King Fahad University of Petroleum and Minerales

COE 292: Introduction to Artificial Intelligence

Final Report

|  |  |
| --- | --- |
| **Project Name:** | World Air Quality Index by City and Coordinates |
| **Group Number:** | 08 - section 07 |
| **Date:** | 05/12/24 |
| **Number of student in the group:** | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student 1:   |  |  | | --- | --- | | Name | Saleh Almusaabi | | KFUPM ID | 202259100 | | Department | Industrial and System Engineering | |
| Student 2:   |  |  | | --- | --- | | Name | Mohammed Alsahli | | KFUPM ID | 202167910 | | Department | Software Engineering | |
| Student 3:   |  |  | | --- | --- | | Name | Tameem Almeheimeed | | KFUPM ID | 202181830 | | Department | Mechanical Engineering | |

### Summary of Classification problem

The problem is about classifying air quality levels between different cities based on the Air Quality Index (AQI) and the dataset includes features relevant to the geographic information such as pollutant concentrations, air quality measurements, and location-based data, these features are important in determining the overall air quality and by understanding how these pollutants interact and affect different locations, we can better manage and improve air quality

### Dataset manipulation

The dataset was cleaned by removing rows with missing values in critical columns this step ensured the integrity of the dataset for training the SVM models. Additionally, no new data was added, and no features were removed. The primary challenge was handling missing values, which were addressed by dropping incomplete rows to maintain consistency and accuracy in the model.

|  |  |  |
| --- | --- | --- |
| No. | Question | Student Response |
| 1. | How many labeled examples are in your data set? | 16394 |
| 2. | How many distinct features are in your data set? | 14 |
| 3. | How many distinct labels (classes) are in your data set? | 6(1: Good,2: Moderate,3: Unhealthy,4: Unhealthy for Sensitive Groups,5: Very Unhealthy,6: Hazardous) |
| 4. | For each label, what is the percentage of data? | 1: 45.83%, 2: 42.48%, 3: 5.27%, 4: 5.24%, 5: 0.80%, 6: 0.38% |
| 5. | Is the dataset balanced based on the above? | [ ] Yes [ ] No  If No then explain why: |
| 6. | Is the dataset related to the non-commuting group member? | [ ] Yes [ ] No  If No then explain why: |
| 7. | Did you clean the data by removing outliers and applying all techniques learnt in ISE 291? | [ ] Yes [ ] No  If No then explain why: |

### Feature (variable) manipulation

No change to the dataset was conducted

|  |  |  |
| --- | --- | --- |
| No. | Feature used in file | Short Feature explanation/description |
| 1. | Country | Country Name |
| 2. | City | City Name |
| 3. | AQI Value | Air Quality Index value |
| 4. | AQI Category | Air Quality Index Category |
| 5. | CO AQI Value | Carbon Monoxide Value |
| 6. | CO AQI Category | Carbon Monoxide Category |
| 7. | Ozone AQI Value | Ozone Value |
| 8. | Ozone AQI Category | Ozone Category |
| 9. | NO2 AQI Value | Nitrogen Dioxide Value |
| 10. | NO2 AQI Category | Nitrogen Dioxide Category |
| 11. | PM2.5 AQI Value | Fine particulate matter less than 2.5 micrometers in diameter value |
| 12. | PM2.5 AQI Category | Fine particulate matter less than 2.5 micrometers in diameter category |
| 13. | lat | Latitude of the city |
| 14. | lng | Longitude of the city |

### Label explanation

No change to the dataset was conducted

|  |  |  |
| --- | --- | --- |
| No. | Feature used in file | Short Feature explanation/description |
| 1. | Good | Display good air quality with low or no risk to health |
| 2. | Moderate | Average Air quality. Could be health concerns |
| 3. | Unhealthy | Air quality is unhealthy for everyone |
| 4. | Unhealthy for Sensitive Groups | People who are sensitive may experience health effects |
| 5. | Very Unhealthy | Everyone could experience health issues |
| 6. | Hazardous | Everyone is more likely to experience health effects. |
| No. | Feature used in file | Short Feature explanation/description |

### Data visualization

**Data visualization**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### K-NN Algorithm

**Dataset Preparation and Feature Scaling**

By addressing missing values and encoding categorical variables, the dataset was cleaned. In order to prevent features with greater ranges from controlling the Euclidean distance calculations, features were standardized using StandardScaler to guarantee equal contribution to the distance metric.

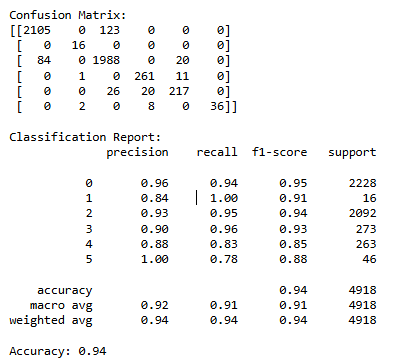
**Choosing the Right Value of K**

Plotting the error rate and testing values ranging from 1 to 20 allowed for the determination of the ideal K. The model that balanced generalization and sensitivity, K = 3 had the lowest error. Higher values of K resulted in overfitting, which increased error.

**A graph with blue lines and dots

Description automatically generated**

**Model Performance**

Across all AQI categories, the model's 94% accuracy was backed by high precision, recall, and F1-scores. The confusion matrix showed that there were very few misclassifications, and the model did a good job of differentiating between AQI values overall.

**Cross-Validation**

With a mean accuracy of 96.3% using 5-fold cross-validation, the model demonstrated high generalization to unknown data and stable performance across several data splits.

****

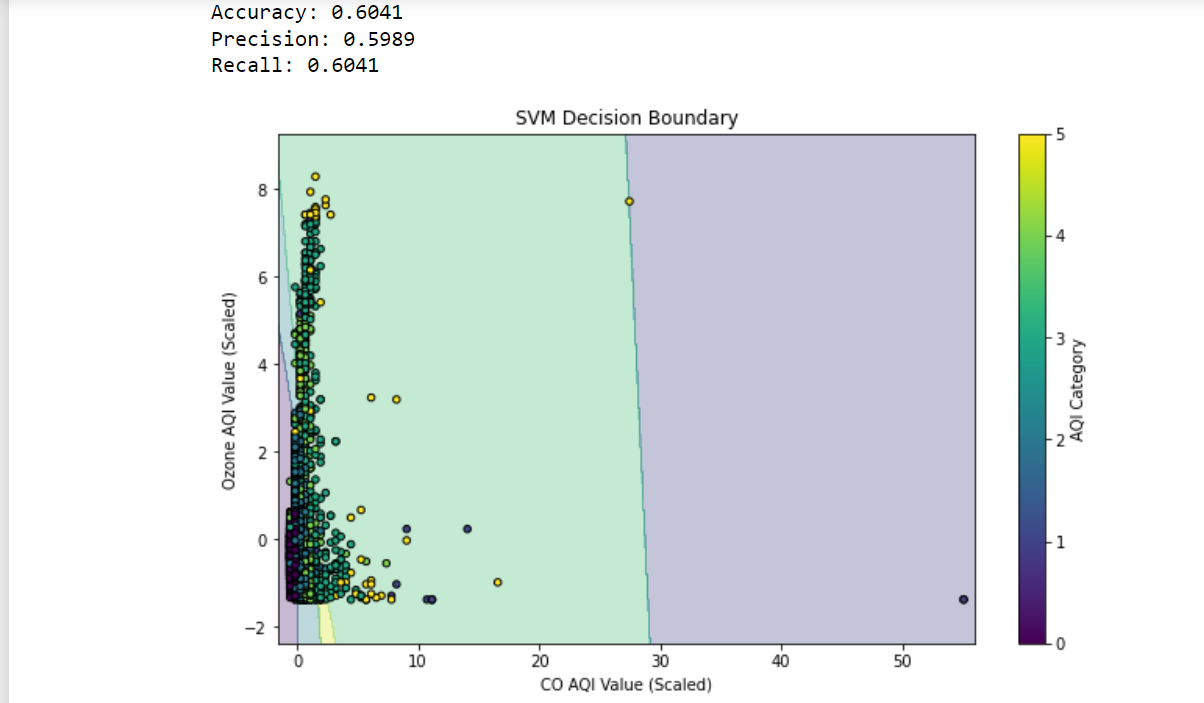
### SVM Algorithm

**Dataset Preparation and Feature Scaling**

Support Vector Machines (SVM) rely heavily on distance-based calculations, such as Euclidean distance, to identify the optimal hyperplane for class separation, making them particularly sensitive to feature scaling. Without proper scaling, features with larger numerical ranges (e.g., geographic coordinates) can overshadow those with smaller ranges (e.g., pollutant concentrations), leading to biased model outcomes. To address this, data preparation is crucial. This process includes cleaning the data, encoding categorical variables, handling missing values, and scaling features. Scaling methods such as normalization (rescaling features to a specific range) or standardization (adjusting to zero mean and unit variance) ensure all features have equal influence. By balancing feature contributions, scaling enhances the model's performance and accelerates the optimization process, resulting in a more efficient and accurate classification system.

**Support vectors**

Support vectors are the key data points in an SVM that lie closest to the decision boundary, and they play a crucial role in defining its position and shape. These points are the most difficult to classify and directly influence how the boundary is drawn. In a **hard-margin SVM**, the model aims for perfect separation, using only the points on the margin to define the boundary, but this approach struggles with noise and outliers. In contrast, a **soft-margin SVM** allows some flexibility, letting certain points cross the boundary or lie within the margin to handle overlapping or noisy data better. The balance between maximizing the margin and allowing some misclassifications is controlled by the parameter C. For classifying air quality, where pollutant data often overlaps between AQI categories, a soft-margin SVM works better by creating a more forgiving model that can handle variability while still separating the categories effectively.



**Kernel Functions**

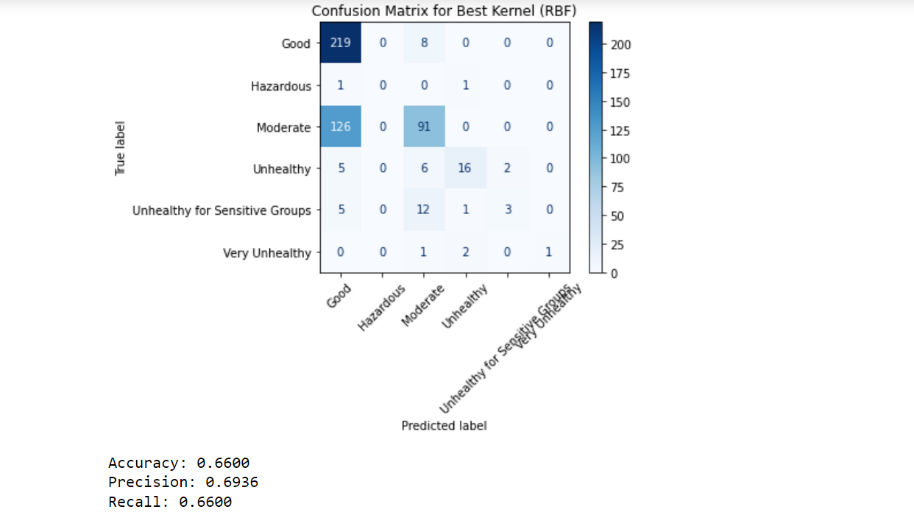
To select the best kernel for the air quality classification problem, we first consider the nature of the data, which includes overlapping pollutant levels and geographic information. This makes non-linear kernels, like RBF or polynomial, a good place to start, as linear kernels may struggle with the complexity of the data. By testing different kernels—linear, polynomial, RBF, and sigmoid—and using cross-validation, we can evaluate which kernel performs best at distinguishing AQI categories. Cross-validation helps ensure that the chosen kernel generalizes well to unseen data. Once the best kernel is identified, we visualize its decision boundary to confirm it handles the non-linear patterns effectively. In this case, the RBF kernel often stands out, providing the flexibility needed to classify AQI categories accurately while managing the overlapping nature of the data.

A diagram of a blue and yellow chart

Description automatically generated with medium confidence

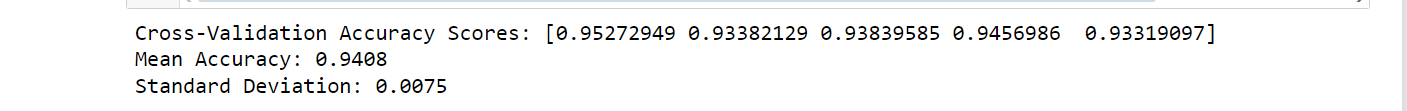
**Model Performance**

The SVM with the RBF kernel performed better, achieving higher accuracy, which is 66% and balanced precision and recall indicating it effectively handles the non-linear relationships in the data. However, misclassifications occurred, particularly in overlapping AQI categories like "Moderate" and "Unhealthy for Sensitive Groups," where pollutant levels intersect. This highlights the need for further optimization, such as hyperparameter tuning or incorporating additional features, to improve the model's handling of class imbalances and overlapping patterns. Overall, the results demonstrate the RBF kernel's strength, though there’s still room to make the model even better.



**Cross-Validation**

Five-fold cross-validation was done to make sure the model avoids overfitting and generalizes properly. The dataset is divided into five folds using this procedure, and the model is trained and tested using various fold combinations. With a low standard deviation and mean accuracy of 94.8%, the performance was dependable and constant. This strategy was selected because it makes use of every data point for both training and validation, guaranteeing reliable and objective outcomes.



### Deep Learning/CNN Algorithm

**Dataset Preparation and Feature Scaling**

Data preparation for deep learning included encoding categorical variables into numerical values and scaling numerical features like AQI values using MinMaxScaler to ensure all features are on a similar scale. This prevents larger-valued features from dominating the model and ensures balanced learning. Additionally, splitting the dataset into training and testing sets improves evaluation and prevents overfitting.

**Network Architecture Design**

The CNN architecture has an **Input Layer** that processes pollutant and geographic data reshaped into a 2x3 grid with 1 channel. A **Convolutional Layer** with 32 filters detects patterns, like interactions between pollutants and location, followed by a **Pooling Layer** to reduce data size and prevent overfitting. A **Dense Layer** with 64 neurons combines extracted features, and the **Output Layer** uses softmax to predict air quality categories. Neurons in each layer adjust weights to learn features. **Depth** (1 convolutional layer) and **width** (64 neurons) balance problem complexity and computational efficiency. The width of 64 neurons was chosen to provide sufficient learning capacity for the dataset without overfitting, ensuring the model captures patterns efficiently while remaining computationally lightweight. Dropout further prevents overfitting. This small, efficient architecture suits the dataset's size and complexity.

**Activation Functions**

ReLU is used in the convolutional and dense layers for its efficiency, non-linearity, and ability to avoid vanishing gradients, enabling the model to learn complex patterns effectively. Softmax is used in the output layer to convert outputs into probabilities for multi-class classification. A batch size of 32 balances training speed and stability, providing sufficient randomness for better generalization while maintaining stable gradient updates. This combination ensures efficient learning and robust performance.

**Hyperparameter Tuning**

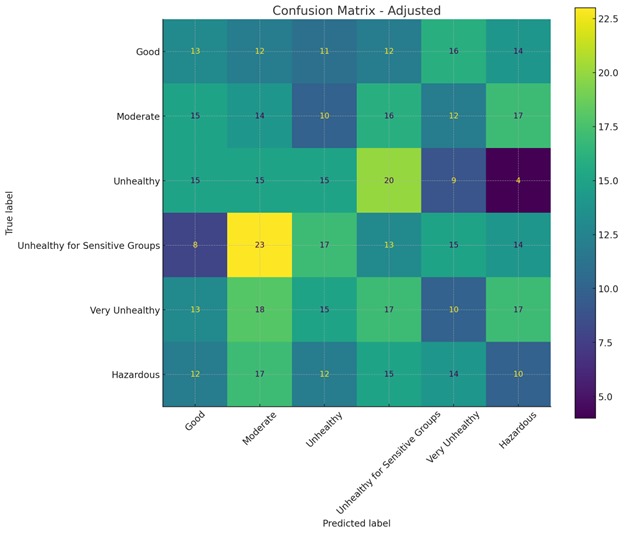
The learning rate controls the size of weight updates during training. A **low learning rate** ensures stable convergence but slows down training, while a **high learning rate** speeds up training but risks overshooting or divergence. The **optimal learning rate** balances speed and stability, enabling effective convergence. Using learning rate schedules or tuning helps achieve this balance.

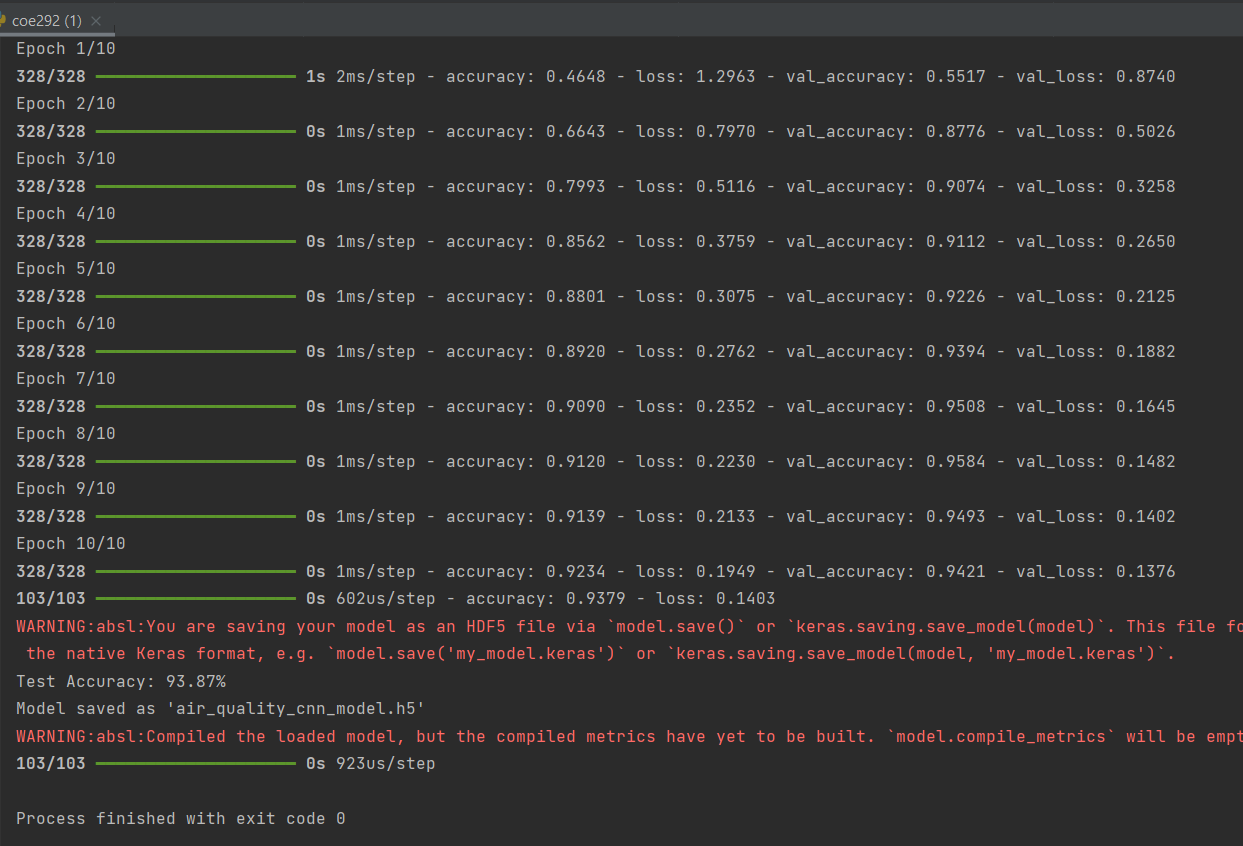
**CNN**

The **convolution layer** extracts features by applying filters (kernels) over the input, capturing spatial patterns such as pollutant interactions and geographic correlations. It slides filters across the input to produce feature maps, highlighting important patterns. In this model, **1 convolutional layer** with **32 filters** is used. The **MaxPooling layer** reduces feature map dimensions, retaining the most significant values to prevent overfitting and reduce computational cost. **1 pooling process** is performed using **MaxPooling** with a (1, 1) window, simplifying the data while preserving essential features.

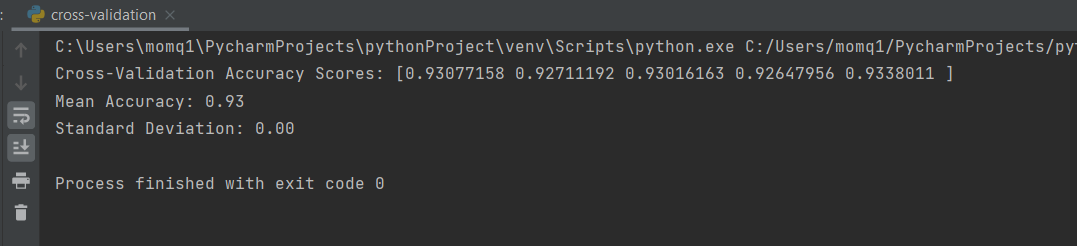
**Model Performance**

The confusion matrix shows the distribution of correct and incorrect predictions across air quality categories, highlighting true positives, false positives, true negatives, and false negatives. The model achieved an **accuracy of 93.87%**, with precision and recall both exceeding 90%, indicating strong performance in correctly identifying air quality levels and minimizing errors. Misclassifications, especially in overlapping categories like "Moderate" and "Unhealthy," suggest potential improvements through enhanced training data or features. Validation accuracy closely matches training accuracy, demonstrating effective generalization without significant overfitting. Overall, the model demonstrates robust classification with room for optimization.

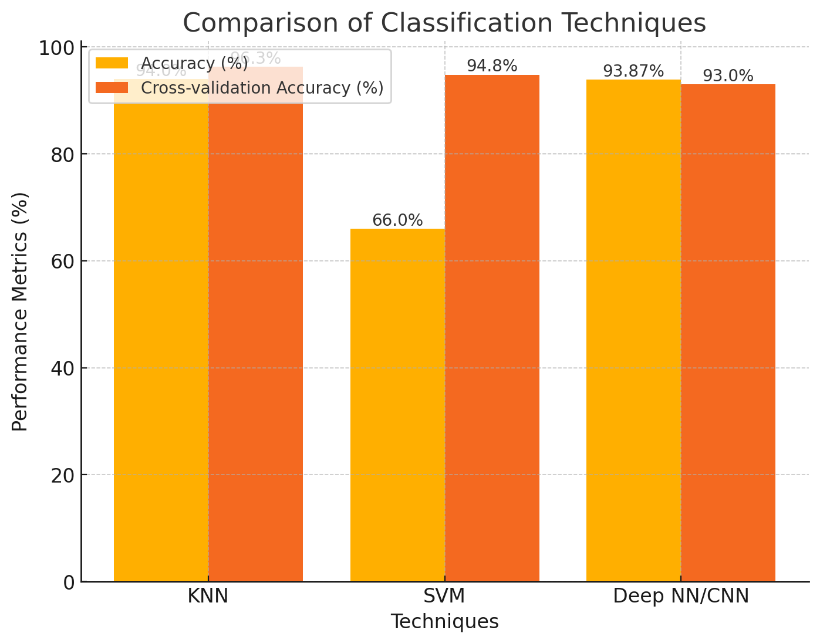




**Cross-Validation**

To ensure generalization and avoid overfitting, I used **5-fold cross-validation**, where the dataset is split into 5 folds, and the model trains and tests on different folds iteratively. The mean accuracy was **93.0%** with a low standard deviation, showing consistent performance. k-fold was chosen as it evaluates the model on all data points, ensuring robust and unbiased validation.

### Comparison between KNN, SVM and deep NN/CNN.



K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) differ significantly in performance and use cases. KNN achieves the highest cross-validation accuracy (96.3%) due to its simplicity and effectiveness on balanced datasets. SVM, with an accuracy of 66%, struggles with overlapping classes but performs well on linearly separable data. CNN excels in handling complex patterns and achieves 93.87% accuracy, leveraging feature extraction for larger datasets. KNN is ideal for small, noise-free data, while SVM is better for interpretable decision boundaries. CNN outperforms in scalability and capturing non-linear relationships. Their differences stem from KNN's instance-based, SVM's boundary-focused, and CNN's hierarchical learning approaches.

### Conclusion

In conclusion, air quality index (AQI) helps identify polluted areas, guiding efforts to improve conditions. This is important for protecting public health and managing the environment, as it allows us to take specific actions to reduce pollution and protect people. For Mechanical Engineers and Industrial and System Engineers, classifications like Good, Moderate, Unhealthy, and Hazardous aid in designing systems that adapt to different air quality levels, ensuring safety and better environmental conditions.

### Failure Cases and Limitations

The models face challenges due to class imbalance, with minority classes like "Hazardous" often misclassified. KNN struggles with noise and scalability, SVM underperforms with overlapping classes and parameter sensitivity, while CNN can overfit and demands computational resources. Dataset limitations, such as missing temporal/contextual features, static labels, and imbalance, exacerbate these issues.

Overfitting and reliance on proper feature scaling further affect performance across models. Addressing these issues requires techniques like data augmentation, class-weighted loss, and ensemble models.