Original Research Article

Enhancing Pollen Allergy Severity Predictions Through Machine Learning

Arnav Bansal

Dougherty Valley, 10550 Albion Rd, San Ramon, CA 94582, United States

ABSTRACT

Pollen allergies significantly impact global health, with rising prevalence and severity exacerbated by climate change. These conditions reduce quality of life and increase healthcare costs. Traditional pollen monitoring techniques are slow, labor-intensive, costly, and lack timely, location-specific accuracy, while general forecasts are often unreliable. This research develops a real-time prediction model for pollen allergy severity using environmental and meteorological data combined with advanced machine learning methods. Four models were evaluated: a baseline Random Forest, XGBoost, tuned Random Forest with time series cross-validation, and an advanced Random Forest incorporating cyclical date features, lagged pollen values, and rolling averages. The final model achieved the best performance with an R^2 value of 0.78. Significantly surpassing the approximately 50% accuracy typically achieved by prior forecasts. Results demonstrate that integrating environmental and seasonal features can substantially enhance the accuracy of pollen allergy severity predictions. Future work should aim to improve model generalizability across diverse regions and to expand the availability and temporal resolution of training data.

Keywords: Pollen Allergy; Machine Learning; Pollen Forecast; Random Forest; Prediction Model

INTRODUCTION

Pollen allergies are a major global health issue, affecting sixty million Americans each year and causing symptoms such as sneezing, congestion, and increased asthma attacks, which collectively lead to reduced productivity and significant health care costs (1). According to the U.S. Centers for Disease Control and Prevention, medical expenses related to allergic conditions exceed three billion dollars annually (1). Pollen exposure also inflames mucous membranes, increasing susceptibility to respiratory infections (2).

Compounding this problem, climate change is lengthening pollen seasons and intensifying pollen production. Warmer conditions and higher atmospheric CO_2 concentrations stimulate plant growth and allergen release, thereby elevating the frequency and severity of allergic reactions (2, 3).

Current approaches to managing pollen allergies primarily rely on anti-allergenic and anti-inflammatory medications, which often provide rapid but temporary relief (4). Studies indicate that 34.6% of individuals who track pollen counts do so for preparation and planning, 32.9% for guiding outdoor activities, and 28.2% for medication decisions (5). However, these strategies depend heavily on the availability and accuracy of pollen forecasts. The standard methodology for pollen monitoring typically provides results only after a delay of three to nine days (6), limiting the ability to make timely decisions. Commonly used forecasting tools,

Corresponding author: Arnav Bansal, E-mail: wisearnav@gmail.com.

Accepted October 27, 2025
https://doi.org/10.70251/HYJR2348.36101108

such as pollen.com, offer broad regional data but fail to reflect local variations (7). Personalized, location-specific estimates are therefore crucial for effective allergy management and public health planning (7) (8). Yet, conventional statistical methods employed by these systems often struggle to model the complex, non-linear relationships found in large environmental datasets (9, 10).

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), offers a promising alternative to conventional methods (7, 9). ML algorithms can learn intricate patterns and relationships within data, yielding superior predictive performance compared to traditional linear methods (9, 11). These methods have been successfully applied to environmental forecasting tasks using architectures such as Deep Neural Networks (DNN), Random Forests, Light Gradient Boosting Machine (LightGBM), and Artificial Neural Networks (ANN). For example, Rojo *et al*. achieved an R^2 value of 0.75 by combining Generalized Additive Models (GAM) with LightGBM and ANN algorithms to predict pollen concentrations (10). Machine learning methods have demonstrated resilience to noisy and non-linear data, providing a more reliable framework for forecasting environmental variables than conventional methods (7, 12). Despite these advantages, AI models in allergy immunology remain in early development and are often tested in a single location, raising concerns about their generalizability. Furthermore, differentiating morphologically similar pollen grains, such as those from fir, spruce, and pine, remains technically challenging (13).

To address these limitations, the present study aims to develop a more accurate, real-time prediction model for pollen allergy severity by integrating diverse environmental and meteorological datasets. This includes regional pollen counts, weather parameters (temperature, humidity, wind speed), and air quality indices, which together capture the environmental context influencing pollen levels. By applying advanced machine learning techniques to this combined dataset, this research seeks to provide more precise, localized, and actionable pollen forecasts that can enhance both personal allergy management and public health response.

METHODS AND MATERIALS

This study used only publicly available environmental and meteorological data. Therefore, no human subjects or identifiable information were involved, and Institutional Review Board approval was not required.

Data Sources

This study utilized two primary datasets: weather data and pollen count data, both corresponding to Raleigh, North Carolina. Figure 1 presents an overview of total pollen counts across the timespan during which the data was collected.

Weather Data Acquisition

Weather data were obtained from Visual Crossing, a global weather data provider. The data spanned from January 1, 2023, to August 18, 2025, and corresponded to the coordinates 4403 Reedy Creek Road, Raleigh, NC 27607. The raw weather dataset, initially named *weather_nc*, comprised 33 columns, including *name, datetime, tempmax, tempmin, temp, feelslikemax,*

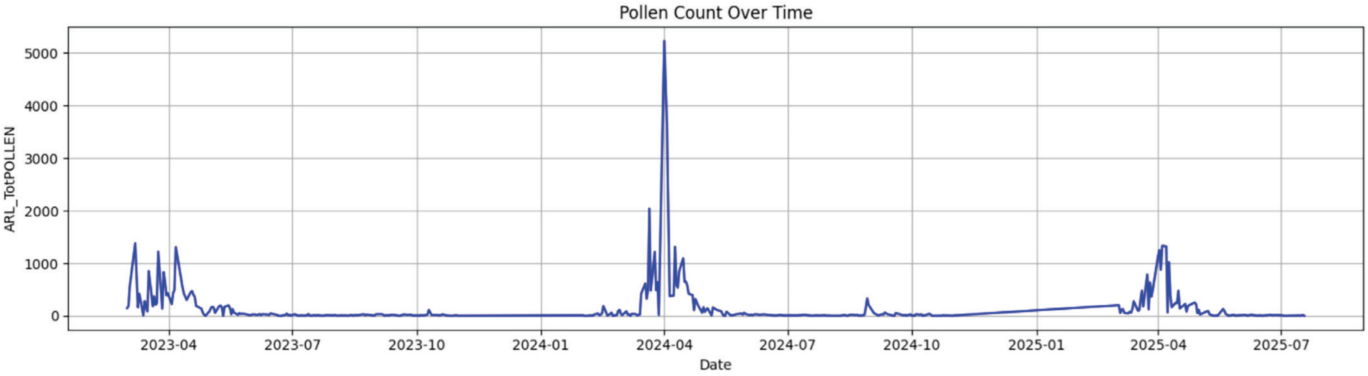


Figure 1. Daily pollen counts recorded in Raleigh, NC (2023-2025), showing distinct seasonal peaks corresponding to spring months.

feelslikemin, feelslike, dew, humidity, precip, precipprob, precipcover, preciptype, snow, snowdepth, windgust, windspeed, winddir, sealevelpressure, cloudcover, visibility, solarradiation, solarenergy, uvindex, severerisk, sunrise, sunset, moonphase, conditions, description, icon, and *stations*.

Pollen Data Acquisition

Pollen count data were obtained from the North Carolina Division of Air Quality (DAQ), which operates the state's only pollen sampler at the same location (4403 Reedy Creek Road, Raleigh, NC 27607). Although DAQ does not provide prospective pollen forecasts, it provides historical counts through its *Pollen Trends Report* dating back to 1999. Sampling occurs Monday through Friday, excluding state holidays, between late February and mid-November. Each record represents 24 hours from 9 a.m. to 9 a.m. the next day. Pollen is collected for one minute every ten minutes, and samples are analyzed microscopically by DAQ technicians to determine daily counts.

The reported *ARL_TotPOLLEN* value represents the total pollen count, defined as the sum of individual pollen grains from trees, grasses, and weeds per cubic meter of air. Figure 1 presents an overview of total pollen counts across the full study period, highlighting clear seasonal peaks during spring months. To illustrate year-specific variability and data collection gaps, Figure 2 focuses on the 2023 pollen season, showing missing intervals corresponding to weekends and holidays. These discontinuities emphasize the challenges of continuous data collection and motivate the modeling of cyclical temporal patterns in later analyses.

Data Preprocessing

All data preprocessing was conducted in a Python environment using Google Colab, employing libraries such as *pandas*, *numpy*, *matplotlib*, and *seaborn*. The methodology included data acquisition, merging, cleaning, imputation, and feature engineering before model development.

Data Loading and Filtering

The pollen data (*PCNC.csv*) were loaded into a DataFrame (*pollen_nc*) containing two columns (*Date, ARL_TotPOLLEN*) and 433 records. Weather data (*WDNC.csv*) were loaded into *weather_nc* with 961 entries and 33 columns. From these, a subset of relevant weather variables was retained, including datetime, tempmax, tempmin, dew, humidity, precipcover, windspeed, winddir, sealevelpressure, cloudcover, solarradiation, solarenergy, uvindex, and moonphase. The datetime field was converted to Date to align with pollen observations.

Data Merging and Reindexing

Both filtered datasets were merged on Date via a left join, yielding 433 rows and 19 columns. The *ARL_TotPOLLEN* variable was converted to numeric using *pd.to_numeric()*. Missing dates were reintroduced using a complete date range (*pd.date_range()*), expanding the merged dataset to 871 rows. Missing values (NaN) appeared for dates without recorded pollen or weather data. A correlation heatmap (Figure 3) was generated to visualize relationships among numeric variables, showing the strongest correlations between pollen counts and temperature, humidity, and solar radiation.

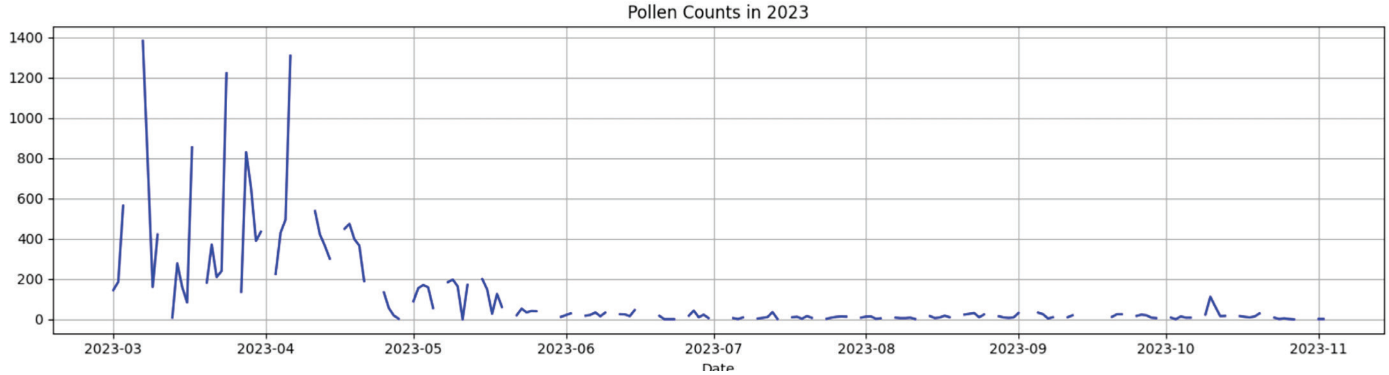
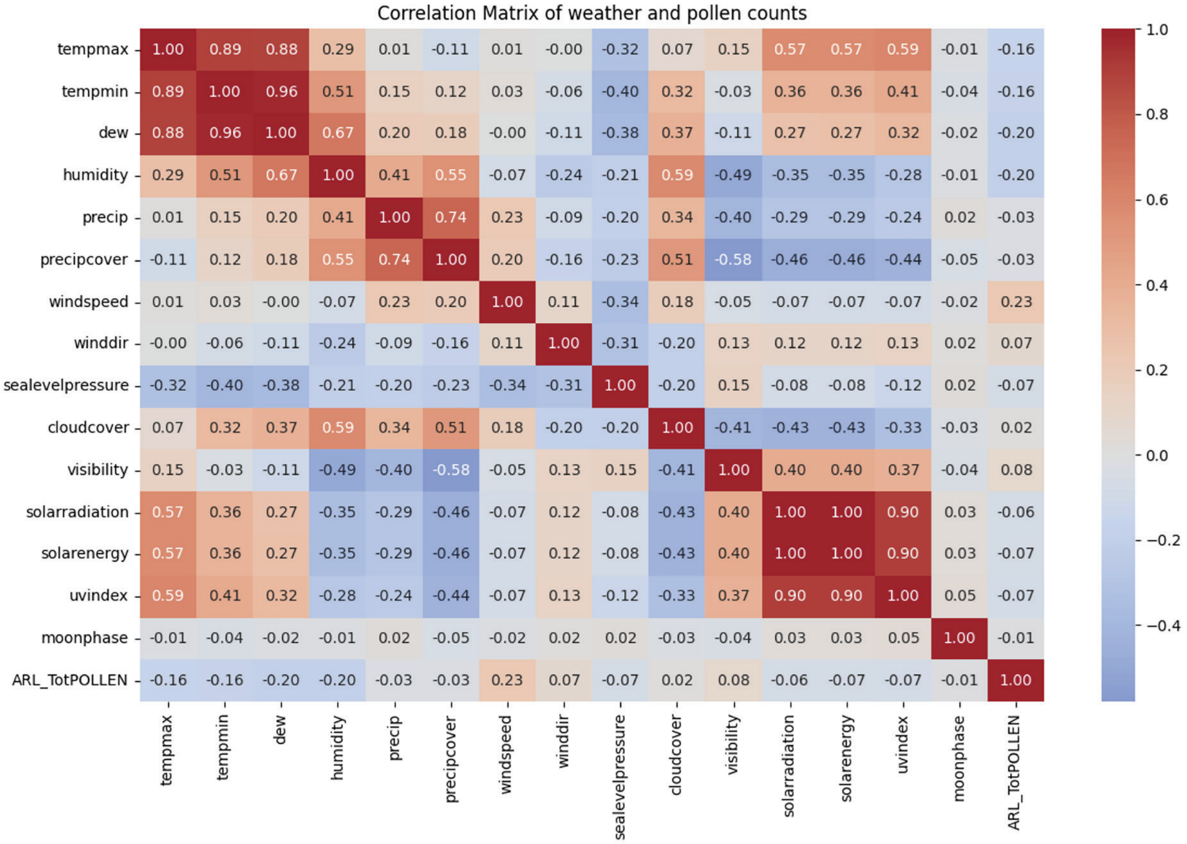


Figure 2. Daily pollen counts for Raleigh, NC (2023), illustrating year-specific variation and short gaps in monitoring during non-sampling periods (weekends and holidays).



	tempmax	tempmin	dew	humidity	precip	precipcover	windspeed	winddir	sealevelpressure	cloudcover	visibility	solarradiation	solarenergy	uvindex	moonphase	ARL_TotPOLLEN
tempmax	1.00	0.89	0.88	0.29	0.01	-0.11	0.01	-0.00	-0.32	0.07	0.15	0.57	0.57	0.59	-0.01	-0.16
tempmin	0.89	1.00	0.96	0.51	0.15	0.12	0.03	-0.06	-0.40	0.32	-0.03	0.36	0.36	0.41	-0.04	-0.16
dew	0.88	0.96	1.00	0.67	0.20	0.18	-0.00	-0.11	-0.38	0.37	-0.11	0.27	0.27	0.32	-0.02	-0.20
humidity	0.29	0.51	0.67	1.00	0.41	0.55	-0.07	-0.24	-0.21	0.59	-0.49	-0.35	-0.35	-0.28	-0.01	-0.20
precip	0.01	0.15	0.20	0.41	1.00	0.74	0.23	-0.09	-0.20	0.34	-0.40	-0.29	-0.29	-0.24	0.02	-0.03
precipcover	-0.11	0.12	0.18	0.55	0.74	1.00	0.20	-0.16	-0.23	0.51	-0.58	-0.46	-0.46	-0.44	-0.05	-0.03
windspeed	0.01	0.03	-0.00	-0.07	0.23	0.20	1.00	0.11	-0.34	0.18	-0.05	-0.07	-0.07	-0.07	-0.02	0.23
winddir	-0.00	-0.06	-0.11	-0.24	-0.09	-0.16	0.11	1.00	-0.31	-0.20	0.13	0.12	0.12	0.13	0.02	0.07
sealevelpressure	-0.32	-0.40	-0.38	-0.21	-0.20	-0.23	-0.34	-0.31	1.00	-0.20	0.15	-0.08	-0.08	-0.12	0.02	-0.07
cloudcover	0.07	0.32	0.37	0.59	0.34	0.51	0.18	-0.20	-0.20	1.00	-0.41	-0.43	-0.43	-0.33	-0.03	0.02
visibility	0.15	-0.03	-0.11	-0.49	-0.40	-0.58	-0.05	0.13	0.15	-0.41	1.00	0.40	0.40	0.37	-0.04	0.08
solarradiation	0.57	0.36	0.27	-0.35	-0.29	-0.46	-0.07	0.12	-0.08	-0.43	0.40	1.00	1.00	0.90	0.03	-0.06
solarenergy	0.57	0.36	0.27	-0.35	-0.29	-0.46	-0.07	0.12	-0.08	-0.43	0.40	1.00	1.00	0.90	0.03	-0.07
uvindex	0.59	0.41	0.32	-0.28	-0.24	-0.44	-0.07	0.13	-0.12	-0.33	0.37	0.90	0.90	1.00	0.05	-0.07
moonphase	-0.01	-0.04	-0.02	-0.01	0.02	-0.05	-0.02	0.02	0.02	-0.03	-0.04	0.03	0.03	0.05	1.00	-0.01
ARL_TotPOLLEN	-0.16	-0.16	-0.20	-0.20	-0.03	-0.03	0.23	0.07	-0.07	0.02	0.08	-0.06	-0.07	-0.07	-0.01	1.00

Figure 3. Correlation matrix between weather variables and total pollen counts, showing the strongest relationships with temperature, humidity, and solar radiation.

Missing Value Imputation

Missing numeric values were imputed sequentially using linear interpolation (*method = 'linear'*), followed by backward and forward filling to address leading or trailing gaps. This ensured complete data coverage before model training.

Machine Learning Model Development

Four machine learning models were implemented to progressively improve pollen prediction accuracy, moving from untuned baselines to tuned, seasonally aware models. In all cases, the input features (*X*) excluded categorical variables (*description, icon*), and the target variable (*y*) was *ARL_TotPOLLEN.* The dataset was split into training (80%) and testing (20%) sets, with shuffle=False to preserve temporal order. Table 1 summarizes the four models evaluated in this study, which progress from a simple baseline to more advanced time-aware and feature-engineered configurations. Model performance was then evaluated using mean absolute error (MAE), mean squared error (MSE), and R^2. These metrics were computed on the held-out test set to assess predictive accuracy.

Machine Learning Model Descriptions

The four models developed in this study provided a systematic progression from basic to advanced approaches. Each model incorporated increasingly complex design elements to capture temporal and environmental dynamics in pollen concentration.

Random Forest (Baseline Model)

The baseline Random Forest model served as an initial benchmark for performance. It did not include any feature engineering (no lagged features or rolling averages), and hyperparameter tuning was not applied. The model was implemented using the RandomForestRegressor with 100 trees and was trained directly on the raw training data.

XGBoost (Baseline Model)

The XGBoost regressor was developed to compare

Table 1. Summary of machine learning models developed for pollen allergy severity prediction

Model Name	Special Features	Tuning Method
Random Forest (Baseline)	No feature engineering or tuning	None
XGBoost	Gradient boosting, no engineered features	None
Random Forest + Time Series CV	Time-aware validation (avoids leakage)	RandomizedSearchCV with TimeSeriesSplit
Random Forest + Cyclical Features	Cyclical date encoding, lagged, and rolling averages	RandomizedSearchCV (5-fold CV)

performance between ensemble tree-based methods. While it similarly excluded engineered features, it used the gradient-boosting framework, which can outperform Random Forests when sufficient data and tuning are available. The model was implemented with XGBRegressor using squared-error loss and default parameters, without additional optimization.

Random Forest with Time-Series Cross-Validation

To improve evaluation reliability, a time-aware validation strategy was employed using Time-SeriesSplit. This approach prevents data leakage by ensuring that training always precedes testing chronologically. Hyperparameters were optimized via RandomizedSearchCV, which tested randomized combinations of parameters such as the number of trees, tree depth, and maximum features. Validation was conducted over five sequential folds. This version retained the same feature set as the baselines but incorporated systematic tuning and validation appropriate for time-series data.

Random Forest with Cyclical and Lagged Features

The final model integrated explicit temporal structure through engineered features. Date values were transformed into cyclical sine and cosine components to represent the repeating annual pollen cycle. Additional lagged features were generated from pollen counts on the one, two, and three previous days, while seven-day rolling averages and rolling standard deviations were computed to capture short-term fluctuations. Rows containing missing values introduced by these lag or rolling operations were removed. The model was trained on this enhanced dataset, excluding non-predictive variables such as Date, description, and icon. The RandomizedSearchCV procedure was again used for tuning with five-fold cross-validation, maintaining the chronological order of the data. This approach allowed the model to learn complex, non-linear, and seasonal relationships between weather variables and pollen counts.

Handling Missing Values after Feature Engineering

Following feature engineering, any rows with NaN values were dropped to prevent bias during training. The data was then split into training (80 %) and testing (20 %) subsets with shuffle=False to preserve temporal sequence. Hyperparameter tuning was applied using the same parameter grid as in the time-series cross-validation model.

Together, these models formed a structured pipeline progressing from simple baselines to tuned, feature-engineered frameworks that better captured temporal and environmental dependencies in pollen data. Model performance metrics are reported and compared in the Results section (Table 2).

Table 2. Comparative model performance metrics for predicting pollen allergy severity

Model	Date Features	# Of Lagged Features	# of Rolling Avg Features	Tuning Method	MAE	MSE	R²
Random Forest	Not Included	0	0	N/A	187	104,875	-0.44
XG Boost	Not Included	0	0	N/A	154	72,125	0.010
Random Forest (Tuned, TS CV)	Time Series Split (5)	0	0	Randomized. Search CV	59	18,031	0.76
Random Forest (Cyclical + Lagged + Rolling)	Cyclical (sin, cos)	3	2	Randomized Search CV	61	15,979	0.78

RESULTS

Table 2 presents the performance metrics for all four models. The baseline Random Forest and XGBoost models, both trained without tuning or engineered features, performed poorly. The baseline Random Forest yielded a Mean Absolute Error (MAE) of 187, a Mean Squared Error (MSE) of 104,875, and an R^2 of –0.44, indicating that it performed worse than simply predicting the mean pollen count. The XGBoost model achieved slightly better accuracy, with an MAE of 154, MSE of 72,125, and R^2 of 0.01, yet still demonstrated limited predictive capability.

Substantial improvement occurred with the tuned Random Forest using time-series cross-validation. This model achieved an MAE of 59, MSE of 18,031, and an R^2 of 0.76, demonstrating that hyperparameter tuning and appropriate time-aware validation significantly enhanced performance.

The best-performing model incorporated cyclical data, lagged, and rolling features. It achieved an MAE of 61, MSE of 15,979, and an R^2 of 0.78, capturing 78% of the variance in pollen counts. Although its MAE was slightly higher than the tuned time-series model, its lower MSE and higher R^2 indicate better generalization and reduced overall error. These results collectively demonstrate that model sophistication, particularly time-series validation and temporal feature engineering, substantially improved predictive accuracy.

DISCUSSION

The findings confirm that generalized, untuned models are inadequate for accurately forecasting pollen levels, consistent with prior research indicating that standard approaches often fail to capture local variability (12). The marked improvement observed in the tuned Random Forest models emphasizes the necessity of model optimization and validation strategies that preserve temporal order, preventing data leakage and yielding more realistic performance estimates.

The Random Forest models consistently outperformed XGBoost, likely due to their robustness to smaller datasets and noisy environmental variables. Random Forests effectively capture complex, non-linear interactions without extensive parameter tuning, while XGBoost typically requires larger datasets and more precise optimization to reach comparable accuracy.

Feature engineering played a key role in improving predictions. The cyclical date features successfully represented the repeating seasonal patterns of pollen release, while lagged pollen variables reflected the autoregressive nature of pollen data, where recent counts are strong indicators of near-future concentrations. Rolling averages further smoothed short-term fluctuations, improved stability, and reduced overfitting.

Deep learning approaches, such as convolutional neural networks (CNNs) or recurrent architectures (RNNs, LSTMs), represent plausible next steps. These models can capture complex temporal dependencies across longer pollen seasons, though their effectiveness will depend on the availability of more extensive multi-year, multi-location datasets.

From a practical perspective, the model's success demonstrates the feasibility of developing location-specific, data-driven tools for predicting pollen severity. This can empower individuals with allergies to make informed decisions about medication and outdoor activity, while public health agencies can use such models for targeted alerts and preparedness. Additionally, mobile and web-based applications could incorporate these models to deliver personalized pollen forecasts, improving accuracy beyond current estimates of roughly 50% (11).

Limitations

This study's scope was limited to data from Raleigh, North Carolina, which may affect generalizability to other regions. Environmental models often show location-specific behavior due to differences in vegetation, climate, and air quality. Expanding to additional cities and incorporating region-specific variables would improve robustness.

The model's high R^2 value (0.78) indicates strong predictive performance, yet the remaining unexplained variance (22%) likely reflects unpredictable meteorological variation and unobserved biological factors. As climate change alters pollen dynamics, lengthening seasons and increasing pollen intensity, the model may require retraining or adaptive updates to remain accurate.

Missing pollen counts on weekends and holidays also introduce potential bias by smoothing short-term fluctuations. While imputation mitigates this issue, it cannot fully substitute for real observations. Integrating supplementary datasets, such as satellite vegetation indices or air quality sensor data, could help fill these gaps.

Finally, confidence intervals for model metrics were not computed. While point estimates of MAE, MSE, and R² provide a basis for comparison, they do not quantify uncertainty or variability across resamples. Future work should incorporate bootstrapping or time-blocked cross-validation to produce confidence intervals, thereby providing more rigorous uncertainty quantification and enhancing interpretability.

CONCLUSION

This research demonstrates the potential of machine learning to improve real-time pollen allergy severity predictions using integrated environmental data. By leveraging feature engineering and time-aware validation, the study achieved substantial performance gains over conventional baselines. These models could significantly enhance individual symptom management and public health forecasting.

Further work is required to validate generalizability across regions, expand data sources, and include uncertainty estimates for performance metrics. Continued development in this area could form the foundation for highly accurate, personalized pollen forecasting systems that adapt to changing environmental conditions and support proactive allergy management.

FUNDING SOURCES

The author has no funding sources that supported the research and preparation of this article.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest related to this work

REFERENCES

1. Allergens and pollen. Climate and Health [Online]. (2024 Mar 2). Available from: https://www.cdc.gov/climate-health/php/effects/allergens-and-pollen.html (accessed on 2025-09-28)
2. Pollen. European Climate and Health Observatory [Online]. (2025 Mar 17). Available from: https://climate-adapt.eea.europa.eu/en/observatory/evidence/health-effects/aeroallergens (accessed on 2025-09-21)
3. Moore A. Allergy season is getting worse - thanks to climate change. *College of Natural Resources News [Online]*. (2023). Available from: https://cnr.ncsu.edu/news/2023/04/allergy-season-climate-change/ (accessed on 2025-09-25)
4. Bergmann K, *et al*. Nonpharmacological measures to prevent allergic symptoms in pollen allergy: A critical review. *Allergol Select [Online]*. 2021; 5 (1): 349-360. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8638355/. https://doi.org/10.5414/ALX02294E
5. Medek D, *et al*. Aerobiology matters: Why people in the community access pollen information and how they use it. *Clin Transl Allergy [Online]*. 2025 Jan 24; 15 (1): e70031. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11761002/. https://doi.org/10.1002/clt2.70031
6. Buters J, *et al*. Automatic detection of airborne pollen: an overview. Aerobiologia [Online]. 2022; 40 (1): 13-37. Available from: https://link.springer.com/article/10.1007/s10453-022-09750-x. https://doi.org/10.1007/s10453-022-09750-x
7. Zhu X, *et al*. Floating in the air: forecasting allergenic pollen concentration for managing urban public health. *Int J Digit Earth [Online]*. 2024; 17 (1). Available at: https://www.tandfonline.com/doi/full/10.1080/17538947.2024.2306894#abstract. https://doi.org/10.1080/17538947.2024.2306894
8. Ridolo E, *et al*. Current treatment strategies for seasonal allergic rhinitis: where are we heading? *Clin Mol Allergy [Online]*. 2022; 20 (1). Available at: https://clinicalmolecularallergy.biomedcentral.com/articles/10.1186/s12948-022-00176-x. https://doi.org/10.1186/s12948-022-00176-x
9. MacMath D, *et al*. Artificial Intelligence: Exploring the Future of Innovation in Allergy Immunology. *Curr Allergy Asthma Rep [Online]*. 2023; 23 (6): 351-362. Available at: https://pubmed.ncbi.nlm.nih.gov/37160554/. https://doi.org/10.1007/s11882-023-01084-z
10. Cordero J, *et al*. Predicting the Olea pollen concentration with a machine learning algorithm ensemble. *Int J Biometeorol [Online]*. 2020; 65 (4): 541-554. Available at: https://www.researchgate.net/publication/346440443_Predicting_the_Olea_pollen_concentration_with_a_machine_learning_algorithm_ensemble. https://doi.org/10.1007/s00484-020-02047-z
11. Lops Y, *et al*. Real-time 7-day forecast of pollen counts using a deep convolutional neural network. *Neural Comput Appl [Online]*. 2019; 32 (15): 11827-11836. Available at: https://www.researchgate.net/publication/337780830_Real-time_7-day_forecast_of_pollen_counts_using_a_deep_convolutional_neural_network. https://doi.org/10.1007/s00521-019-04665-0
12. Breugel M, *et al*. Current state and prospects of

artificial intelligence in allergy. *Allergy [Online].* 2023; 78 (10): 2623-2643. Available at: https://onlinelibrary.wiley.com/doi/10.1111/all.15849. https://doi.org/10.1111/all.15849

13. Goudarzi G, *et al*. Prediction of airborne pollen concentrations by artificial neural networks and their relationship with meteorological parameters and air pollutants. *J Environ Health Sci Eng [Online]*. 2022; 20 (1): 251-264. Available at: https://pubmed.ncbi.nlm.nih.gov/35669831/. https://doi.org/10.1007/s40201-021-00773-z