Continuous monitoring of onsite water recycling systems using machine learning based soft-sensors

Hsiang-Yang (Gary) Shyu[1],

R.A. Bair[1], C.J. Castro[1], Q. Lu[1], D.H. Yeh[1]

[1] University of South Florida

UNIVERSITY of SOUTH FLORIDA

Disclaimer:

The materials being presented represent the speaker's own opinions and do NOT reflect the opinions of NOWRA.

Onsite Water Reuse Systems

Opportunities in Water Reuse

- Treat wastewater onsite for non-potable reuse
- Enable local water recycling for applications like toilet flushing and irrigation

•The Monitoring Imperative for Safe Reuse

- Water quality from these systems must be consistent and reliable to ensure safety
- System reliability is a known challenge
- 14% of conventional household sewage treatment systems in Ohio were estimated failing

Why Monitoring Matters in Decentralized Systems

Public Health Risk

- Inadequately treated water can spread pathogens
- Delayed fault detection increases health risks
- Safe reuse requires consistent treatment performance

Monitoring Challenges

- Lab-based testing is slow, expensive, and labor-intensive
- Real-time monitoring unavailable for small-scale systems
- Frequent sampling is hard to sustain in low-resource settings

•Need for Solutions

- Cost-effective, real-time monitoring options
- Scalable tools for decentralized and off-grid applications

Soft Sensors: A Real-Time Monitoring Solution

Data-Driven Soft Sensors

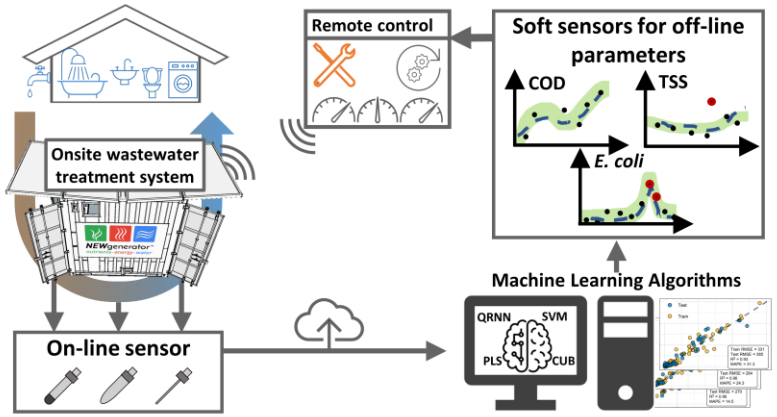
- Estimate key water quality parameters using ML algorithms
- Use real-time data from in-line sensors (e.g., pH, EC, turbidity)
- Proven effective in centralized treatment plants
- Rarely applied in onsite or decentralized systems

Challenges in Small Systems

- Limited historical data for model training
- High variability in system operation and influent quality
- Few documented case studies for OWTS applications

•Study Objectives

- Develop soft sensors for COD, TSS, and *E. coli*
- Use 56 weeks of field data from a decentralized reuse system



What is a Soft Sensor?

Data

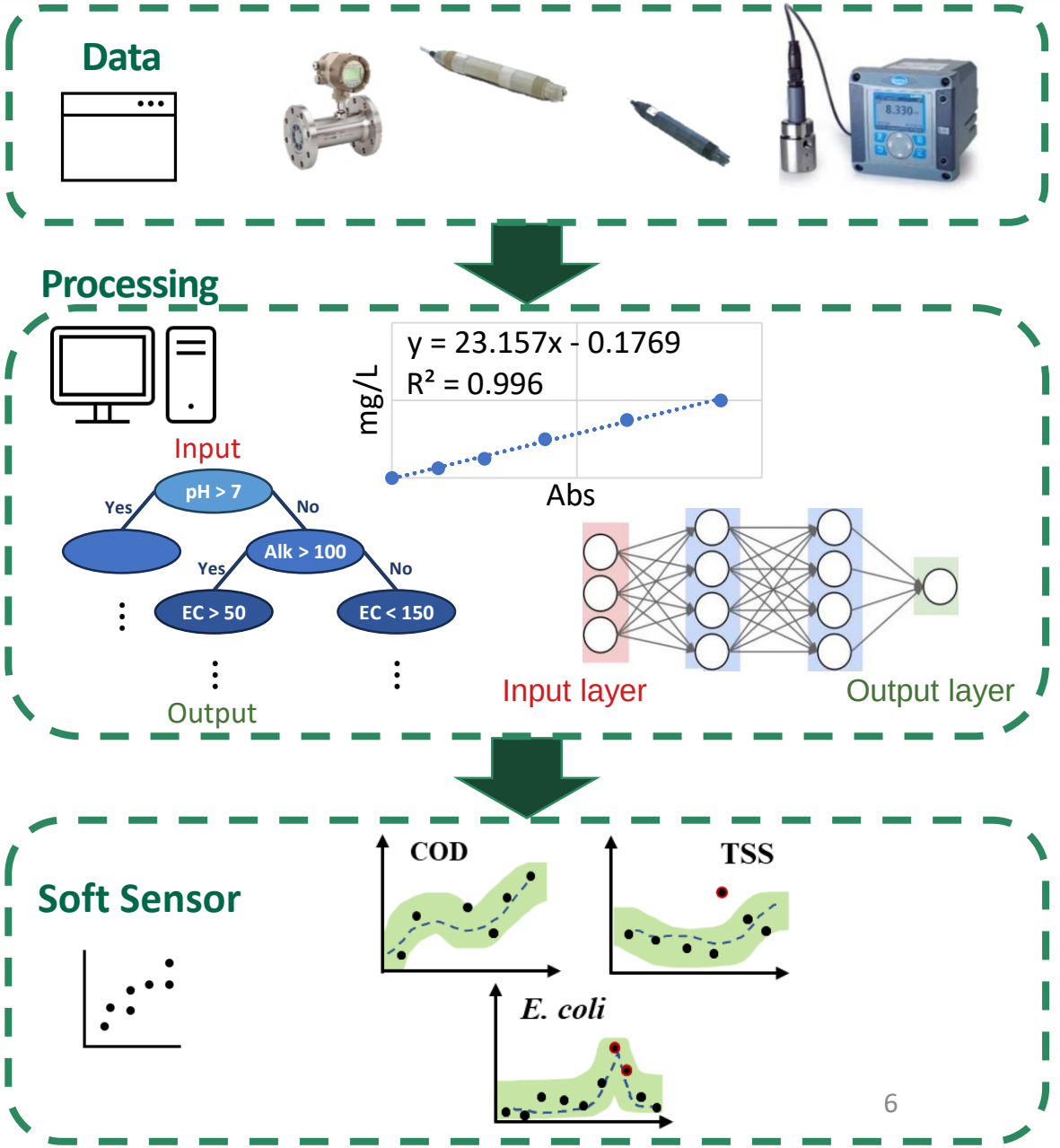
- Real-time monitoring with in-line sensors:
  - pH, EC, turbidity, temperature
  - Ion-selective electrodes, etc.

Data Processing

- Traditional statistical models:
  - Linear, logarithmic, polynomial
- Machine learning models:
  - Decision trees, neural networks, etc.

•Soft Sensor

- Real-time probes data → Trained model → Real-time estimate of hard-to-measure parameters



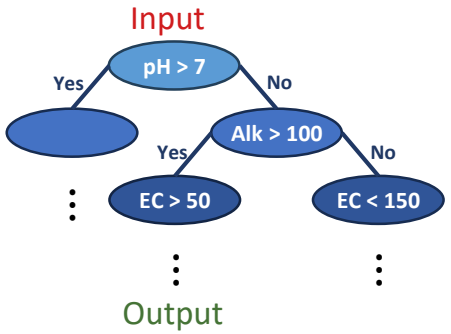
What is Machine Learning?

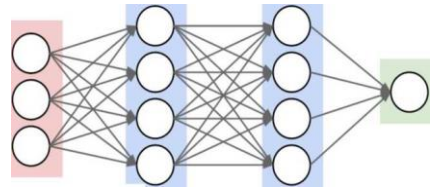
Core concepts

- Data-driven: no fixed equations or manual rules
- Trained on historical input-output relationships
- Predicts complex outcomes in real time
- Effective even with noisy or nonlinear data

Applied in this study

- Algorithms used:
  - Partial Least Squares Regression (PLS)
  - Support Vector Regression (SVR)
  - Cubist Regression (CUB)
  - Quantile Regression Neural Network (QRNN)
- Dataset:
  - 56 weeks of field data





Source of Data: NEWgenerator Field Trial

NEWgenerator system

- Anaerobic membrane bioreactor (AnMBR)
- Decentralized treatment in an informal settlement
- Treated blackwater for non-potable reuse (e.g., toilet flushing)
- Over four years of field deployment in South Africa

Treatment train

- **AnMBR** – Removes organics and suspended solids
- **Nutrient Capture System (NCS)** – Additional COD removal + nutrient recovery
- **Electrochlorinator** – Provides final disinfection

Model Development

Dataset (1.5 yr, n=56 week)

Sampling location: Influent, AnMBR, Permeate, Post-NCS, Effluent
Input: Sampling location, pH, EC, Turb, Color, NH_4^+, NO_3^-, Temp
Output: COD, TSS, *E.coli*

(i) Statistical analysis & data cleaning

- Z-score used to identify outliers
- Standardization of dataset

(ii) Input variable selection

Pearson Correlation		Recursive Feature Elimination		
Strong	**Weak**	**COD**	**TSS**	***E. coli***
Turb	pH	Turb	Turb	Turb
Color	Temp	Samp	Color	Samp
NO_3^-		Color	Samp	Color
		NO_3^-	NO_3^-	EC
		NH_4^+	NH_4^+	NO_3^-
		pH	pH	NH_4^+
		EC	EC	
		Temp		

Selected variables: SL, pH, EC, Turb, Color, NH_4^+, NO_3^-

Dataset (n=52)

Training dataset (70%) | **Testing dataset (30%)**

(iii) Model development (Tuning & Cross-Validation)

PLS | CUB | SVR | QRNN | LM | RF | KNN

(iv) Testing dataset prediction performance analysis

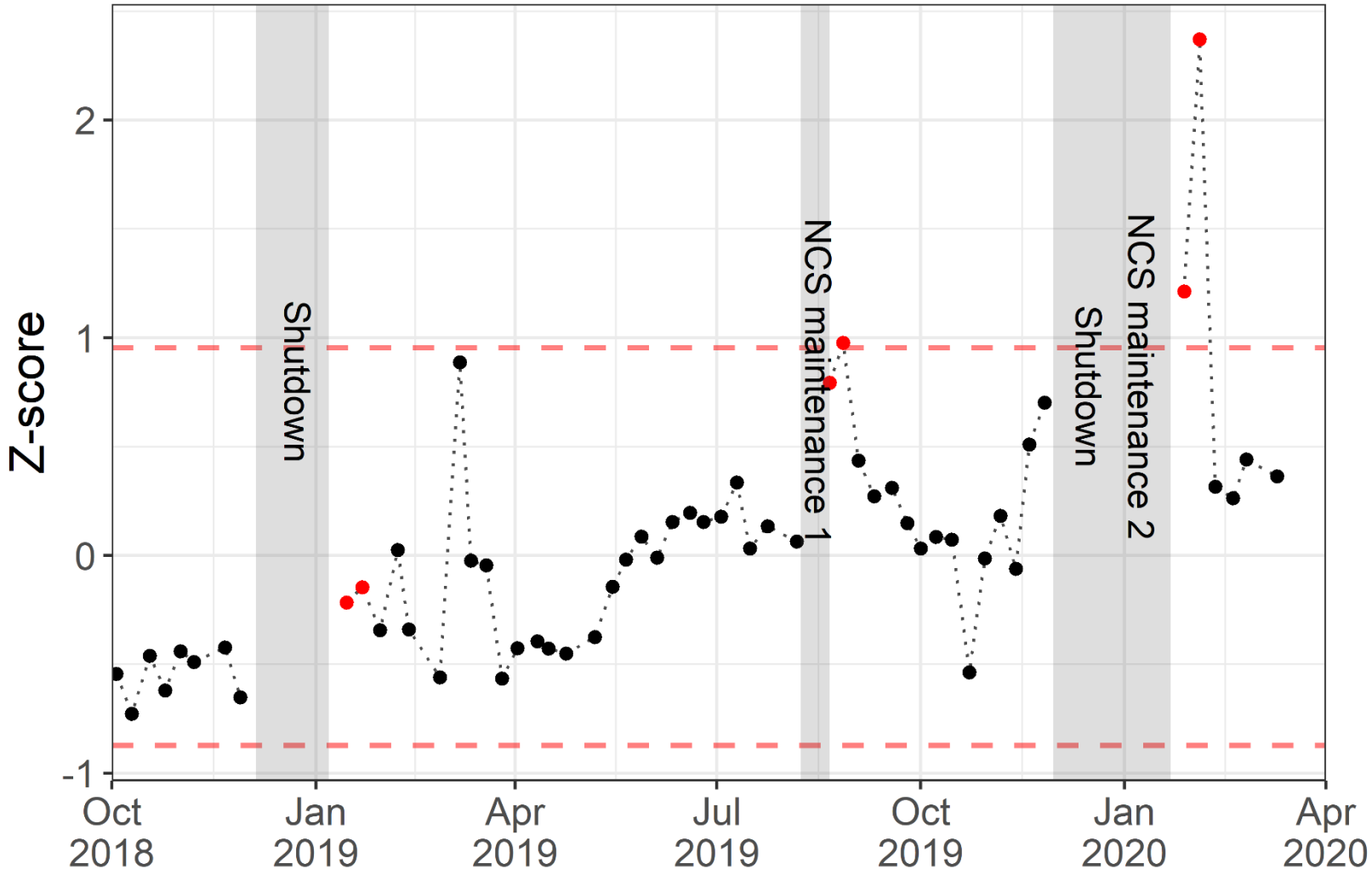
Reject | Accept

(v) Soft sensors are ready to monitor COD, TSS, *E. coli*

Data Pretreatment: Z-Score Filtering

Z-score filtering

- Identify outliers by measuring how far values deviated from the mean
- Three major restart or maintenance events identified
- Operational disruptions negatively impacted data quality
- Two weeks of post-event data excluded to ensure model stability



Data Pretreatment

Pearson Correlation

- Measures the strength and direction of a linear relationship between two variables
- pH and temperature showed weak correlations with output variables
- Turbidity, color, and NO_3^- were strongly correlated with COD and TSS
- Results used to guide variable selection for model development

Result

COD prediction

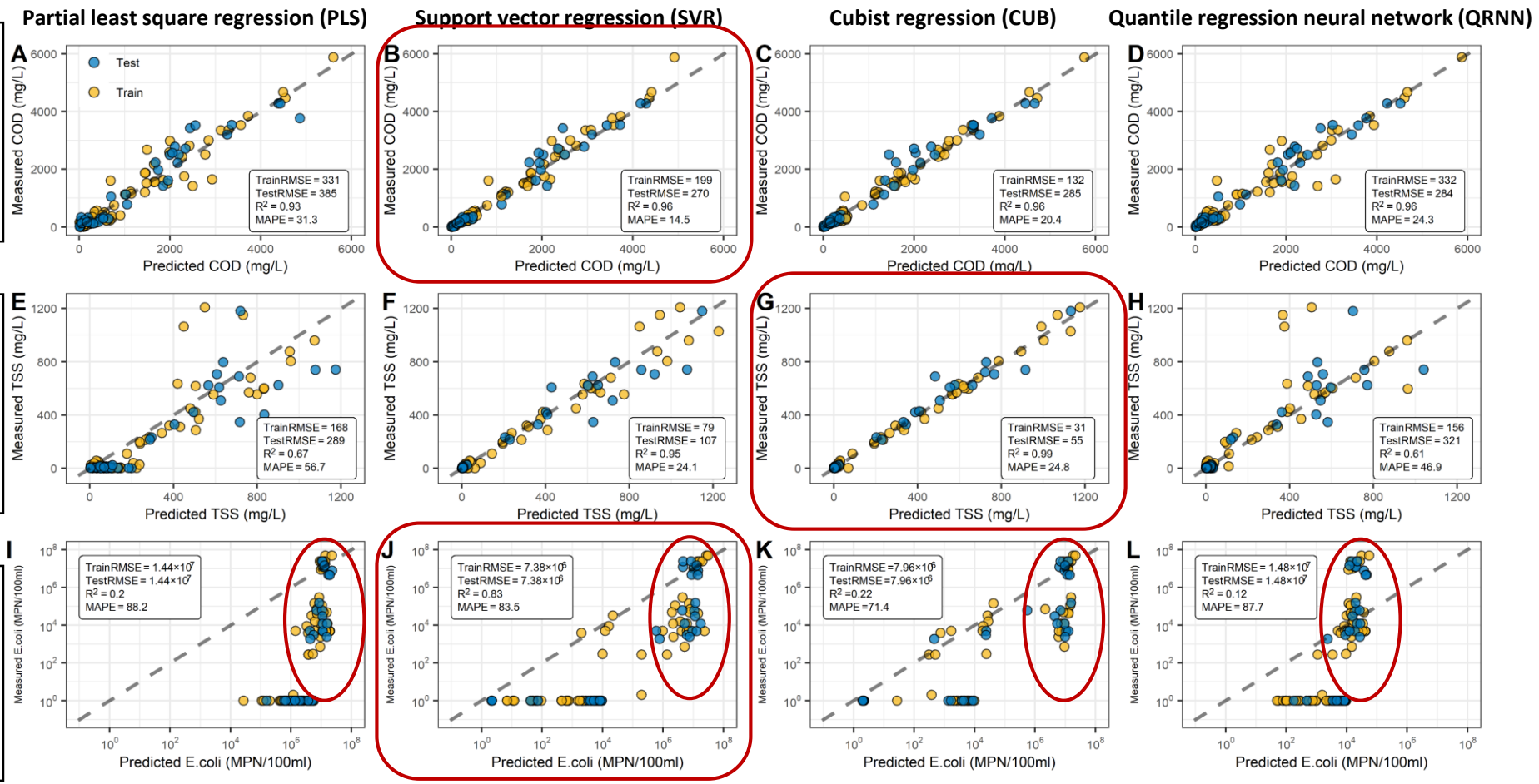
- SVR showed the best
- RMSE: 270; R^2: 0.96
- MAPE: 14.5%

TSS prediction

- CUB was the best
- RMSE: 55; R^2: 0.99
- MAPE: 24.8%

E. coli prediction

- SVR had the lowest RMSE
- But obvious inaccuracies
- Regression model not ideal



E. coli Classification

E. coli classification

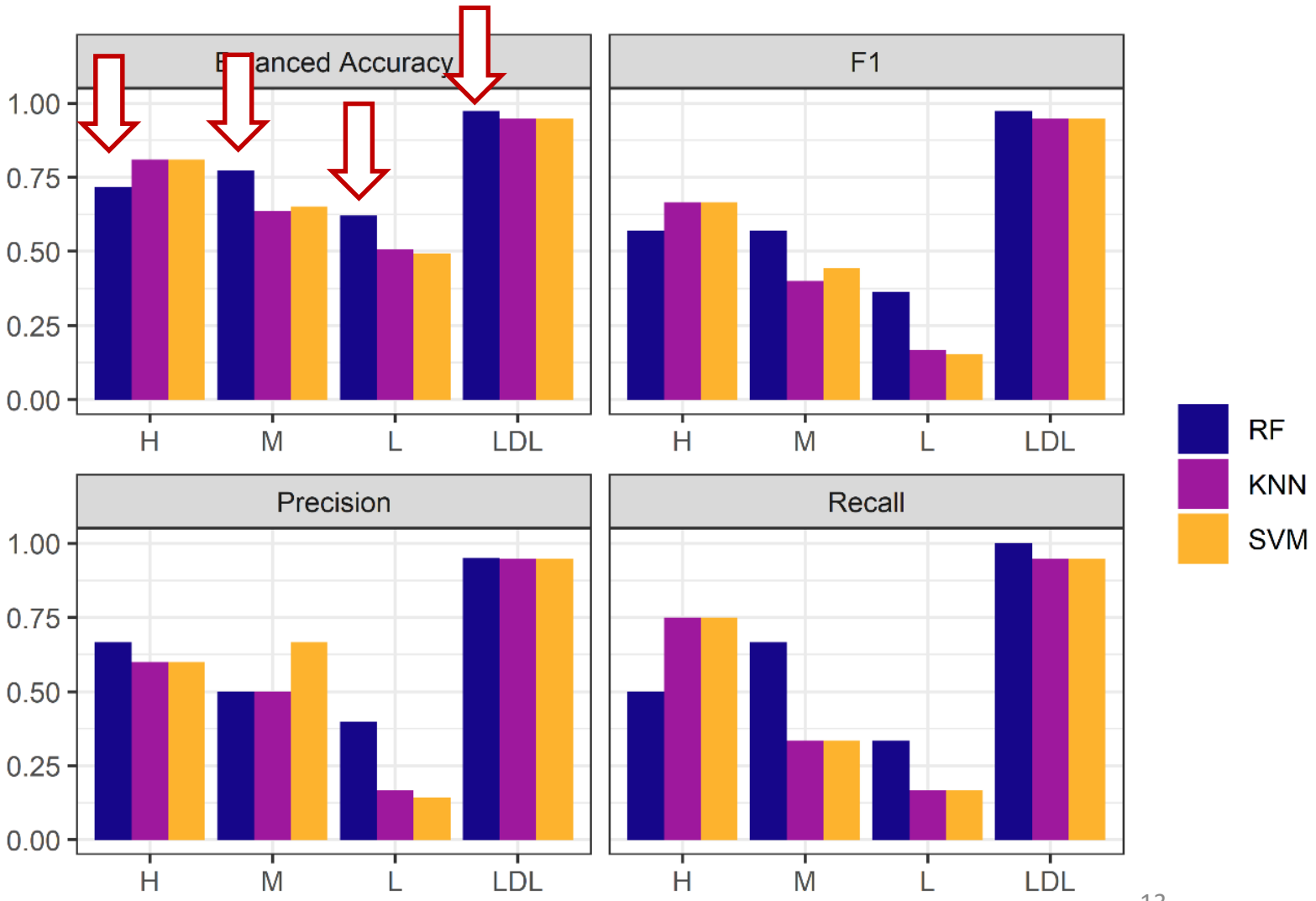
- Broken down into four ranges:
- **High (H)**: > 10^7 MPN/100mL
- **Medium (M)**: ≤ $10^7$and > 10^2 MPN/100mL
- **Low (L)**: ≤ $10^2$and > 1 MPN/100mL
- **LDL:** Lower than the detection limit

Classification algorithms

- Support vector machine (SVM)
- Random forest (RF)
- K-nearest neighbors (KNN)

Results

- RF: Best accuracy of over all 74.36%



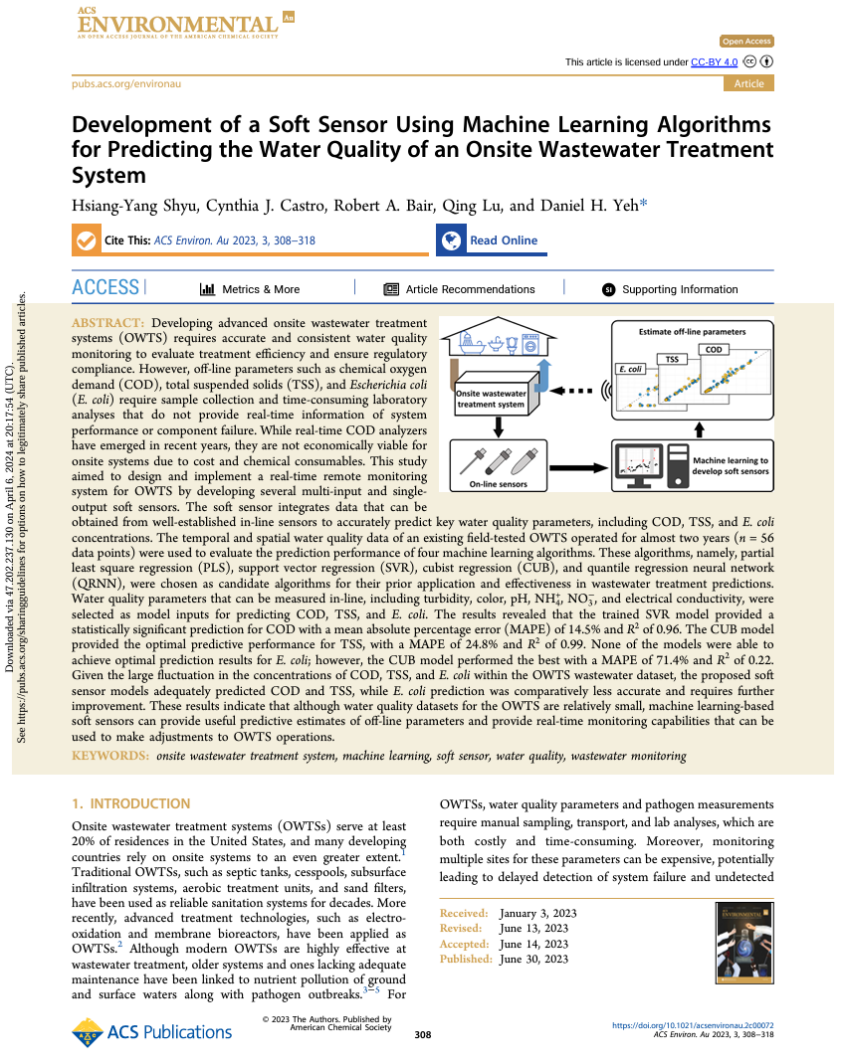
Summary

Summary & Conclusions

- Potential of using in-line sensors and ML to estimate offline parameters (COD, TSS, *E. coli*)
- SVR performed best for COD; CUB was most effective for TSS
- Classification approach showed potential for *E. coli* estimation
  - Further improvements needed
- Limited dataset (56 weeks) still yielded promising performance

Subsection: Future Work

- Expand dataset to improve model robustness
- Estimate pCOD and sCOD
- Add parameters: HRT, flow rate, ORP, DO, DOC, UV254



ACS ENVIRONMENTAL
Open Access
This article is licensed under CC-BY 4.0
pubs.acs.org/environau
Article

Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System

Hsiang-Yang Shyu, Cynthia J. Castro, Robert A. Bair, Qing Lu, and Daniel H. Yeh*

Cite This: ACS Environ. Au 2023, 3, 308–318
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Developing advanced onsite wastewater treatment systems (OWTS) requires accurate and consistent water quality monitoring to evaluate treatment efficiency and ensure regulatory compliance. However, off-line parameters such as chemical oxygen demand (COD), total suspended solids (TSS), and *Escherichia coli* (*E. coli*) require sample collection and time-consuming laboratory analyses that do not provide real-time information of system performance or component failure. While real-time COD analyzers have emerged in recent years, they are not economically viable for onsite systems due to cost and chemical consumables. This study aimed to design and implement a real-time remote monitoring system for OWTS by developing several multi-input and single-output soft sensors. The soft sensor integrates data that can be obtained from well-established in-line sensors to accurately predict key water quality parameters, including COD, TSS, and *E. coli* concentrations. The temporal and spatial water quality data of an existing field-tested OWTS operated for almost two years ($n = 56$ data points) were used to evaluate the prediction performance of four machine learning algorithms. These algorithms, namely, partial least square regression (PLS), support vector regression (SVR), cubist regression (CUB), and quantile regression neural network (QRNN), were chosen as candidate algorithms for their prior application and effectiveness in wastewater treatment predictions. Water quality parameters that can be measured in-line, including turbidity, color, pH, NH_4^+, NO_3^-, and electrical conductivity, were selected as model inputs for predicting COD, TSS, and *E. coli*. The results revealed that the trained SVR model provided a statistically significant prediction for COD with a mean absolute percentage error (MAPE) of 14.5% and R^2 of 0.96. The CUB model provided the optimal predictive performance for TSS, with a MAPE of 24.8% and R^2 of 0.99. None of the models were able to achieve optimal prediction results for *E. coli*; however, the CUB model performed the best with a MAPE of 71.4% and R^2 of 0.22. Given the large fluctuation in the concentrations of COD, TSS, and *E. coli* within the OWTS wastewater dataset, the proposed soft sensor models adequately predicted COD and TSS, while *E. coli* prediction was comparatively less accurate and requires further improvement. These results indicate that although water quality datasets for the OWTS are relatively small, machine learning-based soft sensors can provide useful predictive estimates of off-line parameters and provide real-time monitoring capabilities that can be used to make adjustments to OWTS operations.

KEYWORDS: *onsite wastewater treatment system, machine learning, soft sensor, water quality, wastewater monitoring*

1. INTRODUCTION

Onsite wastewater treatment systems (OWTSs) serve at least 20% of residences in the United States, and many developing countries rely on onsite systems to an even greater extent.[1] Traditional OWTSs, such as septic tanks, cesspools, subsurface infiltration systems, aerobic treatment units, and sand filters, have been used as reliable sanitation systems for decades. More recently, advanced treatment technologies, such as electro-oxidation and membrane bioreactors, have been applied as OWTSs.[2] Although modern OWTSs are highly effective at wastewater treatment, older systems and ones lacking adequate maintenance have been linked to nutrient pollution of ground and surface waters along with pathogen outbreaks.[3−5] For OWTSs, water quality parameters and pathogen measurements require manual sampling, transport, and lab analyses, which are both costly and time-consuming. Moreover, monitoring multiple sites for these parameters can be expensive, potentially leading to delayed detection of system failure and undetected

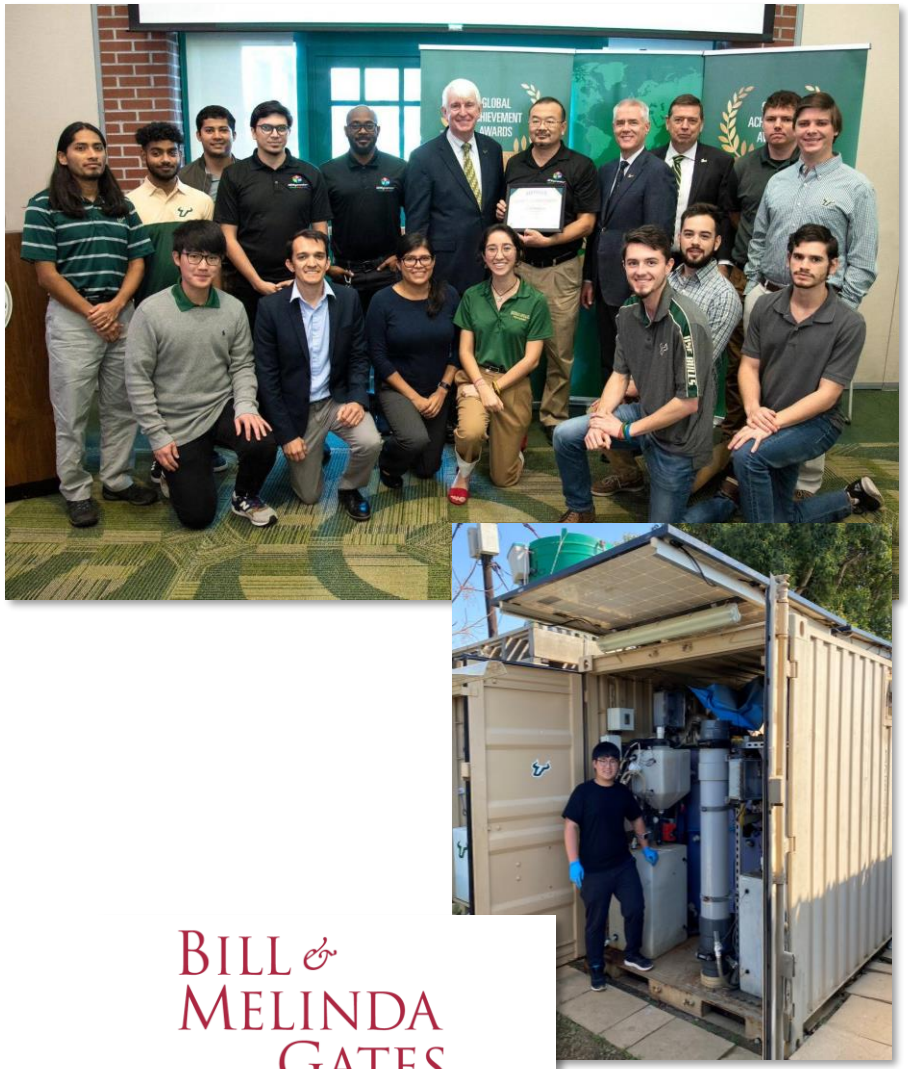
Received: January 3, 2023
Revised: June 13, 2023
Accepted: June 14, 2023
Published: June 30, 2023

https://doi.org/10.1021/acsenvironau.2c00072
ACS Environ. Au 2023, 3, 308−318

Key Takeaways

Why Soft Sensors Matter

- Real-time insights without expensive lab testing
- Transforms common sensor data into actionable predictions
- Supports safe water reuse and system optimization
- Scalable for decentralized and resource-limited systems
- Promising tool for advancing smart sanitation technologies



Thank you!

BILL & MELINDA GATES foundation

Questions
&
Follow up

Speaker

- Hsiang-Yang (Gary), Shyu
- hsiangyang@usf.edu

PI

- Dr. Daniel Yeh
- dhyeh@usf.edu

UNIVERSITY of SOUTH FLORIDA