PREDICTING MALARIA OUTBREAKS IN RURAL LIBERIA USING MACHINE LEARNING

FTL Liberia Group Six (6)

ABSTRACT

This project proposes to build a machine learning model to predict malaria outbreaks in rural Liberia using environmental, demographic, and health data. The goal is to create an early warning system that helps manage resources and save lives, supporting Sustainable Development Goals (SDGs) 3 and 1.

Group Members

1. Nehemiah Kemayah
2. Thomas Adonis Marwolo
3. Robert Bright Yekeh
4. Joel J. Barclay

Machine Learning Project Documentation

1. Overview

The model refinement phase represents a critical transition from initial exploration to optimized prediction capability for our malaria outbreak early warning system in rural Liberia. Following our initial model exploration that identified XGBoost, Random Forest, and Logistic Regression as viable candidates, this phase focused on systematically improving model performance through hyperparameter optimization, feature engineering refinement, and validation strategy enhancement. The primary objectives were to maximize prediction accuracy while maintaining interpretability for healthcare stakeholders, ensure robust generalization across diverse rural districts, and optimize the balance between sensitivity and specificity to minimize both false alarms and missed outbreaks.

2. Model Evaluation

Our initial model exploration phase established baseline performance metrics across three algorithms. The evaluation revealed both strengths and areas requiring improvement:

Model	Baseline Accuracy	Key Issues Identified
XGBoost	87.3%	Slight overfitting on training data, suboptimal recall for outbreak detection
Random Forest	84.7%	Lower sensitivity for minority class, needed feature importance optimization
Logistic Regression	79.2%	Limited capacity to capture non-linear climate-health relationships

Critical Findings from Initial Evaluation:

- Class imbalance significantly impacted minority class detection, with outbreak cases representing only 32% of the dataset
- Feature importance analysis revealed that 3-month lagged rainfall and temperature-humidity interactions were underweighted in initial models
- Temporal validation showed performance degradation on future time periods, indicating overfitting to training temporal patterns
- District-level performance varied significantly, with Montserrado and Nimba showing lower prediction accuracy than other regions

3. Refinement Techniques

3.1 Class Imbalance Handling

To address the critical challenge of outbreak underdetection, we implemented multiple class balancing strategies:

- SMOTE (Synthetic Minority Over-sampling Technique): Generated synthetic outbreak examples by interpolating between existing minority class instances, increasing outbreak representation from 32% to 45% while maintaining data diversity
- Class weight adjustment: Implemented scale_pos_weight parameter in XGBoost (ratio of 2.1:1) to penalize false negatives more heavily than false positives
- Threshold optimization: Moved decision threshold from default 0.5 to 0.42 to increase outbreak detection sensitivity based on cost-benefit analysis of false alarms versus missed outbreaks

Impact: These combined techniques improved outbreak detection recall from 68% to 89% while maintaining overall accuracy above 85%.

3.2 Advanced Feature Engineering

Building on initial feature set, we developed sophisticated engineered features informed by malaria transmission biology:

- Climate suitability index: Composite metric combining temperature (20-30°C optimal range), humidity (60-90% optimal), and rainfall patterns weighted by literature-derived coefficients
- Vector breeding potential: Calculated as interaction between 2-month lagged rainfall and current temperature, capturing mosquito life cycle dynamics
- Seasonal risk modulation: Binary indicators for high-risk months (May-October) combined with year-over-year outbreak clustering patterns
- District vulnerability scores: Encoded historical outbreak frequency, population density, and healthcare access indicators specific to each district
- Climate variability metrics: 3-month rolling standard deviations capturing extreme weather events that disrupt normal transmission patterns

Impact: Feature importance analysis showed climate suitability index and vector breeding potential ranked as top 2 predictors, improving model discrimination by 7.3%.

3.3 Ensemble Method Development

Recognizing that different algorithms capture complementary patterns, we developed a weighted ensemble combining XGBoost and Random Forest predictions. The ensemble weights (0.65 for XGBoost, 0.35 for Random Forest) were optimized through grid search on validation data. This approach leveraged XGBoost's superior handling of complex interactions while benefiting from Random Forest's robustness to overfitting, resulting in a 2.8% accuracy improvement over the best individual model.

4. Hyperparameter Tuning

Systematic hyperparameter optimization was performed using Bayesian optimization with 5-fold cross-validation to identify optimal configurations for each algorithm:

4.1 XGBoost Optimization

Initial configuration was based on literature recommendations, then refined through 200 iterations of Bayesian search:

- Learning rate: Reduced from 0.1 to 0.03 to improve generalization (slower learning prevents overfitting to training temporal patterns)

- Max depth: Optimized to 6 (from initial 8) balancing model complexity with interpretability requirements for health stakeholders
- Number of estimators: Increased to 500 with early stopping (patience=50) based on validation loss
- Subsample ratio: Set to 0.8 to introduce stochasticity and reduce overfitting
- Colsample by tree: Optimized to 0.7 (feature sampling per tree) to improve model diversity
- Gamma: Set to 1.0 (minimum loss reduction) to control tree complexity and prevent overspecialization

Performance Impact: Hyperparameter tuning improved XGBoost validation accuracy from 87.3% to 91.7%, with cross-validation standard deviation reduced from 4.2% to 2.1%, indicating more stable predictions.

4.2 Random Forest Optimization

- Number of trees: Increased from 100 to 300 based on out-of-bag error stabilization analysis
- Max features: Optimized to sqrt(n_features) for optimal variance-bias tradeoff
- Min samples split: Set to 10 to prevent overfitting on small temporal patterns
- Min samples leaf: Optimized to 4, balancing tree depth with minimum leaf size requirements
- Class weight: Applied balanced mode to automatically adjust weights inversely proportional to class frequencies

5. Cross-Validation Strategy Enhancement

Our initial 10-fold cross-validation approach was enhanced to address time-series specific challenges and district heterogeneity:

5.1 Time-Series Cross-Validation

Implemented expanding window time-series cross-validation to respect temporal ordering. Training always occurred on earlier time periods with validation on future periods, using 6 expanding windows with 6-month increments. This approach revealed temporal drift in model performance and guided implementation of seasonal adaptation mechanisms.

5.2 Spatial Cross-Validation

Added leave-one-district-out cross-validation to assess spatial generalization. Models trained on 14 districts and validated on the held-out district, rotating through all 15 districts. This validation identified district-specific features requiring separate handling and confirmed that climate-based features generalize well across spatial boundaries while demographic features showed district-specific patterns.

5.3 Nested Cross-Validation

Implemented nested cross-validation with outer loop for performance estimation and inner loop for hyperparameter tuning. This approach provided unbiased performance estimates by ensuring hyperparameter selection did not leak information from test folds, resulting in more conservative but reliable performance metrics.

6. Feature Selection and Importance Analysis

Systematic feature selection was performed to optimize the feature set while maintaining interpretability:

- **Hyperparameter Optimization:** Grid and randomized search for key parameters (e.g., number of estimators, learning rate, max depth).
- **Ensemble Learning:** Combination of **Random Forest** and **XGBoost** using a weighted voting classifier improved overall robustness.
- **Data Rebalancing:** Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to reduce class imbalance.
- **Feature Lagging:** Introduced 1–3 month lag variables for rainfall and humidity to capture delayed malaria transmission effects.
- **Normalization & Regularization:** Improved model convergence and reduced overfitting.

6.1 Feature Importance Ranking

Using XGBoost's gain-based feature importance combined with permutation importance for validation:

Feature	Importance Score	Interpretation
Climate Suitability Index	0.187	Composite climate favorability
3-month Lag Rainfall	0.164	Mosquito breeding cycle
Vector Breeding Potential	0.142	Rainfall-temperature interaction
Population Density	0.118	Exposure risk factor
2-month Lag Temperature	0.103	Parasite development period

6.2 Feature Selection Results

Recursive feature elimination with cross-validation identified that 18 of the original 27 features provided optimal performance. Removing 9 low-importance features (primarily single-month climate values without lag transformations) improved model interpretability while maintaining 98.7% of original performance. The final feature set balanced predictive power with stakeholder comprehensibility, a critical requirement for operational deployment in healthcare settings.

Test Submission

1. Overview

The test submission phase evaluated our refined malaria outbreak prediction model on held-out temporal data representing future prediction scenarios. This phase assessed model generalization to unseen time periods, validated prediction reliability under operational conditions, and established performance benchmarks for real-world deployment. The test dataset comprised 30% of our temporal data (2022-2023), representing 540 district-months across all 15 districts in our study area. This evaluation simulated operational deployment where the model predicts future outbreaks based on historical training.

2. Data Preparation for Testing

Test data preparation followed identical preprocessing pipelines established during training to ensure consistency:

2.1 Data Preprocessing Pipeline

- Missing value imputation: Applied same monthly average imputation parameters learned from training data
- Feature scaling: Used StandardScaler parameters fitted on training data to transform test features
- Lag feature generation: Created 1-3 month lagged climate variables using proper temporal boundaries
- Engineered features: Computed climate suitability index and vector breeding potential using training-derived coefficients
- Rolling window features: Calculated 3-month and 6-month rolling statistics respecting temporal boundaries

2.2 Data Quality Verification

Before model application, we verified test data quality through statistical validation. Distribution comparisons confirmed that test data climate ranges aligned with training data (no distribution shift). Temporal continuity checks ensured proper lag feature calculation across train-test boundary. Feature correlation patterns in test data matched training expectations, confirming data integrity.

3. Model Application

The refined ensemble model was applied to test data following operational deployment protocols:

3.1 Prediction Generation Process

Each test instance was processed through the ensemble pipeline:

- XGBoost model generated probability predictions using optimized hyperparameters
- Random Forest model produced independent probability predictions
- Ensemble aggregation combined predictions with optimized weights (0.65 XGBoost, 0.35 Random Forest)
- Classification threshold (0.42) converted probabilities to binary outbreak predictions
- Prediction confidence intervals were calculated using ensemble disagreement as uncertainty measure

3.2 Implementation Code Structure

The prediction implementation maintained strict separation between training and test data. Models were loaded from saved state files to ensure exact reproducibility. Feature engineering transformations were applied using fitted preprocessors from training phase. Predictions were generated in batch mode for computational efficiency, processing all 540 test instances. Output format included binary predictions, probability scores, confidence intervals, and district-month identifiers for operational deployment.

4. Test Metrics and Performance Evaluation

Comprehensive evaluation on held-out test data demonstrated strong model performance:

Metric	Test Performance	Validation (Baseline)
Overall Accuracy	**89.8%**	91.7%
Sensitivity (Recall)	**86.4%**	89.0%
Specificity	**91.7%**	93.2%
Precision	**84.2%**	87.5%
F1-Score	**85.3%**	88.2%
ROC AUC Score	**0.947**	0.956

4.1 Performance Analysis

Key observations from test performance:

- Minimal generalization gap: Test accuracy of 89.8% versus validation accuracy of 91.7% demonstrates excellent generalization with only 1.9% degradation
- Strong outbreak detection: 86.4% sensitivity means model successfully identifies 86 of every 100 actual outbreaks, critical for early warning system effectiveness
- Low false alarm rate: 91.7% specificity translates to only 8.3% false positive rate, maintaining stakeholder confidence
- Excellent discrimination: ROC AUC of 0.947 indicates model effectively separates outbreak from non-outbreak cases across all threshold values
- Balanced performance: F1-score of 85.3% demonstrates good balance between precision and recall, avoiding extreme bias toward either metric

4.2 District-Level Performance Breakdown

Analysis of test performance across districts revealed generally consistent accuracy (range 84.7% to 92.3%) with strongest performance in districts with historical data quality and slightly lower accuracy in Montserrado and Nimba due to complex urban-rural dynamics. This spatial validation confirms model reliability across diverse geographic and demographic contexts.

4.3 Temporal Performance Patterns

Month-by-month analysis showed model maintained consistent performance throughout the test period, with no significant performance degradation over time. Performance was slightly higher during rainy season months (May-October) when outbreak signals are stronger, and marginally lower during dry season transitions when outbreak probability is naturally lower and climate signals are weaker.

5. Comparison with Literature Benchmarks

Our test results compare favorably with published malaria prediction studies in Sub-Saharan Africa. The Gambia study achieved 87.2% accuracy using similar meteorological predictors. The Burkina Faso operational system reported 82.5% accuracy for 13-week forecasts. Our 89.8% test accuracy with enhanced feature engineering and ensemble methods represents competitive performance while addressing Liberia's specific data challenges. The ROC AUC of 0.947 exceeds the typical 0.85-0.92 range reported in literature, demonstrating superior discrimination capability.

6. Model Deployment Considerations

6.1 Operational Deployment Strategy

Based on test performance, we recommend phased deployment:

- Phase 1 (Months 1-3): Pilot deployment in 3 districts with strongest historical data quality (Bong, Lofa, Grand Bassa) for validation and stakeholder feedback
- Phase 2 (Months 4-6): Expansion to 8 additional districts with integration into existing DHIS2 reporting workflows
- Phase 3 (Months 7-12): National rollout across all 15 districts with established monitoring and continuous improvement protocols

6.2 Integration with Health Systems

Deployment will integrate with existing infrastructure:

- DHIS2 platform: Monthly automated prediction generation using current climate data and previous month case reports
- Early warning dashboard: Visual interface showing district-level outbreak probabilities with 3-month forecast horizon
- Alert system: Automated notifications to district health officers when outbreak probability exceeds 60% threshold
- Resource planning module: Integration with LMIS for predictive commodity allocation based on outbreak forecasts

6.3 Model Monitoring and Maintenance

Continuous performance monitoring will track prediction accuracy against observed outcomes, feature drift detection identifying changes in climate or demographic patterns, quarterly model retraining incorporating recent data, and annual comprehensive evaluation with potential architecture updates. Performance degradation triggers (accuracy dropping below 85% for two consecutive months) will initiate immediate investigation and corrective action.

7. Code Implementation

The complete model refinement and test submission implementation comprises several integrated components:

7.1 Key Implementation Components

- Data preprocessing pipeline: Handles missing values, feature scaling, lag generation, and engineered feature computation
- Feature engineering module: Implements climate suitability index, vector breeding potential, and seasonal risk factors

- Model training framework: Hyperparameter optimization using Bayesian search with cross-validation
- Ensemble prediction system: Weighted combination of XGBoost and Random Forest with confidence interval estimation
- Evaluation framework: Comprehensive metrics calculation including temporal and spatial validation
- Deployment utilities: Model serialization, batch prediction, and integration interfaces

Technology Stack:

- Python 3.9+ with scikit-learn 1.0+, XGBoost 1.6+, pandas, numpy
- Bayesian optimization: scikit-optimize for hyperparameter tuning
- Class balancing: imbalanced-learn library for SMOTE implementation
- Visualization: matplotlib, seaborn for performance analysis
- Deployment: joblib for model serialization, Flask for API interface

Conclusion

The model refinement and test submission phase successfully transformed our initial exploration models into a robust, deployment-ready malaria outbreak prediction system for rural Liberia. Through systematic hyperparameter optimization, advanced feature engineering, and comprehensive validation strategies, we achieved test accuracy of 89.8% with strong outbreak detection sensitivity of 86.4%. The model demonstrates excellent generalization across both temporal and spatial dimensions, with minimal performance degradation from validation to test data.

Key Achievements

- Improved outbreak detection from 68% to 86.4% recall through class balancing and threshold optimization
- Enhanced overall accuracy from 87.3% baseline to 89.8% test performance through ensemble methods
- Developed interpretable engineered features (climate suitability index, vector breeding potential) that align with malaria transmission biology
- Validated spatial generalization across 15 districts and temporal generalization across 18-month test period
- Established operational deployment framework with phased rollout strategy and continuous monitoring protocols

Challenges Encountered

- Class imbalance required multiple mitigation strategies (SMOTE, class weights, threshold optimization) to achieve satisfactory outbreak detection
- District heterogeneity necessitated spatial cross-validation revealing performance variations requiring district-specific feature adaptations
- Temporal validation revealed slight performance degradation during season transitions, addressed through seasonal adaptation mechanisms

- Balancing model complexity with interpretability requirements for healthcare stakeholders required careful feature selection
- Hyperparameter optimization computational costs necessitated Bayesian optimization over exhaustive grid search

Final Performance Summary

Our ensemble model achieved test performance exceeding published benchmarks from comparable Sub-Saharan African studies. The ROC AUC of 0.947 indicates excellent discrimination capability, while the balanced performance across sensitivity (86.4%) and specificity (91.7%) demonstrates optimal trade-offs for operational early warning systems. The model successfully addresses the core challenge of providing actionable 3-month advance warning for malaria outbreaks while maintaining stakeholder confidence through low false alarm rates.

Impact and Next Steps

This predictive system has potential to transform Liberia's malaria control approach from reactive response to proactive prevention. By enabling 3-month advance warning of district-level outbreaks, health authorities can optimize resource allocation, implement targeted interventions, and potentially reduce malaria morbidity and mortality in vulnerable rural populations. The phased deployment strategy ensures careful validation while building stakeholder confidence and operational capacity.

Future enhancements will focus on extending prediction horizons beyond 3 months, incorporating additional data sources such as vector surveillance and intervention coverage, developing mobile interfaces for field-level access, and establishing feedback mechanisms for continuous model improvement based on operational experience. This work directly supports Sustainable Development Goal 3 (Good Health and Well-being) and demonstrates the potential of machine learning to address critical public health challenges in resource-constrained settings.

References

- Balogun, A.L., et al. (2021). Prediction of malaria incidence using climate variability and machine learning. Informatics in Medicine Unlocked, 22, 100508.
- Jaiteh, F., et al. (2024). Predicting malaria outbreak in The Gambia using machine learning techniques. PLOS One, 19(5), e0304289.
- Martineau, P., et al. (2022). Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa. Frontiers in Public Health, 10, 962377.
- Merkord, C.L., et al. (2021). Predicting malaria epidemics in Burkina Faso with machine learning. PLOS One, 16(6), e0253302.
- Woldegiorgis, A.B., et al. (2023). Machine Learning Techniques for Predicting Malaria in Sub-Saharan Africa. In Artificial Intelligence and Machine Learning for Healthcare, Springer.
- Liberia Ministry of Health DHIS2 System - National Malaria Control Program surveillance data
- World Bank Climate Change Knowledge Portal - Liberia climate datasets
- Liberia Institute of Statistics - Population and demographic data