

Disentangling Causality from Correlation in Data-Driven Systems: An Empirical Study of Spurious Associations in Air Quality and Respiratory Health

V. Sai Abhiram¹

¹Department of Computer Science and Engineering, Neil Gogte Institute of Technology, Hyderabad, Telangana, India.

ABSTRACT-The increasing reliance on data-driven systems has amplified concerns regarding the misinterpretation of statistical relationships as causal effects. While machine learning models excel at identifying patterns within observational data, they often fail to distinguish between correlation and true causation, leading to the learning of spurious associations. This paper presents an empirical investigation into this limitation through the lens of environmental health analytics, specifically examining the relationship between air quality and respiratory risk. Using a real-world air quality dataset, a series of controlled experiments are designed to evaluate how predictive models respond to confounding variables such as seasonal variation. A Random Forest-based framework is employed to compare model performance under different feature configurations, including isolated pollutant indicators and enriched contextual variables. The results demonstrate that models trained solely on correlated features achieve strong predictive performance but fail to generalize under distributional shifts, revealing their dependence on non-causal patterns. Further analysis highlights the significant influence of hidden confounders, which distort feature importance and inflate perceived causal contributions. By systematically exposing these limitations, the study underscores the critical gap between predictive accuracy and causal validity in modern machine learning systems. The findings emphasize the necessity of incorporating causal reasoning, robustness evaluation, and context-aware modeling in high-stakes domains such as healthcare and environmental policy. This work contributes to a deeper understanding of spurious pattern learning and provides practical insights into improving the reliability and interpretability of predictive models.

Key Words: Causality, Correlation, Spurious Correlations, Machine Learning, Confounding Variables, Air Quality Index (AQI), Respiratory Health, Random Forest

1. INTRODUCTION

Data-driven systems and machine learning models are increasingly used to support decision-making in domains such as healthcare and environmental monitoring. While these models are highly effective at identifying patterns in large datasets, they primarily rely on statistical associations and often fail to distinguish between correlation and

causation. This limitation is critical, as incorrect causal interpretation can lead to misleading insights and unreliable predictions.

A fundamental principle in statistics states that correlation does not imply causation [11]. However, most machine learning models optimize for predictive accuracy rather than causal understanding, leading to the learning of spurious correlations—patterns that exist in data but do not reflect true causal relationships. This issue becomes more pronounced in the presence of confounding variables, which influence both input features and outcomes, thereby distorting model behaviour [1], [3].

Recent research highlights that machine learning models often rely on shortcut learning, where they exploit non-causal but highly predictive features to minimize error [6], [7]. Although this improves performance on training data, it reduces robustness and generalization when data distributions change [4], [5].

In the context of environmental health, air quality is commonly associated with respiratory risk [9], [10], but this relationship is influenced by factors such as seasonal variation. These confounders can lead models to overestimate the causal role of air quality.

This paper investigates how machine learning models interpret such relationships and demonstrates that high predictive performance does not necessarily imply causal validity.

2. BACKGROUND AND THEORETICAL FOUNDATION

Understanding the distinction between correlation and causation is essential for interpreting the behavior of modern data-driven systems. While machine learning models are designed to identify patterns within data, they often operate on observational relationships without accounting for underlying causal mechanisms. This limitation can lead to the learning of misleading associations, especially in the presence of hidden variables and complex data dependencies. To address this, it is important to examine key theoretical concepts such as confounding variables, spurious correlations, and the inherent limitations of traditional predictive models. This section provides the necessary conceptual foundation to analyze how and why such issues arise in machine learning systems

2.1 Correlation vs Causation

Correlation refers to a statistical relationship where two variables change together, whereas causation implies that one variable directly influences another. Although correlated variables may appear related, such relationships do not confirm a cause-and-effect link.

In observational data, correlations often arise due to hidden factors or indirect relationships, which can mislead analysis. As highlighted in causal inference literature, relying only on statistical associations can result in incorrect conclusions about causality [1], [11].

Machine learning models, which are primarily designed for prediction, tend to capture these correlations without distinguishing whether they are causal. This makes it essential to carefully interpret model outputs, especially in domains where understanding true causal relationships is critical.

2.2 Confounding Variables

Confounding variables are hidden factors that influence both the input features and the target outcome, creating a misleading association between them. In such cases, a model may incorrectly interpret a correlation as a direct causal relationship, even though the observed effect is driven by an external variable.

In observational data, confounding is a common issue, as not all influencing factors are explicitly measured or included in the dataset. As discussed in causal inference frameworks, failing to account for confounders can lead to biased conclusions and incorrect model interpretations [1], [3], [8]. Machine learning models are particularly vulnerable to this problem, as they rely on available data without understanding underlying causal structures. As a result, they may assign importance to features that are only indirectly related to the outcome, reducing the reliability and generalizability of predictions. Recognizing and addressing confounding variables is therefore essential for improving the robustness of data-driven systems.

2.3 Spurious Correlations in Machine Learning

Spurious correlations refer to patterns learned by machine learning models that appear predictive but do not reflect true underlying relationships. These correlations often arise when models rely on easily identifiable features that are indirectly associated with the target variable rather than causally relevant factors.

Recent studies show that machine learning systems frequently exhibit shortcut learning, where models exploit non-causal signals to achieve high predictive accuracy [6], [7]. While such patterns may improve performance on training data, they fail to generalize under changing conditions, leading to unreliable outcomes.

This issue highlights a key limitation of data-driven models: they prioritize statistical associations over causal

understanding. As a result, models may perform well in controlled settings but struggle in real-world scenarios where underlying relationships differ. Addressing spurious correlations is therefore critical for building robust and trustworthy machine learning systems.

2.4 Limitations of Traditional Machine Learning Models

Traditional machine learning models are primarily designed to optimize predictive accuracy by learning patterns from observational data. However, they do not inherently distinguish between causal relationships and statistical correlations. As a result, these models may rely on features that are predictive but not causally meaningful.

This limitation becomes evident when models are exposed to distributional changes, where the learned correlations no longer hold. In such cases, model performance can degrade significantly, indicating a lack of robustness. Research in causal machine learning highlights that standard predictive approaches often fail to capture stable, invariant relationships across different environments [4], [5].

Consequently, while traditional models can achieve high accuracy under specific conditions, their inability to account for causality limits their reliability in real-world applications. This underscores the need for incorporating causal reasoning to improve generalization and interpretability in data-driven systems.

3. PROBLEM FORMULATION

Air quality has long been associated with adverse respiratory outcomes, with numerous studies indicating a strong relationship between pollution levels and health risks [9], [10]. However, this observed relationship is often derived from observational data, where multiple external factors may influence both air quality and respiratory conditions simultaneously. As a result, the apparent association between Air Quality Index (AQI) and respiratory risk may not fully represent a direct causal effect.

One of the key challenges in such settings is the presence of confounding variables, such as seasonal variation, which can simultaneously impact pollution levels and health outcomes. For instance, certain seasons are characterized by both higher pollution concentrations and increased respiratory issues, leading to a situation where AQI and health risk appear strongly correlated even when influenced by a common underlying factor. This raises the possibility that predictive models trained on such data may misinterpret correlation as causation.

Based on this observation, the central hypothesis of this study is that machine learning models trained on observational air quality data may rely on spurious correlations arising from hidden confounders, rather than capturing true causal relationships. To investigate this, the objective of this work is to empirically evaluate how

predictive models behave under different feature configurations and distributional conditions. Specifically, the study aims to analyse whether models trained on correlated features can generalize across varying conditions or if their performance degrades due to reliance on non-causal patterns.

By framing the problem in this manner, the study provides a structured approach to examine the gap between predictive accuracy and causal validity in data-driven systems.

4. DATASET AND FEATURE ENGINEERING

This study uses a real-world air quality dataset from the Central Pollution Control Board (CPCB), containing daily pollutant measurements across Indian cities. For this analysis, data from a single city (Delhi) is selected to maintain consistency and capture clear temporal patterns. Key attributes include date, Air Quality Index (AQI), and pollutant concentrations such as PM2.5 and PM10.

Basic preprocessing is performed by removing records with missing AQI values and handling remaining missing data using forward filling. The date field is converted into datetime format to enable feature extraction.

To incorporate contextual information, the month is extracted from the date and mapped to seasonal categories (Winter, Summer, Monsoon, and Post-Monsoon), allowing seasonal variation to act as a potential confounding factor.

As direct health data is unavailable, a proxy target variable called the *Respiratory Risk Index* is constructed using normalized AQI and PM2.5 values, adjusted with season-based weighting. This design introduces controlled confounding, enabling analysis of how models respond to hidden influences.

The final dataset includes both original and engineered features, forming the basis for evaluating spurious correlations in predictive models.

5. METHODOLOGY

This study adopts a structured experimental approach to evaluate how machine learning models interpret relationships between air quality and respiratory risk. A Random Forest regression model is used due to its ability to capture complex patterns and provide feature importance insights. To systematically analyze the presence of spurious correlations, three experimental configurations are designed, each varying in feature composition and data distribution. This setup enables a comparative evaluation of model behavior under different conditions.

Pseudocode:

Algorithm: Detection of Spurious Correlations in Air Quality Prediction

Input: Air quality dataset D

Output: Model performance and correlation analysis

1. Load dataset D
2. Perform preprocessing and handle missing values

3. Extract temporal features (month, season)
4. Construct target variable (Respiratory Risk Index)
5. Define evaluation metrics (MSE, R²)
6. Experiment-1: Train model using AQI only
Evaluate performance
7. Experiment-2: Train model using AQI, PM2.5, and Season
Evaluate performance
Extract feature importance
8. Experiment-3: Train model on Winter data
Test model on non-Winter data
Evaluate generalization performance
9. Compare results across experiments
10. Analyse presence of spurious correlations

End

5.1 Experiment 1 – AQI Only

In the first experiment, the model is trained using AQI as the sole input feature to predict the Respiratory Risk Index. This setup evaluates the extent to which AQI alone can explain variations in the target variable. The objective is to establish a baseline performance and observe whether a single correlated feature can produce strong predictive results.

5.2 Experiment 2 – AQI, PM2.5, and Season

In the second experiment, additional features including PM2.5 and season are incorporated into the model. This configuration introduces contextual and confounding variables, allowing analysis of how feature importance and model performance change when more information is provided. The goal is to identify whether the model relies solely on AQI or distributes importance across multiple factors.

5.3 Experiment 3 – Seasonal Generalization

In the third experiment, the model is trained exclusively on data from the winter season and evaluated on data from non-winter periods. This setup tests the model’s ability to generalize across different data distributions. A significant drop in performance would indicate reliance on seasonal correlations rather than stable causal relationships.

6. RESULTS AND ANALYSIS

Table 1: Model Performance Comparison across Experiments

Experiment	Features Used	MSE	R ² Score
Experiment 1	AQI	0.00396	0.8788
Experiment 2	AQI, PM2.5, Season	0.00007	0.9978
Experiment	Winter →	0.00855	

3	Non-Winter	0.5510
---	------------	--------

The performance of the predictive model across different experimental configurations is summarized in Table 1. In the first experiment, where AQI is used as the sole feature, the model achieves a relatively high R^2 score of 0.8788. This indicates a strong correlation between AQI and the constructed Respiratory Risk Index, suggesting that AQI alone appears to be a reliable predictor.

In the second experiment, the inclusion of additional features such as PM2.5 and seasonal information significantly improves model performance, resulting in an R^2 score of 0.9978. This near-perfect fit indicates that the model benefits from additional contextual information. However, this also suggests that the model may be capturing combined effects of multiple variables, including potential confounding factors. The third experiment provides critical insight into model behaviour under distributional changes. When the model is trained exclusively on winter data and tested on non-winter data, the performance drops significantly, with the R^2 score decreasing to 0.5510. This decline indicates that the model struggles to generalize beyond the conditions it was trained on.

These results highlight the presence of spurious correlations in the learning process. While the model performs well under consistent conditions, its reduced performance under seasonal shifts suggests reliance on patterns that are not causally stable. In particular, the model appears to implicitly learn seasonal dependencies rather than isolating the true effect of air quality on respiratory risk.

Overall, the findings demonstrate that high predictive accuracy does not necessarily imply causal validity. The observed performance variations across experiments emphasize the importance of evaluating model robustness and accounting for confounding variables in data-driven systems.

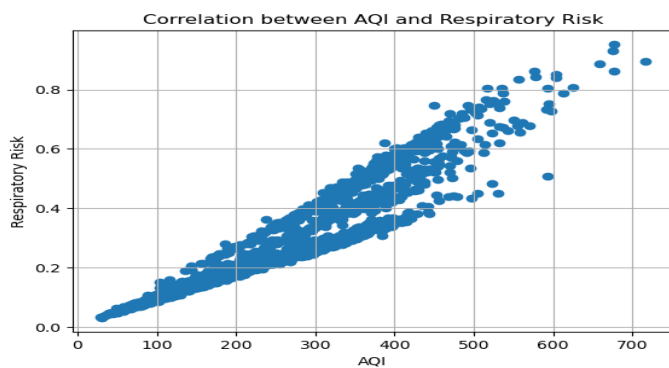


Fig-1: Correlation between AQI and Respiratory Risk

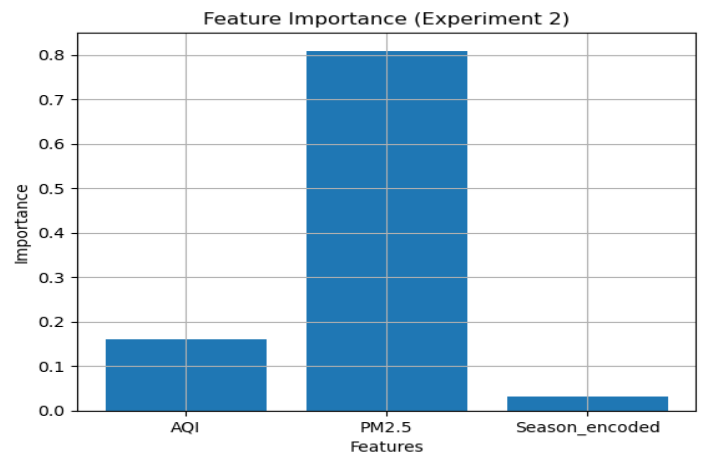


Fig-2: Feature Importance (Exp 2)

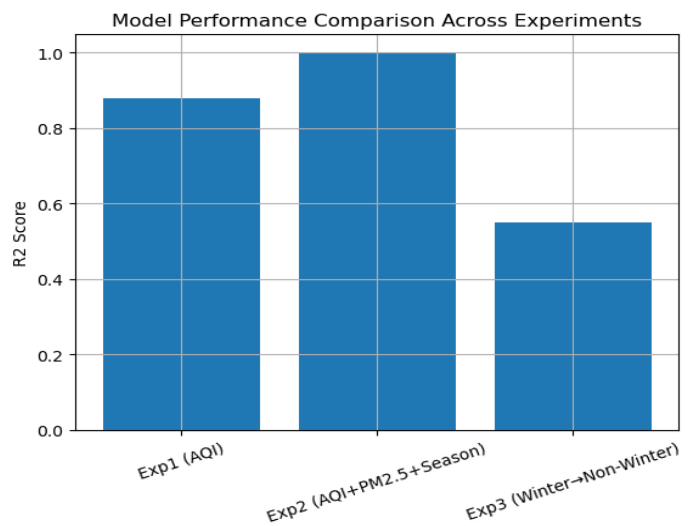


Fig-3: Model Performance Comparison Across Experiments

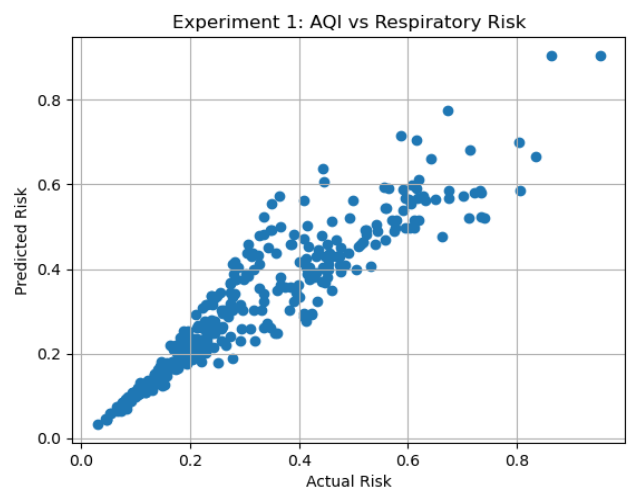


Fig-4: Actual vs Predicted Respiratory Risk for Experiment

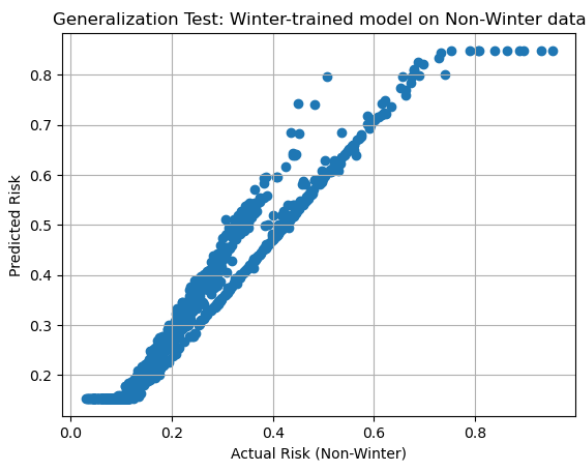


Fig-5: Generalization Plot

7. DISCUSSION

The results demonstrate a clear gap between predictive performance and causal understanding in machine learning models. Although the model achieves high accuracy in Experiments 1 and 2, this performance is largely driven by strong correlations rather than true causal relationships. The improvement observed after adding additional features indicates that the model benefits from contextual information, but it also suggests reliance on multiple variables, including potential confounders.

Feature importance analysis shows that the model does not depend solely on AQI, but instead utilizes other factors such as PM2.5 and seasonal information. This behaviour reflects shortcut learning, where models exploit easily available patterns to improve prediction accuracy without capturing the underlying causal structure [6], [7]. As a result, the model may perform well under consistent conditions but remain sensitive to hidden dependencies.

The generalization experiment provides the strongest evidence of spurious correlation. The significant drop in performance when the model is tested on non-winter data indicates that it fails to learn stable relationships across different conditions. This confirms that the model relies on season-specific patterns rather than causal effects, highlighting the importance of incorporating causal reasoning to improve robustness and reliability in data-driven systems [4], [5].

8. CONCLUSION

This study examined the distinction between correlation and causation in data-driven systems through an empirical analysis of air quality and respiratory risk. The results demonstrate that while machine learning models can achieve high predictive accuracy, such performance does not necessarily reflect true causal understanding. In particular, the experiments reveal that models often rely on correlated

features and hidden confounding factors, leading to the learning of spurious patterns.

The observed decline in performance under distributional changes highlights the limitations of traditional predictive approaches in capturing stable relationships. This emphasizes that models trained on observational data may fail to generalize when underlying conditions vary, thereby reducing their reliability in real-world applications.

Overall, the findings reinforce the importance of incorporating causal reasoning into machine learning frameworks. By addressing spurious correlations and improving model robustness, data-driven systems can become more interpretable, reliable, and suitable for critical domains such as healthcare and environmental decision-making.

9. FUTURE WORK

Future research can extend this study by incorporating real-world clinical or hospital-based respiratory health datasets. While this work utilizes a constructed proxy variable, integrating actual patient-level data would enable a more accurate assessment of causal relationships between air quality and health outcomes. This would also allow for validation of the observed patterns in a more realistic and domain-specific context.

Another important direction is the application of advanced causal inference techniques, such as structural causal models and intervention-based analysis. Methods like invariant risk minimization and causal graph-based approaches can help identify stable relationships that persist across different environments. Incorporating such techniques would improve the robustness of predictive models and reduce their dependence on spurious correlations.

Additionally, future work can explore model behaviour across multiple cities and diverse environmental conditions to assess generalizability at a broader scale. Expanding the analysis to different geographic regions and temporal settings would provide deeper insights into how confounding factors vary across contexts, further strengthening the reliability and applicability of data-driven systems in real-world scenarios.

REFERENCES

- [1] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge, U.K.: Cambridge University Press, 2009.
- [2] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA, USA: MIT Press, 2017.
- [3] M. A. Hernán and J. M. Robins, *Causal Inference: What If*, Boca Raton, FL, USA: Chapman & Hall/CRC, 2020.

[4] B. Schölkopf, "Causality for Machine Learning," arXiv preprint arXiv:1911.10500, 2019.

[5] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," arXiv preprint arXiv:1907.02893, 2019.

[6] R. Geirhos et al., "Shortcut Learning in Deep Neural Networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, 2020.

[7] M. Srivastava et al., "Learning the Difference that Makes a Difference," arXiv preprint arXiv:2007.06661, 2020.

[8] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[9] C. A. Pope III et al., "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution," *New England Journal of Medicine*, vol. 287, no. 2, pp. 1132–1141, 2002.

[10] D. W. Dockery et al., "An Association between Air Pollution and Mortality in Six U.S. Cities," *New England Journal of Medicine*, vol. 329, no. 24, pp. 1753–1759, 1993.

[11] D. A. Freedman, "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, vol. 14, no. 3, pp. 243–258, 1999.