International Journal of Science Engineering and Advance Technology, IJSEAT, Vol. 14, Issue 2, Special Issue on Artificial Intelligence, Machine Learning and Smart Intelligent Systems

ISSN 2321-6905
2026

MACHINE LEARNING-BASED FORECASTING OF AIR QUALITY INDEX USING DATA ANALYTICS

Mrs. K. Mounika[1],*, K. Shathakshari[2], M. Ravi[2], M. Manish[2], Mohd Adnan[2]
[1]Assistant Professor, Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
[2]B.Tech (Scholar), Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
Correspondence: kmounika@tkrcet.com

ABSTRACT

Air pollution is a growing environmental issue that significantly affects human health and ecological balance. The Air Quality Index (AQI) is commonly used to indicate pollution levels, but most existing systems only provide current or past information without predicting future conditions. To address this limitation, this study presents a machine learning-based approach for forecasting AQI using historical pollutant data and environmental parameters. Data from various sources is processed and analyzed to identify patterns that influence air quality. The proposed system applies machine learning models to predict future AQI values and analyze pollution trends over time. It also includes visualization techniques to present results in a clear and understandable manner. By providing early predictions and meaningful insights, the system helps individuals and authorities take timely actions to reduce health risks and manage air pollution effectively. The system is designed to be scalable and adaptable across different regions using diverse data sources. It ensures efficient performance with minimal computational overhead, making it suitable for real-time applications. Overall, the approach enhances proactive environmental monitoring and supports sustainable decision-making.

KEYWORDS: Air Quality Index (AQI), Air Pollution Prediction, Machine Learning, Data Analytics, Environmental Monitoring, Predictive Modeling, Ensemble Learning, Time-Series Analysis, Data Visualization, Public Health, Pollution Trend Analysis, Forecasting Models

I. INTRODUCTION

Air pollution has become one of the most serious environmental challenges in today's world, affecting both developed and developing nations. The rapid growth of industrialization, urbanization, and transportation has significantly increased the emission of harmful pollutants into the atmosphere. Activities such as fuel combustion, industrial processes, construction work, and vehicular emissions release pollutants like particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) [1].

These pollutants degrade air quality and disrupt the natural balance of the environment. As a result, air pollution has emerged as a major global concern requiring immediate attention and effective management strategies [2]. The impact of poor air quality on human health is severe and widespread. Continuous exposure to polluted air can lead to respiratory diseases such as asthma, bronchitis, and chronic obstructive pulmonary disease (COPD) [3]. It also increases the risk of cardiovascular problems, lung cancer, and weakened immune systems. In addition to health issues, air pollution affects ecosystems, reduces agricultural productivity, and contributes to climate change. According to various environmental studies, millions of premature deaths occur every year due to air pollution-related illnesses. This highlights the urgent need for monitoring, analyzing, and controlling air pollution levels to ensure a healthier living environment.

To simplify the representation of complex pollution data, the Air Quality Index (AQI) is widely used as a standardized measure. AQI converts multiple pollutant concentrations into a single numerical value and categorizes it into levels such as good, moderate, unhealthy, or hazardous. This makes it easier for the general public to understand air quality conditions and take necessary precautions. However, most traditional AQI monitoring systems are limited to providing real-time or historical data. While such information is useful, it does not help in anticipating future air quality conditions. The absence of predictive capability restricts proactive decision-making and delays the implementation of preventive measures [4].

In recent years, advancements in computational

ISSN 2321-6905
2026

technologies have opened new opportunities for addressing environmental challenges. Machine Learning (ML) and Data Analytics have emerged as powerful tools capable of processing large volumes of data and extracting meaningful insights. These techniques can identify complex patterns and relationships between pollutants and environmental factors that are often difficult to detect using traditional statistical methods. By leveraging historical air quality data along with meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure, machine learning models can predict future AQI values with considerable accuracy [5-6].

The application of machine learning in air quality prediction has gained significant attention due to its ability to handle non-linear relationships and dynamic environmental conditions. Algorithms such as Linear Regression, Random Forest, and XGBoost have proven to be effective in modeling air pollution trends. These models can learn from past data and continuously improve their prediction performance.Additionally, data analytics techniques help in understanding seasonal variations, identifying pollution hotspots, and analyzing long-term trends in air quality. This combination of prediction and analysis provides a comprehensive approach to managing air pollution [7].

Furthermore, accurate AQI forecasting plays a vital role in supporting public health and environmental policies. Early prediction of pollution levels allows individuals to take preventive actions such as reducing outdoor exposure, using protective equipment, or adjusting daily activities. It also enables government authorities to implement timely control measures, such as traffic regulation,
industrial emission control, and public awareness campaigns. In smart city initiatives, predictive air quality systems can be integrated with IoT sensors and real-time data platforms to enhance monitoring and response mechanisms [8].

This project focuses on developing a machine learning-based framework for forecasting the Air Quality Index using historical pollutant and meteorological data. The system integrates data preprocessing, feature analysis, and predictive modeling to generate accurate AQI forecasts. Multiple machine learning algorithms are implemented and evaluated to identify the most effective model for prediction. In addition, the system incorporates visualization techniques to present AQI trends and predictions in a clear and user-friendly manner.

II. LITERATURE SURVEY

Air quality prediction has gained significant attention in recent years due to the increasing impact of air pollution on human health and the environment [9]. Many researchers have explored different techniques using data analytics and machine learning to improve the accuracy of Air Quality Index (AQI) prediction relationship between pollutants and AQI [10-12].

Random Forest, Support Vector Machines (SVM), and Neural Networks have been widely used for AQI prediction [13]. These models are capable of handling large datasets and identifying non-linear relationships between variables like temperature, humidity, and pollutant concentrations [14]. Among these, Random Forest and Neural Networks have shown better performance in terms of prediction accuracy.

Some researchers have also focused on using deep learning techniques, such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) models, which are particularly useful for time-series data [15]. These models can learn from past trends and provide more accurate future predictions compared to traditional approaches.

In addition, data preprocessing techniques like handling missing values, normalization, and feature selection play a crucial role in improving model performance. Many studies emphasize the importance of clean and well-structured data for achieving reliable results [16].

Overall, the literature shows that machine learning and data analytics provide effective solutions for AQI predi ction. However, there is still scope for improvement in terms of accuracy, real-time prediction, and handling dynamic environmental conditions. This project aims to build upon these existing approaches and develop a more efficient and reliable AQI prediction system [17-19].

Air quality prediction has become an important research area due to the increasing impact of pollution on human health and the environment [20]. Earlier studies mainly focused on statistical methods such as linear regression to analyze the

relationship between different air pollutants and the Air Quality Index (AQI) [21]. Although these methods are simple and easy to implement, they often fail to capture complex and non-linear patterns present in environmental data [22].

Researchers have highlighted the limitations of traditional approaches, especially when dealing with large datasets and time-dependent variations in air quality [23]. With the advancement of technology, machine learning and deep learning techniques have been widely adopted for AQI prediction [24].

III. PROPOSED METHODOLOGY

This section outlines the structured approach used to predict Air Quality Index (AQI) using machine learning models and interactive data visualization. The methodology follows a sequential pipeline, beginning with data acquisition and ending with spatial analysis, as illustrated in the system flow diagram and demonstrated through the web-based interface.

3.1 Data Collection

The system begins by collecting air quality and meteorological data from reliable sources such as public APIs and environmental monitoring platforms. The dataset includes pollutant parameters like PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with weather factors such as temperature, humidity, wind speed, and pressure. This combined data helps in understanding both pollution levels and environmental influences affecting AQI.

3.2 Data Preprocessing

The collected data is cleaned and prepared to ensure accuracy and consistency. Missing values are handled using appropriate techniques, and noisy or irrelevant data is removed. Feature scaling and normalization are applied to maintain uniformity across all variables, which improves the performance and stability of machine learning models.

3.3 FeatureAnalysis and Model Development

Relevant features that significantly impact AQI are selected and analyzed. Exploratory data analysis is performed to identify patterns, trends, and correlations among pollutants and environmental parameters. Based on these insights, machine learning models such as Linear Regression, Random Forest, and XGBoost are developed, with emphasis on ensemble methods for better prediction.

3.3 FeatureAnalysis and Model Development

Relevant features that significantly impact AQI are selected and analyzed. Exploratory data analysis is performed to identify patterns, trends, and correlations among pollutants and environmental parameters. Based on these insights, machine learning models such as Linear Regression, Random Forest, and XGBoost are developed, with emphasis on ensemble methods for better prediction.

3.4 Model Training and Evaluation

The dataset is divided into training and testing sets, typically using an 80:20 ratio. The models are trained using historical data and evaluated on unseen data to ensure proper generalization. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score are used to compare models and select the best-performing one.

3.5 AQI Prediction and Visualization

The selected model is used to predict future AQI values based on input data. The predicted results are presented using visualizations such as graphs and dashboards to enhance understanding. This enables users to interpret air quality trends easily and supports early decision-making for environmental management.

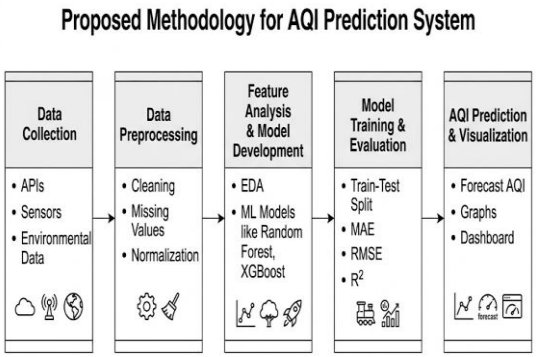


Fig.3.1. Flowchart representing the proposed methodology for AQI prediction using machine learning.

ISSN 2321-6905
2026

IV. ARCHITECTURE

The proposed system architecture for Air Quality Index (AQI) prediction is designed as a modular and scalable framework that integrates data collection, preprocessing, machine learning modeling, and visualization components. The architecture ensures efficient handling of environmental data and accurate prediction of air quality levels. It consists of multiple layers that work together to transform raw data into meaningful insights.

The process begins with the data acquisition layer, where air quality and meteorological data are collected from various sources such as public APIs, environmental monitoring systems, and sensor-based platforms. The collected dataset includes pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with weather parameters like temperature, humidity, wind speed, and atmospheric pressure. This diverse set of features enables a comprehensive understanding of factors influencing air quality. The collected data is then passed to the data preprocessing module, which plays a crucial role in improving data quality. In this stage, missing values are handled using appropriate techniques, and noisy or inconsistent data is removed. Feature scaling and normalization are applied to ensure uniformity across all input variables. This step ensures that the data is clean, structured, and suitable for machine learning algorithms.

Following preprocessing, the data is processed in the feature analysis and selection module, where important features contributing to AQI prediction are identified. Exploratory data analysis is performed to understand relationships between pollutants and environmental factors. Correlation analysis and statistical methods are used to select relevant features and eliminate redundant ones. This step enhances the efficiency and accuracy of the predictive model.

The core component of the system is the machine learning module, where multiple algorithms such as Linear Regression, Random
Forest, and XGBoost are implemented. These models are trained using historical data to learn patterns and relationships among variables. Ensemble learning techniques are emphasized due to their ability to handle complex and non-linear data effectively. The models are optimized to achieve high prediction accuracy and reliability.

After training, the models undergo evaluation in the model evaluation layer, where their performance is assessed using standard metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The model with the best performance is selected for generating final AQI predictions. This ensures that the system provides accurate and reliable forecasting results.

The predicted AQI values are then passed to the visualization layer, where results are displayed using graphs, charts, and interactive dashboards. This layer helps users easily understand air quality trends and predictions. Visual representations improve interpretability and support better decision-making for both individuals and authorities.

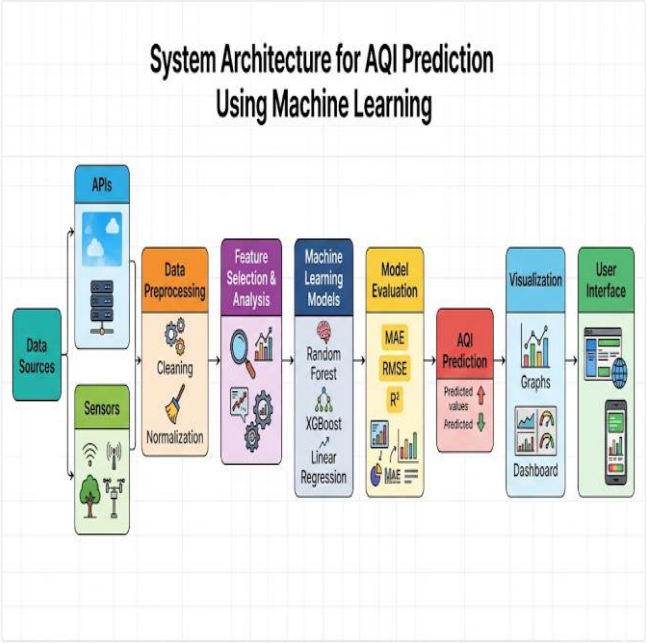


Fig.4.1. System architecture of the proposed AQI prediction system.

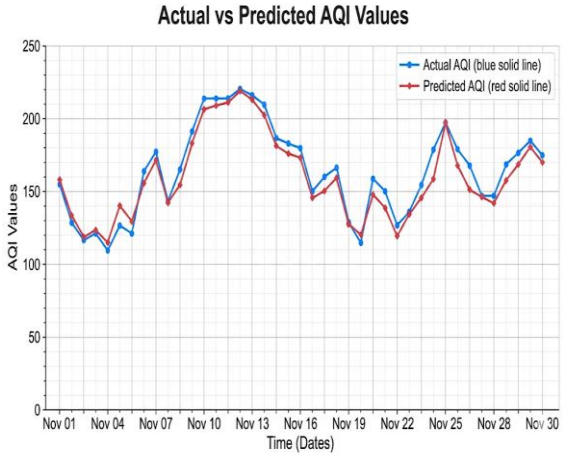
Finally, the system includes a user interface layer, which allows users to interact with the system by inputting data and viewing predictions. The interface is designed to be simple, user-friendly, and accessible across different devices. The overall architecture is flexible and scalable, enabling integration with real-time data sources and deployment in smart city environments.

ISSN 2321-6905
2026

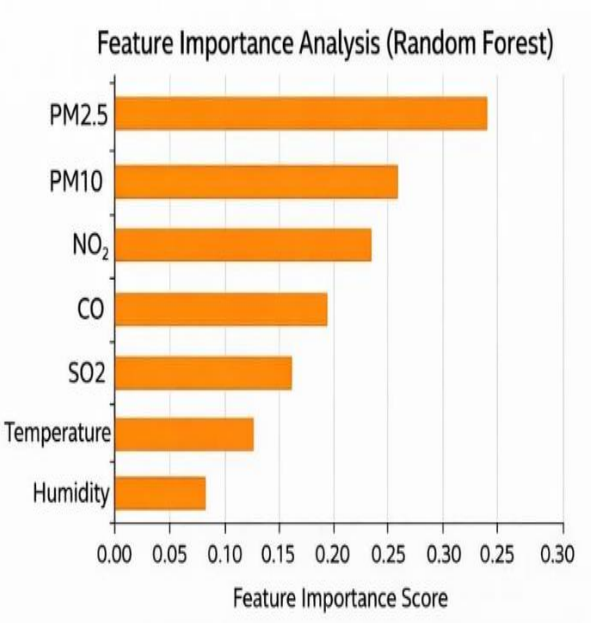
V. RESULT

The implementation of the Air Quality Index (AQI) prediction system using machine learning models produced effective and reliable results. Different algorithms such as Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), LSTM, GRU, and the hybrid LSTM-GRU model were trained and tested on the processed dataset to evaluate their performance.

After training the models, their predictions were compared with actual AQI values using evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2). These metrics helped in understanding how accurately each model performed and how well it could generalize to unseen data.



The results showed that advanced models like LSTM, GRU, and the hybrid LSTM-GRU performed better compared to traditional models such as Linear Regression and KNN. This is because deep learning models are more effective in capturing complex patterns and
time-dependent relationships present in air quality data.



Overall, the system successfully predicted AQI values with good accuracy, proving that machine learning and data analytics can be effectively used for air quality prediction. This will allow continuous monitoring and instant prediction of AQI levels.

VI. CONCLUSION & FUTURE SCOPE

In this project, an effective system for predicting the Air Quality Index (AQI) using machine learning models and data analytics has been developed. The approach involved collecting and preprocessing environmental data, followed by applying multiple machine learning algorithms such as Linear Regression,

K-Nearest Neighbors (KNN), Support Vector Machine (SVM), LSTM, GRU, and hybrid LSTM-GRU models.

The experimental results indicate that advanced models like LSTM, GRU, and LSTM-GRU outperform traditional techniques due to their ability to capture complex and time-dependent patterns in air quality data. Proper data preprocessing and feature selection also played a key role in improving prediction accuracy. The system successfully provides reliable AQI predictions, which can help individuals and authorities take timely actions to reduce the impact

of air pollution

Although the system achieves good prediction accuracy, there are several opportunities for further enhancement. In the future, real-time data from IoT sensors can be integrated to provide live AQI monitoring and prediction. This will make the system more practical and useful for real-world applications.

Advanced deep learning models and hybrid techniques can be explored to further improve prediction performance. The system can also be extended by incorporating geographical and satellite data for more precise location-based.

Additionally, developing a user-friendly web or mobile application can make the system accessible to the public, allowing users to check AQI levels and receive alerts. Integration with government platforms can further support environmental monitoring and policy-making. These improvements can make the system more scalable, accurate, and impactful in the field of air quality prediction.

The present study successfully developed a machine learning-based system for predicting the Air Quality Index (AQI) using historical environmental data. Various models were implemented and compared to analyze their prediction capabilities.

The results clearly show that machine learning techniques can effectively model the relationship between pollutants and AQI levels.

Among the applied models, deep learning approaches such as LSTM and GRU provided better accuracy due to their ability to learn temporal patterns in data. The system not only improves prediction accuracy but also helps in understanding pollution trends.

ACKNOWLEDGEMENT

We sincerely acknowledge the Management of **TKR College of Engineering & Technology (TKRCET)** for providing the necessary permissions, resources, and support that enabled the successful execution of this research work.

We express our deep sense of gratitude to our respected Principal, **Dr. D. V. Ravi Shankar, M.Tech., Ph.D.,** for his constant encouragement and guidance, which have been invaluable throughout our academic journey.

We extend our heartfelt appreciation to **Dr. V. Krishna, M.Tech., Ph.D.**, Head of the Department of CSE (Data Science), TKRCET, for his valuable suggestions, insightful guidance, and constructive feedback that greatly enhanced the quality and direction of this project.

We would also like to sincerely thank **Mr. M. Arokia Muthu, M.E., (Ph.D.)**, Assistant Professor and Project Coordinator, Department of CSE (Data Science), TKRCET, for his continuous support, technical guidance, and motivation during the course of this work.

A special note of gratitude is extended to our Internal Guide, **Mrs. Ms.K.Mounika** Assistant Professor, Department of CSE (Data Science), TKRCET, whose constant support, expert guidance, and encouragement played a vital role in the successful completion of this project.

REFERENCES

[1] Liu, Chunhao, Pan, Guangyuan, Song, Dongming, and Wei, Hao. “Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine.” IEEE Access, vol. 11, 2023.

[2] Krishna, V., Rajyalakshmi, P., Naresh, P., & Ramesh, V. (2019). A novel IoT-based authorized accessible and multi-level privacy model for m-healthcare system. Journal of Xi'an University of Architecture & Technology, 11(11).

[3] Krishna, V., Raju, Y. D. S., Raghavendran, C. V., Naresh, P., & Rajesh, A. (2022). Identification of nutritional deficiencies in crops using machine learning and image processing techniques. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). IEEE.

[4] Krishna, V., Sumalatha, C., Raju, Y. D. S., & Mohan, K. V. M. (2022). Analysis of heart disease prediction using machine learning classification algorithms. Journal of Optoelectronics Laser.

[5] Krishna, V., Raghavendran, C. V., & Faruk, S. K. U. (2024). Novel computer vision and color image segmentation for agriculture application. In Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems. CRC Press.

[6] Muthu, M. A., & Prakash, B. (2025). Efficient privacy-preserving mHealth framework using crisscross AES and FCFS-NDPPP in hybrid cloud. Ingénierie des Systèmes d’Information (ISI).

[7] Muthu, M. A. (2025). Integrated healthcare management and analytics. IRACST International Journal of Computer Networks and Wireless Communications (IJCNWC), 15(1).

[8] Latchoumi, T. P., Parthiban, L., Balamurugan, K., Raja, K., Vijayaraj, J., & Parthiban, R. (2023). A framework for low energy application devices using blockchain-enabled IoT in WSNs. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 121-132). Cham: Springer International Publishing

[9] Balamurugan, K., Deepthi, T., Subramanian, A. K., Banerjee, A., Agarwal, D., Biswas, A., & Sinha, A. (2023). A study on the mechanical properties of rare earth-based aluminium composite. Journal of The Institution of Engineers (India): Series D, 104(1), 15-25

[10] Arunkarthikeyan, K., & Balamurugan, K. (2020). Studies on the effects of deep cryogenic treated WC–Co insert on turning of Al6063 using multi-objective optimization. SN applied Sciences, 2(12), 2103.

[11] Pavan, M. V., Balamurgan, K., & Balamurgan, P. (2021). Wear experiments on PLA-Cu composite filament printed in different FDM conditions. Turkish Journal of Computer and Mathematics Education, 12(9), 2245-2251

[12] Sneha, P., Balamurugan, K., & Kalusuraman, G. (2021). Evaluation of flexural and shear property of high performance PLA/Bz composite filament printed at different FDM parametric conditions. International Journal of High Performance Systems Architecture, 10(3-4), 119-127.

[13] samples printed using fused filament extrusion by response surface method. Progress in Additive Manufacturing, 7(5), 957-969.Balamurugan, K., Pavan, M. V., & Balamurugan, P. (2022). Wear parametric analysis on PLA/Cu filament

[14] Sneha, N., & Balamurugan, K. (2022, October). Micro-drilling optimization study using RSM on PLA-bronze composite filament printed using FDM. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) (pp. 1-5). IEEE.

[15] Deepthi, T., Balamurugan, K., & Uthayakumar, M. (2021). Simulation and experimental analysis on cast metal runs behaviour rate at different gating models. International Journal of Engineering Systems Modelling and Simulation, 12(2-3), 156-164.

[16] Pruthviraju, G., Dambhare, S. G., Pathri, B. P., Ramakrishna, M., Gokulanathan, L., Balamurugan, K., & Shumet, W. (2022). Mechanical Test on Aluminum Alloy with Maximal Soluble SiC Reinforcement. Advances in Materials Science and Engineering, 2022(1), 9848928

[17] Sreenivasa Reddy, K., & Jadhav, P. P. (2023). Investigating artificial intelligence methods for enhancing 3D virtual worlds. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3).

[18] Sreenivasa Reddy, K., & Jadhav, P. P. (2023). Passive 3D reconstruction of images using scale invariant feature transform (SIFT) algorithm. European Chemical Bulletin, 12(S3), 4645–4654.

[19] Geetha, L. S., El-Ebiary, Y. A. B., Srinivasa Rao, B., Rautrao, R. R., Mastan Rao, T. S., Venkata Naga Ramesh, J., & Al-Omari, O. (2025). Challenges and solutions in agile software development: A managerial perspective on implementation practices. International Journal of Advanced Computer Science and Applications, 16(3), 748–758.

[20] Jaya Rama Krishna, V. V., Srinivasa Rao, B., Veeraiah, D., Subba Raju, S., Al Answari, M. S., & Kaur, C. (2024, February). Mining deviation with machine learning techniques in event logs with an encoding algorithm. Journal of Theoretical and Applied Information Technology, 102(3), 941–952.

[21] Sneha, P., & Balamurugan, K. (2022). Investigation on wear characteristics of a PLA-14% bronze composite filament. In Recent Trends in Product Design and Intelligent Manufacturing Systems: Select Proceedings of IPDIMS 2021 (pp. 453-461). Singapore: Springer Nature Singapore

[22] Latchoumi, T. P., Parthiban, L., Balamurugan, K., Raja, K., Vijayaraj, J., & Parthiban, R. (2023). A framework for low energy application devices using blockchain-enabled IoT in WSNs. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 121-132). Cham: Springer International Publishing

[23] Parthiban, L., Latchoumi, T. P., Balamurugan, K., Raja, K., & Parthiban, R. (2023). Cognitive computing for the internet of medical things. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 85-100). Cham: Springer International Publishing

[24] Afrin Tisha, Faria, Nahiduzzaman, Md., and Kibria, Hafsa Binte. "Real-Time AQI Estimation: A Smart and Lightweight AI-Based System Using Gas Sensors." *Proc. Int. Conf. on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 2025.