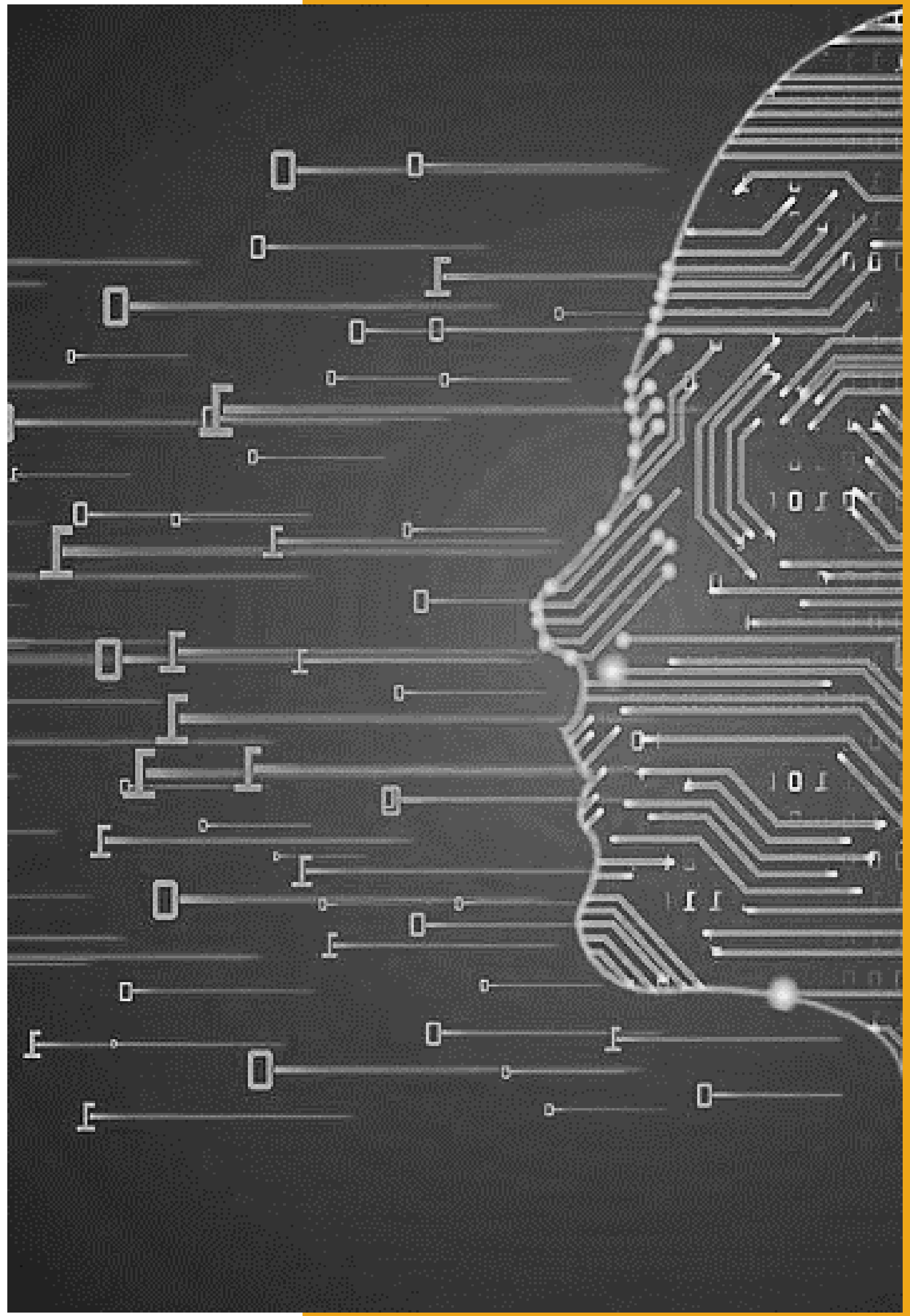
Data Science Without Borders

APHRC African Population and Health Research Center
Transforming lives in Africa through research.

Hackathon Orientation: Materials & Methods

Co-creating AI Solutions for Douala General Hospital



11 Jully, 2025 | FOMAZOU TCHINDA CYRILLE BRICE

Overview of Tools

Concern	Python	Java	PHP
Service Framework	**FastAPI** (async, auto-docs)	**Spring Boot** (standalone JARs)	**Laravel Lumen** (micro-framework)
Dependency Mgmt	pip / Poetry	Maven / Gradle	Composer
Containerization	**Docker**	**Docker**	**Docker**
Local Orchestration	**Docker Compose**	**Docker Compose**	**Docker Compose**
Database	**PostgreSQL, MongoDB**	**PostgreSQL, MongoDB**	**PostgreSQL, MongoDB**
API Gateway / Proxy	**Traefik**	**Traefik**	**Traefik**
CI / CD	**GitHub Actions**	**GitHub Actions**	**GitHub Actions**

Materials Provided

- GitHub Organization & Starter Repositories
- Deployment Accounts on Vercel, Netlify, Heroku
- Access to Cloud Resources (APIs, Storage)
- Documentation & Starter Templates

Methods & Workflow

- • Fork & Clone Repositories
- • Branching Strategy: Feature Branches & Pull Requests
- • Automated Deployments via CI/CD Pipelines
- • Monitoring & Logging Setup

Synthetic Dataset Overview

- • Files: patient_feedback, clinical_summaries, blood_bank_records (CSV)
- • Record Count: 50,000 entries each (12-month period: Jul 2024–Jun 2025)
- • Variables: Demographics, vitals, wait/resolution times, department/sites, hemoglobin levels, etc.
- • Data Quality: 10% missing values injected randomly
- • Outliers: 1% extreme values in numeric fields
- • Formats: CSV files with JSON metadata descriptor

Checkpoints Schedule

- Module 1 (Patient Feedback Analytics): Due Fri, 18 July 2025
- Module 2 (LLM Chatbot Integration): Due Fri, 25 July 2025
- Module 3 (Blood Bank Dashboard): Due Fri, 1 August 2025
- Minimum requirements per module must be met
- Non-compliance => Disqualification

Next Steps & Q&A

- Confirm repository access & team formation
- Download & inspect synthetic dataset
- Plan feature engineering & modeling approach
- Prepare deliverables for first checkpoint

Presentation outline

01 Background and Problem Statement

02 Research Questions and Objectives

03 Methodology and Available/Potential

04 Summary of Progress So Far

05 Perspectives and Next Steps

06 Conclusion

07 References

Introduction

Data Science Without Borders

Global Health Challenges:

- Malaria, tuberculosis (TB), and anemia are major health challenges in low-resource settings.
- In 2023, malaria alone accounted for approximately **597,000 deaths** worldwide (WHO, 2023).
- TB caused about **1.5 million deaths** in 2020 (WHO, 2021).
- Anemia affects roughly **1.62 billion people** globally (24.8% of the population), with prevalence as high as **42% in children under five** and **30% in women of reproductive age** (WHO, 2021).

Diagnostic Limitations:

- Current methods require specialized equipment and trained personnel, which are often lacking in resource-limited settings.

AI in Healthcare:

- Deep learning models show promise but are typically disease-specific.
- Integrated approaches for multi-disease diagnosis are lacking.

Co-occurrence and Complications:

- Malaria and anemia frequently co-occur, with malaria being a leading cause of anemia in endemic regions.
- Co-infection with TB further complicates clinical presentations and treatment strategies.

Rresearch Question & Objectives

Rresearch Question

Can a unified deep learning framework accurately diagnose malaria, TB, and anemia from medical images in resource-limited settings?

Primary Objective

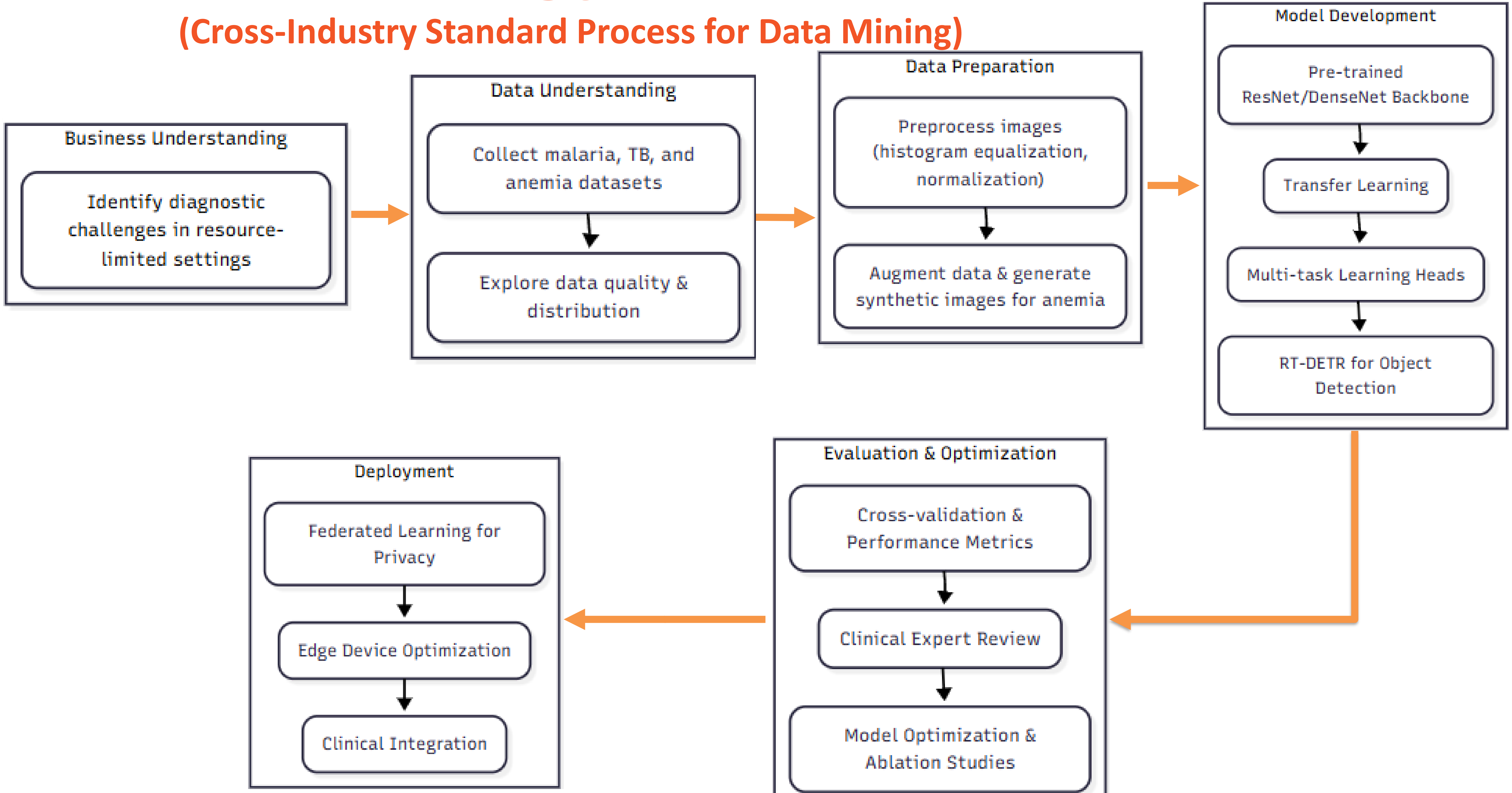
Develop an integrated deep learning framework that diagnoses malaria, TB, and anemia using multi-task learning and synthetic data augmentation.

Specific Objectives

1. Develop and optimize a multi-task learning model that integrates detection of malaria, TB, and anemia using microscopy and radiographic images.
2. Enhance feature extraction and localization by integrating transformer-based architectures (e.g., RT-DETR).
3. Generate high-fidelity synthetic blood smear images using GANs, VAEs, and Diffusion Models to address anemia data scarcity.
4. Implement federated learning and fairness-aware AI strategies for privacy preservation and bias mitigation.
5. Deploy and evaluate the framework in real-world settings, benchmarking against state-of-the-art models with clinical validation

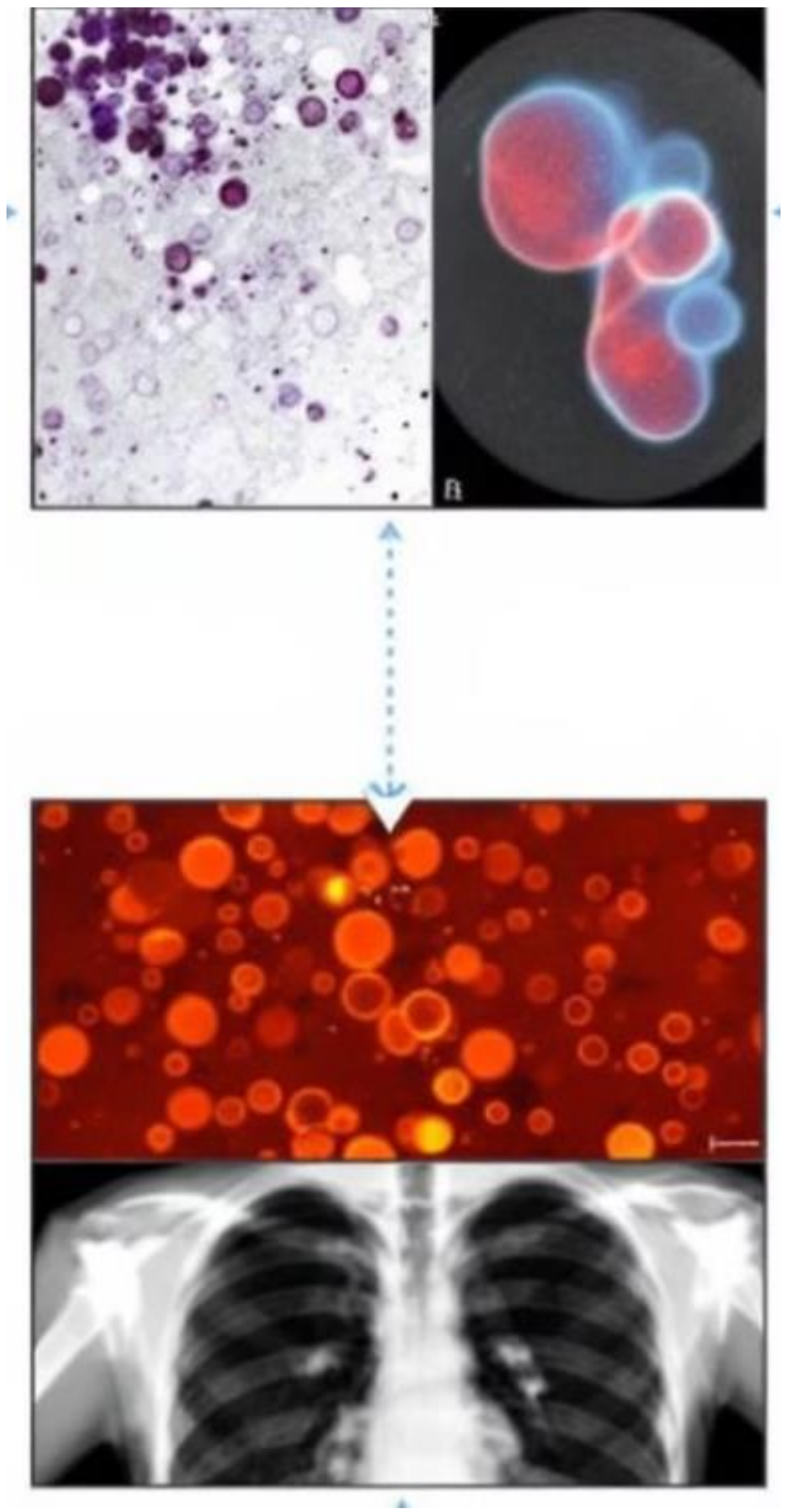
Methodology: CRISP-DM

(Cross-Industry Standard Process for Data Mining)



Potential Datasets

Dataset	Description	Source / Access
NIH Malaria Dataset	Microscopy images of thick and thin blood smears showing parasitized and uninfected red blood cells; includes parasite detection and, when available, stage and species annotations.	National Library of Medicine (NLM): NIH Malaria Project
Shenzhen Chest X-ray Dataset	Chest X-ray images, including normal and TB-positive cases, with annotations indicating TB-related abnormalities.	Publicly available; often sourced from Kaggle or institutional repositories.
Montgomery County Chest X-ray Dataset	A smaller dataset with annotated chest X-rays from a U.S. county, including TB-positive and normal images.	Publicly available; accessible via platforms such as Kaggle.
Synthetic Anemia Dataset	Synthetic blood smear images generated using GANs, VAEs, and Diffusion Models to simulate anemia conditions.	To be generated as part of this project; validation through clinical Turing tests and quantitative metrics (FID, SSIM, Dice).



Summary of Progress So Far

Malaria Module Development:

Multi-task model for malaria detection—including parasite detection, stage classification, and species identification—is nearly complete, with initial testing showing promising accuracy.

Literature Review

Preliminary reviews for tuberculosis (TB) and anemia diagnosis have been conducted, identifying key methodologies and research gaps.

Dataset Exploration

Malaria and TB datasets have been collected and preprocessed.

Exploration for synthetic data generation for anemia is ongoing to address data scarcity.

Model Prototyping

Early prototypes of the multi-task learning framework have been implemented, with CNN-based feature extraction successfully integrated.

Integration with transformer-based architectures (RT-DETR) is in progress.

Perspectives and Next Steps

Short-Term Goals

- Finalize and integrate the TB and synthetic anemia modules with the existing malaria module.
- Complete and validate the synthetic data generation pipeline for anemia using GANs/VAEs/Diffusion Models.
- Enhance transformer-based model integration (RT-DETR) for improved localization and feature extraction.
- Conduct further internal testing and ablation studies to fine-tune model components.

Mid-Term Goals

- Implement federated learning protocols to enable collaborative training across multiple healthcare centers while preserving privacy.
- Expand dataset exploration and preprocessing to include additional clinical data for TB and anemia.
- Initiate pilot testing in collaboration with clinical partners to gather real-world feedback.

Long-Term Goals

- Deploy the complete diagnostic framework on edge devices for real-time, point-of-care applications.
- Iteratively refine the system using an active learning loop with continuous clinician feedback.
- Pursue large-scale clinical trials to benchmark the framework against existing diagnostic methods and secure regulatory approvals.

Conclusion

This research introduces a **next-generation AI diagnostic framework** that combines **RT-DETR (transformer-based architectures)**, **synthetic data generation**, and **multi-task learning** to enable real-time, simultaneous diagnosis of **malaria, tuberculosis (TB), and anemia**. By tackling critical challenges such as **data scarcity**, **ethical AI deployment**, and **real-world scalability**, this framework has the potential to **transform global health diagnostics**, particularly in resource-limited settings. Its innovative approach not only enhances diagnostic accuracy but also ensures accessibility, privacy, and fairness, paving the way for more equitable and effective healthcare solutions worldwide.

References

Reference	Description
Gao, X., et al. (2020) Multi-task learning for diagnosing multiple diseases from a single chest X-ray image. IEEE Access, 8, 55588–55599.	Demonstrates a multi-task approach for integrated disease diagnosis from X-rays.
Rajpurkar, P., et al. (2020) CheXpedition: Investigating generalized deep learning for chest X-ray diagnosis. arXiv:2002.11379.	Explores generalized deep learning for chest X-ray analysis, supporting transfer learning strategies.
Carion, N., et al. (2020) End-to-end object detection with transformers (DETR). ECCV Proceedings.	Introduces transformer-based object detection, forming the basis for RT-DETR integration.
Dosovitskiy, A., et al. (2020) An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929.	Establishes the potential of Vision Transformers for complex image recognition tasks.
Dhariwal, P., & Nichol, A. (2021) Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 34, 8780–8794.	Provides evidence that diffusion models can generate high-fidelity synthetic images, crucial for augmenting scarce datasets (e.g., for anemia).
Mehrabi, N., et al. (2021) A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.	Addresses bias and fairness challenges in AI, essential for ethical, equitable model deployment.
Kairouz, P., et al. (2021) Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.	Discusses federated learning techniques for privacy-preserving collaborative model training.
Rieke, N., et al. (2020) The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1–7.	Demonstrates practical applications of federated learning in healthcare, ensuring data privacy during collaborative training.

Partners

wellcome

Ahri
Armauer Hansen Research Institute
Since 1970
... for better health

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

IRESSEF
Institut de Recherche en Santé de Surveillance Epidémiologique et de Formation

AfricaCDC
Centres for Disease Control and Prevention

OSPO now

HOPITAL GENERAL DE DOUALA
DOUALA GENERAL HOSPITAL

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

APHRC
African Population and Health Research Center
Transforming lives in Africa through research.

Makerere AI Lab

Thank You

Data Science
Without Borders