Practicum Sprint 2 - Project Planning

Matthew Agard
magard3@gatech.edu

1 PROJECT TOPIC

1.1 Background

The U.S. Census Bureau projects that by 2050, about half of all patients seen in the U.S. healthcare system will have skin of color (SOC). [1] These SOC patient groups include African-Americans, Asians, Hispanics, Native Americans and Pacific Islanders. As these communities continue to grow, so will the likelihood of dermatologists encountering cutaneous (skin) diseases that happen more frequently in SOC patients, happen exclusively in SOC patients, and/or present differently in SOC patients than their White counterparts. [1] I believe aiding dermatologists in their understanding, and more importantly diagnoses, of SOC patients' cutaneous disease presentations is paramount to delivering life-saving quality of care to these communities. As a result, my practicum project will explore the use of computer vision for detection of cutaneous diseases in SOC patients.

1.2 Justification

The racial disparities in skin cancer outcomes is an apt example for highlighting the importance of this project. "The 5-year melanoma survival [rate] is 74.1% for blacks compared to 92.9% for whites", despite melanoma being more common in non-Hispanic whites than blacks and hispanics. [2] Additionally, blacks tend to present with later stage or more aggressive non-melanoma skin cancer (NMSC), such as squamous cell carcinoma, than whites despite having lower incidence of NMSC. [2]

1.3 Solution

My project will be comprised of two deliverables. The first will be a data workflow in AWS, beginning with data ingestion into the cloud environment and ending with warehousing the cloud data in preparation for use by my computer vision model. My model training, testing, and evaluation will be conducted for display in an AWS SageMaker research notebook; this will serve as the second and final project deliverable.

2 TECHNICAL DESIGN

2.1 Tools/Technology

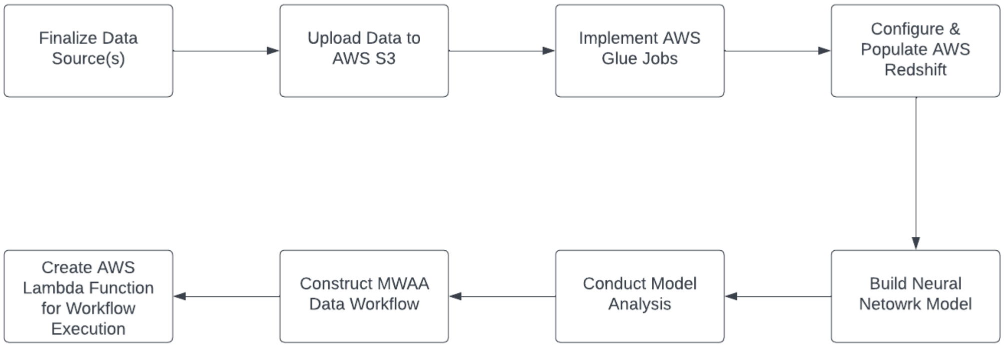
- AWS S3
- AWS Glue
- PySpark / Pandas & NumPy
- AWS Redshift
- AWS SageMaker
- Tensorflow
- Apache Airflow (MWAA) / AWS Step Functions
- AWS Lambda

2.2 Datasets & Data Sources

My dataset of choice is the Diverse Dermatology Images (DDI) Dataset [3] curated by the Stanford University School of Medicine. It contains 656 images of various skin diseases and how they present across the Fitzpatrick Skin Types (FST): fairer skin tones (Types I-II), medium skin tones (Types III-IV), and darker skin tones (Types V-VI).

Neural network models, which are used for computer vision tasks, generalize better on larger datasets, and I fear the original 656-record dataset doesn't meet this criteria. However, I intend to address this concern using Tensorflow's ImageDataGenerator function to augment the original images and artificially enhance the dataset's size. 70% of the dataset's records will be used for model training, with the remaining 30% being utilized for model testing and evaluation.

2.3 Architecture Diagram


Finalize Data Source(s)
Upload Data to AWS S3
Implement AWS Glue Jobs
Configure & Populate AWS Redshift
Build Neural Netowrk Model
Conduct Model Analysis
Construct MWAA Data Workflow
Create AWS Lambda Function for Workflow Execution

2.4 Screen Mock-up

As stated previously, the intended deliverables of this project are an AWS data workflow and an AWS SageMaker research notebook. Given the absence of any UI component, no mock-up will be provided in this document.

3 IMPLEMENTATION PLAN

Timeline	Task	Needs/Risks
Week 1	Finalize data source(s), upload to AWS S3	Need to ensure dataset(s) of choice is readily accessible; Dataset may have imbalanced skin color representation
Week 2	Develop AWS Glue jobs	Potential for complexity in learning PySpark framework, fallback plan is to use Pandas & NumPy
Week 3	Develop AWS Glue jobs	Potential for complexity in learning PySpark framework, fallback plan is to use Pandas & NumPy
Week 4	Configure AWS Redshift data warehouse	Redshift won't be needed if data used comes from a single source. If needed, there's potential for complexity in learning Redshift technology
Week 5	Build neural network model in AWS SageMaker	Perform exploratory data analysis (EDA), build predictive model with Tensorflow library (high probability of complexity in learning Tensorflow)
Week 6	Conduct model analysis in AWS SageMaker	Learning of Tensorflow library to build model may spill over into Week 6. If model construction & analysis is completed with time to spare, exploration of additional evaluation metrics/statistical tests will be included
Week 7	Construct MWAA data workflow for previous tasks	High probability of complexity in learning MWAA, fallback plan is to use AWS Step Functions
Week 8	Create AWS Lambda function to dynamically execute MWAA workflow	Potential for complexity in learning AWS Lambda service, fallback plan is to forgo use of AWS Lambda and manually execute MWAA workflow

4 REFERENCES

1. Berg, S. (2017, July 19). In dermatology, health disparities can be skin deep. American Medical Association. Retrieved February 25, 2024, from https://www.ama-assn.org/delivering-care/patient-support-advocacy/dermatology-health-disparities-can-be-skin-deep

2. Buster, K. J., Stevens, E. I., & Elmets, C. A. (2013). Dermatologic Health Disparities. Dermatologic Clinics, 30(1), 53–59. https://doi.org/10.1016/j.det.2011.08.002

3. Stanford University School of Medicine (2022). Diverse Dermatology Images: A biopsy-proven skin disease dataset with diverse skin tone representation. https://ddi-dataset.github.io/