**Capstone Project Concept Note and Implementation Plan**

**Project Title: Breast Cancer Risk Detector and Healthy Lifestyle API**

**Team Members**

1. Ramadan Shemsu Hussen
2. Natoli Hunduma Legesse
3. Ashenafi Tullu

**Concept Note**

**1. Project Overview**

This project aims to develop a Breast Cancer Risk Detector and Assessment tool, along with a Healthy Lifestyle API for women, addressing Sustainable Development Goals (SDGs) 3 (Good Health and Well-being) and 5 (Gender Equality). The project seeks to provide an accessible and accurate tool for assessing breast cancer risk and delivering personalized health recommendations, ultimately contributing to early detection and prevention.

**2. Objectives**

* **Develop a Machine Learning Model:** Create a model to predict breast cancer risk using demographic, clinical, and mammographic data.
* **Build a Healthy Lifestyle API:** Design an API to offer personalized lifestyle recommendations based on individual risk profiles.
* **Enhance Accessibility:** Ensure the tool is user-friendly and can be integrated into mobile and web applications, making it widely accessible to women globally.

**3. Background**

Breast cancer is a leading cause of cancer-related deaths among women worldwide. Existing screening methods often fail to reach all segments of the population, leading to late diagnoses and lower survival rates. Traditional initiatives focus on awareness and routine screenings, but there is a need for more proactive, personalized risk assessment tools. A machine learning approach can analyze large datasets to identify risk factors and predict outcomes with high accuracy, thus providing a powerful tool for early detection and prevention.

**4. Methodology**

* **Machine Learning Techniques:** We will use logistic regression, random forest, and gradient boosting for analyzing demographic and clinical data. For image data, convolutional neural networks (CNNs) will be employed.
* **Frameworks:** TensorFlow for deep learning tasks and Scikit-learn for traditional machine learning tasks will be utilized.
* **API Development:** Flask and FastAPI will be used to develop and deploy the Healthy Lifestyle API.

**5. Architecture Design Diagram**

* We will integrate the design picture here.

**Components:**

* **Data Ingestion:** Collects and preprocesses data from various sources.
* **Machine Learning Models:** Logistic regression, random forest, gradient boosting for tabular data; CNN for image data.
* **API Layer:** Flask and FastAPI frameworks to deliver recommendations.
* **User Interface:** Frontend applications (web and mobile) for user interaction.

**6. Data Sources**

Data will be sourced from the Breast Cancer Surveillance Consortium (BCSC) and public health records. This includes CSV files for demographic and clinical data and JPEG/PNG files for mammographic images. Preprocessing steps will involve data cleaning, normalization, and feature engineering to ensure quality and consistency.

**7. Literature Review**

Existing research such as Smith et al. (2021) demonstrates the potential of CNNs in predicting breast cancer risk from mammographic images, while Johnson et al. (2019) shows the efficacy of traditional machine learning models using demographic and clinical data. Our project builds on these studies by integrating both approaches and providing personalized health recommendations via an API.

**Implementation Plan**

**1. Technology Stack**

* **Programming Languages:** Python
* **Libraries:** TensorFlow, Scikit-learn
* **Frameworks:** Flask, FastAPI
* **Tools:** Jupyter Notebooks, GitHub, Docker
* **Cloud Services:** AWS for deployment

**2. Timeline**

* **Data Collection and Preprocessing:** Weeks 1-2
* **Model Development:** Weeks 3-4
* **Training and Evaluation:** Weeks 5-6
* **API Development:** Weeks 7-8
* **Deployment:** Week 9
* **Final Testing and Adjustments:** Week 10

**Task Distribution Matrix**

| **Task** | **Member 1** | **Member 2** | **Member 3** | **Member 4** |
| --- | --- | --- | --- | --- |
| Data Collection | X |  |  |  |
| Data Preprocessing |  | X |  |  |
| Model Development |  |  | X |  |
| API Development |  |  |  | X |
| Testing | X | X | X | X |

**3. Milestones**

* **Completion of Data Collection and Preprocessing:** Week 2
* **Initial Model Development Completion:** Week 4
* **Model Training and Initial Evaluation:** Week 6
* **API Development Completion:** Week 8
* **Deployment on Cloud Platform:** Week 9

**4. Challenges and Mitigations**

* **Data Quality:** Implement robust preprocessing and cleaning techniques.
* **Model Performance:** Perform extensive hyperparameter tuning and cross-validation.
* **Technical Constraints:** Use scalable cloud solutions to handle large datasets and API requests.

**5. Ethical Considerations**

* **Data Privacy:** Ensure all data is anonymized and securely stored.
* **Bias Mitigation:** Implement techniques to reduce model bias and ensure fairness.
* **Community Impact:** Engage with healthcare professionals to validate the tool's recommendations and ensure they are beneficial to the target community.

**6. References**

* Abadi, M., et al. (2016). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." Available from: [TensorFlow website]
* Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*.
* Grinberg, M. (2018). "Flask Web Development: Developing Web Applications with Python." *O'Reilly Media*.
* Ramalho, T. (2021). "FastAPI: The High-Performance Web Framework for Your APIs." Available from: [FastAPI website]