Data-Driven Analysis and Prediction using Regression Models on Iot Based Drainage Monitoring System

Telugu Maddileti, Anthony David Victor Raj, Vinayak Neemkar, R.K.Pongiannan

Abstract*: *India with an ambitious goal has stated of coming up with a project of making over a hundred smart cities. The government needs to take several key parameters into consideration few of which are intelligent water systems, intelligent electrical systems, intelligent transport, smart homes etc. The intention of this paper is to come up with a solution using modern technologies of IOT and Machine learning to get a detailed exploration of the data collected through various IOT sensors. The data is processed and used for training the Machine learning models which help in further predicting the safety of future drain data by recognizing patterns and gathering insights using visualizations. Thus, helping to identify and analyze problems related to drains in a more efficient and optimized manner.

Keywords: IOT, Machine learning, sensors, processing, smart, safety.

I. INTRODUCTION

The sewer system plays a vital role in an urban infrastructure. Towards our goal to make our drainage systems better and also observe that everything is in limit. In India, most of the management and maintenance of underground drainage is completely manual and therefore, not efficient enough to have an optimized solution for cleaning as well tackling the various dreadful difficulties faced by the laborers involved in the working of these confined spaces. Because of which, the government personnel find it complex to pinpoint accurately and tackle the problems related to manholes [1].

It is therefore, essential and urgent to come up with a solution that can manage underground drainage efficiently without the workers having to risk their life on a daily basis. Sensors play a crucial role in this scenario to collect data on various parameters such as pressure, toxic gas concentration, temperature, rate of flow, depth of drainage. All the data of parameters are collected over a period of time and a dependent parameter is defined for every record to determine if it is in a specific limit and is assigned a safety grade for a set of sensor values i.e. "Grade". After collecting the data, it is organized using Data Frames and using various machine learning regression algorithms. The best technique is used to predict the 'Grade' of future data to help in having a brief analysis of the drainage and thus, take accurate and problem-specific required measures.

Revised Manuscript Received on October 10, 2019

* Correspondence Author

Telugu Maddileti, Assistant Professor, Electronics and Computers Department, Sreenidhi Institute of Science and Technology, HYD, Telangana-501301 Email:madhu14283@gmail.com

Anthony David Victor Raj, Electronics and Computers Department, Sreenidhi Institute of Science and Technology, HYD, Telangana-501301.

Vinayak Neemkar, Electronics and Computers Department, Sreenidhi Institute of Science and Technology, HYD, Telangana-501301.

R.K. Pongiannan, Professor, Department of EEE, SRM Institute of Science and Technology, Kattankulathur, (Tamil Nadu), India.

II. EXISTING SYSTEM

The Indian drainage system is ancient and there are no further advancements in terms technology and development. Hence whenever a problem surfaces such as a block in the drain pipe it is challenging to diagnose and identify the issue precisely. This has many drastic consequences if not handled in time. Hence the process of handling and managing drainages is time consuming. The problem goes out of hand especially in the situation where the drain pipes are blocked completely. Due to such failures and difficulties in the drainage line, the workers face a lot of problems, sometimes even death [2]. Some of the primary problems are as mentioned below.

• Contains or may contain a hazardous atmosphere.
• Contains materials that may lead to potential engulfment.
• Most of the drains have a design which could cause possible entrapment or even suffocation.
• Serious health risks and consequences if exposed to toxic gases.

- A gas acquires ideal conditions to burn once it has reached its lower Explosion Limit LEL [2]. Concentrations below this limit have lesser chances to burn.
- The highest concentration of a gas at which it is potentially combustible is called the Upper Explosion Limit (UEL). Any concentration above this limit causes ideal conditions for intensive combustion.

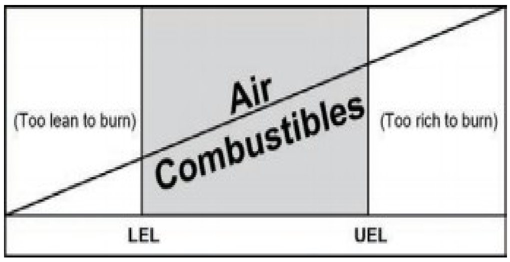

Figure 1. Lower Explosion Limit (LEL) Vs. Upper Explosion Limit (UEL).

In a typical drain, it is important that one takes samples from different distances from the bottom in order to find traces of gases and vapors with varying concentrations. Those with higher concentrations settle at the top most or lower most parts of the drain depending on their densities and concentrations. Gases and vapors with lower concentrations spread uniformly all over the drain. It is important therefore, to collect samples at distances away from the opening. This is important as it may lead to collection of inconsistent data caused due to air interference. The worker has to carry out continuous monitoring which is not feasible considering the situation in Indian drains. Drains have a consistently changing environment and its parameters can change without prior notice or warning.

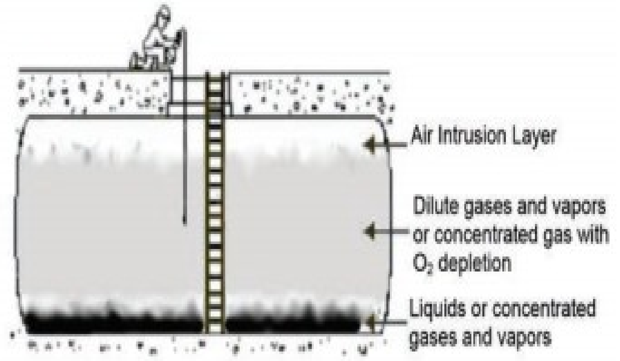

Figure 2.Cross-Sectional diagram of a typical Drain.

Life threatening effects of gases such as CO and H2S:

Ppm	Time(hrs)	Effects
30	8	Tolerable limit
180	3	Slight migraine and uneasiness
350	2	Severe migraine and uneasiness
580	1	Intense migraine and uneasiness
800-2000	2	Confusion, migraine and uneasiness
800-2000	0.5-1	Tendency to stagger
800-2000	0.5	Minor heart problems
2000-2800	0.5	Catalepsy
4000	>1	Deadly

Table1: Effects of Carbon Monoxide.

Ppm	Time(hrs)	Effects
10	8	Tolerable limit
40-110	1	Irritation in eyes and respiration
180-300	1	Irritation in eyes and respiration
500-700	0.5	Catalepsy, passing
>1000	<1	Catalepsy, passing

Table 2: Effects of Hydrogen Sulphide.

proposed system

To achieve an efficient and optimized drainage analysis system which is cost-effective and feasible for conditioning, monitoring and maintaining drains in the city.

The main objectives of this system are:

- Detecting the concentration of inflammable and toxic gases.
- Use of flow sensors to detect the variations in the flow.
- Sensing the temperature and Pressure and updating it in real time through IOT.
- Using ultrasonic sensors to detect the height and level of the drainage.

The Node MCU(ESP-8266) collects the data from various sensors and using HTTP post requests (every 30 seconds); the IOT firmware transmits the sensor data to the domain and using PHP scripts the data is then added to the MySQL database. The table data is then queried and converted into Data Frames for further processing using python and later used for training and testing the machine learning algorithms.

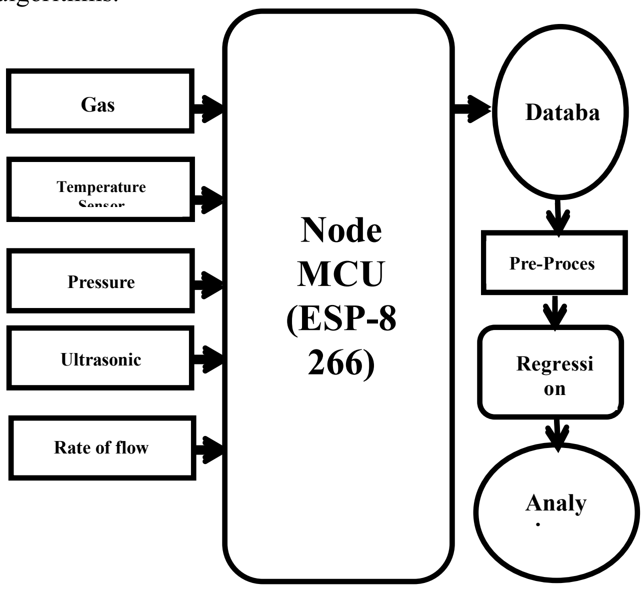

Figure 3. System Architecture.

III. DATA COLLECTION

The sensors play a crucial role in collection of significant data that is used in the process of training the learning models. The required sensors used in this system are explained below [3].

A. MQ9 GAS SENSOR

The sensor uses an analog resistance for the output. It has good sensitivity to CO throughout a wide concentration range and has other benefits such as long lifespan, low cost, and simple drive circuit etc and its Range is 0-1000 ppm.

B. LM 35 TEMPERATURE SENSOR

This sensor is cost effective due to trimming and calibration at the water level. The LM35's low output, precise inherent calibration, impedance and linear output make it easier to interface with the readout or control circuitry. It is also power efficient as it consumes only 60µA and has negligible internal heating and its range is -155 to +150 C.

C. BPM 180 PRESSURE SENSOR

The pressure sensor normally serves as a transducer and measures the force acting per unit area. The output signals are generated depending on the measure of applied pressure and its range is 0-500psi.

D. HC-SR04 LEVEL SENSOR

Ultrasonic sensors are used in this system to detect the level of the sewage. Level monitoring and detection is not hindered by the presence of sludge or other granular elements that may be present in a drain and its range is 0-10ft.

IV. DATA STORAGE

MySQL is the database that is used in the system. Along with several tools and features, it has high capability to scale and has several mechanisms for fault tolerance making it one of the best databases when considering IOT-based applications [5].

Once the data is collated, it needs to be transported to a local database such as a MySQL server for further analysis and insights on the sensor data.

There are four key goals to consider when addressing availability and storage of the data:

- Recover from storage media failure.
- Quickly recover from database failure.
- Improve performance.
- Achieve zero loss of data collection.

The MySQL database has functions that help with monitoring, backup and recovery, redundancy, scaling and fault tolerance of the Data collected.

DATA FRAME CONVERSION AND PREPROCESSING:

The data is collected from all the sensors and stored in the online database in a structured manner. The pre-processing in Python is easy to perform due to various functions provided by Pandas and NumPy. Some of the steps are given below:

- Converting the online database into a Data Frame using pandas so that handling of the data is done more efficiently.
- Converting the data that is in different scales or units.
- Handling null values and using different techniques to fill the "NaN" values of the Data Frame so that data is not lost and consistency is maintained.
- Hot encoding any categorical values which may affect the training of the model
- Dropping columns that is not useful for analysis
- Removing the redundant values or data.

V. DATA MODELING AND TESTING

The scikit library offers several machine learning models that are very easy to implement in code. To use machine learning, one must remember to properly install NumPy and import the desired model from Scikit-Learn. After training the model, it is tested with the same instance and save the results obtained.

Data is tested for efficiency and performance with different regression models like Decision Tree Regressor, Random Forest Regressor and XGBoost.

Decision Tree

Decision trees are constructed by the recursive partitioning, starting from the root node, known as the first parent, each node can be split into left and right child nodes [6]. These nodes can then be further split and they themselves become parent nodes of their resulting children nodes.

From the root, the data is split based on a function generating the largest Information Gain (IG). In an iterative process, splitting operation is repeated on each child node until the leaves are pure [10]. Here the goal is to maximize the information gain each time the tree is split, which we define as follows:

$$IG(D_p, f) = I(D_p) - \left(\frac{N_{left}}{N_p} I(D_{left}) + \frac{N_{right}}{N_p} I(D_{right})\right)$$

Where f the feature and Dp, Dleft, Dright are the Datasets of parent and child nodes and I is the Impurity measure.

Random Forest

This modeling technique is based on an ensemble set and is not a boosting technique. Execution of the individual trees is done in parallel and so there is no sort of communication between the trees when creating the model.

Entropy, Information gain, Reduction in Variance and Gini Index are some of the key parameters needed to be considered when tuning the model [7]. Some of the other significant considerations are as mentioned below:

1. Gini Index or information Gain favors a large partition whereas Entropy favors a small count but many distinct values.
2. A random sample is drawn by the model from the record each time it generates divisions and adds another random element that prevents any over fitting.

XGBoost

This model is a boosting algorithm and so is mentioned as XGBoost. The term "Gradient Boosting" comes from the article "Approximation of Gourmet Functions: a Gradient Reinforcement Machine" by Friedman [8]. It is the most commonly used model when considering supervised learning.

The model is highly scalable with respect to all cases and parameters. It is significantly faster compared to present solutions on a single machine. The model trains a bunch of individual models in such a way that it learns from the mistakes made by the previous model thus, increasing the model accuracy.

Here the hyper parameters include max depth of the tree which is Maximum depth of the tree for base learners, n estimators which is the count of the trees used to fit, learning rate for Boosting the eta and n jobs is the number of parallel threads used to run the XGBoost.

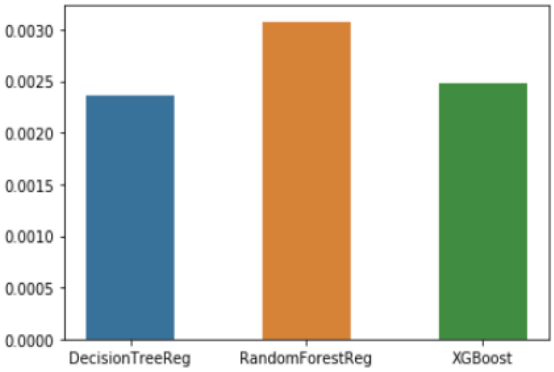

Figure 4. Comparing the performance of various models against their respective Test Mean Errors.

VI. DATA VISUALIZATION OF CORRELATION BETWEEN THE INPUT PARAMETERS AND SAFETY GRADE

A linear relationship in the sample data between Drain level in meters and the Safety Grade is shown in the figure 5 and should be maintained within the required level to avoid risks.

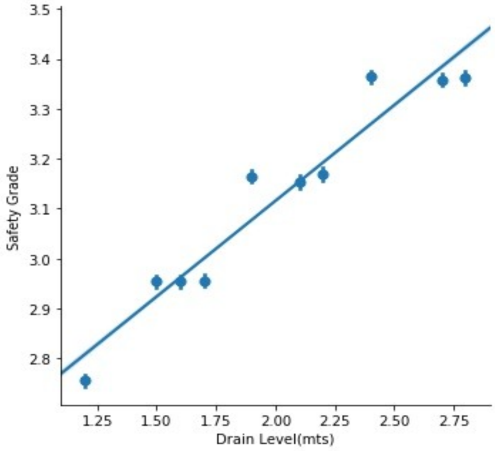

Figure 5. Drain level vs Safety grade

From the figure 6, it can be seen that pressure in psi has an inverse relation with the safety of the drain and should not exceed the permissible threshold level.

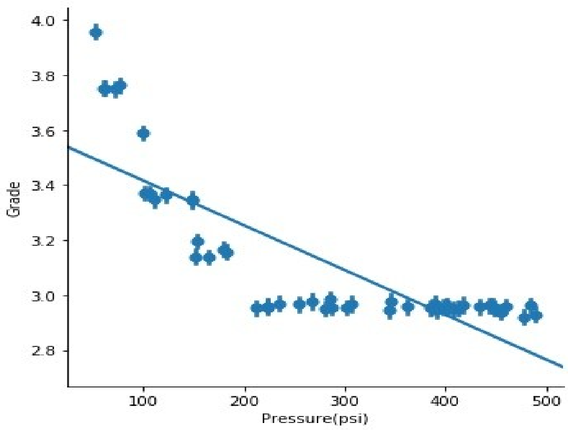

Figure 6. Pressure vs Safety grade

Flow Rate of the drain has a positive impact on the safety of the drain as shown in figure 7 and hence, should be monitored continuously.

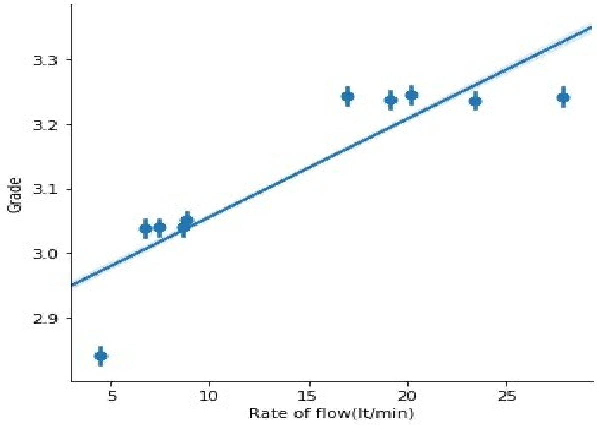

Figure 7. Flow rate vs Safety grade

Temperature has an inverse realtionship with the safety and should be maintained within the limit and is shown in figure 8.

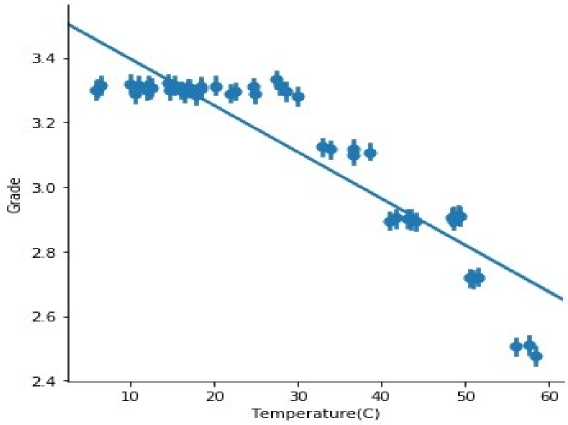

Figure 8. Temperature vs Safety grade

The figure 9 shows the effect of harmful gas concentration in ppm on the safety of the drain and hence factors such flammability and toxicity can be determined.

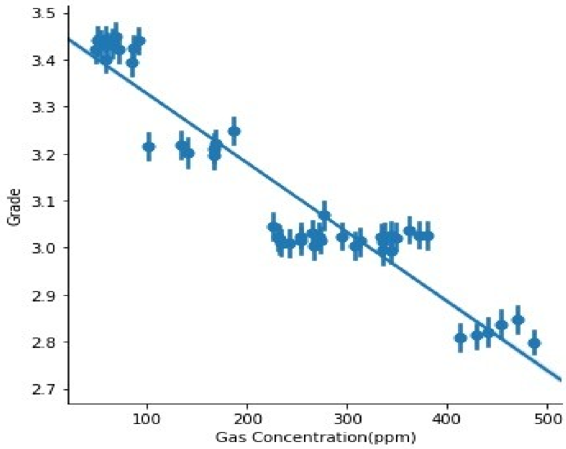

Figure 9. Gas Concentration vs Safety grade

VII. FUTURE ENHANCEMENT

The amount of data collected from the sensors is limited and taken in a small scale to train the models. In the future, this data will increase to a large extent so that model would perform better in terms of accuracy and higher throughput.

Many of the new machines learning models that have been launched recently need to be radically changed to take over. Some might argue of one model's advantages and disadvantages over the other. But more efficient tools to deal with the problems associated with Big Data need to be developed. The tools to be developed must include provisions for dealing with noisy and unbalanced data, uncertainties and inconsistencies. Hence, in the future with better models and approaches our system can be boosted and scaled for higher performance and accuracy.

VIII. CONCLUSION

In most developing countries the drainage system has posed several problems to the environment and posing a threat to human life especially to the municipality workers. In the light of this issue this paper has significantly come up with the solution combining modern technologies such as IOT and Machine learning models, thus helping to analyze and predict the safety of a drain using a data driven approach.

REFERENCES

1. https://www.hyderabadwater.gov.in/en/files/4514/7833/4368/EOI-_GEO_TAGGING_-_Web_site.pdf
2. https://www.raesystems.com/sites/default/files/content/resources/Application-Note-206_Guide-To-Atmospheric-Testing-In-Confined-Spaces_04-06.pdf
3. https://www.ijariit.com/manuscripts/v4i1/V4I1-1207.pdf
4. https://www.irjet.net/archives/V6/i3/IRJET-V6I3471.pdf
5. https://www.mysql.com/
6. https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart
7. https://scikit-learn.org/stable/modules/ensemble.html#random-forests
8. https://xgboost.readthedocs.io/en/latest/parameter.html
9. Chang, A. Y., Yu, C. S., Lin, S. C., Chang, Y. Y., & Ho, P. C. (2009, August). Search, identification and positioning of the underground manhole with RFID ground tag. In *2009 Fifth International Joint Conference on INC, IMS and IDC* (pp. 1899-1903). IEEE.
10. Chen, D. R., & Chiang, K. F. (2015). Cloud-based power estimation and power-aware scheduling for embedded systems. *Computers & Electrical Engineering*, *47*, 204-221.
11. Liu, S., Zhang, Z., & Cao, X. (2016). Information integration for ground-based cloud classification using joint consistent sparse coding in heterogeneous sensor network. *Signal Processing*, *126*, 165-172.
12. . Peddi, S. V. B., Kuhad, P., Yassine, A., Pouladzadeh, P., Shirmohammadi, S., & Shirehjini, A. A. N. (2017). An intelligent cloud-based data processing broker for mobile e-health multimedia applications. *Future Generation Computer Systems*, *66*, 71-86.
13. Pokharel, B., Geerts, B., Jing, X., Friedrich, K., Ikeda, K., & Rasmussen, R. (2017). A multi-sensor study of the impact of ground-based glaciogenic seeding on clouds and precipitation over mountains in Wyoming. Part II: Seeding impact analysis. *Atmospheric Research*, *183*, 42-57.
14. Badawi, H. F., Dong, H., & El Saddik, A. (2017). Mobile cloud-based physical activity advisory system using biofeedback sensors. *Future Generation Computer Systems*, *66*, 59-70.
15. Madsen, H. M., Brown, R., Elle, M., & Mikkelsen, P. S. (2017). Social construction of stormwater control measures in Melbourne and Copenhagen: A discourse analysis of technological change, embedded meanings and potential mainstreaming. *Technological Forecasting and Social Change*, *115*, 198-209.
16. Ju, J., Kim, M. S., & Ahn, J. H. (2016). Prototyping business models for IoT service. *Procedia Computer Science*, *91*, 882-890.
17. Grieco, L. A., Rizzo, A., Colucci, S., Sicari, S., Piro, G., Di Paola, D., & Boggia, G. (2014). IoT-aided robotics applications: Technological implications, target domains and open issues. *Computer Communications*, *54*, 32-47.
18. Bulu, M., Önder, M. A., & Aksakalli, V. (2014). Algorithm-embedded IT applications for an emerging knowledge city: Istanbul, Turkey. *Expert Systems with Applications*, *41*(12), 5625-5635.
19. Basanta-Val, P., & García-Valls, M. (2015). A library for developing real-time and embedded applications in C. *Journal of Systems Architecture*, *61*(5-6), 239-255.
20. Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356-365.
21. Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70*, 308-317.
22. Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment?. *International Journal of Production Economics*, *182*, 113-131.
23. Han, K. K., & Golparvar-Fard, M. (2017). Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. *Automation in Construction*, *73*, 184-198.
24. Ortiz-Partida, J. P., Lane, B. A., & Sandoval-Solis, S. (2016). Economic effects of a reservoir re-operation policy in the Rio Grande/Bravo for integrated human and environmental water management. *Journal of Hydrology: Regional Studies*, *8*, 130-144.
25. Han, M., Hashemi, S., Joo, S. H., & Kim, T. (2016). Novel integrated systems for controlling and prevention of mosquito-borne diseases caused by poor sanitation and improper water management. *Journal of Environmental Chemical Engineering*, *4*(4), 3718-3723.
26. Walsh, B. P., Bruton, K., & O'Sullivan, D. T. J. (2017). The true value of water: A case-study in manufacturing process water-management. *Journal of cleaner production*, *141*, 551-567.
27. Ziolkowski, B., & Ziolkowska, J. R. (2017). Product, Process, and Organizational Innovations in Water Management. In *Competition for Water Resources* (pp. 403-422). Elsevier.
28. Girisrinivaas, R., & Parthipan, V. (2017, September). Drainage overflow monitoring system using IoT (DOMS). In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 2133-2137). IEEE.

AUTHORS PROFILE

Mr. Telugu Maddileti is currently working as Assistant Professor at SNIST and Research Scholar at JNTUA, Ananthapuramu. He has a total of 14 years experience in academics. He completed B.Tech degree (Electronics and communication Engineering) in the year 2005 and M.Tech in 2009 from JNT University, Hyderabad. He is enthusiastic to work in the field of VLSI. His area of interest includes Low power VLSI domain, optimization of analog and mixed signal circuits in addition to digital system design using CAD tools. He has presented more than 10 International/National Technical Papers in journals & conferences.

Anthony David Victor Raj completed his B.Tech degree from JNTU, Hyderabad. His areas of interest include IOT based designs and applications. He has presented 4 International/National Technical Papers in journals & conferences.

Vinayak Neemkar completed his B.Tech degree from JNTU, Hyderabad. His areas of interest include IOT based designs and applications. He has presented 5 International/National Technical Papers in journals & conferences.