ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

Machine Learning for Predicting Diagnostic Test Discordance in Malaria Surveillance: A Gradient Boosting Approach with SHAP Interpretation

May Stow[1]

[1]Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State, Nigeria.

[1]Orcid ID: https://orcid.org/0009-0006-8653-8363)

Publication Date: 2026/01/30

Abstract: Malaria remains a critical public health challenge in Nigeria, where accurate diagnosis is essential for effective disease management and resource allocation. Discordance between rapid diagnostic tests (RDTs) and microscopy poses significant challenges for malaria surveillance programs, potentially leading to misdiagnosis and inappropriate treatment decisions. This study aimed to develop a machine learning model for predicting diagnostic test discordance between RDT and microscopy in malaria surveillance data from Bayelsa State, Nigeria. A dataset comprising 2,100 monthly observations from eight Local Government Areas spanning January 2019 to December 2024 was analyzed. The methodology incorporated Bland Altman agreement analysis, feature engineering with climate and health system variables, and gradient boosting classification with class weight balancing to address data imbalance. Model interpretation was achieved through SHapley Additive exPlanations (SHAP) analysis. The Bland Altman analysis revealed a mean difference of negative 2.33 percentage points between RDT and microscopy, with limits of agreement spanning negative 19.28 to positive 14.62 percentage points. The LightGBM classifier achieved an area under the receiver operating characteristic curve of 0.901, with precision of 0.67, recall of 0.74, and F1 score of 0.703. SHAP analysis identified rainfall, climate index, geographic location, and humidity as the most influential predictors of diagnostic discordance. This study contributes an interpretable machine learning framework for identifying conditions associated with diagnostic disagreement, potentially informing quality assurance protocols and targeted interventions in malaria endemic regions.

Keywords: *Malaria Diagnosis, Diagnostic Discordance, Machine Learning, SHAP, Rapid Diagnostic Test, Microscopy.*

How to Cite: May Stow (2026) Machine Learning for Predicting Diagnostic Test Discordance in Malaria Surveillance: A Gradient Boosting Approach with SHAP Interpretation. *International Journal of Innovative Science and Research Technology,* 11(1), 2235-2247. https://doi.org/10.38124/ijisrt/26jan131

I. INTRODUCTION

Malaria continues to represent one of the most significant infectious disease burdens globally, with the World Health Organization estimating 249 million cases and 608,000 deaths in 2022 alone [1]. The African continent bears a disproportionate share of this burden, accounting for approximately 94 percent of global malaria cases and 95 percent of malaria related deaths [2]. Within this context, Nigeria stands as the most affected nation, contributing roughly 27 percent of global malaria cases and 31 percent of global malaria deaths, making it the epicenter of the worldwide malaria epidemic [3]. The implications of this burden extend beyond immediate health consequences to encompass substantial socioeconomic impacts, including reduced productivity, increased healthcare expenditure, and hindered economic development across affected regions [4].

Accurate and timely diagnosis forms the cornerstone of effective malaria management and control strategies. The World Health Organization recommends parasitological confirmation of all suspected malaria cases before treatment, primarily through microscopy or rapid diagnostic tests (RDTs) [5]. Microscopy, involving the examination of Giemsa stained blood smears, has traditionally served as the gold standard for malaria diagnosis due to its ability to detect and quantify parasites while identifying species [6]. However, microscopy requires trained personnel, functional equipment, and consistent quality assurance, making it challenging to implement in resource limited settings. Rapid diagnostic tests emerged as an alternative diagnostic approach, offering ease of use, rapid results, and suitability for deployment in remote areas lacking laboratory infrastructure [7].

Despite the widespread adoption of both diagnostic methods, significant discordance between RDT and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

microscopy results has been documented across various epidemiological settings [8]. Studies conducted in Nigeria have reported varying levels of agreement between these diagnostic approaches, with sensitivity and specificity of RDTs ranging considerably depending on factors such as parasite density, test brand, storage conditions, and operator expertise [9]. Analysis of data from the 2015 Nigeria Malaria Indicator Survey found that while significant agreement existed between RDT and microscopy outcomes, the discriminatory accuracy of RDT was weak, with positive predictive values particularly low in certain populations [10]. Similarly, RDT accuracy below 70 percent compared to microscopy has been reported in Nigerian community settings, raising concerns about sole reliance on RDTs for malaria diagnosis [11].

The consequences of diagnostic discordance extend beyond individual patient management to affect surveillance data quality, intervention planning, and resource allocation decisions. When RDT and microscopy yield conflicting results, healthcare workers face uncertainty regarding appropriate treatment decisions, potentially leading to either unnecessary antimalarial treatment or missed malaria cases [12]. At the population level, systematic differences between diagnostic methods can distort estimates of malaria burden, complicating efforts to monitor transmission trends and evaluate intervention effectiveness [13]. Understanding the factors that contribute to diagnostic discordance is therefore essential for improving malaria surveillance systems and ensuring accurate disease monitoring.

Machine learning approaches have demonstrated considerable promise in addressing various challenges in malaria research, including case prediction, outbreak forecasting, and risk mapping [14]. Gradient boosting algorithms have been applied to predict malaria incidence using climate variability across endemic African countries, demonstrating the potential of ensemble learning methods for malaria related prediction tasks [15]. Deep learning architectures, including Long Short Term Memory (LSTM) networks, have been employed for temporal forecasting of malaria cases, capturing complex seasonal and climatic patterns influencing transmission dynamics [16]. An interpretable early warning system for malaria outbreaks in Bayelsa State using deep learning and climate data has demonstrated the feasibility of predictive modeling for malaria surveillance in the Niger Delta region [17]. However, the application of machine learning specifically to predict diagnostic test discordance remains largely unexplored, representing a significant gap in the literature.

Furthermore, while machine learning models often achieve high predictive accuracy, their adoption in healthcare settings requires interpretability to enable clinical understanding and trust [18]. Explainable artificial intelligence (XAI) techniques, particularly SHAP (SHapley Additive exPlanations), provide model agnostic approaches for understanding feature contributions to predictions [19]. Empirical analysis of SHAP stability under data corruption across datasets and model architectures has provided methodological guidance for reliable interpretation of machine learning predictions [20]. Additionally, the integration of SHAP and LIME for transparent decision making has demonstrated practical approaches for interpreting complex models [21]. The integration of SHAP analysis with machine learning models offers opportunities to identify which environmental, temporal, and health system factors most strongly influence diagnostic agreement, potentially informing targeted quality improvement strategies.

This study addresses the identified gap by developing and evaluating a machine learning framework for predicting diagnostic test discordance in malaria surveillance. The research focuses on Bayelsa State, located in the Niger Delta region of Nigeria, which experiences year round malaria transmission due to its tropical climate and extensive water bodies that support mosquito breeding [22]. The specific objectives of this study are threefold: first, to quantify the level of agreement between RDT and microscopy derived test positivity rates using Bland Altman analysis; second, to develop a gradient boosting classification model capable of predicting high discordance events; and third, to identify the key predictors of diagnostic discordance through SHAP based interpretation.

The contributions of this research include: (1) the first application of machine learning to predict diagnostic test discordance in malaria surveillance at the subnational level in Nigeria; (2) a comprehensive agreement analysis between RDT and microscopy using established statistical methods; (3) an interpretable prediction framework that identifies actionable factors associated with diagnostic disagreement; and (4) insights that may inform quality assurance protocols and resource allocation for malaria diagnostic services. The findings have potential implications for improving surveillance data quality and supporting evidence based decision making in malaria control programs.

II. LITERATURE REVIEW

- *Malaria Diagnostic Methods and Performance Evaluation*

The evaluation of malaria diagnostic performance has generated substantial research attention, particularly in endemic African settings where accurate diagnosis directly impacts treatment outcomes and surveillance quality. Microscopy examination of Giemsa stained blood films has historically served as the reference standard for malaria diagnosis, enabling parasite detection, species identification, and quantification of parasitemia [23]. However, the reliability of microscopy depends heavily on technician expertise, slide preparation quality, and equipment maintenance, factors that vary considerably across healthcare facilities [24]. Studies examining microscopy performance under field conditions have documented significant inter observer variability, with concordance rates between expert and routine microscopists often falling below optimal levels [25].

Rapid diagnostic tests based on immunochromatographic detection of malaria antigens,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

particularly histidine rich protein 2 (HRP2), have been extensively evaluated against microscopy across diverse epidemiological contexts. Assessment of RDT performance in Nigerian communities found sensitivity of 94.3 percent but specificity of only 41.6 percent compared to microscopy, indicating substantial false positive rates [26]. The study highlighted that RDT performance varied with parasite density and transmission intensity, suggesting that a single performance benchmark may not apply across all settings. Evaluation of nested PCR, microscopy, and RDT for falciparum malaria detection in southwestern Nigeria reported near perfect agreement between microscopy and PCR but lower concordance for RDT based diagnosis [27].

The phenomenon of diagnostic discordance has been examined from multiple perspectives. Investigation of factors affecting RDT and microscopy agreement in Ogun State, Nigeria identified parasite density as a critical determinant, with agreement declining substantially at lower parasitemia levels [28]. Comparison of RDT performance across different endemicity levels found that accuracy metrics varied significantly by transmission setting, with implications for national diagnostic policies [29]. These studies collectively establish that diagnostic agreement is not uniform but rather influenced by biological, environmental, and operational factors that warrant systematic investigation.

- *Machine Learning Applications in Malaria Research*

Machine learning methodologies have been increasingly applied to malaria related prediction problems, demonstrating advantages over traditional statistical approaches for capturing complex nonlinear relationships. Extreme gradient boosting (XGBoost) has been employed to predict malaria incidence across six endemic African countries using climate variables, achieving strong predictive performance and highlighting the influence of temperature, rainfall, and humidity on transmission patterns [15]. The study demonstrated that ensemble tree based methods could effectively model the relationship between environmental conditions and malaria burden at national scales.

Deep learning approaches have been explored for temporal prediction of malaria cases. LSTM networks have been applied to forecast malaria incidence in the Brazilian Amazon, finding that recurrent architectures captured seasonal patterns more effectively than traditional autoregressive models [16]. Similarly, research employing gated recurrent units achieved comparable performance while requiring less computational resources for training [30]. These temporal modeling studies emphasize the importance of accounting for seasonality and lagged relationships when predicting malaria related outcomes.

Clinical prediction models using machine learning have also been developed for individual level malaria diagnosis. Demographic and environmental features have been used to predict malaria positivity in Nigerian patients, employing penalized logistic regression with sequential feature selection [31]. The study achieved area under the curve of 0.83, demonstrating that patient characteristics and environmental exposure factors could inform diagnostic predictions. Multiple machine learning algorithms including random forest and support vector machines have been applied to predict malaria prevalence using demographic health survey data, finding that random forest regression achieved coefficient of determination above 0.99 for prevalence estimation [32]. An explainable machine learning framework for income prediction with class imbalance optimization demonstrated methodological approaches applicable to imbalanced healthcare datasets such as malaria surveillance data [33].

- *Explainable Artificial Intelligence in Healthcare*

The adoption of machine learning in healthcare settings necessitates interpretability to ensure clinical trust and enable actionable insights. SHAP values, based on cooperative game theory, provide theoretically grounded feature importance measures that satisfy desirable properties including local accuracy, missingness, and consistency [34]. It has been demonstrated that SHAP unifies several existing explanation methods under a single framework, enabling consistent interpretation across different model types [19]. The approach has been successfully applied in diverse healthcare contexts, from disease diagnosis to treatment response prediction.

Explainable artificial intelligence models for malaria risk prediction in Kenya have emphasized the value of interpretable predictions for informing public health interventions [35]. The study employed SHAP analysis to identify key risk factors and demonstrated how feature importance rankings could guide resource allocation decisions. Systematic evaluation of SMOTE based techniques on medical datasets provided evidence that synthetic oversampling approaches may degrade model performance in certain healthcare contexts, with implications for handling class imbalance in diagnostic prediction tasks [36]. Similarly, research applying explainable methods to infectious disease prediction has shown that understanding model reasoning enhances both clinical acceptance and practical utility of machine learning systems [37].

- *Research Gap and Study Positioning*

The reviewed literature reveals several gaps that this study addresses. First, while substantial research has examined diagnostic test performance, few studies have applied predictive modeling to understand when and why discordance occurs rather than simply quantifying its extent. Second, machine learning applications in malaria research have focused predominantly on case prediction and forecasting rather than diagnostic quality assessment. Third, the integration of explainable AI methods with diagnostic discordance prediction remains unexplored, limiting opportunities for identifying actionable factors that could inform quality improvement interventions.

This study contributes to the literature by developing an interpretable machine learning framework specifically designed to predict diagnostic test discordance in malaria surveillance. By combining gradient boosting classification with SHAP based explanation, the research provides both predictive capability and mechanistic insights regarding

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

factors associated with diagnostic disagreement. The focus on Bayelsa State, Nigeria, addresses the need for subnational evidence from high burden settings where understanding diagnostic variability has direct programmatic implications.

III. METHODOLOGY

➢ *Study Area and Data Source*

This study utilized malaria surveillance data from 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 state experiences a tropical climate characterized by high rainfall, elevated humidity, and temperatures conducive to year round malaria transmission. The Niger Delta ecology, with extensive creeks, rivers, and wetlands, supports abundant mosquito breeding habitats, contributing to the high malaria burden in the region.

The dataset comprised monthly aggregated malaria surveillance records spanning January 2019 through December 2024, totaling 2,100 observations across the eight LGAs. Data elements included test positivity rates from both microscopy and rapid diagnostic tests, climate variables (rainfall in millimeters, temperature in degrees Celsius, and relative humidity percentage), health system indicators (healthcare worker density, facility reporting rates, antimalarial stock levels), intervention coverage metrics (insecticide treated bed net coverage, indoor residual spraying coverage), and malaria case counts differentiated by severity. Table 1 presents the summary characteristics of the dataset.

Table 1 Dataset Characteristics And Summary Statistics

Characteristic	Value
Study Period	January 2019 – December 2024
Geographic Coverage	Bayelsa State, Nigeria
Total Observations	2,100
Number of LGAs	8
Observations per LGA	72 – 478 (median: 249)
TPR Microscopy – Mean (SD)	28.5% (12.0%)
TPR Microscopy – Range	8.3% – 72.2%
TPR RDT – Mean (SD)	26.2% (10.4%)
TPR RDT – Range	8.2% – 60.9%
Absolute Discordance – Mean (SD)	6.8% (5.8%)
Absolute Discordance – Median (IQR)	5.3% (2.2% – 9.4%)
Low Discordance (≤10%)	1,619 (77.1%)
High Discordance (>10%)	481 (22.9%)

➢ *Proposed Methodology Framework*

The proposed methodology comprises five sequential phases: data preprocessing, discordance quantification and agreement analysis, feature engineering, model development with class imbalance handling, and model interpretation through SHAP analysis. Figure 1 illustrates the complete methodological framework adopted in this study.

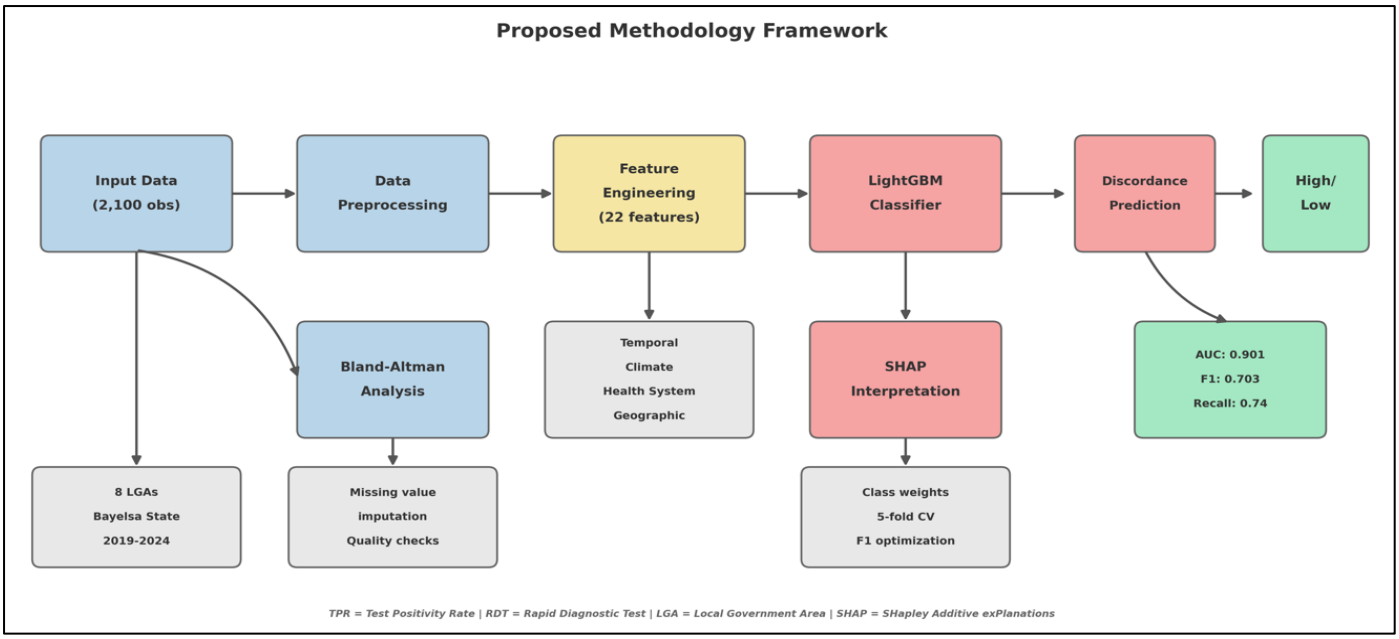

Fig 1 Proposed Methodology Architecture for Diagnostic Discordance Prediction.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

➢ *Data Preprocessing*

Data preprocessing involved several sequential steps to ensure data quality and analytical readiness. Records with non distinct LGA identifiers were removed, retaining only observations from the eight recognized LGAs of Bayelsa State. Missing values in climate and health system variables were imputed using temporal spatial interpolation, whereby missing monthly values were estimated based on the median of the same month across available years within the same LGA. The Facility column, identified as containing entirely empty values, was excluded from analysis. Following preprocessing, the final analytical dataset contained 2,100 complete observations with zero missing values.

➢ *Discordance Definition and Agreement Analysis*

Diagnostic discordance was operationally defined as the absolute difference between RDT derived and microscopy derived test positivity rates. The discordance value for each observation was calculated as:

$$\text{Absolute Discordance} = |\text{TPR\_RDT} - \text{TPR\_Microscopy}|$$

A threshold of 10 percentage points was established to classify observations into high discordance (absolute discordance greater than 10 percent) and low discordance (absolute discordance less than or equal to 10 percent) categories. This threshold was informed by clinical significance considerations, as differences exceeding 10 percentage points may influence treatment decisions and resource allocation at the facility level.

Agreement between the two diagnostic methods was assessed using Bland Altman analysis, which provides graphical and statistical evaluation of measurement agreement [38]. The mean difference (bias) and 95 percent limits of agreement were calculated as:

$$\text{Mean Difference } (\bar{d}) = \Sigma(\text{TPR\_RDT} - \text{TPR\_Microscopy}) / n$$

$$\text{Limits of Agreement} = \bar{d} \pm 1.96 \times SD$$

Additionally, Lin concordance correlation coefficient (CCC) was computed to quantify the agreement between the two measurements while accounting for both precision and accuracy [39]. The CCC ranges from negative one to positive one, with values closer to one indicating stronger agreement.

➢ *Feature Engineering*

Feature engineering was performed to create informative predictors from the raw surveillance data. Temporal features included cyclical encoding of month using sine and cosine transformations to capture seasonal patterns, binary wet season indicator (April through October), and normalized year to represent temporal trends. Climate features included a composite climate index calculated as the product of normalized rainfall and humidity to capture combined environmental favorability for mosquito breeding. Health system features were retained in their original form, including healthcare worker density, facility reporting rates, and intervention coverage percentages. Categorical LGA identifiers were encoded numerically to enable model processing. The complete feature set comprised 22 variables spanning temporal, climate, health system, and intervention domains.

➢ *Model Development*

The prediction task was formulated as binary classification, with the target variable indicating high discordance (class 1) versus low discordance (class 0). The dataset exhibited class imbalance with 77.1 percent low discordance and 22.9 percent high discordance observations. To address this imbalance, class weight balancing was employed rather than synthetic oversampling techniques, as preliminary experiments indicated that SMOTE based approaches led to substantial overfitting with cross validation to test set performance gaps exceeding 0.17. This finding aligns with recent evidence that SMOTE based techniques may degrade performance on medical datasets [36].

The data were partitioned into training (80 percent) and test (20 percent) sets using stratified random sampling to maintain class proportions. Features were standardized using z score normalization based on training set statistics to ensure consistent scaling while preventing data leakage.

LightGBM (Light Gradient Boosting Machine) was selected as the primary classifier due to its efficiency with tabular data, native support for class weights, and strong performance on imbalanced classification tasks [40]. Hyperparameter optimization was conducted using randomized search with five fold stratified cross validation, exploring the parameter space defined in Table 2. The optimization objective was F1 score to balance precision and recall given the class imbalance context.

Table 2 Hyperparameter Search Space And Optimal Values

Parameter	Search Space	Optimal Value
n_estimators	[50, 100, 150]	100
max_depth	[2, 3, 4, 5]	5
learning_rate	[0.01, 0.05, 0.1]	0.1
subsample	[0.6, 0.7, 0.8]	0.6
colsample_bytree	[0.6, 0.7, 0.8]	0.6
min_child_samples	[10, 20, 30, 50]	10
reg_alpha	[0.1, 0.5, 1.0, 2.0]	0.5
reg_lambda	[0.1, 0.5, 1.0, 2.0]	0.1
class_weight	['balanced']	balanced

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

➢ *Model Evaluation*

Model performance was evaluated using multiple metrics appropriate for imbalanced classification. The area under the receiver operating characteristic curve (AUC ROC) measured overall discriminative ability. Precision, recall (sensitivity), and F1 score assessed performance on the minority high discordance class. Specificity and balanced accuracy provided additional perspective on classification performance across both classes. The gap between cross validation and test set performance was monitored to assess generalizability and detect potential overfitting.

➢ *SHAP Analysis for Model Interpretation*

Model interpretation was performed using SHAP (SHapley Additive exPlanations) values, which provide consistent and locally accurate feature attributions based on game theoretic principles [19]. For each prediction, SHAP values quantify the contribution of each feature to the deviation from the expected output. The TreeExplainer algorithm, optimized for tree ensemble models, was employed to compute exact SHAP values efficiently. Summary plots visualizing feature importance and directional effects were generated to identify the most influential predictors of diagnostic discordance. The stability of SHAP interpretations was considered in light of recent methodological guidance on SHAP reliability [20]. Figure 2 presents the dataset summary and distribution characteristics used in the analysis.

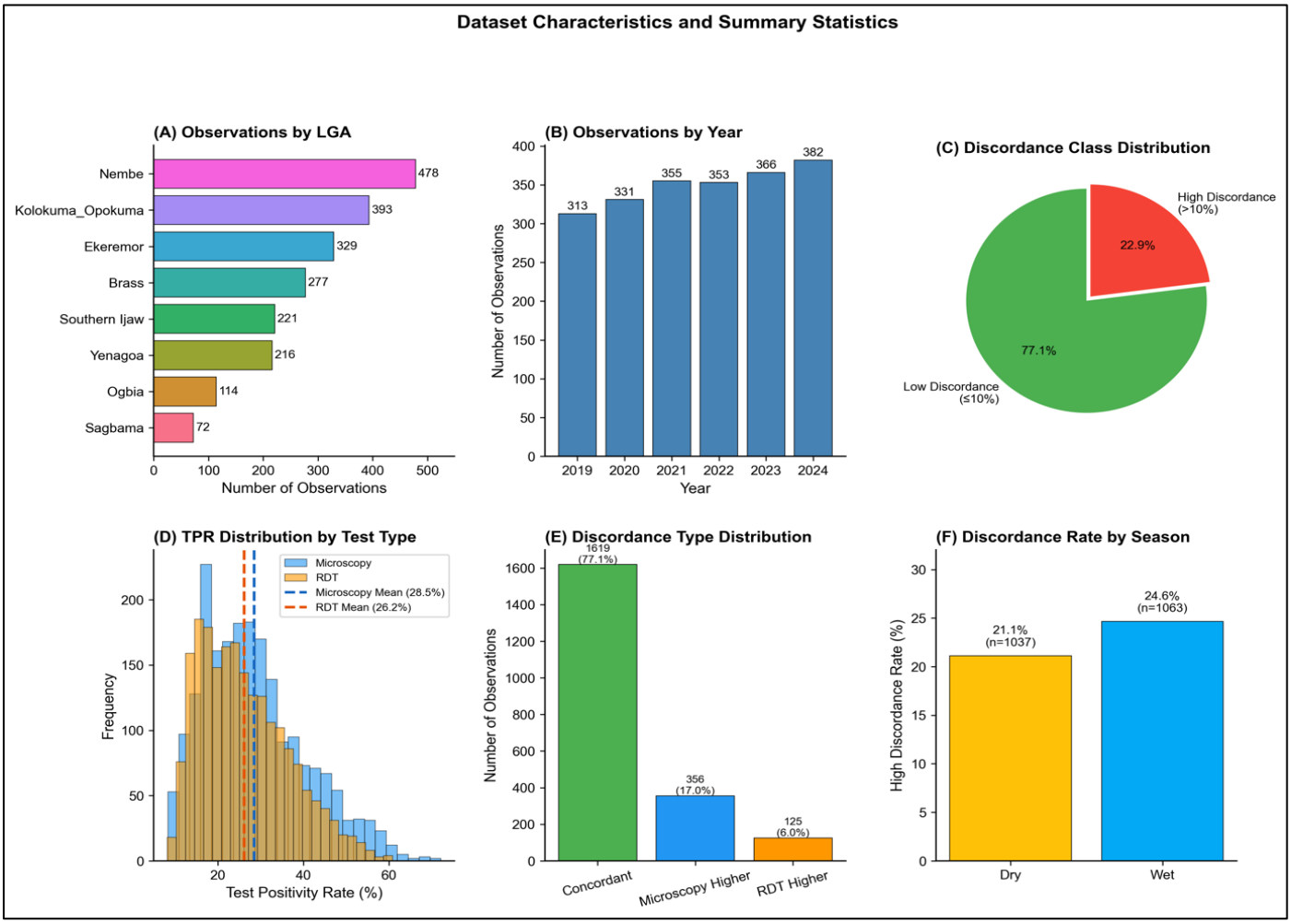

Fig 2 Dataset Characteristics and Distribution Summary.

IV. RESULTS

➢ *Bland Altman Agreement Analysis*

The Bland Altman analysis revealed systematic differences between RDT and microscopy derived test positivity rates. The mean difference was negative 2.33 percentage points (95 percent confidence interval: negative 2.72 to negative 1.94), indicating that RDT consistently yielded lower positivity rates compared to microscopy on average. The standard deviation of differences was 8.64 percentage points, resulting in 95 percent limits of agreement spanning from negative 19.28 to positive 14.62 percentage points. This wide range indicates substantial variability in the agreement between the two diagnostic methods across observations.

Of the 2,100 observations, 93.9 percent fell within the limits of agreement, with 6.1 percent exhibiting extreme discordance beyond these bounds. The Lin concordance correlation coefficient was 0.689, indicating moderate

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

agreement between the methods. The Pearson correlation coefficient of 0.711 (p less than 0.001) suggested reasonable linear association, though the substantial deviation from perfect concordance (CCC equals 1.0) indicates that the two methods cannot be considered interchangeable for surveillance purposes. Figure 3 displays the Bland Altman plot alongside the concordance scatter plot.

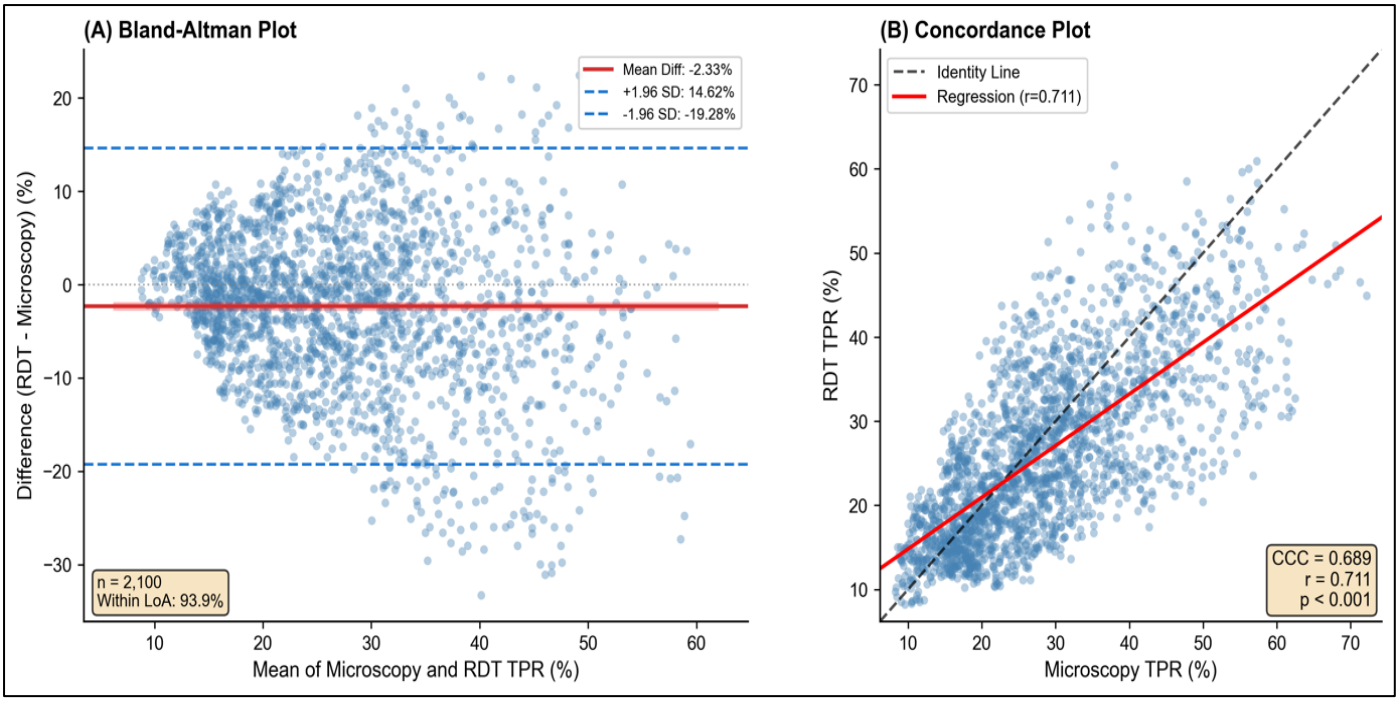

Fig 3 Bland Altman Agreement Analysis Showing (a) Difference Plot and (b) Concordance Scatter Plot.

Analysis of discordance direction revealed that when diagnostic disagreement occurred, microscopy yielded higher positivity rates nearly three times more frequently than RDT. Specifically, microscopy was higher in 356 observations (17.0 percent of total), while RDT was higher in only 125 observations (6.0 percent). The remaining 1,619 observations (77.1 percent) were classified as concordant based on the 10 percentage point threshold.

➢ *Classification Model Performance*

The LightGBM classifier with class weight balancing achieved strong discriminative performance on the held out test set. Table 3 presents the comprehensive evaluation metrics. The model attained an AUC ROC of 0.901, indicating excellent ability to distinguish between high and low discordance cases. The F1 score of 0.703 reflects reasonable balance between precision (0.67) and recall (0.74) for the high discordance class. Importantly, the model demonstrated no evidence of overfitting, with test performance slightly exceeding cross validation performance (CV F1: 0.666, Test F1: 0.703), yielding a negative CV test gap of 0.037.

Table 3 Model Performance Metrics on Test Set

Metric	Value
AUC ROC	0.901
Accuracy	0.857
Precision (High Discordance)	0.670
Recall (High Discordance)	0.740
Specificity	0.892
F1 Score	0.703
Balanced Accuracy	0.816
CV Test Gap	−0.037

The confusion matrix analysis revealed that the model correctly identified 289 of 324 low discordance cases (89.2 percent) and 71 of 96 high discordance cases (74.0 percent). The model produced 35 false positives (low discordance cases incorrectly predicted as high) and 25 false negatives (high discordance cases incorrectly predicted as low). The relatively higher false positive rate compared to false negative rate suggests the model errs on the side of flagging potential discordance, which may be preferable in quality assurance applications where missing true discordance is more costly than investigating false alarms. Figure 4 presents the ROC curve and confusion matrix visualization.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

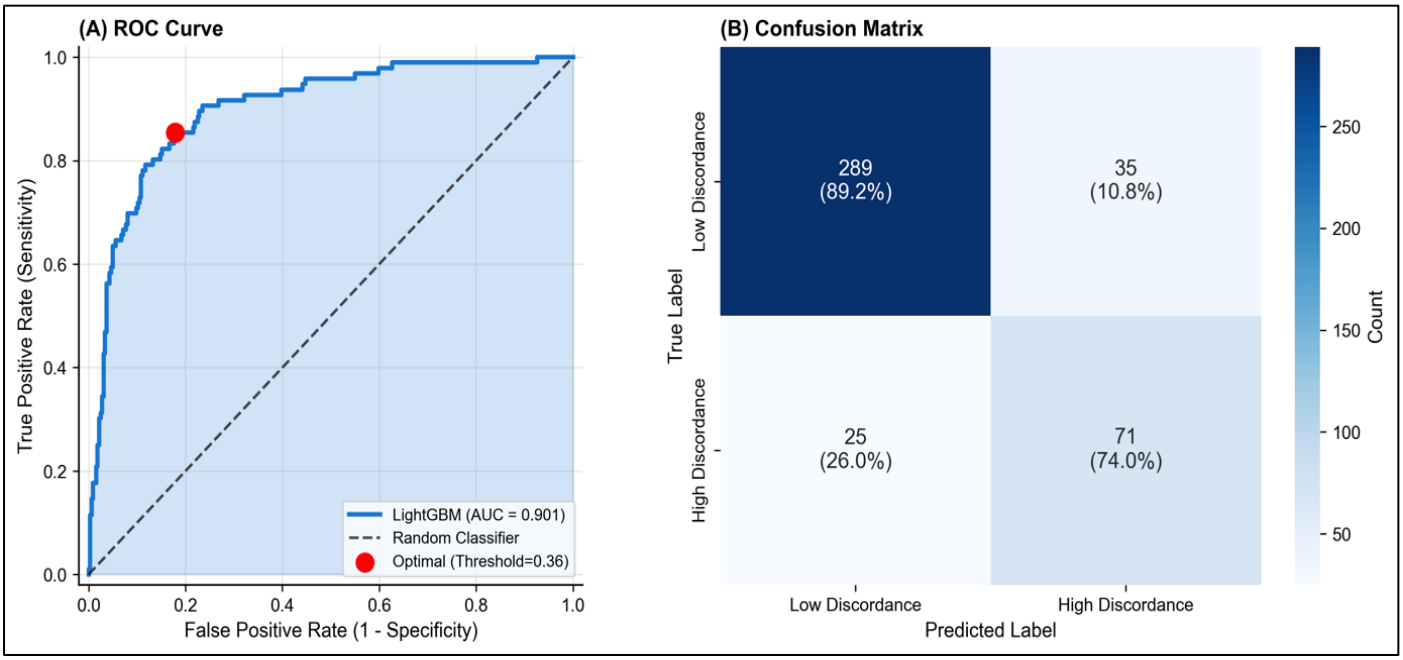

Fig 4 Model Performance Visualization: (a) ROC Curve and (b) Confusion Matrix.

➢ *Feature Importance Analysis*

SHAP analysis identified climate and geographic factors as the dominant predictors of diagnostic discordance. Table 4 presents the top ten features ranked by mean absolute SHAP value. Rainfall emerged as the most influential predictor with mean absolute SHAP value of 0.379, followed by the composite climate index (0.322) and LGA encoded geographic identifier (0.275). Humidity and population density also contributed substantially to predictions, with mean absolute SHAP values of 0.246 and 0.241 respectively. Intervention coverage variables including indoor spraying and bed net coverage ranked lower, suggesting that environmental conditions exert stronger influence on diagnostic agreement than intervention program characteristics.

Table 4 Top 10 Predictive Features Ranked by Shap Importance

Rank	Feature	Mean |SHAP|
1	Rainfall (mm)	0.379
2	Climate Index	0.322
3	LGA (Geographic)	0.275
4	Humidity (%)	0.246
5	Population Density	0.241
6	Temperature (°C)	0.240
7	Antimalarial Stock (%)	0.175
8	Month (cosine)	0.140
9	Indoor Spraying Coverage (%)	0.110
10	Bed Net Coverage (%)	0.103

The SHAP summary plot revealed directional relationships between features and discordance predictions. Higher rainfall values were associated with increased probability of high discordance, as indicated by positive SHAP values for high rainfall observations. Similarly, elevated humidity and climate index values pushed predictions toward the high discordance class. The geographic variable showed distinct patterns across LGAs, with certain locations consistently associated with higher discordance risk regardless of other factors. Seasonal patterns captured by the cosine transformed month variable indicated elevated discordance risk during the wet season months. Figure 5 displays the complete SHAP analysis results.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

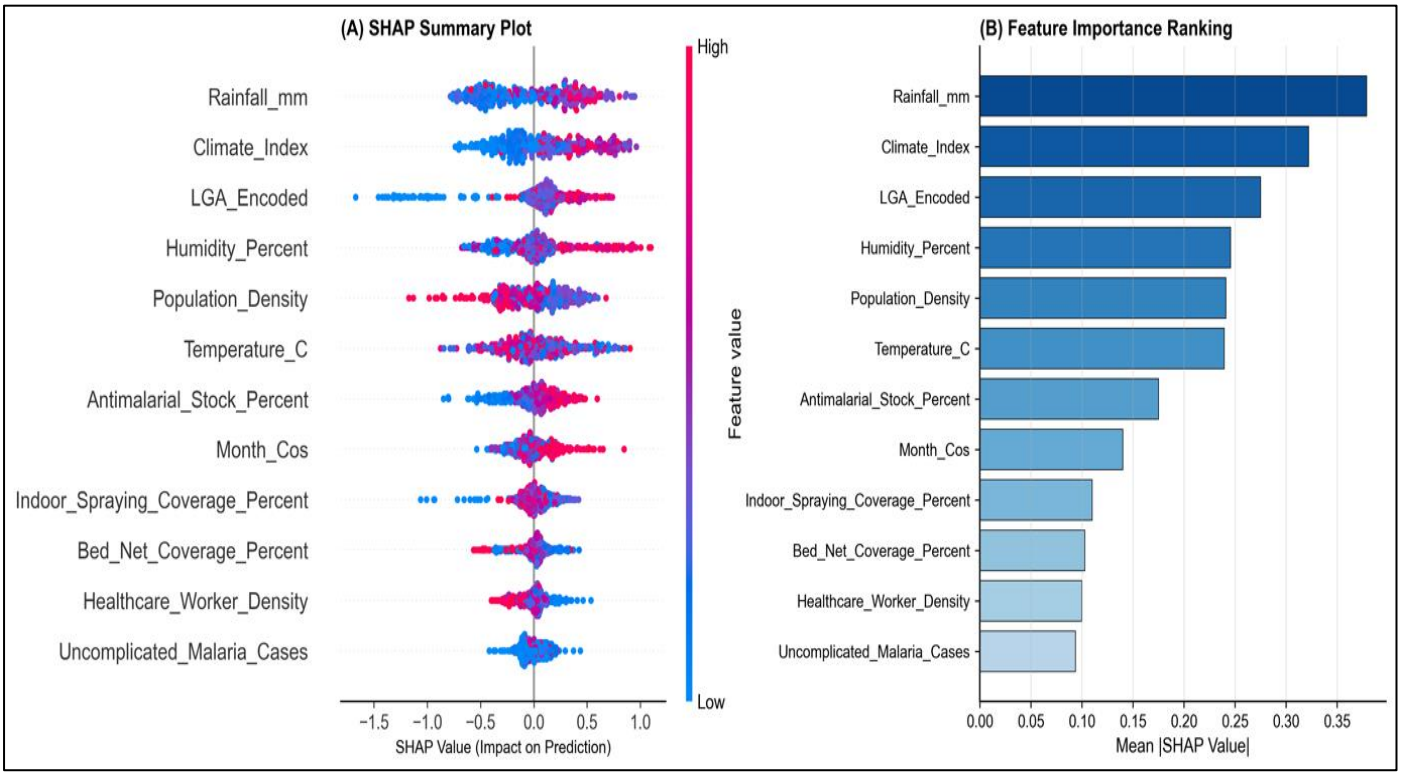

Fig 5 SHAP Feature Importance Analysis: (a) Beeswarm Plot Showing Directional Effects and (b) Summary Bar Plot.

➢ *Spatial and Temporal Patterns*

Spatial analysis revealed heterogeneous discordance rates across the eight LGAs of Bayelsa State. High discordance rates ranged from 13.0 percent in Brass to 37.5 percent in Sagbama, indicating substantial geographic variation in diagnostic agreement. Notably, Sagbama exhibited discordance rates nearly three times higher than Brass, suggesting location-specific factors affecting diagnostic concordance. Seasonal analysis demonstrated that discordance rates were higher during the wet season (24.7 percent) compared to the dry season (17.2 percent), with peak discordance observed in August (37.6 percent) and September (37.0 percent). This seasonal pattern is consistent with the identified importance of climate variables in the SHAP analysis. Figure 6 presents the spatial and temporal pattern visualizations.

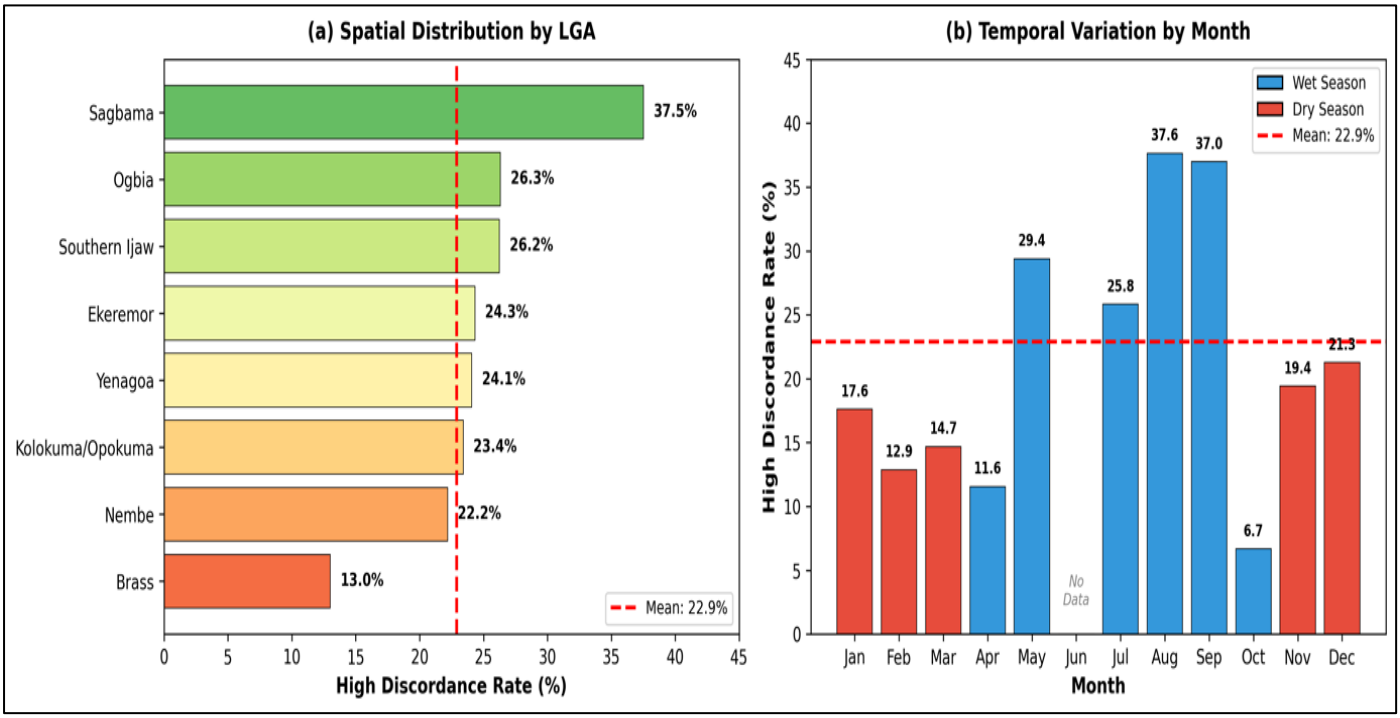

Fig 6 Spatial and Temporal Discordance Patterns: (a) LGA Distribution and (b) Seasonal Variation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

V. DISCUSSION

This study developed and evaluated a machine learning framework for predicting diagnostic test discordance between RDT and microscopy in malaria surveillance data from Bayelsa State, Nigeria. The findings contribute novel insights into both the extent of diagnostic disagreement and the factors that influence when such discordance occurs.

The Bland Altman analysis revealed a systematic bias of negative 2.33 percentage points, indicating that RDT consistently underestimates test positivity rates compared to microscopy. This finding aligns with observations from previous Nigerian studies that have documented variable RDT sensitivity depending on epidemiological conditions [10], [11]. The wide limits of agreement (negative 19.28 to positive 14.62 percentage points) indicate that while the average difference is modest, individual observations can exhibit substantial discordance in either direction. The concordance correlation coefficient of 0.689 represents moderate agreement, falling short of the threshold typically considered acceptable for method interchangeability. This has practical implications for surveillance programs that may switch between or combine diagnostic methods, as systematic adjustments may be necessary when aggregating data from different sources.

The observation that microscopy yielded higher positivity rates than RDT approximately three times more frequently than the reverse pattern warrants consideration. This asymmetry may reflect several biological and technical factors. RDT sensitivity is known to decline at lower parasite densities, potentially missing infections that microscopy can detect [26]. Additionally, the persistence of HRP2 antigen following treatment can cause false positive RDT results, though this would increase rather than decrease RDT positivity. Storage conditions and test kit quality in field settings may also contribute to reduced RDT sensitivity [27]. The predominance of microscopy exceeding RDT suggests that quality assurance efforts should prioritize ensuring adequate RDT sensitivity rather than specificity in this setting.

The LightGBM classifier achieved an AUC of 0.901, demonstrating excellent discriminative ability for identifying high discordance events. This performance compares favorably with machine learning applications in related malaria prediction contexts. Similar AUC values have been achieved for malaria incidence prediction using gradient boosting methods [15], while AUC of 0.83 has been reported for individual malaria diagnosis prediction [31]. The strong performance obtained here suggests that diagnostic discordance, despite its apparent complexity, exhibits learnable patterns that can be captured through supervised learning approaches.

The use of class weight balancing rather than synthetic oversampling proved critical for model generalizability. Initial experiments with SMOTE and related techniques achieved higher cross validation scores but exhibited substantial performance degradation on the held out test set, with CV test gaps exceeding 0.17. The class weight approach yielded a negative CV test gap, indicating that the model performs at least as well on unseen data as during training. This finding has methodological implications for similar imbalanced classification problems in health surveillance contexts, where overfitting to synthetic samples may produce misleadingly optimistic performance estimates. This observation is consistent with recent systematic evaluations showing that SMOTE based techniques may degrade model performance on medical datasets [36].

The SHAP analysis identified climate variables as the dominant predictors of diagnostic discordance, with rainfall, humidity, and the composite climate index occupying the top positions in feature importance rankings. This finding has biological plausibility, as climate conditions affect both malaria transmission intensity and potentially RDT performance. High humidity and temperature can degrade RDT reagents if storage conditions are suboptimal, while increased rainfall is associated with higher mosquito abundance and parasite transmission [41]. The elevated discordance during wet season months (24.7 percent versus 17.2 percent in dry season) provides corroborating evidence for the climate influence on diagnostic agreement.

The prominence of the geographic (LGA) variable in the feature importance rankings indicates that location specific factors beyond climate contribute to discordance patterns. This could reflect differences in laboratory quality, health worker training, RDT storage practices, or local parasite characteristics across the eight LGAs. While this finding highlights the importance of spatial factors, it also represents a limitation in that the model may be learning location specific patterns that do not generalize to other geographic contexts. Future work should investigate the specific mechanisms underlying geographic variation in diagnostic agreement.

Notably, intervention coverage variables including bed net distribution and indoor residual spraying ranked relatively low in feature importance despite their importance for malaria transmission reduction. This suggests that intervention program characteristics have limited direct influence on diagnostic agreement, which is instead driven primarily by environmental and facility level factors. This finding has programmatic relevance, indicating that improving diagnostic concordance requires targeted investments in laboratory quality assurance and RDT storage infrastructure rather than general malaria control intensification.

The interpretable nature of the SHAP based analysis distinguishes this work from black box prediction approaches that may achieve similar accuracy without providing actionable insights. By identifying rainfall and climate conditions as key drivers of discordance, the analysis suggests that surveillance programs could implement season adjusted quality control protocols, with enhanced monitoring during wet season months when discordance risk is elevated. Similarly, the identification of geographic hotspots enables

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

targeted capacity building efforts in LGAs with consistently high discordance rates.

VI. LIMITATIONS

Several limitations should be considered when interpreting these findings. First, the study utilized aggregated monthly data rather than individual patient level records, which may mask within month variability and preclude analysis of individual level factors influencing diagnostic agreement. Second, the dataset exhibited uneven sampling across LGAs, with Nembe contributing 22.8 percent of observations compared to only 3.4 percent from Sagbama. This imbalance could bias results toward patterns observed in more heavily sampled locations. Third, the prominence of the geographic variable in feature importance rankings suggests that the model may have learned location specific patterns that may not generalize to other states or regions of Nigeria.

Fourth, the weak correlations between individual features and discordance (all below 0.25) indicate that diagnostic disagreement is influenced by complex interactions rather than single dominant factors, limiting the explanatory power of any predictive model. Fifth, the study lacked information on specific RDT brands, lot numbers, or storage conditions, factors known to influence RDT performance that could not be incorporated into the analysis.

Fifth, while microscopy is treated as the reference standard, we acknowledge known inter-reader variability and operational errors, which may introduce label noise into discordance classification.

Sixth, random train–test splitting may partially overestimate generalization due to spatial and temporal correlation in malaria transmission. However, the objective is operational screening within similar epidemiological contexts rather than geographic extrapolation. Future work will explore spatially blocked validation. Finally, the cross sectional nature of the evaluation prevents assessment of model performance under prospective deployment conditions, where data distributions may shift over time.

VII. CONCLUSION AND RECOMMENDATIONS

➢ *Conclusion*

This study developed an interpretable machine learning framework for predicting diagnostic test discordance between RDT and microscopy in malaria surveillance. The Bland Altman analysis demonstrated moderate agreement between the diagnostic methods, with systematic bias toward lower RDT positivity rates and wide limits of agreement that preclude treating the methods as interchangeable. The LightGBM classifier achieved excellent discriminative performance with an AUC of 0.901 while maintaining generalizability as evidenced by the absence of overfitting. SHAP analysis identified climate variables, particularly rainfall and humidity, as the primary drivers of diagnostic discordance, with geographic location also contributing substantially to prediction.

The findings contribute to the limited literature on applying machine learning to diagnostic quality assessment in malaria surveillance. By providing both predictive capability and mechanistic insights, the framework offers potential utility for informing quality assurance protocols and resource allocation decisions. The dominance of environmental factors in driving discordance suggests opportunities for season adjusted monitoring strategies that intensify quality control during periods of elevated discordance risk.

➢ *Recommendations*

Based on the study findings, the following recommendations are proposed for practice and future research. For malaria surveillance programs, implementing enhanced quality control protocols during wet season months when discordance risk is elevated could improve data quality. Investment in climate controlled RDT storage facilities may reduce environmentally driven performance degradation. Geographic areas identified as discordance hotspots should receive prioritized laboratory capacity building and microscopist training.

For future research, external validation of the prediction model in other Nigerian states and across different epidemiological settings would establish generalizability bounds. Investigation of individual level factors through patient record linkage could identify additional predictors not captured in aggregated data. Integration of RDT brand and storage condition information would enable more granular analysis of test performance factors. Development of real time decision support tools incorporating the prediction model could facilitate prospective quality monitoring. Finally, health economic evaluation of targeted quality assurance strategies informed by model predictions would support resource allocation decisions.

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," Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240086173

[2]. R. W. Snow, "Global malaria eradication and the importance of *Plasmodium falciparum* epidemiology in Africa," *BMC Med.*, vol. 13, p. 23, 2015, doi: 10.1186/s12916-014-0254-7.

[3]. A. M. Noor, D. K. Kinyoki, C. W. Mundia, C. W. Kabaria, J. W. Mutua, V. A. Alegana, I. S. Fall, and R. W. Snow, "The changing risk of *Plasmodium*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

falciparum malaria infection in Africa: 2000–10," *Lancet*, vol. 383, no. 9930, pp. 1739–1747, 2014, doi: 10.1016/S0140-6736(13)62566-0.

[4]. J. L. Gallup and J. D. Sachs, "The economic burden of malaria," *Am. J. Trop. Med. Hyg.*, vol. 64, no. 1–2, pp. 85–96, 2001, doi: 10.4269/ajtmh.2001.64.85.

[5]. World Health Organization, "Guidelines for malaria," Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/publications/i/item/guidelines-for-malaria

[6]. I. Bates, V. Bekoe, and A. Asamoa-Adu, "Improving the accuracy of malaria-related laboratory tests in Ghana," *Malar. J.*, vol. 3, p. 38, 2004, doi: 10.1186/1475-2875-3-38.

[7]. C. K. Murray, R. A. Gasser, A. J. Magill, and R. S. Miller, "Update on rapid diagnostic testing for malaria," *Clin. Microbiol. Rev.*, vol. 21, no. 1, pp. 97–110, 2008, doi: 10.1128/CMR.00035-07.

[8]. D. Bell, C. Wongsrichanalai, and J. W. Barnwell, "Ensuring quality and access for malaria diagnosis: How can it be achieved?," *Nat. Rev. Microbiol.*, vol. 4, pp. S7–S20, 2006, doi: 10.1038/nrmicro1525.

[9]. A. Endeshaw, T. Gebre, J. Ngondi, P. M. Graves, E. B. Shargie, Y. Ejigsemahu, B. Ayele, G. Yohannes, T. Ber, A. Byber, A. Lemon, and F. O. Richards, "Evaluation of light microscopy and rapid diagnostic test for the detection of malaria under operational field conditions," *Malar. J.*, vol. 7, p. 118, 2008, doi: 10.1186/1475-2875-7-118.

[10]. A. F. Fagbamigbe, "On the discriminatory and predictive accuracy of the RDT against the microscopy in the diagnosis of malaria among under-five children in Nigeria," *Malar. J.*, vol. 18, no. 1, p. 46, 2019, doi: 10.1186/s12936-019-2678-1.

[11]. O. O. Oladosu and W. A. Oyibo, "Performance evaluation of a popular malaria RDT in Nigeria compared with microscopy," *J. Parasitol. Res.*, vol. 2020, p. 3650848, 2020, doi: 10.1155/2020/3650848.

[12]. C. I. R. Chandler, C. Jones, G. Boniface, K. Juma, H. Reyburn, and C. J. M. Whitty, "Guidelines and mindlines: Why do clinical staff over-diagnose malaria in Tanzania?," *Malar. J.*, vol. 7, p. 53, 2008, doi: 10.1186/1475-2875-7-53.

[13]. D. J. Kyabayinze, C. Asiimwe, D. Nakanjako, J. Naber, H. Counihan, J. K. Tibenderana, and S. G. Staedke, "Use of RDTs to improve malaria diagnosis and fever case management at primary health care facilities in Uganda," *Malar. J.*, vol. 9, p. 200, 2010, doi: 10.1186/1475-2875-9-200.

[14]. E. Mbunge, S. G. Fashoto, J. Odun-Ayo, and C. Metfula, "Application of machine learning and deep learning for malaria diagnosis: A systematic literature review," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, p. 101680, 2023, doi: 10.1016/j.jksuci.2023.101680.

[15]. O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," *Inform. Med. Unlocked*, vol. 22, p. 100508, 2021, doi: 10.1016/j.imu.2020.100508.

[16]. P. Martineau, L. Cardenas, D. Kappel, L. Lorenz, V. L. R. G. Fiaccone, K. V. Braga, and P. G. S. Florentino, "Prediction of malaria using deep learning models: A case study on city clusters in the state of Amazonas, Brazil," *Sci. Rep.*, vol. 12, p. 12437, 2022, doi: 10.1038/s41598-022-16455-3.

[17]. M. Stow and E. C. M. Obasi, "An interpretable early warning system for malaria outbreaks in Bayelsa State using deep learning and climate data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 14, no. 8, pp. 38–46, 2025, doi: 10.17148/IJARCCE.2025.14806.

[18]. 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.

[19]. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4765–4774, 2017.

[20]. M. Stow and A. A. Stewart, "Empirical analysis of SHAP stability under data corruption across datasets and model architectures," *Int. Adv. Res. J. Sci. Eng. Technol.*, vol. 12, no. 8, pp. 92–110, 2025, doi: 10.17148/IARJSET.2025.12810.

[21]. M. Stow and A. A. Stewart, "Interpreting machine learning predictions with SHAP and LIME for transparent decision making," *Int. J. Comput. Sci. Math. Theory*, vol. 11, no. 8, pp. 22–49, 2025, doi: 10.56201/ijcsmt.vol.11.no8.2025.pg22.49.

[22]. Federal Ministry of Health Nigeria, "National malaria strategic plan 2021–2025," Abuja, Nigeria, 2020.

[23]. A. Moody, "Rapid diagnostic tests for malaria parasites," *Clin. Microbiol. Rev.*, vol. 15, no. 1, pp. 66–78, 2002, doi: 10.1128/CMR.15.1.66-78.2002.

[24]. P. L. Alonso, G. Brown, M. Arevalo-Herrera, F. Binka, C. Chitnis, F. Collins, O. K. Doumbo, B. Greenwood, B. F. Hall, M. M. Levine, K. Mendis, R. D. Newman, C. V. Plowe, M. H. Rodriguez, R. Sinden, L. Slutsker, and M. Tanner, "A research agenda to underpin malaria eradication," *PLoS Med.*, vol. 8, no. 1, p. e1000406, 2011, doi: 10.1371/journal.pmed.1000406.

[25]. K. Wafula, C. J. Mwangi, G. S. Amwayi, and C. M. Mureithi, "Quality of malaria microscopy diagnosis by laboratory personnel in a clinical setting," *Pan Afr. Med. J.*, vol. 32, p. 51, 2019, doi: 10.11604/pamj.2019.32.51.17576.

[26]. O. A. Mokuolu, M. T. Ajayi, M. L. Ntadom, C. N. Ezeiru, O. T. Bakare, H. O. Musa, C. O. Falade, and S. K. Ojurongbe, "Malaria rapid diagnostic tests and malaria microscopy for guiding malaria treatment of uncomplicated fevers in Nigeria," *Clin. Infect. Dis.*, vol. 63, pp. S290–S297, 2016, doi: 10.1093/cid/ciw631.

[27]. O. B. Awosolu, Z. S. Yahaya, and M. T. Farah Haziqah, "Performance evaluation of nested PCR, light microscopy, and PfHRP2 RDT in the detection of falciparum malaria in Nigeria," *Pathogens*, vol. 11, no. 11, p. 1312, 2022, doi: 10.3390/pathogens11111312.

[28]. O. T. Oyeyemi, A. F. Ogunlade, and I. O. Oyewole, "Comparative assessment of microscopy and rapid diagnostic test for malaria diagnosis in southwestern Nigeria," *J. Parasitol. Vector Biol.*, vol. 7, no. 2, pp. 34–41, 2015.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan131

[29]. O. I. Ita, A. E. Udo, N. E. Usang, N. O. Adegunloye, E. I. Archibong, and J. U. Akpan, "A diagnostic performance evaluation of rapid diagnostic tests and microscopy for malaria diagnosis using nested PCR as reference standard," *Niger. J. Clin. Pract.*, vol. 23, no. 3, pp. 355–361, 2020, doi: 10.4103/njcp.njcp_539_19.

[30]. J. Zhu, K. Wang, S. Li, and Z. Chen, "Stacking ensemble method for malaria incidence prediction," *PLoS ONE*, vol. 16, no. 7, p. e0253545, 2021, doi: 10.1371/journal.pone.0253545.

[31]. T. A. Ojurongbe, O. A. Mokuolu, O. Oyelami, S. S. Okekunle, O. O. Abioye-Kuteyi, T. A. Adeyemo, and O. A. Ojurongbe, "Prediction of malaria positivity using patients' demographic and environmental features," *Malar. J.*, vol. 22, p. 372, 2023, doi: 10.1186/s12936-023-04805-x.

[32]. P. U. Eze, C. I. Okagbue, E. A. Ogbonnia, and I. N. Okafor, "Application of machine learning models in predicting malaria prevalence in Nigeria," *J. Parasit. Dis.*, vol. 48, pp. 1–12, 2024, doi: 10.1007/s12639-025-01880-6.

[33]. M. Stow, "Explainable machine learning framework for income prediction with class imbalance optimization," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 14, no. 8, Art. no. 14801, 2025, doi: 10.17148/IJARCCE.2025.14801.

[34]. 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," *Nat. Mach. Intell.*, vol. 2, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[35]. D. K. Muriithi, G. O. Odhiambo, and F. M. Mwangi, "Explainable artificial intelligence models for predicting malaria risk in Kenya," *Int. J. Environ. Res. Public Health*, vol. 20, no. 13, p. 6257, 2023, doi: 10.3390/ijerph20136257.

[36]. M. Stow, "When data augmentation hurts: A systematic evaluation of SMOTE-based techniques on medical datasets," *Int. J. Adv. Res. Comput. Sci.*, vol. 16, no. 4, pp. 14–33, 2025, doi: 10.26483/ijarcs.v16i4.7313.

[37]. B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.

[38]. J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 327, no. 8476, pp. 307–310, 1986, doi: 10.1016/S0140-6736(86)90837-8.

[39]. L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989, doi: 10.2307/2532051.

[40]. 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146–3154, 2017.

[41]. M. C. Thomson, S. J. Mason, T. Phindela, and S. J. Connor, "Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana," *Am. J. Trop. Med. Hyg.*, vol. 73, no. 1, pp. 214–221, 2005, doi: 10.4269/ajtmh.2005.73.214.