Real Time Noise Pollution Prediction In A City Using Machine Learning And IOT

Abstract

The study focuses on addressing the growing concern about noise pollution due to increasing traffic, construction, etc. Noise pollution is a challenging environmental problem in dense urban areas and requires a holistic understanding of its causes and mitigation processes. Effective solutions are needed to minimize effects of noise pollution. The study includes various noise regression models and forecasting models to estimate noise pollution across the city. A dataset comprising various landscape configurations is collected from various IOT sensing devices (acoustic sensors) and the four regression models (Decision Tree, Gradient Boosting, Support Vector, Linear Regression with Grid Search) along with three forecasting models (ARIMA, Holt-Winters Exponential Smoothing, SARIMA) to predict noise pollution

I. Introduction

The rapid urbanization of cities worldwide has brought about numerous challenges, with noise pollution emerging as a significant issue. As urban areas continue to expand and become more densely populated, effectively managing noise pollution is crucial for the well-being of residents and the sustainability of urban development. This study explores a novel approach to addressing this challenge by using Machine Learning (ML) and the Internet of Things (IoT) to predict real-time noise pollution levels with in Vijayawada city.

Increasing demand for construction and transportation has created significant traffic congestion and safety challenges, which can negatively impact economic growth and quality of life. According to De Coensel, Brown and Tomerini, the frequency, duration and intensity of noise events from road traffic can contribute to noise pollution. Therefore, it is important to consider road traffic as the main source of noise pollution and take measures to reduce its impact. The negative effects of noise pollution on physical health, well-being and even mortality are widely recognized. Excessive sound exposure at home, school, work and other environments can affect concurrent activities and performance and also have long-term effects on human health and development.

Area/ Zone	Category of Area / Zone	Limits in dB(A) Leq Day Time (from 6.00 a.m. to 10.00 p.m.)	Limits in dB(A) Leq Night Time (from 10.00 p.m. to 6.00 a.m.)
(A)	Industrial Area	75	70
(B)	Commercial Area	65	55
(C)	Residential Area	55	45
(D)	Silence Zone	50	40

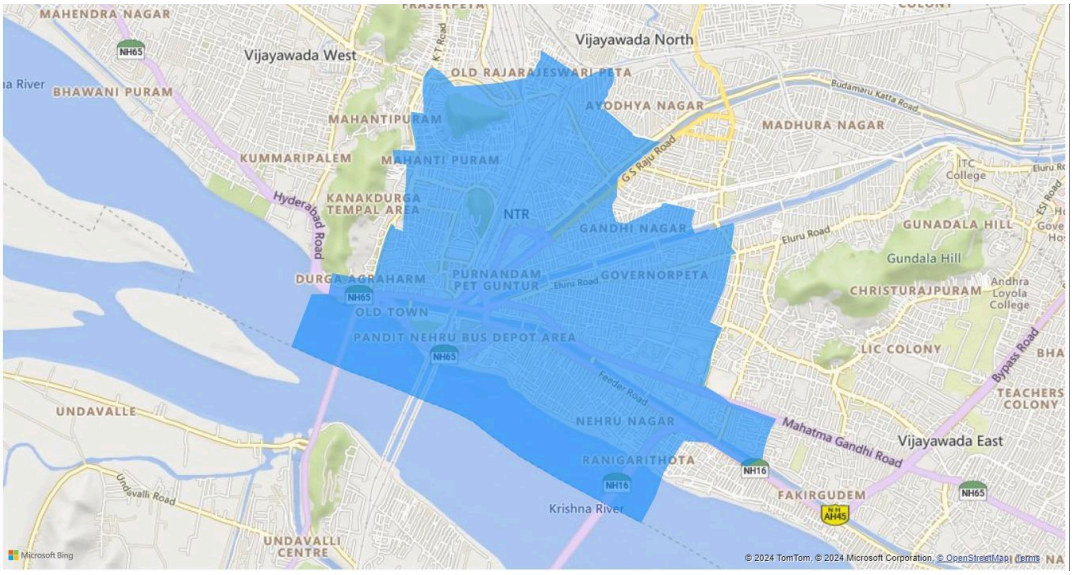
Developing efficient traffic management strategies to mitigate the effects of road traffic on the community necessitates precise road traffic data analysis. Evaluating noise pollution stemming from road traffic is essential for evaluating the environmental quality of urban areas and the well-being of their inhabitants.

By strategically placing IoT-enabled noise sensors throughout the urban landscape, continuous and detailed data on ambient noise levels can be collected. These sensors, capable of measuring various sound frequencies and intensities, form the basis of a comprehensive system designed to monitor and predict noise pollution in real-time. When combined with advanced ML algorithms, this system not only provides immediate insights into noise levels but also forecasts potential future trends, allowing city planners and residents to proactively address the impact of excessive noise on public health and urban quality of life.

Accurate noise event measurement and analysis can guide the formulation of effective noise reduction measures, such as noise barriers, traffic management tactics, and vehicle emission regulations. These initiatives can result in numerous advantages, including enhanced quality of life for residents, decreased stress and health issues, and a more sustainable urban setting. The ongoing research delves into the influence of various factors like the presence of heavy vehicles, time of day, landscape type, and distance from the road to the receiver point on traffic noise in diverse urban layouts.

This study delves into the various components of such a system, starting with the deployment and integration of IoT-enabled noise sensors across the city. The data collected is processed and stored in a centralized database, laying the groundwork for training and implementing ML models. These models, developed to predict noise levels based on factors like time, environmental conditions, and historical data, contribute to a dynamic and adaptable noise pollution prediction system.

By harnessing machine learning capabilities, the study seeks to forecast the impact of traffic noise pollution in Vijayawada. Urban noise levels have been categorized based on traffic composition for environmental noise evaluation. Through the utilization of expert systems, machine learning, and IoT, the research has showcased the potential of ML applications in addressing noise pollution concerns and collecting data for more informed decision-making against urban traffic noise via IoT.

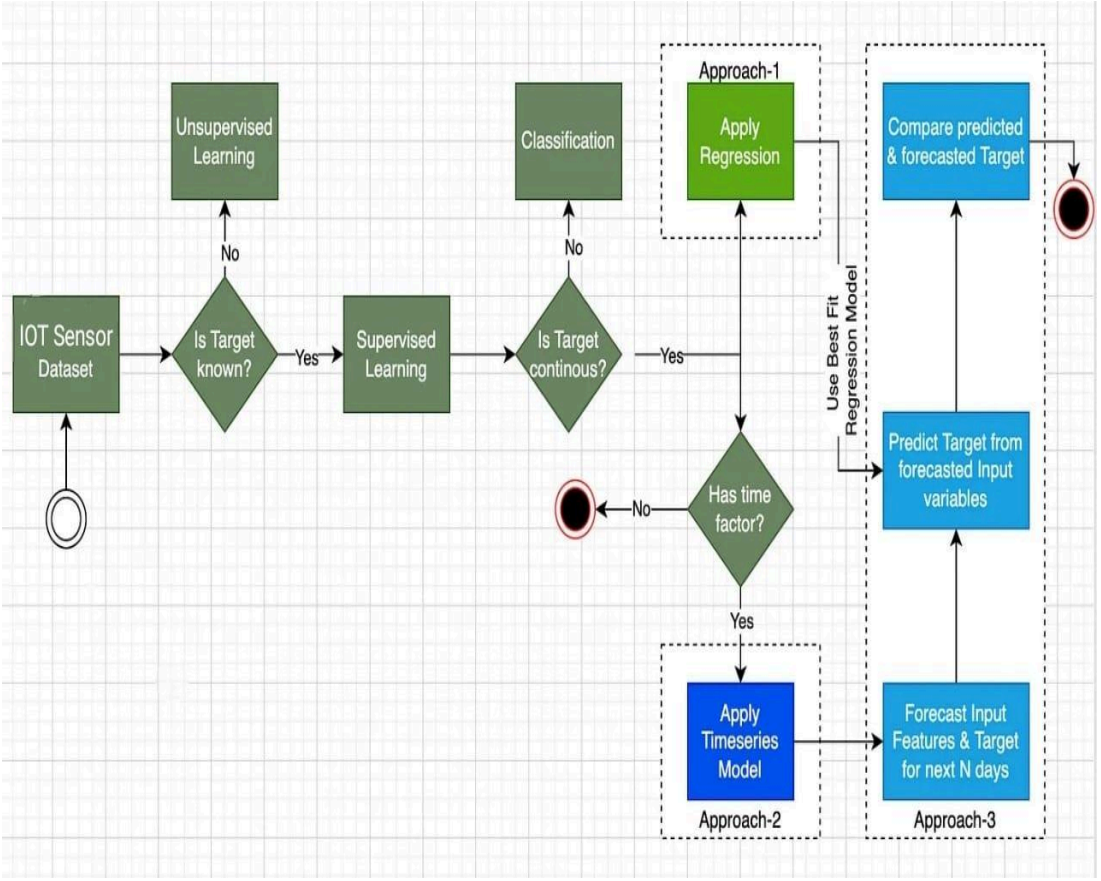


To accomplish this objective, a distinct dataset was curated specifically for the study to anticipate the effects of traffic noise in the city using IoT sound sensors that provide real-time data. The study compared the outcomes of seven machine learning algorithms for their accuracy in predicting traffic noise: Decision Tree, Gradient Boosting, Support Vector, Linear Regression with grid search, along with ARIMA, Holt-Winters Exponential Smoothing, and SARIMA. The models were fine-tuned by adjusting various parameters to identify the most effective approach.

The significance of this research lies in its ability to revolutionize urban noise management, providing decision-makers with actionable insights to implement targeted interventions. By leveraging the potential of ML and IoT technologies, cities can enhance their capacity to create more sustainable urban environments.

2. Method

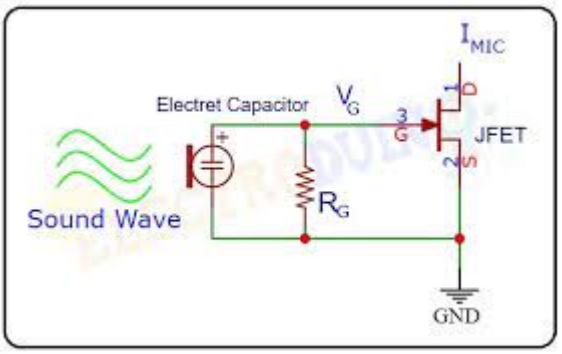
The purpose of this section is to outline the method used to collect the data and set the configurations needed to train and test the prediction models. This will involve the generation of the dataset, as well as the fine-tuning of the model parameters, to ensure the models are equipped with the information they need to accurately predict traffic noise



2.1 Description On IOT Dataset

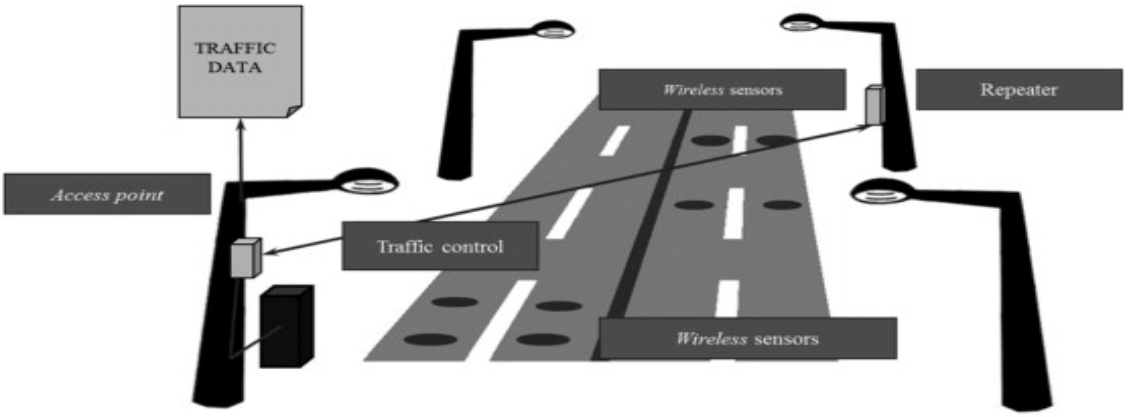
In the case study, simulations were conducted to evaluate how accounting for realistic power distributions in vehicle noise affects estimates of sound event characteristics. To achieve this goal, strategically place sound sensors in areas of interest, such as city streets or public spaces, to capture relevant sound data like noise levels, frequencies, timestamps, and location.

A sound sensor is a simple, user-friendly, and cost-effective device used to detect sound waves in the air. It measures sound intensity and converts it into an electrical signal that can be interpreted by a microcontroller. The sound sensor module has 4 pins: VCC, GND, Digital Out, and Analog Out. The AO pin provides analog readings, while the DO pin gives digital output.



The pinout of the sound sensor is as follows: VCC is the power supply pin, which can be connected to a 3.3V or 5V supply. The analog output varies based on the supply voltage. GND is the ground pin, DOUT is the Digital output pin indicating sound detection, and AOUT provides analog readings directly from the sensor.

Establish a connection between the sound sensors that have been deployed and an IoT platform. Each sensor should be equipped with communication capabilities to transmit data to the central IoT platform. This connection can be set up using various communication protocols. Configure the IoT platform to receive and manage data from the sound sensors, including setting up the necessary channels, protocols, and security measures for secure data transmission.



ThingSpeak is a commonly used IoT Cloud platform. ThingSpeak allows sensors, devices, and online platforms to transmit data to the cloud for storage in either a private or public channel. By default, ThingSpeak stores data in private channels, although public channels can be utilized for data sharing purposes. Ensure that the sound sensors regularly send real-time data to the IoT platform, including information such as sound intensity, frequency spectrum, sensor location, and timestamps. Implement checks to verify the accuracy and integrity of the collected sound data, which may involve validating sensor readings, handling missing or corrupted data, and maintaining data consistency.ThingSpeak is important to synchronize timestamps across all sensors to enable precise temporal analysis of the sound data, as inconsistent timestamps can lead to errors in interpreting the chronological order of events.

Features	Values	Number of Observations	Target Variable
Distance	15, 30, 45, 60, 75, 90 (m)	6480	Noise
Time	Day, Night		
Landscape	None, Tree, Wall		
Road surface	Asphaltic concrete, Uneven surface		
Vehicles/h	10, 20, 40, 50, 100, 200, 400, 500, 1000, 2000		
Speed limit	60, 80, 100, 120, 140 (km/h)		
Percentage of heavy vehicles	5, 10, 20 (%)		

2.2 Regression Models

The research aims to predict noise pollution levels generated by road traffic by analyzing various factors, including the distance from the road to receiver points, time of day, landscape barriers, road surface material, number of vehicles per hour, speed limit, and the percentage of heavy vehicles. To achieve this objective, the study employs and compares the results of four distinct regression models:Decision Tree Regressor, Gradient Boosting Regressor, Linear Regression with Grid Search, Support Vector Regressor.

a) Decision Tree Regressor

Decision Tree is a non-parametric supervised learning method used for classification and regression tasks. It works by recursively partitioning the dataset into subsets based on the value of attributes. The decision-making process involves selecting the best attribute at each step to create a tree-like structure, allowing for easy interpretation. Decision Trees are advantageous due to their interpretability, ability to handle both numerical and categorical data, and robustness against outliers. They are commonly used in various fields such as finance, healthcare, and customer relationship management.

b) Gradient Boosting Regressor

Gradient Boosting Regressor is an ensemble learning method that combines the predictions of several base estimators, typically decision trees, to improve accuracy. It works by sequentially fitting new models to the residuals of the previous models, optimizing a differentiable loss function. The main advantages of Gradient Boosting Regressor include its high predictive power, ability to handle complex interactions in the data, and robustness to outliers. This algorithm is often utilized in areas such as web search ranking, anomaly detection, and ecological modeling.

c) Linear Regression with Grid Search

Linear Regression with Grid Search is a basic yet effective regression method that finds the best hyperparameters through an exhaustive search over a specified parameter grid. It fits a linear relationship between the independent and dependent variables by minimizing the residual sum of squares. The advantages of this method lie in its simplicity, ease of implementation, and suitability for large datasets. It is commonly used in fields like economics, social sciences, and engineering.

d) Support Vector Regressor

Support Vector Regressor is a supervised learning algorithm that uses support vector machines for regression tasks. It works by finding the hyperplane that best fits the data while maximizing the margin between the data points and the hyperplane. The advantages of Support Vector Regressor include its effectiveness in high-dimensional spaces, robustness to outliers, and ability to capture complex relationships in the data. It is commonly used in areas such as stock market forecasting, medical diagnosis, and geophysical data analysis.

Support Vector Regressor is a supervised learning algorithm that uses support vector machines for regression tasks. It works by finding the hyperplane that best fits the data while maximizing the margin between the data points and the hyperplane. The advantages of Support Vector Regressor include its effectiveness in high-dimensional spaces, robustness to outliers, and ability to capture complex relationships in the data. It is commonly used in areas such as stock market forecasting, medical diagnosis, and geophysical data analysis.

These algorithms offer a diverse set of tools for addressing regression tasks, each with its own advantages and potential use cases in various domains. When applied thoughtfully, they can contribute significantly to the development of accurate and robust predictive models.

2.3 Time Series Models

ARIMA

Autoregressive Integrated Moving Average (ARIMA) is a popular time series analysis and forecasting method that models the next step in the sequence as a linear function of the differenced observations and lagged error terms. ARIMA models are denoted with the notation ARIMA(p, d, q), where p, d, and q are the order of the autoregressive, differencing, and moving average terms, respectively. The "AR" part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The "I" (for "integrated") indicates that the data values have been differenced - i.e., the value of a time series data point is subtracted from its preceding value. The "MA" part indicates that the regression error is a linear combination of error terms whose values occurred contemporaneously and at various times in the past. ARIMA models are widely used for forecasting future data points in various fields such as finance, economics, and environmental science. Due to its flexibility and capability to handle a wide range of time series behaviors, ARIMA remains a fundamental tool in the field of time series analysis and forecasting.

Holt-Winters Exponential Smoothing

Holt-Winters exponential smoothing is a popular time series forecasting method that captures the level, trend, and seasonality in the data. This method is particularly useful for data with a trend and/or seasonal patterns. The Holt-Winters method involves three smoothing equations: one for the level, one for the trend, and one for the seasonal component. The level equation updates the current level of the series, the trend equation updates the current trend, and the seasonal equation updates the seasonal indices. By updating these components, the method can provide accurate forecasts for future time points. Holt-Winters exponential smoothing has found applications in various domains such as finance, inventory management, and demand forecasting due to its ability to effectively capture and forecast time series data exhibiting trend and seasonal patterns. This method has proven to be a valuable tool for analysts and forecasters seeking to make accurate predictions based on historical time series data.

SARIMA

Seasonal Autoregressive Integrated Moving-Average, often denoted as SARIMA, is a time series forecasting model that extends the capabilities of the ARIMA model to account for seasonality in the data. SARIMA models are suitable for analyzing and forecasting time series data that exhibit both non-seasonal and seasonal patterns. The model is specified with the notation SARIMA(p, d, q)(P, D, Q)s, where (p, d, q) are the non-seasonal components, (P, D, Q) are the seasonal components, and s represents the length of the seasonal cycle. The non-seasonal components (p, d, q) represent the autoregressive, differencing, and moving average terms, while the seasonal components (P, D, Q)s capture the seasonal autoregressive, seasonal differencing, and seasonal moving average terms, respectively. SARIMA models are widely used in fields such as economics, finance, and climate science, where the data exhibits

both trend and seasonal patterns. By incorporating both non-seasonal and seasonal parameters, SARIMA provides a flexible framework for modeling and forecasting complex time series data, making it a valuable tool for analysts and researchers seeking to accurately forecast seasonal time series data.

2.4. Model Evaluation

The study assesses the performance of various regression models in predicting traffic noise levels at the New Klang Valley Expressways in Shah Alam, Malaysia. Through the use of six evaluation criteria, the research identifies the Random Forest (RF) model as the most effective for this purpose. Additionally, the study underscores the potential of Land Use Regression (LUR) models for noise mapping in scenarios with limited noise data. It also emphasizes the importance of proper validation against long-term traffic data and introduces a methodology that can be expanded to improve the accuracy of traffic noise prediction maps, while acknowledging the study's limitations and scope.

Three performance metrics are employed to measure the models' effectiveness: mean square error (MSE), root mean square error (RMSE), and the coefficient of determination (R2). These metrics offer a thorough evaluation of the models' accuracy and assist in identifying the best model among the Regression Algorithms Decision Tree, Gradient Boosting Regressor, Lasso Regressor, Linear Regression with Grid Search, Ridge Regression, and Support Vector Regressor and Time Series Analysis ARIMA, Holt-Winters Exponential Smoothing, and SARIMA. Moreover, these measures have been extensively used in similar studies, enabling a comparison of our results with state-of-the-art research.

MSE

The Mean Square Error (MSE) is a metric that quantifies the average of the squared variances between actual and estimated values. In the context of noise-pollution prediction models, a lower MSE suggests that the model's predictions are more closely aligned with the actual values, indicating higher accuracy.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

RMSE

The Root Mean Square Error (RMSE) provides insight into the typical deviation of the model's predictions from the actual values in the same units as the response variable. This makes it a valuable measure for understanding the practical predictive performance of the model.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

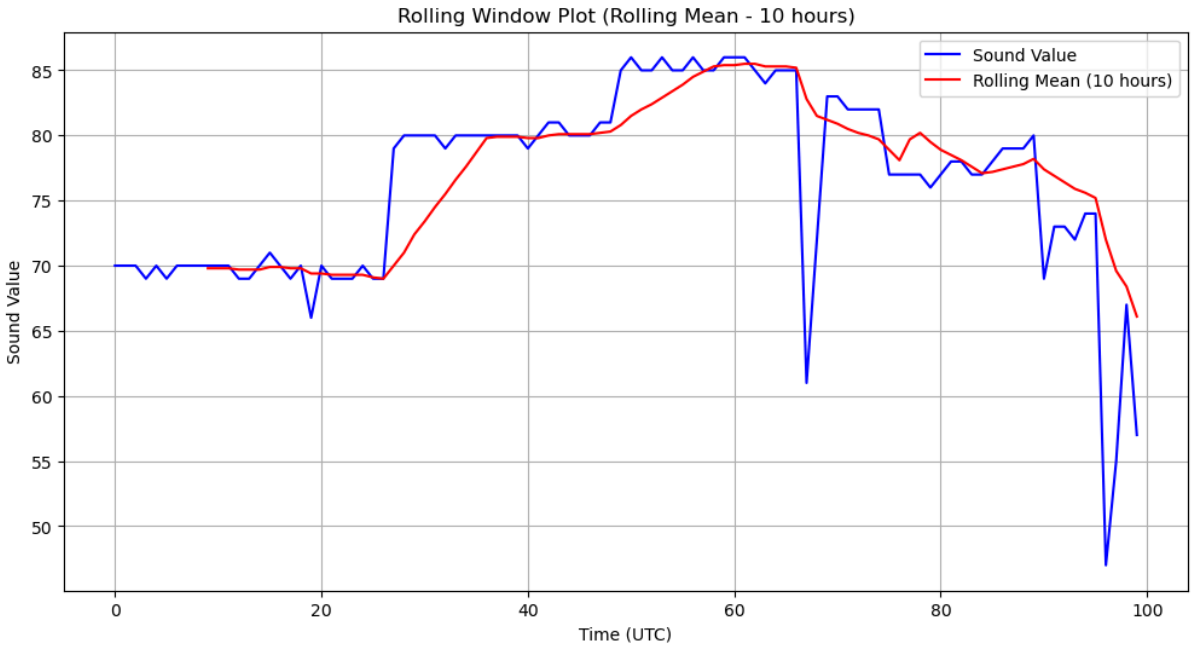
R SQUARE:

The R-squared (R2) value, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that is accounted for by the independent variable(s) in a regression model. In noise-pollution prediction models, a higher R-squared value signifies a stronger fit between the model and the observed data.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where: y_i - true values of y
$\hat{y}_i$ - predicted values of y
$\bar{y}$ - average value of y

3. Results and Analysis:



This details the results of each Regression model, encompassing the assessment of Mean Square Error (MSE), Root Mean Square Error (RMSE), and the coefficient of determination (R2) for each Regression model.

MODEL	Sum of MSE	Sum of RMSE	Sum of ACCURACY	Sum of PREDICTED NOISE
Decision Tree Regressor	4.73	2.18	0.97	69.67
Gradient Boosting Regressor	4.63	2.15	0.97	69.70
Lasso Regressor	28.71			
Linear Regression with grid search	25.39	5.04	0.93	74.68
Ridge Regressor	28.71			
Support Vector Regressor	6.43	2.15	0.97	70.30
Voting Regressor	15.54			

This details the results of each Time Series Analysis Models,, encompassing the assessment of Mean Square Error (MSE), Root Mean Square Error (RMSE), and the coefficient of determination (R2) for each model.

MODEL	Sum of RMSE	Sum of R^2 Score
ARIMA	10.26	-0.34
Holt-Winters Exponential Smoothing	9.43	-0.13
SARIMA	10.46	-0.39

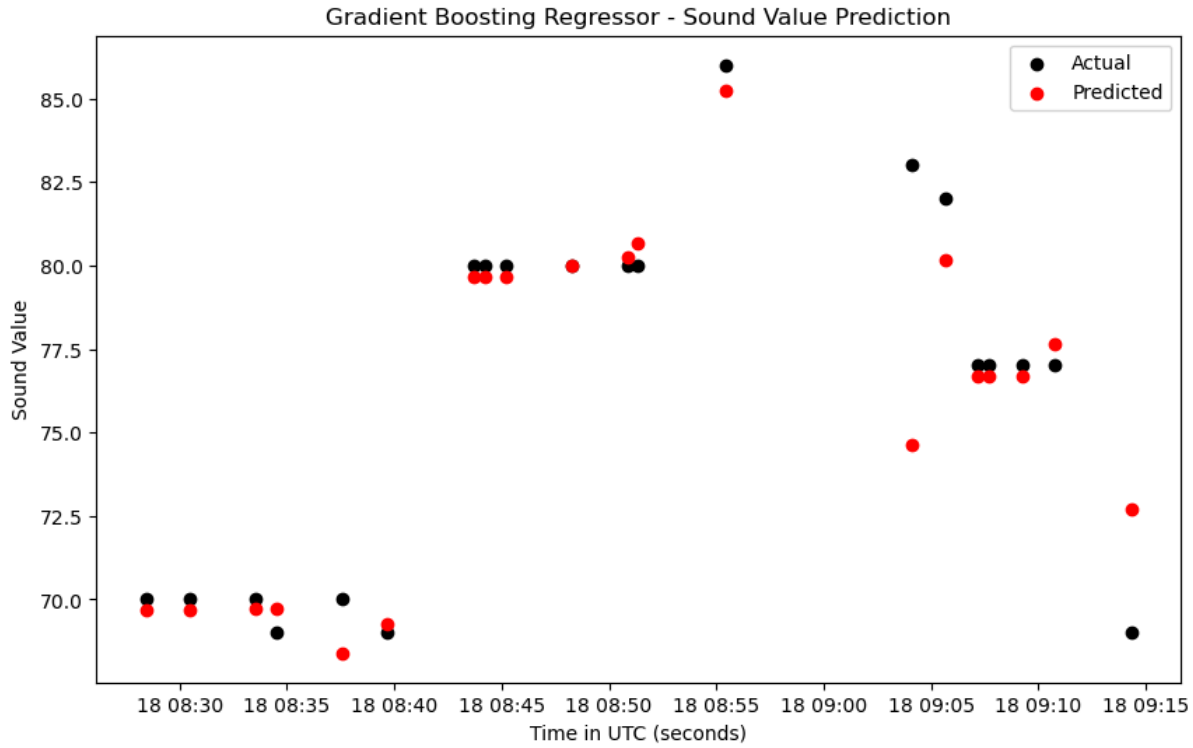


Fig 1: RMSE=2.15,MSE=4.63,ACCURACY=0.97

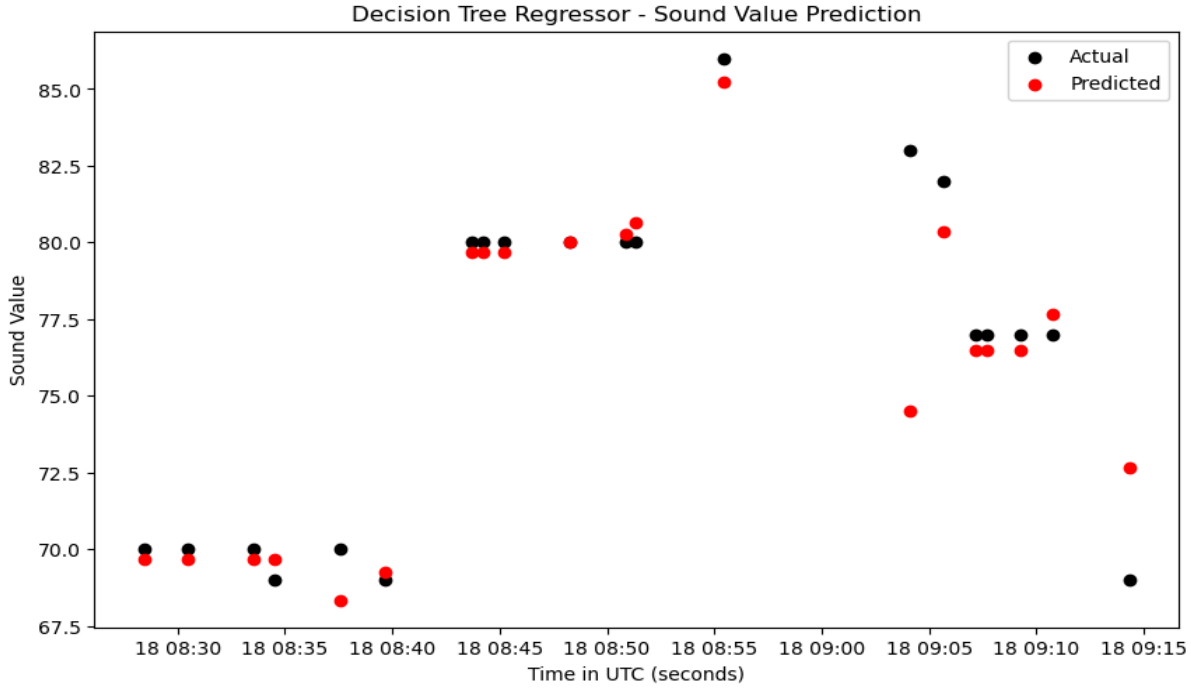


Fig 2: RMSE=2.18,MSE=4.73,ACCURACY=0.97

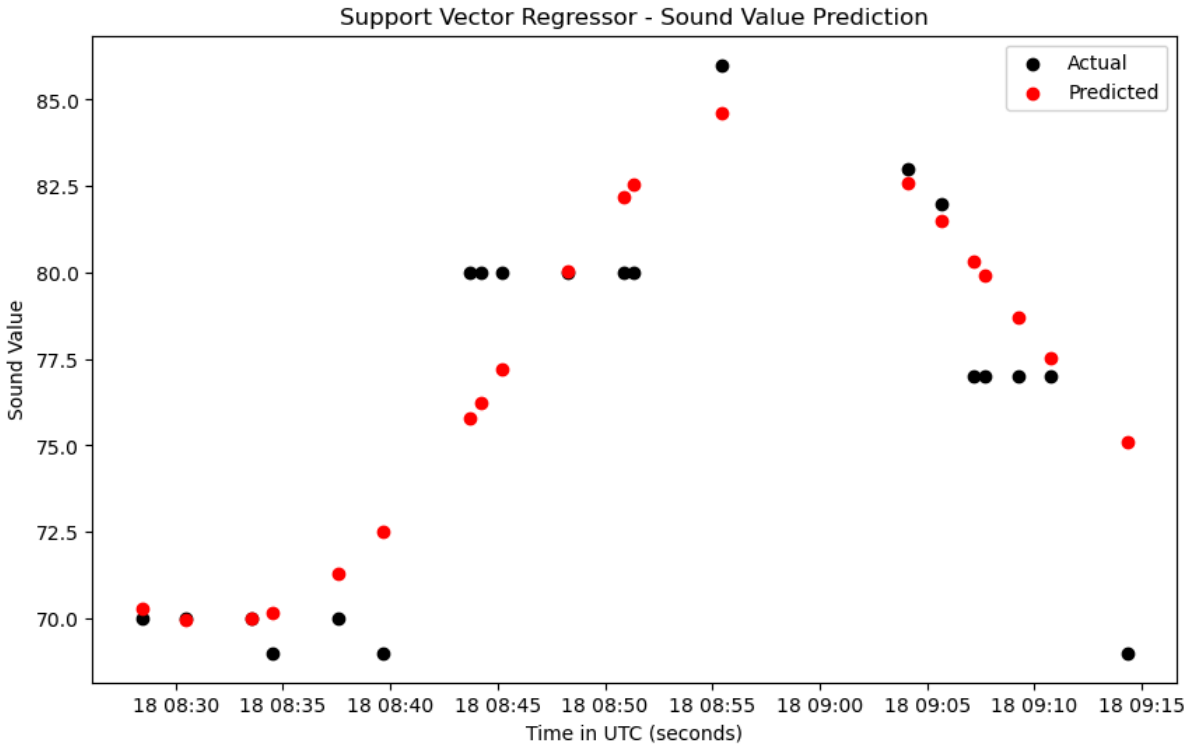


Fig 3: RMSE=2.15,MSE=6.43,ACCURACY=0.97

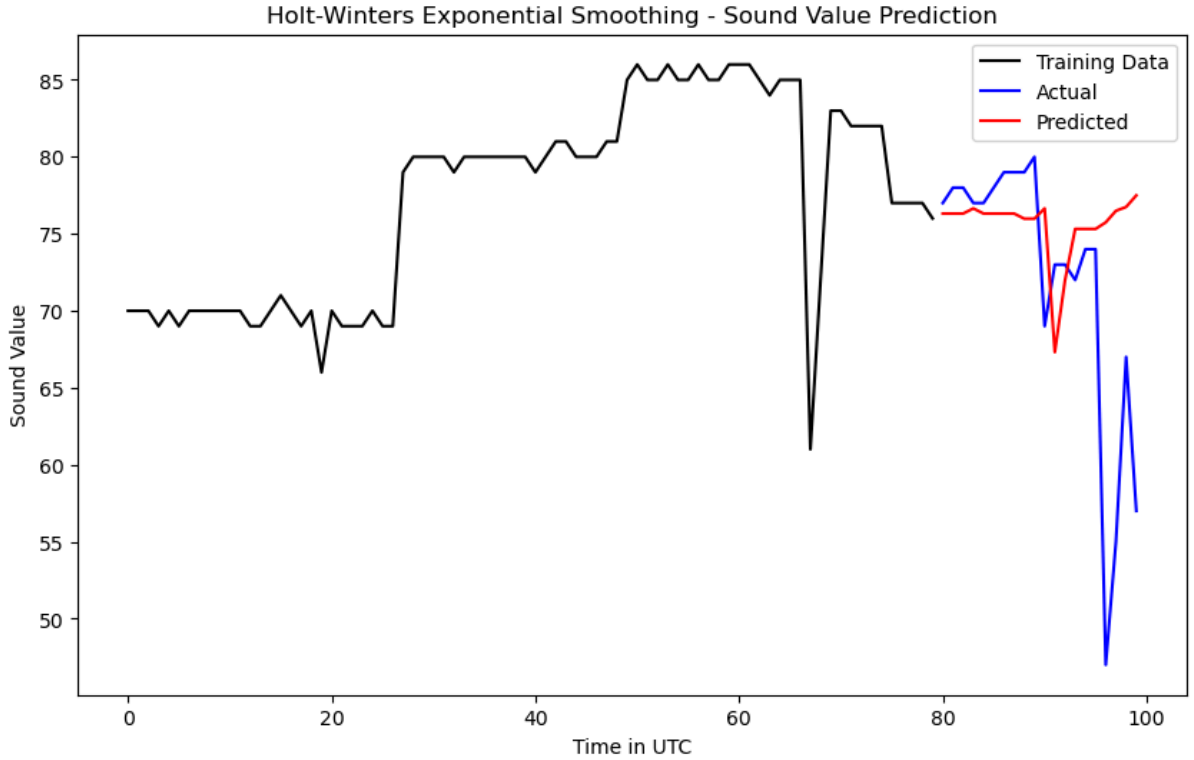


Fig 4: RMSE=9.43,R-Squared= -0.13

CONCLUSION:

In this study, we examined the estimation of noise pollution levels caused by road traffic by analyzing various factors that contribute to it. Our objective was to develop accurate predictive models by considering factors such as the distance between the road and receiver points, time of day, landscape barriers, road surface material, number of vehicles per hour, speed limit, and the percentage of heavy vehicles.

We compared four different regression models - Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Gradient Boosting Regression - to determine the most effective method for predicting noise pollution. Each model provided valuable insights into the relationships between predictor variables and noise pollution levels, offering unique advantages and considerations.

Our findings emphasize the complexity of noise pollution dynamics and the importance of comprehensive modeling techniques in understanding and forecasting its effects. While Linear Regression offered simplicity and interpretability, Random Forest Regression, SVR, and Gradient Boosting Regression demonstrated superior predictive capabilities, particularly in capturing non-linear relationships and interactions among variables.

The choice of utilizing the Gradient Boosting regressor model should be based on the specific characteristics of the dataset, with a goal of achieving a predictive accuracy of 97.18% while also taking into account the interpretability of the findings. It is important to acknowledge that there is an anticipated 29.51% increase in noise levels for the next year. By leveraging the findings of this study, urban planners, policymakers, and environmental scientists can better tackle noise pollution challenges and strive towards developing more vibrant and sustainable communities.

REFERENCES

[1] J. Shin, Y. Kim, S. Yoon, and K. Jung, ”Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification,” 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 2018, pp. 491-494, doi: 10.1109/BigComp.2018.00079.

[2] W. Lu, Y. Duan, and Y. Song, ”Self-Attention-Based Convolutional Neural Networks for Sentence Classification,” 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 2020, pp. 2065-2069, doi: 10.1109/ICCC51575.2020.9345092.

[3] J. Yang and J. Yang, ”Aspect Based Sentiment Analysis with SelfAttention and Gated Convolutional Networks,” 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2020, pp. 146-149, doi: 10.1109/ICSESS49938.2020.9237640.

[4] Feng, X. Zhang, and X. Song, ”Unrestricted Attention May Not Be All You Need–Masked Attention Mechanism Focuses Better on Relevant Parts in Aspect-Based Sentiment Analysis,” in IEEE Access, vol. 10, pp. 8518-8528, 2022, doi: 10.1109/ACCESS.2022.3142178.

[5] Pamungkas, E.W., Basile, V., Patti, V. ”Towards multidomain and multilingual abusive language detection: a survey.” Pers Ubiquit Comput 27, 17–43 (2023). https://doi.org/10.1007/s00779-021-01609-1.

[6] R. Amrutha and K. R. Bindu, ”Detecting Hate Speech in Tweets Using Different Deep Neural Network Architectures,” 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 923-926, doi: 10.1109/ICCS45141.2019.9065763.

[7] Asudani, D.S., Nagwani, N.K. Singh, P. ”Impact of word embedding models on text analytics in deep learning environment: a review.” Artif Intell Rev 56, 10345–10425 (2023). https://doi.org/10.1007/s10462-023-10419-1.

[8] F. A. Rawther and G. Titus, ”Transformer Models for Recognizing Abusive Language An investigation and review on Tweeteval and SOLID dataset,” 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 1-6, doi: 10.1109/ICEEICT56924.2023.10157848.

[9] Gampala, Veerraju Vallapuneni, Jaideep Kumar, Ande Indurthi, Ravindra Nichenametla, Rajesh. (2021). ”Comparative Study on Telugu text Classification using Machine Learning and Deep Learning models.” 1393-1398. 10.1109/ICOEI51242.2021.9453040.

[10] L. Ketsbaia, B. Issac and X. Chen, ”Detection of Hate Tweets using Machine Learning and Deep Learning,” 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 2020, pp. 751-758, doi: 10.1109/TrustCom50675.2020.00103.

[11] B. Pariyani, K. Shah, M. Shah, T. Vyas and S. Degadwala, ”Hate Speech Detection in Twitter using Natural Language Processing,” 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1146-1152, doi: 10.1109/ICICV50876.2021.9388496.

[12] A. Salehgohari, M. Mirhosseini, H. Tabrizchi and A. V. Koczy, ”Abusive Language Detection on Social Media using Bidirectional Long-Short Term Memory,” 2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES), Georgioupolis Chania, Greece, 2022, pp. 000243-000248, doi: 10.1109/INES56734.2022.9922628.

[13] R. M. Alhejaili, W. M. S. Yafooz and A. A. Alsaeedi, ”Hate Speech and Abusive Language Detection In Twitter and Challenges: Review,” 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2022, pp. 86-94, doi: 10.1109/CISES54857.2022.9844317.

[14] Marreddy, Mounika Oota, Subba Vakada, Sireesha Chinni, Venkata Charan Mamidi, Radhika. (2022). ”Multi-Task Text Classification using Graph Convolutional Networks for Large-Scale Low Resource Language.” 1-8. 10.1109/IJCNN55064.2022.9892105.

[15] M. Marreddy, S. R. Oota, L. S. Vakada, V. C. Chinni and R. Mamidi, ”Clickbait Detection in Telugu: Overcoming NLP Challenges in Resource-Poor Languages using Benchmarked Techniques,” 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534382.

[16] Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., Mamidi, R. (2022). ”Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP tasks in Telugu Language.” Transactions on Asian and Low-Resource Language Information Processing. ACM New York, NY.
[17] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, “A review of irregular time series data handling with gated recurrent neural networks,” Neurocomputing, vol. 441. Elsevier BV, pp. 161–178, Jun. 2021, doi: 10.1016/j.neucom.2021.02.046