Computational Methods, 3(2):7–17, July-December: 2026.
DOI: https://doi.org/10.58614/cm322

Computational Methods
ISSN (Online): 3048-7447

Journal homepage: https://jpub.org/journal-details.php?journal-id=43

Research Article

Human-Centered Risk Analytics for Household Malaria Stratification: An Explainable Decision-Support Framework

Aniekan Brown[1], Aniefiok Ukommi[2], Philip Asuquo[3], Bliss Utibe-Abasi Stephen[3*] and Emmanuel Abraham Udoh[4]

[1]*Criminology Department, University of Uyo, Uyo, Nigeria.*
[2]*University of Uyo, Uyo, Nigeria.*
[3]*TETFund Centre of Excellence in Computational Intelligence Research, University of Uyo, Uyo, Nigeria.*
[4]*Computer Engineering Department, University of Uyo, Uyo, Nigeria.*
[*]*Corresponding author: blissstephen@uniuyo.edu.ng*

Article Info

Keywords: *Household malaria risk profiling, Machine learning, Explainable artificial intelligence, Human-in-the-loop decision support, Public health governance.*

AMS: *62P10, 62H30, 68T05, 92C60.*

Received: *15.07.2026;*
Accepted: *28.08.2026;*
Published: *04.09.2026*

© 2026 by the author's. The terms and conditions of the Creative Commons Attribution (CC BY) license apply to this open access article.

Abstract

Household-level malaria risk profiling is a high-impact decision-support problem because public health teams must allocate prevention, surveillance, and follow-up resources under uncertainty. This paper develops and evaluates a human-centered artificial intelligence and machine learning framework for classifying household malaria infection risk while preserving expert oversight. The study uses the supplied project outputs from a synthetic household dataset designed to emulate variables commonly collected in demographic and malaria indicator surveys, including housing materials, bed-net availability and use, proximity to water, household size, education, income, eave openness, flooring, window screening, and binary infection status. The pipeline imputes missing values, transforms skewed environmental distance, encodes ordinal and nominal attributes, balances the training data with synthetic minority over-sampling, trains interpretable and ensemble classifiers, evaluates performance on a stratified hold-out set, and produces global and local explanations using feature-attribution visualizations. Logistic regression achieved the strongest overall accuracy (0.805), area under the receiver operating characteristic curve (0.87), and calibration score (Brier score 0.143), whereas the decision tree achieved the highest infected-household recall (0.756), indicating its usefulness when missed infections are operationally costly. Explainability outputs identified bed-net use, income level, proximity to water, flooring, bed-net availability, window screening, mud walls, household size, and thatch roofs as influential risk drivers. The paper contributes a reproducible AI/ML workflow, a human-in-the-loop review mechanism based on calibrated risk tiers and disagreement triggers, and an ethical governance framework for fairness, accountability, and deployment monitoring. The findings support the thesis that artificial intelligence should augment public health expertise rather than replace human judgment.

1. Introduction

Malaria control requires decisions that are geographically and socially granular. Global and regional mapping studies have shown that malaria transmission is heterogeneous across space and time, and that interventions become more effective when they are aligned with local ecological, demographic, and behavioral risk patterns [1–4]. At the household level, the same logic becomes more operational: households

differ in housing quality, bed-net access, net use, proximity to mosquito breeding habitats, household size, socioeconomic status, and exposure-reducing features such as screened windows and closed eaves. These factors do not determine infection status mechanically, yet they influence exposure and vulnerability in ways that can help public health teams prioritize outreach, education, environmental management, and follow-up testing. Systematic reviews and multi-country analyses indicate that improved housing and structural barriers can reduce malaria risk, while hotspot-oriented thinking highlights the importance of identifying concentrated transmission patterns rather than distributing limited resources uniformly [5–7].

The operational significance of household malaria risk profiling is substantial. In malaria-endemic settings, field teams often make decisions under constraints: diagnostic resources may be limited, vector-control teams cannot visit every household with equal frequency, and surveillance officers must distinguish households needing urgent intervention from those requiring routine monitoring. Traditional approaches use rule-based risk categories, expert judgment, household survey indicators, and sometimes coarse spatial maps. These approaches remain indispensable, but they can struggle to combine many interacting predictors, quantify uncertainty, and update decision rules reproducibly when data distributions change. At the same time, high-quality household survey systems such as Demographic and Health Surveys provide a methodological foundation for standardized individual and household variables that can support micro-epidemiological modelling [8]. The challenge is not simply to predict infection risk, but to transform prediction into transparent, auditable, and equitable decision support.

Artificial intelligence and machine learning offer tools for this challenge because they can learn nonlinear relationships, estimate individual-level probabilities, rank features, and reveal cases where the model is uncertain or where model classes disagree. In health and medicine more broadly, AI systems are increasingly framed as tools that can expand clinical and public health capacity, provided that they are evaluated rigorously and embedded into human workflows [9–11]. The promise is particularly relevant for malaria programs: a household risk model can support triage, guide resource allocation, and make tacit risk logic explicit. However, predictive performance alone is insufficient. If a model cannot be interpreted, monitored, challenged, or contextualized by public health workers, it may produce brittle or inequitable recommendations. Human-centered AI emphasizes reliability, safety, transparency, and meaningful human control, and these principles are central to this study's design [12].

This paper investigates how AI and machine learning can augment human decision-making within household malaria risk profiling. The study uses the supplied project artifacts as primary empirical evidence: a synthetic dataset of 1,000 households, classification reports for three models, normalized confusion matrices, receiver operating characteristic (ROC) and precision–recall curves, a calibration diagram, a visualized decision tree, SHAP global importance plots, and local explanations for high-risk households. The research objective is to develop an explainable AI/ML framework that classifies household malaria infection risk while strengthening, rather than replacing, expert judgment. Three contributions follow: a reproducible modelling workflow, a human-in-the-loop governance mechanism based on calibrated risk tiers, and an ethical risk-management framework for fairness, documentation, monitoring, and accountability.

2. Related Work and Conceptual Basis

The introduction established household malaria profiling as an augmentation problem: models must improve triage while remaining interpretable, governable, and accountable. The literature therefore needs to explain both the epidemiological basis for household risk and the AI design principles that keep prediction connected to human decision-making.

Malaria risk assessment has historically relied on epidemiological surveillance, geospatial modelling, household surveys, vector ecology, and expert-defined intervention strategies. Global endemicity maps demonstrated that malaria risk is not spatially uniform and that mapping can guide intervention planning and evaluation [1, 2]. Later spatial-temporal studies connected intervention coverage with changing parasite prevalence and disease burden, strengthening the case for data-driven control programs [3, 4]. These studies operate at larger spatial scales than household risk profiling, but they establish an important principle: risk stratification is central to malaria governance.

Household-level approaches complement broad mapping by representing local exposure and vulnerability. Housing quality, eave structure, wall and roof material, floor type, and window screening affect mosquito entry and indoor exposure. A systematic review found consistent evidence that improved housing is associated with lower malaria outcomes, and a multi-country analysis of survey data linked improved housing to reduced malaria risk among children in sub-Saharan Africa [5, 6]. Hotspot research similarly argues that concentrated transmission can justify targeted interventions, while also warning that hotspots are complex and require careful operational validation [7]. These findings motivate models that can synthesize multiple household indicators into actionable risk estimates.

Machine learning expands the analytical toolkit for risk prediction by estimating patterns that may be difficult to encode manually. Logistic regression remains attractive because coefficients and odds ratios are comparatively transparent for binary outcomes [13]. Decision trees provide rule-like structures that align with operational reasoning, and ensemble methods such as random forests reduce variance by aggregating multiple randomized trees [14]. Gradient boosting represents a related ensemble family that builds predictive functions sequentially and has influenced many contemporary tabular prediction pipelines [15]. In imbalanced health classification problems, methods such as synthetic minority over-sampling can improve the representation of positive cases during model training [16].

Evaluation must match the decision problem. Accuracy can obscure failures when the positive class is less frequent or more costly to miss. ROC analysis summarizes discrimination across thresholds, while precision–recall curves are especially informative when class prevalence is uneven [17, 18]. Calibration is equally important because public health decisions often depend on probability thresholds rather than labels alone; a model that discriminates well but produces unreliable probabilities can mislead triage [19]. These methodological concerns explain why the supplied project outputs include classification reports, confusion matrices, ROC curves, precision–recall curves, and calibration curves rather than a single metric.

Human-in-the-loop AI systems are designed so that automated predictions remain subject to human interpretation, contestation, and contextualization. Public health teams know whether a household recently received nets, whether water proximity reflects seasonal or permanent breeding habitat, whether local vector behavior has changed, and whether social constraints affect intervention feasibility. Human-centered AI therefore requires interfaces that clarify model capabilities, communicate uncertainty, support correction, and preserve user agency [12, 20]. The literature on automation warns that decision support can generate misuse, disuse, or over-reliance [21, 22]. These risks imply that malaria risk systems should avoid presenting predictions as commands; instead, predictions should be paired with explanations, thresholds, confidence information, and escalation rules. Model explanations such as LIME and SHAP can help users inspect why a prediction was made, although interpretability claims must be treated carefully [23–25]. For high-stakes domains, some scholars argue

that interpretable models should be preferred where they can achieve adequate performance [26].

Ethical deployment requires more than technical accuracy. Health algorithms can reproduce inequities when labels, proxies, missingness, or access patterns encode historical disadvantage. Evidence from population health management shows that biased proxy outcomes can systematically misallocate resources [27]. Fairness research further emphasizes that technical abstractions can fail when they ignore sociotechnical context, including how predictions are used, who bears errors, and who can challenge decisions [28]. Documentation practices such as model cards and dataset datasheets help address these risks by describing intended use, performance, limitations, and data provenance [29, 30]. Reporting guidance for prediction models similarly encourages transparent description of data, preprocessing, validation, and performance [31]. These requirements motivate the methodology that follows, where predictive modelling is deliberately coupled with calibration, explanation, and review triggers.

3. Methodology

Consistent with the review, the empirical workflow treats prediction, interpretation, and oversight as linked design requirements rather than separate tasks. The dataset contains 1,000 household observations and 12 columns, including 11 predictors and a binary target where 1 denotes infected and 0 denotes uninfected. The predictors are grouped into four categories: structural housing variables, prevention behavior, environmental exposure, and socioeconomic or demographic context. Structural variables include wall type, roof type, floor type, eave openness, and the number of screened windows. Prevention variables include bed-net availability and bed-net use on the previous night. Environmental exposure is represented by proximity to a water body in meters. Demographic and socioeconomic variables include household size, education of the household head, and income level. Table 1 summarizes the empirical data profile used by the pipeline.

Table 1: Dataset and modelling profile derived from the supplied project artifacts.

Element	Observed value
Records and predictors	1,000 households; 11 predictors plus binary target
Target distribution	612 uninfected (61.2%); 388 infected (38.8%)
Missingness	238 missing cells in education of household head; no duplicate records
Numeric ranges	Water 0.02–1995.64 m; size 1–15; screened windows 0–6
Validation design	Stratified 80/20 split; hold-out test set of 200 households
Test support	122 uninfected and 78 infected households
Training balance	SMOTE applied to training data after preprocessing

Table 1 shows that the problem is moderately imbalanced and contains household attributes that are both operationally meaningful and ethically sensitive. Data preprocessing was designed to preserve the semantics of these variables while producing a numeric feature matrix. Missing categorical, ordinal, and binary values were imputed using the mode, while missing numerical values would be imputed using the median. In the supplied dataset, missingness occurred in the education variable and was handled through this procedure. Proximity to water was transformed using $\log(1+x)$ to reduce skewness while preserving monotonic interpretation. Education and income were encoded ordinally, reflecting ordered categories. Nominal housing variables were one-hot encoded, producing explicit indicators such as mud wall, concrete wall, thatch roof, earth floor, and cement floor.

Let x_i denote the processed feature vector for household i, $y_i \in \{0,1\}$ denote infection status, and $\widehat{p}_i = f_\theta(x_i)$ denote the model-estimated infection probability. The final decision support output is not only a class label but also a probability and a review tier:

$$T_i = \begin{cases} \text{Low}, & \widehat{p}_i < 0.30, \\ \text{Medium}, & 0.30 \le \widehat{p}_i \le 0.60, \\ \text{High}, & \widehat{p}_i > 0.60. \end{cases} \tag{1}$$

Equation 1 formalizes the human-in-the-loop pathway: high-risk cases, medium-risk cases with contextual concern, and cases with model disagreement are routed to expert review before action.

The modelling stage trained three evaluated classifiers: logistic regression, a decision tree, and a random forest. Logistic regression was configured with class balancing and maximum iterations sufficient for convergence, providing an interpretable linear baseline for binary classification [13]. The decision tree used a maximum depth of five and balanced class weights, producing a visual rule set that can be inspected by public health experts. The random forest used 100 trees and balanced class weights, providing an ensemble benchmark with nonlinear feature interactions [14]. Class imbalance was addressed only within the training split using SMOTE [16]. This order avoids contaminating the hold-out test set with synthetic examples while improving the model's exposure to the infected class during learning.

Evaluation metrics were selected to reflect both statistical and operational goals. Accuracy summarizes overall correctness. Precision for the infected class measures how often predicted high-risk households are truly infected. Recall for the infected class measures how many infected households are found, a critical metric when missed infection risk is costly. F1-score balances precision and recall. ROC AUC assesses threshold-independent ranking performance, and precision–recall curves emphasize positive-class behavior under imperfect prevalence balance [17, 18]. Calibration curves and Brier scores assess whether predicted probabilities can be interpreted as risk estimates [19]. Explainability was implemented at two levels: global SHAP summaries for population-level feature influence and local waterfall explanations for high-risk households [24]. This design makes the model an input to decision-making rather than a replacement for decision-makers.

4. Results and Analysis

The preceding workflow generated quantitative and explanatory outputs that directly test the augmentation thesis. The hold-out results show that no single model dominates every operational criterion. Table 2 reports the supplied classification metrics and figure-derived AUC and

Brier scores.

Table 2: Hold-out test performance for evaluated household malaria risk classifiers. Positive-class metrics refer to infected households.

Model	Acc.	Prec.	Recall	F1	AUC	Brier
Logistic regression	0.805	0.791	0.679	0.731	0.87	0.143
Decision tree	0.765	0.678	0.756	0.715	0.82	0.176
Random forest	0.750	0.741	0.551	0.632	0.84	0.163

Table 2 indicates that logistic regression achieved the highest overall accuracy at 0.805, the strongest ROC AUC at 0.87, and the best Brier score at 0.143. It also produced the highest infected-class precision at 0.791. These results indicate that the logistic model is the most reliable general-purpose risk scorer among the evaluated models, particularly when calibrated probability estimates and stable performance are important. The decision tree achieved lower accuracy but higher infected-household recall at 0.756, meaning that it identified a larger share of infected households at the cost of more false positives. The random forest had a higher AUC than the decision tree but weaker positive-class recall at the default threshold, illustrating why threshold selection and operational objectives are crucial.

The same trade-off is visible at the level of error counts, so Table 3 reports true negatives, false positives, false negatives, and true positives for each classifier.

Table 3: Confusion-count analysis derived from the supplied classification reports. Rows indicate true class.

Model	TN	FP	FN	TP
Logistic regression	108	14	25	53
Decision tree	94	28	19	59
Random forest	107	15	35	43

Table 3 clarifies the error trade-offs. Logistic regression missed 25 infected households and falsely flagged 14 uninfected households. The decision tree reduced false negatives to 19 but doubled false positives to 28. The random forest preserved a low false-positive count of 15 but missed 35 infected households. For household malaria surveillance, these differences map directly onto policy choices. If interventions are inexpensive and the priority is to avoid missed infection risk, the decision tree or a lower logistic-regression threshold may be preferred. If intervention capacity is scarce and unnecessary visits must be minimized, logistic regression's precision and calibration are attractive.

Figure 1 shows the supplied decision tree. Although less accurate overall than logistic regression, the tree is valuable because it exposes conditional rules that can be reviewed by epidemiologists. Its structure helps identify whether the model is using plausible splits, such as bed-net behavior, household conditions, and environmental exposure, or whether it is relying on fragile artifacts.

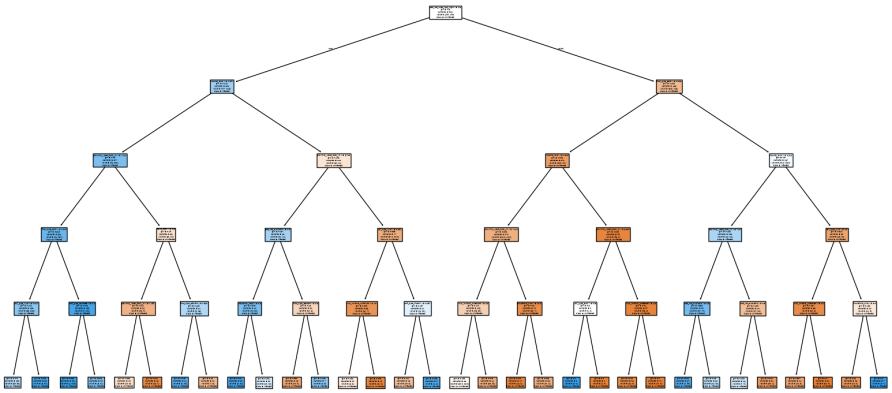

Figure 1: Visualized decision tree used as an interpretable classifier for household malaria infection risk. The tree provides rule-level evidence for expert review and operational communication.

The tree figure makes the model's rule structure auditable: reviewers can see whether splits correspond to plausible exposure and prevention variables before accepting its high-recall operating point. Figure 2 reports discrimination, precision–recall behavior, and calibration for the three evaluated classifiers.

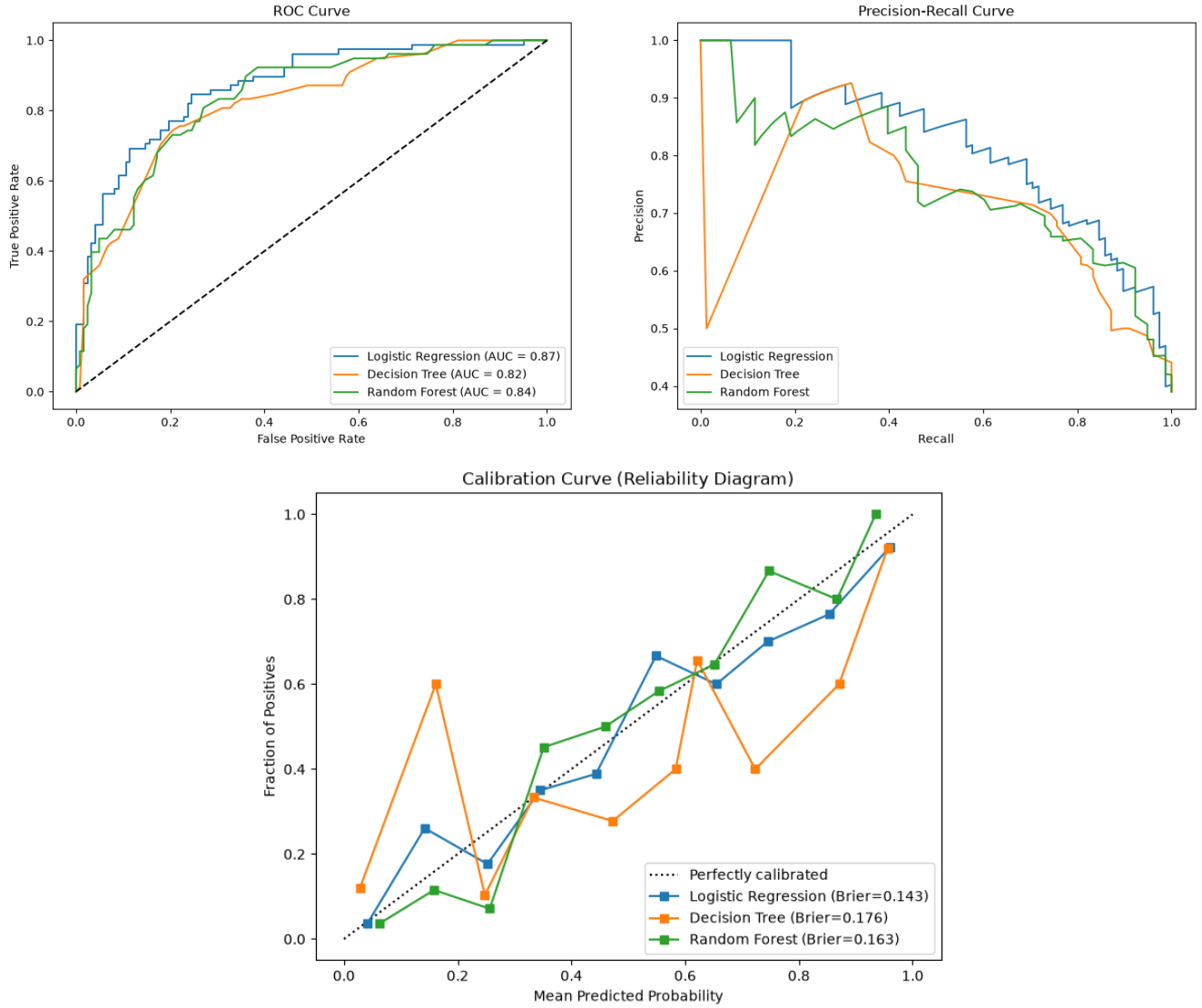

Figure 2: Ranking and probability-quality diagnostics: ROC curve, precision–recall curve, and calibration curve. Logistic regression shows the best AUC and Brier score, while precision–recall behavior highlights threshold-dependent trade-offs

Figure 2 confirms that logistic regression ranks infected and uninfected households best across thresholds, followed by random forest and decision tree. The precision–recall curve reveals that precision declines as recall approaches complete case-finding, emphasizing the operational cost of aggressive screening. The calibration plot shows that logistic regression is closest to reliable probability estimation overall, while the random forest and decision tree show greater deviations in several probability bins. Because the proposed human-review mechanism depends on probability tiers, calibration is not a secondary concern; it affects how often households are routed to low-, medium-, or high-risk workflows.

Figure 3 presents the supplied normalized confusion matrices. These plots make class-specific behavior visible without requiring readers to infer it from scalar metrics.

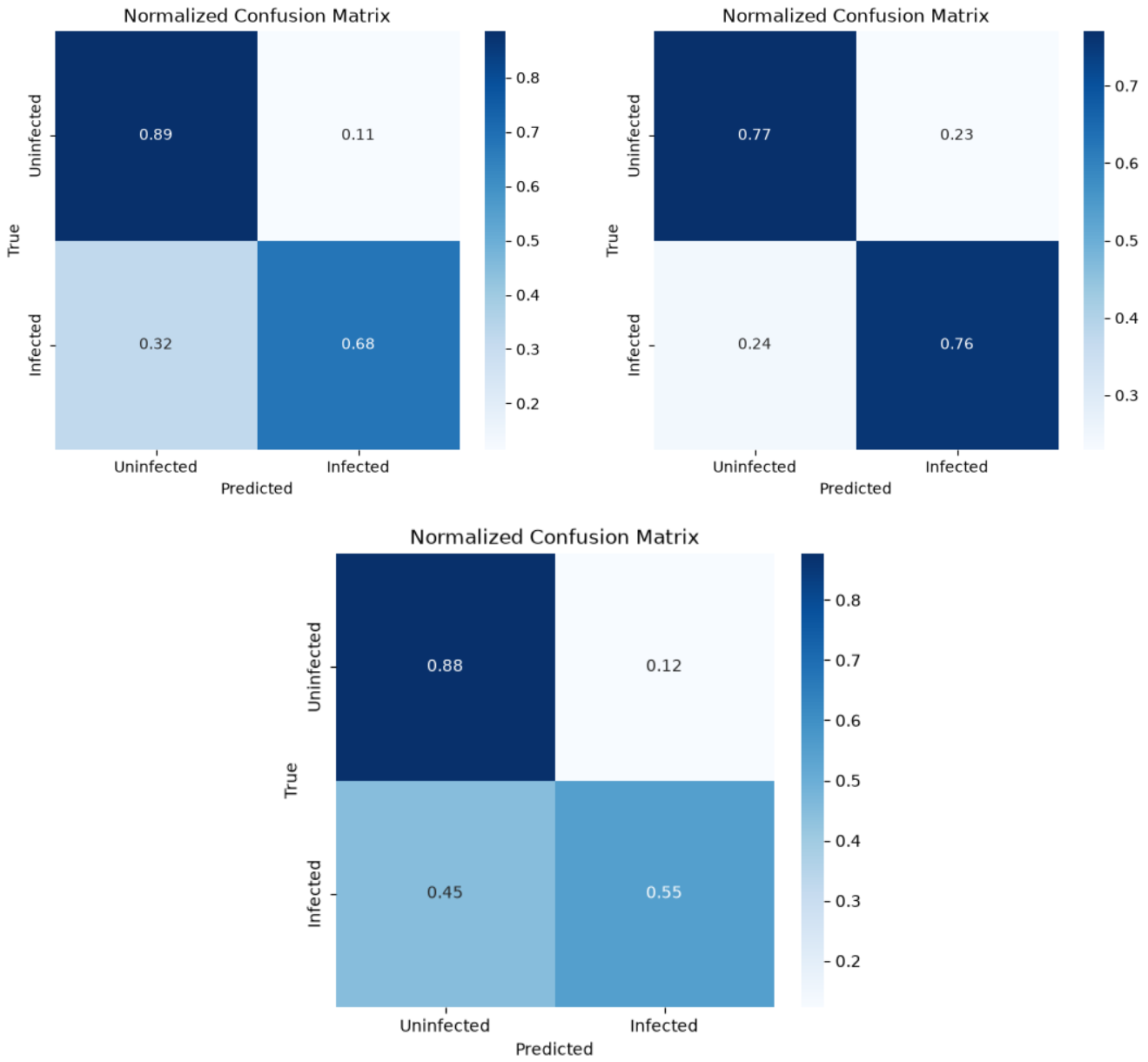

Figure 3: Normalized confusion matrices for logistic regression, decision tree, and random forest. The matrices show the central operating trade-off between infected-household recall and false-positive burden

Figure 3 shows that logistic regression correctly classifies 0.89 of uninfected households and 0.68 of infected households. The decision tree sacrifices uninfected specificity, correctly classifying 0.77 of uninfected households, but improves infected-household detection to 0.76. Random forest correctly classifies 0.88 of uninfected households but only 0.55 of infected households. These matrices translate scalar metrics into operational consequences, making the false-negative burden and false-positive workload visible before explanation is considered.

Explainability results identify plausible risk drivers. Figure 4 shows that the largest average random-forest SHAP contributions are associated with bed-net use on the previous night, income level, proximity to a water body, earth flooring, bed-net availability, cement flooring, screened windows, mud walls, household size, and thatch roofs.

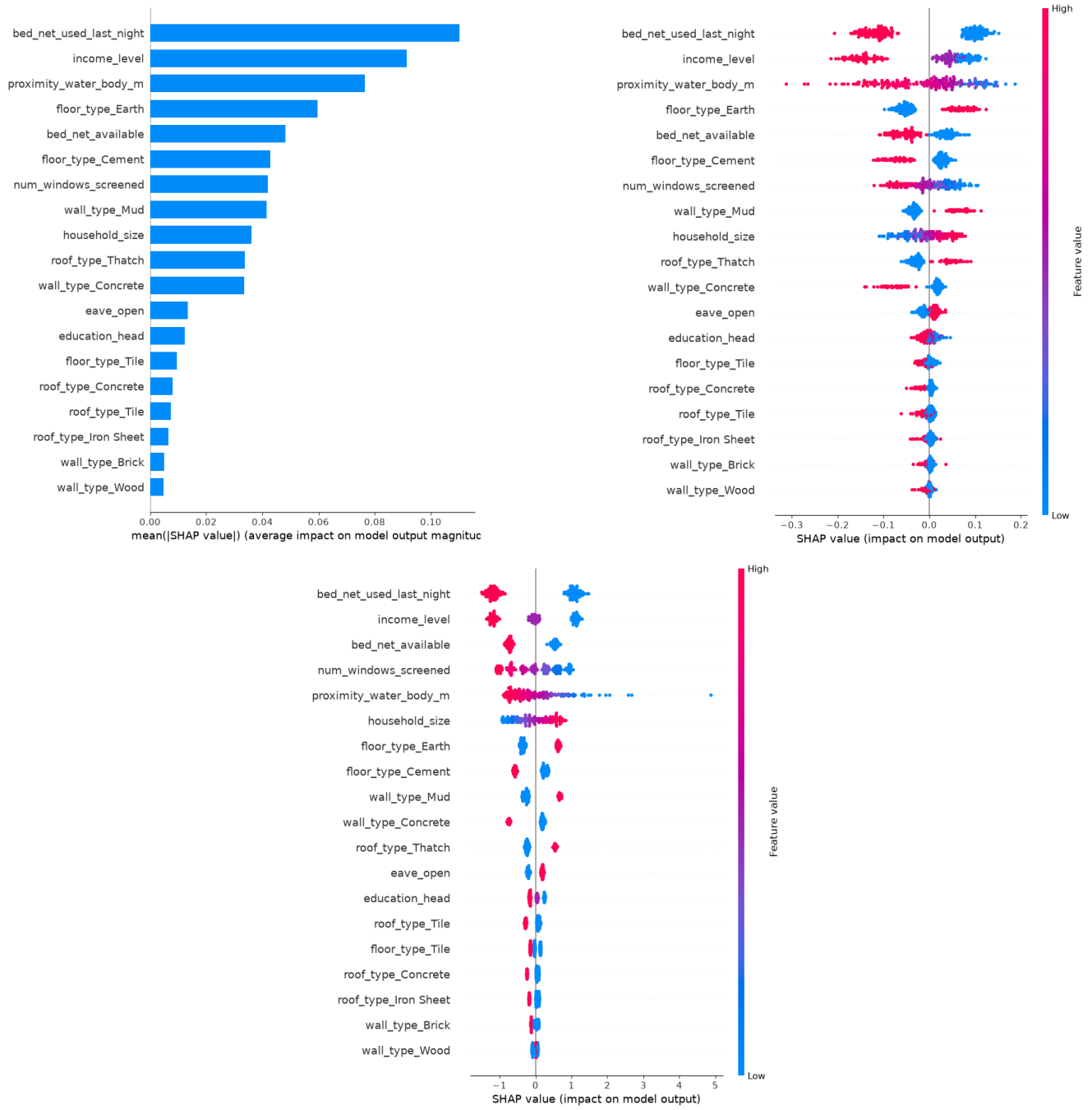

Figure 4: Global explainability outputs: random-forest mean absolute SHAP ranking, random-forest SHAP beeswarm, and logistic-regression SHAP beeswarm. The compact composite keeps the explanatory evidence adjacent while preserving readable feature labels

The global explanations connect model performance to substantive risk patterns: the strongest features are not abstract proxies alone, but variables that a reviewer can inspect and, in many cases, address through prevention or follow-up. The direction and dispersion of the beeswarm plot indicate that features do not merely matter in isolation; their values push risk upward or downward depending on household context. The logistic-regression SHAP summary provides an interpretable baseline for comparison.

Figure 5 gives local explanations for three high-risk households. These plots are especially relevant to the augmentation thesis because they convert an opaque score into a reviewable case narrative.

The local explanations complete the evidence chain by turning a ranked case into an auditable case narrative, which is exactly the form needed for human-in-the-loop review. A field officer can inspect whether a high-risk prediction is driven by non-use of a bed net, short distance to water, earth flooring, open eaves, large household size, or other variables. The reviewer can then decide whether the appropriate action is household education, net replacement, environmental inspection, diagnostic follow-up, or no immediate intervention because local information contradicts the model. Taken together, the results support a decision-support interpretation rather than an automation interpretation. Logistic regression is the strongest default scorer, the decision tree is useful when high recall is prioritized, the random

forest supplies nonlinear explanations but requires threshold adjustment for screening, and SHAP explanations expose the feature structure underlying risk estimates.

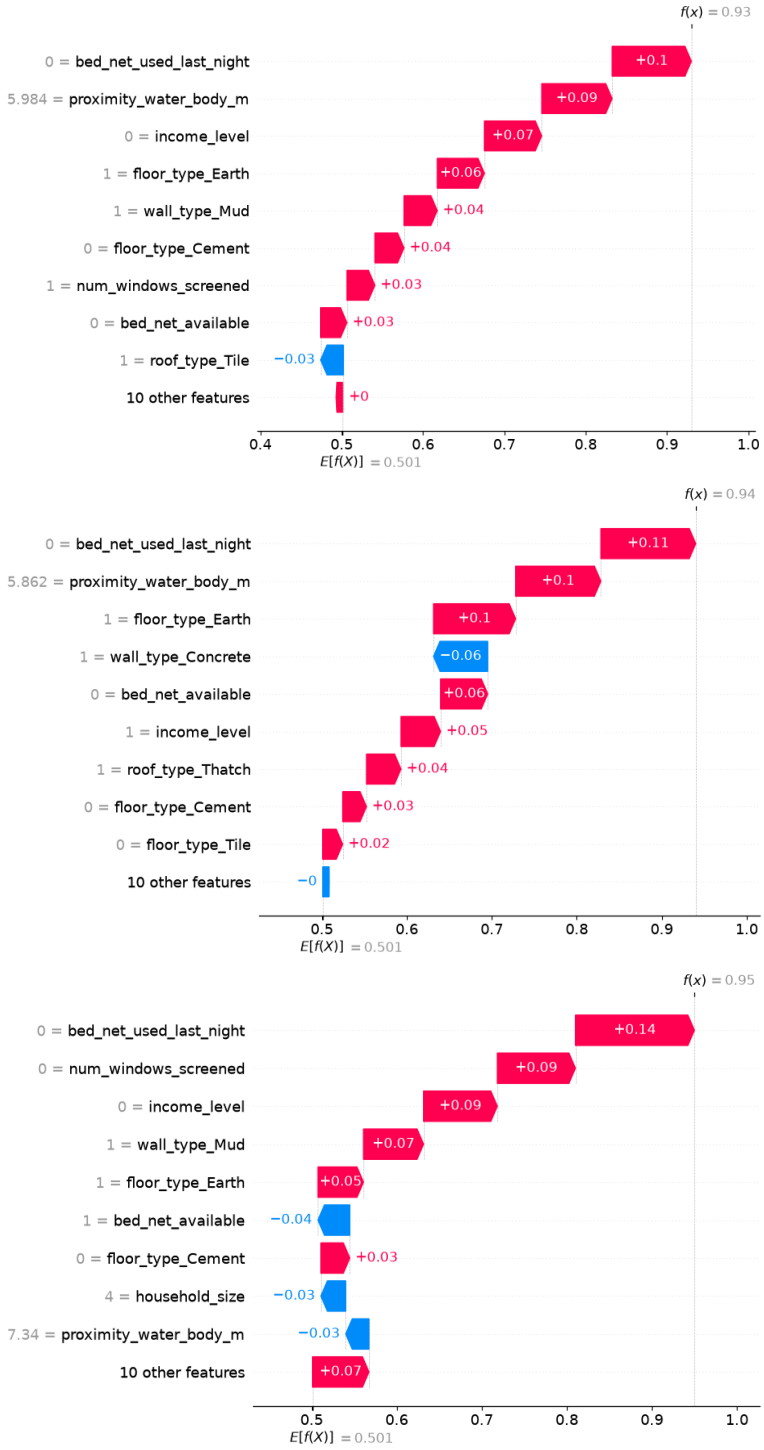

Figure 5: Local explanations for three high-risk household profiles. The stacked waterfall plots decompose predictions into feature contributions that reviewers can compare with field knowledge before action

5. Discussions

The empirical pattern just reported moves the paper from model comparison to deployment interpretation: the question is not which classifier should act alone, but how each output can strengthen accountable public health judgement. The findings show that AI can improve household malaria decision-making by organizing evidence, quantifying risk, and revealing patterns that are difficult to maintain in manual rules. The strongest model, logistic regression, achieved a useful balance of discrimination, calibration, and interpretability. This matters because public health teams need probabilities that can support prioritization, not just binary labels. At the same time, the decision tree's higher infected-class recall demonstrates that model choice depends on intervention policy. A program facing an outbreak may prefer a sensitive screening configuration; a program with limited visit capacity may prefer calibrated precision. AI therefore expands the decision

space, but human leaders must still choose the operating point.

Practical implications follow directly. A household malaria risk system could be deployed as a dashboard or scoring tool that ranks households for review, flags high-risk profiles, and identifies the features most responsible for each score. Such a tool could reduce administrative burden and support consistent triage across teams. However, implementation should begin with prospective local validation rather than direct operational substitution. The supplied dataset is synthetic and useful for pipeline development, but real-world deployment would require representative survey data, temporal validation, geospatial linkage, and evaluation across demographic and ecological subgroups. Reporting standards for prediction models encourage this transparency because model utility depends on data provenance, inclusion criteria, missingness, and validation design [31].

Policy implications center on equitable allocation. Variables such as income, education, and housing quality can identify vulnerability, but they also encode structural inequality. If a model uses these variables to direct supportive interventions, it may improve equity by finding underserved households. If it is used to restrict services, penalize households, or stigmatize communities, it can deepen inequity. The fairness literature warns that algorithmic systems must be evaluated as sociotechnical systems, not merely as mathematical predictors [28]. Evidence of bias in health algorithms further demonstrates that even high-performing systems can misallocate resources when proxy labels or design choices encode inequitable assumptions [27]. Therefore, the governance framework should require subgroup performance audits, fairness-sensitive review of proxy-driven predictions, and community accountability.

Human-AI collaboration is the central design principle. The model should make public health expertise more effective by surfacing risk signals quickly, but it should not override local knowledge. Reviewers may know that a household recently received nets, that nearby water is treated or seasonal, that a recorded income value is outdated, or that a case cluster requires immediate response despite a medium predicted score. Interface guidelines for human-AI interaction recommend communicating system capabilities, uncertainty, and recoverability when errors occur [20]. Automation-bias research reinforces the need for training and workflow design that encourage users to question predictions rather than accept them passively [21, 22].

Explainability strengthens trust when used appropriately. SHAP summaries showed that bed-net behavior, income, proximity to water, flooring, window screening, wall type, household size, and roof type were influential. These features are plausible, but plausibility is not proof of causality. Explanations should be used to inspect model behavior, detect unexpected patterns, support communication, and guide audit questions. They should not be used to claim that changing a single feature will necessarily change infection status. This distinction is particularly important for socioeconomic variables: the model may identify poverty-related risk, but ethical intervention requires addressing vulnerability rather than blaming households.

The study has limitations. The dataset is synthetic, so performance estimates should be interpreted as evidence for pipeline feasibility rather than field effectiveness. The hold-out test set contains 200 households, which is adequate for demonstration but insufficient for definitive subgroup fairness evaluation. The evaluated random forest used the default threshold, so its lower infected recall may be improved through threshold tuning or cost-sensitive optimization. The project artifacts do not include temporal validation, external validation, or intervention outcome data. Future work should evaluate the framework on real DHS/MIS-linked data where permissible, include spatial and seasonal variables, compare threshold policies under realistic resource constraints, measure subgroup calibration, and test how public health workers use explanations in simulated or prospective workflows.

Despite these limitations, the results support the augmentation thesis. AI provides speed, consistency, ranking, calibration, and explanation. Humans provide contextual reasoning, ethical judgment, governance, and accountability. The best system is therefore not an autonomous classifier but a structured partnership in which machine learning narrows attention, explains risk patterns, and prompts review while public health experts retain authority over action. The conclusion synthesizes this partnership into the paper's final claim about AI as an augmentation layer.

6. Conclusions

This paper developed a complete human-centered AI/ML research framework for household malaria risk profiling using the supplied project outputs as empirical evidence. The methodology combined survey-like household predictors, mode and median imputation, logarithmic transformation of environmental proximity, ordinal and one-hot encoding, SMOTE-based class balancing, stratified hold-out validation, comparative model evaluation, calibration assessment, and SHAP-based global and local explanations. The results show that logistic regression provided the strongest overall performance, with accuracy of 0.805, ROC AUC of 0.87, and Brier score of 0.143, while the decision tree offered higher infected-household recall and the random forest supplied nonlinear feature-attribution evidence. The explainability analysis identified bed-net use, income, water proximity, flooring, bed-net availability, window screening, wall material, household size, and roof type as major contributors to predicted infection risk. The study's three contributions are a reproducible machine learning workflow for household risk classification, a human-in-the-loop governance mechanism based on calibrated risk tiers and review triggers, and an ethical risk-management framework for fairness, documentation, monitoring, and accountability. More broadly, the paper shows that predictive models are most useful when they make public health decisions more transparent and better prioritized without pretending to settle the social, epidemiological, and ethical questions that surround intervention. Because the dataset is synthetic and externally unvalidated, deployment would require local data, subgroup audits, calibration monitoring, and prospective evaluation with public health teams. Nevertheless, the evidence demonstrates that well-governed AI can help decision-makers detect household risk patterns, allocate attention efficiently, and communicate the basis for intervention choices. Artificial Intelligence can fundamentally improve the efficiency, intelligence, and effectiveness of decision-making within household malaria risk profiling, but only as an augmentation layer rather than a replacement for human expertise, governance, accountability, and ethical oversight.

Article Information

Funding: No external funding was received for this study.

Author Contributions: All authors contributed to the conception, methodology, analysis, and preparation of the manuscript and approved the final version.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability Statement: Data available on reasonable request.

Disclaimer (Artificial Intelligence): The author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.), and text-to-image generators have been used during writing or editing of manuscripts.

References

[1] Simon I. Hay, Carlos A. Guerra, Peter W. Gething, Anand P. Patil, Andrew J. Tatem, Abdisalan M. Noor, Catherine W. Kabaria, Benson H. Manh, Iqbal R. F. Elyazar, Simon Brooker, David L. Smith, Rana A. Moyeed, and Robert W. Snow. A World Malaria Map: Plasmodium falciparum Endemicity in 2007. *PLoS Medicine*, 6(3):e1000048, 2009. URL https://doi.org/10.1371/journal.pmed.1000048.

[2] Peter W. Gething, Anand P. Patil, David L. Smith, Carlos A. Guerra, Iqbal R. F. Elyazar, Geoffrey L. Johnston, Andrew J. Tatem, and Simon I. Hay. A New World Malaria Map: Plasmodium falciparum Endemicity in 2010. *Malaria Journal*, 10:378, 2011. URL https://doi.org/10.1186/1475-2875-10-378.

[3] Samir Bhatt, Daniel J. Weiss, Ewan Cameron, Daniel Bisanz, Bonnie Mappin, Ursula Dalrymple, Katherine E. Battle, Catherine L. Moyes, Amelia Henry, Philip A. Eckhoff, Edward A. Wenger, Olivier Bri"et, Melissa A. Penny, Thomas A. Smith, Adam Bennett, Joshua Yukich, Thomas P. Eisele, Jamie T. Griffin, Catherine A. Fergus, Molly Lynch, Fiona Lindgren, Justin M. Cohen, Christopher L. J. Murray, David L. Smith, Simon I. Hay, Richard E. Cibulskis, and Peter W. Gething. The Effect of Malaria Control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526(7572):207–211, 2015. URL https://doi.org/10.1038/nature15535.

[4] Daniel J. Weiss, Tim C. D. Lucas, Michele Nguyen, Anita K. Nandi, Donal Bisanzio, Katherine E. Battle, Ewan Cameron, Katherine A. Twohig, Daniel A. Pfeffer, J. A. Rozier, Harry S. Gibson, Pratik C. Rao, Daniel Casey, Amelia Bertozzi-Villa, Emma L. Collins, Ursula Dalrymple, Nicole Gray, Joseph R. Harris, Rosalind E. Howes, Soo T. Kang, Susan H. Keddie, Daniel May, Susan F. Rumisha, Michelle P. Thorn, Ryan M. Barber, Nancy Fullman, Chantal K. Huynh, Xie Rachel Kulikoff, Michael J. Kutz, Alan D. Lopez, Ali H. Mokdad, Mohsen Naghavi, Grant Nguyen, Katya A. Shackelford, Theo Vos, Haidong Wang, David L. Smith, Stephen S. Lim, Christopher J. L. Murray, Samir Bhatt, Simon I. Hay, and Peter W. Gething. Mapping the Global Prevalence, Incidence, and Mortality of Plasmodium falciparum, 2000–17: A Spatial and Temporal Modelling Study. *The Lancet*, 394(10195):322–331, 2019. URL https://doi.org/10.1016/S0140-6736(19)31097-9.

[5] Lucy S. Tusting, Matthew M. Ippolito, Barbara A. Willey, Immo Kleinschmidt, Grant Dorsey, Roly D. Gosling, and Steve W. Lindsay. The Evidence for Improving Housing to Reduce Malaria: A Systematic Review and Meta-Analysis. *Malaria Journal*, 14:209, 2015. URL https://doi.org/10.1186/s12936-015-0724-1.

[6] Lucy S. Tusting, Christian Bottomley, Harry Gibson, Immo Kleinschmidt, Andrew J. Tatem, Steve W. Lindsay, and Peter W. Gething. Housing Improvements and Malaria Risk in Sub-Saharan Africa: A Multi-Country Analysis of Survey Data. *PLoS Medicine*, 14(2): e1002234, 2017. URL https://doi.org/10.1371/journal.pmed.1002234.

[7] Teun Bousema, Jamie T. Griffin, Robert W. Sauerwein, David L. Smith, Thomas S. Churcher, Willem Takken, Azra Ghani, Chris Drakeley, and Roly Gosling. Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLoS Medicine*, 9(1): e1001165, 2012. URL https://doi.org/10.1371/journal.pmed.1001165.

[8] Daniel J. Corsi, Melissa Neuman, Jocelyn E. Finlay, and S. V. Subramanian. Demographic and Health Surveys: A Profile. *International Journal of Epidemiology*, 41(6):1602–1613, 2012. URL https://doi.org/10.1093/ije/dys184.

[9] Eric J. Topol. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, 25(1):44–56, 2019. URL https://doi.org/10.1038/s41591-018-0300-7.

[10] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Medicine*, 17:195, 2019. URL https://doi.org/10.1186/s12916-019-1426-2.

[11] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in Health and Medicine. *Nature Medicine*, 28(1):31–38, 2022. URL https://doi.org/10.1038/s41591-021-01614-0.

[12] Ben Shneiderman. Human-Centered Artificial Intelligence: Reliable, Safe and Trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. URL https://doi.org/10.1080/10447318.2020.1741118.

[13] David R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B*, 20(2):215–242, 1958.

[14] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

[15] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. URL https://doi.org/10.1214/aos/1013203451.

[16] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. URL https://doi.org/10.1613/jair.953.

[17] Tom Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. URL https://doi.org/10.1016/j.patrec.2005.10.010.

[18] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3):e0118432, 2015. URL https://doi.org/10.1371/journal.pone.0118432.

[19] Alexandru Niculescu-Mizil and Rich Caruana. Predicting Good Probabilities with Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005. URL https://doi.org/10.1145/1102351.1102430.

[20] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. URL https://doi.org/10.1145/3290605.3300233.

[21] Raja Parasuraman and Victor Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2):230–253, 1997. URL https://doi.org/10.1518/001872097778543886.

[22] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012. URL https://doi.org/10.1136/amiajnl-2011-000089.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. URL https://doi.org/10.1145/2939672.2939778.

[24] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

[25] Zachary C. Lipton. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10):36–43, 2018. URL https://doi.org/10.1145/3233231.

[26] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. URL https://doi.org/10.1038/s42256-019-0048-x.

[27] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464):447–453, 2019. URL https://doi.org/10.1126/science.aax2342.

[28] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019. URL https://doi.org/10.1145/3287560.3287598.

[29] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. URL https://doi.org/10.1145/3287560.3287596.

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92, 2021. URL https://doi.org/10.1145/3458723.

[31] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, 13:1, 2015. URL https://doi.org/10.1186/s12916-014-0241-z.