Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250401012
Volume 4, Issue1 (January-April 2025), PP: 73-78.
www.indjcst.com
ISSN No: 2583-5300

Identifying and Forecasting Wastewater Pollutions Wring IOT & NLP

Dr D J Samatha Naidu[1], P. Sowjanya[2]

[1,2] *Department of MCA, Annamacharya PG College of Computer Studies, Rajampet Andhra Pradesh, India.*

To Cite this Article: *Dr D J Samatha Naidu[1], P. Sowjanya[2], "Identifying and Forecasting Wastewater Pollutions Wring IOT & NLP", Indian Journal of Computer Science and Technology, Volume 04, Issue 01 (January-April 2025), PP: 73-78.*

Abstract: *The detection of contaminants in several environments (e.g., air, water, sewage systems) is of paramount importance to protect people and predict possible dangerous circumstances. Most of existing works do this using classical Machine Learning tools that act on the acquired measurement data. The main disadvantage of the Existing work approach is that it relies on knowing the injection time, i.e., the instant in time when the contaminant is injected into the wastewater. The proposed work introduces two main elements: a low-cost platform to acquire, pre-process, and transmit data to classify contaminants in wastewater; and a novel classification approach to classify contaminants in wastewater, based on deep learning and the transformation of raw sensor data into natural language metadata. The proposed solution presents clear advantages against state-of-the-art systems in terms of higher effectiveness and reasonable efficiency. For this reason, the developed system also includes a finite state machine tool able to infer the exact time instant when the substance is injected. The entire system is presented and discussed in detail. Furthermore, several variants of the proposed processing technique are also presented to assess the sensitivity to the number of used samples and the corresponding promptness/computational burden of the system. The lowest accuracy obtained by our technique is 91.4%, which is significantly higher than the 81.0% accuracy reached by the best baseline method.*

.

Keywords: *data acquisition, cloud based analytics, Machine Learning models*

1. ***Sensors:*** *Measure parameters like pH, turbidity, BOD, COD, heavy metals, temperature, and dissolved oxygen.*
2. ***Data Transmission:*** *Wireless networks (e.g., LoRaWAN, NB-IoT) for real-time data transfer.*
3. ***Edge Computing:*** *Process data locally to reduce latency.*

I.INTRODUCTION

Wastewater pollution is a critical environmental issue that affects ecosystems, public health, and economic development worldwide. With the growing industrialization, urbanization, and population density, managing wastewater pollution has become a complex and urgent challenge. Traditional methods of monitoring wastewater quality often rely on manual sampling and laboratory analysis, which can be time-consuming, costly, and may lack real-time insights.To address these challenges, the integration of advanced technologies such as Internet of Things (IoT) and Natural Language Processing (NLP) presents an innovative approach to monitor, identify, and forecast wastewater pollution.IoT for Wastewater Pollution DetectionThe Internet of Things (IoT) refers to a network of interconnected devices that can collect, exchange, and process data in real-time. In the context of wastewater management, IoT sensors can be deployed to monitor various parameters such as temperature, pH levels, dissolved oxygen, turbidity, and the presence of harmful chemicals or pollutants

II. MATERIAL AND METHODS

1 Material

1.1 Data Collection

Limited by effectiveness and lack of integration with advanced technologies.

- Focuses on environments such as air, water, and sewage systems.
- Does not fully exploit raw sensor data or advanced data transformation methods.
- Lower accuracy compared to modern techniques

Data Preprocessing:

The raw data from the IoT sensors is often noisy or incomplete. Preprocessing steps may involve:
Data cleaning: Removal of noise, handling of missing values, and outlier detection.
Normalization or scaling: Standardizing values to a particular range or unit of measurement.
Time-series analysis: Identifying patterns over time in the sensor data.

3. Machine Learning for Identification and Forecasting: Algorithms are used to analyze sensor data to identify current levels of pollution and forecast future pollution trends. Common techniques included.

- **Supervised learning algorithms:** Such as decision trees, random forests, support vector machines (SVM), and neural networks for classification and regression tasks.

- **Unsupervised learning algorithms:** Used for anomaly detection, clustering pollution data to detect unusual patterns or predict emerging pollution levels (e.g., K-means clustering or DBSCAN).

Time-series forecasting models: Such as ARIMA (AutoRegressive Integrated Moving Average), LSTM (Long Short- Term Memory) networks, and Prophet, to predict future pollution trends based on historical data1q

1.1. Tools and Libraries

Programming Languages: python R

1. **Machine Learning Libraries**:

- **Scikit-learn**: A well-liked package for putting fundamental machine learning methods like clustering, regression, and classification into practice.
- **XG Boost / Light GBM**: Libraries for gradient boosting are used to solve classification and regression issues.
- **TensorFlow / Keras**: Libraries for methods based on deep learning.
- **Pandas**: to manipulate and preprocess data.
- **NumPy / SciPy**: For calculations involving numbers.
- **Matplotlib / Seaborn:** for the visualization of data.

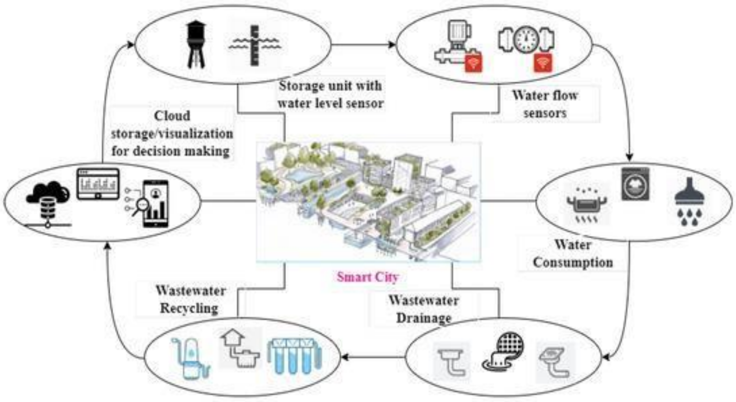


Figure 2.2.1: system architecture

1.3 Hardware Requirements

Processor	:	I3 or higher
Speed	:	2.9 GHz
RAM	:	4 GB (min)
Hard Disk	:	160 GB

2. Methods

2.1. Data Preprocessing

Prior to using machine learning methods, data pretreatment is an essential step. Among the actions involved are:

- **Data Cleaning**: addressing missing data, eliminating anomalies, and fixing dataset mistakes.
- **Feature Engineering**: Making the raw data more helpful for machine learning models by adding extra features.
- **Normalization/Standardization**: Features are scaled to prevent any one feature from taking over the model because of its size.
- **Categorical Data Encoding**: Converting soil type and other category features.

2.2. Feature Selection

- **Correlation Matrix**: removing strongly connected aspects by analyzing feature correlations.
- **Principal Component Analysis (PCA)**: A method for keeping as much variance in the data as feasible while reducing its dimensionality.
- **Random Forest Feature Importance**: assessing each feature's significance in forecasting water quality suitability using tree-based models.

2.3. Model Selection

- **Decision Trees:** used to categorize crops according to a variety of input characteristics, such as climate and soil type.
- **Random Forest**: A decision tree ensemble approach that works well for problems involving regression and classification, particularly when working with structured data such as soil and farm data.
- **Support Vector Machines (SVM)**: can be applied to classification problems, particularly when there are distinct boundaries between the various classes in the data.
- **K-Nearest Neighbors (KNN)**: classification based on a data point's "nearness" to other like points using a non-parametric approach.

2.4. Model Training

Splitting the Data: Make training and testing sets out of the dataset (e.g., 80% training, 20% testing).
Cross-Validation: Use cross-validation (such as k-fold) to make sure the model is reliable and works well with unknown data.
Hyperparameter Tuning: For best results, adjust hyperparameters using strategies like Random Search.

2.5. Types of Software developing life cycles (SDLC): V- Model

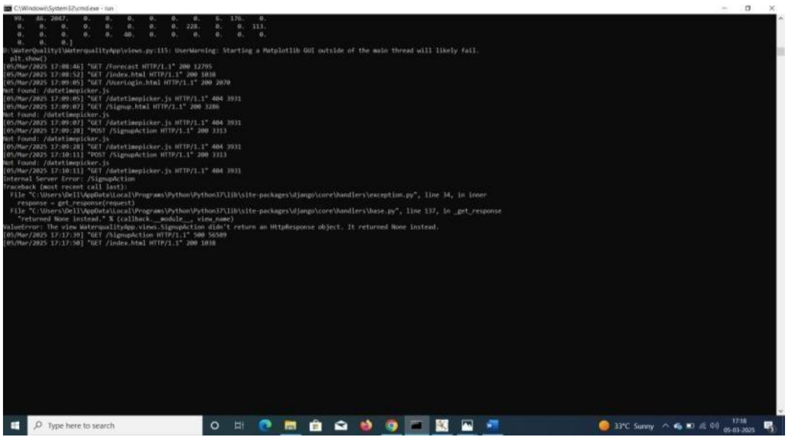
The V-Model (Verification and Validation Model) is a software development methodology that follows a sequential process, similar to the Waterfall Model, but with an emphasis on testing at each stage. It is called the V-Model because the development and testing phases are arranged in a V-shape, where each development phase has a corresponding testing phase. The process consists of Requirement Analysis, System Design, Architectural Design, Module Design, Implementation, Unit Testing, Integration Testing, System Testing, and Acceptance Testing. Each phase must be completed before moving to the next, ensuring a structured approach with early defect detection. The V-Model is best suited for projects with well-defined requirements, as changes in later stages can be costly. While it enhances software quality through rigorous validation, it lacks flexibility for projects with evolving needs.

Module Design

The module design phase focuses on the detailed design of each software component. The system is broken down into smaller, manageable modules, and each module's internal structure, logic, and data flow are defined. This phase ensures that individual components are well-planned, making the coding of phase smoother.

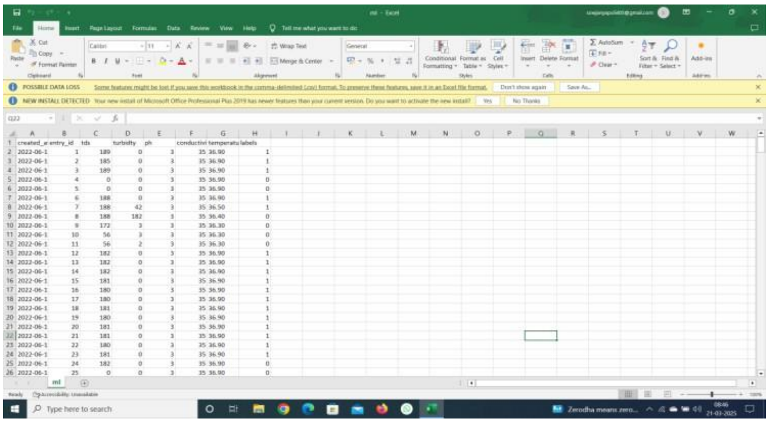
III.RESULT

Data cleaning and processing, missing value analysis, exploratory analysis, and model creation and evaluation were all part of the analytical process. The best accuracy on a public test set will be discovered, as will the highest accuracy score. This application can assist in determining the current state of water quality. The conductivity acts as a sensor gateway. The sensor input are sent to the pi4, a edge level processor (personal computer) where in the K Means, a machine learning algorithm is used for predicting the quality of water. The predicted water quality data are stored in Cloud server for future access. The predicted data is sent to the water controller unit for further action.



```
   99.   46. 2047.    0.    0.    0.    0.    0.    0.    6.  176.    0.
    0.    0.    0.    0.    0.    0.    0.    0.  228.    0.    0.  111.
    0.    0.    0.    0.    0.   40.    0.    0.    0.    0.    0.    0.
    0.    0.    0.]
D:\WaterQuality1\WaterqualityApp\views.py:115: UserWarning: Starting a Matplotlib GUI outside of the main thread will likely fail.
  plt.show()
[05/Mar/2025 17:08:46] "GET /Forecast HTTP/1.1" 200 12795
[05/Mar/2025 17:08:52] "GET /index.html HTTP/1.1" 200 1038
[05/Mar/2025 17:09:05] "GET /UserLogin.html HTTP/1.1" 200 2070
Not Found: /datetimepicker.js
[05/Mar/2025 17:09:05] "GET /datetimepicker.js HTTP/1.1" 404 3931
[05/Mar/2025 17:09:07] "GET /Signup.html HTTP/1.1" 200 3286
Not Found: /datetimepicker.js
[05/Mar/2025 17:09:07] "GET /datetimepicker.js HTTP/1.1" 404 3931
[05/Mar/2025 17:09:28] "POST /SignupAction HTTP/1.1" 200 3313
Not Found: /datetimepicker.js
[05/Mar/2025 17:09:28] "GET /datetimepicker.js HTTP/1.1" 404 3931
[05/Mar/2025 17:10:11] "POST /SignupAction HTTP/1.1" 200 3313
Not Found: /datetimepicker.js
[05/Mar/2025 17:10:11] "GET /datetimepicker.js HTTP/1.1" 404 3931
Internal Server Error: /SignupAction
Traceback (most recent call last):
  File "C:\Users\Dell\AppData\Local\Programs\Python\Python37\lib\site-packages\django\core\handlers\exception.py", line 34, in inner
    response = get_response(request)
  File "C:\Users\Dell\AppData\Local\Programs\Python\Python37\lib\site-packages\django\core\handlers\base.py", line 137, in _get_response
    "returned None instead." % (callback.__module__, view_name)
ValueError: The view WaterqualityApp.views.SignupAction didn't return an HttpResponse object. It returned None instead.
[05/Mar/2025 17:17:39] "GET /SignupAction HTTP/1.1" 500 56509
[05/Mar/2025 17:17:50] "GET /index.html HTTP/1.1" 200 1038
```

Screen1: CMD Running Process

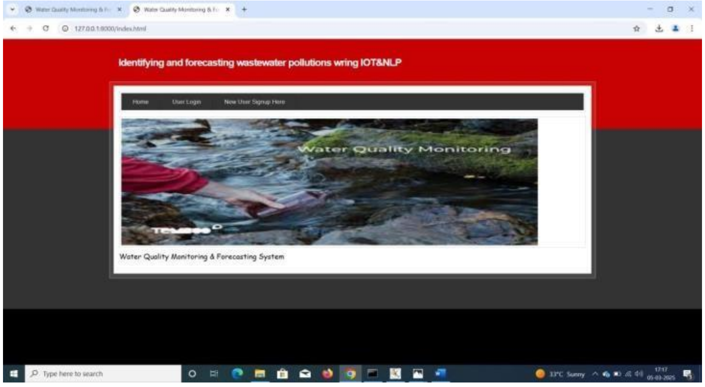


POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

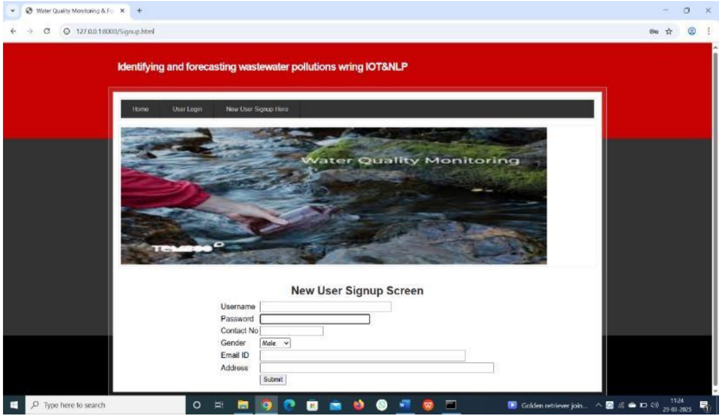
NEW INSTALL DETECTED Your new install of Microsoft Office Professional Plus 2019 has newer features than your current version. Do you want to activate the new install?

	A	B	C	D	E	F	G	H
1	created_a	entry_id	tds	turbidty	ph	conductivi	temperatu	labels
2	2022-06-1	1	189	0	3	35	36.90	1
3	2022-06-1	2	185	0	3	35	36.90	1
4	2022-06-1	3	189	0	3	35	36.90	1
5	2022-06-1	4	0	0	3	35	36.90	0
6	2022-06-1	5	0	0	3	35	36.90	0
7	2022-06-1	6	188	0	3	35	36.90	1
8	2022-06-1	7	188	42	3	35	36.50	1
9	2022-06-1	8	188	182	3	35	36.40	0
10	2022-06-1	9	172	3	3	35	36.30	0
11	2022-06-1	10	56	3	3	35	36.30	0
12	2022-06-1	11	56	2	3	35	36.30	0
13	2022-06-1	12	182	0	3	35	36.90	1
14	2022-06-1	13	182	0	3	35	36.90	1
15	2022-06-1	14	182	0	3	35	36.90	1
16	2022-06-1	15	181	0	3	35	36.90	1
17	2022-06-1	16	180	0	3	35	36.90	1
18	2022-06-1	17	180	0	3	35	36.90	1
19	2022-06-1	18	181	0	3	35	36.90	1
20	2022-06-1	19	180	0	3	35	36.90	1
21	2022-06-1	20	181	0	3	35	36.90	1
22	2022-06-1	21	181	0	3	35	36.90	1
23	2022-06-1	22	180	0	3	35	36.90	1
24	2022-06-1	23	181	0	3	35	36.90	1
25	2022-06-1	24	182	0	3	35	36.90	0
26	2022-06-1	25	0	0	3	35	36.90	0

Screen 2: Showing Data Set


127.0.0.1:8000/index.html
Identifying and forecasting wastewater pollutions wring IOT&NLP
Home
User Login
New User Signup Here
Water Quality Monitoring
Water Quality Monitoring & Forecasting System

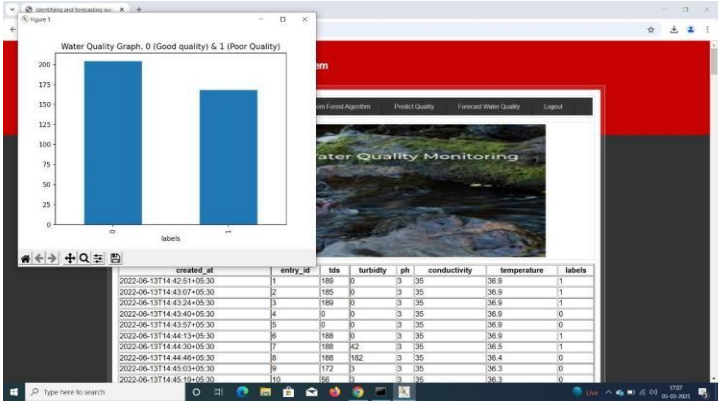
Screen 3: Show The Home Page


127.0.0.1:8000/Signup.html
Identifying and forecasting wastewater pollutions wring IOT&NLP
Home
User Login
New User Signup Here
Water Quality Monitoring
Submit

New User Signup Screen

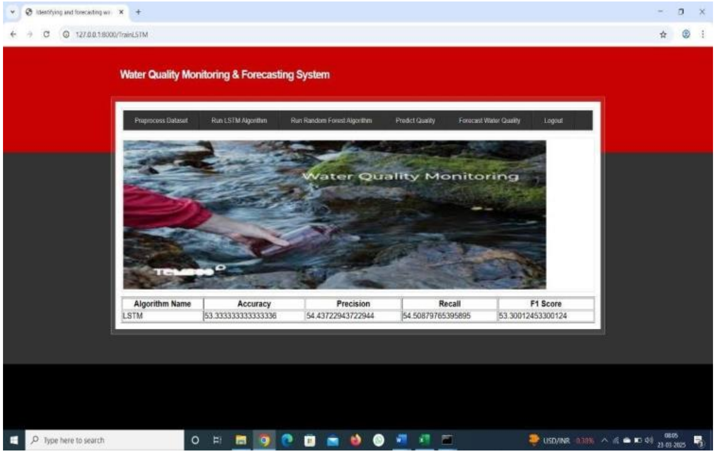
Username
Password
Contact No
Gender Male
Email ID
Address

Screen 5: Fill Detail For user sign up



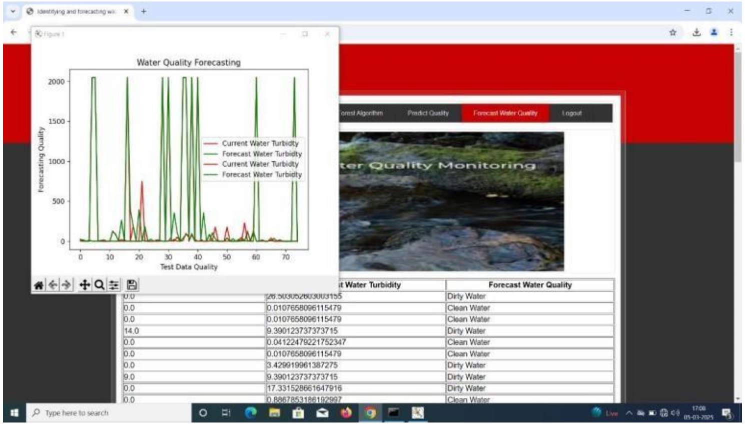
created_at	entry_id	tds	turbidty	ph	conductivity	temperature	labels
2022-06-13T14:42:51+05:30	1	189	0	3	35	36.9	1
2022-06-13T14:43:07+05:30	2	185	0	3	35	36.9	1
2022-06-13T14:43:24+05:30	3	189	0	3	35	36.9	1
2022-06-13T14:43:40+05:30	4	0	0	3	35	36.9	0
2022-06-13T14:43:57+05:30	5	0	0	3	35	36.9	0
2022-06-13T14:44:13+05:30	6	188	0	3	35	36.9	1
2022-06-13T14:44:30+05:30	7	188	42	3	35	36.5	1
2022-06-13T14:44:46+05:30	8	188	182	3	35	36.4	0
2022-06-13T14:45:03+05:30	9	172	3	3	35	36.3	0
2022-06-13T14:45:19+05:30	10	56	3	3	35	36.3	0

Screen 6: wastewater quality checking



Algorithm Name	Accuracy	Precision	Recall	F1 Score
LSTM	53.333333333333336	54.43722943722944	54.50879765395895	53.30012453300124

Screen 7: accuracy of LSTM Algorithm



	t Water Turbidity	Forecast Water Quality
0.0	26.503052603003155	Dirty Water
0.0	0.0107658096115479	Clean Water
0.0	0.0107658096115479	Clean Water
14.0	9.390123737373715	Dirty Water
0.0	0.04122479221752347	Clean Water
0.0	0.0107658096115479	Clean Water
0.0	3.429919961387275	Dirty Water
9.0	9.390123737373715	Dirty Water
0.0	17.331528661647916	Dirty Water
0.0	0.8867853186192997	Clean Water

Screen 8: water quality forecasting

IV. DISCUSSION

Wastewater pollution is a growing environmental concern, impacting aquatic ecosystems, public health, and water resources. Traditional monitoring systems often rely on manual sampling and laboratory testing, which are time-consuming and may not provide real-time insights. To address these challenges, the integration of IoT (Internet of Things) and NLP (Natural Language Processing) has emerged as a transformative approach. This combination facilitates real-time monitoring, predictive analytics, and automated reporting, enabling proactive environmental management.

V. CONCLUSION

In this paper, the study's findings show that machine learning models, the integration of Internet of Things (IoT) and Natural Language Processing (NLP) presents a promising and innovative approach to identifying and forecasting wastewater pollution. By leveraging IoT, real-time monitoring of water quality parameters such as pH levels, temperature, dissolved oxygen, turbidity, and chemical composition can be continuously tracked and analyzed. IoT sensors placed at strategic locations in wastewater systems provide accurate and timely data, enabling better detection of contamination and pollution levels.On the other hand, Natural Language Processing (NLP) can be utilized to process and analyze large volumes of textual data, including reports, scientific articles, and regulatory documents, to extract valuable insights and trends related to wastewater quality. NLP techniques help in detecting patterns and emerging concerns from unstructured data, which could be missed by traditional methods. The collaboration of IoT for data collection and NLP for data analysis leads to a more holistic and efficient solution for managing and mitigating wastewater pollution. This approach can significantly contribute to environmental protection, public health, and the sustainability of water resources by providing actionable insights that help in the timely detection, monitoring, and prevention of wastewater contamination.

References

1. *L. T. Lee and E. R. Blatchley, ''Long-term monitoring of water and air quality at an indoor pool facility during modifications of water treatment,'' Water, vol. 14, no. 3, p. 335, Jan. 2022. [Online]. Available: https://www.mdpi.com/2073-4441/14/3/335*
2. *H. Chojer, P. T. B. S. Branco, F. G. Martins, M. C. M. Alvim-Ferraz, and S. I. V. Sousa, ''Development of low-cost indoor air quality monitoring devices: Recent advancements,'' Sci. Total Environ., vol. 727, Jul. 2020,Art. no. 138385. [Online]. Available: https://www.sciencedirect.com/ science/article/pii/S0048969720318982*
3. *K. Farkas, L. S. Hillary, S. K. Malham, J. E. McDonald, and D. L. Jones, ''Wastewater and public health: The potential of wastewater surveillance for monitoring COVID-19,'' Current Opinion Environ. Sci. Health, vol. 17,pp. 14–20, Oct. 2020.*
4. *Trubetskaya, W. Horan, P. Conheady, K. Stockil, S. Merritt, and S. Moore, ''A methodology for assessing and monitoring risk in the industrial wastewater sector,'' Water Resour. Ind., vol. 25, Jun. 2021,Art. no. 100146.*

Authors Profile:

Dr D J Samatha Naidu, completed MCA from S V University, Tirupati , MPhil computer science from Madurai Kamaraj University Madurai, MTech in Computer Science and Engineering in JNTUA, Anantapur, PhD in Computer Science from Vikrama Simha Puri University, Nellore, currently working as Professor and Principal Annamacharya PG College of computer studies, Rajampet since 20 years, 2 years industrial experience as network support engineer, 12 years research experience, Completed consultancy and major projects like AICTE and other IT industry.150 international Research journal papers published,100 national and international conferences are attended and presented papers.10 National And International Design Grant Patents, Utility Patents, Copy Rights, Patents Are Published. 12 Text Books are published.8 Theory and Lab Manuals are designed for MCA and MBA students, 22 national and international professional bodies Life member, associate member, fellow member for

Edunix research university USA, ISTE, IE, IACSIT, IAENG, IMRF, IRDP, NITTE, GLOBAL PROFESSOR FOR ALUMINI ASSOCIATION, HRPC, UAE, EAI, KALA'S LIFE MEMBERSHIP, COUNCIL OF TEACHERS EDUCATION MEMBER, GLOBOL TEACHERS ASSOCIATE MEMBER,INSTITUTE OF GREEN ENGINEERS. Research papers are Reviewed as Editorial Member and Reviewer Member, 25 National and international awards are received from USA, MALAYSIA, Andhra Pradesh, Telangana, Tamilnadu state organizations, I received prestigious university best teacher award received from JNTUA Anantapur for 2022. Am very much honour getting second time award as university principal award for 2024

P.Sowjanya, Earned her Bachelor's degree in Computer Science (B.Sc.) from Sri Vaishnavi degree college,rajampet Y.V.U University in 2022, where she developed a strong foundation in programming, data structures, and software development. Currently, she is pursuing a Master of Computer Applications (MCA) at Annamacharya PG College of Computer Studies,rajampet. further enhancing her technical expertise and research skills. This publication marks her first contribution to the academic community, reflecting her keen interest in the intersection of technology and healthcare. Her primary research focuses on the identifying and forecasting wastewater pollutions wring IOT&NLP.