Universidade Estadual Paulista
"Júlio de Mesquita Filho"
Programa Interunidades

unesp

Mestrado
Engenharia Civil e Ambiental

ABAYOMI OLUWATOBILOBA BANKOLE

MACHINE LEARNING FRAMEWORK FOR OPTIMIZATION OF FLOCCULATION PROCESS OF WATER TREATMENT

Bauru
2023

ABAYOMI OLUWATOBILOBA BANKOLE

MACHINE LEARNING FRAMEWORK FOR OPTIMIZATION OF FLOCCULATION PROCESS OF WATER TREATMENT

Dissertation presented as a requirement for obtaining the title of Master in Civil and Environmental Engineering at the São Paulo State University "Júlio de Mesquita Filho", Area of Concentration Sanitation.

Advisor: Prof. Dr. Rodrigo Moruzzi

Bauru
2023

B218m

Bankole, Abayomi Oluwatobiloba

Machine Learning Framework for Optimization of Flocculation Process of Water Treatment / Abayomi Oluwatobiloba Bankole. -- Bauru, 2023

104 f. : tabs., fotos

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Engenharia, Bauru

Orientadora: Rodrigo Braga Moruzzi

1. Floc length evolution. 2. Flocculation. 3. Machine Learning. 4. Neural Network. 5. Smart water treatment. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Engenharia, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE ABAYOMI OLUWATOBILOBA BANKOLE, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL E AMBIENTAL, DA FACULDADE DE ENGENHARIA - CÂMPUS DE BAURU.

Aos 02 dias do mês de junho do ano de 2023, às 10:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de ABAYOMI OLUWATOBILOBA BANKOLE, intitulada **MACHINE LEARNING FRAMEWORK FOR OPTIMIZATION OF FLOCCULATION PROCESS OF WATER TREATMENT**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. RODRIGO BRAGA MORUZZI (Orientador(a) - Participação Virtual) do(a) Departamento de Engenharia Civil e Ambiental / Faculdade de Engenharia de Bauru / UNESP, Prof. Dr. SOROOSH SHARIFI (Participação Virtual) do(a) Department of Civil Engineering, School of Engineering, University of Birmingham / University of Birmingham, Prof. Dr. ROGERIO GALANTE NEGRI (Participação Virtual) do(a) Departamento de Engenharia Ambiental / Instituto de Ciencia e Tecnologia do Campus de Sao Jose dos Campos / UNESP. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: APROVADO . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. RODRIGO BRAGA MORUZZI

DEDICATION

I dedicate this dissertation to God Almighty, my beautiful wife, Afolashade Racheal Bankole, and my amazing daughter, Rejoice Camila Bankole, for their perseverance.

ACKNOWLEDGEMENT

My sincere gratitude to my saviour, the Amighty God for the grace and privilege to successfully complete this master's degree program.

This great achievement and triumphant completion of my M. Sc. program and successful writing of my dissertation would not have been a reality without the steadfastness and supports I received from numerous wonderful people all around me, which aided my transition from Water Resources Management to Engineering; translating into quality outputs and expositions. Therefore, with plesaure in my heart, I would like to express my immense appreciation for the unwavering supports I received from those people, I consider, as parts of God's blessings for the success of this enviable journey. Firstly, I would like to appreciate the management of my home University (Federal University of Agriculture, Abeokuta) for considering me worthy of nomination for the Agricultural Research and Innovation Fellowship for Africa (ARIFA) scholarship, and the Tertiary Education Trust Fund (TETFund) and the Forum for Agricultural Research in Africa (FARA) for the partnership initiative. My sincere appreciation goes to my highly supportive wife and daughter (Afolashade And Rejoice Bankole) for their unquantifiable supports and encouragement all through the sleepless nights and several times spent away from home. My late father (M.A.O. Bankole) of blessed memory and my mother (Olaide Bankole), brothers (Imole, Sunday, and Damilare, Bankole) and sisters (Abiola Adebogun, Omobolanle, Yinka and Boluwatife, Bankole) for their tireless encouragements, care and love, albeit miles away.

Moreover, my gratitude to my supervisor, Professor Rodrigo Moruzzi Braga, is immensely immeasurable. His mentoring, motivation and encouragement to ensure my transformation into an independent researcher, as I have emerged a smart and innovative thinker through this process. Beyond his supervisory role, he saw to the wellbeing of my entire family. Additionally, the role of Professor Rogerio Gelante during the computation of my Machine Learning algorithms and transformation of the developed models into a relatable framework has shappened both my machine learning and modelling versatility and prowess. Equally, Professor Adriano Reis's unwavering supports and attention strongly influence my level of critical thinking, further bringing out the giant in me, in the field of coagulation and flocculation.

Importantly, I appreciate all the families and individuals that have invested into my education and career growth, Late Coach Victor Oshuntolu, Mr & Mrs Apelehin, Pastor and Dcns. Bode Maxwell, Dr. Olabanji Samuel, the entire FUNAAB Directorate of Sports, and my most referred mentor; Professor Grace Oluwasanya, I say a very big thank you for your tireless

supports and for believing in me. Also, I would like to appreciate Maria Victoria, Gabriela Santos for tirelessly assisting us all through our coursework phase with translation of lecture materials. The supports of Glauco Perpetuo, Thalita Santos, Eduardo Miguel, Caroline Pompei, Caroline Calil, and particularly Pedro for their exposition while traversing the laboratory experiments also built my confidence in handling varieties of laboratory duties. Worthy of mention are my colleagues from Nigeria, Abraham James and Emmanuel Babajide. The brotherly love, moral supports and practically helping out through the journey emboldened my poise to pull through.

RESUMO

A implementação do modelo de aprendizado de máquina (ML) que poderia melhorar tanto a eficácia como a sustentabilidade do sistema de tratamento de água é um grande problema no setor de água, com a otimização do processo de floculação sendo um grande obstáculo. Neste estudo, desenvolvemos o primeiro modelo ML para monitoramento da evolução do comprimento dos flocos e uma estrutura para sua potencial adoção em tratamento de água em larga escala. Modelos de Rede Neural Artificial (ANN) e Memória de Curto e Longo Prazo (LSTM), juntamente com o modelo tradicional de séries temporais: Média Móvel Integrada Regressiva Automática (ARIMA), foram explorados para prever os dados de evolução do comprimento dos flocos obtidos por análise de imagem não intrusiva de um ensaio de teste em lote e modelar o processo otocinético. Os dados do ensaio em lote, com dois gradientes de velocidade (Gf 20 seg^{-1} e 60 seg^{-1}) e tempo de floculação de 3 horas, foram divididos em 5 intervalos para faixas de comprimento de floco de 0,27 a 3,5 mm e otmizados usando o método linear. Os resultados mostraram que o modelo ARIMA não é adequado para prever o número de flocos, com uma acurácia de teste negativa (R^2). A ANN registrou R^2 de 0,86 – 1,0 para treinamento e 0,84 – 0,99 para teste, em Gf 20 seg^{-1} e Gf 60 seg^{-1}. O modelo LSTM tem a melhor precisão de previsão de 98 – 100% para Gf 20 seg -1 e previsão perfeita do número de flocos em todas os intervalos e Gfs. Nosso estudo comprovou que a estrutura desenvolvida pode ser replicada em tratamento de água em larga escala e promoverá a aplicação de tecnologia inteligente em tratamento de água/esgoto em larga escala.

Palavras-chave: Evolução do Comprimento dos Flocos; Floculação; Aprendizado de Máquina; Rede Neural; Tratamento Inteligente de Água.

Abstract

The implementation of machine learning (ML) model that could improve both the effectiveness and sustainability of water treatment system is a major problem in the water sector, with the optimization of flocculation process being a major setback. In this study, we have developed the first ML model for floc length evolution monitoring and a framework for its potential adoption in large-scale water treatment. Artificial Neural Network (ANN) and Long-Short Term Memory (LSTM) models, and traditional time series model; Auto Regressive Integrated Moving Average (ARIMA) were explored to predict floc length evolution data that was obtained through non-intrusive image analysis from a jar test batch assay and model the orthokinetic process. Batch assay data of two velocity gradient (G*f* 20 sec^{-1} and 60 sec^{-1}) and flocculation time of 3hrs were partitioned into 5 bins for floc length range 0.27 – 3.5 mm and upscaled using linear method. Results showed that ARIMA model is not suitable for predicting number of flocs with a negative test accuracy (R^2). ANN recorded R^2 of 0.86 – 1.0 for training and 0.84 – 0.99 for testing, across G*f* 20 sec^{-1} and G*f* 60 sec^{-1}. LSTM model has the best prediction accuracy of 98 – 100% for G*f* 20 sec^{-1} and perfect prediction of number of flocs across all bins and G*f*s. Our study has proven that the developed framework can be replicated in large scale water treatment and will promote application of smart technology in large-scale water/wastewater treatment.

Keywords: Floc length evolution; Flocculation; Machine Learning; Neural Network; Smart water treatment.

List of Figures

List of Table

List of Acronyms

AI	Artificial Intelligence
ANFIS	Adaptive fuzzy neural inference system
ANN	Artificial Neural Network
BPANN	Back propagation artificial neural network
BOD	Biological oxygen demand
COD	Chemical oxygen demand
CLD	Chord length distribution
DIA	Dynamic Image Analysis
DL	Deep learning
DNN	Deep neural network
EANN	Elman artificial neural network
ERT	Extremely randomized tree
FF	Feed forward
FBRM	Focus beam reflectance measurement
GRNN	Generalized regression neural network
GSGMDH	Generalized structure of group method of data handling
KNN	K-Nearest neighbor
LPSA	Laser particle size analyzer
LSSVM	Least square support vector machine
LSTM	Long short-term memory
MAE	Mean absolute error
ML	Machine Learning
MLP	Multilayer perceptron
MLR	Multiple linear regression
MSE	Mean Squared Error
NTU	Nephelometric turbidity unit
PDA	Particle dispersion analyzer
PDP	Partial dependence plot
PSD	Particle size description
PVM	Particle vision measurement
RBFNN	Radial basis function neural network

RF	Random forest
RMSE	Randomized mean square error
SDG	Sustainable Development Goal
SOM	Self-organizing map
SED	Statistical experimental design
TSS	Total suspended solid
VIM	Variable importance measure
WWTP	Wastewater treatment plant
W-WWTP	Water and Wastewater Treatment Plant

List of Symbols

A	aggregate area
β	power law slope coefficient
β_0	intercept
$\beta_{1...n}$	slope (unknown)
Dip	geometric average of aggregate range per scale
ε	regression error
G*f*	velocity gradient
K	density coefficient (dimensionless)
L	length of aggregate
D*f*	fractal dimension
η	kolmogrov microscale
R^2	coefficient of determination (r squared)
v	viscosity of water (m^2s^{-1})
T*f*	time gradient

TABLE OF CONTENTS

1 INTRODUCTION

Globally, demand for safe water for domestic uses and suitable water for agricultural purposes increases annually. Stressing the need for sustainable and ecofriendly treatment of water and wastewater to ensure adequate removal of pathogens and promote public health safety. Whereas the increasing effect of Climate Change is aggravating these demands, technological advancements to enhance the performance of conventional water treatment including Forward and Reverse Osmosis systems (Altaee et al., 2017; Y. Kim et al., 2017; Ibrar et al., 2019), Nano filtration (Cortés Muñoz et al., 2013; Kim et al., 2017), graphene technology (Park et al., 2018; Ganguli et al., 2020; Parsa et al., 2020) among others, with potential to treat even high strength wastewater still have major shortcomings. These include the non-availability of materials, high energy consumption, technicalities, and cost-effectiveness.

On the other hand, the coagulation and flocculation process is an essential part of water and wastewater treatment (Kurniawan et al., 2022), either in the conventional water treatment process or advanced treatment like Nano-technology (Abu-Dalo et al., 2022) and graphene technologies (Parsa et al., 2020). Coagulation/flocculation is a primary water treatment phase with high efficiency in suspended particles/pollutants removal (Abu Bakar et al., 2021; Kurniawan et al., 2022). Flocculation has been in practice for over ten decades, with the sole principle of inducing shear stress to water that is dosed with coagulant/flocculants, to form aggregates (flocs) that can settle and be removed with time (Jiang, 2015). Essentially, the performance of the flocculation phase and flocs properties determine the output and efficiency of most water and wastewater separation technology within treatment plants (W-WWTP) (Xiao et al., 2011), particularly the widely used conventional treatment technology.

Great advancement has been made in both coagulation process (coagulant type and dosage) (Rajala et al., 2020; B. Singh & Kumar, 2020; Y. Huang et al., 2021; Ortiz et al., 2021; Tegladza et al., 2021) and flocculation phase of water and wastewater treatments. This includes the modeling and simulation of flocs characteristics, fractal dimensions, and aggregates (Rong et al., 2013; Moruzzi et al., 2017; H. Zhang et al., 2019; Moruzzi et al., 2019; Hadiyanto et al., 2021; Teixeira et al., 2022). Despite the progress recorded in both laboratory and pilot studies of water and wastewater treatment, typical conventional water treatment facility faces the problem of source water quality monitoring, to determine the coagulant dosing and timing of the flocculation stage. This often leads to man-hour loss, energy wastage (due to multiple jar tests), resources wastage (resulting from multiple water quality testing), and ultimately protracted access to timely informations on adequate correctional changes that could enhance the treatment effectiveness (water quality).

Smart solutions such as machine learning (ML) and artificial intelligence (AI) have been explored in recent time to mitigate these operational problems in drinking water and wastewater treatment facilities. The influx of research on AI implementation in real-time water treatment problems and their potential as transformational solution include pilot and full-scale treatment facilities. For instance, Guo et al. (2015) applied two robust (high prediction accuracy and sensitivity) machine learning models to predict One (1) day of total nitrogen at the wastewater treatment facility in Ulsan, Korea. The study established that the non-linear time series model is efficient to forecast the 1-day nitrogen content in the treated water quality. Additionally, ML models have been applied to optimize different stages within the water/wastewater treatment technology, including coagulant dosage (Zhang et al., 2013; Kim & Parnichkun, 2016), water quality (Granata et al., 2017; Igwegbe et al., 2021), among others.

Although the application of ML in the expression and forecasting of flocs characteristics are very limited. Li et al. (2021) established that despite recent advancements in the application of both online and model-based ML/AI principles to enhance water treatment performances, little information is available on the use of machine learning to model phenomena within flocculation processes, including flocs characteristics and optimum timing for jar test/flocculation process. The gaps considered are relative to the nonlinearity of hydraulic or hydrodynamic-based models relating to flocculation processes, making the tedious laborious jar test and loss of man-hours persist.

Therefore, this study seeks to develop a machine learning framework to model the flocculation process of water/wastewater treatment, with a focus on developing a robust user-friendly model to predict flocs characteristics and the optimum timing of the flocculation phase in water and wastewater treatment facilities.

1.1 Justification of Research

Understanding of flocculation process requires detailed examination of all the phenomena involved and their guiding principles. Non-intrusive Dynamic Image Analysis (DIA) is the best flocs characterization method (fractal aggregate and dimensions) that provides the best information about the flocculation process, without tampering with the jar test setup or causing alteration in the flocs integrity (Liang et al., 2015). Modeling and simulation of flocs characteristics (fractal aggregate, breakage, and re-aggregation), flocculation time, and effect of shear (gradient) velocity have impressively been explored to promote understanding of the flocculation process in the past decades (Moruzzi et al., 2017, 2019; Zhang et al., 2019). Although the introduction of diverse approaches to simplify the interaction between flocs properties and shear velocity/time varies among authors, primarily due to the nonlinear relationship between variables.

Machine learning models with the capability to solve nonlinear relationships with few lines of code and model optimization time has been applied to demystify the water and wastewater treatment process in several studies. For instance, Wang et al. (2021) used a novel adaptation of the Random Forest model (RF), Deep Neural Network (DNN), Partial Dependence Plot analysis (PDP), and Variable Importance Measure (VIM) to improve lag time within treatment processes, water quality and operational cost in the Umea Wastewater Treatment Plant (WWTP), Sweden. Elman Artificial Neural Network (EANN) was compared with Multiple Linear Regression (MLR), Least Square Support Vector Machine (LSSVM), and Radial Basis Function Neural Network (RBFNN) by Wang et al. (2022) to establish the robustness of hybridized Artificial Neural Network (ANN) based models over MLR, SVM, and RBFNN. While the model achieved predicting coagulant dosing accuracy that could be applied in a drinking water treatment facility.

Great advancement has been made in the use of nonlinear machine learning and artificial neural network models to predict coagulant dosage and water quality. Jayaweera et al. (2019) developed an ANN model for the prediction of coagulant dosage with a

training time of less than 30 seconds. This shows the applicability, robustness, and efficiency of ANN in predicting coagulant dosage. Nevertheless, fewer information exists on the use of ML or AI-based model to predict and establish the floc evolution in a flocculation process. Particularly on the relationship between floc lengths, shear velocity and flocculation time, and the optimum time for the flocculation process that could further reduce man-hour loss and energy consumption in treatment facilities. Such models could provide better prediction of flocs evolution and flocs characteristics, aggregation, breakage, and re-aggregation.

In addition, lack of adequate data, nonlinearity of the relationship between phenomena in the flocculation process, and application of complex hydraulic/hydrodynamic principles are the main setbacks that drifted attention away from modeling flocs characteristics. For instance, DIA data (Alum floc) from a water treatment facility in Finland was modeled by Juntunen et al. (2012) using a Multiple Linear Regression and Self Organizing Map (SOM) for online characterization of flocs surfaces and floc length aggregation, breakage, and re-aggregation. The authors affirmed that the DIA setup is cost-saving and applicable in full-scale water and wastewater treatment (conventional) facilities. The image data acquired during the flocculation phase could be integrated into an ML and/or AI framework to form a seamless process of analyzing data and further promote understanding. This would be a major leap to have an early warning prediction tool, consequently reducing man-hour loss and management operational costs.

Advances in the modeling of the flocculation process have been on the possibility of integrating first principle or kinetic models with ML/AI. Such an approach was applied in the study of Nazemzadeh et al. (2021), population balance model and mass balance was used as the first principle models, integrated with a machine learning framework to predict the flocculation process for flocs aggregation and breakage. The author established that the machine-learning framework enhanced the prediction accuracy of the end-of-batch in the flocs' distribution. Combining the data of physical observations (floc lengths) during flocculation with a machine learning model could give a more robust approach to predicting flocculation behavior during water treatment (Nazemzadeh et al., 2022).

However, studies that applied only machine learning models in demystifying flocculation phenomena are limited. Most importantly, machine learning-based models that could provide insight into flocs characteristics (aggregation, breakage, and reaggregation),

effects of induced share velocity (G*f*) per given gradient time (T*f*), which could enhance treatment operation and drastically reduce the complexities in monitoring flocculation process of water treatment. Therefore, this study aims to develop a machine learning framework to establish an underlying relationship between floc lengths, time gradient (T*f*), and velocity gradient, and predict the floc evolution and optimum time for a flocculation process, toward achieving a cost-efficient, time-saving and sustainable conventional water treatment facility.

2 OBJECTIVES

This study aims to develop a machine learning model framework to predict the floc particle evolution and the relationship between the gradient velocity and time gradient of the flocculation process and predict the optimum time for a flocculation process.

2.1 Specific Objective

1. Model and predict the particle floc length evolution using different machine learning algorithms.
2. Develop a predictive Machine Learning Framework for the modeling of floc evolution process of flocculation phase in a water and wastewater treatment facility.

3 LITERATURE REVIEW

3.1 Flocculation Process and Application

Application of flocculation transcends beyond water and wastewater treatment as its importance in the pharmaceutical, papermaking, and mineral processing industry for enhancement of tailing settling rate cannot be overemphasized. Generally, addition of coagulant at a certain dosage and induced shear velocity (G*f*) creates the interactions of colloid particles to form larger particles regarded as flocs, with higher settling ability. The aggregation of flocs into larger sizes that can settle translate into the quality of the treated water (Jiao et al., 2016), which is measured in terms of turbidity, pH, and other water quality indicators (Wang et al., 2011; Zhang et al., 2019; Li et al., 2021).

Several studies have established the effectiveness of flocculation in the removal of emerging contaminants including; leachate (Cheng et al., 2020; Chaouki et al., 2021; Cheng et al., 2021; Igwegbe et al., 2021; Reddy et al., 2022), Polystyrene Micro and Nano particles (Chen et al., 2021; Li et al., 2021), Polyethylene Microplastics (Li et al., 2022), metals (Nyström et al., 2020; Zhang et al., 2020), etc. Though the effectiveness of water treatment is the quality of the treated water or effluent, an adequate understanding of the complex interactions between the stages of water treatment guarantees better output (quality) and timesaving.

Since floc growth has been recognized as the main principle of flocculation, appreciable progress has been made in identifying the drivers of fractal aggregate (Bushell, 2005; Moruzzi et al., 2017, 2018), the active agent in the coagulants/flocculants materials and effectiveness (Abu Bakar et al., 2021; Jiao et al., 2016; Kurniawan et al., 2022; Teh et

al., 2014), effects of shear velocity (G*f*) on floc evolution, and simulation of the processes (Du et al., 2021; Lawrence et al., 2022; Moruzzi, Bridgeman, et al., 2020; Moruzzi et al., 2019; Qasim et al., 2020, 2021).

For instance, Kurniawan et al. (2022) comprehensively reviewed current practices and applications of natural (Bio-coagulant/Bio-flocculant) and synthetic coagulants/flocculants. Authors established active compounds in flocculants (carboxyl, amine, hydroxyl, and protein) as well as working mechanisms; sweep coagulation, particle bridging, and ionic layer. Moruzzi et al. (2017) studied the influence of shear velocity (G*f*) on two-dimensional fractal dimensions (D*f*) of large aggregate (flocs > 540µm) from a simulated Kaolin suspension jar test experiment. The study also characterized the floc lengths using a non-intrusive image analyzer, particle size distribution (PSD) and varying G*f*. Authors used existing hydrodynamic principles for the continuous function of PSD (Eqn. 1.), power law for fractal dimensions (Eqn. 2.), and Kolmogorov microscale (Eqn. 3.). Findings showed that different dynamic steady-state exists for different G*f*, and interactions between particles and clusters differ with certain shear velocity gradient (G*f*) (i.e., particle-cluster and cluster-cluster interactions).

Lopez-Exposito et al. (2019) used another expression for the scaling of power-law to derive fractal dimensions of microalgae flocs cultured from *chlorella sorokiniana* using a photobioreactor. The study also correlates D_f with the chord length distribution (CLD) of the flocs using a Random Forest (RF) machine-learning model.

$$\frac{dN}{d(d_p)} = k(d_p)^{-\beta} \tag{1}$$

where: d_p is the geometric average of aggregate range per scale, k is the power law density coefficient (dimensionless), and β is the power law slope coefficient.

$$A \sim l^{D_f} \tag{2}$$

where: A is the area of aggregates (mm^2); l is the characteristic length of the aggregate (mm); D_f is the two-dimensional fractal dimension which is dimensionless.

$$\eta = \left(\frac{v}{G}\right)^{\frac{1}{2}} \tag{3}$$

where: η is the Kolmogrov microscale (mm); v is the kinematic viscosity of water (m^2s^{-1}); G is the velocity gradient (s^{-1}).

Zhang et al. (2019) studied the effects of shear velocity gradient on flocs evolution of microalgae cultivated water using aluminum chloride coagulant. A high-speed microscopic camera was used to capture fractal images and analyzed them to find the floc lengths, fractal dimension, floc strength, breakage, and aggregation. Floc evolution was characterized by the first decrease in fractal dimensions before a rise in dimensions then followed by a steady-state fractal dimension under a given shear rate (G*f*). Similarly, the trend was observed for varying G*f* but optimum floc length was found under a steady shear rate of $9s^{-1}$. Furthermore, the authors established that general floc growth (sizes and fractal dimensions) and time gradient (T*f*) depends on the shear velocity gradient G*f*, which corroborates the findings of previous studies such as Moruzzi (2017, 2019, and 2021).

While those studies use different approaches to examine floc evolution during the flocculation process, similar conclusions were made regarding the influence of shear velocity gradient and time gradient. However, past studies mainly use log-log functions to establish the graphical relationship between floc lengths, time gradient (T*f*), and velocity gradient to derive floc strength but could not establish the variables with the highest importance and their level of significance in the flocculation process. In addition, information on the prediction of scenarios surrounding flocs evolution under prolonged time/velocity, which could serve as early warning mechanisms and source of insight into variation in source water characteristics (pollution load) for a treatment facility is lacking.

Notably, different coagulants are generally used for different types of source water and laboratory scale studies, which implies that variables considered by studies vary but insight into floc lengths and their measurement remains crucial for a better understanding of flocs evolution.

3.2 Floc length Measurements

The complexity of the flocculation process has led to different approaches to measuring floc dimensions, particularly floc lengths. Considerations on the accuracy of individual methods and the integrity of the flocs aggregate are critical issues that many studies have dealt with. Generally, floc length measurement can be divided into two broad groups: non-intrusive method and sampled method. The difference between the two

approaches is mainly the contact with the flocs; the in-situ method measures without physical contact with the flocculation setup, while the flocs may be distorted or sheared during the sampled method.

A review by Liang et al. (2015) detailed modern approaches in floc length measurement, classifying the current practices into five methods, namely; Laser Particle Size Analyzer (LPSA), Microscopy, Focus Beam Reflectance Measurement (FBRM), Particle Vision Measurement (PVM), and Direct Image Analysis (DIA). The LPSA uses a light beam that is channeled through a particle field, the rays are reflected at certain angles when it comes in contact with flocs/particles or move in a straight line when no contact with floc is made and the scattering angles (differences between the incident and reflected angles) is negatively correlated against particle size. Optical detectors are positioned around the particle field to gather information on the scattering angles and intensity of the reflected light rays. A notable advantage of this method is the speed, high accuracy (in terms of light rays), and the ability to quickly repeat the process, but major shortcomings are tendencies of flocs breakage through the ultrasonic treatment and circulating pumps. While its floc lengths result is inaccurate, unlike the image analysis.

FBRM is one of the best approaches in floc length measurement. The principle acquires floc chord length by multiplying the speed of the beam scanner and the time between recorded reflections which is a reflection from the two opposite edges of a particle/floc. FBRM operating principle is very close to the LPSA, but the difference is the non-disruptive setup of the FBRM. A projected beam moves around the circumference of a sapphire window at a given speed (<6m/s) and the scattered light by particles is recorded through the optical detector. Avoidance of floc breakage and adequate measurement of floc growth (including breakage and re-aggregation) are major advantages of the FBRM. Though major setbacks are the accuracy of the optical properties, focal points, shape of floc property, and particularly the presence of the probe. The flocs' movement around the probe could easily be altered, which could affect the measurement accuracy. PVM is also similar to the FBRM as both methods can be combined to determine the structures of slurry and particle size distribution (Nasser & Salhi, 2015; Qi et al., 2015). However, they have low resolution compared to the FBRM.

The Microscopy method has received lesser attention compared to other particle size measurement approaches. Although the reason for this is not known. Application of

the microscopy method has been adopted in flocculation studies earlier than the study of Suhr et al. (1995) which reflects the potential of the approach. In the microscopy method of particle size measurement, a microscope of a high-resolution microscope that has a light source fixed to the camera lens is mounted on a hole/small port of the experimental setup (jar test/bioreactor) to capture an image of the primary particles at intervals. Suhr et al. (1995) improvised the approach to study cell characterizations in a bioreactor by generating still images that were further processed by the image algorithm. The approach was used to introduce online monitoring of cell behaviors in the bioreactor.

Santos Nunes et al. (2022) adopted the microscopy approach to characterize flocs evolution (sizes) during a jar test experiment. A light source was connected to a camera with an objective lens to capture images of floc particles in suspension through a given window in the setup. A unique advantage of this approach is the ability to capture smaller particles or flocs that could be accounted as primary particles in a flocculation process. However, the small window (coverage) of the images captured limits their effectiveness in accurate floc lengths evolution. Nevertheless, this approach could be improved upon, particularly since it could allow easy and online processing of data, making full automation of the flocculation process a possibility.

The DIA method is also close to the LPSA method except that images are taken in Laser images which are also close to the common microscopy method. Though images are captured in a dynamic mode where the shutter speed is high and multiple images can be captured at different intervals within the experiment. The flow of water through the observation pipe of a DIA could be different from the real flow within the flocculation chamber. Nevertheless, the high processing speed of the DIA and the ability to process information fast through a computer system gives the DIA an edge. Particularly considering the technological advancements in the general water treatment sector. An image of a DIA set up according to Liang et al. (2015) is presented in Figure 1 below.

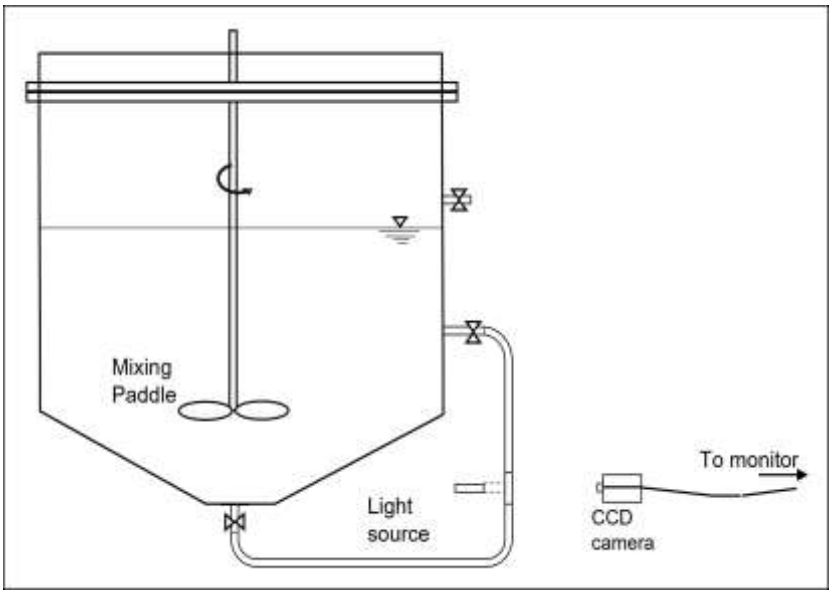

Figure 1. A schematic setup of a DIA image analysis system for particle size measurement.

Several studies have attempted to hybridize the setups explained above. For instance, Moruzzi et al. (2019) combined a photometric dispersion analyzer (PDA) as an intrusive method with a non-intrusive image analysis measurement of floc characteristics (size and strength). Although the authors affirmed that the PDA could not give exact particle size measurements of the flocs rather the information could be used to determine floc aggregation and flocs strength, while the image analysis is more reliable. Integration of the floc length measurement processes into a machine learning algorithm could present better means of evaluating flocs evolution within the flocculation process of water and wastewater treatment.

3.3 Modeling of Flocculation Process

Modeling of the flocculation process either in the water treatment or the paper-making industry has received research attention in recent years. The phenomena of the flocculation process have been keenly studied by researchers, to promote an understanding of the complex interaction between the principal variables (gradient velocity, time gradient, floc length). In past decades, kinetic models have been introduced to establish the importance of individual phenomena of the flocculation process. For instance, theories such as the power law and orthokinetic models have been applied to interpret jar test experiment data.

The population balance equation/model (PBE/PBM) is one of the most applied kinetic models for particle size evolution. PBM is a mass balance approach for the distribution of discrete particle size with an interest in the influence of shear stress (G*f*) on aggregation and breakage rate on the population of particle size(s) (Singh et al., 2022). PBM was introduced by Hounslow et al. (1988) and later modified by Spicer & Pratsinis (1996), afterwards, several researchers modified the PBM to develop interesting kinetic models that have improved the knowledge of the flocculation process (Costa et al., 2007; Nazemzadeh et al., 2021; Ruan et al., 2022; Seghir et al., 2022; Singh et al., 2022). PBE equation has been applied in many fields aside from water and wastewater treatment, such as polymerization (Ahamed et al., 2020), material synthesis (Buddhiraju & Runkana, 2012), and crystallization modeling (Costa et al., 2007) among others.

Ruan et al. (2022) adopted the PBM model to develop a kinetic model to express the flocculation phase in a polymer bridging process of total tailing. A regression model was applied after splitting the PBM equation into aggregation kernel (collision frequency and efficiency) and breakage kernel (breakage rate and distribution function). The modeled PBM was achieved by combining a computational fluid dynamics (CFD) model with the PBM equation. The kinetic model developed was cross validated by comparing the modeled result with primary data. The floc length distribution data obtained from an FBRM setup was compared with the modeled PBM results to ascertain its high accuracy.

Mathematical expressions have also been used to model the flocculation process of water treatment, aside from the PBM model. Argaman & Kaufman (1970) developed a mathematical expression (diffusion model) for the orthokinetic flocculation in a steady state (further referred to as the Argaman-Kaufman model). The Argaman-Kaufman model has been adopted in studies and regarded as a principal model in the orthokinetic flocculation in a steady state (Eqn. 4). A continuous stirring tank reactor (CSTR) was used to examine floc formation and assumed that primary particle collides to form bigger particles (flocs), which can settle and be removed during sedimentation/clarification.

The mathematical expression of the Argaman-Kaufman model is given below.

$$\frac{\partial n}{\partial t} = k_B n G^m - k_A n_D G \tag{4}$$

where, n is the number of primary particles per unit volume of water (m^{-3}), t is the time (s), k_B is the floc breakup rate coefficient (s), G is the velocity gradient (s^{-1}), m is a parameter, k_A is the floc aggregation rate coefficient (-), and n_D is the number of floc particles formed per unit volume of water (m^{-3}).

The integrated and rearranged form of Equation 4, according to Bratby (2016) is presented in equation 5, below.

$$\frac{n_o}{n_i} = \left[\frac{K_B}{K_A} G + \left(1 - \frac{K_B}{K_A} G\right) e^{-K_A GT}\right]^{-1} \quad (5)$$

The theory was validated through experimental studies by Haarhoff & Joubert (1997). A unique contribution of the study by Haarhoff & Joubert (1997) besides the validation is the clarification of the critical things to note in the flocculation process, such as the settling time, and possible variation in flocculation constant when flocculation is applied, among others. Akinmolayan et al. (2015) developed a mathematical model for a complete water treatment process, following a simple conventional water treatment plant design (coagulation/flocculation – clarification – filtration phases). Authors adopted the Argaman-Kaufman model in modeling the flocculation phase of the treatment design and further validate their model results with the experimental datasets achieved by Haarhoff & Joubert (1997).

Despite the invention of different kinetic models in expressing flocs evolution/floc length measurements during the flocculation process, and the effects of G*f* and T*f* on floc lengths and shapes, the real-time problem still includes the laborious work in getting an accurate result that is useful for a treatment facility. Therefore, simplifying the process or integrating the models to achieve a time-effective and early warning system, such that would make the water/wastewater treatment process faster, cost-effective, and sustainable would be a game changer. The evolution of machine learning and artificial intelligence could have shown that ML/AI is the game changer that could drastically reduce the existing limitations.

3.4 Application of Machine Learning and Artificial Intelligence in Flocculation Process

The universality of machine learning (ML) and artificial intelligence (AI) has become obvious with the implantation of ML/AI-based technological solutions to relatively all human problems. Water and wastewater treatment has never been left behind in the integration of smart information-based technology. In the last decade, massive progress has been recorded in the integration of AI/ML into the water and wastewater treatment systems both on a lab scale or pilot scale and in full-scale treatment facilities (Ghaed Rahmati et al., 2021; Li et al., 2021). The broad spectrum of ML/AI models made integration easier since data could be acquired both online and offline.

Full-scale water and wastewater treatment facilities in both developed and developing countries are speedily adopting smart solutions to improve conventional water treatment systems. About a decade ago, Nasr et al. (2012) developed an artificial intelligence-based system to predict the water quality of the El-Agamy wastewater treatment plant (WWTP) in Egypt. The study used a feed forward (FF) artificial neural network to monitor and predict the Chemical Oxygen Demand (COD), Total Suspended Solids (TSS), and Biological Oxygen Demand (BOD). The model achieved a high accuracy with a correlation coefficient of $R^2 = 0.90$ for the observed variable against the predicted variable.

Khatri et al. (2019) also used a feed forward (FF) ANN algorithm to predict the water quality of a sequential batch reactor in a municipal wastewater treatment facility in Jamnagar, India. The water quality data for pH, BOD, COD, TSS, ammonium nitrogen and Total Kjeldahl Nitrogen, and total phosphorus of the influent were used to predict the effluent quality. Authors emphasized that FF models are compatible with modeling the effluent water quality in a full-scale water/wastewater treatment facility. Also, the ability to optimize the treatment efficiency makes treatment plant management easier and improves treatment reliability.

Mundi et al. (2021) applied multiple linear regression (MLR) and generalized structure of group method of data handling (GSGMDH) ANN black box model to predict the treated water quality of thirteen wash water from fourteen wastewater facilities in Ontario, Canada. The study achieved about 83 percent accuracy for the MLR model and up to 99 percent for the GSGMDH model, respectively. Li et al. (2021) reviewed recent

advancements in the application of ML/AI in water and wastewater treatment, the study highlighted advances in coagulation and flocculation modeling by recognizing different ML/AI models that have successfully been deployed, including Back Propagation Artificial Neural Networks (BPANN), multilayer perceptron (MLP), and generalized regression neural network (GRNN), which have been used to predict optimum coagulant dosage. The study also highlighted that GRNN is efficient for modeling with limited data while MLP is suitable for modeling full-scale water facility.

Also, models such as adaptive fuzzy neural inference system (ANFIS) were tried in online mode in comparison with the ANN model (Heddam & Dechemi, 2015), with the ability to update the training data set (input) immediately as they were available. Zhang et al. (2013) combined K-Nearest Neighbor (KNN) with a Support Vector Machine (SVM) to predict the coagulant dosage in three different scales of water treatment facilities: large, medium, and small scale. The author established that KNN-SVM outperformed the only SVM algorithm in a large and medium-scale treatment facility. While KNN alone was better for small-scale facilities. Further stressing that it is difficult to affirm the best ML model for a particular scenario unless they are tested.

This has been the situation with the ML models applied to coagulant dosage in different studies, although model accuracy differs. Random Forest (RF) models have recently been explored by researchers in the modeling of coagulant dosage and water quality of both water and wastewater treatment facilities. RF is an ensemble technique that decorrelate a simple decision tree to enhance accuracy in the algorithm. RF can create a relationship between the predictor variables and other variables based on set rules (trees) (as shown in Figure 2).

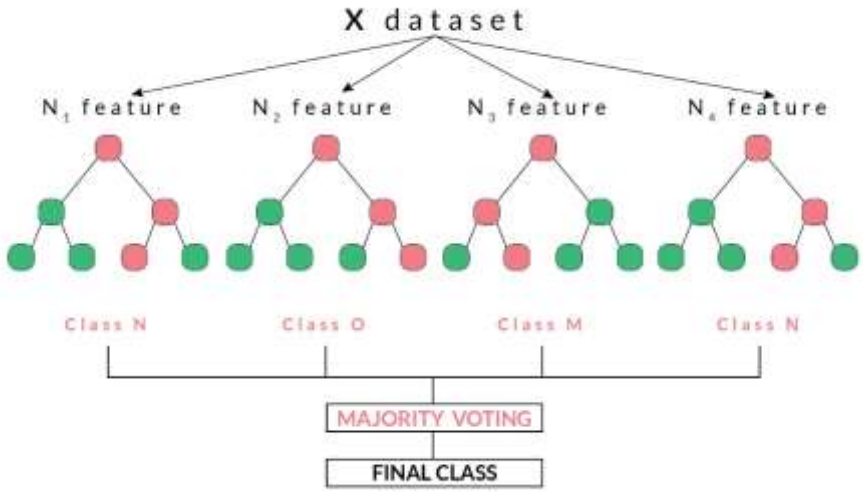

Figure 2. A simple Random Forest Algorithm. Source: (Quantinsti, 2019)

Heddam (2021) applied extremely random tree (ERT) and random forest (RF) models to predict coagulant dosage and concentration of dissolved oxygen, pH, turbidity, specific conductance, and temperature of water quality in Boudouaou water treatment facility, Algeria. Both RF and ERT showed high accuracy results with low randomized mean square error (RMSE) and mean absolute error (MAE), but ERT performed better than RF (R^2 up to 0.999) and works well with both large data and smaller data sets. The efficiencies of ERT and RF were compared to MLR, and both models (ERT and RF) outperformed the MLR model.

The metal removal efficiency of a chitosan-based flocculant was predicted using an RF algorithm in the study of Lu et al. (2022). The authors explored an 80:20 train-test data ratio with a random state of 10, the model was cross-validated with an ANN model. The RF model performed better than the ANN model, having an R^2 of 0.9354. The author also emphasized that the ability of the RF framework to predict complex processes, where a lesser understanding of both input and output parameters exist are the major advantages of the RF model and their choice of RF model for the study. Although the ANN algorithm is also built to demystify complex and hidden relationships within processes such as the flocculation process, even the entire water and wastewater treatment.

A more robust (accurate) RF framework with boosting; Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) was explored by Wang, et al. (2022) to improve their RF model on the management of wastewater treatment process.

Findings showed that the XGBoost model performs best while RF has the least performance. The introduced model elucidates underlying process-based problems that could enhance the performance of the treatment facility. XGBoost algorithm operates with the principle of gradient boosting decision trees (GBDT) which is a sequential iterative process that utilizes residuals from a trained tree to regress to a new tree. An example of XGBoost framework is presented in Figure 3 below.

XGBoost uses a loss function and controls model overfitting through the hyperparameters tuning. The outstanding performance of XGBoost over other tree-based and non-tree-based algorithms like SVM have also been established in studies (Ching et al., 2022; Didavi et al., 2021; Pan, 2018; Sahin, 2020). Despite the prospect of the tree-based algorithm in modeling of flocculation phenomenon, no research has explored the possibility of modeling floc length evolution, which is an essential aspect of the flocculation process, using any of the tree-based models.

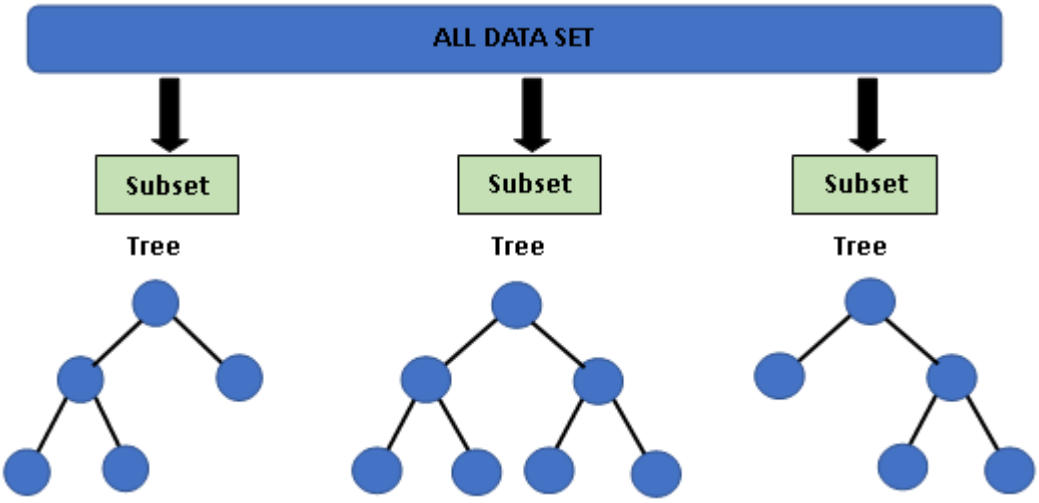

Figure 3. An XGboost model framework.

Neural networks (NN) are algorithms developed by an inspiration of neurosciences, with an aim to mimic human brain function. The inspiration of NN stems from the fact that biological organisms' (including human) decision and response to his environment are solely guided by the previous information at their disposal. Therefore, it is believed that machines can learn from data and make appropriate decisions. Considering human limitations with several complex problems such as in the case of extremely long mathematical calculation/operations. Computers are believed to maximize information through the leaning (train dataset) process and make accurate prediction of the expected

outcome from unseen data (test dataset) at a faster rate. Neural networks are the building block of Artificial Intelligence and many technological innovations in our world today. Leveraging on their ability to train on varieties of data (image, text, audio, numerical data) and make reliable prediction and seamless delivering of tasks.

Just as the name implies "neural network", NN processes and computes information through the "neuron" and screens the generated results through the "weight". The weight is a scale given to each processed data from one neuron to another. Each architecture (MLP, CNN, GRN etc.) has its merits and demerits, creating trade-offs in their usage. Thus, different algorithm/configuration are used for different purposes, while some algorithms have many purposes such as the LSTM (with ability to process convolutional data; image, time series/sequential vector data, etc.).

FF is a simple ANN architecture with one directional input processing direction, also called Multi-Layer Perceptron (MLP) model. An artificial neural network is governed by a simple mathematical expression (as shown in Equation 6 below) and a simple framework that takes in data through the input and processes it through the hidden perceptron to generate an output (Gharabaghi & Sattar, 2019).

The Equation for FF is presented as shown below:

$$y(x) = f\left(\sum_{i=1\ to\ n} w_i \times x_i + b\right) \tag{6}$$

where; y(x) is the output of the neural network for a given data (x), *f*() is the activation function used for the sum of input variables. N is the number of neurons, w_i is the weight assigned to the neurons, x_i is the input to the *i*th neuron, and b is the bias added to the weighted sum. As earlier discussed in this work, ANN has been explored to optimize different parts of water and wastewater treatment process and has recorded tremendous prospect in flocculation studies, including flocs evolution process. An example of a simple FF is presented in Figure 4 below.

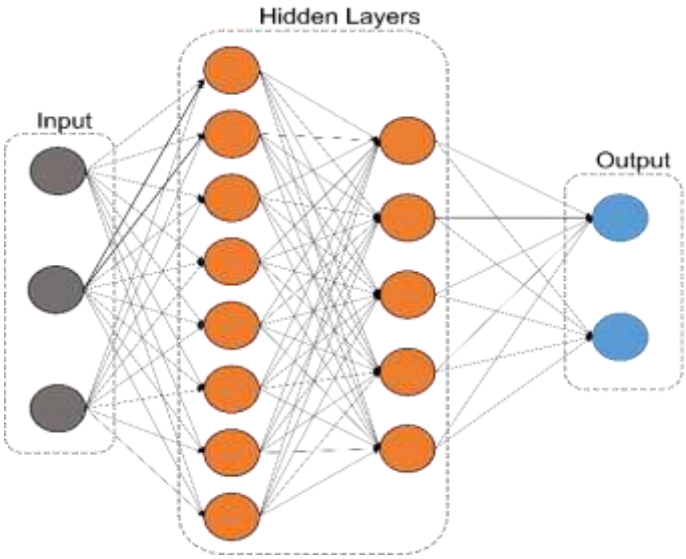

Figure 4. A simple Artificial Neural Network Framework.

Notably, Recurrent Neural Network (RNN); a deep learning neural network model with the capability to interact with both images, qualitative and quantitative data, is the least explored deep learning (DL) algorithm by studies on ML/AI and its application in water and wastewater treatment. Long Short-Term Memory (LSTM) is a type of RNN that works with the principle of recurrent memory of information (stored at the controlled recurrent gate), particularly when dealing with big dataset. The process of recurring information is called Backpropagation (BP), information is passed through a gate to recall the trend in the training data (information) that the system has processed. An example of the structure of the hidden layer cell of LSTM is presented in Figure 5 below.

Huang et al. (2021) affirmed that RNN has more prediction accuracy than GRNN and outperforms GRNN, even when trained with a small dataset. Furthermore, the authors affirmed that RNN with the LSTM algorithm is the best RNN-based model for time series data. This assertion corroborates the findings of El-Rawy et al. (2021) that applied various deep neural network models with back propagation and feed forward algorithms to predict and forecast effluent quality in a full-scale treatment plant. The study asserted that deep learning model with backpropagation are the most effective ML algorithm for diverse time series forecasting tasks.

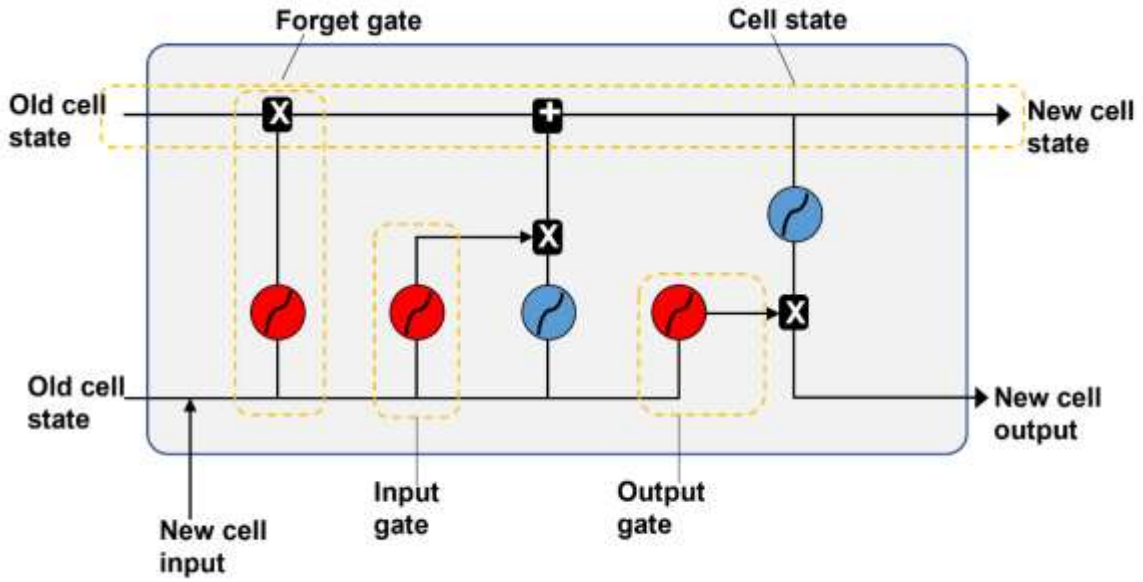

Figure 5. A Recurrent Neural Network Framework (RNN).

LSTM has been explored in the finance industry for stock prediction (Chen et al., 2022; Moghar & Hamiche, 2020; Teng et al., 2022; Yadav et al., 2019), waste management prediction (Cubillos, 2020), and energy industry (Feng et al., 2021; Li & Wang, 2020; Qiao et al., 2022). Also, LSTM models have been applied in the prediction of wastewater treatment effluent quality in some studies (Farhi et al., 2021; Pisa et al., 2020; Taloba, 2022), but the application of LSTM in the flocculation process of water and wastewater treatment is very limited. It is strongly believed that LSTM would be able to serve as an early warning tool for a treatment facility (Oliveira et al., 2020). While a robust model could be used to elucidate the complex flocculation problem that has resulted in multiple jar test analysis by water treatment facilities.

Considering the importance of Time gradient Tf in the flocculation process, which has been duly reported in past studies, coupled with the kinetic models that have shown the relationship of Tf to floc length evolution, it could be assumed that a robust time series model would be able to effectively learn (train) and predict the floc evolution process. Furthermore, the ability of LSTM (time series-based model) to analyze image-based data could be a groundbreaking success by directly analyzing data from non-intrusive floc image setup. Therefore, this study seeks to optimize the flocculation process of water and wastewater treatment using robust ML algorithms (LSTM and ANN), towards establishing an ML framework with the capability to serve as an early warning system for the

flocculation process of water treatment. In addition, ML models have the potential to improve the treatment process, resulting in a cost-effective, time-efficient, and sustainable water treatment.

3.4.1 Basic Neural Network Terminologies

Neuron: This is the building block of neural network and functions like the human brain cell, it receives information in form of vectors or scalars and processes them to generate an output. Neurons have the capability to act in series which makes the development of layers of neurons possible during computation. In an instance of many hidden layers of neurons, information is processed by each neuron and its layer and passed forward to the next layer until the desired accuracy is achieved. The interpretation of inputs to output in every neuron is guided by the weight and bias, giving the neurons direction and ability to recognize non-linearity.

Weight: This is a neural network parameter of scalar nature (values) generated during the learning process of a neuron. Weights are used to correct the learning process of each neuron to achieve satisfactory output and improve the model generalization capability. **Weight constraint** is a method of guiding/restricting the weight generated during the neuron's learning process. It is a regularization approach used to prevent model overfitting by determining the magnitude of the generated weight. Weight constraint also helps to avoid exploding gradient during the training process, causing the model to generalize more during the training process and improve prediction accuracy at the testing phase.

Bias: They are also scalar parameters added to the weighted sum of the neuron's input data during the learning process. They are used to control the model capability to generalize on the input data and make accurate predictions on an unseen dataset. Both weight and bias are controlled during the model training process.

Under fitting: This refers to a state where the model is unable to adequately learn from the input data, resulting in a poor generalization and prediction accuracy of the generated model.

Over fitting: This is a state where a model fails to generalize the input data by closely fitting the model on the dataset, thus causing a poor prediction accuracy on the

unseen data. This is a serious problem in machine learning model training and often limits/reduces the adoption potential of the model. An indicator of overfitting is an instance where a very high training accuracy is observed but test accuracy is very poor. This can be improved through steps such as regularization, early stopping, increase of training data, hyper-parameter tunning, and cross-validation.

Activation function: Neural networks learn complex patterns at the output of the model and regularizes through the help of a mathematical function known as “activation function”. The activation function introduces non-linearity to the output layer of the model so that the model could learn from complex processes. The choice of activation function is often determined by the kind of data (discrete or continuous data, time series, etc.) and the expected outcomes (images, timeseries, regression, etc.). Neural networks have different activation functions for different purposes; sigmoid, ReLu, SoftMax, Tanh, etc.

Epochs: This is the process by which the training dataset passes through the neural network. At each passage of the training data (epoch), the neural network computed the weight or each neuron, bias, and the prediction error (loss). The weight and bias are updated at each epoch during the training process. It is essential that the epochs determine the model fitting, too few epochs could cause underfitting while too many epochs could cause over fitting. The appropriate epochs for each neural network can be predetermined through the model hyperparameter tuning process.

Batch size: This is the number of training data processed per model epoch, and an essential hyperparameter of neural network models. The batch size determines how the training dataset is divided before entry through the neural layers. This updates the weight and bias of the model based on the prediction loss at every batch during the model training phase. The batch size of the model is determined by the size of the dataset, neural network structure, and the computing capacity of the training system.

Optimizers: This uses the gradient of the model loss function (e.g., mean squared error) for the model parameters to update the parameters in such a way that it reduces the prediction error. The parameters of machine learning models are controlled by the optimizers. There are different types of optimizers, such as Adaptive moment estimation (Adam), Stochastic Gradient Decent, etc. The choice of optimizer is also determined by the type of problem and model architecture.

Cross validation: This is a process by which the prediction accuracy of model on small dataset is evaluated. The data is first divided into training and validation set, and the training set is further partitioned into different parts. The model is trained on each dataset and validated, and the overall performance is evaluated to determine its generalization capabilities. Cross validation is a very good way of controlling model overfitting.

Hyperparameter tunning: Tunning is the process of adjusting the parameters of ML/NN to achieve the best configuration per each model architecture. The parameters of the models range from batch size, epochs, number of neurons and layers, activation functions, weight constraints etc., The hyperparameter tuning is used to determine the best sets of the parameters for the training of the model, such that the best prediction accuracy and generalization capability is achieved.

4 MATERIALS AND METHOD

4.1 Data Sources

Secondary data was collected from the UNESP Laboratory in Rio Claro Campus under the supervision of Professor Rodrigo Moruzzi. All data sets for this study are already available, as experiments have been performed and presented by Moruzzi & Reali (2014) and Moruzzi et al. (2017). The experimental data of a batch assay (laboratory jar test analysis) that was carried out using synthetic water was used for this study. The data collected are jar test gradient of velocity (G*f*), time gradient (T*f*), floc length, perimeter, aspect ratio, and area. The collected data were in two forms; raw image data (presented in supplementary data), and processed information in tabular representation. The images were used to verify the total number of images analyzed by the software. Description of the experiment conducted to generate the jar test data is described in detailed in the supplementary material after the conclusion section of this thesis. **The extracted flocs properties used for this study are floc length, gradient of velocity (G*f*), and Time gradient (T*f*).**

4.2 Data Treatment

The harvested processed experimental data from the images were compiled on an MS excel spreadsheet. The compiled data sets were filtered to organize the dataset according to G*f* values. The data per each G*f* has over 1,500,000 count of flocs, and outliers in each dataset were confirmed by comparing the floc length with areas of the same floc length corresponding to the time gradient (T*f*). The range of floc lengths were determined through the statistical summary (largest floc lengths; G*f* $20sec^{-1}$ = 3.5, G*f* $60sec^{-1}$ = 5.439). The selection of the value ranges was determined by the maximum floc lengths in the data. Afterward, floc lengths were grouped into “bin” (floc length interval)

following the principle of particle size distribution. An interval of 0.646 mm was selected to divide each dataset into five (5) groups (0.27–0.916, 0.916–1.562, 1.526–2.208, 2.208–2.854, and 2.854–3.5 mm). The choice of 0.646mm was based on the floc length range (0.27 – 3.5). Smaller floc length interval was selected to ensure the evolution of smaller floc length ranges are adequately captured. Notably, maximum floc length of 3.5 mm was chosen for G*f* 60 sec^{-1} because larger flocs were found in only two time-steps, thus the model would not be able to accurately predict the groups. The "count" of particle sizes within each class was carried out to create a new dataset. This was repeated for each G*f*s (as shown in Table 1 below).

Table 1: Sample spreadsheet of the grouped dataset for Gf 60 sec^{-1}.

Time (min)	GR1	GR2	GR3	GR4	GR5
2	58308	4144	827	297	86
5	46619	5771	537	62	8
10	38610	3299	150	5	0
15	40501	3548	165	9	0
20	45004	4356	251	17	0
25	43418	3163	130	9	0
30	43246	2873	99	1	0
35	43963	2658	69	2	0
40	48235	3394	150	17	5
45	50598	4022	158	8	1
50	49866	3442	138	9	0
⋮	⋮	⋮	⋮	⋮	⋮
180	51435	2642	92	4	0

NB: The new dataset is utilized for the methodology optimization (modeling) phase of the analysis.

4.3 Methodology Optimization (Selection and Modeling of ML Algorithm)

The following iterative steps (Statistical Experimental Design - SED) were used as a guide in the model selection, development, data training, and deployment.

i. Recognition of variables and levels.
ii. Selection of models - design of the run sheet via concurrent alteration of independent variables

iii. Analysis of the result and establishment of the selected models (e.g., time series analysis).

iv. Evaluation of model accuracy and curvature in the analysis using performance indicators (RMSE, R^2, and MAE) to identify model fitting and over-fitting.

v. Extension to deep learning model (when necessary).

vi. Model validation using test data sets (testing for model accuracy).

vii. Model prediction.

4.4 Modeling of particle (floc) size evolution using machine learning models.

The trends in flocs particle evolution have the ability to elucidate the flocs growth and breakage, and the importance of the time gradient at a given velocity gradient in a flocculation process. Since the flocculation process is time-dependent, a time series model is considered the most appropriate for understudying the flocs behavior and predicting the optimum time for the process, based on the particle evolution. In this study, three types of time series models were developed and compared to understand the extent of complexity in the computation of flocs evolution and select the most accurate model for the prediction. The selected time series models were the traditional Auto-Regressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), and Long-Short Term Memory (LSTM) models.

After the computation, each model's accuracy was evaluated by standard indicators in relation to the machine learning algorithms and statistical analysis. The selected indicators for the model evaluation were Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Residual Mean Squared Error (RMSE). The choice of R^2, MAE, MSE, and RMSE is their universality in statistical computation and applicability to this study (quantitative data), and to determine the global and local mean differences in the model prediction. Furthermore, the model Loss validation (Loss Curve) was assessed (plotted) to determine the convergence level of the deep learning models (ANN and LSTM). The superiority of the models was further evaluated with a regression plot (scatter plot) and R^2.

The mathematical expressions for R^2, MAE, MSE, and RMSE are given below.

$$R^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2} \quad (7)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (8)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (10)$$

where: y_i is the observed value of *ith* observation, $\hat{y}_i$ is the estimated value of the *ith* parameter, N is the number of observations, $\bar{y}$ is the mean of the observed variable.

4.4.1 Time Series Analysis

4.4.1.1 Auto Regressive Integrated Moving Average Model (ARIMA Model)

ARIMA model has been used in the forecasting of water and wastewater-related events (Elkiran et al., 2019; Safeer et al., 2022; Wen et al., 2022), particularly due to their effectiveness in handling simple and multi-step time series problems. To model the flocs evolution using the ARIMA model, the following iterative process was followed.

The datasets were first upscaled using the time series principle to increase the number of timesteps to an interval of 5 secs per time step. The time stamp was set to "Seconds" to increase the data time steps. The resample function creates a data point with the null value "NAN". The linear option of the interpolation function was used to create values with a linear and equal interval between each time step with values (original data points). The upscaled samples were exported for further analysis.

After data upscaling, the compatibility of the data with ARIMA model were evaluated by checking the dataset stationarity and seasonality. Generally, ARIMA model is designed for non-stationary and non-seasonal data, while the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is designed to handle seasonal data. Although an alternative and advanced form of the two models (SARIMAX; SARIMA with an exogenous variable) has the capability to model seasonal or non-seasonal and stationary

datasets. An Augmented Dickey-Fuller (ADF) test was carried out to test the stationarity of the data. The significance level was set at 0.05, with values above 0.05 ($p > 0.05$) indicating the data is not stationary.

After the ADF test, each group of the dataset was optimized with Auto-ARIMA algorithm derived from Hyndman-Khandakar (2008) to determine the best hyperparameter for the model computation. Generally, ARIMA models are in three tiers: Auto regressive (p), moving average (q), and the differences (d). The effectiveness of the hyperparameter tuning (auto arima) were cross validated by plotting the autocorrelation graph for the values of 'p' and partial autocorrelation 'q.'

Afterwards, data training and testing splitting ratio of 70: 30 were done with the cautiousness of the time step, to ensure continuity in time step. Afterward, the best (recommended) hyperparameter (p, d, and q) were used to train the ARIMA model and subsequently test the model accuracy. The model fitness and the summary were generated before the prediction of the test dataset was made. Finally, the training and testing data were predicted with the model and visualized. The model accuracy was evaluated using the R^2, MAE, MSE, and RMSE measures. Similar iterative procedure was repeated for each group of the datasets (G*f*).

4.4.1.2 Neural Network Model Development (ANN & LSTM)

Neural network model has been explored by researchers in the field of water and wastewater treatment, including flocculation studies (Ammar et al., 2021; Arab et al., 2022; Bagheri et al., 2019; Graça et al., 2022; Igwegbe et al., 2021; Khatri et al., 2020; Nasr et al., 2012). However, none of the studies considered particle size (flocs) evolution, particularly with a time series approach. Therefore, neural network models (ANN & LSTM) were selected for this study to model the floc length evolution following the SDE procedure highlighted above (section 3.3).

4.4.1.2.1 Model data preprocessing

Data preprocessing such as conversion into an array and creation of compatible time steps (sequence) were first carried out before the hyperparameter tunning. Neural network algorithms are designed with a default compatibility to a regression-based problem (dependent (Y) and independent (X) variables). Therefore, time step sequences were created by a function to generate the dependent and independent variables (sequence

= 5) used for the computation of the model (ANN & LSTM). The upscaled dataset was first partitioned into training and testing datasets (for each group) using the same ratio (70: 30) before they were normalized. The partitioning of dataset before normalizing were done to avoid data leakage (Gharib & Davies, 2021), thereby making the test data unseen to the model. Given that the datasets were in time series format (T*f* and numbers of flocs), they were converted into an array and normalized using the MinMaxScaler function. Notably, both training and testing data were normalized separately to avoid data leakage, and further reshaped to a 2-Dimensional array for compatibility with the ANN algorithm input format. For the LSTM, the data were reshaped to convert the data into a vector time series format (samples, timesteps, and features).

4.4.1.2.2 Hyperparameter tunning and cross-validation

The hyperparameters of neural network model are vital and major determinants of the model accuracy and fitness. Overfitting or underfitting can easily be avoided with adequate hyperparameter tunning (Gharib & Davies, 2021; Nielsen et al., 2020; Uddin et al., 2022). This is to maximize the resources embedded in the algorithm. For this study, hyperparameter tunning was performed in two phases for each bin of the dataset using the Grid search Cross-Validation (Grid SearchCV) algorithm. First, the model activation function, optimizer and neuron were tuned. Second, the model features (neurons, epoch, batch size, dropout rate, weight constraint, and validation split) were tuned at the second phase. Adam optimizer and RELU activation function were recommended for the model for obvious reasons such as the universality of Adams over other optimizers (AdaGrad, SGD, Adagrad, etc.).

The choice of ReLU activation function is due to its superiority to other activation functions such as sigmoid and tanh. Sigmoid has several problems which include vanishing and exploding gradient, slow convergence, and particularly because they are not zero-centered. Tanh activation function is good but is mainly compatible with models suitable for the range between -1 to 1 and suffers from vanishing gradient, unlike the ReLU activation function. MSE and MAE were used as the model loss and accuracy indicator. The summary of tuned hyperparameters for the models is presented in Table 2 below. The best configuration for each model was used to train the model, evaluate the accuracy before testing the model for prediction, and final evaluation of the prediction accuracy (R^2, MAE, and RMSE). The model training and validation loss curve were further plotted to understudy the model convergence and potential under/overfitting.

The framework of the developed model was designed for the adoption by subsequent studies and its usage in full scale water/wastewater treatment facilities. The ML framework is presented in Figure 6 below.

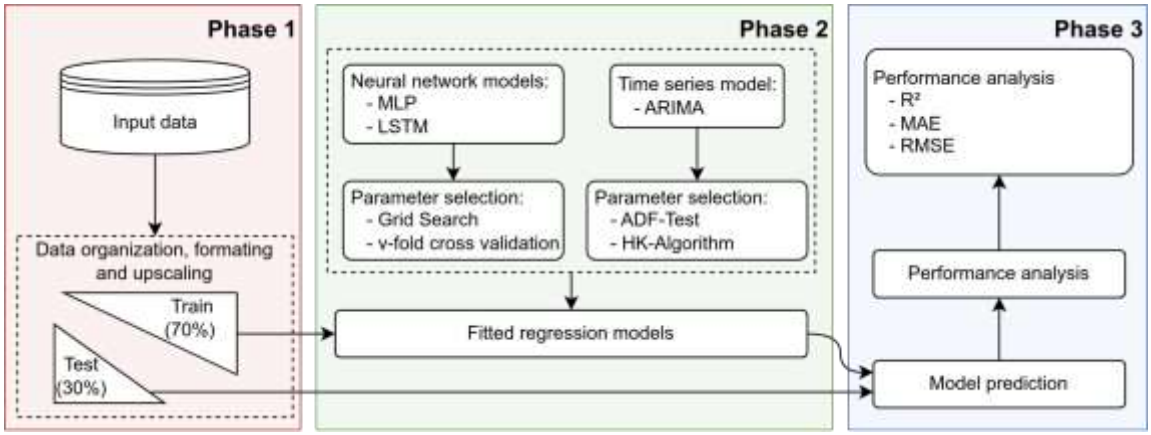

Figure 6: Machine Learning framework for modeling flocs evolution of flocculation process.

Table 2: Summary of deep learning models hyper-parameter tunning for best model predictor.

Dataset	Dataset	Number of neuron(s)	Number of hidden Layer(s)	Batch size	Epochs	Dropout rate	Validation Split	Weight constraint	Recurrent constraint
				ANN					
Gf 20 sec^{-1}	20_Bin 1	128: 64: 32: 8: 1	4	200	50	0.1	0.01	1	-
	20_ Bin 2	128: 64: 32: 8: 1	4	200	50	0.3	0.01	3	-
	20_Bin 3	128: 64: 32: 8: 1	4	200	30	0.3	0.01	1	-
	20_Bin 4	128: 64: 32: 8: 1	4	100	50	0.3	0.01	1	-
	20_Bin 5	32:16:8:1	3	100	10	0.01	0.01	1	-
Gf 60 sec^{-1}	60_Bin 1	128: 64: 32: 8: 1	4	200	10	0.1	0.01	1	-
	60_Bin 2	64:32:8:1	3	100	30	0.1	0.01	1	-
	60_Bin 3	10: 1	1	200	30	0.1	0.01	1	-
	60_Bin 4	10: 1	1	100	30	0.01	0.01	1	-
	60_Bin 5	10: 1	1	100	30	0.01	0.01	1	-
				LSTM Model					
Gf 20 sec^{-1}	20_Bin 1	128: 64: 32: 8: 1	4	200	50	0.3	0.01	2	2
	20_ Bin 2	128: 64: 32: 8: 1	4	200	50	0.1	0.01	3	3
	20_Bin 3	128: 64: 32: 8: 1	4	200	30	0.1	0.01	1	1
	20_Bin 4	128: 64: 32: 8: 1	4	100	50	0.1	0.01	1	1
	20_Bin 5	32: 16: 8: 1	3	100	10	0.01	0.01	1	1
Gf 60 sec^{-1}	60_Bin 1	128: 64: 32: 8: 1	4	200	20	0.1	0.01	1	1
	60_Bin 2	64: 32: 8: 1	3	100	30	0.1	0.01	1	1
	60_Bin 3	64: 32: 8: 1	3	200	50	0.01	0.01	1	1
	60_Bin 4	64: 32: 8: 1	3	200	50	0.01	0.01	1	1
	60_Bin 5	64: 32: 8: 1	3	200	50	0.01	0.01	1	1

5 ADDITIONAL INFORMATION ON EXPERIMENTAL SETUP AND IMAGE PROCESSING

5.1.1 Experimental Setup

This section provide additional information on the experimental setup used for the jar test batch assay and the image processing phase for the extraction of floc length data used for this research, as performed and presented by Moruzzi & Reali (2014) and Moruzzi et al. (2017).

The jar test analysis was carried out using Aluminum sulfate ($Al_2(SO_4)_3.14H_20$; later denoted as Al^{3+}) as the coagulant. Kaolin stock suspension with high turbidity (5000 ± 200 NTU) was prepared by introducing Forty grams of Kaolin powder into a 1L deionized water. The mixture was performed using magnetic stirrer for 2hrs at velocity of 1000 sec^{-1}, the settling time of 12hours was observed before 850 mL of the supernatant was removed as the stock solution. 2ml of the stock solution was added to deionized water to make a sampled synthetic water with turbidity of 25 ± 2NTU. 2mg Al^{3+} coagulant dosage was used at a constant rate for the experiment, rapid mix was carried out at 800 sec^{-1} shear velocity for 10 seconds using a PoliControl jar test equipment. The pH at the end of coagulation was 7.5, the particle size distribution of the Kaolin used for the experiment is presented in Figure 7b below.

The flocculation speed (*Gf*) were 20 sec^{-1} and 60 sec^{-1} respectively. Each flocculation shear velocity was maintained for 3 hours on jar test (batch assay) and images were taken initially at 1-minute intervals (2 min, 3 min, …,9min) and increased to 5mins intervals after the first 10mins of observation (i.e., 10min, 15min, 20min…180min). All flocculation stages (G*f*) used the same coagulant dosage and pH. Hybridized non-intrusive dynamic image analysis was coupled to a 2-liter jar test setup (Moruzzi et al., 2017) as shown in Figure 7a below.

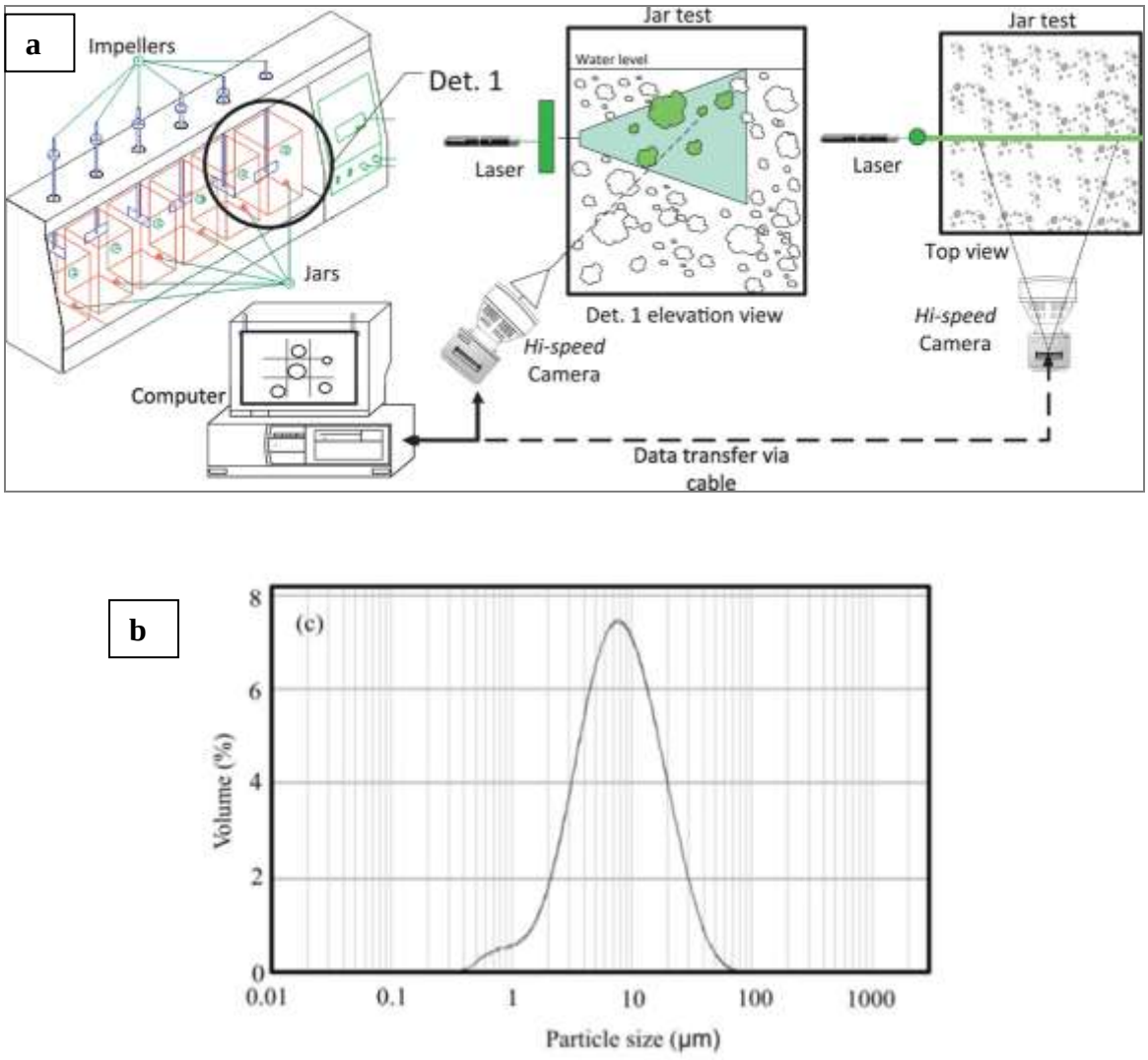

Figure 7: Research (a) experimental setup and (b) kaolin particle size distribution, after).

A camera shutter speed was set to 800 μs, and images were captured at the frequency of 10 Hz during 10 seconds per observation, resulting in 100 images per observation. A laser sheet of 532 nm was used for illumination, with the aim of enhancing the contrast for the flocs on the camera focus. The image resolution resulted in a pixel size of 270 μm. Further details can be seen in the above-mentioned references.

5.1.2 Floc Image Processing

Flocs images captured at the resolution of 840 by 640 pixels were further processed by the Image Pro-Plus® software to transform from the original 8-bit grayscale representation (Figure 8) to 1-bit binary form (i.e., black and white), and performed the enhancement and segmentation for subsequent measurement of traits such as floc area, perimeter, and length. Errors such as underestimation and overestimation in the

measurement of floc area were addressed by using areas produced by the floc outlines and the area of the elementary pixel of an image, as fully described in Moruzzi et al. (2018). Furthermore, the extracted flocs data were subjected to statistical analysis at 95% level of significance to ensure the integrity of the data. Detailed information on experimental and image process analysis were documented in Moruzzi & Reali, (2010); Moruzzi et al., (2017, 2018, 2020, 2020).

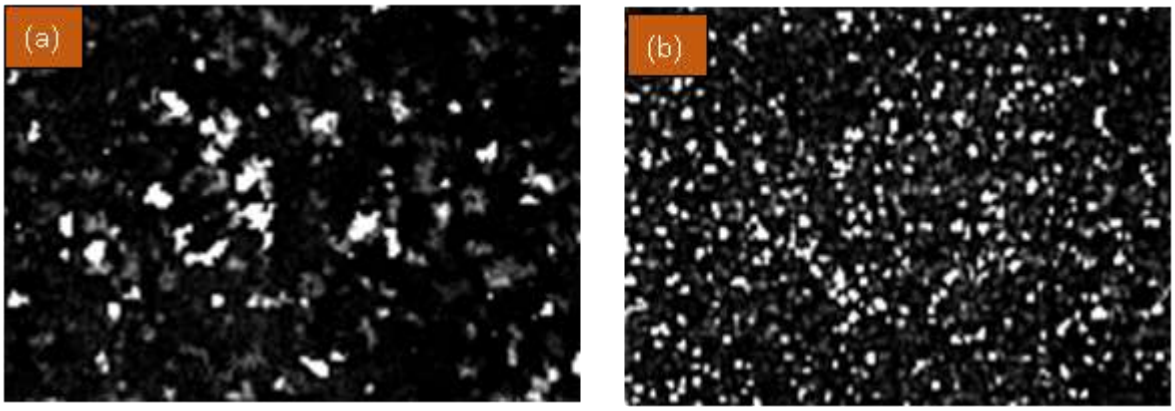

Figure 8: Floc image at given flocculation time during jar test experiment (10 minutes) (a) $Gf = 20\ sec^{-1}$, (b) $Gf = 60\ sec^{-1}$

6 RESULTS AND DISCUSSION

6.1 Floc Lengths Evolution

Adequate knowledge of floc particle size evolution is essential in the determination of flocculation effectiveness in water and wastewater treatment. This section presents the results for the ARIMA, ANN, and LSTM models trained and tested with the G*f* = 20 sec^{-1} and 60 sec^{-1} datasets, and the generalization effectiveness of the models in predicting the particle size evolution during the flocculation process. The trend in floc length evolution across different groups and G*f* is explored to understand the impact of G*f* variation on flocs evolution.

6.1.1 First floc length group (0.27- 0.916 mm)

The accuracy of the ARIMA, ANN, and LSTM models and their effectiveness in predicting the floc length evolution for the first group (floc length = 0.27- 0.916 mm) is presented in this section. The AD Fuller test result for the dataset stationarity has ADF test statistics of -6.1697 and p-value of 6.85E-08. P-value less than 0.05 implies that the dataset is stationary, which proves that the dataset satisfies the stationarity criteria for an Arima model. The output of the auto Arima algorithm optimization is presented in Figure 9 below. The ARIMA hyperparameters (number of autoregression; p, number of differences; d, and number of moving averages; q) recommended were p = 1, d = 1, and q = 1 for autoregression.

The auto Arima model recommended the SARIMAX model because the model (SARIMAX) has the capability to handle data with exogenous variables. The Arima model was used to compute the final model since the dataset compatibility has been established. Table 3 presents the summary of the prediction evaluators (R^2, MAE, and RMSE) for the ARIMA, ANN, and LSTM models in predicting the particle floc length evolution during a

flocculation process. The ARIMA model recorded a very poor test prediction accuracy (R^2 = -0.36) but an excellent training accuracy R^2 = 0.97. This led to a very high-test data prediction error (MAE = 2222.61 and RMSE = 2622.56). The poor prediction accuracy is obvious by the model's inability to follow the trend variations in the particle size evolution, as shown in Figure 10a. This shows that ARIMA model is unable to learn from the provided dataset and understand the underlying relation between the floc length evolution and gradient velocity, and time gradient, during a flocculation process of water treatment. The failure to predict the testing dataset shows that ARIMA algorithm cannot handle the underlying complex phenomenon in the flocculation process, thus the high accuracy recorded at the training phase could be attributed to the fact that it was trained with the data, so the model is already exposed to the dataset.

SARIMAX Results

Dep. Variable:	y	No. Observations:	2161
Model:	SARIMAX(1, 1, 1)	Log Likelihood	-8065.096
Date:	Sat, 03 Sep 2022	AIC	16138.192
Time:	03:36:47	BIC	16160.904
Sample:	01-01-2022	HQIC	16146.499
	- 01-01-2022	Covariance Type:	opg

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.5678	0.259	-2.195	0.028	-1.075	-0.061
ar.L1	0.9931	0.003	317.632	0.000	0.987	0.999
ma.L1	0.0007	0.298	0.003	0.998	-0.583	0.585
sigma2	98.9118	0.441	224.125	0.000	98.047	99.777

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	39266011.15
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.50	Skew:	19.16
Prob(H) (two-sided):	0.00	Kurtosis:	662.41

Figure 9: Auto Arima model output for ARIMA model hyperparameter tuning and identification of p, d, q.

Also, the assumption of an overfitting could not be assumed in this scenario because the predicted numbers of flocs totally differ from the observed numbers of particles, coupled with the fact that the prediction could not follow any of the changes in the evolution trend. Therefore, it could be assumed that ARIMA model is not a suitable model for computing/modeling flocs evolution phenomenon of the flocculation process of water and wastewater treatment.

Table 3: Summary of models (ARIMA, ANN, & LSTM) evaluation for floc length evolution during the flocculation process.

		Gf 20 sec^{-1}						Gf 60 sec^{-1}					
Model	Dataset	R^2 Train	R^2 Test	MAE Train	MAE Test	RsMSE Train	RMSE Test	R^2 Train	R^2 Test	MAE Train	MAE Test	RMSE Train	RMSE Test
ARIMA	Bin 1	0.97	-0.36	35.94	2222.61	1339	2622.56	0.88	-12.54	39.69	3495.8	1499.56	3830.77
	Bin 2	1	-0.98	0.33	322.1	3.96	460.71	0.96	-1.96	3.25	350.88	106.61	411.12
	Bin 3	1	-4.07	0.03	78.57	0.22	94.7	0.97	-6.57	0.58	93.8	21.27	101.54
	Bin 4	1	-2.23	0	2.68	0.04	3.33	0.96	-11.08	0.20	9.98	7.64	11.73
	Bin 5	1	-4.93	0	0.28	0.01	0.3	0.95	-0.16	0.06	0.09	2.21	0.25
ANN model	Bin 1	0.99	0.93	281.43	365.81	783.76	612.85	1	0.99	119.76	105.19	192.7	125.92
	Bin 2	0.86	0.92	92.8	70.8	108.17	90.85	0.96	0.94	86.59	46.24	100.96	59.41
	Bin 3	0.98	0.98	9.89	3.58	13.52	5.16	1	0.95	3.32	1.93	5.09	2.65
	Bin 4	0.96	0.92	2.07	0.4	3.71	0.5	1	0.84	0.41	0.36	0.85	0.45
	Bin 5	1	0.99	0.06	0	0.13	0.01	1	0.99	0.08	0.01	0.33	0.02
LSTM model	Bin 1	1	0.98	60.1	185.7	104.28	308.89	1	0.98	113.38	118.68	180.25	143.39
	Bin 2	0.99	0.99	19.99	29.34	30.35	36.54	0.99	0.94	32.07	43.56	43.53	56.94
	Bin 3	1	0.99	2.46	3.25	2.51	4.28	1	0.83	2.59	4.4	3.61	4.97
	Bin 4	1	0.95	0.37	0.32	0.61	0.42	1	0.95	0.39	0.24	0.72	0.25
	Bin 5	1	1	0.08	0.01	0.14	0.01	1	0.98	0.06	0.01	0.19	0.03

Meanwhile, both ANN and LSTM models recorded a better performance in flocs evolution modeling compared to the ARIMA model.

The ANN has training and testing prediction accuracy (R^2) of 0.99 and 0.93. The prediction errors were MAE of 281.43 and 365.81 for training and testing, and RMSE of 783.76 and 612.85 for train and test errors, respectively. The high accuracy of the ANN model compared to the ARIMA model is expected considering that neural network is more robust and can understand complex relationship within data variables. Also, the prediction accuracy proves that machine learning algorithms stand a high chance in learning and predicting flocs evolution process. The high-test prediction accuracy (93%) is also an indication that the model did not demonstrate any overfitting. This is justified by the nearly perfect overlay of the predicted numbers of flocs across the flocculation process as shown in Figure 10b. Although a minor overestimation is observed at 90 mins and underprediction at 175 mins, other precise estimations justify that the model has lots of prospect in flocculation process modeling, particularly, flocs evolution.

Similarly, the LSTM model recorded training accuracy and errors (R^2 = 1.0, MAE = 60.1, and RMSE =104.28) and testing accuracy and errors (R^2 = 0.98, MAE = 185.7, and RMSE = 308.89), respectively. The high prediction accuracy (almost 100%) established that the LSTM model is a robust time series compatible neural network model and has an edge over the ANN model. This is obvious through the perfect prediction of the number of flocs at all T*f* (as shown in Figure 10c above). Also, the prediction error of the LSTM model is almost doubled in the ANN model prediction error (as shown in Table 3).

The observed pattern and stability of the LSTM and ANN model is expected because the two deep learning models are more robust compared to the ARIMA model. The performance (R^2, MAE, and RMSE) of the ANN and LSTM model demonstrated in this study a high level of generalization (model hyperparameters could fit almost different variant of G*f* dataset for the same particle size range). The ANN accuracy corroborates the findings of Oliveira et al. (2018) on the modeling of fractal dimensions using ANN. The study achieved 99 percent prediction accuracy, which is almost what is achieved in this study, with R^2 = 0.93 – 0.99. Notably, the limited research works that applied ARIMA model in water treatment studies could probably be traced to their inability to adequately handle complex scenarios with non-linear relationships such as the flocculation process.

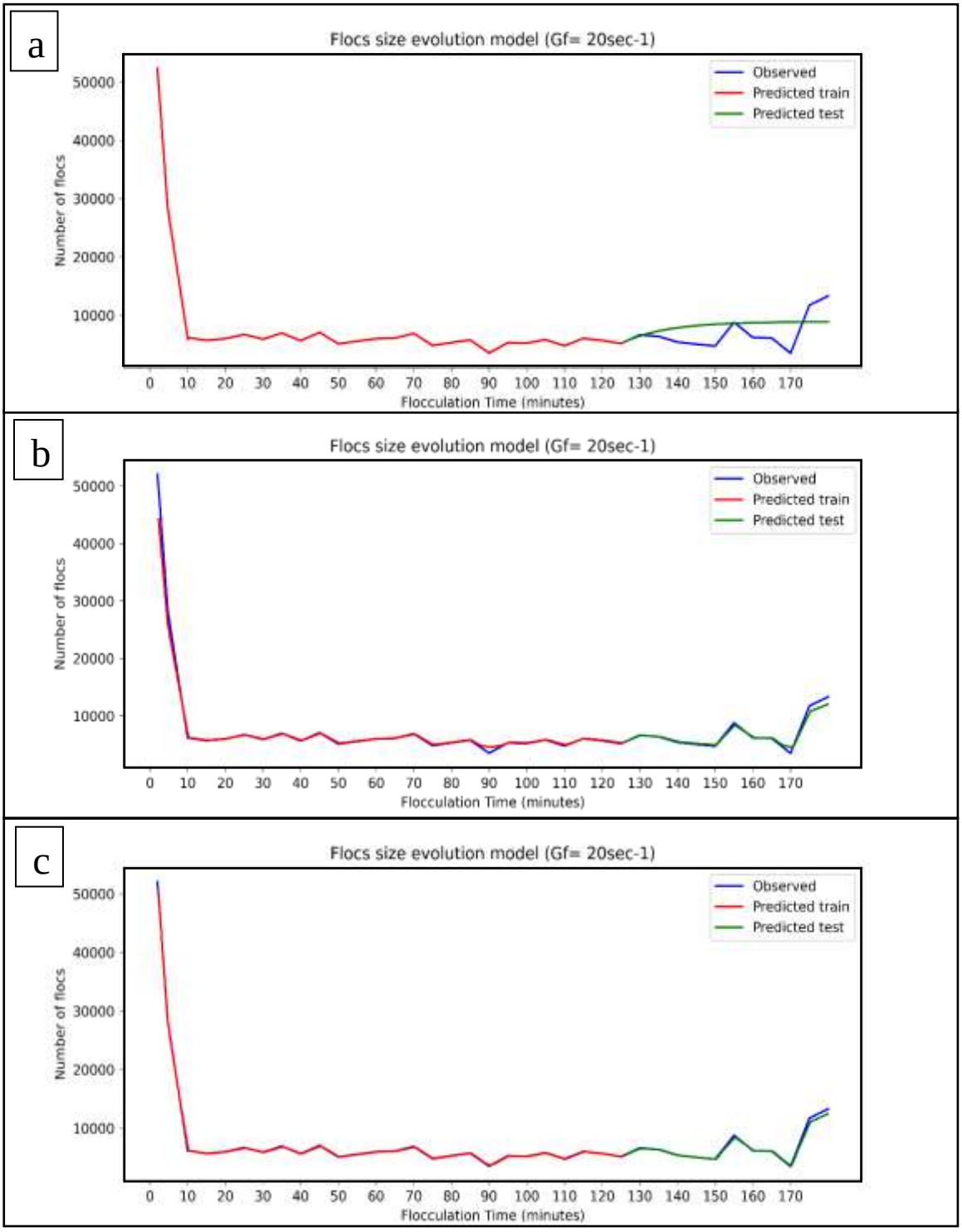

Figure 10. Comparison of different model prediction and observed number of flocs within the first floc length group under G*f* 20 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

The ANN has training and testing prediction accuracy (R^2) of 0.99 and 0.93. The prediction errors were MAE of 281.43 and 365.81 for training and testing, and RMSE of 783.76 and 612.85 for train and test errors, respectively. The high accuracy of the ANN

model compared to the ARIMA model is expected considering that neural network is more robust and can understand complex relationship within data variables. Also, the prediction accuracy proves that machine learning algorithms stand a high chance in learning and predicting flocs evolution process. The high-test prediction accuracy (93%) is also an indication that the model did not demonstrate any overfitting. This is justified by the nearly perfect overlay of the predicted numbers of flocs across the flocculation process as shown in Figure 10b. Although a minor overestimation is observed at 90 mins and underprediction at 175 mins, other precise estimations justify that the model has lots of prospect in flocculation process modeling, particularly, flocs evolution.

Similarly, the LSTM model recorded training accuracy and errors (R^2 = 1.0, MAE = 60.1, and RMSE =104.28) and testing accuracy and errors (R^2 = 0.98, MAE = 185.7, and RMSE = 308.89), respectively. The high prediction accuracy (almost 100%) established that the LSTM model is a robust time series compatible neural network model and has an edge over the ANN model. This is obvious through the perfect prediction of the number of flocs at all T*f* (as shown in Figure 10c above). Also, the prediction error of the LSTM model is almost doubled in the ANN model prediction error (as shown in Table 3).

The observed pattern and stability of the LSTM and ANN model is expected because the two deep learning models are more robust compared to the ARIMA model. The performance (R^2, MAE, and RMSE) of the ANN and LSTM model demonstrated in this study a high level of generalization (model hyperparameters could fit almost different variant of G*f* dataset for the same particle size range). The ANN accuracy corroborates the findings of Oliveira et al. (2018) on the modeling of fractal dimensions using ANN. The study achieved 99 percent prediction accuracy, which is almost what is achieved in this study, with R^2 = 0.93 – 0.99. Notably, the limited research works that applied ARIMA model in water treatment studies could probably be traced to their inability to adequately handle complex scenarios with non-linear relationships such as the flocculation process.

The smallest floc length which constitutes the major part of the first group of floc lengths (floc length range of 0.27 – 0.916mm) can be regarded as the primary particle present in the raw water, since the camera lens could only detect particles above the lens size (Moruzzi et al., 2017, 2018). Therefore, the decrease in their evolution (total count) at few minutes after the start of the flocculation process is expected since bigger floc lengths are expected to be formed. The long stretch of gentle alternation in the upward and

downward trend of the floc's growth could be considered as the particle restructuring phase. The particle restructuring phase consists of interaction between particle-particle and particle-cluster, which could easily be favored by the low velocity gradient *Gf* of the flocculation phase.

This is in tandem with the assertion of Moruzzi et al. (2017) that lower gradient velocity (shear rate) favors the particle cluster formation from an hyperdispersed primary particles of kaolin coagulated with aluminum (Al^{3+}). The long trend is considered possible as suspended primary particles in the water could loosely attach to aggregated flocs (clustered flocs), which could further lead to particle-cluster restructuring during the experiment. This illustration was adequately discussed by past studies on fractal aggregates (Moruzzi, Bridgeman, et al., 2020; Moruzzi, Campos, et al., 2020; Moruzzi et al., 2017; Wang et al., 2011; Yu et al., 2022; H. Zhang et al., 2019). Therefore, it could be asserted that the floc aggregation is reflected through the floc length evolution modeling with ML algorithms (ANN and LSTM).

The floc length evolution trend (though unsteady) which later rises towards the end of the experiment (T*f* at 150mins to 170mins) is also an indicator of particle breakage and re-aggregation. Therefore, this study holds lots of prospects in modeling floc length evolution and forecasting the flocculation process. It is worthy to note that a unique advantage of this time series approach is the ability to upscale or downscale the datasets, to achieve desired timesteps that could enable full scale treatment facilities to make predictions with limited data.

The result of the ARIMA, ANN, and LSTM models for the floc evolution trend of primary particles (floc length = 0.27 – 0.916 mm) under G*f* 60 sec^{-1} is presented in Figure 11a, b & c below. The DF test result shows that the first group dataset is stationary with p-value = 0.0097 ($p < 0.05$) and DF statistics = -3.438. The DF test for other groups and G*f* is presented in the supplementary document. Also, the auto Arima algorithm recommended a SARIMAX model with p, d, & d of 1, 1, 0, respectively (as presented in the supplementary material). The output of the model also follows the poor prediction accuracy presented above, with training accuracy and error (R^2 = 0.88, MAE = 39.69, and RMSE = 1499.56) and testing prediction accuracy and error (R^2 = -12.54, MAE = 3495.80, and RMSE = 3830.77), respectively. Such poor and negative testing accuracy was never expected at first, considering the partitioning ratio (70:30) used in the model development. However,

the complex interactions that produce flocs evolution could be understood as the reason for the poor prediction performance (as shown in Figure 11a below).

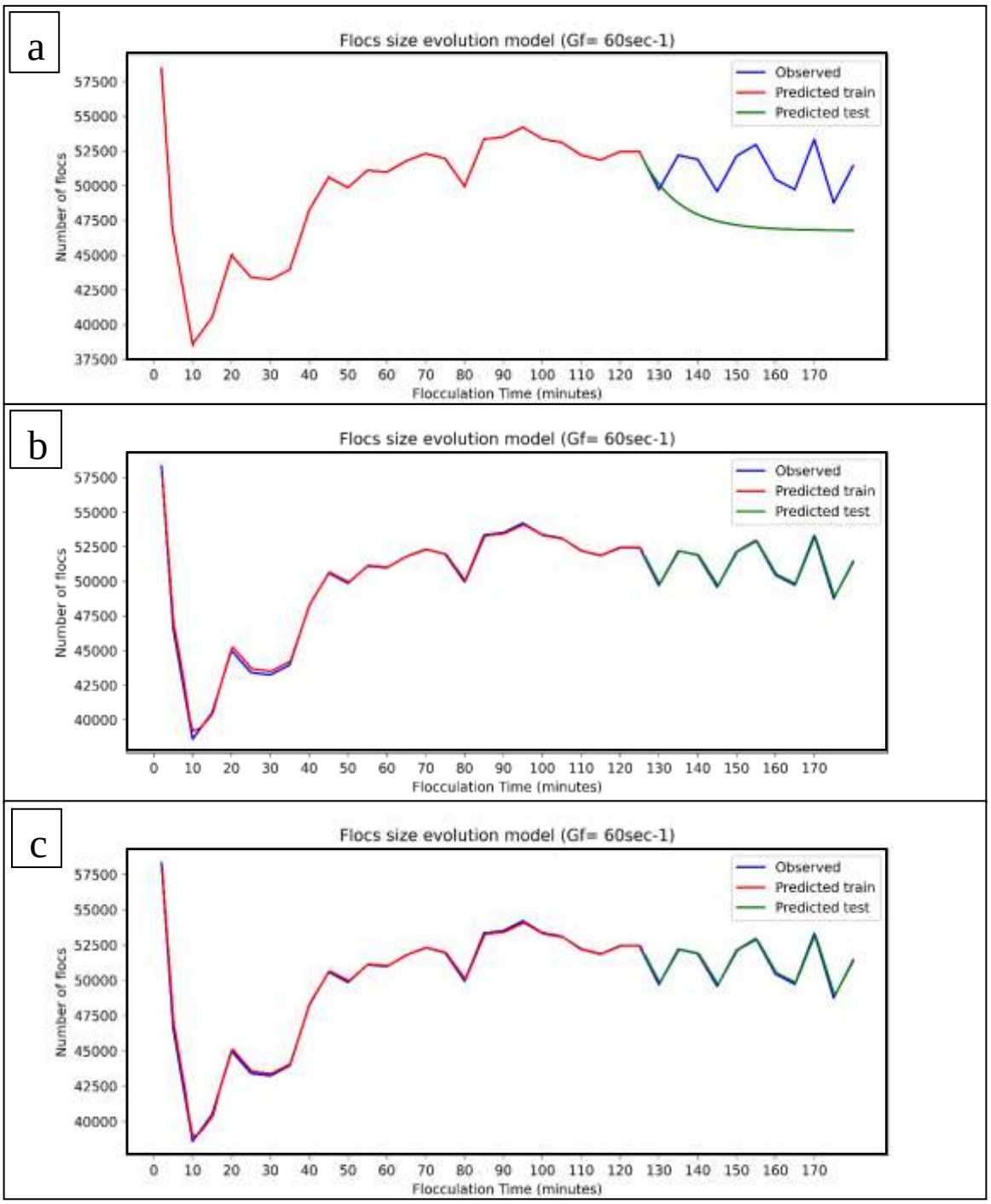

Figure 11: Comparison of different model prediction and observed number of flocs within the first floc length group under G*f* 60 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

High accuracy and generalization potential of the ANN model in predicting particle size evolution is shown by the training and testing R^2 values of 1.0 and 0.99, respectively. The prediction errors for training (MAE = 119.76 and RMSE = 192.7) and testing (MAE =

105.19, and RMSE = 125.92) also show a close range between the training and testing predictions, further proving the model robustness and ability to achieve high prediction accuracy without overfitting. In addition, the accuracy is confirmed by the perfect overlay of the predicted training and testing lines on the observed numbers of flocs in the first group of G*f* 60 sec^{-1} (as shown in Figure 11b below).

The LSTM model recorded training and testing prediction accuracy of R^2 = 1.00 and 0.98, respectively. The MAE and RMSE of the model were 0.113.38 and 180.25 for training, and 118.68 and 143.39, for the test prediction (as shown in Table 3 above). The LSTM algorithm demonstrated a robust and extremely high accuracy and generalization potential. This is visible in the prediction chart of the floc's evolution presented in Figure 11c, below. Furthermore, the precise overlay of the predicted numbers of floc lengths at the training and testing phases, which overlays perfectly than the ANN predicted numbers of flocs. Although, prediction accuracy of the ANN and LSTM models are the same (100%), a slight gap in the predicted values by the ANN model is visible between 10 – 35 mins, whereas the LSTM predicted values perfectly fit the whole trend.

The sharp drop in the number of primary particles between 2 mins and 10 mins shows a typical trend of floc aggregation, which is expected (Marques & Filho, 2022). Meanwhile, the sharp upward trend between 10 mins and 20 mins shows that about 25,000 primary particles have evolved through the breakage process, which further increased to achieve a peak breakage point at 95 mins before the unstable aggregation and breakage trend (125 – 180 mins). The high breakage rate experienced shows the impact of high velocity gradient (*Gf* 60sec^{-1}) on the floc length evolution, compared to the long steady particle restructuring phase identified in the G*f* 20 sec^{-1} (as shown in Figure 10). This is in accordance with the assertion of Bratby (2016); Filho et al. (2000) and Moruzzi et al. (2018) on the influence of high velocity gradient on floc length aggregation and breakage.

6.1.2 ***Second floc length group (0.916 – 1.562 mm)***

The evolution of floc lengths within the second group (0.916 – 1.562 mm) of the *Gf* 20 sec^{-1} in comparison with the prediction of the ARIMA, ANN, and LSTM models are presented in Figures 12a, b, and c below. The DF stationarity test shows that the second group dataset are stationary. The reason for testing each group of the dataset is to confirm if there are particle size range that has non-stationary data and ensures that model response

is not misinterpreted. P-value = 8.516 E-07 and DF statistics = -5.68, respectively. The auto Arima algorithm also suggested 1,1,0 for the ARIMA hyperparameter (p, d, & q). Meanwhile, the ARIMA model output demonstrates poor performance like the previous output. The model train prediction accuracy and errors (R^2 = 1.0, MAE = 0.33, and RMSE = 3.96) were recorded and testing accuracy and error (R^2 = -0.98, MAE = 322.1, and RMSE = 460.71) were recorded. This can also be considered as proof of the inability of ARIMA algorithm to learn from the floc evolution process (as shown in Figure 12a).

The ANN algorithm displayed a very good level of accuracy, although with a good training prediction accuracy. The model achieved an average accuracy of 90% for the training and testing floc length predictions (R^2 = 0.86 for training and R^2 = 0.92 for testing prediction). There were underpredictions at some parts of the training dataset (5 – 85 mins) but the accuracy improved at the testing phase (as shown in Figure 12b below).

This is evident in the R^2 values, considering that the testing phase recorded higher prediction accuracy than the training phase of the model, which is also reflected in the floc evolution chart (Figure 12b). This further proved that the model is capable of adequately learning from the floc evolution process and improve its prediction accuracy as the experiment progresses, though an underprediction is observed at 175 mins.

Figure 12c (shown above) presented the LSTM model prediction of number of flocs within the second group of the G*f* 20 sec^{-1}. Both training and testing prediction accuracy of the model were 0.99, with training and testing MAE and RMSE of 19.99 and 30.35, and 29.34 and 36.54, respectively. The model demonstrates the superiority of the algorithm and the relevance of hyperparameters such as the recurrent rate that helps the algorithm to keep the memory of past trend in the learning process, which is lacking in the ANN model. The prediction accuracy is similarly evident in the accurate overlay of the predicted number of flocs over the observed number of flocs.

The number of flocs rapidly increased between 2 mins to the peak (2,500) at 5 mins before a sharp depression to below 1,500 at 10 minutes. This sudden rise in the number of flocs within the second group could be traced to the transitional phase of flocs aggregate formation as established by Moruzzi et al. (2018). The authors also stated that the dramatic phase of aggregate transition phase occurs at the early stage of the flocculation process (often $\leq$ 5 mins), which is similar to the trend observed in Figures 12a, b, and c above. The

gentle fluctuation in the number of flocs between 10 mins and 120 mins could be traced to either an increase in the number of larger flocs aggregate or fragmentation of larger flocs.

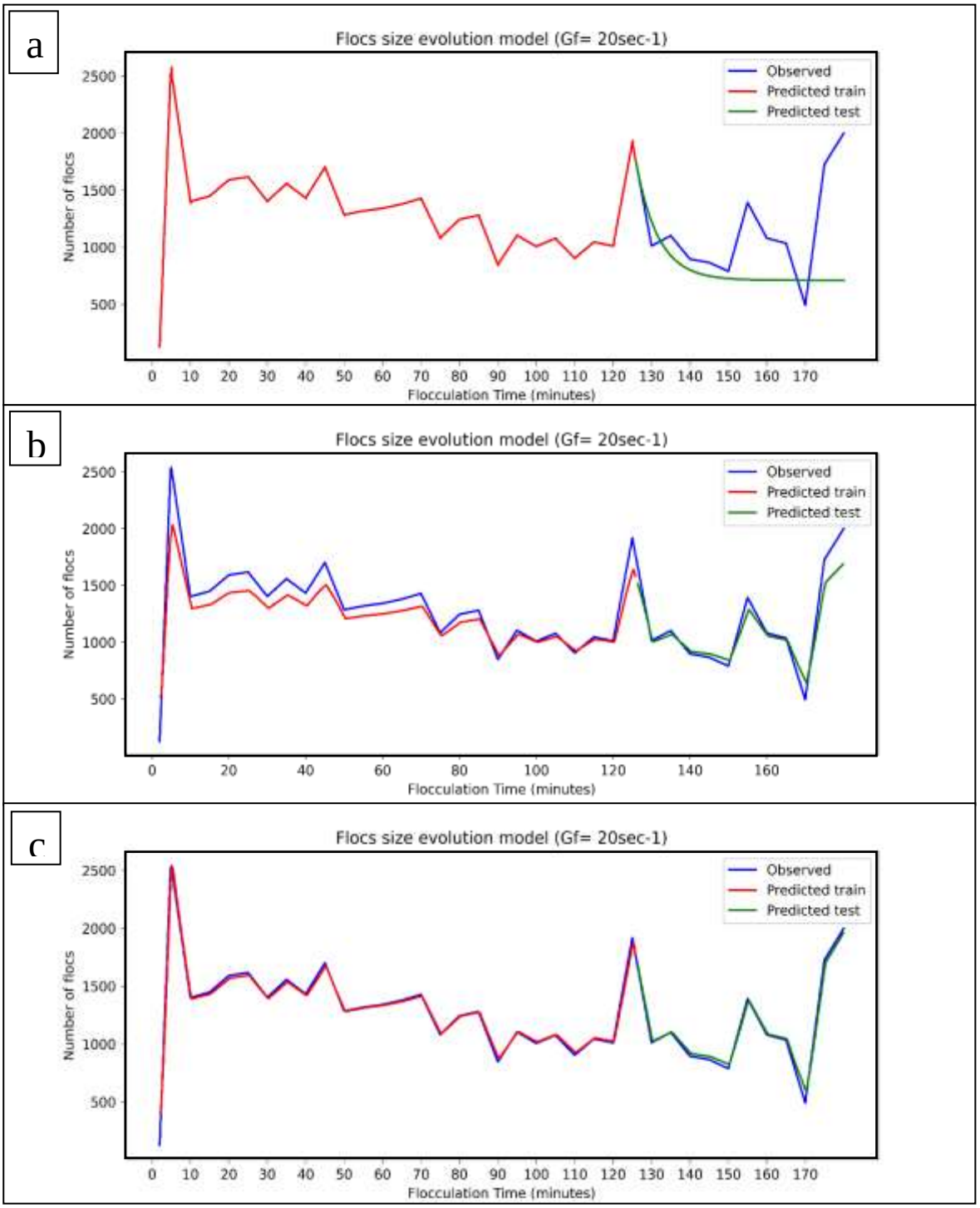

Figure 12: Comparison of different model prediction and observed number of flocs within the second floc length group under G*f* 20 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

This is considered in line with the pattern observed in the evolution of primary particles (as shown in Figure 10 above). The second peak in number of flocs at 125 mins shows that the restructuring phase has produced more clustered flocs within the second group, but these clustered flocs could be considered as loosely packed because the number of flocs immediately decreased at 130 mins to 150 mins, with the lowest count at 170 mins. This assertion is in correlation with the study conducted by Li et al. (2007) and Moruzzi et al. (2019) that identified flocs formed under low gradient velocity to have lesser resistance to floc breakage and weak floc strength index (SF). Although floc strength index was not determined in this study, findings have proven that resistance and strength of flocs under high gradient velocity are higher since loosely bonded flocs are easily fragmented due to the high induced shear rate. This was also evident in the study of Moruzzi et al. (2019) that found Kaolin flocs formed from Al^{3+} coagulant under G*f* 20 sec^{-1} to have 33.33% SF, compared to the 58.00% and 85.23% recorded under G*f* 60 sec^{-1} and 120 sec^{-1}, respectively. Therefore, it can be concluded that weakly bonded flocs could be formed and maintained under a low shear rate (G*f*).

The G*f* 60 sec^{-1} model prediction and its comparison with the observed numbers of flocs is presented in Figure 13a, b, and c below. The dataset is stationary with p-value = 0.00011 and DF test statistics = -4.645 (as presented in supplementary material). Similarly, the ARIMA model recorded high prediction accuracy and error (R^2 = 0.96 and -1.96, MAE = 3.25 and 350.88, and RMSE = 106.61 and 411.12) for training and testing phases, respectively. This also corroborates previous results of the ARIMA model on both the first group and second group of *Gf* 20sec^{-1}. The poor prediction accuracy is also reflected in Figure 14a that compares the predicted numbers of flocs against the observed numbers.

The ANN algorithm prediction accuracy of floc from G*f* 60 sec^{-1} is presented in Figure 13b below. The ANN model accuracy (R^2 and RMSE) was 0.96 and 100.96 for training phase, and 0.94 and 59.41 for testing phase, respectively. Although minor underprediction is observed during the training phase of the model (as shown in Figure 13b), the accuracy later improved with a very precise test prediction that captures all changes in floc evolution pattern. The improved testing accuracy is obvious through the reduced residual mean squared error (RMSE test is half of RMSE training). This is an indication of the high generalization potential of the model and the capability of the ANN algorithm to learn the floc evolution process and make accurate prediction.

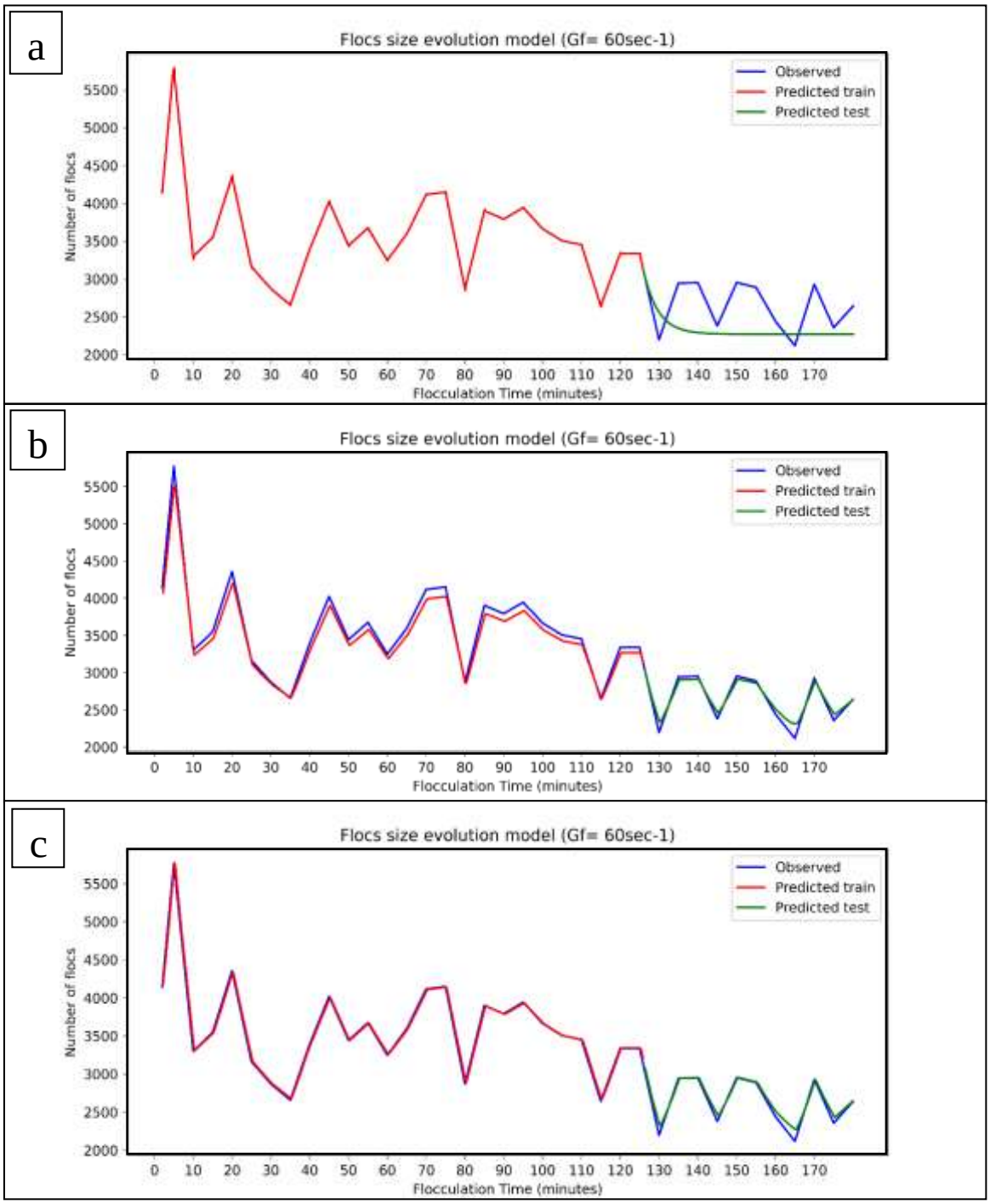

Figure 13: Comparison of different model prediction and observed number of flocs within the second floc length group under G*f* 60 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

The LSTM algorithm demonstrated similar accuracy with almost perfect training prediction accuracy (R^2 = 0.99) and very impressive testing prediction of the number of second group flocs (R^2 = 0.94). The superiority of the algorithm against the ANN model is demonstrated through the perfect fit of the predicted numbers of flocs as shown in Figure

13c above. This is also evident in the MAE and RMSE values of the model (32.07 and 43.53 for training, and 43.56 and 56.94 for testing stage). Although a minor underprediction is observed at the testing phase and is expected considering the gap in R^2 for train and testing phases. The model is considered not to overfit as the margin is very small. Besides, the model perfectly predicts all the sharp changes in the trend of flocs evolution. Therefore, the LSTM model demonstrates a good generalization capability, and can adequately predict particles/flocs within 0.916 – 1.562 mm without any limit to parameter features such as G*f*.

The rapid peak achieved at 5 mins of the evolution in addition with the number of particles at T*f* = 2 mins also justify the position of Moruzzi et al. (2018) that found flocs aggregate at higher G*f* to achieve their peak of the transition phase faster than the lower shear rates (G*f*). Generally, a down-trend is observed across the entire flocculation time (T*f*), though with a consistent up and down drift. This is considered as the steady phase of the aggregate formation process and possibly the influence of high G*f* on floc evolution, against what is observed in G*f* 20 sec^{-1}, particularly since weakly parked flocs may not be able to form for longer time due to fragmentation and erosion of flocs (small water eddies caused by shear stress). Also, the formation of larger floc aggregate could cause a steady downtrend in floc length evolution.

6.1.3 Third floc length group (1.562 – 2.208 mm)

The evolution of floc lengths within the third group (1.562 – 2.208 mm) of G*f* 20sec^{-1} predicted by the ARIMA, ANN, and LSTM models are presented in Figures 14a, b, and c, respectively. The third group of the G*f* 20 sec^{-1} is non-stationary with p value = 0.07 and DF test statistic = -2.691. The ARIMA model also recorded low testing accuracy with similar patterns (high training accuracy and negative test accuracy) as first and second groups. The ARIMA model's poor prediction is also obvious through its inability to recognize changes in floc evolution pattern as seen in Figure 14a.

The ANN model recorded prediction accuracy (R^2) of 0.98 for both testing and training dataset, with training and testing MAE and RMSE of 9.89 and 13.52, and 3.58 and 5.16, respectively. Similar pattern of accurate prediction at the beginning of the training set with some level of under prediction after 15mins is observed. The accuracy improves at the later instance of the training process and achieved a nearly perfect prediction towards the end of the training phase. This further improved during the testing dataset prediction with a

perfect overlay of the predicted and actual trend of flocs evolution (as shown in Figure 14b).

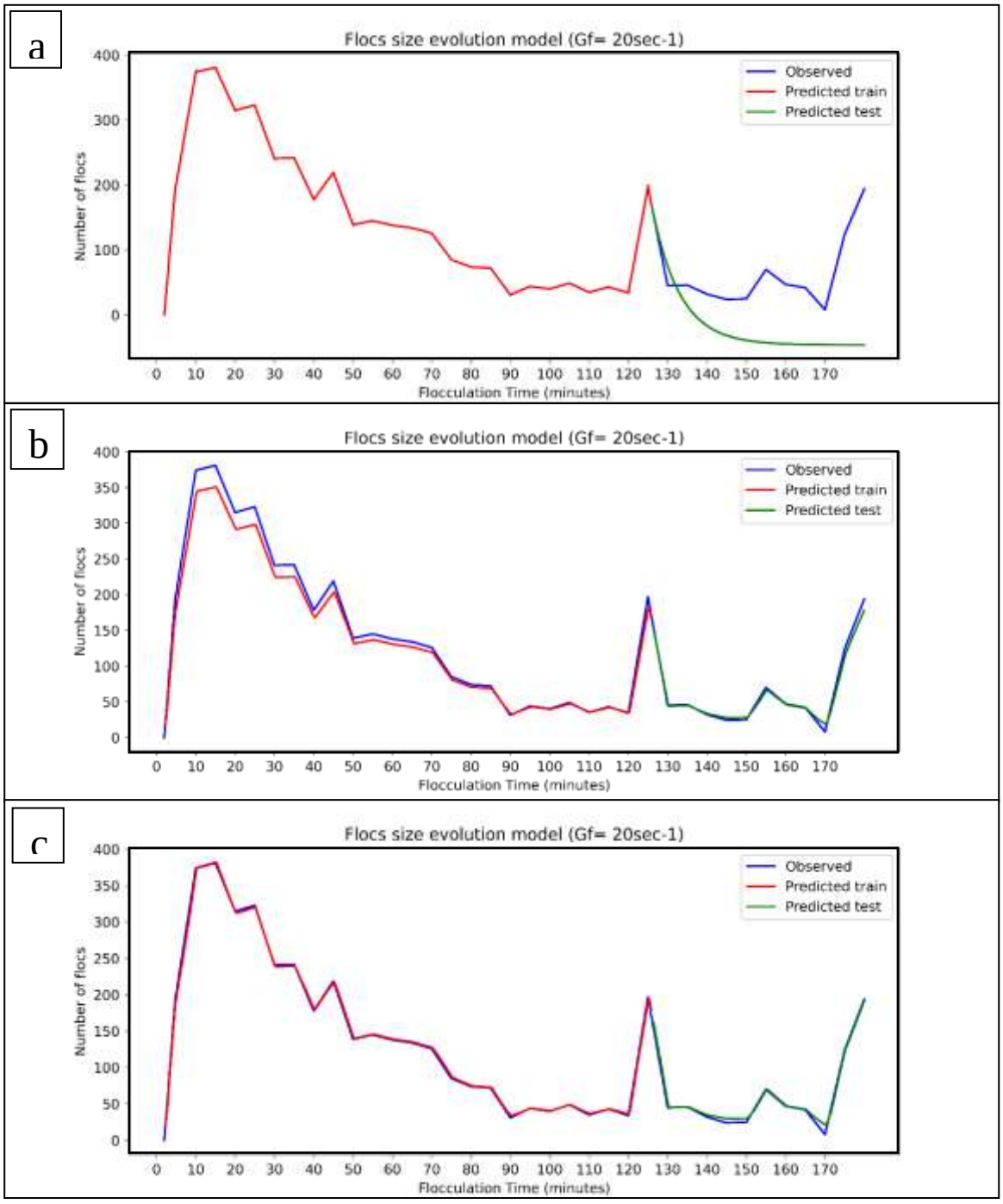

Figure 14: Comparison of different model prediction and observed number of flocs within the third floc length group under G*f* 20 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

The LSTM algorithm perfectly learned from the third group flocs evolution process of the G*f* 20 sec^{-1} dataset. The accuracy (R^2) of the model were 1.00 and 0.95, which indicates that the model accurately predicted the number of flocs in the group and its reliability. Also, the prediction error (MAE and RMSE) for the training phase were 0.37 and 0.61, and testing phase were 0.32 and 0.42, respectively. The perfect fit of the predicted number of flocs at both training and testing phase (as shown in Figure 14c above) reaffirm the effectiveness of the LSTM algorithm in modeling flocculation process, particularly floc evolution.

The peak of the floc length evolution is recorded at 15 mins after a sharp increase between 2 mins and 10 mins. A gentle decline in the number of larger flocs is observed before the first lowest count of flocs at 90 mins. This consistent downtrend could be attributed to both decrease in floc lengths due to the steady phase phenomenon or continuous breakage of aggregate particles. However, a continuous breakage in particles could have favored an increase in particular group size, which is not very evident except in the second group that has upward and downward pattern (increase and decrease in flocs) within specific numbers of flocs (average number of 1,500 flocs between 10 – 45 mins). Furthermore, the gentle fluctuation between 90 mins and 120 mins could be considered as flocs restructuring phase. This is considered possible due to weak bonding that may exist between fragmented flocs and the readily formed compact clusters of flocs. Also, the short time interval of the gentle fluctuation, followed by an increase in the flocs counts suggested that particle-cluster restructuring would be the dominant activities within the period.

The evolution of flocs within the third group of *Gf* 60 sec^{-1} dataset is shown in Figure 15a, b, and c below. The third group is also stationary with p-value and DF test statistics of 5.59 E-07 and -5.76, respectively. The Hyperparameter combination (p, d, and q) of 1,1,1 was used based on the auto Arima result. The ARIMA model performed poorly (as shown in Figure 15a) like the previous prediction, R2 = 0.97, MAE = 0.58, RMSE = 21.27 for training, and R2 = -6.57, MAE = 93.80, and RMSE = 101.54, were recorded for the testing phase. The ANN model achieved a perfect prediction accuracy for the training phase (R^2 = 1.0) and a nearly perfect testing accuracy (R^2 = 0.95), with a predicted training and testing error (MAE and RMSE) of 3.32 and 5.09, and 1.93 and 2.65, respectively. This high accuracy is also represented in the perfect overlay of the predicted number of flocs in

the third group on the observed trend of the third group floc length evolution (as shown in Figure 15b below). This further proves the effectiveness of the ANN model in accurate prediction of floc length evolution at different gradient velocity and particle sizes, which emphasizes its robustness and generalization potential for proper adoption in the flocculation studies and real-time forecasting at municipal water treatment facility.

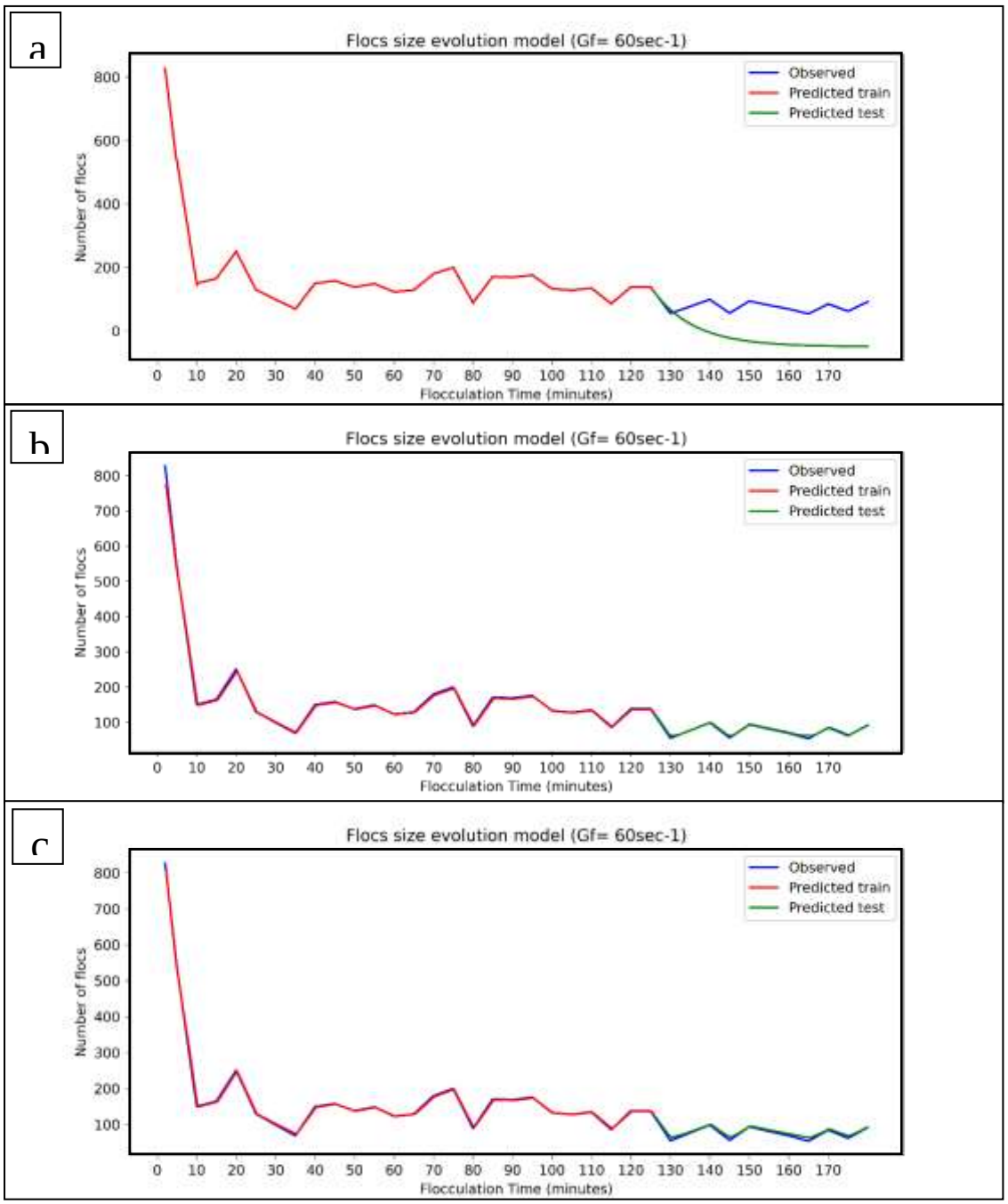

Figure 15: Comparison of different model prediction and observed number of flocs within the third floc length group under G*f* 60 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

The LSTM model achieved a perfect prediction at the training phase (R^2 = 1.0) and a strong testing prediction accuracy (R^2 = 0.83). The floc evolution prediction error (MAE and RMSE) was 2.59 and 3.61 for training stage, testing stage error (MAE and RMSE) were 4.40 and 4.97, respectively. The testing accuracy is lower compared to the ANN model, though model training prediction recorded the same accuracy. Nonetheless, the margin of prediction error is very close (MAE = 2.47 and RMSE = 2.32). This is obvious through Figure 15c that show the predicted number of flocs at testing phase to nearly overlay the observed numbers of flocs, with all drastic changes in floc evolution adequately captured by the model. Therefore, the LSTM algorithm is capable of accurately predicting the number of flocs within the groups across the G*f*.

Unlike the trend observed in G*f* 20 sec^{-1}, the number of particles rapidly decreased from the peak (above 800 flocs) at 2 mins to less than 200 at 10 mins. Though a small increase in the count was recorded by 20 mins, an unsteady up and downtrend was observed all through the entire experiment. Notably, the number of flocs within the third group decreases to a little above and below 100 from 130 mins to 180 mins. These could be assumed as the impact of high shear stress on the formation of larger floc particles, since high G*f* often favors the formation of smaller flocs with higher resistance to fragmentation and erosion mechanisms induced through the high G*f*. This aligns with the description of Moruzzi et al. (2019) that described smaller floc lengths to have higher shear stress (G*f*) resistant, and the strength factor of Al-Kaolin flocs and the floc lengths maintains similar behavior.

6.1.4 Fourth Floc length group (2.208 – 2.854 mm)

Floc length evolution within the fourth group (2.208 – 2.854 mm) under G*f* 20 sec^{-1} and the prediction output for ARIMA, ANN, and LSTM models are presented in Figure 16a, b, and c below. Just like other groups of the dataset, the ARIMA model recorded the least performance among the three models. Although the fourth group dataset is not stationary (p-value = 0.062 and DF test statistics = -2.772), the auto Arima algorithm recommended SARIMAX model (p = 1, d = 1, and q = 0), yet the prediction accuracy is poor (as shown in Figure 16a below).

The ANN model achieved a prediction accuracy (R^2) of 0.96 and 0.92 for training and testing phases. Prediction errors (MAE = 2.07 and RMSE = 3.71) were recorded during the training stage, and testing errors of 0.40 and 0.50 for MAE and RMSE,

respectively (as shown in Table 3 above). The model underpredicted the early part of the training phase (10 – 45 minutes) but accurately predicted the later part (as shown in Figure 16b above). This is similar to the underpredictions observed for the second and third group sizes of the G*f* 20 sec^{-1} (as shown in Figures 12 and 14 above), though the underpredicted time interval (T*f*) for group two is more. Similarly, the prediction accuracy improved with time, as observed in previous patterns. Therefore, it could be assumed that the ANN has the tendency of underpredicting the early phase of the larger floc lengths (groups) for G*f* dataset. Although, this will require further study to critically investigate the effect of varying G*f* on ANN prediction accuracy. Nonetheless, the improvement demonstrated during the test prediction shows that the algorithm is robust enough to make accurate prediction of any floc lengths from any shear stress.

Figure 16c shows the prediction accuracy of the LSTM model at both training and testing phases. Prediction accuracy of R^2 = 1.0 for training and R^2 = 0.95 for testing phase were recorded. The prediction error for the training phase were MAE = 0.37 and RMSE = 0.61, and testing phase; MAE = 0.32 and RMSE = 0.42, respectively. A very marginal error (less than 1.0) was recorded throughout the prediction phase, which proves the precision level of the algorithm and its capability to learn from the complex process and make accurate prediction. The LSTM also recorded higher accuracy than the ANN model with both R^2 and error margin (MAE and RMSE).

The steady number of flocs between 10 mins and 15 mins after the rapid increase due to floc aggregation can be considered as stability in the floc formation process. There is a gentle decline in the floc evolution trend with an unstable pattern between 30 – 55 mins. The region with wavy pattern is observed to have similar pattern with group two, which could imply that particle-cluster relationship at this phase is weak considering the differences between the particle sizes. The number of fourth group flocs reduced until a minor increase is recorded at 125 mins, before an almost flat period. It is essential to note that the number of flocs at this phase in the experiment is very small, compared to other group sizes, particularly the second group.

The prediction accuracy of the fourth group particle size evolution of G*f* 60 sec^{-1} by the ARIMA, ANN, and LSTM models are presented in Figure 17a, b, and c. The performance of the ARIMA model remains almost the same as previous datasets.

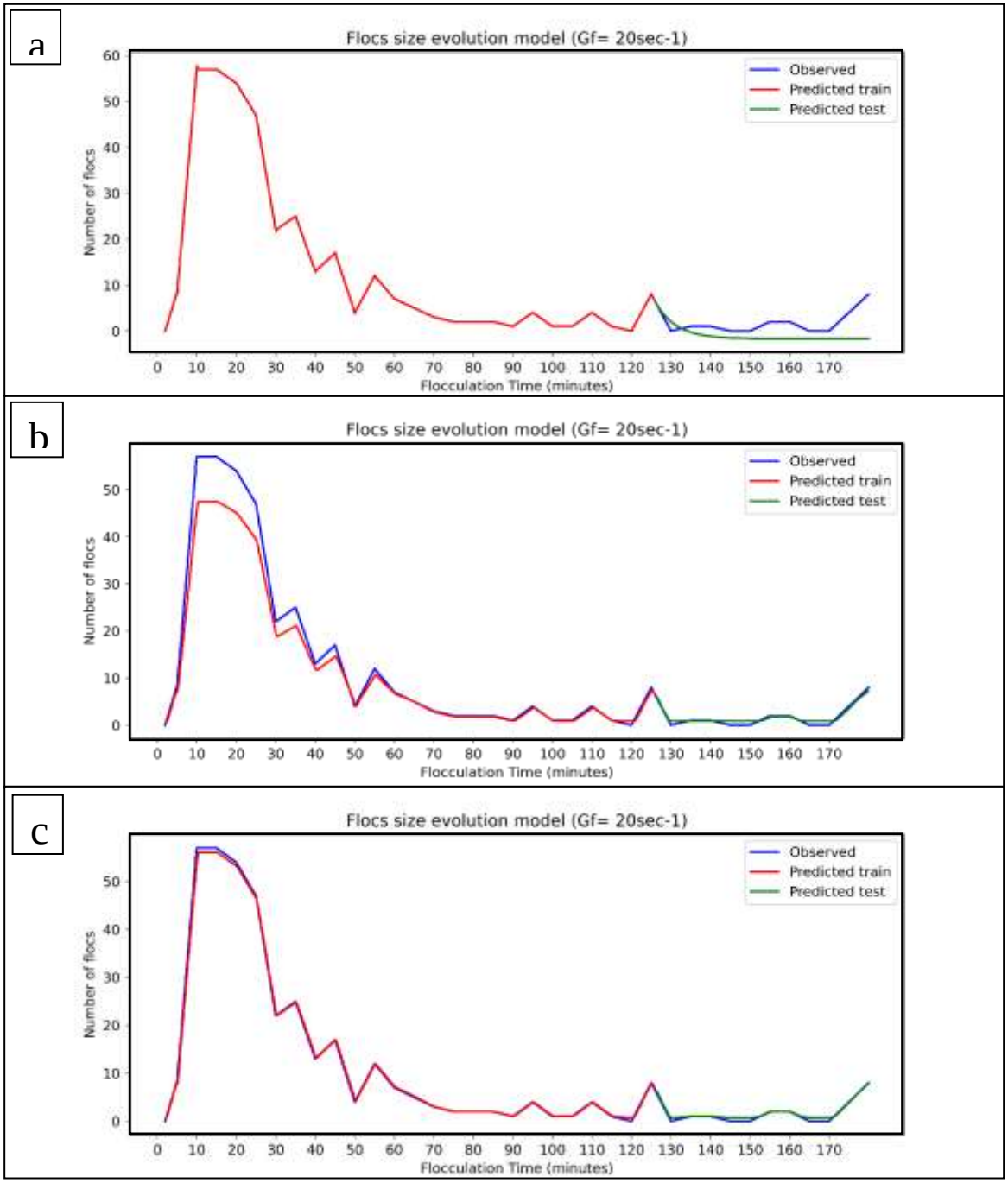

Figure 16: Comparison of different model prediction and observed number of flocs within the fourth floc length group under G*f* 20 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

Prediction accuracy for the training and testing phases were R^2 = 0.95 and -11.08, with prediction errors (MAE = 0.20 and RMSE = 7.64) for training and testing errors were (MAE = 9.98 and RMSE = 11.73). Thought a small prediction error were achieved at both phases (training and testing) the model still had a very poor prediction accuracy by

estimating negative values for number of flocs at some time steps (T*f*), which is considered impossible for a typical flocculation process and differs from other models' prediction.

The ANN model recorded prediction accuracy (R^2) of 1.0 and 0.84 for training and testing datasets. Prediction errors (MAE = 0.49 and RMSE = 1.11) were recorded at the training phase, testing prediction errors (MAE = 0.43 and RMSE = 0.43) were recorded, respectively.

The extremely high prediction accuracy of the floc evolution is a true demonstration of the algorithm's capability to adequately understudy the complex flocculation process (also evident in Figure 17b above). It could also be asserted that the ANN algorithm has the ability to improve in learning as the floc lengths increases i.e., accuracy of group1= group 5 > group 3 > group 4 > group 2

The LSTM model prediction of the number of flocs in the fourth group for the G*f* 60 sec^{-1} dataset recorded a training and testing accuracy (R^2) of 1.0 and 0.95, respectively. Also, training prediction error (MAE = 0.39 and RMSE = 0.72) were recorded, and testing error (MAE = 0.24 and RMSE = 0.25), respectively. The algorithm also perfectly predicted the number of flocs within the fourth group and recorded the best prediction accuracy, with almost zero error (as shown in Table 3). This is anticipated considering the prediction accuracy maintained by the algorithm in predicting other group sizes and G*f*. The model also predicted the numbers of floc in the fourth group better than the ANN model. Though both models showed very good accuracy with perfect overlay of the observed numbers of flocs against the predicted training and testing (Figures 17 b & c).

The sharp drop in the number of large flocs of 2.208 – 2.854 mm size is expected because of the high gradient velocity. The impact of G*f* variation on flocs evolution is evident through the distinct opposite trend and number of flocs recorded across the groups. Also, the observed trend (almost zero numbers of large flocs after 40 mins) proves that most floc within this size range has an extremely poor resistance to shear stress (probably because the floc length in the fourth group is very big). It could be inferred that the aggregation of primary particles in G*f* 60 sec^{-1} has a small particle size distribution (group 2 – 4) compared to the evolution of particles under G*f* 20 sec^{-1}, this was linked to the short interval of the transition phase as described by He et al. (2012), Moruzzi et al. (2020) and Moruzzi et al. (2018).

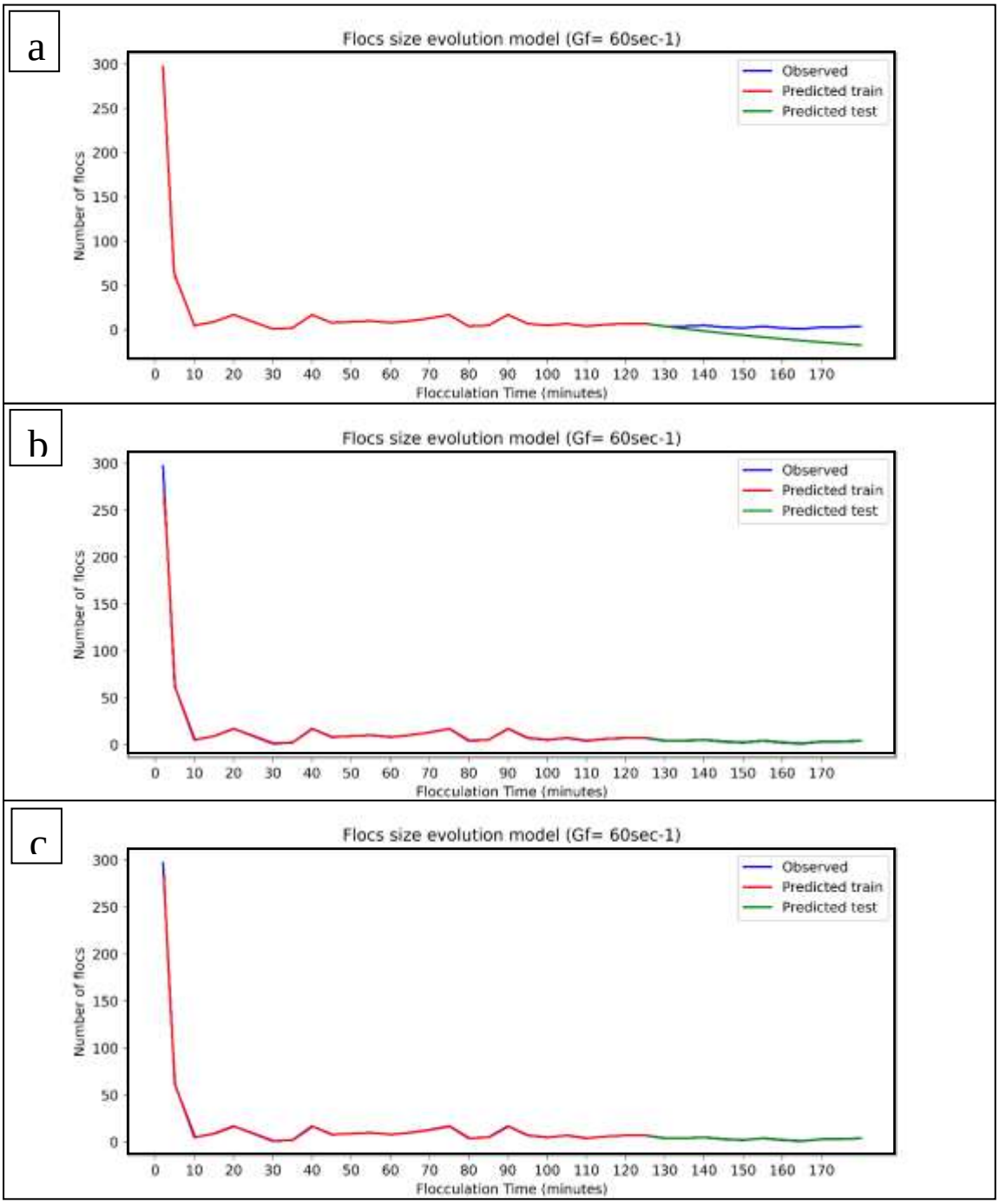

Figure 17: Comparison of different model prediction and observed number of flocs within the fourth floc length group under G*f* 60 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

Therefore, G*f* determines the particle aggregation process and their evolution with time; increase in G*f* causes decrease in floc lengths (Li & Logan, 1997; Thomas et al., 1999). This also aligns with the description of particle aggregation for both transition and steady phases, considering that fractal aggregates diminish in sizes with time (He et al.,

2012). This is in tandem with the findings of other studies on the fractal aggregation of Kaolin suspension using Al^{3+} coagulant, with varying G*f* (Argaman & Kaufman, 1970; Bratby, 2016; Haarhoff & Joubert, 1997; Spicer & Pratsinis, 1996; Yang et al., 2013).

6.1.5 Fifth Floc length group (2.854 – 3.500 mm)

The evolution of floc lengths of 2.854 – 3.500 mm under the G*f* 20 sec^{-1} typically follows the pattern demonstrated in the fourth group, with shorter interval of both transition and steady phases of flocs evolution (5 – 45 mins). The ARIMA model prediction of this group is similar to the fourth group, presented above. The comparison of the predicted number of flocs by ARIMA, ANN, and LSTM models against the observed number of flocs in the fifth group is presented in Figure 18a, b, and c. The DF test statistics and p-value were -3.52 and 0.0074 (i.e., dataset is stationary). The model was computed with p= 1, d = 1, q = 0 based on the auto Arima simulation. The poor accuracy of the model (R^2 = 1.0 and -4.93 for train and test dataset) and high prediction error all through the groups (1 – 5) of the G*f* 20 sec^{-1} (as shown in Table 3) further proves that the ARIMA model is not a suitable algorithm for modeling flocs evolution.

Nonetheless, the ANN algorithm was able to learn from few available data and precisely predict the numbers of flocs in the group size with time. A perfect training and testing accuracy (R^2 = 1.0) was achieved with training and testing prediction error (MAE) of 0.06 and 0.00, and RMSE of 0.13 and 0.01, respectively. Also, the LSTM model recorded similar high accuracy (R^2 = 1.0) for both training and testing phase, the MAE for training and testing phases were 0.08 and 0.01, and RMSE of 0.14 and 0.01 for training and testing phase, respectively. The prediction accuracy of the ANN and LSTM models are reflected in their perfect fit of the observed and predicted numbers of flocs (as shown in Figure 18a and b).

Both models (ANN & LSTM) have the overall best prediction accuracy in the fifth group, which could be traced to the lesser complexity in the evolution pattern, due to lack of values at many times step (T*f* = 45 – 55 mins; 65 – 120 mins, and 130 – 180 mins). This is an envisaged pattern of flocs evolution at the very large sizes, considering that floc lengths decrease with increased T*f* due to fragmentation and erosion of large particles.

Also, larger flocs tend to have weaker resistance strength, thus, the reduction in their number as flocculation progresses. The observed pattern of evolution at T*f* ≤ 40 mins

(rapid peak before decrease in number of flocs) shows that transition and steady phases of flocs in fifth group followed the description of Moruzzi et al. (2018) on the fractal evolution during flocculation process.

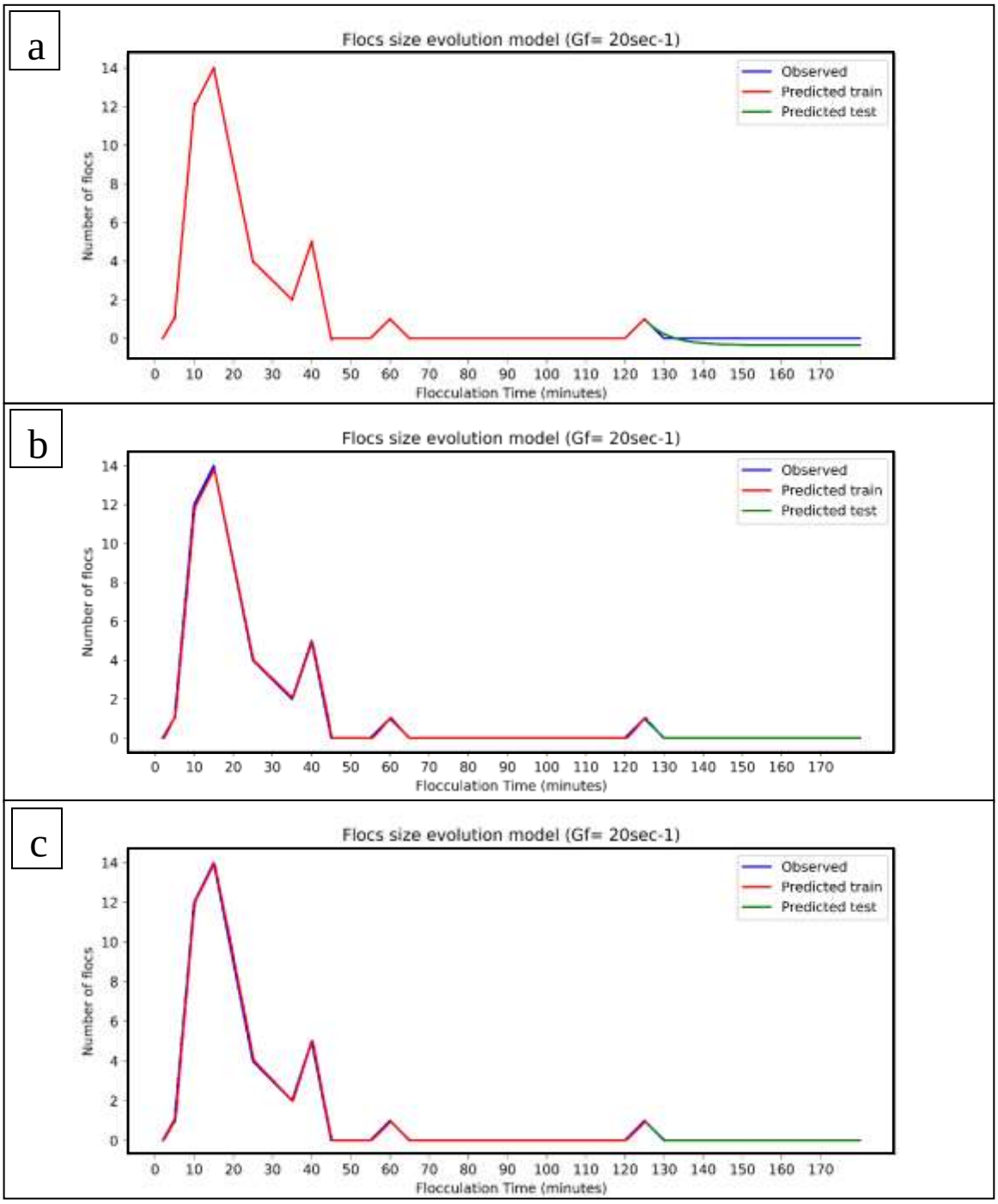

Figure 18: Comparison of different model prediction and observed number of flocs within the fifth floc length group under G*f* 20 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

Similarly, flocs evolution at the fifth group of G*f* 60 sec^{-1} followed the pattern of the fourth group (as shown in Figure 17a, b, and d). The ARIMA model recorded prediction accuracy R^2 = 0.95 for training and R^2 = -0.016 for testing phase, errors (MAE = 0.06 and RMSE = 2.21) were recorded for training, and testing error (MAE = 0.09 and RMSE = 0.25) were recorded, respectively. Notably, the fifth group dataset is stationary with p-value and DF test statistics of 3.74 E-18 and -10.28, respectively. Although the prediction errors were low, the model still recorded some negative values at some time step, which is similar to other groups in the dataset. The comparing of the model with the observed is presented in Figure 19a).

The ANN model recorded a very high accuracy of R^2 = 1.00 for training and R^2 = 0.99 for testing phase. Prediction errors were MAE = 0.08 and RMSE = 0.33 for the training, and MAE = 0.01 and RMSE = 0.02 for testing. Also, the LSTM model maintains similar accuracy (R^2) of 1.00 and 0.98 for training and testing, with training MAE and RMSE of 0.06 and 0.19, respectively. Model testing prediction errors were MAE = 0.01 and RMSE = 0.03. The flocs evolution is in the form of the fourth group floc length evolution of G*f* 20 sec^{-1}. The number of flocs drastically reduced from the peak at 2 mins to Zero at 10 mins. Afterwards, the number of flocs in the fifth group remained zero, except at 40 mins, 70 mins, 110 mins, and 140 mins. This is considered to be a contributing factor to the prediction accuracy and errors, as justified above (also, as presented in Figure 16a and b, and Figure 19a and b).

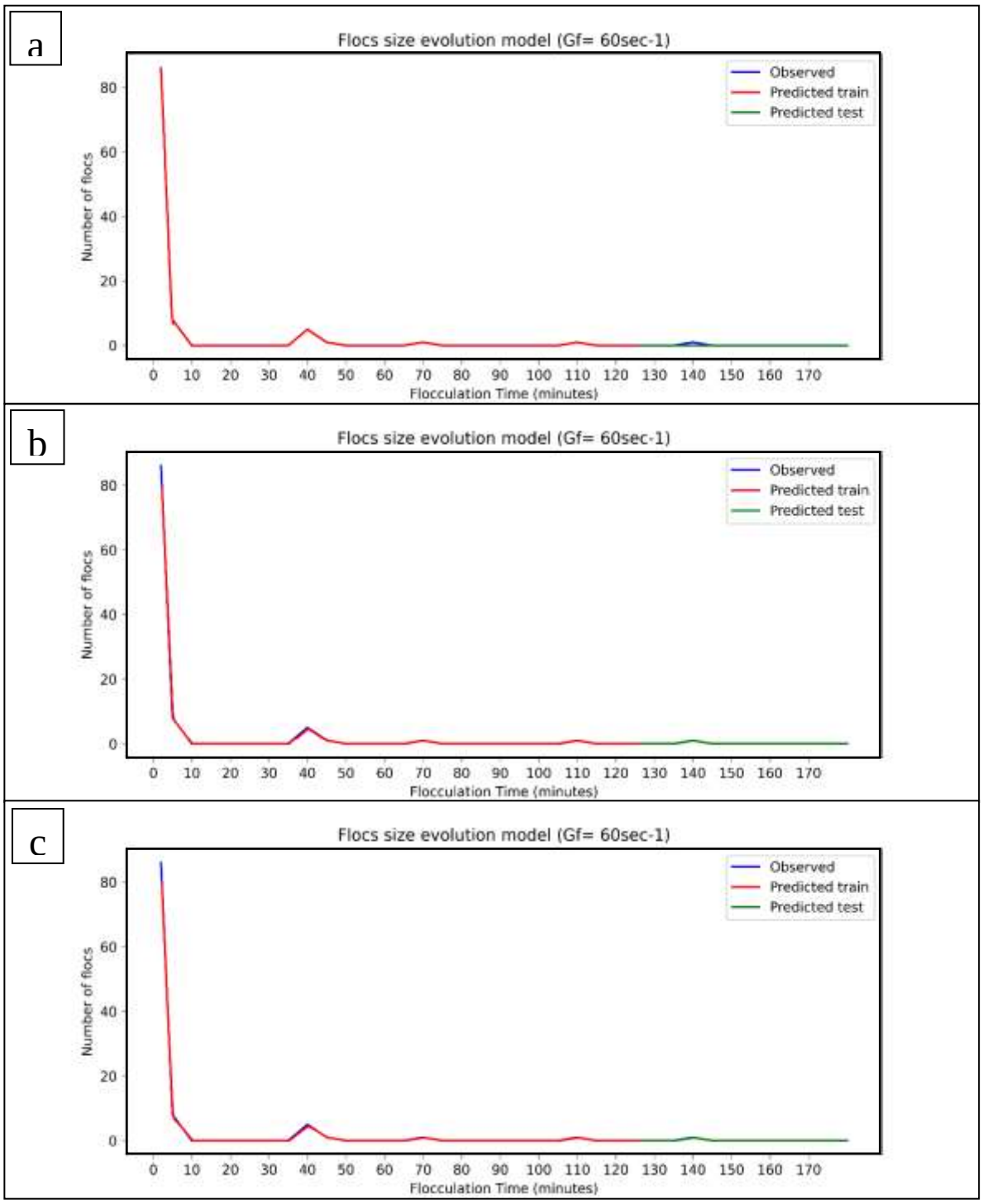

Figure 19: Comparison of different model prediction and observed number of flocs within the fifth floc length group under G*f* 60 sec^{-1} (a) ARIMA model, (b) ANN model, (c) LSTM model.

6.1.6 Deep learning Model Training and Validation Loss

The behavior of neural networks during training phase is vital to understand its extent of generalization and potential level of accuracy on newly introduced data. Training and validation loss are vital indicators of the model performances on both seen (training data) and unseen dataset (testing data). The learning process of the model empowers the neural network model to learn background patterns and trends in the dataset and make accurate prediction on similar information. The model minimizes the training loss by adjusting the weight and biases to make better predictions on the training data. Meanwhile, validation loss is the error between the predicted outputs and the actual observation of the testing dataset/validation dataset (Goodfellow et al., 2016). This is used to evaluate the model's performance on new datasets aside from the training set and to check for the models' level of overfitting.

The validation loss also gives insight on the model generalization capability, particularly, the prediction response to new dataset (unseen data). If the validation loss increases while the model decreases, it may indicate that the model is overfitting. Meanwhile, if the model validation and training loss keep decreasing as the model learn from the data, and converge to a low value, it implies that the model effectively learnt from the training data and generalized well on the unseen data, leading to improved model accuracy (Goodfellow et al., 2016; Hochreiter & Schmidhuber, 1997; LeCun et al., 2015; Srivastava et al., 2014). Therefore, it is essential that training and validation loss converge to ensure that the minimum error level is attained, while the best model accuracy is achieved.

The ANN and LSTM model training and validation loss during the process of learning the evolution of flocs from the G*f* 20 sec^{-1} data is presented in Figure 20 below. Notably, both models (ANN and LSTM) demonstrated an outstanding and impressive learning and generalization capability with respect to floc evolution of the flocculation process of water and wastewater treatment. The hyperparameter tunning process of the models as related to the floc lengths (groups) are presented in Table 2 above. The best hyperparameter combination for the model also displays a pattern in the model's ability to master the evolution process.

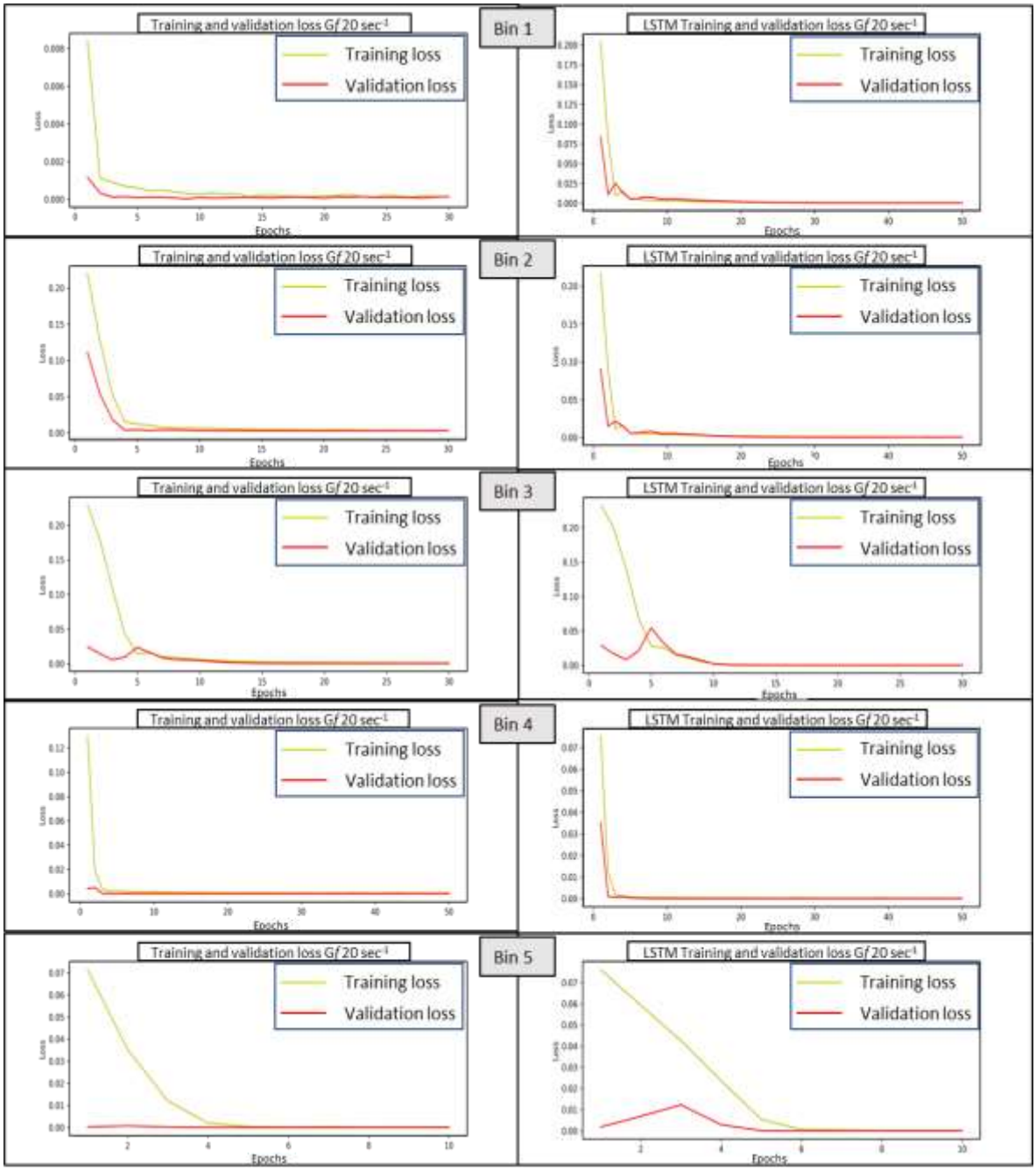

Figure 20: Deep learning model (ANN and LSTM) training and validation loss on prediction of number of flocs in different groups of G*f* 20 sec^{-1} dataset.

For instance, both the LSTM and ANN models maintained similar numbers of hidden layers (4) and dense neurons (128: 64: 32: 8: 1) for Group 1 – 4 of the G*f* 20 sec^{-1} dataset, which are the main group with the greatest number of flocs. Similar number of epochs was observed except for the third group that has 30 epochs due to early stopping. The early stopping is a regularization approach adopted to avoid model overfitting (Nielsen et al., 2020), because the model converge early and tends to diverge at about 35 epochs .

Both models converge at the early stage of the learning process and achieved almost 0.0002 loss at less than 10 epochs, except for Group 5 where each model has late convergence. This is attributed to the lack of flocs at many time steps due to the nature of the floc evolution and the dataset (fifth group). Also, the LSTM model has a very sharp drop in the training loss and converges with the validation loss for the first and second groups, compared to the gentle exponential-like pattern of the ANN model at group 2 & 3.

The similarity observed in the pattern of neuron and hidden layers above (G*f* 20 sec^{-1}) is also observed in the LSTM model for the G*f* 60 sec^{-1} data, with group 2 – 5 utilizing 64: 32: 8: 1 dense neuron. ANN used 10:1 dense neuron to effectively predict flocs evolution of group 4 & 5, respectively. Both models recorded an extremely high accuracy with very minimal error (< 0.02) for both training and validation prediction, which proves that neural network model can effectively predict floc evolution of the flocculation process under any given shear stress (G*f*). Although the ANN model has delayed convergence at group 3, the LSTM converge at less than 10 epochs for all the groups (as shown in Figure 21 below).

Despite the paucity of adequate findings on the implementation of neural networks or machine learning models in flocculation studies, particularly modeling of data from non-intrusive method of fractal dimensions. Our current result has proven that little data could be used to generate resourceful information on flocs behavior during the flocculation process and could improve the treatment efficiency. It is also worthy to note that our model has achieved the best accuracy with least training and validation loss (prediction error) compared to other studies.

For instance, Zhu et al. (2022) developed a tensor diagram from a convolutional neural network model to predict the pollutant migration during flocculation process. Although their algorithm used the convolutional neural network layer for flocculation image data and achieved a very high prediction accuracy of 92 – 98%, their training and validation loss recorded an ultimate loss of 0.25 (which is high compared to <0.02 achieved in our study). Additionally, the model validation loss did not converge after the first minor overlap at about 10 epochs, which indicates a potential overfitting of their model.

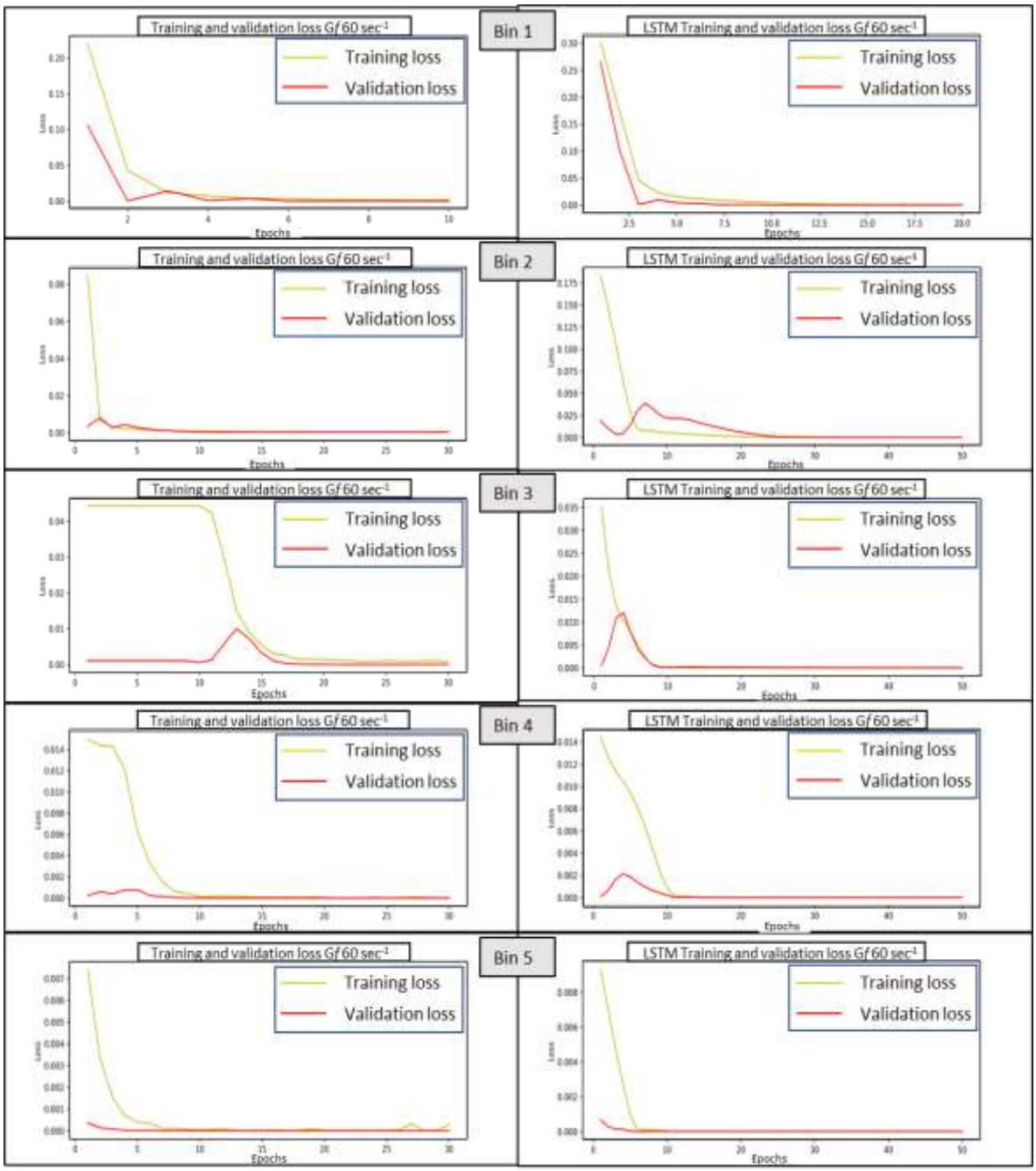

Figure 21: Deep learning model (ANN and LSTM) training and validation loss on prediction of number of flocs in different groups of G*f* 60 sec^{-1} dataset.

Our model accuracy also achieved better output when compared to that of Nielsen et al. (2020) that hybridized machine learning based soft sensors and kinetic model (population balance model) to predict particle agglomeration and breakage during the flocculation process. The authors recorded 32% accuracy which is lower compared to the accuracy recorded by our model, and model runtime of 15.6 minutes for 134 epochs, which is also higher than our average runtime of less than 3 mins for up to one hundred epochs.

This also proves that the ANN and LSTM model are effective in modeling flocculation process.

Nazemzadeh et al. (2021) also recorded higher training and validation loss (loss > 5) in their study that integrated the first principles model (Population balance and Mass balance models) with neural network algorithm to predict the silica flocs agglomeration and breakage during a laboratory flocculation experiment. Interestingly, their study approached the flocculation process as a time series based model, while Nielsen et al. (2020) used the regression combined with kinetic model approach. This justifies our choice of the LSTM model and the adaptation of ANN as a time series model, although other non-linear regression models (random forest, support vector machine, XG Boost, etc.) also stands a chance of adoption for time series forecasting.

Water resources modeling studies have used the coefficient of determination (R^2) to evaluate machine learning algorithms' effectiveness and superiority by juxtaposing the predicted and observed variables (Chicco et al., 2021; El-Rawy et al., 2021; Uddin et al., 2022). Scatter plots and R^2 were used to validate the effectiveness of both ANN and LSTM models. Figure 22 and 23 shows the regression plot of observed and predicted numbers of flocs by ANN and LSTM models for G*f* 20 sec^{-1} and 60 sec^{-1}, respectively. Both Figures confirmed that the two models are well suitable for modeling flocs evolution with almost 100 % accuracy ($R^2 \geq 0.99$). Generally, it could be seen that LSTM displayed a slight superiority above the ANN, based on the results of each group discussed above.

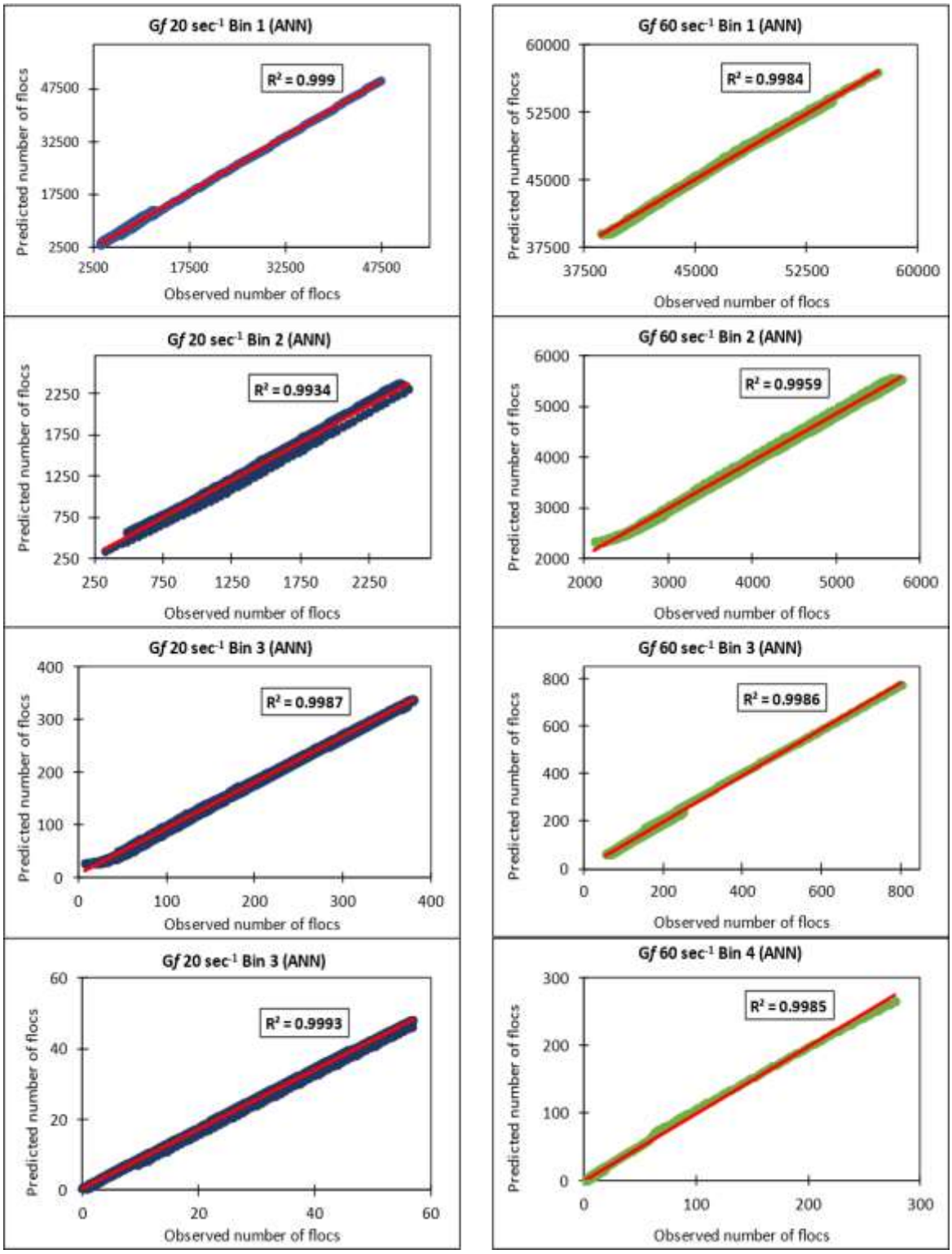

Figure 22: ANN model results showing the observed number of flocs from the experiment versus the predicted number of flocs per group and the regression factor (R^2) value.

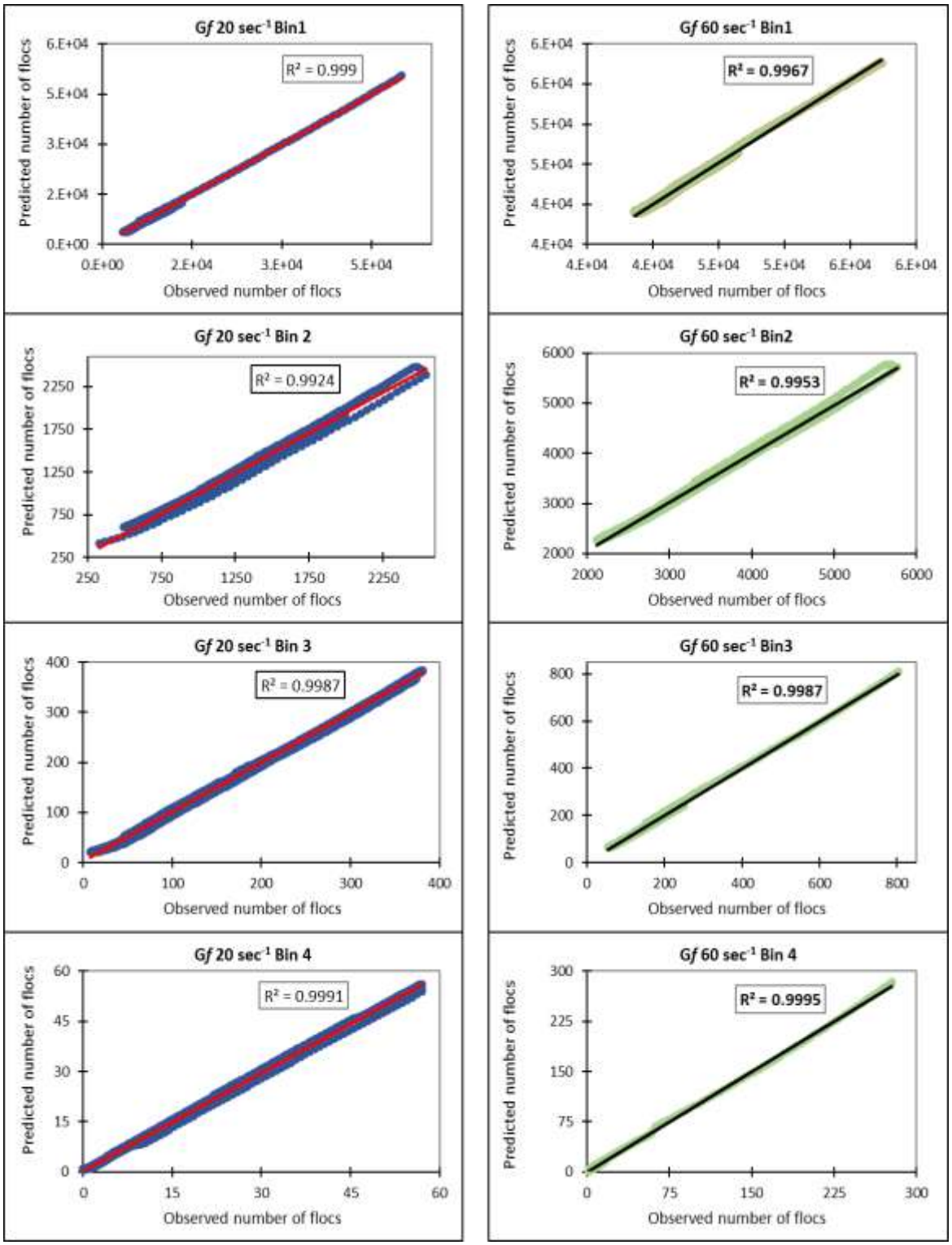

Figure 23: LSTM model results showing the observed number of flocs from the experiment versus the predicted number of flocs per group and the regression factor (R^2) value.

7 CONCLUSIONS AND BENEFITS OF THE RESEARCH

This study developed a machine learning framework for optimizing the flocculation process of water and wastewater treatment. The study focus was in two tiers: modeling of floc length evolution and particle aggregation and breakage. Two ML algorithms (ANN & LSTM) and a traditional time series model (ARIMA) were explored to model and predict floc length evolution data that was obtained through non-intrusive image analysis from a jar test batch assay and model the aggregation and breakage process. Results showed that ARIMA model is not a suitable algorithm for modeling floc lengths, with a negative prediction on numbers of flocs and R^2 for all testing data predictions across all groups.

ANN algorithm performed excellent with floc length evolution prediction R^2 of 0.86 – 1.0 for training and 0.84 – 0.99 for testing, across G*f* 20sec^{-1} and G*f* 60 sec^{-1}. floc length evolution prediction by ANN recorded slight training underprediction at similar *Tf* on for group 2, 3, and 4 of G*f* 20 sec^{-1} dataset, but with improved performance on the testing datasets. Similar pattern is observed for the second group of G*f* 60 sec^{-1} dataset. LSTM model displayed slight superiority with training and testing prediction R^2 of 0.99 – 1.0 for training and R^2 of 0.83 – 1.0 for testing, across both G*f* 20 sec^{-1} and 60 sec^{-1}, respectively. Notably, LSTM has an excellent test prediction accuracy of 98 – 100 % across all groups in G*f* 20 sec^{-1} dataset, and least accuracy of 83% for G*f* 60 sec^{-1}. Also, LSTM showed a perfect prediction of the number of flocs across all groups and G*f*s. The prediction perfectly reflects the underlying factors responsible for the evolution such as the transition and steady phases, and the impact of fragmentation and erosion caused by shear stress on flocs formation.

Model training loss and validation plots proved that all models (ANN and LSTM) converged very early and did not diverge by the end of epochs, except for G*f* 60 sec^{-1} first group that early stopping was used. Both ANN and LSTM displayed architectural stability

in modeling the groups of each data set (128: 64: 32: 8 : 1 neurons produced best result for group 1 – 4 of both G*f* 20 sec^{-1} and 60 sec^{-1}). The regression plot of observed and predicted numbers of flocs per group were in the order of accuracy with overall accuracy of 99% for both ANN & LSTM models.

This study has proven that ANN and LSTM models, and the developed framework are highly suitable for modeling floc evolution during flocculation process of water treatment.

7.1 Benefits of this research

Adoption of the framework developed from the findings of this research holds lots of prospect in large-scale water and wastewater treatment facilities and promote future learning on ML implementation in flocculation process. Adoption of this framework could promote the application of smart technology in full-scale water/wastewater treatment facility and facilitates the realization of the Sustainable Development Goal 6. Further aiding the attainment of target 1, 2, 3, 11, 13, and 14.

The following specific sustainable benefits are also envisaged.

a. Research-based benefits:

I. Developed an approach to model flocculation with **small data.**
II. Developed framework can be adopted in flocculation studies at Pilot scale.
III. Resolve the complexity of hydraulic-based principles in flocculation modeling.
IV. Potential adoption for online/real-time floc evolution monitoring.

b. Real-time operational benefit:

V. Aid the integration of smart technology in flocculation phase of water treatment.
VI. Reduced man-hour loss and equipment overburden.
VII. Reduced operational and maintenance costs of water and wastewater treatment process.
VIII. Provide an early warning system to change in flocs characteristics and forecast optimum time and shear velocity per treatment process.
IX. promote better understanding and data presentation to policymakers.
X. Promote quality assurance and quality control of water and wastewater treatment.

8 BIBLIOGRAPHIC REFERENCES

Abu Bakar, S. N. H., Abu Hasan, H., Abdullah, S. R. S., Kasan, N. A., Muhamad, M. H., & Kurniawan, S. B. (2021). A review of the production process of bacteria-based polymeric flocculants. *Journal of Water Process Engineering*, *40*, 101915. https://doi.org/10.1016/j.jwpe.2021.101915

Abu-Dalo, M., Abdelnabi, J., Al-Rawashdeh, N. A. F., Albiss, B., & Al Bawab, A. (2022). Coupling coagulation-flocculation to volcanic tuff-magnetite nanoparticles adsorption for olive mill wastewater treatment. *Environmental Nanotechnology, Monitoring & Management*, *17*, 100626. https://doi.org/10.1016/j.enmm.2021.100626

Ahamed, F., Singh, M., Song, H.-S., Doshi, P., Ooi, C. W., & Ho, Y. K. (2020). On the use of sectional techniques for the solution of depolymerization population balances: Results on a discrete-continuous mesh. *Advanced Powder Technology*, *31*(7), 2669–2679. https://doi.org/10.1016/j.apt.2020.04.032

Akinmolayan, F., Thornhill, N., & Sorensen, E. (2015). A Detailed Mathematical Modelling Representation of Clean Water Treatment Plants. In K. V. Gernaey, J. K. Huusom, & R. Gani (Eds.), *Computer Aided Chemical Engineering* (Vol. 37, pp. 2537–2542). Elsevier. https://doi.org/10.1016/B978-0-444-63576-1.50117-5

Al Nasser, W. N., & Al Salhi, F. H. (2015). Kinetics determination of calcium carbonate precipitation behavior by inline techniques. *Powder Technology*, *270*, 548–560. https://doi.org/10.1016/j.powtec.2014.05.025

Altaee, A., Millar, G., Zaragoza, G., & Sharif, A. (2017). Energy efficiency of RO and FO-RO system for high-salinity seawater treatment. *Clean Technologies and Environmental Policy*, *19*(1), 77–91. https://eprints.qut.edu.au/103709/

Ammar, Y., Cognet, P., & Cabassud, M. (2021). ANN for hybrid modelling of batch and fed-batch chemical reactors. *Chemical Engineering Science*, *237*, 116522. https://doi.org/10.1016/j.ces.2021.116522

Arab, M., Akbarian, H., Gheibi, M., Akrami, M., Fathollahi-Fard, A. M., Hajiaghaei-Keshteli, M., & Tian, G. (2022). A soft-sensor for sustainable operation of coagulation and flocculation units. *Engineering Applications of Artificial Intelligence*, *115*, 105315. https://doi.org/10.1016/j.engappai.2022.105315

Argaman, Y., & Kaufman, W. J. (1970). Turbulence and Flocculation. *Journal of the Sanitary Engineering Division*, *96*(2), 223–241. https://doi.org/10.1061/JSEDAI.0001073

Bagheri, M., Akbari, A., & Mirbagheri, S. A. (2019). Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: A critical review. *Process Safety and Environmental Protection*, *123*, 229–252. https://doi.org/10.1016/j.psep.2019.01.013

Bratby, J. (2016). Coagulation and Flocculation in Water and Wastewater Treatment. *Water Intelligence Online*, *15*(0), 9781780407500–9781780407500. https://doi.org/10.2166/9781780407500

Buddhiraju, V. S., & Runkana, V. (2012). Simulation of nanoparticle synthesis in an aerosol flame reactor using a coupled flame dynamics–monodisperse population balance model. *Journal of Aerosol Science*, *43*(1), 1–13. https://doi.org/10.1016/j.jaerosci.2011.08.007

Bushell, G. (2005). Forward light scattering to characterise structure of flocs composed of large particles. *Chemical Engineering Journal*, *111*(2), 145–149. https://doi.org/10.1016/j.cej.2005.02.021

Chaouki, Z., Hadri, M., Nawdali, M., Benzina, M., & Zaitan, H. (2021). Treatment of a landfill leachate from Casablanca city by a coagulation-flocculation and adsorption process

using a palm bark powder (PBP). *Scientific African*, *12*, e00721. https://doi.org/10.1016/j.sciaf.2021.e00721

Chen, Y., Wu, J., & Wu, Z. (2022). China's commercial bank stock price prediction using a novel K-means-LSTM hybrid approach. *Expert Systems with Applications*, *202*, 117370. https://doi.org/10.1016/j.eswa.2022.117370

Chen, Z., Huang, Z., Liu, J., Wu, E., Zheng, Q., & Cui, L. (2021). Phase transition of Mg/Al-flocs to Mg/Al-layered double hydroxides during flocculation and polystyrene nanoplastics removal. *Journal of Hazardous Materials*, *406*, 124697. https://doi.org/10.1016/j.jhazmat.2020.124697

Cheng, S. Y., Show, P.-L., Juan, J. C., Chang, J.-S., Lau, B. F., Lai, S. H., Ng, E. P., Yian, H. C., & Ling, T. C. (2021). Landfill leachate wastewater treatment to facilitate resource recovery by a coagulation-flocculation process via hydrogen bond. *Chemosphere*, *262*, 127829. https://doi.org/10.1016/j.chemosphere.2020.127829

Cheng, S. Y., Show, P.-L., Juan, J. C., Ling, T. C., Lau, B. F., Lai, S. H., & Ng, E. P. (2020). Sustainable landfill leachate treatment: Optimize use of guar gum as natural coagulant and floc characterization. *Environmental Research*, *188*, 109737. https://doi.org/10.1016/j.envres.2020.109737

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Ching, P. M. L., Zou, X., Wu, D., So, R. H. Y., & Chen, G. H. (2022). Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environmental Research*, *210*, 112953. https://doi.org/10.1016/j.envres.2022.112953

Cortés Muñoz, J. E., Calderón Mólgora, C. G., Martín Domínguez, A., Espino de la O, E. E., Gelover Santiago, S. L., & Hernández Martínez, C. L. (2013). Endocrine Disruptors in Water Sources: Human Health Risks and EDs Removal from Water Through Nanofiltration. In *International Perspectives on Water Quality Management and Pollutant Control* (Vol. 2). Intech Open. http://dx.doi.org/10.5772/54482

Costa, C. B. B., Maciel, M. R. W., & Filho, R. M. (2007). Considerations on the crystallization modeling: Population balance solution. *Computers & Chemical Engineering*, *31*(3), 206–218. https://doi.org/10.1016/j.compchemeng.2006.06.005

Cubillos, M. (2020). Multi-site household waste generation forecasting using a deep learning approach. *Waste Management*, *115*, 8–14. https://doi.org/10.1016/j.wasman.2020.06.046

Didavi, A. B. K., Agbokpanzo, R. G., & Agbomahena, M. (2021). Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system. *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, 1–5. https://doi.org/10.1109/BioSMART54244.2021.9677566

Du, P., Li, X., Yang, Y., Fan, X., Zhang, T., Wang, N., Li, H., Ji, S., & Zhou, Z. (2021). Effect of rapid-mixing conditions on the evolution of micro-flocs to final aggregates during two-stage alum addition. *Environmental Technology*, *42*(20), 3122–3131. https://doi.org/10.1080/09593330.2020.1723710

Elkiran, G., Nourani, V., & Abba, S. I. (2019). Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology*, *577*, 123962. https://doi.org/10.1016/j.jhydrol.2019.123962

El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering*, *44*, 102380. https://doi.org/10.1016/j.jwpe.2021.102380

Farhi, N., Kohen, E., Mamane, H., & Shavitt, Y. (2021). Prediction of wastewater treatment quality using LSTM neural network. *Environmental Technology & Innovation*, *23*, 101632. https://doi.org/10.1016/j.eti.2021.101632

Feng, J., Yang, J., Li, Y., Wang, H., Ji, H., Yang, W., & Wang, K. (2021). Load forecasting of electric vehicle charging station based on grey theory and neural network. *Energy Reports*, *7*, 487–492. https://doi.org/10.1016/j.egyr.2021.08.015

Ferreira Filho, S. S., Hespanhol, I., & Moreira, H. A. (2000). Flocculation kinetics of colloidal suspensions: Effects of metallic coagulant dosage and primary particle

concentration on the breakup and aggregation contants. *Chemical water and wastewater treatment : proceedings*. https://repositorio.usp.br/item/001137423

Filho, S. S. F., Hespanhol, I., & Moreira, H. A. (2000). Flocculation Kinetics of Colloidal Suspensions: Effects of Metallic Coagulant Dosage and Primary Particle Concentration on the Breakup and Aggregation Constants. In H. H. Hahn, E. Hoffmann, & H. Ødegaard (Eds.), *Chemical Water and Wastewater Treatment VI* (pp. 101–109). Springer. https://doi.org/10.1007/978-3-642-59791-6_10

Ganguli, A., Ganguly, P., Das, P., & Saha, A. (2020). Integral approach for the treatment of phenolic wastewater using gamma irradiation and graphene oxide. *Groundwater for Sustainable Development*, *10*, 100355. https://doi.org/10.1016/j.gsd.2020.100355

Ghaed Rahmati, M., Tishehzan, P., & Moazed, H. (2021). Determining the best and simple intelligent models for evaluating BOD5 of Ahvaz wastewater treatment plant. *DESALINATION AND WATER TREATMENT*, *209*, 242–253. https://doi.org/10.5004/dwt.2021.26481

Gharabaghi, B., & Sattar, A. M. A. (2019). Empirical models for longitudinal dispersion coefficient in natural streams. *Journal of Hydrology*, *575*, 1359–1361. https://doi.org/10.1016/j.jhydrol.2017.01.022

Gharib, A., & Davies, E. G. R. (2021). A workflow to address pitfalls and challenges in applying machine learning models to hydrology. *Advances in Water Resources*, *152*, 103920. https://doi.org/10.1016/j.advwatres.2021.103920

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. In *Deep Learning* (pp. 120–122). MIT press. http://alvarestech.com/temp/deep/Deep%20Learning%20by%20Ian%20Goodfellow,%20Yoshua%20Bengio,%20Aaron%20Courville%20(z-lib.org).pdf

Graça, N. S., Ribeiro, A. M., & Rodrigues, A. E. (2022). Modeling and optimization of a continuous electrocoagulation process using an artificial intelligence approach. *Water Supply*, *22*(1), 643–658. Scopus. https://doi.org/10.2166/ws.2021.249

Granata, F., Papirio, S., Esposito, G., Gargano, R., & De Marinis, G. (2017). Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. *Water*, *9*(2), Article 2. https://doi.org/10.3390/w9020105

Gregory, J. (2009). Monitoring particle aggregation processes. *Advances in Colloid and Interface Science*, *147–148*, 109–123. https://doi.org/10.1016/j.cis.2008.09.003

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J., Kim, J. H., & Cho, K. H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences*, *32*, 90–101. https://doi.org/10.1016/j.jes.2015.01.007

Haarhoff, J., & Joubert, H. (1997). Determination of aggregation and breakup constants during flocculation. *Water Science and Technology*, *36*(4), 33–40. https://doi.org/10.1016/S0273-1223(97)00416-2

Hadiyanto, H., Christwardana, M., Widayat, W., Jati, A. K., & Laes, S. I. (2021). Optimization of flocculation efficiency and settling time using chitosan and eggshell as bio-flocculant in Chlorella pyrenoidosa harvesting process. *Environmental Technology & Innovation*, *24*, 101959. https://doi.org/10.1016/j.eti.2021.101959

He, W., Nan, J., Li, H., & Li, S. (2012). Characteristic analysis on temporal evolution of floc size and structure in low-shear flow. *Water Research*, *46*(2), 509–520. https://doi.org/10.1016/j.watres.2011.11.040

Heddam, S. (2021). 24 - Extremely randomized tree: A new machines learning method for predicting coagulant dosage in drinking water treatment plant. In P. Samui, H. Bonakdari, & R. Deo (Eds.), *Water Engineering Modeling and Mathematic Tools* (Vol. 24, pp. 475–489). Elsevier. https://doi.org/10.1016/B978-0-12-820644-7.00013-X

Heddam, S., & Dechemi, N. (2015). A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modelling coagulant dosage (Dos): Case study of water treatment plant of Algeria. *Desalination and Water Treatment*, *53*(4), 1045–1053. Scopus. https://doi.org/10.1080/19443994.2013.878669

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hounslow, M. J., Ryall, R. L., & Marshall, V. R. (1988). A discretized population balance for nucleation, growth, and aggregation. *AIChE Journal*, *34*(11), 1821–1832. https://doi.org/10.1002/aic.690341108

Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021). Machine learning in natural and engineered water systems. *Water Research*, *205*, 117666. https://doi.org/10.1016/j.watres.2021.117666

Huang, Y., Zhang, B., Liu, B., Su, S., Han, G., Wang, W., Guo, H., & Cao, Y. (2021). Clean and deep separation of molybdenum and rhenium from ultra-low concentration solutions via vapidly stepwise selective coagulation and flocculation precipitation. *Separation and Purification Technology*, *267*, 118632. https://doi.org/10.1016/j.seppur.2021.118632

Ibrar, I., Altaee, A., Zhou, J. L., Naji, O., & Daoud, K. (2019). Challenges and potentials of forward osmosis process in the treatment of wastewater. *Critical Reviews in Environmental Science and Technology*. https://doi.org/10.1080/10643389.2019.1657762

Igwegbe, C. A., Onukwuli, O. D., Ighalo, J. O., & Menkiti, M. C. (2021). Bio-coagulation-flocculation (BCF) of municipal solid waste leachate using Picralima nitida extract: RSM and ANN modelling. *Current Research in Green and Sustainable Chemistry*, *4*, 100078. https://doi.org/10.1016/j.crgsc.2021.100078

Jayaweera, C., Othman, M., & Aziz, N. (2019). Improved predictive capability of coagulation process by extreme learning machine with radial basis function. *Journal of Water Process Engineering*, *32*, 100977. https://doi.org/10.1016/j.jwpe.2019.100977

Jiang, J.-Q. (2015). The role of coagulation in water treatment. *Current Opinion in Chemical Engineering*, *8*, 36–44. https://doi.org/10.1016/j.coche.2015.01.008

Jiao, R., Fabris, R., Chow, C. W. K., Drikas, M., van Leeuwen, J., & Wang, D. (2016). Roles of coagulant species and mechanisms on floc characteristics and filterability. *Chemosphere*, *150*, 211–218. https://doi.org/10.1016/j.chemosphere.2016.02.030

Juntunen, P., Liukkonen, M., Lehtola, M., & Hiltunen, Y. (2012). Characterization of Alum Floc by Image Analysis in Water Treatment Processes. *IFAC Proceedings Volumes*, *45*(2), 959–963. https://doi.org/10.3182/20120215-3-AT-3016.00169

Khatri, N., Khatri, K. K., & Sharma, A. (2019). Prediction of effluent quality in ICEAS-sequential batch reactor using feedforward artificial neural network. *Water Science and Technology*, *80*(2), 213–222. https://doi.org/10.2166/wst.2019.257

Khatri, N., Khatri, K. K., & Sharma, A. (2020). Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. *Journal of Water Process Engineering*, *37*, 101477. https://doi.org/10.1016/j.jwpe.2020.101477

Kim, C. M., & Parnichkun, M. (2016). MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant. *Journal of Water Supply: Research and Technology-Aqua*, *66*(1), 49–61. https://doi.org/10.2166/aqua.2016.022

Kim, J. E., Phuntsho, S., Chekli, L., Hong, S., Ghaffour, N., Leiknes, T., Choi, J. Y., & Shon, H. K. (2017). Environmental and economic impacts of fertilizer drawn forward osmosis and nanofiltration hybrid system. *Desalination*, *416*, 76–85.

Kim, Y., Woo, Y. C., Phuntsho, S., Nghiem, L. D., Shon, H. K., & Hong, S. (2017). Evaluation of fertilizer-drawn forward osmosis for coal seam gas reverse osmosis brine treatment and sustainable agricultural reuse. *Journal of Membrane Science*, *537*, 22–31.

Kurniawan, S. B., Imron, M. F., Chik, C. E. N. C. E., Owodunni, A. A., Ahmad, A., Alnawajha, M. M., Rahim, N. F. M., Said, N. S. M., Abdullah, S. R. S., Kasan, N. A., Ismail, S., Othman, A. R., & Hasan, H. A. (2022). What compound inside biocoagulants/bioflocculants is contributing the most to the coagulation and flocculation processes? *Science of The Total Environment*, *806*, 150902. https://doi.org/10.1016/j.scitotenv.2021.150902

Lawrence, T. J., Carr, S. J., Wheatland, J. A. T., Manning, A. J., & Spencer, K. L. (2022). Quantifying the 3D structure and function of porosity and pore space in natural sediment flocs. *Journal of Soils and Sediments*. https://doi.org/10.1007/s11368-022-03304-x

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

Li, B., Zhao, J., Ge, W., Li, W., & Yuan, H. (2022). Coagulation-flocculation performance and floc properties for microplastics removal by magnesium hydroxide and PAM. *Journal of Environmental Chemical Engineering*, *10*(2), 107263. https://doi.org/10.1016/j.jece.2022.107263

Li, C., Busquets, R., Moruzzi, R. B., & Campos, L. C. (2021). Preliminary study on low-density polystyrene microplastics bead removal from drinking water by coagulation-flocculation and sedimentation. *Journal of Water Process Engineering*, *44*, 102346. https://doi.org/10.1016/j.jwpe.2021.102346

Li, J., & Wang, J. (2020). Forcasting of energy futures market and synchronization based on stochastic gated recurrent unit model. *Energy*, *213*, 118787. https://doi.org/10.1016/j.energy.2020.118787

Li, L., Rong, S., Wang, R., & Yu, S. (2021). Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal*, *405*, 126673. https://doi.org/10.1016/j.cej.2020.126673

Li, T., Zhu, Z., Wang, D., Yao, C., & Tang, H. (2007). The strength and fractal dimension characteristics of alum–kaolin flocs. *International Journal of Mineral Processing*. https://www.semanticscholar.org/paper/The-strength-and-fractal-dimension-characteristics-Li-Zhu/51902ee512f261bd0dcc7417add26729fc8cfe39

Li, X., & Logan, B. E. (1997). Collision Frequencies between Fractal Aggregates and Small Particles in a Turbulently Sheared Fluid. *Environmental Science & Technology*, *31*(4), 1237–1242. https://doi.org/10.1021/es960772o

Liang, L., Peng, Y., Tan, J., & Xie, G. (2015). A review of the modern characterization techniques for flocs in mineral processing. *Minerals Engineering*, *84*, 130–144. https://doi.org/10.1016/j.mineng.2015.10.011

Lopez-Exposito, P., Negro, C., & Blanco, A. (2019). Direct estimation of microalgal flocs fractal dimension through laser reflectance and machine learning. *Algal Research*, *37*, 240–247. https://doi.org/10.1016/j.algal.2018.12.007

Lu, C. F., & Spielman, L. A. (1985). Kinetics of floc breakage and aggregation in agitated liquid suspensions. *Journal of Colloid and Interface Science*, *103*(1), 95–105. https://doi.org/10.1016/0021-9797(85)90080-3

Lu, C., Xu, Z., Dong, B., Zhang, Y., Wang, M., Zeng, Y., & Zhang, C. (2022). Machine learning for the prediction of heavy metal removal by chitosan-based flocculants. *Carbohydrate Polymers*, *285*, 119240. https://doi.org/10.1016/j.carbpol.2022.119240

Marques, R. de O., & Ferreira Filho, S. S. (2017). Flocculation kinetics of low-turbidity raw water and the irreversible floc breakup process. *Environmental Technology*, *38*(7), 901–910. https://doi.org/10.1080/09593330.2016.1236149

Marques, R. de O., & Filho, S. S. F. (2022). Further investigation of the irreversible floc breakup in flocculation kinetics modelling. *Water Supply*, *22*(4), 3814–3823. https://doi.org/10.2166/ws.2022.023

Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Computer Science*, *170*, 1168–1173. https://doi.org/10.1016/j.procs.2020.03.049

Moruzzi, R. B., Bridgeman, J., & Silva, P. A. G. (2020). A combined experimental and numerical approach to the assessment of floc settling velocity using fractal geometry. *Water Science and Technology*, *81*(5), 915–924. https://doi.org/10.2166/wst.2020.171

Moruzzi, R. B., Campos, L. C., Sharifi, S., da Silva, P. G., & Gregory, J. (2020). Nonintrusive investigation of large Al-kaolin fractal aggregates with slow settling velocities. *Water Research*, *185*, 116287. https://doi.org/10.1016/j.watres.2020.116287

Moruzzi, R. B., da Silva, P. G., Sharifi, S., Campos, L. C., & Gregory, J. (2019). Strength assessment of Al-Humic and Al-Kaolin aggregates by intrusive and non-intrusive methods. *Separation and Purification Technology*, *217*, 265–273. https://doi.org/10.1016/j.seppur.2019.02.033

Moruzzi, R. B., de Oliveira, A. L., da Conceição, F. T., Gregory, J., & Campos, L. C. (2017). Fractal dimension of large aggregates under different flocculation conditions. *Science of The Total Environment*, *609*, 807–814. https://doi.org/10.1016/j.scitotenv.2017.07.194

Moruzzi, R. B., & de Oliveira, S. C. (2013). Mathematical modeling and analysis of the flocculation process in chambers in series. *Bioprocess and Biosystems Engineering*, *36*(3), 357–363. https://doi.org/10.1007/s00449-012-0791-4

Moruzzi, R. B., Oliveira, A. L. de, & Almeida, T. de. (2018). Fractal Aggregates Evolution During Flocculation. *Brazilian Journal of Chemical Engineering*, *35*, 1203–1210. https://doi.org/10.1590/0104-6632.20180354s20170231

Moruzzi, R. B., & Reali, M. A. P. (2010). Characterization of micro-bubble size distribution and flow configuration in DAF contact zone by a non-intrusive image analysis system and tracer tests. *Water Science and Technology*, *61*(1), 253–262. https://doi.org/10.2166/wst.2010.784

Moruzzi, R. B., & Reali, M. A. P. (2014). The influence of floc size and hydraulic detention time on the performance of a dissolved air flotation (DAF) pilot unit in the light of a mathematical model. *Bioprocess and Biosystems Engineering*, *37*(12), 2445–2452. https://doi.org/10.1007/s00449-014-1221-6

Mundi, G., Zytner, R. G., Warriner, K., Bonakdari, H., & Gharabaghi, B. (2021). Machine learning models for predicting water quality of treated fruit and vegetable wastewater. *Water (Switzerland)*, *13*(18). Scopus. https://doi.org/10.3390/w13182485

Nasr, M. S., Moustafa, M. A. E., Seif, H. A. E., & El Kobrosy, G. (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. *Alexandria Engineering Journal*, *51*(1), 37–43. https://doi.org/10.1016/j.aej.2012.07.005

Nazemzadeh, N., Malanca, A. A., Nielsen, R. F., Gernaey, K. V., Andersson, M. P., & Mansouri, S. S. (2021). Integration of first-principle models and machine learning in a modeling framework: An application to flocculation. *Chemical Engineering Science*, *245*, 116864. https://doi.org/10.1016/j.ces.2021.116864

Nazemzadeh, N., Olivé, J. S., Nielsen, R. F., Gernaey, K. V., Andersson, M. P., & Mansouri, S. S. (2022). A combinatorial tool for monitoring flocculation processes: Using non-invasive measurements and hybrid deep learning assisted modelling. In Y. Yamashita & M. Kano (Eds.), *Computer Aided Chemical Engineering* (Vol. 49, pp. 811–816). Elsevier. https://doi.org/10.1016/B978-0-323-85159-6.50135-4

Nielsen, R. F., Nazemzadeh, N., Sillesen, L. W., Andersson, M. P., Gernaey, K. V., & Mansouri, S. S. (2020). Hybrid machine learning assisted modelling framework for particle processes. *Computers & Chemical Engineering*, *140*, 106916. https://doi.org/10.1016/j.compchemeng.2020.106916

Nyström, F., Nordqvist, K., Herrmann, I., Hedström, A., & Viklander, M. (2020). Removal of metals and hydrocarbons from stormwater using coagulation and flocculation. *Water Research*, *182*, 115919. https://doi.org/10.1016/j.watres.2020.115919

Oliveira, A. da S., Lopes, V. dos S., Filho, U. C., Moruzzi, R. B., & de Oliveira, A. L. (2018). Neural network for fractal dimension evolution. *Water Science and Technology*, *78*(4), 795–802. https://doi.org/10.2166/wst.2018.349

Oliveira, P., Fernandes, B., Aguiar, F., Pereira, M. A., Analide, C., & Novais, P. (2020). A Deep Learning Approach to Forecast the Influent Flow in Wastewater Treatment Plants. In C. Analide, P. Novais, D. Camacho, & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2020* (pp. 362–373). Springer International Publishing. https://doi.org/10.1007/978-3-030-62362-3_32

Ortiz, A., García-Galán, M. J., García, J., & Díez-Montero, R. (2021). Optimization and operation of a demonstrative full scale microalgae harvesting unit based on coagulation, flocculation and sedimentation. *Separation and Purification Technology*, *259*, 118171. https://doi.org/10.1016/j.seppur.2020.118171

Pan, B. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, *113*(1), 012127. https://doi.org/10.1088/1755-1315/113/1/012127

Park, C. M., Heo, J., Wang, D., Su, C., & Yoon, Y. (2018). Heterogeneous activation of persulfate by reduced graphene oxide–elemental silver/magnetite nanohybrids for the oxidative degradation of pharmaceuticals and endocrine disrupting compounds in water. *Applied Catalysis B: Environmental*, *225*, 91–99.

Parsa, M. M., Pourfakhar, H., & Baghdadi, M. (2020). Application of graphene oxide nanosheets in the coagulation-flocculation process for removal of Total Organic Carbon (TOC) from surface water. *Journal of Water Process Engineering*, *37*, 101367. https://doi.org/10.1016/j.jwpe.2020.101367

Pisa, I., Morell, A., Vicario, J. L., & Vilanova, R. (2020). LSTM-based IMC approach applied in Wastewater Treatment Plants: Performance and stability analysis. *IFAC-PapersOnLine*, *53*(2), 16569–16574. https://doi.org/10.1016/j.ifacol.2020.12.782

Qasim, M., Park, S., & Kim, J.-O. (2021). A model to determine the drag coefficient of aggregated nonspherical flocs in the ballasted flocculation. *Journal of Water Process Engineering*, *44*, 102409. https://doi.org/10.1016/j.jwpe.2021.102409

Qasim, M., Park, S., Moon, Y., & Kim, J.-O. (2020). Developing a model to determine the settling velocity of ballasted flocs. *Journal of Environmental Chemical Engineering*, *8*(6), 104515. https://doi.org/10.1016/j.jece.2020.104515

Qi, L., Meng, X., Zhang, R., Liu, H., Xu, C., Liu, Z., & Klusener, P. A. A. (2015). Droplet size distribution and droplet size correlation of chloroaluminate ionic liquid–heptane dispersion in a stirred vessel. *Chemical Engineering Journal*, *268*, 116–124. https://doi.org/10.1016/j.cej.2015.01.009

Qiao, D., Wu, S., Li, G., You, J., Zhang, J., & Shen, B. (2022). Wind speed forecasting using multi-site collaborative deep learning for complex terrain application in valleys. *Renewable Energy*, *189*, 231–244. https://doi.org/10.1016/j.renene.2022.02.095

Rajala, K., Grönfors, O., Hesampour, M., & Mikola, A. (2020). Removal of microplastics from secondary wastewater treatment plant effluent by coagulation/flocculation with iron, aluminum and polyamine-based chemicals. *Water Research*, *183*, 116045. https://doi.org/10.1016/j.watres.2020.116045

Reddy, C. V., Rao, D. S., & Kalamdhad, A. S. (2022). Combined treatment of high-strength fresh leachate from municipal solid waste landfill using coagulation-flocculation and fixed bed upflow anaerobic filter. *Journal of Water Process Engineering*, *46*, 102554. https://doi.org/10.1016/j.jwpe.2021.102554

Rong, H., Gao, B., Li, J., Zhang, B., Sun, S., Wang, Y., Yue, Q., & Li, Q. (2013). Floc characterization and membrane fouling of polyferric–polymer dual/composite coagulants in coagulation/ultrafiltration hybrid process. *Journal of Colloid and Interface Science*, *412*, 39–45. https://doi.org/10.1016/j.jcis.2013.09.013

Ruan, Z., Wu, A., Bürger, R., Betancourt, F., Ordoñez, R., Wang, J., Wang, S., & Wang, Y. (2022). A Population Balance Model for Shear-Induced Polymer-Bridging Flocculation of Total Tailings. *Minerals*, *12*(1), Article 1. https://doi.org/10.3390/min12010040

Safeer, S., Pandey, R. P., Rehman, B., Safdar, T., Ahmad, I., Hasan, S. W., & Ullah, A. (2022). A review of artificial intelligence in water purification and wastewater treatment: Recent advancements. *Journal of Water Process Engineering*, *49*, 102974. https://doi.org/10.1016/j.jwpe.2022.102974

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1308. https://doi.org/10.1007/s42452-020-3060-1

Santos Nunes, G., Sammarro Silva, K. J., Souza Freitas, B. L., Belini, V. L., & Sabogal-Paz, L. P. (2022). In-situ microscopy investigation of floc development during coagulation-flocculation with chemical and natural coagulants. *Separation Science and Technology*, *57*(14), 2312–2322. https://doi.org/10.1080/01496395.2022.2056055

Seghir, S., Hasseine, A., & Rasteiro, M. G. (2022). Describing the flocculation of PCC particles using population balance modelling approaches. *Chemical Engineering Research and Design*, *186*, 638–646. https://doi.org/10.1016/j.cherd.2022.08.038

Singh, B., & Kumar, P. (2020). Pre-treatment of petroleum refinery wastewater by coagulation and flocculation using mixed coagulant: Optimization of process parameters using response surface methodology (RSM). *Journal of Water Process Engineering*, *36*, 101317. https://doi.org/10.1016/j.jwpe.2020.101317

Singh, M., Ranade, V., Shardt, O., & Matsoukas, T. (2022). Challenges and opportunities concerning numerical solutions for population balances: A critical review. *Journal of Physics A: Mathematical and Theoretical*, *55*(38), 383002. https://doi.org/10.1088/1751-8121/ac8a42

Spicer, P. T., & Pratsinis, S. E. (1996). Coagulation and fragmentation: Universal steady-state particle-size distribution. *AIChE Journal*, *42*(6), 1612–1620. https://doi.org/10.1002/aic.690420612

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Suhr, H., Wehnert, G., Schneider, K., Bittner, C., Scholz, T., Geissler, P., Jähne, B., & Scheper, T. (1995). In situ microscopy for on-line characterization of cell-populations in bioreactors, including cell-concentration measurements by depth from focus. *Biotechnology and Bioengineering*, *47*(1), 106–116. https://doi.org/10.1002/bit.260470113

Taloba, A. I. (2022). An Artificial Neural Network Mechanism for Optimizing the Water Treatment Process and Desalination Process. *Alexandria Engineering Journal*, *61*(12), 9287–9295. https://doi.org/10.1016/j.aej.2022.03.029

Tegladza, I. D., Xu, Q., Xu, K., Lv, G., & Lu, J. (2021). Electrocoagulation processes: A general review about role of electro-generated flocs in pollutant removal. *Process Safety and Environmental Protection*, *146*, 169–189. https://doi.org/10.1016/j.psep.2020.08.048

Teh, C. Y., Wu, T. Y., & Juan, J. C. (2014). Potential use of rice starch in coagulation–flocculation process of agro-industrial wastewater: Treatment performance and flocs characterization. *Ecological Engineering*, *71*, 509–519. https://doi.org/10.1016/j.ecoleng.2014.07.005

Teixeira, M. S., Speranza, L. G., da Silva, I. C., Moruzzi, R. B., & Silva, G. H. R. (2022). Tannin-based coagulant for harvesting microalgae cultivated in wastewater: Efficiency, floc morphology and products characterization. *Science of The Total Environment*, *807*, 150776. https://doi.org/10.1016/j.scitotenv.2021.150776

Teng, X., Zhang, X., & Luo, Z. (2022). Multi-scale local cues and hierarchical attention-based LSTM for stock price trend prediction. *Neurocomputing*, *505*, 92–100. https://doi.org/10.1016/j.neucom.2022.07.016

Thomas, D. N., Judd, S. J., & Fawcett, N. (1999). Flocculation modelling: A review. *Water Research*, *33*, 1579–1592. https://doi.org/10.1016/S0043-1354(98)00392-3

Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, *321*, 115923. https://doi.org/10.1016/j.jenvman.2022.115923

Wang, D., Chang, X., & Ma, K. (2022). Predicting flocculant dosage in the drinking water treatment process using Elman neural network. *Environmental Science and Pollution Research*, *29*(5), 7014–7024. https://doi.org/10.1007/s11356-021-16265-4

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management*, *301*, 113941. https://doi.org/10.1016/j.jenvman.2021.113941

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of The Total Environment*, *784*, 147138. https://doi.org/10.1016/j.scitotenv.2021.147138

Wang, D., Wu, R., Jiang, Y., & Chow, C. W. K. (2011). Characterization of floc structure and strength: Role of changing shear rates under various coagulation mechanisms. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *379*(1), 36–42. https://doi.org/10.1016/j.colsurfa.2010.11.048

Wen, J., Yang, J., Li, Y., & Gao, L. (2022). Harmful algal bloom warning based on machine learning in maritime site monitoring. *Knowledge-Based Systems*, *245*, 108569. https://doi.org/10.1016/j.knosys.2022.108569

Xiao, F., Lam, K. M., Li, X. Y., Zhong, R. S., & Zhang, X. H. (2011). PIV characterisation of flocculation dynamics and floc structure in water treatment. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *379*(1), 27–35. https://doi.org/10.1016/j.colsurfa.2010.11.053

Yadav, K. K., Kumar, S., Pham, Q. B., Gupta, N., Rezania, S., Kamyab, H., Yadav, S., Vymazal, J., Kumar, V., Tri, D. Q., Talaiekhozani, A., Prasad, S., Reece, L. M., Singh, N., Maurya, P. K., & Cho, J. (2019). Fluoride contamination, health problems and remediation methods in Asian groundwater: A comprehensive review. *Ecotoxicology and Environmental Safety*, *182*. Scopus. https://doi.org/10.1016/j.ecoenv.2019.06.045

Yang, Z., Yang, H., Jiang, Z., Huang, X., Li, H., Li, A., & Cheng, R. (2013). A new method for calculation of flocculation kinetics combining Smoluchowski model with

fractal theory. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *423*, 11–19. https://doi.org/10.1016/j.colsurfa.2013.01.058

Yu, J., Xu, H., Yang, X., Sun, H., Jin, Z., & Wang, D. (2022). Floc formation and growth during coagulation removing humic acid: Effect of stirring condition. *Separation and Purification Technology*, *302*, 122084. https://doi.org/10.1016/j.seppur.2022.122084

Zhang, H., Yang, L., Zang, X., Cheng, S., & Zhang, X. (2019). Effect of shear rate on floc characteristics and concentration factors for the harvesting of Chlorella vulgaris using coagulation-flocculation-sedimentation. *Science of The Total Environment*, *688*, 811–817. https://doi.org/10.1016/j.scitotenv.2019.06.321

Zhang, K., Achari, G., Li, H., Zargar, A., & Sadiq, R. (2013). Machine learning approaches to predict coagulant dosage in water treatment plants. *International Journal of Systems Assurance Engineering and Management*, *4*(2), 205–214. Scopus. https://doi.org/10.1007/s13198-013-0166-5

Zhang, Q., Ye, X., Li, H., Chen, D., Xiao, W., Zhao, S., Xiong, R., & Li, J. (2020). Cumulative effects of pyrolysis temperature and process on properties, chemical speciation, and environmental risks of heavy metals in magnetic biochar derived from coagulation-flocculation sludge of swine wastewater. *Journal of Environmental Chemical Engineering*, *8*(6), 104472. https://doi.org/10.1016/j.jece.2020.104472

Zhu, G., Lin, J., Fang, H., Yuan, F., Li, X., Yuan, C., & Hursthouse, A. S. (2022). A flocculation tensor to monitor water quality using a deep learning model. *Environmental Chemistry Letters*, *20*(6), 3405–3414. https://doi.org/10.1007/s10311-022-01524-8