Environ Health Perspect, DOI: 10.1289/ehp15937, Accepted 5/14/2025

An integrated machine learning framework to understand zoonotic spillover emergence across anthropogenically modified landscapes

Yinsheng Zhang[1,2], Jinchen Wang[1], Luqi Wang[1], Linxuan Miao[1], Yifan Sun[1], Xin Yang[1], Ruying Fang[1,2], Yiyang Guo[1], Sophie Vanwambeke[3], Sen Li[1,4]

1. School of Environmental Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei, P.R. China

2. Institute of Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, Hubei, P.R. China

3. Earth and Life Institute, Université Catholique de Louvain (UCLouvain), B1348 Louvain-la-Neuve, Belgium

4. Hubei Key Laboratory of Multi-media Pollution Cooperative Control in Yangtze Basin, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, 430074, China

Author for correspondence: Sen Li, School of Environmental Science and 18 Engineering, Huazhong University of Science and Technology, Wuhan, 430074, P.R. China. Email: senli@hust.edu.cn

Competing Financial Interests

The authors declare they have no conflicts of interest related to this work to disclose.

Abstract

BACKGROUND: Anthropogenic land modification influences human-livestock-wildlife interactions and zoonotic spillover emergence. However, the extent of this impact remains unclear and could be better understood through the collaborative use of advanced predictive and explanatory analytical tools, alongside, an up-to-date dataset on zoonotic spillover events.

OBJECTIVE: The main objective is to develop and evaluate an integrated modeling framework to predict and explain spatial patterns and nonlinear relationships of zoonotic spillover events, using updated datasets and the human modification index to differentiate anthropogenic pressures.

METHODS: Our study expanded the historical datasets to include recent spillover events and a comprehensive set of predictors. By combining robustness of finely-tuned stacking algorithms with structural equation modeling, we considered global heterogeneity in relative reporting adequacy and mapped spillover patterns at different scales. Using the human modification index, we disentangled anthropogenic processes modifying natural ecosystem across land modification gradients, and described their linkages to spillover occurrence over the past three decades.

RESULTS: This integrated approach effectively improved the model's predictive and explanatory power. Our analysis reveals that the intermediate levels of human pressure facilitated the zoonotic spillover. The indirect effects of anthropogenic pressure, mediated by specific cropping intensity, are strongly associated with zoonoses emergence. Livestock distribution serves as an indicator of spillover hotspots, acting as effective proxies for distinctive landscapes.

DISCUSSION: Our findings identify high zoonotic spillover risks present across geographically and socioeconomically diverse regions worldwide, extending beyond tropical areas, including extensive regions experiencing high-intensity human modification. These insights support targeted surveillance in areas with potentially high relative risk or uncertainty, and demonstrate how zoonotic spillover responds to complex human-environment interactions.

Introduction

Over the past century, human interactions with the environment and animals have intensified significantly. These interactions have exerted considerable, and sometimes irreversible, impacts on both society and ecosystem[1]. Land use for agriculture, animal domestication, and livestock farming have reshaped habitat structures and human-animal contact patterns. These changes have disproportionately increased the risk of zoonotic infections in humans[2]. Identifying processes that drive zoonotic transmission would support effective pandemic surveillance and targeted investments, aiding in the management of global health over time and space.

The transmission of zoonotic pathogens is a complex process that begins with their circulation in animal reservoirs, crosses species barriers to infect human (known as spillover), and can potentially lead to human-to-human transmission[3]. Zoonotic diseases caused by these pathogens account for most recent pandemics and pose a growing threat to global economic stability[4]. The mechanisms underlying zoonotic spillover are intricate, influenced by opportunities for pathogens exchange, and shaped by environmental changes, landscape modifications, and shifts in the geographic distributions of hosts and vectors[5,6]. These changes are often driven by human activity and animal synanthropy, with human-induced modifications typically playing a dominant role[7,8]. Advances in high-resolution datasets and computing power have enabled mapping of human-animal interactions and landscape changes, promoting interdisciplinary collaborations and exploration of their public health consequences. Despite these advances, our ability to predict new spillover events and identify how human activities drive zoonotic emergence remains limited.

The patterns of land use, such as urban, rural, and forested areas, along with the distribution of livestock or wildlife, have been extensively used to study zoonotic diseases. These factors, which can be monitored on a fine scale via satellite imagery, are often used to characterize the impact of human activities and enhance our predictive models. However, relying solely on these factors is inadequate for discerning the complex relationships between anthropogenic pressures and the risk of zoonotic emergence, as they do not always distinguish between pressures that degrade versus those that transform natural ecosystems[9]. Moreover, the interplay between these factors and

biophysical or climatic processes may impede our ability to accurately evaluate and quantify their relationships with zoonotic outbreaks. The land modification patterns and intensity across different regions and biomes show significant variability over time and space[10]. The majority of landscapes are subject to one or more anthropogenic pressures. The cumulative effects of these pressures from multiple development projects within a region can result in complex and unpredictable outcomes for ecosystems and the animal-human interface[10,11]. This highlights the necessity of differentiating between multiple anthropogenic pressures, and understanding the operational patterns and pathways of emerging spillover that have similar impacts worldwide.

The human modification index serves as a comprehensive measure that threshold along a continuous gradient of land modification values to evaluate landscape structure, quantifying the extent of anthropogenic impact on natural terrestrial systems, referred to as intensity. Although evidence suggests that human modifications can affect infection rates in wildlife, transmission risks in rodent-borne diseases, and incidence rates in vector-borne diseases[9,12], a holistic analysis that connects zoonotic spillovers with multiple anthropogenic pressures across continuous gradients of land modification under large spatio-temporal scales has yet to be conducted using robust predictive methodologies. Given the variations in land use, climate patterns, biodiversity, and human disturbance across different regions and biomes, the characteristics of landscapes and animal-human interfaces also exhibit substantial variation. In this context, human modification provides a critical perspective for unraveling the complexity of zoonotic emergence on a global scale. This index evaluates the diverse impacts of anthropogenic pressures on land and wildlife, allowing for systematic identification and disentanglement of common predictors effective across various pathogen systems, thereby assessing the multidimensional influence on zoonotic emergence worldwide across continuous gradients of land modification.

Mechanistic models, while effective at simulating and explaining zoonotic transmission, face challenges in parameterization and rely heavily on parameter assumptions. In contrast, statistical and machine learning (ML) models are often favored as spatio-temporal data can be directly used to estimate parameters from observations[13]. However, these ML models are often described as

"black box" due to their complexity and lack of interpretability, posing challenges for deriving healthcare practices. Additionally, oversimplifying variables can obscure complex and nonlinear dynamics in zoonotic systems. Recent advances in artificial intelligence (AI) have introduced innovative public health approaches, with high-performance machine learning (ML) techniques, such as bagging and boosting algorithms, showing potential for their efficiency and adaptability in identifying critical drivers and correlations for specific regional diseases[14,15]. Although previous studies have mapped hotspots of emerging diseases using ML techniques[11], the causal pathways by which key predictors influence disease risk remain largely untapped. Moreover, distinguishing direct and indirect impacts is difficult, hindering effective interventions and resource allocation. Combining explanatory analytical tools with robust ML in a predictive framework is promising for enhancing reliability and interpretability, thereby elucidating the mechanisms of zoonotic emergence in complex scenarios.

This study aims to develop an integrated modeling framework to predict spatial patterns of emerging zoonotic spillover risk and evaluate anthropogenic impacts across land-modification gradients (Fig. 1). Our research addresses three key questions: (1) How effectively do advanced ensemble machine learning approaches enhance spatial prediction of spillover events? (1) What are the correlation and relative contributions of refined factors in predicting emerging spillover risk? (2) How does human modification influence zoonotic spillover risk across anthropogenic landscape gradients through direct and indirect pathways? To investigate these questions, we expanded the historical zoonotic spillover database to include recent events from both wildlife and domestic sources. Our analytical framework incorporates various predictors, including climate, landscapes, animal hosts, human population and modification-while accounting for global heterogeneity in surveillance and relative reporting adequacy. Using the human modification index, we differentiated between natural and anthropogenic influences on disease transmission patterns. We implemented a stacked machine learning approach integrated with explainable artificial intelligence (XAI) to enhance predictive power and identify key factor correlations with spillover events. Finally, we employed structural equation modeling (SEM) to elucidate the linkages between emerging zoonotic spillover and key predictor variables, enabling assessment of anthropogenic effects across land-modification

gradients.

Methods

Compiling historical spillover event distribution record

Data on emerging zoonotic spillover events were sourced from an updated version of the database initially developed by ref. [2,16], which cataloged emerging pathogens from 1940 to 2004. These baseline data were compiled through extensive reviews of emerging disease literature, institute of medicine reports, and texts of human infectious disease. We primarily employed species-level classification to minimize biases from organisms with extensive subspecies variation. However, recognizing the significance of certain pathogens, we expanded our database to include specific subtypes for organisms like Ebolavirus and Influenza A virus, particularly when they demonstrated substantial global disease burden or appeared on WHO priority disease lists.

For post-2004 emerging pathogens, we integrated data from multiple authoritative sources, including the National Institute of Allergy and Infectious Diseases (NIAID) Biodefense Pathogens database, the World Health Organization's Priority Pathogen List, and the United Kingdom public sector information platform's emerging infectious diseases list. These data were further supplemented by datasets from previous literature[17], which encompassed updated reported events through 2008. Additionally, the Emerging Infectious Disease Repository (https://eidr.ecohealthalliance.org/) provided comprehensive data through 2013, enabling further validation of our pathogen inventory through its extensive documentation.

The identification of the zoonotic pathogens origins in humans depends on timely detection and extensive reporting, which are influenced by technical and regional variances[43]. This study focused on the first emergence of zoonotic pathogens in human population and their subsequent documentation (termed "reported emerging zoonotic spillover"). We defined emergence events as "newly identified zoonotic pathogens in human populations with well-documented geographical and temporal origin evidence". These events include diseases that emerged due to expanded distribution, increased incidence, or enhanced virulence. By clearly defining the scope of analysis

and event data, we aim to minimize potential ambiguities and complexities in subsequent interpretations. Emergence was usually confirmed through more than one peer-reviewed papers or, for previously known diseases, reports considered significantly above background. This criterion ensured the inclusion of only those events with substantial evidence regarding geographical and temporal origins.

To ensure data quality, we employed a validation process, any discrepancies were resolved through additional literature review and expert consultation. All emergence events expanded in our analysis are supported by literature sources, which are provided in the Table S10. We systematically compiled data on 226 zoonotic spillover events, documenting the location, occurrence year, pathogen genus and species, transmission type, and zoonotic origin for each event. The specificity of the geographic distribution of each event varied, depending on the details available in initial literature and profiles, ranging from precise coordinates to broader regions, such as municipalities, counties, districts, or even countries and partial continents (Excel Table S1).

Assessing anthropogenic and environmental impact data

Data on human modification and population density

Human modification index (HMc) was used as the primary measure of anthropogenic pressure, providing an ideal and single metric to separate and quantify either human pressure or natural processes on animals and landscapes. The HMc quantifies the cumulative human modification of global terrestrial lands, building upon established approaches[18]. Detailed methodological documentation are available in the original Global Human Modification of Terrestrial Systems, v1 (2016)[10]. The index calculation incorporates the spatial extent and impact intensity of 13 anthropogenic stressors (Table 1), the produced cumulative HMc was calculated using a " fuzzy algebraic sum" approach[19], which combines 13 stressor indicator layers (HMs) through the following equation:

$$HMc = 1.00 - \prod_{s=1}^{n} (1-(HM_s)) \qquad (1)$$

This method generates HMc values from 0.00 (unmodified land) to 1.00 (highly modified land). The algorithm assigns higher modification values to areas affected by multiple stressors compared

to those influenced by a single stressor. Importantly, while the HMc value equals or exceeds the highest individual stressor indicator value, the incremental contribution of additional stressors diminishes as values overlap, approaching but never exceeding 1.00.

Following thresholds established in the previous study[10], the HMc values were categorized into four distinct classes: low ($0.00 \leqslant HMc \leqslant 0.10$), moderate ($0.10 < HMc \leqslant 0.40$), high ($0.40 < HMc \leqslant 0.70$), and very high ($0.70 < HMc \leqslant 1.0$). HMc = 0.4 marks the transition from a moderately modified to a highly modified state. This threshold also corresponds to areas of low-intensity agriculture and signifies the beginning of a human-dominated state.

The global human population density dataset was acquired from HYDE3.3 at a resolution of 5' from 1940 to 2020. We quantified temporal dynamics through spatial change layer, calculated as differences between consecutive five-year intervals (1940-2020). This layer (Population density change) was temporally aligned with emerging zoonotic events using nearest-year matching[20], with final dataset integration based on spatial grid identifiers and event year.

Data on climate and land-use categories

Climate influences disease transmission and distribution through multiple pathways, affecting reproduction, development, behavior, and dynamics of host and vector populations [9,21]. We extracted climate data from the Copernicus Climate Change Service ERA5 database (https://cds.climate.copernicus.eu/), which provides comprehensive global climate and weather data since 1940. Using a re-gridded subset at the original resolution, we sampled monthly reanalysis data at five-year intervals (1940-2020). We focused on three key variables: temperature at 2 meters, total precipitation, and surface pressure. These variables were selected for their weak to moderate correlation ($r < 0.7$) at the global scale, which aligns with established thresholds for variables selection and enables clear interpretation of individual variable contributions[5,9,22]. For each time point, we calculated annual averages and seasonality metrics—standard deviation for temperature and coefficient of variation for precipitation and surface pressure. We also quantified temporal dynamics through spatial change layers, computed as differences between consecutive five-year

intervals. Climate parameters were temporally aligned with emerging zoonotic events using nearest-year matching (1940-2020), with final dataset integration based on spatial grid identifiers and year.

Land use categories effectively reflect the extent to which humans modify the environment, either through large-scale conversion of natural habitat or localized practices, each of which can alter the habitat structures, species distributions, and interactions[23,24], thereby increasing human exposure to diverse pathogen assemblages. They may serve as broad predictors of zoonotic spillover emergence, encompassing ecological and socioeconomic contexts. Historical land-use data from 1940 to 2015 were obtained from the Land Use Harmonization 2 (LUH2) project at a 0.25° spatial resolution[25,26]. These categories encompass urban areas, agricultural lands (including C3 annual crops, managed pasture, and C4 perennial crops), forested regions (e.g. forested primary land and potentially forested secondary land), and other non-forested lands (e.g., non-forested primary land). The LUH2 dataset reconstructs and predicts changes in land use across 12 categories. Through our variable screening process (described in subsequent sections). we selected eight variables for subsequent analysis (with the complete predictors presented in Table S1 and Figure S7). Similarly, we generated the spatial change layer by calculating differences between successive five-year intervals. This approach enabled us to capture spatiotemporal patterns in land-use transitions, which correlate with anthropogenic modifications (e.g. tree harvesting for forested-related land). We temporally aligned land-use variables with documented zoonotic emergence events using nearest-year matching, integrating the dataset based on spatial grid identifiers and temporal indices. In addition, we selected six static land cover classes from the EarthEnv dataset[57], quantifying diverse natural ecosystems as a percentage per grid cell (Table S1). Moreover, we obtained the Global Multi-resolution Terrain Elevation Data (GMTED2010) with a 30” resolution, which provides detailed topographic information of elevation.

Data on animal hosts

The domestication of animals, grazing, and habitat changes have impacted humans by altering contact patterns and creating transmission opportunities[27]. Many zoonotic pathogens, originating in wildlife, are transmitted to humans via domestic animals, which often serve as the intermediate or

amplifier hosts. The shift towards more intensive livestock production complicates the global trend in the emergence of novel zoonoses due to increased animal density, transformed patterns of animal-human interaction, and land-use changes. Specific sentinel animal species may be key indicators of potential spillover risks. We obtained eight predictors from the Gridded Livestock of the World (GLW) dataset, including headcounts of chicken, duck, horse, goats, buffaloes, cattle, sheep, and pigs[28].

The diversity and prevalence of zoonotic pathogens in animal reservoirs are critical for maintaining pathogen circulation and sustaining epidemics. However, there is a lack of spatial data on global pathogen diversity, and the estimation of pathogen diversity is often underrated[29,30]. We therefore used mammal biodiversity, quantified by the number of mammal species per grid cell, as a well-known correlate of zoonotic spillover risk to indicate pathogen species richness[2,29,31,32]. We extracted the global mammal richness dataset from the NASA Socioeconomic Data and Application Center, and incorporated the species distribution grids from the International Union for the Conservation of Nature (IUCN) Red List status categories[33].

Estimating relative reporting adequacy

Data on reporting adequacy in zoonotic spillover events

The spatial distribution of reported zoonotic spillover events is significantly influenced by inconsistencies in detection and reporting, known as reporting or observational bias[34]. A previous study analyzed the PubMed Central Open-Access Subset using the GeoNames database for geographical entity recognition, and employed a Poisson boosted regression tree model to generate reporting effort layer and address missing data[17]. In this study, we used existing reporting efforts data, while assuming that high reporting efforts could indicate either superior surveillance abilities or high disease burden. However, high disease burden with proportionate reporting efforts does not necessarily indicate significant reporting bias. Instead, reporting bias emerges and becomes pronounced when reporting efforts are misaligned with human disease burden (e.g. where high disease burden coincides with inadequate surveillance or reporting). To account for this, we defined the "relative reporting adequacy", which estimates the alignment between a region's

surveillance/reporting capacity and its human disease burden. This metric is quantified as the ratio of reporting efforts to disability-adjusted life years (DALYs) caused by infectious disease. The integration of these spatially-matched reporting effort estimates with disability-adjusted life years (DALYs) data enabled us to estimate reporting and observational bias at a fine spatial scale (see Adjusting for observation bias). For high-resolution estimates of reporting effort, we used the 'pubs_superhires' data from the EcoHealth Alliance's pubcrawler2hotspot project (github.com/ecohealthalliance/pubcrawler2hotspots). This dataset was selected due to its optimal spatial resolution, which aligns well with our analytical scale. Per capita Disability-Adjusted Life Year (DALY) rates from infectious disease was obtained from the Global Burden of Disease Study 2021 (GBD 2021, 1990-2015) and converted to the study grid weighted by grid cell population[35].

Predictors and modeling in relative reporting adequacy

The raw reporting efforts layer comprised numerous grid cells with zero values. These zero values, however, may not necessarily indicate an absence of reporting but rather potential data gaps, as suggested by previous research[17]. To address this limitation and smooth noise in the raw data, we used a CatBoost model[36] to impute the relative reporting adequacy index across all grid cells. This approach generated a layer that effectively represented the underlying data while maintaining consistency with the spatial resolution (0.25°) of other variables. We used predictors including health expenditure, accessibility, urban land, Gross Domestic Product (GDP), and healthcare infrastructure indicators[17]. These variables were selected as proxies for socioeconomic development and healthcare system capacity, factors hypothesized to influence relative reporting adequacy. The predictors for creating the relative reporting adequacy layer were analyzed at grid-cell resolution and included urban/built-up land, derived from the EarthEnv dataset (1992-2006, 0°0'30" resolution), and average travel time to urban centers with populations exceeding 50,000 (2015, 0°0'30" resolution)[37]. Additional predictors, available at sub-country or country levels, were first converted to per capita measures and weighted by grid-cell population to ensure spatial consistency. These variables include economic development (per capita GDP, 1990-2015 average, from sub-country datasets[38]), healthcare investment (external and out-of-pocket expenditure per capita in US$, 2000–2020 average, from the World Health Organization [WHO]), and healthcare accessibility and

infrastructure (Healthcare Access and Quality [HAQ] Index[39], number of hospital beds, and number of medical doctors and nurses, 1990-2020 average, from WHO).

Unlike previous study's parametric approach[17], our analysis adopts a posterior-driven approach using CatBoost. As an advanced gradient boosting framework, CatBoost's ordered boosting principle reduces prediction shift through unbiased gradient estimation estimation, preventing target leakage during training[36]. To enhance model robustness, we employed ten-fold cross-validation, achieving mean R^2 values of 0.72 in cross-validation. This data-driven approach avoids potential bias from parametric assumptions, and provides robust estimates of model generalization through multiple independent test sets[40]. The CatBoost algorithm was fitted using "CatBoostRegressor()" from the Python library "catboost", hyperparameter tuning was performed through Bayesian optimization employing the Tree of Parzen Estimators (TPE) method available in the "hyperopt" library[47].

Preprocessing data and developing the Machine learning framework

Preprocessing of predictors and spillover events

This study collected 58 variables from natural and anthropogenic factors, categorizing them into climate, landscapes, animal hosts, human population and modification (see full details in Table S1). Prior to developing the modeling algorithms and workflow, we performed data rescaling and aggregation procedures to ensure format consistency and compatibility across all datasets. For spatial alignment, we initially obtained data at their original resolutions (see Table S1, Resolution). Subsequently, we used a 0.25° land area grid template derived from the Gridded Population of the World, Version 4 (GPWv4): Land and Water Area database. We aggregated all predictors with bilinear interpolation to 0.25° spatial resolution (WGS84) to match with the land area data resolution. For temporal alignment, our dynamic predictors (e.g., climate variables and land use categories), cover the period from 1950 to 2020 at five-year intervals. We matched each spillover event to the nearest sampling year to ensure predictors accurately reflected conditions at the time of occurrence (detailed in Excel Table S1, with events matched to their nearest five-year interval).

Although collinearity does not affect the performance of machine learning models, it can distort variable importance and bias interpretations[9]. The relationships between variables were evaluated using Spearman's rank and Pearson's correlation tests (Figure 2, Figure S4). We set a correlation coefficient threshold of greater than 0.7 to identify pairs of variables with high collinearity[41]. For each pair that exceeded this threshold, one variable was retained based on its relevance to our research objectives and a priori empirical knowledge. This selection process resulted in a final set of 45 variables for our modeling framework.

To spatially represent each of the 226 documented spillover events, we developed a systematic approach for polygon construction. Events with accurate coordinates were assigned a 5-km radius circular buffer centered on their coordinates, rather than confining them to an administrative boundary. For locations lacking exact coordinates but identified at sub-national levels, we employed administrative boundary data from the Global Administrative Areas (GADM) Version 3.6 database. We identified specific administrative areas using GADM's unique geographic identifiers (GID). Events only reported at the country- or continent-level were considered coarse and excluded to decrease geographic uncertainty. The resulting polygons were used to clip the template land area data, and the extracted candidate grid cells for each event were subsequently used in our bootstrap resampling workflow.

Bootstrap resampling and stacking framework

We employed a systematic resampling strategy to evaluate the predictive power of our models, address spatial uncertainty in spillover events, and generate quantile intervals that capture both sampling and spatial uncertainty. The procedure began with bootstrap sampling (sampling with replacement) of all events to obtain training event sets and out-of-bag (OOB) validation sets. For each documented event (presence samples) within these sets, we randomly selected one presence grid cell from the candidate grids within the event polygon boundary. For absence samples, we randomly sampled from grids outside the event's candidate grids, maintaining a balanced 1:1 case-control ratio. This sampling strategy is well-established in environmental and ecological modeling due to its demonstrated high specificity[42]. To account for bias from land area in WGS84 grid cells, we weighted grid cell selection probability proportionally to land area for both presence and absence

samples (Figure S2a). The entire resampling process was repeated 600 times to account for spatial uncertainty in historical records, generating 600 unique bootstrap resampling datasets, each containing 226 presence grids and 226 pseudo-absence grids.

We employed an ensemble stacking pipeline[43], combining various machine learning techniques for prediction. The pipeline comprised two levels: base learners were fitted to form "level-one" predictions, while a meta learner was fitted to find the optimal combination of these "level-one" predictions. The base learners include eXtreme Gradient Boosting (XgBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost)), Adaptive Boosting (AdaBoost), and Gradient Boosting Machine (GBM). The Meta-estimator comprised Logistic Regression (LoR), Decision Tree (DT) and Bagging Classifier (BC) (see Supplement 1 for model selection details). We initially divided bootstrap datasets (n =600) into six splits, five folds for base-learners training and one fold for meta-learner training. A grid search process was conducted to identify the optimal meta-learner and hyperparameters (Tables S4 for hyperparameter optimization ranges). We also evaluated individual base learner, these single models are compared with the stacking pipeline to validate robustness of performance (Table 2).

Validating and evaluating ML model performance

We evaluated our model's performance using multiple metrics and validation approaches. The Area Under the Curve (AUC) served as the primary metric, measuring the model's ability to correctly rank presence over absence sites[44,45]. However, since AUC is threshold-independent and may include unrealistic probability thresholds, we complemented it with the True Skill Statistic (TSS), to provide a more robust evaluation of probability predictions[42,46]. For hyperparameter optimization, we employed advanced search algorithms—Bayesian optimization and Tree of Parzen Estimators (TPE)—which efficiently identify optimal parameters with minimal computational cost[47,48].

For each bootstrap resampling dataset, we selected one hyperparameter configuration and conducted 10-fold cross-validation, calculating the mean Area Under the Curve (AUC) and True Skill Statistic (TSS) scores from the test fold. To address potential overfitting from hyperparameter tuning, we

employed out-of-bag (OOB) datasets—consisting of events excluded during replacement-based resampling—to calculate OOB AUC and TSS scores. This validation method provided an unbiased estimate of model performance without requiring a separate test set[49].

Cross-validation tests demonstrated robust performance across all models (Table 2, Figure S6). However, when evaluated on out-of-bag (OOB) test sets, individual models exhibited inferior performance, indicating potential overfitting. The Stacking models effectively mitigated this issue, with the Stacking3 configuration achieving superior OOB performance metrics (AUC: 0.88, TSS: 0.57; Table 2) and minimal prediction uncertainty (Figure S12). Based on its superior predictive ability, we selected the Stacking3 model for transmission-specific analyses. We extracted and categorized zoonotic spillover events according to their primary transmission types (direct, vector-borne, oral, and airborne, Figure S2c). Subsequently, we applied the Stacking3 architecture to develop transmission-specific sub-models following the previously described training procedures. These transmission-specific models also demonstrated good performance, with average AUC values ranging from 0.86 to 0.95 and average TSS values from 0.54 to 0.76 (Table S5).

Adjusting for observation bias

Based on the superior predictive performance demonstrated by the Stacking3 model, we selected this architecture for risk assessment and subsequent predictor analysis. However, imbalances in infectious disease reporting and surveillance have long been recognized, with significant variations linked to regional socioeconomic conditions and healthcare infrastructure investments[34]. These reporting imbalances constrain our ability to accurately assess spillover emergence risk. We hypothesized that documented spillover events are correlated with reporting adequacy distribution patterns, suggesting that our model outputs likely reflect reported risk patterns rather than actual emerging spillover burden, with discrepancies attributable to reporting bias[34,50]. Addressing this reporting bias is critical for identifying potentially high-risk regions with limited reporting and healthcare resources. Therefore, we employed optimized stacking model to predict reported spillover risk (Figure 3a) and generated risk maps adjusted for relative reporting adequacy (Figure

3c), with relative reporting adequacy serving as a proxy for observational and reporting bias estimates.

Our analysis advances previous work by using disease burden as theoretical optimal distribution of reporting effort, enabling observation bias quantification through a relative reporting adequacy index. Our analysis assume that reporting effort comprises two primary components: observation bias and the multiplicative effect of disease burden on spillover risk. We assumed that optimal reporting effort should proportionally align with population disease burden, although socioeconomic factors may cause deviations from this theoretical distribution[34]. While additional unmeasured factors likely influence this relationship, our assumption enables calculating relative reporting adequacy as the ratio of reporting effort to disease burden per cell, representing "observation bias" between actual and theoretically optimal reporting efforts.

$$\text{Relative reporting adequacy} \propto \frac{\text{Reporting efforts}}{\text{Human disease burden}} \quad (2)$$

When bias can be quantified, the underlying true distribution can be estimated through bias correction known as "factoring bias out[50,51]". This typically involves dividing by the measured sampling effort, assuming optimal sampling effort should be uniformly distributed across space[17]. Our analysis hypothesized that the optimal reporting effort distribution for spillover events should be proportionate to human disease burden. Therefore, we factored out reporting effort layer and then incorporated the theoretically optimal effort.

$$\text{Adjusted relative risk} \propto \frac{\text{Reported relative risk}}{\text{Relative reporting adequacy}} \quad (3)$$

By adjusting for reporting adequacy, we generate a hypothetical risk index that better reflects the true distribution of spillover events.

$$\text{Adjusted relative risk} \propto \text{Reported relative risk} \times \frac{\text{Human disease burden}}{\text{Reporting efforts}} \quad (4)$$

Identifying key predictors, correlations and pathways affecting zoonotic spillover

Our main model employed bootstrap resampling to fit 600 replicate stacking models. These models were used to calculate Partial Dependence with 95% confidence intervals and employed Shaply Additive Explanation (SHAP) values —a game-theoretic AI interpretability framework[52]—to

identify key predictors (Figure 2a, Figure S7). SHAP values quantify variable importance by evaluating model predictions across all possible variable combinations, with variables ranked by their mean contribution to individual observations[13]. Partial Dependence plots with empirical 95% confidence intervals were generated to illustrate the relationships between spillover events and predictors (Figure 2b, Figure S8). This approach demonstrates the model's response to an indicidual variable while maintaining all other predictors at constant values[17]. Moreover, given the critical role of administrative entities in zoonotic disease surveillance and prevention, we analyzed adjusted spillover risk across regions and countries, investigated the spatial overlapping between human modification and adjusted risk patterns (Figure 4), and examined relationships between key factors and adjusted spillover risk at the national scale (Figure S9).

Previous study noted the accelerating natural habitat loss between 1990-2015 due to anthropogenic pressures[53], which reshaped habitat structures and human-animal interactions, raising zoonotic infections risk in humans. In this context, we used a Structural Equation Model (SEM) to disentangle direct and human-mediated pathways influencing zoonotic spillover emergence (Fig. 5), The analysis incorporated key predictors identified through high SHAP importance in our stacking model and strong correlations in partial dependence plots (e.g. human modification, urban, horse, elevation). Using confirmatory factor analyses, the SEM evaluated our pre-defined hypothetical model (meta SEM model) within a meta-pathway framework (Figure S10) to jointly estimated all predictors.

Our main SEM model examined the hypothesized direct and indirect effects of key predictors on zoonotic spillover emergence from 1990-2020 using the bootstrap resampling dataset. The model was constructed using maximum likelihood estimation in AMOS 26 software to fit the covariance matrix. After model fitting, we removed non-significant pathways ($p > 0.05$) to obtain the final main model (Figure 5, Table S7). Land-use modification gradients, ranging from low to high intensity, were characterized by human-environment interactions mediated through vegetation and forest, livestock farming, cropping intensity, and urbanization. We established connections between high- and low-intensity land modifications to investigate potential land conversion processes affecting

zoonotic spillover risk. To control for reporting bias, we included a relative reporting adequacy index as a covariate. We assessed model fit using standardized coefficients, Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMSR), with good fit indicated by RMSEA < 0.05 and SRMSR < 0.08 for models with sample sizes exceeding 200 (n > 200)[8]. The indirect effects of human modification on emerging zoonotic spillover, mediated by livestock farming, vegetation and forest, cropping intensity, and urban (Figure 5b), are detailed in Table S6. To investigate the explanatory role of human modification on new zoonotic spillover, we also developed a comparative SEM model excluding human modification (Figure S11, Table S8). The final main model demonstrated good fit (SRMSR = 0.021; RMSEA = 0.061, 90% CI: 0.060-0.062), with RMSEA slightly exceeding the suggested 0.05 threshold. Therefore, we proceeded with interpreting the final main model relationships and coefficients without further pathway adjustments.

Results

Global patterns of emerging zoonotic spillover events

Our study analyzed 226 documented zoonotic spillover events worldwide between 1940-2020 (Figure S2). Spatial analysis revealed a non-uniform distribution, with major events reported in north America (particularly the United States), western Europe, southern Brazil, and eastern Asia. From 1940 to 2020, spillover events occurred continuously, with significant increases observed in the late 20th centuries, showing distinct geographical and temporal heterogeneity across regions (Figure S2b). Regarding transmission modes, direct transmission was predominant, followed by vector-borne transmission, with oral and airborne routes also playing substantial roles (Figure S2 c, middle panel). While wildlife-derived pathogens constituted the majority of events, highlighting their critical role in zoonotic emergence, a substantial proportion remained etiologically unclassified (Figure S2 c, right panel). Regional analysis confirmed North America's predominance across all categories, with significant contributions from Asia and Europe. These findings suggest complex interactions between geographical distribution, pathogen characteristics, and transmission mechanisms in shaping global zoonotic spillover patterns.

Impact of anthropogenic, environmental, and climatic factors on zoonotic

spillover risk

Our analysis revealed anthropogenic factors as the predominant drivers of spillover events. Areas with spillover presence typically exhibit higher levels of human modification, urbanization, and C3 annual crop coverage (Figure S5). These findings align with our machine learning analyses, where human modification and related factors — including urban land use, changes in population density, change in C3 annual cropping intensity, and livestock (e.g., horse and goat) headcount— emerged as crucial predictors, evidenced by their high mean SHAP values (Figure 2a). In addition, environmental and topographic factors, such as 2-meter temperature and elevation 2-meter temperature seasonality, also contributed significantly to the model's predictions, although their overall impact was less pronounced to human-related factors. We further investigated the model's response to these predictors using partial dependence plots (Figure 2b). In general, the predicted risk increased gradually with the human modification index but accelerated sharply at moderate modification levels. Urban land use, human modification index, population density change, horse headcount, evergreen/deciduous needleleaf trees, and total precipitation had an overall positive correlation with the model's response. By contrast, elevation and 2-meter temperature seasonality exhibited a negative correlation. The temperature at 2 meters and intensity of C3 annual crops, and C3 annual crops change showed a specific trend, with higher risk values occurring at median values of the variables.

Predicted global distribution of zoonotic spillover risk

The reported risk of emerging zoonotic spillover showed strong spatial correlation with developed regions and major population centers, particularly in the eastern United States, Europe, South Korea, and Japan (Figure 3a). Significant risks were also identified in other regions with varying levels of economic development, including the Caribbean, Central America, the Sahel, Ethiopian highlands and the Rift Valley, southern Africa, northern India, eastern China, coastal regions of maritime Southeast Asia (e.g. southern Thailand, Java, and parts of the Philippines). After adjusting for relative reporting adequacy, the Caribbean, central America, Andean Latin America, eastern China, India, and extensive regions in sub-Saharan Africa and Southeast Asia showed higher adjusted spillover risks (Figure 4c). Within these regions, the adjusted risk distribution remained

heterogeneous (Figure 4a). Despite varying degrees of correlation strength across different super-regions (Figure 4b, divided by epidemiological similarity and geographic closeness)[54], human modification generally showed a positive relationship with adjusted risk patterns. This correlation was particularly strong in regions experiencing intensive human pressure, which exhibited higher adjusted spillover risk (Figure 4c, Figure 4d). The relationship between adjusted risk patterns and anthropogenic interactive factors across land-modification gradients was also apparent. The response curve suggested that increases in the human modification index, horse and goat headcount, and intensity of C3 annual crops were associated with elevated zoonotic spillover risk at the national scale (Figure S9).

Pathways by which human pressure affects zoonotic spillover risk

The stacking method identified critical factors influencing zoonotic spillover events over the past three decades, including environment and topography, human modification, vegetation and forest, livestock headcount, cropping intensity and urbanization (Figure 5a). These factors represented a spectrum of anthropogenic interactions with the environment and animals, ranging from low to high intensity. Our main SEM model improved the explanatory power for factors of anthropogenic interaction and zoonotic spillover emergence compared to the SEM model excluding human modification (Table S9). Human modification affected the emergence of zoonotic spillover by increasing urban land use, C3 annual crops, Deciduous Broadleaf Trees, and horse headcount. The standardized indirect effects of human modification mediated by C3 annual crops had the greatest influence on the emergence of zoonotic spillover (Figure 5b). In contrast, livestock headcount and vegetation variables representing relative lower-intensity human modification, exhibited relative lower correlations (standardized indirect effects <0.05). In addition, human modification had a pronounced negative effect on the relative reporting adequacy index, which ultimately had a negative correlation with zoonotic spillover risk. Climatic variables also significantly affected zoonotic emergence, with total precipitation and 2-meter temperature seasonality showing strong associations and high pathway coefficients (0.15 for total precipitation and -0.15 for 2-meter temperature seasonality, Table S7). Although urban land use extent positively impacted population density change (0.40), its correlation with zoonotic emergence was weel (0.05).

Discussion

Spatial risk patterns across different intensity gradients of anthropogenic modification

Our analyses demonstrate a positive correlation between spillover risk and multiple anthropogenic pressures across intensity gradients of land modification. Zoonotic spillover emergence was strongly mediated by medium to high-intensity land modifications, with both cropping intensity and urbanization showed higher path coefficient (Figure 5a). The adjusted spillover risk map reflects this pattern, identifying extensive regions worldwide where high human modification coincides with emerging risks (Figure 4). Rapid changes in human-animal interfaces, driven by increasing human influence on agricultural landscapes, are likely the primary cause of zoonotic spillover[55]. Although our partial dependence plots indicate a modest positive correlation between C3 annual crops and reported spillover risk, this relationship becomes more pronounced after adjusting for relative reporting adequacy (Figure S9). In our main SEM model, this predictor emerges as the most significant anthropogenic variable contributing to zoonotic spillover emergence across land modification gradients (Figure 5). These findings suggests that specific cropping intensity may serve as an effective proxy for pathogen biological characteristics and cross-species transmission, mediated by ecosystem disturbances or alterations in host-human interaction patterns[56,57].

Rapid deforestation, agricultural expansion, and urbanization pose significant ecological risks by fragmenting natural habitats and profoundly impacting trophic structures and native wildlife species. Animal-human interactions may change depending on how animal habitats are affected by anthropogenic pressures. Evidence indicates that wildlife species that serve as pathogen hosts are relatively more abundant in managed landscapes than in nearby undisturbed sites[58]. Across gradients of anthropogenic landscape modification, these ecological relationships manifest in diverse forms and generate impacts of varying intensities. For example, the establishment of pastures, plantations, or intensive livestock farms near forest edges may increase the transmission of pathogens from wildlife to humans[59,60]. The processes of deforestation and forest fragmentation reshape the dynamics of wildlife communities, potentially leading to the extinction of habitat-specialist species while allowing generalist and pathogen-hosting species to thrive. Moreover, resource shortages,

large-scale migration, social disparity, environmental pollution, and poor living conditions arising from unplanned urbanization result in habitat and resource provision variation, disproportionately amplifying the uncertainty of zoonotic spillover[12]. These findings underscore the critical necessity of considering various human activities and adopting responsive management strategies for landscapes, animals, and vectors associated with these changes. Effective strategies should integrate landscape management with public health initiatives to mitigate the risks of zoonotic spillover.

Socio-ecological drivers of emerging zoonotic disease

Our modeling framework effectively identified critical socio-ecological predictors and quantified their relative contributions to prediction accuracy. The SHAP (SHapley Additive exPlanations) analysis indicated that variables with broader ranges often indicate a significant contribution of variables and notable variability in the models' response to feature values. Specific human-related factors (e.g., human modification, population density change, C3 annual crop change) and environmental predictors (temperature at 2 meters and elevation) had high mean SHAP values and wide distribution ranges, indicating their high contribution and multidimensional influence on spatial risk assessment. In contrast, variables with narrow SHAP value distributions demonstrated more constrained responses to feature variations, suggesting specific or limited impacts. Our analysis revealed that certain variables exhibited more concentrated SHAP values within specific feature ranges. For example, higher elevations were associated with lower SHAP values, whereas human modification, population density change, and horse headcount displayed opposite patterns.

In our machine learning analyses, incorporating both SHAP analysis and partial dependence plots, population density change emerged as crucial predictor strongly correlated with zoonotic spillover emergence. However, the Structural Equation Modeling (SEM) results indicated a positive but modest influence of population density change (Figure 5). Notably, urbanization independently exerted a significant positive effect on the emergence of zoonotic diseases while simultaneously influencing population density changes. This interconnected relationship may explain the strong association between population density changes and zoonotic emergence observed in the machine learning models. Our SEM model attempted to disentangle these complex relationships, which had

not been adequately addressed previously. Moreover, the adverse effects of human modification on the relative reporting adequacy index in the SEM model may imply a disproportionately high disease burden or insufficient reporting practices. The former could alter the course of pathogen transmission, while the latter may distort the spatial risk pattern.

Our findings indicated that factors related to livestock and poultry were critical, although the degree and tendency of correlation varied across livestock species (Figure 2, Figure S7). As human societies have developed, each era of the livestock revolution has introduced new health challenges and opportunities for the emergence of zoonotic pathogens. The interactions between domestic animals and humans across rural and urban landscapes provide extensive opportunities for pathogen transmission[12]. This is particularly evident in informal settlements characterized by rapid urbanization and high population density, where inadequate livestock management practices can result in environmental degradation and elevated disease transmission risks[61]. Integrating a broader range of sentinel animal species into the modeling framework could enhance predictive power iteratively. Disease surveillance in domestic animal populations is an early indicator of potential emergence hotspots, enabling the identification of new pathogens of concern[62]. However, continuous monitoring of multiple species, which may function as sentinel animals, hosts, or vectors, requires substantial resources and must be prioritized based on time and environmental changes[63].

Although our stacking model performed well, researchers should cautiously approach the interpretation of key livestock drivers identified by this modeling framework. Notably, while horse and goat headcount emerged as a significant predictors (Figure 2, Figure S7), their influence on zoonotic spillover emergence became slight when complex factors were considered in the SEM model. This could be attributable to the relatively specific distribution of these animals or confounding geographic association of zoonotic spillover with land use or climatic factors. For example, the distribution of horses is specific to certain global regions, mirroring diverse uses across different areas. Horse may be proxies for other factors, such as specific farming systems, cultural practices, transportation modes, or socioeconomic activities[64].

Climatic predictors also distinctly improved the ML model's predictive power. Previous studies have illustrated that climate change restricts the distribution and biology of hosts and vectors, influencing the first encounters and viruses' transmission between hosts[5,65,66]. However, the complex response curve of climate predictors (Figure 2a, such as 2m temperature) highlighted the necessity of detailed, region-specific studies on climate impacts. The interplay between human activities and climate change is complex, and pinpointing precise mechanisms is challenging, given that climatic factors are confounded by the geographic associations of different pathogens with land and other variables[9]. Our study did not explore specific mechanisms through which climate change affects the emergence of animal-human interfaces and spatial risk variation; instead, climate variables were treated as covariates in the ML model, assuming direct influences on boto anthropogenic factors and zoonotic spillover emergence to reduce interpretive complexities.

Implications for prevention policy of emerging zoonoses

This study enhances the prediction and explanation of zoonotic spillover emergence in several essential ways. We incorporated advanced machine learning algorithms and spatial data mining techniques, focusing on refined anthropogenic factors. Our analysis also accounted for human disease burden and reporting efforts, leading to the creation of a relative reporting adequacy index that addresses observational and reporting bias in the spatial risk pattern of "reported new zoonotic spillover", enabling us to identify cost-effective areas for surveillance. By combining advanced stacking algorithms with comprehensive predictor variables and explainable AI techniques, our modeling framework enhances zoonotic spillover prediction and identifies critical correlations. The stacking models achieved superior predictive performance, with the Stacking3 model attaining AUC scores above 0.90 in both cross-validation and out-of-bag validation sets, while maintaining robust performance (0.86-0.93) in transmission-specific predictions (Table S5). These results represent a notable improvement than that used individual machine learning algorithms.

Our analysis reveals previously unidentified spatial patterns of zoonotic spillover risk, including high levels of "reported spillover risk" in regions with varying degrees of economic development, such as northern India, eastern Asia and eastern Europe—areas that extend the traditionally

identified tropical regions (Figure 3a). After accounting for observational and reporting bias, our findings suggest that actual spillover risks remain substantially underestimated in extensive regions within continents such as sub-Saharan Africa and Southeast Asia, highlighting the need for enhanced surveillance in these regions. Areas experiencing high-intensity human modification are likely to present substantially high spillover risks (Figure 4c, Figure 4d). The complexity underlying zoonotic spillover emergence is underscored by the impacts of various environmental and anthropogenic determinants, which contribute to specific and distinct risks across different regions[67].

Our confirmative pathway analyses also provide quantitative evidence for pathogen spillover risks across gradients of human-modified landscape. While previous research noted higher proportions of human-pathogen hosts in modified environments[58], our study quantifies these relationships by integrating multiple observational factors, including vegetation and forest, agricultural cropping intensity, livestock farming, and urbanization. Through comprehensive analysis of these anthropogenic factors, we provide empirical evidence that elucidates disease spillover risk patterns, demonstrating that moderate to high levels of human modification significantly facilitates novel pathogen spillover events.

There are several limitations to our approach that future work should address. First, despite accounting for reporting bias and attempting to establish temporal relationships between predictors and events, the new zoonotic spillover events extracted from published literature and reports remain subject to generalizability constraints. Additionally, the predictors of disease spillover emergence may not fully align with the drivers of disease outbreak and human disease burden[9]. The relationship between reporting adequacy metrics and actual reporting bias for spillover events requires further validation. These discrepancies could affect our adjustment of reported spillover risk, potentially impacting pandemic monitoring and prevention across different regions. Second, the inherent uncertainty and stochastic nature of zoonotic spillover emergence, combined with limitations in predictor availability, restricted our ability to precisely forecast risk pattern changes at fine temporal scales. Third, while our study incorporated comprehensive variables and accounted for temporal effects, these variables are subject to measurement errors and inter-variable autocorrelation. This

limitation is compounded by the likelihood of misdiagnosed and unreported cases, particularly in cases of asymptomatic infections[41]. Emerging zoonotic spillover events may involve undiscovered confounding factors, and the complex interactions among variables may necessitate additional methodological approaches to adequately capture underlying mechanisms. Lastly, as noted in previous research, each pathogen possesses unique biological and ecological characteristics[67]. These pathogen-specific differences introduce taxonomic specificity to our analysis, potentially complicating models' identification of processes for new spillover events.

Future research should focus on understanding the specific mechanisms and dominant causal relationships that drive zoonotic spillover events. While our pathway-specific analyses help address knowledge gaps in zoonotic emergence and inform targeted control measures, more robust evidence requires interventional or randomized controlled studies across distinct pathogen taxa. Integrating findings across pathogen groups can reveal common mechanisms underlying spillover events that extend beyond observational data. Models of zoonotic spillover emergence should strike a balance between capturing complex infection dynamics and maintaining efficiency, interpretability, and adaptability. This necessitates integrating machine learning frameworks with mechanistic models, as the latter reveal causal relationships by isolating specific pathways through refined, continuously updated parameters, thereby avoiding reliance on correlations between observed occurrences (which may themselves be inherently biased by e.g. observation effort), environmental covariates, and their statistical extrapolations[68,69]. Moreover, since spillover determinants may vary temporally, robust models require comprehensive longitudinal surveillance data, iterative projections, and continuous updates as new events emerge[68]. Urgent research is needed to examine current trends in zoonotic spillover emergence under various scenarios and fine-scale changes in governing rules. Such understanding will enable precise design of human development and environmental interaction pathways, identifying areas with elevated spillover risks.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (No. 42477465, No. 42311530697)

All spatial data processing, data matching, machine learning model construction, hyperparameter tuning, and SHAP plots were completed using Python 3.9.13. R 4.4.1 was used to generate partial dependence plots. SEM models and performance reports were generated using Amos 26. The datasets used in model training are included in Supplemental Excel File, the shape files of new zoonotic spillover events and code are available from the corresponding author on reasonable request.

Reference

1. Ellis EC, Gauthier N, Klein Goldewijk K, Bliege Bird R, Boivin N, Díaz S, et al. 2021. People have shaped most of terrestrial nature for at least 12,000 years. Proc Natl Acad Sci U S A 118: e2023483118.
2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. 2008. Global trends in emerging infectious diseases. Nature 451:990-993.
3. Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JR, Dobson AP, et al. 2009. Epidemic dynamics at the human-animal interface. Science 326:1362-1367.
4. Pike J, Bogich T, Elwood S, Finnoff DC, Daszak P. 2014. Economic optimization of a global strategy to address the pandemic threat. Proc Natl Acad Sci U S A 111:18519-18523.
5. Carlson. CJ, Albery. GF, Merow. C, Trisos. CH, Zipfel. CM, Eskew. EA, et al. 2022. Climate change increases cross-species viral transmission risk. Nature 607:555-562.
6. Gregory F. Albery, Evan A. Eskew 1, Noam Ross 1, Olival KJ. 2020. Predicting the global mammalian viral sharing network using phylogeography. Nat Commun 11:2260.
7. Warmuth VM, Metzler D, Zamora-Gutierrez V. 2023. Human disturbance increases coronavirus prevalence in bats. Sci Adv 9:eadd0688.
8. Ecke F, Han BA, Hörnfeldt B, Khalil H, Magnusson M, Singh NJ, et al. 2022. Population fluctuations and synanthropy explain transmission risk in rodent-borne zoonoses. Nat Commun 13: 7532.
9. Skinner EB, Glidden CK, MacDonald AJ, Mordecai EA. 2023. Human footprint is associated with shifts in the assemblages of major vector-borne diseases. Nat Sustain 6:652-661.
10. Kennedy CM, Oakleaf JR, Theobald DM, Baruch-Mordo S, Kiesecker J. 2019. Managing the middle: A shift in conservation priorities based on the global human modification gradient. Glob Chang Biol 25:811-826.
11. Schmiedel I, Bergmeier E, Culmsee H. 2015. Plant species richness patterns along a gradient of landscape modification intensity in lower saxony, germany. Landscape Urban Plan 141:41-51.
12. Hassell JM, Bettridge JM, Ward MJ, Ogendo A, Imboma T, Muloi D, et al. 2021. Socio-ecological drivers of vertebrate biodiversity and human-animal interfaces across an urban landscape. Glob Chang Biol 27:781-792.
13. Farooq Z, Rocklov J, Wallin J, Abiri N, Sewe MO, Sjodin H, et al. 2022. Artificial intelligence to predict west nile virus outbreaks with eco-climatic drivers. Lancet Reg Health Eur 17:100370.

14. Peters DPC, McVey DS, Elias EH, Pelzel-McCluskey AM, Derner JD, Burruss ND, et al. 2020. Big data-model integration and ai for vector-borne disease prediction. Ecosphere 11:e03157.
15. Parselia E, Kontoes C, Tsouni A, Hadjichristodoulou C, Kioutsioukis I, Magiorkinis G, et al. 2019. Satellite earth observation data in epidemiological modeling of malaria, dengue and west nile virus: A scoping review. Remote Sensing 11:1862.
16. Taylor LH, Woolhouse MEJ, Dye C, Latham SM, woolhouse MEJ. 2001. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci 356:983-989.
17. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. 2017. Global hotspots and correlates of emerging zoonotic diseases. Nat Commun 8:1124.
18. Theobald DM. 2013. A general model to quantify ecological integrity for landscape assessments and us application. Landscape Ecol 28:1859-1874.
19. Bonham-Carter G. 1994. Geographic information systems for geoscientists: Modelling with GIS. 1995. Oxford, UK: Pergamon/Elsevier.
20. Goldewijk KK. 2001. Estimating global land use change over the past 300 years: The hyde database. Global Biogeochem Cy 15:417-433.
21. Kilpatrick AM, Randolph SE. 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. Lancet 380:1946-1955.
22. Taylor C, Brodbelt DC, Dobson B, Catchpole B, O'Neill DG, Stevens KB. 2021. Spatio-temporal distribution and agroecological factors associated with canine leptospirosis in great britain. Prev Vet Med 193:105407.
23. Rondinini C, Di Marco M, Chiozza F, Santulli G, Baisero D, Visconti P, et al. 2011. Global habitat suitability models of terrestrial mammals. Philos Trans R Soc Lond B Biol Sci 366:2633-2641.
24. Jenkins CN, Pimm SL, Joppa LN. 2013. Global patterns of terrestrial vertebrate diversity and conservation. Proc Natl Acad Sci U S A 110:E2602-2610.
25. Hurtt GC, Chini L, Sahajpal R, Frolking S, Bodirsky BL, Calvin K, et al. 2020. Harmonization of global land use change and management for the period 850–2100 (luh2) for cmip6. Geosci Model Dev 13:5425-5464.
26. Hurtt GC, Chini LP, Frolking S, Betts RA, Feddema J, Fischer G, et al. 2011. Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands. Climatic Change 109:117-161.
27. Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, et al. 2012. Ecology of zoonoses: Natural and unnatural histories. Lancet 380:1936-1945.
28. Robinson TP, Wint GR, Conchedda G, Van Boeckel TP, Ercoli V, Palamara E, et al. 2014. Mapping the global distribution of livestock. PLoS One 9:e96084.
29. Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, et al. 2013. A strategy to estimate unknown viral diversity in mammals. mBio 4:e00598-00513.
30. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. 2017. Host and viral traits predict zoonotic spillover from mammals. Nature 546:646-650.
31. Marcolin L, Tonelli A, Di Marco M. 2024. Early-stage loss of ecological integrity drives the risk of zoonotic disease emergence. J R Soc Interface 21:20230733.
32. Dunn RR, Davies TJ, Harris NC, Gavin MC. 2010. Global drivers of human pathogen richness and prevalence. Proc Biol Sci 277:2587-2595.

33. 2015. Socioeconomic data and applications center (sedac). 2015 release (2013). Global mammal richness grids. Available at: https://sedac.Ciesin.Columbia.Edu/data/set/species-global-mammal-richness-2015.
34. Yang K, LeJeune J, Alsdorf D, Lu B, Shum CK, Liang S. 2012. Global distribution of outbreaks of water-associated infectious diseases. PLoS Negl Trop Dis 6:e1483.
35. Global burden of disease collaborative network. Global burden of disease study 2021 (gbd 2021) results. Seattle, United States: Institute for health metrics and evaluation (ihme), 2022. Available from https://vizhub.Healthdata.Org/gbd-results/.
36. Ostroumova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. 2018. Catboost: Unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada:Curran Associates Inc., 6639-6649.
37. Nelson A, Weiss DJ, van Etten J, Cattaneo A, McMenomy TS, Koo J. 2019. A suite of global accessibility indicators. Sci Data 6:266.
38. Kummu M, Taka M, Guillaume JHA. 2018. Gridded global datasets for gross domestic product and human development index over 1990-2015. Sci Data 5:180004.
39. Barber RM, Fullman N, Sorensen RJD, Bollyky T, McKee M, Nolte E, et al. 2017. Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990-2015: A novel analysis from the global burden of disease study 2015. Lancet 390:231-266.
40. Shmueli G. 2010. To explain or to predict? Stat Sci 25:289-310.
41. Ratner B. 2009. The correlation coefficient: Its values range between +1/−1, or do they? J Target Meas Anal Mark 17:139-142.
42. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods Ecol Evol 3:327-338.
43. Wolpert DH. 1992. Stacked generalization. Neural Networks 5:241-259.
44. Liu C, White M, Newell G. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. Ecography 34:232-243.
45. Fawcett T. 2006. An introduction to roc analysis. Pattern Recogn Lett 27:861-874.
46. Lobo JM, Jiménez-Valverde A, Real R. 2008. Auc: A misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17:145-151.
47. Watanabe S. 2023. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. ArXiv abs/2304.11127.
48. Bergstra J, Yamins D, Cox DD. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. Atlanta, GA, USA:JMLR.org, I–115–I–123.
49. Breiman L. 2001. Random forests. Mach Learn 45:5-32.
50. Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecol Appl 19:181-197.
51. Dudík M, Schapire RE, Phillips SJ. 2005. Correcting sample selection bias in maximum entropy density estimation. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada:MIT Press, 323-330.
52. Wieland R, Kuhls K, Lentz HHK, Conraths F, Kampen H, Werner D. 2021. Combined climate

and regional mosquito habitat model based on machine learning. Ecol Model 452:109594.

53. Theobald DM, Kennedy C, Chen B, Oakleaf J, Baruch-Mordo S, Kiesecker J. 2020. Earth transformed: Detailed mapping of global human modification from 1990 to 2017. Earth Sys Sci Data 12:1953-1972.
54. Bhattacharjee NV, Schumacher AE, Aali A, Abate YH, Abbasgholizadeh R, Abbasian M, et al. 2024. Global fertility in 204 countries and territories, 1950-2021, with forecasts to 2100: A comprehensive demographic analysis for the global burden of disease study 2021. Lancet 403:2057-2099.
55. Hassell JM, Begon M, Ward MJ, Fevre EM. 2017. Urbanization and disease emergence: Dynamics at the wildlife-livestock-human interface. Trends Ecol Evol 32:55-67.
56. Rulli MC, D'Odorico P, Galli N, Hayman DTS. 2021. Land-use change and the livestock revolution increase the risk of zoonotic coronavirus transmission from rhinolophid bats. Nat Food 2:409-416.
57. Jones BA, Grace D, Kock R, Alonso S, Rushton J, Said MY, et al. 2013. Zoonosis emergence linked to agricultural intensification and environmental change. Proc Natl Acad Sci U S A 110:8399-8404.
58. Gibb R, Redding DW, Chin KQ, Donnelly CA, Blackburn TM, Newbold T, et al. 2020. Zoonotic host diversity increases in human-dominated ecosystems. Nature 584:398-402.
59. Bloomfield LSP, McIntosh TL, Lambin EF. 2020. Habitat fragmentation, livelihood behaviors, and contact between people and nonhuman primates in africa. Landscape Ecol 35:985-1000.
60. Zhou P, Fan H, Lan T, Yang X-L, Shi W-F, Zhang W, et al. 2018. Fatal swine acute diarrhoea syndrome caused by an hku2-related coronavirus of bat origin. Nature 556:255-258.
61. Hassell JM, Ward MJ, Muloi D, Bettridge JM, Phan H, Robinson TP, et al. 2019. Deterministic processes structure bacterial genetic communities across an urban landscape. Nat Commun 10:2643.
62. Duggett N, Ellington MJ, Hopkins KL, Ellaby N, Randall L, Lemma F, et al. 2021. Detection in livestock of the human pandemic escherichia coli st131 fimh30(r) clone carrying blactx-m-27. J Antimicrob Chemother 76:263-265.
63. Oeschger TM, McCloskey DS, Buchmann RM, Choubal AM, Boza JM, Mehta S, et al. 2021. Early warning diagnostics for emerging infectious diseases in developing into late-stage pandemics. Acc Chem Res 54:3656-3666.
64. Newsome D, Smith A, Moore SA. 2008. Horse riding in protected areas: A critical review and implications for research and management. Curr Issues Tour 11:144-166.
65. Mordecai EA, Caldwell JM, Grossman MK, Lippi CA, Johnson LR, Neira M, et al. 2019. Thermal biology of mosquito-borne disease. Ecol Lett 22:1690-1708.
66. Gage KL, Burkot TR, Eisen RJ, Hayes EB. 2008. Climate and vectorborne diseases. Am J Prev Med 35:436-450.
67. Loh EH, Zambrana-Torrelio C, Olival KJ, Bogich TL, Johnson CK, Mazet JA, et al. 2015. Targeting transmission pathways for emerging zoonotic disease surveillance and control. Vector Borne Zoonotic Dis 15:432-437.
68. Roberts M, Dobson A, Restif O, Wells K. 2021. Challenges in modelling the dynamics of infectious diseases at the wildlife-human interface. Epidemics 37:100523.
69. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, et al. 2019. Integrating machine learning and multiscale modeling-perspectives, challenges, and

opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med 2: 115.

Tables

Table 1 Description of stressors used in the global Human modification index (HMc)[10]

Class	Stressor	Description
Human settlement	Population density	Population density derives from the Gridded Population of World dataset with 2015 UN adjusted values, which were log[X+1] transformed, producing values ranging from 0 to 3.628.
	Built-up areas	Building density per cell (0-255, representing 0-100% coverage), using the Global Human Settlements Layer (GHSL)
Agriculture	Cropland	Percentage of annually cultivated areas at a 250-m resolution, based on harmonizing global, regional, and national datasets.
	Livestock	Livestock data derives from the Gridded Livestock of the World v2 database, converting cattle, sheep, and goat populations to standardized livestock units (LU) per km² using global average coefficients.
Transportation	Major roads, minor roads, and two-tracks	Transportation infrastructure derives from OpenStreetMap data (OSM). categorizing roads as major (motorway, trunk, primary, secondary), minor (tertiary, unclassified, residential), and two-track. Road density calculations (km/km²) incorporate typical widths for each category. Railroad density follows similar methodology, using a 0.010 km width factor per km².
Mining and energy production	Mining, wind turbines and oil wells	This component utilizes OSM data for mines, oil wells, and wind turbines. Mining evaluates the proportion of 1-km² cells designated as “quarry” or “industrial” areas. Energy infrastructure calculations combine oil well (0.014 km²) and wind turbine (0.0014 km²) spatial footprints per cell.

Electrical infrastructure	Powerlines and nighttime lights	Above-ground powerlines and night-time lights. Powerline assessment follows transportation methodology, using 0.015 km width per km². Nighttime illumination utilizes Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) satellite data, with digital number values (0-63) log[X+1] transformed to 0-1.806 and max-normalized.

Table 2 Area Under the Curve (AUC) and True Skill Statistic (TSS) score for 10-fold cross-validation (CV) and out-of-bag (OOB) datasets, including the mean, 5% and 95% quantile within the bootstrap assemblage

Model	AUC (OOB)	TSS(OOB)	AUC (CV)	TSS(CV)
Adaptive Boosting (AdaBoost)	0.83 (0.76-0.87)	0.51 (0.42-0.61)	0.87 (0.84-0.90)	0.58 (0.51-0.64)
Gradient Boosting Machine (GBM)	0.85 (0.79-0.90)	0.53 (0.43-0.63)	0.90 (0.87-0.93)	0.63 (0.56-0.70)
eXtreme Gradient Boosting (XGBoost)	0.84 (0.80-0.89)	0.52 (0.44-0.61)	0.88 (0.84-0.91)	0.60 (0.52-0.66)
Categorical Boosting (CatBoost)	0.85 (0.81-0.90)	0.53 (0.43-0.62)	0.90 (0.88-0.92)	0.64 (0.58-0.70)
Light Gradient Boosting Machine (LightGBM)	0.84 (0.80-0.89)	0.52 (0.42-0.63)	0.89 (0.87-0.92)	0.62 (0.56-0.68)
Stacking algorithm 1[a]	0.89 (0.86-0.93)	0.62 (0.55-0.71)	0.90 (0.88-0.93)	0.65 (0.56-0.72)
Stacking algorithm 2[b]	0.90 (0.87-0.93)	0.64 (0.57-0.73)	0.92 (0.89-0.94)	0.68 (0.61-0.74)
Stacking algorithm 3[c]	0.90 (0.87-0.94)	0.65 (0.57-0.73)	0.92 (0.90-0.94)	0.69 (0.62-0.75)

Stacking algorithms 1-3 represent different combinations of base learners with automated meta-learner selection, where Stacking algorithm 1 uses three base learners (GBM, LightGBM, XgBoost), Stacking algorithm 2 incorporates four base learners (CatBoost, GBM, LightGBM, XgBoost), and Stacking algorithm 3 employs five base learners (AdaBoost, CatBoost, GBM, LightGBM, XgBoost). All three configurations utilize automated meta-learner selection from BaggingClassifier, Logistic Regression, or Decision Tree. The specific configurations are as follows:

[a] Stacking algorithm 1: base learner: GBM, LightGBM, XgBoost; meta learner: auto select from BaggingClassifier, Logistic Regression or Decision Tree

[b] Stacking algorithm 2: base learner: CatBoost, GBM, LightGBM, XgBoost; meta learner: auto select from BaggingClassifier, Logistic Regression or Decision Tree

[c] Stacking algorithm 3: base learner: AdaBoost, CatBoost, GBM, LightGBM, XgBoost; meta learner: auto select from BaggingClassifier, Logistic Regression or Decision Tree

Figure legends

Figure 1.

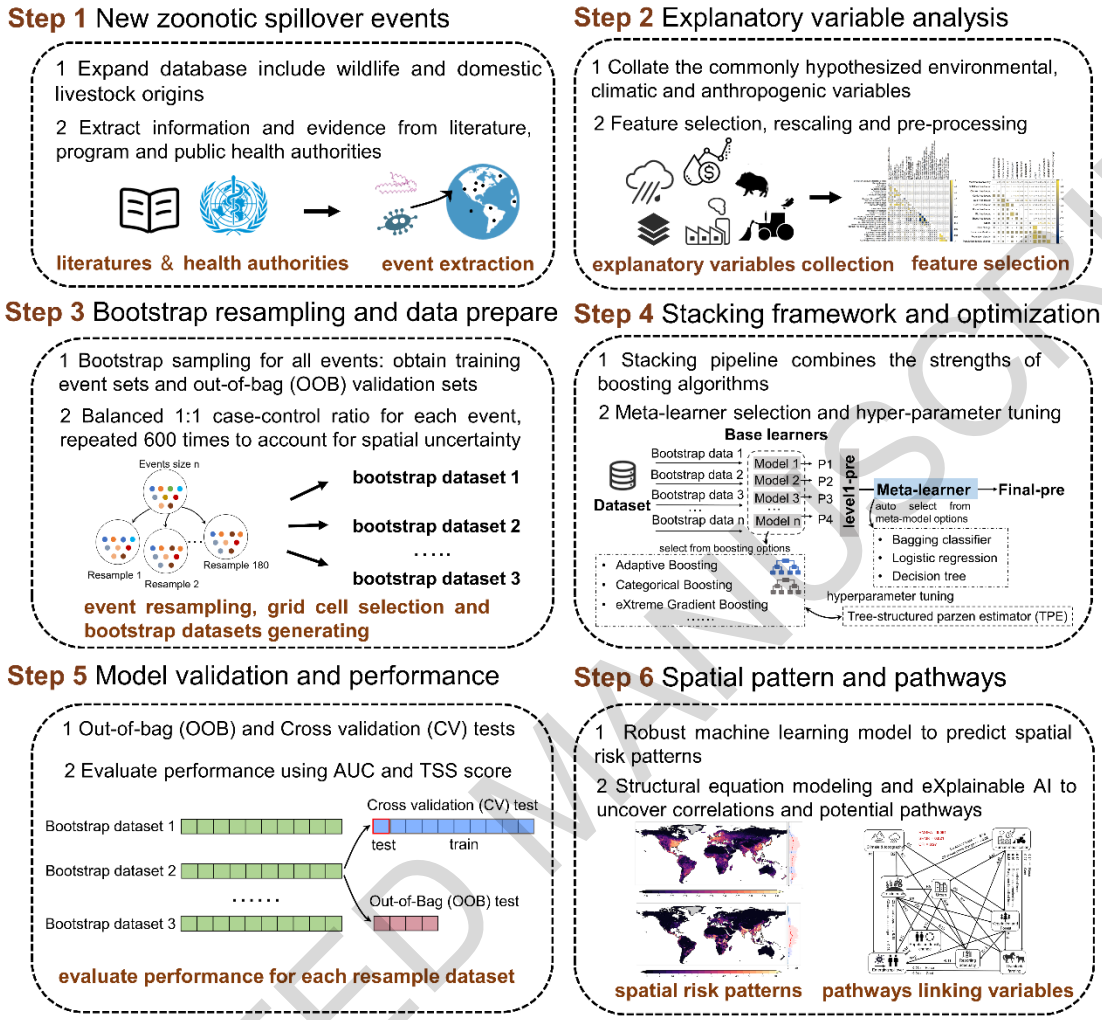

Figure 1. Workflow for predicting the risk of new zoonotic spillover. This figure outlines six main steps, each containing detailed sub-steps that collectively describe the methodology proposed for effective prediction and management of zoonotic disease. Credit: icons, iconfont.cn.

Figure 2.

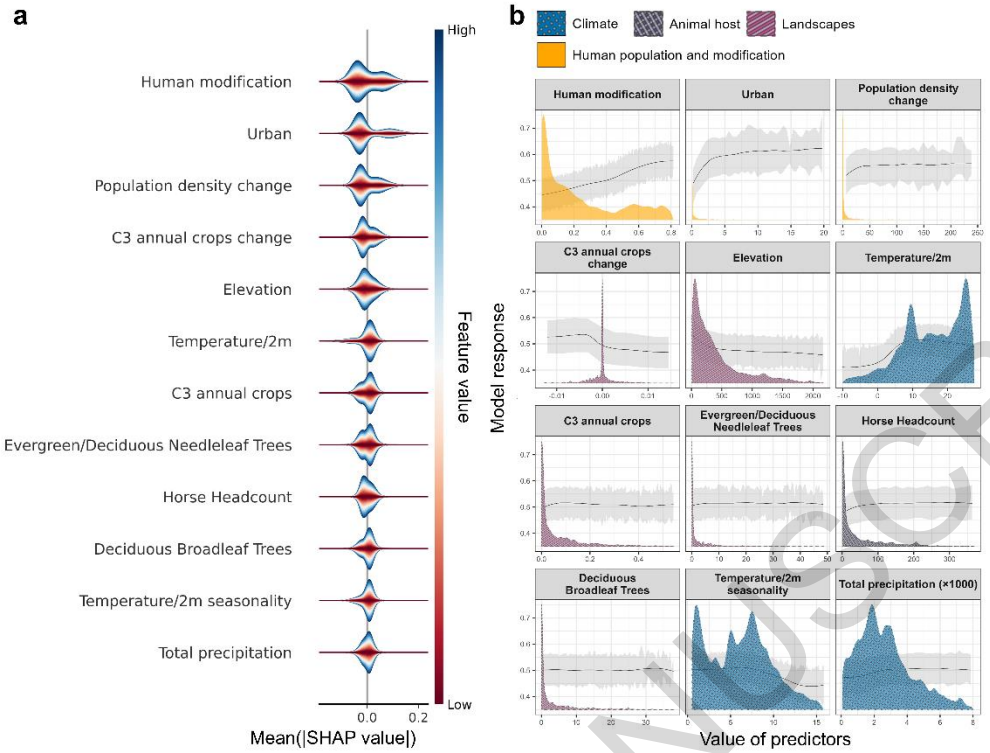

Figure 2. Key predictors and dependence in Stacking3 model analysis. (a) The top 12 predictors, ranked by SHapley Additive exPlanations (SHAP) importance, are presented in descending order along the y-axis based on their mean absolute SHAP values. The x-axis represents SHAP values, which indicate the change in log odds of spillover risk, calculated from the test sets across bootstrap resampling datasets during 10-fold cross-validation. The ranking of predictor importance is quantified in detail through Relative SHAP influence value (Figure S7) (b) Partial dependence plots: X-axes span the 5th to 95th percentiles of predictor values, with colored density plots showing predictor distributions. Y-axes represent the effect on new zoonotic spillover risk from that variable. Black lines represent median values, while the gray bands indicate 95% confidence intervals, derived through bootstrap resampling that incorporates uncertainty in spillover event locations.

Figure 3.

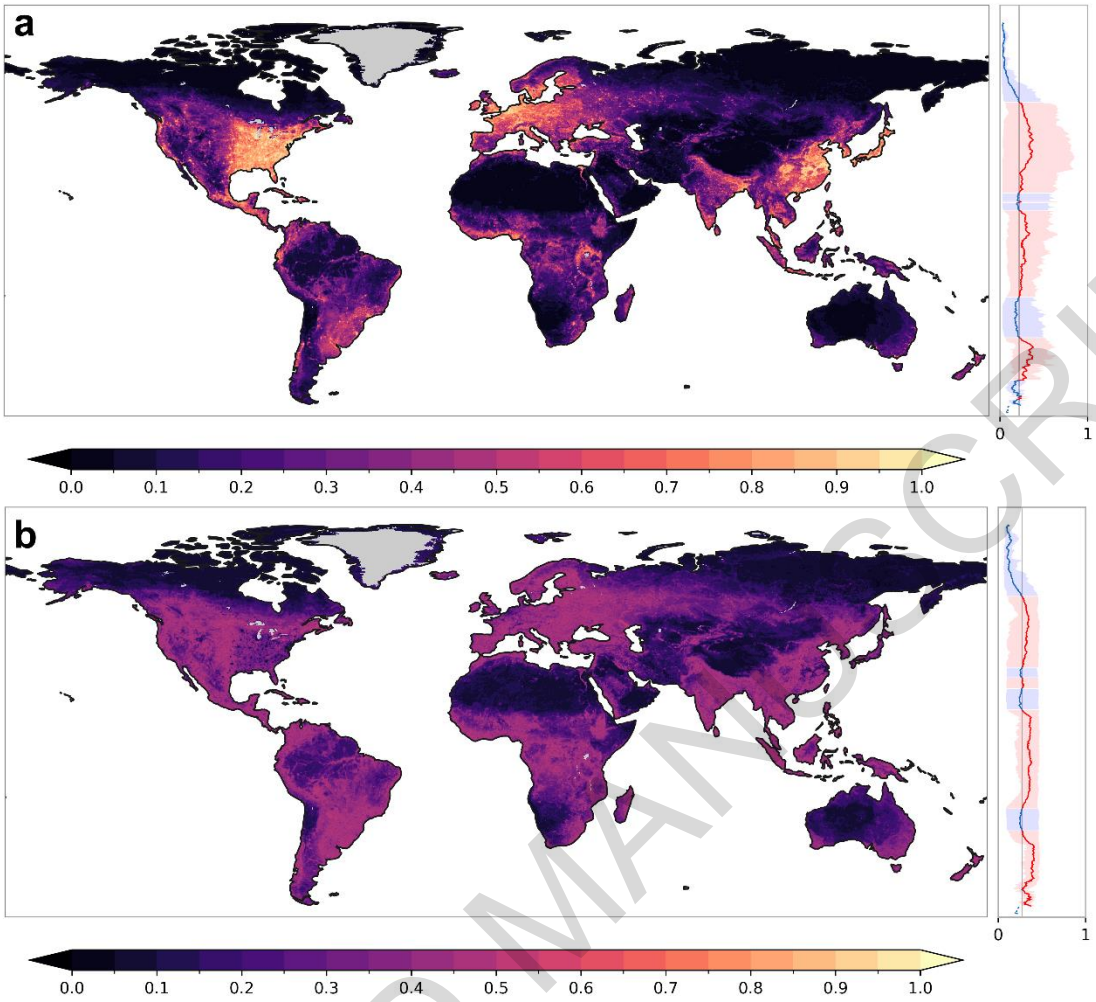

Figure 3. Predicted spatial pattern and uncertainty assessment of new zoonotic spillover risk (a) The spatial distribution of predicted relative risk for novel zoonotic spillover events and their subsequent reporting (reported new zoonotic spillover risk). The values represent the binomial probability that a grid cell sampled at that location will contain and report a new zoonotic spillover event, as compared to background sampling data. (b) The uncertainty assessment of "reported new zoonotic spillover risk" is quantified using the 10th-90th percentile range. The right panels display the latitudinal bins of average model response (a. predicted relative risk, b. uncertainty), where the y-axis represents latitude and the x-axis shows the zonally averaged model response, corresponding to the displayed color scale. All spatial visualizations were generated using Python 3.9.13 with Matplotlib 3.8.0.

Figure 4.

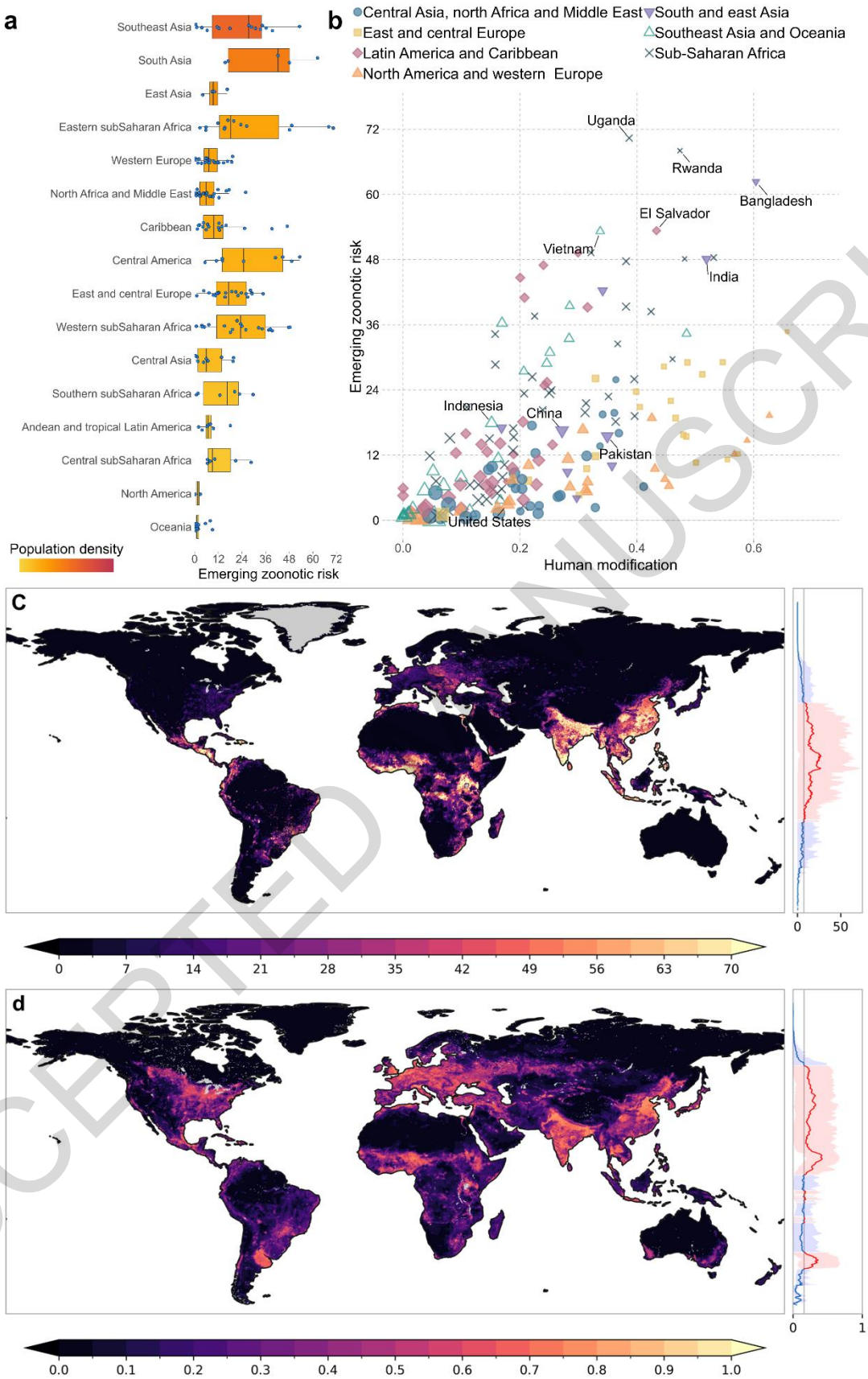

Figure 4. Spatial pattern of adjusted zoonotic spillover risk and associations with human modification

(a) Regional distribution of adjusted zoonotic spillover risk across 16 distinct geographic regions, with individual data points representing country-specific values. (b) Correlation between adjusted zoonotic spillover risk and human modification index, where points represent different countries, with point sizes proportional to the variance coefficient of adjusted risk. Colors denote different super-regions, which were delineated based on epidemiological similarities and geographic closeness. (c) Spatial pattern of new zoonotic spillover risk after accounting for relative reporting adequacy index (adjusted new zoonotic risk). (d) Aggregated human modification index at 0.25-degree spatial resolution. All spatial visualizations were generated using Python 3.9.13 with Matplotlib 3.8.0.

Figure 5.

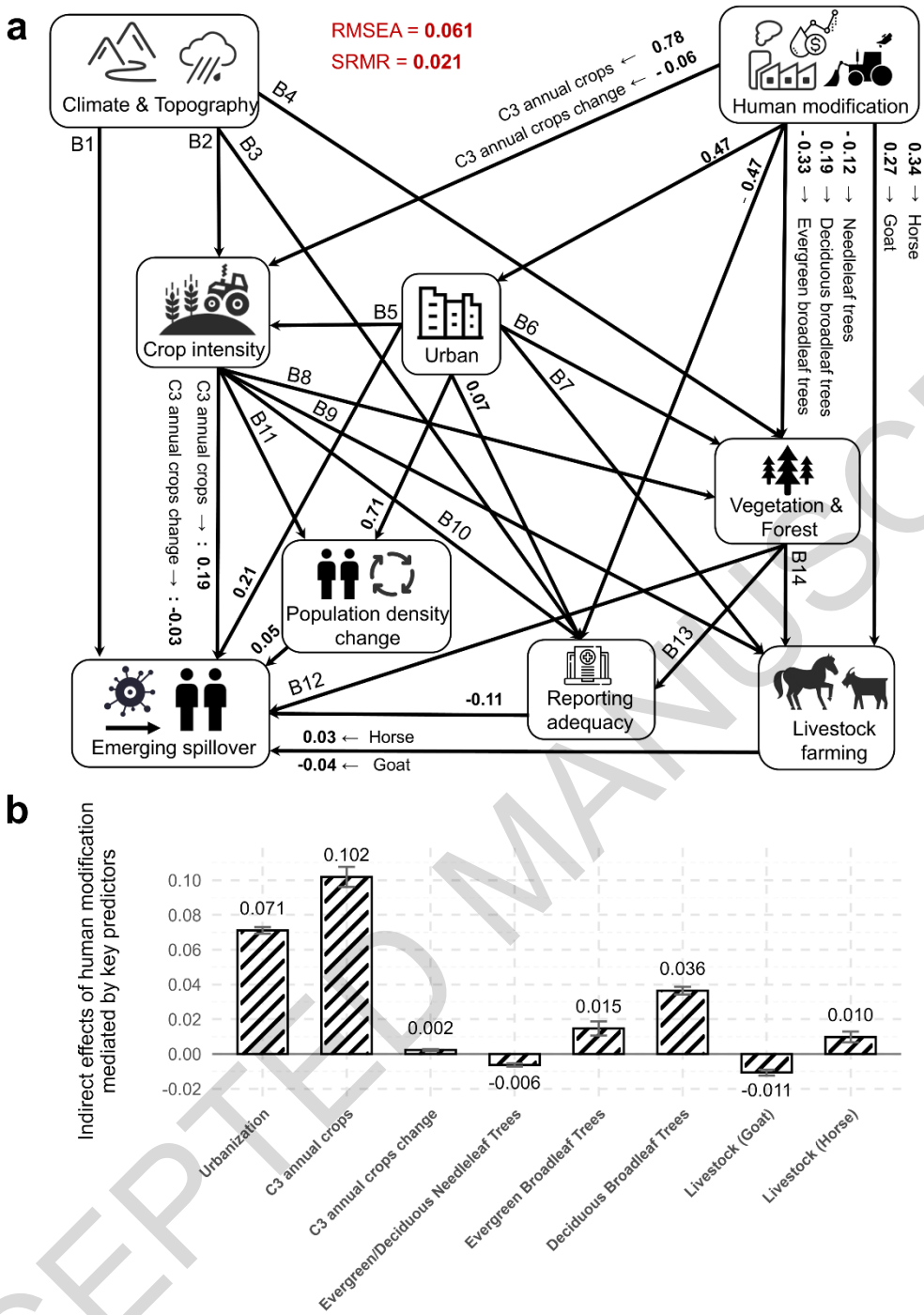

Figure 5. The relationship between zoonotic disease emergence and critical factors identified through stacking algorithm.

(a) Using the main structural equation model (SEM), we disentangled the anthropogenic effects on zoonotic disease emergence across land-modification gradients. We also accounted for important climatic and topographic predictors (2-meter temperature, 2-meter temperature seasonality, total precipitation, and elevation) in the main SEM model. For graphic simplicity, we grouped the different climatic and topographic, vegetation and forest, livestock farming, and crop intensity factors into the same box (B1-B14); however, it does not represent latent variables. One-sided arrows represent significant pathways, numbers adjacent to arrows are the standardized effect size. All paths are detailed in Table S7. (b) The standardized indirect effects of human modification on

the emergence of zoonotic spillover mediated by livestock farming, vegetation and forest, cropping intensity, and urbanization (95% CI). These factors, identified as significant through our stacking model and confirmed by SEM analysis, represent varying intensities of human modification. All indirect effect calculations are presented in Table S6. Credit: icons, iconfont.cn.