Air Quality Prediction Using Machine Learning

Ekampreet Kaur[1] **and Sandeep Singh Sandha**[2]

[1]**Dr. B.R.Ambedkar National Institute of Technology,Jalandhar**
[2]**Punjab AI Excellence**

ABSTRACT

This study presents a machine learning-based approach for forecasting air quality by predicting Air Quality Index (AQI) values and their corresponding health-related categories—'Good', 'Moderate', and 'Unhealthy'. Leveraging environmental and pollutant data, the system aims to provide early warnings about hazardous air conditions and associated health risks, thereby enabling proactive decision-making for individuals and authorities. Two distinct datasets were utilized for model development: the UCI Air Quality dataset and a city-specific dataset of Delhi sourced via Kaggle. A Random Forest Regressor was employed for Air Quality Index value prediction, while a TensorFlow-based multi-layer neural network was developed for air quality classification. The model focused on key atmospheric pollutants, including carbon monoxide (CO), nitrogen dioxide (NO_2), and benzene (C_6H_6), along with ambient temperature. The classification model achieved a test accuracy of 72% on the UCI dataset and 76% on the Delhi dataset, with balanced precision and recall across all classes. This study also addresses challenges such as class imbalance, dataset variability, and computational efficiency. Furthermore, the trained model is optimized for real-time deployment on edge devices using TensorFlow Lite, offering a scalable and practical solution for smart air quality monitoring systems. This research contributes to the development of AI-driven tools for environmental monitoring and public health protection.

Keywords: Air Quality,Machine Learning,Data Preprocessing,Random Forest Regression,Tensor flow Neural Networks,Pollution Forecasting, Environmental AI,$PM_{2.5}$,CO,NO_2

1.INTRODUCTION

Air pollution has become a global health threat. In cities, levels of $PM_{2.5}$, NO_2, and CO frequently exceed safe limits, contributing to serious health conditions such as asthma,bronchitis, lung cancer, pneumonia, and various others. According to research released by the Blacksmith Institute in 2008, indoor air pollution and urban air quality are two of the most pressing pollution challenges the world is facing today. In addition to endangering human health, these pollutants have a detrimental effect on ecosystems and the environment at large. The harmful impacts of air pollution range widely, from short-term issues such as respiratory system irritation to long-term issues such as lung cancer and heart disease. The pollutants listed above, combined with other factors, play a crucial role in the rapid and significant increase in respiratory diseases, cardiovascular diseases, and other chronic health problems. In addition, air pollution contributes to the thinning of ozone layer, which worsens the damaging effects of UV radiation and produces acid rain, which destroys ecosystems. Making use of the big data analytics and machine learning techniques, different ways to measure air quality have been possible in recent years due to technological and scientific breakthroughs. By using complex algorithms and large data sets to model and forecast air quality, these methods have the potential for increased precision and predictive power.

Unfortunately,current quality monitoring systems provide real-time data of prevalent conditions, but lack the capabilities to provide futuristic details that would be much better used by the public and authorities in advance. An effective air quality prediction system allows one to monitor existing pollutant data and identify pollutants that passively but deeply affect the environment and one's own health. Such a nicely fit model enables early mornings and proactive measures,allowing societies to minimize health risks, regulate emissions, and improve environmental outcomes.

For people with respiratory diseases such as asthma,allergic disorders, etc., having prior warnings about the prevalent aerial conditions helps them adopt preventive measures such as staying indoors,wearing

masks , or using other preventive equipment. For healthcare systems, anticipating pollution surges helps in planning resources and related events, preventing overcrowding in hospitals and public places, and offering better care to those affected. On a larger scale,governments can benefit from it by managing traffic, controlling industrial emissions, and releasing timely public health alerts.Overall,it leads to a better and cleaner environment.

As an individual who has experienced allergic disorders, especially those associated with dust particles in the air since childhood, this project has personal significance beyond its academic value. Having a first hand awareness of the health impacts of poor air quality has been a strong motivating factor in taking up this work. The challenges associated with environmental exposure have inspired the development of a system aimed at supporting individuals who face similar concerns. Through this research, the goal is to leverage the rapidly advanced Artificial Intelligence technology to promote safer and healthier living environments, with a specific focus on improving the quality of the air we breathe.

2.LITERATURE REVIEW

Air pollution is a critical environmental issue that affects both public health and urban infrastructure. Traditional air quality monitoring systems provide real-time data on pollutant levels but lack predictive capabilities that could allow proactive measures. This has led to increasing interest in the use of machine learning (ML) and artificial intelligence (AI) to develop models that can predict Air Quality Index based on historical pollutant data and meteorological conditions.

Some studies carried out by previous researchers in the same field have demonstrated the deployment of machine learning in Air Quality Index prediction.For example, Adhikari et al.(2025) explored Air Quality Index prediction using machine learning , facilitating the use of Random Forest and SVM by classifying air quality into respective categories.But the study neglected the class imbalance in the data set and implementation of deep learning architectures,which are more relevant for real-time predictions. Another work on outlier Framework detection by Dongre et al. combined linear regression with ensemble models. Their approach highlighted the impact on data quality on air quality prediction,but did not focus on the Air Quality Index classification or the feature selection aspect,crucial for model's improved generalization across different datasets.

So,this project was basically built keeping in mind the central idea of feature selection,tackling class imbalance, and optimizing model generalization while maintaining the overall model accuracy.

3.PROBLEM STATEMENT

Air quality prediction systems are essential for mitigating the adverse health and environmental effects of pollution, especially in urban regions. However, existing models often face significant limitations, including class imbalance in air quality categories, complex feature sets that hinder usability, and limited accessibility for public use. In addition, many models lack the ability to interpret in terms of understanding how individual pollutants contribute to overall air quality when other factors remain constant, which is critical for informed decision-making by government bodies and event organizers.

This project aims to develop a user-friendly and efficient AI-based air quality prediction system that addresses these challenges. By applying feature selection techniques, the model focuses on a minimal yet impaction set of pollutants (e.g., CO, NO_2, C_6,H_6, Temperature), ensuring reduced computational load while maintaining accuracy. To handle class imbalance, techniques such as SMOTE are employed, improving the model's ability to classify all AQI categories fairly. In addition, the model is designed to incorporate selection of features , allowing stakeholders to analyze the depth of effect of individual pollutants on air quality. This allows for data-driven planning, such as issuing health advisories, organizing outdoor events, or implementing traffic regulations. The model is also designed to be extensible, allowing future integration of more features and datasets for enhanced generalization and broader public application.

4.METHODOLOGY AND BACKGROUND

In this context,the proposed model, though basic in nature, still incorporates the classical machine learning and deep learning techniques to deliver reliable predictions even with limited data,thus ensuring a wider application range.

4.1 TOOLS AND LIBRARIES USED

The following tools and libraries were utilized for data preprocessing, model development, evaluation, and visualization:

(i)Python 3.x-The primary programming language used for data preprocessing, model development, evaluation and visualization.

(ii)NumPy – Used for numerical computations and efficient handling of arrays during data manipulation.

(iii)Pandas – Utilized for data cleaning, preprocessing, and manipulation of tabular air quality data.

(iv)Matplotlib and Seaborn – Used for data visualization, including pollutant distributions, correlation matrices, and model performance plots.

(v)Scikit-learn – Used for:

-Splitting the dataset into training and testing subsets.

-Implementing the Random Forest regression model.

-Evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), and R2 score.

(vi)TensorFlow/Keras – Applied for developing a neural network-based classifier to predict air quality categories ('Good', 'Moderate', 'Unhealthy') using a lightweight feedforward model.

(vii)UCI Air Quality data set and Kaggle Delhi data set – The dataset was sourced from publicly available repositories for reproducible experimentation.

(viii)Google Colab – The development environment used to write and execute the model, offering GPU acceleration and interactive data exploration.

4.2 STEP-WISE APPROACH

The proposed approach follows a systematic methodology that involves data collection from reputable public sources,followed by holistic preprocessing to ensure data quality and preserve its consistency.An exploratory data analysis (EDA) phase is conducted to identify key patterns, correlations, and trends, which guide the feature selection process. Based on the processed data, a set of predictive models—including both machine learning and deep learning algorithms—are developed and evaluated.A detailed description of all the steps involved is as follows:-

- **(i)Data Collection:** Data was sourced from multiple credible sources, including the UCI Air Quality Dataset and Kaggle relevant datasets. These datasets contain hourly measurements of prominent pollution parameters such as $PM_{2.5}$,NO,NO_2,CO,C_6H_6 and meteorological parameters such as temperature and humidity. Collecting data from varied and verified sources promises robustness of the data and improves the generalization of the model.

- **(ii)Data Preprocessing:** It was performed to ensure data consistency,accuracy and readiness for modeling. It included the following sub-steps:

- **Handling missing values:**The raw dataset consisted of several missing values,particularly in the pollutant readings.These were addressed using mean imputation method,and in case of columns with a large number of incomplete entries ,corresponding records were removed to maintain data integrity.

- **Timestamp processing:**The original dataset included separate date and time columns,which were combined into a single datetime column for the ease of analysis.This unified timestamp facilitated temporal tracking of pollution levels with more accuracy.As a result,the original separate date and time columns were eradicated from the dataset to discard duplicacy for the sake of better model performance.

- **Feature extraction from Timestamp:**From the combined datetime column,additional time-based informational features such as day,hour and month were extracted.These features were incorporated into the dataset to assist the model in capturing daily and seasonal variations in air quality.These

features were not utilised in the initial model building but are considered to be crucial to broaden the scope as well as scale of the model for future applications.

- **Feature Selection:**Feature selection was carried out to retain only the most relevant attributes affecting air quality. A correlation heatmap was used to analyze the relationships between various pollutants and the target variable (AQI or AQI Category). Features that showed low or redundant correlation were excluded from modeling. Additionally, domain knowledge was used to prioritize pollutants known to significantly impact air quality, such as CO, NO_2, and $PM_{2.5}$. This step helped reduce model complexity and improve performance.

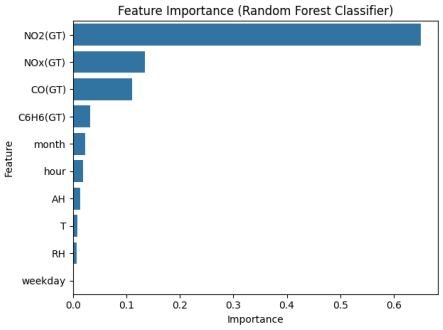

Figure 1. Feature selection for UCI dataset

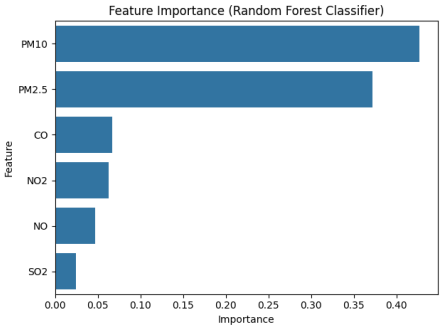

Figure 2. Feature selection for Delhi Dataset

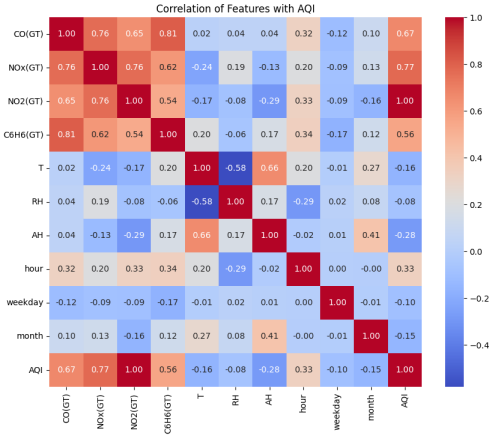

Figure 3. Correlation heatmap of UCI Dataset

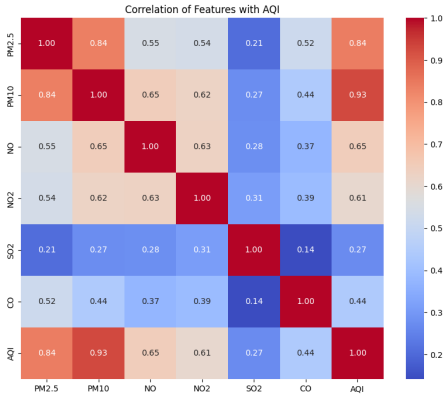

Figure 4. Correlation heatmap of Delhi dataset

(iii)Calculation of AQI from Standard formula: The Air Quality Index (AQI) is a standardized indicator used to convey the quality of air and its potential health effects. In this study, AQI was calculated using a sub-index-based approach, where pollutant-specific sub-indices were first computed for major air pollutants such as $PM_{2.5}$, PM_{10}, NO_2, CO, and SO_2 for the Delhi dataset and CO(GT) ,NO_2(GT) and C_6H_6 for the UCI dataset. The sub-index for each pollutant was calculated using the linear segmented formula defined by the Central Pollution Control Board (CPCB), which involves mapping the concentration of each pollutant to its corresponding AQI breakpoint range. The final AQI for a given time instance was determined as the maximum of all the individual sub-indices. This method ensures that the most critical pollutant at any time governs the overall AQI, thereby reflecting the most significant health risk.

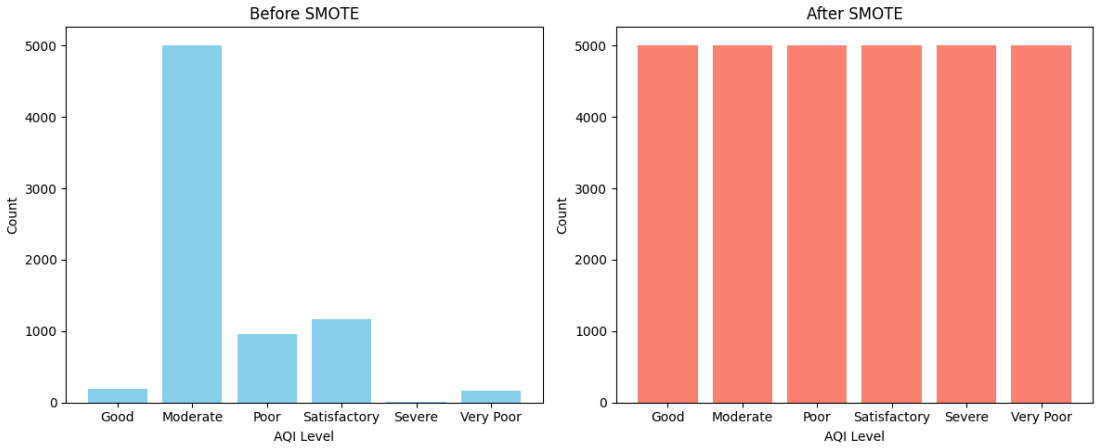

Figure 5. Tackling class imbalance for UCI dataset

To compute the sub-index for a pollutant, the following linear interpolation formula is used:

$$I = \left(\frac{I_{\text{Hi}} - I_{\text{Lo}}}{C_{\text{Hi}} - C_{\text{Lo}}} \right) \times (C - C_{\text{Lo}}) + I_{\text{Lo}}$$

where:
I = AQI sub-index C = Concentration of the pollutant $C_{\text{Lo}}, C_{\text{Hi}}$ = Breakpoints that surround C
$I_{\text{Lo}}, I_{\text{Hi}}$ = AQI values corresponding to those breakpoints

(iv)Handling class imbalance using SMOTE technique: The dataset initially exhibited a significant class imbalance across the six AQI categories — Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. Such imbalance can lead to biased model predictions, where the classifier favors majority classes and performs poorly on minority classes.To mitigate this, the **Synthetic Minority Oversampling Technique (SMOTE)** was applied exclusively to the training dataset. SMOTE generates synthetic samples for minority classes by selecting a sample and introducing new data points along the line segments connecting it to its k-nearest neighbors within the same class. This approach ensures that the minority classes are represented more equally without merely duplicating existing samples. By applying SMOTE, the dataset achieved a balanced class distribution across all six AQI categories, enabling the model to learn effectively from each class and improving its overall classification performance and contributing to increased generalization.This enabled the model to be more suited for real-world applications where impartial knowledge about every aspect concerned with the model is required.

- **(v)Model Selection:**Two basic forms of models,namely Random Forest and Tensorflow based Neural Networks were explored for the model building.The selection process involved evaluating each method's strengths, weaknesses, and suitability for handling the dataset's challenges such as class imbalance, multiple features, and the need for accurate classification across different AQI categories.

  (i)Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It works by averaging the results of several trees to reduce overfitting and improve generalization. In the initial experiments, the Random Forest model was used for regression-based AQI prediction due to its ability to handle non-linear relationships and mixed feature types. While it performed reasonably well, the classification results were uneven, with higher accuracy for certain AQI categories but poor generalization for others. This was primarily due to the imbalanced dataset and the model's tendency to overfit to the majority class.

- **For Regression**,the prediction of an input sample is given by:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

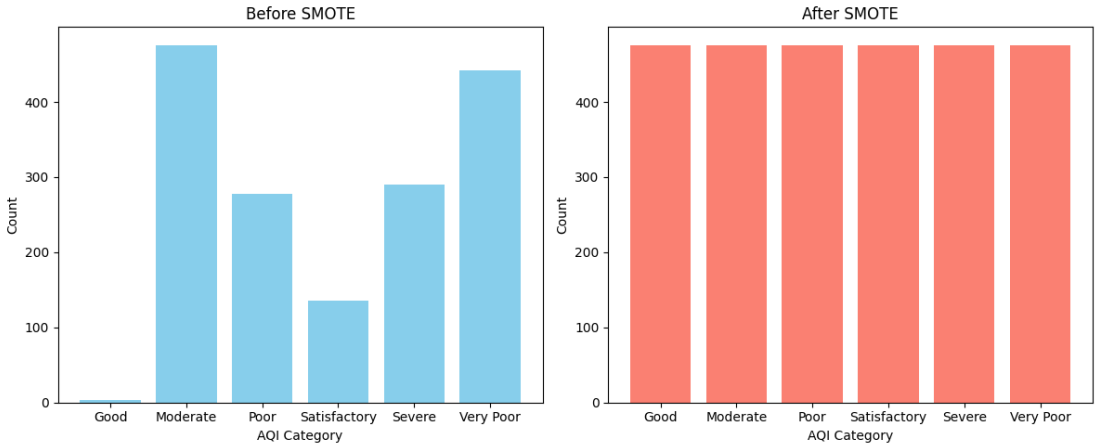

Figure 6. Tackling class imbalance for Delhi dataset

where:

- $\hat{y}$: final prediction of the Random Forest
- T : total number of trees in the forest
- $h_t(x)$: prediction from the t^{th} decision tree

- **Gini impurity** has been used for classification tasks

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

where:

- C : number of classes
- p_i : proportion of samples belonging to class i in the node

(ii)Tensor Flow based Neural Networks: To improve the performance of AQI classification, the initial Random Forest model—chosen for its robustness and ability to provide feature importance—was replaced with a TensorFlow-based deep learning model due to uneven accuracy across AQI categories and its limited capacity to capture complex non-linear relationships. The deep learning architecture consisted of multiple dense layers with ReLU activation, dropout layers to mitigate overfitting, and a Softmax output layer for multi-class probability distribution. Class imbalance was addressed using SMOTE, ensuring fair representation of all AQI categories, while SHAP-based feature selection retained only the most influential variables such as CO, NO_2, C_6H_6(GT), $PM_{2.5}$, and PM_{10}. This refined approach produced more consistent accuracy across classes and offered better scalability for integrating additional features or larger datasets in future applications.

Forward Propagation Equation:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$$

$$a^{(l)} = f(z^{(l)})$$

where W and b are the weights and biases, and f is the activation function (e.g., ReLU). **ReLU Activation Function:**

$$f(x) = \max(0, x)$$

Softmax for Multi-class Classification:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Categorical Cross-Entropy Loss:

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic})$$

where y_{ic} is the true label (one-hot encoded) and $\hat{y}_{ic}$ is the predicted probability.

(vi)Model Building: Initially,the Random Forest Regressor/Classifier from scikit-learn was used as the baseline model.

```
x=df[['T','C6H6(GT)']]
y=df['AQI']
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=20)
reg_model=RandomForestRegressor(n_estimators=100,random_state=42)
reg_model.fit(x_train,y_train)
y_pred=reg_model.predict(x_test)
# Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
# Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)
# Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)
# R² Score
r2 = r2_score(y_test, y_pred)
# Print results
print(f"Mean Absolute Error (MAE): {mae:.4f}")
print(f"Mean Squared Error (MSE): {mse:.4f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")
print(f"R² Score: {r2:.4f}")
```

Figure 7. Random Forest Regressor Model

```
def predict_aqi( temp,c6h6):
    input_data = pd.DataFrame([[ temp,c6h6]], columns=[ 'T','C6H6(GT)'])
    return reg_model.predict(input_data)[0]
#--- Classify AQI Levels ---
def classify_aqi(aqi):
  def classify_aqi(aqi):
    return ('Good' if aqi <= 50 else 'Satisfactory' if aqi <= 100 else
            'Moderate' if aqi <= 200 else 'Poor' if aqi <= 300 else
            'Very Poor' if aqi <= 400 else 'Severe')

#Giving the inputs to the model
temp_value = 25.0
c6h6_value = 9.2
# Predict AQI
predicted_aqi = predict_aqi(temp_value,c6h6_value)
# Classify AQI category
aqi_category = classify_aqi(predicted_aqi)
# Show result
print(f" ◆  Predicted AQI: {predicted_aqi:.2f}")
print(f" ◆  AQI Category: {aqi_category}")
```

Figure 8. Random Forest Classifier model

Although the model was provided satisfactory results for accuracy and precision,it lagged in generalizing the results.So,a Tensor Flow based Neural Networks model was used due to uneven accuracy across AQI categories and its limited capacity to capture complex non-linear relationships.

DATASET AND KNOWLEDGE SOURCES

The datasets that have been incorporated for model training and testing were derived from trusted and reliable sources such as:

1. **UCL Air Quality Dataset:** The data set consists of 9,358 hourly average measurements collected from an air quality chemical multisensor device deployed at road level in a polluted Italian city from March 2004 to February 2005. It contains responses from five metal oxide gas sensors along with ground truth concentrations for CO, NMHC, benzene (C_6H_6), NO_x, and NO_2, measured using certified analyzers. The data, recorded over a year, capture multivariate and time series features,

```
from tensorflow.keras.models import Sequential
from tensorflow.keras import layers
def create_model():
  model = Sequential([
    layers.Input(shape=(6,)),
    layers.Dense(units=64,activation='relu'),
    layers.Dropout(0.1),
    layers.Dense(units=32,activation='relu'),
    layers.Dropout(0.1),
    layers.Dense(units=16,activation='relu'),
    layers.Dropout(0.1),
    layers.Dense(units=6,activation='softmax'),
  ])
  model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
  return model
```

Figure 9. Tensor Flow deep learning model

with some missing values marked as -200. This dataset, which exhibits cross-sensitivity and sensor drift effects, is intended solely for research purposes and is widely used for air quality monitoring and prediction studies.

2. **Kaggle Delhi data set:**The air pollution data set for Delhi contains hourly air quality measurements from multiple monitoring stations in the National Capital Territory of Delhi, India. It records key pollutant parameters such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO and NO, providing a detailed temporal and spatial view of the city's air quality. The data set is used primarily for regression, forecasting, trend analysis, and modeling air quality indices to support public health planning and environmental policy decisions. It is available on Kaggle under standard dataset usage terms, generally allowing noncommercial use for research and analysis.

EXPECTED BENEFITS AND APPLICATIONS

The proposed model i has been built in the simplest way.But the underlying goal is to extend it to real-life applications so that the common people and government-related authorities can benefit from it and plan the course of future events accordingly.

- **Citizens** – Can use air quality forecasts to plan outdoor activities, avoid areas of high pollution, and reduce exposure to harmful pollutants.
- **Healthcare professionals** – can anticipate increases in respiratory or cardiovascular cases, allowing better resource allocation and timely interventions.
- **Government and municipal authorities** – Can issue early warnings, enforce short-term pollution control measures, and plan traffic diversions to minimize exposure.
- **Urban planners and smart city developers** – Can integrate the model into real-time monitoring dashboards, intelligent transportation systems, and zoning strategies.
- **Technology integration**– The model can be further developed into a web-based application or mobile tool, providing live AQI predictions, personalized alerts, and location-specific recommendations for the public.

Future Approach- In the future, the vision is to extend the model into an integrated, real-time decision support system by connecting it with on-ground sensor networks that continuously collect live atmospheric data such as temperature, humidity, pollutant concentrations, and wind patterns. The AI-ML model can then be retrained periodically using these live streams, supplemented with comprehensive historical datasets that merge environmental parameters with relevant socio-economic indicators.

For instance, datasets from agricultural markets (mandis) containing purchase, sale, and pricing information for vegetables and other produce can be combined with air quality and weather data. Such integration would enable the system to not only forecast pollution levels but also predict their potential impacts on crop sales, market demand, and distribution logistics.

This extension would require:

- **Deployment of IoT-based air quality sensors** in strategic urban and rural locations to ensure accurate and localized data collection.
- **Data integration pipelines** to merge atmospheric readings with external datasets (e.g. mandi transactions, weather forecasts, transportation data).
- **Advanced model architectures** (such as hybrid deep learning models) capable of handling multi source, heterogeneous data streams.
- **User-facing platforms**(mobile applications or dashboards) that provide real-time AQI, market trend predictions, and actionable recommendations for farmers, traders, and policy-makers.

  Such a system could empower agricultural stakeholders with data-driven insights, improve market efficiency, and strengthen the link between environmental monitoring and economic decision making.

EXPERIMENTAL RESULTS

The experimental results obtained from two machine learning approaches—Random Forest and a TensorFlow-based neural network—applied to two distinct datasets: the UCI Air Quality dataset and a region-specific Delhi air quality dataset. For each model–dataset combination, we evaluate performance using the confusion matrix, classification report, and a comparison of training and testing accuracies. The results are organized sequentially, beginning with the Random Forest model on the UCI dataset, followed by the TensorFlow model on the same dataset, and subsequently repeated for the Delhi dataset.

- **Visualizing Predictions:**The plots for the comparison of actual and predicted AQI values for the Random Forest Regressor model and that for the training and testing accuracies of the Tensor Flow Neural networks are shown at the top.

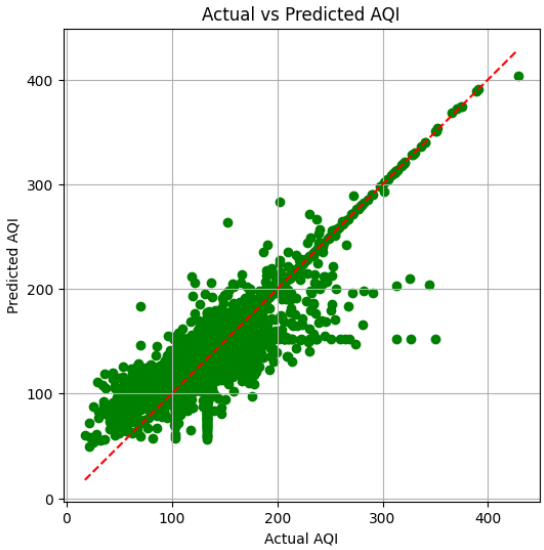

Figure 10. Actual v/s Predicted AQI values

- **UCI data set:**
- Training accuracy remains high while validation accuracy fluctuates heavily — suggests possible overfitting or small validation set variance.
- Validation loss stays erratic and higher than training loss, indicating unstable generalization.
- **Kaggle Delhi data set:**

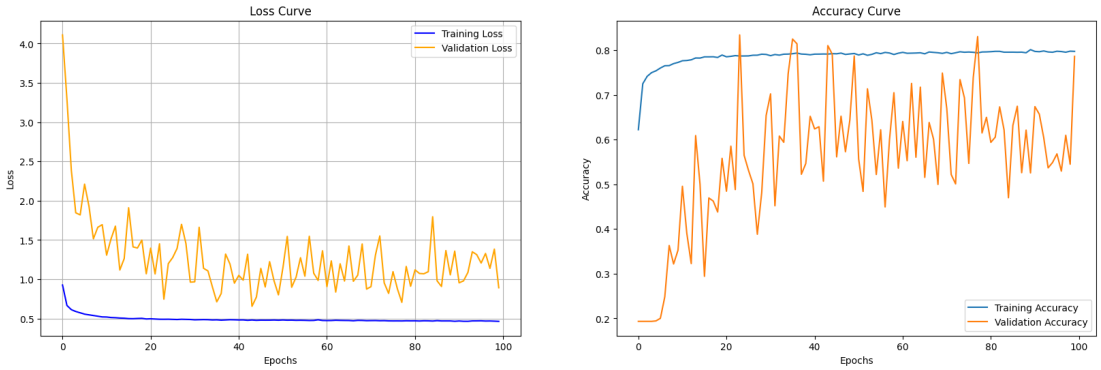

Figure 11. Training v/s Testing losses and accuracies for Tensor Flow deep learning model(UCI data set)

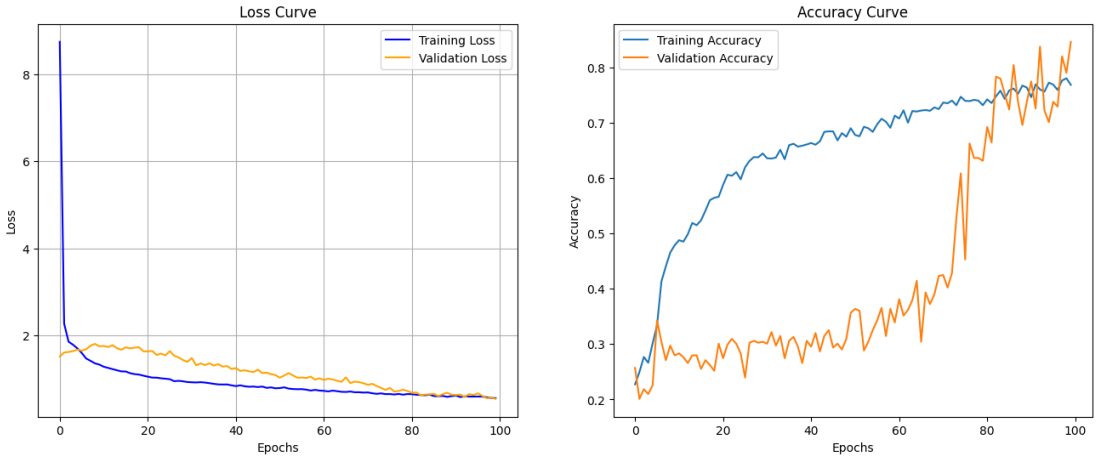

Figure 12. Training v/s Testing losses and accuracies for Tensor Flow deep learning model(Delhi data set)

- Validation accuracy improves steadily after initial fluctuations, eventually matching training accuracy — indicates better generalization over time.
- Loss curves converge, showing the model adapts well with prolonged training.
- **Evaluation Metrics:**The performance of Random Forest model was analyzed by evaluating it through standard parameters as listed in the table below.

Metric	Numerical value
Mean Absolute Error(MAE)	21.4715
Mean Squared Error(MSE)	899.5884
Root Mean Squared Error(RMSE)	29.9931
R^2 Score	0.7295

Table 1. Performance of Random Forest on validation dataset.

- **Confusion Matrix:**The confusion matrix provides a tabular representation of the model's predictions compared to the actual class labels, enabling a clear view of correctly and incorrectly classified instances.
  Conclusion: The classification results varied across datasets and models.
- **UCI Dataset – Random Forest**
  (i)High accuracy for Moderate and Poor categories.
  (ii)Frequent misclassifications between Good and Satisfactory; Severe and Very Poor underrepresented and rarely predicted correctly.
- **Delhi Dataset – TensorFlow**
  (i)Strong performance on Moderate, Poor, Satisfactory, and Very Poor categories.
  (ii)Struggled with Good and Severe due to low representation and class imbalance.
- **UCI Dataset – TensorFlow**
  (i)Predictions heavily biased toward Moderate.
  (ii)Misclassifications often between Moderate and Satisfactory, indicating overlapping feature boundaries.
- **Classification report:**The classification performance is summarized in the first figures using a heatmap of the classification report, which presents the precision, recall, and F1-score for each AQI category along with macro and accuracy scores. To facilitate a clearer visual comparison of these

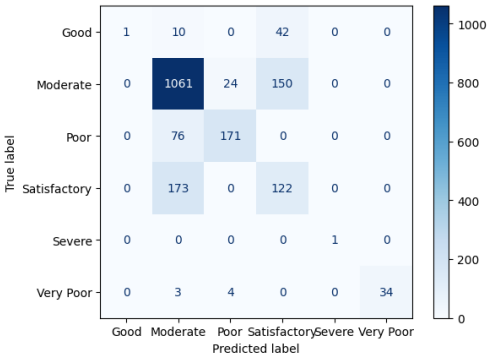

Figure 13. Confusion Matrix for Random Forest Model(UCI data set)

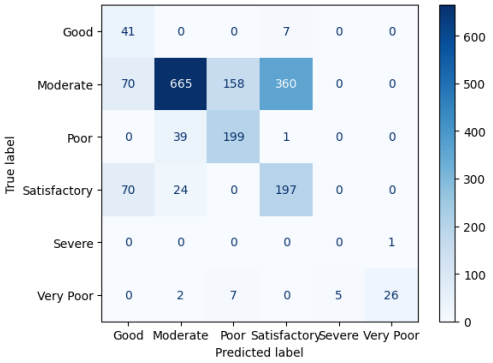

Figure 14. Confusion Matrix for Tensor Flow Deep learning model(UCI data set)

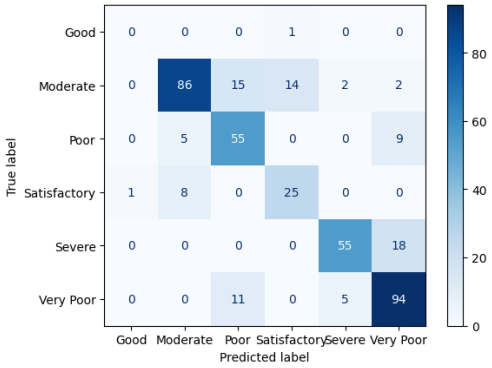

Figure 15. Confusion Matrix for Tensor Flow Deep learning model(Kaggle Delhi data set)

metrics across categories, figures plotted later depict the same data in a grouped bar chart format.

Conclusion:

(i)Random Forest (UCI dataset):

(i)Achieves perfect precision and recall for "Severe" and high scores for "Moderate" and "Very Poor," but performs poorly for "Good" (very low recall and F1-score).

(ii)Overall accuracy is 74%, with balanced macro-average scores (0.64–0.66) indicating moderate performance across classes.

(ii)TensorFlow Neural Network (UCI dataset):

(i)Shows stronger performance for "Satisfactory" and "Severe," but lower precision and recall for "Good" and "Poor."

(ii)Overall accuracy is slightly lower (63%), with macro-average precision, recall, and F1-scores below 0.6, showing greater variability in class-wise predictions.

(iii)TensorFlow Neural Network (Delhi dataset):

(i)Performs consistently well across most classes, especially "Very Poor" and "Moderate," though fails completely for "Good" (precision and recall = 0).

(ii)Achieves 78% overall accuracy and balanced macro-average (0.64–0.66), showing more stable multi-class predictions than the UCI TensorFlow model.

RELATED WORK

- **IIT Delhi – Real-time Air Quality Monitoring with ML:** IIT Delhi developed a real-time AQI prediction system focused on Delhi's urban environment using Random Forest and Gradient Boosting algorithms. The model was trained on CPCB air quality data combined with meteorological parameters. It was deployed with IoT sensors placed around the campus, enabling live AQI updates every 10 minutes. This architecture demonstrated high accuracy ($R^2 \approx 0.92$) and validated the

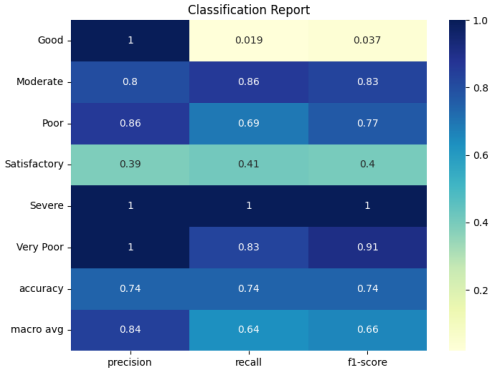

Figure 16. Classification report(heatmap) for Random Forest(UCI dataset)

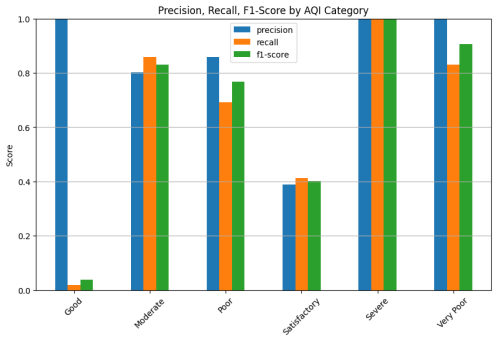

Figure 17. Classification report(bar graph) for Random Forest(UCI data set)

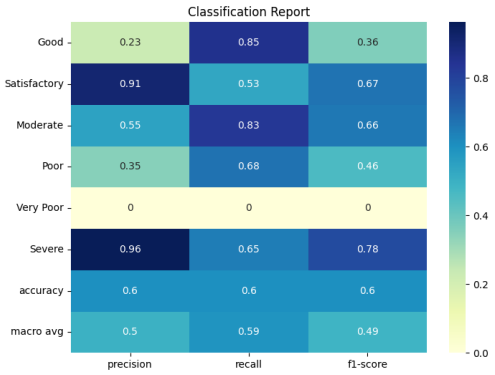

Figure 18. Classification report(heatmap) for Tensor Flow Neural networks(UCI data set)

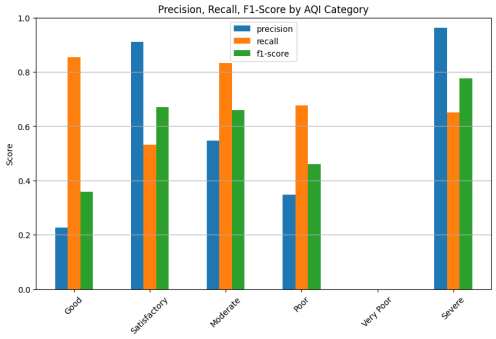

Figure 19. Classification report(bar graph) for Tensor Flow Neural networks(UCI data set)

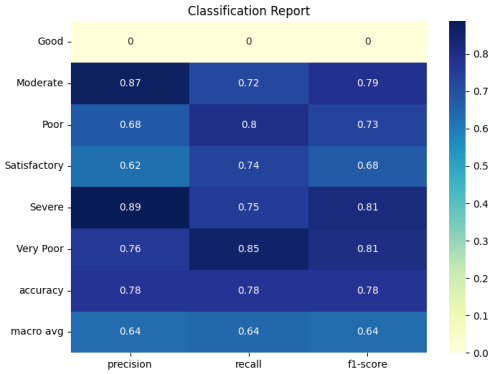

Figure 20. Classification report(heatmap) for Tensor Flow Neural networks(Delhi data set)

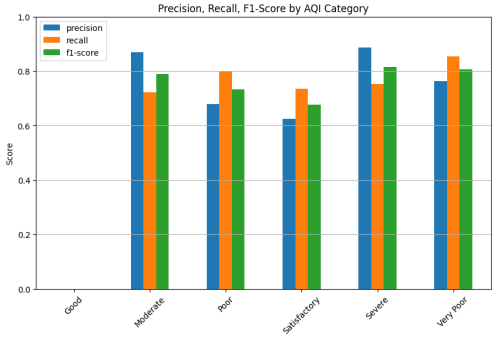

Figure 21. Classification report(bar graph) for Tensor Flow Neural Networks (Delhi data set)

feasibility of integrating ML with low-cost sensing hardware.

- **BreezoMeter API – Commercial AQI Forecasting Platform:** BreezoMeter offers a commercial API providing hyperlocal AQI predictions worldwide. It combines government station data, satellite imagery, meteorological forecasts, and proprietary sensor networks. The system uses gradient boosting and spatio-temporal interpolation to achieve a fine spatial resolution of 1 km². Its real-time updates and health advisories are integrated into popular weather apps and air quality dashboards used by millions globally.

- **OpenAQ + University of Chicago – Global Air Quality Data Platform:** OpenAQ, in collaboration with the University of Chicago, operates a large-scale open-source platform for aggregating real-time and historical air quality data from multiple sources worldwide. Its architecture uses a scalable AWS-based ingestion pipeline, allowing researchers and developers to build predictive models. The platform has been used in academic projects to train ML models for AQI forecasting, making it a vital bridge between raw environmental data and applied machine learning research.

CONCLUSION

The primary objective of this project was to design and implement an accurate and efficient Air Quality Index (AQI) prediction system, capable of performing well on both standardized datasets and real-world, location-specific data. Beginning with clear predefined goals—such as achieving high prediction accuracy, ensuring adaptability to varying data sources, and addressing dataset imbalance—the workflow was systematically executed. The developed framework incorporated essential technical features including a robust preprocessing pipeline (missing value imputation, outlier handling, and normalization), feature selection for pollutants like CO(GT) and NO_2(GT), simplified AQI sub-index calculation, and comparative evaluation between a Random Forest regression model and a TensorFlow-based neural network. Evaluation metrics such as R^2 score, RMSE, accuracy, and confusion matrix analysis were employed for comprehensive assessment.

The models successfully met the preset objectives, with the Random Forest model achieving an accuracy of 75% on the test set, thereby delivering strong baseline performance. For the TensorFlow-based neural network, performance varied across datasets:

UCI dataset — Training Accuracy: 76.52%, Testing Accuracy: 62.02%

Delhi dataset — Training Accuracy: 85.12%, Testing Accuracy: 77.59%

These results highlight the adaptability of the neural network model to diverse datasets, while confirming the Random Forest's reliability as a robust baseline.

Key challenges—such as incomplete data, variability in pollutant concentration ranges, and the need for scalable feature handling—were effectively addressed through careful preprocessing and hyperparameter tuning. The results confirmed that the system can perform robust AQI predictions and adapt to datasets of different scales and characteristics without significant performance degradation.

Looking ahead, the future scope involves extending the model for real-time applications by integrating it with live IoT sensor feeds, enriching datasets with meteorological parameters, and incorporating advanced architectures such as LSTM for temporal trend forecasting. Furthermore, cross-domain integration with datasets like agricultural market trends could enable broader applications beyond environmental monitoring. These enhancements will be carried out via incremental pipeline development, sensor-network integration, and iterative retraining, ensuring the system evolves into a fully operational, scalable, and impactful predictive platform.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my mentor,Dr.Sandeep Singh Sandha, for their invaluable guidance, constructive feedback, and constant encouragement throughout the course of this project. I am also thankful to the team of Punjab AI Excellence for providing the necessary resources, technical support, and academic environment that facilitated the successful completion of this work. Additionally, I acknowledge the contributions of open-source data providers, including the UCI Machine Learning Repository and Kaggel-hosted Delhi Air Quality datasets, which were essential for model development and evaluation.

REFERENCES

1.)Brostow, G. J., Fauqueur, J.,& Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2), 88–97.

2.)Janai, J., Güney, F., Behl, A.,& Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. Foundations and Trends® in Computer Graphics and Vision.

3.)Ronneberger, O., Fischer, P.,& Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI).

4.)UCI dataset:https://archive.ics.uci.edu/dataset/360/air+quality
5.)Kaggle Delhi dataset:https://www.kaggle.com/datasets/hrithikpm/air-pollution-dataset