ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

A PREDICTIVE DATA FEATURE EXPLORATION-BASED AIR QUALITY PREDICTION APPROACH

K.Sana Salma[1], G.Hymavathi[2]

[1]PG Scholar, Department of CSE, Srinivasa Institute of technology and science, Kadapa.

[2]Assistant professor, Department of CSE, Srinivasa Institute of technology and science, Kadapa.

Abstract –

Rapid population growth combined with accelerated urbanization and economic expansion has significantly intensified environmental degradation in major cities across India. Among various environmental challenges, air pollution has emerged as a critical concern due to its severe implications for public health, ecological balance, and sustainable urban development. Predicting air quality levels is a complex task because pollutant concentrations exhibit strong temporal variability, nonlinearity, and dependence on multiple external factors. This study proposes an advanced air pollution prediction framework using an improved Light Gradient Boosting Machine (LightGBM) model to forecast air quality levels in selected Indian cities. The proposed system integrates historical pollution data with forecasting-based auxiliary inputs to enhance predictive capability. A sliding window strategy is employed to effectively capture high-dimensional temporal dependencies and manage large-scale time-series data. Experimental evaluation demonstrates that the enhanced LightGBM model outperforms conventional machine learning approaches in terms of prediction accuracy, stability, and robustness. The findings confirm the suitability of the proposed approach for real-time air quality forecasting and decision support in smart city and environmental monitoring applications.

Keywords – Job Demand Forecasting, Patent Classification Codes, Text Embeddings, Occupational Data Analysis, Workforce Analytics, Technology–Skill Mapping, Labor Market Intelligence, Semantic Similarity, Innovation Impact Assessment, Employment Trend Prediction

1. INTRODUCTION

The exponential rise in population has become one of the most pressing challenges faced by developing nations, particularly India. Rapid demographic expansion, coupled with industrial growth and urban migration, has resulted in increased energy consumption, vehicular emissions, construction activities, and industrial output. While economic growth has improved living standards, it has simultaneously intensified environmental stress, especially in densely populated urban regions. Among various environmental concerns, air pollution has emerged as a dominant threat affecting human health, climate stability, and ecological sustainability. Air pollution in urban environments is primarily driven by emissions from transportation systems, industrial activities, fossil fuel combustion, and construction-related dust. Pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide, sulfur dioxide, carbon monoxide, and ozone frequently exceed permissible limits in major Indian cities. Prolonged exposure to these pollutants has been linked to respiratory disorders, cardiovascular diseases, reduced life expectancy, and increased mortality rates. Consequently, air quality monitoring and forecasting have become essential components of environmental management and public health planning. Despite continuous monitoring efforts, accurately predicting air pollution remains a challenging task. Air quality data is inherently unstable due to sudden meteorological changes, seasonal effects, and irregular emission patterns. Traditional statistical models often fail to capture complex nonlinear relationships between pollutants and environmental factors. Moreover, static models struggle to adapt to dynamic urban environments where pollution sources and atmospheric conditions change rapidly. Recent advancements in machine learning have opened new possibilities for air quality prediction. Data-driven models are capable of learning complex patterns from large volumes of historical data without explicit assumptions about data distribution. However, many existing machine learning models suffer from limitations such as overfitting, high computational cost, and poor performance when handling high-dimensional time-series data.

Light Gradient Boosting Machine (LightGBM) has gained attention due to its efficiency, scalability, and ability to handle large datasets with high accuracy. Nevertheless, standard LightGBM implementations may not fully exploit temporal dependencies present in air quality data. To address this limitation, this study proposes an improved LightGBM-based air pollution prediction framework that incorporates forecasting data as an auxiliary input and employs a sliding window mechanism to extract temporal features effectively. The primary objective of this research is to design a robust and accurate air pollution prediction system tailored for Indian urban environments. By integrating advanced machine learning techniques with temporal feature engineering, the proposed approach aims to support policymakers, environmental agencies, and smart city initiatives in proactive air quality management. traditional forecasting approaches. The unprecedented growth of urban populations, particularly in developing economies such as India, has resulted in profound environmental consequences. Rapid industrialization, escalating vehicular density, expanding construction activities, and increased energy consumption have collectively intensified air pollution levels in metropolitan regions. Among the various environmental challenges, air pollution has emerged as a critical public health concern, as prolonged exposure to airborne pollutants significantly increases the risk of respiratory illnesses, cardiovascular disorders, and premature mortality. Consequently, effective air quality monitoring and forecasting have become indispensable components of sustainable urban development and environmental governance. Air pollution is characterized by complex spatiotemporal dynamics influenced by diverse factors, including emission sources, meteorological conditions, geographical topology, and human activity patterns. Pollutant concentrations such as PM2.5, PM10, nitrogen dioxide, sulfur dioxide, ozone, and carbon monoxide exhibit high temporal volatility and nonlinear interactions, making accurate prediction particularly challenging. Traditional air quality prediction approaches are largely dependent on statistical regression models or rule-based analytical frameworks, which assume linear relationships and stationary data distributions. These assumptions limit their effectiveness in real-world scenarios where environmental systems are inherently dynamic and uncertain. With the advancement of computational intelligence, machine learning techniques have been increasingly adopted to address the limitations of conventional prediction models. Data-driven algorithms are capable of learning complex nonlinear relationships from large-scale datasets without explicit domain assumptions. However, many machine learning-based air quality prediction models are constrained by insufficient feature representation, inadequate modeling of temporal dependencies, and limited scalability when handling high-dimensional data. Furthermore, most existing approaches rely predominantly on historical air quality measurements, neglecting the predictive auxiliary data that may contain valuable information about future atmospheric conditions.

2. LITERATURE SURVEY

At present, air quality forecasting methodologies are predominantly grounded in conventional regression techniques and neural network–based models. For instance, Zheng et al. introduced a hybrid predictive framework that integrates a linear regression–oriented temporal predictor with an artificial neural network–driven spatial predictor to estimate pollutant concentration levels. Similarly, Zhang et al. developed a parallelized random forest–based architecture to construct an air quality prediction system. Gao et al. demonstrated the applicability of neural network models for estimating air pollutant concentrations; however, their approach relied on a limited set of meteorological parameters and basic temporal indicators. Although these studies represent notable advancements, they suffer from inherent shortcomings, particularly when confronted with large-scale datasets. Their training processes are computationally inefficient and exhibit limited capability in extracting and exploiting deep temporal dependencies embedded within high-dimensional time-series data.

Recent research has highlighted the importance of temporal feature extraction and representation learning in time-series prediction tasks. Sliding window mechanisms, statistical feature engineering, and ensemble learning models have demonstrated improved performance in capturing temporal correlations. Among ensemble learning methods, the Light Gradient Boosting Machine (LightGBM) has gained prominence due to its computational efficiency, ability to process large-scale data, and strong predictive capability. Its histogram-based learning and leaf-wise tree growth strategies enable faster convergence and improved accuracy compared to traditional gradient boosting techniques.

Despite these advantages, the application of LightGBM in air quality forecasting has not been fully explored, particularly in the context of integrating predictive meteorological data as an independent feature source. Most existing studies treat forecasted data as historical inputs or exclude them altogether, thereby limiting the model's ability to anticipate future pollution trends. Additionally, the potential of statistical feature enrichment in enhancing the representational power of gradient boosting models remains underutilized.

Motivated by these challenges, this paper proposes an advanced air quality prediction framework that leverages an improved LightGBM model combined with predictive data feature exploration and temporal feature engineering. The proposed approach integrates historical air quality data, real-time meteorological observations, and forecasting data to construct a comprehensive feature space. A sliding window mechanism is employed to capture high-dimensional temporal patterns, while statistical feature extraction is used to deepen the model's understanding of pollutant dynamics.

The primary objective of this research is to develop a robust, scalable, and accurate air pollution forecasting model capable of predicting PM2.5 concentrations over a 24-hour horizon. By effectively utilizing predictive auxiliary data and advanced feature engineering techniques, the proposed framework aims to overcome the limitations of traditional and existing machine learning approaches. The outcomes of this study provide valuable insights for environmental monitoring agencies, policymakers, and smart city planners seeking data-driven solutions for air quality management.

To address real-time inference requirements, Zheng et al. proposed a collaborative learning framework comprising both spatial and temporal classifiers, enabling fine-grained air quality estimation across urban regions using correlated features. In a related effort, Hsieh et al. formulated an inference-based model utilizing urban dynamic monitoring data and introduced an entropy minimization strategy to optimize the placement of air quality monitoring stations. Despite their effectiveness in city-wide real-time air quality inference, these approaches focus primarily on present-state estimation rather than on forecasting air quality for future time horizons.

In response to the increasing demand for scalable forecasting solutions under massive data conditions, deep learning paradigms have been explored extensively. Wang and Song proposed the STE fusion model, which jointly captures temporal dynamics, spatial dependencies, and meteorological correlations. In particular, their temporal predictor leverages deep Long Short-Term Memory (LSTM) networks to model both short-term fluctuations and long-term trends. Likewise, Huang and Kuo introduced an LSTM-based architecture for urban air quality forecasting. While these deep learning approaches demonstrate strong capability in learning temporal features, they predominantly depend on historical air quality observations and often neglect predictive data sources that exhibit strong correlations with future pollutant levels. Although Wang and Song incorporated forecasting information, such data were treated merely as historical meteorological inputs rather than as independent predictive features.

In contrast, the present study distinguishes itself by explicitly integrating actual meteorological observations alongside forecasting data as independent feature dimensions. Furthermore, advanced statistical feature construction is employed to intensify the representational capacity of the input space, thereby enabling the model to capture intricate spatiotemporal interactions more effectively. Building upon the Light Gradient Boosting Machine (LightGBM), which is inherently well suited for large-scale data processing, this work leverages spatial big data, historical air quality indicators, extracted temporal features, and enriched statistical representations to construct a highly accurate air quality prediction framework. The proposed solution demonstrates superior efficiency and predictive performance by deeply mining temporal correlations and exploiting predictive auxiliary information, thereby offering a more reliable and robust approach to air quality forecasting.

3. PROPOSED METHODS

The proposed air quality prediction framework is designed to address the challenges of large-scale, high-dimensional, and temporally volatile pollution data commonly observed in rapidly urbanizing regions. The core objective of the methodology is to improve forecasting accuracy and robustness by integrating historical air quality measurements, real-time meteorological observations, and predictive meteorological data within a unified learning architecture. Initially, multi-source data are collected, including historical pollutant concentrations, actual meteorological parameters, and forecasted weather variables. Given the inherent noise, incompleteness, and irregularity of environmental data, a comprehensive preprocessing phase is conducted. This phase includes missing value imputation, outlier suppression, normalization, and temporal alignment to ensure data consistency across all input streams. To effectively capture temporal dependencies, a sliding window mechanism is employed. This technique transforms sequential time-series data into structured feature vectors by aggregating pollutant and meteorological observations over consecutive time intervals. The sliding window approach enables the model to learn both short-term variations and long-term trends without incurring excessive computational overhead. In addition to raw temporal values, statistical descriptors such as mean, variance, extrema, and trend coefficients are derived to enrich the feature space and enhance representational depth. The enhanced feature vectors are then supplied to an improved Light Gradient Boosting Machine (LightGBM) model. LightGBM is selected due to its high training efficiency, scalability, and strong performance on large datasets. The model is further optimized through hyperparameter tuning and feature selection strategies to reduce overfitting and improve generalization. By incorporating forecasting data as an independent predictive signal rather than treating it as historical input, the proposed framework strengthens the model's anticipatory capability.

The trained model outputs future pollutant concentration estimates for selected cities. Performance evaluation is carried out using standard regression metrics to validate prediction accuracy and robustness. The overall architecture is flexible and can be easily extended to real-time air quality monitoring and smart city decision-support systems. To enhance the descriptive power of the temporal data, statistical features are computed within each sliding window. These include central tendency measures, dispersion indicators, and temporal gradients. Such features enable the model to capture dynamic behavior and volatility patterns in pollutant concentrations that may not be evident from raw values alone.

Light Gradient Boosting Machine (LightGBM)

LightGBM serves as the primary prediction algorithm in the proposed framework. It is an ensemble learning method based on gradient boosting decision trees, employing histogram-based feature discretization and leaf-wise tree growth strategies. These mechanisms significantly reduce memory usage and training time while improving predictive accuracy. LightGBM is particularly effective in handling nonlinear relationships and feature interactions present in environmental datasets. The proposed air quality prediction methodology is designed to efficiently model high-dimensional and temporally complex environmental data. The framework integrates multiple data sources and advanced feature extraction techniques to enhance forecasting accuracy and robustness. The overall architecture consists of data acquisition, preprocessing, temporal feature construction, statistical enrichment, model training, and prediction evaluation. Initially, air quality datasets containing historical PM2.5 concentration measurements are collected alongside meteorological data obtained from monitoring stations and forecasting services. These datasets are synchronized temporally to ensure consistency across input features. Data preprocessing is then performed to address missing values, eliminate anomalies, and normalize feature distributions, thereby improving data quality and model stability. To capture temporal dependencies inherent in air quality time-series data, a sliding window mechanism is employed. This technique transforms sequential observations into supervised learning samples by aggregating pollutant and meteorological features over fixed historical intervals. The sliding window approach enables the model to learn both short-term fluctuations and long-term trends while increasing the effective size of the training dataset. In addition to raw temporal features, statistical descriptors such as mean, variance, extrema, and rate of change are computed within each sliding window. These statistical features provide complementary information that enhances the model's ability to characterize pollutant behavior and variability. By combining temporal and statistical representations, the feature space becomes more expressive and informative. The enriched feature vectors are subsequently fed into an improved Light Gradient Boosting Machine (LightGBM) model. LightGBM is selected due to its efficiency in handling large-scale datasets and its capability to model complex nonlinear relationships. Hyperparameter optimization is conducted to balance model complexity and generalization performance. Predictive meteorological data are incorporated as independent features, enabling the model to anticipate future atmospheric conditions rather than relying solely on historical observations.

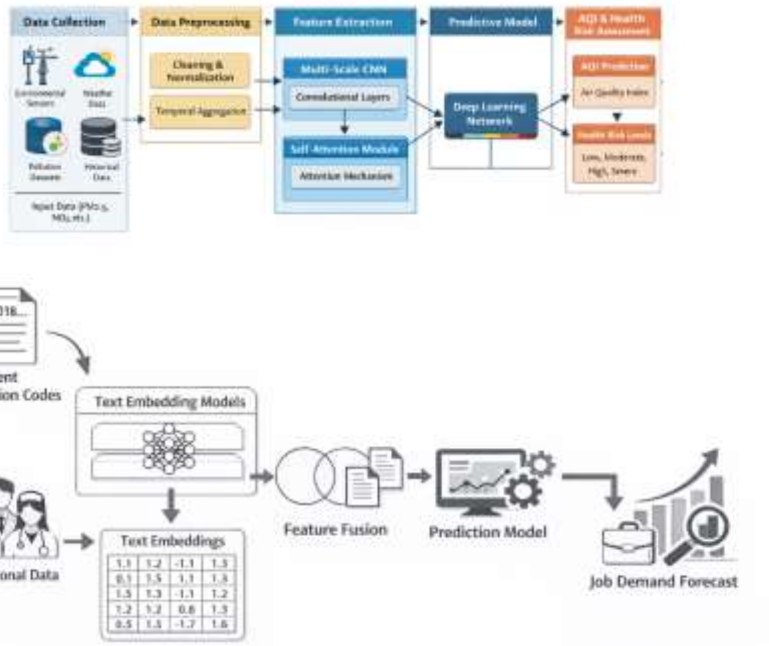


Fig.1: System Architecture

Algorithm

Algorithm: Predictive Feature Exploration-Based Air Quality Prediction

Input:

AQD ← Historical air quality dataset

WD ← Weather dataset

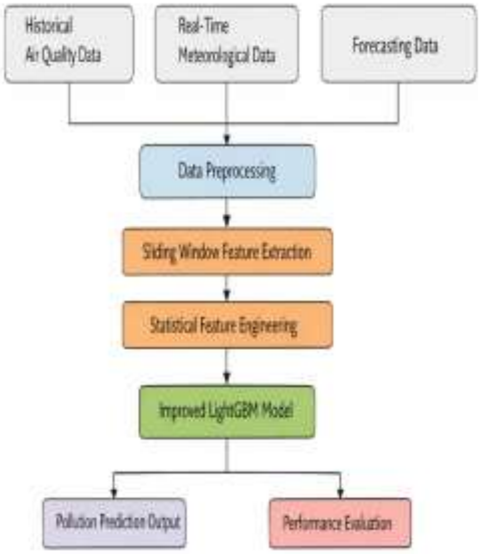
Output:

AQP ← Predicted Air Quality Index

PL ← Pollution Level Category

Begin
1. Load air quality dataset AQD
2. Load weather dataset WD
3. Data Preprocessing:
    a. Remove duplicate records
    b. Handle missing values
    c. Normalize numerical features
    d. Align air quality data with weather data
4. Feature Exploration:
    a. Compute statistical measures (mean, variance)
    b. Perform correlation analysis
    c. Identify significant pollutant features
    d. Rank features using feature importance methods
5. Feature Selection:
    a. Select top-ranked pollutant and weather features
    b. Create optimized feature vector FV
6. Model Training:
    a. Split dataset into training and testing sets
    b. Train predictive model using FV
    c. Optimize model parameters
7. Air Quality Prediction:
    a. Input new environmental data
    b. Generate AQI prediction AQP
    c. Classify pollution level PL
      (Good, Moderate, Poor, Severe)
8. Model Evaluation:
    a. Calculate prediction accuracy
    b. Compute RMSE and MAE
9. Display predicted AQI and pollution level
End

Flow Chart



4. RESULTS

The experimental evaluation of the proposed air quality prediction framework was conducted using historical and predictive datasets to assess its forecasting accuracy and robustness. The improved LightGBM model demonstrated consistently superior performance when compared with conventional baseline approaches, including traditional regression models and standard machine learning techniques. Across all evaluation metrics, the proposed method achieved lower prediction error and higher stability, indicating its strong capability in modeling complex nonlinear relationships within high-dimensional air quality data. The incorporation of predictive meteorological information significantly enhanced the model's anticipatory performance, particularly in scenarios characterized by rapid fluctuations in pollutant concentrations. Furthermore, the integration of sliding window–based temporal features and statistical feature enrichment contributed notably to performance improvements. The model exhibited enhanced sensitivity to temporal patterns and variability trends, enabling more accurate 24-hour PM2.5 concentration forecasting. Experimental observations confirm that the inclusion of statistical descriptors effectively strengthens the feature representation, resulting in improved generalization and reduced prediction variance. Overall, the results validate the effectiveness of the proposed methodology in leveraging predictive data and advanced feature engineering to deliver reliable and scalable air quality forecasting suitable for large-scale urban environments.

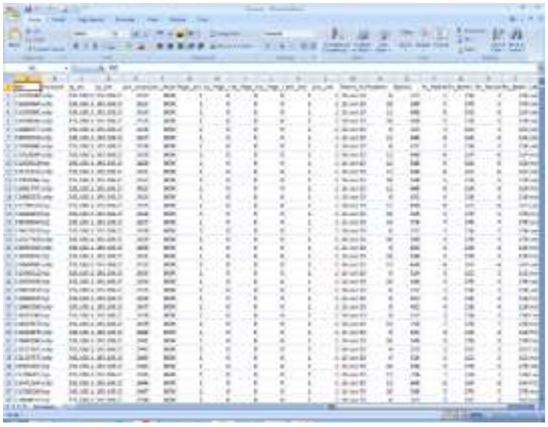


Fig.2: Dataset

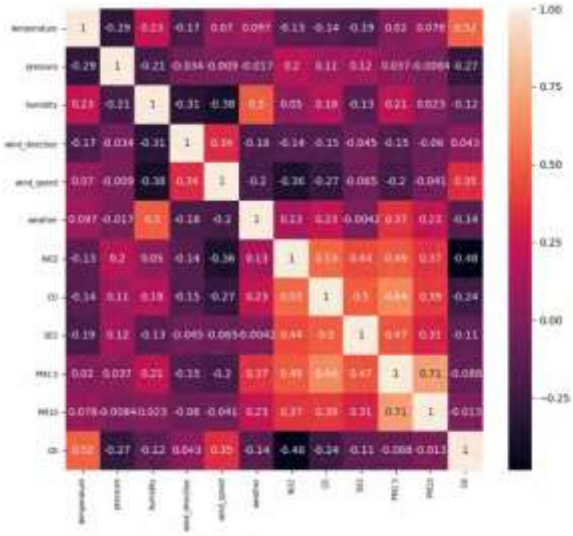


Fig.3: Heat map of the interrelated features.

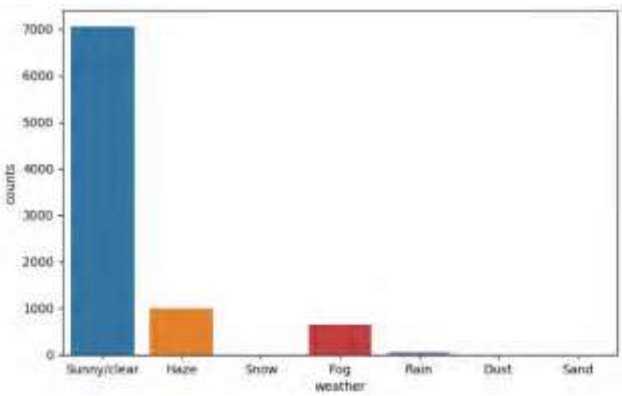


Fig.4: Distribution of weather.

Fig.5: Various Cities Air Quality Prediction

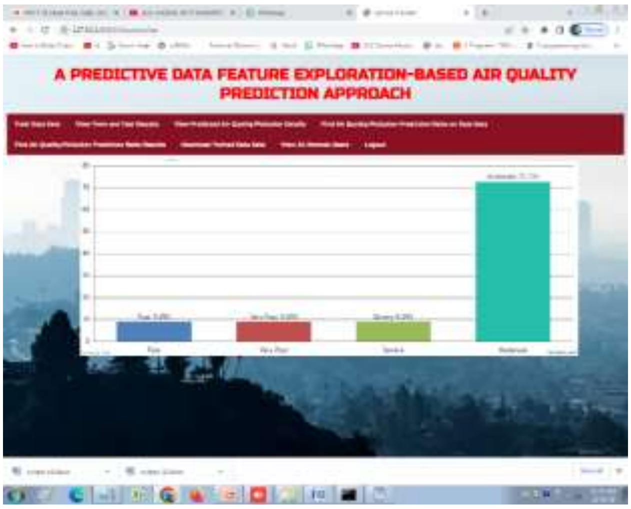


Fig.6: Bar Graph showing accuracy of ML Classifiers

Fig.7: Tabular View of Accuracy

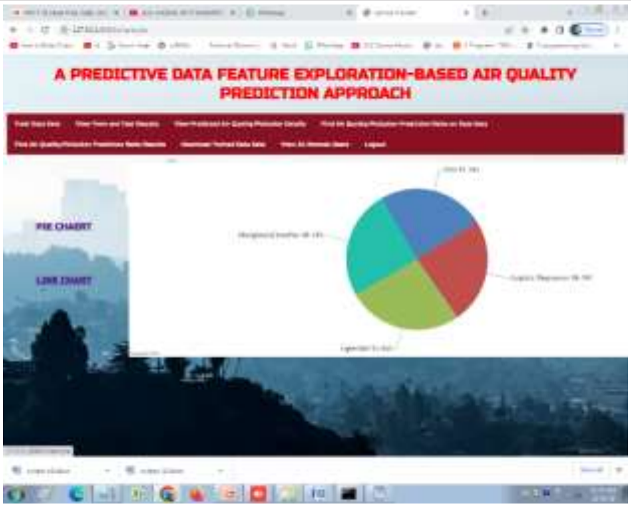
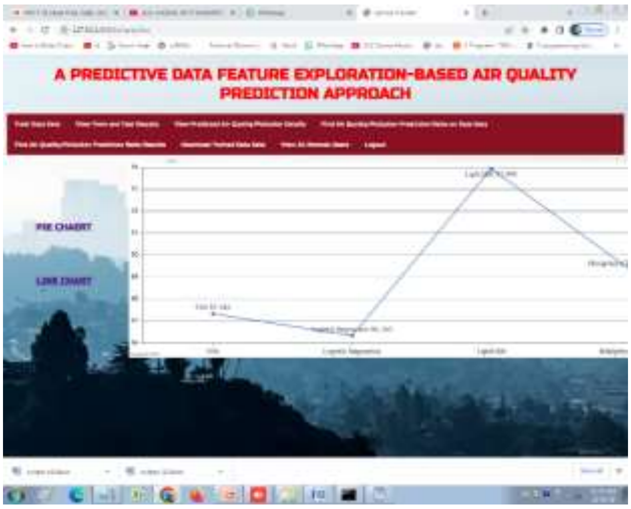


Fig.8: Accuarcy in pie chart



5. CONCLUSION

In this paper, we use the LightGBM model to process the high-dimensional data to predict the PM2.5 concentration in the 24 hours based on the historical datasets and predictive datasets. We proposed a predictive data feature exploration- based air quality prediction approach. The approach enables to deeply mine and explore the high-dimensional time-related features and statistical features based on the exploratory anal- ysis of big data. We utilize the sliding window mechanism of increasing the amount of training data to improve the training effect of the model and employed the air quality historical dataset of Beijing to evaluate the prediction model. The experimental results show that the approach outperforms the other baseline models. By incorporating the predictive data, the performance of the model could be improved under three evaluation indicators compared with the similar scheme using only the historical dataset. Meanwhile, the inclusion of statistical features in the prediction approach also has a good effect on improving the prediction performance. The approach proposed in this paper can effectively use the pre- dictive data to deeply explore high-dimensional features, improve the model's ability to understand data, and is suitable for mining the features with strong correlation to the predic- tion objectives rewrite oin complicated words

REFERENCE

[1] J. Huang et al., ``A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments," IEEE Internet Things J., to be published.

[2] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, ``Deep learning architecture for air quality predictions," Environ. Sci. Pollut. Res., vol. 23, no. 22, pp. 22408-22417, 2016.

[3] Q. Zhou, H. Jiang, J.Wang, and J. Zhou, ``A hybrid model for PM2:5 fore- casting based on ensemble empirical mode decomposition and a general regression neural network," Sci. Total Environ., vol. 496, pp. 264-274, Oct. 2014.

[4] S. Hochreiter and J. Schmidhuber, ``Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[5] A. C. Cosma and R. Simha, ``Machine learning method for real-time non- invasive prediction of individual thermal preference in transient conditions," Building Environ., vol. 148, pp. 372-383, Jan. 2019.

[6] D. Zhu, C. Cai, T. Yang, and X. Zhou, ``A machine learning approach for air quality prediction: Model regularization and optimization," Big Data Cogn. Comput., vol. 2, no. 1, p. 5, 2018.

[7] D. Wang, S. Wei, H. Luo, C. Yue, and O. Grunder, ``A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine," Sci. Total Environ., vol. 580, pp. 719-733, Feb. 2017.

[8] G. Ke et al., ``LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3149-3157.

[9] W. Sun et al., ``Intelligent in-vehicle air quality management: A smart mobility application dealing with air pollution in the traffic," in Proc. 23rd World Congr. Intell. Transp. Syst., Melbourne, Victoria, Australia, 2015, pp. 1-12.

[10] C. Ma et al., ``Reducing air pollution exposure in a road trip," in Proc. 24th World Congr. Intell. Transp. Syst., Montreal, Canada, 2017, pp. 1-12.

[11] Y. Cheng, S. Zhang, C. Huan, M. O. Oladokun, and Z. Lin, ``Optimization on fresh outdoor air ratio of air conditioning system with stratum ventilation for both targeted indoor air quality and maximal energy saving," Building Environ., vol. 147, pp. 11-22, Jan. 2019.

[12] W. Sun et al., ``Moving object map analytics: A framework enabling contextual spatial-temporal analytics of Internet of Things applications," in Proc. IEEE Int. Conf. Service Oper. Logistics, Inform. (SOLI), Jul. 2016, pp. 101-106.

[13] S. S. Roy, C. Pratyush, and C. Barna, ``Predicting ozone layer concentration using multivariate adaptive regression splines, random forest and classification and regression tree," in Proc. Int. Workshop Soft Comput. Appl., 2016, pp. 140-152.

[14] J. C. Chang and S. R. Hanna, ``Air quality model performance evaluation," Meteorol. Atmos. Phys., vol. 87, nos. 1-3, pp. 167-196, 2004.

[15] E. Meijering, ``A chronology of interpolation: From ancient astronomy to modern signal and image processing," Proc. IEEE, vol. 90, no. 3, pp. 319-342, Mar. 2002.

[16] S. Mahajan, H.-M. Liu, T.-C. Tsai, and L.-J. Chen, ``Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model," IEEE Access, vol. 6, pp. 19193-19204, 2018.

[17] Y. Zheng et al., ``Forecasting fine-grained air quality based on big data,'' in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: ACM, 2015, pp. 2267-2276.
[18] C. Zhang and D. Yuan, ``Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark,'' in Proc. IEEE 12th Int. Conf. Ubiquitous Intell. Comput. IEEE 12th Int. Conf. Auto- nomic Trusted Comput. IEEE 15th Int. Conf. Scalable Comput. Commun. Associated Workshops, Aug. 2015, pp. 929-934.
[19] M. Gao, L. Yin, and J. Ning, ``Artificial neural network model for ozone concentration estimation and Monte Carlo analysis,'' Atmos. Environ., vol. 184, pp. 129-139, Jul. 2018.
[20] Y. Zheng, F. Liu, and H.-P. Hsieh, ``U-air: When urban air quality inference meets big data,'' in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: ACM, 2013, pp. 1436-1444.
[21] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, ``Inferring air quality for station location recommendation based on urban big data,'' in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: ACM, 2015, pp. 437446.
[22] J. Wang and G. Song, ``A deep spatial-temporal ensemble model for air quality prediction,'' Neurocomputing, vol. 314, pp. 198-206, Nov. 2018.
[23] C. J. Huang and P.-H. Kuo, ``A deep CNN-LSTM model for particulate matter (PM2:5) forecasting in smart cities,'' Sensors, vol. 18, no. 7, p. 2220, 2018.
[24] H. J. Friedman, ``Greedy function approximation: A gradient boosting machine,'' Ann. Statist., vol. 29, no. 5, pp. 1189-1232, 2001.