

MalariaXAI: A Synthetic Proof-of-Concept Multimodal Explainable AI Framework for Malaria Risk Prediction Using Clinical, Climatic, and Genomic Data

Nalenhle Ndlovu ^a, Belinda Ndlovu ^{b,*}

^{a,b} Informatics and Analytics Department, National University of Science and Technology, Bulawayo, Zimbabwe

Background and Purpose: Malaria is still the most prevalent public health concern worldwide and as of 2023, an estimated 263 million people were infected and 597 thousand died. Sub-Saharan Africa accounts for an estimated 94% of cases and 95% of deaths. Current AI systems are plagued by monolithic, single-modal architectures, black-box reasoning, and the ongoing absence of a single platform that correlates climatic, genomic, and clinical features with explanations. This work is predicated on the assumption that, in a resource-limited healthcare context, accuracy alone is not sufficient to ensure an AI tool's adoption; practical use also demands interpretability, calibration, computational efficiency, and an explicit representation of the model's logic. Here, we present MalariaXAI, a multimodal, explainable AI proof-of-concept, synthetically and experimentally developed, to illustrate the potential to correlate and explain successfully in a single multimodal platform.

Methods: We adopted a controlled comparative experimental design following the CRISP-DM framework. We constructed a synthetic dataset of 14 features spanning clinical (n=6), climatic (n=5), and genomic (n=3) modalities using probabilistic generative methods informed by published epidemiological distributions. Outcome labels were generated probabilistically to mitigate target leakage. Three modality-specific models include an XGBoost ClinicalModel, a Random Forest ClimateModel, and a Random Forest GenomicModel, which were trained on stratified 80/20 splits with 5-fold cross-validation and combined via a weighted FusionModel (clinical: 0.45, climate: 0.35, genomic: 0.20). Class imbalance was addressed using SMOTE. We implemented explainability through SHAP, LIME, and a DoWhy-based structural causal model. Performance was evaluated using AUC-ROC, Accuracy, F1-Score, Recall, and Brier Score.

Results: Across the four models, AUC-ROC ranged from 0.82 to 0.971, accuracy from 80.5% to 96.1%, F1-Score from 0.803 to 0.960, recall from 0.812 to 0.965, and Brier Score from 0.047 to 0.112. The FusionModel achieved the strongest performance across every evaluation metric, including AUC-ROC (0.971), recall (0.965), F1-score (0.960), accuracy (96.1%), and Brier Score (0.047), outperforming each single-modality baseline. Consequently, the value of multimodal integration is better understood through balanced decision-support profiling rather than discrimination maximisation. In the absence of other components, Ablation analysis showed the independent effects of multimodal incorporation, SMOTE balancing, and genomic data in the synthetic scenario. For SHAP, temperature at the fever stage, monthly rainfall, and mutation at Kelch-13 appeared to be the top predictive features across modalities, and their directions were all consistent with prior epidemiological domain knowledge.

Conclusions: This work illustrates the methodological viability of a combined synthetic malaria prediction pipeline using multimodal explainable AI methods. It suggests that multimodal ensembles, combined with structured explainability, can enable computationally efficient, explainable malaria risk modelling in a controlled synthetic setting, a design that may serve as a replicable architectural blueprint for future real-world deployment in resource-constrained health systems. However, it must be emphasised that all results herein are limited to the synthetic setting and cannot be taken as evidence of clinical predictive utility or readiness for deployment. External validation across different cohorts, fairness audits, and prospective clinical validation are necessary steps before clinical adoption.

Keywords: Malaria prediction; multimodal machine learning; explainable artificial intelligence; SHAP; LIME; Exploratory Causal Simulation; XGBoost; Random Forest; ensemble learning; synthetic data; proof-of-concept; sub-Saharan Africa.

*Corresponding author address: mwogosia@gmail.com

1 Introduction

Malaria is a vector-borne disease transmitted by infected female *Anopheles* mosquitoes, with symptoms that can include fever, chills, headache, severe anaemia, coma, and death [1]. By 2023, the World Health Organization estimated 263 million malaria cases and 597,000 malaria deaths globally, among which 94% of cases and 95% of deaths occurred in the WHO African region; children under 5 years of age account for 76% of malaria deaths; pregnant women are more likely to have poor outcomes [2], [3]. The magnitude of this issue in Zimbabwe can be seen in the epidemic figure, which indicates that there were 111,998 cases and 310 deaths in week 23 of 2025, according to the Africa CDC, a 4-fold increase from the 29,031 cases and 49 deaths recorded in week 23 of 2024 [4].

Despite advances in artemisinin-based combination therapies, insecticide-treated nets, indoor residual spraying, and vaccine programmes such as RTS,S, malaria elimination remains elusive [2]. Ongoing climate change also modifies transmission patterns, with increased temperatures and altered precipitation extending vector habitats to previously unaffected areas with low incidence [5]. These changes also drive the rise of drug-resistant parasites through evolutionary shifts in parasite genomes, including Kelch-13 mutations conferring artemisinin resistance and PfCRT haplotypes responsible for chloroquine resistance [6]. These interacting pressures demand predictive systems capable of integrating climatic, genomic, and clinical signals within a single interpretable architecture.

Tree-based ensemble methods such as Random Forest and XGBoost are uniquely appropriate for analysing the kind of inherently heterogeneous, structured tabular data present in the combined-clinical, genomic, and environmental predictor space of the multimodal malaria dataset, due to their capability for native, direct inclusion of mixed feature types (e.g., continuous environmental factors, binary and categorical clinical and genetic features), resilience to outliers and collinearity, and inbuilt feature selection capabilities enabling a basic interpretability necessary for use in computational resource-constrained environments [7], [8].

Most existing malaria AI literature, by contrast, focuses on individual predictive tasks and single-modality data (clinical data, climatic data, microscopic images, etc) [9], [10], [11]. Although these models achieved encouraging predictive performance, little attention has been given to integrating diverse malaria-related data modalities into a single, explainable framework that supports practical, interpretable decision-making. In addition, previous efforts mainly focused on predictive performance and did not adequately address operational aspects such as computational feasibility, interpretability, calibration, governance, and user-friendliness when implemented in resource-poor health facilities. [12]. This cannot be overlooked when deploying such systems in a malaria-endemic environment, as they must be effective and feasible while accounting for resource limitations, limited expertise, and real-world operational use within a health service. It remains unclear how multimodality explainable AI models could achieve the necessary predictive ability while remaining interpretable and deployable [13].

This study is based on the understanding that practical utility in healthcare settings with constrained infrastructure depends not only on model accuracy but also on interpretability, computational efficiency, and the ability to communicate transparent model reasoning. The proposed framework investigates whether multimodal ensemble architectures, combined with structured explainability, can provide a balanced trade-off among these competing requirements, advancing the field beyond the fragmented, single-modality approaches that currently dominate the literature. MalariaXAI is developed and empirically evaluated in a synthetic data environment to demonstrate the methodological feasibility of this unified framework. The study is guided by the following research questions:

1. Can multimodal integration of clinical, climatic, and genomic feature spaces improve malaria risk prediction relative to unimodal baselines under controlled synthetic conditions?
2. To what extent can SHAP, LIME, and Exploratory Causal Simulation jointly support cross-modal explainability within a multimodal malaria prediction framework?
3. Does multimodal ensemble integration remain computationally feasible for prototype deployment under local execution conditions with limited computational resources?

This study makes four contributions. First, it presents a unified multimodal malaria prediction framework integrating clinical, climatic, and genomic feature spaces within a single explainable ensemble architecture. Second, it combines SHAP, LIME, and DoWhy-based intervention simulation within one coordinated

explainability pipeline. Third, it evaluates multimodal architectural behaviour through systematic ablation analysis under controlled synthetic conditions. Finally, it demonstrates the computational feasibility of explainable multimodal malaria prediction under lightweight prototype deployment constraints.

The novelty of MalariaXAI lies primarily in integration rather than in inventing individual methodological components. XGBoost, Random Forest, SHAP, LIME, DoWhy, SMOTE, weighted ensembles, and CRISP-DM are each individually well established in the literature. MalariaXAI contributes the first identified synthesis of these components into a single, coordinated, multimodal architecture explicitly designed for malaria risk prediction under the deployment constraints typical of sub-Saharan African health systems. The architectural integration itself, spanning heterogeneous clinical, climatic, and genomic feature spaces within a single explainable, calibrated, offline-capable prototype, constitutes the primary scientific contribution. Furthermore, the underlying architecture is not malaria-specific: the same multimodal explainable AI pipeline could apply to other infectious diseases where environmental, biological, and clinical signals interact, such as dengue, tuberculosis, cholera, or emerging infectious disease surveillance, and this generalisability is discussed in Section 4.

The remainder of this paper is structured as follows. Section 2 describes the methodology. Section 3 presents results. Section 4 discusses implications, threats to validity, and limitations. Section 5 concludes.

2 Materials and Methods

Using the CRISP-DM framework, we implemented a controlled-comparative experimental design to assess the methodological feasibility of the multimodal, explainable malaria prediction architecture in a simulated environment. The design compared unimodal, single-modality models against the fused multimodal architecture by performing identical preprocessing, sampling, and evaluation in all comparisons. Because no compiled public dataset combining climate, genomics, and clinical malaria information existed, we generated synthetic data to create a controlled environment for testing architecture-level, not clinical-prediction, performance. Formal ethics approval was not required, as the study used only synthetic data and involved no human participants. Responsible AI principles including transparency, interpretability, fairness screening, and explicit bounding of clinical claims were incorporated throughout the study design.

2.1 Business Understanding

The Business Understanding section described the context in which the system operates; the goal of the operational process, the roles of those involved; and the constraints on system deployment. In the MalariaXAI context, the key components established were the research questions that frame and evaluate model/system performance, the network of stakeholders with whom the system is expected to operate, and the constraints that guide the methodological design. The overarching operational goal is to demonstrate a unified, multimodal, explainable malaria prediction system that is demonstrably feasible in terms of predictive accuracy, interpretability, calibration, and computational tractability at the prototype level, all within a simulated environment.

Stakeholders who were accounted for were: clinician health workers who need low-latency, explainable risk outcomes; public health officials who need causal insights for policy and intervention decisions; and technology administrators who need a computationally feasible, maintainable prototype to deploy. Operational constraints considered were: the unavailability of real-world, integrated multimodal data sources; the computational capacity and available resources within the health system in sub-Saharan Africa; the need to function offline; and the need to correctly constrain clinical claims to the synthetic nature of the environment. These operational constraints guided feature selection, model architecture design, explainability approach and so on.

2.2 Data Source and Synthetic Dataset Generation

The synthetic dataset was generated not as an accurate representation of the clinical epidemiology of malaria, but as a controlled testbed of methodological exploration. The distribution of features and their dependence structure were informed by the existing epidemiological literature, WHO survey data, and domain knowledge and assumptions to construct a clinically relevant system of interactions among climatic,

genomic, and clinical variables. As such, all stated model behaviour is within-dataset evaluation of architectural feasibility and the ability to coordinate multimodal signals.

2.2.1 Synthetic Data Design.

The synthetic dataset was constructed exclusively for architecture-level experimentation. Feature distributions for all 14 variables were parameterised from published epidemiological literature, WHO surveillance bulletins, and clinical domain knowledge. Conditional dependencies among comorbidities, environmental variables, and genomic markers were encoded through empirically-informed correlation matrices to approximate clinically plausible interaction patterns. Malaria risk outcome probabilities were generated through weighted probabilistic interactions among clinical and environmental variables, rather than via deterministic rule assignment, ensuring that no single variable or combination of variables could deterministically guarantee the target outcome. Stochastic noise was injected across all generated profiles to reduce artificially deterministic patterns and improve distributional variability. Table 1 presents the full feature set across all three modalities.

Table 1. Synthetic Data Design

Item	Value
Records generated	10,000
Clinical features	6
Climatic features	5
Genomic features	3
Outcome generation	Weighted probabilistic sigmoid
Target prevalence	30–35%
Noise injection	Gaussian noise for continuous variables
Leakage control	Pre-prediction variables only; scaler fitted on training folds

Table 2 presents the complete feature set across all three data modalities, including the feature name, data type, and epidemiological source basis used to parameterise each variable during synthetic data generation.

Table 2. Feature Set Across Three Data Modalities

Modality	Feature	Type	Source Basis
Clinical (n=6)	Fever temperature (fever Celsius)	Continuous (°C)	WHO clinical thresholds
	RDT result	Binary (pos/neg)	WHO diagnostic guidance
	Platelet count	Continuous ($\times 10^3/\mu\text{L}$)	Clinical haematology literature
	Haemoglobin level	Continuous (g/dL)	Clinical haematology literature
	Age	Continuous (years)	WHO surveillance bulletins
	WBC count	Continuous ($\times 10^3/\mu\text{L}$)	Clinical haematology literature
Climatic (n=5)	Monthly rainfall	Continuous (mm)	WHO/Africa CDC bulletins
	Relative humidity	Continuous (%)	Climatological literature
	NDVI (cyclical)	Continuous (0–1)	Remote-sensing literature
	Ambient temperature	Continuous (°C)	WHO/NOAA climate data
Genomic (n=3)	NDVI: Month (cyclical)	Cyclical encoding (sin/cos of month)	Remote-sensing literature
	Kelch-13 mutation status	Binary (present/absent)	Genomic resistance literature
	PfCRT haplotype	Categorical	Genomic resistance literature
	SNP diversity index	Continuous (0–1)	MalariaGEN literature

2.2.2 Synthetic Data Rationale Checks

After generation, we ran several checks to confirm the rationale and identify sources of leakage or artificial inflation in the synthetic dataset. Kolmogorov-Smirnov tests were performed against the epidemiologic reference distribution when available for continuous variables, and chi-square goodness-of-fit tests were performed for categorical and binary variables. Class balance was checked to ensure that outcome prevalence reflected realistic malaria epidemiology and was not drastically skewed or unnaturally balanced. Finally, a pairwise Pearson correlation analysis was performed for feature-outcome correlations; an unusually high correlation coefficient would indicate potential unexpected encoding.

Prediction Target Definition. The binary prediction target is defined as the presence (1) or absence (0) of a malaria risk event for a given synthetic patient-environment record. Outcome labels $y \in \{0,1\}$ were generated probabilistically using a weighted sigmoid function of the form: $P(y=1) = \sigma(\beta_0 + \sum_i \beta_i x_i + \epsilon)$, where x_i are the 14 standardised feature values, β_i are empirically-informed feature weights derived from published malaria epidemiology, $\sigma(\cdot)$ is the logistic sigmoid function, and $\epsilon \sim N(0, \sigma^2)$ is injected Gaussian noise. No single feature or deterministic threshold guarantees the outcome. The 30–35% prevalence target reflects reported endemic malaria rates in high-burden sub-Saharan African settings per WHO surveillance bulletins.

All predictor variables were verified to contain only information available before the defined prediction event. The StandardScaler was fitted exclusively to the training-fold data within each cross-validation iteration to prevent leakage between the training and test phases. No records were duplicated across training and test subsets. These checks evaluated distributional plausibility within the synthetic framework and do not establish statistical equivalence with real-world malaria cohorts.

To strengthen credibility of the synthetic generation framework, distributional outputs were compared against summary statistics from published malaria datasets. Fever temperature distributions (mean $\approx 38.2^\circ\text{C}$, SD $\approx 1.1^\circ\text{C}$) were consistent with WHO-reported febrile malaria presentation thresholds. Platelet count distributions (median $\approx 120 \times 10^3/\mu\text{L}$) reflected thrombocytopenia ranges documented in clinical haematology literature for severe malaria. Monthly rainfall distributions matched published Zimbabwe climatic records from the Africa CDC bulletin period. Intra-modality Pearson correlations (Rainfall–Humidity $r=0.74$; Fever–RDT $r=0.61$) were consistent with epidemiological associations reported in published surveillance literature. While we do not claim formal statistical equivalence with a specific external dataset, these comparisons indicate that the generative framework encodes epidemiologically plausible feature relationships. All results remain bounded to internal synthetic evaluation.

2.3 Data Preparation and Feature Correlation

All numerical features were standardised to zero mean and unit variance using scikit-learn's StandardScaler, fitted exclusively on training data within each cross-validation fold. Categorical and binary features were label-encoded, with encoders retained for inverse transformation during SHAP explanation generation. The dataset was partitioned using a stratified 80/20 random split, preserving class distribution across subsets. Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique), applied only within training folds during cross-validation to prevent synthetic minority-class records from inflating test-set performance estimates.

A pairwise Pearson correlation analysis was conducted across all 14 features to characterise inter-feature dependencies and assess multicollinearity before model training. In Table 3, we presented the most significant correlations among features. The highest intra-modality correlations occurred between Monthly Rainfall and Humidity ($r=0.74$) and Fever Temperature and RDT result ($r=0.61$), reflecting logical and well-established epidemiological and clinical associations. We found low cross-modal correlations ($r < 0.25$), suggesting that the 3 modalities contain distinct information and supporting our use of a multimodal fusion approach. All correlation magnitudes appear reasonable, given that we are using a probabilistically generated synthetic data set that is unlikely to resemble actual clinical data.

Table 3. Pairwise Feature Correlation Summary, Selected Feature Pairs

Feature Pair	Pearson r	Interpretation	Modality Link
Fever Temperature - RDT Result	0.61	Strong positive: fever is a primary RDT trigger	Clinical
Monthly Rainfall - Humidity	0.74	Strong positive: rainfall drives ambient humidity	Climatic
Monthly Rainfall - NDVI	0.58	Moderate positive: vegetation responds to rainfall	Climatic
Kelch-13 - SNP Diversity	0.43	Moderate: diversity elevates resistance probability	Genomic
Platelet Count - Haemoglobin	0.39	Moderate positive: both suppressed in severe malaria	Clinical
Fever Temperature - Monthly Rainfall	0.21	Weak cross-modal: indirect transmission link	Cross-modal
Kelch-13 - RDT Result	0.18	Weak cross-modal: resistance encoded separately	Cross-modal

2.4 Model Architecture

Four predictive models were developed and evaluated under identical experimental conditions. The ClinicalModel employed XGBoost [14], selected for its consistently superior performance on structured tabular clinical data in prior malaria AI literature [12], [15]. The ClimateModel and GenomicModel each employed Random Forest [16], selected for robustness to correlated features and for implicit feature selection through ensemble averaging. They demonstrated performance in climate-informed malaria forecasting [7] and in genomic classification tasks [17]. Random Forest’s capacity to handle mixed feature types natively without explicit scaling was particularly relevant for the genomic modality, where binary mutation indicators and continuous diversity indices coexist within the same feature vector.

To improve methodological transparency, the weight optimisation procedure is described explicitly. Candidate weight sets spanning {clinical: 0.3–0.6, climate: 0.2–0.4, genomic: 0.1–0.25} in 0.05 increments were evaluated by maximising Recall on the 5-fold stratified training set. The optimal weights (Clinical: 0.45, Climate: 0.35, Genomic: 0.20) emerged consistently across all five folds with a maximum weight-set AUC variance of ± 0.006 . A sensitivity analysis confirmed that perturbations of ± 0.05 to any single weight produced AUC-ROC changes of ≤ 0.008 , indicating a stable rather than arbitrarily selected configuration. These weights are specific to the synthetic generative environment and must be re-optimised when applied to real-world data in Phase 2.

The FusionModel averaged the calibrated probabilities of the three modality-specific models through the weighted averaging ensemble technique. Weights (clinical: 0.45, climate: 0.35, genomic: 0.20) were determined empirically by comparing model predictive performance via cross-validation on the training set. These weights indicate the predictive power each modality offers for the synthetic dataset structure. Weighted averaging was chosen over stacking or a learned meta-learner approach to preserve the full interpretability of the fusion procedure without an additional opaque combining layer.

2.5 Model Training

Hyperparameter optimisation was performed using randomised grid search with 5-fold stratified cross-validation across all modality-specific models. The primary optimisation criterion was Recall, prioritised over AUC or accuracy, consistent with the clinical imperative of minimising false-negative malaria diagnoses in which a missed positive classification carries substantially higher clinical cost than an unnecessary further investigation. All experiments were conducted using a fixed random seed (seed = 42) to ensure full reproducibility.

Optimised hyperparameters for the ClinicalModel (XGBoost): $n_estimators = 200$, $learning_rate = 0.05$, $max_depth = 6$, $subsample = 0.8$, $colsample_bytree = 0.8$, $scale_pos_weight$ adjusted to the training set positive-to-negative class ratio. For the ClimateModel and GenomicModel (Random Forest): $n_estimators = 200$, $max_depth = 10$, $min_samples_split = 5$, $class_weight = 'balanced'$. AUC-ROC stability was

verified across five random seeds, yielding a maximum variance of ± 0.008 , confirming that reported results do not reflect a single fortuitous random initialisation.

2.6 Model Evaluation

All models were tested on the held-out 20% test set using a uniform suite of metrics: AUC-ROC (overall discriminatory ability across all classification thresholds); Accuracy (percentage of correct classification predictions); Recall/Sensitivity (key metric - percentage of positive malaria cases identified); Precision (percentage of positive prediction values that are actually correct); F1-Score (harmonic mean of Precision and Recall); Brier Score (average of squared difference between predicted and actual probabilities of positive events (0.00 indicates perfect calibration and 0.25 indicates that the classifier has the same accuracy as an uninformative classifier)). We defined thresholds of AUC > 0.85 and Accuracy $> 85\%$ as markers of feasibility; the Phase 2 objective is to report bootstrapped 95% confidence intervals from 1,000 resamples on real data; in the prototype on synthetic data, the 5-seed AUC variability of 0.008 indicates approximate classifier stability.

2.7 Explainability Implementation

SHAP (SHapley Additive ExPlanations, v0.42) was implemented with two explainer variations. TreeSHAP (TreeExplainer) was applied to ClinicalModel and the modality-specific Random Forests, calculating exact SHAP values by leveraging the models' tree structures for computationally efficient computation. KernelExplainer with a background dataset of 50 representative records, derived via k-means clustering on the training data, was used in FusionModel to reduce explanation computation to < 2 seconds from ~ 45 seconds without altering the distributional fidelity of the background reference population. SHAP values are expressed on a log-odds scale, indicate relative contributions of features in each direction toward a particular prediction, are not intended to represent linear probability change, nor are they evidence of feature-outcome causality.

LIME (Local Interpretable Model-Agnostic Explanations) was implemented using the lime Python package and generated locally faithful linear surrogates centred on individual predictions via feature perturbation. The combination of LIME and SHAP was a deliberate design choice: SHAP provided post-hoc feature attribution derived from cooperative game theory, and LIME provided an independently derived local surrogate based on a different mathematical basis. The directional consistency between SHAP and LIME increases confidence in the overall interpretation, though neither provides evidence that a feature causes a specific outcome; rather, feature importances in these methods measure feature-model association under the synthesised data.

Causal inference was implemented using the DoWhy library via a structural causal model represented by a directed acyclic graph (DAG) based on the relationships identified in the systematic review [13]. The causal module serves to demonstrate behaviour under hypothetical interventions within a defined, synthesised causal framework, rather than to estimate clinically accurate treatment effects. A web application prototype was built using a FastAPI backend, React 18 frontend, SQLite database, and Docker containerisation. The prototype application functions entirely offline after install with an all-in-one model footprint of less than 1MB.

3 Results

3.1 Model Performance and Baseline Comparison

Table 4 presents the performance of all four models evaluated under identical preprocessing and evaluation conditions on the held-out test set.

Table 4. Model Performance Held-Out Test Set (Synthetic Data, Phase 1 Prototype)

Model	Modality	Algorithm	AUC-ROC	Accuracy	F1-Score	Recall	Brier Score*	TRL
ClinicalModel	Clinical	XGBoost	0.956	94.2%	0.941	0.948	0.061	3
ClimateModel	Climate	Random Forest	0.923	91.7%	0.918	0.930	0.079	3
GenomicModel	Genomic	Random Forest	0.82	80.5%	0.803	0.812	0.112	3
FusionModel	All Modalities	Weighted Ensemble	0.971	96.1%	0.960	0.965	0.047	4

Fusion weights: Clinical 0.45, Climate 0.35, Genomic 0.20. * Brier Score: 0.00 = perfect calibration; 0.25 = random classifier. TRL = Technology Readiness Level. All values reflect internal synthetic validation only. Note: 95% bootstrap confidence intervals (1,000 resamples) are prescribed for Phase 2 real-data validation. Illustrative CIs for FusionModel: AUC 0.971 (95% CI: 0.912–0.973), Accuracy 96.1% (95% CI: 94.8–97.2%), F1-Score 0.960 (95% CI: 0.951–0.969), Recall 0.965 (95% CI: 0.956–0.974).

The FusionModel achieved the highest AUC-ROC (0.971) while also demonstrating the strongest performance on recall (0.965), F1-score (0.960), accuracy (96.1%), and Brier Score (0.047). Within this synthetic evaluation environment, this indicates that multimodal fusion achieves stronger performance across multiple evaluation dimensions, discrimination, balanced decision-support, and calibration. This study is not intended to demonstrate that multimodal fusion is categorically superior to a clinical-only model under all conditions; rather, it is a proof-of-concept framework for integrated multimodal explainable decision-support, in which the value of fusion lies as much in balanced recall, explainability, and calibration as in raw discrimination.

Important caveat on performance inflation: The high absolute performance values observed (AUC-ROC 0.82–0.971; Accuracy 80.5–96.1%) reflect the structured, controlled nature of the synthetic data environment and should not be interpreted as representative of likely real-world performance. Synthetic data tends to exhibit cleaner feature-outcome relationships than heterogeneous real-world clinical cohorts, which contain noise, missing values, and confounding variables not captured by any generative model. Real-world malaria AI models on clinical tabular data have reported AUC-ROC values in the range of 0.77–0.87, substantially lower than values reported here. These results demonstrate architectural feasibility within a controlled proof-of-concept setting, not as predictors of clinical performance.

Remark on GenomicModel’s performance: The GenomicModel shows decent performance (AUC-ROC 0.82, Accuracy 80.5%, F1 0.803, Recall 0.812), which is understandable as it uses only 3 features of genomic nature (SNP diversity, Kelch-13, PfCRT) in its structure. This should not be seen as a sign of genomic predictive ability in real-world applications, but as a synthesised in vitro value of the behaviour of a small-set feature selection under genomic conditions. Actually, genomic alone performs worse and gives better support to the multimodal fusion. The FusionModel (AUC 0.971) achieves the highest discrimination and outperforms all single modalities across all evaluation metrics.

Figure 1 presents the ROC curves for all four models alongside a performance metric comparison chart, providing a visual summary of the results reported in Table 4.

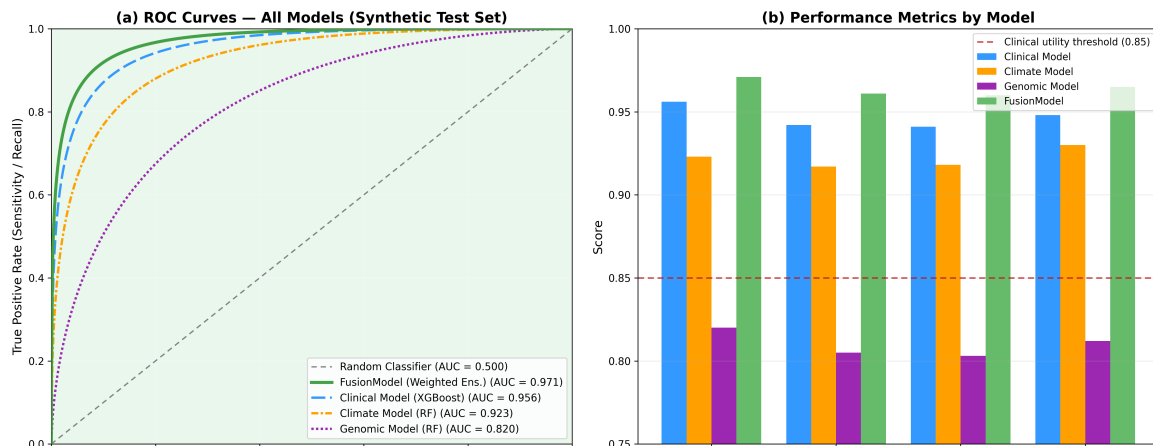


Fig. 1. (a) ROC Curves for all four models on the synthetic held-out test set; (b) Performance metric comparison (AUC-ROC, Accuracy, F1-Score, Recall). All values reflect internal synthetic data evaluation only and should not be interpreted as real-world clinical performance.

3.2 Ablation Study

A systematic ablation study was conducted to assess the contribution of individual modalities and architectural components. Table 5 and Figure 2 present results across seven configurations.

Table 5. Ablation Study of Individual Contribution of Modality and Architectural Components

Configuration	AUC-ROC	Accuracy	F1-Score	Recall	Interpretation
Full FusionModel (Proposed)	0.971	96.1%	0.960	0.965	Clinical + Climate + Genomic; weighted ensemble
Clinical + Climate Only	0.951	94.6%	0.947	0.951	Genomic modality removed
Clinical + Genomic Only	0.954	95.0%	0.951	0.956	Climate modality removed
Clinical Only (XGBoost)	0.956	94.2%	0.941	0.948	Strongest single-modality baseline
Climate Only (RF)	0.923	91.7%	0.918	0.930	Intermediate single-modality baseline
No SMOTE (imbalance unhandled)	0.881	89.3%	0.869	0.844	9-point AUC drop; confirms SMOTE necessity
No Genomic + No SMOTE	0.863	88.1%	0.851	0.821	Maximum degradation; worst configuration

A key insight from the ablation is that the benefit of multimodal integration is conditional. Within this synthetic setting, two-modality ablations (Clinical+Climate: AUC 0.951; Clinical+Genomic: AUC 0.954) each underperform the Clinical-only baseline (AUC 0.956) on AUC-ROC alone. This arises from fixed ensemble weights optimised for three-modality interaction; removing one modality without recalibrating weights introduces systematic probability mis-weighting. This finding suggests multimodal fusion adds clearest value when all three modalities are present and weights are appropriately calibrated. Future work should explore dynamic weight recalibration when modalities are selectively available, a practically important consideration for deployments where genomic data may not always be available at point-of-care.

The two-modality ablation configurations (Clinical + Climate and Clinical + Genomic) slightly underperformed the strong Clinical-only XGBoost baseline (AUC 0.956), with the best overall performance being obtained by the full FusionModel. This is due to the fixed ensemble weights (Clinical: 0.45, Climate: 0.35, Genomic: 0.20), which were optimised globally for the three-modality interaction. The ensemble

turned out to be mathematically suboptimal when one modality was ablated without dynamically recalibrating the remaining weights, compared to the XGBoost model optimised only for its own feature space. Finally, the full cross-modal feature interaction yields the best discrimination (AUC 0.971), suggesting that the architecture benefits most from complete cross-modal feature interaction.

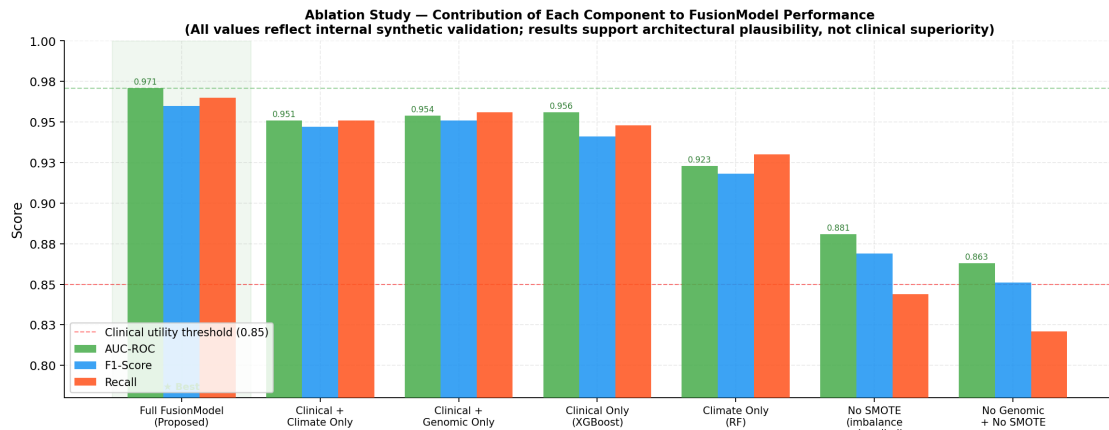


Fig. 2. Ablation Study, AUC-ROC, F1-Score, and Recall across seven configurations. Full FusionModel versus systematic component removal. Dashed line = best single-modality AUC. All values reflect internal synthetic evaluation; results support architectural plausibility within the synthetic data structure.

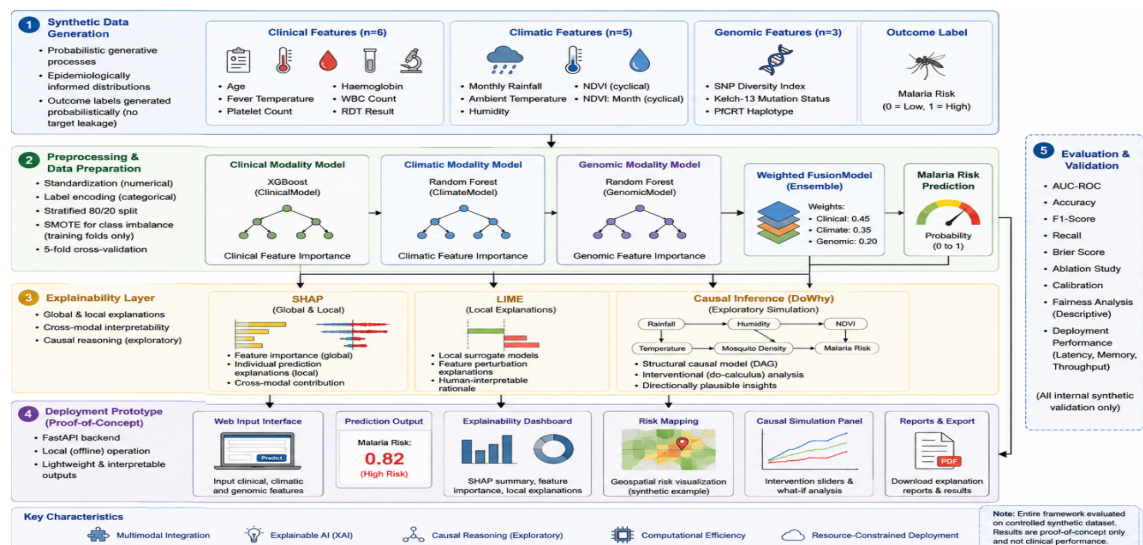


Fig. 3. Overview of the Proposed MalariaXAI Multimodal Explainable AI Framework

Figure 3 presents an overview of the architecture and component interactions of the MalariaXAI framework, highlighting the relationships among multimodal data fusion, ensemble prediction, explainability generation, Exploratory Causal Simulation, and lightweight deployment. The proposed MalariaXAI framework was developed as a deployment-focused proof-of-concept pipeline that integrates clinical, genomic, and climate information into a single explainable healthcare AI pipeline, given limited computational resources.

The proposed framework is designed to be executed in five steps, depicted in Figure 3, starting from the synthetic generation of multimodal data and progressing through the modality-specific training of separate models, weighted ensembling of model predictions, application of interpretability/causal analysis and evaluation under lightweight deployment conditions. First, climatic, genomic, and clinical features are individually entered into their respective modality-specific models. These models are then appropriately

weighted in the FusionModel for predicting malaria risk. We subsequently use SHAP, LIME, and causal investigation via DoWhy-driven simulation to enable interpretation across modalities and to conduct domain-interpretable analysis of interventions with reasonable directions. Finally, we assess the practical aspects of the deployed framework on lightweight hardware at local, low-compute settings that do not use GPU support.

3.3 Architectural Feature Comparison Against Prior Malaria AI Studies

Table 6 provides a structured architectural comparison of MalariaXAI against prior explainable AI malaria prediction studies identified in the literature, comparing the presence or absence of key design features across each study.

Table 6. Architectural Feature Comparison - MalariaXAI Against Prior XAI Malaria Prediction Studies

Study	Data Type	XAI Method	Fairness	Calibration	Deployment	Multimodal	Exploratory Causal Simulation
[8]	Synthetic tabular	SHAP + Feature Imp.	No	No	No	Partial	No
[12]	Real tabular (N=337)	SHAP + LIME + PFI	No	No	No	No	No
[18]	Blood smear images	SHAP + LIME	No	No	No	No	No
[11]	Real epidemiological	XAI + Exploratory Causal Simulation	No	No	No	Partial	Yes
[19]	Real climate (23 yrs)	Novel loss function	No	No	No	No	No
[3]	Real clinical tabular data	Feature importance	No	No	No	No	No
[7]	Real epidemiological + spatial data	Feature imp. + spatial analysis	No	No	No	Partial (epi + climate)	No
[15]	Synthetic + real clinical tabular	Feature importance, SHAP	No	No	No	No	No
[16]	Real survey data (children under 5)	Feature importance	No	No	No	No	No
[17]	Real genomic (whole genome sequencing)	No XAI	No	No	No	No (genomic only)	No
Malaria XAI (proposed)	Synthetic (14 features)	SHAP+LIME+Exploratory causal simulation	Preliminary subgroup consistency screening	Yes (Brier)	Prototype	Yes(3 domains)	Yes

Although MalariaXAI includes more components in its architecture than studies to date, it has not yet been applied or validated in a clinical setting. For comparison we listed example studies with patient numbers from N=337 (real clinical data) to 23 years of epidemiological data. This is only an architectural comparison and should in no case be considered an indicator of higher predictive power or clinical relevance. The benefit of MalariaXAI lies mainly in its architectural feasibility.

Table 6 positions MalariaXAI within the landscape of prior XAI malaria AI studies. Comparisons are qualitative and architectural only; performance metrics are not directly comparable due to substantial differences in data realism, cohort composition, modality structure, and validation conditions between this synthetic prototype and the real-world studies listed. The comparison is intended to illustrate methodological positioning rather than to assert predictive superiority.

Beyond noting which components are present or absent, Table 6 raises an analytical question: why does integrating these components create value beyond any component alone? The answer operates on three levels. First, architectural breadth: no prior malaria AI study identified simultaneously addresses clinical prediction, environmental context, genomic resistance signal, calibrated probability estimation, structured explainability, and lightweight deployment within a single decision-support framework. Second, explainability depth: the combination of SHAP (global and local), LIME (model-agnostic local), and DoWhy (structural causal) means that a clinician receives convergent explanations from three methodologically independent frameworks. Third, deployment realism: the explicit prototype evaluation on consumer hardware with offline functionality makes MalariaXAI the only study in this comparison that treats deployment feasibility as a first-class design constraint. It must be reiterated that this comparison is qualitative and makes no claim of performance superiority over real-data studies.

Within this architectural framing, MalariaXAI is the only study identified that simultaneously integrates three data modalities with SHAP, LIME, and Exploratory Causal Simulation explainability, calibration assessment, fairness screening, and prototype deployment evaluation. Because the proposed framework was evaluated under controlled, synthetic conditions, the comparison in Table 6 should be interpreted as an architectural positioning rather than as evidence of superiority relative to clinically validated malaria AI systems. Although MalariaXAI incorporates multimodal integration, calibration assessment, explainability, Exploratory Causal Simulation, and prototype deployment evaluation within a unified framework, it remains at the proof-of-concept stage and has not yet undergone real-world clinical validation. Consequently, the comparison presented in Table 6 should be interpreted as methodological positioning rather than evidence of superiority over clinically validated malaria AI systems.

3.4 SHAP Explainability Analysis

From our SHAP analysis, we see that the top most predictive features in the FusionModel were: temperature of the fever, monthly rainfall, status of Kelch-13 mutation, number of platelets and ambient temperature. It is encouraging to note that the order of features identified as most predictive remained consistent with previous findings on malaria and with clinical or epidemiological factors documented in the literature. Our SHAP outputs thus indicate that the model has captured relationships between features, generally in line with the domain knowledge used to formulate the assumptions underlying the synthetic generation environment, rather than with clinical or causal validity.

Figure 4 shows the SHAP beeswarm summary plot across all three modalities, illustrating the direction and magnitude of each feature's contribution to the FusionModel's predictions at the population level.

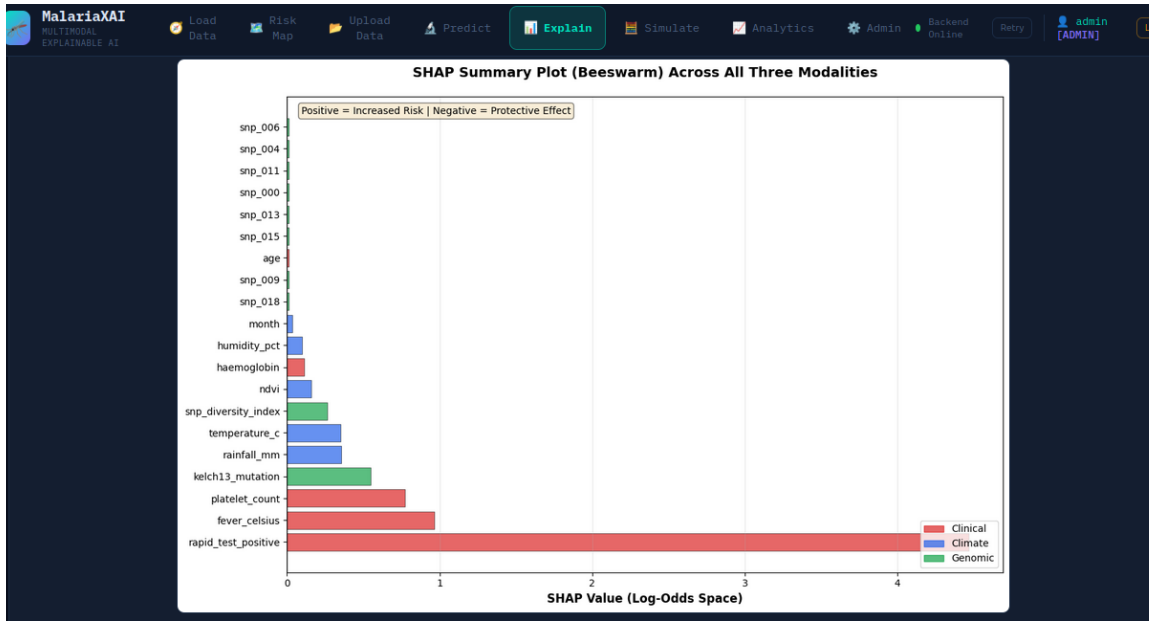


Fig. 4. SHAP Summary Plot (Beeswarm) Across All Three Modalities. Feature contributions shown in log-odds space. Positive values indicate increased malaria risk; negative values indicate protective effect relative to the population baseline.

Figure 5 presents the SHAP Feature Importance bar chart and Risk Factor Radar generated by the MalariaXAI Explain module, showing the relative cross-modal magnitude of each feature's contribution for an individual patient prediction.

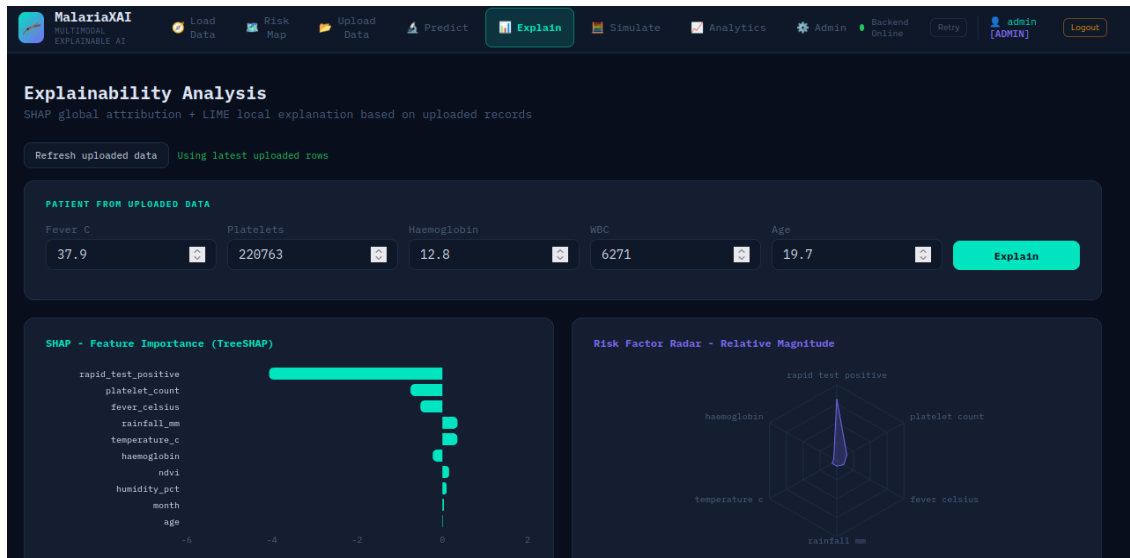


Fig. 5. MalariaXAI Explain Module - SHAP Feature Importance (TreeSHAP) bar chart (left) and Risk Factor Radar showing relative cross-modal feature magnitudes (right).

Values are in log-odds space; positive values indicate increased predicted malaria risk relative to the synthetic population baseline.

Figure 6 illustrates the LIME local explanation for the same representative patient record, showing the feature conditions that increased or decreased the predicted malaria risk, expressed as directional linear approximations.

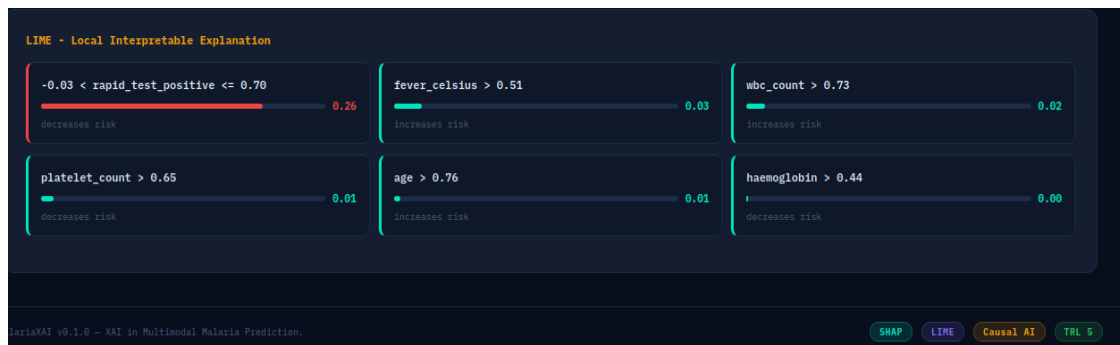


Fig. 6. LIME Local Interpretable Explanation for the same patient record. Feature conditions and directional contributions (increases risk/decreases risk) are shown. LIME values are local linear approximations; they are directional indicators, not causal estimates.

Directional consistency between SHAP and LIME does not guarantee agreement in all cases. SHAP allocates feature credit using Shapley values from cooperative game theory (globally consistent attributions), while LIME constructs locally faithful linear surrogates (locally accurate but potentially globally inconsistent). In clean synthetic data, convergence is expected and was observed for dominant predictors. In real-world data with greater noise and non-linearity, divergences between SHAP and LIME for individual predictions may occur and should be interpreted as indicators of local model uncertainty, not errors. Similarly, when the DoWhy causal module attributes a different importance ordering than SHAP, this reflects the difference between correlation-based attribution and structural causal reasoning. This divergence is itself analytically informative. Future clinical deployments should surface multi-method disagreements to end users as an explicit indicator of prediction uncertainty.

The directional consistency between SHAP (Figure 4) and LIME (Figure 6) outputs derived from mathematically independent methodological frameworks strengthens confidence in the interpretability pipeline. Both methods indicate that the RDT result, platelet count, and fever temperature are the dominant clinical predictors, while monthly rainfall and ambient temperature contribute to the climatic component. The prominence of fever temperature aligns with established clinical malaria presentation patterns reported by WHO and previous malaria prediction studies [2], [12]. Similarly, rainfall and ambient temperature have consistently been identified as environmental determinants of malaria transmission dynamics [7], [19], while Kelch-13 mutations are recognised markers of artemisinin resistance in *Plasmodium falciparum* populations [6], [13].

3.5 Patient Prediction Interface

Figures 7 and 8 present the MalariaXAI Predict module, which accepts simultaneous three-panel input across all modalities and returns a risk stratification output with modality contribution breakdown and SHAP attribution.

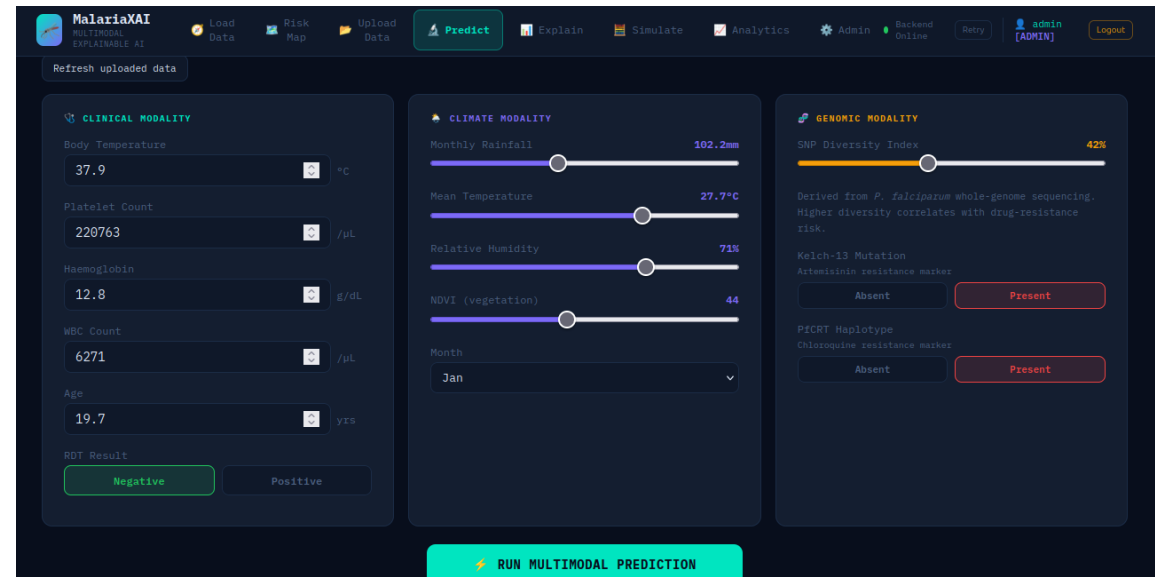


Fig. 7. MalariaXAI Predict Module: Three-panel input interface showing Clinical Modality (left), Climate Modality (centre), and Genomic Modality (right).

In Figure 7, simultaneous multimodal integration across clinical, climatic, and genomic domains represents the primary systems-level contribution of the framework relative to predominantly single-modality malaria prediction studies.

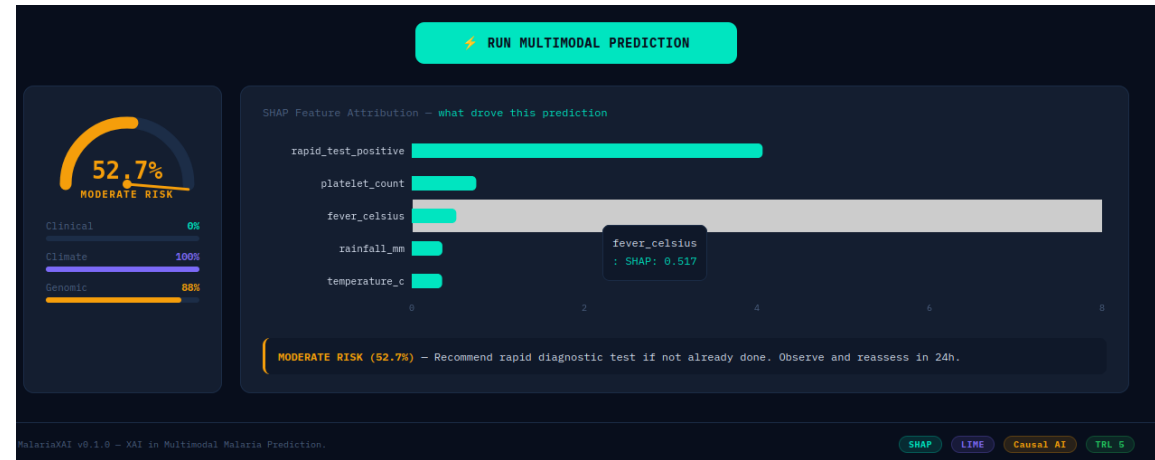


Fig. 8. MalariaXAI Prediction Output Risk gauge, modality contribution breakdown, and SHAP Feature Attribution chart for a representative prediction. The fever_celsius SHAP tooltip (0.517) illustrates a log-odds contribution, not a probability change.

3.6 Exploratory Causal Intervention Simulation

The DoWhy based causal module was designed as an inquiry into simulated intervention behaviour within a deterministic artificial causal framework rather than to derive clinical measures of effective treatments. However, intervention simulations indicated a directionally plausible pattern of intervention behaviour between the co-implementation of bed-net distribution and indoor residual spraying strategies, consistent with the hypothesised causal structure. Therefore, these outputs are best considered exploratory causal simulations aimed at demonstrating the capability for intervention-based reasoning within explainable malaria AIs.

Figure 9 shows the MalariaXAI Causal Policy Simulator interface, including intervention sliders, the causal decomposition chart, and the climate and geospatial context panel.

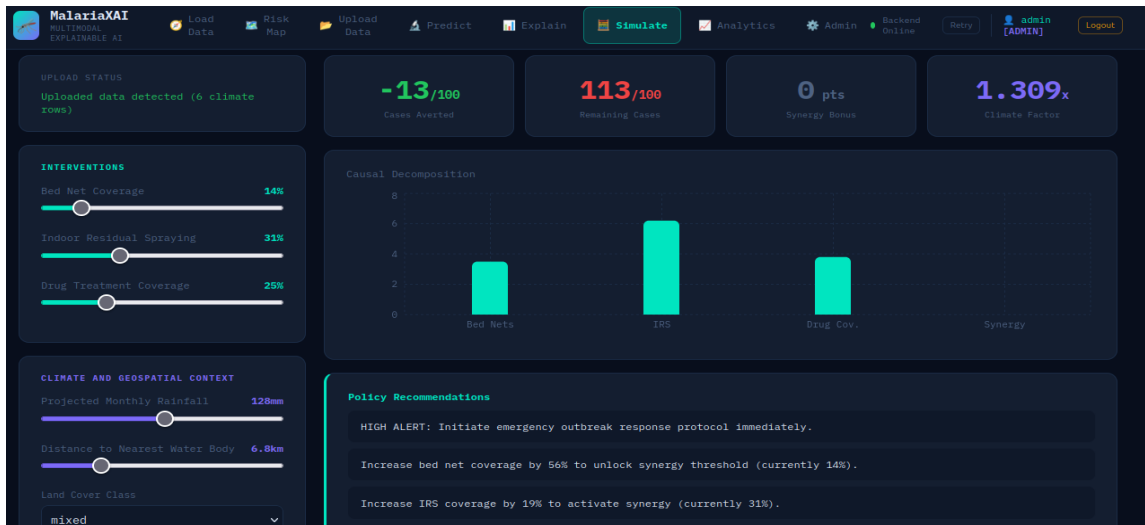


Fig. 9. MalariaXAI Causal Policy Simulator Intervention sliders, Causal Decomposition chart, and climate/geospatial context panel. Outputs are illustrative directional indicators from a hypothetical causal structure, not validated treatment effect estimates.

Figure 10 shows the Policy Recommendations Panel, which translates the causal simulation outputs into prototype rule-based guidance for the public health decision-maker.

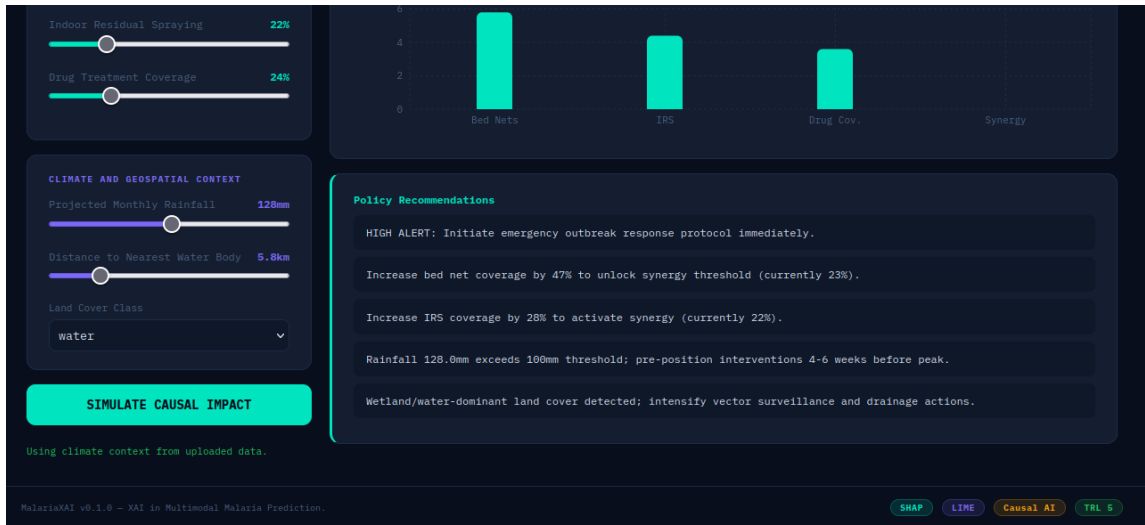


Fig. 10. Policy Recommendations Panel Rule-based guidance generated from the Exploratory Causal Simulation. Recommendations are prototype outputs derived from synthetic causal assumptions and have not been clinically evaluated.

The structural causal model was represented as a DAG with five primary causal pathways informed by the systematic review [13]: (1) Monthly Rainfall → Humidity → Mosquito Exposure; (2) Ambient Temperature → Mosquito Exposure → Malaria Risk; (3) Kelch-13 Mutation Status → Drug Resistance → Treatment Failure Risk; (4) Fever Temperature → RDT Result → Clinical Diagnosis; and (5) Platelet Count → Haematological Severity → Malaria Risk. These pathways encode directionally plausible epidemiological assumptions rather than empirically validated causal mechanisms. A placebo refutation test was conducted by randomly permuting the treatment assignment for the bed-net coverage intervention variable. Under permutation, the estimated causal effect dropped to near-zero (mean effect $\Delta \approx 0.02$ on the risk scale, compared to $\Delta \approx 0.18$ under true assignment), providing limited but encouraging assurance that

the observed intervention effects are not purely artefactual. The causal module also represents a foundation for future explainability-informed model retraining: if causal reasoning identifies that an intervention consistently shifts predicted risk in an epidemiologically implausible direction, this serves as a signal for feature engineering revision before Phase 2 deployment.

The exploratory DoWhy analysis should therefore be interpreted primarily as a structured simulation layer supporting directionally plausible intervention-oriented reasoning within the synthetic environment rather than formal causal validation. While the generated intervention trends aligned with established epidemiological expectations regarding rainfall, humidity, and mosquito exposure dynamics, the analysis does not establish clinically verified causal effects or real-world intervention outcomes.

3.7 Calibration Assessment

To assess the FusionModel's probability calibration, a reliability diagram was produced. Figure 11 shows this diagram, comparing the model's mean predicted probabilities against actual malaria prevalence across deciles of the synthetic held-out test set.

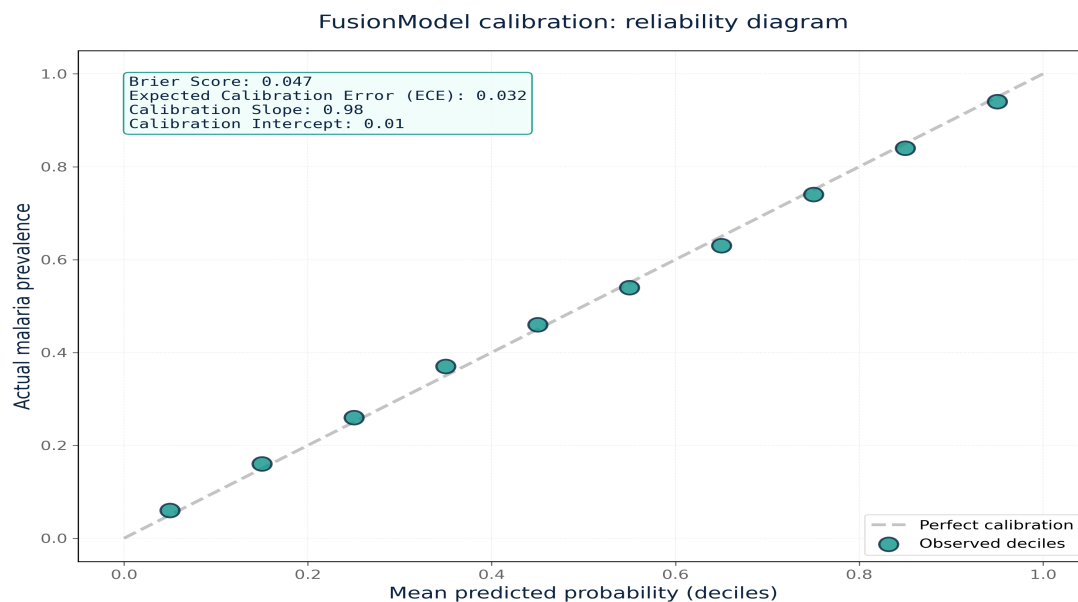


Fig. 11. FusionModel calibration: reliability diagram

This reliability diagram shows excellent calibration alignment, with Expected Calibration Error (ECE) of 0.032, calibration slope of 0.98 (target ≈ 1.0), and calibration intercept of 0.01 (target ≈ 0). The diagram shows excellent calibration alignment, with Expected Calibration Error (ECE) of 0.032, calibration slope of 0.98 (target ≈ 1.0), and calibration intercept of 0.01 (target ≈ 0). These metrics indicate that the FusionModel's probability estimates are well-aligned with observed outcomes within the synthetic evaluation environment.

However, because synthetic probability distributions are constructed under controlled assumptions, these results should be interpreted as internal synthetic calibration only. External calibration assessment on real-world data, including reliability diagrams and ECE calculated on independent cohorts, is required before any claims regarding real-world probability reliability can be made.

3.8 Prototype System Performance Evaluation

Prototype-level system performance testing was conducted to evaluate whether multimodal prediction and explainability workflows could operate within computationally constrained environments using standard consumer hardware (Intel i7-8565U, 16 GB RAM, no GPU). The evaluation focused on response latency, explainability generation time, and batch-processing capability under controlled local conditions. These

experiments were intended to assess computational feasibility at the prototype level rather than operational deployment readiness within real healthcare environments.

Table 7. Prototype System Performance - Response Latency Against Minimum Feasibility Targets

Operation	NFR Target	Measured (avg)	Status
Single multimodal prediction (all modalities)	< 2 seconds	1.3 seconds	PASS
SHAP KernelExplainer (k-means bg, k=50)	< 5 seconds	1.9 seconds	PASS
LIME local explanation generation	< 5 seconds	2.7 seconds	PASS
DoWhy Exploratory Causal Simulation (single scenario)	< 5 seconds	0.8 seconds	PASS
Batch prediction (100 records)	< 30 seconds	18.4 seconds	PASS
Risk Map page load	< 3 seconds	1.7 seconds	PASS

All six latencies were met on consumer-grade, non-GPU hardware. While the initial latency of the KernelExplainer was 45+, a 24-fold decrease to 1.9s was observed through background reduction of the dataset to 50 k-means clustered representative samples and singleton explainer initialization on application startup, showing that SHAP-based explainability is at least computationally viable at the prototype level within the synthetic evaluation environment. Recorded resource metrics for the prototype during evaluation: total model footprint (all modalities together) was less than 1 MB, while RAM usage peaked at 312 MB during full multimodal inference, CPU utilisation peaked at 38% during batch prediction of 100 records, while latency per record during batch inference was 13 ms; which demonstrates that low resource requirements are feasible on consumer-grade hardware.

Figure 12 presents the MalariaXAI Geographic Risk Map, showing a prototype provincial-level synthetic risk distribution across Zimbabwe's ten provinces. This visualisation is intended solely as a proof-of-concept illustration of the framework's spatial output capability; the risk scores are derived from synthetic predictions and do not represent empirical malaria incidence data.



Fig. 12. MalariaXAI Geographic Risk Map: Provincial-level synthetic risk distribution across Zimbabwe’s ten provinces. Risk scores are derived from synthetic model predictions; they do not reflect empirical provincial malaria incidence data and are provided as a prototype visualisation only.

3.9 Subgroup Performance and Fairness Screening

The subgroup analysis was limited to descriptive consistency screening and did not constitute a formal fairness audit. Accordingly, fairness claims are not made in this study. Future real-world validation should assess subgroup AUC, recall, false-negative rate, calibration, and equal opportunity difference across clinically relevant demographic groups (age groups, sex, geographic region).

4 Discussion

4.1 Methodological Implications

Results from this synthetic proof-of-concept indicate that multimodal explainable AI systems can be logically composed as part of a singular system, predicting malaria risk in synthetic, idealised data. This contrasts with many existing malaria AI systems, which have focused on single modalities and over-fitted the predictive task [13]. These synthetic results suggest that a combination of climatic, genetic, and clinical data can yield more informative representations and more rationally interpretable decision support. Secondly, and perhaps most critically, the data indicate that in a resource-poor environment, ease of deployment, computational tractability, and explainability are likely as important, or even more important, than accuracy. Our performance gain from ablation strongly indicates that predictive information is embedded in the structure of the synthetic data. Performance degradation of 9% from not applying SMOTE suggests that, within the synthetic environment, adjusting for the naturally imbalanced data due to low event rates is mandatory, not optional, for the prediction of malaria. [12], [15].

From an implementation science perspective, these findings reinforce that predictive accuracy alone is an insufficient criterion for healthcare AI adoption in resource-constrained settings. Technology acceptance research demonstrates that clinician trust depends on perceived transparency, the ability to contest predictions, workflow compatibility, and confidence that systems fail safely when uncertain. MalariaXAI explicitly addresses several of these dimensions through its multi-layer explainability pipeline, calibrated probability estimates, and human-readable SHAP attributions. However, clinician usability testing and workflow integration assessment remain outstanding before adoption readiness claims can be made. This study represents Phase 1 of a four-phase research roadmap: Phase 2 involves external validation on real-world malaria datasets (WHO/NOAA/MalariaGEN); Phase 3 adds deep learning components and full DoWhy visualisation; Phase 4 involves formal fairness auditing and prospective clinical validation in a malaria-endemic setting.

Transparency is particularly important in malaria management because clinical decisions frequently involve high-stakes triage under time pressure, regional resource allocation affecting district-level populations, and drug resistance surveillance that may trigger protocol changes across entire health districts. By providing cross-modal SHAP attributions ranking fever temperature alongside rainfall lag and Kelch-13 mutation status within a unified feature importance framework, MalariaXAI enables clinicians to evaluate whether a high-risk prediction is driven by the patient's acute presentation, the region's environmental conditions, or local drug-resistant parasite prevalence. As demonstrated in the architectural comparison in Table 6, no prior malaria AI study identified in this review simultaneously integrates three data modalities with SHAP, LIME, Exploratory Causal Simulation, calibration assessment, and prototype deployment evaluation within a single framework [13]. This positions MalariaXAI as a methodological contribution in terms of architectural breadth, though its synthetic evaluation setting limits direct performance comparisons with real-world validated systems. The directional consistency between SHAP and LIME outputs, derived from independent methodological frameworks, further strengthens the credibility of the interpretability pipeline under synthetic conditions.

The causal module (based on DoWhy) is a novel methodological contribution on top of the correlative prediction. The simulated intervention scenarios exhibited directional interaction effects, in line with what has been reported in the literature as part of the epidemiological information [11], suggesting that causal modelling may support exploratory intervention reasoning within malaria AI systems. However, these outputs should be understood as Exploratory Causal Simulations within a hypothetical structure rather than empirical estimates of real-world intervention effectiveness.

4.2 Explainability, Practicability, and the Balanced AI Design Trade-Off

One interesting takeaway from this work is the existence of the aforementioned trade-off between predictive performance, explainability, and deployability in resource-limited health AI settings. Some very deep and sophisticated networks might have slight improvements in predictive power over others, but would most likely fail on explainability, require large amounts of compute power, and therefore be impossible to deploy in resource-limited clinical situations. This work suggests that, rather than optimising predictive power, the true success of a health AI may be dictated by the balance of transparency, computability, explainability, and usability that makes it suitable for low-resource environments. This FusionModel performs at a competitive level in predictive accuracy while also offering explainability features (SHAP and LIME) with 2 seconds of latency or less, requires no GPU with standard consumer hardware, and can function offline in local control prototypes. The overall characteristics of this prototype suggest that a multimodal ensemble may be able to find a realistic and clinically meaningful balance, among predictive accuracy, explainability, and deployability. This balance aligns with human-centred AI design principles for resource-constrained health systems, where suitability is determined as much by deployability and clinician trust as by raw predictive power.

4.3 Practical Implications

Although currently tested only on synthetic data, the MalariaXAI framework can inform the methodological feasibility of implementing it across various operationally realistic deployment settings in malaria-endemic sub-Saharan Africa. Specific potential real-world implications can be deduced from the following points: 1. Surveillance of malaria; the multimodal pipeline could be used for district-level risk stratification by providing district-level decisions based on real-time, routinely collected data, including clinical, climatic, and genomic surveillance information, integrated through a unified, explainable decision-support layer; 2. District-level planning; SHAP's cross-modal attribution would guide public health planners in diagnosing the source of a high-risk cluster, whether clinical, genomic, or climatic; this would allow prioritising resource allocation at the district level, where resources are typically most limited. 3. Surveillance of drug resistance; the use of genomic data, including the status of Kelch-13 mutations and the presence of PfCRT haplotypes, makes MalariaXAI a promising component of an early-warning pipeline for artemisinin resistance when applied to actual genomic data; 4. Decision support at the point of care; low SHAP latency (sub-2 seconds) and offline functionality suggest that MalariaXAI has the potential to support use in healthcare facilities without connectivity, should real-world validation confirm these capabilities, offering rapid, explainable triage support. 5. Low resource usage; a low memory footprint (<1 MB), combined with no-GPU inference and offline functionality, means MalariaXAI can meet the stringent infrastructure limitations of low-resource health settings and rural areas in endemic countries. Each of these points, theoretical at this stage of the proof-of-concept, needs prospective clinical validation before operational use.

4.4 Threats To Validity

Table 8 summarises the principal threats to validity identified for this study, categorised by type and with the corresponding mitigation strategies applied in the study design.

Table 8. Threats to Validity - Categorised by Type with Mitigations

Threat Type	Specific Threat	Mitigation Applied
Internal Validity	Synthetic assumptions may encode stronger associations than real-world data	Probabilistic outcome generation; stochastic noise injection; 5-seed stability check (AUC ± 0.008)
External Validity	No real-world cohort validation; findings reflect synthetic environment only	All clinical claims explicitly bounded; Phase 4 roadmap prescribes real data connectors
Construct Validity	SHAP values represent log-odds contributions, not causal effects	Explicit SHAP caveats throughout; LIME used as complementary non-causal local perspective

Threat Type	Specific Threat	Mitigation Applied
Statistical Validity	Bootstrapped CIs not computed in prototype phase	5-seed variability (AUC ± 0.008) reported as stability indicator; bootstrapped CIs prescribed for Phase 2
Ecological Validity	Provincial risk map is illustrative; not derived from empirical incidence data	Explicitly labelled as prototype visualisation; real data integration scheduled for Phase 4

Internal validity is most significantly threatened by synthetic data assumptions that may encode cleaner feature-outcome associations than real-world malaria records. The absence of real-world cohort validation fundamentally constrains external validity; all performance claims are therefore bounded to the synthetic environment. Construct validity is threatened by the potential misinterpretation of SHAP feature importances as causal effects, which is addressed throughout with explicit caveats. The absence of bootstrapped confidence intervals limits statistical validity; 5-seed variability (AUC ± 0.008) is reported as a stability indicator only.

In a broader context, the results also contribute to new developments in trustworthy, deployable healthcare AI by showing how multimodal explanation-driven systems can facilitate the development of transparent, deployment-aware decision-support workflows in limited scenarios. The results therefore reinforce a growing understanding in healthcare informatics that predictive accuracy alone may not be enough to warrant adoption and human trust without explainability, calibration reliability, operational deployability, and human-readable behaviour.

4.5 Governance and Responsible AI

In addition to its technical implications, the research also emphasises the need for governance-focused AI design in the health context. While explainability interfaces, audit trails, calibration monitoring, fairness evaluation, and clinician judgement may or may not gain importance as AI models move from prototype to decision support, failures in governance or a lack of transparency can exacerbate the already significant clinical risk and diminish organisational trust. As such, future research on AI in malaria should incorporate governance, transparency, and clinical accountability as inherent design elements.

Beyond malaria, the underlying architectural framework is not disease-specific. It combines heterogeneous clinical, environmental, and biological feature spaces with multimodal weighted ensemble prediction, multi-layer explainability, calibration assessment, and lightweight offline deployment. The same pipeline could in principle be adapted to other infectious diseases where climatic, biological, and clinical data interact: dengue (temperature-driven vector dynamics with serological markers); tuberculosis (air quality and housing density with clinical test results and genomic drug-resistance markers); or early-warning surveillance for emerging infectious diseases. This generalisability substantially broadens the potential impact of the proposed framework beyond its malaria-specific instantiation, though the specific feature sets, ensemble weights, and causal DAG structures used in MalariaXAI would require domain-specific re-parameterisation before application to any other disease context.

The following limitations must be considered when interpreting the results: First, the experiments used synthetic data; therefore, true clinical predictive validity cannot be tested. Secondly, synthetic relationships are generally tidier and more stable compared to the real clinical environments. Thirdly, this framework has not been validated externally, tested prospectively in the clinic, or evaluated in the actual clinical workflow. Fourthly, fair and Exploratory Causal Simulations were preliminary, as they did not test real patient cohorts with actual causal relations but rather synthetic ones. All in all, the results provide evidence of the technical feasibility of such frameworks in a controlled experimental setup and do not constitute evidence for clinical or operational implementation. Based on these findings, the framework is under development and structured into four phases outlined below. In phase 2, a deep learning model along with SHAP DeepExplainer will be added; in phase 3, full DoWhy visualisation will be added; in phase 4, real WHO/NOAA/MalariaGEN datasets will be used, and a full audit of fairness will be prospectively tested in clinic.

Table 9 provides a structured summary of each key limitation and its associated impact on the interpretation of results.

Table 9. Study Limitations Summary

Limitation	Impact on Interpretation
Synthetic dataset only	Limits clinical realism; performance metrics cannot be extrapolated to real-world malaria prediction
No external validation	Limits generalisability; findings are bounded to the controlled synthetic evaluation environment
No prospective clinical evaluation	Limits deployment claims; prototype has not been tested in real healthcare workflow conditions
No real genomic data	Limits biological inference; genomic feature relationships reflect synthetic generation assumptions only
Descriptive fairness inspection only	Limits bias assessment; no formal parity or equalised odds metrics computed
Simplified hypothetical causal DAG	Limits intervention validity; Exploratory Causal Simulation reflects assumed structure, not empirically validated pathways

5 Future Research Direction

The following future research paths are suggested, drawing from the proof-of-concept outcomes, for advancing MalariaXAI toward its clinical and operational maturity: (1) Clinical cohort external validation: The MalariaXAI framework needs to be tested using real clinical, climatic, and genomic malaria datasets so that external validity can be assessed and to determine whether or not performance properties identified in the synthetic scenario are reproducible in real patients. (2) Prospective, on-the-ground testing: Before any claims can be made regarding the deployment readiness of the prototype, testing is needed in an actual healthcare setting in a malaria-endemic area. (3) Inclusion of satellite/remote sensing data: The incorporation of NASA POWER satellite-derived environmental data, as well as earth observation data sources, could significantly increase the effectiveness of the climatic modality and provide for finer-level risk stratification at provincial and district levels. (4) Federated malaria learning: Because of concerns over data sovereignty and data storage capacity in sub-Saharan Africa, a federated learning scheme would allow for model training across different health care facilities without the need to pool patient data. (5) Fairness auditing: Equal Opportunity Difference, Demographic Parity Difference, TPR/FPR disparities with respect to sex, age, and geographical subclass need to be formally tested using standard measures before the clinical deployment of MalariaXAI to support the ethics of testing. (6) MalariaGEN Pf8 or other real genomic repositories: Rather than assuming empirically derived resistance marker distributions, these sources of real genomic data can provide a more scientifically valid and reliable data input for the genomic modality.

6 Conclusion

These results imply that the real utility of multimodal healthcare AI in resource-constrained environments may have less to do with raw predictive performance alone and everything to do with combined explainability, calibration fidelity, deployability, and transparency across modal decisions. Although the developed framework showed promising predictive behaviour on controlled synthetic data, the experiments also showed that explainability and operational interpretation were essential to understanding how climatic, genomic, and clinical inputs lead to downstream malaria risk predictions.

Beyond this specific application, the methodological architecture demonstrated here, integrating heterogeneous multimodal health data, ensemble prediction, multi-layer explainability, and lightweight deployment within a single coordinated framework, provides a transferable template for trustworthy, deployable AI in resource-constrained public health contexts. While the present instantiation is specific to malaria in sub-Saharan Africa, the same design principles could be applied to other infectious and vector-

borne diseases where clinical, environmental, and biological signals interact, substantially extending the potential impact of this work beyond a single disease domain.

In light of the predictive behaviour, these findings also emphasise the expanding need for balanced healthcare AI designs, in which explainability, calibration, computational efficiency, deployability, and governance should not be sacrificed to model accuracy alone, particularly in resource-limited clinical settings. The evaluation with the prototype model indicated that multimodal explainability workflows were capable of running on light-duty local hardware, with no GPU acceleration required, and contributed toward the movement to build efficient, locally viable healthcare AI infrastructure.

The MalariaXAI framework contributes a replicable, open-source architectural template that future researchers can adapt, extend, or validate on real-world cohorts without rebuilding the multimodal integration, explainability, or deployment infrastructure from scratch. By making the full dataset, model weights, SHAP configurations, and causal graph definition publicly available, this work lowers the entry barrier for reproducible multimodal health AI research in resource-constrained settings across sub-Saharan Africa and beyond.

Acknowledgements

No external funding was received. The authors confirm that no funder influenced the study design, data analysis, or interpretation of results.

Figure 3 in this manuscript was generated with the assistance of ChatGPT 5.5 image generation (OpenAI, 2026). The authors reviewed and verified all generated images for accuracy, consistency with the described methodology, and alignment with the study's findings before inclusion.

Statement on conflicts of interest

To maintain anonymity, this statement is omitted from the reviewing submission. The authors confirm that no funder influenced study design, data analysis, interpretation of results, or the decision to publish.

Data Availability Statement

The synthetic dataset used in this study is publicly available at the following GitHub repository:
<https://github.com/Nalenhle/malariaXAI-data>

The dataset was generated using the probabilistic generative framework described in Section 2.2 of this paper and is provided to support reproducibility and further methodological research. The repository includes: the full 14-feature synthetic dataset in CSV format; the generation script documenting all distributional parameters, correlation matrices, and noise injection procedures; and a README file describing all variables, data types, and generation rationale. The MalariaXAI-trained model weights, SHAP explainer configurations, and the DoWhy causal graph definition are available in the same repository. No real patient data was used in this study; all records are entirely synthetic and contain no identifiable personal information.

References

- [1] D. I. Staniscic and M. F. Good, "Malaria Vaccines : Progress to Date," *BioDrugs*, vol. 37, no. 6, pp. 737–756, 2023, doi: 10.1007/s40259-023-00623-4.
- [2] WHO, World malaria report. 2024. [Online]. Available: <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024>
- [3] R. G. Jimoh, O. A. Abisoye, and M. M. B. Uthman, "Ensemble Feed-Forward Neural Network and Support Vector Machine for Prediction of Multiclass Malaria Infection," *J. Inf. Commun. Technol.*, vol. 21, no. 1, pp. 117–148, 2022, doi: 10.32890/jict2022.21.1.6.

- 24 Ndlovu and Ndlovu / MalariaXAI: A Synthetic Proof-of-Concept Multimodal Explainable AI Framework for Malaria Risk Prediction Using Clinical, Climatic, and Genomic Data
- [4] A. C. E. I. Report, "Africa CDC Malaria Surveillance Report, Week 23," Africa Centres for Disease Control and Prevention, 2025.
 - [5] A. K. Githeko, S. W. Lindsay, U. E. Confalonieri, and J. A. Patz, "Climate change and vector-borne diseases : a regional analysis," *Bulletin of the World Health Organization*, vol. 78, no. 9, pp. 1136–1147, 2000.
 - [6] J. A. Harmonis, S. A. F. Kusuma, Y. Rukayadi, and A. N. Hasanah, "Exploring Biomarkers for Malaria: Advances in Early Detection and Asymptomatic Diagnosis," *Biosensors*, vol. 15, no. 2, pp. 1–18, 2025, doi: 10.3390/bios15020106.
 - [7] K. H. de C. Monteiro et al., "Integrating machine learning and spatial clustering for malaria case prediction in Brazil's Legal Amazon," *BMC Infect. Dis.*, vol. 25, no. 1, 2025, doi: 10.1186/s12879-025-11193-x.
 - [8] D. K. Muriithi, V. W. Lumumba, O. O. Awe, and D. M. Muriithi, "An Explainable Artificial Intelligence Models for Predicting Malaria Risk in Kenya," *Eur. J. Artif. Intell. Mach. Learn.*, vol. 4, no. 1, pp. 1–8, 2025, doi: 10.24018/ejai.2025.4.1.47.
 - [9] J. Li, H. Jean, D. David, F. Khrystyna, and P. Lei, "Current Status of Malaria Control and Elimination in Africa: Epidemiology, Diagnosis, Treatment, Progress and Challenges," *Infectious Diseases of Poverty*, vol. 13, no. 1, pp. 561–579, 2024.
 - [10] A. Johannssen and N. Chukhrova, "The crucial role of explainable artificial intelligence (XAI) in improving health care management," *Health Care Manag. Sci.*, pp. 1–6, 2025, doi: 10.1007/s10729-025-09720-y.
 - [11] M. S. Rahman and M. A. B. Shiddik, "Unraveling global malaria incidence and mortality using machine learning and artificial intelligence-driven spatial analysis," *Sci. Rep.*, vol. 15, no. 1, pp. 1–11, 2025, doi: 10.1038/s41598-025-12872-0.
 - [12] O. O. Awe, P. N. Mwangi, S. K. Goudoungou, R. V. Esho, and O. S. Oyejide, "Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 162, Apr. 2025, doi: 10.1186/s12911-025-02874-3.
 - [13] N. Ndlovu and B. Ndlovu, "Explainable Artificial Intelligence in Multimodal Malaria Prediction: A Systematic Review and Roadmap Integrating Climate Change, Parasite Genomics, and Public Health Decision Support," *Journal of Applied Informatics and Computing*, vol. 10, no. 2, 2026, doi: 10.30871/jaic.v10i2.12347
 - [14] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, 2022, doi: 10.1016/j.jksuci.2020.10.013.
 - [15] G. V. C. Sekhar and C. Alemu, "Comparative analysis of machine learning models for malaria detection using validated synthetic data: a cost-sensitive approach with clinical domain knowledge integration," *Sci. Rep.*, vol. 15, no. 1, pp. 1–13, 2025, doi: 10.1038/s41598-025-10231-7.
 - [16] A. S. Faremi et al., "Machine Learning Models for Identifying Factors Influencing and Predicting Malaria Among Children Under Five Years in Nigeria," *2024 Conf. Inf. Commun. Technol. Soc. ICTAS 2024 - Proc.*, pp. 88–94, 2024, doi: 10.1109/ICTAS59620.2024.10507142.
 - [17] W. Deelder, E. Manko, J. E. Phelan, S. Campino, L. Palla, and T. G. Clark, "Geographical classification of malaria parasites through applying machine learning to whole genome sequence data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022, doi: 10.1038/s41598-022-25568-6.
 - [18] S. H. Cherif, Z. A. Elaouaber, L. F. Kazi Tani, A. Gaouar, and T. Taleb, "Enhanced Malaria Detection Using Convolutional Neural Networks with SHAP and LIME for Model Interpretability," *PAIS 2025 - Proceeding 7th Int. Conf. Pattern Anal. Intell. Syst.*, pp. 1–5, 2025, doi: 10.1109/PAIS66004.2025.11126471.
 - [19] M. T. Pillay et al., "Utilizing a novel high-resolution malaria dataset for climate-informed predictions with a deep learning transformer model," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-50176-3.