**ADS508 Data Science With Cloud Computing**

**Design Document**

**Authors:** Arjun Venkatesh, Darren Chen, Vinh Dao  
**Company Name:** AdvanceHC Solution  
**Company Industry:** Healthcare AI  
**Company Size:** Startup

**Abstract:**

AdvanceHC Solution is developing an AI-powered diagnostic tool to detect breast cancer at an early stage using advanced machine learning models. The challenge we face is that our product is still in the testing phase, making it difficult to gain traction with healthcare providers and regulatory bodies.

**Problem Statement:**

Breast cancer remains one of the leading causes of mortality among women worldwide. Early detection significantly improves survival rates, yet many current screening methods rely on manual radiology analysis, which can be time-consuming, prone to human error, and inaccessible in certain regions. Our company, AdvanceHC Solution, aims to develop an AI-powered tool that can detect breast cancer at an advanced level with higher accuracy and efficiency. However, as a startup, we face a major challenge: our AI model is still in the trial phase, and hospitals and healthcare providers are hesitant to adopt a product that has not been fully validated.

To overcome this, we need a robust dataset and validation process that proves the reliability of our model. Additionally, we must address regulatory concerns, secure partnerships with medical institutions, and ensure our AI system remains explainable and ethical in decision-making.

**Goals:**

1. Develop and test a machine learning model that can detect breast cancer from mammogram images with high accuracy.
2. Validate our AI model against existing radiologist diagnoses to prove its reliability.
3. Store and process data securely while adhering to healthcare privacy regulations.
4. Build trust with healthcare providers by demonstrating transparent AI decision-making.
5. Create a strategy for commercializing the product, focusing on partnerships with hospitals and research institutions.

**Non-Goals:**

1. We are not developing an AI system that replaces doctors but rather one that assists them in diagnosis.
2. The focus is on breast cancer detection only, not treatment recommendations.
3. We are not handling direct patient interactions or building a patient-facing application.
4. No real-time diagnosis during the testing phase—this is a research-oriented project.
5. Regulatory approvals (e.g., FDA, CE marking) are outside the scope of this phase.

**Data Sources:**

* **Primary Source:** Kaggle breast cancer imaging datasets.
  + <https://www.kaggle.com/datasets/piotrgrabo/breastcancerproteomes/data>
  + (*Breast Cancer Proteomes*, n.d.)
* **Data Storage:** AWS S3 for secure storage and easy access. (*AWS Educate - Cloud Skills for Education- AWS*, n.d.)
* **Risks:** Data quality and potential bias in training sets (e.g., underrepresentation of certain demographics).

**Data Exploration:**

* Check for missing or corrupted images and remove any anomalies.
* Analyze dataset distribution (age, ethnicity, cancer stages) to avoid bias.
* Identify key image features and metadata (e.g., tumor size, shape, density).
* Ensure ethical considerations are met in data usage.

**Storage & Tools:**

* Data ingestion and exploration using **AWS SageMaker Studio Notebooks**.
* Code repository stored in **GitHub** for collaboration.

**Data Preparation:**

* **Data Scrubbing:** Remove duplicate or low-quality images.
* **Feature Selection:** Focus on image characteristics and exclude irrelevant metadata.
* **Feature Creation:** Augment data (e.g., rotation, contrast enhancement) to improve generalization.
* **Feature Transformation:** Normalize pixel values and apply image preprocessing techniques.
* **Balancing:** Address class imbalance using oversampling techniques.
* **Splitting:** 80% training, 10% validation, 10% testing.

**Model Training:**

* **Training Approach:** Use AWS SageMaker built-in algorithms and fine-tune pre-trained CNN models (e.g., ResNet, EfficientNet).
* **Parameters:** Learning rate, batch size, number of epochs optimized using hyperparameter tuning.
* **Instance Size:** GPU-based SageMaker instances for efficient model training.
* **Evaluation:** Performance measured using accuracy, precision, recall, and AUC-ROC score.

**Measuring Impact:**

* **Accuracy Improvement:** Increase detection accuracy beyond 90% compared to human radiologists.
* **Reduction in Diagnostic Time:** Decrease time required for diagnosis by at least 50%.

**Security Checklist, Privacy, and Risks:**

* **PHI Data Handling:** No direct PHI storage; images anonymized before processing.
* **PII Data Handling:** Ensure compliance with HIPAA/GDPR for data privacy.
* **Bias Consideration:** Regular audits for dataset fairness and transparency in AI decision-making.
* **Ethical Concerns:** AI should not make independent decisions but assist radiologists.

**Future Enhancements:**

1. Expand dataset sources to include real-world hospital imaging data.
2. Integrate explainable AI techniques to provide justifications for model decisions.
3. Develop a real-time AI-assisted diagnostic tool for clinical use.

**References:**

*Breast Cancer Proteomes*. (n.d.). Retrieved March 9, 2025, from<https://www.kaggle.com/datasets/piotrgrabo/breastcancerproteomes>

*AWS Educate—Cloud Skills for Education- AWS*. (n.d.). Retrieved March 9, 2025, from<https://aws.amazon.com/education/awseducate/>