Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022)
IEEE Xplore Part Number: CFP22BC3-ART; ISBN: 978-1-6654-9710-7

A Comparison of Oversampling and Undersampling Methods for Predicting Air Quality in Metropolitan Region

Deepali Javale, Pooja Pillai, Purvang Patel, Sushmita Jagtap
School of Computer Engineering and Technology
MIT World Peace University
Pune, Maharashtra, India
E-mail: deepali.javale@mitwpu.edu.in, 1032210350@mitwpu.edu.in, 1032210074@mitwpu.edu.in, 1032211696@mitwpu.edu.in

2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9793084

***Abstract*— Air pollution is a growing environmental concern, especially in big cities. The effects of air pollution are harmful to both living beings and the environment. Air quality can be predicted using techniques like probability, statistics but these methods are complex to predict. Machine learning is a better approach to air quality prediction. Air quality forecasting is a crucial step to protect public health by providing an early warning against harmful air pollutants. Prediction of air quality will assist in initiating emergency measures to reduce the discharge of pollutants and mitigate the consequences. It was analysed that imbalance class distribution give inaccurate predictions. As a result, two well-known resampling techniques are used: Synthetic Minority Oversampling Technique (SMOTE) for oversampling and Neighbourhood Cleaning Rule (NCR) for undersampling. The effects of these approaches and their combination (SMOTE+NCR) on prominent machine learning classifiers K-Nearest Neighbours (KNN) and Naive Bayes are compared. The presented results demonstrate that KNN performed better on resampled data using SMOTE+NCR and Naive Bayes performed better on undersampled data using NCR.**

***Keywords*— Naïve Bayes, SMOTE, NCR, KNN, Air Pollution, Undersampling, Oversampling.**

I. Introduction

Humans rely on air to survive. Monitoring and understanding its quality are crucial to our well-being. In recent years, one of India's most pressing issues has been growing air pollution. Air pollution is the most prevalent type of pollution faced by the world today and as per the Air Quality Report 2021, India ranks fifth in world's most polluted countries index. The most common cause for rising air pollution is burning of fossil fuels like coal to generate power for electricity and transportation. Another major source is from burning of agricultural waste which releases chemicals like methane and ammonia which causes respiratory illnesses like asthma.

Toxic air is a major public health and environmental issue. Due to reasons such as rapid urbanisation, industry, and population increase, air quality in India has worsened significantly over the previous few years. Human health is highly impacted by air pollution causing various health issues like difficulty in breathing, skin diseases, cancer and many more [1].

During such high-risk circumstances, it is advised to avoid any outdoor activities. Air quality monitoring stations record significant amount of data on air pollutant concentrations data across cities, which must be appropriately analysed in order to accurately predict air quality. [25] This data is typically in an imbalanced form and this paper intends to propose an innovative technique in handling such data. This paper presents a comparison of the proposed resampling techniques on two popular machine learning classifiers, K-Nearest Neighbor and Naïve Bayes. The performance of the models is evaluated on the basis of the proposed resampling techniques to determine which technique is optimal for predicting air quality.

II. Literature Review

Du et al. [2] emphasizes on building an innovative real-time warning system that predicts air pollution in China consisting of four major modules: clustering, pre-processing, forecasting, and assessment. The experimental findings of eighteen data sets from three cities are provided, which gives adequate air quality information, which is necessary for controlling air pollution. The hybrid model developed performs better than the single models and other hybrid models proposed previously.

Boonphun et al. [3] employs several machine learning algorithms to investigated the likelihood of PM2.5 surpassing a predefined safety threshold. SMOTE is used to handle the imbalance class distribution in the original dataset. It is used to oversample the minority population. A random point is selected between the minority class sample and its k-nearest neighbors. In this way the synthetic copies of minority samples are created in order to balance the class distribution in the dataset. After oversampling by SMOTE, classification models are built to estimate the risk of PM2.5 levels which are above the safety threshold. Algorithms used for implementation include Logistic Regression, Naïve Bayes, Neural Networks and Random Forest. Among these Random Forest performs best.

Srivastava et al. [4] employs different classification and regression techniques to predict AQI of major pollutants. The model is evaluated several statistical measures. The findings indicate that SVR and Neural Networks produce best results.

Mohammed et al. [5] found that oversampling outperforms undersampling for many classifiers and achieves higher scores in various assessment criteria. For distinct machine learning classifier models, oversampling outperforms undersampling. When the random undersampling method was employed instead of the random oversampling technique, the score for the classifier models was lower.

Agustianto et al. [6] used Neighborhood Cleaning Rule (NCL) to balance educational data to obtain Precision Student Modeling. The model used for classification is Decision Tree C4.5 algorithm. Results show that NCL achieved an accuracy of 91.37%

III. Objective

- Air Quality Prediction: By analysing data its air quality in metropolitan Indian cities is evaluated.
- To analyse between different machine learning classification models.
- Understanding alternative approaches to pre-processing unbalanced data and evaluating how they affect model's performance.
- Compare performance of machine learning models before and after resampling using various evaluation techniques

IV. Methodology

The architectural block diagram is as shown in Fig 1.

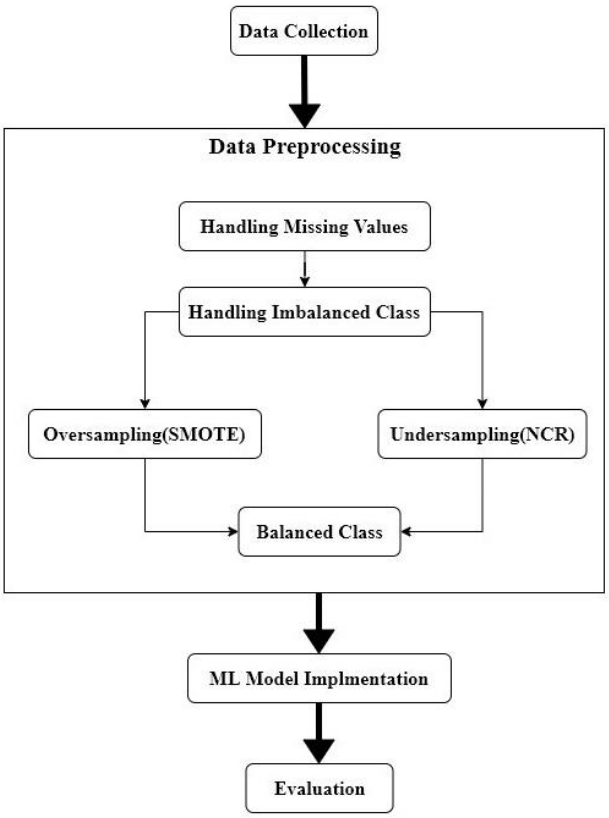

Fig. 1. Architectural block diagram

A. Data Collection

The dataset comprises pollutant levels and the computed AQI (Air Quality Index) at a daily basis from several stations across multiple cities in India. The data was collected via Kaggle, which was retrieved from the Central Pollution Control Board site. Dataset comprises pollution levels of NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, PM10, PM2.5 [12] and NO measured from 2015-2020 across 26 cities. It consists of 29,531 rows and 16 columns.

The cities included are major metropolitan cities of India like Mumbai, Delhi, Calcutta, Chennai, Bangalore etc.

Link : https://www.kaggle.com/rohanrao/air-quality-data-in-india

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN

Fig. 2. Air Quality Data (2015-2020)

B. Data Preprocessing

Data pre-processing involves two steps in this implementation.

a) Handling missing values: When working with real time data, it is particularly prevalent for missing values to occur. Handling missing values is critical in every machine learning problem since the model's performance is dependent on the data we have.[13] Hence, before training any machine learning algorithm, we must clean the data and handle missing values. Fig 3 illustrates the percentage of missing/null values in the data initially.

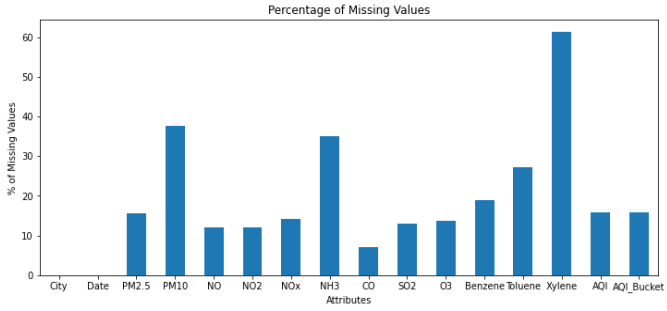

Fig. 3. Percentage of null values

Except for AQI and AQI Bucket, interpolate strategy was used to handle missing values for NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, PM10, PM2.5 and NO. Interpolate method uses many interpolation approaches for estimating unknown values like linear, time, index to mention a few. We have used 'linear' approach to fill missing values for the above listed attributes. Interpolation technique works well for time series data.

We employed Linear Regression to handle null values in 'AQI'. The 'AQI' attribute is taken as the dependent variable and the pollutant levels are taken as independent variables. Complete data instances are used to build the regression equation, which is then used to predict the missing values for incomplete instances. [12][13] In this manner, 3307 null values of the 'AQI' attribute were substituted. Subsequently the missing values in 'AQI Bucket' are filled on the basis of pre-defined buckets of AQI as illustrated in the Fig 4.

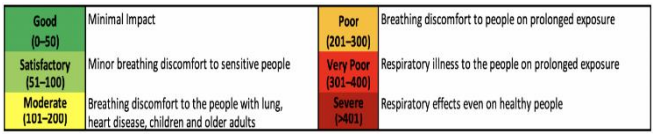

Fig. 4. Predefined Buckets of AQI

The AQI Bucket has categorical values. For model implementation, it is necessary to convert the categorical data into numerical values. Hence the values in AQI Bucket are encoded as follows:

TABLE I. Encoded categorical values

Categorical Values	Encoded Values
Good	0
Satisfactory	1
Moderate	2
Poor	3
Very Poor	4
Severe	5

b) Handling imbalance distribution: The imbalanced distribution of the target variable is handled by sampling methods as described below. We have used Synthetic Minority Over Sampling Technique (SMOTE) for oversampling and Neighbourhood Cleaning rule (NCL) for under sampling.

1) *Synthetic Minority Oversampling Technique (SMOTE):*This is the most widely used approach to oversample the minority classes by creating synthetic examples.[17][22] SMOTE begins by picking a minority class instance at random and then locating its k closest minority class neighbours. It uses the concept of K nearest neighbor. It follows the following equation:

$$Xnew = X = rand(0,1) * (X' - X) \quad (1)$$

where
X is minority class
X' is one of K nearest neighbor

2) *Neighbourhood Cleaning Rule:* NCL [7] is an undersampling strategy that reduces data based on cleaning to overcome the imbalanced class distribution. NCL focuses not just on data reduction but also on data cleaning. The data cleaning method is meant not only for samples from the majority class, but also for samples from the minority class. The NCL principle is a technique for carefully reducing classes. It categorises data into four groups: noise, borderline, redundant, and safety sample.

C. Model Implementation

In order to achieve the desired objective, we have used the below Supervised Machine Learning Techniques for model training, testing, and predicting the output. We have taken 60% of the data for training purpose and 40% for testing purpose.

a) Naïve Bayes: Naive Bayes classification algorithm is one of the simple machine learning algorithms and helps in building models quickly and produces effective results. [11] It is a classification algorithm which classifies using the concept of basic probability. It has a fast computation speed for predicting the class of test data and is well suited for multiclass prediction.

b) K Nearest Neighbors: K Nearest Neighbour [11] is a Supervised Machine Learning Algorithm. It is considered to be the most simple, easier to implement, and identified as one of the popular algorithms for predictions. It is employed in classification and regression. The input in both situations consists of the k closest training samples. As KNN is more suitable for large datasets, we have used KNN for our prediction and analysis. It is straightforward to implement and robust to noisy training data.

D. Evaluation

Comparative analysis is done for different evaluation techniques on machine learning models with and without oversampling (SMOTE) and under sampling (NCR) methods as well as their combined approach (SMOTE+NCR). We have used Accuracy, F1 Score, Precision, Recall and ROC curve as evaluation criteria [16].

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \quad (3)$$

$$Accuracy\ Rate = 100 * (TP + TN)/FN((TP + 1) + (1 + TN)) \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

where,
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

V. Result and Discussion

A. Handling imbalance data

Fig 5 shows the distribution of AQI Bucket before applying the balancing technique. It has an uneven distribution, as shown below. As the data has imbalanced class distribution,

there is a need to resample the dataset so that the classifier gives appropriate results. There are two techniques to resamples the data, they are Undersampling and Oversampling. In this paper we present a comparison of both techniques on machine learning models and present the comparative results. For oversampling, the technique chosen is SMOTE. SMOTE is one of the most popular oversampling techniques. SMOTE does not create duplicate samples rather it creates synthetic samples that differ slightly from the original samples [17][21]. Fig 6 shows balanced data distribution after resampling by SMOTE(Oversampling).

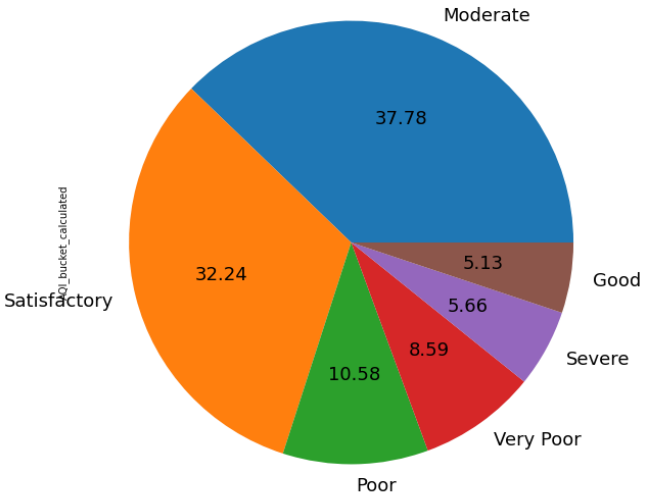

Fig. 5. Distribution of AQI bucket before sampling technique.

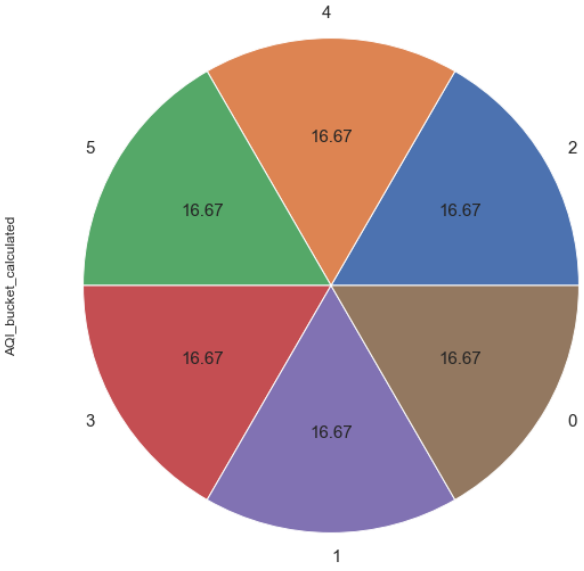

Fig. 6. Distribution of AQI Bucket after Oversampling (SMOTE)

B. Model implentation

The table below shows a comparison of the performance of various sampling techniques on machine learning algorithms. Weighted average is considered for measuring F1 Score, precision and recall. Table II shows comparison between various evaluation metrics and proposed techniques. The results show that Naïve Bayes performs better after under sampling by NCR as compared to other sampling methods proposed. Fig 7 shows the graphical representation of various performance evaluation technique.

TABLE II. Result of Naïve Bayes

Method	**Naïve Bayes**			
	Accuracy	***F1 Score***	***Precision***	***Recall***
Initial	61.73%	60.09%	63.86%	61.73%
SMOTE	63.3%	62.52%	63.25%	63.3%
NCR	**73.26%**	**73.51%**	**75.37%**	**73.26%**
SMOTE+NCR	71.06%	70.89%	71.05%	71.06%

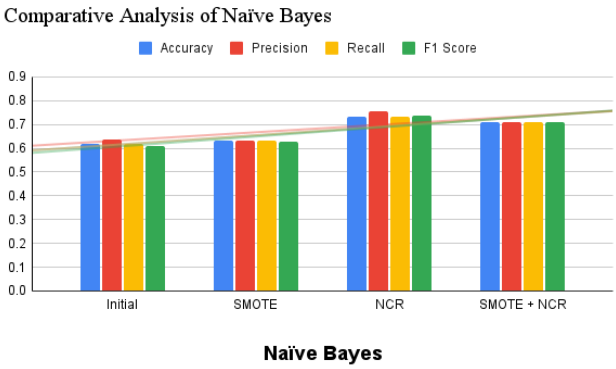

Fig. 7. Analysis of Naïve Bayes on different sampling techniques

Table III represents the comparison between various performance evaluation techniques on KNN. Results show that performance of KNN is highest on combination of SMOTE and NCR. The graphical representation of results of evaluation metrics on KNN is shown in Fig 8.

TABLE III. Result of KNN

Method	**KNN**			
	Accuracy	***F1 Score***	***Precision***	***Recall***
Initial	77.44%	77.2%	77.44%	77.22%
SMOTE	86.83%	86.69%	86.83%	86.63%
NCR	92.7%	92.64%	92.7%	92.59%
SMOTE+NCR	**94.75%**	**94.75%**	**94.75%**	**94.72%**

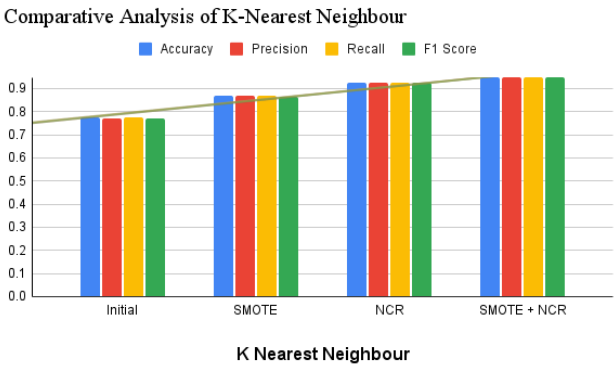

Fig. 8. Analysis of KNN on different sampling techniques

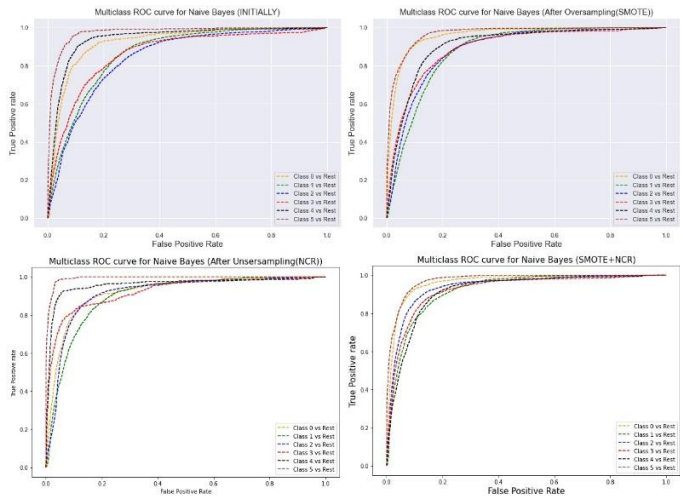

Fig 9. ROC curve depicting performance of Naïve Bayes initially and then on SMOTE, NCR and SMOTE+NCR

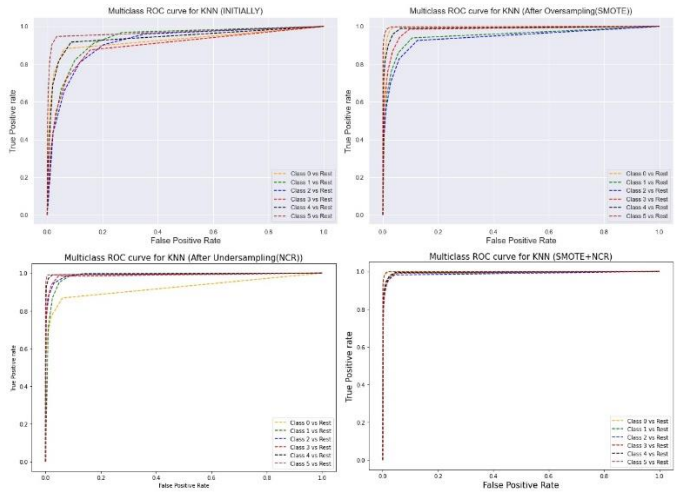

Fig 10. ROC curve depicting performance of KNN initially and then on SMOTE, NCR and SMOTE+NCR

One of the most essential evaluation measures for classification problems is ROC (Receiver Operating System). The probability curve is represented by ROC, while the degree of separability is represented by AUC, ranging between 0 and 1. If the AUC is close to one, it indicates that it has a high degree of separability. AUC close to 0 indicates that the model is not good at predicting labels.

The results reveal that the classifier's performance improves before and after resampling. Fig 9 shows the ROC curve presenting performance of Naïve Bayes initially and then compares it with SMOTE, NCR and SMOTE+NCR. AUC for Naive Bayes is initially 0.9, but after resampling with SMOTE+NCR, it rises to 0.94. The AUC score for KNN initially is 0.93, however after balancing the imbalance dataset with SMOTE+NCR, it is 0.99 as shown in Fig 10. KNN outperforms Naive Bayes in terms of performance.

VI. Future Work and Conclusion

Humans are highly affected by air pollution, causing acute and chronic diseases. Hence it is critical to monitor the changes in the air quality and to keep a keen eye on places with high and rising air pollution.

We have demonstrated in this paper that combination of SMOTE and NCR will provide higher performance and can provide better classification for imbalanced dataset. The resampling techniques used in our experiment are critical for effective data classification The proposed technique for classifying air quality will be useful for boosting existing preventive initiatives as well as improving the capabilities of effective emergency response in the worst pollution situation. Future work includes developing a wearable early warning system for persons who are extremely vulnerable to air pollution. If the device detects that a patient is suffering from a specific ailment caused by air pollution, he or she will be alerted to limit exposure to such regions.

References

[1] Kampa, Marilena, and Elias Castanas. "Human health effects of air pollution." *Environmental pollution* 151, no. 2 (2008): 362-367.

[2] Du, Zongjuan, Jiani Heng, Mingfei Niu, and Shaolong Sun. "An innovative ensemble learning air pollution early-warning system for China based on incremental extreme learning machine." *Atmospheric Pollution Research* 12, no. 9 (2021): 101153.

[3] Boonphun, Jirat, Chalat Kaisornsawad, and Papis Wongchaisuwat. "Machine learning algorithms for predicting air pollutants." In *E3S Web of Conferences*, vol. 120, p. 03004. EDP Sciences, 2019.

[4] Srivastava, Chavi, Shyamli Singh, and Amit Prakash Singh. "Estimation of air pollution in Delhi using machine learning techniques." In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 304-309. IEEE, 2018.

[5] Mohammed, Roweida, Jumanah Rawashdeh, and Malak Abdullah. "Machine learning with oversampling and undersampling techniques: overview study and experimental results." In *2020 11th international conference on information and communication systems (ICICS)*, pp. 243-248. IEEE, 2020.

[6] Agustianto, Khafidurrohman, and Prawidya Destarianto. "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling." In *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, pp. 86-89. IEEE, 2019.

[7] Gore, Ranjana Waman, and Deepa S. Deshpande. "An approach for classification of health risks based on air quality levels." In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, pp. 58-61. IEEE, 2017.

[8] Ketu, Shwet, and Pramod Kumar Mishra. "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare." *Complex & Intelligent Systems* 7, no. 5 (2021): 2597-2615.

[9] Ghaemi, Z., A. Alimohammadi, and M. Farnaghi. "LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran." *Environmental monitoring and assessment* 190, no. 5 (2018): 1-17.

[10] Aini, Nurul, and M. Syukri Mustafa. "Data mining approach to predict air pollution in Makassar." In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1-5. IEEE, 2020.

[11] Yarragunta, SriramKrishna, and Mohammed Abdul Nabi. "Prediction of Air Pollutants Using Supervised Machine Learning." In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1633-1640. IEEE, 2021.

[12] Harishkumar, K. S., K. M. Yogesh, and Ibrahim Gad. "Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models." *Procedia Computer Science* 171 (2020): 2057-2066.

[13] Ameer, Saba, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, and Muhammad Nabeel Asghar. "Comparative analysis of machine learning techniques for predicting air quality in smart cities." *IEEE Access* 7 (2019): 128325-128338.

[14] Razavi-Termeh, Seyed Vahid, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi. "Effects of air pollution in spatio-temporal modeling of asthma-prone areas using a machine learning model." *Environmental Research* 200 (2021): 111344.

[15] Hable-Khandekar, Varsha, and Pravin Srinath. "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring." In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1-6. IEEE, 2017.

[16] Xi, Xia, Zhao Wei, Rui Xiaoguang, Wang Yijie, Bai Xinxin, Yin Wenjun, and Don Jin. "A comprehensive evaluation of air pollution prediction improvement by a machine learning method." In *2015 IEEE international conference on service operations and logistics, and informatics (SOLI)*, pp. 176-181. IEEE, 2015.

[17] Rok, Blagus, and L. Lusa. "SMOTE for high-dimensional class-imbalanced data." *BMC Bioinformatics* 14, no. 1 (2013): 106-121.

[18] Liang, Yun-Chia, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez. "Machine learning-based prediction of air quality." *Applied Sciences* 10, no. 24 (2020): 9151.

[19] Su, Yuelai. "Prediction of air quality based on Gradient Boosting Machine Method." In *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, pp. 395-397. IEEE, 2020.

[20] Chen, Hongqian, Mengxi Guan, and Hui Li. "Air Quality Prediction Based on Integrated Dual LSTM Model." *IEEE Access* 9 (2021): 93285-93297.

[21] Yan, Yuanting, Ruiqing Liu, Zihan Ding, Xiuquan Du, Jie Chen, and Yanping Zhang. "A parameter-free cleaning method for SMOTE in imbalanced classification." *IEEE Access* 7 (2019): 23537-23548.

[22] Kim, Kyoungok. "Noise Avoidance SMOTE in Ensemble Learning for Imbalanced Data." *IEEE Access* 9 (2021): 143250-143265.

[23] Zhang, Ying, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, and Linyan Huang. "A predictive data feature exploration-based air quality prediction approach." *IEEE Access* 7 (2019): 30732-30743.

[24] Gu, Ke, Junfei Qiao, and Weisi Lin. "Recurrent air quality predictor based on meteorology-and pollution-related factors." *IEEE Transactions on Industrial Informatics* 14, no. 9 (2018): 3946-3955.

[25] Shaban, Khaled Bashir, Abdullah Kadri, and Eman Rezk. "Urban air pollution monitoring system with forecasting models." *IEEE Sensors Journal* 16, no. 8 (2016): 2598-2606.

[26] Chen, Hongqian, Mengxi Guan, and Hui Li. "Air Quality Prediction Based on Integrated Dual LSTM Model." *IEEE Access* 9 (2021): 93285-93297.

[27] Zhang, Dan, and Simon S. Woo. "Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network." *IEEE Access* 8 (2020): 89584-89594.

[28] Zhou, Yuchao, Suparna De, Gideon Ewa, Charith Perera, and Klaus Moessner. "Data-driven air quality characterization for urban environments: A case study." *IEEE Access* 6 (2018): 77996-78006.