**Capstone Project Concept Note and Implementation Plan**

**Project Title: Machine Learning in Breast Cancer Diagnosis**

**Team Members**

1. M Zaker Humayoon AMIN
2. Hamda ABDI

**Concept Note**

#### 1. Project Overview

Breast cancer remains a significant global health challenge, with over 2.3 million new cases diagnosed annually, making it the second deadliest cancer worldwide [1], [2]. This capstone project addresses the critical need for early diagnosis and effective treatment to mitigate the risk of cancer recurrence. The proposed solution leverages machine learning (ML) to develop a predictive model for accurately classifying breast masses as benign or malignant. By analyzing features extracted from digitized fine needle aspirate (FNA) images, our project aims to contribute to early detection, allowing for timely intervention and improved health outcomes.

#### 2. Objectives

* Develop a predictive model using machine learning techniques to classify breast masses as benign or malignant.
* Improve diagnostic accuracy, enabling early detection of breast cancer.
* Contribute to the global effort to strengthen breast cancer diagnosis and treatment strategies.
* Mitigate the risk of cancer recurrence through timely and accurate classification.

#### 3. Background

Breast cancer poses a significant health threat, necessitating innovative approaches for early detection and treatment. Traditional diagnostic methods have limitations, prompting the exploration of machine learning as a complementary tool. Previous studies, such as Shetty's master's thesis [7] and Nguyen et al.'s research [8], showcase the effectiveness of machine learning algorithms, including Support Vector Machines and Random Forest Classifiers, in breast cancer analysis. This project builds upon the existing knowledge, focusing on digitized FNA images for enhanced diagnostic capabilities.

#### 4. Methodology

Machine Learning Techniques: Our project employs machine learning approaches, specifically Support Vector Machines, Random Forests, and Logistic Regression. These models are chosen for their effectiveness with relatively smaller datasets and their interpretability—critical factors in medical contexts. By mapping features extracted from FNA images to diagnostic outcomes, the machine learning model aims to learn and predict breast mass classification accurately.

#### 5. Architecture Design Diagram

#### 

Input Layer:

* **Description:** Dataset containing numerical representations of FNA images of breast masses.
* **Processing:** No processing required.
* **Representation:** Each image is represented as a set of numerical features.

Preprocessing Layer:

* **Description:** Preprocessing the dataset before model development (Processing).
* **Processing:** Data analysis and exploration, handling missing values and outliers, normalization, feature selection and etc.
* **Representation:** Cleaned data ready for processing.

Processing Layer:

* **Description:** Utilize Support Vector Machines (SVM), Random Forests, and Logistic Regression for classification.
* **Processing:**
  + **Support Vector Machines (SVM):** Train SVM models to find the optimal hyperplane that separates benign and malignant classes based on extracted features from FNA images.
  + **Random Forests:** Develop an ensemble of decision trees to collectively classify breast masses by considering various features from the digitized images.
  + **Logistic Regression:** Train a logistic regression model to predict the probability of a breast mass being malignant based on the selected features.

Output Layer:

* **Description:** Binary classification output - Benign or Malignant.
* **Processing:** Based on the models' predictions, classify each breast mass as either benign or malignant.
* **Representation:** Binary output indicating the classification result.

**6. Data Sources**

The data source that will be used is Breast Cancer Wisconsin (Diagnostic) Data Set [3]. This Dataset is tailored for predicting the benign or malignant nature of breast cancer masses. It encompasses features derived from digitized fine needle aspirates (FNAs) images, detailing characteristics of cell nuclei present in the images

The relevance of this dataset to the project lies in its pivotal role for predicting the benign or malignant nature of breast cancer masses. These features, ranging from radius and texture to concavity and fractal dimension, provide crucial insights into the composition of breast masses. As part of preprocessing, it is imperative to conduct exploratory data analysis (EDA) to understand feature distributions and relationships. Moreover, given the imbalanced class distribution with 357 benign and 212 malignant samples, careful consideration of model evaluation metrics, such as precision, recall, and F1-score, is recommended to ensure the robustness of the breast cancer detection model. The absence of missing attribute values and the well-documented nature of the dataset contribute to its data integrity, making it a valuable resource for machine learning endeavors in breast cancer diagnosis.

**7. Literature Review**

The existing literature in the field of breast cancer diagnosis using machine learning, as explored in Shetty's [4] master's thesis and the work by Cuong Nguyen et al. [5], emphasizes the significance of predictive models in early-phase identification and treatment. From those studies our project aims to leverage machine learning techniques, specifically Support Vector Machines, Random Forests, and Logistic Regression, to develop a predictive model for classifying breast masses as benign or malignant. The methodology aligns with the interpretability requirements in medical contexts, crucial for transparent decision-making. the work by Cuong Nguyen et al., which focuses on feature selection and the use of a Random Forest classifier, contributes insights into optimizing breast cancer diagnosis and prognosis. Our project extends this work by concentrating on digitized FNA images, providing a more detailed and nuanced analysis of cell nuclei features for enhanced diagnostic accuracy. In summary, our project integrates and extends the key findings from existing research to contribute to the advancement of machine learning-based breast cancer diagnosis.

**Implementation Plan**

**1. Technology Stack**

* Programming Language:
  + Python will serve as the primary programming language for our project.
* Libraries:
  + NumPy: Utilized for efficient data storage and manipulation.
  + Pandas (Dataframe Library): Employed for structured data handling and analysis.
* Machine Learning Framework:
  + Scikit-learn (sklearn): Chosen for creating, training, and evaluating machine learning models.
* Visualization:
  + Matplotlib, Plotly, Seaborn: These libraries will be integrated for data visualization, aiding in the interpretation of model results and insights.
* Code Editor:
  + Visual Studio Code (VSCode): Selected as the code editor for its robust features and versatility.
* Notebook Environment:
  + Jupyter Notebook: To facilitate interactive and collaborative development and documentation of the project.

**2. Timeline**

Data Collection and Preprocessing

* **Tasks:**
  + Acquire Breast Cancer Wisconsin (Diagnostic) Data Set.
  + Perform initial data exploration and analysis.
  + Handle any missing values and outliers.
  + Normalize or standardize features as necessary.
  + Correlation Analysis and Univariate Feature Selection

Model Development

* **Tasks:**
  + Split the dataset into training and testing sets.
  + Implement machine learning models (Support Vector Machines, Random Forests, Logistic Regression).
  + Train and validate models using the training set.
  + Fine-tune hyperparameters for optimal performance (Grid Search).
  + Conducting further feature selection using techniques such as Regularization for Logistic Regression and Tree-Based Methods by utilizing models such as Random Forests.

Training and Evaluation

* **Tasks:**
  + Evaluate models using the testing set.
  + Assess performance metrics such as accuracy, precision, recall, and F1-score.
  + Iteratively refine models based on evaluation results.

Deployment

* **Tasks:**
  + Select the best-performing model for deployment.
  + Develop a deployment strategy, considering scalability and interpretability.
  + Create documentation for model deployment.

**3. Milestones**

* Identify key milestones in your project's development.
  + These could be related to the completion of specific tasks, the achievement of certain performance metrics, or the successful implementation of key features.
* Data pre-processing tasks are all done.
* Other tasks from Model Development and Training and Evaluation are also done. The current task at hand is iteratively improving the model.

**4. Challenges and Mitigations**

* The biggest challenge is not having too big of a dataset. Although what we have is clean and the models that have been developed so far has shown good results with high accuracy, there is just not enough data points to test the models.

**5. Ethical Considerations**

* The dataset doesn’t have any data regarding sex, race or ethnicity. It gives no personal information about the patients to be breached or misused. And the dataset is public.

**6. References**

[1] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, “Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review,” *Cancers 2021, Vol. 13, Page 4287*, vol. 13, no. 17, p. 4287, Aug. 2021, doi: 10.3390/CANCERS13174287.

[2] S. Paul, P. P. Solanki, U. P. Shahi, and S. Srikrishna, “Epidemiological Study on Breast Cancer Associated Risk Factors and Screening Practices among Women in the Holy City of Varanasi, Uttar Pradesh, India,” *Asian Pac J Cancer Prev*, vol. 16, no. 18, pp. 8163–8171, 2015, doi: 10.7314/APJCP.2015.16.18.8163.

[3] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

[4] Shetty, S. (2020). *Breast Cancer Analysis and Prognosis Using Machine Learning* (Doctoral dissertation, Dublin, National College of Ireland).

[5] Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. doi: [10.4236/jbise.2013.65070](https://doi.org/10.4236/jbise.2013.65070)