowledge of the physical, chemical, and biological water quality parameter impacts on algal dynamic proves necessary for optimal system comprehensibility and management. One measure of interest implies the examination of dynamic algae criteria within the algae domain where it is pivotal to disclose the input-output variable link: an essential algae ecological study element, particularly for precise prediction.

Ecological data typically appear in time-series format[34]. Time-series problems add to the order

complexity or frequently encompass temporal dependency, which causes two otherwise identical time points to fall under distinct classes or forecast different behaviors, unlike simpler classification and regression problems[35]. This attribute could prove challenging given the necessity for specialized data management in model fitting and assessment where the time series data need to be formatted or framed as a supervised machine learning pre-forecasting, hence increasing data evaluation complexities.

The ML techniques could function optimally on more intricate time-series forecasting issues with various input variables, intricate non-linear associations, and missing data. Such approaches frequently require hand-engineered features from domain experts or practitioners with domain backgrounds for enhanced performance. To date, time-series forecasting with DL serves to gauge temporal dependence from the data, efficiently determine past pivotal observations, and grasp their relevance to the present prediction process[35]. In this vein, crucial information could be derived from the input and dynamically shift the context as required. Feature engineering also constitutes one of the ML disciplines that could transform and engineer raw data into the fitting format for time-series forecasting.

As this section only highlights current algae ecological data complexities, the aforementioned methods would be extensively reviewed in another section. The current section serves to examine regression problems or concerns involving real-value predictions and empirical works that employ raw numerical static and time-series data. This review does not cover research on spatial data image-processing as the methodology constitutes distinctive terms of feature extraction or selection, which slightly differs from previous descriptions for easier comparison.

## 3. Data-driven prediction model

The algorithms currently incorporated into algal bloom prediction are categorized into data- and process-driven models. Process-driven model typically requires several parameters, such as initial conditions and ecological variables. The models, which occasionally encounter the uncertainty of kinetic coefficients, require optimal system knowledge[12] although process-driven models reflect highly-precise predictions. Data derivation intricacies in the simulation process have instigated drawbacks in process-driven method implementation.

Past research documented the successful implementation of the data-driven AI-based method. Essentially, data-driven models rely on computational astuteness and ML techniques[36]. An ML algorithm serves to identify the system input-output association with a training dataset that characterizes all system behaviors. The trained model could be subsequently tested with an independent dataset to ascertain the extent to which it could be generalized across unseen data. Algae prediction would be more precise upon identifying the optimal parameter level through past ecological data insights.

Appropriate FS is mandatory in this case. The ML for unsupervised techniques, such as clustering could facilitate data discovery or elicit useful insights rather than merely relying on domain knowledge. As such, unsupervised ML was also reviewed. Sections for the algae prediction model through ML (presented in distinctive parts) are classified into unsupervised and supervised ML and time-series forecasting.

### 3.1. Algal growth prediction with unsupervised ML

Traditional clustering approaches are unsupervised given the absence of outcome variables and knowledge of the associations between the dataset observations[37]. Ecological data classification facilitates notable pattern and feature identification. Despite the abundance of clustering methods, such as hierarchical clustering, self-organizing map (SOM), and K-means, both SOM and K-means denote the

common clustering technique within this domain. Kohonen's[38] SOM, a useful feature extraction tool with a series of known patterns[39], represents multi-dimensional data in a relatively lower (one or two) dimensional space. Neurons on the grid would eventually merge around areas with high-density data points using multiple iterations. Overall, SOM denotes an efficient instrument in high-dimensional data visualization that proves adequate for the data comprehension stage in the knowledge discovery process.

Following Li et al.'s[40] recent research on SOM implementations in algae analysis and forecasting, the proposed SOM perceivably selects the most influential input variables for Chl-a. The K-means approach was subsequently integrated to define the cluster boundaries. Resultantly, SOM and GA-BPNN functioned as an FS instrument and efficiently performed clustering and predictions. In the study of Malek et al.[41], SOM was also utilized as an analysis tool of limnological time-series in the Putrajaya Lake and wetlands for algae growth identification. An expert system was subsequently established following the rules elicited from SOM for algal growth modeling and prediction.

Generally, K-means clustering is incorporated into datasets where all the variables are quantitative and the distance between observations is evaluated with the squared euclidean distance. **Table 4** presents past studies with SOM and K-means that are applied in this domain and the number of employed features (#F), water source types, method, and factors category connected to **Table 3**.

**Table 4.** Unsupervised ML prediction method.

| Article (s) | #F | Sources | Method | Factors category | | | |
|---|---|---|---|---|---|---|---|
| | | | | BF | PF | CF | MF |
| [27] | 5 | Lake | SOM-fuzzy | / | / | / | - |
| [39] | 6 | Coastal | SOM | / | / | / | - |
| [40] | 24 | Lake, reservoir | SOM, K-means GABPN | / | / | / | - |
| [42] | 4 | Lake | SOM-FL | - | / | / | - |
| [43] | 13 | Lake | RNN-SOM | / | / | / | - |
| [44] | 11 | Lake | SOM | / | / | / | - |

Coastal dataset clustering is few compared to other water sources (see **Table 4**), thus implying the lack of SOM and K-means application to coastal ecological datasets. The SOM benefits and effectiveness regarding information extraction without background knowledge of the examined ecosystem reflect a potential unsupervised learning approach.

## 3.2. Algal growth prediction with supervised ML

Supervised learning implies the ML task of learning a function that outlines input to output following sample input-output pairs. A function is inferred from the labeled training data encompassing a set of training samples. Methods involving regression and time-series analysis and AI are implemented to evaluate the historical dataset for algal bloom prediction[27]. Such approaches adequately model the HAB dynamics[45]. This review, which emphasized previous studies on the algal growth prediction model, is sequenced based on ML approaches (fuzzy, ANN, and SVM) and includes hybrid techniques alongside a group of other approaches classified under the 'other' category: genetic algorithm (GA), naïve bayes (NB), RF, evolutionary algorithm (EA), hybrid evolutionary algorithm (HEA), multilayer perceptron (MLP) and DT. Lastly, the TSF method was thoroughly reviewed in a separate section. **Table 5** demonstrates the trend of algae prediction approaches between 2014 and 2020.

**Table 5.** The trend of algae prediction method (2014–2020).

| Article (s) | Fuzzy | ANN | SVM | Hybrid | Other | TSF |
|---|---|---|---|---|---|---|
| [7] | - | / | - | - | - | / |
| [12] | - | / | - | - | - | - |
| [15] | - | - | - | - | / | - |
| [17] | - | / | - | - | / | - |
| [23] | - | - | - | - | - | / |
| [28] | / | - | - | - | / | - |
| [29] | - | - | - | - | - | / |
| [34] | - | - | - | - | - | / |
| [41] | - | / | / | - | - | - |
| [46] | - | - | - | - | / | - |
| [47] | - | - | - | - | - | / |
| [48] | - | - | / | - | - | - |
| [49] | - | / | - | - | - | - |
| [50] | - | / | - | - | - | - |
| [51] | - | / | - | - | - | - |
| [52] | - | / | - | - | - | - |
| [53] | - | / | - | - | - | - |
| [54] | - | - | / | - | - | - |
| [55] | - | / | - | - | - | - |
| [56] | - | / | - | - | - | / |
| [57] | - | / | - | - | - | - |
| [58] | - | / | / | / | / | - |
| [59] | - | - | - | - | - | / |
| [60] | - | - | - | - | - | / |
| [61] | - | - | - | - | - | / |
| [62] | - | / | - | - | - | / |
| [63] | - | / | - | - | / | / |
| [64] | - | - | / | - | / | - |
| [65] | - | - | - | - | / | - |
| [66] | - | - | - | - | - | / |
| [67] | - | - | - | - | - | - |
| [68] | - | / | / | - | / | / |
| [69] | - | - | - | / | / | - |
| [70] | - | / | / | - | - | / |
| [71] | - | / | - | - | - | / |
| [72] | - | - | - | - | - | / |

### 3.2.1. Fuzzy approach

The fuzzy approach successfully resolves ecological data ambiguity due to its dynamic nature. Zadeh's[72] fuzzy set theory in 1965 follows an extension of the classical connotation of the term 'set', which enables the processing of fuzzy premises in the 'IF-THEN' form with fuzzy sets in the premise and

conclusive parts. The implementation of algal bloom prediction was scarcely documented despite the theory prevalence in uncertainty analysis. Most of the studies emphasized data clustering into specified classes or categories for pre-prediction rather than actual forecasting. In this regard, the fuzzy approach complements classification as opposed to regression and time-series prediction problems. Nevertheless, this method could establish comprehensible rules or offer useful insights owing to logical descriptions of the FL system action. Only three empirical works have been documented to date.

Chen and Mynett[73] employed nine parameters (pH, conductivity, BOD, DO, $NH^{4+}$, $NO^{3-}$, $NO^{2-}$, TIP, Chl-a) with the fuzzy approach in lakes. The SOM sought appropriate fuzzy set connotations and explicit inference rules that are supported by heuristic knowledge amidst data unavailability. Chen et al.[27] utilized the fuzzy method through SOM to ascertain the multivariate structure and provide insights into the spatial-temporal dynamics of algal blooms. Notwithstanding, both techniques only offered a one-way procedure that focused on understanding instead of actual prediction and disregarded model output feedback for further optimization. The study of Malek et al.[42] highlighted four variables for lake-based research: pH, SD, dissolved oxygen, and nitrate nitrogen. The method might not prove successful for large datasets and intricate features given the absence of empirical evaluation despite its practicality and efficiency in managing insufficient datasets with complex correlations and ambiguous interconnections.

### 3.2.2. Artificial neural network (ANN)

The ANN implies a computational non-linear model entailing artificial neurons or processing elements that is organized in three interconnected layers: input, hidden layer (middle). Each neuron constitutes weighted inputs (synapses), an activation function (denotes the output given an input), and one output. The weighted input sum generates the activation signal transferred to the activation function in obtaining one neuron output. The extensively employed activation functions encompass linear, step, sigmoid, tanh, and rectified linear unit (ReLu) functions. Backpropagation, which computes the loss function gradient, is the most commonly utilized approach to identify the error contribution of every neuron. In 1997, Recknagel et al.[74] pioneered the modeling of algal blooms in freshwaters through ANNs. Backpropagation was employed in training where inputs and outputs imply palpable water quality parameters and the biomass quantities of particular algal groups, respectively. **Table 6** summarizes other algal growth studies using ANN and its ensuing variants.

**Table 6.** Supervised ML prediction using ANN.

| Article (s) | #F | Sources | Method | Results | Factors category | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BF | PF | CF | MF |
| [12] | 8 | Reservoir | FFBP-ANN | RMSE: 0.108 | - | / | / | - |
| [17] | 1 | Reservoir | ANN | MSE: 1.303<br>R: 0.774 | / | - | - | - |
| [49] | 9 | Reservoir | MLP<br>TDNN | MSE: 1.76 | - | / | - | - |
| [50] | 12 | River | ANN<br>Sensitivity analysis | $R^2$: 0.82 | / | / | / | / |
| [52] | 11 | Coastal | RBMDBN | RSME: 0.0475<br>MAE: 18.72% | / | / | / | - |
| [53] | 17 | Fresh water | ELM | RSME: 0.3013<br>MAE: 0.2366<br>$R^2$: 0.8322 | / | / | / | - |

**Table 6.** (*Continued*).

| Article (s) | #F | Sources | Method | Results | Factors category | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BF | PF | CF | MF |
| [56] | 12 | Coastal | DBN-ARIMA | RMSE: 0.154<br>MAE: 0.123<br>MAPE: 17.21<br>R: 0.798 | / | / | / | - |
| [57] | 1 | Lake | DDBN<br>TDBN<br>DBN | RMSE: 1.48<br>RMSE: 1.53<br>RMSE: 1.55 | / | - | - | - |
| [62] | 7 | River | ELM<br>ANFIS<br>LR | RMSE: 13.8<br>RMSE: 16.7<br>RMSE: 17.5 | / | | / | / |

### 3.2.3. Support vector machine (SVM)

The SVM denotes a linear decision boundary that operates by mapping data to high-dimensional feature space for data point classification despite not being linearly isolated. Data would be transformed in such a way that the separator could be drawn as a hyperplane upon identifying a distinction between the categories. In SVR, it serves to forecast the output by fitting the maximum number of output points (from the training data) between the boundary lines and simultaneously remaining as flat as possible[75]. Essentially, SVM regression is a non-parametric approach following its dependence on the function of kernel: the employment of a linear classifier to address a non-linear classification task. Several works recommended SVM application in algae forecasting. Xie et al.[75] suggested an SVM-oriented prediction to internalize and predict a dynamic algae population shift in freshwater reservoirs and resolve the complex non-linearity of water variables with their interactions. Resultantly, SVM generates high prediction accuracy despite utilizing a small sample number. The modeling outcomes demonstrated that SVM outperformed ANN. **Table 7** highlights previous studies on algal growth prediction with SVM.

**Table 7.** Supervised ML prediction using SVM.

| Article (s) | #F | Sources | Method | Results | Factors category | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BF | PF | CF | MF |
| [47] | 9 | Coastal | SVM<br>GRNN | RMSE: 5.436<br>RMSE: 9.966 | / | / | / | - |
| [54] | 10 | Fresh water | SVM | $R^2$: 067 | / | / | / | - |
| [69] | 9 | Reservoir | SVM | RMSE: 1.04<br>$R^2$: 0.71<br>MAE: 0.40 | / | / | / | - |
| [75] | 16 | Reservoir | SVM | $R^2$: 0.863<br>RMSE: 0.264<br>MAE: 0.226 | - | / | / | - |
| [76] | 11 | Lake | SVM<br>BPNN<br>MRS | RMSE: 13.4822<br>RMSE: 14.8427<br>RMSE: 15.3446 | / | / | / | - |

### 3.2.4. Other approaches to algal growth prediction

The hybrid method integrates multiple algorithms for high performance. Specifically, the incorporation of various ML algorithms could significantly enhance the overall outcome by refining one another, generalizing, or adapting to unknown tasks as most of the algorithms are developed for a

particular dataset or task[76,77]. The NN was integrated with another approach in the study by Wang et al.[78] where a hybrid model constituting BPNN, rough decision model, and decision rule were structured to forecast cyanobacteria bloom. The rough reduction method omits irrelevant characteristics without losing pivotal knowledge by only choosing key neural network determinants. Intriguing results were outlined in the research of Li et al.[61] who recommended an approach to forecast algae bloom with FS: minimum redundancy maximum relevance (mRMR) with RF, which is resistant to overfitting problems. Likewise, the study by Serry et al.[55] primarily employed FS for algae bloom prediction in the improvement phase. Regrettably, the RMSE and correlation coefficient outcomes remained low. The SVM algorithm could be further enhanced with a metaheuristic approach, such as GA in line with Wang et al.[58]. Several other methods as in **Table 8** are also portrayed similar performance.

**Table 8.** Other supervised approaches in algal growth prediction.

| Article (s) | #F | Sources | Method | Results | Factors category | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BF | PF | CF | MF |
| [6] | 5 | Lake | ADHDP-AGM | RMSE: 1.0363 | / | / | / | - |
| [52] | 7 | Mortar surface | LS-SVR | RMSE: 4.55<br>$R^2$: 0.94 | - | - | - | - |
| [55] | 4 | River | Meta-learning: CFS & GA | RMSE:0.2<br>MAPE:0.14<br>Corr.:0.8 | - | / | / | - |
| [57] | 8 | Freshwater | GA-BP<br>GA-LSSVM<br>BP<br>TS | Error: 40.7<br>Error: 35.4<br>Error 64.2<br>Error:116.9 | / | / | / | - |
| [60] | 11 | Satellite coastal data | Multi-variate regression | Accuracy<br>1-day: 65.6<br>2-day: 72.1<br>3-day: 71.9 | / | / | / | / |
| [61] | 24 | Lakes, reservoir | GA-BPNN | RMSE<br>CI: 0.0030<br>CII: 0.0006<br>CIII: 0.012<br>CIV: 0.0040 | / | / | / | - |
| [62] | 13 | Lake | RF with mRMR (FS) | CE:0.33<br>RMSE:2.12<br>MAE:7.57 | / | / | / | - |
| [64] | 1 | Coastal | Wavelet transform multistep 1–20 | RMSE: 2.010-4.696,<br>MAPE:0.375-1.266 | / | - | - | - |
| [65] | 14 | Lake, reservoir | ABC-RBF-SVM | RMSE:0.0030 MAE: 0.0020 $R^2$:90 | / | / | / | - |
| [66] | 9 | Reservoir | DCCPI, PCA, cusp catastrophe | R: 0.873 | / | / | / | - |
| [68] | 34 | Freshwater | Hybrid moth search algorithm (MSA) (RVFL) | RMSE:0.187<br>Data Partition (50:50)<br>RMSE: 0.0446<br>Data Partition (70:30) | / | / | / | - |
| [79] | 8 | Lake | SMR-GP | RMSE:37.9 | / | / | / | / |

The GA denotes an approach to resolving both constrained and unconstrained optimization problems in line with natural selection, which catalyzes biological evolution. This algorithm reiteratively

refines a population of individual solutions and arbitrarily chooses individuals from the present population to be parents at every step to produce children for the next generation. In this regard, the population 'evolves' towards an optimal solution over successive generations. Essentially, GAs could address various optimization problems that do not complement standard optimization algorithms, including counterparts in which the objective function is discontinuous, not differentiable, stochastic, or highly non-linear. Wang et al.[57] incorporated GA to enhance the BP network and least squares SVM generalization capacity. Meanwhile, Wang et al.[57] recommended a prediction method of bloom combined with the time series and intelligent non-linear models to optimize the error caused by time-series analysis. The influencing factors and forecasting data of chlorophyll-a time-series prediction error were modeled by BP, GA-BP, LSSVM, and GA-LSSVM. Consequently, the suggested model optimized time-series prediction. Wang et al.[58] eventually integrated SVM with GA and a relevance vector machine (GA-RVM) to forecast the abundance of phytoplankton in association with algal blooms at a Macau freshwater reservoir and compare their performance with an ANN model. The GA-SVM models outperformed other approaches. Evidently, GA-oriented research is applied to or integrated with SVM. Further studies are necessary to perceive whether GA could still demonstrate a competitive performance for other algal blooms prediction.

### 3.3. Time series forecasting (TSF)

Although current literature on prediction techniques elaborates on supervised ML, including simple classification or regression problems, the processes could not forecast algal growth beyond or prior to the present period, which is pivotal in the early prevention of HAB disasters or potential outbreaks. Based on the review, the ML models recommended in this domain failed to represent temporal data attributes, which proves essential when the time dimension adds explicit ordering to data points that should be conserved given their provision of additional or vital information to learning algorithms. Furthermore, Xie et al.[75] indicated the forecasting model to demonstrate higher performance than the prediction counterpart. Following the research outcomes, the algal bloom is a complex, non-linear, and dynamic system that is impacted by water variables in previous and current months. Such problems should be resolved with time series.

A time series implies a series of chronologically indexed (listed or graphed) data points. Generally, time series denotes a successive sequence of discrete-time data taken at equally-spaced points in time. Time series encompassing univariate and multivariate forms have evolved across various disciplines, specifically in hydrological and ecological modeling and oceanography. Univariate time series constitutes a series with a single time-dependent variable while the multivariate counterpart depicts multiple single time-dependent variables that rely on past values and other variables. Notably, this dependence serves to predict future values.

Much algae prediction research encompasses various intricate variables. Time series data could be broadly categorized into (i) stationary time series and (ii) non-stationary time series. In stationary time series, statistical components resembling mean value or variance prove constant over time and stay in relative equilibrium based on their corresponding mean values as opposed to their non-stationary counterpart. Time series data could be framed as supervised learning through the value at the previous time-step to forecast the value at the following time-step[80]. The time series forecast horizons are defined as follows: short-, medium-, and long-term forecast ranges from one hour to one week, one week to one year, and over a year, respectively. A one-step prediction only forecasts the training dataset of the following day, whereas multi-step TSF predicts multiple time-steps in the future. Multistep-ahead TSF

enables the prediction of algae growth duration for the following year and the maximum and minimum temperature for algal growth in the following month or years. Typically, multivariate TSF models are sensitive to multi-step (short-, mid-, and long-term) horizons as in-depth predictions lead to the complex modeling of multi-step forecasting following accumulated errors and low performance[81].

Traditional direct, recursive strategies, hybrid and multiple input multiple output (MIMO) strategies were employed for multi-step forecasting[82]. The recursive strategy aims to train a model that exclusively emphasizes a one-step-ahead prediction. The predictions are recursively forecasted post-model training. In other words, intermediate predictions are utilized as inputs to forecasting the following values until the time horizon prediction[83].

The direct strategy establishes a set of different $N$ models for various time steps with the same input data employed to feed all the models, unlike the recursive counterpart that utilizes one model. Meanwhile, MIMO implies a multiple output strategy where the prediction model output denotes a vector of future values forecasted with only one model. The MIMO strategy could conserve the temporal stochastic dependency of sequential data to address the drawbacks of recursive and direct methods given that the objective function during model training simultaneously alleviates the prediction errors on multiple horizons[83]. The computational costs of MIMO are also lower than that of the direct strategy following its prerequisite of only one model to be trained.

Some scholars from other disciplines who compared distinctive multi-type forecasting types[84–86] proved that such variations elicited multiple outcomes. Nevertheless, current forecasting in the algal growth prediction domain typically emphasized a single-step prediction. Research on multi-step forecasting has failed to thoroughly describe the aforementioned approach. Traditional time series and DL in time series would be extensively discussed in the following subsections.

### 3.3.1. Traditional TSF

Stochastic time series models, such as ARIMA that constitute subclasses of other models (AR, MA, and ARMA) are one of the most renowned and extensively utilized time series techniques. Box and Jenkins suggested a fairly successful variation of the ARIMA model, such as the seasonal ARIMA (SARIMA) for seasonal TSF. This model has garnered much scholarly attention following its versatility in representing several time series variations with simplicity and the associated Box-Jenkins methodology for robust model development. Nevertheless, the models encountered specific drawbacks in terms of pre-assuming a linear form of the associated time-series data, which proves inadequate in various practical circumstances[10], and the inability to determine complex interactions from non-linear data[87]. The AI and DL approaches are becoming increasingly common in empirical studies. Based on the literature review, RNN and LSTM denote two DL techniques that demonstrate a more optimal performance compared to other algorithms in TSF.

### 3.3.2. Deep TSF

Deep learning is an ML subdiscipline that concerns algorithms, such as ANN that are inspired by the brain structure and function. Several DL model types, such as RNN and its ensuing variant (LSTM) are typically employed in TSF. The RNN, which entails a network with feedback connections from hidden and output layers to the preceding counterparts, is recommended when managing dynamic datasets[83]. In this regard, sequential data dynamics could be ascertained with previous pattern memories retained through network cycles. Meanwhile, LSTM implies a novel form of neural network that performs predictions based on the data derived from previous times.

The LSTM is a specified RNN architecture developed to model temporal sequences and their long-range dependencies more precisely than conventional RNNs. Notably, LSTM does not utilize activation functions in its recurrent components. The stored values are not altered while the gradient is retained during RNN-oriented training. The LSTM units are implemented in 'blocks' with several units with three or four 'gates' (input, forget, and output) that regulate the information flow based on the logistic function.

This architecture facilitates the learning of longer-term dependence. The GRUs resemble LSTMs albeit with a more simplified structure and utilize a set of gates to control information flow despite not employing separate memory cells and incorporating fewer gates[88]. Although the recently-evolved LSTM has been implemented across various disciplines, specifically in TSF, only a few studies adopted the LSTM algorithm for algae prediction. For example, Lee and Lee[7] incorporated the LSTM model involving algal bloom prediction for a short-term (one week) prediction on newly-constructed water quality on 16 rivers. Wang et al.[60] employed the time series non-linear model to rectify the error induced by traditional time series analysis. Although LSTM has reflected much improvement and undergone multiple integrations with other DL approaches (Bi-LSTM, Encoder-Decoder, and CNN-LSTM) to date, the algorithms are yet to be examined in terms of algal prediction. **Table 9** presents past studies within the TSF domain.

**Table 9.** Time series with DL forecasting.

| Author (s) | #F | Sources | Method | Results | Factors Category | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BF | PF | CF | MF |
| [7] | 10 | River | MLP<br>RNN<br>LSTM | RMSE: 9.28<br>RMSE: 7.93<br>RMSE: 7.67 | / | / | / | - |
| [22] | 10 | River | LSTM | RMSE<br>1-D Pred.: 0.04868<br>4-D Pred.: 0.08015 | / | / | / | - |
| [28] | 6 | River | MPUM | RMSE: 16.89<br>$R^2$: 0.74 | / | / | / | - |
| [33] | 9 | Coastal | CCM | MAE: 0.55–0.35 | - | / | / | / |
| [43] | 13 | Lake | RNN-SOM | RMSE: 19.0<br>R: 0.7<br>Accuracy: 87% | / | / | / | - |
| [56] | 12 | Coastal | DBN-ARIMA | RMSE: 0.154<br>MAE: 0.123<br>MAPE: 17.21<br>R: 0.798 | / | / | / | - |
| [60] | 7 | Reservoir | GLM | R: 0.71 | / | - | / | - |
| [66] | 12 | River | Merge LSTM | RMSE: 0.0459 | / | / | / | / |
| [70] | 13 | Coastal | RNN | RMSE: 1.269<br>MAE: 0.79 | / | / | / | / |
| [71] | 4 | Reservoir | LSTM | RMSE 1–10: 27–16 (decrease) | - | / | - | / |

## 4. Analysis and discussion

This study review has outlined several unresolved concerns and knowledge gaps for optimal prediction performance. The first issue denotes FS where most of the selections are performed arbitrarily or based on domain knowledge following Rahman and Shahriar[89]. Despite the implementation of other

approaches, such as MA, mRMR, and influence matrix, the techniques only emphasized vital FS without insights into the reason underpinning pivotal FS selection. The aforementioned complexities garnered much scholarly attention when some researchers began employing sensitivity analysis or the clustering approach for data discovery with SOM and K-means. Despite a rise in the utilization of other clustering techniques, only two counterparts appear to be extensively employed by relevant researchers. This knowledge gap has led the current research to examine other unsupervised ML approaches.

Features that are regarded as extremely high or low could lead to model fitting intricacies and performance fluctuations following the arbitrary FS. Arguments on model fitting and performance correspond to McGowan et al.[33], Li et al.[40], and Lu et al.[16] where dataset arbitrariness has instigated model overfitting and prediction errors. Inappropriate FS techniques for time series might also hinder or degrade the time-series forecasting performance. The employed features differ in number with a maximum of 34 indicators and a minimum of one. This research classified the number of utilized indicators into the following categories: low (1–11), medium (12–22), and high (over 22). **Figure 1** depicts the percentage of studies under the aforementioned categories. Specifically, a low number of (parameters) or indicators reflected over 50%, followed by the incorporation of a medium number of indicators, and a high number of indicators at 7%.

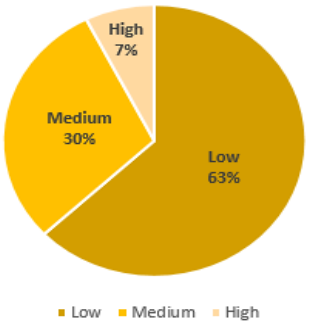


**Figure 1.** Range of indicators percentage.

Correlational research could be performed between parameter behaviors and the amount of Chl-a or algal growth predictors with an unsupervised method (clustering), which is typically regarded as the fundamental notion in pattern discovery. This conundrum has left a grey area where insights into the forecasting approach based on associations in an intricate ecological time-series data might catalyze forecasting-oriented decision-making.

Based on the current research, most of the water sources included in past studies involved lakes and reservoirs. As such, future works prove necessary for other sources, such as freshwater bodies, coastal areas, and estuaries as presented in **Figure 2**. Most of the studies only emphasized one water source at a time following their distinctive attributes and variations owing to hydrologic, geographic, climatic, morphologic, physical, chemical, geochemical, and biological aspects. Accommodating all water source types would imply the customization of study indicators to specified sources. Assumably, past empirical works were primarily reliant on data availability.

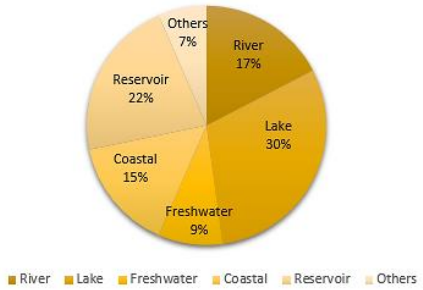


**Figure 2.** Trend of type of water sources.

Insufficient data following frequency updates would deter the forecasting process. Some high-volume data entail update frequencies that are consistently taken in minutes, hours, or on a daily basis. Nevertheless, sporadic data prevents pattern identification when (i) data is only gathered once a month, (ii) a substantial amount of data is missing, and (iii) the data is on a small scale. Such complexities could be resolved through large-scale data to complement the learning and training process. Concerning the factors category, this review has only observed a few studies that fully employed the four aforementioned factors despite the necessity of climates under the meteorological factor. Several open issues or gaps in managing different data size ranges and integrating data with adequate approaches remain unaddressed.

The third issue denotes algorithm performance where measurements that disregard water source types demonstrated overall or average TSF performance with DL compared to basic ML. Based on the empirical outcomes, current ML data-driven models could not sufficiently extract multi-factor timing data features with most of the models not depicting temporal data attributes. The LSTM has consistently outperformed other approaches with minimal prediction errors.

In terms of analysis method, specific performance measures, such as RMSE, the coefficient of determination ($R^2$), correlation coefficient (R), the mean absolute error (MAE), mean-square-error (MSE), and mean absolute percentage error (MAPE) were utilized with RMSE as the most favored and extensively utilized counterpart to assess the forecasted model-actual data value variance. The prevalence of past literature that employed RMSE has catalyzed the comparison process across domains. Much research has proven the deep time-series performance with RNN and LSTM to determine and depict temporal data attributes following the reviewed evaluation techniques.

Although data-driven methods offer versatility in FS to make predictions, this liberty to perform variable selection could instigate over-fitting and under-fitting complexities if carelessly ascertained. Discussions on feature engineering were not extensively throughout this review, specifically in the time-series prediction approach, following the need for non-trivial and time-consuming efforts[90] although feature engineering is pivotal in developing lag value and minimum or maximum horizon to forecast in TSF.

The fourth issue concerns improvement. The LSTM was primarily applied to river- and reservoir-oriented data despite its overall efficiency. As such, further works prove crucial to examine other water source types. Fluctuations in LSTM performance following the number of employed indicators and method-based shortcomings could also be observed. Although LSTM is capable of retaining information in the long run, the sequence-to-sequence LSTM architecture can only receive the input sequence to a fixed-length internal representation owing to the categorization of specific knowledge into small parts for easy remembrance. The LSTMs are impacted by multiple random weight initializations and behave akin to the feed-forward neural network where small weight initializations are favored. In this vein, other open issues require examination to resolve current circumstances. Feature engineering with LSTM and a

comprehensive understanding of the temporal aspect entailing algal growth data could induce optimal performance. Additional parameter tuning and learning approaches could similarly enhance present LSTM performance.

The fifth and final concern constitutes the engagement or data integration of various categories as one dataset is a complex task. A different data update and intricacies regarding the frequency taken would also vary. Such differences would result in the incorporation of multiple pre-processing approaches in data cleaning. Future studies should consider different factor categories as data fusion is primarily disregarded based on the aforementioned reasons. Different factors were classified following specific attributes: CF, BF, PF, and MF (see **Table 3**). Although several past studies did not include specific categorical factors, particularly MF, recent research from 2016 has incorporated MF following much scholarly attention. **Figure 3** depicts the overall trend of past studies regarding factor (BF, PF, CF, and MF) usage between 2009 and 2020.

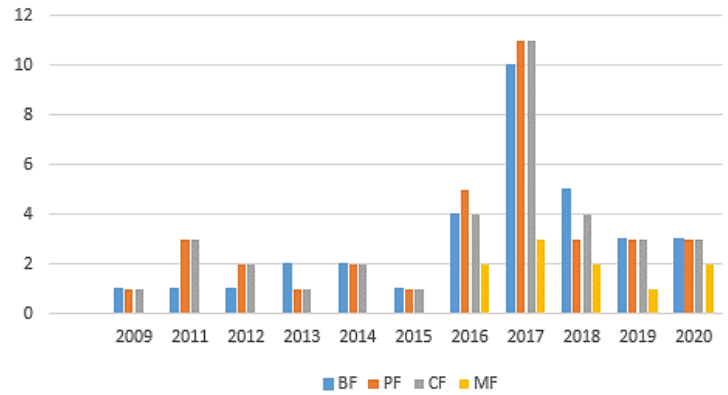


**Figure 3.** Trend of using categorical factors from 2009–2020.

The MF factor category, which only began considering and employing algae prediction from 2016 onwards, is palpably illustrated in **Figure 3**. Additionally, different predictors or factors began increasing between 2016 and 2020. This phenomenon might be associated with the advent of Internet of Things (IoT), which evolved with the development of sensors and utilization of multiple technologies that catalyze the data acquisition process. Such assumptions corroborated with Ande et al.[91] who disclosed that over 450 organizations have provided IoT platforms specializing in end-to-end solutions, system security, application enablement, device management, analytics, cloud storage, and back-end connectivity in 2017. These reasons further strengthen the notions underpinning multiple factor utilization with large-scale data based on the different factors gathered through various sensors. Overall, most of the employed factors attained the highest peak in 2017 or were significantly regarded from 2016 to 2018.

The benefits of data integration and dataset design are comprehensively performed in feature engineering, which typically requires expert knowledge to develop designs in terms of data and temporal elements, such as minimum or maximum past value prediction. Feature engineering provides such contributions based on the type of information that proves crucial from the input, which is vital for mapping and dynamic shifts as contextually required. Feature engineering implies one of the ML domains that could convert and engineer raw data into a fitting format for the prediction process, particularly in TSF[61].

Previous scholars emphasized single-step multivariate prediction with little research on modeling

the multi-step forecasting technique and no in-depth descriptions of multi-step methodology in their respective works although a long-term forecast could prove advantageous in future outbreak prediction. Various multi-step methods that were performed across different disciplines yielded distinctive outcomes. This inconsistency has resulted in another knowledge gap concerning the most adequate method for multi-step algal growth prediction. This approach proves more challenging as opposed to the normal counterpart given the presence of performance issues, particularly on cumulative errors and low performance in the wake of extended prediction. Multi-step performance in TSH, which is crucial in preventing algal growth, requires further investigation. Specific studies on enhancing the multi-step forecasting approach have been duly identified. Regarding the time-series domain, Venkatraman[92] incorporated several stages (prediction and optimization) in the predictive model build. Relevant works to resolve the aforementioned intricacies remain lacking despite the challenges encountered in advocating this multi-step method.

## 5. Conclusion

Prediction constitutes the core research concern in HAB-oriented research. Early prevention and awareness are pivotal following the HAB outbreaks and the increase in algae growth. The present prediction process could be enhanced through DL with time series by evaluating specific open issues that must be resolved based on comprehension and prediction performance given the high capacity in managing the non-linearity, ambiguities, and dynamics of algal growth. This method could be optimized by examining the prediction part and considering key features by improving the present selection approach and revealing the factor interconnections between the factors for a robust predictive algorithm. The capacity to forecast blooms (even if just a week in advance) by fully incorporating the multi-step method could enable public health authorities to address human health issues[30,31] and provide adequate time for water facilities to shut down before the equipment is damaged. The current study has holistically reviewed contemporary algal growth forecasting techniques. Particular open issues were also indicated for future research. Summarily, in-depth examination proves necessary to develop workable strategies in the future.

## Funding

This study, which is part of the collaboration between Universiti Sains Malaysia and the University of Florida under the facilitation of the Pacific Rim Application and Grid Middleware Assembly (PRAGMA), is funded by the Malaysian Ministry of Higher Education through the Transdisciplinary Research Grant Scheme (TRGS/1/2018/USM/01/5/4 - 203.PKOMP.67612).

## Conflict of interest

The authors declare no conflict of interest.

## References

1. Anderson DM. Approaches to monitoring, control and management of harmful algal blooms (HABs). *Ocean & Coastal Management* 2009; 52(7): 342–347. doi: 10.1016/j.ocecoaman.2009.04.006
2. McCormick PV, Cairns J. Algae as indicators of environmental change. *Journal of Applied Phycology* 1994; 6(5–6): 509–526. doi: 10.1007/BF02182405
3. Recknagel F, Michener WK. *Ecological Informatics: Data Management and Knowledge Discovery*. Springer; 2017.
4. Wong KTM, Lee JHW, Hodgkiss IJ. A simple model for forecast of coastal algal blooms. *Estuarine, Coastal and Shelf Science* 2007; 74(1–2): 175–196. doi: 10.1016/j.ecss.2007.04.012
5. Sun Y, Li J, Liu J, et al. Using causal discovery for feature selection in multivariate numerical time series.

*Machine Learning* 2015; 101(1–3): 377–395. doi: 10.1007/s10994-014-5460-1
6. Zhang H, Hu B, Wang X, et al. An action dependent heuristic dynamic programming approach for algal bloom prediction with time-varying parameters. *IEEE Access* 2020; 8: 26235–26246. doi: 10.1109/ACCESS.2020.2971244
7. Lee S, Lee D. Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *International Journal of Environmental Research and Public Health* 2018; 15(7): 1–15. doi: 10.3390/ijerph15071322
8. Huo S, He Z, Su J, et al. Using artificial neural network models for eutrophication prediction. *Procedia Environmental Sciences* 2013; 18: 310–316. doi: 10.1016/j.proenv.2013.04.040
9. Yang X, Wu X, Hao H, He Z. Mechanisms and assessment of water eutrophication. *Journal of Zhejiang University SCIENCE B* 2008; 9(3): 197–209. doi: 10.1631/jzus.B0710626
10. Adhikari R, Agrawal RK, Kant L. PSO based neural networks vs. traditional statistical models for seasonal time series forecasting. In: Proceedings of the 2013 3rd IEEE International Advance Computing Conference (IACC); 22–23 February 2013; Ghaziabad, India. pp. 719–725.
11. Radmer RJ. Algal diversity and commercial algal products. *BioScience* 1996; 46(4): 263–270. doi: 10.2307/1312833
12. Bui MH, Pham TL, Dao TS. Prediction of cyanobacterial blooms in the Dau Tieng Reservoir using an artificial neural network. *Marine and Freshwater Research* 2017; 68(11): 2070–2080. doi: 10.1071/MF16327
13. Whigham PA, Recknagel F. An inductive approach to ecological time series modelling by evolutionary computation. *Ecological Modelling* 2001; 146(1–3): 275–287. doi: 10.1016/S0304-3800(01)00313-1
14. Wells ML, Trainer VL, Smayda TJ, et al. Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful Algae* 2015; 49: 68–93. doi: 10.1016/j.hal.2015.07.009
15. Huang JD, Zheng H. Current trend of metagenomic data analytics for cyanobacteria blooms. *Journal of Geoscience and Environment Protection* 2017; 5(6): 198–213. doi: 10.4236/gep.2017.56018
16. Lu J, Huang T, Hu R. Data mining on algae concentrations (chlorophyll) time series in source water based on wavelet. In: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery; 18–20 October 2008; Ji'nan, China. pp. 611–616.
17. Tian W, Liao Z, Zhang J. An optimization of artificial neural network model for predicting chlorophyll dynamics. *Ecological Modelling* 2017; 364: 42–52. doi: 10.1016/j.ecolmodel.2017.09.013
18. Zellweger F, De Frenne P, Lenoir J, et al. Advances in microclimate ecology arising from remote sensing. *Trends in Ecology & Evolution* 2019; 34(4): 327–341. doi: 10.1016/j.tree.2018.12.012
19. Kearney MR, Porter WP. NicheMapR-an R package for biophysical modelling: The microclimate model. *Ecography* 2017; 40(5): 664–674. doi: 10.1111/ecog.02360
20. Amsler CD, Reed DC, Neushuli M. The microclimate inhabited by macroalgal propaguies. *British Phycological Journal* 1992; 27(3): 253–270. doi: 10.1080/00071619200650251
21. Shi K, Zhang Y, Zhou Y, et al. Long-term MODIS observations of cyanobacterial dynamics in Lake Taihu: Responses to nutrient enrichment and meteorological factors. *Scientific Reports* 2017; 7(1): 1–16. doi: 10.1038/srep40326
22. Cho H, Choi UJ, Park H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *WIT Transactions on Ecology and the Environment* 2018; 215: 157–163. doi: 10.2495/EID180141
23. Mathulamuthu SS, Asirvadam VS, Dass SC, et al. Predicting dengue incidences using cluster based regression on climate data. In: Proceedings of the 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE); 25–27 November 2016; Penang, Malaysia. pp. 245–250.
24. Mustaffa Z, Sulaiman MH, Emawan F, et al. Dengue outbreak prediction: Hybrid meta-heuristic model. In: Proceedings of 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD); 27–29 June 2018; Busan, Korea (South). pp. 271–274.
25. Zhu G, Hunter J, Jiang Y. Improved prediction of dengue outbreak using the delay permutation entropy. In: Proceedings of the 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData); 15–18 December 2016; Chengdu, China. pp. 828–832.
26. Džeroski S. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling* 2001; 146(1–3): 263–273. doi: 10.1016/S0304-3800(01)00312-X
27. Chen Q, Rui H, Li W, Zhang Y. Analysis of algal bloom risk with uncertainties in lakes by integrating self-organizing map and fuzzy information theory. *Science of the Total Environment* 2014; 482–483: 318–324. doi: 10.1016/j.scitotenv.2014.02.096
28. Kim S. A multiple process univariate model for the prediction of chlorophyll-a concentration in river systems. *International Journal of Limnology* 2016; 52: 137–150. doi: 10.1051/limn/2016003
29. Egerton TA, Morse RE, Marshall HG, Mulholland MR. 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* 2014; 2(1): 33–57. doi: 10.3390/microorganisms2010033
30. Rostam NAP, Ahamed Hassain Malim NH, Abdullah R. Development of a low-cost solar powered & real-time water quality monitoring system for Malaysia seawater aquaculture: Application & challenges. In: Proceedings of the 2020 4th International Conference on Cloud and Big Data Computing; 26–28 August 2020; United Kingdom. pp. 86–91.
31. Caron DA, Garneau ME, Seubert E, et al. Harmful algae and their potential impacts on desalination operations off southern California. *Water Research* 2010; 44(2): 385–416. doi: 10.1016/j.watres.2009.06.051
32. Lewitus AJ, Horner RA, Caron DA, et al. Harmful algal blooms along the North American west coast region: History, trends, causes, and impacts. *Harmful Algae* 2012; 19: 133–159. doi: 10.1016/j.hal.2012.06.009
33. McGowan JA, Deyle ER, Ye H, Carter ML, et al. Predicting coastal algal blooms in southern California. *Ecology* 2017; 98(5): 1419–1433. doi: 10.1002/ecy.1804
34. Pennekamp F, Iles AC, Garland J, et al. The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs* 2019; 89(2): e01359. doi: 10.1002/ecm.1359
35. Gamboa JCB. Deep learning for time-series analysis. *arXiv* 2017; arXiv:1701.01887. doi: 10.48550/arXiv.1701.01887
36. Jung NC, Popescu I, Kelderman P, et al. Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. *Journal of Hydroinformatics* 2010; 12(3): 262–274. doi: 10.2166/hydro.2009.004
37. Bair E. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews Computational Statistics* 2013; 5(5): 349–361. doi: 10.1002/wics.1270
38. Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 1982; 43(1): 59–69. doi: 10.1007/BF00337288
39. Wu ML, Zhang YY, Dong JD, et al. Identification of coastal water quality by self-organizing map in Sanya Bay, South China Sea. *Aquatic Ecosystem Health & Management* 2011; 14(3): 291–297. doi: 10.1080/14634988.2011.604273
40. Li X, Sha J, Wang ZL. Chlorophyll-a prediction of lakes with different water quality patterns in China based on hybrid neural networks. *Water* 2017; 9(7): 1–13. doi: 10.3390/w9070524
41. Malek S, Salleh A, Ahmad SMS. Analysis of algal growth using Kohonen self-organizing feature map (SOM) and its prediction using rule based expert system. In: Proceedings of the 2009 International Conference on Information Management and Engineering; 3–5 April 2009; Kuala Lumpur, Malaysia. pp. 501–504.
42. Malek S, Syed Ahmad SM, Singh SKK, et al. Assessment of predictive models for chlorophyll-a concentration of a tropical lake. *BMC Bioinformatics* 2011; 12(Suppl 13): S12. doi: 10.1186/1471-2105-12-S13-S12
43. Malek S, Salleh A, Milow P, et al. Applying artificial neural network theory to exploring diatom abundance at tropical Putrajaya Lake, Malaysia. *Journal of Freshwater Ecology* 2012; 27(2): 211–227. doi: 10.1080/02705060.2011.635883
44. Voutilainen A, Arvola L. SOM clustering of 21-year data of a small pristine boreal lake. *Knowledge and Management of Aquatic Ecosystem* 2017; 418: 36. doi: 10.1051/kmae/2017027
45. Nitin M, Kwok-wing C. Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence* 2007; 20(6): 735–744. doi: 10.1016/j.engappai.2006.11.016
46. Obenour DR, Gronewold AD, Stow CA, Scavia D. Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. *Water Resources Research* 2014; 50(10): 7847–7860. doi: 10.1002/2014WR015616
47. Knoll LB, Hagenbuch EJ, Stevens MH, et al. Predicting eutrophication status in reservoirs at large spatial scales using landscape and morphometric variables. *Inland Waters* 2015; 5(3): 203–214. doi: 10.5268/IW-5.3.812
48. Li X, Yu J, Jia Z, Song J. Harmful algal blooms prediction with machine learning models in Tolo Harbour. In: Proceedings of the 2014 International Conference on Smart Computing; 3–5 November 2014; Hong Kong, China. pp. 245–250.
49. Aria SH, Asadollahfardi G, Heidarzadeh N. Eutrophication modelling of Amirkabir Reservoir (Iran) using an artificial neural network approach. *Lakes & Reservoirs: Research and Management* 2019; 24(1): 48–58. doi: 10.1111/lre.12254
50. Guallar C, Delgado M, Diogene J, Fernandez-Tejedor M. Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of Karlodinium and Pseudo-nitzschia. *Ecological Modelling* 2016; 338: 37–50. doi: 10.1016/j.ecolmodel.2016.07.009
51. Tran TH, Hoang ND. Estimation of algal colonization growth on mortar surface using a hybridization of

machine learning and metaheuristic optimization. *Sādhanā* 2017; 42(6): 929–939. doi: 10.1007/s12046-017-0652-6
52. Zhang Z, Peng G, Guo F, et al. The key technologies for eutrophication simulation and algal bloom prediction in Lake Taihu, China. *Environmental Earth Sciences* 2016; 75(18): 1295. doi: 10.1007/s12665-016-6106-3
53. Lou I, Xie Z, Ung WK, Mok KM. Freshwater algal bloom prediction by extreme learning machine in Macau Storage Reservoirs. In: Sun F, Toh KA, Romay M, et al. (editors). *Extreme Learning Machines 2013: Algorithms and Applications. Adaptation, Learning, and Optimization*. Springer, Cham; 2014. Volume 16. pp. 95–111.
54. Fan J, Wu J, Kong W, et al. Predicting bio-indicators of aquatic ecosystems using the support vector machine model in the Taizi River, China. *Sustainability* 2017; 9(6): 892. doi: 10.3390/su9060892
55. Serry H, Hassanien AE, Zaghlou S, Hefny HA. Predicting algae growth in the Nile River using meta-learning techniques. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017; 9–11 September 2017; Cairo, Egypt. pp. 745–754.
56. Qin M, Li Z, Du Z. Red tide time series forecasting by combining ARIMA and deep belief network. *Knowledge-Based Systems* 2017; 125: 39–52. doi: 10.1016/j.knosys.2017.03.027
57. Wang L, Wang X, Jin X, et al. Analysis of algae growth mechanism and water bloom prediction under the effect of multi-affecting factor. *Saudi Journal of Biological Sciences* 2017; 24(3): 556–562. doi: 10.1016/j.sjbs.2017.01.026
58. Wang Y, Xie Z, Lou IC, et al. Algal bloom prediction by support vector machine and relevance vector machine with genetic algorithm optimization in freshwater reservoirs. *Engineering Computations* 2017; 34(2): 664–679. doi: 10.1108/EC-11-2015-0356
59. Karki S, Sultan M, Elkadiri R, Elbayoumi T. Mapping and forecasting onsets of harmful algal blooms using MODIS data over coastal waters surrounding Charlotte County, Florida. *Remote Sensing* 2018; 10(10): 1–19. doi: 10.3390/rs10101656
60. Wang H, Zhu R, Zhang J, et al. A novel and convenient method for early warning of algal cell density by chlorophyll fluorescence parameters and its application in a highland lake. *Frontiers in Plant Science* 2018; 9: 1–3. doi: 10.3389/fpls.2018.00869
61. Li X, Sha J, Wang ZL. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. *Environmental Science and Pollution Research* 2018; 25(20): 19488–19498. doi: 10.1007/s11356-018-2147-3
62. Yi HS, Park S, An KG, Kwak KC. Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. *International Journal of Environmental Research and Public Health* 2018; 15(10): 2078. doi: 10.3390/ijerph15102078
63. Du Z, Qin M, Zhang F, Liu R. Multistep-ahead forecasting of chlorophyll *a* using a wavelet nonlinear autoregressive network. *Knowledge-Based Systems* 2018; 160: 61–70. doi: 10.1016/j.knosys.2018.06.015
64. Nieto PG, García-Gonzalo E, Fernández JA, Muñiz CD. Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain). *Ecological Modelling* 2019; 404: 91–102. doi: 10.1016/j.ecolmodel.2019.03.009
65. Tian Y, Zheng B, Shen H, et al. A novel index based on the cusp catastrophe theory for predicting harmful algae blooms. *Ecological Indicators* 2019; 102: 746–751. doi: 10.1016/j.ecolind.2019.03.044
66. Cho H, Park H. Merged-LSTM and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast. In: *IOP Conference Series: Earth and Environmental Science*, Proceedings of the 2019 International Conference on Advances in Civil and Ecological Engineering Research; 1–4 July 2019; Kaohsiung, Taiwan. IOP Publishing; 2019. Volume 351.
67. Hussein AM, Elaziz MA, Wahed MSA, Sillanpää M. 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. *Journal of Hydrology* 2019; 575: 852–863. doi: 10.1016/j.jhydrol.2019.05.073
68. Hill PR, Kumar A, Temimi M, Bull DR. HABNet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2020; 13: 3229–3239. doi: 10.1109/JSTARS.2020.3001445
69. Mamun M, Kim JJ, Alam MA, An KG. Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. *Water* 2020; 12(1): 30. doi: 10.3390/w12010030
70. Wang X, Xu L. Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion. *Future Internet* 2020; 12(2): 34. doi: 10.3390/fi12020034
71. Song C, Zhang H. Study on turbidity prediction method of reservoirs based on long short term memory neural network. *Ecological Modelling* 2020; 432: 109210. doi: 10.1016/j.ecolmodel.2020.109210
72. Zadeh LA. Fuzzy sets. *Information and Control* 1965; 8(3): 338–353. doi: 10.1016/S0019-9958(65)90241-X
73. Chen Q, Mynett AE. Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling

of eutrophication in Taihu Lake. *Ecological Modelling* 2003; 162(1–2): 55–67. doi: 10.1016/S0304-3800(02)00389-7
74. Recknagel F, French M, Harkonen P, Yabunaka KI. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* 1997; 96(1–3): 11–28. doi: 10.1016/S0304-3800(96)00049-X
75. Xie Z, Lou I, Ung WK, Mok KM. Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. *Mathematical Problems in Engineering* 2012; 2012: 397473. doi: 10.1155/2012/397473
76. Abdelrahim M, Merlosy C, Wang T. Hybrid machine learning approaches: A method to improve expected output of semi-structured sequential data. In: Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC); 4–6 February 2016; Laguna Hills, CA, USA. pp. 342–345.
77. Liu J, Zhang Y, Qian X. Modeling chlorophyll-a in Taihu Lake with machine learning models. In: Proceedings of the 2009 3rd International Conference on Bioinformatics and Biomedical Engineering; 11–13 June 2009; Beijing, China. pp. 8–13.
78. Wang Z, Huang K, Zhou P, Guo H. A hybrid neural network model for cyanobacteria bloom in Dianchi Lake. *Procedia Environmental Sciences* 2010; 2: 67–75. doi: 10.1016/j.proenv.2010.10.010
79. Daghighi A. *Harmful Algae Bloom Prediction Model for Western Lake Erie Using Stepwise Multiple Regression and Genetic Programming* [Master's thesis]. Cleveland State University; 2017.
80. Hota HS, Handa R, Shrivas AK. Time series data prediction using sliding window based RBF neural network. Available online: https://www.semanticscholar.org/paper/Time-Series-Data-Prediction-Using-Sliding-Window-Hota-Handa/91037f01fd4b845eadca0b53f5dc00d9f61ac493 (accessed on 22 June 2023).
81. Yin J, Rao W, Yuan M, et al. Experimental study of multivariate time series forecasting models. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 3–7 November 2019; Beijing, China. pp. 2833–2839.
82. Taieb SB, Bontempi G, Atiya AF, Sorjamaa A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications* 2012; 39(8): 7067–7083. doi: 10.1016/j.eswa.2012.01.039
83. Nguyen HP, Liu J, Zio E. A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators. *Applied Soft Computing* 2020; 89: 106116. doi: 10.1016/j.asoc.2020.106116
84. An NH, Anh DT. Comparison of strategies for multi-step-ahead prediction of time series using neural network. In: Proceedings of the 2015 International Conference on Advanced Computing and Applications (ACOMP); 23–25 November 2015; Ho Chi Minh City, Vietnam. pp. 142–149.
85. Taieb SB, Sorjamaa A, Bontempi G. Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing* 2010; 73(10–12): 1950–1957. doi: 10.1016/j.neucom.2009.11.030
86. Taieb SB, Hyndman RJ. *Recursive and Direct Multi-Step Forecasting: The Best of Both Worlds.* Monash University; 2012.
87. Divina F, Torres MG, Vela FAG, Noguera JLV. A comparative study of time series forecasting methods for short term electric energy consumption prediction in smart buildings. *Energies* 2019; 12(10): 1–23. doi: 10.3390/en12101934
88. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* 2014; arXiv:1412.3555. doi: 10.48550/arXiv.1412.3555
89. Rahman A, Shahriar MS. Algae growth prediction through identification of influential environmental variables: A machine learning approach. *International Journal of Computational Intelligence and Applications* 2013; 12(2): 1–19. doi: 10.1142/S1469026813500089
90. Yin J, Rao W, Yuan M, et al. Experimental study of multivariate time series forecasting models. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 3–7 November 2019; Beijing, China. pp. 2833–2839.
91. Ande R, Adebisi B, Hammoudeh M, Saleem J. Internet of Things: Evolution and technologies from a security perspective. *Sustainable Cities and Society* 2020; 54: 101728. doi: 10.1016/j.scs.2019.101728
92. Venkatraman A, Hebert M, Bagnell JA. Improving multi-step prediction of learned time series models. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 25–30 January 2015; Austin, Texas, USA.