Enhancing Airborne Disease Prediction: Integrating Deep Infomax and Self-Organizing Maps for Risk Factor Identification

Bhakti S. Pimpale, Dr. Anala A. Pandit
Department of Computer Application
Veermata Jijabai Technological Institute (V.J.T.I)
Mumbai, India

***Abstract*—Asthma poses a significant global public health concern, particularly in urban centers where environmental pollutants and variable weather patterns contribute to heightened prevalence and symptom exacerbation. The Deonar dumping ground, one of Mumbai's largest landfills, releases a complex mix of particulate matter and hazardous gases, posing a serious threat to local respiratory health. Despite the urgency for comprehensive research integrating patient-specific data with localized weather and air quality metrics, such studies remain limited. This study addresses the critical research gap by investigating asthma risk factors near the Deonar dumping ground. Integrating detailed patient records with precise local weather and air quality measurements, our research aims to unravel the intricate relationship between environmental exposure and respiratory health outcomes. The findings provide crucial insights into the specific risk factors influencing asthma incidence and severity in this region, informing the development of targeted interventions and mitigation strategies. Employing a novel ensemble Deep Info Max - Self-Organizing Map (DIM-SOM) technique, our study compares its performance with various clustering algorithms, including SOM, K-Means, Bisecting K-Means, DBSCAN, and others. The novel ensemble DIM-SOM demonstrated superior performance, achieving a significantly higher Silhouette Score of 0.9234, a lower Davies-Bouldin Score of 0.1276, and a more favorable Calinski-Harabasz Score of 389723.6225 compared to other algorithms. These findings underscore the efficacy of the novel ensemble DIM-SOM approach in generating dense, well-separated, and meaningful clusters, emphasizing its potential to enhance clustering performance compared to traditional algorithms. The study further emphasizes the need for proactive mitigation measures and tailored healthcare interventions based on the identified environmental risk factors.**

Keywords*—*Asthma; deepinfomax; self organizing map; risk factors; air pollution

I. Introduction

Asthma, a chronic respiratory condition marked by airway inflammation and hyper-responsiveness, has been a focal point of global health concerns. Urban environments, characterized by a dynamic interplay of environmental pollutants and ever-changing weather patterns, witness a rising prevalence of asthma and an escalation of related symptoms [1] [2] [3] [4]. Mumbai, home to over 20 million residents, grapples with the intricate relationships between air pollution, meteorological factors, and public health. The Deonar dumping ground, among India's largest landfills, stands as a substantial contributor to air pollution, raising pertinent concerns about potential health risks, particularly respiratory ailments such as asthma.

In light of the advancements and challenges faced in this domain, this study seeks to contribute to the existing body of knowledge by investigating the risk factors associated with asthma in the vicinity of the Deonar dumping ground. Existing literature as discussed in Section II highlights the complexities of understanding the intricate relationship between environmental exposure and respiratory health outcomes. Despite a growing body of research, there remains a critical gap in studies that integrate patient-specific data with localized weather and air pollution metrics. This research endeavors to bridge this gap.

The importance of this study lies in its potential to offer crucial insights into the specific risk factors contributing to the incidence and severity of asthma in the specified region. By merging detailed patient records with precise measurements of local weather patterns and air quality indices, this research aims to unravel complex relationships, providing a nuanced understanding of the environmental determinants of respiratory health. The novelty of this approach is underscored by the proposed ensemble Deep Info Max - Self-Organizing Map (DIM-SOM) technique, offering a sophisticated methodology for asthma risk factor analysis.

While existing studies provide valuable information, this research distinguishes itself through its comprehensive integration of patient-specific data and localized environmental metrics, emphasizing the specificity of the Deonar dumping ground's impact on respiratory health. By shedding light on the complex interplay between environmental exposure and asthma outcomes, this study aims to provide actionable insights for local authorities and healthcare providers. Although such insights are essential for the development of targeted interventions, mitigation strategies, and informed healthcare protocols.

The subsequent sections of this paper provide a structured approach to the study. Section I introduces the significance of the study, emphasizing the global health concern posed by asthma, particularly in urban areas with high environmental pollutant levels. Section II offers a comprehensive overview of existing literature, summarizing related work in the field. Section III details the dataset used, incorporating patient-specific data, localized weather patterns, and air quality metrics. In Section IV, the proposed ensemble Deep Info Max - Self-Organizing Map (DIM-SOM) technique is introduced, outlining the methodology for asthma risk factor analysis. Section V describes the experimental setting, including environmental setup and machine setup. Following this, Section VI presents the results obtained from the clustering approach and initiates a comprehensive discussion on their implications. In the concluding Section VII, key findings are summarized, conclusions are drawn, and avenues for future research are suggested.

II. Related Work

Asthma's global prevalence and its association with urban environments, air pollution, and complex weather patterns have been extensively studied [5] [6] [7] [8]. The Deonar dumping ground's impact on the respiratory health of Mumbai's inhabitants highlights the critical need for an in-depth exploration of the specific risks involved in this context. Understanding the interplay of environmental factors and respiratory health outcomes in this setting is essential for the development of effective intervention strategies.

The study conducted in Helsinki [9] on the inhabitants of blockhouses built on a former dump area shed light on the potential health risks associated with landfill exposure. The findings suggested a slightly increased incidence of cancer, particularly in males following prolonged residence on the dump site. Moreover, the relative risk of chronic diseases, including asthma and chronic pancreatitis, exhibited a notable elevation. The implications of these results led to the subsequent demolition of the affected houses by the Helsinki City Council.

In a research report published by Shally Awasthi, Priya Tripathi, and Rajendra Prasad, [10] the authors aimed to identify asthma risk factors, including environmental exposures like motor vehicle air pollution, industrial smoke, active and passive smoking, and exposure to environmental tobacco smoke (ETS). The analysis, conducted according to Global Initiative for Asthma guidelines, categorized participants into two age groups, revealing significant associations between the studied environmental factors and asthma. The results indicated that motor vehicle air pollution, industrial smoke, and exposure to ETS were associated with asthma in participants aged 1–15 years. Additionally, the study highlighted the role of hospitalization in asthma severity.

Nonhlanhla Tlotleng, et al from Johannesburg [11] highlighted the health risks faced by waste recyclers, emphasizing the prevalence of acute respiratory symptoms in the population. Findings revealed that exposure to waste containing chemical residues significantly increased the likelihood of respiratory symptoms. Moreover, discrepancies in symptom prevalence were observed across different landfill sites, underscoring the need for improved occupational health and safety measures. Recommendations included providing appropriate protective gear and promoting hygiene practices to mitigate health hazards associated with waste sorting among informal workers.

ShriKant Singh, Praveen Chokhandre, Pradeep S. Salve, Rahul Rajak [12] aimed to evaluate the health effects of a dumping site on the nearby community, emphasizing potential risk factors associated with solid waste management. Utilizing a case comparison design, the research identified a notable increase in respiratory illnesses, eye irritation, and stomach problems among the exposed group in comparison to the non-exposed group. The findings underscore the significant impact of exposure to the dumping site on the prevalence of respiratory illness and eye infections, as highlighted by both the Propensity Score Matching (PSM) method and multivariate analysis. Furthermore, the assessment of air-quality-index indicated concerning levels of PM10 and PM2.5 during a fire outbreak at the Deonar dumping site, reinforcing the urgency for effective waste management strategies.

The study conducted by Seung-Woo Shin et al. [13] delves into the impact of air pollution levels on the severity of asthma exacerbations. Over a ten-year period, the research collected data from 143 adult asthmatics who experienced 618 exacerbation episodes, analyzing the influence of air pollutants such as O3, SO2, and NO2. The findings reveal significant associations between asthma exacerbations and increased pollutant levels during summer and winter, emphasizing similar relative risks for moderate and severe exacerbations. These results underscore the potential risks posed by specific air pollutants on asthma severity, particularly during distinct seasons, contributing valuable insights to the field of respiratory health. The study initiated by Kebalepile MM, Dzikiti LN, Voyi K. [14] investigated acute respiratory outcomes, particularly asthma, in relation to environmental exposures using self-organizing maps (SOM), a computational intelligence paradigm of artificial neural networks (ANNs). Utilizing air quality data such as nitrogen dioxide, sulphur dioxide, and particulate matter, along with clinical and socio-demographic information, the SOM effectively classified asthma outcomes. Notably, age emerged as a significant factor, with older patients exhibiting a higher likelihood of asthma. The study highlighted the importance of SO2 as a critical pollutant requiring attention. The SOM model demonstrated a low quantization error, suggesting its efficacy in studying asthma outcomes with multidimensional data. The overall accuracy of the model was found to be 59%.

In this study [15], the authors address the existing gap in mHealth applications for asthma self-management by proposing an optimized Deep Neural Network Regression (DNNR) model. Integrating weather, demography, and asthma tracking, the model demonstrates significant potential, achieving a score of 0.83 with Mean Absolute Error (MAE) of 1.44 and Mean Squared Error (MSE) of 3.62. The authors further enhance the model's accuracy through an optimization process, resulting in a remarkable 94% accuracy rate with MAE of 0.20 and MSE of 0.09.

Based on the literature review, it is evident that existing studies have explored the relationship between environmental factors and asthma prevalence and severity. However, there is a distinct gap in research that delves into personalized risk factor identification. While prior studies have provided valuable insights into the general associations between air pollution, climatic conditions, and asthma, this research seeks to address this gap by focusing on personalized risk, thus providing a more targeted approach to asthma management and prevention. This study's approach holds promise for identifying individualized risk factors and tailoring interventions for patients facing air pollution-induced asthma exacerbations.

III. Dataset

The data for this research was collected from multiple sources to facilitate a comprehensive analysis of the risk factors

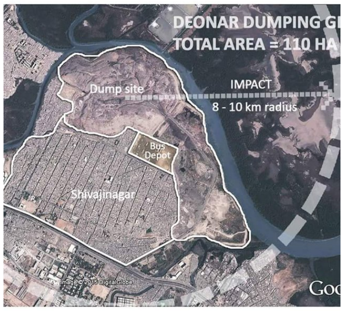


Fig. 1. Location of deonar dumping ground.

associated with asthma in the vicinity of Mumbai's Deonar dumping ground as shown in Fig. 1. The most affected areas near dumping ground are Bainganwadi, Deonar village, Govandi village, Shivajinagar, KamalaRamannagar, Rafiqnagar, Sanjaynagar and Shantinagar. Patients' data, air pollution and weather data is mainly collected from above selected areas.

A. Data Collection

1) Patient Data: Patient data, crucial for understanding the health profiles and medical histories of individuals, was obtained from Shatabdi Hospital, a government healthcare facility situated in Govandi, Mumbai. The patient data, comprising a comprehensive set of health metrics and demographic details, including Date of Visit, Age, Gender, Location, Smoking Status, Blood pressure, Body temperature, Height, Weight, Allergy status, and Symptoms of asthma (like cough intensity, sputum colour, wheezing, rales, rhonchus, etc.), spans the period from Jan 2015 to March 2020, encompassing a total of 46158 patients of asthma. The dataset contains total 76 features including physical characteristics and symptoms as discussed above. This dataset forms the basis for the study's examination of asthma incidence, symptomatology, and individual risk factors, enabling a detailed analysis of the interplay between air pollution and patient health.

2) Air Pollution Data: Air pollution data was sourced from the Copernicus Atmosphere Monitoring Service (CAMS) via the CAMS Global Reanalysis EAC4 dataset [16]. This dataset offers detailed information on air quality and atmospheric conditions. The collection period extends from Jan 2015 to March 2020, encompassing various pollutants and their concentrations, which are pivotal to assessing their impact on respiratory health in the study area. The dataset includes various key pollutants such as SO2, NO2, PM10, PM2.5, CO, O3, CH4, NH3, and dust aerosols, all of which were measured in kilograms per kilogram (kg/kg).

3) Weather Data: To understand the influence of meteorological factors on asthma risk, weather data was gathered from the NASA's Power Data Access Viewer [17]. The dataset provides a comprehensive overview of weather variables including temperature, humidity, Uv index, wind speed, dew point and rainfall etc. Data collection for weather variables aligns with the time-frame from Jan 2015 to March 2020, ensuring a holistic examination of weather patterns in the study region.

From Table I, it is clear that several parameters in the dataset display a positively skewed distribution, indicating a prevalence of higher values, while others exhibit a symmetrical skew, highlighting a more balanced distribution of values. The positively skewed parameters, including SO2, CO, NO2, O3, PM10, PM2.5, NH3, Dust Aerosol concentrations, UV index, wind speed, dew point, rainfall, and maximum temperature, exhibit a distribution where the tail of the data points extends towards higher values. This skewness indicates that, most of these parameters have relatively lower values, while a small number of data points have significantly higher values. Understanding the skewness of these parameters is crucial for assessing their impact on asthma exacerbation, as it suggests that elevated levels of these pollutants might be associated with more severe asthma symptoms. In contrast, parameters such as CH4 (methane), temperature, relative humidity, and minimum temperature demonstrate a more symmetrical distribution, where the data is relatively evenly distributed around the mean. This symmetric distribution implies that the values of these parameters do not have a strong skew towards higher or lower values. For these parameters, it is essential to explore their impact on asthma exacerbation by considering their overall levels rather than the skewness of their distribution.

B. Data Pre-processing

Prior to the analysis, a series of data pre-processing steps were applied to ensure the quality and integrity of the dataset.

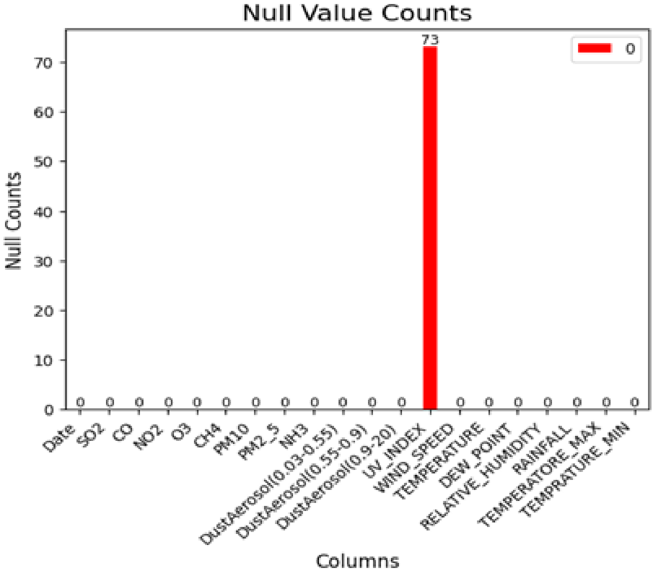


Fig. 2. Null value count of air pollution and weather data.

1) Missing values: All the datasets were first examined for missing values. In the case of the air pollution dataset, missing values were handled using the KNN Imputer [18], which imputes missing values based on the values of the nearest neighbors in the feature space. This approach enabled to account for the complex interdependencies among the pollutant variables, ensuring a more accurate representation of the air quality indicators. Air pollution data contains only 73 missing values in UV index column shown in Fig. 2. A customized function was developed, to fill the missing values

TABLE I. Statistical Representation of Air Pollution and Weather Data

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
SO2	2162.0	1.474753e-08	8.211345e-09	2.500000e-09	6.155000e-09	1.540000e-08	2.190000e-08	3.510000e-08
CO	2162.0	7.954561e-07	6.059048e-07	1.130000e-07	2.202500e-07	7.110000e-07	1.230000e-06	4.330000e-06
NO2	2162.0	1.247824e-08	6.983532e-09	2.100000e-09	5.760000e-09	1.200000e-08	1.800000e-08	3.680000e-08
O3	2162.0	8.128904e-08	4.077776e-08	2.770000e-08	4.110000e-08	7.870000e-08	1.130000e-07	2.560000e-07
CH4	2162.0	1.006156e-06	7.618521e-09	9.870000e-07	1.000000e-06	1.010000e-06	1.010000e-06	1.020000e-06
PM10	2162.0	2.339296e-07	1.675897e-07	3.510000e-08	8.315000e-08	1.775000e-07	3.670000e-07	8.600000e-07
PM2_5	2162.0	1.645160e-07	1.189400e-07	2.370000e-08	5.630000e-08	1.260000e-07	2.590000e-07	6.060000e-07
NH3	2162.0	4.738377e-10	3.942314e-10	3.530000e-11	1.310000e-10	4.320000e-10	6.490000e-10	2.300000e-09
DustAerosol(0.03-0.55)	2162.0	2.675921e-09	2.607848e-09	5.260000e-12	5.785000e-10	2.060000e-09	3.850000e-09	1.730000e-08
DustAerosol(0.55-0.9)	2162.0	5.110430e-09	5.022425e-09	6.930000e-12	9.210000e-10	3.790000e-09	7.790000e-09	3.210000e-08
DustAerosol(0.9-20)	2162.0	6.185520e-09	7.339973e-09	0.000000e+00	4.822500e-10	2.980000e-09	9.795000e-09	4.140000e-08
UV_INDEX	2162.0	1.752590e+00	5.417236e-01	2.100000e-01	1.330000e+00	1.700000e+00	2.200000e+00	3.100000e+00
WIND_SPEED	2162.0	2.509676e+00	1.006681e+00	8.600000e-01	1.830000e+00	2.230000e+00	2.900000e+00	8.270000e+00
TEMPERATURE	2162.0	2.677056e+01	2.874391e+00	1.725000e+01	2.507000e+01	2.638500e+01	2.881000e+01	3.436000e+01
DEW_POINT	2162.0	1.868569e+01	5.950174e+00	-2.540000e+00	1.438000e+01	1.998000e+01	2.410000e+01	2.590000e+01
RELATIVE_HUMIDITY	2162.0	6.707998e+01	1.847839e+01	1.681000e+01	5.320500e+01	6.565500e+01	8.619000e+01	9.525000e+01
RAINFALL	2162.0	7.148363e+00	1.765512e+01	0.000000e+00	0.000000e+00	5.000000e-02	4.400000e+00	1.438900e+02
TEMPERATORE_MAX	2162.0	3.289481e+01	4.051823e+00	2.487000e+01	2.948000e+01	3.208000e+01	3.619000e+01	4.491000e+01
TEMPRATURE_MIN	2162.0	2.216858e+01	3.766420e+00	1.051000e+01	1.912000e+01	2.339500e+01	2.502000e+01	2.891000e+01

of blood pressure and temperature in the patients' dataset. The function utilizes age-specific ranges for blood pressure and temperature, using linear interpolation to estimate missing values within the normal ranges. This approach ensures that the filled values remain within the expected physiological limits for each patient, minimizing potential data inaccuracies. Table II shows the missing value count in patients' dataset.

TABLE II. Missing Values in Patients' Data

Column	Null Count
Date	0
Diastolic Blood Pressure	3304
Systolic Blood Pressure	3320
Temperature	10141
.	.
.	.
Disease	0

2) Label encoding: To facilitate the integration of the Body Mass Index (BMI) information into the analysis, a categorical BMI variable was derived from the patients' body weight and height measurements. The BMI values were categorized into distinct groups, namely 'Underweight,' 'Normal,' and 'Overweight,' based on predefined threshold values. To incorporate this categorical information into the dataset, a one-hot encoding technique was applied. This process involved transforming the categorical variable into a set of binary variables, with each representing a specific BMI category.

3) Data normalization: To ensure consistent scaling in the dataset, an extensive analysis of various scaling methods was conducted. Initially, several known standardization techniques were explored, including the standard scaler and normalization, along with the application of the MinMaxScaler for assessing its efficacy in preserving data distribution within a specific range. However, after thorough experimentation and evaluation, the MaxAbsScaler [19] technique emerged as the most suitable choice. Results of all normalization techniques shown in Table V.

The MaxAbsScaler normalization method was selected for its remarkable ability to retain the data's inherent sparsity and preserve the relative relationships among the features. By rescaling each feature based on its maximum absolute value, the method ensured that the range of each feature fell within the [-1, 1] range. This approach not only maintained the dataset's unique characteristics but also enabled a fair comparison and interpretation of the features. Consequently, the adoption of the MaxAbsScaler technique not only provided a balanced and optimal scaling approach but also contributed to the robustness and reliability of subsequent analyses.

4) Data integration: The integration of the patient data and air pollution data involved a merging procedure based on the 'Date of Visit' and 'Date' variables from the respective datasets. This operation facilitated the alignment of the patients' symptomatology records with the corresponding air pollution data, providing insights into the potential associations between symptom onset and ambient air quality. A unified dataset was created, laying the groundwork for a detailed analysis of the interdependencies between air pollution exposure and asthma incidence.

5) Data correlation: The heatmap shown in Fig. 3 revealed several notable correlations within the dataset. Parameters such as runny nose, high cough, chest tightness, throat irritation, itchy eyes, and watery eyes exhibited a positive correlation with CO, O3, PM10, and PM2.5, indicative of potential associations between these respiratory symptoms and elevated levels of air pollutants. Conversely, these symptoms demonstrated a negative correlation with minimum temperature and Dew Point. The negative correlation indicate that cooler and more humid weather conditions could be linked to an increase in the severity of the mentioned respiratory issues.

IV. Proposed Model

In the initial stages of the study, the consolidated dataset, which included the merged information from air pollution, weather, and patient health records, underwent an extensive application of various clustering algorithms. K-means [20] [21], Self-Organizing Maps (SOM) [22], Mini-Batch K-means [23],FuzzyCmeans [24], Birch clustering [25], DBSCAN clustering [26], Agglomerative clustering [27], Spectral Clustering, Bisecting Kmeans clustering and Gaussian Mixture models [28] were each individually applied across lag 2 to 15, encompassing a comprehensive analysis of the dataset at different time intervals. Despite these efforts, the outcomes did not meet the predetermined benchmarks, indicating a need for an alternative approach. DeepInfomax algorithm [29], a deep learning

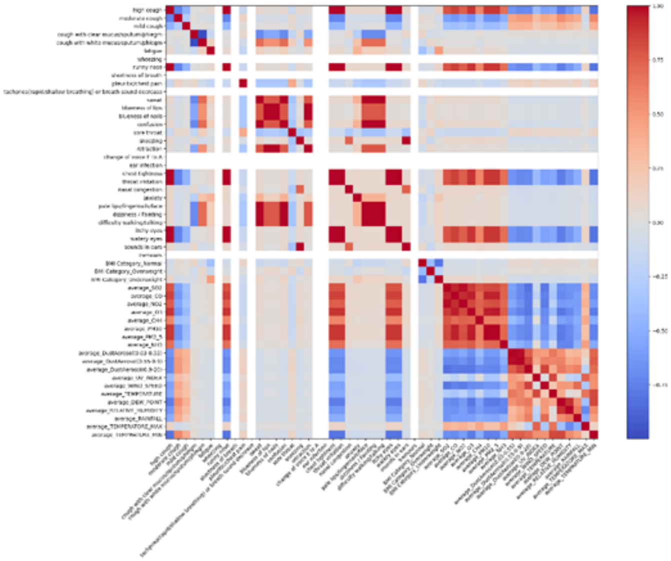


Fig. 3. HeatMap.

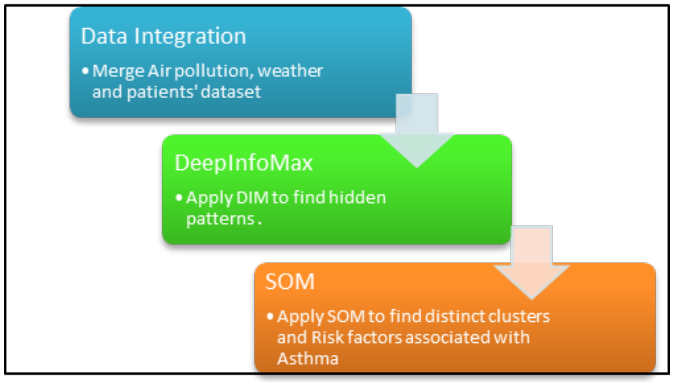

Data Integration
• Merge Air pollution, weather and patients' dataset
DeepInfoMax
• Apply DIM to find hidden patterns .
SOM
• Apply SOM to find distinct clusters and Risk factors associated with Asthma

Fig. 4. Proposed DIM-SOM Model.

technique was implemented. Although it was mainly used for images. This algorithm is capable of extracting intricate features and uncovering latent patterns within complex dataset. In the case of DIM, the model is trained to predict certain parts of the input data from other parts, without requiring explicit supervision from labeled data.

Using the power of the DeepInfomax technique, intricate associations and previously unnoticed patterns within the combined dataset numeric in nature were observed. This helps to better understand the complex relationships within the dataset, which traditional clustering techniques couldn't capture well.

Later the Self-Organizing Map (SOM) technique was used to further refine the clustering outcomes and establish distinct clusters and associated labels within the dataset. This combined methodology yielded a notable enhancement in the evaluation metric. This integrated approach not only improved the precision of the clustering process but also facilitated a holistic understanding of the multifaceted factors contributing to the incidence and patterns of asthma in relation to the interconnected influences of air pollution and weather conditions over the specified lag. Fig. 4 shows the proposed model with three different stages like data integration, pattern recognition and clustering followed by risk factor analysis.

V. Experimental Settings

The experiments were performed on an Intel Core i5 @ 1.19 GHz and 8 GB of memory. Python software was used for model creation and prototyping because it includes publicly accessible library sets for machine learning and statistical methods like Scikit-learn, and Matplotlib. Modeling tests were run to confirm the efficacy of the proposed model. Data and results were plotted using the Python 2D graphing tool included in the Matplotlib. The effectiveness of the model was examined using Sklearns cluster metrics, a Python module for performance measurements of machine learning models. All experiments were also conducted using Google Colab, a cloud-based Jupyter notebook environment that enabled seamless integration with Google Drive and provided access to high-performance computing resources. Validated our results using earlier method.

VI. Result and Discussion

The aim of this study was to identify clusters that could analyze the impact of air pollution on patients living near a dumping ground along with the associated risk factors. To accomplish this, determining the precise time-frame between disease onset and patient treatment was crucial. Multiple methods were utilized to determine the most efficient approach, and the results were documented and are presented in Table III and Table IV for further analysis.

However, the results obtained from these different methods did not reveal clear or meaningful patterns in the data. This unexpected outcome posed a challenge in selecting the appropriate lag for subsequent analysis. Therefore, identifying the best lag remains a crucial aspect of the research, prompting to explore alternative approaches to address this issue.

To determine the optimal lag, the novel technique of creating an ensemble using the Deep Info Max (DIM) initially, followed by the application of the self-organizing map (SOM) was implemented. The choice of the SOM algorithm was based on its ability to maintain the topological characteristics of the data, thereby uncovering the fundamental structure and interrelationships among data points. This capability becomes particularly pertinent when dealing with intricate, high-dimensional datasets, where preserving the original data structure is paramount. Unlike the K-Means algorithm, SOM excels in capturing non-linear associations within the data, proving invaluable for datasets where clusters may lack clear boundaries or exhibit complex configurations. Additionally, SOM demonstrates greater resilience to outliers and noise, thereby minimizing the influence of such irregularities on the clustering outcomes. Moreover, SOM can function as a dimensionality reduction technique during the clustering process, which proves especially advantageous when dealing with datasets featuring numerous dimensions(in this case 97 features). This reduction can simplify data exploration and comprehension, facilitating a more accessible understanding of the data's underlying patterns and relationships. Architecture of DIM is shown in Fig. 5 The DIM neural network architecture comprises an input layer designed to accommodate data

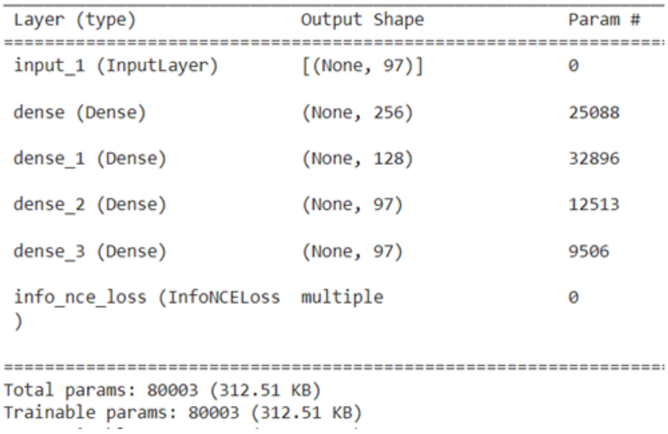
TABLE III. SILHOUETTE SCORES FOR KMEANS, MINIBATCH KMEANS, SOM, GAUSSIAN MIXTURE AND FUZZYCMEANS

Silhouette Index					
Lag	Kmeans	Minibatch Kmeans	Self-organizing maps	Gaussian Mixture	FuzzyCmeans
2	0.199	0.272	0.260	0.181	0.242
3	0.270	0.196	0.259	0.180	0.242
4	0.270	0.203	0.260	0.121	0.242
5	0.271	0.193	0.260	0.186	0.244
6	0.272	0.186	0.259	0.184	0.245
7	0.272	0.203	0.259	0.184	0.245
8	0.274	0.269	0.259	0.178	0.246
9	0.276	0.199	0.259	0.176	0.248
10	0.278	0.204	0.260	0.177	0.249
11	0.279	0.237	0.261	0.176	0.250
12	0.280	0.203	0.262	0.152	0.250
13	0.281	0.207	0.262	0.192	0.251
14	0.281	0.200	0.264	0.179	0.251
15	0.282	0.211	0.265	0.190	0.252

TABLE IV. SILHOUETTE SCORES FOR BIRCH, DBSCAN, AGGLOMERATIVE, SPECTRAL AND BISECTING KMEANS CLUSTERING

Silhouette Index					
Lag	Birch	DBSCAN	Agglomerative	Spectral	Bisecting kmeans
2	0.153	-0.0935	0.317	0.318	0.298
3	0.157	-0.1047	0.319	0.321	0.301
4	0.159	-0.0914	0.321	0.320	0.301
5	0.162	-0.1105	0.321	0.321	0.301
6	0.163	-0.0999	0.322	0.324	0.300
7	0.163	-0.0999	0.322	0.324	0.300
8	0.169	-0.1067	0.327	0.325	0.298
9	0.171	-0.0987	0.327	0.328	0.297
10	0.173	-0.1031	0.329	0.329	0.302
11	0.218	-0.0981	0.330	0.328	0.308
12	0.219	-0.1179	0.327	0.330	0.310
13	0.220	-0.1145	0.333	0.330	0.310
14	0.219	-0.0992	0.331	0.331	0.312
15	0.218	-0.0976	0.333	0.331	0.312

instances of 97 features. This is followed by a sequence of four densely connected layers, namely, 'dense', 'dense1', 'dense2', and 'dense3', responsible for hierarchically processing the input data. These layers utilize various mathematical operations, including weighted summation and activation functions, to extract and transform the input data into meaningful representations. Furthermore, the inclusion of a custom 'InfoNCELoss' layer underscores the network's reliance on the Information Noise-Contrastive Estimation (InfoNCE) loss function. By leveraging this tailored layer, the model is adept at maximizing the agreement between representations of related instances while minimizing the agreement between representations of unrelated instances. This approach, commonly associated with self-supervised learning, facilitates the acquisition of effective data representations, consequently enabling the model to find intricate patterns and relationships within the dataset.



```
Layer (type)                Output Shape              Param #
=================================================================
input_1 (InputLayer)        [(None, 97)]              0

dense (Dense)               (None, 256)               25088

dense_1 (Dense)             (None, 128)               32896

dense_2 (Dense)             (None, 97)                12513

dense_3 (Dense)             (None, 97)                9506

info_nce_loss (InfoNCELoss  multiple                  0
)

=================================================================
Total params: 80003 (312.51 KB)
Trainable params: 80003 (312.51 KB)
```

Fig. 5. Architecture of DIM.

Let $\mathbf{X}$ denote the input data with dimensions $(N, 97)$, where N represents the batch size. The neural network architecture can be symbolically represented as follows:

$$\text{Input Layer:} \quad \mathbf{X} \in \mathbf{R}^{N \times 97} \tag{1}$$

$$\text{Dense Layer 1:} \quad \mathbf{H}^{(1)} = \sigma(\mathbf{X} \cdot \mathbf{W}^{(1)} + \mathbf{b}^{(1)}) \tag{2}$$

$$\text{Dense Layer 2:} \quad \mathbf{H}^{(2)} = \sigma(\mathbf{H}^{(1)} \cdot \mathbf{W}^{(2)} + \mathbf{b}^{(2)}) \tag{3}$$

$$\text{Dense Layer 3:} \quad \mathbf{H}^{(3)} = \sigma(\mathbf{H}^{(2)} \cdot \mathbf{W}^{(3)} + \mathbf{b}^{(3)}) \tag{4}$$

$$\text{Dense Layer 4:} \quad \mathbf{H}^{(4)} = \sigma(\mathbf{H}^{(3)} \cdot \mathbf{W}^{(4)} + \mathbf{b}^{(4)}) \tag{5}$$

$$\text{InfoNCELoss Layer:} \quad \text{Custom layer utilizing InfoNCE} \tag{6}$$

In the above representation, σ denotes the activation function, such as the LeakyReLU function, which introduces non-linearities into the network. $W^{(i)}$ represents the weight matrix and $b^{(i)}$ denotes the bias vector for the ith dense layer.

The results of the novel ensemble Deep Info Max - Self-Organizing Map (DIM-SOM) technique, presented in Table VI, reflect the evaluation based on prominent clustering metrics, including the Silhouette Score [30], Calinski Harabasz Score [31], and Davies Bouldin Score [32]. The findings illustrate the complex dynamics involved in choosing the ideal lag value.

Among the considered lag values, the analysis spotlights lag 12 as a prominent contender with an impressive Silhouette Score of 0.9234, indicating distinct and well-defined clusters within the dataset. Moreover, the notably high Calinski-Harabasz Score of 389723.6225 for lag 12 further signifies the presence of dense and well-separated clusters, in producing meaningful clustering outcomes. Complementing these findings, the relatively low Davies-Bouldin Score of 0.1276 for lag 12 underlines the improved cluster separation compared to other lag values. Hence novel ensemble DIM-SOM technique, emphasizing its potential to generate compact, well-separated, and distinct clusters was established.

Further insight into the clustering performance was gained through a comparison of the self-organizing map (SOM) and the novel ensemble DIM-SOM. Table VII illustrates the comparison, highlighting the superior performance of the novel ensemble DIM-SOM over the standard SOM algorithm. Specifically, the novel ensemble DIM-SOM algorithm attained a significantly higher Silhouette Score of 0.9234, a lower Davies-Bouldin Score of 0.1276, and a more favorable Calinski-Harabasz Score of 389723.6225, surpassing the respective metrics achieved by the SOM algorithm (Silhouette Score: 0.2619, Davies-Bouldin Score: 1.6610, Calinski-Harabasz Score: 3040.4192). These findings underscore the efficacy of the novel ensemble DIM-SOM approach in generating dense, well-separated, and meaningful clusters, further emphasizing its potential to enhance clustering performance in comparison to the traditional SOM algorithm.

TABLE V. Result of Different Normalization Methods using DIM-SOM

Novel ensemble DIM - SOM using MinMaxScaler			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	0.64	14955.70	0.64
3	0.47	13708.11	0.76
4	0.47	13708.11	0.76
5	0.76	21635.34	0.44
6	0.57	19263.01	0.59
7	0.88	151696.15	0.17
8	0.45	8659.86	1.01
9	0.72	44186.74	0.39
10	0.62	9825.42	0.56
11	0.59	19719.31	0.64
12	0.34	4541.49	1.36
13	0.44	7187.50	0.99
14	0.49	8625.06	0.98
15	0.57	21352.39	0.62
Novel ensemble DIM - SOM using Standard Scaler			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	0.64	24194.60	0.49
3	0.60	15814.63	0.60
4	0.57	18366.48	0.55
5	Nan	Nan	Nan
6	Nan	Nan	Nan
7	0.56	17819.22	0.57
8	Nan	Nan	Nan
9	0.57	13829.26	0.62
10	Nan	Nan	Nan
11	Nan	Nan	Nan
12	Nan	Nan	Nan
13	Nan	Nan	Nan
14	Nan	Nan	Nan
15	Nan	Nan	Nan
Novel ensemble DIM - SOM using Normalizer(L1)			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	0.97	806.28	1.07
3	0.45	905.65	1.04
4	0.35	750.41	1.09
5	0.94	1466.52	1.09
6	0.70	1000.87	1.05
7	-0.06	865.39	1.06
8	0.56	895.44	1.09
9	0.72	799.64	1.06
10	-0.06	937.34	1.05
11	0.10	996.51	1.05
12	0.68	791.87	1.05
13	**1.00**	**1017.05**	**1.06**
14	0.06	920.00	1.07
15	0.40	1127.97	1.04
Novel ensemble DIM - SOM using Normalizer(L2)			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	-0.06	1806.77	1.03
3	-0.14	956.99	1.08
4	0.41	650.60	1.08
5	Nan	Nan	Nan
6	Nan	Nan	Nan
7	Nan	Nan	Nan
8	-0.06	1132.53	1.05
9	0	1227.77	1.08
10	**1**	**754.18**	**1.03**
11	0.29	649.79	1.04
12	0.37	852.65	1.03
13	-0.26	558.91	1.04
14	Nan	Nan	Nan
15	0.61	16888.15	0.53
Novel ensemble DIM - SOM using Normalizer(Max)			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	Nan	Nan	Nan
3	0.57	13159.36	0.63
4	0.56	1148.30	1.04
5	-0.05	920.83	1.05
6	0.00	682.11	1.02
7	0.60	17444.69	0.54
8	0.64	17009.87	0.55
9	Nan	Nan	Nan
10	**1.00**	**932.70**	**1.09**
11	0.92	875.73	1.04
12	0.62	19727.16	0.52
13	0.00	1040.48	1.12
14	Nan	Nan	Nan
15	Nan	Nan	Nan

TABLE VI. Novel Ensemble DIM - SOM using MaxAbsScaler

Novel ensemble DIM - SOM using MaxAbsScaler			
Lag	**Silhouette Score**	**Calinski-Harabasz Score**	**Davies-Bouldin Score**
2	0.7441	58304.0288	0.3750
3	0.8456	87980.6821	0.2492
4	0.3255	3466.9936	1.5385
5	0.5793	20339.4434	0.5563
6	Nan	Nan	Nan
7	Nan	Nan	Nan
8	0.4056	8932.7228	0.9797
9	0.6715	33287.8864	0.4172
10	Nan	Nan	Nan
11	Nan	Nan	Nan
12	**0.9234**	**389723.6225**	**0.1276**
13	0.4699	12359.7120	0.7685
14	0.5233	15293.3356	0.6500
15	0.6336	19533.1530	0.6197

TABLE VII. Comparison of Novel Ensemble DIM-SOM and SOM

Comparison of Novel Ensemble DIM-SOM and SOM			
Model	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
DIM-SOM	**0.9234**	**389723.6225**	**0.1276**
SOM	0.2619	3040.4192	1.6610

Risk factor analysis The role of environmental factors in exacerbating asthma symptoms remains a critical area of investigation. The analysis focused on exploring the influence of air pollution, meteorological conditions, and obesity on two distinct clusters of asthma patients. The feature importance analysis is shown in Fig. 7 which notably indicate the varying impact of these factors on the identified clusters, describing important insights into the complex interactions between health outcomes and environmental conditions. The visualization from Fig. 6 clearly indicates the division of the data into two distinct clusters. Subsequent analysis enabled the classification of these clusters as “Asthma Aggravated by Air Pollution” and “Asthma Not Significantly Affected by Air Pollution”.

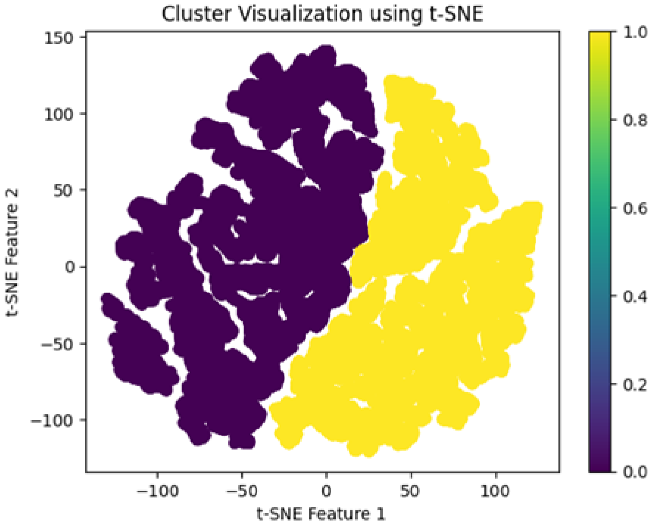


Fig. 6. Cluster visualization.

Asthma Aggravated by Air Pollution Cluster: The analysis revealed a significant association between air pollution and the severity of asthma symptoms in this cluster. The feature importance analysis highlighted the substantial impact of various pollutants, including average SO2, average CO, average NO2, average O3, average CH4, average PM10, and average PM2.5. Additionally, meteorological factors such as average TEMPERATURE, average DEW POINT, and average RELATIVE HUMIDITY exhibited a relationship with air pollutants, amplifying the adverse effects on asthma exacerbation. Asthma Not Significantly Affected by Air Pollution Cluster: In this cluster, the influence of air pollution on asthma symptoms appeared notably milder compared to the first cluster. Nevertheless, specific pollutants, particularly average PM10 and average PM2.5, still demonstrated a discernible impact, albeit to a very lesser extent. Moreover, meteorological parameters, including average TEMPERATURE, average WIND SPEED, and average RELATIVE HUMIDITY, played a prominent role, indicating their potential contribution to the manifestation of asthma symptoms. Obesity has long been recognized as a risk factor for various health conditions, including asthma. The feature importance analysis in both clusters elucidated the noteworthy impact of different BMI categories, emphasizing the crucial role of weight status in the severity and exacerbation of asthma symptoms. The analysis revealed the following insights: In both clusters, the feature importance values for both BMI Category Overweight and BMI Category Underweight suggested a similar negative impact, indicating that being either underweight or overweight may contribute to the aggravation of asthma symptoms (Table VIII).

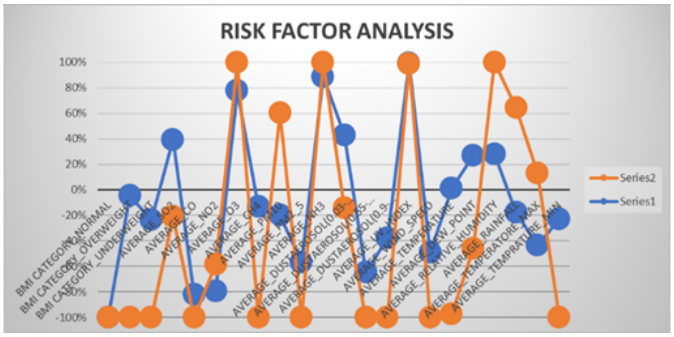


Fig. 7. Risk factor analysis.

TABLE VIII. Comparison of Studies and Proposed Model

Study	Focus	Key Findings	Results
Helsinki[9]	Landfill Exposure	Increased cancer incidence; elevated relative risk of chronic diseases, including asthma.	High correlation of air pollutants and pulmonary diseases.
Awasthi et al.[10]	Asthma Risk Factors	Significant associations between motor vehicle air pollution, industrial smoke, and asthma in participants aged 1–15 years.	Children are more vulnerable to the adverse effects of high air pollution.
Tlotleng et al.[11]	Waste Recyclers' Health	Increased respiratory symptoms in waste recyclers due to exposure to waste containing chemical residues.	Noticable correlation of respiratory symptoms and air pollution in the vicinity of dumping ground.
Singh et al.[12]	Dumping Site Impact	Notable increase in respiratory illnesses, eye irritation, and stomach problems among the exposed group as compared to controlled group.	High correlation of symptoms and air pollution.
Shin et al.[13]	Air Pollution and Asthma	Significant associations between asthma exacerbations and increased pollutant levels during summer and winter.	High correlation of air pollutants and weather parameters with asthma symptoms.
Kebalepile et al.[14]	SOM for Asthma Outcomes	SOM classified asthma outcomes based on air quality data and socio-demographic information. Age was a significant factor.	Classification Accuracy: 59%
Model for regression[15]	mHealth Application	Optimized DNNR model for asthma self-management.	Achieved 94% accuracy.
Proposed Model DIM-SOM	**Clustering Performance,for Asthma risk factor analysis**	**Higher Silhouette Score, lower Davies-Bouldin Score, and favorable Calinski-Harabasz Score for DIM-SOM.**	**Silhouette Score:0.92, Davies-Bouldin Score:0.12, Calinski-Harabasz Score:389723.62**

VII. Conclusion

This study has undertaken a comprehensive exploration of the intricate interplay between environmental factors and their impact on respiratory health, particularly in the context of the Deonar dumping ground in Mumbai. By employing a sophisticated novel ensemble model, which combines patient-specific data, localized weather patterns, air quality metrics, and the innovative ensemble Deep Info Max - Self-Organizing Map (DIM-SOM) technique, the multifaceted risk factors contributing to the incidence of asthma within this region were identified. The findings have unveiled the pivotal role played by environmental pollutants and meteorological variations in exacerbating(synonym) respiratory ailments, highlighting the heightened vulnerability of the local population to asthma. The novel ensemble DIM-SOM algorithm, introduced in this literature has demonstrated its efficacy in generating meaningful clusters, signifying its potential for accurate and interpretable insights into the complex relationship between environmental exposure and respiratory health. In the wake of these insights, this research calls for the implementation of proactive mitigation measures and tailored healthcare interventions to address the specific challenges posed by the Deonar dumping ground. By incorporating the results into policy planning and public health initiatives, local authorities and healthcare stakeholders can design effective strategies to alleviate the impact of environmental hazards on respiratory well-being in the affected communities.

Moreover, this study offers the potential for developing sustainable, evidence-based interventions that can safeguard public health in regions confronted with similar environmental challenges, further reinforcing the importance of ensemble methodologies in complex data analysis and pattern recognition within the context of public health.

While this study provides crucial insights into the complex interplay between environmental factors and respiratory health in the Deonar dumping ground area, it is important to acknowledge certain limitations. The reliance on retrospective data poses constraints on the establishment of causal relationships, warranting further longitudinal investigations to ascertain the temporality and directionality of the identified associations. Additionally, the study's focus on a specific geographical location necessitates caution in generalizing the findings to broader populations, emphasizing the need for multi-site studies to enhance the external validity of the results.

In terms of future scope, the incorporation of genetic and epigenetic data in conjunction with environmental factors could offer a more nuanced understanding of the individual susceptibility to asthma in polluted environments. Moreover, creating specific plans to help the community get involved and actively participate could be very effective in dealing with the many challenges caused by the environmental issues and respiratory health in similar areas.

Acknowledgment

We would like to express our sincere gratitude to Dr. Priyanka Kulkarni BAMS, MD - Ayurveda, and, Dr. Sonali Renushe BAMS, MD - Ayurveda for their essential guidance and support in assisting us to comprehend the medical component of our research work. Our research has been substantially enhanced by their insights into it.

References

[1] Y. Zhang, X. Yin, and X. Zheng, *"The relationship between PM2.5 and the onset and exacerbation of childhood asthma: a short communication," Frontiers in Pediatrics,Frontiers Media SA,* vol. 11. August, 2023.

[2] M. Barnthouse, BL. Jones,*"The impact of environmental chronic and toxic stress on asthma," Clin Rev Allergy Immunol.* vol.57,no.3, pp.427–38,2019.

[3] Lara Joanna Macedo Borges, *"The Effect Of Biomedical Waste Disposal In Surrounding Areas Of Deonar Dumping Ground," International Journal Of Legal Developments And Allied Issues*, Vol. 8, no. 1,pp. 128-165, 2022.

[4] P. Tripathy and C. McFarlane, *"Perceptions of atmosphere: Air, waste, and narratives of life and work in Mumbai"*, *Environment and Planning D: Society and Space*, vol. 40, no. 4, pp. 664-682, 2022.

[5] A. Abbah, S. Xu, and A. Johannessen, *"Long-term exposure to outdoor air pollution and asthma in low- and middle-income countries: A systematic review protocol," PLoS One*, vol. 18, no. 7, 2023.

[6] D. Singh, I. Gupta, and A. Roy, *"The association of asthma and air pollution: Evidence from India"*, *Economics and Human Biology*, vol. 51, p. 101278, 2023.

[7] C. Hoffmann, M. Maglakelidze, E. von Schneidemesser et al., *"Asthma and COPD exacerbation in relation to outdoor air pollution in the metropolitan area of Berlin, Germany"*, *Respir Res*, vol. 23, p. 64, 2022.

[8] J. Chatkin, L. Correa, U. Santos, *"External Environmental Pollution as a Risk Factor for Asthma"*, *Clinical Reviews in Allergy & Immunology*, vol. 62, pp. 1-18, 2022.

[9] E. Pukkala, A. Pönkä, *"Increased incidence of cancer and asthma in houses built on a former dump area"*, *Environ Health Perspect*, vol. 109, no. 11, pp. 1121-1125, 2001.

[10] A. Lotfata, M. Moosazadeh, M. Helbich, and B. Hoseini,*"Socioeconomic and environmental determinants of asthma prevalence: a cross-sectional study at the U.S. County level using geographically weighted random forests."*, *International Journal of Health Geographics*, Springer Science and Business Media LLC, vol. 22, no. 1, 2023.

[11] N. Tlotleng et al., *"Prevalence of Respiratory Health Symptoms among Landfill Waste Recyclers in the City of Johannesburg, South Africa."*, *International journal of environmental research and public health*, vol. 16, no. 21, p. 4277, 2019.

[12] S. Singh, P. Chokhandre, P. Salve, and R. Rajak *"Open dumping site and health risks to proximate communities in Mumbai, India: A cross-sectional case-comparison study."*, *Clinical Epidemiology and Global Health*, vol. 9, 21, pp. 34-40, 2020.

[13] S. Shin et al., *"Effects of air pollution on moderate and severe asthma exacerbations."*, *The Journal of asthma : official journal of the Association for the Care of Asthma*, vol. 57, no. 8, pp. 875-885, 2020.

[14] M. Kebalepile, L.Dzikiti, and K. Voyi, *"Supervised Kohonen Self-Organizing Maps of Acute Asthma from Air Pollution Exposure."*, *International Journal of Environmental Research and Public Health*, MDPI AG, vol. 18, no. 21. p. 11071,2021.

[15] Haque R et al., *"Optimised deep neural network model to predict asthma exacerbation based on personalised weather triggers."*, *F1000Res*, vol. 10, pp. 911, 2021.

[16] CAMS. Available online: https://atmosphere.copernicus.eu/data (accessed on 28 August 2023).

[17] NASA. Available online: https://power.larc.nasa.gov/data-access-viewer/ (accessed on 28 August 2023).

[18] Murti, D. Prawidya, U. Pujianto, A. Wibawa, and M. Akbar. *"K-Nearest Neighbor (K-NN) based Missing Data Imputation," 2019 5th International Conference on Science in Information Technology (ICSITech)*,pp. 83-88.

[19] K. N. Abd Halim, A. S. Mohd Jaya, and A. F. A. Fadzil, *"Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing," International Journal of Innovative Technology and Exploring Engineering,* vol. 9, no. 3, pp. 272–277, Jan. 30, 2020.

[20] S. P. Lloyd, *Least squares quantization in PCM.* Technical Report RR-5497, Bell Lab, September 1957.

[21] J. B. MacQueen, *"Some methods for classification and analysis of multivariate observations".* In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, California: University of California Press,Vol. 1, pp. 281–297 , 1967.

[22] T. Kohonen, "The self-organizing map," *in Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, Sept. 1990.

[23] D. Sculley, *"Web-scale k-means clustering," Proceedings of the 19th international conference on World wide web.* ACM, Apr. 26, 2010.

[24] J. Bezdek *"Fuzzy C-means cluster analysis", Scholarpedia,* vol. 6, no. 7, pp. 2057, 2011.

[25] T. Zhang, R. Ramakrishnan, and M. Livny,*"BIRCH, " Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96.* ACM Press, 1996.

[26] M. Ester, H. Kriegel, J. Sander, and Xiaowei Xu, *"A density-based algorithm for discovering clusters in large spatial databases with noise," In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96).* AAAI Press, pp.226–231.1996.

[27] F. Murtagh and P. Legendre, *"Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," Journal of Classification,* Springer Science and Business Media LLC Oct. vol. 31, no. 3, pp. 274–295, 2014.

[28] D. Reynolds, *"Gaussian Mixture Models," Encyclopedia of Biometrics.* Springer US, pp. 659–663, 2009.

[29] R. D. Hjelm et al., *"Learning deep representations by mutual information estimation and maximization." arXiv*, pp. 1-24, 2018.

[30] P. J. Rousseeuw,*"Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics*,Elsevier BV, vol. 20 , pp. 53–65, Nov. 1987.

[31] T. Calinski and J. Harabasz,*"A dendrite method for cluster analysis," Communications in Statistics - Theory and Methods*, vol. 3, no. 1. Informa UK Limited, pp. 1–27, 1974.

[32] D. L. Davies and D. W. Bouldin, *"A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence*,Institute of Electrical and Electronics Engineers (IEEE), vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.