Air Quality Prediction Using Machine Learning Algorithms

Ayush Gupta[1], Bhagyashree Maharana[2], Dr Yojna Arora[3], Dr Ashima Rani[4]

[1]Department of Computer Science and Engineering, Sharda School of Engineering and Technology, Greater Noida, U.P., India. ayugup1999@gmail.com

[2] Department of Computer Science and Engineering, Sharda School of Engineering and Technology, Greater Noida, U.P., India; bhagyashreemaharana4@gmail.com

[3]Department of Computer Science and Engineering, Sharda School of Engineering and Technology, Greater Noida, U.P., India. yojana183@gmail.com

[4]Associate Professor, SGT University, Gurugram; Ashima.ashugambhir@gmail.com

***Corresponding Author:** Yojna Arora (yojana183@gmail.com)

Abstract

As a result of rapid urbanization coupled with an increase in population in India, there is a substantial increase in air pollution that poses a risk to public health as well as environmental sustainability. This paper work looks into how the Air Quality Index (AQI) is estimated machine learning algorithms like Random Forest, Linear Regression, and XGBoost. The research employs a sophisticated dataset with metrics on pollution related parameters such as PM2.5, NO2, and SO2. In addition, novel preprocessing techniques are executed and the accuracy of the model is evaluated. With the ability to accurately predict short-term, long-term, and sudden changes in Air Quality index (AQI), these models help in making policy decisions in a timely manner. This research elucidates the gaps this paper tries to fill in the existing studies concerning techniques that improve air quality prediction as well as combat urban air pollution. Integrating machine learning with air quality management provides great insights into defeating public health challenges alongside making reliable and dynamic forecasting of AQI. Using and developing techniques to combat different facets of urbanization is vital for proper environmental management.

Keywords

Air Quality Index (AQI), Machine Learning, Random Forest, XGBoost and Environmental Forecasting

1. Introduction

The increasing population in India has resulted in an unprecedented level of defilement (mostly including CO2, pm2.5 particulate matter, amongst others) and other dangerous pollutants. The air quality in a particular area surrounding a state or a country is a rough estimate of the impact of the defilement in a particular zone. According to the contours of Indian air pollution standards, contaminants are defined by PM10 and PM2.5 Transmittable Suspended Particulate Matter, these air quality indexes serve as a reference to the level of pollutant in the airspace. Different gases emit from various sources which deposits on the air and contaminates the level of pollution of the surrounding [1]. Each type of pollution has different ports of origin and must be measured using different equipment. The primary air pollutants NO2 and SO2, RSPM (Respirable Suspended Particulate Matter) and other primary pollutants allows the determination of AQI. it is possible to determine the values of air quality index from seperate data. Anticipating the air quality index enables the determination of the major pollution causing contaminant and the area across India that is suffering from severe defilement. A large portion of developing countries suffers greatly from air pollution due to the rapid expansion of urban regions [2]. No matter where in the world, all government bodies have a common responsibility - forecasting future changes in air quality. As part of our research work, we seek to project air quality at each air quality monitoring station by utilizing the existing data. We hold the air, realm, and all the contaminants in the atmosphere in a neural network (NN) based approach, which integrates local alternation, a deep distribution network, and Unified. The former alters the coarse grained air quality data into translation equivalents or pollutants while the latter, which inputs neural shifted proportions, processes the complex data within the city simultaneously to attain the weather elements that influence air quality [3]. Every hour, our system has smart information for 300++ Chinese cities. The advantages of Deep Air in terms of data results from three to nine years for a city test show the first ten ways that were not available. With limited accuracy forecasting short, long term, and sudden alterations, we improve for the prior cyberspace technique by 2.4%, 12.2%, and 63.2% respectively.

1.1 Problem Statement

- As the population and urban areas grow at a much faster pace in India, the amount of air pollution caused by CO2, PM2.5, NO2, SO2, and RSPM emissions has gone up significantly. This pollution is a threat to both human life and the environment, which is why reliable methods to predict air quality are required.
- There seems to be a major gap in the previously developed approaches pertaining to air quality prediction as they fail to provide the necessary accuracy and efficacy to combat the problems revolving around urban air pollution. The concern rises because there is an imminent need for modern prediction models which can compute the short-term, long-term changes, and even sudden shifts in air quality so that the policymakers know how to address the ever-changing air pollution challenges.

2. Literature Review

Author	Year	Title	Technique used	Findings
TinkuSingh, NikhilSharma, Satakshi &Manish Kumar	05 Jan 2023	Analysis and forecasting of air quality index based on satellite data	The study utilized satellite data from Google Earth Engine to analyze AQI changes during COVID-19 lockdowns in India, employing Holt-Winter and LSTM variants for short-term AQI forecasting, with Holt-Winter yielding the lowest MAPE scores.	1.AQI ranged from 100 to 300, showing moderate to very poor air quality. 2. The lockdown in 2020 led to the most substantial reduction in AQI. 3. Short-term AQI forecasting with Holt-Winter was the most accurate, with the lowest MAPE scores.
Dong-HerShih, To ThiHien, Ly Sy Phu Nguyen, Ting-Wei Wu. Yen-Ting Lai	25 Aug 2022	“AModified γ-Sutte Indicator for Air Quality Index Prediction”	The technique used in this study is time series analysis, focusing on predicting Air Quality Index (AQI) values. It introduces and compares different versions of the Sutte indicator (α-Sutte, β-Sutte, and γ-Sutte) for this purpose, alongside ensemble modeling and comparison with other methods like ARIMA.	1. A modified γ-Sutte indicator is proposed for air quality prediction, balancing computational efficiency and superior prediction performance compared to α-Sutte and β-Sutte indicators. The γ-Sutte indicator shows strong predictive ability across different evaluation metrics and geographical areas, suggesting its effectiveness as an air quality predictor.
DipshaParesh Shah, Dr.Piyushkumar Patel	4 Nov 2021	comparison between national air quality index, India and composite air quality indexAhmedabad, India.	1. Pollutant Dynamics: Ahmedabad sees higher winter pollutant levels, notably PM10, PM2.5, and CO, exceeding standards due to weather conditions. Indexing Systems: CAQI outperforms NAQI by considering multiple pollutants simultaneously, aiding	1. Ahmedabad sees higher winter pollutant levels, mainly PM10, PM2.5, and CO, exceeding NAAQS limits due to weather. Summers stay within limits. 2. The fuzzy-based Composite Air Quality Index (CAQI)

			pollution control efforts.	outperforms the National Air Quality Index (NAQI), offering a better understanding of multiple pollutant exceedances and aiding pollution control efforts.
SandroRodriguez Garzon, Marcel Reppenhagen,Marcel Müller	25 Feb 2022	What if Air Quality DictatesRoad Pricing? Simulation of an Air Pollution-based Road Charging Scheme	1. Simulation Analysis: Employed to assess the effects of a dynamic air quality-based charging scheme for Berlin's metropolitan region, particularly focusing on road usage charges and vehicle emissions. Integrated Approach Recommendation: Suggests integrating the dynamic charging scheme with broader context parameters such as traffic volume, road type, and vehicle occupancy to enhance its effectiveness in reducing urban air pollution.	1. Dynamic Charging Scheme: The study examined a dynamic air quality-based toll system for Berlin, revealing a moderate emissions rise with rerouted vehicles due to varying pollution levels within toll areas. Future Considerations: While initial simulations offered insights into charges and infrastructure effects, further investigation is needed to gauge emissions reduction success. Integrating the scheme with broader traffic and vehicle factors could enhance its effectiveness.
Quang Cuong Doan, Chen Chen, Shenjing He, Xiaohu Zhang	2 Jan 2024	How urban air quality affects land values: Exploring non-linear and threshold mechanism using explainable artificial intelligence	1. Variable Selection: Stepwise OLS regression and multicollinearity analysis were used to select relevant variables affecting land values in New York City. Modeling:Thestudy employed OLS regression, spatial lag model (SLM), and spatial error model (SEM) to explore the non-linear relationship between air pollution and house prices, providing insights into the impact of pollution on property values.	1. Air Pollution and House Prices: The study observed a non-linear connection between air pollution and house prices in New York City, indicating that fluctuations in pollution levels impact housing values. Model Comparison: By employing OLS regression, multicollinearity analysis, SLM, and SEM, the study assessed different modeling techniques' performance in addressing spatial auto-correlation when

				predicting house prices.
Giovanni Gualtieri, Lorenzo Brilli,FedericoCarotenuto, Carolina Vagnoli, Alessandro Zaldei , Beniamino Gioli,	23 Sept 2020	Quantifying road traffic impact on air quality in urban areas: A Covid19-induced lockdown analysis in Italy	The techniques used include comparative analysis of air quality and traffic data during lockdown, data analysis of daily time series, and normalization with correlation to ensure representativeness.	Lockdown Effect: Italy's Covid-19 lockdown slashed road traffic, notably cutting nitrogen dioxide (NO2) levels across urban, rural areas, and motorways. Pollution Complexity: Despite traffic drops, fine particulate matter (PM2.5) and PM10 levels saw little change, emphasizing the intricate nature of pollution and the need for ongoing efforts to improve air quality and public health.
SethA.Horn,PurnenduK.Dasgupta	5 Oct 2023	The Air Quality Index (AQI) in historical and analytical perspective a tutorial review	1. Comparative analysis of air quality data using the Air Quality Index (AQI) and reference measurement methods. Examination of air quality data for the Dallas-Fort Worth (DFW) area as a case study.	1. Air quality data, when presented without context, may lack meaningful interpretation regarding health risks associated with pollutants. The AQI serves as a valuable tool in communicating air quality information to the public by condensing complex data into a single index with a color-coded format.
Mariantonietta Ruggieri, Antonella Plaia	15 March 2012	An aggregate AQI: Comparing different standardizations and introducing a variability index	The techniques used in this study involve the development of a relative index of variability associated with an aggregate Air Quality Index (AQI) and the evaluation of different standardization methods for air pollution data.	A proposed variability index clarifies individual pollutants' influence on AQI, emphasizing chronic health effects and long-term environmental damage. Study applies indices to simulated and real air pollution data, offering insights for policymakers and the public.
Abdelfettah Benchrif, Ali Wheida, Mounia Tahri, Ramiz	14July 2021	Air quality during three covid-19 lockdown phases: AQI, PM2.5 and NO2	Techniques included statistical analysis, correlation studies between pollutant levels and lockdown measures, and examination of local emission sources and policy responses to assess the impact on PM2.5,	1. COVID-19 lockdowns significantly reduced air pollution, shifting AQI from high to mild pollution levels compared to pre-lockdown conditions.

M. Shubbar, Biplab Biswas		assessment in cities with more than 1 million inhabitants	NO2, and AQI levels.	2. PM2.5 levels were notably higher before lockdown in half of the cities studied, indicating substantial reductions during lockdown periods.
S. Dubey, M. K. Singh, P. Singh, and S. Aggarwal,	17 Aug 2020	Waste Management of Residential Society using Machine Learning and IoT Approach	techniques including SVM, NB, RF, DT, and KNN for predicting waste management alerts.	IoT and machine learning-based waste management system in residential societies improves waste management efficiency, achieving an 85.29% accuracy rate in predicting waste management alerts.
D. Zhang and S. S. Woo,	11 May2020	Real Time Localized Air Quality Monitoring and Prediction through Mobile and Fixed IoT Sensing Network,	techniques like SVM, NB, RF, DT, and KNN are employed to predict waste management alerts, with the RF algorithm achieving the highest accuracy of 85.29%.	The RF algorithm achieved the highest accuracy of 85.29% in predicting waste management alerts using machine learning classification techniques.
Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici	18 jan 2021	Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction,	The techniques used in this research include a multi-point spatio-temporal deep learning model based on ConvLSTM for air pollution forecasting and uncertainty quantification techniques to enhance reliability.	A ConvLSTM model outperforms seven methods for air pollution prediction, with uncertainty estimates enhanced by techniques like MC dropout and quantile regression.

3. Methodology

3.1 Data Set Overview:

The dataset in this research comprises the information on the Air Quality Index (AQI) of different states. It consists of 16 attributes: city, day, PM2.5, PM10, NO, NO2, NH3, CO, SO2, etc. The features mentioned here are the basis on which the AQI values depend, and the dataset contains 18314 records with 10 columns. The dataset was divided into 80% training and 20% testing, and the data underwent several preprocessing methods such as imputation and normalization for the purpose of analysis.

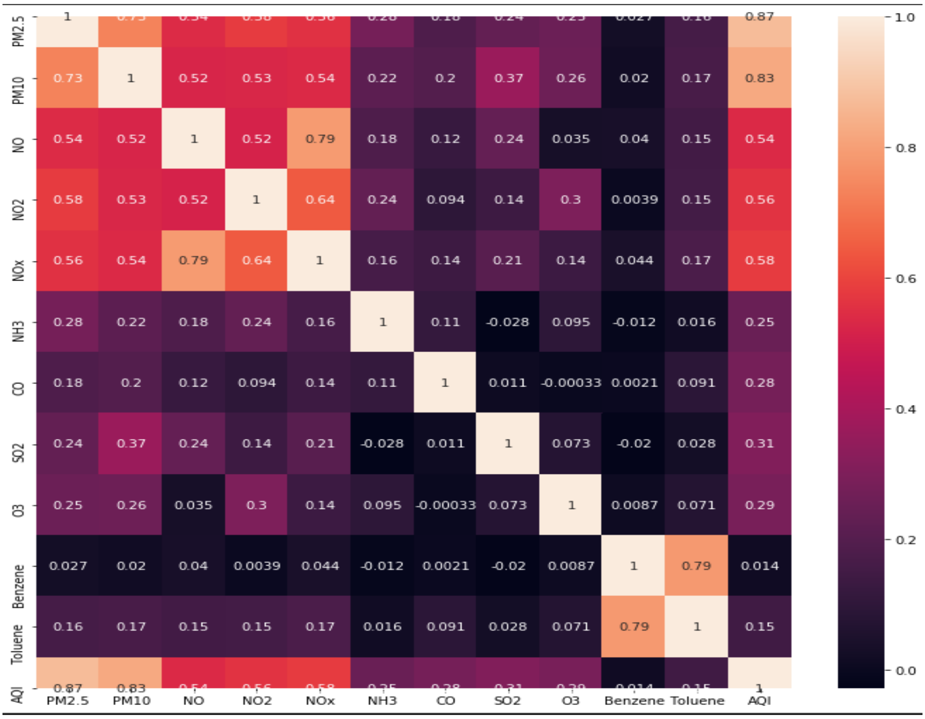

Fig1. DatasetOverview

3.2 Algorithms Used:

3.2.1 Random Forest:

Random Forest is a type of ensemble learning that builds multiple decision tress when training. Training is done with random selections from the dataset. The outputs are then given by taking an average or a vote from the decision from each of the trees [4]. It increases accuracy by reducing overfitting that may happen in individual models. Random Forest has feature importance measures that quantify the value of features for explaining the target. It is strong, salable, and tolerant to noise which makes it good when dealing with large datasets with lots of features [5].

3.2.2 Linear Regression:

Linear regression is a model in which a linear equation is fitted to related variables. This model assumes linearity between dependent and independent variables where the latter predicts the former [6]. The objective is to determine a straight line that fits best to the data in the deterministic sense. This line is defined with its slope and y-intercept. Linear regression is commonly applied for predictions, inferences, and other relationships in economics, finances and social sciences [7].

3.2.3 XGBoost:

In essence, XGBoost or Extreme Gradient Boosting has become an unrivalled name in the realm of machine learning algorithms [8]. It is a form of gradient boosted tree and it has an unrivaled combination of speed and performance. XGBoost works by building multiple decision trees in an iterative manner, with each new tree correcting the errors made by the previous ones [9]. Mechanisms of regularization are employed in order to optimize the algorithms for speed and accuracy without overfitting the model. Its efficiency and versatility has made it a staple XGBoost for numerous applications and in over the XGBoost has gained a lot of popularity in machine learning competitions [10].

3.2.4 Method Used:

Using the dataset from data.gov.in which contained factors related to the AQI, we carried out air quality prediction. The dataset underwent preprocessing via Python in Jupyter Notebook, such as the replacing of missing data and removal of uncorrelated features that were not important [11]. The dataset was separated into training (80%) and testing (20%) portions, this also normalized the data. To make predictions with respect to the AQI, Linear Regression, XG Boost and Random Forest algorithms were implemented [12]. Models were built and tested in order to measure the accuracy of the AQI values for the dataset.

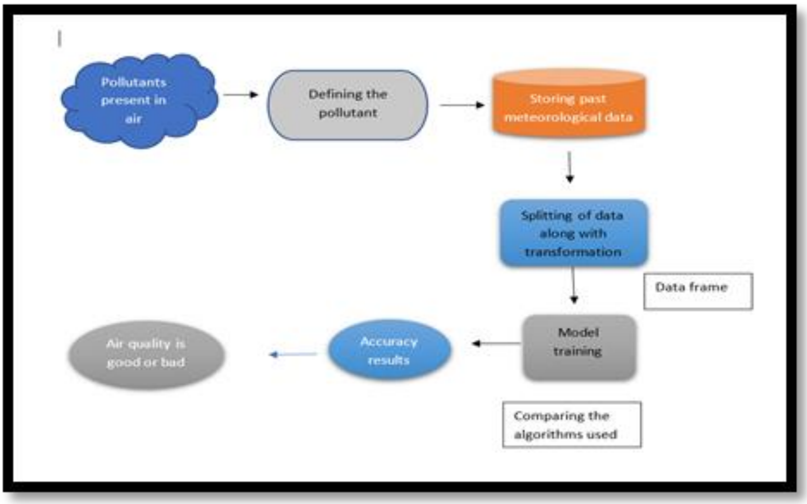
Pollutants present in air
Defining the pollutant
Storing past meteorological data
Splitting of data along with transformation
Data frame
Model training
Accuracy results
Air quality is good or bad
Comparing the algorithms used

Fig 2 Workflow of each method

4. Conclusion

This research stems from the growing need to accurately predict air quality, especially in high population regions like India with ever-growing air pollution. The problem is being addressed by using machine learning programs like Random Forest, Linear Regression and others. They test how well these algorithms are able to predict Air Quality Index (AQI) from a large dataset that has many pollutants. The predictive models that are offered by this research and that are claimed is their ability to predict medium, long, and abrupt changes of air quality, that is needed for decisions to be made and policies to be drafted. This study also outlines algorithms along with step by step breakdown, showing the techniques that different scientists and researchers use to predict air quality and develop strategies for controlling pollution and improving public health. In addition, by comparing this research with previous studies, it provides insights into the progress that has been made in the field as well as the problems that remain to be addressed, pointing future research towards achieving better and more accurate methods of predicting air quality.

5. References

[1]. Tinku SinghORCID Icon,Nikhil Sharma,Satakshi &Manish Kumar,"Analysis and forecasting of air quality index based on satellite data" ,tandfonline 05 Jan 2023,https://doi.org/10.1080/08958378.2022.2164388

[2]. Dong-Her Shih 1ORCID,To Thi Hien 2,3ORCID,Ly Sy Phu Nguyen 2,3ORCID,Ting-Wei Wu 1,* andYen-Ting Lai 1, "A Modified γ-Sutte Indicator for Air Quality Index Prediction"mdpi, 2022, 10(17), 3060; https://doi.org/10.3390/math10173060

[3]. Dipsha Paresh Shah a b, Dr. Piyushkumar Patel c,"A comparison between national air quality index, india and composite air quality index for Ahmedabad, India", Science direct,Volume 5, December 2021,https://doi.org/10.1016/j.envc.2021.100356

[4]. Sandro Rodriguez Garzon, Marcel Reppenhagen, Marcel Müller,"What if Air Quality Dictates Road Pricing? Simulation of an Air Pollution-based Road Charging Scheme ",Scicence direct, Volume 2, December 2022, https://doi.org/10.1016/j.urbmob.2022.100018

[5]. Quang Cuong Doan a, Chen Chen a, Shenjing He a b c, Xiaohu Zhang a c, " How urban air quality affects land values: Exploring non-linear and threshold mechanism using explainable artificial intelligence",Volume 434, 1 January 2024,https://doi.org/10.1016/j.jclepro.2023.140340

[6]. Giovanni Gualtieri, Lorenzo Brilli, Federico Carotenuto, Carolina Vagnoli, Alessandro Zaldei, Beniamino Gioli, " Quantifying road traffic impact on air quality in urban areas: A Covid19-induced lockdown analysis in Italy",Science Direct Volume 267, December 2020,https://doi.org/10.1016/j.envpol.2020.115682

[7]. Seth A. Horn, Purnendu K. Dasgupta ,"The Air Quality Index (AQI) in historical and analytical perspective a tutorial review",Sciencedirect Volume 267, 15 January 2024, https://doi.org/10.1016/j.talanta.2023.125260

[8]. Mariantonietta Ruggieri, Antonella Plaia "An aggregate AQI: Comparing different standardizations and introducing a variability index ", Sciencedirect, Volume 420, 15 March 2012,https://doi.org/10.1016/j.scitotenv.2011.09.019

[9]. Abdelfettah Benchrif a, Ali Wheida b, Mounia Tahri a, Ramiz M. Shubbar c, Biplab Biswas d"Air quality during three covid-19 lockdown phases: AQI, PM2.5 and NO2 assessment in cities with more than 1 million inhabitants" ,SCIENCEDIRECT,Volumes 488–489, November 2021https://doi.org/10.1016/j.scs.2021.103170

[10]. Suhrab, Muhammad, Chen Pinglu, Radulescu Magdalena, Jahangeer Ahmed Soomro, and Surjeet Dalal. "The impact of AI and automation on income inequality in BRICS countries and the role of structural factors and women's empowerment." Industrial Quantum Computing: Algorithms, Blockchains, Industry 4.0 (2024): 155.

[11]. Mahmoud, Amena, Surjeet Dalal, and Umesh Kumar Lilhore. "Advancing healthcare through the opportunities and challenges of quantum computing." Industrial Quantum Computing: Algorithms, Blockchains, Industry 4.0 3 (2024): 239.

[12]. Seth, Bijeta, Surjeet Dalal, and Umesh Kumar Lilhore. "Quantum computing in drug and chemical." Industrial Quantum Computing: Algorithms, Blockchains, Industry 4.0 101000, no. 1011010110 (2024): 255.

[13]. Lilhore, Umesh Kumar, Surjeet Dalal, Vishal Dutt, and Magdalena Radulescu, eds. Industrial Quantum Computing: Algorithms, Blockchains, Industry 4.0. Walter de Gruyter GmbH & Co KG, 2024.

[14]. Dalal, S., Shaheen, M., Lilhore, U. K., Kumar, A., Sharma, S., & Dahiya, M. (2024, December). Traffic forecasting using LSTM and SARIMA models: A comparative analysis. In AIP Conference Proceedings (Vol. 3217, No. 1). AIP Publishing.

[15]. Dalal, S., Jaglan, V., Agrawal, A., Kumar, A., Joshi, S. J., & Dahiya, M. (2024, December). Navigating urban congestion: Optimizing LSTM with RNN in traffic prediction. In AIP Conference Proceedings (Vol. 3217, No. 1). AIP Publishing.

[16]. S. Dubey, M. K. Singh, P. Singh, and S. Aggarwal, "Waste Management of Residential Society using Machine Learning and IoT Approach," 2020 Int. Conf. Emerg. Smart Comput. Informatics, ESCI 2020, pp. 293–297, 2020, doi: 10.1109/ESCI48226.2020.9167526.

[17]. D. Zhang and S. S. Woo, "Real Time Localized Air Quality Monitoring and Prediction through Mobile and Fixed IoT Sensing Network," IEEE Access, vol. 8, pp. 89584–89594, 2020, doi: 10.1109/ACCESS.2020.2993547.

[18]. Das1.Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici, "Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction," IEEE Access, vol. 9, pp. 14765–

14778, 2021, doi: 10.1109/ACCESS.2021.3052429.

[19]. Tarunim Sharma, Aman Jatain, Shalini Bhaskar Bajaj, Kavita Pabreja, “An empirical analysis of feature selection techniques for Software Defect Prediction”, Journal of Autonomous Intelligence (2024) Volume 7 Issue 3, pp. 1-17

[20]. Bhavna Galhotra, Aman Jatain, Shalini Bhaskar Bajaj, Vivek Jaglan, “E WALLET: PAYMENT MECHANISM AND ITS SECURITY MODEL”, Eur. Chem. Bull. 2023, 12(Special Issue 10), pp.3505 –3510

[21]. Bhavna Galhotra, Aman Jatain, Shalini Bhaskar Bajaj, Vivek Jaglan, “GEN Z’S DIGITAL PAYMENTS: DISRUPTIVE OR USEFUL FOR ONLINE SHOPPING IN SECURITY ASPECT”, China Petroleum Processing and Petrochemical Technology, Volume 23, Issue 2, September 2023, pp. 563-574

[22]. T Sharma, A Jatain, S Bhaskar, K Pabreja, “Ensemble Machine Learning Paradigms in Software Defect Prediction”, Procedia Computer Science, 2023, 218, 199-209

[23]. Nishu Sethi, Shalini Bhaskar Bajaj, “Neural Network Based image detection and tracking for security and surveillance”, Journal of Discrete Mathematical Sciences and Cryptography, April 2023, 26(3), pp. 939-949