Analyzing Health Risk Based on Air Pollution using Advanced Machine Learning Algorithms

Asif Imtiaz Chowdhury
Department of CSE
BRAC University
Dhaka, Bangladesh

Israt Jahan Hira
Department of CSE
BRAC University
Dhaka, Bangladesh

Sadman Taufiq Mahin
Department of CSE
BRAC University
Dhaka, Bangladesh

I. Abstract

Air pollution significantly impacts public health, with severe consequences ranging from respiratory diseases to cardiovascular illnesses. This study aims to predict health risks based on air quality using machine learning (ML) methodologies. The dataset, containing 5,811 records, includes air quality metrics (e.g., AQI, PM2.5, PM10, NO2, SO2, O3), meteorological data (temperature, humidity, wind speed), and health indicators (respiratory cases, cardiovascular cases, hospital admissions). The target variable categorizes health impact into five classes based on a Health Impact Score. We implemented Decision Tree, Random Forest, AdaBoost, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models. AdaBoost achieved the highest accuracy of 95%, demonstrating its efficacy in handling complex, non-linear relationships. This study contributes to the growing field of environmental health by leveraging ML for proactive health risk mitigation.

II. Introduction

Air pollution is one of the most pressing environmental health challenges, contributing to millions of deaths annually. Fine particulate matter (PM2.5), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ground-level ozone (O3) are among the primary pollutants causing respiratory and cardiovascular diseases. According to the World Health Organization (WHO), nearly 90% of the global population breathes air that exceeds safe pollution levels, leading to severe health outcomes and economic losses.

Traditional methods for assessing air quality and health risks rely on statistical models, which often fail to capture the complex, non-linear relationships between pollutants, meteorological factors, and health outcomes. Advances in machine learning (ML) offer new opportunities to enhance predictive accuracy and provide actionable insights. ML methods, such as Random Forest, Support Vector Machines (SVM), and neural networks, have been successfully applied to air quality forecasting, pollutant concentration prediction, and health risk classification.

This study explores the use of ML techniques to predict health risks based on air quality metrics. The dataset includes comprehensive environmental and health data, enabling the development of robust models to classify health impact into five severity levels. By comparing multiple ML algorithms, we aim to identify the most effective approach for predicting health risks and informing public health interventions.

III. Literature Review

Air Pollution and Health Impact Numerous studies have documented the adverse health effects of air pollution. Chen et al. (2014) demonstrated a significant correlation between the Air Quality Health Index (AQHI) and emergency department visits for ischemic stroke, particularly among the elderly【

7†source】. Similarly, Sicard et al. (2010) highlighted the increased risk of respiratory and cardiovascular diseases due to urban pollution in the Provence-Alpes-Côte d'Azur region【8†source】. These findings underscore the need for accurate health risk prediction models that incorporate air quality metrics.

Machine Learning in Air Quality Prediction ML methodologies have shown great promise in environmental science. Liang et al. (2020) employed Random Forest and AdaBoost to predict AQI levels, achieving high accuracy across multiple regions in Taiwan【9†source】. Castelli et al. (2020) utilized Support Vector Regression (SVR) with a radial basis function kernel to forecast pollutant concentrations in California, achieving a classification accuracy of 94.1% for AQI categories【9†source】. Such studies demonstrate the potential of ML in capturing the temporal and spatial variability of air pollutants.

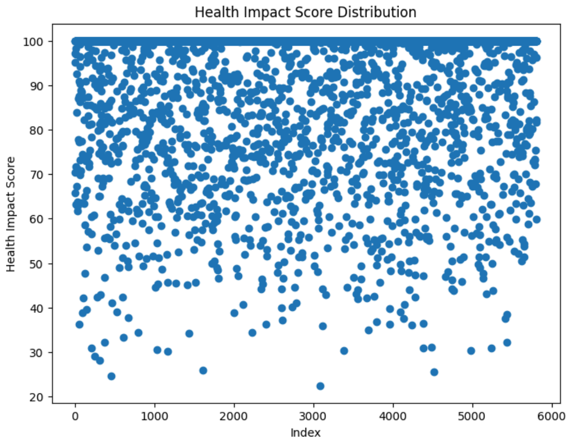
Deep Learning for Health Risk Prediction Advanced deep learning architectures, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have been applied to air pollution forecasting. Bekkar et al. (2021) proposed a hybrid CNN-LSTM model for PM2.5 prediction in Beijing, outperforming traditional ML methods【9†source】. Similarly, Maxwell et al. (2017) demonstrated the efficacy of deep neural networks (DNNs) in multi-label classification of chronic diseases, highlighting their ability to extract complex patterns from medical data 【9†source】.

Challenges and Opportunities Despite their success, ML models face challenges such as data quality, model interpretability, and generalizability. Goldstein et al. (2016) emphasized the importance of addressing biases in electronic health record (EHR) data, including missing values and loss to follow-up【9†source】. Ensuring the robustness and fairness of ML models is critical for their adoption in public health.

Building on existing research, this study integrates air quality metrics, meteorological data, and health indicators to develop ML models for health risk prediction. By comparing the performance of Decision Tree, Random Forest, AdaBoost, KNN, and SVM, we aim to advance the field of environmental health and provide tools for proactive risk management.

IV. Dataset Preprocessing and Model Description:

Data preprocessing is a crucial step in the data analysis pipeline that involves cleaning, transforming, and organizing raw data into a format suitable for analysis and modeling. It includes tasks such as handling missing values, removing duplicates, standardizing data formats, encoding categorical variables, and scaling numerical features. Data preprocessing is important because it ensures the quality and reliability of the data used for analysis, which directly impacts the accuracy and effectiveness of machine learning models. By preparing the data properly, data preprocessing helps improve the model's performance, reducing bias, and enhancing the interpretability of the results. It also aids in identifying patterns, relationships, and insights that can drive informed decision-making and actionable outcomes.



The dataset we used for our project was already preprocessed. All the features used numerical values and there were no duplicates or null fields. As such, further preprocessing was not required for us to be able to apply our models on the dataset.

We generated the following machine learning models to predict the impact certain factors on our health by taking a subset of features present in our dataset:

A. Decision Tree (DT)

A decision tree is a supervised learning algorithm that uses a tree-like structure to split data into branches based on feature thresholds. Each node represents a decision, and leaf nodes indicate outcomes. It’s simple, interpretable, and effective for classification and regression but prone to overfitting without pruning.

B. **Random Forest (RF)**

Random Forest is an ensemble learning algorithm combining multiple decision trees to improve accuracy and robustness. Each tree is trained on random subsets of data and features, and predictions are aggregated through majority voting (classification) or averaging (regression). It reduces overfitting, increases stability, and works well on diverse datasets.

C. **Adaptive Boost (AdaBoost)**

AdaBoost is a boosting algorithm that combines weak learners, typically decision stumps, to create a strong classifier. It iteratively adjusts sample weights, emphasizing misclassified examples in each round. This approach improves overall performance but can be sensitive to noisy data and requires careful parameter tuning for optimal results.

D. **K-Nearest Neighbors (KNN)**

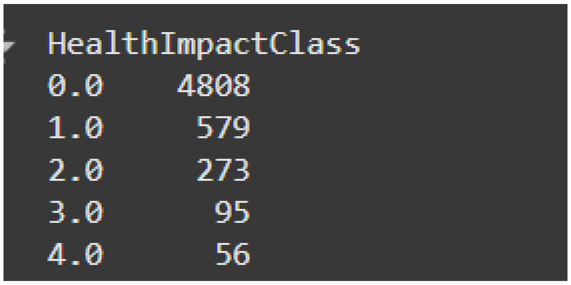
KNN is a simple, non-parametric algorithm used for classification and regression. It assigns a class or predicts a value based on the majority class or average of the kkk-nearest points in feature space. Effective for low-dimensional data, it’s computationally expensive with large datasets and sensitive to feature scaling.

E. **Support Vector Machines (SVM)**

SVM is a supervised learning algorithm that finds the optimal hyperplane to separate data classes with maximum margin. It works well in high-dimensional spaces and can handle non-linear boundaries using kernel functions. SVM is effective for small to medium-sized datasets but computationally intensive and sensitive to hyperparameter selection.

V. RESULTS AND DISCUSSION

Our objective was to predict the health impact class based on the air pollution data using KNN, SVM,AdaBoost, Decision Tree and Random Forest. The dataset included various air quality factors like AQI, PM10, PM2.5, NO2, SO2, O3 etc. Moreover we can observe an uneven distribution of the results.



```
HealthImpactClass
0.0    4808
1.0     579
2.0     273
3.0      95
4.0      56
```

Fig.2.Health Class Distribution(a).

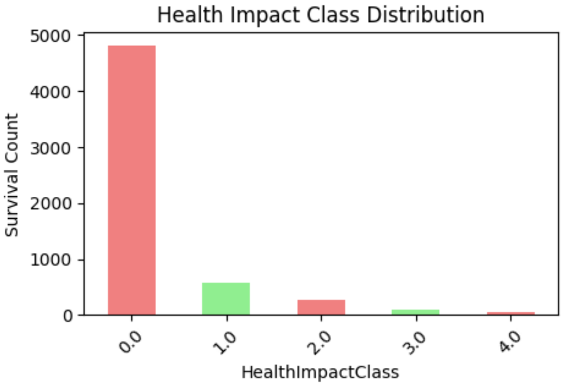


Fig.3.Health Class Distribution(b)

Accuracy Comparison

If we observe the accuracy of the models used, we see that among the models, AdaBoost demonstrated the highest accuracy of all the other machine learning models used here scoring around 95%.However, Random Forest and Decision Tree perform around the same criterion. Closely followed, with the accuracy of 94.92% and 94.66% respectively. Nonetheless, KNN, while being relatively simple, attained an accuracy of 92%,which was lower than others but still notable.

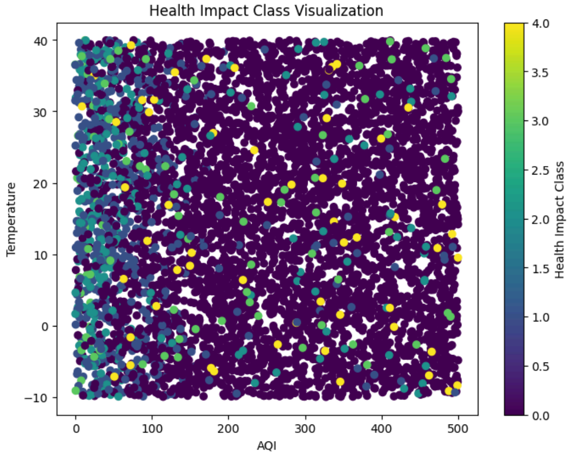


Fig.4.ScatterPlot for KNN

Feature Contribution

Among other features AQI, PM10, PM2.5 and NO2 played a significant part in determining the health impact class as they are strongly correlated with respiratory and cardiovascular health issues.On The other hand, the environmental factors like temperature, humidity and wind speed combined helped to make better prediction for the models.

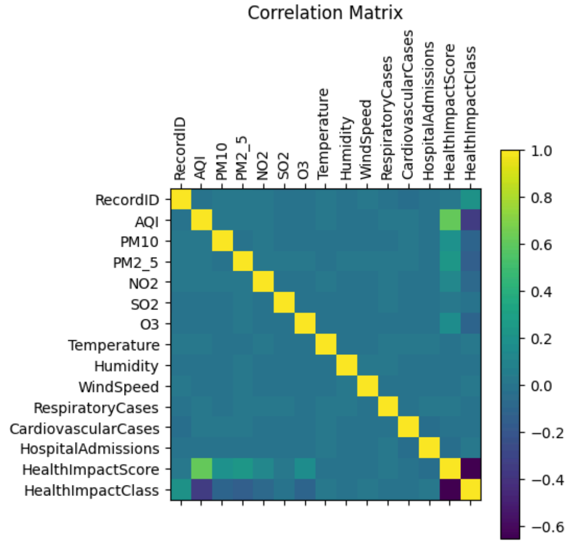


Fig.5. Heatmap of the dataset

Limitations

This research encountered several constraints that could impact the findings. An important issue was the imbalance in specific features like RespiratoryCases, CardiovascularCases and HospitalAdmissions potentially introducing biases into the predictive models. Additionally, some models, such as KNN, were affected by their reliance on feature scaling and sensitivity to noisy data, while SVM exhibited challenges in capturing complex non-linear relationships due to the dataset's multidimensional nature. Limited hyperparameter optimization was conducted, which might have restricted the full potential of certain models, particularly SVM and Random Forest. The dataset's relatively small size for certain health-related features also posed challenges in drawing generalized conclusions, and the models' applicability to other regions or datasets with different pollution-health dynamics remains unexplored. Lastly, the absence of advanced feature engineering, such as leveraging temporal patterns or derived variables, may have limited the models' ability to uncover deeper insights. Future studies should address these challenges to enhance the robustness and real-world relevance of the findings.

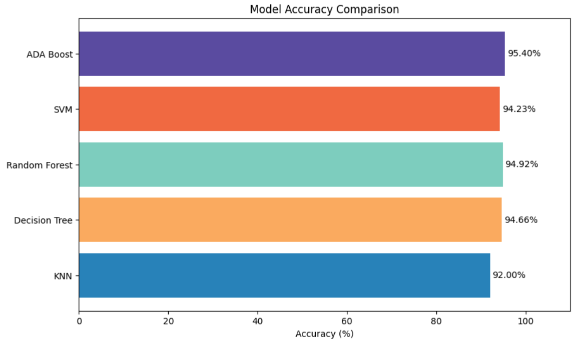


Fig.2. Performance Comparison

VI. Conclusion

This study demonstrated the effectiveness of machine learning models in predicting health impact classes using air pollution and health-related data. AdaBoost achieved the highest accuracy (95.40%), followed by Random Forest and Decision Tree, highlighting the strength of ensemble methods in handling complex datasets. The findings emphasize the importance of integrating environmental and health metrics for better public health predictions.Future work should address feature imbalances, incorporate advanced feature engineering, and explore deeper models to improve accuracy and generalizability. These insights provide a data-driven framework for policymakers to design targeted interventions, mitigating the health impacts of air pollution.

References:

1. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. *Complexity*, *2020*, 1–23. https://doi.org/10.1155/2020/8049504
2. Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00548-1
3. Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022). Data-Driven Machine learning in Environmental Pollution: Gains and problems. *Environmental Science & Technology*, *56*(4), 2124–2133. https://doi.org/10.1021/acs.est.1c06157
4. Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., ... & Zhang, C. (2017). Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC bioinformatics*, *18*, 121-131.
5. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *24*(1), 198.
6. Dimopoulos, A. C., Nikolaidou, M., Caballero, F. F., Engchuan, W., Sanchez-Niubo, A., Arndt, H., ... & Panagiotakos, D. B. (2018). Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC medical research methodology*, *18*, 1-11.
7. Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *applied sciences*, *10*(24), 9151.
8. Sicard, P., Lesne, O., Alexandre, N., Mangin, A., & Collomp, R. (2011). Air quality trends and potential health effects–development of an aggregate risk index. *Atmospheric environment*, *45*(5), 1145-1153.
9. Chen, L., Villeneuve, P. J., Rowe, B. H., Liu, L., & Stieb, D. M. (2014). The Air Quality Health Index as a predictor of emergency department visits for ischemic stroke in Edmonton, Canada. *Journal of exposure science & environmental epidemiology*, *24*(4), 358-364.