Received: 29 March 2023 | Revised: 16 August 2023 | Accepted: 22 August 2023

DOI: 10.1111/ele.14307

ECOLOGY LETTERS cnrs WILEY

METHOD

Predicting the fine-scale spatial distribution of zoonotic reservoirs using computer vision

Nathan C. Layman[1,2] | **Andrew J. Basinski**[2] | **Boyu Zhang**[2] | **Evan A. Eskew**[2] | **Brian H. Bird**[3] | **Bruno M. Ghersi**[3,4] | **James Bangura**[5] | **Elisabeth Fichet-Calvet**[6] | **Christopher H. Remien**[7] | **Mohamed Vandi**[8] | **Mohamed Bah**[9] | **Scott L. Nuismer**[10]

[1]EcoHealth Alliance, New York, New York, USA
[2]Institute for Interdisciplinary Data Sciences, University of Idaho, Moscow, Idaho, USA
[3]One Health Institute, School of Veterinary Medicine, University of California—Davis, Davis, California, USA
[4]Tufts University, Medford, Massachusetts, USA
[5]University of Makeni and University of California, Davis One Health Program, Makeni, Sierra Leone
[6]Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany
[7]Department of Mathematics and Statistical Science, University of Idaho, Moscow, Idaho, USA
[8]Ministry of Health and Sanitation, Freetown, Sierra Leone
[9]Ministry of Agriculture and Forestry, Freetown, Sierra Leone
[10]Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA

Correspondence
Scott L. Nuismer, Department of Biological Sciences, University of Idaho, Moscow, ID 83843, USA.
Email: snuismer@uidaho.edu

Funding information
Defense Advanced Research Projects Agency, Grant/Award Number: D18AC00028; National Institutes of Health, Grant/Award Number: R01GM122079; National Science Foundation, Grant/Award Number: DEB 2028162

Editor: Barbara A Han

Abstract

Zoonotic diseases threaten human health worldwide and are often associated with anthropogenic disturbance. Predicting how disturbance influences spillover risk is critical for effective disease intervention but difficult to achieve at fine spatial scales. Here, we develop a method that learns the spatial distribution of a reservoir species from aerial imagery. Our approach uses neural networks to extract features of known or hypothesized importance from images. The spatial distribution of these features is then summarized and linked to spatially explicit reservoir presence/absence data using boosted regression trees. We demonstrate the utility of our method by applying it to the reservoir of Lassa virus, *Mastomys natalensis*, within the West African nations of Sierra Leone and Guinea. We show that, when trained using reservoir trapping data and publicly available aerial imagery, our framework learns relationships between environmental features and reservoir occurrence and accurately ranks areas according to the likelihood of reservoir presence.

KEYWORDS
artificial intelligence, emerging infectious disease, Lassa virus, spillover, zoonoses

INTRODUCTION

Zoonotic pathogens, those that transmit from animals to humans, account for almost two-thirds of emerging infectious diseases (Jones et al., 2008). Because pathogen spillover into the human population is facilitated by contact between humans and animals, anthropogenic activities such as agriculture, land clearing and urban expansion may increase disease risk in some systems (e.g. rodent hosts of Lassa virus (Fichet-Calvet et al., 2007), bat hosts of Nipah virus (Chua et al., 2002) and SARS-related coronaviruses (Rulli et al., 2021), mosquito vectors of dengue virus (Su Yin et al., 2021)). Indeed, the apparent increase in recent disease emergence events may be attributable in part to increased human expansion into natural areas (Allen et al., 2017; Rohr et al., 2019). Hence, there is a growing need to understand how wildlife reservoirs and their pathogens are distributed, as well as how these distributions are changing with human activity.

Several modelling frameworks have been developed to understand and predict how the risk of pathogen spillover changes over space and time. Often, these models use remotely sensed environmental features such as precipitation, temperature and land cover as predictor variables (e.g. Lassa (Basinski et al., 2021; Fichet-Calvet & Rogers, 2009; Mylne et al., 2015), Ebola (Pigott et al., 2014, 2016; Rulli et al., 2017) and dengue (Bhatt et al., 2013)). Although some modelling frameworks are trained using only data on human infection, others integrate data on the distribution of reservoir species to better capture the multiple steps involved in zoonotic spillover (Basinski et al., 2021; Bhatt et al., 2013; Plowright et al., 2017; Redding et al., 2016; Rulli et al., 2021). Because of their more mechanistic underpinnings, these latter approaches are better suited to predicting how anthropogenic changes such as habitat destruction and climate change alter the risk of spillover (Carlson et al., 2022; Klitting et al., 2022). A remaining limitation, however, is dependence on remotely sensed data products that may lack key environmental variables and may be available only at relatively coarse spatial or temporal scales. As a consequence, these approaches may struggle to accurately predict spillover by ignoring the local distribution of important reservoir species and how downstream spillover risk will respond to anthropogenic change at local scales.

To fill this gap, we have developed a software pipeline that learns to predict the fine-scale distribution of a reservoir species using the spatial distribution of ecologically relevant features extracted directly from aerial imagery. Specifically, our approach first uses computer vision to extract visible features of possible importance to a specific reservoir species. Next, the spatial distribution of these features is summarized across a range of scales for use as predictor variables. Finally, these predictors are used in a regression method to predict the probability that the focal wildlife reservoir is present within a $50\,m^2$ cell. Because predictions can be traced back to associations with human-interpretable features in aerial imagery, our pipeline's output can be used to reveal important aspects of reservoir host ecology (Redding et al., 2016). Additionally, because our pipeline is modular, it allows users to easily substitute their preferred models for the computer vision or regression components. Although similar approaches have been developed, for example to improve the performance of dengue risk models by automatically extracting features from GoogleStreetMaps imagery (Su Yin et al., 2021), these efforts do not allow customized features of known or hypothesized ecological relevance to be directly extracted from aerial imagery.

To demonstrate the utility of our method for a disease system of major relevance to human health, we have applied it to Lassa fever. Lassa fever is a zoonotic disease endemic to West Africa that causes tens of thousands of deaths each year (Gibb et al., 2017; McCormick et al., 1987). Humans are infected by Lassa virus when they come into contact with the faeces or urine of the primary rodent reservoir, *Mastomys natalensis*. Previous ecological studies have demonstrated that the fine-scale spatial and temporal distribution of *M. natalensis* within rural town sites is influenced by the proximity of human habitations, cropland management strategies and rainfall patterns (Fichet-Calvet et al., 2007). Applying our method to this system allows ecologically relevant predictor variables (e.g. human habitations, crops) to be rapidly extracted and summarized directly from aerial images. Boosted regression trees (BRT) trained using these predictors and data from rodent trapping studies conducted in Sierra Leone and Guinea are able to rank areas within novel test towns according to the likelihood of reservoir occurrence.

METHODS

General methodology

Our software pipeline consists of two discrete modules: (1) computer vision and (2) regression. The computer vision module operates on aerial imagery, classifying land cover using semantic segmentation and buildings using instance segmentation. The regression module develops a predictive model by regressing geotagged reservoir trapping data against visible features identified and summarized in the first module (Figure 1).

Land cover classification using semantic segmentation

Geotagged aerial imagery was segmented using a combination of a custom Convolution Neural Network (CNN) classifier and the DeepLabv3 (Chen et al., 2017) semantic segmentation architecture (see Appendix S1: 'Land cover classification using semantic segmentation'). In this context, pixel segmentation is the process of classifying each pixel as a specific land cover type, for example rice cultivation, bare land or forest. The model is trained by manually identifying and labelling contiguous clusters of a uniform feature type. This piece of the computer vision module allows key features relevant to a specific reservoir species to be identified and quantified.

Building classification using instance segmentation

Buildings were identified and classified using an instance segmentation model. Specifically, we used the Mask R-CNN model (He et al., 2017) trained with the

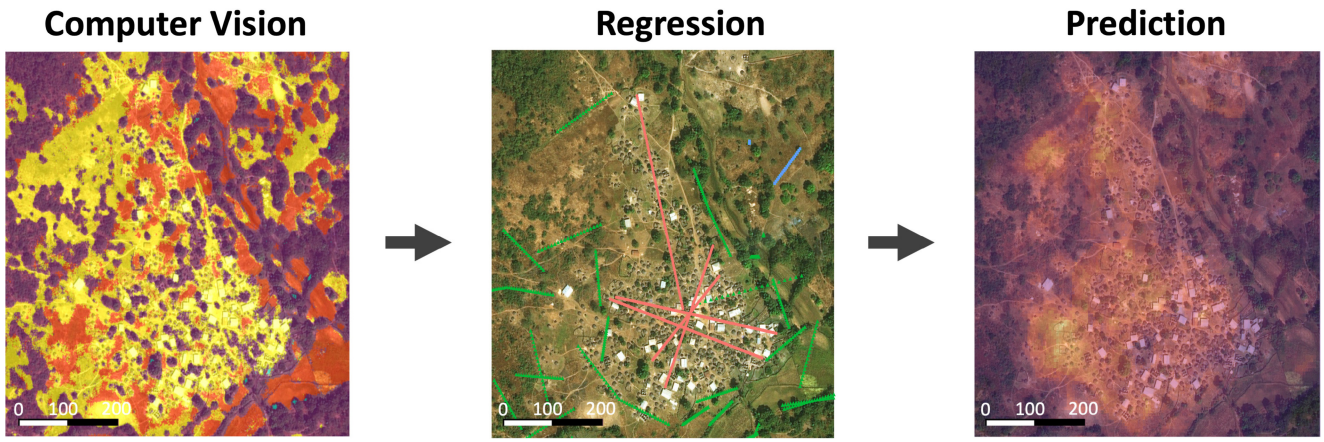

FIGURE 1 Overview of our modelling pipeline. The first module of our pipeline uses computer vision to identify features of hypothesized relevance to the reservoir within aerial images. The spatial distribution of these features (colours) is then summarized over various spatial scales. The second module of our pipeline then regresses these summarized features against geotagged reservoir trapping data (coloured lines indicate locations of traps) to develop a predictive model for reservoir trapping success. Learned relationships between visible features and geotagged trapping data allow the distribution of the reservoir to be predicted. The black and white scale bars indicate 200 m.

building outlines derived from OpenStreetMap data to detect and classify buildings (see Appendix S1: 'Building classification using instance segmentation'). This allows us to identify building type, size and distribution from aerial photographs of study sites alone. We use this information to create predictors describing the density of both traditional and modern buildings surrounding each pixel across several spatial scales (from a radius of 25 to 1000 m). At our study sites, traditional buildings are generally round and constructed of mud and thatch. Modern buildings are generally square and constructed of masonry and metal. These differences in construction may be associated with permeability to rodents and suitability as rodent habitat (Bonwitt et al., 2017). This piece of the computer vision module allows us to quantify how fine-scale characteristics of human housing impact the distribution of the reservoir.

Predicting reservoir distribution using regression

To predict reservoir distribution, we generate a full predictor set that combines land cover features and building predictors with additional environmental data, such as average monthly precipitation (de Sousa et al., 2020). These predictors—and spatially explicit reservoir presence/absence data—are then aggregated into spatial blocks for prediction. Aggregation into spatial blocks allows predictors to be averaged over scales relevant to the biology of the focal reservoir species. Using this feature-set to train regression models allows our method to learn the relationships between reservoir presence/absence (e.g. trap outcome) and local patterns of land cover, building characteristics and environmental features. This module allows us to generate high-resolution predictions for the spatial distribution of the reservoir species and thus a clearer and more accurate picture of spillover risk and its drivers.

Applied methodology: *Mastomys natalensis* in Sierra Leone and Guinea

Data description

The datasets we use to demonstrate the utility of our method and evaluate its performance stem from a combination of aerial photographs and rodent trapping studies from three towns in West Africa. Aerial photographs of the towns were sourced from Bing and Google very high-resolution (VHR) tile servers (Microsoft, 2022, Bing Maps Tile System, https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system; Google, 2022, Google Maps Static API, https://developers.google.com/maps/documentation/maps-static). *M. natalensis* trapping data came from previously published studies conducted in Guinea between 2003 and 2005 (Fichet-Calvet et al., 2007, 2014) and a recently completed study conducted in Sierra Leone between 2019 and 2020 (see Appendix S1: 'Rodent trapping datasets').

The trapping study datasets describe geotagged trap locations and whether or not a trap captured a *M. natalensis* rodent. Data from Sierra Leone come from a single town site (SLTA), whereas data from Guinea come from two town sites (GTA and GTB). Town names have been replaced with these codes to preserve, as much as is possible, site privacy. Because the precise location of each trap is not known for sites within Guinea, we aggregated trapping data into date-stamped spatial blocks by overlaying a 50 m^2 grid within each town. This spatial aggregation also better reflects the fact that reservoir animals range more broadly than suggested by the precise

location where they were trapped. For each spatial block and date, we calculate trap success and the total number of traps laid on a given visit start date and trap night. Because trap success is generally higher inside buildings than outside, we use each spatial block location and trapping date to generate two spatial blocks that describe trap success for indoor traps and outdoor traps. To remove any artefacts that result from a particular spatial grid, we perform the aggregation procedure for five randomly jittered 50 m^2 spatial grids.

Land cover classification using semantic segmentation

Land cover classification used aerial images downloaded from the Bing Map server with a resolution of ~0.22 m^2 per pixel. We focused on identifying bare ground, tree cover, rice fields, natural grass, mound agriculture (e.g. cassava), burned ground,\ and water as all are human-recognizable in aerial images and thought to influence the distribution of *M. natalensis*. In addition to these ecologically relevant land cover types, we included an additional category, cloud, to cover the cases where the land is obscured by clouds.

We used semantic segmentation to assign a land cover category to each pixel of our sites' aerial images (Long et al., 2015). Because models used for semantic segmentation generally include millions of parameters, they must be trained using a large number of images labelled at the level of individual pixels (Guo et al., 2018). This poses a challenge for remote areas like our study sites in West Africa where semantic labels are not widely available for aerial images. We used two methods to overcome this hurdle. First, we used transfer learning to borrow knowledge gained from training general models on other data sources to fine-tune models specific to our use case (Pan & Yang, 2010). Second, we developed a weakly supervised learning framework that minimizes the number of training samples required to achieve accurate models. Additional detail about procedures used for land cover classification are provided in the Appendix S1.

Building classification using instance segmentation

Aerial imagery for building classification was sourced from both Google and Bing tile servers through Quantum GIS and stored as rasters (QGIS Development Team, 2021). We employed the Mask R-CNN (Region-Based CNN) model (He et al., 2017) to detect buildings in the aerial images of our sites. The Mask R-CNN model uses multi-task learning (Caruana, 1997) with multiple branches that perform detection, segmentation and classification simultaneously. Through an independent forward derivation process, the Mask R-CNN outputs the building bounding boxes, the building category and the foreground/background label at the pixel level in the bounding boxes. Additional details about procedures used for building classification are provided in the Appendix S1.

Predicting reservoir distribution using boosted regression trees

The local spatial density of land cover and building features was calculated for each pixel for successively larger circles with radii of 25, 50, 100, 200, 500 and 1000 m, resulting in 36 building features (6 radii × scaled/unscaled × modern/traditional/all) and 42 land cover features (6 radii × 7 land cover types not including clouds). These features were extracted and averaged across trap locations within each date-stamped spatial block. In addition to the features extracted directly from the aerial images, we also included multiple features of presumed relevance to our sampling regime and the ecology of *M. natalensis*. First, as trapping studies have shown that reproductive cycles of *M. natalensis* are influenced by rainfall (Fichet-Calvet et al., 2008; Leirs et al., 1994), we included predictors that describe rainfall history over the preceding year. For each date-stamped spatial block, we computed the monthly average rainfall for each month in the year preceding that block's visit date. Thus, these predictors contain information about seasonal rainfall history relative to when each block's traps were set. This results in 12 predictors that are termed 'rainfall lags'. Second, because previous work has demonstrated that *M. natalensis* are more likely to be found within human habitations, we also included a binary 'trap placement' predictor that details whether a spatial block describes traps placed inside a building (0) or outdoors (1). Finally, to account for potential decreases in trap success over consecutive nights of trapping, we included an integer-valued predictor 'night' that describes how many consecutive nights of trapping occurred on or before a spatial block's date. In total, our framework used 92 predictors (building: 36, land cover: 42, rainfall: 12, trap placement: 1, night: 1).

We trained BRT models to learn the relationship between the trap success within each date-stamped 50 m^2 spatial block and its corresponding environmental feature set. BRTs are ensemble models composed of many simple decision trees that each have a small effect on prediction. Trees are added to the model in an iterative manner to reduce a loss function that measures the discrepancy between model predictions and observed trap success outcomes in the training data. BRTs are widely used in spatial ecology because they are relatively robust to overfitting but simultaneously capable of learning nonlinear relationships that may be present in ecological data (Elith et al., 2008). We fit BRTs using the XGBoost package in Python (Chen & Guestrin, 2016).

We assessed our framework's ability to learn relationships between rodent captures and environmental features at one location, and subsequently used these relationships to accurately predict trap outcomes at a novel location. For these model validation steps, we split the dataset into training, validation and test datasets that are each comprised of one town's trapping data. Because there are three towns, this results in six unique training, validation and test town combinations. Next, for each of these six combinations, models with a given hyperparameter combination are fit to a training dataset comprised of a single town's data aggregated over the five randomly jittered 50 m^2 spatial grids. For model training, we use the root mean square error as the loss function. Next, the best hyperparameter set is selected by evaluating the fitted model on the town chosen for validation. Specifically, we apply the fitted model separately to each of the five randomly jittered grids from the validation town. This results in five estimates of mean absolute error (MAE) as applied to the validation dataset.

The MAE estimates from the validation dataset were then averaged together into a single validation MAE score that was ultimately used to select the best hyperparameter combination. To minimize overfitting, we implement early stopping on the number of tree-fitting iterations. Specifically, for each hyperparameter combination, we limit the number of tree-fitting iterations to that which reduces the validation MAE over each 100-iteration cycle. For each train-validation-test fold, we chose the hyperparameter combination that resulted in the lowest validation MAE score as the best hyperparameter combination. The hyperparameter combinations we tested are described in the Appendix S1.

Finally, to determine how the models perform on a never-before-seen town, the fitted models were assessed on the test town dataset comprised of 25 randomly jittered sets of spatial blocks. We used 25 sets of spatial blocks instead of five in order to develop robust estimates of model performance. For a given test town, this assessment is performed on those two models that performed best on the remaining training and validation towns. To help increase the ability of the final model predictions to generalize to new areas, we implement a model averaging approach when generating predictions from the two best models. For example, when the town SLTA was used as a test town, we generated predictions using the average of the two models that were trained on GTA and validated on GTB and vice versa.

To determine whether the best models selected using the procedure described above have meaningful predictive power when applied to a novel site, we quantify the model's ability to correctly determine which block has higher trapping success in pairs of 50 m^2 spatial blocks that are randomly sampled from the test town. We focus on the models' ability to rank spatial blocks because it is easily interpretable and has direct implications for research on zoonotic diseases. For instance, knowing which areas are more likely to harbour large numbers of reservoir animals allows viral surveillance to be focused on those areas rather than wasted in areas where the reservoir animal is likely to be rarer. Similarly, distribution of edible baits containing vaccines for reservoir animals can be made more efficient by focusing on areas where the reservoir animal is more likely to be present in large numbers. Many other examples exist where the ability to rank areas according to predicted reservoir trapping success facilitates the efficient detection and management of zoonotic diseases. In these applied research contexts, accurate prediction of absolute reservoir density is likely to be of less relevance than relative ranking, and thus, we evaluate models using such ranks.

For context, we compare the ranking ability of the models developed using our full pipeline (Figure 1) with two other simpler models. The first is a null model that ranks spatial blocks by assigning a random score between zero and one to each block. This represents a totally naive prediction regarding relative *M. natalensis* trap success and is completely independent of our pipeline, using no information from the computer vision or regression modules. We will refer to this as the 'Random' model. The second model assigns a score to each spatial block that is equal to the density of buildings within a 100 m radius of the block's centre point. The building density measurement, in turn, comes directly from the output of the computer vision module. Thus, this second model uses only the first module of our pipeline (computer vision) to quantify the spatial distribution of a feature of hypothesized ecological importance. We chose to focus on the density of buildings because decades of ecological studies show *M. natalensis* rodents tend to be associated with human habitation (Bangura et al., 2021; Brouat et al., 2007). Thus, in this particular system we expect this simple model that combines computer vision with an a priori hypothesis regarding a key visible feature to perform quite well. In other systems that lack such a singular and well-vetted visible predictor, this simple approach is not feasible and the regression module will be required. We will refer to this second model as the 'Buildings' model.

RESULTS: *MASTOMYS NATALENSIS* IN SIERRA LEONE AND GUINEA

We evaluated the utility of our approach by quantifying its performance when applied to *M. natalensis*, the primary reservoir of Lassa virus within West Africa. We begin by assessing our method's ability to accurately classify land cover and building features. Next, we present associations our method learns between features extracted from the aerial imagery and trapping data on the reservoir host. Finally, we assess how well our method

predicts the spatial distribution of the reservoir within novel test towns not used in the model training process.

Methodological performance: Feature classification

When applied to our study sites in West Africa, our method learned to identify features of hypothesized importance to *M. natalensis* (Figure 2). Quantitatively, the performance of our methodology was good for both land cover and building classification (see Appendix S1). The lone exception was grass which was difficult for our method to accurately distinguish from other land cover types. Overall, our land cover classifier had an accuracy of 95%, precision of 96% and recall of 96% across features. Similarly, our building classifier had an overall accuracy of 96%, a precision of 90% and a recall of 98% across different building types. The resulting maps for each site, showing land cover and buildings along with locations of traps, are provided in the Appendix S1 (Figures 1–3).

Methodological performance: Predicting reservoir distribution

Boosted regression tree models were able to predict trap success for *M. natalensis* inside each date-stamped 50 m^2 spatial block. Table 1 shows the performance of the six models with the lowest MAE between observed trap success and model predictions at a validation site. The MAE scores on validation data indicate that models with a maximum tree depth of one (no interaction effects among predictors) perform best on two of the six validation folds, while models with a tree complexity of four perform better on the remaining four validation folds.

The trained BRT models learn to associate trap success in a 50 m^2 spatial block with anthropogenic features of the environment. Generally, variable importance output from the fitted models, which describe the typical reduction in training loss for a given hyperparameter combination due to each focal predictor, indicated that building density within the surrounding 25–100 m radius, land cover and precipitation lags were key features associated with trap success (Figure 3). In addition, trap location (inside vs. outside) and trap night influenced trap success, albeit to a lesser degree. As implied by the strong relationships learned by the models between *M. natalensis* and building density, trap success is generally predicted to be highest near the centre of a town and lower in areas farther away from buildings (Figure 4).

To determine whether the six best models shown in Table 1 have meaningful predictive power when applied to a novel site, we compared their predictions to the 'Random' null model and the 'Buildings' model. Averaged across test sites, we find that the BRT, Buildings and Random models achieved pairwise ranking metrics of 0.787, 0.778 and 0.506, respectively (Figure 5; panel a). This demonstrates that our method, when using BRT in the regression module, is significantly better at ranking areas with respect to rodent trapping success than is the null model that randomly ranks spatial blocks. Surprisingly, our method using BRT in the regression module only marginally outperforms the 'Buildings' model which foregoes a secondary regression module in favour of a simple a priori model that predicts trapping success from building density, as calculated using only the first computer vision module of our pipeline. We attribute the strong performance of this simpler approach in this case to the quality of extensive on-the-ground ecological research identifying building density as an important predictor of *M. natalensis* occurrence.

Although our BRT models are good at ranking the relative trapping success of paired 50 m^2 spatial blocks, both the BRT models and the 'Buildings' model show curious bimodal patterns (Figure 5; panel a). To investigate whether this bimodality arises through differences in these models' ability to predict trapping success inside and outside of houses, we incorporate a predictor that describes whether a spatial block corresponds to inside-house or outside-house trap success. Similarly, we assess the 'Buildings' and 'Random' models independently on inside-house and outside-house spatial blocks. The town GTA did not have enough inside trap data to generate meaningful measures of ranking ability inside houses.

Analysing the ranking performance of the models separately for trapping success inside and outside of houses demonstrated that differences in predictive performance in these two contexts were likely the source of the bimodal patterns. Both BRT and 'Buildings' models do a poor job predicting trapping success within houses. Specifically, the ability of these models to rank spatial areas by trap success is not significantly greater than a random ranking. This lack of predictive ability suggests that relative trap success within houses cannot be predicted for a new location using information on building density and land cover features. In contrast, when applied to spatial blocks that describe trap success outside of houses (Figure 5; panel B), both the BRT and 'Buildings' model demonstrate performance superior to that of the 'Random' model. Generally, however, the performance of our framework is similar to that of the 'Buildings' model that uses only the first module of our pipeline (identification of visible features using computer vision) to generate predictions.

To further explore the predictive ability of our method, we investigated its ability to predict trapping success per se, rather than simply ranking areas according to trapping success as we report above. Specifically, we plotted the predicted vs. observed trapping success within each 50 m^2 spatial block for each combination of training, validation and testing sites (Figure S16). This analysis demonstrates that our method generates predictions with significant positive correlations between predicted and

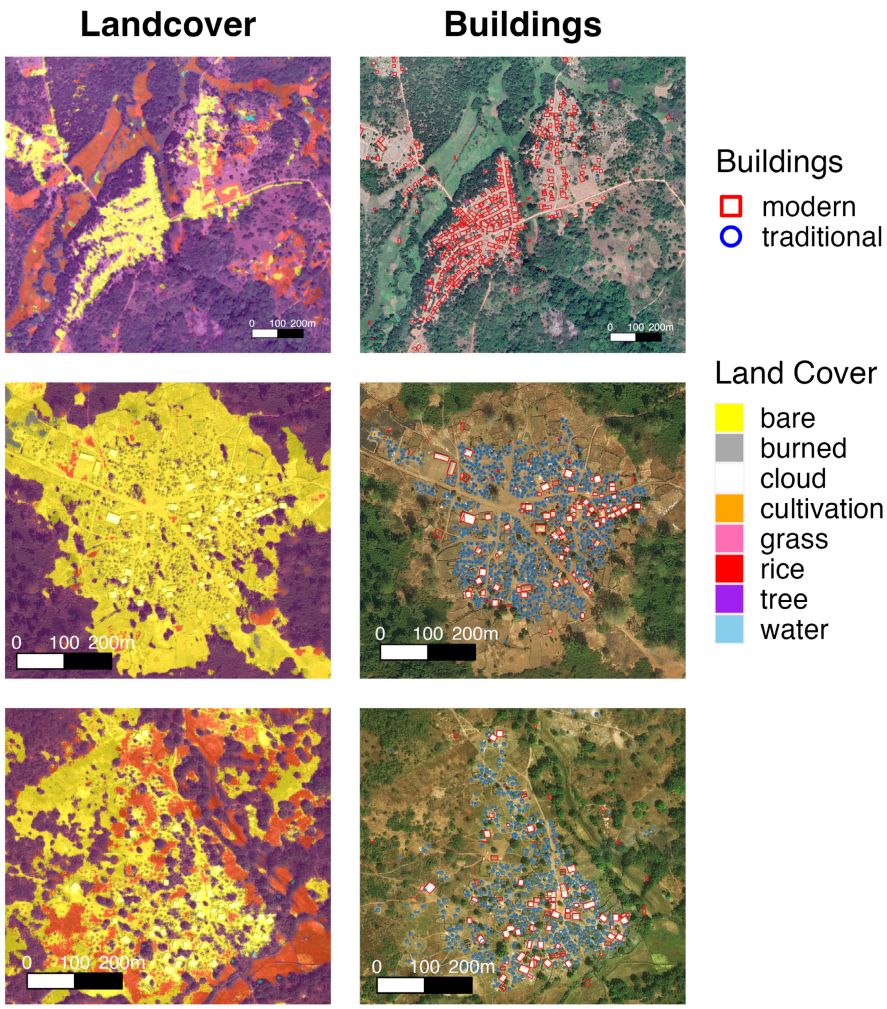

FIGURE 2 Computer vision algorithms applied to SLTA (top), GTA (middle) and GTB (bottom). The algorithms classify aerial imagery into different land cover features (left) and building features (right). Each pixel in the land cover image reflects the most likely feature classification within 25 m of that point. The black and white scale bars indicate 200 m.

observed values but that the absolute accuracy of predictions is rather poor (Figure S16). This general result varies depending on which towns were used for training, validation and testing. Most notably, cases where SLTA is used for training tend to generate much better quantitative predictions than cases where SLTA is used for testing. We suspect this result arises because locations of individual traps within SLTA were recorded precisely rather than inferred from the locations where trap lines began and ended as was done for GTA and GTB. Thus, SLTA offers a richer stream of data for training than do GTA and GTB; by the same token, it also represents a more challenging target for prediction.

DISCUSSION

Zoonotic spillover occurs at the intersection of human and animal populations and thus depends on the granular distribution of reservoir animals within areas of human activity. Despite this, most existing methods for predicting spillover risk focus on broad geographic areas and rely on coarse-grained, remotely sensed predictors (Allen et al., 2017; Basinski et al., 2021; Becker, Washburne, et al., 2019; Klitting et al., 2022; Mylne et al., 2015; Redding et al., 2016). Here, we advance our understanding of spillover risk by developing a methodology that predicts the spatial distribution of a reservoir species at the fine spatial scales where reservoir animals and humans interact. These predictions can be updated as rapidly as new aerial imagery becomes available and so can, in principle, predict how reservoir distribution will change as visible anthropogenic disturbance proceeds. Our approach capitalizes on advances in computer vision to extract features of hypothesized relevance to the reservoir animal directly from aerial images. These features are then fed into boosted regression trees, along with other relevant ecological predictors, to predict the spatial pattern of reservoir abundance. Because our method is modular, rapidly advancing computer vision methods can be easily substituted for those we use here as they become available (Kirillov et al., 2023). The same is

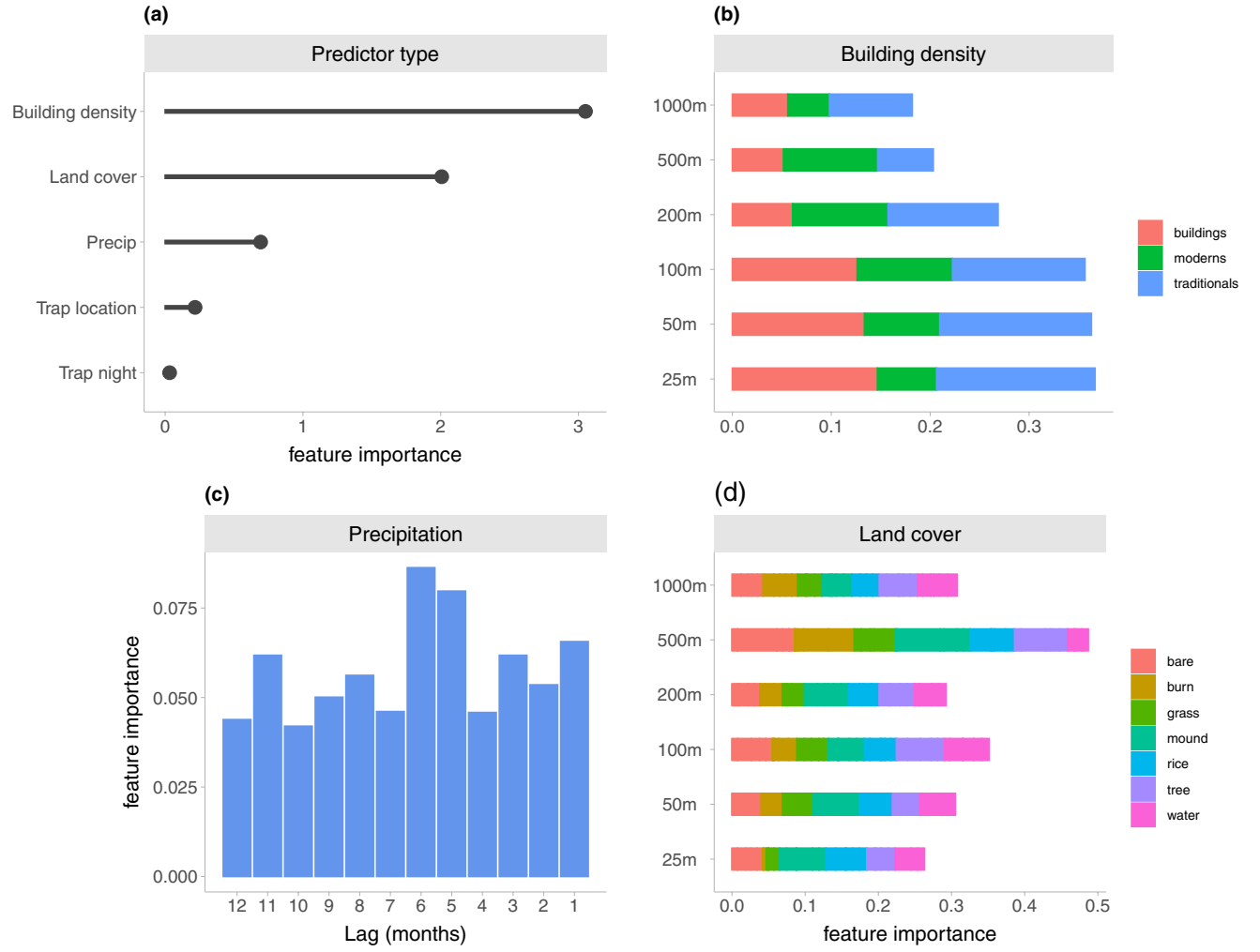

FIGURE 3 Relative variable importance plots averaged across the best-performing models (see Table 1). Variable importance was calculated as the gain, or relative increase in model accuracy, provided by each of the 92 predictors (see Appendix S1). To ease interpretation, variable importance was aggregated by (a) predictor type, (b) the density of buildings within a given distance in meters, (c) the average precipitation for each month between one and 12 months prior and (d) the composition of the surrounding land cover. Trap location indicates whether traps were placed inside or outside of houses. The 'buildings' category in (b) refers to total building density regardless of building type.

true for the regression module. Applying our methodology to *M. natalensis*, the reservoir of Lassa virus within West Africa, demonstrates three important advantages of our method.

The first advantage of our method is that it enables the spatial distribution of a pathogen reservoir to be predicted at much finer spatial scales than was previously possible (Basinski et al., 2021; Fichet-Calvet & Rogers, 2009; Mylne et al., 2015; Redding et al., 2016). Our method's ability to accurately rank (50 m^2) grid cells according to their likelihood of harbouring a reservoir animal creates opportunities for targeted interventions within areas of human activity. For instance, these rankings could be used to position bait stations that distribute poison, sterilizing medications or wildlife vaccines within areas where the reservoir animal is most likely to be found. There is increasing interest in primary prevention as a potentially cost-effective means to prevent zoonotic spillover (Bernstein et al., 2022), and optimization of intervention allocation across space could overcome some of the remaining challenges inherent in this approach (Mariën et al., 2019; Nuismer et al., 2020). Although our method was able to accurately rank (50 m^2) grid cells within rural town sites in Sierra Leone and Guinea with respect to their likelihood of harbouring the rodent reservoir of Lassa virus, *M. natalensis*, its ability to predict absolute values of trap success within cells was weak (see Figure S16). Thus, when tested using the sparse data currently available for *M. natalensis* in West Africa, our method can reliably be used to decide where the reservoir is most likely to occur, but not to predict the actual density of the animal in any given location.

A second advantage of our method is its ability to transparently identify key features determining the reservoir host distribution. Despite integrating sophisticated machine learning algorithms, our software pipeline maintains interpretability by summarizing the final BRT results in ways that reveal ecological relationships and

TABLE 1 Performance of boosted regression tree models on training and validation datasets. MAE stands for mean absolute error.

				Hyperparameters				
Train	**Validation**	**Train MAE**	**Val. MAE**	**Max depth**	**ColSample**	**Subsample**	**Iterations**	**Gamma**
SLTA	GTA	0.059	0.037	4	0.05	0.5	1000	2
SLTA	GTB	0.057	0.032	4	0.05	0.5	1050	0.5
GTA	SLTA	0.016	0.086	1	0.1	0.1	1600	0
GTA	GTB	0.012	0.029	4	0.3	0.05	850	0
GTB	SLTA	0.016	0.088	1	0.05	0.25	1900	0
GTB	GTA	0.013	0.028	4	0.2	0.025	1750	1

Note: The hyperparameters describe the maximum allowed tree depth (Max Depth), the proportion of columns sampled at each tree-fitting iteration (ColSample), the proportion of training data used to generate each tree (Subsample), the number of tree-fitting iterations (Iterations) and minimum loss reduction required to make a further partition on a leaf node of the tree (Gamma).

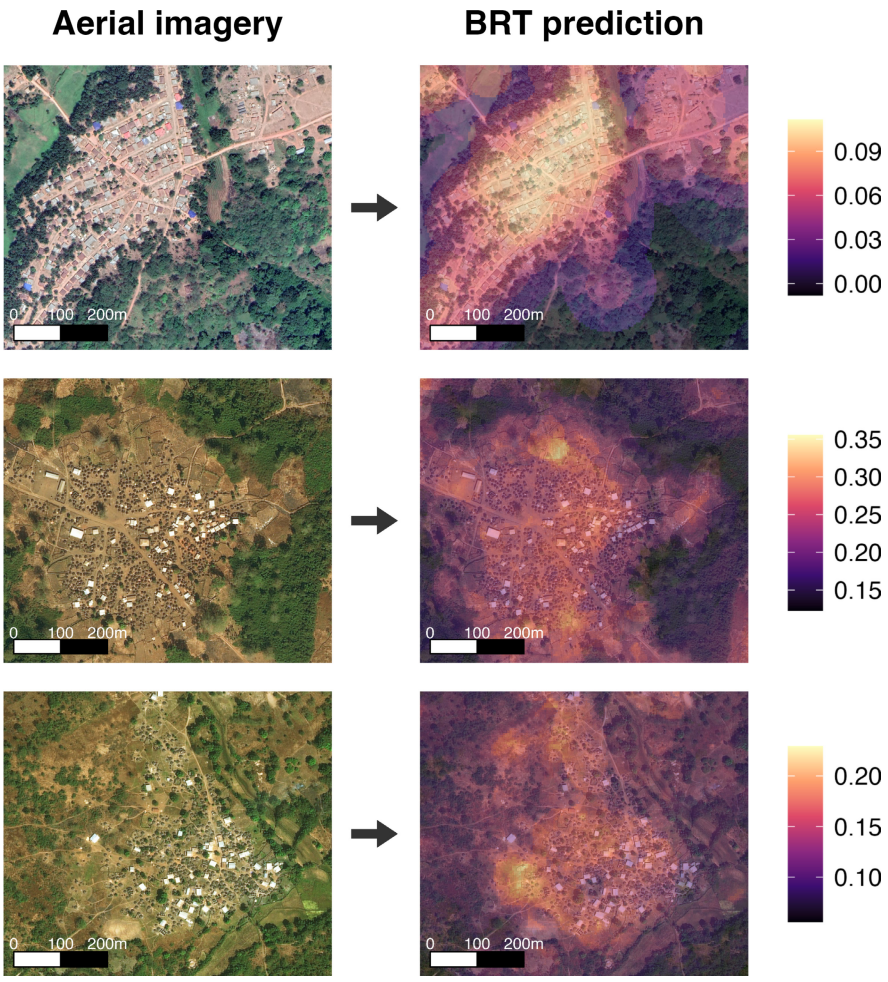

FIGURE 4 Imagery (left-hand panels) and model predictions for trapping success (right-hand panels) in each of the three towns in our dataset: SLTA (top), GTA (middle) and GTB (bottom). Each prediction was generated by averaging the model output of the two best-boosted regression tree models that were trained and validated on data with that test site omitted. Predictions shown were generated for traps placed outside of houses. Predictions for traps placed within houses show a similar spatial pattern but have uniformly elevated predictions for trapping success. The black and white scale bars indicate 200 m.

avoid opaque 'black box' prediction (Elith et al., 2008). For example, our feature importance plots indicate that building density is vitally important to predicting *M. natalensis* trapping success. Though this is commonly known for the commensal rodent *M. natalensis*, our framework was able to automatically learn this important relationship from aerial imagery and rodent trapping data and clarify the spatial scale over which this relationship operates. Applying our method to other less well-studied disease systems could rapidly identify previously

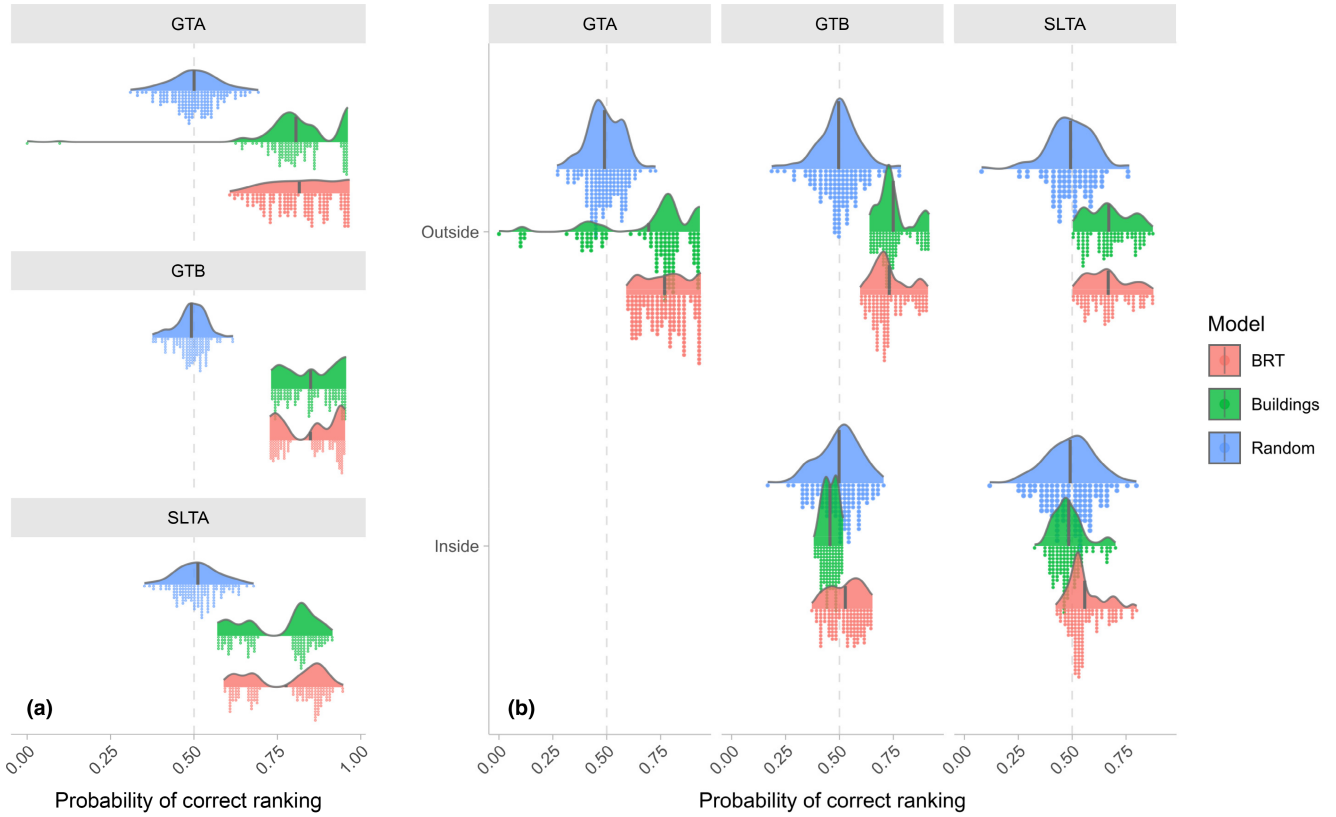

FIGURE 5 Performance of models at ranking pairs of 50 m^2 spatial blocks by trap success, across towns (a) and decomposed into traps placed indoors and outdoors (b). The different model types are BRT (fitted models used to rank trap success at each 50 m^2 spatial block), Buildings (trap success is ranked solely by comparing building densities within a 100 m radius between each pair of cells) and Random (trap success is ranked randomly). Vertical lines indicate distribution means. In panel (b), mean values were 0.5 for the random model regardless of trap location or town. For GTA, the distributional means were BRT: 0.77 and Buildings: 0.7 (outside). For GTB, the means were BRT: 0.73, Buildings: 0.75 (outside) and BRT: 0.53, Buildings 0.46 (inside). For SLTA, the means were BRT: 0.67, Buildings: 0.67 (outside) and BRT: 0.56, Buildings: 0.48 (inside).

unknown associations between environmental features and spillover risk.

A third advantage of our method is its ability to rapidly update predictions for reservoir distribution as new aerial imagery becomes available. Thus, our method can be used to predict how reservoir distribution should change in response to visible anthropogenic disturbance as rapidly as it is possible to update aerial imagery. Because our method generates predictions by identifying features in aerial imagery, it can also be used to predict how the spatial distribution of a reservoir species is likely to shift in response to planned anthropogenic activity. For instance, our method can be used to predict how construction of additional human housing will influence reservoir distribution by simply adding houses to an aerial image. Similarly, the likely consequences of deforestation can be predicted by erasing forest from the image and replacing it with grass or bare ground. In addition to enabling the impact of anthropogenic activity to be predicted, this capability enables alternative plans for human development to be rationally assessed with respect to their likely consequences for spillover risk.

Although our methodology provides significant advantages over existing approaches, it does have limitations that were revealed when applied to the Lassa virus system. One fundamental limitation is the need for high-resolution aerial imagery. Although this is not likely to be a problem for geographic regions with regularly updated imagery, it may be a significant limitation for understudied regions like our study sites in West Africa. For example, aerial imagery of sufficiently high resolution was not available for our study sites in Guinea between 2003 and 2005 when rodent trapping data were collected. Instead, we were forced to use more contemporary imagery collected during and after 2014. Given this temporal separation between imagery and trapping data, it is both promising and remarkable that our framework was able to accurately predict relative trap success at a novel site. We anticipate that these limitations will become less and less important as datasets containing high-resolution aerial imagery become increasingly available. A second limitation revealed by applying our method to the Lassa virus system is more fundamental. Specifically, although our method proved quite capable of predicting the distribution of *M. natalensis* outside of houses, it was no better than a random model at predicting their distribution within houses. In retrospect, this is an understandable outcome given the

potential for houses to differ wildly in their suitability/desirability for *M. natalensis* based on internal features not captured by aerial imagery. For instance, differences in food storage practices, permeability to rodents and tolerance for rodents within the house may all play important roles but be entirely invisible in aerial imagery. This demonstrates the need to complement imagery-based approaches like the method developed here with on-the-ground investigations of cultural practices and individual behaviours that influence spillover risk (Bonwitt et al., 2017; Martinez et al., 2022; PREDICT Consortium et al., 2021).

Finally, although our methodology can be used to predict how changes to the visible environment will impact the distribution of an animal reservoir, doing so assumes that the system moves directly from one steady state to another. Future methods could use time series methods such as ARIMA (Hyndman & Khandakar, 2008), exponential smoothing (De Livera et al., 2011) or a suite of emerging deep learning architectures (Lim & Zohren, 2021) to develop a formal forecasting framework. The central roadblock to implementing these methods is their thirst for high-resolution spatial data that has been repeatedly collected from many points in time. Although the dataset we have studied here does include samples taken from different points in time, the amount of data available for any given time point are far too sparse to be useful without pooling over time. This limitation is not unique to our dataset, instead it reflects the challenges associated with collecting data from wild animals that serve as reservoirs for infectious disease (Becker, Crowley, et al., 2019). Developing new technologies that facilitate regular data collection from hard-to-study wildlife reservoirs will be critical to removing this bottleneck to developing robust forecasting methodologies for zoonotic infectious diseases.

Methods like ours are a vital next step to predicting how the distribution of animal reservoirs will change with human alterations to the environment over the fine spatial scales where spillover actually occurs. Across the world, people are rapidly encroaching on previously wild areas and engaging in land-clearing activities that displace wildlife and facilitate contact with humans (Eby et al., 2023; Plowright et al., 2021). Forecasting models that directly use aerial imagery have the potential to help us understand and predict the consequences of these actions in near real-time, enabling policymakers to act on measures that will reduce the health burden from zoonotic diseases.

AUTHOR CONTRIBUTIONS

AJB and SLN conceived of the study. AJB, BZ and NCL performed analyses. BHB, BMG, JB and EFC collected data. MV and MB facilitated field studies. AJB, SLN, BZ, NCL and EAE drafted the manuscript. All authors edited the manuscript.

ACKNOWLEDGEMENTS

This work was supported by NIH grant R01GM122079, NSF grant DEB 2028162 and DARPA grant D18AC00028. The funders had no role in the study design, data collection and analysis, the decision to publish or the preparation of the manuscript.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ele.14307.

DATA AVAILABILITY STATEMENT

Key portions of the project files, including analysis scripts, are available on GitHub at https://github.com/Boyu-Zhang-UOI/Machine-Vision-for-Zoonotic-Reservoir-Density-at-Fine-Resolution. The full project repository is archived on Zenodo at https://doi.org/10.5281/zenodo.7717286.

ORCID

Nathan C. Layman https://orcid.org/0000-0003-2238-6584
Scott L. Nuismer https://orcid.org/0000-0001-9817-0056

REFERENCES

Allen, T., Murray, K.A., Zambrana-Torrelio, C., Morse, S.S., Rondinini, C., Di Marco, M. et al. (2017) Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8(1), 1124.

Bangura, U., Buanie, J., Lamin, J., Davis, C., Bongo, G.N., Dawson, M. et al. (2021) Lassa virus circulation in small mammal populations in Bo District, Sierra Leone. *Biology*, 10(1), 28.

Basinski, A.J., Fichet-Calvet, E., Sjodin, A.R., Varrelman, T.J., Remien, C.H., Layman, N.C. et al. (2021) Bridging the gap: using reservoir ecology and human serosurveys to estimate Lassa virus spillover in West Africa. *PLoS Computational Biology*, 17(3), e1008811.

Becker, D.J., Crowley, D.E., Washburne, A.D. & Plowright, R.K. (2019) Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biology Letters*, 15(12), 20190423.

Becker, D.J., Washburne, A.D., Faust, C.L., Mordecai, E.A. & Plowright, R.K. (2019) The problem of scale in the prediction and management of pathogen spillover. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1782), 20190224.

Bernstein, A.S., Ando, A.W., Loch-Temzelides, T., Vale, M.M., Li, B.V., Li, H. et al. (2022) The costs and benefits of primary prevention of zoonotic pandemics. *Science Advances*, 8(5), eabl4183.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L. et al. (2013) The global distribution and burden of dengue. *Nature*, 496(7446), 504–507.

Bonwitt, J., Sáez, A.M., Lamin, J., Ansumana, R., Dawson, M., Buanie, J. et al. (2017) At home with *Mastomys* and *Rattus*: human-rodent interactions and potential for primary transmission of Lassa virus in domestic spaces. *The American Journal of Tropical Medicine and Hygiene*, 96(4), 935–943.

Brouat, C., Loiseau, A., Kane, M., Bâ, K. & Duplantier, J.-M. (2007) Population genetic structure of two ecologically distinct multimammate rats: the commensal *Mastomys natalensis* and the wild *Mastomys erythroleucus* in southeastern Senegal. *Molecular Ecology*, 16(14), 2985–2997.

Carlson, C.J., Albery, G.F., Merow, C., Trisos, C.H., Zipfel, C.M., Eskew, E.A. et al. (2022) Climate change increases cross-species viral transmission risk. *Nature*, 607(7919), 555–562.

Caruana, R. (1997) Multitask learning. *Machine Learning*, 28(1), 41–75.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. (2017) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.

Chen, T.Q. & Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 785–794. https://doi.org/10.1145/2939672.2939785

Chua, K.B., Chua, B.H. & Wang, C.W. (2002) Anthropogenic deforestation, El Niño and the emergence of Nipah virus in Malaysia. *Malaysian Journal of Pathology*, 24(1), 15–21.

De Livera, A.M., Hyndman, R.J. & Snyder, R.D. (2011) Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496), 1513–1527.

de Sousa, K., Sparks, A., Ashmall, W., van Etten, J. & Solberg, S. (2020) Chirps: API client for the CHIRPS precipitation data in R. *Journal of Open Source Software*, 5(51), 2419.

Eby, P., Peel, A.J., Hoegh, A., Madden, W., Giles, J.R., Hudson, P.J. et al. (2023) Pathogen spillover driven by rapid changes in bat ecology. *Nature*, 613(7943), 340–344.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813.

Fichet-Calvet, E., Becker-Ziaja, B., Koivogui, L. & Günther, S. (2014) Lassa serology in natural populations of rodents and horizontal transmission. *Vector-Borne and Zoonotic Diseases*, 14(9), 665–674.

Fichet-Calvet, E., LeCompte, E., Koivogui, L., Daffis, S. & Meulen, J.T. (2008) Reproductive characteristics of *Mastomys natalensis* and Lassa virus prevalence in Guinea, West Africa. *Vector-Borne and Zoonotic Diseases*, 8(1), 41–48.

Fichet-Calvet, E., Lecompte, E., Koivogui, L., Soropogui, B., Doré, A., Kourouma, F. et al. (2007) Fluctuation of abundance and Lassa virus prevalence in *Mastomys natalensis* in Guinea, West Africa. *Vector-Borne and Zoonotic Diseases*, 7(2), 119–128.

Fichet-Calvet, E. & Rogers, D.J. (2009) Risk maps of Lassa fever in West Africa. *PLoS Neglected Tropical Diseases*, 3(3), e388.

Gibb, R., Moses, L.M., Redding, D.W. & Jones, K.E. (2017) Understanding the cryptic nature of Lassa fever in West Africa. *Pathogens and Global Health*, 111(6), 276–288.

Guo, Y., Liu, Y., Georgiou, T. & Lew, M.S. (2018) A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2), 87–93.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017) Mask R-CNN, IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hyndman, R.J. & Khandakar, Y. (2008) Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 27(3), 1–22.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. et al. (2008) Global trends in emerging infectious diseases. *Nature*, 451(7181), 990–993.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L. et al. (2023) Segment Anything. *arXiv*, 2304.02643.

Klitting, R., Kafetzopoulou, L.E., Thiery, W., Dudas, G., Gryseels, S., Kotamarthi, A. et al. (2022) Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. *Nature Communications*, 13(1), 5596.

Leirs, H., Verhagen, R. & Verheyen, W. (1994) The basis of reproductive seasonality in *Mastomys* rats (Rodentia: Muridae) in Tanzania. *Journal of Tropical Ecology*, 10(1), 55–66.

Lim, B. & Zohren, S. (2021) Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200209.

Long, J., Shelhamer, E. & Darrell, T. (2015) Fully convolutional networks for semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Mariën, J., Borremans, B., Kourouma, F., Baforday, J., Rieger, T., Günther, S. et al. (2019) Evaluation of rodent control to fight Lassa fever based on field data and mathematical modelling. *Emerging Microbes & Infections*, 8(1), 640–649.

Martinez, S., Sullivan, A., Hagan, E., Goley, J., Epstein, J.H., Olival, K.J. et al. (2022) Living safely with bats: lessons in developing and sharing a global one health educational resource. *Global Health: Science and Practice*, 10(6), e2200106.

McCormick, J.B., Webb, P.A., Krebs, J.W., Johnson, K.M. & Smith, E.S. (1987) A prospective study of the epidemiology and ecology of Lassa fever. *The Journal of Infectious Diseases*, 155(3), 437–444.

Mylne, A.Q., Pigott, D.M., Longbottom, J., Shearer, F., Duda, K.A., Messina, J.P. et al. (2015) Mapping the zoonotic niche of Lassa fever in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 109(8), 483–492.

Nuismer, S. L., C. H. Remien, A. J. Basinski, T. Varrelman, N. Layman, K. Rosenke, B. Bird, M. Jarvis, P. Barry, P. W. Hanley, and others, 2020: Bayesian estimation of Lassa virus epidemiological parameters: Implications for spillover prevention using wildlife vaccination. *PLoS Neglected Tropical Diseases*, 14(9), e0007920.

Pan, S.J. & Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.

Pigott, D.M., Golding, N., Mylne, A., Huang, Z., Henry, A.J., Weiss, D.J. et al. (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife*, 3, e04395.

Pigott, D. M., A. I. Millear, L. Earl, C. Morozoff, B. A. Han, F. M. Shearer, D. J. Weiss, O. J. Brady, M. U. Kraemer, C. L. Moyes, and others, 2016: Updates to the zoonotic niche map of Ebola virus disease in Africa. *eLife*, 5, e16412.

Plowright, R.K., Parrish, C.R., McCallum, H., Hudson, P.J., Ko, A.I., Graham, A.L. et al. (2017) Pathways to zoonotic spillover. *Nature Reviews Microbiology*, 15(8), 502–510.

Plowright, R.K., Reaser, J.K., Locke, H., Woodley, S.J., Patz, J.A., Becker, D.J. et al. (2021) Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. *The Lancet Planetary Health*, 5(4), e237–e245.

PREDICT Consortium, Saylors, K., Wolking, D.J., Hagan, E., Martinez, S., Francisco, L. et al. (2021) Socializing one health: an innovative strategy to investigate social and behavioral risks of emerging viral threats. *One Health Outlook*, 3(1), 11.

QGIS Development Team. (2021) *QGIS geographic information system*. QGIS Association.

Redding, D.W., Moses, L.M., Cunningham, A.A., Wood, J. & Jones, K.E. (2016) Environmental-mechanistic modelling of the impact of global change on human zoonotic disease emergence: a case study of Lassa fever. *Methods in Ecology and Evolution*, 7(6), 646–655.

Rohr, J.R., Barrett, C.B., Civitello, D.J., Craft, M.E., Delius, B., DeLeo, G.A. et al. (2019) Emerging human infectious diseases and the links to global food production. *Nature Sustainability*, 2(6), 445–456.

Rulli, M.C., D'Odorico, P., Galli, N. & Hayman, D.T.S. (2021) Land-use change and the livestock revolution increase the risk of zoonotic coronavirus transmission from rhinolophid bats. *Nature Food*, 2(6), 409–416.

Rulli, M.C., Santini, M., Hayman, D.T.S. & D'Odorico, P. (2017) The nexus between forest fragmentation in Africa and Ebola virus disease outbreaks. *Scientific Reports*, 7(1), 41613.

Su Yin, M., Bicout, D.J., Haddawy, P., Schöning, J., Laosiritaworn, Y. & Sa-Angchai, P. (2021) Added-value of mosquito vector breeding sites from street view images in the risk mapping of dengue incidence in Thailand. *PLoS Neglected Tropical Diseases*, 15(3), e0009122.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Layman, N.C., Basinski, A.J., Zhang, B., Eskew, E.A., Bird, B.H., Ghersi, B.M. et al. (2023) Predicting the fine-scale spatial distribution of zoonotic reservoirs using computer vision. *Ecology Letters*, 26, 1974–1986. Available from: https://doi.org/10.1111/ele.14307