Utilizing Polynomial Regression and XGBoost for Optimized Lignocellulosic Derivative Adsorption in Water Treatment with Ficus Nitida

1st Abdullah Gamal M. Elafifi
20221238@stud.fci-cu.edu.eg

2nd Adam Ahmed
Adam_20220050@dci.helwan.edu.eg

3rd Mohamed Sabry A. Ebrahim
mohamed2023609@m-eng.helwan.edu.eg

4th Ahmed Ayman
ahmed123abk@gmail.com

***Abstract*—Today, the water crisis is growing overall in the world, not only in terms of the small amount of water but also in the misuse of water resources. The Paper and pulp industry is a major contributor to Egypt's water crisis and environmental pollution. It consumes a massive amount of water, approximately 17,000 gallons/ton of Paper. This project aims to develop efficient water treatment procedures in paper and pulp industries via the adsorption of lignocellulosic derivatives in wastewater by utilizing Machine Learning (ML) Models and controlling the whole process using an automated mechanical unit for water consumption monitoring. Furthermore, regression and classification models were trained on different datasets to determine the adsorption of lignocellulosic derivatives and the potability of the water. During the training of over 5 and 8 different models for the adsorption and the potability classification, accuracy metrics were utilized to evaluate the performance of algorithms. In the testing phase, Polynomial Regression (PR) with a degree of 5 outperformed by achieving: 0.817, 1.031, 1.015, & 0.961 in MAE, MSE, $RMSE$, & R^2 respectively, and XGBoost Classifier achieved 67.37% accuracy after Hyperparameter Optimization. These results could be considered promising solutions for worldwide climate action and production by minimizing the amount of CO2 emissions burning wood in paper and pulp manufacture and treating the wastewater realized from the industrial process.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Wastewater-related diseases kill a child every eight seconds and are responsible for 80 percent of all illnesses and deaths in the developing world [1]. 3900 children die every day from water borne diseases [2]. Industrial practices are responsible for water crisis and pollution especially paper and pulp industry. In 2015, paper industry released 174,000 tons of emissions to air, water, and land (or 5.3%) out of a total of 3.3 million tons of emissions released by all industries in Canada which are the third largest amount [3]. Moreover, In the United States the pulp and paper industry are the sixth largest in the industrial pollution, as it released about 79, 000 tons or about 5% of all industrial pollutant. Paper and pulp wastewater is treated commonly using the activated sludge process, which depends on microorganisms to get rid of most of the lignocellulosic derivative's biomasses. However, its inflexible in operation, large surface area required for sludge disposal, and high operation cost make it not the appropriate method of treatment [4]. The reusing of treated paper wastewater will face water shortage and achieve water sustainability and can be considered as an additional value for environmental protection. An integration of the chemical treatment and mechanical system could offer an efficient solution for the water consumption and pollution of the paper and pulp industry. The paper wastewater could be chemically treated by the Modified Ficus Nitida, while the mechanical system would serve as an automated control unit to address the water consumption, and water treatment, and achieve the reusability concept.

II. Problem

As the environment sustainability considering of the most crucial gaols in the Sustainable Development Goals by the United Nations especially numbers 13 and 12 which is have a very significant impact on different fields. The production of activated charcoal (carbon) from Ficus Nitida produce large amount of harmful gaseous

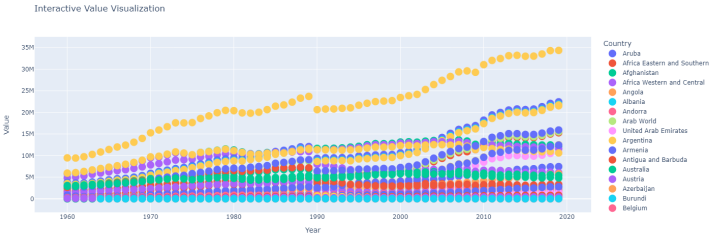


Fig. 1: The increasing of emissions from human activities (Burning wood) on exponential growth over the years.

emissions such as (not limited to): CO, CO2, different Nitric Oxides, and other volatile organic compounds (VOCs). Moreover, there is a clear difficulty in terms of time, cost, as well as practically of determining the adsorbed capacity and the capability to adsorb more pollutants in the industrial paper and pulp wastewater. In the terms of finding the capacity of adsorption capacity there is a common method is used the isotherms models such as Langmuir, Freundlich, Temken and

Dubinin-Radushkevich (D-R), and more. which leads to get rid of the current carbon, so more harmful emissions and cost and increase the impact on the environment. Moreover, there is a lack of clear overview of the complex relation of the environmental standards parameters that involve in water treatment and recycling in the frame of artificial intelligence technology. So, developing effective ANN AI model is significant for mitigating harmful emissions by determining the available capacity on the activated charcoal which leads to minimize the production of new activated charcoals. Also, the AI model will have the ability to determine the parameters that have an impact on the adsorption of wastewater pollutants (Lignocellulosic Derivatives) which leads to force the conditions to make the carbon acts optimally along. Finally, determine when the carbon will reach the optimum condition in treating the wastewater as well as the number of upcoming cycles.

III. Literature Review

In terms of determining the adsorption capacity of any charcoal the chemical practical method should be involved. Moreover, the Dublin-Radushkevich (D-R) isotherm model is a very popular Isotherm model like Langmuir and Freundlich. (D-R) is an empirical adsorption model that is generally applied to express the adsorption mechanism with Gaussian energy distribution onto heterogeneous surfaces [5] This isotherm is only suitable for an intermediate range of adsorbate concentrations because it exhibits unrealistic asymptotic behavior and does not predict Henry's laws at low pressure [6]. Also, (D-R) has many limitations: For complicated adsorption systems such as adsorption lignocellulosic derivatives using active charcoal made from Ficus Nitida may not always hold valuable values as it requires uniform surface energy distribution and monolayer adsorption, but in the activated carbon absorption mechanism, it consists of Multilayer Adsorption due to its high porosity and large surface area as well as a Surface energy distribution which have heterogeneous surface. This heterogeneity results in a non-uniform surface energy distribution which leads to work inefficiently and inaccurately. Especially when dealing with lignocellulosic derivatives, but our machine learning model solves this problem based on showing a strong relationship between the adsorption percentage and the model parameters that appear clearly in the wastewater of the paper and pulp industry including temperature, turbidity, Total Dissolved Solids (TDS), and the conductivity.

Furthermore, the Dubinin-Radushkevich isotherm model is temperature dependent; and in our case in finding the optimum condition as well as determining and monitoring the adsorption capacity for Ficus Nitida activated carbon to work in we are depending on the temperature and other factors mentioned before. In brief, the D-R isotherm model is not the best fitted in our case, however, our AI model solves this problem by adjusting the model to the surface energy distribution of the activated carbon (heterogeneous surface energy - non-uniform distribution) and also finds the best approach for the optimum condition, also based on the data from the databases we also founded the relation between the multilayer adsorption and the activated carbon itself which helped us in finding the optimum condition for the Ficus Nitida activated carbon.

IV. Contributions & Significance

Reduce Emissions: Ficus Nitida makes a huge amount of gases that harm the environment and climate change, which are CO, CO2, Nitritic Oxides, and volatile organic compounds (VOCs), and implementing the Machine Learning Model to get the optimal use of Ficus Nitida can reduce the emissions by minimizing the production cycles.

Efficient Utilization of Ficus Nitida: The prior methods of determining adsorption capacity like the isotherm model, lead to an increase the emissions and costs, and the goal of ML model is to optimize the utilization of Ficus Nitida based on finding and calculating the adsorption capacity, which enhances its effectiveness in wastewater treatment.

Optimized Wastewater treatment: Finding the relation between the environmental factors and wastewater treatment is important for the effectiveness of the process, and the proposed ML model provides the required information for the process.

Minimization of Carbon Footprint: Ficus Nitida contributes to the carbon footprint, and the project can reduce it.

Cost-Effective: The traditional methods of determining the adsorption capacity and Ficus Nitida effectiveness are expensive and time-consuming, and the provided ML model is cost-effective.

Predictive and resource planning: Predicting when Ficus Nitida reaches the optimum condition enables proactive maintenance and reduces downtime.

V. Methodology

In this project, the proposed model is an analysis of predictive models built on machine learning (ML) and deep learning (DL) algorithms, based on training, validating, and then testing to determine the adsorption of Ficus Nitida and the potability of water.

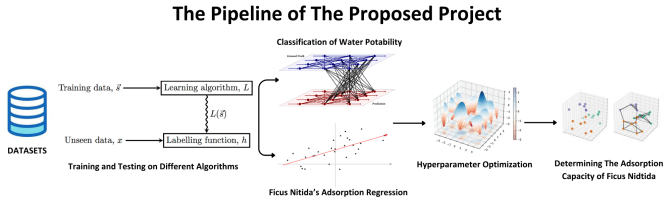
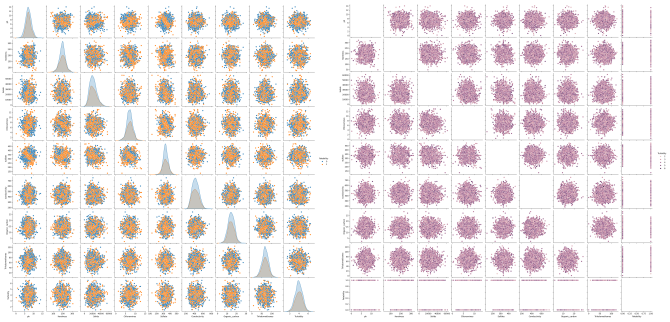


Fig. 2: A comprehensive overview of the proposed pipeline, explaining the different stages of analyzing the inputs, classifying, finding the relation, and predicting the adsorption capacity of Ficus Nitida.

In this chapter, five regression models and eight classification models are being trained and analyzed for based on different accuracy metrics. For the univariate models, the objective is to develop an effective model to mitigating

harmful emissions by determining the available capacity of the Ficus Nitida which leads to minimizing the production of new Ficus Nitida. That model will have the ability to determine the parameters that have an impact on the adsorption of wastewater pollutants (Lignocellulosic Derivatives) which leads to force the conditions to make the carbon act optimally along. Determine when the carbon will reach the optimum condition in treating the wastewater as well as the number of upcoming cycles. In this method, the training data is utilized for constructing the models. The models are then used for predicting the adsorption of Ficus Nitida based on the coming stages of filtration.

As shown in Fig(1), the pipeline provides an overview of the proposed model, as it shows starting with implementing the dataset and doing Exploratory data analysis (EDA). The datasets contain the raw data of different parameters that affect Ficus Nitida adsorption and turbidity. Pre-processing the dataset for training and data cleaning to find any missing values, finding the shape of the dataset & mean of the parameters. For the missing values, the mean of the dataset is calculated for filling in place of null values.



(a) The correlation between the factors and potability.

(b) The correlation between the factors and turbidity.

Fig. 3: The correlation between different factors.

During EDA, heatmaps describe the dimensionality reduction to check which feature is less activated.

Experimenting the data on different algorithms: regression ones involve five techniques, Polynomial Regression, SVR, XGB, Random Forest Regressor, and Artifical Neural Network (ANN), and in eight classification, Decision Tree, KNN, Logistic Regression, Rnadom Forest Classifier, XGBoost Classifier, Gaussian Naive Bayes, SVM, and AdaBoost. Moreover, Polynomial Regression and XGBoost Classifier outperformed and achieved the highest in multiple accuracy metrics.

The Mathematical Annotation: Suppose that prediction of Y from a vector (inputs) X of k features. As Y is a continuous scalar in the regression models, and Y is an indicator vector in terms of $Y_i = 1, Y_j = 0$ for $j \neq i$. In regression, the prediction of the values is dependent on intialized data, where Y is conditioned on X as $r(t) = E(Y|X = t)$, and the function $r(t)$ is the probability of different classes in classification;

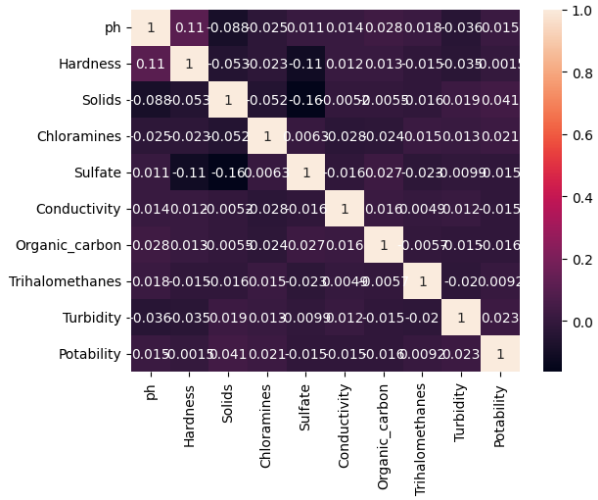


Fig. 4: The Heatmap correlation between parameters to see the activation.

it has to be approximated from the sample data whether parametrically or non parametrically.

A. ML Algorithms: Theoretical Implementation

Polynomial Regression: The polynomial regression is applied for finding the relation between temperature, MAC, and adsorption of Ficus Nitida to determine the optimal point for the charcoal based on the change in the volume of pollutants, given that:

$$y = w_0 + w_1 x + ... + w_{k-1} x^{k-1} + e = w^{\intercal} v(x) + e \quad (1)$$

which y is the predicted variable for our model and the independent variable, x is the dependent one, and $w := [w_0, ..., w_{k-1}]$ are the coefficients for the polynomial with different degrees, and $v(x)$ represents:

$$\mathbf{V}(\mathbf{x}) := \begin{bmatrix} 1 & x_0 & \cdots & x_0^{K-1} \\ 1 & x_1 & \cdots & x_1^{K-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N-1} & \cdots & x_{N-1}^{K-1} \end{bmatrix}$$

And this is the representation of the implemented ploynomial regression with a degree of five:

$$\begin{pmatrix} n & \sum x_i & \sum {x_i}^2 & \sum {x_i}^3 & \sum {x_i}^4 & a_0 \\ \sum x_i & \sum {x_i}^2 & \sum {x_i}^3 & \sum {x_i}^4 & \sum {x_i}^5 & a_1 \\ \sum {x_i}^2 & \sum {x_i}^3 & \sum {x_i}^4 & \sum {x_i}^5 & \sum {x_i}^6 & a_2 \\ \sum {x_i}^3 & \sum {x_i}^4 & \sum {x_i}^5 & \sum {x_i}^6 & \sum {x_i}^7 & a_3 \\ \sum {x_i}^4 & \sum {x_i}^5 & \sum {x_i}^6 & \sum {x_i}^7 & \sum {x_i}^8 & a_4 \\ \sum {x_i}^5 & \sum {x_i}^6 & \sum {x_i}^7 & \sum {x_i}^8 & \sum {x_i}^9 & a_5 \end{pmatrix}$$

That is, implementing Polynomial regression with a degree of 5 turned out to achieve the highest accuracy for determining the adsorption of the Ficus Nitida and finding the optimum point it could be reached at temperature of 50.

XGBoost: One of the distributed gradient boosting is XGBoost, which is an optimized iterative decision tree algorithm algorithm, where is every tree learning from the residuals of

all previous ones. **Explaning why it outperformed over the algorithms especially random forest** as it takes the sum of all the results for predition rather than adopting most voting in Random Forest:

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), f_k \in F, \tag{2}$$

where F means the combination of trees, f_k indicates the tree, hence $f_k(x_i)$ is the output of tree k, and $\hat{y}_i$ is the predicted value which is in this case, the potability of water. The goal of XGBoost is:

$$Obj(\theta) = g(\theta) + \omega(\theta), \tag{3}$$

as $g(\theta) = \sum_{i=1}^{n}(y_i, \hat{y}_i$ is the loss function for calculating the error. The architecture can be illustrated as:

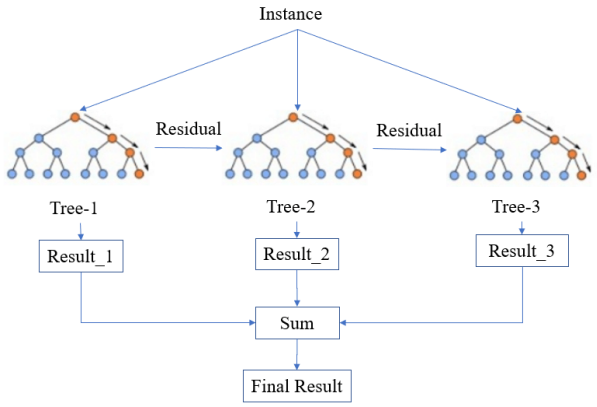


Fig. 5: The architecture of XGBoost Classifier

For water potability classification, XGBoost classifier determines the prediction $\hat{y}$ as water potability based on the sum of Solids, ph, Sulfate, Turbidity, etc.

VI. Results

In evaluating the performance of algorithms, various accuracy metrics were applied to calculate the accuracy of regression and classification. The following equations represent the different metrics in evaluating regression tasks:

$$MeanAbsoluteError(MAE) = \sum_{i=1}^{D} |x_i - y_i| \tag{4}$$

$$MeanSquaredError(MSE) = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{5}$$

$$RootMeanSquaredError(RMSE) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2} \tag{6}$$

Based on the provided formulas, the calculations are based on finding the error and difference between the actual data and predicted-finding the correlation of data. However, the metrics of evaluating classification are different based on finding the number of correct labels identified and the number of wrong ones, which are called True Positive(TP), True Negative (TN), False Positive (FP), and False Negative (FN):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

Regression: Showing the numerical results to assess the validity of the algorithms in a practical sense. Comparing between five different models in regression as shown in the Table(1) and Fig(6) in term of R, R2, RMSE, MSE, MAE:

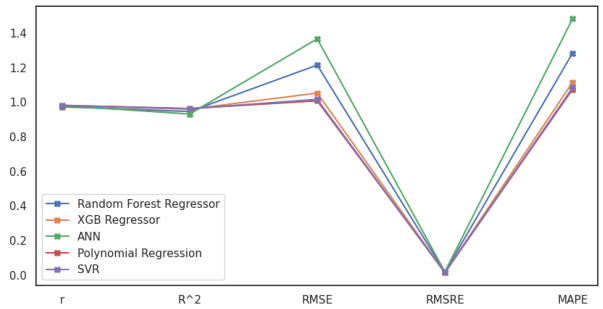


Fig. 6: Representation of Accuracy Metrics for the different algorithms.

TABLE I: Numerical results for the algorithms in regression.

Regression Evaluation	MSE	RMSE	MAE	R2
Random Forest Regressor	1.472	1.213	0.969	0.944
XGBoost Regressor	1.105	1.051	0.841	0.958
Artificial Neural Network (ANN)	1.864	1.365	1.110	0.930
Polynomial Regression	1.014	1.007	0.810	0.962
Support Vector Regressor (SVR)	1.031	1.015	0.817	0.961

Based on the results, Polynomial regression algorithm outperformed in all aspects and utilized to find the optimum value and condition of the Ficus Nitida's adsorption as shown in Fig(7) and Fig(8) as it represents the data after giving the inputs of the volume of pollutants and plotting the adsorption values:

Classification:

TABLE II: The accuracy metrics of XGBoost Classifier before Hyperparameter Optimization

XGBoost Classifier	Precision	Recall	F1-score	Support
Y_j	0.69	0.86	0.77	680
Y_i	0.60	0.34	0.43	402
Accuracy			0.67	1082
Macro Avg.	0.64	0.60	0.60	1082
Weighted Avg.	0.65	0.67	0.64	1082

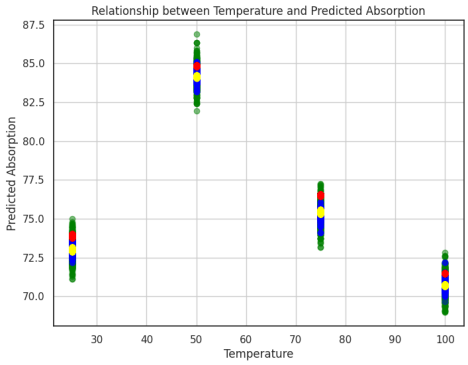


Fig. 7: The relation between Adsorption of Ficus Nitida and Temperature on different levels.

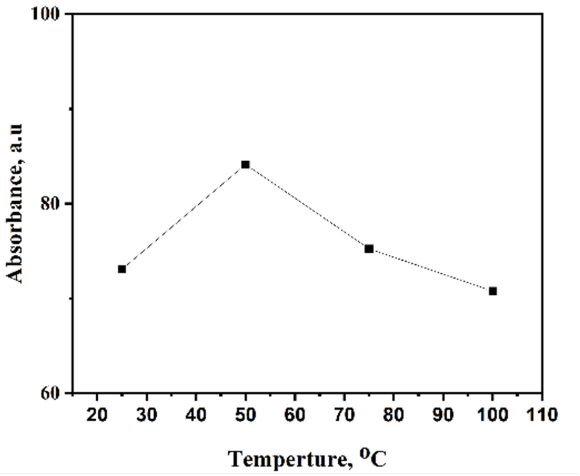


Fig. 8: The regression of Adsorption of Nitida and Temperature by changing the pollutants.

Testing the classification of the water potability on different 8 models, and XGBoost Classifier achieved the highest accuracy based on the reasons that are mentioned earlier in the theoretical section, as shown in Fig(9). Before Hyperparameter Optimization, XGBoost Classifier was set on 8, 125, 0, 0.04, and 5 in max_depth, n_estimators, random_state, learning_rate, and n_jobs with 66.913% accuracy, respectively, as shown in Table(2). After Hyperparameter Optimization using a Randomized Search with 30 iterations and Cross-validation (CV) of 5, XGBoost Classifier is optimized based on the best parameters, 0.8, 150, 1, 8, 0.01, 0, 0.9 in subsample, n_estimators, min_child_weight, max_depth, learning_rate, gamma, colsample_bytree, respectively, to higher benchmarks with 67.37% accuracy. Moreover, hyperparameter Optimization is applied to four classification algorithms, leading to an increase in their accuracy, as the Random Forest Classifier ranks second in accuracy at 66.82% after the XGBoost Classifier with the utilization of Bayesian Optimization; tuning on n_estimator ranges from 100 to 600, 100 max_eval, criterion of gini and entropy.

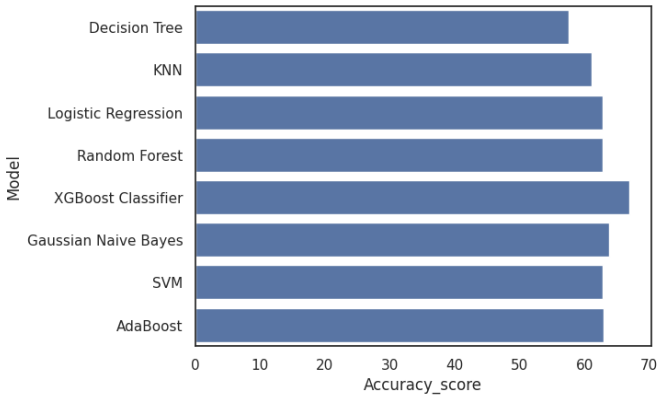


Fig. 9: The comparison between 8 different algorithms in Classification.

VII. Limitations

The provided public datasets on the internet aren't huge and precise enough to determine their effectiveness and effect on the model. Thus, we contacted **Laboratory of Analytical and Environmental Chemistry, Faculty of Science, Helwan University. All these data Are collected under the supervision of Associate Professor Farida Saad El-Din.** to get a review of the collected data and compare the predicted with the real-life application. Hyperparameter Optimization is time and computational consumption, which needs comprehensive trials for reaching the best parameters to achieve the highest benchmarks. Hence, applying hyperparameter optimization on the proposed algorithms in regression and classification.

VIII. Conclusion & Future Plans

In conclusion, XGBoost Classifier and Polynomial Regression with a degree of 5 achieved the highest accuracy over all the algorithms, leading to determine the adsorption of Ficus Nitida and the optimum point for the coming stage. For the future plans, we are looking forward to combine large language models (LLMs) with Chain-of-Thoughts or Retrivel Augmented Generation (RAG) to give more information about the Ficus Nitida, and collecting the data from the real scale project for optimal representation and effectivness on the real world.

References

[1] "water-related diseases responsible for 80 per cent of all illnesses, deaths in developing world', says secretary- general in environment day message — UN press," United Nations, https://press.un.org/en/2003/sgsm8707.doc.htm.

[2] "Facts and figures about water - worldwatercouncil.org," 6th World Water Forum Kick-Off

[3] E. and C. C. C. Government of Canada, "2015 summary report: Reviewed facility-reported data," Environment and Climate Change Canada - Pollution and Waste,

[4] A. VANHAANDEL, G. EKAMA, and G. MARAIS, "The activated sludge process?3 single sludge denitrification," Water Research, vol. 15, no. 10, pp. 1135–1152, 1981. doi:10.1016/0043-1354(81)90089-0XO. Çelebi, Ç. Üzüm, T. Shahwan, and H. N. Erten, "A radiotracer study of the adsorption behavior of aqueous Ba2+ ions on nanoparticles of zero-valent iron," Journal of Hazardous Materials, vol. 148, no. 3, pp. 761–767, 2007.

[5] C. Theivarasu and S. Mylsamy, “Removal of malachite green from aqueous solution by activated carbon developed from cocoa (Theobroma Cacao) shell—A kinetic and equilibrium studies,” E-Journal of Chemistry, vol. 8, no. 1, pp. S363–S371, 2011.

[6] I. Klyuzhin, A. Symonds, J. Magula, and G. H. Pollack, “New method of water purification based on the particle- exclusion phenomenon,” Environmental Science ;amp; Technology, vol. 42, no. 16, pp. 6160–6166, 2008. doi:10.1021/es703159q

[7] J. Grimm et al., “Review of electro-assisted methods for water purification,” Desalination, https://www.sciencedirect.com/science/article/abs/pii/S0011916498000472

[8] Priyesh Wagh a et al., “A new technique to fabricate high-performance biologically inspired membranes for water treatment,” Separation and Purification Technology, https://www.sciencedirect.com/science/article/abs/pii/S1383586615303221

[9] Chaohai Wang a b 1 et al., “In-situ fabrication of nanoarchitectured MOF filter for water purification,” Journal of Hazardous Materials, https://www.sciencedirect.com/science/article/abs/pii/S0304389420301527

[10] S. Bolisetty and R. Mezzenga, “Amyloid–carbon hybrid membranes for universal water purification,” Nature News, https://www.nature.com/articles/nnano.2015.310