UC Irvine
UC Irvine Previously Published Works

Title

Mapping Potential Malaria Vector Larval Habitats for Larval Source Management in Western Kenya: Introduction to Multimodel Ensembling Approaches.

Permalink

https://escholarship.org/uc/item/2zv6z8z5

Journal

American Journal of Tropical Medicine and Hygiene, 110(3)

Authors

Zhou, Guofa
Lee, Ming-Chieh
Wang, Xiaoming
et al.

Publication Date

2024-03-06

DOI

10.4269/ajtmh.23-0108

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Am. J. Trop. Med. Hyg., 110(3), 2024, pp. 421–430
doi:10.4269/ajtmh.23-0108
Copyright © 2024 American Society of Tropical Medicine and Hygiene

Mapping Potential Malaria Vector Larval Habitats for Larval Source Management in Western Kenya: Introduction to Multimodel Ensembling Approaches

Guofa Zhou,[1]* Ming-Chieh Lee,[1] Xiaoming Wang,[1] Daibin Zhong,[1] Andrew K. Githeko,[2] and Guiyun Yan[1]

[1]*Program in Public Health, University of California, Irvine, California;* [2]*Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya*

Abstract. Identification and mapping of larval sources are a prerequisite for effective planning and implementing mosquito larval source management (LSM). Ensemble modeling is increasingly used for prediction modeling, but it lacks standard procedures. We proposed a detailed framework to predict potential malaria vector larval habitats by using multimodel ensemble modeling, which includes selection of models, ensembling method, and predictors, evaluation of variable importance, prediction of potential larval habitats, and assessment of prediction uncertainty. The models were built and validated based on multisite, multiyear field observations and climatic/environmental variables. Model performance was tested using independent field observations. Overall, we found that the ensembled model predicted larval habitats with about 20% more accuracy than the average of the individual models ensembled. Key larval habitat predictors in western Kenya were elevation, geomorphon class, and precipitation for the 2 months prior. Additional predictors may be required to increase the predictive accuracy of the larva-positive habitats. This is the first study to provide a detailed framework for the process of multimodel ensemble modeling for malaria vector habitats. Mapping of potential habitats will be helpful in LSM planning.

INTRODUCTION

Malaria is still the most serious mosquito-borne infectious disease in the tropics, especially in Africa.[1] The scale-up of indoor interventions such as long-lasting insecticidal nets (LLINs) and indoor residual insecticide spraying (IRS), together with effective treatment, has led to a substantial reduction in the malaria burden.[1] Nonetheless, malaria control faces increased challenges due to vector resistance to insecticides and outdoor residual transmission.[1,2] Larval source management (LSM) has become a viable choice for further reducing malaria transmission and is recommended by the WHO for use under certain conditions for malaria control and elimination.[3,4] Larval control complements LLINs and IRS and controls both indoor and outdoor transmission.[3,5,6] Previous studies in several countries where malaria is endemic prove that larval source reduction and larviciding can significantly reduce both indoor and outdoor vector density and malaria infections.[5,7–10] Several African countries have adopted LSM as a key vector control tool parallel to LLINs and IRS or as a supplementary strategy.[11–13] For example, the Kenyan government has planned to target all larval sources for LSM by 2023, although the plan is overambitious.[11] Because larval sources can be anywhere after rain, the currently recommended LSM strategy is targeted LSM with environmental management and larviciding.[3] Large-scale implementation of LSM requires a carefully designed strategy and effective planning, especially the identification and mapping of larval sources prior to any field operations.[11–14] However, effective larval habitat identification and mapping are lacking.

Many climate- and environment-based models have been used to predict the potential distribution of malaria vector larval habitats, for example, ecological/environmental niche models,[15,16] logistic regression,[17–19] and machine learning methods such as artificial neural network and random forest models.[17,18] Clearly, different modeling methods are likely to produce different results and select different risk factors.[17,18,20] With so many models available, it will be difficult to find the robust model with the best predictions. More importantly, since different models end up with different groups of risk factors, how should the key risk factors be determined?

Recently, multimodel ensemble approaches have been used increasingly for predictions in various field studies.[21–28] According to Kotu and Deshpande, "Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in one final prediction for the unseen data."[29] The reasons for employing ensemble methods in building a model are to enhance the overall performance of the model, minimize the error rate that can be caused by using individual models, and reduce the overall uncertainty of predictions.[22,29,30] There are different ways to ensemble the models, including most votes, simple average, weighted average (linear or nonlinear), boosting, and stacking.[22,24,27,31–33] In mosquito studies, ensemble modeling has been used to predict the global expansion of *Aedes* mosquitoes and the invasion of *Anopheles stephensi* in Africa.[26,34,35] Very recently, ensembled modeling techniques have also been adopted to predict potential mosquito larval habitats.[20,25,36] For example, Wieland et al. used the average of dry/wet/normal year predictions as the ensembled prediction, which makes sense biologically for reducing over- or underestimation due to variations in precipitation,[36] although we do not know whether the dry and wet years have similar intensities (negative/positive) of impact on larval habitat availability and productivity. In studies by Rhodes et al. and Beeman et al., the final ensembled model was fitted using only those models with area-under-curve (AUC) scores of ≥ 0.7, and each model was weighted proportionally to its AUC score.[20,25] The assumption is that only models with AUC scores > 0.7 are considered to be well-performing models, and better-performing models should be given higher weight.[20,24,25,27] These methods lack a biological

*Address correspondence to Guofa Zhou, 3501 Hewitt Hall, Irvine, CA 92697. E-mail: zhoug@hs.uci.edu

(and/or statistical) basis for both the model and weight selections; the selection of an AUC $>$ 0.7 is somewhat arbitrary, and the weight selection may not lead to robust estimations. More importantly, feature importance or risk factor analysis is essential for larval habitat prediction and for LSM.[21,36] Bose et al. used rank order (i.e., most votes) for feature selection.[21] Rhodes et al. and Sinka et al. did not specify how the variables were selected for the ensembled model.[25,26] Beeman et al. displayed the variable importance for each individual model but not for the ensembled model.[20] Overall, no standard method currently exists for variable selection and relative importance evaluation for ensembled models.[24,27,29,37]

The aim of this study was to build a habitat prediction model using historical observations of aquatic and larva-positive habitats in western Kenya and weighted multimodel ensembling. We used recent field observations and observations from different areas to test the model. Finally, we proposed a method to assess risk factors and to measure the uncertainty of model predictions. The predicted map of habitat and larval distribution in western Kenya will be beneficial for LSM planning.

MATERIALS AND METHODS

Study area, field data collection, and data assignment for modeling. Field aquatic habitats and mosquito larval surveys were conducted in four sentinel sites in Kakamega (Iguhu site) and Vihiga (Mbale, Emakakha, and Emutete sites) counties in western Kenya (Figure 1). Western Kenya is the last malaria transmission hot spot persisting in Kenya.[38–43] The study sites included places with seasonal malaria transmission. Malaria larval distribution, larval ecology, and parasite transmission in these sites have been studied extensively over the past 20 years.[38–52] The elevation in the study area ranges from 1,420 m to 1,670 m. The study area has two rainy seasons: a long rainy season that usually starts in March and lasts until June, and a short rainy season between October and November, with two dry seasons in between.[43] Annual precipitation reaches around 1,400 mm. Aquatic habitats and mosquito larval infestations were surveyed in 2002, 2003, 2005, 2008, 2010–2012, 2017, and 2018 (Figure 1). In most years, field habitat surveys were carried out in February (dry season), May (long rainy season), August (dry season), and November (short rainy season). Global positioning system (GPS) locations of all habitats were recorded, the sizes of habitats were measured, and the availability (yes/no) of immature *Anopheles* mosquitoes was checked in most years. Overall, we surveyed about 50,000 aquatic habitats, and *Anopheles* larval infestation status was available in about 40,000 habitats (Supplemental Table S1).

Since all aquatic habitats had water at the time they were surveyed, to identify predictors of aquatic habitats, we artificially generated about 13,500 pseudohabitats in randomly selected locations with known residential houses and no habitats whatsoever[25,36] (Supplemental Table S1). Pseudohabitats were selected at least 50 m away from any known aquatic habitats. These pseudohabitats will always have an aquatic status of "no water" regardless of sampling season. The pseudohabitats were generated using ArcGIS 10.0 (ESRI, Redlands, CA).

We used 2002–2012 field observations to train and validate the models. A 70:30 random splitting of data was used for model training and validation (Supplemental Table S1). To produce an unbiased evaluation of the final model, we used field data collected from the same area in 2017 and 2018 as testing data, i.e., an independent data set.[53,54] About two-thirds of the pseudohabitats were randomly

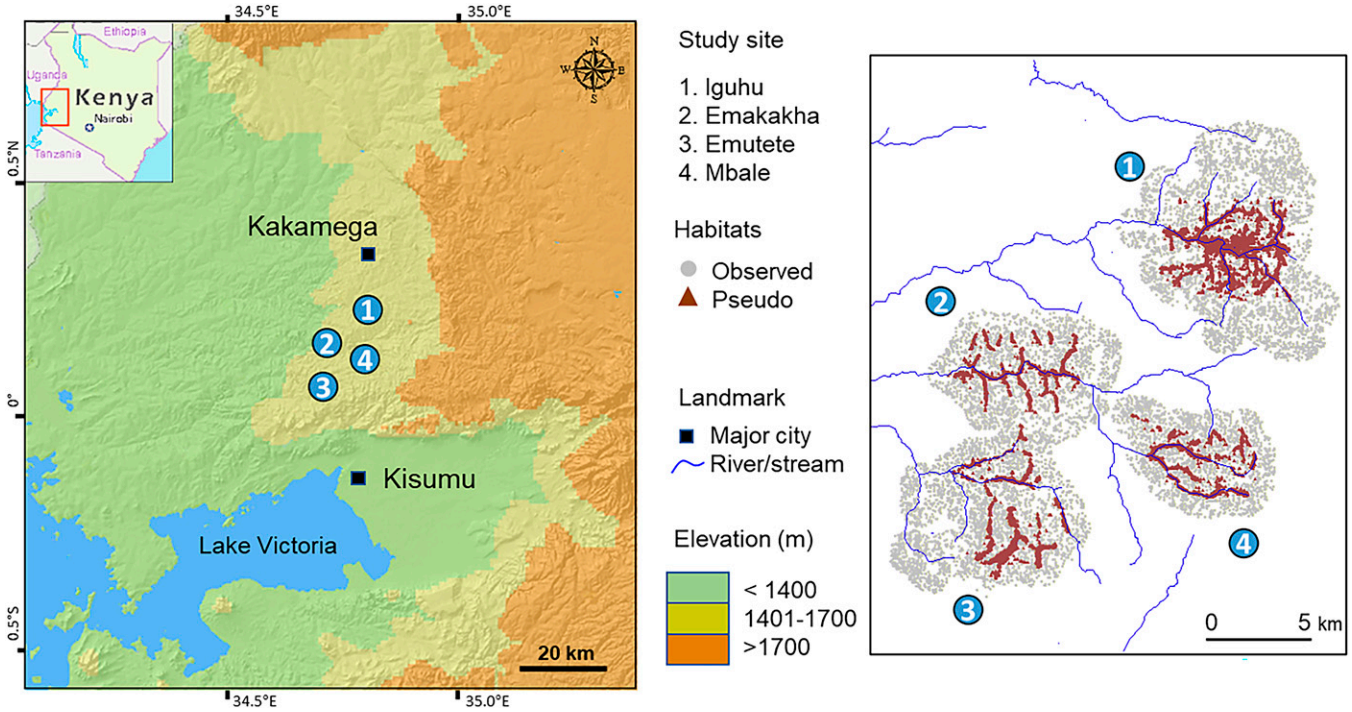

FIGURE 1. Study sites (left) and distribution of observed aquatic habitats (2003–2018) and pseudohabitats in the four study sites in western Kenya (right).

assigned (with equal probability) to February, May, August, and November 2012 samples, and one-third were assigned as 2017 samples.

Climatic and environmental data. Climatic and environmental data included about 220 variables and were described in a previous study.[43] Briefly, environmental data included altitude, topographic/geomorphologic features (e.g., slope, aspect, land surface roughness, etc.), land use land cover, and tree coverages from different years. Climatic data included monthly average temperature and cumulative precipitation. Satellite image-derived data included normalized difference vegetation index, normalized difference water index, land surface temperature, and evapotranspiration, among others. Climatic and satellite image-derived data were all monthly based and covered the study period from 2000 to 2019. To account for the time-lagged effect, we selected only the data from the 6 months prior to the habitat surveys, assuming that the habitat aquatic status and larval availability were affected only by climatic conditions during the last 6 months.

The assignments of climatic and environmental data to each survey point were done using ArcGIS 10.0.

Modeling process. In this study, we used 2002–2012 data for model training and validation and 2017–2018 data (serving as independent data) for model testing. After data dimension reduction, we separated the two data sets, i.e., training/validation and testing (Figure 2). Once models were validated, all models were subjected to testing. After model testing, the combined data set was used to build the final models for ensemble modeling (Figure 2). Data dimension reduction was done using principal component analysis and was described in previous studies.[29,43] Individual model training/validation, testing, and prediction are rather straightforward; however, there is no standard method for model ensembling or risk factor determination in ensemble modeling.[20,25,26,29,30,36,55]

Model specification and modeling process. Many models can be used to identify mosquito larval habitats.[17,18,20,25] There is no standard method for model selection in ensemble modeling. Many studies have used the R package *biomod2* for model selection and ensembling.[20,25,26] In principle, model diversity and independence are key to ensemble modeling[29,37] because models built using similar methods may perform similarly (for example, gradient-boosted logistic and gradient-boosted tree models), so ensembling them may provide limited additional and possibly biased information for risk factor analysis and limited improvement in model performance due to redundancy. Including conventional models such as logistic and decision tree models and modern machine learning methods such as neural network increases model diversity. We selected 10 models for classification analyses in this study: five of them are conventional methods (ordinary logistic regression, Bayesian logistic regression, regular classification tree, naïve Bayes tree, k-nearest neighbor [kNN] classifier), and five are machine learning methods (gradient-boosted [GB] logistic regression, support vector machine [SVM] logistic regression, extreme gradient-boosting [XGB] tree, random forest, regular neural network [NNW] tree) (Supplemental Table S2).

Two of the 10 models require careful prior specifications: Bayesian logistic regression and NNW. For Bayesian logistic regression, one needs to specify a joint distribution for the outcome(s) and all the unknown parameters, which typically takes the form of a marginal prior distribution for the unknowns multiplied by a likelihood for the outcome(s) conditional on the unknowns. In this study, since we had yes/no outcomes, we used a conditional binomial model with a logit link function. Since we had little a priori confidence that the parameters would be close to zero, we chose to use Student's t distribution as the prior distribution for parameter estimation, i.e., heavier tails than the normal shape. For the NNW model, we used a logistic activation function for the output layer and tangens hyperbolicus for the hidden layer(s). We also tested different hidden layer structures for the NNW, ranging from one to three layers and two to five neurons for each layer. All models need some prior specifications, but most of the settings are straightforward and likely do not severely affect the classification results, for example, the number of repeats of cross-validations (10 in this study) and the number of trees (as long as we specify a large number).

We used stepwise variable selection in ordinary logistic regression and projection predictive variable selection in Bayesian logistic regression.[56,57] Variable selection for other models was based on either a significance test ($P < 0.05$) or relative importance score ($< 1\%$).[29] In addition to using a

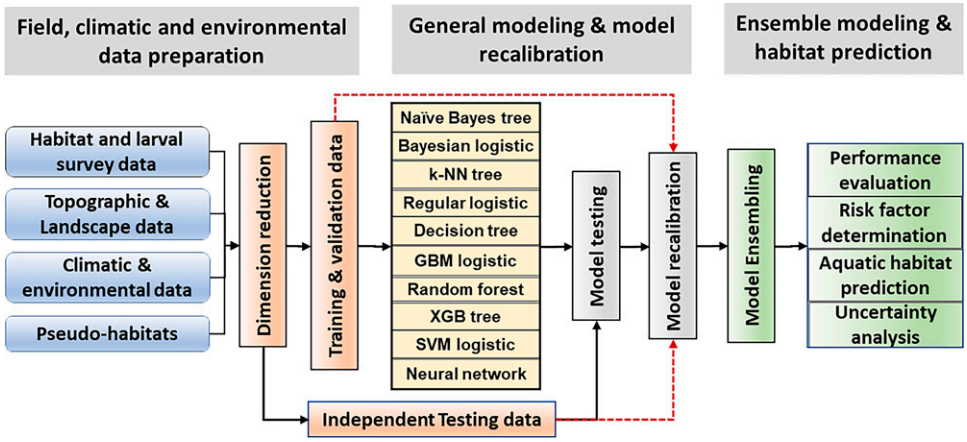

FIGURE 2. Flowchart of the modeling and habitat prediction process. GBM = gradient-boosted machine; kNN = kth nearest neighbor tree; Logistic = logistic regression; SVM = support vector machine; XGB = extreme gradient boosted.

separate data set for validation, to further reduce predictive bias and uncertainty (i.e., variance of performance estimates), we used 10-fold cross-validation for the training of all models except NNW, since we had a large training data set.[29,53,58,59]

For aquatic habitat identification, we used both field-observed aquatic habitats (yes) and pseudohabitats (no). For the identification of larva-positive habitats, we excluded the pseudohabitats.

Comparisons of model performance. We used several criteria to evaluate the performance of different models. The overall prediction accuracy was calculated for each model using both validation and testing data sets. Validation was done using the original model built on the training data set. Prediction accuracy of the final model was assessed using the testing data set, and the model was calibrated using the validation data set. Since we had binary outcomes, we also calculated the sensitivity, specificity, AUC, positive/negative likelihood ratios, and positive/negative predictive values for each model.[37,59,60] Agreement between observed and predicted records was measured using the kappa statistic.[61,62]

To assess other differences in model performance, we checked the relative importance of variables selected for model predictions. Variable relative importance was measured using scaled relative importance ranging from 0 (least important) to 100 (most important).

Ensembling of models. Since different models may select different variables for predictions (i.e., harness different aspects of the data), the models likely do not contribute equally to the ensembled model. To obtain robust predictions, the ensembling process can be treated as a classic case of simplex optimization. Assuming p_i is the output from each model and w_i is the weight for each model such that $\sum_{i=1}^{n} w_i = 1$, where n is the total number of models, the predicted value using the ensembled model is

$$p = \sum_{i=1}^{n} w_i p_i.$$

The objective function is a maximum likelihood function (MLF), i.e.,

$$Maximize \prod p^{y}(1-p)^{1-y},$$

where y is the response flag for the observations. This is a classic logistic regression type of MLF. Thus, the weights, w_i, can be estimated using logistic regression analysis. We also examined the weight estimates by using the simple average (equal weight for all models), the most votes (predictions by the most models), and neural network models (to measure potential nonlinearity).

The ensembled model used predictions based on all data sets; i.e., models were calibrated by all data sets rather than the training data set alone.

Ensembling of risk factors. Currently, there is no standard method for examining variable importance in ensembled models.[20,25,29,60] We proposed the following approach. For each model, the top 20 most important (by relative influence) risk factors (predictors) were selected. For any model (e.g., logistic regression) that ended up with < 20 significant risk factors after variable selection, only the significant risk factors were selected. The risk factors were ranked based on the votes of the models (i.e., how many models selected each risk factor), and the top 20 ranked risk factors were selected as the important risk factors. The variable importance (*RI*) was measured as the weighted average of the relative influence from each individual model:

$$RI = \sum_{i=1}^{n} w_i \overline{RI}_i,$$

where $\overline{RI}_i$ was the standardized (scaled) relative importance RI_i from model i, i.e.,

$$\overline{RI}_i = \frac{RI_i - \min_i RI_i}{\max_i RI_i - \min_i RI_i}.$$

The weights, w_i, were the same as estimated by the ensembled model.

Assessing prediction uncertainty. Model prediction uncertainty has also been less studied or not mentioned in most previous studies. We proposed to use the square-rooted mean squared error against the ensembled model predictions to measure the prediction uncertainty,[29,37] i.e.,

$$MSE^2 = \sum_{i=1}^{n} (p_i - p)^2 / n$$

where p was the probability predicted by the ensembled model and p_i was the probability predicted by each individual model.

All data analyses were conducted using R 4.0.3 (The R Foundation for Statistical Computing, Vienna, Austria). The following packages were used in this study: *caret* for training and validation data splitting and for the kNN tree; *factoextra*, *Rcpp*, *FactoMineR*, *ggplot2*, *Hmisc*, and *reshape2* for data dimension reduction; *dplyr*, *ROCR*, *caTools*, *mlbench*, *MLmetrics*, *MASS*, *plyr*, and *tidyverse* for logistic regression; *mboost*, *gbm*, and *cvAUC* for GBM logistic; *arm*, *logicFS*, *LogicREG*, and *mcbiopi* for SVM logistic; *GGally*, *bayesplot*, *rstanarm*, *loo*, *projpred*, and *reportROC* for Bayesian logistic; *party*, *rpart*, and *pROC* for decision tree; *randomForest* for random forest; *Metrics* and *xgboost* for XGB tree; *neuralnet*, *devtools*, and *usethis* for NN models; *deepnet* for DeepNN models; *e1071* for Naïve Bayes classifier; and *vip* for measuring variable importance. The above-mentioned packages are listed in the order in which they appeared in the code. Some packages were used for multiple models; for example, *caret* was used for data partitioning and in logistic regression, SVM logistic, GBM tree, and XGB tree, among others.

RESULTS

This study collected data from 49,261 aquatic habitats, of which 38,693 had larval survey results (Supplemental Table S1). For aquatic habitat identification, about 42,600 aquatic habitat records were used for model training and validation and 6,600 were used for model testing. For larva-positive habitat identification, about 32,000 records were used for model training and validation and 6,000 were used for model testing.

Prediction accuracy. For aquatic habitat identification, prediction accuracy for the testing data varied substantially among different models (Figure 3A); the overall prediction accuracy varied from 58% (neural network) to 82% (ensembled model) (Supplemental Table S3). For predicting the aquatic habitats, the ensembled model (accuracy,

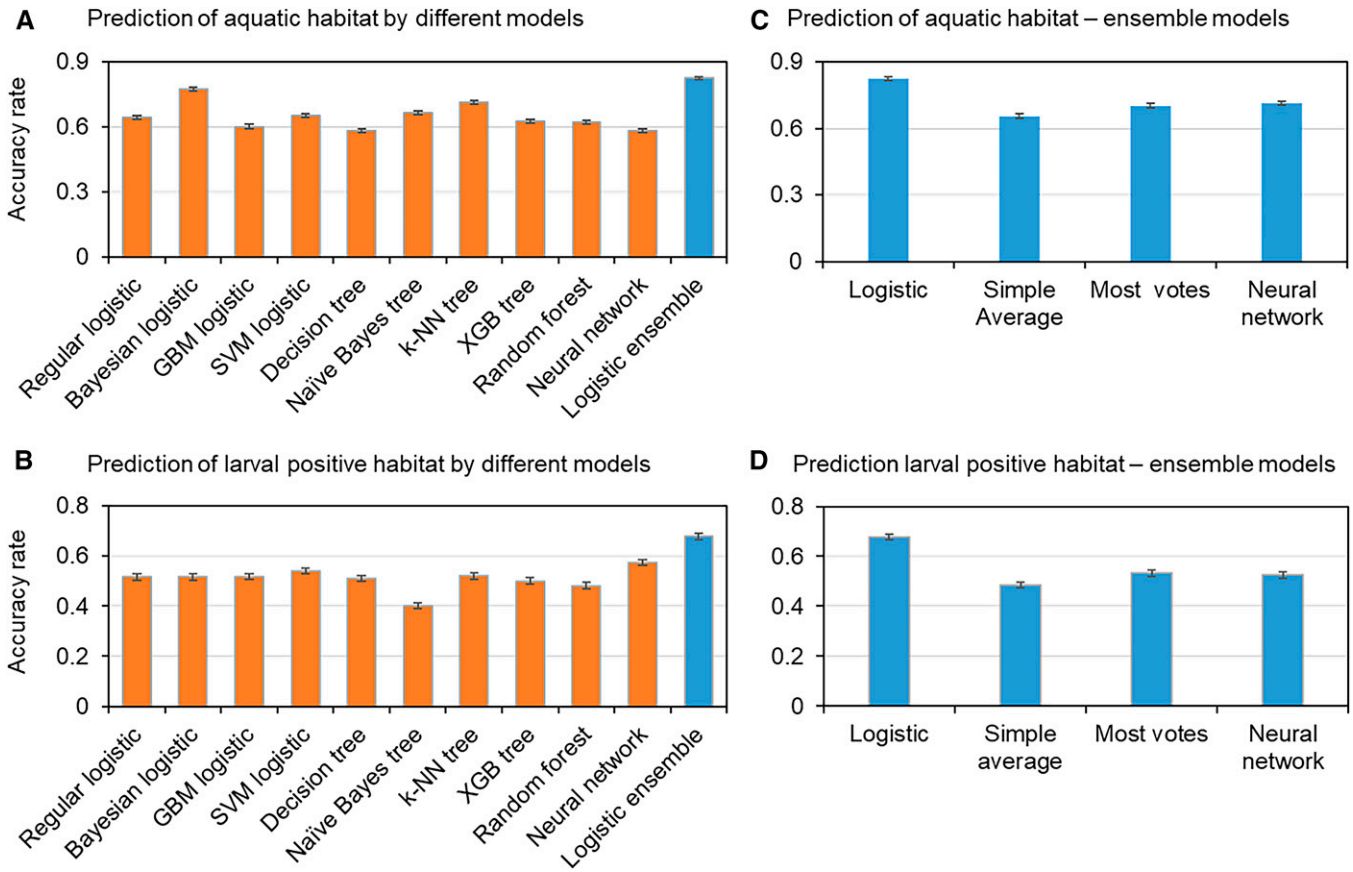

FIGURE 3. Accuracy of predictions of different models. (**A**) Prediction of aquatic habitats by different models. (**B**) Prediction of larva-positive habitats by different models. (**C**) Ensembled models for the prediction of aquatic habitats. (**D**) Ensembled model for the prediction of larva-positive habitats.

89.7%) produced the best accuracy, although it was only 5% better than that of the Bayesian logistic model (Figure 3A). For the overall prediction of aquatic habitats, logistic regression ensembling outperformed simple average, most-votes, and neural network-ensembled models (Figure 3C).

Similarly, for larval infestation identification, prediction accuracy varied substantially among different models (Figure 3B), and all models had $<$ 60% prediction accuracy for the testing data except the ensembled model (accuracy, 68%) (Supplemental Table S4). The logistic regression ensembled model overperformed other models (Figure 3B). Similarly, logistic regression ensembling overperformed simple average, most-votes, and neural network ensembled models (Figure 3D).

On average, the logistic regression ensembled model had about 20% greater accuracy than each individual model for the prediction of aquatic and larva-positive habitats (Figure 3, Supplemental Tables S3 and S4).

Agreement between observations and predictions. For the aquatic habitat identification, prediction sensitivity of the 10 models ranged from 33% to 100%, specificity ranged from 0% to 99%, and the overall kappa agreement was low ($<$ 0.6) for all models (Supplemental Table S3). The ensembled model had a kappa of 0.62 (i.e., moderate agreement), 88% sensitivity, 73% specificity, and 80% AUC, that is, reasonably good accuracy compared with that of pure chance (AUC of 0.5) (Supplemental Table S3). The ensembled model had about 30% higher sensitivity than the average of individual models.

For the larva-positive habitats, all models had relatively low sensitivity and specificity, and all had an AUC $<$ 60% (Supplemental Table S4). Kappa agreement between the observed and predicted larva-positive habitats was poor ($<$ 0.2 for all). The ensembled model had a sensitivity similar to the average of individual models ($\sim$50%), but the ensembled model had 50% higher specificity than the average of individual models (Supplemental Table S4).

For the final model, aquatic habitat prediction had an AUC of 97.9%, a kappa agreement of 0.81 (95% CI, 0.805, 0.822), a sensitivity of 95.6%, and a specificity of 85.0%. For the larva-positive habitat prediction, the AUC was 74.1%, the kappa agreement was 0.35 (95% CI, 0.334, 0.365), the sensitivity was 50.4%, and the specificity was 83.2%.

Identification of risk factors. In most cases, different models selected different groups of important risk factors because of the difference in variable selection algorithms (Supplemental Tables S5 and S6). In some models, a few risk factors had clear, high relative influences, for example, the GBM logistic and XGB tree models for identifying aquatic habitats (Supplemental Table S5). In other models, many factors influenced the model predictions, for example, random forest for the prediction of larva-positive habitats (Supplemental Table S6).

If we take the ensembled model as the final model, the digital elevation model (relative influence [rel.inf], 100%), geomorphon class (relative influence scaled to 0% to 100% [rel.inf], 72.9%), and the amount of precipitation 2 months prior to the survey (rel.inf, 45.9%) were the top three risk factors determining the aquatic habitats (Table 1). The relative influence of the other factors was significantly lower than that of the top three (Table 1). Although many factors may have affected larval infestation (Table 2), the top three risk factors were maximum temperature 4 months prior to the survey, the amount of precipitation 3 months prior to the survey, and the distance to the river/streams; however, northness and other factors also had high influence (Table 2). Regardless, the top 10 risk factors were all related to temperature and precipitation with the exception of northness, which is somewhat related to the amount of sunlight received (Table 2).

Mapping potential larval habitats and uncertainty assessment. Risk mapping is rather straightforward. Figure 4 showed the ensembled model-predicted probability of potential aquatic habitats and larva-positive habitats in the study area and the uncertainty of the predictions measured as mean squared error. Prediction of real (not pseudo) aquatic habitat had the lowest uncertainty in the Iguhu area, while the Mbale area had the lowest uncertainty for the prediction of larva-positive habitats (Figure 4).

DISCUSSION

Larval source management is a potentially viable supplement to the currently implemented first-line malaria control tools for use under certain conditions for malaria control and elimination.[1,3,4] The WHO has recommended LSM, and a number of African countries where malaria is endemic have adopted LSM as a key vector control tool parallel to LLINs and IRS or as a supplementary strategy for malaria control and elimination.[3,4,11–14] Implementation of LSM requires a carefully designed strategy and effective planning; identification and mapping of larval sources is a prerequisite for LSM.[11,12,14] The use of climatic/environmental data, especially satellite monitoring data and mathematical models, is a common approach for larval source identification.[17–20,25,36] Ensemble modeling provides high-accuracy larval source predictions; however, standard procedures are lacking, especially for the ensembling method and risk factor selection and evaluation. Here, we proposed a framework for larval source prediction using multimodel ensemble approaches, including methods for model selection, model ensembling, risk factor selection and evaluation, model prediction assessment, larval source mapping, and uncertainty analysis. To illustrate the procedure of the proposed approach, we used 10 years of multisite field observations of larval habitat surveys for model training and validation and independent field data for model testing. Using ensembles of 10 models, we identified three major predictors of aquatic habitat in western Kenya: elevation, geomorphon class, and amount of precipitation 2 months prior. We provided a map of potential malaria vector larval habitats in the study area. The aquatic habitat risk map will be valuable for LSM planning.

Model selection in ensemble modeling. Data mining scientists recommend using diverse and independent models for ensemble modeling.[29,37] Most medical and biological application studies have not mentioned how individual models were selected for ensemble modeling.[20,21,24–26,32,60] However, model selection may affect the eventual predictions of the ensembled model. For example, in this study, we tested both GBM logistic and GBM classification models. The two models selected exactly the same group of variables with minimal difference in variable importance and very similar prediction results (results not shown); note that the two models utilize almost exactly the same algorithm for variable selection. If we included both models in the ensembled model and used accuracy or AUC as the weight,[24,25] we would likely overweight (create bias toward) these models. Indeed, stepwise selection selected only the

TABLE 1
Relative influence of top 20 risk factors for prediction of aquatic habitats

Variables	Votes*	Raw.inf	Rel.inf
Altitude	9	21.81	100
Geomorphon classes	9	16.51	72.9
Precipitation 2 months prior	4	11.24	45.9
Minimum temperature 1 month prior	4	7.62	27.4
Soil adjusted vegetation index	6	6.90	23.7
Maximum temperature 2 months prior	4	5.99	19.0
Land surface temperature daytime quantile 4	4	5.96	18.9
Land surface temperature daytime quantile 2	4	5.15	14.7
Distance to river	5	5.09	14.4
Modified soil adjusted vegetation index	7	4.69	12.4
Maximum temperature 1 month prior	4	4.45	11.1
Normalized difference built-up index	4	4.20	9.9
Curvature	8	4.13	9.5
Average enhanced vegetation index	6	3.99	8.8
Minimum temperature 4 months prior	5	3.64	7.0
Minimum temperature 2 months prior	5	3.61	6.8
Normalized difference vegetation index	5	3.30	5.2
Population density	4	3.19	4.7
Land surface temperature daytime quantile 3	6	2.82	2.8
Distance to major road	5	2.27	0

Raw.inf = original weighted relative influence of the 10 models; Rel.inf = relative influence scaled to 0% to 100%.
*Votes, number of votes over the 10 models.

TABLE 2
Relative influence of top 20 risk factors for prediction of larva-positive habitats

Variables selected	Votes*	Raw.inf	Rel.inf
Maximum temperature 4 months prior	5	54.4	100
Precipitation 3 months prior	5	39.0	69.1
Distance to river/stream	7	35.0	61.2
Northness	6	33.2	57.6
Average daytime land surface temperature	6	30.1	51.3
Precipitation 2 months prior	5	29.4	49.9
Maximum temperature 5 months prior	5	26.9	45.0
Precipitation 4 months prior	5	24.1	39.3
Nighttime land surface temperature S3	6	22.1	35.3
Precipitation variability in February	8	17.2	25.5
DEM	8	14.7	20.5
Minimum temperature 3 months prior	6	14.3	19.6
Maximum temperature 2 months prior	6	12.7	16.6
Daytime land surface temperature S2	5	11.4	13.8
Average enhanced vegetation index	5	9.6	10.3
Normalized difference vegetation index	5	7.2	5.6
Maximum temperature 1 month prior	5	6.3	3.7
Valley bottom flatness	7	5.8	2.7
Daytime land surface temperature S4	5	5.1	1.4
Precipitation 1 month prior	5	4.4	0

S1 to S4 means quantile 1 to quantile 4. DEM = digital elevation model; Raw.inf = original weighted relative influence of the 10 models; Rel.inf = relative influence scaled to 0% to 100%.
*Votes, number of votes over the 10 models.

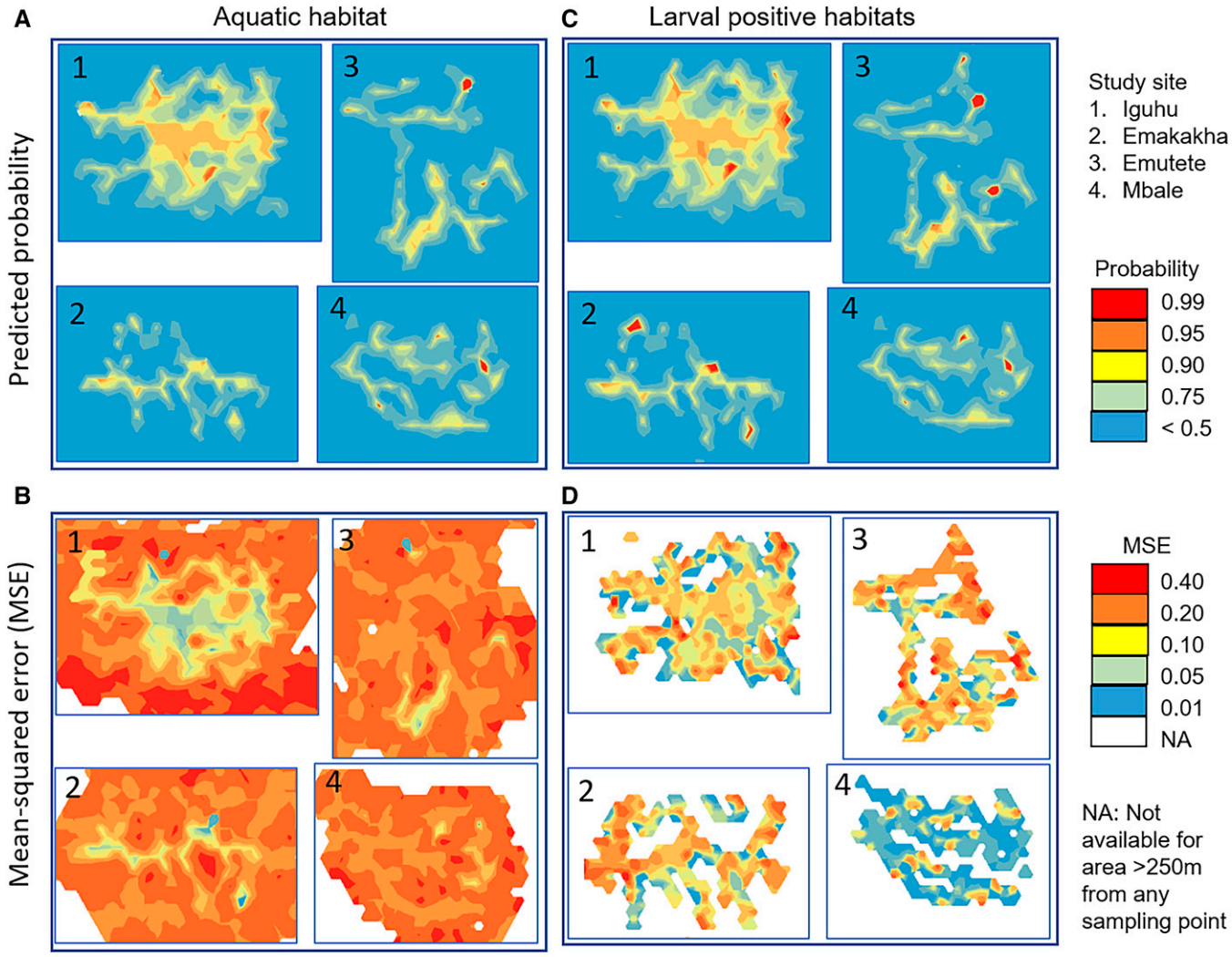

FIGURE 4. Predicted probability (top panels) and mean squared error (MSE) (bottom panels) of aquatic habitats (left panels) and *Anopheles* larva-positive habitats (right panels).

GBM logistic model (results not shown); i.e., the GBM classification model may be redundant. In addition, if we used accuracy or AUC $>$ 0.7 for model selection in this study,[24,25] we would actually end up with no model for larva-positive habitat prediction. This is a reminder that the combination of artificially selected weights/models may severely affect the results of the ensembled model.

Relative importance of risk factors. Currently, there is no standard method in ensemble modeling for selecting the important predictors and estimating the relative influence of these predictors.[29,37] In cancer research, Liu et al. used the sum of relative importance across all models,[60] which is equivalent to the simple average over all models, i.e., treating all models equally. Bose et al. used "frequency of occurrence" for variable selection,[21] which is essentially the most-votes method. In many other studies, researchers have listed the variable importance of all models but have not provided an overall measure of the relative influence of predictors.[18,20,24,25,36] Because different models perform quite differently in risk prediction and in variable selection, the most-votes method is a viable way to select the important variables, as most models selected these variables. Similarly, we may not want to treat all models equally in evaluating variable influence; i.e., unequal-weight weighted average of relative importance may be more reasonable. In this study, we found that logistic regression-estimated weights outperformed simple average and most-votes methods in terms of prediction accuracy. It is important to note that the significance of variables as determinants of larval habitat suitability may vary depending on the landscape settings.[25,43]

Number of risk factors. The selection of the top 20 risk factors in this study was arbitrary. In our previous study of a multi-indicator approach for assessing malaria risks, we started with $>$ 200 variables, and the final GBM logistic model selected only 19 significant variables.[43] Similarly, in a study by Zheng et al. assessing monthly distributions of *Aedes albopictus* in China, the authors started with a similar $>$ 200 variables and the final model used only 17 variables.[63] Risk analyses by Solano-Villarreal et al. using boosted regression also ended up with 18 significant variables.[64] Several other disease risk analysis studies selected $<$ 20 predictors in their final models.[65,66] Practically, if we want to find the key risk factors, we should limit the number of

candidate factors. In this context, we think the top 20 candidate predictors may be enough. In this study, each model selected 20 variables, and there were about 50 candidate predictors in the ensembled model, but not all of them were equally important. In fact, in several models, such as the SVM logistic and XGB tree models, there were fewer than five key predictors based on their relative importance in the model, and the ensembled model had only three to five key predictors of habitats. Practically, selecting the top 20 important variables may be enough.

Larva-positive habitat identification. The accuracy of larva-positive habitat predictions and the kappa agreement between predictions and observations were low in this study. This is likely due to the selection of predictors at the initial stage. We used climatic and satellite-observed environmental variables to predict the larva-positive habitats. However, the presence of larvae in a habitat depends on two major factors: attractants for female breeding and food and environment for larval survival and development. Physical and chemical cues allow female mosquitoes to assess the suitability of potential larval habitats for breeding and hence influence the acceptance of oviposition sites.[67–71] Physical cues originate from vegetation (land cover type and density), water temperature, sunlight, and texture of the substrate, and other biotic factors such as the existence of certain algae are crucial for larval development.[72–76] For example, Munga et al. found that land cover type affects *Anopheles* female oviposition.[49] Sumba et al. found that *A. gambiae* oviposition may be regulated by the daily light-dark cycle.[77] Eneh et al. found that water temperature also affects female oviposition site selection.[78] Factors such as vegetation cover, light-dark cycle, and water temperature may be monitored by ground observations or satellite monitoring.[73,74] However, studies also found that certain biotic cues such as habitat microorganisms (e.g., bacteria) and volatile profiles (e.g., grass volatiles) affect female oviposition habitat selection.[77–81] These factors cannot be monitored through satellite monitoring or simple ground measurements. Biotic variables such as bacteria and grass types may vary from habitat to habitat and change over time. Therefore, it is entirely possible that while a certain aquatic habitat is suitable for female oviposition and larval development at this time, it will not be attractive for female oviposition and/or not be suitable for larval development next time. Thus, predicting larva-positive habitats is more difficult than predicting aquatic habitats. In fact, very few studies have conducted larva-positive habitat prediction.[19]

The major limitation of this study is the selection of models for ensemble modeling. We do not have a strategy for model selection, although data mining experts suggest selecting diverse and independent models for ensemble modeling.[29,37] We used diverse models in this study.[17,25,26] However, it was difficult to decide which models were independent or, more loosely, unrelated. We tried to select models that were not related to each other; we may also want to include other less-related models such as multiple adaptive regression splines, among others.[25,26] The second limitation is the selection of sites using field observations. Ideally, we should include diverse study sites. However, it is difficult to cover a wide variety of aquatic habitats with different ecological backgrounds because of the scope of scientific research; census-style surveillance may include more diverse ecological settings. Including more diverse ecological areas may increase the accuracy of model predictions and the generalizability of the model. For example, in western Kenya, most areas are hilly, and therefore elevation and valley bottom flatness are key factors in determining the suitability of larval habitat, as shown in our results. However, in plain areas, elevation and topographic factors may not be as important as in hilly areas.[43] Future modeling needs to consider balanced samples across different ecological settings. Models need to be recalibrated and revalidated when landscape setting changes or data from other settings are included, but the modeling process can be the same as described in this study. The third limitation is the determination of predictor relative importance in ensemble modeling. As mentioned earlier, there is no standard method for calculating the relative importance in ensemble modeling. The method we described here was purely empirical, and further refinement of predictor selection and relative influence assessment methods as well as standardization are required.

In conclusion, this is the first study to provide a detailed framework for the process of multimodel ensemble modeling for malaria vector habitats, including selection of models, estimation of model weights (statistically optimized), determination of key predictors (most votes), evaluation of the relative influence of key predictors, risk mapping, and prediction uncertainty assessment, including several key components of ensemble modeling that have not been addressed in previous studies. We hope the modeling process we have proposed will be useful for other studies.

Received February 15, 2023. Accepted for publication November 3, 2023.

Published online February 13, 2024.

Note: Supplemental material appears at www.ajtmg.org.

Financial support: This study was funded by the National Institutes of Health (grant numbers D43 TW001505 and U19 AI129326).

Disclosure: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' addresses: Guofa Zhou, Ming-Chieh Lee, Xiaoming Wang, Daibin Zhong, and Guiyun Yan, Program in Public Health, University of California, Irvine, CA, E-mails: zhoug@hs.uci.edu, mingchil@uci.edu, xiaomiw1@uci.edu, xiaomiw1@hs.uci.edu, xiaomiw1@uci.edu, and guiyuny@uci.edu. Andrew K. Githeko, Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya, E-mail: githeko@yahoo.com.

REFERENCES

1. WHO, 2021. *World Malaria Report 2021.* Geneva, Switzerland: World Health Organization.
2. WHO, 2018. *Global Report on Insecticide Resistance in Malaria Vectors: 2010–2016.* Geneva, Switzerland: World Health Organization.
3. WHO, 2013. *Larval Source Management—A Supplementary Measure for Malaria Vector Control: An Operational Manual.* Geneva, Switzerland: World Health Organization. Available at: http://apps.who.int/iris/bitstream/10665/85379/1/9789241505604_eng.pdf.
4. WHO, 2015. *Global Technical Strategy for Malaria 2016–2030.* Geneva, Switzerland: World Health Organization.
5. Afrane YA, Mweresa NG, Wanjala CL, Gilbreath TM 3rd, Zhou G, Lee MC, Githeko AK, Yan G, 2016. Evaluation of long-lasting microbial larvicide for malaria vector control in Kenya. *Malar J 15:* 577.

6. Zhou G, Lo E, Githeko AK, Afrane YA, Yan G, 2020. Long-lasting microbial larvicides for controlling insecticide resistant and outdoor transmitting vectors: a cost-effective supplement for malaria interventions. *Infect Dis Poverty 9:* 162.
7. Choi L, Majambere S, Wilson AL, 2019. Larviciding to prevent malaria transmission. *Cochrane Database Syst Rev 8:* CD012736.
8. Dambach P, Winkler V, Bärnighausen T, Traoré I, Ouedraogo S, Sié A, Sauerborn R, Becker N, Louis VR, 2020. Biological larviciding against malaria vector mosquitoes with *Bacillus thuringiensis israelensis* (Bti)—long term observations and assessment of repeatability during an additional intervention year of a large-scale field trial in rural Burkina Faso. *Glob Health Action 13:* 1829828.
9. Fillinger U, Ndenga B, Githeko A, Lindsay SW, 2009. Integrated malaria vector control with microbial larvicides and insecticide treated nets in the western Kenyan highlands: a controlled trial. *Bull World Health Organ 87:* 655–665.
10. Imbahale SS, Githeko A, Mukabana WR, Takken W, 2012. Integrated mosquito larval source management reduces larval numbers in two highland villages in western Kenya. *BMC Public Health 12:* 362.
11. Kenya Ministry of Health, 2019. *Kenya Malaria Strategy 2019–2023.* Nairobi, Kenya: National Malaria Control Programme, Ministry of Health.
12. Ethiopia Ministry of Health, 2020. *Malaria Elimination Strategic Plan: 2021–2025.* Addis Ababa, Ethiopia: Ethiopia Ministry of Health.
13. South Africa Department of Health, 2019. *Malaria Elimination Strategic Plan 2019–2023.* Pretoria: National Malaria Control Programme, Ministry of Health, South Africa.
14. Mozambique Ministry of Health, 2017. *Malaria Strategic Plan 2017–2022.* Maputo, Mozambique: Ministry of Health.
15. Lippi CA, Stewart-Ibarra AM, Loor MEFB, Zambrano JED, Lopez NAE, Blackburn JK, Ryan SJ, 2019. Geographic shifts in *Aedes aegypti* habitat suitability in Ecuador using larval surveillance data and ecological niche modeling: implications of climate change for public health vector control. *PLoS Negl Trop Dis 13:* e0007322.
16. Shoraka HR, Sofizadeh A, Mehravaran A, 2020. Larval habitat characteristics and predicting the distribution of *Culex tritaeniorhynchus* using maximum entropy (MaxEnt) model in Golestan Province (north of Iran). *J Vector Borne Dis 57:* 259–267.
17. Li L, Bian L, Yakob L, Zhou G, Yan G, 2011. Analysing the generality of spatially predictive mosquito habitat models. *Acta Trop 119:* 30–37.
18. McCann RS, Messina JP, MacFarlane DW, Bayoh MN, Vulule JM, Gimnig JE, Walker ED, 2014. Modeling larval malaria vector habitat locations using landscape features and cumulative precipitation measures. *Int J Health Geogr 13:* 17.
19. Byrne I, Aure W, Manin BO, Vythilingam I, Ferguson HM, Drakeley CJ, Chua TH, Fornace KM, 2021. Environmental and spatial risk factors for the larval habitats of *Plasmodium knowlesi* vectors in Sabah, Malaysian Borneo. *Sci Rep 11:* 11810.
20. Beeman SP, Morrison AM, Unnasch TR, Unnasch RS, 2021. Ensemble ecological niche modeling of West Nile virus probability in Florida. *PLoS One 16:* e0256868.
21. Bose S, Das C, Banerjee A, Ghosh K, Chattopadhyay M, Chattopadhyay S, Barik A, 2021. An ensemble machine learning model based on multiple filtering and supervised attribute clustering algorithm for classifying cancer samples. *PeerJ Comput Sci 7:* e671.
22. De Gooijer JG, Hyndman RJ, 2006. 25 years of time series forecasting. *Int J Forecast 22:* 443–473.
23. Khanna D, Rana PS, 2020. Improvement in prediction of antigenic epitopes using stacked generalisation: an ensemble approach. *IET Syst Biol 14:* 1–7.
24. Osamor VC, Okezie AF, 2021. Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis. *Sci Rep 11:* 14806.
25. Rhodes CG et al., 2022. *Anopheles albimanus* (Diptera: Culicidae) ensemble distribution modeling: applications for malaria elimination. *Insects 13:* 221.
26. Sinka ME, Pironon S, Massey NC, Longbottom J, Hemingway J, Moyes CL, Willis KJ, 2020. A new malaria vector in Africa: Predicting the expansion range of *Anopheles stephensi* and identifying the urban populations at risk. *Proc Natl Acad Sci USA 117:* 24900–24908.
27. Tao X, Chi O, Delaney PJ, Li L, Huang J, 2021. Detecting depression using an ensemble classifier based on Quality of Life scales. *Brain Inform 8:* 2.
28. Xian P et al., 2019. Current state of the global operational aerosol multi-model ensemble: an update from the International Cooperative for Aerosol Prediction (ICAP). *Q J R Meteorol Soc 145 (Suppl 1):* 176–209.
29. Kotu V, Deshpande B, 2019. *Data Science: Concepts and Practice*, 2nd ed. Amsterdam, The Netherlands: Morgan Kaufman Publishers.
30. Yoo BH, Kim J, Lee BW, Hoogenboom G, Kim KS, 2020. A surrogate weighted mean ensemble method to reduce the uncertainty at a regional scale for the calculation of potential evapotranspiration. *Sci Rep 10:* 870.
31. Adhikari R, Agrawal RK, 2014. Performance evaluation of weights selection schemes for linear combination of multiple forecasts. *Artif Intell Rev 42:* 529–548.
32. Oidtman RJ et al., 2021. Trade-offs between individual and ensemble forecasts of an emerging infectious disease. *Nat Commun 12:* 5379.
33. Shahhosseini M, Hu G, Khaki S, Archontoulis SV, 2021. Corn yield prediction with ensemble CNN-DNN. *Front Plant Sci 12:* 709008.
34. Monaghan AJ, Eisen RJ, Eisen L, McAllister J, Savage HM, Mutebi JP, Johansson MA, 2019. Consensus and uncertainty in the geographic range of *Aedes aegypti* and *Aedes albopictus* in the contiguous United States: multi-model assessment and synthesis. *PLOS Comput Biol 15:* e1007369.
35. Proestos Y, Christophides GK, Ergüler K, Tanarhte M, Waldock J, Lelieveld J, 2015. Present and future projections of habitat suitability of the Asian tiger mosquito, a vector of viral pathogens, from global climate simulation. *Philos Trans R Soc Lond B Biol Sci 370:* 20130554.
36. Wieland R, Kuhls K, Lentz HHK, Conraths K, Kampen H, Werner D, 2021. Combined climate and regional mosquito habitat model based on machine learning. *Ecol Modell 452:* 109594.
37. Nisbet R, Miner G, Yale K, 2018. *Handbook of Statistical Analysis and Data Mining Applications*, 2nd ed. London, United Kingdom: Elsevier, Inc.
38. Kapesa A, Kweka EJ, Atieli H, Kamugisha E, Zhou G, Githeko AK, Yan G, 2017. Why some sites are responding better to anti-malarial interventions? A case study from western Kenya. *Malar J 16:* 498.
39. Kapesa A, Kweka EJ, Atieli H, Afrane YA, Kamugisha E, Lee MC, Zhou G, Githeko AK, Yan G, 2018. The current malaria morbidity and mortality in different transmission settings in Western Kenya. *PLoS One 13:* e0202031.
40. The Malaria Atlas Project, 2020. *Plasmodium falciparum* PR 2-10 version 2020. *Plasmodium falciparum* parasite rate in 2–10 year olds globally, 2000–2019. Available at: https://data.malariaatlas.org/maps?layers=Malaria:202206_Global_Pf_Parasite_Rate. Accessed March 3, 2022.
41. Zhou G, Afrane YA, Vardo-Zalik AM, Atieli H, Zhong D, Wamae P, Himeidan YE, Minakawa N, Githeko AK, Yan G, 2011. Changing patterns of malaria epidemiology between 2002 and 2010 in Western Kenya: the fall and rise of malaria. *PLoS One 6:* e20318.
42. Zhou G, Lee MC, Githeko AK, Atieli HE, Yan G, 2016. Insecticide-treated net campaign and malaria transmission in Western Kenya: 2003–2015. *Front Public Health 4:* 153.
43. Zhou G, Zhong D, Lee MC, Wang X, Atieli HE, Githure JI, Githeko AK, Kazura J, Yan G, 2021. Multi-indicator and multi-step assessment of malaria transmission risks in Western Kenya. *Am J Trop Med Hyg 104:* 1359–1370.
44. Derua YA, Kahindi SC, Mosha FW, Kweka EJ, Atieli HE, Wang X, Zhou G, Lee MC, Githeko AK, Yan G, 2018. Microbial larvicides for mosquito control: impact of long lasting formulations of *Bacillus thuringiensis* var. *israelensis* and *Bacillus sphaericus* on non-target organisms in western Kenya highlands. *Ecol Evol 8:* 7563–7573.

45. Himeidan YE, Zhou G, Yakob L, Afrane Y, Munga S, Atieli H, El-Rayah el-A, Githeko AK, Yan G, 2009. Habitat stability and occurrences of malaria vector larvae in western Kenya highlands. *Malar J 8:* 234.
46. Kweka EJ, Zhou G, Munga S, Lee MC, Atieli HE, Nyindo M, Githeko AK, Yan G, 2012. Anopheline larval habitats seasonality and species distribution: a prerequisite for effective targeted larval habitats control programmes. *PLoS One 7:* e52084.
47. Minakawa N, Munga S, Atieli F, Mushinzimana E, Zhou G, Githeko AK, Yan G, 2005. Spatial distribution of anopheline larval habitats in Western Kenyan highlands: effects of land cover types and topography. *Am J Trop Med Hyg 73:* 157–165.
48. Minakawa N, Omukunda E, Zhou G, Githeko A, Yan G, 2006. Malaria vector productivity in relation to the highland environment in Kenya. *Am J Trop Med Hyg 75:* 448–453.
49. Munga S, Minakawa N, Zhou G, Barrack OO, Githeko AK, Yan G, 2005. Oviposition site preference and egg hatchability of *Anopheles gambiae*: effects of land cover types. *J Med Entomol 42:* 993–997.
50. Munga S, Minakawa N, Zhou G, Mushinzimana E, Barrack OO, Githeko AK, Yan G, et al., 2006. Association between land cover and habitat productivity of malaria vectors in western Kenyan highlands. *Am J Trop Med Hyg 74:* 69–75.
51. Mushinzimana E et al., 2006. Landscape determinants and remote sensing of anopheline mosquito larval habitats in the western Kenya highlands. *Malar J 5:* 13.
52. Zhou G et al., 2013. Modest additive effects of integrated vector control measures on malaria prevalence and transmission in western Kenya. *Malar J 12:* 256.
53. Kuhn M, Johnson K, 2013. *Applied Predictive Modeling.* New York, NY: Springer.
54. Russell S, Norvig P, 2021. *Artificial Intelligence: A Modern Approach*, 4th ed. London, United Kingdom: Pearson Education, Inc.
55. Sarnovsky M, Kolarik M, 2021. Classification of the drifting data streams using heterogeneous diversified dynamic class-weighted ensemble. *PeerJ Comput Sci 7:* e459.
56. Piironen J, Paasiniemi M, Vehtari A, 2018. Projective inference in high-dimensional problems: prediction and feature selection. arXiv preprint. Available at: https://arxiv.org/abs/1810.02406.
57. Vehtari A, Gelman A, Gabry J, 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput 27:* 1413–1432.
58. Avali VR, Cooper GF, Gopalakrishnan V, 2014. Application of Bayesian logistic regression to mining biomedical data. *AMIA Annu Symp Proc 2014:* 266–273.
59. Xiao W, Jing L, Xu Y, Zheng S, Gan Y, Wen C, 2021. Different data mining approaches based medical text data. *J Healthc Eng 2021:* 1285167.
60. Liu MM, Wen L, Liu YJ, Cai Q, Li LT, Cai YM, 2018. Application of data mining methods to improve screening for the risk of early gastric cancer. *BMC Med Inform Decis Mak 18 (Suppl 5):* 121.
61. McHugh ML, 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb) 22:* 276–282.
62. Waugh SM, He J, 2019. Inter-rater agreement estimates for data with high prevalence of a single response. *J Nurs Meas 27:* 152–161.
63. Zheng X, Zhong D, He Y, Zhou G, 2019. Seasonality modeling of the distribution of *Aedes albopictus* in China based on climatic and environmental suitability. *Infect Dis Poverty 8:* 98.
64. Solano-Villarreal E et al., 2019. Malaria risk assessment and mapping using satellite imagery and boosted regression trees in the Peruvian Amazon. *Sci Rep 9:* 15173.
65. Bui Q-T, Nguyen Q-H, Pham VM, Pham MH, Tran AT, 2018. Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers. *Geocarto Int 34:* 1300–1314.
66. Kabaria CW, Molteni F, Mandike R, Chacky F, Noor AM, Snow RW, Linard C, 2016. Mapping intra-urban malaria risk using high resolution satellite imagery: a case study of Dar es Salaam. *Int J Health Geogr 15:* 26.
67. Blackwell A, Johnson SN, 2000. Electrophysiological investigation of larval water and potential oviposition chemoattractants for *Anopheles gambiae* s.s. *Ann Trop Med Parasitol 94:* 389–398.
68. Herrera-Varela M, Lindh J, Lindsay SW, Fillinger U, 2014. Habitat discrimination by gravid *Anopheles gambiae* sensu lato—a push-pull system. *Malar J 13:* 133.
69. Lindh JM, Okal MN, Herrera-Varela M, Borg-Karlson AK, Torto B, Lindsay SW, Fillinger U, 2015. Discovery of an oviposition attractant for gravid malaria vectors of the *Anopheles gambiae* species complex. *Malar J 14:* 119.
70. Warburg A, Faiman R, Shtern A, Silberbush A, Markman S, Cohen JE, Blaustein L, 2011. Oviposition habitat selection by *Anopheles gambiae* in response to chemical cues by *Notonecta maculata*. *J Vector Ecol 36:* 421–425.
71. Wondwosen B, Hill SR, Birgersson G, Seyoum E, Tekie H, Ignell RA, 2017. (maize)ing attraction: gravid *Anopheles arabiensis* are attracted and oviposit in response to maize pollen odours. *Malar J 16:* 39.
72. Antonio-Nkondjio C, Ndo C, Costantini C, Awono-Ambene P, Fontenille D, Simard F, 2009. Distribution and larval habitat characterization of *Anopheles moucheti, Anopheles nili*, and other malaria vectors in river networks of southern Cameroon. *Acta Trop 112:* 270–276.
73. Christiansen-Jucht C, Parham PE, Saddler A, Koella JC, Basáñez MG, 2014. Temperature during larval development and adult maintenance influences the survival of *Anopheles gambiae* s.s. *Parasit Vectors 7:* 489.
74. Getachew D, Balkew M, Tekie H, 2020. *Anopheles* larval species composition and characterization of breeding habitats in two localities in the Ghibe River Basin, southwestern Ethiopia. *Malar J 19:* 65.
75. Olayemi IK, Ojo VO, 2013. Immature development of the malaria vector mosquito, *Anopheles gambiae* S.L. (Diptera: Culicidae), in relation to soil-substrate organic matter content of larval habitats in northcentral Nigeria. *Pak J Biol Sci 16:* 135–140.
76. Tuno N, Okeka W, Minakawa N, Takagi M, Yan G, 2005. Survivorship of *Anopheles gambiae* sensu stricto (Diptera: Culicidae) larvae in western Kenya highland forest. *J Med Entomol 42:* 270–277.
77. Sumba LA, Guda TO, Deng AL, Hassanali A, Beier JC, Knols BG, 2004. Mediation of oviposition site selection in the African malaria mosquito *Anopheles gambiae* (Diptera: Culicidae) by semiochemicals of microbial origin. *Int J Trop Insect Sci 24:* 260–265.
78. Eneh LK, Fillinger U, Borg Karlson AK, Kuttuva Rajarao G, Lindh J, 2019. *Anopheles arabiensis* oviposition site selection in response to habitat persistence and associated physicochemical parameters, bacteria and volatile profiles. *Med Vet Entomol 33:* 56–67.
79. Asmare Y, Hill SR, Hopkins RJ, Tekie H, Ignell R, 2017. The role of grass volatiles on oviposition site selection by *Anopheles arabiensis* and *Anopheles coluzzii. Malar J 16:* 65.
80. Konopka JK, Task D, Afify A, Raji J, Deibel K, Maguire S, Lawrence R, Potter CJ, 2021. Olfaction in *Anopheles* mosquitoes. *Chem Senses 46:* bjab021.
81. Suh E, Choe DH, Saveer AM, Zwiebel LJ, 2016. Suboptimal larval habitats modulate oviposition of the malaria vector mosquito *Anopheles coluzzii. PLoS One 11:* e0149800.