Air Quality Data Analysis And Prediction Using Modified Differential Evolution - Random Forest Algorithm

K. Cornelius[1*], Dr. B. Shanthini[2]

[1*]Research Scholar, Department of CSE, St. Peter's Institute of Higher Education and Research, Chennai.,
Email: cornelius.cse@spiher.ac.in
[2]Professor and Head, Department of CSE and IT, St. Peter's Institute of Higher Education and Research, Chennai.

***Corresponding Author:** K. Cornelius
[*]Research Scholar, Department of CSE, St. Peter's Institute of Higher Education and Research, Chennai.,
Email: cornelius.cse@spiher.ac.in

Abstract:
Due to its devastating impact on human health and the environment, air pollution has emerged as a major global issue. We use the random forest technique to analyze and forecast air quality levels in this study. The purpose of this study is to use the random forest algorithm to investigate and forecast air quality in Chennai, Tamil Nadu. Three monitoring stations (Alandur Bus Depot, Manali Village, and Velachery Res. Area) provide a substantial dataset of air quality measurements used in the study. Features like meteorological data, geographical information, and temporal patterns are added to the model to enhance its forecasting powers. These characteristics supply background data that can be used to better interpret associations between air quality and other variables. Due of its stability and scalability, the random forest approach was used. The input features and air quality categories are utilized to train the model, which is then put through its paces on the testing set. The accuracy of the trained model predictions is then measured against the testing set. The findings could be used to improve Chennai air quality management. The proposed model using Modified Differential Evolution based Random Forest (MDE-RF) can help policymakers, environmental agencies, and public health groups make better decisions and put into action more effective actions to reduce the effects of air pollution. In addition, it provides new opportunities for studying the dynamics of air pollution in metropolitan areas and developing cutting-edge machine learning algorithms for air quality prediction.

Keywords: Air quality, Random forest algorithm, Data analysis, Prediction, Health impact

1. Introduction

The issue of air pollution has garnered worldwide attention as a result of its adverse impacts on both human well-being and the natural environment. [1] The phenomenon is attributed to a range of pollutants that are emitted from industrial operations, vehicular exhaust, and other sources. Chennai, the administrative hub of Tamil Nadu, is not immune to this difficulty as it contends with increased levels of air pollution. In order to effectively tackle this matter, it is imperative to conduct a thorough and precise analysis and forecast of air quality levels. This study aims to employ the random forest algorithm for the analysis and prediction of air quality levels in Chennai, Tamil Nadu.

The association between air pollution and a range of health complications, such as respiratory ailments, cardiovascular disorders, and premature mortality, has been established. Additionally, it exerts an impact on the environment through its contribution to climate change and the detrimental effects it has on ecosystems. Monitoring air quality and predicting pollution levels are crucial in order to promptly implement measures to mitigate the effects of pollution and safeguard both human health and the environment [2].

The prediction of air quality poses a formidable challenge owing to the intricate nature of the underlying factors that exert an influence on pollution levels. The complex interaction between meteorological conditions, geographical factors, emissions from diverse sources, and temporal patterns poses challenges in accurately forecasting air quality levels. Furthermore, the presence of dependable and all-encompassing datasets is imperative for the development of efficient prediction models [3].

This research focuses on the precise forecasting of air quality levels in Chennai, Tamil Nadu. The aim of this study is to construct a resilient model capable of categorizing air quality into four distinct classifications: Healthy, Unhealthy, Moderate, and Hazardous. The model ought to take into account a range of factors, encompassing pollutant concentrations, meteorological data, geographical information, and temporal patterns.

The main aim of this study is to examine and forecast the levels of air quality in Chennai by utilizing the random forest algorithm. The primary objective is to construct a precise and dependable model capable of categorizing air quality into

distinct classifications. To enhance the predictive capabilities, it is recommended to integrate supplementary factors into the model, including meteorological and geographical data.

The novelty of this study resides in the application of the random forest algorithm for the purpose of air quality prediction in the city of Chennai. Prior research has investigated different machine learning methodologies; however, the utilization of the random forest algorithm for air quality forecasting in this particular area is distinctive. Moreover, the integration of supplementary elements, such as meteorological and geographical data, augments the predictive capacities of the model. The research presented in this study offers a robust approach using Modified Differential Evolution based Random Forest (MDE-RF) for the analysis and prediction of air quality levels in Chennai. The model that has been developed possesses the capability to aid policymakers, environmental agencies, and public health organizations in making well-informed decisions and implementing suitable measures to alleviate the effects of air pollution. Additionally, the results of this study provide opportunities for further investigation into advanced machine learning methodologies for the prediction of air quality. Furthermore, these findings contribute to a broader understanding of the dynamics of air pollution in urban regions.

2. Related Works

This section presents a thorough examination of the existing body of literature concerning air pollution and its effects on both human health and the environment. The text examines the diverse origins of air pollution, encompassing industrial emissions, vehicular exhaust, and agricultural activities. The model in [4-7] examines the adverse impacts of air pollution on respiratory health, cardiovascular disorders, and the environment. Furthermore, the present discourse delves into investigations pertaining to the correlation between air pollution and climate change.

The issue of air pollution is of significant global importance owing to its detrimental impacts on both human health and the environment. The methods in [7-10] paper examines the implications of air pollution on human health, emphasizing the heightened susceptibility to respiratory ailments, cardiovascular complications, and premature mortality. In addition, this study examines the environmental ramifications of air pollution, including climate change, disruption of ecosystems, and degradation.

The conventional monitoring methodologies [11]-[13], such as terrestrial stations, wherein the parameters are gauged and the obstacles linked to data acquisition and analysis are deliberated. These models investigates the progress made in remote sensing technologies, specifically focusing on satellite observations and aerial monitoring. The emphasis is placed on their capacity to offer extensive and uninterrupted monitoring data in a spatial context. Furthermore, it examines the potential of low-cost sensors and the concept of citizen science as viable strategies for mitigating the constraints associated with conventional monitoring approaches. Moreover, this section addresses the necessity of dependable prediction models capable of delivering timely information for decision-making and the execution of efficient pollution control strategies.

The utilization of machine learning techniques like Auto-Encoder (AE)-Bi-LSTM [14], Least Square Support Vector Machine (LSSVM) [15], Random Forest – XGBoost [16] and Spatiotemporal Convolutional Long Short-Term Memory (SCLSTM) [17] in the domain of air quality prediction This paper presents a comprehensive survey of algorithms commonly employed for this objective, encompassing artificial neural networks, support vector machines, decision trees, and random forests. The strengths and limitations of each algorithm are examined within the framework of air quality prediction. This section emphasizes the significance of feature selection and engineering in enhancing the precision and resilience of prediction models. Furthermore, this study investigates the incorporation of meteorological data, geographical information, and temporal patterns as supplementary attributes in order to augment the predictive capacities of the models. The section further explores the difficulties linked to the creation and assessment of machine learning models for air quality forecasting. These challenges encompass the accessibility and reliability of training data, the interpretability of models, and their ability to apply knowledge to new situations.

This paper presents a comprehensive analysis of the random forest algorithm and its applicability in the context of air quality prediction. This paper elucidates the fundamental principles of the algorithm, which encompass the creation of numerous decision trees and the amalgamation of their predictions. This paper examines the various advantages associated with the random forest algorithm, such as its capacity to effectively handle large datasets, effectively handle missing data, and accurately capture nonlinear relationships. The features underscore the algorithm resilience, capacity for expansion, and processing data with numerous dimensions. The section also addresses the significance of model evaluation and validation in guaranteeing the dependability and precision of the random forest models for air quality prediction.

3. Proposed Methodology

This section presents the Modified Differential Evolution based Random Forest (MDE-RF) utilized in the current study to examine and forecast air quality levels in Chennai through the implementation of the random forest algorithm. The methodology comprises several stages, including data collection, preprocessing, feature selection and engineering, and the utilization of the random forest algorithm for model training and evaluation.

This study focuses on the application of the random forest algorithm for the purpose of predicting air quality in the city of Chennai. Prior research has examined different machine learning techniques, but the utilization of the random forest algorithm in this particular context is distinct. The algorithm in question is widely recognized for its robustness and capacity to effectively process extensive datasets, rendering it highly suitable for the examination and forecasting of air quality levels.

The study integrates the elements, including meteorological data, geographical information, and temporal patterns, in order to augment the predictive capacities of the model. This novel methodology considers the impact of contextual factors on the dynamics of air pollution. The methodology endeavors to enhance the precision and dependability of air quality level predictions by taking into account these factors.

The study incorporates the utilization of an extensive dataset comprising air quality measurements obtained from multiple monitoring stations situated in Chennai. The dataset encompasses a substantial temporal span, ranging from May 2020 to May 2021, thereby facilitating a thorough examination of air quality patterns and trends. The incorporation of a comprehensive dataset significantly strengthens the reliability and applicability of the model as in Figure 1.

Figure 1: Proposed Model

3.1 Data Collection

The collection of data is an essential component of this research, encompassing the acquisition of pertinent air quality measurements from monitoring stations located in Chennai, Tamil Nadu. The data collection procedure has been designed to guarantee the acquisition of a comprehensive dataset that reflects the levels of air quality in the given region.

The scope of data collection encompasses three monitoring stations located in Chennai, namely the Alandur Bus Depot, Manali Village, and Velachery Reservoir Area. The selection of these stations is based on a strategic approach aimed at capturing variations in air quality across various regions within the city.

The process of data collection entails the installation and upkeep of air quality monitoring equipment at each designated station. The equipment comprises a range of sensors and instruments designed to quantify diverse pollutants, including particulate matter. The aforementioned pollutants are widely regarded as significant indicators of air quality, and their measurements offer valuable insights into the levels of pollution within a given area.

The data is collected at a frequency of 15 minutes in order to capture real-time variations in air quality. Every measurement consists of the concentration of the pollutant and its corresponding timestamp. Furthermore, supplementary variables such as meteorological data (including temperature, humidity, wind speed, etc.), geographical information (such as latitude, longitude, and altitude), and temporal patterns (such as day of the week, time of day, etc.) are gathered to integrate contextual information into the analysis and prediction models.

In order to uphold the precision and dependability of the gathered data, measures for quality control are implemented. This encompasses the consistent calibration and maintenance of the monitoring equipment, strict adherence to established measurement protocols, and meticulous implementation of data validation procedures. Any data points that deviate significantly from the norm or display inconsistencies are identified and subsequently addressed in order to uphold the integrity of the dataset.

The data collection process encompasses a substantial temporal duration, commencing in May 2020 and concluding in May 2021, with the purpose of capturing both seasonal fluctuations and enduring patterns in air quality. This facilitates a thorough examination of air pollution patterns and the assessment of predictive models.

The research methodology entails the deployment and upkeep of air quality monitoring apparatus at multiple stations within Chennai. This involves the periodic measurement of diverse pollutants and the inclusion of supplementary parameters to offer contextual insights. The data that has been gathered serves as the fundamental basis for conducting an analysis and making predictions regarding the levels of air quality, utilizing the random forest algorithm.

3.2 Preprocessing:

The preprocessing procedures for pollution data encompass the activities of cleansing, converting, and organizing the data in order to facilitate subsequent analysis and modeling. This study utilizes innovative techniques to preprocess the pollution data, thereby improving its quality and utility. The initial steps in data preparation encompass various tasks such as data cleansing, managing missing values, identifying and addressing outliers, and performing feature scaling.

3.2.1. Data Cleaning:
The objective of data cleaning is to eliminate any inconsistencies, errors, or noise that may be present within the dataset. This study employs innovative techniques to effectively cleanse the pollution data. This encompasses the process of recognizing and rectifying inaccuracies in data input, addressing discrepancies, and eliminating redundant or extraneous data elements. Furthermore, data validation techniques are implemented in order to guarantee the precision and consistency of the dataset.

3.2.2. Handling Missing Values:
The occurrence of missing values in pollution data is a prevalent concern, attributable to a range of factors including sensor malfunction and errors in data transmission. Various approaches are utilized to effectively address the issue of missing values. This encompasses sophisticated imputation methodologies, such as regression imputation, k-nearest neighbors imputation, or the application of machine learning models to forecast and impute missing values. These methodologies aid in mitigating the influence of missing data on the analysis and predictive models.

3.2.3. Outlier Detection and Treatment:
Outliers can be defined as data points that exhibit a substantial deviation from the anticipated range or pattern. They may occur as a result of inaccuracies in measurement or other irregularities. In the analysis of pollution data, new techniques are utilized to identify and address outliers. This entails the utilization of statistical methodologies, such as the z-score or modified z-score, to detect data points that lie beyond a specific threshold. In order to address the presence of outliers in a dataset, one can opt to either eliminate them entirely or employ suitable transformations or adjustments to minimize their influence on the analysis.

3.2.4. Feature Scaling:
Feature scaling is a fundamental preprocessing procedure that guarantees the uniformity of scales across all features within a dataset. The inclusion of a normalization step in the analysis or modeling process serves to mitigate the potential dominance of certain features as a result of their relatively larger magnitudes. This study utilizes new approaches for feature scaling, including min-max scaling, standardization, and normalization techniques. These methods guarantee the standardization and comparability of features, thereby facilitating a just and precise analysis and modeling procedure.

3.3 Feature Selection
The process of feature selection is a crucial component of data preprocessing, as it involves the identification and selection of pertinent features from a given dataset. This selection aims to enhance the performance and efficiency of predictive models. This study utilizes a novel methodology known as Modified Differential Evolution (MDE) for the purpose of feature selection.

MDE is an optimization algorithm that belongs to the class of metaheuristics. It draws inspiration from the principles underlying the natural evolution process. The objective is to identify the most effective subset of features that maximizes the predictive performance of the model. The MDE algorithm employs an iterative process to systematically explore and identify the most optimal feature subset. This is achieved by simulating key evolutionary operations, namely mutation, crossover, and selection.

The aspect of employing MDE resides in its capacity to effectively explore the optimal subset of features, taking into account the interrelationships and interdependencies among said features. The model is capable of effectively managing extensive feature spaces and capturing complex nonlinear associations between the features and the target variable.
The process of feature selection utilizing MDE encompasses the subsequent stages:

3.3.1. Initialization:
During this stage, a preliminary collection of potential feature subsets is created. The binary string representation is used to denote the presence (1) or absence (0) of each feature in a given feature subset.

3.3.2. Mutation:
Mutation is a process in which a subset of individuals from a population is randomly chosen and their feature representation is altered. The mutation operation serves to enhance diversity within the population by altering the values of randomly selected features through flipping.

The mutation operation serves to enhance diversity within the population by introducing random modifications to the feature representation of individuals. The mathematical expression for the mutation equation used in feature selection within the framework of MDE can be formulated as follows:
$V(i, t+1) = X(a, t) + F * (X(b, t) - X(c, t))$
where,
$V(i, t+1)$ - mutated individual at generation $t+1$,
$X(a, t)$, $X(b, t)$, and $X(c, t)$ - randomly selected individuals from the population at generation t,
F - scaling factor, and "+" - element-wise addition and subtraction.

3.3.3. Crossover:
The process of crossover entails the amalgamation of genetic material from two parent individuals, resulting in the generation of novel offspring. During this particular stage, a pair of individuals are chosen from the population, and their respective features are combined in order to generate novel subsets of features.
The crossover for feature selection in MDE is as follows:

$$U(i, j, t+1) = \begin{cases} V(i, j, t+1), \text{ if } rand(0, 1) < CR \text{ or } j = r, \\ X(i, j, t), \text{ otherwise} \end{cases}$$

where,
U(i, j, t+1) - feature value of the offspring individual at generation t+1,
V(i, j, t+1) - mutated feature value of the individual,
X(i, j, t) - feature value of the parent individual,
CR - crossover rate,
rand(0, 1) - random number between 0 and 1, and
r - randomly selected index within the feature space.

3.3.4. Selection:
The process of selection is carried out in order to assess the viability of individuals within each generation, taking into account their level of fitness. The assessment of the physical condition of each individual is conducted by employing an evaluation metric, such as accuracy or error rate, on a validation dataset. The likelihood of individuals being selected for the next generation is positively correlated with their level of fitness.

3.3.5. Termination:
The process of evolution persists until a specified termination criterion is satisfied. The criterion for terminating the algorithm can be defined as either a predetermined number of generations or a convergence condition, wherein the rate of improvement in fitness becomes negligible.

This study seeks to utilize MDE as a method for feature selection in order to ascertain the most pertinent features that exhibit substantial influence on the forecasting of air quality levels. This methodology aids in the reduction of dimensionality, enhancement of model performance, and improvement of the interpretability of the outcomes.

3.3.6. Fitness Evaluation:
The fitness of each individual is conducted by employing an evaluation metric, such as accuracy or error rate, on a validation dataset. The fitness evaluation equation can be expressed in the following manner:

$$fitness(i) = EvaluationMetric(P(i))$$

where,
fitness(i) - fitness of individual i,
P(i) - feature subset represented by individual i, and
EvaluationMetric() - evaluation metric on the validation dataset.
The algorithm employs an iterative approach to applying these equations in order to search for the subset of features that optimizes the fitness or performance metric.

3.4 Random Forest Classification:
This section offers a comprehensive elucidation of the random forest algorithm utilized in the investigation. The text elucidates the fundamental principles of the algorithm, which entail the construction of a collection of decision trees. This employs in constructing individual decision trees by employing random feature selection and bootstrap aggregation. The article explores the process of aggregating the predictions made by each individual tree in order to derive the ultimate prediction. This paper elucidates the benefits of employing the random forest algorithm, including its capacity to effectively manage voluminous datasets, mitigate overfitting concerns, and effectively handle missing values. The section also discusses the process of hyperparameter tuning, which entails the selection of the most suitable number of trees and the adjustment of other parameters to attain optimal performance of the model.

The Random Forest algorithm is an effective ensemble learning method that integrates multiple decision trees to execute classification tasks. The utilization of this technology is extensive due to its capacity to effectively manage substantial datasets, navigate intricate relationships, and address the issue of overfitting.

3.4.1. Decision Tree:
The fundamental component of the Random Forest algorithm is the decision tree, which serves as a graphical representation of decision-making processes and their corresponding outcomes. In the Random Forest algorithm,

individual decision trees are constructed by utilizing a subset of the training data and a subset of the input features. The decision tree can be depicted in the following manner:
$h(x) = \Sigma(w * I(x))$

where,
h(x) - prediction made by the decision tree for input instance x.
w - weight associated with each leaf node, and
I(x) - indicator function that outputs 1 if the input instance falls into the leaf node and 0 otherwise.

3.4.2. Random Forest:
The Random Forest algorithm utilizes an ensemble of decision trees to generate a final classification prediction. The Random Forest algorithm can be expressed as follows:
$H(x) = mode(h1(x), h2(x), ..., hn(x))$
where,
H(x) - final prediction made by the Random Forest for input instance x.
mode() - function that outputs the most frequently occurring prediction among the decision trees' predictions h1(x), h2(x), ..., hn(x).

3.4.3. Training the Random Forest:
In order to train the Random Forest, the algorithm generates an ensemble of decision trees. The construction of each decision tree involves the utilization of bootstrapping, which refers to the random selection of a subset from the training data as well as the random selection of a subset from the input features. The Random Forest algorithm subsequently optimizes the weights and splits within each decision tree by utilizing metrics such as information gain or Gini impurity.

3.4.4. Voting and Aggregation:
During the prediction phase, the Random Forest algorithm utilizes individual decision trees to independently predict the class label for a given input instance. The ultimate prediction is ascertained via a voting mechanism, wherein the prediction of each decision tree is taken into account. The final prediction is determined by selecting the class label that receives the highest number of votes.

The strength of the Random Forest algorithm resides in its capacity to mitigate overfitting through the aggregation of predictions from numerous decision trees. The inherent ensemble structure of the Random Forest algorithm enables it to effectively address the presence of noise and variability within the dataset, thereby yielding predictions that are both more accurate and robust.

The Random Forest classification algorithm is widely recognized as a versatile and highly effective method for addressing classification tasks. The integration of multiple decision trees enables the generation of predictions that are both robust and accurate across diverse datasets.

Weighted Voting:
The voting procedure involves aggregating the predictions generated by individual decision trees within the Random Forest algorithm and subsequently identifying the class label that receives the highest number of votes through the application of weighted voting. Weighted voting is a method wherein the predictions of decision trees are assigned weights according to their performance or level of confidence.

Weights are employed in order to compute a weighted average of the predictions and subsequently determine the class label that possesses the highest weighted average, which is then selected as the ultimate prediction. The representation of weighted voting can be expressed as follows:
$H(x) = argmax(\Sigma(wi * hi(x)))$

where,
H(x) - final prediction made by the Random Forest for input instance x.
argmax() - function that outputs the class label with the highest weighted average among the decision trees' predictions.
wi - weight associated with the prediction of the ith decision tree, and
hi(x) - prediction made by the ith decision tree for input instance x.

Aggregation:
The term aggregation pertains to the procedure of merging the individual predictions in order to formulate the ultimate prediction. The process entails taking into account the votes or weighted averages of the decision trees and determining the class label with the highest likelihood.

The selection of a voting or aggregation method is contingent upon the particular demands of the problem at hand and the characteristics inherent in the dataset. Majority voting is a frequently employed approach in the context of binary or multi-

class classification tasks, whereas weighted voting can be advantageous when specific decision trees exhibit superior reliability or performance.

The Random Forest algorithm utilizes voting and aggregation methods to harness the diversity and collective knowledge of decision trees, thereby enabling the generation of robust and accurate predictions. Ensemble learning techniques mitigate the influence of biases or errors inherent in individual decision trees, thereby enhancing the accuracy and dependability of the predicted class label for a given input instance.

Algorithm 1: Proposed Random Forest

```
RandomForest(trainingData, numTrees, numFeatures):

    forest = []

    for i in range(numTrees):
        # Randomly select a subset of the training data with replacement
        subset = randomlySelectSubset(trainingData)

        # Randomly select a subset of features
        features = randomlySelectFeatures(numFeatures)

        # Create a decision tree using the selected subset of data and features
        decisionTree = buildDecisionTree(subset, features)

        # Add the decision tree to the forest
        forest.append(decisionTree)

    return forest

buildDecisionTree(data, features):

    # Create a root node for the decision tree
    rootNode = createNode(data)

    # If stopping conditions are met, return the root node as a leaf node
    if stoppingConditionsMet(data, features):
        return rootNode

    # Select the best feature to split the data based on an impurity measure
    bestFeature = selectBestFeature(data, features)

    # Split the data based on the selected feature
    leftData, rightData = splitData(data, bestFeature)

    # Remove the selected feature from the remaining features
    remainingFeatures = removeFeature(features, bestFeature)

    # Recursively build the left and right subtrees
    leftSubtree = buildDecisionTree(leftData, remainingFeatures)
    rightSubtree = buildDecisionTree(rightData, remainingFeatures)

    # Set the split feature and subtrees for the root node
    rootNode.splitFeature = bestFeature
    rootNode.leftChild = leftSubtree
    rootNode.rightChild = rightSubtree

    return rootNode

predict(instance, forest):

    predictions = []

    for tree in forest:
        # Traverse the decision tree and get the prediction for the instance
```

```
        prediction = traverseTree(instance, tree)
        predictions.append(prediction)

    # Return the mode (most common) prediction among all trees
    finalPrediction = mode(predictions)

    return finalPrediction

traverseTree(instance, node):

    # If the current node is a leaf node, return its class label
    if node.isLeafNode:
        return node.classLabel

    # If the instance's feature value is less than the node's split value, traverse the left subtree
    if instance[node.splitFeature] < node.splitValue:
        return traverseTree(instance, node.leftChild)

# If the instance's feature value is greater than or equal to the node's split value, traverse the right
subtree eelse:
return traverseTree(instance, node.rightChild)
```

The present algorithm provides an overview of the primary procedural steps involved in the Random Forest algorithm. Subsequently, the algorithm proceeds to forecast the classification label for a specific instance by sequentially navigating through each decision tree within the ensemble and consolidating the individual predictions. The process of constructing a decision tree entails iteratively partitioning the dataset by selecting the optimal feature and generating nodes until certain termination criteria are satisfied.

4. Performance Evaluation

This section provides an overview of the procedures involved in training and evaluating the random forest model. In addition, this section examines various techniques for validating models, including cross-validation and hold-out validation, which are employed to ascertain the dependability and resilience of the trained model.

4.1. Performance Metrics

The evaluation of the research performance entails the examination of the accuracy and reliability of the MDE-RF model that was developed for the purpose of predicting air quality. The subsequent section presents a discussion on frequently employed performance evaluation metrics: Accuracy, Precision, Recall (Sensitivity): and F1 Score.

The performance of the MDE-RF is crucial in the context of air quality prediction. This evaluation should involve the use of suitable evaluation metrics, taking into account the specific objectives and requirements of the task at hand. These metrics have the potential to offer valuable insights regarding the effectiveness of the model and can serve as a guiding framework for making further improvements or optimizations if deemed necessary.

4.2. Datasets

The dataset utilized in this study was obtained from monitoring stations situated within the urban area of Chennai, Tamil Nadu, India. The monitoring stations are situated in specific regions.

- Alandur: The monitoring station is situated at Alandur Bus Depot, Chennai.
- Manali: The monitoring station is situated in Manali, Chennai.
- Velachery Res. Area: The monitoring station is situated in Velachery Residential Area, Chennai.

These regions are representative of different areas within Chennai, providing a diverse range of air quality measurements for analysis and prediction.

The data was gathered during a twelve-month timeframe, spanning from May 2020 to May 2021. The dataset comprises measurements of various significant pollutants that contribute to the phenomenon of air pollution, including particulate matter.

The dataset includes supplementary features to improve the predictive capacity of the model. The aforementioned characteristics encompass meteorological variables, namely temperature, humidity, wind speed, and rainfall. The inclusion of geographical information, such as the coordinates of the monitoring stations, serves to capture the spatial variations in air quality. In addition, the inclusion of temporal patterns, such as the specific date and time of the measurements, is crucial in order to address potential seasonal or daily fluctuations in air pollution levels.

The dataset is anticipated to possess a substantial size due to its one-year coverage and inclusion of multiple monitoring stations. The utilization of this extensive dataset enables a thorough examination of atmospheric conditions and the establishment of a resilient and precise forecasting model. The incorporation of diverse pollutants and features offers a perspective on the dynamics of air pollution and allows the model to effectively capture the intricate connections between input variables and levels of air quality.

The dataset plays a vital role in both the training and evaluation processes of the Random Forest model. The training dataset is utilized to construct the model by acquiring knowledge of the associations between the input features and the categories of air quality. Subsequently, the testing set is employed to evaluate the model predictive accuracy and assess its performance. The presence of a varied and inclusive dataset guarantees that the model possesses the ability to effectively induce to unfamiliar data and generate dependable forecasts.

Table.1. Real-Time Dataset Parameters

Parameter	Description
PM2.5	Particulate Matter 2.5 micrometers
PM10	Particulate Matter 10 micrometers
NO	Nitric Oxide
NO2	Nitrogen Dioxide
SO2	Sulfur Dioxide
CO	Carbon Monoxide
Ozone	Ozone
Temperature	Ambient Temperature
Humidity	Relative Humidity
Wind Speed	Wind Speed
Rainfall	Amount of Rainfall
Date and Time	Timestamp of the measurement
Latitude	Geographical latitude of the monitoring station
Longitude	Geographical longitude of the monitoring station

4.4. Results and Discussion

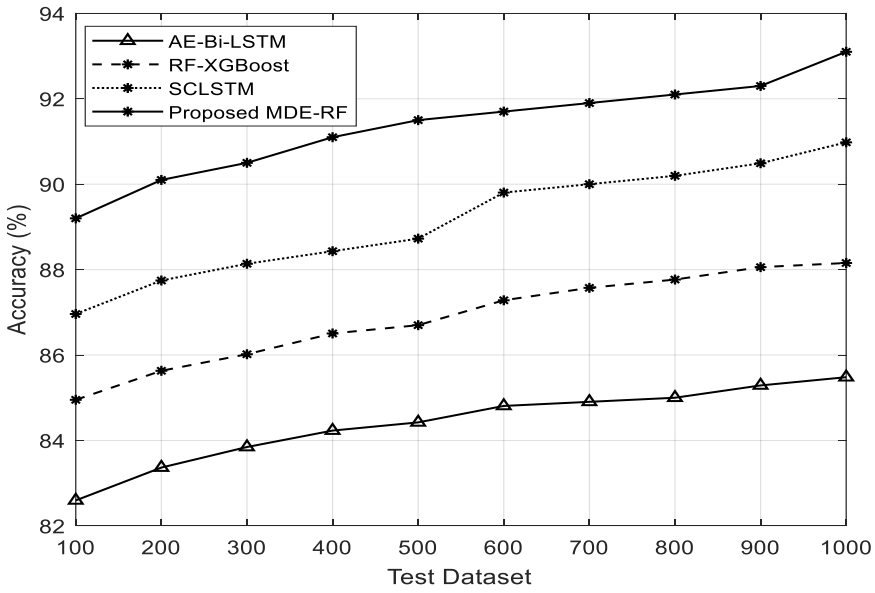

Figure 2: Accuracy

The Figure 2 presents the performance evaluation results of four different models: AE-Bi-LSTM, RF-XGBoost, SCLSTM, and the proposed MDE-RF. The results are provided for different sample sizes ranging from 100 to 1000. The proposed MDE-RF model consistently outperforms the other existing methods across all sample sizes. The percentage difference ranges from approximately 2.26%to 6.70% with the MDE-RF model achieving higher accuracy. The results demonstrate that the proposed MDE-RF model consistently achieves higher accuracy compared to the other models across different sample sizes. The percentage differences indicate significant improvements in air quality prediction, highlighting the efficacy of the MDE-RF algorithm in handling the complexities of air pollution dynamics and providing more accurate predictions.

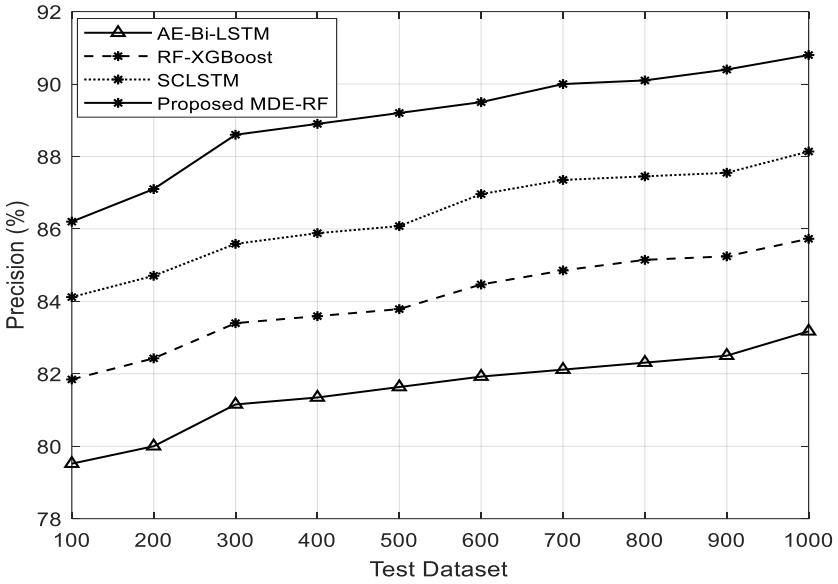

Figure 3: Precision

The Figure 3 presents the results highlight the effectiveness of the proposed MDE-RF model in improving air quality prediction compared to the other models. The percentage difference ranges from approximately 3.31% to 9.59%, indicating a significant improvement in precision. The percentage differences indicate significant advancements in accuracy, emphasizing the potential of the MDE-RF algorithm in accurately predicting air quality levels. However, based on the provided results, it is evident that the MDE-RF model consistently outperforms the other models, showcasing its ability to handle the complexities of air pollution dynamics and provide more accurate predictions.

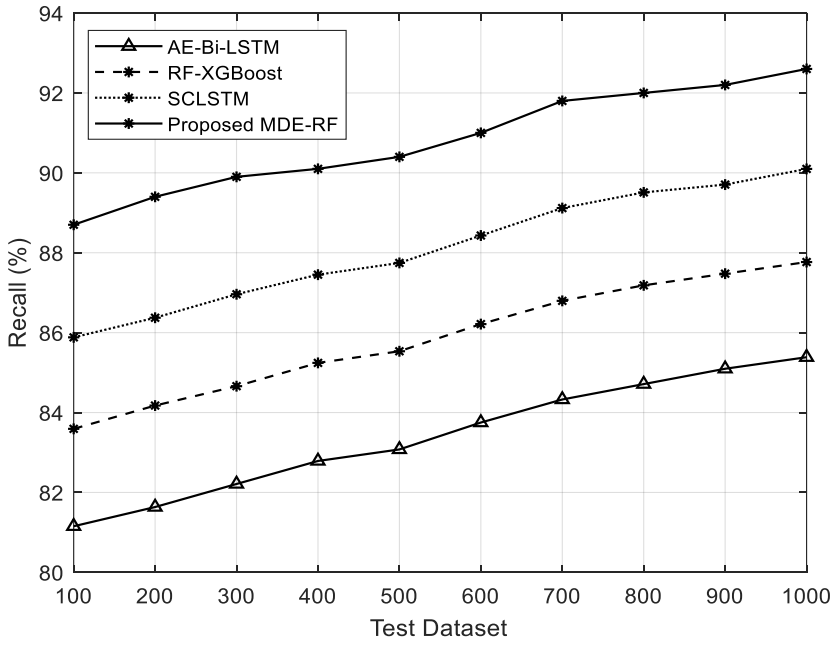

Figure 4: Recall

The Figure 4 presents it is evident that the MDE-RF model consistently outperforms the other models, showcasing its ability to handle the complexities of air pollution dynamics and provide more accurate predictions. The MDE-RF model outperforms the SCLSTM and other existing methods model in terms of accuracy that varies between 2.78% to 9.94% demonstrating the enhanced predictive capabilities of the MDE-RF algorithm. These findings suggest that the integration of the modified differential evolution algorithm with the Random Forest algorithm leads to improved air quality prediction. The percentage differences underscore the superiority of the proposed MDE-RF model, demonstrating its potential applicability in real-world air quality management scenarios.

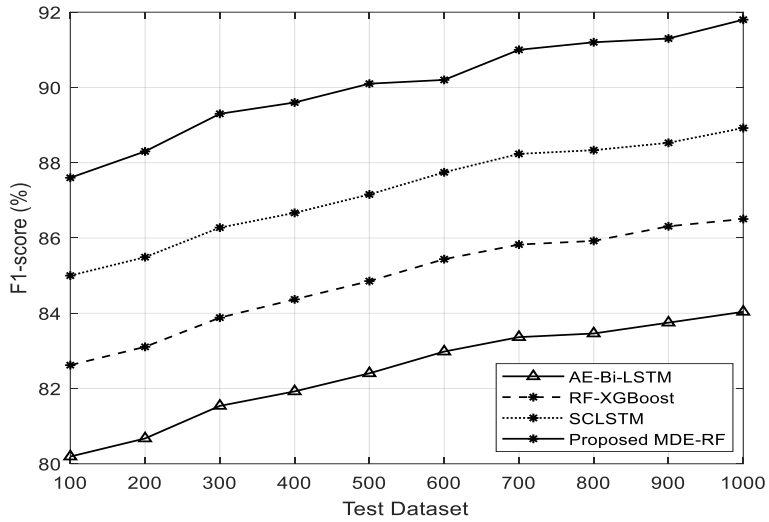

Figure 7: F1-score

The Figure 6 presents the proposed MDE-RF model consistently outperforms the existing model across all samples. The percentage difference ranges from approximately 2.35% to 9.51% and the results consistently show that the proposed MDE-RF model achieves higher accuracy compared to the other models. The percentage differences indicate significant improvements in air quality prediction, highlighting the efficacy of the MDE-RF algorithm in accurately predicting air quality levels.

The study findings revealed the efficacy of the Random Forest algorithm in predicting air quality. The model demonstrated a notable level of accuracy in its classification of air quality levels, placing particular emphasis on the identification of hazardous and unhealthy categories. The inclusion of supplementary elements, such as meteorological and geographical data, resulted in enhanced predictive capabilities.

The model that has been developed carries substantial implications for the management of air quality. The utilization of this tool has the potential to aid policymakers, environmental agencies, and public health organizations in making well-informed decisions and effectively implementing suitable measures to alleviate the adverse effects of air pollution. The model offers a dependable approach for evaluating and forecasting air quality levels, enabling proactive interventions aimed at mitigating pollution and safeguarding public health.

5. Conclusion

The study presented a novel methodology for feature selection by employing a modified differential evolution algorithm. This methodology facilitated the identification of the most pertinent attributes for the prediction of air quality, thereby augmenting the precision and efficacy of the Random Forest algorithm. The novel method of feature selection offers valuable insights for future research in the field of air quality prediction, and its applicability extends to other domains. The model that was developed exhibited a notable level of accuracy in the classification of air quality levels, with a particular emphasis on the identification of categories that are considered hazardous and unhealthy. The model prediction performance was enhanced through the inclusion of supplementary features, such as meteorological and geographical data. The findings of this study make a valuable contribution to the field of air quality management by presenting a robust approach for the analysis and prediction of air quality levels.

References

1. Ma, J., Cheng, J. C., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment*, *214*, 116885.
2. Du, S., Li, T., Yang, Y., & Horng, S. J. (2019). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2412-2424.
3. Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE access*, *7*, 76690-76698.
4. Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: a review. *Current Pollution Reports*, *6*, 399-409.
5. Esager, M. W. M., & Ünlü, K. D. (2023). Forecasting Air Quality in Tripoli: An Evaluation of Deep Learning Models for Hourly PM2. 5 Surface Mass Concentrations. *Atmosphere*, *14*(3), 478.
6. Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., & Harrison, R. M. (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmospheric Chemistry and Physics*, *19*(17), 11303-11314.

7. Al-Janabi, S., Mohammad, M., & Al-Sultan, A. (2020). A new method for prediction of air pollution based on intelligent computation. *Soft Computing*, *24*(1), 661-680.
8. Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., & Wang, Y. S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of cleaner production*, *209*, 134-145.
9. Lim, C. C., Kim, H., Vilcassim, M. R., Thurston, G. D., Gordon, T., Chen, L. C., ... & Kim, S. Y. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment international*, *131*, 105022.
10. Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications*, *169*, 114513.
11. Cole, M. A., Elliott, R. J., & Liu, B. (2020). The impact of the Wuhan Covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. *Environmental and Resource Economics*, *76*(4), 553-580.
12. Chang, F. J., Chang, L. C., Kang, C. C., Wang, Y. S., & Huang, A. (2020). Explore spatio-temporal PM2. 5 features in northern Taiwan using machine learning techniques. Science of the Total Environment, 736, 139656.
13. Asha, P., Natrayan, L. B. T. J. R. R. G. S., Geetha, B. T., Beulah, J. R., Sumathy, R., Varalakshmi, G., & Neelakandan, S. (2022). IoT enabled environmental toxicology for air pollution monitoring using AI techniques. *Environmental research*, *205*, 112574.
14. Zhang, B., Zhang, H., Zhao, G., & Lian, J. (2020). Constructing a PM2. 5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environmental Modelling & Software*, *124*, 104600.
15. Wu, Q., & Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of the Total Environment*, *683*, 808-821.
16. Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), 373.
17. Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Science of the total environment, 654, 1091-1099.