AIR QUALITY INDEX FORECASTING VIA GENETIC ALGORITHM BASED IMPROVED EXTREME LEARNING MACHINE

J. David Sukeerthi Kumar[1], P. Rizwana[2], M. Pushpanjali[3], Y. Suma Mallika[4], P. Akshaya Sree[5], T. Meghana[6]

[1]Assistant Professor in Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

[2,3,4,5,6]Student, Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

DOI: https://www.doi.org/10.58257/IJPREMS33523

ABSTRACT

Air quality has always been one of the most important environmental concerns for the general public and society. Using machine learning algorithms for Air Quality Index (AQI) prediction is helpful for the analysis of future air quality trends from a macro perspective. In order to effectively address this problem, a genetic algorithm-based improved extreme learning machine (GA-KELM) prediction method is enhanced. First, a kernel method is introduced to produce the kernel matrix which replaces the output matrix of the hidden layer. To address the issue of the conventional limit learning machine where the number of hidden nodes and the random generation of thresholds and weights lead to the degradation of the network learning ability, a genetic algorithm is then used to optimize the number of hidden nodes and layers of the kernel limit learning machine. The thresholds, the weights, and the root mean square error are used to define the fitness function. Genetic algorithms are able to find the optimal solution in the search space and gradually improve the performance of the model through an iterative optimization process. In order to verify the predictive ability of GA-KELM, based on the collected basic data of long-term air quality forecast at a monitoring point in a city in China, the optimized kernel extreme learning machine is applied to predict air quality (SO2, NO2, PM10, CO, O3, PM2.5 concentration and AQI), with comparative experiments based CMAQ (Community Multiscale Air Quality), SVM (Support Vector Machines) and DBN-BP (Deep Belief Networks with Back-Propagation). The results show that the proposed model trains faster and makes more accurate predictions. As extension we have experimented with BI-LSTM algorithm which will optimize features weight in both forward and backward direction.

Keywords: Air Quality Index (AQI), Forecasting, Genetic Algorithm, Pollution Prediction

1. INTRODUCTION

Air pollution is a prevalent environmental problem in the twenty-first century. In light of the rapid industrialization and urbanization, air pollution is getting worse, which greatly affects our living environment and health. Li et al. came to the conclusion that outdoor physical activity poses numerous health risks due to ambient air pollution in China. According to the Chinese Ambient Air Quality Standards (GB3095-2012), there are six conventional air pollutants used to measure air quality: sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter with a particle size less than 10 microns (PM10), particulate matter with a particle size less than 2.5 microns (PM2.5), ozone (O3), and carbon monoxide (CO). These pollutants have adverse effects on human health. The International Energy Agency estimates that air pollution causes 6.5 million premature deaths per year, while long-term exposure to pollutants, such as fine particles (e.g.,PM2.5) or traffic-related pollutants, is linked to higher rates of lung cancer, coronary heart disease, and other illnesses. Therefore, studies on air quality prediction are particularly important and are considered a key factor for environmental protection. In order to more comprehensively assess the health effects of air pollution, numerous air quality monitoring stations have been set up in major cities. Air quality predictions can be made based on the data collected from these stations. Air quality monitoring, modeling, and accurate predictions are important for having a clear understanding of future pollution levels and their associated health risks.

2. LITERATURE REVIEW

a. Variational Bayesian Network with Information Interpretability Filtering for Air Quality Forecasting:

Air quality plays a vital role in people’s health, and air quality forecasting can assist in decision making for government planning and sustainable development. In contrast, it is challenging to multi-step forecast accurately due to its complex and nonlinear caused by both temporal and spatial dimensions. Deep models, with their ability to model strong nonlinearities, have become the primary methods for air quality forecasting. However, because of the lack of mechanism-based analysis, uninterpretability forecasting makes decisions risky, especially when the government makes decisions. This paper proposes an interpretable variational Bayesian deep learning model with information self-screening for PM2.5 forecasting. Firstly, based on factors related to PM2.5 concentration, e.g., temperature, humidity, wind speed, spatial distribution, etc., an interpretable multivariate data screening structure for PM2.5 forecasting was established to catch as much helpful information as possible. Secondly, the self-screening layer was implanted in the deep learning network to optimize the selection of input variables. Further, following implantation of the screening layer, a variational Bayesian gated recurrent unit (GRU) network was constructed to overcome the complex distribution of PM2.5 and achieve accurate multi-step forecasting. The high accuracy of the proposed method is verified by PM2.5 data in Beijing, China, which provides an effective way, with multiple factors for PM2.5 forecasting determined using deep learning technology.

b. Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM:

Modeling and predicting air pollution concentrations is important to provide early warnings about harmful atmospheric substances. However, uncertainty in the dynamic process and limited information about chemical constituents and emissions sources make air-quality predictions very difficult. This study proposed a novel deep-learning method to extract high levels of abstraction in data and capture spatiotemporal features at hourly and daily time intervals in NEOM City, Saudi Arabia. The proposed method integrated a residual network (ResNet) with the convolutional long short-term memory (ConvLSTM). The ConvLSTM method was boosted by a ResNet model for deeply extracting the spatial features from meteorological and pollutant data and thereby mitigating the loss of feature information. Then, health risk assessment was put forward to evaluate PM10 and PM2.5 risk sensitivity in five districts in NEOM City. Results revealed that the proposed method with effective feature extraction could greatly optimize the accuracy of spatiotemporal air quality forecasts compared to existing state-of-the-art models. For the next hour prediction tasks, the PM10 and PM2.5 of MASE were 9.13 and 13.57, respectively. The proposed method provides an effective solution to improve the prediction of air-pollution concentrations while being portable to other regions around the world.

c. Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast System version 16:

A new dynamical core, known as the Finite-Volume Cubed-Sphere (FV3) and developed at both NASA and NOAA, is used in NOAA's Global Forecast System (GFS) and in limited-area models for regional weather and air quality applications. NOAA has also upgraded the operational FV3GFS to version 16 (GFSv16), which includes a number of significant developmental advances to the model configuration, data assimilation, and underlying model physics, particularly for atmospheric composition to weather feedback. Concurrent with the GFSv16 upgrade, we couple the GFSv16 with the Community Multiscale Air Quality (CMAQ) model to form an advanced version of the National Air Quality Forecasting Capability (NAQFC) that will continue to protect human and ecosystem health in the US. Here we describe the development of the FV3GFSv16 coupling with a “state-of-the-science” CMAQ model version 5.3.1. The GFS–CMAQ coupling is made possible by the seminal version of the NOAA-EPA Atmosphere–Chemistry Coupler (NACC), which became a major piece of the next operational NAQFC system (i.e., NACC-CMAQ) on 20 July 2021. NACC-CMAQ has a number of scientific advancements that include satellite-based data acquisition technology to improve land cover and soil characteristics and inline wildfire smoke and dust predictions that are vital to predictions of fine particulate matter (PM2.5) concentrations during hazardous events affecting society, ecosystems, and human health. The GFS-driven NACC-CMAQ model has significantly different meteorological and chemical predictions compared to the previous operational NAQFC, where evaluation of NACC-CMAQ shows generally improved near-surface ozone and PM2.5 predictions and diurnal patterns, both of which are extended to a 72 h (3 d) forecast with this system.

d. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework:

Air quality forecasting has been regarded as the key problem of air pollution early warning and control management. In this article, we propose a novel deep learning model for air quality (mainly PM2.5) forecasting, which learns the spatial-temporal correlation features and interdependence of multivariate air quality related time series data by hybrid deep learning architecture. Due to the nonlinear and dynamic characteristics of multivariate air quality time series data, the base modules of our model include one-dimensional Convolutional Neural Networks (1D-CNNs) and Bi-directional Long Short-term Memory networks (Bi-LSTM). The former is to extract the local trend features and spatial correlation features, and the latter is to learn spatial-temporal dependencies. Then we design a jointly hybrid deep learning framework based on one-dimensional CNNs and Bi-LSTM for shared representation features learning of multivariate air quality related time series data. We conduct extensive experimental evaluations using two real-world datasets, and the results show that our model is capable of dealing with PM2.5 air pollution forecasting with satisfied accuracy.

e. Multivariate regression analysis of air quality index for Hyderabad city: Forecasting model with hourly frequency:

The present study gives a description on air quality index (AQI) for the major city of India i.e. Hyderabad. Major parameter considered for AQI computation are NO, NO2, NOx, SO2, Ambient temperature, relative humidity, bar pressure, solar radiation, wind speed, wind direction, benzene, toluene, xylene, PM2.5 and rack temperature. An approach to assess and represent air quality status through an Air Quality Index (AQI), in Hyderabad city is done by representing the variation of AQI with a multivariate regression model. The utility of this study lies for major metropolitan cities, where different types of activities, viz. industrial, commercial and residential are in progress, on a short and a long term basis this model can be useful for better forecasting of air quality parameters. To make the index more informative, air quality status is classified into five different categories, viz. Clean, Moderate, Poor, Bad and Dangerous. Long term air quality indices are then calculated using hourly basis data for Hyderabad city which is a metropolitan city of India and is fast developing in terms of different climatologically defined features. The included application of this formal analysis includes accounting for atmospheric processes, ambient measurements, emissions characterization, air quality modeling of emissions to ambient concentrations, and characterization of human and ecological responses to ambient pollutant exposure. There is a need for new management strategy that would expand the current practice of accountability that relates emission reductions and attainment of air quality derived from air quality criteria and standards. Conceptually, achievement of accountability would establish goals optimizing risk reduction associated with pollution management.

3. METHODOLOGY

In this section, the methodology was adopted in order to predict Air Quality.

1. Understanding the Problem Domain:

- Familiarize yourself with AQI and its determinants like PM2.5, PM10, Ozone, CO, SO2, and NO2.

- Understand the significance of forecasting AQI for environmental management and public health.

2. Data Collection and Preprocessing:

- Collect historical AQI data from reliable sources such as government agencies or environmental monitoring stations.

- Gather meteorological data (temperature, humidity, wind speed, etc.) and other relevant factors that influence air quality.

- Preprocess the data, including cleaning, filtering outliers, handling missing values, and normalizing to ensure consistency and accuracy.

3. Feature Selection and Engineering:

- Select relevant features that contribute significantly to AQI variations. Use domain knowledge and statistical techniques for feature selection.

- Engineer new features if necessary, such as lagged variables or interactions between variables, to capture complex relationships.

4. Model Selection: Extreme Learning Machine (ELM):

- ELM is chosen as the base model due to its simplicity, fast learning speed, and good generalization performance.

- Understand the working principle of ELM, which involves randomly assigning input weights and analytically determining output weights.

5. Improvement using Genetic Algorithm (GA):

- Integrate GA to optimize the parameters of the ELM model for better performance.

- Define the chromosome representation, genetic operators (crossover, mutation), and fitness function tailored to the ELM parameters.

- Implement the GA to evolve the population of candidate solutions over multiple generations, selecting the best-performing individuals.

6. Model Training and Evaluation:

- Split the dataset into training, validation, and testing sets.

- Train the ELM model using the training data and optimize its parameters using GA on the validation set.

- Evaluate the model's performance using appropriate metrics (e.g., RMSE, MAE, R-squared) on the testing set to ensure its generalization capability.

7. Forecasting AQI:

- Once the model is trained and validated, deploy it to forecast future AQI values.

- Use real-time or forecasted meteorological data as input to predict AQI for upcoming time intervals (e.g., hours, days).

8. Model Monitoring and Maintenance:

- Continuously monitor the model's performance and update it periodically with new data to maintain its accuracy and relevance.

- Consider retraining the model if significant changes occur in the environment or data distribution.

4. IMPLEMENTATION AND ANALYSIS

In this section, the implementation details are mentioned to detect Air Quality. It contains the model selection, and the analysis that has done, and its accuracy is shown.

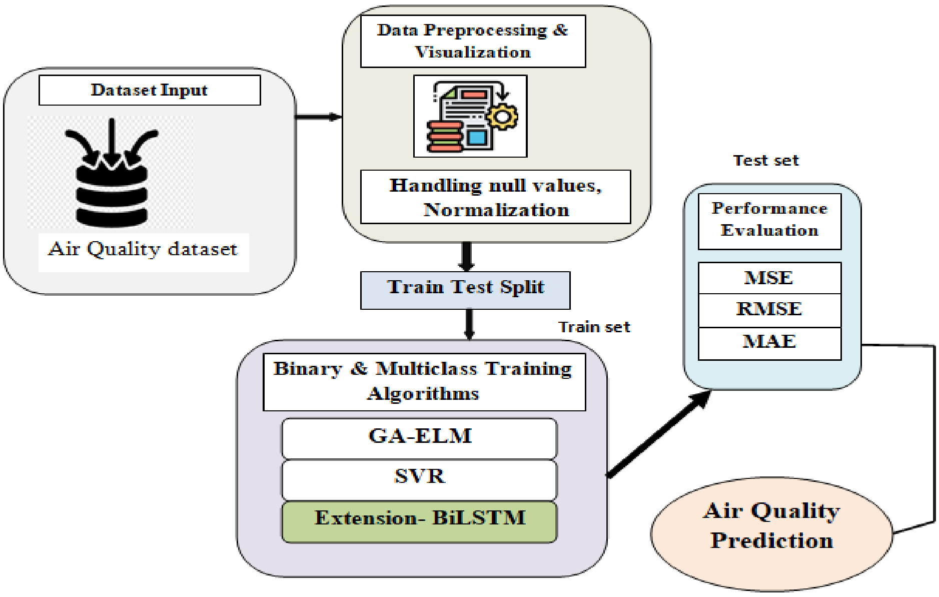


Figure 1: System Architecture

Model-1:Data loading: Using this module we are going to import the dataset. Once the data is preprocessed and split, it needs to be loaded into memory in a format suitable for training or inference. This involves converting the data into appropriate data structures such as NumPy arrays, PyTorch tensors, or TensorFlow tensors.

Model-2: Data Processing: Using the module we will explore the data. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing.

Model-3: Splitting data into Train & Test: Using this module data will be divided into train & test. Splitting data into training and test sets is essential for assessing model performance and ensuring that machine learning models generalize well to new, unseen data. Careful consideration of data splitting techniques and validation strategies helps build more robust and reliable models.

Model-4: Model Generation: Model building - SVR - GA-KELM - DBN-BP(NN with Back-Propagation) - BiLSTM. Algorithms accuracy calculated. The goal is to develop a model that can accurately predict outcomes or make decisions based on input data, while also being robust and generalizable to new, unseen data. Effective model generation requires careful consideration of the problem domain, choice of algorithms, data preprocessing techniques, hyperparameter tuning, and thorough evaluation to ensure the model's effectiveness and reliability in real-world applications.

Model-5: User signup & login: Using this module will get registration and login. The signup process is often the very start of the user journey, while the login process is an ongoing part of it. Making both as smooth as possible increases user conversion and retention. If users run into obstacles while signing up, they're likely to abandon the process entirely.

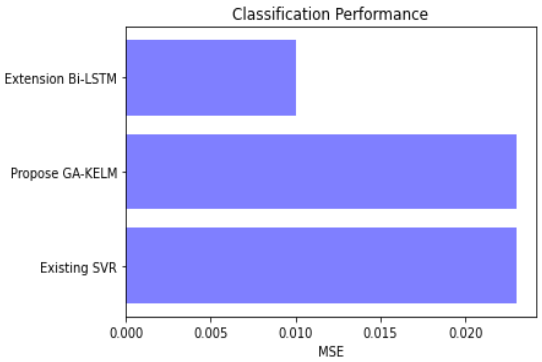
Model-6: User input: Using this module will give input for prediction. User input is the information or commands provided by users to computer systems or applications. It comes in various forms such as text, numbers, selections, gestures, voice commands, or biometric data. Handling user input involves validation, sanitization, error handling, and data conversion to ensure correctness, completeness, and security.

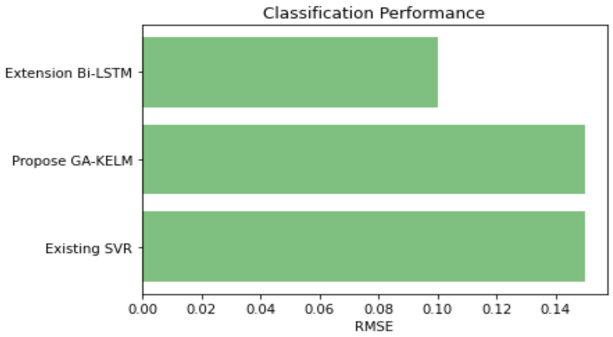
Model-7: Prediction: Final predicted displayed.

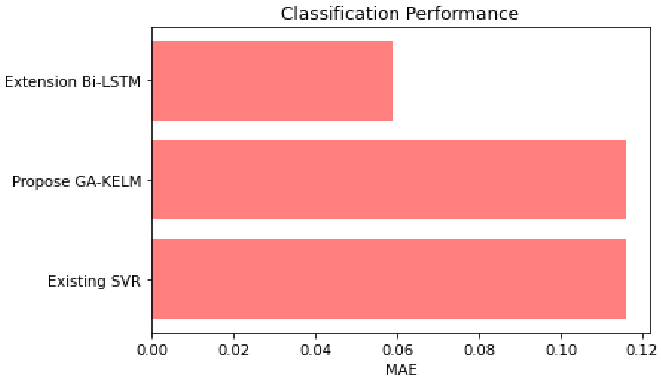
5. RESULTS

Analysis of data is done by using deep Learning Models with different data balancing techniques and recurrent models which helps us to choose the best model in order to predict Air Quality. We have used different Models such as Genetic Algorithm with Extreme Learning Machine, Support Vector Regressor , Extension- BiLSTM. Genetic Algorithm with Support Vector Regression, Genetic Algorithm with Extension BiLSTM, Genetic Algorithm with GA-KELM.

In the bar plot, the accuracy of each model has been plotted.







- The Accuracy score for MSE Existing SVR is 0.023
- The Accuracy score for MSE Propose GA-KELM is 0.023
- The Accuracy score for MSE Extension Bi-LSTM is 0.010
- The Accuracy score for RMSE Existing SVR is 0.15
- The Accuracy score for RMSE Propose GA-KELM is 0.15
- The Accuracy score for RMSE Extension Bi-LSTM is 0.10
- The Accuracy score for MAE Existing SVR is 0.116
- The Accuracy score for MAE Propose GA-KELM is 0.116
- The Accuracy score for MAE Extension Bi-LSTM is 0.059

	ML Model	MSE	RMSE	MAE
0	Existing SVR	0.023	0.15	0.116
1	Propose GA-KELM	0.023	0.15	0.116
2	**Extension Bi-LSTM**	**0.010**	**0.10**	**0.059**

6. CONCLUSION

In this project the economic development achieved by the country through rapid urbanization is polluting the environment in an alarming way and putting people's lives in danger. Therefore, a correct analysis and accurate prediction of air quality remains a primary condition to achieve the objective of sustainable development. Therefore, future research should explore the underlying significance and value of combinatorial intelligence optimization algorithms such as the Limit Learning Machine.

7. REFERENCES

[1] Mahammad, F. S., & Viswanatham, V. M. (2020). Performance Analysis Of Data Compression Algorithms For Heterogeneous Architecture Through Parallel Approach. The Journal Of Supercomputing, 76(4), 2275-2288.

[2] Karukula, N. R., & Farooq, S. M. (2013). A Route Map For Detecting Sybil Attacks In Urban Vehicular Networks. Journal Of Information, Knowledge, And Research In Computer Engineering, 2(2), 540-544.

[3] Farook, S. M., & Nageswarareddy, K. (2015). Implementation Of Intrusion Detection Systems For High Performance Computing Environment Applications. Inter National Journal Of Scientific Engineering And Technology Research, 4(0), 41.

[4] Sunar, M. F., & Viswanatham, V. M. (2018). A Fast Approach To Encrypt And Decrypt Of Video Streams For Secure Channel Transmission. World Review Of Science, Technology And Sustainable Development, 14(1), 11-28.

[5] Mahammad, F. S., & Viswanatham, V. M. (2017). A Study On H. 26x Family Of Video Streaming Compression Techniques. International Journal Of Pure And Applied Mathematics, 117(10), 63-66.

[6] Devi,S M. S., Mahammad, F. S., Bhavana, D., Sukanya, D., Thanusha, T. S., Chandrakala, M., & Swathi, P. V. (2022)." Machine Learning Based Classification And Clustering Analysis Of Efficiency Of Exercise Against Covid-19 Infection." Journal Of Algebraic Statistics, 13(3), 112-117.

[7] Devi, M. M. S., & Gangadhar, M. Y. (2012)." A Comparative Study Of Classification Algorithm ForPrinted Telugu Character Recognition." International Journal Of Electronics Communication And Computer Engineering, 3(3), 633-641.

[8] Devi, M. S., Meghana, A. I., Susmitha, M., Mounika, G., Vineela, G., & Padmavathi, M. Missing Child Identification System Using Deep Learning.

[9] V. Lakshmi Chaitanya. "Machine Learning Based Predictive Model For Data Fusion Based Intruder Alert System." Journal Of Algebraic Statistics 13, No. 2 (2022): 2477-2483.

[10] Chaitanya, V. L., & Bhaskar, G. V. (2014). Apriori Vs Genetic Algorithms For Identifying Frequent Item Sets. International Journal Of Innovative Research &Development, 3(6), 249-254.

[11] Chaitanya, V. L., Sutraye, N., Praveeena, A. S., Niharika, U. N., Ulfath, P., & Rani, D. P. (2023). Experimental Investigation Of Machine Learning Techniques For Predicting Software Quality.

[12] Lakshmi, B. S., Pranavi, S., Jayalakshmi, C., Gayatri, K., Sireesha, M., & Akhila, A. Detecting Android Malware With An Enhanced Genetic Algorithm For Feature Selection And Machine Learning.

[13] Lakshmi, B. S., & Kumar, A. S. (2018). Identity-Based Proxy-Oriented Data Uploading And Remote Data Integrity Checking In Public Cloud. International Journal Of Research, 5(22), 744-757.

[14] Lakshmi, B. S. (2021). Fire Detection Using Image Processing. Asian Journal Of Computer Science And Technology, 10(2), 14-19.

[15] Devi, M. S., Poojitha, M., Sucharitha, R., Keerthi, K., Manideepika, P., & Vasudha, C. Extracting And Analyzing Features In Natural Language Processing For Deep Learning With English Language.

[16] Kumar Jds, Subramanyam Mv, Kumar Aps. Hybrid Chameleon Search And Remora Optimization Algorithm-Based Dynamic Heterogeneous Load Balancing Clustering Protocol For Extending The Lifetime Of Wireless Sensor Networks. Int J Commun Syst. 2023; 36(17):E5609. Doi:10.1002/Dac.5609

[17] David Sukeerthi Kumar, J., Subramanyam, M.V., Siva Kumar, A.P. (2023). A Hybrid Spotted Hyena And Whale Optimization Algorithm-Based Load-Balanced Clustering Technique In Wsns. In: Mahapatra, R.P., Peddoju, S.K., Roy, S., Parwekar, P. (Eds) Proceedings Of International Conference On Recent Trends In Computing. Lecture Notes In Networks And Systems, Vol 600. Springer, Singapore. Https://Doi.Org/10.1007/978-981-19-8825-7_68

[18] Murali Kanthi, J. David Sukeerthi Kumar, K. Venkateshwara Rao, Mohmad Ahmed Ali, Sudha Pavani K, Nuthanakanti Bhaskar, T. Hitendra Sarma, “A Fused 3d-2d Convolution Neural Network For Spatial-Spectral Feature Learning And Hyperspectral Image Classification,” J Theor Appl Inf Technol, Vol. 15, No. 5, 2024, Accessed: Apr. 03, 2024. [Online]. Available: Www.Jatit.Org

[19] Prediction Of Covid-19 Infection Based On Lifestyle Habits Employing Random Forest Algorithm Fs Mahammad, P Bhaskar, A Prudvi, Ny Reddy, Pj Reddy Journal Of Algebraic Statistics 13 (3), 40-45

[20] Machine Learning Based Predictive Model For Closed Loop Air Filtering System P Bhaskar, Fs Mahammad, Ah Kumar, Dr Kumar, Sma Khadar, ...Journal Of Algebraic Statistics 13 (3), 609-616

[21] Kumar, M. A., Mahammad, F. S., Dhanush, M. N., Rahul, D. P., Sreedhara, K. L., Rabi, B. A., & Reddy, A. K. (2022). Traffic Length Data Based Signal Timing Calculation For Road Traffic Signals Employing Proportionality Machine Learning. Journal Of Algebraic Statistics, 13(3), 25-32.

[22] Kumar, M. A., Pullama, K. B., & Reddy, B. S. V. M. (2013). Energy Efficient Routing In Wireless Sensor Networks. International Journal Of Emerging Technology And Advanced Engineering, 9(9), 172-176.

[23] Kumar, M. M. A., Sivaraman, G., Charan Sai, P., Dinesh, T., Vivekananda, S. S., Rakesh, G., & Peer, S. D. Building Search Engine Using Machine Learning Techniques.

[24] “ Providing Security In Iot Using Watermarking And Partial Encryption. Issn No:

a. 2250-1797 Issue 1, Volume 2 (December 2011)

[25] The Dissemination Architecture Of Streaming Media Information On Integrated Cdn And P2p, Issn 2249-6149 Issue 2, Vol.2 (March-2012)

[26] Provably Secure And Blind Sort Of Biometric Authentication Protocol Using Kerberos, Issn: 2249-9954, Issue 2, Vol 2 (April 2012)

[27] D.Lakshmaiah, Dr.M.Subramanyam, Dr.K.Satya Prasad,” Design Of Low Power 4- Bit Cmos Braun Multiplier Based On Threshold Voltage Techniques”, Global Journal Of Research In Engineering, Vol.14(9),Pp.1125-1131,2014.

[28] R Sumalatha, Dr.M.Subramanyam, “Image Denoising Using Spatial Adaptive Mask Filter”, Ieee International Conference On Electrical, Electronics, Signals, Communication &Amp; Optimization (Eesco-2015), Organized Byvignans Institute Of Information Technology, Vishakapatnam, 24 Th To 26th January 2015. (Scopus Indexed)

[29] P.Balamurali Krishna, Dr.M.V.Subramanyam, Dr.K.Satya Prasad, “Hybrid Genetic Optimization To Mitigate Starvation In Wireless Mesh Networks”, Indian Journal Of Science And Technology,Vol.8,No.23,2015. (Scopus Indexed)

[30] Y.Murali Mohan Babu, Dr.M.V.Subramanyam,M.N. Giri Prasad,” Fusion And Texure Based Classification Of Indian Microwave Data – A Comparative Study”, International Journal Of Applied Engineering Research, Vol.10 No.1, Pp. 1003-1009, 2015. (Scopus Indexed)