



## OPEN Investigating environmental determinants and spatiotemporal dynamics of highly pathogenic avian influenza H5N1 outbreaks in India through machine learning

Suresh Kuralayanapalya Puttahonnappa<sup>1✉</sup>, Jayashree Anandakumar<sup>1</sup>, Nagendra Nath Barman<sup>2</sup>, Raaga Rajkumar<sup>1</sup>, Krishnamoorthy Paramanandham<sup>1</sup>, Sharanagouda S. Patil<sup>1</sup>, Sanjay Lamba<sup>3</sup>, Archana Veeranagouda Patil<sup>1</sup> & Baldev Raj Gulati<sup>1</sup>

Avian Influenza (AI), caused by highly pathogenic strains of influenza viruses, poses a significant threat to poultry populations and public health worldwide. This study offers a comprehensive evaluation of the spatial and temporal dynamics of HPAI outbreaks in India, employing a multidisciplinary approach that integrates geospatial analysis, machine learning modelling, remote sensing, and environmental risk factor assessment. The resulting incidence map provides a powerful visual representation, offering valuable insights into the distribution and concentration of HPAI (Highly Pathogenic Avian Influenza) outbreaks across the country. The study identifies a peak in HPAI outbreaks during the winter and spring seasons. Critical environmental variables such as air temperature, enhanced vegetation index (EVI), leaf area index (LAI), potential evapotranspiration (PET), rain precipitation rate, specific humidity, and wind speed are identified as significant predictors of HPAI risk. Ensemble technology was adopted by integrating high-performing random forest (RF) and classification tree (CT) models for HPAI risk assessment. The risk map generated and estimated basic reproduction numbers ( $R_0$ ) indicate, the southern and north-eastern regions of India are vulnerable to HPAI. The findings presented provide a holistic perspective essential for effective surveillance, strategic planning for resource allocation and policy development for disease management, aimed at safeguarding both avian and human populations from the looming threat of influenza outbreaks.

**Keywords** Highly pathogenic avian influenza, Outbreak, Environmental and remote sensing parameters, Spatiotemporal analysis, Risk map, Basic reproduction number

Highly pathogenic avian influenza (HPAI) viruses, specifically the H5N1 and H5N8 strains, pose substantial zoonotic threats with significant implications for public health and the global economy<sup>1–4</sup>. These viruses have a demonstrated capacity to cause severe morbidity and mortality in avian populations, leading to disruptions in poultry industries worldwide. Moreover, their ability to infect humans through close contact with infected poultry underlines their pandemic potential, with human cases exhibiting respiratory illnesses ranging from mild to severe, often with fatal outcomes<sup>5</sup>.

The H5N1 virus, first identified in Guangdong Province, China, in 1996, marked a turning point in avian influenza epidemiology. Its subsequent transmission to humans through live poultry markets in Hong Kong in 1997<sup>6</sup> emphasized the zoonotic potential of HPAI viruses. Initially confined to Southeast Asia, the virus underwent genetic diversification, resulting in multiple clades that facilitated its spread to Europe, India, Africa, and the Middle East by 2005<sup>7</sup>. India's first H5N1 outbreak in January 2006 in Maharashtra's Nawapur Sub-district<sup>8</sup> highlighted the country's vulnerability, with subsequent outbreaks documented across various states. Migratory waterfowl, especially those along the Central Asian migratory flyway, play a pivotal role in HPAI transmission,

<sup>1</sup>ICAR-National Institute of Veterinary Epidemiology & Disease Informatics, 4HG6+5QW, Ramagondanahalli, Unnamed Road, Yelahanka, Bengaluru, Karnataka 560064, India. <sup>2</sup>Veterinary Microbiology, College of Veterinary Science, Assam Agricultural University, Khanapara, Guwahati, Assam, India. <sup>3</sup>Tata Institute for Genetics and Society, Bengaluru, Karnataka, India. ✉email: suresh.kp@icar.org.in

linking India's wetlands to global spread dynamics<sup>9–11</sup>. These birds, congregating in India's wetlands during winter, facilitate cross-species transmission to poultry, with seasonal outbreaks peaking between October and March<sup>12</sup>.

Despite advancements in surveillance and control measures, the recurring outbreaks of HPAI highlight the need for a deeper understanding of the environmental and spatiotemporal factors influencing its transmission. Research has consistently demonstrated the role of environmental determinants such as temperature, precipitation, humidity, and proximity to water bodies in the spread and persistence of HPAI viruses<sup>13–15</sup>. While spatiotemporal analyses have provided useful insights into outbreak dynamics, existing studies often face limitations, including short observation periods, geographically restricted data, and a narrow focus on selected environmental variables. Previous studies, such as those on the spatial aggregation of H5N1 outbreaks in Bangladesh<sup>16</sup>, landscape suitability and HPAI outbreaks in India<sup>17</sup>, and the spatiotemporal distribution and seasonality of HPAI outbreaks in India from 2006 to 2021 have been valuable. Other research, such as modeling HPAI outbreaks in West Bengal, India<sup>18</sup>, and risk factor modeling of spatiotemporal HPAI patterns focusing only on anthropogenic variables<sup>19</sup> also contributed to understanding outbreak patterns. Studies on HPAI spread in France<sup>20</sup> and temperature-driven transmission patterns across Eurasia<sup>2,21</sup> further illustrate regional patterns. Similarly, studies in Vietnam during 2003–2007<sup>22</sup>, China in 2013–2014<sup>23</sup>, Southern China in Early 2004<sup>24</sup>, and Vietnam, 2015–2018<sup>25</sup> have provided key insights into the spatiotemporal spread of the virus. Despite these valuable contributions, many of these studies lack a comprehensive integration of diverse environmental datasets, often focusing only on spatiotemporal analysis with limited data and traditional methodologies. This limits the thorough understanding of the complex factors driving HPAI outbreaks. Traditional methods provide foundational insights but struggle with the scale and complexity of available data. Machine learning (ML) techniques, however, can process large, diverse datasets, uncover hidden patterns, and make accurate predictions. Recent ML advancements in infectious disease research highlight their potential for identifying environmental determinants and predicting outbreak risks<sup>26</sup>. The present study, HPAI outbreaks spanning from 2006 to 2024 in India, aims to address these gaps by incorporating meteorological, remote sensing, geographical and anthropogenic parameters into a risk modeling framework, utilizing machine learning approaches for identifying vulnerabilities, which offers a visualized and comprehensive view of likelihood and impact of disease in a given region, helps the policy makers and planners to improve risk management and risk governance by prioritizing risk management efforts.

This interdisciplinary framework represents a significant advancement over traditional approaches by addressing the limitations of modelling efforts and regional constraints, by incorporating a broader spectrum of environmental variables and utilization of machine learning algorithms. By providing actionable insights into the dynamics of HPAI outbreaks, this work has the potential to inform evidence-based strategies for prevention and control, enhancing India's preparedness for future outbreaks. Furthermore, the integration of ML with spatiotemporal analysis offers a scalable model for application in other regions, thereby contributing to the global understanding and management of HPAI risks.

## Results

### Spatial distribution

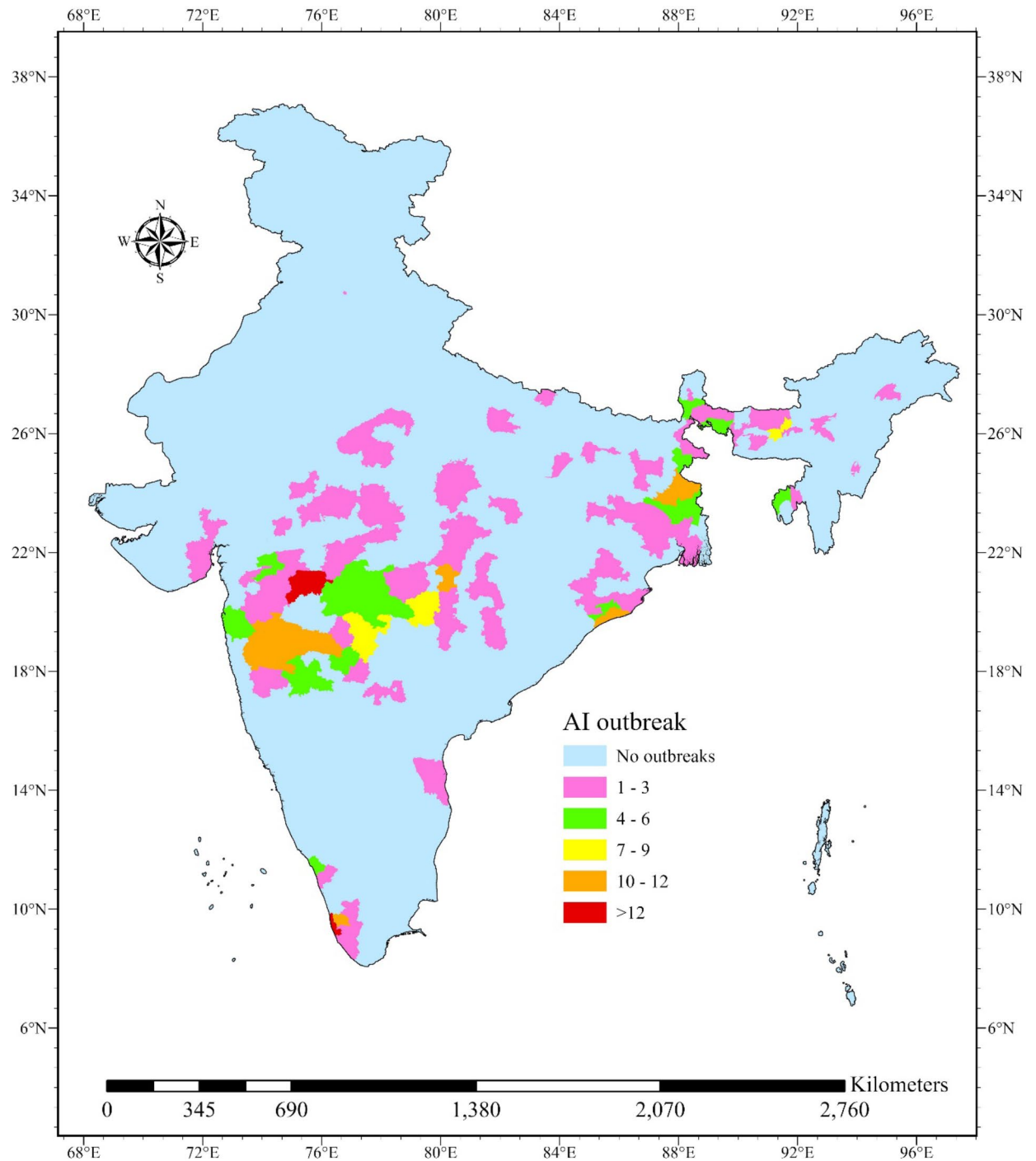
Over an 18-year period from 2006 to 2024, a total of 312 outbreaks of HPAI were reported across different parts of India. The Alappuzha district of Kerala experienced the highest number of outbreaks (39), followed by Jalgaon district in Maharashtra with 26 outbreaks. Kottayam district in Kerala and Puri district in Odisha each reported 12 outbreaks each. Additionally, 10 outbreaks were reported in the districts of Ahmednagar, Beed, Gondiya, and Pune in Maharashtra, as well as in Birbhum and Murshidabad districts of West Bengal. The majority of districts reported outbreaks between 1 and 3. Figure 1 illustrates the geographical distribution of HPAI outbreaks across India.

### Temporal distribution

The temporal analysis of HPAI outbreaks from 2006 to 2024 reveals monthly and yearly trends. A total of 312 (100%) outbreaks were recorded over 18 years, with varying distribution across months and years. Figure 2 provides a visual representation of the temporal trends in HPAI outbreaks. Monthly Distribution: Outbreaks peaked in January (32.64%) and December (13.76%), with relatively high occurrences also in March (16.96%) and February (11.52%). The lowest activity was observed during the monsoon months, particularly in July (0.96%) and August (0.96%), indicating seasonal patterns in outbreak occurrences. Yearly Distribution: The highest percentage of outbreaks occurred in 2021 (23.04%), followed by 2008 (22.72%) and 2006 (9.28%). Conversely, the years 2007, 2010, and 2023 reported the least outbreaks, with only 0.32 to 1.6% of cases recorded. Enhanced surveillance during high-risk months, pre-monsoon vaccination campaigns, strategic resource allocation, strengthening biosecurity, monitoring migratory bird activity educating stakeholders can significantly mitigate HPAI risks.

### Spatial auto-correlation

A total of 46 districts spanning Jharkhand (3 districts), Kerala (5 districts), Madhya Pradesh (3 districts), Maharashtra (23 districts), Odisha (4 districts), Telangana (1 district), and West Bengal (7 districts) were recognized as HPAI hotspots (Fig. 3A, Table S1) based on positive Z-scores determined using the Getis–Ord  $G_i^*$  index. A hotspot identified through positive Z-scores in the Getis–Ord  $G_i^*$  index indicates a statistically significant clustering of high disease cases in a specific area. This helps pinpoint regions with increased transmission risk, allowing for targeted surveillance and control measures. The identification of HPAI hotspots has critical public health implications, necessitating targeted interventions to mitigate the risks. Enhanced monitoring enables



**Fig. 1.** Spatial distribution of HPAI outbreaks in India, generated using ArcGIS Pro 3.2.1 (<https://www.esri.com>).

early detection and response, reducing human exposure and zoonotic transmission risks. Strategic resource allocation, including vaccines and diagnostics, ensures effective outbreak management in high-risk areas.

#### Space-time cluster analysis

A total of 11 clusters, encompassing 210 districts, were identified during the period from 2006 to 2024 (Fig. 3B; Table S2). Of these, 7 clusters (Clusters 1, 3–7, and 10) were classified as high-incidence clusters, covering 184 districts with a high relative risk of HPAI ranging from 2.84 to 41.77. The remaining 4 clusters (Clusters 2, 8, 9, and 11), comprising 26 districts, were categorized as low-incidence clusters with a relative risk ranging from 0.04 to 0.38.

0.64	0	3.52	5.12	0	0	0	0	0	0	0	0	9.28	2006
0	0	0	0	0	0	0.32	0	0	0	0	0	0.32	2007
13.76	0	1.6	0.96	0.64	0	0	0	0	0	1.6	4.16	22.72	2008
1.6	0.64	0.64	0	0.32	0	0	0	0	0	0	0	3.2	2009
0.64	0	0	0	0	0	0	0	0	0	0	0	0.64	2010
0	0	0	0	0	0	0	0.32	0.32	0	0	0	0.64	2011
0.64	0.32	0.32	0	0	0	0	0	0	0	0	0	1.28	2012
0	0	0	0	0	0	0.64	0	0	0	0	0	0.64	2013
0.32	0	0	0	0	0	0	0	0	0	1.6	0.32	2.24	2014
0.32	0	0.96	0.64	0	0	0	0	0	0	0	0	1.92	2015
0.32	0	0	0	0.32	0	0	0	0	0	0	0.64	1.28	2016
0.96	0	0	0	0	0	0	0	0	0	0	0	0.96	2017
0	0.32	0.32	0	0	0	0	0	0	0	0.32	4.16	5.12	2018
2.56	0.96	0.64	0	0	0	0	0	0	0	0	0.32	4.48	2019
0.32	0	3.2	0	0	0	0	0	0	0	0	1.6	5.12	2020
8.96	8.32	5.76	0	0	0	0	0	0	0	0	0	23.04	2021
0	0.64	0	0	0	0	0	0	0	0.64	2.56	2.56	6.4	2022
1.6	0	0	0	0	0	0	0	0	0	0	0	1.6	2023
0	0.32	0	0	2.56	4.16	0	0.64	1.28	0	0	0	8.96	2024
32.64	11.52	16.96	6.72	3.84	4.16	0.96	0.96	1.6	0.64	6.08	13.76	100	Total
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	

Fig. 2. Temporal distribution of outbreaks (in percentage) of HPAI in India.

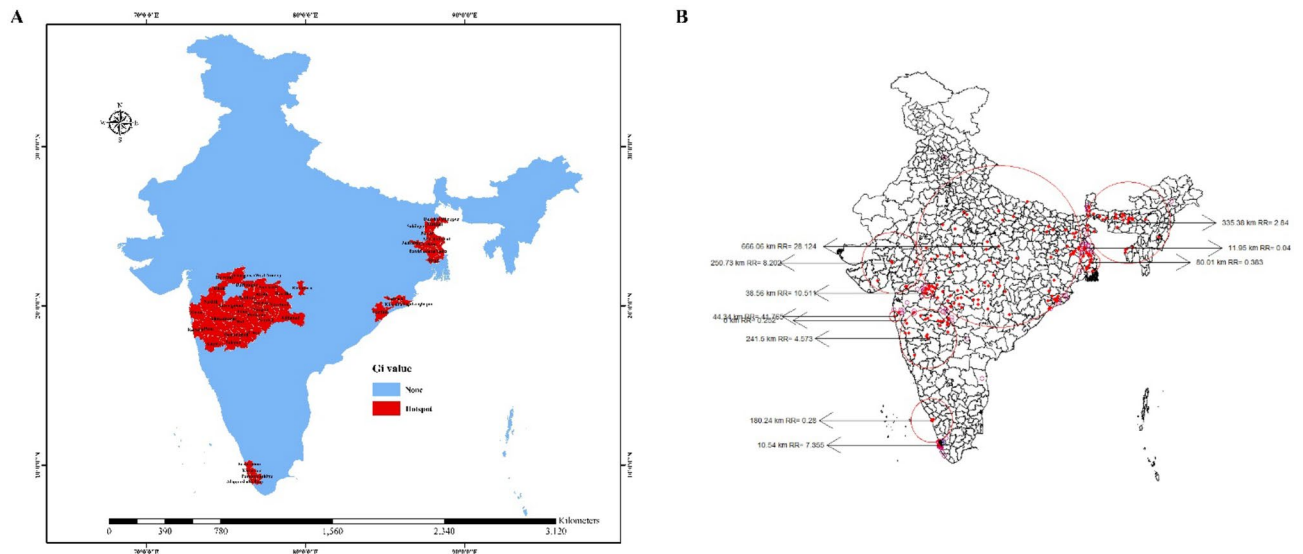


Fig. 3. (A) Hot spot map of HPAI in India identified through spatial autocorrelation, generated using ArcGIS Pro 3.2.1 (<https://www.esri.com>). (B) Spatial-temporal clusters of HPAI incidences in India, generated using SaTScan v10.1.2 (<https://www.satscan.org>) and R statistical software version 4.3 (<https://www.r-project.org>).

### Environmental risk factors identified through linear discriminant analysis (LDA)

The study identified the crucial environmental risk factors that play a significant role in the HPAI disease outbreak. Environmental factors demonstrating a p-value of 0.05 or less were recognized as strongly correlated with HPAI. According to LDA results, air temperature, enhanced vegetation index (EVI), leaf area index (LAI), potential evapotranspiration (PET), potential evaporation rate, rain precipitation rate, specific humidity, and wind speed emerged as the variables exhibiting the strongest associations with HPAI occurrence (Table 1). Consequently, these variables were considered for subsequent HPAI risk modelling and mapping.

Parameters	Mean	SD	F-value	p-value	95% CI
Air temperature	24.34	3.68	14.31	$1.82 \times 10^{-4*}$	23.96–24.72
Potential evaporation rate	212.82	81.84	14.88	$1.36 \times 10^{-4*}$	204.38–221.26
Rainfall precipitation rate	1.84	3.07	43.45	$1.60 \times 10^{-10*}$	1.52–2.16
Specific humidity	0.01	0.00	5.30	$2.19 \times 10^{-2*}$	0.009–0.01
Surface pressure	98403.93	2895.99	0.40	0.53	98105.43–98702.43
Wind speed	2.31	0.76	6.53	$1.10 \times 10^{-2*}$	2.23–2.39
Soil moisture	23.77	5.18	0.88	0.35	23.24–24.3
LST	29.18	6.11	0.56	0.46	28.55–29.81
NDVI	0.47	0.16	0.09	0.76	0.45–0.49
EVI	0.26	0.10	12.67	$4.23 \times 10^{-4*}$	0.25–0.27
PET	759.1	1237.48	10.91	$1.05 \times 10^{-3*}$	632.05–887.15
LAI	0.38	0.77	16.18	$7.07 \times 10^{-5*}$	0.29–0.45
Water bodies	2288.98	2979.58	1.46	0.23	1981.86–2596.1

**Table 1.** Critical environmental risk factors of HPAI identified using LDA. *SD* standard deviation, *CI* confidence interval. \*Significant at 5 per cent level of significance.

### Risk assessment and Estimation

The study employed climate-disease modelling based on the significant environmental risk factors identified through LDA. The distribution pattern of HPAI -affected (case) and unaffected (control) regions was visually represented in Figure S1. The case-control data underwent fitting to integrated machine learning models, and the best-fitted models were discerned based on a comprehensive set of statistical assessment criteria. The risk maps were generated utilizing machine learning modelling, with random forest (RF) and classification tree (CT) models emerging as the most effective models for simulating spatial risk based on disease data and relevant prediction factors, meeting the defined evaluation criteria as detailed in methodology. The RF model proved superior performance with the cohen's kappa (*Kappa*) (0.68), receiving operating characteristic (ROC) (1.00), true skill statistics (TSS) (0.96), area under the ROC curve (AUC) (1.00), accuracy (0.98), F1 score (0.99), error rate (0.02), and logistic loss (LOGLOSS) (0.12), followed by CT model with *Kappa* (0.66), ROC (0.93), TSS (0.74), AUC (0.93), accuracy (0.85), F1 score (0.93), error rate (0.15), and LOGLOSS (0.34). The models, RF and CT, were selected for their ability to handle complex, non-linear relationships and provide high predictive accuracy. RF is well-suited for large datasets with many variables, while CT offers interpretability and decision rules. Together, these models complement each other by balancing accuracy and interpretability, enhancing the overall prediction of HPAI outbreaks. To enhance confidence in the predictive outcomes, an innovative ensemble approach was employed by averaging the scores of the RF and CT models (Table 2). This technique refined the prediction accuracy, enabling the delineation of HPAI risk within the study area. The combined prediction outcomes of ensemble models were used in the current study, which are in the scale of 0 (low risk) to 1 (high risk) and the resulting risk map clearly indicates that HPAI is most likely to spread in the southern and north-eastern regions of India (Fig. 4A).

### Transmission dynamics

Following the risk mapping, the basic reproduction number ( $R_0$ ) was computed and subsequently overlaid onto the risk map.  $R_0$  values exceeding 1.00 signify regions or districts experiencing an upward trend in disease incidence, while values below 1.00 indicate a decline. The  $R_0$  values depicted in Fig. 4B ranged from 0.75 to 3.30, indicating a higher likelihood of HPAI outbreaks in districts situated in the north-eastern and southern regions. Significantly, 52 districts exhibited  $R_0$  values surpassing 1, with several in Assam, Maharashtra, and Kerala. Specifically, districts Ahmednagar of Maharashtra (3.30), followed by Bongaigaon, Dibrugarh, Kamrup Metropolitan, Nagaon, Baksa, and Chirang in Assam, as well as Nayagarh in Odisha, alongside Chandigarh, recorded  $R_0$  values 2.08, indicating a notably elevated risk of HPAI incidence in these specific areas (Table S3). The estimated vaccination coverage for these districts ranges from 10.55 to 69.74 per cent, indicating the critical insight derived from the  $R_0$  assessment underscores the importance of targeted intervention strategies in these high-risk districts to mitigate the potential spread of the disease. To understand the importance of wetland bodies in HPAI disease occurrence risk map was superimposed on wetland locations (Fig. 4C). It is evident that the regions near wetland locations associated with suitable environmental variables were identified as high-risk regions.

### Discussion

Highly Pathogenic Avian Influenza outbreaks in India pose a significant threat to both poultry industries and public health. Our findings highlight the identification of potential risk zones, seasonal patterns, and environmental risk factors that influence the occurrence and transmission of HPAI, providing valuable insights for future efforts to mitigate the spread of HPAI in India. It is evident from the Fig. 1 that the reported outbreaks were more in Kerala, Maharashtra, and north-eastern states of India that aligns with findings from studies in Southeast Asia and Europe, which similarly identified hotspots in areas with high poultry density, wetland proximity, and substantial human-animal interfaces with trade activities<sup>27–29</sup>. Kerala and north eastern states are

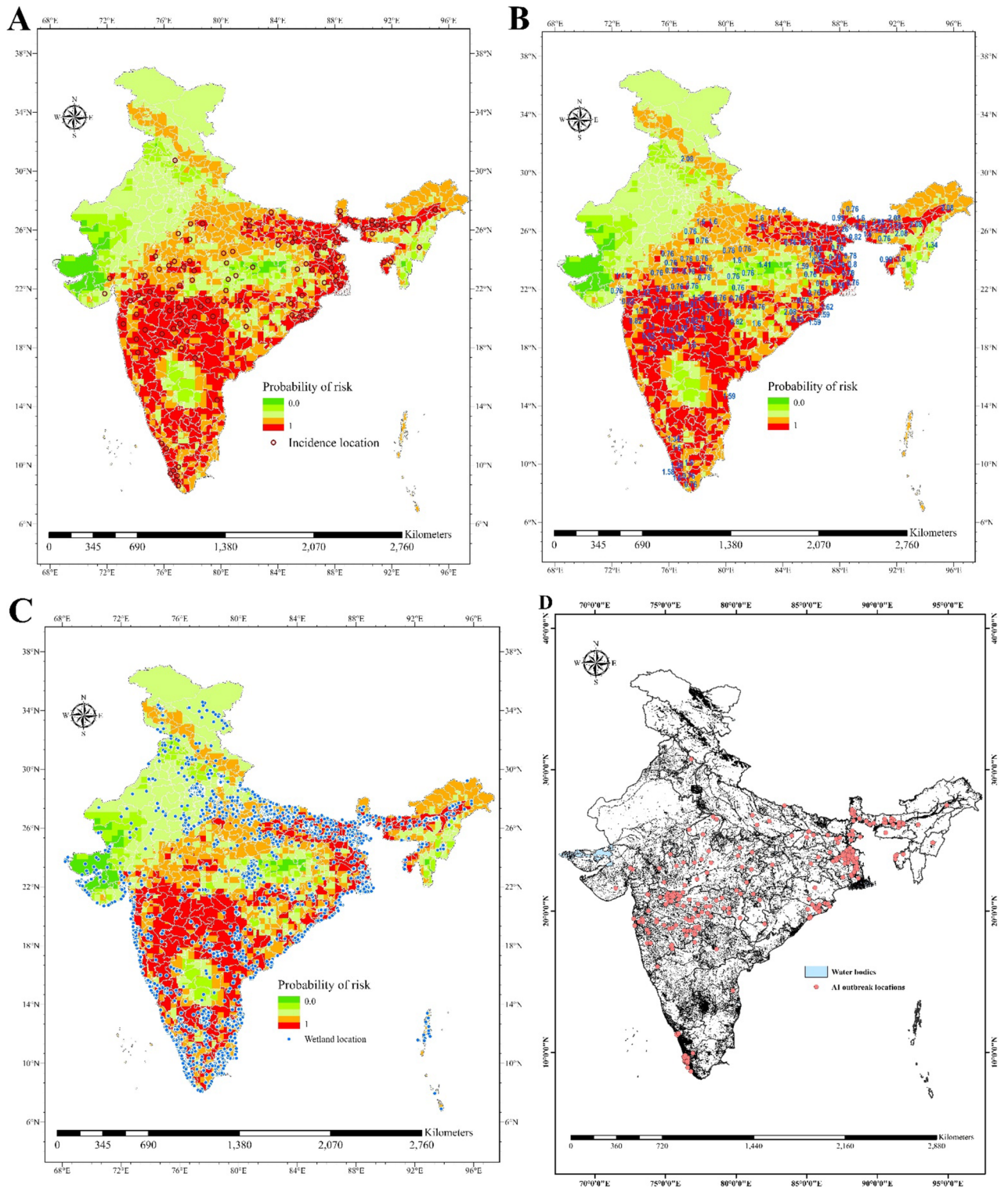


Sl. No	Models	Model Specification	KAPPA	ROC	TSS	AUC	Accuracy	ERROR RATE	F1 SCORE	LOG LOSS
1.	GLM	$E(Y X) = \mu = g^{-1}(X\beta)$ Y- Expected Value, X-Conditional, $X\beta$ - Linear Predictor, g-Link Function	0.34	0.80	0.49	0.80	0.75	0.25	0.89	0.43
2.	GAM	$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_i(x_i)$ Y-Response Variable, g-Link Function, $f_i$ -Specified Parametric Form, $x_i$ -Predictor Variable	0.34	0.80	0.49	0.80	0.75	0.25	0.89	0.43
3.	RF	$Y = \sum_{i=1}^n f(t_n)$ Y- Average of aggregated predictions of the multiple decision trees, $t_n$ - multiple decision trees trained on different subset of the same training data	0.68	1.00	0.96	1.00	0.98	0.02	0.99	0.12
4.	GBM	$f(x) = \arg \min_{\theta} \sum_{i=1}^n L(y_i, \theta) + \sum_{m=1}^M \eta \rho_m \varphi_m(x)$ m- Iteration, $\eta$ -Learning Rate, $\rho_m$ - Step length	0.50	0.93	0.72	0.93	0.84	0.16	0.92	0.31
5.	NNET	$Y = f(\sum_{i=1}^n x_i w_i) + b$ Y-Output, $x_i$ -Inputs, $w_i$ -Weights, b- Bias	0.00	0.50	0.00	0.50	0.25	0.75	0.00	25.90
6.	MARS	$\hat{f}(x) = \sum_{i=1}^k c_i B_i(x)$ $c_i$ - Constant Coefficient, $B_i(x)$ - Basis Function	0.41	0.87	0.60	0.87	0.80	0.21	0.91	0.34
7.	FDA	$\eta_l(x) = X^T \beta_l$	-0.01	0.50	-0.01	0.50	0.79	0.21	0.88	7.28
8.	CT	$f(x) = \sum_{j=1}^T w_j I(x \in R_j)$	0.66	0.93	0.74	0.93	0.85	0.15	0.93	0.34
9.	SVM	$\{x : f(x) = x^T \beta + \beta_0 = 0\}$	0.51	0.87	0.68	0.87	0.85	0.15	0.91	0.53
10.	NB	$P(c x) = \frac{P(x c)P(c)}{P(x)}$ $P(c x)$ -Posterior Probability $P(x c)$ -Likelihood, $P(c)$ -Class Prior Probability, $P(c)$ -Predictor Prior Probability	-0.20	0.72	-0.18	0.72	0.30	0.70	0.47	3.55
11.	ADA	$F_T(x) = \sum_{t=1}^T f_t(x)$ $f_t$ - Weak Learner, x- Input, T- $T^{\text{th}}$ Positive or Negative Classifier	0.73	0.83	0.66	0.83	0.92	0.08	0.95	2.71

**Table 2.** Evaluation metrics of machine learning models employed for risk prediction of HPAI. Generalized Linear Models (GLM), Generalized Additive Models (GAM), Random Forest (RF), Gradient Boosting Machine (GBM), Artificial Neural Network (NNET), Multiple Adaptive Regression Splines (MARS), Flexible Discriminant Analysis (FDA), Classification Tree Analysis (CT), Support Vector Machine (SVM), Naive Bayes (NB), Adaptive Boosting (ADA), Receiving Operating Characteristic (ROC) curve, True Skill Statistics (TSS), Area Under the ROC Curve (AUC), Logistic Loss (LOGLOSS).

known for its extensive wetland ecosystems and migratory bird activity, provides ecological niches conducive to viral maintenance and spread. Moreover, Maharashtra's prominence in poultry trade and farming may amplify transmission dynamics through the movement of infected birds or contaminated products. These observations are consistent with studies highlighting the role of trade networks in the spatial clustering of HPAI<sup>30</sup>. In 2016, the Food and Agriculture Organization (FAO) identified legal movements of live poultry as a crucial pathway for the spread of HPAI in the Middle East<sup>31</sup>. Research by Awada<sup>32</sup> identified two global major peaks of HPAI in 2006 and 2016 due to sudden surge in the annual average trade of live poultry.

We identified a distinct yearly seasonal trend, with higher instances in 2006, 2008, and in 2021 and with the peak of HPAI occurring between November and March (Fig. 2). The seasonal variations in outbreaks were closely tied to bird migration patterns, and weather conditions. The timing of these outbreaks often aligned with peak bird migration seasons, underscoring the need for heightened vigilance during these times. The pronounced winter peak in outbreaks, particularly in January, is consistent with global patterns where low temperatures and humidity enhance the environmental stability of the influenza virus<sup>33,34</sup>. The minimal activity during monsoon months may reflect unfavourable climatic conditions for virus survival, such as higher temperatures and rainfall leading to dilution of viral particles in aquatic systems. Similar to our findings, earlier research has shown that Northern Temperate or Subtropical countries, including Bangladesh, China, Egypt, and Turkey, as well as tropical nations like Cambodia, Indonesia, and Vietnam, experienced the highest H5N1 outbreak peaks between January and March<sup>35</sup>. Consistent with our findings, Egypt, an avian influenza virus (AIV)-endemic region, exhibited increased odds of positive H5N1 virus infections in live bird markets during winter<sup>36</sup>. Park and Glass<sup>37</sup> also noted a heightened risk of avian and human influenza in East and Southeast Asia during the winter months. Similarly, studies in Japan<sup>38</sup> and Korea<sup>39</sup> have reported a higher prevalence of H5N1 infections in both wild and domestic birds during winter, corroborating the seasonal trend observed in our study. Magalhaes<sup>40</sup> found that the seasonal pattern of HPAI was linked to increased poultry stock and sales during Chinese New Year celebrations, which supports our current findings. Furthermore, a study by Chen<sup>41</sup> corroborated our research by demonstrating that the peak of H5N1 outbreaks occurs in the winter and spring seasons. A significant HPAI H5N1 outbreak was identified in April 2005, affecting various wild bird species at Qinghai Lake, China<sup>42,43</sup>. Qinghai Lake serves as a critical breeding ground for migratory waterfowl within the Central Asian Flyway and the East Asia-Australasia Flyway<sup>44</sup>. The Qinghai nature reserve supports approximately 150,000 birds from over 180 species during the spring and summer seasons<sup>42</sup>. Among these, bar-headed geese (*Anser indicus*), a prominent species at this site, migrate to Southeast Asia and the Indian subcontinent over



**Fig. 4.** (A) Risk prediction map of HPAI showing variations in colour ranging from green (low risk) to red (high risk). (B) Superimposition of basic reproduction number ( $R_0$ ) on the risk map (blue). (C) Superimposition of wetland location (blue dots) on the risk map. (D) Spatial pattern map of water body along with AI outbreak location. All maps were generated using ArcGIS Pro 3.2.1 (<https://www.esri.com>).

the Himalayas in the fall (September) and return to Qinghai Lake in April for the summer<sup>45</sup>. This migratory pattern significantly contributes to the winter and summer peaks of HPAI outbreaks in India. As of December 2021, H5N1 virus outbreaks have been documented on a large scale among poultry and wild birds in more than 77 countries, signalling expanded virus circulation<sup>12,46</sup>. The first human case of H5N1 virus infection in India

was reported in June 2021, coinciding with the monsoon season in the country<sup>47</sup>. Over recent years, numerous subtypes of the HPAI viruses have been circulating in diverse bird populations on a large geographical scale. In particular in 2021, an unprecedented genetic variability of subtypes has been reported in birds, thus creating an epidemiologically challenging landscape<sup>46</sup>.

Our study examined the climatic factors influencing the survivability and persistence of HPAI in the environment and their role in facilitating disease spread. We identified significant associations between HPAI incidence rates with variables such as air temperature, enhanced vegetation index, leaf area index, potential evapotranspiration, rain precipitation rate, specific humidity, and wind speed. Similarly, previous research has demonstrated correlations between environmental parameters, including temperature, humidity, and wind speed, and the transmission of influenza viruses<sup>48</sup>. Research has shown that climatic factors such as temperature and relative humidity significantly influence HPAI virus circulation in live bird markets of Bangladesh (Islam et al., 2023). Previous studies have reported that AIV replication increases at lower temperatures (Siboannan et al., 2013), and colder conditions may prolong viral survival in secretions and faeces of infected poultry<sup>35,49</sup>. Similarly, Paek et al.<sup>50</sup> demonstrated that higher temperatures reduce the likelihood of AIV survival. In Finland, Jaakkola et al.<sup>34</sup> observed that a 1 °C drop in temperature was associated with an 11% increase in the predicted risk of influenza. Seasonal variations in influenza transmission have also been attributed to low absolute humidity, as shown in a study conducted in the continental United States<sup>51</sup>. Conversely, a study in Japan found that high humidity was linked to increased influenza incidence<sup>52</sup>. In Bangladesh, lower temperatures and humidity levels are observed between November and March compared to the rest of the year. Additionally, Chen et al.<sup>41</sup> revealed that higher wind speeds negatively affect the H5N1 epidemic by enhancing ventilation in farms, markets, and other live poultry facilities in China, thereby reducing virus circulation. On the other hand, in Nigeria, low rainfall has been associated with a higher prevalence of AIV<sup>53</sup>. Several studies have also reported wind speed, leaf area index (LAI), and specific humidity as protective factors against H5N1 outbreaks using binary logistic regression models<sup>41</sup>. The occurrence of H5N1 outbreaks is negatively associated with wind speed, as strong winds enhance ventilation in farms, markets, and other live poultry facilities<sup>34</sup>. Si et al.<sup>55</sup> using logistic regression analysis revealed that HPAI H5N1 outbreaks were strongly associated with Normalized Difference Vegetation Index (NDVI) and lower elevation, and outbreaks increased with an increasing human population density combined with proximity to lakes or wetlands.

According to our model, areas at high risk of HPAI transmission were projected to be located along the coastal areas and wetlands of India, spanning from the western coast of southern India to the north-eastern regions (Fig. 4). These regions encompass areas of paddy cultivation and traditional paddy-duck farming, which include numerous rivers, their tributaries, and lakes. Due to the mixture of semi-domestic and wild duck hosts, which are the typical carriers of AIVs, traditional rice-duck farming systems create an excellent setting for AIV spread. Additionally, as these ecosystems serve as resting sites for migratory birds, there is a significant likelihood of a high risk of AIV transmission<sup>56</sup>. The shared habitats between domestic ducks allowed to graze in water bodies with migratory birds is identified as risk factors of HPAI transmission in West Bengal, India<sup>18</sup>. Our findings align with a study conducted in China, which noted that many villages in coastal southern parts of China with duck farming systems are recognized as high-risk areas for AIV transmission. This also includes high-risk areas surrounding the Yangtze River delta, including major tributaries like Dongting Lake and Poyang Lake<sup>13,30,57</sup>. Si and Co-workers (2013) (Si et al., 2013) using logistic regression analysis revealed that HPAI H5N1 outbreaks increased with an increasing human population density combined with proximity to lakes or wetlands. The International Office of Epizootics emphasizes that regions with extensive wetlands and a high density of poultry in contact with wild birds represent significant risk areas<sup>58</sup>. Preventive efforts can be focused on such regions with measures include minimizing interactions between wild birds and domestic ducks by restricting domestic birds from foraging freely near wetland areas. Estimating the basic reproduction number ( $R_0$ ) is crucial for policymakers developing effective disease control strategies. The reported  $R_0$  range of 0.75 to 3.30 suggests the involvement of multiple primary sources in the outbreak, warranting further investigation<sup>18</sup>. Notably, the  $R_0$  value calculated in our study closely aligns with estimates from the 2006 H5N1 outbreak in Romania, where  $R_0$  ranged between 1.95 and 2.68 across 161 affected villages<sup>59</sup>. Additionally, using Exponential Growth (EG) methods,  $R_0$  estimates for subtype H5N1 were between 1.65 and 2.20, while values for subtype H5N8 ranged from 0.03 to 1.56<sup>60</sup>.

The findings from the current study underscore the importance of targeted surveillance, risk mapping, and climate-informed strategies to mitigate outbreaks effectively, particularly in high-risk regions and seasons.

## Materials and methods

As previously stated, the primary focus of this study is to investigate the geospatial dynamics of HPAI in relation to environmental variables, considering them as potential risk factors through the application of machine learning methodologies. This comprehensive framework for disease modelling is depicted in Fig. 5. Within this framework, we meticulously documented the processes of data collection, pre-processing, and feature engineering. Furthermore, we delineated the steps for conducting spatio-temporal analysis, established criteria for model implementation, and evaluated the best-fitting model. Hyper-parameter tuning was carried out to assess potential under fitting and overfitting of the models. Additionally, ensemble techniques were integrated to enhance model accuracy and sensitivity, culminating in the generation of risk maps intended for use in strengthening surveillance and disease management strategies. This framework significantly advanced our understanding of integrating modelling techniques in both analytical and predictive capacities. The modelling approach adopted for this study is thoroughly discussed in the subsequent sections and subsections.



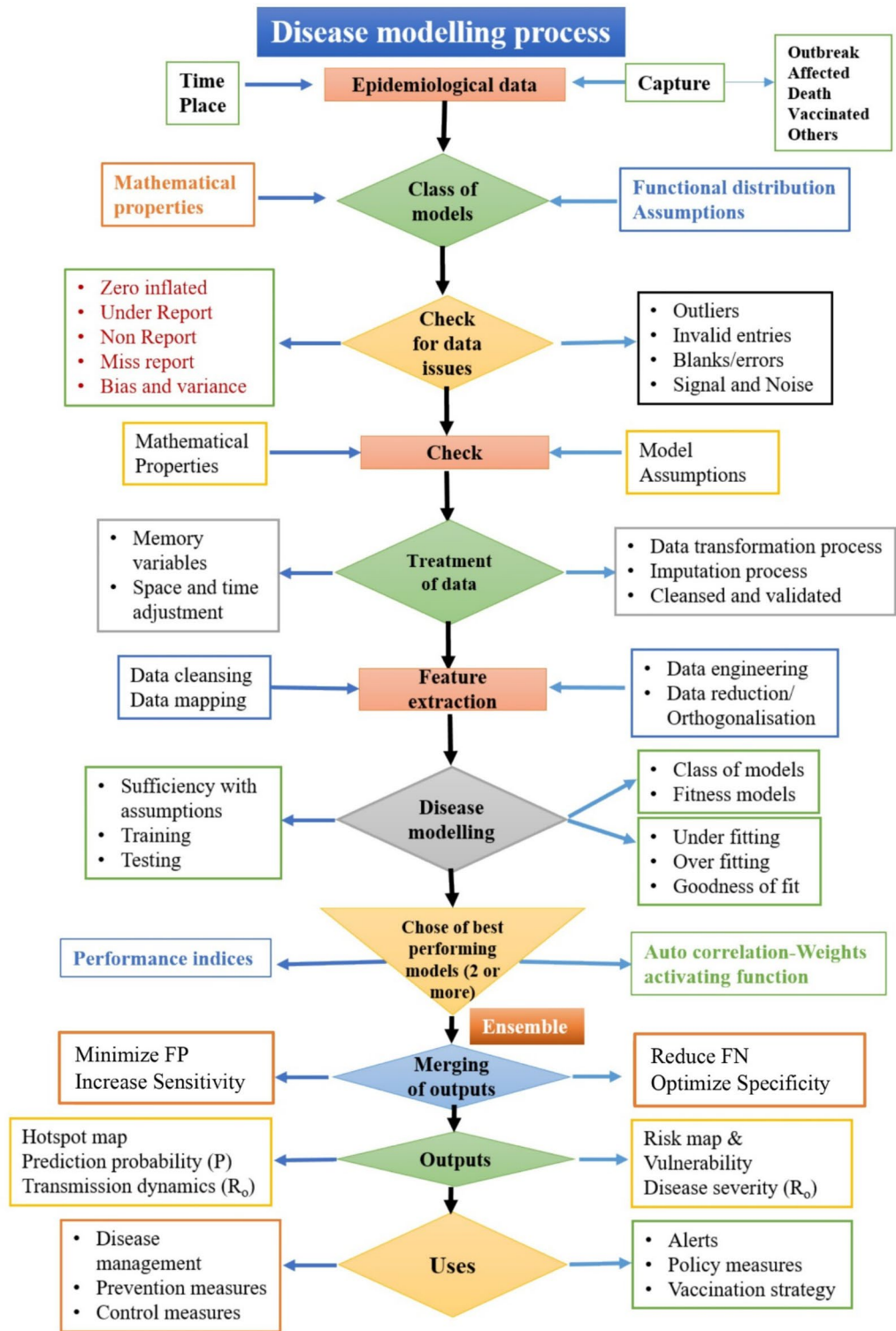


Fig. 5. A comprehensive framework employed in the current study for HPAI disease modeling.

**Study area**

India, a country in South Asia lies between 8°4' N and 37°6' N latitude and 68°7' E to 97°25' E longitude north of the equator. The country has a total area of 3,287,263 sq. km and its climate is made up of six major subtypes, ranging from arid deserts in the west to alpine tundra and glaciers in the north to tropical rainforests in the southwest and islands. The country experiences four distinct seasons: winter (January to February), summer (March to May), the monsoon-rainy (June to September), and the post-monsoon (October to December). The Himalayas serve as a protective barrier against the glacial katabatic winds that originate in Central Asia. As a

Variable	Source	Units/Range	Attribute	Resolution
Meteorological parameters				
Air temperature	GLDAS version 2 <a href="https://ldas.gsfc.nasa.gov/gldas">https://ldas.gsfc.nasa.gov/gldas</a>	k	netCDF	0.25° × 0.25°
Potential evaporation rate		W/m <sup>2</sup>	netCDF	0.25° × 0.25°
Rainfall precipitation rate		kg/m <sup>2</sup> /s	netCDF	0.25° × 0.25°
Specific humidity		kg/kg	netCDF	0.25° × 0.25°
Surface pressure		p.a.	netCDF	0.25° × 0.25°
Wind speed		m/s	netCDF	0.25° × 0.25°
Soil moisture		kg/m <sup>2</sup>	netCDF	0.25° × 0.25°
Remote sensing parameters				
LST	MODIS <a href="https://ladsweb.modaps.eosdis.nasa.gov">https://ladsweb.modaps.eosdis.nasa.gov</a>	°C	Raster	1 km × 1 km
NDVI		-1 to 1	Raster	500 m × 500 m
EVI		-1 to 1	Raster	500 m × 500 m
PET		mm	Raster	500 m × 500 m
LAI		m <sup>2</sup> /m <sup>2</sup>	Raster	500 m × 500 m
Geographical parameters				
Water bodies	OpenStreetMap Data Extracts ( <a href="https://download.geofabrik.de/asia/india.html">https://download.geofabrik.de/asia/india.html</a> )	m	Shape file	NA

**Table 3.** Environmental risk factors data and their sources used in the current study. *netCDF* network common data format, *NA* not available.

result, northern India is warm or mostly moderately cool during the winter, the same phenomena cause India to be relatively hot during the summer. Even though the Tropic of Cancer, the line separating the tropics from the subtropics, runs through the centre of India, the entire nation is regarded as tropical country<sup>61</sup>.

### Disease outbreaks and poultry population data

India possesses a substantial backyard poultry population, totalling 317.07 million, according to data from the 20th Livestock Census conducted by the Department of Animal Husbandry and Dairying (DAHD) in India. The population data were collected at the district level. Information regarding HPAI disease outbreaks in backyard poultry over a span of 18 years (2006–2024) was sourced from the DAHD, India, and The World Organisation for Animal Health (OIE). This data includes details such as the onset of disease (month & year), number of outbreaks and the precise global positioning system (GPS) coordinates of outbreak locations.

### Environmental risk factors data

This section outlines a comprehensive description of remote sensing data and meteorological data used in this study. Data on these risk factors were collected at a grid level over an 18-year period (2006–2024) and information is detailed in (Table 3) and visually represented in Figure S2.

#### Meteorological data

The meteorological variables utilized in this study were sourced from the Global Land Data Assimilation System (GLDAS version 2), available at <https://ldas.gsfc.nasa.gov/gldas><sup>62</sup>. These variables include air temperature (k), potential evaporation rate (W/m<sup>2</sup>), rainfall precipitation rate (kg/m<sup>2</sup>/s), specific humidity (kg/kg), surface pressure (Pa), and wind speed (m/s). The data available at a spatial resolution of 0.25° × 0.25° in the network common data format (netCDF) were extracted. Subsequently, they were converted into CSV files using R Studio with packages ‘raster’, ‘rgdal’, ‘qdap’, ‘data.table’, and ‘ncdf4’ to facilitate further analysis.

#### Remote sensing data

The satellite data utilized in this study were sourced from the Moderate Resolution Imaging Spectroradiometer (MODIS)<sup>63</sup>. These include the enhanced vegetation index (EVI) and potential evapotranspiration (PET) with a 16-day interval at a resolution of 500 m, land surface temperature (LST) with an 8-day interval at a resolution of 1-km, normalized difference vegetation index (NDVI) with a 16-day interval at a resolution of 500 m, and potential leaf area index (LAI) with a 16-day interval at a resolution of 500 m. These parameters were obtained from image products such as MOD16A2, MOD11A2, MOD13A1, and MOD15A2H, which were available in Hierarchical Data Format (HDF) file format with various spatial and temporal resolutions. To process these data, the R packages ‘gdalutils’ and ‘modis’ were utilized to extract information from HDF files and convert them into GeoTIFF files. Subsequently, the R package ‘raster’ was employed to organize all variables into raster (grid) type files, with each predictor represented as a raster layer reflecting a specific variable of interest<sup>64</sup>.

#### Geographical parameters

The current study included data on waterbodies extracted from OpenStreetMap Data Extracts (<https://download.geofabrik.de/asia/india.html>). The data available in a zone-wise manner for India were downloaded in shapefile (shp.) format. These shapefiles were merged using the geopandas and pandas packages in Python. These shapefiles were used for mapping HPAI outbreaks incidence regions over all geographical parameters under

study. QGIS was used to estimate the distance from actual disease outbreak regions to locations of waterbodies in an attempt to understand the influence of this parameter on HPAI incidence.

### Data pre-processing and feature engineering

Data pre-processing is a crucial iterative process that transforms raw data into comprehensible and structured formats, such as CSV, ensuring its suitability for analysis. Raw datasets often contain inconsistencies, missing values, noise, and errors, which can obscure patterns and trends. To address these challenges, data pre-processing was systematically performed to eliminate noise, handle missing data, detect anomalies, and apply label encoding before initiating machine learning modelling. This structured approach encompassed four primary stages: data cleaning, integration, transformation, and reduction, ensuring both quality and relevance for subsequent analysis<sup>65</sup>. Initially, data cleaning involved identifying and correcting errors, handling duplicates, and managing outliers using visualization tools<sup>66</sup>. Missing values were treated through appropriate imputation methods. In the data integration stage, information from multiple sources was merged to create a unified dataset, resolving discrepancies in format and structure. Further, data transformation techniques, consisting of normalization and standardization, were employed to scale numerical values for consistency, while categorical variables were converted using one-hot encoding and label encoding to enhance model compatibility<sup>65</sup>. By following this structured pre-processing methodology, we ensured that the dataset was optimized, error-free, and well-prepared for efficient machine learning modelling, thereby improving prediction accuracy and computational efficiency.

### Spatio-temporal endemicity

Additionally, we meticulously depicted the year-wise and month-wise cumulative outbreaks and accurately depicted the geographic spread of reported cases, detailing information down to the district level, with descriptive spatio-temporal endemicity providing critical insights for analyzing clustering patterns and seasonal variations to aid in surveillance and control measures<sup>67</sup>.

### Spatial autocorrelation

This study used the local Getis–Ord  $G_i^*$  index to identify local autocorrelation and find the differences of the neighbouring cell values for a geographic area. This index was effective in detecting “hot spots” demonstrating positive autocorrelation and “cold spots” indicating negative autocorrelation. The positive  $Z$  score indicates the presence of a hotspot; the negative  $Z$  score, a cold spot<sup>68</sup>.

### Spatio-temporal cluster analysis

The scan statistic test was used to examine the spatio-temporal clustering of HPAI outbreaks using SaTScan software version 10.1<sup>69</sup>. SaTScan employs moving windows of varying diameters to pinpoint spatial clusters within a study area. It also detects temporal clusters and delineates ellipses or circles with dynamically changing sizes across a three-dimensional study region. Clusters were reported for circles with observed values exceeding predicted values. For the SaTScan analysis, longitude and latitude coordinates at the district level were collected to conduct clustering on the dataset for each parameter associated with disease activity (case vs. control), encompassing both temporal and spatial attributes. The method was performed on a year-wise case dataset, using the total cases recorded for each epidemiological unit (district) for that year, while accounting for the overall population of each unit. The significance level was set at ( $p \leq 0.05$ ) for robust cluster identification.

### Discerning climatic risk factors using linear discriminant analysis (LDA)

LDA is a machine learning algorithm based on Fisher’s linear discriminant theory, designed to differentiate between multiple classes. In this study, LDA was employed to thoroughly examine environmental risk parameters of HPAI, establishing a linear relationship among them. However, LDA has certain limitations: it assumes that variables follow a normal distribution and that classes have equal covariance matrices, which may not hold true in real-world datasets. Additionally, it is sensitive to outliers and performs poorly when data are not linearly separable. Despite these limitations, the linear relationships provide a strong basis for evaluating attribute impacts on computation and assessment. In this study, SaTScan was applied for detection of the significant and non-significant space-time clusters to identify risk occurrences. Then LDA was used to examine the variation of the environmental risk factors in these identified regions. The binary clustering status variable was assigned by the clustering status, where status = 1 for clustered regions and status = 0 for non-clustered regions. In the present study, an LDA was carried out with a pre-determined level of statistical significance of  $p \leq 0.05$  for all variables under study.

### Predictive risk assessment through machine learning models

The study employs machine learning algorithms to accurately estimate the effect of significant environmental risk factors on disease prediction. The correlation between environmental risk parameters and disease was analysed to generate a risk map predicting the spatiotemporal occurrences of HPAI. A total of eleven machine learning models, including Naive Bayes (NB), Flexible Discriminant Analysis (FDA), Random Forest (RF), Support Vector Machine (SVM), Multiple Adaptive Regression Splines (MARS), Adaptive Boosting (ADA), Gradient Boosting Machine (GBM), Artificial Neural Network (NNET), Classification Tree Analysis (CT), Generalized Linear Models (GLM), and Generalized Additive Models (GAM), were trained and validated to ascertain the disease risk.

#### *Weighted outbreak score*

For each district and selected month, we implemented a weighted scoring system based on historical outbreak occurrences. Districts with fewer than three outbreaks were assigned a score of 0 and those experiencing between 3 and 6 outbreaks received a score of 1, while districts with more than six outbreaks were allocated a score of 2. This scoring system provides a nuanced representation of outbreak risk, enabling finer-grained analyses in subsequent modelling efforts.

#### *Pseudo-absence data generation*

Pseudo-absence refers to locations assumed to lack a species or condition, used when actual absence data is unavailable. These points help balance datasets by representing areas unlikely to have the species based on environmental or spatial factors. A greater proportion of pseudo-absence to present data may influence model performance either in a positive or negative way. This will add biases to model inter-comparisons, for which incidence should be maintained constant at an intermediate level. To avoid bias in the comparison, pseudo-absence data generation on HPAI were made. In line with past research, a incidence rate was set at 0.5 to ensure a balanced proportion of pseudo-absences in relation to presences in the dataset<sup>70–73</sup>. Additionally, we implemented an exclusion buffer of 10 km around occurrence points to prevent overlap between cells containing presence and pseudo-absence data<sup>74</sup>. Pseudo-absence points were randomly selected from the entire background area, with the exclusion of grid points located within the buffer zone.

#### *Hyper parameterization*

The ability of a model to provide accurate outputs for unseen input data, known as generalization, is a key objective in machine learning. A well-generalized model strikes a balance between under fitting and over fitting. Training and testing data play pivotal roles in regulating model performance. The training data enables the algorithm to discern patterns, cross-validation ensures accuracy, and the test data assesses predictive capability with new information. Over fitting occurs when a model excessively learns noise in the training data, impairing its performance on new data. Non-parametric and non-linear models, while more flexible, are more susceptible to over fitting<sup>75</sup>. Conversely, an under fit model cannot effectively model the training data or generalize to new data. Striking the right balance between memorization and generalization is a common challenge in machine learning algorithms. Regularization techniques were employed to mitigate over fitting<sup>76</sup>. In this study, all models were assessed for over fitting or under fitting, and to optimize coefficient estimation, p-values, and model performance metrics, the data was randomly split into a 70% training set and a 30% testing set. This approach ensures a robust evaluation of model performance.

### **Model evaluation and ensemble techniques**

In this study, a comprehensive set of evaluation metrics including the Receiving Operating Characteristic (ROC) curve, True Skill Statistics (TSS), Cohen's Kappa (Heidke Skill Score), Area Under the ROC Curve (AUC), F1 score, error rate, accuracy, and logistic loss (LOGLOSS) were employed to assess the discriminative capacity of the machine learning models<sup>77,78</sup>. These metrics were utilized to evaluate the accuracy of prediction models based on the presence (1) or absence (0) of data<sup>79,80</sup>. In this study, the outcomes of separate forecasts from multiple model methods were aggregated using a Raster Stack approach<sup>78</sup>. Rather than relying on a single best model, it is recommended to combine predictions from different models, which provide scores ranging from 0 to 1 and averaging these scores yielded the most accurate prediction<sup>77,81</sup>. The average model score was derived by considering models that met the criteria of kappa > 0.60, ROC > 0.70, TSS > 0.70, accuracy > 0.70, F1 score > 0.80, error rate < 0.20 and LOGLOSS < 0.70 for further assessment of disease risk<sup>82</sup>. This approach ensures a robust evaluation and aggregation of predictions for a more accurate risk assessment.

### **Transmission dynamics—estimation of basic reproduction number and vaccination coverage**

The basic reproduction number ( $R_0$ ) serves as a critical measure depicting the rate of disease transmission. It signifies the average number of secondary cases produced by an infectious individual throughout its entire period of infectivity when introduced to a fully susceptible population. The significance of  $R_0$  lies in its threshold value: if  $R_0$  exceeds 1, there is an increased risk of disease propagation, whereas an  $R_0$  below 1 suggests a lower risk. Various approaches such as the Attack Rate (AR)<sup>83</sup>, Exponential Growth rate (EG)<sup>84</sup>, and Maximum Likelihood estimation (ML)<sup>85,86</sup> were employed to gauge  $R_0$  and the method with maximum of the estimates was considered. Specific details regarding the methodology utilized can be found in similar study<sup>59</sup>, where similar techniques were employed for  $R_0$  estimation. The herd immunity threshold (HIT) (vaccination coverage) refers to the proportion of the poultry population that needs to be vaccinated against an infectious HPAI to be stabilized within the herd or population. When this threshold is reached through vaccination, each case leads to precisely one more case, causing the infection to become stable in the population of livestock, i.e.,  $R_0 = 1$ . The HIT was determined according to previous methods<sup>87–89</sup> using the formula  $HIT = 1 - 1/R_0$ .

### **Statistical software**

The statistical analyses, risk mapping, and disease forecasting were conducted utilizing R statistical software version 4.3 (R Foundation for Statistical Computing, Vienna, Austria; version 4.3). R served as a versatile platform for data mining, computation, and graphical representation. Various R packages including plyr<sup>90</sup>, dplyr<sup>91</sup>, rgdal<sup>92</sup>, raster<sup>64</sup>, data.Table<sup>93</sup>, openxlsx<sup>94</sup>, tmap<sup>95</sup>, sp<sup>96</sup>, spdep<sup>97</sup>, sf<sup>98</sup>, BMM tools<sup>99</sup>, foreign<sup>100</sup>, geosphere<sup>101</sup>, MASS<sup>102</sup>, biomod2<sup>103</sup>, dsimo<sup>104</sup>, mgcv<sup>105</sup>, randomforest<sup>106</sup>, mda<sup>107</sup>, gbm<sup>108</sup>, and earth<sup>109</sup> were instrumental in tasks such as data extraction, alignment, annotation, analysis, model fitting, and validation. Additionally, SaTScan v10.1.2, QGIS, and ArcGIS Pro 3.2.1 were used for spatial statistics and mapping.



## Conclusion

This comprehensive study on HPAI in India employs a multi-dimensional approach. The graphical framework of risk estimation of HPAI under the influence of significant environmental parameters by employing machine learning model is depicted in Figure S3. It begins with a vivid illustration of outbreak patterns at the district level, offering a detailed spatial overview of the disease incidence. Accurate geo-positioning ensures the reliability of the data, forming the basis for subsequent analyses to identify high-risk zones and transmission patterns. To address potential underreporting, the study generates data indicating instances of pseudo absence. Machine learning models, bolstered by a suite of environmental covariates and pseudo-absence data, are then employed. These models predict the probability of HPAI occurrence across India, providing a spatially explicit assessment of risk and enabling targeted intervention strategies. Additionally, the  $R_0$  transmission dynamics highlighted the potential for disease spread, aiding in strategic planning for prevention and control. This holistic approach not only enhances our understanding of HPAI dynamics but also lays a foundation for informed and effective disease management strategies in India. These findings underscore the importance of data-driven approaches for effective disease surveillance and resource allocation, supporting targeted interventions to mitigate public health risks associated with HPAI.

## Data availability

The dataset and code used in this study are publicly available on GitHub at <https://doi.org/10.5281/zenodo.15082884>. Researchers and interested parties can access and utilize the data for further analysis and validation. For any additional inquiries, correspondence should be addressed to S.K.P.

Received: 22 February 2024; Accepted: 31 March 2025

Published online: 16 October 2025

## References

1. Beigel, J. H. et al. Avian influenza A (H5N1) infection in humans. *N. Engl. J. Med.* **353**, 1374–1385 (2005).
2. Liu, M., Lu, Q., Zhang, S., Feng, X. & Hossain, M. S. Temporal and Spatial characteristics of highly pathogenic avian influenza outbreaks in China during 2004 to 2015. *Poult. Sci.* **96**, 3113–3121 (2017).
3. Nguyen, L. T. et al. Spatiotemporal and risk analysis of H5 highly pathogenic avian influenza in Vietnam, 2014–2017. *Prev. Vet. Med.* **178**, 104678 (2020).
4. Sealy, J. E. et al. Poultry trading behaviours in Vietnamese live bird markets as risk factors for avian influenza infection in chickens. *Transbound. Emerg. Dis.* **66**, 2507–2516 (2019).
5. Li, K. S. et al. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in Eastern Asia. *Nature* **430**, 209–213 (2004).
6. Lai, S. et al. Global epidemiology of avian influenza A H5N1 virus infection in humans, 1997–2015: a systematic review of individual case data. *Lancet Infect. Dis.* **16**, e108–e118 (2016).
7. Webster, R. G. & Govorkova, E. A. H5N1 influenza—continuing evolution and spread. *N Engl. J. Med.* **355**, 2174–2177 (2006).
8. Chakrabarti, A. K. et al. Characterization of the influenza A H5N1 viruses of the 2008–09 outbreaks in India reveals a third introduction and possible endemicity. *PLOS ONE*. **4**, e7846 (2009).
9. Gilbert, M. et al. Flying over an infected landscape: distribution of highly pathogenic avian influenza H5N1 risk in South Asia and satellite tracking of wild waterfowl. *EcoHealth* **7**, 448–458 (2010).
10. Takekawa, J. Y. et al. Movements of wild Ruddy shelducks in the central Asian flyway and their Spatial relationship to outbreaks of highly pathogenic avian influenza H5N1. *Viruses* **5**, 2129–2152 (2013).
11. Tian, H. et al. Avian influenza H5N1 viral and bird migration networks in Asia. *Proc. Natl. Acad. Sci. U S A*. **112**, 172–177 (2015).
12. Pawar, S. D., Kode, S. S., Keng, S. S., Tare, D. S. & Pande, S. A. Spatio-temporal distribution & seasonality of highly pathogenic avian influenza H5N1 & H5N8 outbreaks in India, 2006–2021. *Indian J. Med. Res.* **158**, 113–118 (2023).
13. Gilbert, M. et al. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proc. Natl. Acad. Sci.* **105**, 4769–4774 (2008).
14. Kilpatrick, A. M. et al. Predicting the global spread of H5N1 avian influenza. *Proc. Natl. Acad. Sci.* **103**, 19368–19373 (2006).
15. Sims, L. D. Lessons learned from Asian H5N1 outbreak control. *Avian Dis.* **51**, 174–181 (2007).
16. Ahmed, S. S. U., Ersbøll, A. K., Biswas, P. K. & Christensen, J. P. The space-time clustering of highly pathogenic avian influenza (HPAI) H5N1 outbreaks in Bangladesh. *Epidemiol. Infect.* **138**, 843–852 (2010).
17. Walsh, M. G., Mor, S. M. & Hossain, S. Highly Pathogenic Avian Influenza (H5N1) landscape suitability varies by wetland habitats and the degree of interface between wild waterfowl and poultry in India. *Viruses* **12**, 1290 (2020).
18. Pandit, P. S., Bunn, D. A., Pande, S. A. & Aly, S. S. Modeling highly pathogenic avian influenza transmission in wild birds and poultry in West Bengal, India. *Sci. Rep.* **3**, 2175 (2013).
19. Gilbert, M. & Pfeiffer, D. U. Risk factor modelling of the spatio-temporal patterns of highly pathogenic avian influenza (HPAIV) H5N1: A review. *Spat. Spatio-Temporal Epidemiol.* **3**, 173–183 (2012).
20. Guinat, C. et al. Spatio-temporal patterns of highly pathogenic avian influenza virus subtype H5N8 spread, France, 2016 to 2017. *Eurosurveillance* **23**, 1700791 (2018).
21. Liu, C. M. et al. Temperature drops and the onset of severe avian influenza A H5N1 virus outbreaks. *PLOS ONE*. **2**, e191 (2007).
22. Minh, P. Q. et al. Spatio-temporal epidemiology of highly pathogenic avian influenza outbreaks in the two deltas of Vietnam during 2003–2007. *Prev. Vet. Med.* **89**, 16–24 (2009).
23. Dong, W., Yang, K., Xu, Q., Liu, L. & Chen, J. Spatio-temporal pattern analysis for evaluation of the spread of human infections with avian influenza A(H7N9) virus in China, 2013–2014. *BMC Infect. Dis.* **17**, 704 (2017).
24. Oyana, T. J., Dai, D. & Scott, K. E. Spatiotemporal distributions of reported cases of the avian influenza H5N1 (Bird Flu) in Southern China in early 2004. *Avian Dis.* **50**, 508–515 (2006).
25. Loth, L., Pham, L. T. & Stevenson, M. A. Spatio-temporal distribution of outbreaks of highly pathogenic avian influenza virus subtype H5N1 in Vietnam, 2015–2018. *Transbound. Emerg. Dis.* **68**, 13–20 (2021).
26. Jayashree, A. et al. Exploring the impact of climate variables on livestock anthrax outbreaks: A machine learning approach. *Int. J. Environ. Clim. Change*. **14**, 494–507 (2024).
27. Gilbert, M. et al. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proc. Natl. Acad. Sci. U S A*. **105**, 4769–4774 (2008).
28. Gilbert, M. et al. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data*. **5**, 180227 (2018).

29. Lim, J. S. et al. Spatial epidemiology of highly pathogenic avian influenza subtype H5N6 in Gyeonggi Province, South Korea, 2016–2017. *Transbound. Emerg. Dis.* **69**, e2431–e2442 (2022).
30. Martin, V. et al. Risk-based surveillance for avian influenza control along poultry market chains in South China: the value of social network analysis. *Prev. Vet. Med.* **102**, 196–205 (2011).
31. Lockhart, C., Kreindel, S., Pittiglio, C. & Escher, M. Highly pathogenic avian influenza (H5N1 HPAI) spread in the middle East: risk assessment. **36** (2016).
32. Awada, L. et al. Global dynamics of highly pathogenic avian influenza outbreaks in poultry between 2005 and 2016-Focus on distance and rate of spread. *Transbound. Emerg. Dis.* **65**, 2006–2016 (2018).
33. Islam, A. et al. Potential risk zones and Climatic factors influencing the occurrence and persistence of avian influenza viruses in the environment of live bird markets in Bangladesh. *One Health.* **17**, 100644 (2023).
34. Jaakkola, K. et al. Decline in temperature and humidity increases the occurrence of influenza in cold climate. *Environ. Health.* **13**, 22 (2014).
35. Durand, L. O. et al. Timing of influenza A(H5N1) in poultry and humans and seasonal influenza activity worldwide, 2004–2013. *Emerg. Infect. Dis.* **21**, 202–208 (2015).
36. ElMasry, I. et al. Avian influenza H5N1 surveillance and its dynamics in poultry in live bird markets, Egypt. *Transbound. Emerg. Dis.* **64**, 805–814 (2017).
37. Park, A. W. & Glass, K. Dynamic patterns of avian and human influenza in East and Southeast Asia. *Lancet Infect. Dis.* **7**, 543–548 (2007).
38. Sakoda, Y. et al. Reintroduction of H5N1 highly pathogenic avian influenza virus by migratory water birds, causing poultry outbreaks in the 2010–2011 winter season in Japan. *J. Gen. Virol.* **93**, 541–550 (2012).
39. Choi, J. G. et al. Characterization of clade 2.3.2.1 H5N1 highly pathogenic avian influenza viruses isolated from wild birds (Mandarin Duck and Eurasian eagle Owl) in 2010 in Korea. *Viruses* **5**, 1153–1174 (2013).
40. Soares Magalhães, R. J. et al. Live poultry trade in Southern China provinces and HPAIV H5N1 infection in humans and poultry: the role of Chinese new year festivities. *PLoS ONE.* **7**, e49712 (2012).
41. Chen, W. et al. Environmental factors and Spatiotemporal distribution characteristics of the global outbreaks of the highly pathogenic avian influenza H5N1. *Environ. Sci. Pollut. Res.* **29**, 44175–44185 (2022).
42. Chen, H. et al. H5N1 virus outbreak in migratory waterfowl. *Nature* **436**, 191–192 (2005).
43. Liu, J. et al. Highly pathogenic H5N1 influenza virus infection in migratory birds. *Science* <https://doi.org/10.1126/science.1115273> (2005).
44. Olsen, B. et al. Global patterns of influenza A virus in wild birds. *Science* **312**, 384–388 (2006).
45. Blem, C. R. & Blem, C. R. Handbook of the birds of the world. 1: ostrich to ducks by Josep Del Hoyo, Andrew Elliott, Jordi Sargatal. *Wilson Bull.* **106**, 575–575 (1994).
46. Pinet, P. WOAHA calls for higher surveillance of avian influenza as outbreaks intensify. *WOAH - World Organisation for Animal Health.* <https://www.woah.org/en/the-world-organisation-for-animal-health-oie-calls-for-increased-surveillance-of-avian-influenza-as-outbreaks-in-poultry-and-wild-birds-intensify/> (2021).
47. Potdar, V. et al. Identification of human case of avian influenza A(H5N1) infection, India. *Emerg. Infect. Dis.* **28** (2022).
48. Firestone, S. M. et al. The influence of meteorology on the spread of influenza: survival analysis of an equine influenza (A/H3N8) outbreak. *PLoS ONE.* **7**, e35284 (2012).
49. Siboonnan, N. et al. A serine-to-asparagine mutation at position 314 of H5N1 avian influenza virus NP is a temperature-sensitive mutation that interferes with nuclear localization of NP. *Arch. Virol.* **158**, 1151–1157 (2013).
50. Paeck, M. R. et al. Survival rate of H5N1 highly pathogenic avian influenza viruses at different temperatures. *Poult. Sci.* **89**, 1647–1650 (2010).
51. Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T. & Lipsitch, M. Absolute humidity and the seasonal onset of influenza in the continental united States. *PLoS Biol.* **8**, e1000316 (2010).
52. Urashima, M., Shindo, N. & Okabe, N. A seasonal model to simulate influenza Oscillation in Tokyo. *Jpn J. Infect. Dis.* **56**, 43–47 (2003).
53. Waziri, M. I., Abdu, P. A., Sa'idu, L. & Bello, M. Seroepidemiology and assessment of risk factors for the spread of avian influenza in birds in two Nigerian States. *Vet. Med. Sci.* **3**, 227–238 (2017).
54. Jonges, M. et al. Wind-Mediated spread of Low-Pathogenic avian influenza virus into the environment during outbreaks at commercial poultry farms. *PLoS ONE.* **10**, e0125401 (2015).
55. Si, Y., De Boer, W. F. & Gong, P. Different environmental drivers of highly pathogenic avian influenza H5N1 outbreaks in poultry and wild birds. *PLoS ONE.* **8**, e53362 (2013).
56. Simms, L. & Jeggo, M. Avian influenza from an ecohealth perspective. *EcoHealth* **11**, 4–14 (2014).
57. Fang, L. Q. et al. Environmental factors contributing to the spread of H5N1 avian influenza in Mainland China. *PLoS ONE.* **3**, e2268 (2008).
58. FAO & Division, A. P. and H. The Global Strategy for Prevention and Control of H5N1 Highly Pathogenic Avian Influenza (2008).
59. Ward, M. P., Maftei, D., Apostu, C. & Suru, A. Estimation of the basic reproductive number ( $R_0$ ) for epidemic, highly pathogenic avian influenza subtype H5N1 spread. *Epidemiol. Infect.* **137**, 219–226 (2009).
60. Kim, W. H. & Cho, S. Estimation of the basic reproduction numbers of the subtypes H5N1, H5N8, and H5N6 during the highly pathogenic avian influenza epidemic spread between farms. *Front. Vet. Sci.* **8**, 597630 (2021).
61. Geography of India. Wikipedia. (2024).
62. Sinkie, Z. M. Modeling And Simulation Study Of Anthrax Attack On Environment. **3** (2016).
63. Justice, C. O. et al. An overview of MODIS land data processing and product status. *Remote Sens. Environ.* **83**, 3–15 (2002).
64. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. 3.6–30 (2010).
65. Saleem, A., Asif, K. H., Ali, A., Awan, S. M. & Alghamdi, M. A. Pre-processing Methods of Data Mining. in *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing* 451–456 (IEEE, 2014). <https://doi.org/10.1109/UCC.2014.57>.
66. Sharma, G. *Complete Guide to Feature Engineering: Zero to Hero.* (2024). <https://www.analyticsvidhya.com/blog/2021/09/complete-guide-to-feature-engineering-zero-to-hero/>
67. Anandakumar, J. et al. Comprehensive Spatial-Temporal and Risk Factor Insights for Optimizing Livestock Anthrax Vaccination Strategies in Karnataka, India. *Vaccines* **12**, 1081 (2024).
68. Mathur, M. Spatial autocorrelation analysis in plant population: an overview. *J. Appl. Nat. Sci.* **7**, 501–513 (2015).
69. Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R. & Mostashari, F. A Space-Time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**, e59 (2005).
70. Hanspach, J., Kühn, I., Schweiger, O., Pompe, S. & Klotz, S. Geographical patterns in prediction errors of species distribution models: patterns in prediction error. *Glob Ecol. Biogeogr.* **20**, 779–788 (2011).
71. Hengl, T., Sierdsema, H., Radović, A. & Dilo, A. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecol. Model.* **220**, 3499–3511 (2009).
72. McPHERSON, J. M., Jetz, W. & Rogers, D. J. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.* **41**, 811–823 (2004).
73. Senay, S. D., Worner, S. P. & Ikeda, T. Novel Three-Step Pseudo-Absence selection technique for improved species distribution modelling. *PLoS ONE.* **8**, e71218 (2013).

74. Chefaoui, R. M. & Lobo, J. M. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecol. Model.* **210**, 478–486 (2008).
75. Cano-Ortiz, S. & Pascual-Muñoz, P. Castro-Fresno, D. Machine learning algorithms for monitoring pavement performance. *Autom. Constr.* **139**, 104309 (2022).
76. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **1168**, 022022 (2019).
77. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223–1232 (2006).
78. White, M. & Newell, G. Measuring the accuracy of species distribution models: a review.
79. Huettmann, F. et al. Use of machine learning (ML) for predicting and analyzing ecological and ‘presence only’ data: an overview of applications and a good outlook. in *Machine Learning for Ecology and Sustainable Natural Resource Management* (eds Humphries, G., Magness, D. R. & Huettmann, F.) 27–61 (Springer International Publishing, Cham, doi:[https://doi.org/10.1007/978-3-319-96978-7\\_2](https://doi.org/10.1007/978-3-319-96978-7_2)). (2018).
80. Liu, C., White, M. & Newell, G. Measuring the accuracy of species distribution models: a review (2009).
81. Huppert, A. & Katriel, G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin. Microbiol. Infect.* **19**, 999–1005 (2013).
82. Abdrakhmanov, S. K. et al. Maximum entropy modeling risk of anthrax in the Republic of Kazakhstan. *Prev. Vet. Med.* **144**, 149–157 (2017).
83. Dietz, K. The Estimation of the basic reproduction number for infectious diseases. *Stat. Methods Med. Res.* **2**, 23–41 (1993).
84. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. Biol. Sci.* **274**, 599–604 (2007).
85. Roberts, M. G. & Nishiura, H. Early Estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in new Zealand. *PLoS One.* **6**, e17835 (2011).
86. Arruda, A. G., Alkhamis, M. A., VanderWaal, K., Morrison, R. B. & Perez, A. M. Estimation of Time-Dependent reproduction numbers for Porcine reproductive and respiratory syndrome across different regions and production systems of the US. *Front. Vet. Sci.* **4**, 46 (2017).
87. Bridges, C. B. et al. Prevention and control of influenza. Recommendations of the advisory committee on immunization practices (ACIP). *MMWR Recomm. Rep. Morb. Mortal. Wkly. Rep. Recomm. Rep.* **52**, 1–34 (2003). quiz CE1-4.
88. Fine, P., Eames, K. & Heymann, D. L. Herd immunity: a rough guide. *Clin. Infect. Dis. Off Publ Infect. Dis. Soc. Am.* **52**, 911–916 (2011).
89. Suresh, K. P. et al. A new methodology to comprehend the effect of El Niño and La Niña Oscillation in early warning of anthrax epidemic among livestock. *Zoonotic Dis.* **2**, 267–290 (2022).
90. Wickham, H. The Split-Apply-Combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
91. dplyr.pdf.
92. rgdal.pdf.
93. Barrett, T. et al. data.table: Extension of ‘data.frame’. 1.16.4. <https://doi.org/10.32614/CRAN.package.data.table> (2006)
94. openxlsx.pdf.
95. Tennekes, M. & tmap Thematic maps in R. *J. Stat. Softw.* **84**, 1–39 (2018).
96. Pebesma, E. & Bivand, R. S. Classes and Methods for Spatial Data: the sp Package.
97. Bivand, R. S. & Wong, D. W. S. Comparing implementations of global and local indicators of Spatial association. *TEST* **27**, 716–748 (2018).
98. Pebesma, E. & Bivand, R. *Spatial Data Science: with Applications in R* (Chapman and Hall/CRC, 2023). <https://doi.org/10.1201/9780429459016>.
99. Rabosky, D. L. et al. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* **5**, 701–707 (2014).
100. R Core Team. foreign: Read Data Stored by ‘Minitab’, ‘S’, ‘SAS’, ‘SPSS’, ‘Stata’, ‘Systat’, ‘Weka’, ‘dBase’, ... <https://doi.org/10.32614/CRAN.package.foreign> (1999).
101. Hijmans, R. J. & geosphere Spherical Trigonometry. 1.5–20. <https://doi.org/10.32614/CRAN.package.geosphere> (2010).
102. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S.
103. Thuiller, W. et al. biomod2: ensemble platform for species distribution modeling. 4.2-5-2. <https://doi.org/10.32614/CRAN.package.biomod2> (2012).
104. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. *Species Distribution Model.* 1.3–16 <https://doi.org/10.32614/CRAN.package.dismo> (2010).
105. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood Estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36 (2011).
106. Liaw, A. & Wiener, M. Classification and Regression by randomForest.
107. Hastie, T. & Tibshirani, R. Mixture and flexible discriminant analysis. 0.5-5 (1998). <https://doi.org/10.32614/CRAN.package.mda>
108. Ridgeway, G. Generalized Boosted Models: A guide to the gbm package.
109. Milborrow, S., Hastie, T. & Tibshirani, R. earth: Multivariate Adaptive Regression Splines. 5.3.4. <https://doi.org/10.32614/CRAN.package.earth> (2007).

## Acknowledgements

We express our gratitude to the Department of Animal Husbandry and Dairying, India, as well as The World Organization for Animal Health (formerly OIE), for providing data on AI outbreaks. We also appreciate the continuous support, guidance, and assistance from the Director General (DG), Deputy Director General (DDG) (AS) of ICAR, and Director of ICAR-NIVEDI throughout our study. Furthermore, we extend our heartfelt thanks to the NADRES v2, ICAR-NICRA and DBT - One Health projects for their assistance in preparing this manuscript.

## Author contributions

Conceptualization of topic, methodology and supervision: S.K.P.; Data curation and formal analysis: S.K.P., A.V.P., P.K. & S.S.P.; Writing original draft: J.A.; Review and editing: S.K.P., N.N.B., P.K., S.S.P., S.L. and R.R.; Editing and funding: B.R.G. Visualization: R.R., A.V.P & J.A. All authors reviewed the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-96765-2>.

**Correspondence** and requests for materials should be addressed to S.K.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025