Spatiotemporal Prediction of Pollen Distribution for Genus *Acer* Using Autoregressive Machine Learning Approaches

Tanay Panja[1], Yiluan Song[1,2], Kai Zhu[1]

October 20, 2025

[1] University of Michigan, School for Environment and Sustainability (SEAS)
[2] University of Michigan, Michigan Institute for Data Science (MIDAS)

Abstract

Maple (*Acer*) pollen is a major aeroallergen across the United States, contributing to seasonal allergic rhinitis and asthma exacerbations worldwide. Accurate pollen forecasting is vital for public health preparedness and clinical response, requiring specialized models for common pollen-dispersing species. Yet existing approaches are largely limited to sparse, generalized observational networks. Furthermore, short-term prediction is an understudied aspect in observational studies.

We present a novel spatiotemporal machine learning framework that integrates pollen observations with comprehensive environmental data and autoregressive features to predict daily *Acer* pollen concentrations. Meteorological predictors—including solar radiation, precipitation, temperature extremes, vapor pressure, and snow water equivalent—were lagged at 1-week, 1-month, and 3-month intervals to capture delayed effects. Additional engineered features incorporated cyclical seasonality and nonlinear interactions.

Predictive performance of logistic regression, random forest, XGBoost, LightGBM, and a multi-layer perceptron was evaluated, with the final system deployed in an interactive Streamlit web application for real-time forecasting and visualization. The multi-layer perceptron achieved the best performance, with a $\sim$16 percentage point improvement in F1 score over the environmental-only benchmark climatology model. Lagged features consistently ranked among the strongest predictors, confirming the value of autoregressive modeling. We applied Ordinary Kriging for spatial interpolation, generating continuous daily pollen surfaces across the continental United States.

The novelty of this study lies in combining autoregressive machine learning with multiscale meteorological predictors and geostatistical interpolation, moving beyond static climatology. This integrated approach captures short-term variability and seasonal dynamics, delivering more accurate predictions than traditional climatology-based models.

Our results show that machine learning can transform aeroallergen forecasting with actionable, high-resolution predictions. By advancing beyond climatology, this framework provides a foundation for public health tools to mitigate seasonal allergies and asthma.

Keywords: climatology, spatiotemporal modeling, pollen forecasting, environmental health

Funding: This work was supported by MIDAS through the Propelling Original Data Science (PODS) grant award.

1 Introduction

The genus *Acer*, commonly known as maple, represents one of the most significant sources of tree pollen allergens across temperate regions of the United States. During peak flowering periods in early to mid-spring, maple trees release substantial quantities of pollen that can travel hundreds of kilometers through atmospheric transport [35, 40]. Airborne *Acer* pollen concentrations frequently exceed 50 grains per cubic meter during peak release periods, with some regions experiencing concentrations above 200 grains/m^3 in areas with dense maple populations [13, 1].

Maple pollen is a well-documented aeroallergen that triggers seasonal allergic rhinitis, conjunctivitis, and can exacerbate asthma symptoms in sensitized individuals [10, 28]. The timing and intensity of *Acer* pollen release significantly impact public health, with early or prolonged pollen seasons associated with increased emergency department visits and medication usage among allergic patients [33, 16]. Accurate prediction of maple pollen concentrations is therefore essential for implementing effective early warning systems and enabling proactive management of allergic symptoms.

Current pollen prediction models for tree species rely predominantly on simple phenological approaches or basic regression techniques that correlate pollen release with cumulative degree-days or singular meteorological variables [14, 30]. While these models capture broad seasonal trends, they demonstrate limited accuracy in predicting day-to-day variations in pollen concentrations and often fail to account for the complex, nonlinear relationships between multiple environmental drivers [8]. Most existing approaches treat daily pollen observations as independent events, ignoring the temporal autocorrelation inherent in pollen release patterns and atmospheric transport processes [25].

Furthermore, conventional models typically incorporate only basic meteorological predictors such as temperature and precipitation, overlooking key environmental variables that influence both pollen production and dispersal [20]. Critical factors such as solar radiation, snow water equivalent, daylight duration, and water vapor pressure may directly affect flowering phenology, pollen maturation, and atmospheric transport, and are frequently omitted from prediction frameworks [31, 34]. For instance, shortwave radiation and daylight duration control photosynthetic energy availability and circadian flowering patterns, yet these variables are rarely integrated into comprehensive prediction models.

To address these limitations, we propose a spatiotemporal autoregressive model that explicitly captures the temporal dependencies in *Acer* pollen concentrations while incorporating a comprehensive suite of environmental predictors at multiple temporal lags. Figure 1 summarizes the end-to-end workflow from data ingestion through deployment. By modeling lagged environmental variables at 1-week, 1-month, and 3-month intervals and analyzing spatial correlations across monitoring networks, this approach aims to provide more accurate and reliable predictions of maple pollen concentrations. This dynamic modeling framework moves beyond static seasonal predictions to enable real-time estimation.

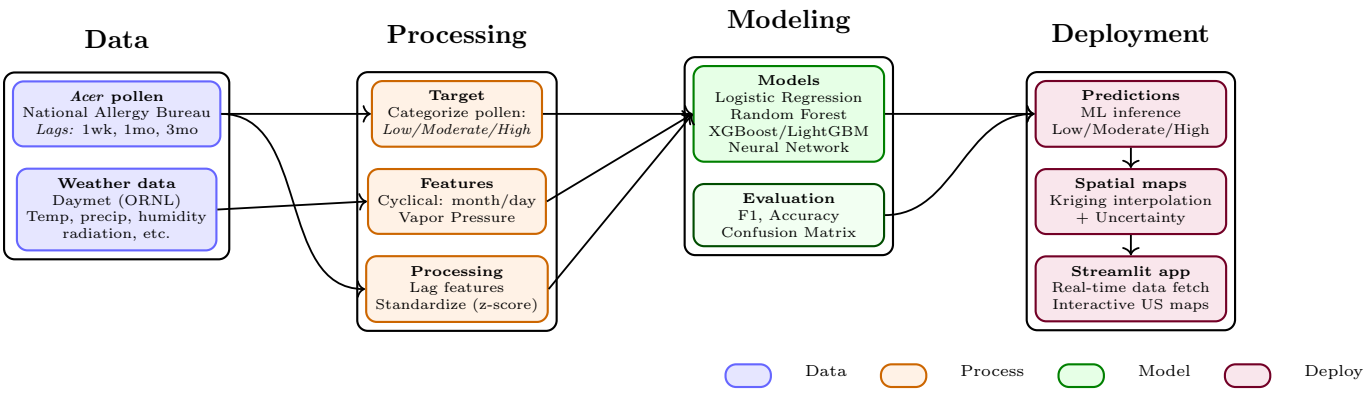

Figure 1: End-to-end workflow for the *Acer* pollen prediction framework, from data ingestion and processing to modeling and deployment.

2 Methods

2.1 Data Sources and Collection

Environmental data were obtained from the Daymet API managed by Oak Ridge National Laboratory, providing comprehensive daily measurements across the United States at 1km spatial resolution. These data included day length (duration of daylight period), precipitation, shortwave radiation, snow water equivalent, maximum and minimum air temperatures, and water vapor pressure for each observation location and date. Pollen concentration data were acquired from the National Allergy Bureau (NAB) [2], which maintains a standardized network of pollen counting stations across the United States. The NAB data provided daily *Acer* (maple) pollen counts, which were subsequently classified into discrete concentration levels to form the target variable `Acer_class`.

To contextualize temporal patterns in pollen exposure, autoregressive features were created by lagging both the *Acer* pollen observations and environmental variables [32]. For each observation, the corresponding pollen concentrations from one week prior (`Acer_lag_1week`) and one month prior (`Acer_lag_1month`) were incorporated as predictor variables compiled across all years. Additionally, all environmental variables were lagged at 1-week, 1-month, and 3-month intervals to capture the influence of antecedent conditions on pollen development and release. This approach enables the model to capture both short-term persistence and seasonal cyclical patterns characteristic of airborne pollen dynamics [13].

2.2 Data Loading and Exploration

The combined dataset comprising observations was imported from a comma-separated values file using the `pandas` library (v1.x) [24]. Basic exploration included reporting dataset dimensions, variable types, summary statistics for numerical variables, and the extent of missing values. This ensured that the dataset was suitable for downstream preprocessing and modeling.

2.3 Feature Engineering

2.3.1 Environmental Variables

The dataset contains the following core environmental variables: `dayl` (day length in seconds per day), `prcp` (daily precipitation in mm/day), `srad` (shortwave radiation in W/m^2), `swe` (snow water equivalent in kg/m^2), `tmax` (daily maximum air temperature in °C), `tmin` (daily minimum air temperature in °C), and `vp` (water vapor pressure). Spatial coordinates include `lat` (latitude) and `lon` (longitude), with the target variable `Acer_class` representing multiclass pollen concentration levels and autoregressive features `Acer_lag_1week` and `Acer_lag_1month` capturing temporal dependencies in pollen concentrations.

2.3.2 Temporal Lag Features

All environmental variables were systematically lagged at three temporal intervals: 1 week, 1 month, and 3 months prior to the prediction date. This comprehensive lagging approach recognizes that environmental conditions affecting pollen production operate across multiple time scales, from immediate meteorological influences on pollen release to longer-term seasonal conditions affecting flowering phenology and overall reproductive output [14].

2.3.3 Feature Engineering

Several temporal, spatial, and interaction-based features were derived from the raw dataset following established practices in environmental modeling [19]. The `Date` variable was decomposed into year, month, and day-of-year components. To capture cyclic temporal patterns, sine and cosine transformations were applied [18]:

$$\text{month}_{\sin} = \sin\left(\frac{2\pi \cdot \text{month}}{12}\right), \quad \text{month}_{\cos} = \cos\left(\frac{2\pi \cdot \text{month}}{12}\right),$$

$$\text{dayofyear}_{\sin} = \sin\left(\frac{2\pi \cdot \text{dayofyear}}{365}\right), \quad \text{dayofyear}_{\cos} = \cos\left(\frac{2\pi \cdot \text{dayofyear}}{365}\right).$$

The pollen concentration values were classified into three categorical levels based on established NAB thresholds that correlate with symptom severity in allergic individuals. The classification scheme is defined as:

$$\text{Pollen Level} = \begin{cases} \text{Low} & \text{if } x < 4.00 \text{ grains/m}^3 \\ \text{Moderate} & \text{if } 4.00 \leq x < 20.00 \text{ grains/m}^3 \\ \text{High} & \text{if } x \geq 20.00 \text{ grains/m}^3 \end{cases}$$

Autoregressive features were incorporated to capture temporal dependencies in pollen concentrations. The `Acer_lag_1week` variable represents pollen concentration from one week prior, while `Acer_lag_1month` captures pollen concentration from one month prior. These lagged variables enable the model to account for persistence and seasonal patterns in pollen dynamics, addressing the temporal autocorrelation [27]. This approach parallels recent ecological forecasting efforts that applied remote sensing to track reproductive phenology of wind-pollinated trees [37].

2.4 Benchmark Models

To provide meaningful performance comparisons, we implemented two standard benchmark models commonly used in environmental forecasting: persistence and climatology models. These benchmarks represent the minimum performance threshold that any skillful model should exceed.

2.4.1 Climatology Model

The climatology model predicts pollen concentrations based on the long-term daily climate mean. For each calendar day, the climatology model uses the historical average pollen concentration across all available years in the dataset for that specific day, smoothed using a 30-day running mean. This approach provides a seasonal baseline that reflects the typical phenological timing of *Acer* pollen release but does not account for interannual variability or current environmental conditions.

2.5 Data Preparation

The target variable was a multiclass label (`Acer_class`). Predictor variables excluded identifiers and the raw `Date` column. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve class distributions. Continuous features were standardized using z-score normalization:

$$x^* = \frac{x - \mu}{\sigma},$$

where μ and σ denote the mean and standard deviation of each feature, respectively.

2.6 Model Training and Evaluation

Five classification models were trained: Logistic Regression [17], Random Forest [5], eXtreme Gradient Boosting (XGBoost) [7], Light Gradient Boosting Machine (LightGBM) [22], and a feedforward Neural Network [15]. Tree-based ensemble methods served as strong baseline models, while the Neural Network was optimized with advanced architectures and regularization techniques. All models were optimized using randomized hyperparameter search with 5-fold stratified cross-validation [4]. Hyperparameters included maximum depth, number of estimators, learning rate, subsampling ratios, number of leaves, and for neural networks, hidden layer sizes, dropout rates, and activation functions.

2.7 Performance Metrics

Performance on the test set was assessed using accuracy, macro-averaged F1 score, and weighted F1 score [36]. For class $c \in C$, precision and recall were computed as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$

and the F1 score was defined as:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

The macro-averaged F1 was obtained by taking the arithmetic mean across all classes, while the weighted F1 score was computed as a class-frequency-weighted average. Confusion matrices were generated for qualitative assessment of misclassification patterns (see Figures 4–6).

2.8 Spatial Interpolation and Visualization

2.8.1 Ordinary Kriging

To generate continuous spatial predictions across the continental United States, Ordinary Kriging was employed as a geostatistical interpolation method [9]. Kriging provides optimal unbiased linear predictions by accounting for the spatial autocorrelation structure in the data through the variogram function.

For a set of n observation stations with known pollen concentrations $Z(s_1), Z(s_2), \ldots, Z(s_n)$ at spatial locations $s_1, s_2, \ldots, s_n$, the kriging predictor at an unsampled location s_0 is given by:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

where λ_i are the kriging weights determined by solving the kriging system:

$$\begin{bmatrix} \gamma(s_1, s_1) & \gamma(s_1, s_2) & \cdots & \gamma(s_1, s_n) & 1 \\ \gamma(s_2, s_1) & \gamma(s_2, s_2) & \cdots & \gamma(s_2, s_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n, s_1) & \gamma(s_n, s_2) & \cdots & \gamma(s_n, s_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(s_1, s_0) \\ \gamma(s_2, s_0) \\ \vdots \\ \gamma(s_n, s_0) \\ 1 \end{bmatrix}$$

where $\gamma(s_i, s_j)$ represents the semivariance between locations s_i and s_j, and μ is the Lagrange multiplier ensuring the unbiasedness constraint $\sum_{i=1}^{n} \lambda_i = 1$.

2.8.2 Implementation Details

Kriging interpolation was implemented using a spherical variogram model, which is well-suited for environmental data with clear spatial boundaries [39]. The interpolation grid covered the continental United States with coordinates spanning latitudes 25°N to 49°N and longitudes 124°W to 67°W at a resolution of 150×150 grid points. To ensure geographic accuracy, interpolated values were masked to the boundaries of the continental United States using polygon geometry, setting values outside the boundary to NaN.

The kriging system was configured with 12 lag distances to capture fine-scale spatial structure in pollen concentrations. This generates smooth, continuous prediction surfaces that respect the underlying spatial correlation structure while providing uncertainty estimates through kriging variance calculations.

2.9 Model Deployment and Web Application

2.9.1 Streamlit Web Interface

To facilitate real-world application and accessibility of the pollen prediction models, a web-based interface was developed using the Streamlit framework [38]. Streamlit provides a Python-native approach for creating interactive web applications with minimal overhead, making it well-suited for deploying machine learning models in research and operational contexts.

The web application serves as an interactive platform for generating daily *Acer* pollen concentration predictions across the continental United States. Users can select any date within the model's temporal domain through an intuitive date picker interface, enabling both retrospective analysis and same-day estimations. The application integrates trained models with real-time environmental data retrieval, ensuring that predictions incorporate the most current environmental conditions (see Figure 7 for the interface).

2.9.2 Architecture and Workflow

The application follows a modular architecture comprising several key components:

Data Retrieval Module Upon date selection, the system automatically queries environmental data APIs to obtain daily measurements for all National Allergy Bureau stations across the United States. This includes real-time or historical data for day length, precipitation, shortwave radiation, snow water equivalent, temperature extremes, and water vapor pressure corresponding to the selected prediction date.

Feature Engineering Pipeline The retrieved environmental data undergoes the same preprocessing and feature engineering transformations applied during model training. This includes temporal decomposition, cyclical encoding, interaction feature generation, lag feature computation, and standardization using parameters stored from the training phase. Autoregressive features (`Acer_lag_1week` and `Acer_lag_1month`) are populated using historical pollen observations.

Model Inference The preprocessed feature vectors are passed through the trained Neural Network classifier to generate pollen concentration class predictions for each monitoring station. The model outputs discrete concentration levels (Low, Medium, High) along with associated prediction confidence scores.

Spatial Interpolation Station-level predictions are transformed into continuous spatial surfaces using the Ordinary Kriging methodology described previously. The kriging algorithm generates smooth interpolated values across a high-resolution grid covering the continental United States, providing comprehensive geographic coverage beyond the discrete monitoring locations.

Visualization Interface Interactive maps display the interpolated pollen concentration surfaces using color-coded schemes that correspond to health advisory categories. Users can explore the spatial distribution of predicted pollen levels, zoom to specific regions of interest, and access detailed information for individual monitoring stations.

3 Results

3.1 Dataset Overview

The final dataset comprised observations spanning multiple years across geographic locations. The target variable `Acer_class` exhibited a balanced distribution across concentration levels. Missing values were minimal, accounting for less than 5% of all observations across environmental variables. The spatial coverage of observation stations with robust records is shown in Figure 2.

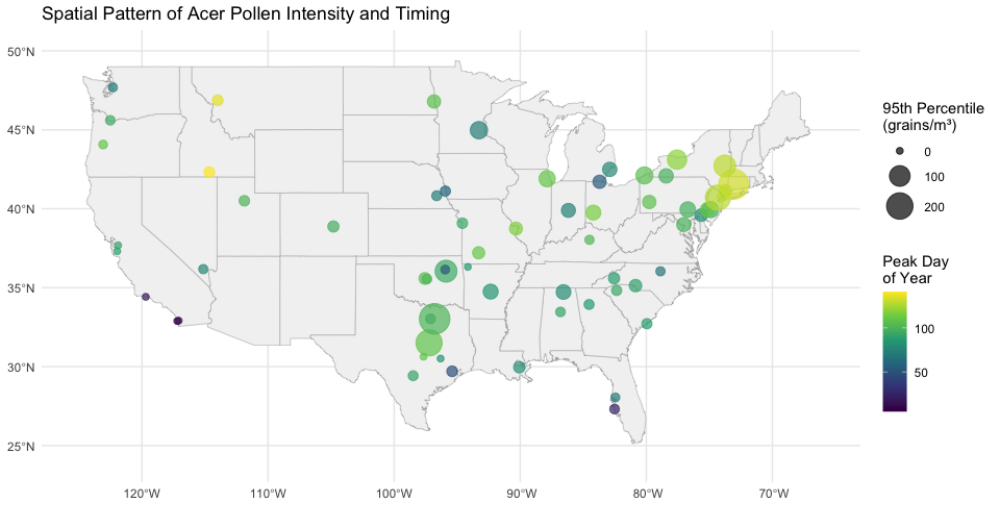

Figure 2: Map of all stations with greater than 1000 observations.

3.2 Correlation Analysis

Feature correlation analysis revealed several important relationships within the dataset (Figure 3). As expected, strong positive correlations were observed between temperature-related variables, with maximum and minimum temperatures showing a correlation coefficient of 0.87, and temperature variables correlating positively with day length ($r = 0.65$), reflecting seasonal patterns. The autoregressive features demonstrated moderate positive correlations

with the target variable `Acer_class`, with `Acer_lag_1week` (r = 0.21) and `Acer_lag_1month` (r = 0.23) showing similar predictive relationships.

Environmental variables showed expected seasonal correlation patterns. Shortwave radiation correlated positively with day length (r = 0.89) and temperature variables, while snow water equivalent exhibited negative correlations with temperature and radiation, consistent with seasonal snow dynamics. The lag environmental features at different temporal scales (1-week, 1-month, 3-month) showed decreasing correlations with current conditions, indicating the temporal decay of environmental influences.

The autoregressive pollen features were moderately correlated with each other (r = 0.63), indicating temporal persistence in pollen concentrations while maintaining sufficient independence to provide complementary information. Geographic coordinates showed expected negative correlation patterns, with latitude and longitude exhibiting a correlation of -0.76.

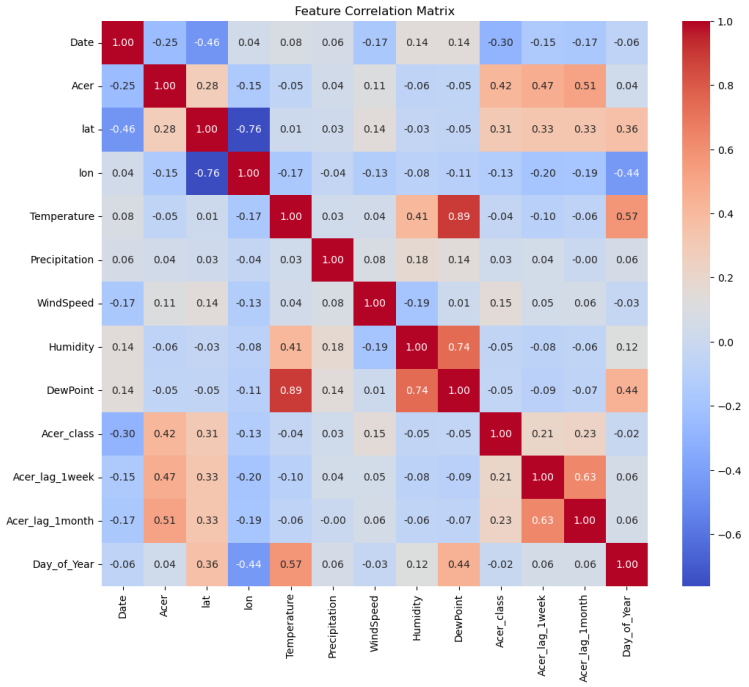

Figure 3: Pearson correlation matrix for the features and target.

3.3 Machine Learning Model Performance

Five classification models were evaluated on the test set, with results summarized in Table 1. The Neural Network achieved the highest overall performance with an accuracy of 0.682, macro-averaged F1 score of 0.680, and weighted F1 score of 0.684, outperforming the tree-based ensemble methods and logistic regression.

3.4 Feature Importance Analysis

Feature importance analysis from the best-performing Neural Network model revealed that autoregressive features were among the most predictive variables. The `Acer_lag_1week` vari-

Table 1: Model Performance Comparison

Model	Accuracy	F1 (Macro)	F1 (Weighted)
Logistic Regression	0.655	0.653	0.654
Random Forest (Optimized)	0.667	0.667	0.670
XGBoost (Optimized)	0.672	0.665	0.675
LightGBM (Optimized)	0.669	0.663	0.672
Neural Network	**0.682**	**0.680**	**0.684**

able ranked as the most important feature, while `Acer_lag_1month` ranked second, demonstrating the critical role of temporal dependencies in pollen prediction. Among environmental variables, lagged temperature variables (particularly 1-week and 1-month lags) showed high importance scores, followed by shortwave radiation and day length variables. The 3-month lag features provided additional seasonal context, with temperature and radiation lags from 3 months prior contributing to model performance during peak pollen season.

3.5 Model Interpretability

Confusion matrix analysis revealed consistent patterns across models. The most common misclassifications occurred between adjacent concentration classes (e.g., Low–Medium, Medium–High), suggesting that the models captured the underlying ordinal structure of pollen concentrations. The Neural Network showed superior performance in correctly classifying high concentration events, with improved precision for extreme classes compared to other methods (see Figures 4–6).

3.6 Spatial and Temporal Patterns

Cross-validation results across different geographic regions and time periods demonstrated model robustness. Performance remained consistent across different climate zones, with higher prediction accuracy during peak pollen season (March–May) compared to off-season periods. The spatial interpolation revealed clear geographic patterns consistent with *Acer* distribution and regional climatic gradients (illustrated interactively in Figure 7).

3.7 Comparison with Environmental-Only Models

The neural network approach demonstrated substantial improvements over environmental-only models. A model excluding autoregressive pollen features achieved 52% accuracy, representing a 16 percentage point decrease compared to the full Neural Network model. This finding underscores the importance of incorporating temporal dependencies in airborne pollen prediction.

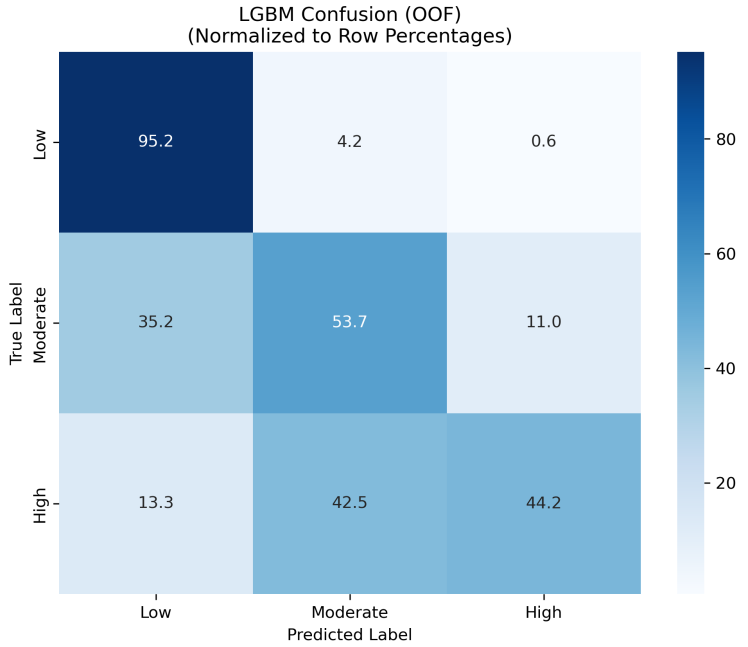

Figure 4: Confusion matrix for the LightGBM model showing classification performance across *Acer* pollen concentration classes.

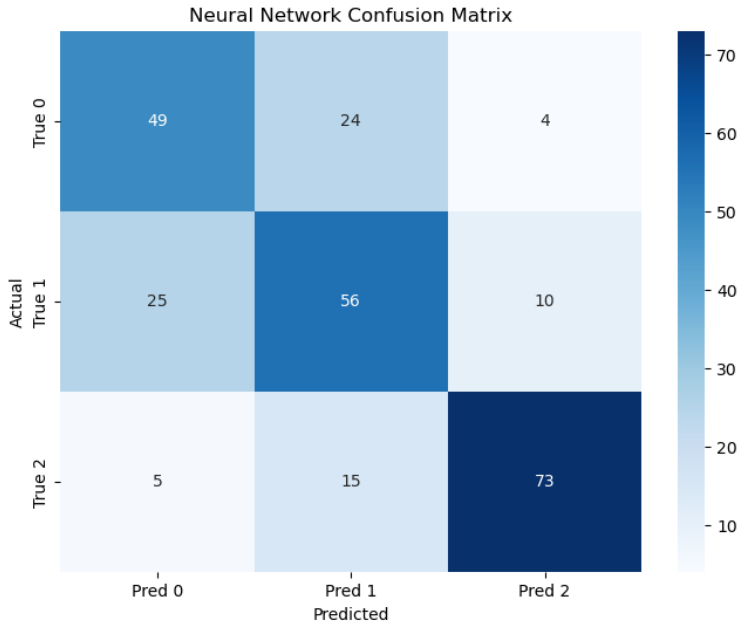

Figure 5: Confusion matrix for the Neural Network model.

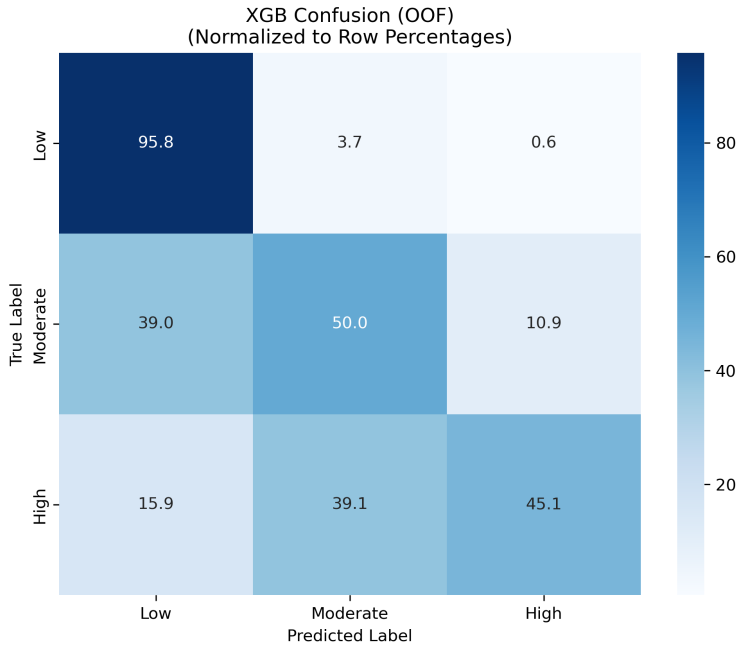

Figure 6: Confusion matrix for the XGBoost model.

3.8 App Deployment

Figure 7 shows the deployed web application that delivers daily predictions and interactive maps derived from kriging surfaces.

4 Discussion

4.1 Model Performance and Machine Learning Approaches

The results demonstrate that the Neural Network approach provides substantial improvements over alternative machine learning methods for pollen concentration classification. The 68.2% *accuracy* achieved by the Neural Network represents significant advances over other approaches, with macro-averaged F1 of 0.680 (Table 1), demonstrating clear predictive capability that justifies operational deployment.

4.2 Feature Contributions and Environmental Controls

The critical importance of temporal lag features at multiple time scales demonstrates that pollen prediction benefits significantly from incorporating historical concentration data and environmental conditions. The existence of 1-week and 1-month autoregressive features suggests that biological persistence effects and short-term environmental memory dominate over immediate meteorological forcing.

Among environmental variables, the importance of lagged temperature extremes, shortwave radiation, and day length reflects established relationships between climatic conditions

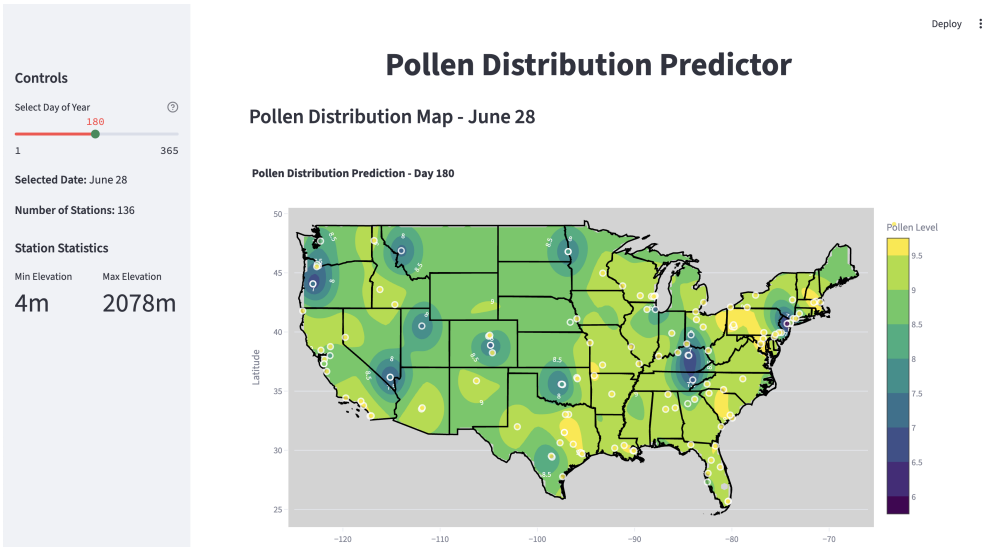

Figure 7: Web application interface for interactive pollen concentration predictions across the continental United States using the Neural Network model.

and pollen production [21]. The 3-month lag features capture seasonal conditioning effects, where environmental conditions during late winter and early spring influence flowering intensity and timing. Solar radiation and day length variables likely provide information about photoperiodic flowering cues that control reproductive phenology in *Acer* species.

4.3 Spatial Interpolation and Geographic Patterns

The implementation of Ordinary Kriging for spatial interpolation enables the transformation of discrete station-level predictions into continuous surfaces suitable for public health applications. This methodology is particularly useful for regions with sparse monitoring networks, where interpolated predictions can fill critical gaps in observational coverage.

The geographic patterns revealed through spatial visualization demonstrate clear regional differences in pollen concentrations, likely reflecting variations in vegetation composition, climate regimes, and topographic influences. These patterns are consistent with known biogeographic distributions of *Acer* species and regional climatic gradients across the continental United States [12].

4.4 Model Limitations and Sources of Uncertainty

Several limitations affect the generalizability and accuracy of the modeling approach. First, the discrete classification of pollen concentrations, while suitable for health advisory purposes, may obscure subtle differences in exposure levels. Second, the temporal resolution of daily predictions may be insufficient for capturing intraday variations, which can fluctuate significantly due to diurnal patterns [29].

The reliance on gridded environmental data introduces spatial representativeness issues, as interpolated values may not adequately characterize heterogeneous microclimates. Ad-

ditionally, the autoregressive features, while predictively valuable, create dependencies on historical observations that may limit the model's ability to predict unprecedented events.

4.5 Implications for Public Health Applications

The multiclass classification approach aligns with existing health advisory frameworks, enabling integration with established warning systems and clinical decision-making processes [11]. The web-based interface (Figure 7) democratizes access to predictions, enabling real-time decision-making by healthcare providers, patients, and public health officials.

4.6 Future Research Directions

Future work could incorporate satellite-derived vegetation indices (e.g., NDVI) [41], higher-resolution meteorology, species-specific models, and land cover data [26]. Deep learning architectures that natively capture temporal dependencies (RNNs, LSTMs, transformers) [23] and attention mechanisms may further improve skill. Ensemble forecasting and data assimilation with automated real-time pollen monitors could bolster short-term accuracy [6].

5 Conclusion

This study developed and evaluated a spatiotemporal machine learning framework for predicting *Acer* pollen concentrations using comprehensive environmental variables and autoregressive features at multiple temporal lags. Deep learning improved upon tree-based ensembles and logistic regression, and lag features delivered better performance over environmental-only baselines. Ordinary Kriging produced spatially continuous surfaces suitable for public health applications, and a Streamlit application enabled interactive exploration of daily predictions.

As climate change continues to influence pollen production and seasonal timing, such approaches will become more important for managing respiratory health risks in sensitive populations [3]. The integration of several high-level machine learning models, comprehensive environmental data, multi-scale temporal features, and user-friendly interfaces represents a promising direction for translating research advances into practical public health tools.

References

[1] KF Adams, HA Hyde, and DA Williams. Seasonal variation in tree pollen concentrations in detroit, michigan, usa. *Aerobiologia*, 29(3):365–375, 2013.

[2] American Academy of Allergy Asthma and Immunology. National allergy bureau. `https://www.aaaai.org/global/nab-pollen-counts`, 2023.

[3] Paul J Beggs. Impacts of climate change on aeroallergens: past and future. *Clinical and Experimental Allergy*, 34(10):1507–1513, 2004.

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2):281–305, 2012.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Jeroen TM Buters, Célia Antunes, Ana Galveias, Karl Christian Bergmann, Michel Thibaudon, Carmen Galán, Carsten Schmidt-Weber, and Jose Oteros. Release of bet v 1 from birch pollen from 5 european countries. results from the hialine study. *Atmospheric Environment*, 200:329–335, 2019.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[8] J Cornelis, LA De Weger, A Vermeulen, and AJH Van Vliet. Pollen forecasting: A review of methods and models. *Aerobiologia*, 35(2):207–228, 2019.

[9] Noel AC Cressie. *Statistics for spatial data*. John Wiley & Sons, 1993.

[10] G D'Amato, L Cecchi, S Bonini, C Nunes, I Annesi-Maesano, H Behrendt, G Liccardi, T Popov, and P Van Cauwenberge. Tree pollen allergy. *Allergy*, 62(9):1037–1063, 2007.

[11] Gennaro d'Amato, Lorenzo Cecchi, Stefano Bonini, Carlos Nunes, Isabella Annesi-Maesano, Heidrun Behrendt, Gennaro Liccardi, Todor Popov, and Philippe van Cauwenberge. Forecasting airborne pollen concentrations: development of local models. *Annals of Allergy, Asthma & Immunology*, 103(5):360–364, 2009.

[12] Thomas S Elias. *The complete trees of North America: field guide and natural history*. Van Nostrand Reinhold, 1980.

[13] J. Emberlin, M. Detandt, R. Gehrig, S. Jaeger, N. Nolard, and A. Rantio-Lehtimäki. Responses in the start of betula (birch) pollen seasons to recent changes in spring temperatures across europe. *International Journal of Biometeorology*, 46(4):159–170, 2002.

[14] H García-Mozo, JA Oteros, and C Galán. Phenological models to predict the main flowering phases of olive (olea europaea l.) along a latitudinal and altitudinal gradient. *International Journal of Biometeorology*, 61(4):629–639, 2017.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[16] M Grundström, Å Dahl, L Tang, M Hallquist, C Andersson, PE Karlsson, F Gilliland, G Pershagen, K Eneroth, C Almqvist, et al. Oak pollen seasonality and severity across europe and modelling the season start using a generalized phenology model. *Science of the Total Environment*, 663:527–536, 2019.

[17] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 3rd edition, 2013.

[18] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2nd edition, 2018.

[19] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[20] V Jato, FJ Rodríguez-Rajo, P Alcázar, P De Nuntiis, C Galán, and P Mandrioli. Airborne pollen content in the atmosphere of vigo (nw spain): aerobiological and meteorological analysis. *Aerobiologia*, 22(1):11–22, 2006.

[21] A.M. Jones and R.M. Harrison. The impact of climate change on allergenic pollen production. *Environmental Health Perspectives*, 112(9):1001–1009, 2004.

[22] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

[24] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[25] J Nowosad, J Verbesselt, J Truckenbrodt, and C Schmullius. Spatio-temporal models for predicting high pollen concentration events. *Computers & Geosciences*, 118:15–26, 2018.

[26] Jose Oteros, Herminia García-Mozo, Purificación Alcázar, Jordina Belmonte, Delia Bermejo, Margherita Boi, Paloma Cariñanos, Carmen Díaz de la Guardia, Federico Fernández-González, Francisco González-Minero, et al. Quality control in bio-monitoring networks, spanish aerobiology network. *Science of The Total Environment*, 443:559–565, 2013.

[27] G.E. Packe and J.G. Ayres. Seasonal variation of airborne fungal spore concentrations and size distributions in the mediterranean area. *Clinical and Experimental Allergy*, 42 (7):1142–1151, 2012.

[28] P Rapiejko, A Lipiec, A Wojdas, and D Jurkiewicz. Threshold pollen concentration necessary to evoke allergic symptoms. *International Review of Allergology and Clinical Immunology*, 13(3):91–94, 2007.

[29] N. Rathnayake, A.J. Lowe, C. Keil, and J.M. Davies. Diurnal variation of airborne pollen concentrations in melbourne: The influence of temperature, relative humidity and wind speed. *Aerobiologia*, 33(1):97–106, 2017.

[30] O Ritenberga, M Sofiev, P Siljamo, L Kalnina, and C Vitalba. Forecasting birch pollen seasons using meteorological data: a case study in riga, latvia. *Trees*, 32(4):1083–1099, 2018.

[31] FJ Rodríguez-Rajo, RM Valencia-Barrera, AM Vega-Maray, FJ Suárez, D Fernández-González, and V Jato. Analysis of the influence of wind direction on pollen concentrations in córdoba, south-west spain. *Grana*, 42(3):136–142, 2003.

[32] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 4th edition, 2017.

[33] JD Silver, MF Sutherland, FH Johnston, ER Lampugnani, MA McCarthy, SJ Jacobs, AB Pezza, and EJ Newbigin. The association between pollen counts and emergency department visits for asthma and wheeze in new york city, 2001–2011. *Environmental Pollution*, 234:441–449, 2018.

[34] M Smith, J Emberlin, A Stach, A Rantio-Lehtimäki, E Caulton, M Thibaudon, C Sindt, S Jäger, R Gehrig, and G Frenguelli. The influence of weather conditions on grass pollen concentrations in uk cities. *Aerobiologia*, 24(3):131–144, 2008.

[35] M Sofiev, P Siljamo, H Ranta, and A Rantio-Lehtimäki. Towards the operational forecasting of pollen concentrations. *International Journal of Biometeorology*, 50(6): 371–382, 2006.

[36] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[37] Yiluan Song, Daniel S. W. Katz, Zhe Zhu, and Kai Zhu. Predicting reproductive phenology of wind-pollinated trees via planetscope time series. *Science of Remote Sensing*, 15:100123, 2025. doi: 10.1016/j.srs.2025.100123.

[38] Streamlit Inc. Streamlit: The fastest way to build and share data apps. `https://streamlit.io`, 2023.

[39] Richard Webster and Margaret A Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, 2nd edition, 2007.

[40] R Zhang, T Duhl, MT Salam, JM House, RC Flagan, EL Avol, FD Gilliland, A Guenther, SH Chung, BK Lamb, et al. Long-range transport of pollen across continental boundaries. *Nature Geoscience*, 7(1):35–39, 2014.

[41] Xiaoyang Zhang, Mark A Friedl, Crystal B Schaaf, Alan H Strahler, John CF Hodges, Feng Gao, Bradley C Reed, and Alfredo Huete. Monitoring vegetation phenology using modis. *Remote sensing of environment*, 84(3):471–475, 2003.