The Effect of Chemical Parameters on Water Quality Index in Machine Learning Studies: A Meta-Analysis

Nur Hanisah Abdul Malek¹, Wan Fairos Wan Yaacob¹*, Syerina Azlin Md Nasir¹ and Norshahida Shaadan²

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kelantang, Bukit Ilmu, Machang, Kelantan, Malaysia
²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

*wnfairos@uitm.edu.my

Abstract. According to the World Health Organization (WHO), approximately 2 billion people worldwide use drinking water sources that are contaminated with faeces. This is a serious issue since contaminated water may lead to certain waterborne diseases such as cholera, hepatitis A, dysentery, jaundice, and typhoid fever. Therefore, many researchers around the world are interested in studying the water quality. One of the most commonly used approaches is by using machine learning. Machine learning approach has grabbed the interest of many researchers since the last several years due to its power to compute complicated mathematical computations on big data analysis. Therefore, this study explored the correlation between different water quality parameters and Water Quality Index (WQI) in water quality studies that used machine learning by using a meta-analysis approach. This study used estimated variance, heterogeneity index, Chi-squared heterogeneity test and the random effects model. Based on the selected articles, pH, dissolved oxygen (DO) and biochemical oxygen demand (BOD) are the parameters commonly used in water quality studies which use a machine learning approach. This study found that pH is the best chemical factor which greatly affects the Water Quality Index since it has the highest mean correlation and lowest estimated variance due to sampling error. The result showed that the correlation between pH and WQI are heterogeneous across studies based on the Chi-squared of heterogeneity, Q value and heterogeneity index, I² value. The 95% confidence interval of effect summary supports the findings that the correlation of pH is different among the studies. This study also found that there is no evidence of publication bias using Egger and Begg’s test. Therefore, in order to ensure good water quality supply, the local authorities and government agencies should give more attention to this parameter since pH of water plays an important role in determining the water quality status.

Keywords: Chi-squared heterogeneity, heterogeneity index, machine learning, meta analysis, Water Quality Index,

1. Introduction

Water quality is a serious issue since contaminated water may lead to certain waterborne diseases such as cholera, hepatitis A, dysentery, jaundice, and typhoid fever. According to the World Health
Organisation (WHO), approximately 2 billion people worldwide use drinking water sources that are contaminated with faeces [1]. Therefore, many researchers around the world are interested in studying the water quality since it has become an alarming issue by local authorities, government agencies and also public communities around the world. There are a lot of factors that can affect the water quality such as erosion, runoff, temperature, decayed organic materials, oil and grease, pesticides, detergents, litter, and rubbish which can alter the level of sediments, nutrients, and salts in the river [2]. Many rivers are facing deteriorating water quality globally with distorted levels of nutrients, salts, and sediments in the river. Hence, researchers from all around the world used different water quality parameters to measure the factors that significantly affect water quality. The parameters of water quality can be characterised into biological, chemical, and physical parameters [3]. The physical parameters are colour, odour, temperature, turbidity, conductivity, salinity, suspended solid and total dissolved solid. Meanwhile, the chemical parameters are pH, dissolved oxygen (DO), nitrates, biochemical oxygen demand (BOD), chlorides, ammoniacal nitrogen, chemical oxygen demand and various heavy metals including copper, lead, nickel, iron, zinc, and cadmium. Biological parameters are pathogens and indicator microorganisms such as total coliforms and faecal coliforms [3]. Water Quality Index is an index that can represent overall status of water quality in a single grade based on some parameters [3]. However, the calculation of WQI is complicated by the calculation of sub-indices within the WQI equation and there is no universally applicable WQI method [4]. To overcome these challenges, many researchers have adopted the artificial intelligence (AI) approach [5–7].

The AI-based modelling eliminates the calculations of sub-indices and quickly produces a WQI value. Besides that, AI-based models have some advantages such as can predict complex phenomena, insensitive to missing data, the non-linear structures and can handle big data which comprise data at dissimilar scales [4]. These advantages have grabbed the interest of many researchers to use the AI-based model or also known as Machine Learning (ML). The commonly used machine learning models are Decision Tree [4,8–12], Artificial Neural Network [10,13–19], Support Vector Machine [6,10,12,17–20], k-Nearest Neighbour [10,12,17,20] and Naïve Bayes [10,11,21]. Still, these basic machine learning models do have certain problems such as overfitting or high bias due to too much variation to be robust. Thus, this led to improvement and advancement of ML models by using ensemble methods such as bagging, boosting, and stacking approaches to cater for the problem [22]. Ensemble models produce more accurate prediction by combining the decisions of multiple base classifiers. Recently, new algorithms of machine learning like Gradient Boosting [12,17,22], Xtreme Gradient Boosting [23], Random Forest [4,8,9,11,12,23], Deep Learning [24], Random Tree [4,9], bagging [4,12,21] and many other algorithms have been developed to predict water quality. Therefore, the purpose of this study is to review papers on water quality studies that use machine learning models.

This study chooses articles [4,25–28] that report the correlation value between water quality variables and Water Quality Index (WQI). Variability estimation of correlation values between the water quality parameter and WQI was gathered and discussed. Results from the analysis of heterogeneity and publication bias were presented and discussed in detail in Section 3. In the last section of this paper, this study was summarized in a conclusion.

2. Methodology
This study used a meta-analysis method to investigate the variation in the correlation between the water quality parameter and the water quality index in several previous studies. Meta-analysis can be defined as analysis of analyses using appropriate statistical analysis based on the collection of results from various studies [29]. It is commonly used to integrate the results of combined studies [30]. In the meta-analysis, the results of the combined studies are the data obtained from the summary statistics derived from the primary analyses of the previous studies. In statistical terms, the data refer to the effect size considered under study. In literature, the term "effect size" has numerous different meanings. First, the effect size may indicate a statistic which estimates the extent of an effect such as mean difference, correlation coefficient, Cohen’s D and regression coefficient. Second, this also means the actual values
calculated on the basis of certain effect statistics (for example: mean difference = 15 or $r = 0.8$) [31]. In most cases, the effect size means this and written as the effect size value. The effect size is a scale-less measure of the strength of the relationship between the dependent and independent variable [32]. The effect size may determine whether the difference is real or caused by a change in factors.

In the meta-analysis study, the availability of effect size is important since the analysis is based on effect size. Studies will be excluded if the effect size is unavailable or if the effect size type used is different. The effect size is measured in many ways such as Pearson correlation, standardized mean difference, Cohen’s D, Glass’ Δ, Hedges’ G, Cramer’s V, Fisher’s Z score and odds ratio. This study selected the Pearson correlation, $r$ as the effect size for determining the strength of the relationship between water quality parameters and the water quality index. The advantage of Pearson correlation is that it is widely used as a statistical descriptor and there are well-developed meta-analytical methods [30]. Hence, the Pearson correlation is the best choice because it aggregates the effect of the relationship between water quality parameters and the WQI value. Pearson's correlation effect value ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation). The negative sign indicates the relationship between the two variables as negative while the positive sign indicates a positive relationship. Meanwhile, the value of the correlation indicates how large the relationship is.

According to Cohen [33], the effect size is small if the value of $r$ varies around 0.1, moderate if $r$ varies around 0.3, and high if $r$ varies more than 0.5. The formula for Pearson correlation is as follows [34]:

$$ r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}} $$

where $n$ is the number of samples, $x$ is the independent variable and $y$ is the dependent variable.

Despite combining the results of the studies, meta-analysis can help to assess the variability of previous studies. Therefore, the meta-analysis adopted in this study seeks to examine variability in the effect size of different water quality parameters on the reported WQI value across studies. The meta-analysis procedure involves several steps as follows [35]:

Step 1: Develop a strategy for determining relevant studies.
Step 2: Set eligibility criteria for data to be included.
Step 3: Set up a standardized form for data collection.
Step 4: Standardize individual findings to compare them across studies.
Step 5: Compute the overall effect by combining the data.

2.1 Search Strategy

The analysis starts by identifying the research strategy and selection criteria. All relevant literature from many online databases such as Web of Science (WOS), SpringerLink, Science Direct, Google Scholar and others were explored for all publication years using the keywords “water quality index”, “water quality parameter”, “machine learning” and “correlation”. The references of the selected papers were also examined. This study managed to identify about 163 papers (Science Direct 44, Springer Link 33, Web of Science 3, Scopus 83) from the searching by using the specified keywords. However, only a few studies reported the value of correlation analysis in their study and some study did not even do the correlation analysis during the pre-processing stage before fitting the parameter into their machine learning models. In machine learning studies, the initial input parameter needs to be properly selected.
in order for the trained model to accomplish satisfactory prediction accuracy. Model oversimplification or redundancy may occur when too few or too many parameters are selected [18].

2.2 Inclusion and Exclusion Criteria

The selection of papers in this study is based on the correlation value since the correlation is selected as the effect size for this study. The papers were selected if they indicated the correlation values between water quality parameter and WQI for the whole year. Studies that did not use the Water Quality Index as an indicator of water quality were not included in the analysis. In addition, papers were also excluded if they indicated a correlation between water quality parameters since this study wanted to study the effect of water quality parameters on WQI. No language restrictions when searching for the papers. Selected papers may be based on time series, cross-sectional or a combination of the two studies. As a result, after deleting the duplicate items and applying the inclusion and exclusion criteria, only 9 of the 163 items in the four databases met the defined criteria.

However, the water quality parameters used in the studies were different. Some used chemical parameters; others used physical parameters while others used biological parameters which make comparison between studies difficult. Therefore, during screening of these 9 articles, the inclusion criteria were the chemical parameter as independent variables, WQI as dependent variable and the parameter value was the Pearson correlation, $r$. Chemical parameters are chosen as inclusion criteria, as most of the nine studies used parameters such as dissolved oxygen (DO), biochemical oxygen demand (BOD), and pH. Whereas the exclusion criteria are articles which do not meet one or more inclusion criteria.

2.3 Data Extraction

Only 5 out of 9 articles that used the same chemical parameters which are DO, BOD and pH were included in this meta-analysis study. Only these 5 studies reported the correlation value between DO, BOD and pH with WQI. The variables extracted from the study included the first author’s name, publication year, location of the study, total sample size, statistical method, studied parameters (DO, BOD, pH) and the value of Pearson correlation for each chemical parameter. Dissolved oxygen (DO), biochemical oxygen demand (BOD) and pH were always reported in most of the previous studies as the significant parameters that affect water quality [14][15][16][36][37][38]. Dissolved Oxygen is defined as the concentration of oxygen presence in water which is produced by photosynthetic plants and dissolved from the air. The presence of dissolved oxygen in rivers is necessary to maintain diversity of all aquatic life and oxygen balanced in the water system [39]. The pH is also an important parameter to determine water quality. It measures the acidity and alkalinity of the water [40]. Generally, if the water sample is below pH 7, it is considered as an acidic while higher than pH 7, it is alkaline. Biochemical oxygen demand shows the amount of oxygen that is needed by the microorganisms in water to decompose the organic substance [41]. It has been known that different areas possess different ranges of DO, BOD and pH.

Hence, this study focused on these three chemical parameters by observing its association with WQI value. This study succeeded in collecting about five studies which fulfilled the criteria of selection. The flow diagram of search strategy and selection of articles [42] are shown in Figure 1. Next, Table 1 shows the variables extracted from the five selected articles for this meta-analysis.
Figure 1. Flow diagram of articles selection procedure

Table 1. Variables Extracted from Selected Articles for Meta-Analysis

| No | Author                  | Year | Location          | Sample Size | Variables                          | Machine Learning          | Correlation values, r |
|----|-------------------------|------|-------------------|-------------|------------------------------------|---------------------------|-----------------------|
| 1  | Asadollah et al., [25]  | 2021 | Lam Tsuen River, Hong Kong | 240         | BOD, COD, DO, Electrical Conductivity, Nitrate-Nitrogen, Nitrite-Nitrogen, Phosphate, pH, Temperature and Turbidity | Extra Tree Regression (ETR), Support Vector Regression (SVR) and Decision Tree Regression (DTR) | EC= -0.0060, pH= -0.0400, Temp= 0.0390, Turb = 0.7880, DO = -0.3840, BOD = 0.7990, COD = 0.5170, NO₃ = -0.0050, NO₂ = 0.3480, PO₄ = 0.7270 |
| No. | Authors                | Year | Location                  | Data Parameters                                                                 | Models                                                                 | Results                                      |
|-----|------------------------|------|---------------------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------|
| 2   | Rezaie et al., [26]    | 2020 | Klang River, Malaysia     | BOD, DO, SS, COD, NH$_3$-N, and pH                                              | Ensemble Kalman Filter-Artificial Neural Network                       | DO = 0.7300, BOD = -0.5900, COD = -0.7800, SS = -0.6400, pH = 0.7600, NH$_3$-N = -0.4900 |
| 3   | Bui et al., [4]         | 2020 | Talar catchment, Northern Iran | BOD, COD, DO, pH, total solids (TS), fecal coliform (FC), phosphate, nitrate, turbidity and electrical conductivity (EC) | Twelve hybrid algorithms (Combinations of standalone with bagging (BA), CV parameter selection (CVPS) and randomizable filtered classification (RFC)) 4 standalone (Random Forest (RF), M5P, Random Tree (RT) and Reduced Error Pruning Tree (REPT)) | BOD = -0.7800, COD = -0.5600, TS = -0.1100, DO = 0.6900, FC = -0.8300, pH = 0.0500, PO4 = -0.2300, NO3 = -0.7200, Turb = -0.1200, EC = -0.5800 |
| 4   | Aldhyani et al., [27]   | 2020 | India                     | Dissolved Oxygen (DO), pH, conductivity, BOD, nitrate, fecal coliform and total coliform. | Support Vector Machine, k-Nearest Neighbour, Naïve Bayes | DO = -0.3836, pH = 0.0466, Cond= -0.2914, BOD = 0.1819, Nitrate = -0.0347, Fecal Coliform = -0.1128, Total Coliform = -0.1536 |
| 5   | Abba et al., [28]       | 2020 | Kinta River, Malaysia     | Suspended Solid, DO, BOD, COD, temperature, Total Solids (TS), pH                | Extreme Gradient Boosting (XGB), Extreme learning machine (ELM), Genetic Programming (GP), Linear Regression (LR), Step-wise-linear regression (SWLR) | DO = 0.7900, COD = -0.5800, BOD = -0.4500, Temp=-0.2800, SS = -0.2600, TS = -0.1000, pH = 0.0300 |
2.4 Data Analysis

Extracted data were collected in Excel 2016 and analysis was done by using STATA software (version 16) to generate graphs and tables. The next section describes the calculation involved.

2.4.1 Variability Estimation. Since this study uses correlation value as the effect size, the weighted mean correlation, $\bar{r}_w$ value is calculated by combining all the correlation coefficients, $r$ for each study. The weighted correlation is associated with the use of weights that can be attributed to subjects in the calculation of the correlation coefficient between the independent and dependent variable [43]. Weights can be naturally available beforehand or selected by the user for special use. For example, if there is a different number of sample sizes for each study, it is accepted to use these figures as weights. In the meta-analysis, the weighted mean correlation may be used for the sampling error variance of a correlation [44]. Therefore, after calculating the weighted mean correlation, the weight estimate of observed variance and estimated variance, due to sampling error, is calculated using the Hunter and Schmidt approach [44]. This is done because the weighting scheme uses the sample size [45]. The idea underlying this approach is to estimate the variability in the effect size distribution, which is the correlation. The observed variance's weight estimate, $\hat{V}_{obs(w)}$ are calculated using equation (2).

$$\hat{V}_{obs(w)} = \sum_{i=1}^{k} \frac{n_i (r_i - \bar{r}_w)^2}{\sum_{i=1}^{k} n_i}$$  \hspace{1cm} (2)

where $\bar{r}_w = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}$; $n_i$ is the sample size of $i^{th}$ study, $r_i$ is the observed Pearson correlation for each study and $\bar{r}_w$ is the weighted mean correlation among studies.

When the value of $\hat{V}_{obs(w)}$ is large, it means that the variability of correlation values among studies is high [46]. Next, the estimated variance on account of sampling error $\hat{V}_s$ is computed using equation (3). The estimated variance is considered for this study as it can be used to estimate the effect size distribution [46].

$$\hat{V}_s = \frac{(1 - \bar{r}^2_w)}{\bar{n} - 1}$$  \hspace{1cm} (3)

where $\bar{n}$ is the mean of sample size among studies.

2.4.2 Heterogeneity Statistics. The variability of the effect size assessed in the various studies is called statistical heterogeneity. It is caused by methodological or clinical diversity across studies. Statistical heterogeneity can be measured using Chi-squared heterogeneity test. In heterogeneity hypothesis tests, the $Q$ value is used in the decision to reject or not reject the null hypothesis. When the null hypothesis
is rejected, heterogeneity occurs among studies. The Chi-squared heterogeneity, $Q$ can be calculated using equation (4).

$$Q = \sum_{i=1}^{k} (W^* e_i^2) - \sum_{i=1}^{k} \left( \frac{(W^* e_i)^2}{\sum w} \right)$$

(4)

where;

$$W = \frac{1}{SE^2} \text{ and } SE = \frac{e_i}{\sqrt{e_i * n}}$$

$W$ is the weight of each study, $e_i$ is the effect size, $SE$ is the standard error and $n$ is the sample size of the study.

Next, heterogeneity index, $I^2$ is calculated in this study. $I^2$ is calculated as the percentage of variation in point estimate that is due to heterogeneity among studies caused by population difference in each study. The anticipated value of chi-squared heterogeneity, $Q$ if there is no heterogeneity, is equal to its degrees of freedom. Without 100, $I^2$ is essentially an intra-class correlation coefficient. It represents the percentage of the total change that is caused by variations between studies. Higgins et al. [47] stated that $I^2$ is preferable to the heterogeneity test when evaluating inconsistency across studies since it does not inherently depend on the number of studies in the meta-analysis. They suggested that if $I^2 = 0\%$ (no heterogeneity), $I^2 = 25\%$ (low heterogeneity), $I^2 = 50\%$ (moderate heterogeneity) and $I^2 = 75\%$ (high heterogeneity). If heterogeneity exists, the analysis of the random effect model should be considered. If there is no heterogeneity, the fixed effect model should be considered. Heterogeneity index, $I^2$ is calculated using equation (5).

$$I^2 = \frac{(Q - df)}{Q} \times 100$$

(5)

where $df$ is the degrees of freedom and $Q$ is the Chi-squared heterogeneity value.

2.4.3 Random Effects Weights. To re-examine the consistency of the heterogeneity test result among studies, the adjusted $I^2$ and adjusted $Q$ is calculated. A constant value, $v$, is computed to adjust the weight for the study individually. The constant value, $v$ and the adjusted weight of each study, $w_v$, are calculated using equation (6).

$$v = \frac{Q - (k - 1)}{\sum_{i=1}^{k} w_i^2 - \frac{\sum_{i=1}^{k} w_i^2}{\sum w}}$$

(6)

$$w_v = \frac{1}{(SE^2 + v)}$$

where $k$ is the number of studies and $w_v = \frac{1}{(SE^2 + v)}$. 
Then, the adjusted Chi-squared heterogeneity, $Q'$ and adjusted heterogeneity index, $I'^2$ are computed using the formula in equation (7).

$$Q' = \sum_{i=1}^{k} (w_v * e_i^2) - \frac{\sum_{i=1}^{k} (w_v * e_i)}{\sum_{i=1}^{k} w_v}$$

$$I'^2 = \left( \frac{(Q' - df)}{Q'} \right) \times 100$$ (7)

2.4.4 Assessing the Effect Summary and Correlation Across Studies. The effect summary of overall effect size, overall standard error and overall variance is calculated to supplement the final analysis. The overall correlation between previous studies is obtained by calculating the 95% CI. The confidence interval is calculated as an indication that the effect of the correlation between previous studies is different. The effect summary for effect size (Equation 8), standard error (equation 9) and variance (equation 10) are calculated using the corresponding equations.

$$ES_e = \frac{\sum_{i=1}^{k} (W_v * e_i)}{\sum_{i=1}^{k} W_v}$$ (8)

$$ES_{SE} = \left( \frac{1}{\sum_{i=1}^{k} W_v} \right)^{\frac{1}{2}}$$ (9)

$$ES_{SE^2} = \left( \frac{1}{\sum_{i=1}^{k} W_v} \right)^{2}$$ (10)

where $ES_e$ is the effect summary for the effect size, $ES_{SE}$ is the effect summary for standard error and $ES_{SE^2}$ is the effect summary for variance.

The 95% confidence intervals are obtained and computed using equation (11).

$$CI = e_i \pm 1.96(SE)$$ (11)

The absence of zero within the limit indicates the existence of the difference and variability of the correlation between the studies and vice versa.
2.4.5 Forest Plot. The forest plot is used to display the results of a meta-analysis or as a tool to identify where a formal meta-analysis can be useful. A forest plot gives a visual indication of the heterogeneity of the study and shows all the estimated common effects in a figure. A forest plot presents the confidence interval and effect size represented by whisker for multiple results within a study in a horizontal orientation [48].

3. Results and Discussions

3.1 Variability Analysis of Chemical Parameter
Variability in correlation values between studies was analysed using the Hunter and Schmidt [44] approach, as the effect size used was the correlation value. First, the weighted mean correlation was calculated by using sample sizes for each study as the weights. Based on the output in Table 2, the weighted mean correlations for dissolved oxygen (DO), biochemical oxygen demand (BOD) and pH are -0.0900, 0.3239 and 0.4082 respectively. These showed that the strength of the relationship between the chemical parameters and Water Quality Index were different. The pH had the highest mean correlation and the DO had the lowest mean correlation. Then, variability between studies was evaluated by examining the value of observed variance and estimated variance due to sampling error. Based on the value of observed variance in Table 2, pH had the lowest value whereas the BOD had the highest value. This means that the BOD correlation values varied greatly between studies. The higher the value of $\hat{\tau}_\text{abs(w)}$, the greater the variability of correlation values between studies. However, based on the estimated variance due to the sampling error, DO had the largest estimated variance and pH had the smallest value of estimated variance. Therefore, since pH had the highest mean correlation and lowest variance for both observed and estimated variance, correlation value between pH and WQI was used in the next analysis for heterogeneity test.

Table 2. Variability Estimation of Pearson’s Correlation

| Study                  | Sample Size | Pearson Correlation, $r$ | DO   | BOD   | pH   |
|------------------------|-------------|--------------------------|------|-------|------|
| Rezaie-Balf (2020)     | 227         |                          | 0.7300 | -0.5900 | 0.7600 |
| Asadollah et al. (2021)| 240         |                          | -0.3840 | 0.7990 | -0.0400 |
| Bui et al. (2020)      | 144         |                          | 0.6900 | -0.7800 | 0.0500 |
| Aldhyani et al. (2020) | 1679        |                          | -0.3836 | 0.6130 | 0.5233 |
| Abba et al. (2020)     | 301         |                          | 0.7900 | -0.4500 | 0.0300 |
| **Weight Mean**        |             |                          | -0.0900 | 0.3239 | 0.4082 |
| **Observed Variance**  |             |                          | 0.2465 | 0.2855 | 0.0618 |
| **Estimate Variance**  |             |                          | 0.0019 | 0.0015 | 0.0013 |

Note: DO = Dissolved Oxygen; BOD = Biochemical Oxygen Demand (BOD)

3.2 Heterogeneity Result
Heterogeneity analysis was carried out by calculating two important measures, namely Chi-squared heterogeneity ($Q$) test and heterogeneity index ($I^2$). The Chi-squared heterogeneity test gives information about the absence or presence of heterogeneity among the five studies. Based on the output in Table 3, $Q$ value is 777.5447 which is higher than chi-squared tabulated (9.4900) from the statistical table. Since $Q_{\text{calculated}} > Q_{\text{table}}$, the null hypothesis was rejected. Therefore, there is sufficient evidence to conclude that the correlation values between the studies are heterogeneous. As heterogeneity is present, the heterogeneity index $I^2$ was calculated to determine the variation percentage due to heterogeneity.
The heterogeneity index $I^2$ is 99.4856 as shown in Table 3, indicating a high variation due to heterogeneity between studies. In other words, pH variability from study to study was high.

Table 3. Heterogeneity Analysis

| Measurement            | Estimate |
|------------------------|----------|
| No of Studies          | 5        |
| Chi-squared Heterogeneity, $Q$ | 777.5447 |
| Heterogeneity Index, $I^2$ | 99.4856  |

As heterogeneity is present in this study, the random effects model was calculated because the effect size varied across studies. The findings are set out in Table 4 below. Chi-square heterogeneity and heterogeneity index had been adjusted in the random effects model by adding the constant value in the calculation. At this stage, these values were adjusted to re-evaluate the consistency of the presence of heterogeneity across studies. Based on the output in Table 4, adjusted $Q'$ is 9.8026, which is higher than $Q$-tabulated (9.4900). Since adjusted $Q'_{\text{calculated}} > Q'_{\text{tabulated}}$, the null hypothesis was rejected. Therefore, there is a heterogeneity in the pH of the random effect models across studies. In addition, the adjusted heterogeneity index $I'^2$ is 59.19% suggesting a moderate heterogeneity of pH in the random effects model for all studies. Therefore, the conclusion based on the results in Table 3 and 4 is coherent.

Table 4. Heterogeneity Analysis (Adjusted Value)

| Measurement                          | Estimate |
|--------------------------------------|----------|
| Constant Value                       | 0.0499   |
| Chi-squared Heterogeneity (Adjusted), $Q'$ | 9.8026   |
| Heterogeneity Index (Adjusted), $I'^2$ | 59.1944  |

3.3 Summary of Meta-analysis Correlation Results (Random Effects Model)

This study then summarized the meta-analysis results based on pH of water including the 95% confidence interval. The findings are set out in Table 5. The point estimate for each study is the effect size value, which is the correlation coefficient. To calculate the confidence interval values for each of the studies, the point estimate, standard error and variance were required. A 95% confidence interval was computed for each study to measure the significant difference in pH correlation between studies. From the results obtained in Table 5, there are no zero values within the lower and upper limits for all studies. For overall study, the lower bound is 0.0611 and the upper bound is 0.4561. Since, zero value is not included within the interval for effect summary, there is significant difference in pH correlation among studies.
Table 5. Meta-Analysis

| Study                  | Effect Size | Standard Error (SE) | Variance (SE^2) | 95% Confidence Interval | Lower Bound | Upper Bound |
|------------------------|-------------|---------------------|------------------|--------------------------|-------------|-------------|
| Rezaie-Balf (2020)     | 0.76        | 0.0579              | 0.0033           | 0.6466 - 0.8734          |             |             |
| Asadollah et al., (2021)| -0.04       | 0.0129              | 0.0002           | -0.0653 - -0.0147        |             |             |
| Bui et al., (2020)     | 0.05        | 0.0186              | 0.0003           | 0.0135 - 0.0865          |             |             |
| Aldhayani et al., (2020)| 0.5233      | 0.0177              | 0.0003           | 0.4887 - 0.5579          |             |             |
| Abba et al., (2020)    | 0.03        | 0.0100              | 0.0001           | 0.0104 - 0.0496          |             |             |
| **Effect Summary (ES)**| **0.2586**  | **0.1008**          | **0.0102**       | **0.0611 - 0.4561**      |             |             |

3.4 Forest Plot

Figure 2 below shows the forest plot of the five studies. The width of the confidence interval line showed the effect estimates of individual studies. From the confidence interval line, we can conclude that there is statistically significant difference in correlation between pH and WQI across the studies because zero values are not included in the interval. The area of the box represents the weight given on the study. Most of the studies are equally weighted. The diamond below the studies represents the overall effect.

![Forest Plot of Effect Size](image)

**Figure 2.** Forest Plot of Effect Size

3.5 Publication Bias

The publication and associated biases in the meta-analysis are often examined by visually checking the asymmetry in the funnel plots. However, it is an inherently subjective visual interpretation. Therefore, this study used Egger and Begg's test to assess if there is an association between the effect estimates and their standard error. A publishing bias exists if the p-value is less than 0.05 for the Egger and Begg tests [49]. However, based on the output in table 6, the probability value of Egger test is 1.000 and Begg's test is 0.4624 which is both greater than 0.05. Therefore, this study concludes that the publication bias is absent.

Table 6. Egger and Begg’s Test

| Meta Bias | p-value | Decision                          |
|-----------|---------|-----------------------------------|
| Egger     | p-value = 1.000 | p-value > 0.05, no publication bias |
| Begg      | p-value = 0.4624 | p-value > 0.05, no publication bias |
4. Conclusion
This study investigates chemical water quality parameters such as DO, BOD and pH as key components of the water quality index. It is intended to study and explore the correlation between chemical parameters and the WQI value. A meta-analysis approach was used for previous studies related to this study that focused on the effect size between the water quality chemical parameter and WQI. Based on the analysis conducted, this study found that pH is the best chemical factor which greatly affects the Water Quality Index since it has the highest mean correlation, lowest observed variance and estimated variance due to sampling error. Moreover, the correlation between pH and WQI was found to be heterogeneous in all studies through the chi-square heterogeneity test, $Q$. This was confirmed by the heterogeneity index, $I^2$, which also reported high heterogeneity among these two variables. Based on 95% confidence interval for effect summary, there is significant difference in correlation between pH and WQI among studies. This was supported by the forest plot of effect size. This study also concluded that the publication bias is absent based on Egger and Begg’s test. The pH of water plays an important role in determining the water quality status since the normal range of pH for surface water systems is 6.5 to 8.5. Any value outside the range must be examined. It is also important to note that the relationship between pH and WQI varies from place to place. This is because pH of water can be affected by natural changes or man-made (eg: acid rain, anthropogenic activity, wastewater discharged). Consequently, a well-planned action should be taken to ensure that the water quality in the river remains in good condition. As a recommendation, more articles should be included in the future study to minimize the standard error. In addition, the study scope should be broader with consideration of other chemical water quality parameters for example Chemical Oxygen Demand (COD), Ammoniacal Nitrogen and physical parameters such as temperature, total suspended solid and biological parameters.

References
[1] WHO 2019 Drinking-water https://www.who.int/news-room/factsheets/detail/drinking-water
[2] Lintern A, Webb J A, Ryu D, Liu S, Bende-Michl U, Waters D, Leahy P, Wilson P, and Western, A W 2018 Key factors influencing differences in stream water quality across space Wiley Interdisciplinary Reviews: Water 5(1) e1260.
[3] Abbasi T and Abbasi S A 2012 Water quality indices Elsevier
[4] Bui D T, Khorasvi K, Tiefenbacher J, Nguyen H and Kazakis N 2020 Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment 721(March) 137612 https://doi.org/10.1016/j.scitotenv.2020.137612
[5] Iticescu C, Georgescu L P, Murarui G, Topa C, Timofti M, Pintilie V, and Arseni M 2019 Lower Danube water quality quantified through WQI and multivariate analysis Water 11(6) 1305
[6] Leong W C, Bahadori A, Zhang J and Ahmad Z 2021 Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM) International Journal of River Basin Management 19(2) 149–56
[7] Yaseen Z M, Ramal M M, Diop L, Jaafar O, Demir V, and Kisi O 2018 Hybrid adaptive neuro-fuzzy models for water quality index estimation. Water Resources Management 32(7) 2227–45
[8] Gakii C and Jepkoech J 2019 A Classification Model for Water Quality Analysis Using Decision Tree. European Journal of Computer Science and Information Technology 7(3) 1–8
[9] Jelhouni M, Toomanian A, and Mansourian A 2020 Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS Water Resources Management 34(1) 139–54
[10] Babbar R and Babbar S 2017 Predicting river water quality index using data mining techniques. Environmental Earth Sciences 76(14) https://doi.org/10.1007/s12665-017-7645-9
[11] Vijay S and Kamaraj K 2019 Ground Water Quality Prediction using Machine Learning Algorithms in R. International Journal of Research and Analytical Reviews 6(1) 743–49
[12] Ahmed U, Mumtaz R, Anwar H, Shah A A, Irfan R and Garcia-Nieto J 2019 Efficient water quality prediction using supervised machine learning Water (Switzerland) 11(11) 1–14
[13] Ahmad Z, Rahim N A, Bahadori A, and Zhang J 2017 Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks *International Journal of River Basin Management* 15(1) 79–87

[14] Aradhanan G and Singh N B 2014 Comparison of Artificial Neural Network algorithm for water quality prediction of River Ganga *Environmental Research Journal* 8(2) 55–63

[15] Gazzaz N M, Yusoff M K, Aris A Z, Juahir H and Ramli M F 2012 Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors *Marine Pollution Bulletin* 64(11) 2409–20

[16] Hameed M, Sharqi S S, Yaseen Z M, Afan H A, Hussain A, and Elshafie A 2017 Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region *Malaysia Neural Computing and Applications* 28(1) 893–905

[17] Xu T, Coco G, and Neal, M 2020 A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning *Water Research* 177 115788. https://doi.org/10.1016/j.watres.2020.115788

[18] Liu M and Lu J 2014 Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environmental Science and Pollution Research* 21(18) 11036–53.

[19] Mohammadpour R, Shaharuddin S, Chang C K, Zakaria N A, Ab Ghani A and Chan N W 2015 Prediction of water quality index in constructed wetlands using support vector machine *Environmental Science and Pollution Research* 22(8) 6208–19.

[20] Sattari, M T, Joudi A R and Kusiak A 2016 Estimation of Water Quality Parameters With Data-Driven Model *Journal-American Water Works Association* 108(4) E232–9

[21] Muhammad S Y, Makhtar M, Rozaimee A, Aziz A A, and Jamal A A 2015 Classification model for water quality using machine learning techniques *International Journal of Software Engineering and Its Applications* 9(6) 45–52

[22] Sharafati A, Asadollah S B H S, and Hosseinzadeh M 2020 The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty *Process Safety and Environmental Protection* 140 68–78

[23] Lu H and Ma X 2020 Hybrid decision tree-based machine learning models for short-term water quality prediction *Chemosphere* 249 126169

[24] Barzegar R, Aalami M T, and Adamowski J 2020 Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment* 34(2) 415–33. https://doi.org/10.1007/s00477-020-01776-2

[25] Asadollah S B H S, Sharafati A, Motta D, and Yaseen Z M 2021 River water quality index prediction and uncertainty analysis: A comparative study of machine learning models *Journal of Environmental Chemical Engineering* 9(1) 104599

[26] Rezaie-Balf M, Attar N F, Mohammadzadeh A, Murti M A, Ahmed A N, Fai C M, Nabipour N, Alaghmand S, and El-Shafie A 2020 Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach *Journal of Cleaner Production* 271 122576.

[27] Aldhyani T H H, Al-Yaari M, Alkahtani H, and Maashi M 2020 Water Quality Prediction Using Artificial Intelligence Algorithms *Applied Bionics and Biomechanics* 2020 6659314 https://doi.org/10.1155/2020/6659314

[28] Abba S I, Hadi S J, Sammen S S, Salih S Q, Abdulkadir R A, Pham Q B, and Yaseen Z M 2020 Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *Journal of Hydrology* 587 124974

[29] Glass G V 1976 Primary, secondary, and meta-analysis of research *Educational Researcher* 5(10) 3–8

[30] Hedges L V and Olkin I 2014 *Statistical methods for meta-analysis* Academic press.

[31] Nakagawa S and Cuthill I C 2007 Effect size confidence interval and statistical significance: a
practical guide for biologists *Biological Reviews* **82**(4) 591–605

[32] Andrej J A, Keith S W, and Leiby B E 2009 Meta-analysis: A Brief Introduction *Clinical and Translational Science* **2**(5) 374–8

[33] Cohen J 1992 Statistical power analysis *Current Directions in Psychological Science* **1**(3) 98–101

[34] Hmoud Al-Adhaileh M and Waselallah Alsaade F 2021 Modelling and Prediction of Water Quality by Using Artificial Intelligence *Sustainability* **13**(8) 4259

[35] Egger M, Smith G D, and Phillips A N 1997 Meta-analysis: principles and procedures *Bmj* **315**(7121) 1533–37

[36] Abyaneh H Z 2014 Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters *Journal of Environmental Health Science and Engineering* **12**(1) 40

[37] Lerios J L and Villarica M V 2019 Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir *International Journal of Mechanical Engineering and Robotics Research* **8**(6)

[38] Sengorur B, Koklu R, and Ates A 2015 Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey *Water Quality, Exposure and Health* **7**(4) 469–490

[39] Matta G 2014 A study on physico-chemical Characteristics to assess the pollution status of river Ganga in Uttarakhand. *Journal of Chemical and Pharmaceutical Sciences* **7**(3) 210–7.

[40] Rahmanian N, Ali S H B, Homayoonfard M, Ali N J, Rehan M, Sadef Y, and Nizami A S 2015 Analysis of physiochemical parameters to evaluate the drinking water quality in the State of Perak Malaysia *Journal of Chemistry* **2015**

[41] Mohamed I, Othman F, Ibrahim A I N, Alaa-Eldin M E, and Yunus R M 2015 Assessment of water quality parameters using multivariate analysis for Klang River basin Malaysia *Environmental Monitoring and Assessment* **187**(1) 4182

[42] Soumaila K I, Niandou A S, Naimi M, Mohamed C, Schimmel K, Luster-Teasley S, and Sheick N N 2019 A systematic review and meta-analysis of water quality indices *Journal of Agricultural Science and Technology B* **9** 1–14.

[43] Costa J F P 2011 *Weighted Correlation BT - International Encyclopedia of Statistical Science* (M. Lovric (ed.); 1653–55). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-04898-2_612](https://doi.org/10.1007/978-3-642-04898-2_612)

[44] Hunter J E and Schmidt F L 1994 Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case *Journal of Applied Psychology* **79**(2) 171

[45] Borenstein M, Hedges L V, Higgins J P T, and Rothstein H R 2010 A basic introduction to fixed-effect and random-effects models for meta-analysis *Research Synthesis Methods* **1**(2) 97–111

[46] Hunter J E and Schmidt F L 2004 *Methods of meta-analysis: Correcting error and bias in research findings* Sage

[47] Higgins J P T, Thompson S G, Deeks J J, and Altman D G 2003 Measuring inconsistency in meta-analyses *Bmj* **327**(7414) 557–560.

[48] Woodall G M 2014 Graphical Depictions of Toxicological Data *Encyclopedia of Toxicology* 786–95

[49] Sutton A J, Song F, Gilbody S M, and Abrams K R 2000 Modelling publication bias in meta-analysis: a review *Statistical Methods in Medical Research* **9**(5) 421–45