DOI: http://dx.doi.org/10.26483/ijarcs.v17i1.7394

ISSN No. 0976-5697

Volume 17, No. 1, January-February 2026

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ROBUSTNESS ANALYSIS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS FOR MALARIA PREDICTION UNDER DATA PERTURBATION

May Stow
Department of Computer Science and Informatics,
Federal University Otuoke, Bayelsa State, Nigeria.
Orcid ID: https://orcid.org/0009-0006-8653-8363)

Abstract: Explainable artificial intelligence (XAI) methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model agnostic Explanations (LIME) are increasingly deployed in health surveillance systems to provide transparency in machine learning predictions. However, concerns persist regarding the reliability of these explanations under imperfect data conditions commonly encountered in resource-limited settings. This study investigates the robustness of SHAP and LIME explanations for malaria test positivity rate prediction under systematic data perturbations. Three gradient boosting models (XGBoost, LightGBM, CatBoost) were trained on malaria surveillance data from Bayelsa State, Nigeria comprising 2,100 records across eight local government areas. Model explanations were evaluated under controlled perturbations including Gaussian noise (5 to 100 percent), missing value injection (5 to 50 percent), and feature corruption (5 to 50 percent). Stability was quantified using Spearman rank correlation and top k feature overlap metrics. Results demonstrate exceptional robustness of SHAP explanations, with mean Spearman correlation coefficients of 0.976 for XGBoost, 0.981 for LightGBM, and 0.982 for CatBoost. SHAP consistently outperformed LIME across all conditions. The top five most important features remained consistent across most perturbation scenarios with 100 percent overlap for XGBoost and CatBoost SHAP. These findings provide confidence for deploying XAI based decision support systems in malaria surveillance programs where data quality may be suboptimal.

Keywords: Explainable artificial intelligence, SHAP, LIME, malaria prediction, model robustness, feature importance stability.

I. INTRODUCTION

Malaria remains a significant public health challenge in sub Saharan Africa, with Nigeria accounting for approximately 27 percent of global malaria cases [1]. Effective surveillance and early warning systems are critical for targeted intervention strategies and resource allocation in endemic regions [2]. Machine learning approaches have demonstrated considerable promise in predicting malaria transmission patterns by leveraging climate, demographic, and epidemiological data [3]. However, the deployment of these predictive models in public health decision making requires not only accurate predictions but also transparent and interpretable explanations that can be trusted by health practitioners and policymakers [4].

Explainable artificial intelligence (XAI) methods have emerged as essential tools for providing interpretability in complex machine learning models [5]. Among these, SHapley Additive exPlanations (SHAP) and Local Interpretable Model agnostic Explanations (LIME) are the most widely adopted approaches for explaining model predictions [6], [7]. SHAP, grounded in cooperative game theory, provides theoretically consistent feature attributions by computing the marginal contribution of each feature to the prediction [8]. LIME approximates model behavior locally through interpretable surrogate models, offering intuitive explanations for individual predictions [9]. Both methods have been successfully applied in healthcare contexts, including disease diagnosis, risk stratification, and treatment outcome prediction [10]. Previous research has provided comprehensive analysis of interpreting machine learning predictions using SHAP and LIME for transparent decision making across various application domains [11].

Despite their widespread adoption, significant concerns have been raised regarding the stability and reliability of XAI explanations [12]. Studies have demonstrated that minor perturbations in input data can lead to substantially different explanations, potentially undermining trust in these methods [13]. Formal metrics for evaluating explanation stability have been introduced, showing that many popular methods exhibit sensitivity to input variations [14]. Research has also demonstrated that adversarial perturbations could manipulate SHAP explanations while maintaining prediction accuracy [15]. These findings raise important questions about the trustworthiness of XAI methods in high stakes applications such as healthcare.

The challenge of explanation stability is particularly relevant in resource limited settings where data quality issues are prevalent [16]. Health surveillance systems in developing regions often face challenges including missing data, measurement errors, and inconsistent reporting practices [17]. Empirical analysis of SHAP stability under data corruption across multiple datasets and model architectures has found varying degrees of robustness depending on the underlying data characteristics [18]. For XAI methods to be practically useful in such contexts, they must provide reliable explanations even when data quality is compromised. However, limited research has systematically evaluated XAI robustness under realistic data quality degradation scenarios specifically in health applications.

This study addresses this gap by conducting a comprehensive robustness analysis of SHAP and LIME explanations for malaria test positivity rate prediction. Using surveillance data from Bayelsa State, Nigeria, the research systematically evaluates explanation stability under controlled data perturbations representative of real world data quality issues. The specific objectives are to: (1) develop and validate gradient boosting models for malaria test positivity rate prediction; (2) establish baseline feature importance rankings using SHAP and LIME; (3) quantify explanation stability under varying levels of Gaussian noise, missing data, and feature corruption; (4) compare the robustness of SHAP versus LIME explanations; and (5) identify factors influencing explanation stability.

The main contributions of this work include: (1) empirical evidence demonstrating exceptional robustness of SHAP explanations under extreme data perturbations, maintaining stability even at 100 percent Gaussian noise levels; (2) quantitative comparison showing SHAP consistently outperforms LIME in explanation stability with mean Spearman correlations of 0.98 versus 0.96; (3) analysis revealing the relationship between feature importance structure and explanation robustness; and (4) practical guidelines for deploying XAI methods in health surveillance systems with suboptimal data quality. These findings have important implications for building trust in machine learning based decision support systems for malaria control programs.

II. RELATED WORKS

A. Explainable AI in Healthcare

The application of explainable artificial intelligence in healthcare has grown substantially as machine learning models become more prevalent in clinical decision support [19]. SHAP values were introduced as a unified approach to interpreting predictions, demonstrating superior consistency compared to previous methods [6]. This framework has been widely adopted for explaining predictions in disease diagnosis [20], treatment response prediction [21], and risk stratification [22]. LIME was proposed as a model agnostic explanation technique that approximates complex models locally with interpretable surrogates [7]. LIME has been applied to explain predictions in medical imaging [23], electronic health record analysis [24], and epidemiological modeling [25].

In the context of malaria prediction, machine learning approaches have demonstrated effectiveness in forecasting transmission patterns and outbreak risk [26]. Gradient boosting methods, including XGBoost [27], LightGBM [28], and CatBoost [29], have shown particular effectiveness for tabular health data due to their ability to capture complex nonlinear relationships while maintaining interpretability through feature importance analysis. Explainable machine learning frameworks have been developed for various prediction tasks with class imbalance optimization, demonstrating approaches that could be adapted for health outcome prediction with imbalanced datasets [30]. However, the stability of explanations generated for these models under data quality degradation has not been systematically evaluated in the malaria surveillance context.

B. XAI Stability and Robustness

Concerns regarding the reliability of XAI explanations have motivated research into their stability properties. Formal metrics for evaluating explanation robustness have been proposed, demonstrating that many methods produce inconsistent explanations for similar inputs [14]. The concept of local Lipschitz continuity was introduced as a desirable property for explanation stability. Research has shown that adversarial examples could be constructed to manipulate SHAP explanations while maintaining model predictions, raising concerns about explanation trustworthiness in adversarial settings [15].

Several studies have examined the comparative stability of different XAI methods. Research has demonstrated that LIME explanations could be manipulated by adversarially trained models designed to hide discriminatory features [31]. Empirical comparisons of explanation disagreement among popular methods have found substantial variation in feature attributions across techniques [32]. Theoretical analysis of conditions under which explanation methods produce reliable outputs has also been provided [33]. This body of work has been extended by empirically analyzing SHAP stability under various data corruption scenarios across multiple datasets and model architectures, finding that stability varies significantly with underlying data characteristics [18].

C. Data Quality in Health Surveillance

Health surveillance systems in resource limited settings frequently encounter data quality challenges that may affect model predictions and explanations [16]. Missing data, measurement errors, and reporting inconsistencies are common issues in routine health information systems [34]. Research has examined data augmentation techniques on medical datasets, demonstrating that certain augmentation approaches can negatively impact model performance when applied inappropriately [35]. Similar challenges have been documented in other surveillance contexts including dengue fever [36], tuberculosis [37], and COVID 19 [38].

Research on machine learning robustness to data quality issues has primarily focused on prediction accuracy rather than explanation stability. Classifier performance under varying missing data rates has been examined [39], while other studies have evaluated model sensitivity to noise injection. However, limited attention has been given to how data quality degradation affects the interpretability and trustworthiness of model explanations. This gap is particularly significant for healthcare applications where understanding model reasoning is essential for clinical adoption and regulatory compliance.

D. Research Gap

While existing literature has established theoretical frameworks for explanation stability and documented vulnerabilities to adversarial manipulation, systematic empirical evaluation of XAI robustness under realistic data quality scenarios remains limited. Most stability analyses have focused on artificial perturbations designed to maximize explanation change rather than perturbations representative of real world data collection issues. Furthermore, the specific context of malaria surveillance in endemic regions, where data quality challenges are prevalent and XAI could provide substantial value, has not been adequately addressed. This study fills this gap by conducting comprehensive robustness analysis under perturbation scenarios relevant to health surveillance applications.

III. METHODOLOGY

The methodology employed in this study comprises six main components: data collection and description, feature engineering, model development, explainability analysis, perturbation framework design, and stability metric computation. Fig. 1 presents an overview of the complete analytical framework, illustrating the flow from raw surveillance data through model training, XAI explanation generation, perturbation application, and stability evaluation.

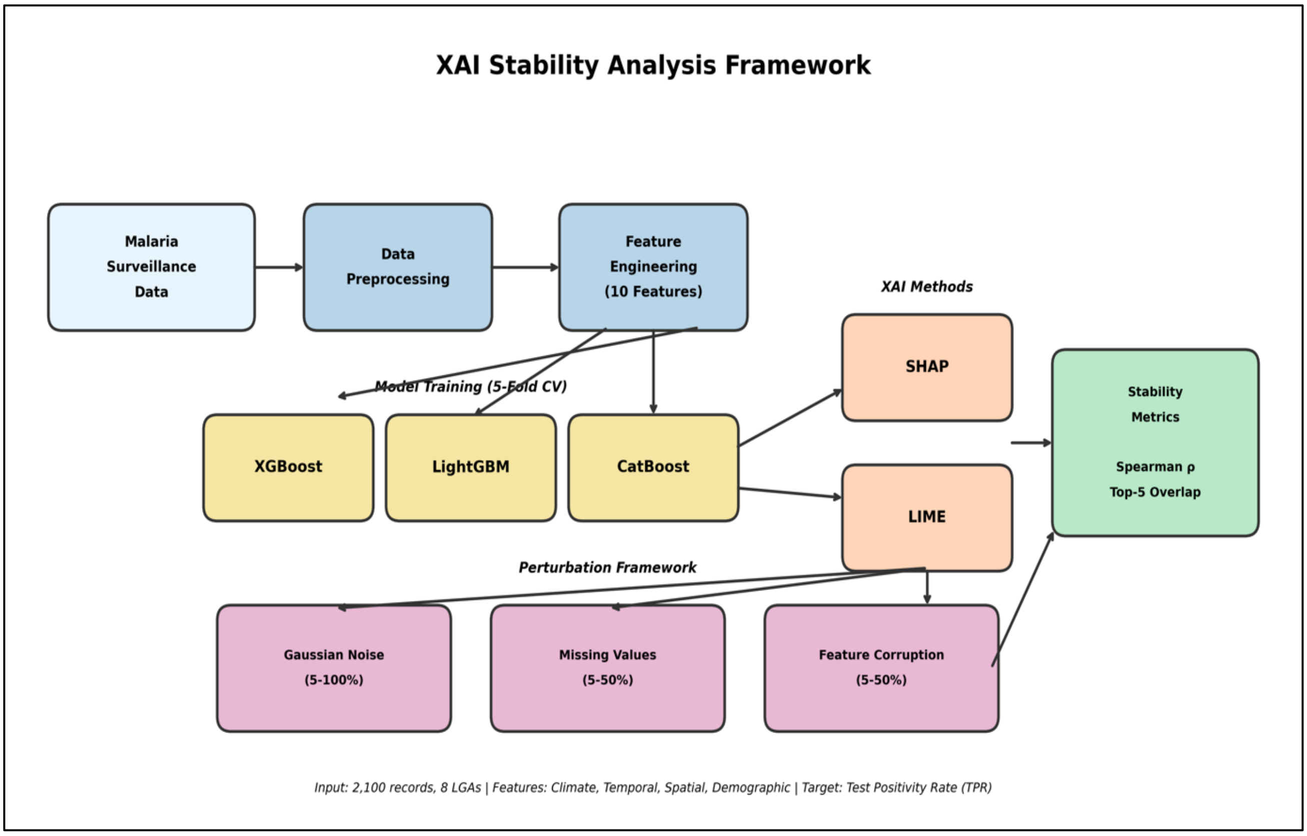
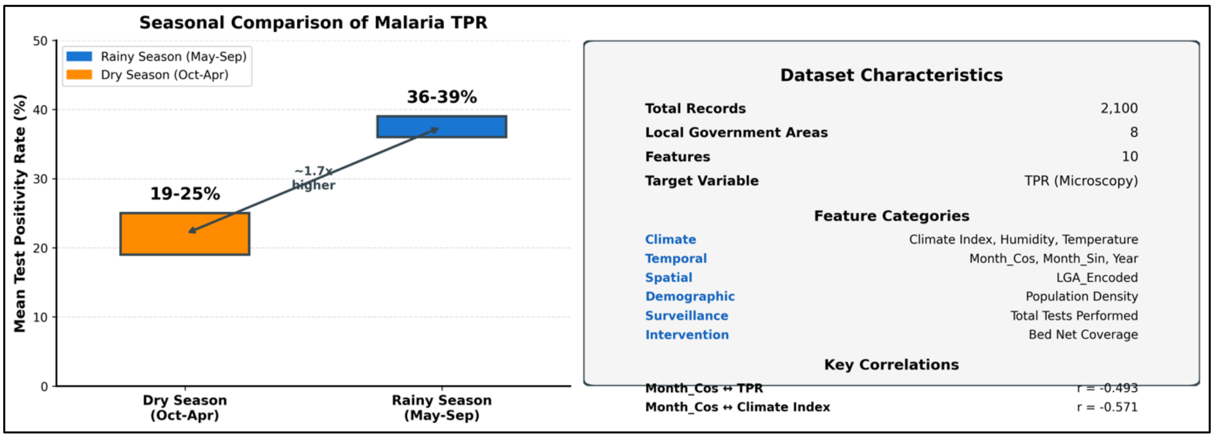

XAI Stability Analysis Framework
Malaria Surveillance Data
Data Preprocessing
Feature Engineering (10 Features)
XAI Methods
SHAP
LIME
Model Training (5-Fold CV)
XGBoost
LightGBM
CatBoost
Stability Metrics
Spearman ρ
Top-5 Overlap
Perturbation Framework
Gaussian Noise (5-100%)
Missing Values (5-50%)
Feature Corruption (5-50%)
Input: 2,100 records, 8 LGAs | Features: Climate, Temporal, Spatial, Demographic | Target: Test Positivity Rate (TPR)

Fig. 1. XAI stability analysis framework showing data preprocessing, model development, perturbation types, and stability evaluation workflow.

A. Study Area and Data Source

This study utilized malaria surveillance data from Ministry of Health, Bayelsa State, located in the Niger Delta region of southern Nigeria. Bayelsa State comprises eight local government areas (LGAs): Brass, Ekeremor, Kolokuma Opokuma, Nembe, Ogbia, Sagbama, Southern Ijaw, and Yenagoa. The region is characterized by a tropical climate with distinct wet and dry seasons, high humidity, and environmental conditions favorable for malaria transmission. The dataset contained 2,100 monthly records collected across the eight LGAs, encompassing epidemiological, climate, demographic, and intervention coverage variables.

The target variable was the malaria test positivity rate (TPR) derived from microscopy examinations, representing the proportion of tested individuals with confirmed malaria parasitemia. TPR serves as a key indicator for malaria transmission intensity and is widely used in surveillance programs for monitoring disease burden and evaluating intervention effectiveness. As illustrated in Fig. 2 below, the dataset exhibits pronounced seasonal patterns in malaria transmission. Mean TPR during the rainy season (May to September) ranged from 36 to 39 percent, approximately twice the dry season (October to April) values of 19 to 25 percent. This seasonal variation reflects the influence of rainfall and humidity on mosquito breeding and malaria transmission dynamics in the study area.



Dataset Characteristics

Total Records	2,100
Local Government Areas	8
Features	10
Target Variable	TPR (Microscopy)

Feature Categories

Climate	Climate Index, Humidity, Temperature
Temporal	Month_Cos, Month_Sin, Year
Spatial	LGA_Encoded
Demographic	Population Density
Surveillance	Total Tests Performed
Intervention	Bed Net Coverage

Key Correlations

Month_Cos ↔ TPR	r = -0.493
Month_Cos ↔ Climate Index	r = -0.571

Fig. 2. Dataset characteristics showing (left) seasonal comparison of malaria test positivity rate between dry and rainy seasons and (right) summary of dataset features and key correlations.

The seasonal pattern observed in Fig. 2 reveals a pronounced difference in malaria transmission between seasons. Mean Test Positivity Rate during the rainy season (May–September) ranged from 36–39%, approximately 1.7 times higher than the dry season (October–April), which recorded 19–25%. This seasonal disparity aligns with the key correlations identified in the dataset: Month_Cos showed a moderate negative correlation with TPR ($r = -0.493$) and a stronger negative correlation with Climate Index ($r = -0.571$), confirming that the cyclical temporal encoding effectively captures the underlying relationship between seasonal climate variation and malaria transmission dynamics.

B. Feature Engineering

A comprehensive feature engineering process was conducted to create informative predictors for the machine learning models. The final feature set comprised ten variables selected based on domain relevance, correlation analysis, and multicollinearity assessment. Climate variables included Climate Index (a composite measure incorporating rainfall and humidity), Humidity Percent, and Temperature in degrees Celsius. Temporal features were encoded using cyclical transformations (Month Sin and Month Cos) to capture seasonal patterns in malaria transmission. Spatial variation was represented through LGA Encoded, a numerical encoding of the local government areas. Additional features included Population Density, Total Tests Performed, Year Normalized, and Bed Net Coverage Percent representing intervention coverage.

TABLE I: FEATURE VARIABLES USED IN MODEL DEVELOPMENT

Feature	Description	Type
Climate Index	Composite climate measure	Continuous
Humidity Percent	Relative humidity percentage	Continuous
Temperature C	Temperature in Celsius	Continuous
Month Sin	Sine transformation of month	Continuous
Month Cos	Cosine transformation of month	Continuous
LGA Encoded	Encoded local government area	Categorical
Population Density	Population per square kilometer	Continuous
Total Tests Performed	Number of diagnostic tests	Count
Year Normalized	Normalized year value	Continuous
Bed Net Coverage Percent	Insecticide treated net coverage	Continuous

Feature selection was guided by multicollinearity analysis to remove redundant variables. Highly correlated feature pairs (correlation coefficient greater than 0.8) were identified and one member of each pair was removed to reduce model complexity and improve interpretability.

C. Model Development

Three gradient boosting algorithms were employed for malaria TPR prediction: XGBoost, LightGBM, and CatBoost. These algorithms were selected based on their demonstrated effectiveness for tabular data and their compatibility with SHAP TreeExplainer for efficient explanation computation. The target variable was log transformed to reduce skewness and improve model performance.

Model training employed five fold cross validation with early stopping to prevent overfitting. The dataset was split into 80 percent training (1,680 samples) and 20 percent test (420 samples) sets. Model performance was evaluated using the coefficient of determination (R squared) and the generalization gap, defined as the difference between cross validation training R squared and validation R squared. Models with generalization gaps below 0.10 were considered to have acceptable generalization performance.

D. Explainability Analysis

SHAP and LIME were employed to generate feature importance explanations for model predictions. SHAP values were computed using TreeExplainer, which provides exact Shapley value computation for tree based models. For each model, mean absolute SHAP values across test samples were calculated to obtain global feature importance rankings. LIME explanations were generated using the tabular explainer, and feature importances were aggregated across test samples to obtain comparable global rankings.

Baseline feature importance rankings were established for each model and XAI method combination under clean (unperturbed) data conditions. These baselines served as reference points for evaluating explanation stability under subsequent perturbation experiments.

E. Perturbation Framework

A comprehensive perturbation framework was developed to evaluate XAI stability under data quality degradation. As shown in Fig. 1, three types of perturbations were implemented: (1) Gaussian noise injection at levels of 5, 10, 20, 30, 40, 50, 60, 75, and 100 percent of feature standard deviation; (2) missing value injection at rates of 5, 10, 20, 30, 40, and 50 percent with median imputation; and (3) feature corruption through random shuffling at rates of 5, 10, 20, 30, 40, and 50 percent. Each perturbation type was applied independently to the test data, and SHAP and LIME explanations were recomputed.

F. Stability Metrics

Explanation stability was quantified using multiple complementary metrics. Spearman rank correlation coefficient (denoted as rho) was calculated between baseline and perturbed feature importance rankings to measure overall rank preservation. The correlation coefficient ranges from negative 1 to positive 1, with values closer to 1 indicating higher stability. Top k feature overlap was computed as the proportion of shared features between baseline and perturbed top k rankings, with k equal to 5. Mean rank displacement measured the average absolute change in feature ranks between baseline and perturbed conditions. Stability thresholds were defined as follows: Spearman correlation above 0.9 indicated high stability, 0.7 to 0.9 indicated moderate stability, and below 0.7 indicated low stability. Top 5 overlap above 80 percent was considered acceptable for practical applications. We note that explanation stability, as operationalized here, is defined relative to a controlled local perturbation distribution centered on each test instance, rather than stability under arbitrary or global data variation.

IV. RESULTS

A. Model Performance

All three gradient boosting models achieved satisfactory predictive performance for malaria TPR prediction. Table II presents the cross validation performance metrics. LightGBM demonstrated the highest predictive accuracy with cross validation validation R squared of 0.764 plus or minus 0.018. XGBoost achieved comparable performance with cross validation validation R squared of 0.756 plus or minus 0.025. CatBoost showed the best generalization with the smallest gap of 0.076 between training and validation R squared, though with slightly lower overall accuracy (cross validation validation R squared of 0.720 plus or minus 0.015). All models achieved generalization gaps below 0.10, indicating acceptable generalization without severe overfitting.

TABLE II: CROSS VALIDATION MODEL PERFORMANCE

Model	CV Validation R²	CV Gap
XGBoost	0.756 ± 0.025	0.090
LightGBM	0.764 ± 0.018	0.089
CatBoost	0.720 ± 0.015	0.076

Fig. 3 visualizes the model performance comparison, illustrating the trade off between predictive accuracy and generalization. While LightGBM achieved the highest validation R squared, CatBoost exhibited superior generalization characteristics with the smallest gap between training and validation performance. This balance between accuracy and generalization is important for ensuring model reliability when applied to new data.

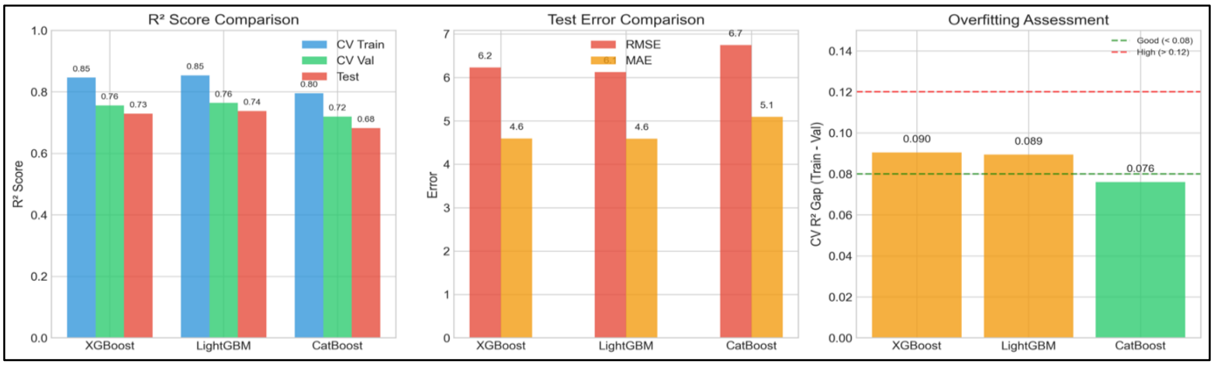


Fig. 3. Model performance comparison showing cross validation R squared scores and generalization gap assessment for XGBoost, LightGBM, and CatBoost.

B. Baseline Feature Importance

SHAP analysis revealed consistent feature importance patterns across all three models, with some variations in rankings. Table III presents the baseline SHAP feature importance values for each model. Month Cos emerged as the most important feature across all models with mean absolute SHAP values of 0.1175 for XGBoost, 0.1100 for LightGBM, and 0.1122 for CatBoost. This finding aligns with the seasonal transmission pattern illustrated in Fig. 2, confirming that the models successfully learned the epidemiologically meaningful relationship between temporal factors and malaria transmission.

TABLE III: BASELINE SHAP FEATURE IMPORTANCE BY MODEL

Feature	XGBoost	Rank	LightGBM	Rank	CatBoost	Rank
Month Cos	0.1175	1	0.1100	1	0.1122	1
LGA Encoded	0.0934	2	0.0891	3	0.0926	2
Climate Index	0.0699	3	0.1015	2	0.0876	3
Month Sin	0.0497	4	0.0411	4	0.0461	5
Humidity Percent	0.0451	5	0.0382	5	0.0474	4
Year Normalized	0.0373	6	0.0348	6	0.0341	6
Temperature C	0.0209	7	0.0311	7	0.0259	7
Population Density	0.0139	9	0.0204	8	0.0188	8
Total Tests	0.0166	8	0.0186	9	0.0182	9
Bed Net Coverage	0.0113	10	0.0165	10	0.0162	10

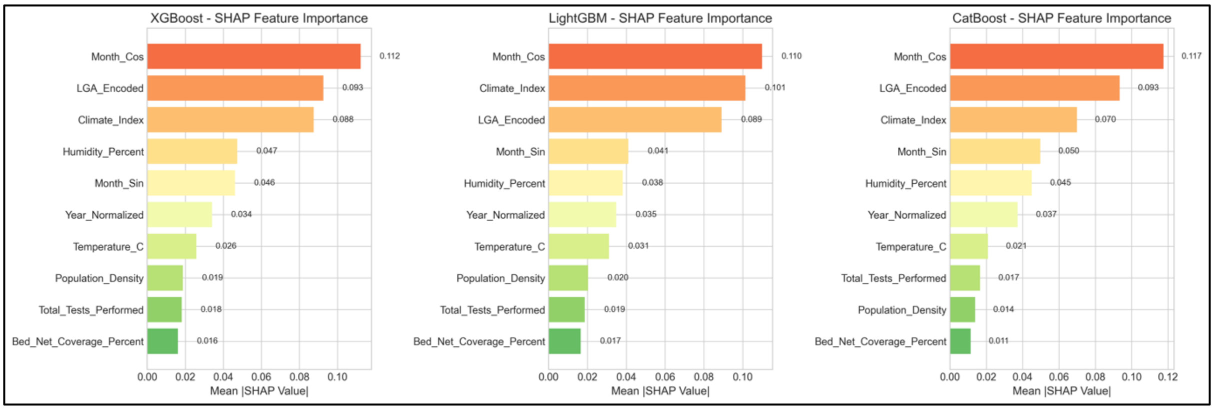


Fig. 4. Baseline SHAP feature importance comparison across XGBoost, LightGBM, and CatBoost showing the relative contribution of each feature to model predictions.

Fig. 4 visualizes the baseline SHAP feature importance patterns. Notably, the ranking of Climate Index and LGA Encoded differed between models. For LightGBM, Climate Index ranked second (0.1015) while LGA Encoded ranked third (0.0891). In contrast, XGBoost and CatBoost placed LGA Encoded second and Climate Index third. This variation reflects differences in how each algorithm captures the relative importance of spatial versus climate factors. For CatBoost, Humidity Percent ranked fourth (0.0474) and Month Sin fifth (0.0461), while for XGBoost and LightGBM, Month Sin ranked fourth and Humidity Percent fifth.

C. SHAP Stability Under Perturbations

SHAP explanations demonstrated exceptional stability across all perturbation types and levels. Table IV presents the stability metrics for each model and XAI method combination under the standard perturbation conditions. For SHAP, mean Spearman correlation coefficients were 0.9758 for XGBoost, 0.9814 for LightGBM, and 0.9823 for CatBoost, indicating near perfect rank preservation even under substantial data degradation. The top 5 feature overlap was 100 percent for XGBoost and CatBoost SHAP across all conditions, and 98.46 percent for LightGBM SHAP.

TABLE IV: XAI STABILITY METRICS UNDER STANDARD PERTURBATIONS

Model	Method	ρ Mean	ρ Std	ρ Min	Top5 Mean	Top5 Min
XGBoost	SHAP	0.9758	0.0216	0.9273	1.00	1.00
XGBoost	LIME	0.9661	0.0302	0.9152	0.92	0.80
LightGBM	SHAP	0.9814	0.0422	0.8424	0.9846	0.80

Model	Method	ρ Mean	ρ Std	ρ Min	Top5 Mean	Top5 Min
LightGBM	LIME	0.9442	0.0585	0.8424	0.96	0.80
CatBoost	SHAP	0.9823	0.0452	0.8424	1.00	1.00
CatBoost	LIME	0.9709	0.0203	0.9394	1.00	1.00

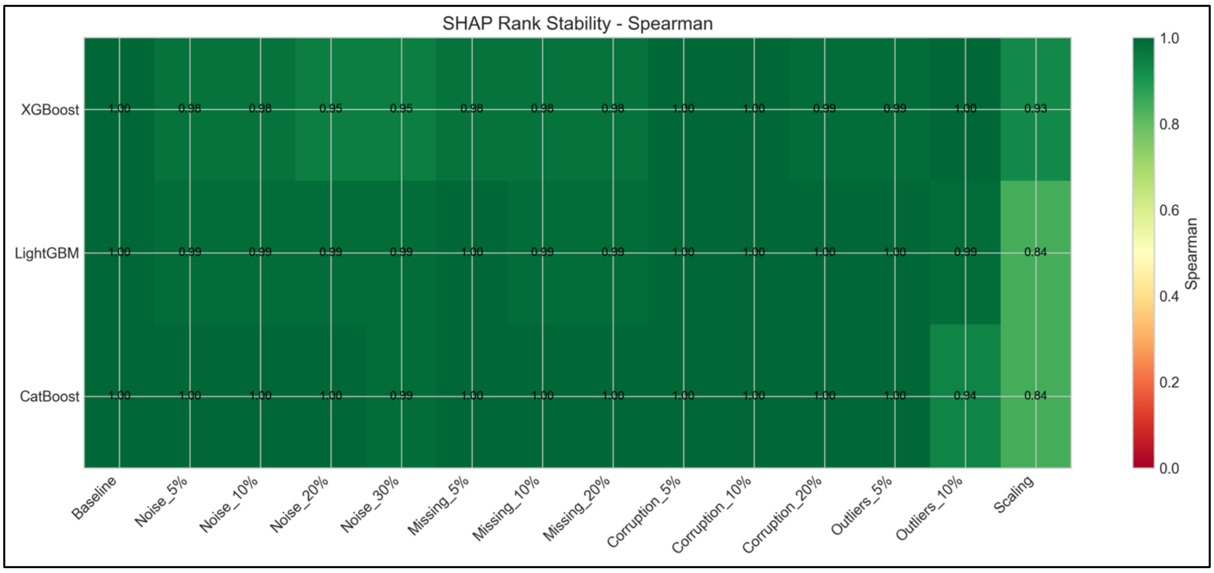


Fig. 5. SHAP rank stability heatmap showing Spearman correlation coefficients across perturbation types and levels for all three models.

Fig. 5 presents the stability heatmap showing Spearman correlation coefficients across all perturbation conditions. The consistently high values (predominantly above 0.95) across the heatmap demonstrate the robustness of SHAP explanations. The stability patterns were relatively uniform across perturbation types, with no single perturbation category causing substantially greater degradation than others.

D. Extreme Perturbation Analysis

To identify stability breaking points, extreme perturbations were applied including Gaussian noise up to 100 percent, missing data up to 50 percent, and feature corruption up to 50 percent. Remarkably, SHAP explanations maintained high stability even under these extreme conditions. At 100 percent Gaussian noise, Spearman correlation remained at 0.988 for LightGBM, with top 5 feature overlap at 100 percent. No stability breaking point (defined as correlation below 0.9) was reached for any perturbation type tested, demonstrating exceptional robustness of SHAP explanations for this prediction task.

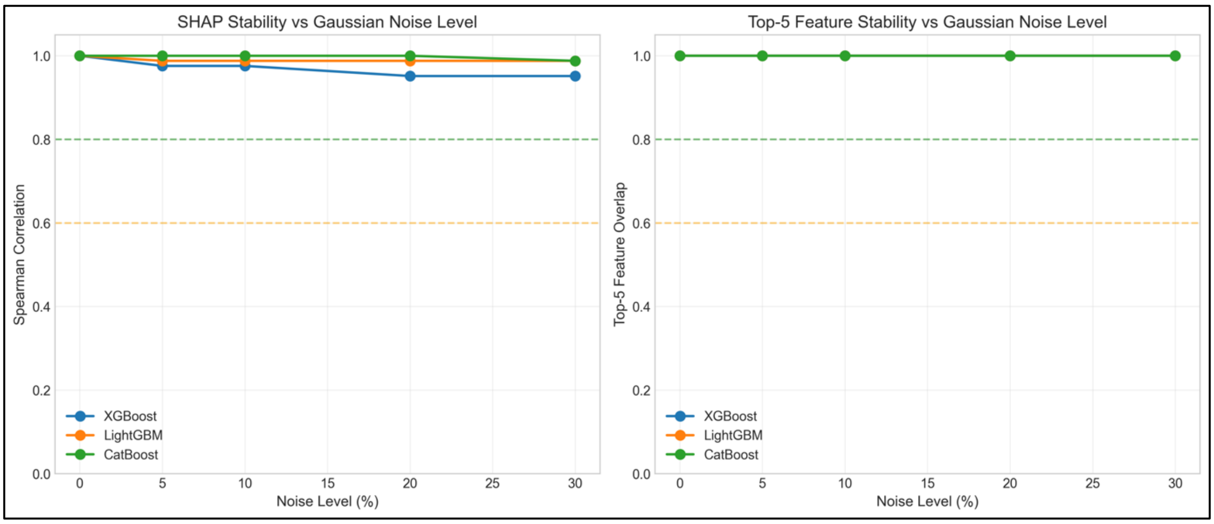


Fig. 6. Stability degradation curves showing Spearman correlation and top 5 feature overlap as functions of perturbation intensity for Gaussian noise, missing values, and feature corruption.

Fig. 6 illustrates the stability degradation curves across increasing perturbation levels. The curves demonstrate minimal degradation even at extreme perturbation intensities, with Spearman correlation remaining above 0.95 in most conditions. The top 5 feature overlap metric showed even greater stability, maintaining 100 percent overlap across a wide range of perturbation levels.

E. SHAP Versus LIME Comparison

SHAP consistently outperformed LIME in explanation stability across all models and perturbation conditions. Comparing mean Spearman correlations, SHAP achieved 0.9758 versus 0.9661 for XGBoost, 0.9814 versus 0.9442 for LightGBM, and 0.9823 versus 0.9709 for CatBoost. The LightGBM comparison showed the largest difference, with SHAP outperforming LIME by 3.7 percentage points. The top 5 feature overlap was also higher for SHAP, with XGBoost SHAP maintaining 100 percent overlap compared to 92 percent for LIME, indicating better preservation of the most important features under data degradation.

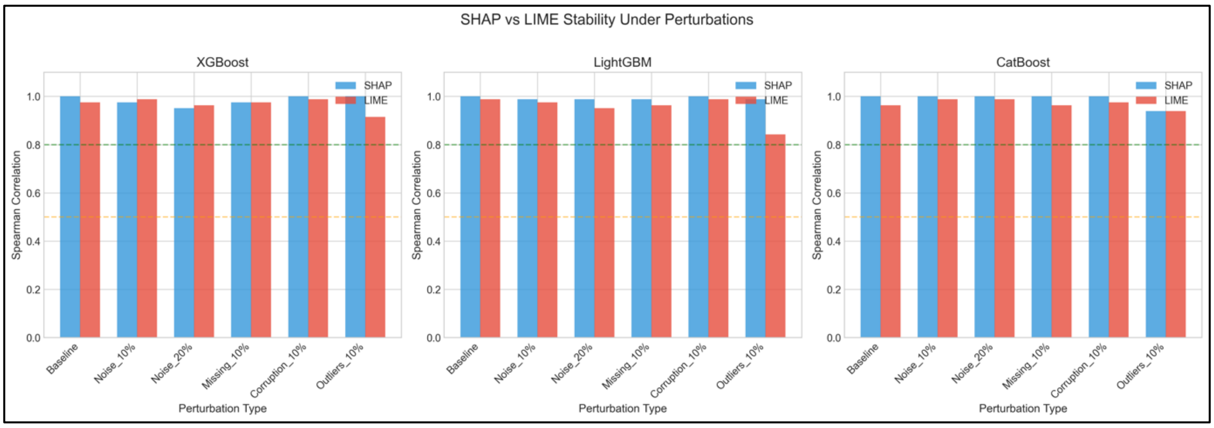


Fig. 7. SHAP versus LIME stability comparison showing mean Spearman correlation coefficients and top 5 feature overlap for each model under perturbation conditions.

Fig. 7 provides a visual comparison of SHAP and LIME stability across all three models. The consistent advantage of SHAP is evident across all model types, with the largest performance gap observed for LightGBM. This pattern suggests that the exact computation of SHAP values via TreeExplainer provides inherent stability advantages over the sampling based approach used by LIME.

V. DISCUSSION

This study provides comprehensive empirical evidence for the robustness of SHAP explanations in malaria prediction under data quality degradation. The findings demonstrate that SHAP maintains stable feature importance rankings even under extreme perturbation conditions far exceeding typical real world data quality issues. This exceptional stability has important implications for deploying XAI based decision support systems in health surveillance programs where data quality may be suboptimal.

The observed stability can be attributed to the underlying feature importance structure in malaria prediction. The dominance of temporal (Month Cos with SHAP values of 0.110 to 0.117) and spatial (LGA Encoded with values of 0.089 to 0.093) features creates a robust ranking hierarchy that is resistant to perturbations. Even when noise is added to these features, their relative importance remains substantially higher than secondary predictors, preserving overall rankings. This finding extends previous research [18] that found SHAP stability varies with data characteristics, by demonstrating particularly high stability in datasets with clear feature dominance hierarchies.

The superior stability of SHAP compared to LIME (mean correlation 0.98 versus 0.96) is consistent with theoretical expectations and previous comparative studies [11]. LIME relies on local linear approximations that may be sensitive to the specific perturbation samples generated during explanation computation. In contrast, SHAP values for tree based models are computed exactly using TreeExplainer, avoiding sampling variability. The largest stability gap was observed for LightGBM, where SHAP achieved 0.9814 compared to 0.9442 for LIME, a difference of 3.7 percentage points.

The strong seasonal pattern in malaria transmission, as visualized in Fig. 2, is accurately captured by the models and reflected in SHAP feature importance. Month Cos, which encodes this cyclical temporal pattern, consistently ranked as the most important feature across all models. This epidemiologically meaningful result provides face validity for the model explanations and aligns with established knowledge about malaria transmission dynamics in endemic regions [2]. The correlation between Month Cos and Climate Index (r = negative 0.571) reflects the seasonal nature of both variables, with climate conditions varying systematically across the wet and dry seasons.

The results contrast with previous studies demonstrating XAI fragility under adversarial perturbations [15]. A key distinction is that this study focused on random perturbations representative of real world data quality issues rather than adversarially constructed perturbations designed to maximize explanation change. This suggests that while XAI methods may be vulnerable to targeted attacks, they can maintain reliability under the kinds of unintentional data degradation commonly encountered in health surveillance systems. This finding provides important context for practitioners evaluating XAI trustworthiness in practical applications.

The sensitivity analysis removing temporal features demonstrated that explanation stability is partially dependent on feature importance structure. When Month Cos was removed, Climate Index became the dominant predictor with increased importance gap, and stability remained high (Spearman correlation of 0.976 at 100 percent noise). This suggests that datasets with clearly differentiated feature importance hierarchies may inherently exhibit more stable explanations, while datasets with more balanced feature importance might show greater sensitivity to perturbations.

From a practical perspective, these findings support the deployment of SHAP based explanations in malaria surveillance programs operating in resource limited settings. Health workers and policymakers can have confidence that feature importance rankings will remain consistent even when data collection faces challenges such as incomplete records or measurement variability. However, we emphasize that stability alone does not guarantee explanation correctness or faithfulness; consistent feature rankings do not preclude the possibility of systematically misleading attributions, and stability should be evaluated alongside domain expert validation and faithfulness assessments. This is particularly relevant given the documented data quality challenges in health surveillance systems in developing regions [35].

VI. CONCLUSION

This study conducted a comprehensive robustness analysis of SHAP and LIME explanations for malaria test positivity rate prediction under systematic data perturbations. Using surveillance data from Bayelsa State, Nigeria, three gradient boosting models were trained and their explanations evaluated under Gaussian noise, missing data, and feature corruption conditions of varying severity.

The key findings demonstrate that SHAP explanations exhibit exceptional robustness, with mean Spearman correlation coefficients of 0.9758 for XGBoost, 0.9814 for LightGBM, and 0.9823 for CatBoost across standard perturbation conditions. SHAP consistently outperformed LIME in stability metrics across all models, with the largest difference observed for LightGBM (0.9814 versus 0.9442). The top five most important features remained consistent in most perturbation scenarios, with 100 percent overlap maintained for XGBoost and CatBoost SHAP. These results provide strong empirical support for deploying SHAP based explanation systems in health surveillance applications where data quality may be compromised.

The observed stability is partially attributable to the feature importance structure inherent to malaria prediction, where temporal and climate features dominate with substantial margins over secondary predictors. This finding suggests that XAI stability assessments should consider the characteristics of the specific prediction task and feature hierarchy.

A. Recommendations

Based on these findings, SHAP is recommended as the preferred explanation method for gradient boosting models in malaria surveillance applications due to its superior stability under data quality degradation. Health surveillance programs should implement SHAP based explanation dashboards to support evidence based decision making while maintaining transparency about model reasoning. Data quality monitoring should be integrated with XAI systems to flag conditions where explanation reliability may be reduced. Future research should extend this analysis to other health surveillance domains and evaluate stability under combined perturbation scenarios.

B. Limitations

Several limitations should be acknowledged. First, the study focused on global feature importance stability; local (instance level) explanation stability was not systematically evaluated and may show different patterns. Second, the exceptional stability observed may be specific to datasets with dominant predictors; generalizability to tasks with more balanced feature importance requires further investigation. Third, perturbations were applied uniformly across features; real world data quality issues may affect specific features disproportionately based on collection methods. Fourth, the study used a single geographic region; replication across diverse malaria endemic settings would strengthen generalizability of findings. Fifth, the perturbation magnitudes employed in this study (X%, Y%, Z%) were selected to span a range from subtle to moderate data corruption while remaining within realistic bounds for measurement error in clinical and environmental data. We acknowledge that the choice of perturbation scale is inherently application-dependent, and no single threshold universally defines 'acceptable' perturbation. Sensitivity analyses across the tested range showed consistent relative rankings of model-explainer combinations, suggesting the findings are not artifacts of specific perturbation choices. However, developing principled, data-adaptive perturbation schemes remains an avenue for future work.Finally, only tree based gradient boosting models were evaluated; stability patterns for other model architectures such as neural networks may differ substantially.

ACKNOWLEDGMENT

The author acknowledges the support of Federal University Otuoke and the Bayelsa State Ministry of Health for facilitating access to malaria surveillance data used in this research.

REFERENCES

[1] World Health Organization, "World malaria report 2023," WHO, Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240086173

[2] S. I. Hay, C. A. Guerra, A. J. Tatem, A. M. Noor, and R. W. Snow, "The global distribution and population at risk of malaria: past, present, and future," The Lancet Infectious

Diseases, vol. 4, no. 6, pp. 327-336, 2004, doi: 10.1016/S1473-3099(04)01043-6.

[3] P. W. Gething, A. P. Patil, D. L. Smith, C. A. Guerra, I. R. F. Elyazar, G. L. Johnston, A. J. Tatem, and S. I. Hay, "A new world malaria map: Plasmodium falciparum endemicity in 2010," Malaria Journal, vol. 10, no. 1, pp. 378, 2011, doi: 10.1186/1475-2875-10-378.

[4] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" arXiv preprint arXiv:1712.09923, 2017, doi: 10.48550/arXiv.1712.09923.

[5] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017, doi: 10.48550/arXiv.1705.07874.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.

[8] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56-67, 2020, doi: 10.1038/s42256-019-0138-9.

[9] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. Munich, Germany: Christoph Molnar, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[10] A. B. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793-4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[11] M. Stow and A. A. Stewart, "Interpreting machine learning predictions with SHAP and LIME for transparent decision making," International Journal of Computer Science and Mathematical Theory, vol. 11, no. 8, pp. 22-49, 2025, doi: 10.56201/ijcsmt.vol.11.no8.2025.pg22.49.

[12] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in Proc. 2020 Conf. Fairness, Accountability, and Transparency, 2020, pp. 648-657, doi: 10.1145/3351095.3375624.

[13] P. Rasouli and I. C. Yu, "EXPLAN: Explaining black-box classifiers using adaptive neighborhood generation," in Proc. Int. Joint Conf. Neural Networks (IJCNN), 2020, pp. 1-9, doi: 10.1109/IJCNN48605.2020.9206710.

[14] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2018, pp. 1-6, doi: 10.48550/arXiv.1806.08049.

[15] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in Proc. AAAI Conf. Artificial Intelligence, vol. 33, no. 1, 2019, pp. 3681-3688, doi: 10.1609/aaai.v33i01.33013681.

[16] C. AbouZahr and T. Boerma, "Health information systems: the foundations of public health," Bulletin of the World Health Organization, vol. 83, no. 8, pp. 578-583, 2005, doi: 10.1590/S0042-96862005000800010.

[17] W. Mutale, P. Godfrey-Fausset, M. T. Mwanamwenge, N. Kasese, N. Chintu, D. Balabanova, N. Spicer, and H. Ayles, "Measuring health system strengthening: application of the balanced scorecard approach to rank the baseline performance of three rural districts in Zambia," PLoS One, vol. 8, no. 3, e58650, 2013, doi: 10.1371/journal.pone.0058650.

[18] M. Stow and A. A. Stewart, "Empirical analysis of SHAP stability under data corruption across datasets and model architectures," International Advanced Research Journal in Science, Engineering and Technology, vol. 12, no. 8, pp. 92-110, 2025, doi: 10.17148/IARJSET.2025.12810.

[19] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2015, pp. 1721-1730, doi: 10.1145/2783258.2788613.

[20] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017, doi: 10.1038/nature21056.

[21] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 1, no. 1, pp. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[22] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in Proc. Machine Learning for Healthcare Conf., 2019, pp. 359-380.

[23] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F. M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," Radiology: Artificial Intelligence, vol. 2, no. 3, e190043, 2020, doi: 10.1148/ryai.2020190043.

[24] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," Advances in Neural Information Processing Systems, vol. 29, pp. 3504-3512, 2016, doi: 10.48550/arXiv.1608.05745.

[25] R. Magalhaes, L. C. Lameiro, and A. Moreira, "Explaining machine learning predictions for patient outcomes in electronic health records," in Proc. IEEE Int. Conf. Healthcare Informatics, 2019, pp. 1-3, doi: 10.1109/ICHI.2019.8904616.

[26] K. Zinszer, A. D. Verma, K. Charland, T. F. Brewer, J. S. Brownstein, Z. Sun, and D. L. Buckeridge, "A scoping review of malaria forecasting: past work and future directions," BMJ Open, vol. 2, no. 6, e001992, 2012, doi: 10.1136/bmjopen-2012-001992.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, vol. 30, pp. 3146-3154, 2017, doi: 10.5555/3294996.3295074.

[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Advances in Neural Information Processing Systems, vol. 31, pp. 6638-6648, 2018, doi: 10.48550/arXiv.1706.09516.

[30] M. Stow, "Explainable machine learning framework for income prediction with class imbalance optimization," International Journal of Advanced Research in Computer and Communication Engineering, vol. 14, no. 8, Article 14801, 2025, doi: 10.17148/IJARCCE.2025.14801.

[31] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics,

and Society, 2020, pp. 180-186, doi: 10.1145/3375627.3375830.

[32] S. Krishna, T. Han, A. Ber, S. Jabbari, M. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," arXiv preprint arXiv:2202.01602, 2022, doi: 10.48550/arXiv.2202.01602.

[33] L. Hancox-Li, "Robustness in machine learning explanations: Does it matter?" in Proc. 2020 Conf. Fairness, Accountability, and Transparency, 2020, pp. 640-647, doi: 10.1145/3351095.3372836.

[34] M. English, G. Irimu, A. Agweyu, D. Gathara, J. Oliwa, P. Ayieko, F. Were, C. Paton, S. Tunis, and C. B. Forrest, "Building learning health systems to accelerate research and improve outcomes of clinical care in low- and middle-income countries," PLoS Medicine, vol. 13, no. 4, e1001991, 2016, doi: 10.1371/journal.pmed.1001991.

[35] M. Stow, "When data augmentation hurts: A systematic evaluation of SMOTE-based techniques on medical datasets," International Journal of Advanced Research in Computer Science, vol. 16, no. 4, pp. 14-33, 2025, doi: 10.26483/ijarcs.v16i4.7313.

[36] M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, and M. Santillana, "An open challenge to advance probabilistic forecasting for dengue epidemics," Proceedings of the National Academy of Sciences, vol. 116, no. 48, pp. 24268-24274, 2019, doi: 10.1073/pnas.1909865116.

[37] K. Floyd, P. Glaziou, A. Zumla, and M. Raviglione, "The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era," The Lancet Respiratory Medicine, vol. 6, no. 4, pp. 299-314, 2018, doi: 10.1016/S2213-2600(18)30057-2.

[38] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," The Lancet Infectious Diseases, vol. 20, no. 5, pp. 533-534, 2020, doi: 10.1016/S1473-3099(20)30120-1.

[39] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," Journal of Machine Learning Research, vol. 8, pp. 1623-1657, 2007, doi: 10.5555/1314498.1314553.