IJARR, 5(6), 2019; 113-124

International Journal of Advanced Research and Review

www.ijarr.in

DETECTION OF WASTEWATER POLLUTION THROUGH NATURAL LANGUAGE GENERATION WITH LOW-COST SENSING PLATFORM

Venumadhava M , Usha. G , Manjunatha K H

Asst. Professor, Asst. Professor, Asst. Professor
venudsp@gmail.com , ushagonal645@gmail.com , manjukh09@gmail.com
Department of ECE, Proudhadevaraya Institute of Technology, Abheraj Baldota Rd, Indiranagar, Hosapete, Karnataka-583225

ABSTRACT

In order to safeguard individuals and anticipate potentially harmful situations, it is crucial to detect pollutants in various settings, such as air, water, and sewage systems. The majority of these studies use traditional Machine Learning techniques to process the collected measurement data. First, a novel classification approach to classify contaminants in wastewater based on deep learning and the transformation of raw sensor data into natural language metadata; second, a low-cost platform to acquire, pre-process, and transmit data for this purpose. Superior efficacy and tolerable efficiency distinguish the suggested solution from state-of-the-art methods. Knowing the injection time—the precise moment when the pollutant is introduced into the wastewater—is crucial to the suggested method, which is its primary drawback. Therefore, a finite state machine tool capable of inferring the precise moment of injection is also included into the proposed system. There is an extensive presentation and discussion of the complete system. In addition, we provide many versions of the suggested processing method to evaluate how the system responds to changes in the amount of samples and the accompanying computing load and speed. While the best baseline approach only managed an accuracy of 81.0%, our strategy achieved a minimum of 91.4%.

INTRODUCTION

The task of accurate environmental monitoring is a pressing worldwide issue which is bound to become increasingly more important in the near future. There are many aspects that should be kept under control and concern the quality of the air, soil, and water [1, 2]. In fact, their continuous monitoring would allow targeted and timely actions aimed at restoring optimal conditions following dangerous events such as the appearance of pollutants. In this context, monitoring wastewater (WW) is particularly important [3]. WW is the water that has already been used for some purpose (civil or industrial uses) and must be subjected to purification before being returned to the natural cycle. To function at their best and effectively, the purification systems must know a priori the type of substances mixed with the water. It follows that a purification system for water for industrial use will be different from a purification plant for water for civil use. Hence, there is a strong need for protocols to promptly detect incompatible substances, to guarantee the correct and effective operation of purification plants[4]. Currently, this is solved by organizing periodic monitoring activities at particular points of the water path, which are carried out by the control institutes in charge using specialized laboratory instruments. Although this is an effective method, the quality of the water between two consecutive checks is unknown, and the checks may be not frequent enough to promptly identify problems.

The ideal solution would combine automated continuous and distributed early warning monitoring, alongside periodic manual checks carried out by the control institutes. To solve the problems of cost and installation of a distributed and continuous monitoring system, it is necessary to resort to low-cost and IoT-ready systems [5], which are able not only to collect environmental data but also to process them relying on centralized data collection and elaboration points. In this context, the data collected from the sensors need to be processed by an algorithm that is used to analyze and forecast the presence (or absence) of polluting substances in the WW. Current state-of-the-art systems for this task rely on machine learning algorithms such as decision trees [6, 7]. In this paper, we propose a novel system based on deep learning, and in particular on causal generative models developed for natural language tasks, for the detection and classification of pollutants in WW, starting from the data collected by a multisensory system based on SENSIPLUS (Sensichips srl, Pisa, Italy) [8]. Note that the present paper does not present the infrastructure necessary for data transport as any solution based, for example, on MQTT or message queuing protocols could be used for this purpose. The effectiveness of the proposed classifier is tested against a set of state-of-the-art baselines on a dataset created in collaboration with Sensichips s.r.l. and made available to the scientific community [9]. Results show that the proposed methodology outperforms the baseline methods and its effectiveness allows for practical usage of the developed methodology.

LITERATURE REVIEW

Long-Term Monitoring of Water and Air Quality at an Indoor Pool Facility during Modifications

of Water Treatment

Previous research has shown that volatile disinfection byproducts (DBPs) can adversely affect the human respiratory system. As a result, swimming pool water treatment processes can play important roles in governing water and air quality. Thus, it was hypothesized that water and air quality in a swimming pool facility can be improved by renewing or enhancing one or more components of water treatment. This study is designed to identify and quantify changes in water and air quality that are associated with changes in water treatment at an indoor chlorinated swimming pool facility. Reductions in aqueous trichloramine (NCl_3) concentration were observed following the use of secondary oxidizer with its activator. This inclusion also resulted in significant decreases in the concentrations of cyanogen chloride (CNCl) in pool water. The concentration of urea, a compound that is common in swimming pools and that functions as an important precursor to NCl_3 formation, as well as a marker compound for the introduction of contaminants by swimmers, was also reduced after the addition of the activator. Concentrations of gas-phase NCl_3 did not decrease after the treatment processes were changed. The collection of long-term water and air quality measurements also allowed for an assessment of the effects of bather load on water and air quality. In general, the concentrations of urea (an NCl_3 precursor), liquid-phase NCl_3, and gas-phase NCl_3 all increased during periods of high swimmer number.

sand filter media to activated filter media, which was monitored for roughly four weeks; the

Development of low-cost indoor air quality monitoring devices: Recent advancements

The use of low-cost sensor technology to monitor air pollution has made remarkable strides in the last decade. The development of low-cost devices to monitor air quality in indoor environments can be used to understand the behaviour of indoor air pollutants and potentially impact on the reduction of related health impacts. These user-friendly devices are portable, require low-maintenance, and can enable near real-time, continuous monitoring. They can also contribute to citizen science projects and community-driven science. However, low-cost sensors have often been associated with design compromises that hamper data reliability. Moreover, with the rapidly increasing number of studies, projects, and grey literature based on low-cost sensors, information got scattered.

Intending to identify and review scientifically validated literature on this topic, this study critically summarizes the recent research pertinent to the development of indoor air quality monitoring devices using low-cost sensors. The method employed for this review was a thorough search of three scientific databases, namely: ScienceDirect, IEEE, and Scopus. A total of 891 titles published since 2012 were found and scanned for relevance. Finally, 41 research articles consisting of 35 unique device development projects were reviewed with a particular emphasis on device development: calibration and performance of sensors, the processor used, data storage and communication, and the availability of real-time remote access of sensor data. The most prominent finding of the study showed a lack of studies consisting of sensor performance as only 16 out of 35 projects performed calibration/validation of sensors. An even fewer number of studies conducted these tests with a reference instrument. Hence, a need for more studies with calibration, credible validation, and standardization of sensor performance and assessment is recommended for subsequent research.

Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19

Pathogenic viruses represent one of the greatest threats to human well-being. As evidenced by the COVID-19 global pandemic, however, halting the spread of highly contagious diseases is notoriously difficult. Successful control strategies therefore have to rely on effective surveillance. Here, we describe how monitoring wastewater from urban areas can be used to detect the arrival and subsequent decline of pathogens, such as SARS-CoV-2. As the amount of virus shed in faeces and urine varies largely from person to person, it is very difficult to quantitatively determine the number of people who are infected in the population. More research on the surveillance of viruses in wastewater using accurate and validated methods, as well as subsequent risk analysis and modelling is paramount in understanding the dynamics of viral outbreaks.

A methodology for assessing and monitoring risk in the industrial wastewater sector

The concept of sustainable risk assessment in industrial wastewater treatment is vital to determine the causes and consequences of plant failure. The potential wastewater-related risks that could hamper the operation of the entire manufacturing facility are currently inadequately defined and under researched. This work proposes a framework that includes the comparison of literature and experimental data to quantify the impact of the significant process parameters on the critical process outputs. From the business perspective, managing and minimising risks will be possible when the number of impact parameters is low and the relationships between different parameters are clearly understood. The results show that even only the evaluation of technical risks can provide an assessment platform template for other risk types. Also, the structured and statistically analyzed data sets applied might be further used in the design and development of machine learning platforms algorithms to inform sustainable process outcomes adjusted for various geographical locations and human factors which significantly affect the industrial water sector globally.

An Intelligent Modular Water Monitoring IoT System for Real-Time Quantitative and Qualitative Measurements

This study proposes a modular water monitoring IoT system that enables quantitative and qualitative measuring of water in terms of an upgraded version of the water infrastructure to sustain operational reliability. The proposed method could be used in urban and rural areas for consumption and quality monitoring, or eventually scaled up to a contemporary water infrastructure enabling water providers and/or decision makers (i.e., governmental authorities, global water organization, etc.) to supervise and drive optimal decisions in challenging times. The inherent resilience and agility that the proposed system presents, along with the maturity of IoT communications and infrastructure, can lay the foundation for a robust smart water metering solution. Introducing a modular system can also allow for optimal consumer profiling while alleviating the upfront adoption cost by providers, environmental stewardship and an optimal response to emergencies. The provided system addresses the urbanization and technological gap in the smart water metering domain by presenting a modular IoT architecture with consumption and quality meters, along with machine learning capabilities to facilitate smart billing and user profiling.

Drinking Water Quality Assessment Using a Fuzzy Inference System Method: A Case Study of Rome (Italy)

Drinking water quality assessment is a major issue today, as it is crucial to supply safe drinking water to ensure the well-being of society. Predicting drinking water quality helps strengthen water management and fight water pollution; technologies and practices for drinking water quality assessment are continuously improving; artificial intelligence methods prove their efficiency in this domain. This research effort seeks a hierarchical fuzzy model for predicting drinking water quality in Rome (Italy). The Mamdani fuzzy inference system is applied with different defuzzification methods. The proposed model includes three fuzzy intermediate models and one fuzzy final model. Each model consists of three input parameters and 27 fuzzy rules. A water quality assessment model is developed with a dataset that considers nine parameters (alkalinity, hardness, pH, Ca, Mg, fluoride, sulphate, nitrates, and iron). These nine parameters of drinking water are anticipated to be within the acceptable limits set to protect human health.

EXISTING SYSTEM

The monitoring of wastewater is a widely discussed topic in the scientific literature. In particular, several kinds of technologies contribute to developing sensors that discriminate and classify undesired substances to ensure an adequate water quality level. Some of the authors developed systems able to monitor both water and air thanks to the SENSIPLUS platform [10], [11], [12], [13]. The monitoring outputs can vary, ranging from a classification of the pollutants to a simple binary decision on the presence of contaminants in general. Precise solutions to specific problems are often preferred to the development of generic monitoring system that can work properly in very wide contexts. As an example, Lim [14] describes a system to detect pollutants in the WW framework, although the distinction between different substances is missing and the technologies appear outdated nowadays. A different approach is taken by Lepot et al. [15], where the presence of illegal connections in the sewage system is monitored using an infrared camera. Ji et al. [16] present an image processing system, intended to estimate the WW amount without taking care of the distinction among substances. The cameras adopted to acquire images do not suffer from sensors' corrosion problems but they require a high energy budget, thus making the system far from the low-cost condition. There are other cases where the classification accuracy is very high but the energy/cost constraints are not taken into account. This is the case of Pisa et al. [17], who developed a system to detect ammonium and total nitrogen based on another one that is more broadly designed to detect all components derived from nitrogen.

Drenoyanis et al. [18] propose an interesting portable device to monitor sewer pumping station pumps in order to generate alarms whenever anomalies are detected. The system is surely of great interest, but it does not include any pollutant classification stage. In terms of processing techniques, to the best of our knowledge, this is the first work leveraging natural language processing techniques, and in particular causal models developed for natural language generation, for the task of detecting WW pollution. Nevertheless, in literature we can find examples of the usage of natural language processing techniques and language models for non-canonical tasks. Language models have been used in the medical domain after the application of a ''reverse encoding'' (i.e., translating codes back to their description) for the classification of diagnostic tests [19], [20], [21] and for diagnostic rule encoding [22]. Furthermore, they have been used with a similar technique for the task of human mobility forecasting [23], [24]. More in general, transformer based models originally designed for NLP tasks have demonstrated successful applications in a wide variety of non-NLP tasks [25], including: images [26], [27], [28], videos [29], [30], [31], speech and audio recognition [32], [33], conversational systems [34], [35], recommender systems [36], [37], reinforcement learning [38], [39], graphs [40], [41], protein structure predictions [42], [43], autonomous driving [44], [45], and anomaly detection problems [46], [47].

Disadvantages

• The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets for Detection of Wastewater Pollution.

• Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.

• Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

Proposed System

In the proposed system, the system proposes a novel system based on deep learning, and in particular on causal generative models developed for natural language tasks, for the detection and classification of pollutants in WW, starting from the data collected by a multisensory system based on SENSIPLUS (Sensichips srl, Pisa, Italy). Note that the present paper does not present the infrastructure necessary for data transport as any solution based, for example, on MQTT or message queuing protocols could be used for this purpose.

Advantages

- Baseline extraction: a baseline signal is extracted to normalize raw data.
- Forwarding decision: for each sample, the FSM decides whether to forward it to the classifier, also providing the injection time.
- The proposed classification module is based on deep learning for natural language processing, and in particular on Transformer-based models.
- The proposed system is end-to-end and contains hardware and software components in

MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of

Water Pollution Type, View Water Pollution Type Ratio, Download Predicted Data Sets, View Water Pollution Type Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

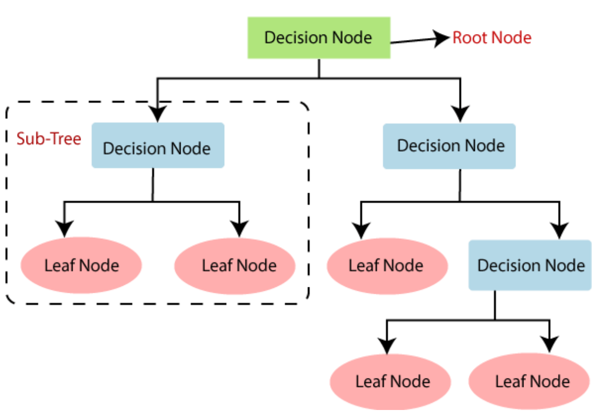
Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT WATER POLLUTION TYPE, VIEW YOUR PROFILE.

ALGORITHMS

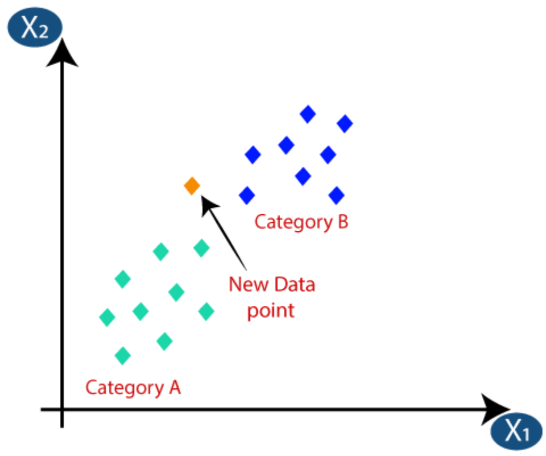
DECISION TREE CLASSIFICATION ALGORITHM

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***



K-NEAREST NEIGHBOR(KNN) ALGORITHM FOR MACHINE LEARNING

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:

CONCLUSION

In this paper we studied the capabilities of natural language processing models, especially generative causal models and more in detail T5, for the task of detecting the presence of polluting substances in wastewater. To this end, differently from up-to-date machine learning models, we textified the input characteristics to make them textual so they could be fed into a generative natural language model. The second part of the system is designed to detect and categorise samples according to their potential contamination levels. We conducted an experimental evaluation of the suggested approach, comparing its performance to a set of benchmarks representing the state-of-the-art, and we calculated its efficiency. In comparison to the baseline approaches, the suggested methodology performs better in experiments, and its efficiency and efficacy make it suitable for deployment and real-world application. In light of the fact that the intended method is unconventional and may seem unusual or counter-intuitive at first glance, we will explain below why this strategy is sound and effective in practice. Some examples of tasks that are not directly related to or expressed naturally using natural language processing include images [26, 27], videos [29], reinforcement learning [39], and graphs [40]. However, recent work has shown that transformers and attention based models can generalise on a large variety of tasks, even ones that the model has not been trained on [60, 61, 62, 63]. The attention mechanism and the almost task-agnostic training approach give transformer-based models their generalisability. Reconstructing a masked or disturbed input item using domain-specific methods or predicting its continuation (if the masked section is the final element of the input) is really what it boils down to in its most basic form. The model is able to learn relevant and, more significantly, broad latent associations in input sequences and how to connect them to the network's output via the combined use of these strategies. For instance, when applied to texts, networks can either generate new text or restore damaged or missing text; when applied to images and videos, they can restore or repair damaged or missing frames; when applied to graphs, they can learn intricate sub-structures, such as arrangements of sets of nodes and edges; and so on. Networks trained with masking or causal goals (i.e., to predict masked regions or the continuation of the input) and based on transformers also exhibit excellent gener-alization skills across tasks and domains, in addition to those particular abilities. We also think that the textual description we get from the sensors and utilise to train our neural network could allows for accurate forecasting predictions for the possible polluting substances present in wastewater.

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, 2022. [Online]. Available: https: //www.mdpi.com/2073-4441/14/3/335

[2] H. Chojer, P. Branco, F. Martins, M. Alvim-Ferraz, and S. Sousa, "Development of low-cost indoor air quality monitoring devices: Recent advancements," *Science of The Total Environment*, vol. 727, p. 138385, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720318982

[3] L. S. Hillary, S. K. Malham, J. E. McDonald, and D. L.Jones, "Wastewater and public health: the potentialof wastewater surveillance for monitoring covid-19," *Current Opinion in Environmental Science & Health*, vol. 17, pp. 14–20, 2020.

[4] A. 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 Resources and Industry*, vol. 25, p. 100146, 2021.

[5] E. Syrmos, V. Sidiropoulos, D. Bechtsis, F. Stergiopou- los, E. Aivazidou, D. Vrakas, P. Vezinias, and I. Vla- havas, "An intelligent modular water monitoring iot system for real-time quantitative and qualitative mea- surements," *Sustainability*, vol. 15, no. 3, p. 2127, 2023.

[6] D. Jalal and T. Ezzedine, "Decision tree and support vector machine for anomaly detection in water distribu- tion networks," in *2020 International Wireless Commu- nications and Mobile Computing (IWCMC)*, 2020, pp. 1320–1323.

[7] D. G. Eliades and M. M. Polycarpou, "Water con-tamination impact evaluation and source-area isolationusing decision trees," *Journal of Water Resources Plan-ning and Management*, vol. 138, no. 5, pp. 562–570, 2012. [Online]. Available: A. Ria, M. Cicalini, G. Manfredini, A. Catania, M. Pi-

otto, and P. Bruschi, “The sensiplus: A single-chip fully programmable sensor interface,” in *Applications in Electronics Pervading Industry, Environment and Society*, S. Saponara and A. De Gloria, Eds. Cham: Springer International Publishing, 2022, pp. 256–261.

[8] M. Molinara, C. Bourelly, L. Ferrigno, L. Gerevini,

M. Vitelli, A. Ria, F. Magliocca, L. Ruscitti, R. Sim- marano, A. Trynda, and P. Olejnik, “A new dataset fordetection of illegal or suspicious spilling in wastewaterthrough low-cost real-time sensors,” in *2022 IEEE In- ternational Conference on Smart Computing (SMART-COMP)*, 2022, pp. 293–298.M. Ferdinandi, M. Molinara, G. Cerro, L. Ferrigno,C. Marrocco, A. Bria, P. Di Meo, C. Bourelly, and R. Simmarano, “A novel smart system for contaminants detection and recognition in water,” in *2019 IEEE In- ternational Conference on Smart Computing (SMART-COMP)*, 2019, pp. 186–191.

[9] C. Bourelly, A. Bria, L. Ferrigno, L. Gerevini, C. Mar-rocco, M. Molinara, G. Cerro, M. Cicalini, and A. Ria,“A preliminary solution for anomaly detection in waterquality monitoring,” in *2020 IEEE International Con- ference on Smart Computing (SMARTCOMP)*, 2020, pp. 410–415.

[10][11] Reddy, Kallem Niranjan, and Pappu Venkata Yasoda Jayasree. "Low Power Strain and Dimension Aware SRAM Cell Design Using a New Tunnel FET and Domino Independent Logic." International Journal of Intelligent Engineering & Systems 11, no. 4 (2018).

[11][12] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Design of a Dual Doping Less Double Gate Tfet and Its Material Optimization Analysis on a 6t Sram Cells."

[12][13] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Low power process, voltage, and temperature (PVT) variations aware improved tunnel FET on 6T SRAM cells." Sustainable Computing: Informatics and Systems 21 (2019): 143-153.

[13][14] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Survey on improvement of PVT aware variations in tunnel FET on SRAM cells." In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 703-705. IEEE, 2017

[14][15] Karne, R. K. ., & Sreeja, T. K. . (2023). PMLC- Predictions of Mobility and Transmission in a Lane-Based Cluster VANET Validated on Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5s), 477–483. https://doi.org/10.17762/ijritcc.v11i5s.7109

[15][16] Radha Krishna Karne and Dr. T. K. Sreeja (2022), A Novel Approach for Dynamic Stable Clustering in VANET Using Deep Learning (LSTM) Model. IJEER 10(4), 1092-1098. DOI: 10.37391/IJEER.100454.