Capstone Project Concept Note and Implementation Plan

Project Title**:**

**Early Detection of Common Diseases System in Afghanistan Using Machine Learning**

Team Members**:**

1. **Samiullah Gulzar**
2. **Homa Akrami**
3. **Hashmatullah Asady**
4. **Fardaws Karimi**
5. **Baiqra Muradi**

Table of Contents

[**Project Title:** 1](#_Toc187222029)

[**Team Members:** 1](#_Toc187222030)

[Concept Note 1](#_Toc187222031)

[1. Project Overview 1](#_Toc187222032)

[2. Objectives 1](#_Toc187222033)

[3. Background 1](#_Toc187222034)

[4. Methodology 1](#_Toc187222035)

[5. Architecture Design Diagram 2](#_Toc187222036)

[6. Data Sources 2](#_Toc187222037)

[7. Literature Review 2](#_Toc187222038)

[Implementation Plan 3](#_Toc187222039)

[1. Technology Stack 3](#_Toc187222040)

[2. Timeline 3](#_Toc187222041)

[3. Milestones 3](#_Toc187222042)

[4. Challenges and Mitigations 3](#_Toc187222043)

[5. Ethical Considerations 4](#_Toc187222044)

[6. References 4](#_Toc187222045)

# Concept Note

## 1. Project Overview

Afghanistan faces critical healthcare challenges, including inadequate access to medical facilities, a shortage of skilled professionals, and high rates of preventable diseases such as malnutrition, diarrhea, and respiratory infections. These issues are exacerbated in rural and underserved regions, where healthcare resources are scarce. To address this, our project aims to develop a machine learning-based system for the early detection of common diseases, focusing on improving public health outcomes and reducing mortality rates. This initiative aligns with **Sustainable Development Goal 3 (Good Health and Well-being)** by leveraging technology to provide accessible and effective healthcare solutions. The proposed system will utilize health data collected over the last decade to identify at-risk populations, enabling timely interventions and efficient resource allocation.

## 2. Objectives

* Develop a machine learning system capable of accurately predicting and detecting common diseases such as malnutrition, diarrhea, and respiratory infections.
* Reduce mortality rates and improve healthcare outcomes by enabling early interventions.
* Facilitate equitable healthcare access in underserved and rural areas of Afghanistan.
* Provide actionable insights to policymakers for improved healthcare planning and resource allocation.

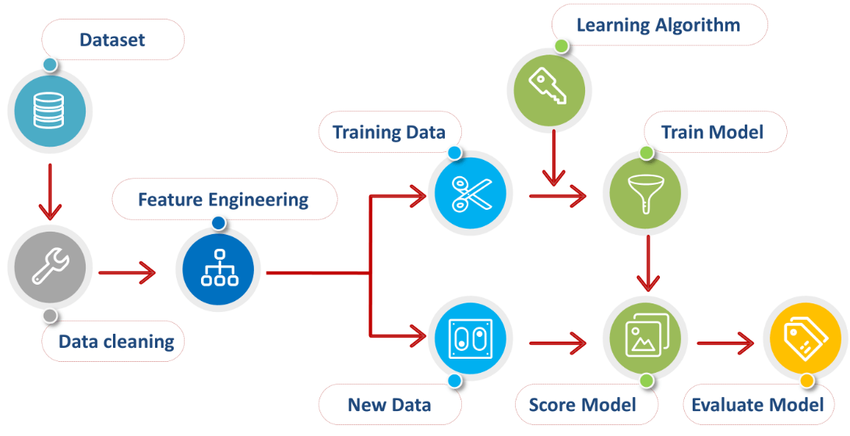
## 3. Background

The healthcare system in Afghanistan struggles to address the widespread prevalence of preventable diseases. Despite global advancements in medical technology, rural and underserved areas continue to face barriers in accessing timely and accurate diagnosis. Existing solutions rely on traditional approaches, which often lack scalability and efficiency. The integration of machine learning into healthcare systems offers a transformative approach by analyzing large datasets to uncover patterns and predict health outcomes. Machine learning not only improves diagnostic accuracy but also optimizes resource allocation, making it a vital tool for addressing the unique challenges of Afghanistan’s healthcare landscape.

## 4. Methodology

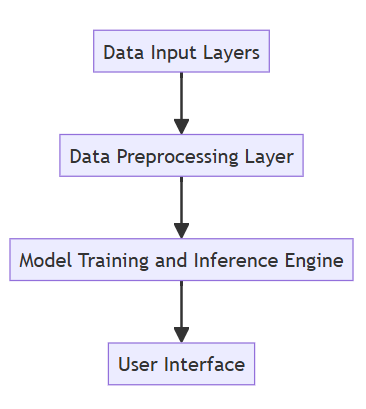
The proposed system will utilize supervised learning techniques such as Random Forest and Logistic Regression, alongside advanced algorithms like Support Vector Machines (SVMs). The methodology includes the following steps:

1. **Data Collection and Preprocessing:** Gather and clean health data from reliable sources, including WHO and OpenMRS.
2. **Feature Engineering:** Extract and select relevant features such as patient demographics, symptoms, and environmental factors.
3. **Model Training and Validation:** Train ML models on labeled datasets, using cross-validation techniques to ensure accuracy and robustness.
4. **Deployment:** Develop a user-friendly interface for healthcare workers and policymakers, enabling real-time predictions and insights.



## 5. Architecture Design Diagram

The system architecture consists of the following components:

1. **Data Input Layer:** Collects data from various sources, including electronic health records and public health surveys.
2. **Data Processing Layer:** Handles data cleaning, normalization, and feature extraction.
3. **Model Training and Inference Engine:** Trains machine learning models and provides disease predictions.
4. **User Interface:** Presents results and insights through dashboards and mobile applications.

## 6. Data Sources

The project will rely on the following data sources:

* **Afghanistan Demographic and Health Survey (DHS):** Provides comprehensive health-related data.
* **World Health Organization (WHO) Datasets:** Includes global and local health statistics.
* **OpenMRS:** Offers patient-level clinical data. These datasets include essential attributes such as patient demographics, symptoms, diagnoses, and outcomes, forming a robust foundation for predictive modeling.

## 7. Literature Review

Existing studies highlight the potential of machine learning in improving healthcare outcomes. For instance, Momand et al. (2023) demonstrated the efficacy of ML models in predicting dehydration among Afghan children, achieving high accuracy. Similarly, Gupta et al. (2022) showcased an ML-based tuberculosis prediction model tailored for resource-constrained settings. These findings underscore the feasibility and transformative potential of leveraging machine learning for early disease detection in Afghanistan.

# Implementation Plan

## 1. Technology Stack

* **Programming Languages:** Python
* **Libraries and Frameworks:** Scikit-learn, TensorFlow, Keras, Pandas, NumPy, Matplotlib
* **Tools:** Jupyter Notebook, Google Colab for development and Docker for deployment.
* **Databases:** SQLite for structured storage of preprocessed data.
* **Hardware Requirements:** GPUs for training deep learning models (e.g., NVIDIA or equivalent)

## 2. Timeline

|  |  |  |
| --- | --- | --- |
| Stage | Tasks | Time line |
| Data Collection and Cleaning | Gather and Preprocess datasets | 2 weeks |
| Feature Engineering | Select and engineer relevant features | 1 week |
| Model Development | Train and validate ML models | 3 weeks |
| Deployment | Build and test user interface | 2 weeks |

**Task Distribution:**

|  |  |
| --- | --- |
| Team Members | Task |
| Fardaws Karimi & Hashmatullah Asady | Focus on data collection and preprocessing and feature engineering |
| Homa Akrami, Samiullah Gulzar and Baiqra Muradi | Responsible for model selection, model training, validation and deployment. |

## 3. Milestones

* Completion of data collection and pre-processing
* Development of initial predictive models
* Validation achieving >85% accuracy on test datasets
* Deployment of a functional user interface

## 4. Challenges and Mitigations

* **Data Collection Problems**: Solve incompleteness and inaccessibility problems in data through the collection of diversified data from authentic sources like WHO and local health agencies.
* **Feature Engineering**: Lessen the difficulty in selecting appropriate features with domain experts' consultation and usage of automated techniques of feature selection.
* **Data Quality**: Clean the data for any missing values or inconsistencies by the use of different imputation techniques and outlier detection methods.
* **Model Performance**: Increase the performance or accuracy of models through hyperparameter tuning, using ensemble methods, and leveraging domain knowledge.
* **Technical Limitations:** Overcome hardware limitations by leveraging cloud computing resources to train larger and more efficient models.

## 5. Ethical Considerations

* **Data Privacy:** Ensure compliance with data protection regulations by anonymizing patient data.
* **Bias Mitigation:** Regularly evaluate models for biases to ensure fair predictions across diverse demographics.
* **Community Impact:** Engage with local stakeholders to ensure the system aligns with cultural and community needs.

## 6. References

* Momand et al. (2023). "Machine Learning for Predicting Dehydration in Children."
* Gupta et al. (2022). "AI-Driven Tuberculosis Prediction for Resource-Constrained Settings."
* World Health Organization (WHO). Health Statistics Database.
* Afghanistan Demographic and Health Survey (DHS).