Capstone Project Concept Note and Implementation Plan

Project Title: AI-Driven Malaria Outbreak Prediction System

Team Members

1. Thomas S. McCay
2. David Paye

1. Project Overview

Malaria is one of the leading public health challenges in Liberia, particularly in rural areas where access to healthcare services is limited. Seasonal fluctuations in rainfall, temperature, and humidity create favorable conditions for malaria transmission, leading to recurrent outbreaks that place significant strain on communities.

This project aims to design an **AI-driven malaria outbreak prediction system** that leverages health, climate, and demographic data to forecast potential outbreaks. By providing early warnings, the system will allow Liberia's Ministry of Health and community health workers to allocate resources more efficiently, plan interventions, and ultimately save lives.

The project directly contributes to:

- **SDG 3: Good Health and Well-being** by supporting early detection, disease surveillance, and outbreak prevention.

- **SDG 1: No Poverty** by reducing the economic burden of malaria on families and local economies

2. Objectives

These are our project objective as follow:

- To develop a machine learning model capable of accurately predicting malaria outbreaks.
- To integrate climate, population, and health surveillance data into a predictive system.
- To design a user-friendly GIS dashboard for visualization and decision-making.
- To contribute to reduced malaria morbidity and mortality by improving early response strategies.

3. Background

Malaria remains endemic in Liberia and accounts for a large proportion of outpatient consultations, especially in rural regions. Despite efforts by the government and NGOs—such as mosquito net distribution and health campaigns—response measures are often reactive rather than proactive.

Previous studies have shown that malaria outbreaks are strongly influenced by **environmental conditions (rainfall, temperature, humidity)** and **population density**, which can be modeled with machine learning for accurate prediction. While global research has demonstrated success, localized AI applications for Liberia are still emerging. Implementing a machine learning–driven system tailored to Liberia's unique conditions is therefore necessary to strengthen outbreak surveillance and response.

4. Methodology

This project will follow a structured machine learning pipeline:

- **Data Collection & Preprocessing:** Obtain malaria case reports, climate data, and population statistics; clean and merge data; handle missing values; normalize and standardize features.

- **Feature Engineering:** Create lag variables (e.g., rainfall in previous weeks), aggregate population density, and seasonal indicators.

- **Model Development:** Train machine learning algorithms including **Random Forests, Gradient Boosting (XGBoost/LightGBM), and Logistic Regression**.

- **Model Evaluation:** Use metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC for evaluation.

- **Deployment:** Build a prototype **dashboard/API** that provides predictions and trend visualizations to stakeholders.

5. Architecture Design Diagram

```
┌─────────────────────────────────────┐
│     Data Sources  │
│ Health (MOH), Climate, Pop │
└─────────────────┬───────────────────┘
                  │
  ┌────────▼────────┐
  │ Preprocessing   │
  │ Cleaning, Merge │
  └───────┬─────────┘
          │
  ┌────── ▼──────┐
```

```
│ Feature Eng.    │
│ Lag, Scaling    │
└────────┬────────┘
         │
┌────────▼────────┐
│ ML Models       │
│ RF, XGBoost, LR │
└────────┬────────┘
         │
┌────────▼────────┐
│ Model Eval.     │
│ Metrics, Tuning │
└────────┬────────┘
         │
┌────────▼────────┐
│ Visualization   │
│ Dashboard / API │
└─────────────────┘
```

- Provide a high-level overview of the architecture of your project.
  - Use a diagram to illustrate the key components and their interactions.
- Briefly describe each component shown in the diagram
  - Highlighting their roles and functionalities within the overall system.

6. Data Sources

- Health Data: Weekly malaria reports from Liberia’s Ministry of Health (CSV).
- Health Data from the National Public Health Institute of Liberia (NPHIL)

- Climate Data: Rainfall and temperature datasets from the World Bank Climate Change Knowledge Portal.

- Population Data: County/district-level statistics from the Liberia Institute of Statistics.

7. Literature Review

Introduction

Malaria continues to cause over 247 million cases and 619,000 deaths annually, mostly in Sub-Saharan Africa (WHO, 2023). Outbreak prediction is vital for timely interventions like mosquito net distribution and spraying. Reviewing past studies ensures our project builds on proven methods while addressing existing gaps.

Thematic Review

1. **Climate & Environment:** Rainfall, humidity, and temperature strongly influence malaria (Pascual et al., 2006; Caminade et al., 2014). However, climate-only models ignore health and demographic data.

2. **Machine Learning & AI:** ML models (e.g., Random Forest, Deep Learning) show higher predictive accuracy than traditional methods (Loha et al., 2019; M’boga et al., 2021) but need large, high-quality datasets.

3. **GIS & Visualization:** Mapping tools (Machault et al., 2011; Gething et al., 2016) help identify hotspots and guide interventions but rely on complete geocoded health data, which is often missing in rural areas.

4. **Challenges & Gaps:** Most models are academic, depend on historical data, and lack real-time, community-level deployment.

Conclusion

Climate factors are strong predictors, AI enhances accuracy, and GIS improves usability. Yet, there is no **integrated, real-time, community-based malaria prediction system**. Our project addresses this by combining **real-time weather data, AI prediction models, and GIS dashboards**, supporting **SDG 3: Good Health and Well-Being**.

Implementation Plan

1. Technology Stack

Technology/Tool	**Use / Purpose**
Python	Main programming language for data processing, ML model development
Pandas & NumPy	Data cleaning, preprocessing, and manipulation

Scikit-learn	Machine learning models (Random Forest, SVM)
XGBoost / LightGBM	Advanced predictive modeling for high accuracy
TensorFlow / PyTorch	Deep learning models (optional, for neural network-based prediction)
Matplotlib / Seaborn	Data visualization and exploratory data analysis
Plotly / Dash / D3.js	Interactive GIS-based dashboards and maps
PostgreSQL + PostGIS	Database for storing processed data and geospatial information
Google Earth Engine / WorldClim	Climate and environmental data acquisition
WHO / DHS / WorldPop datasets	Malaria incidence, population, and demographic data
AWS / Google Cloud Platform	Cloud hosting, data storage, and scalable computation
Git / GitHub	Version control and collaborative coding
Docker	Containerization for deployment
Optional: Arduino / IoT Sensors	If integrating local real-time environmental sensors (temperature, humidity)

2. **Timeline**

- **12 Weeks (3 months)**

Week	Task	Deliverable
1-2	Collect & Preprocess Data	Clean Dataset
3-4	Exploratory Data Analysis	Data Insight
5-6	Feature engineering	Feature set
7-8	Train baseline models	Initial model
9-10	Tune & evaluate models	Optimized results
11	Build Dashboard/API	Prediction Tool

12	Final testing & submission	Project Report

Task Distribution Matrix:
Note: We have been two active members on this initiative in group 7.

- Thomas S. McCay: Data collection, preprocessing, EDA
- David Paye: Model development, dashboard creation
- Both Thomas and David: Evaluation, report writing, presentation

3. Milestones

Project Milestones

1. **Dataset Collected, Cleaned & Prepared**
   - All climate, population, and malaria incidence data collected, missing values handled, units standardized, and dataset ready for modeling.
2. **Baseline Model Trained**
   - Initial ML models (Random Forest, XGBoost) trained using the prepared dataset.
3. **Target Performance Achieved**
   - Optimized model reaches ≥80% accuracy or ROC-AUC score, validated on test data.
4. **Dashboard/API Prototype Developed**
   - Interactive visualization dashboard showing predicted malaria hotspots and trends, integrated with model outputs.
5. **Final Testing, Submission & Presentation**
   - Complete system tested, final report written, and group presentation prepared.

4. Challenges and Mitigations

Challenge	Mitigation Strategy
Data Quality	- Missing or inconsistent data may reduce model accuracy. - Mitigation: Use data cleaning, imputation methods, and cross-validation to handle missing values. Combine multiple reliable sources (NASA, WHO, WorldPop) for robustness.

Model Performance	- Risk of low accuracy or overfitting with limited data. - Mitigation: Use ensemble methods (Random Forest, XGBoost), hyperparameter tuning, and regularization techniques. Evaluate models with multiple metrics (Accuracy, ROC-AUC, Precision, Recall).
Technical Constraints	- Limited computing power or software issues may slow down training. - Mitigation: Use cloud computing platforms (AWS, Google Cloud) for scalable resources. Containerize deployment using Docker to ensure reproducibility.
Time Management	- Personal commitments may delay project tasks. - Mitigation: Follow the 14-week timeline with buffer weeks, parallelize tasks where possible, and hold weekly progress meetings to stay on track.
Dashboard/Visualization Complexity	- Building interactive GIS dashboards may be challenging. - Mitigation: Use prebuilt libraries (Plotly, Dash, D3.js) and focus on core functionality first; extend features iteratively.

5. Ethical Considerations

1. **Data Privacy:**

   - All health and demographic data used will be **anonymized and aggregated** to prevent identification of individuals.

   - Only publicly available datasets (WHO, DHS, WorldPop, NASA) will be used. No private patient records will be accessed.

2. **Bias and Fairness:**

   - Machine learning models can unintentionally favor certain regions or populations if training data is uneven.

   - **Mitigation:** Ensure datasets cover a representative range of communities, validate model predictions across different regions, and regularly check for biases in outputs.

3. **Impact on Target Community:**

   - Predictions may influence public health decisions. **False positives or negatives** could lead to unnecessary panic or missed interventions.

   - **Mitigation:** Present model predictions with confidence levels and disclaimers; design the dashboard for **decision-support**, not as the sole authority for interventions.

4. **Transparency and Accountability:**

- Model methodology, assumptions, and limitations will be clearly documented for users.
- Decision-makers will be trained on interpreting outputs responsibly.

5. **Equity and Accessibility:**
    - Ensure the system is **accessible to rural health workers**, even with low-bandwidth internet or limited computing resources.

6. References

World Health Organization (WHO). (2023). *World Malaria Report 2023.* Retrieved from https://www.who.int/publications/i/item/9789240195423

Pascual, M., Ahumada, J. A., Chaves, L. F., Rodo, X., & Bouma, M. (2006). *Malaria resurgence in the East African highlands: Temperature trends revisited.* Proceedings of the National Academy of Sciences, 103(15), 5829–5834.

Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., ... & Lloyd, S. J. (2014). *Impact of climate change on global malaria distribution.* Proceedings of the National Academy of Sciences, 111(9), 3286–3291.

Loha, E., Deressa, W., & Lindtjørn, B. (2019). *Machine learning methods for predicting malaria incidence in Ethiopia.* Malaria Journal, 18(1), 1–12.

M’boga, P., Liywali, J., & Kisia, D. (2021). *Deep learning-based malaria prediction using climate and population data.* International Journal of Health Geographics, 20(1), 1–12.

Machault, V., Vignolles, C., Pagès, F., Vazeille, M., Failloux, A. B., & Roger, P. (2011). *GIS mapping of malaria transmission hotspots in Africa.* Malaria Journal, 10(1), 1–11.

Gething, P. W., Casey, D. C., Weiss, D. J., & Tatem, A. J. (2016). *Mapping Plasmodium falciparum prevalence globally using spatial-temporal models.* Nature, 536(7615), 203–207.

NASA POWER Project. (2023). *Global Climate Data for Environmental and Health Studies.* Retrieved from https://power.larc.nasa.gov/

WorldPop. (2023). *High-resolution population datasets for Africa.* Retrieved from https://www.worldpop.org/