OncoDetectAI

Project Tracker

📊 Project Overview

Timeline: September 13 - December 12, 2024
Team: Aditi Deshmukh, Siddharth Pawar, Pranali Chipkar
Status: ✅ Completed
Total Sprints: 7
Total Tasks: 80

🎯 Sprint 1: Foundation & Environment Setup

Duration: Sept 13 - Sept 26 (2 weeks)
Status: ✅ Completed
Goal: Establish development environment and complete initial research

Sprint 1 Tasks

Task	Status	Priority	Assignee	Type
Research SCLC literature and identify relevant papers	Done	Medium	Aditi Deshmukh	Research
Review DBT documentation for data transformation	Done	Medium	Pranali Chipkar	Research
Review Snowflake Cortex AI documentation	Done	Medium	Siddharth Pawar	Research
Configure Airflow environment using Docker	Done	Medium	Siddharth Pawar	Setup

Task	Status	Priority	Assignee	Type
Create .env configuration files and Dockerfile for Airflow	Done	Low	Aditi Deshmukh	Setup

Challenges Faced:

- Docker container compatibility issues with different operating systems
- Initial Airflow configuration complexity

🚀 Sprint 2: Data Pipeline Development & Deployment

Duration: Sept 27 - Oct 10 (2 weeks)
Status: ✅ Completed
Goal: Build and deploy multimodal data ingestion pipelines

Sprint 2 Tasks

Task	Status	Priority	Assignee	Type
Set up S3 buckets for PDF storage and image storage	Done	Medium	Pranali Chipkar	Setup
Define DDL schemas for unstructured data tables in Snowflake	Done	Medium	Aditi Deshmukh	Development
Develop DAG to fetch SCLC papers from PubMed and upload to S3	Done	High	Siddharth Pawar	Development
Develop DAG for text embedding pipeline with Snowflake Arctic	Done	High	Pranali Chipkar	Development
Develop DAG for image embedding pipeline with Voyage Multimodal	Done	High	Aditi Deshmukh	Development
Deploy Airflow DAGs to GitHub repository via VS Code	Done	Medium	Siddharth Pawar	Deployment

Challenges Faced:

- Papers being fetched via Airflow took longer than expected due to large file sizes
- XCom errors in Airflow when passing data between tasks
- Batch processing in Airflow implementation to reduce load and processing time
- S3 bucket permission configurations

🗃️ Sprint 3: DBT Data Modeling & Transformation

Duration: Oct 11 - Oct 24 (2 weeks)
Status: ✅ Completed
Goal: Build structured data transformation pipeline using DBT

Sprint 3 Tasks

Task	Status	Priority	Assignee	Type
Create DBT Staging Models for Raw Clinical & Genomic Data	Done	Medium	Pranali Chipkar	Development
Implement Data Quality & Validation Tests in Staging Layer	Done	Medium	Siddharth Pawar	Development
Develop Intermediate Models for Biomarker & Risk Feature Engineering	Done	Medium	Pranali Chipkar	Development
Build Mart-Level ML Feature Tables	Done	Medium	Aditi Deshmukh	Development
Configure Snowflake Connections via profiles.yml	Done	High	Siddharth Pawar	Setup

Challenges Faced:

- Snowflake connection authentication errors via profiles.yml
- Data quality issues with raw clinical data containing null values
- Complex transformations for biomarker feature engineering

📊 Sprint 4: Midterm Deliverables & Subtype Classification

Duration: Oct 25 - Nov 7 (2 weeks)
Status: ✅ Completed
Goal: Complete midterm requirements and build SCLC subtype classification feature

Sprint 4 Tasks

Task	Status	Priority	Assignee	Type
Midterm ppt	Done	Medium	Siddharth Pawar	Documentation
Midterm paper research and code	Done	High	Aditi Deshmukh	Documentation
Midterm demo implementation logic	Done	Medium	Pranali Chipkar	Development
Quick classification logic for sclc subtype	Done	Medium	Aditi Deshmukh	Development
Full analysis logic for sclc subtype results	Done	High	Pranali Chipkar	Development
Setup UI page for Subtype classification	Done	Medium	Siddharth Pawar	Development

Challenges Faced:
- Subtype classification logic required multiple iterations
- Code refactoring needed for demo-ready prototype

💻 Sprint 5: Streamlit Application - User Management & Risk Prediction

Duration: Nov 8 - Nov 21 (2 weeks)
Status: ✅ Completed

Goal: Build user authentication, risk prediction features, and Maps integration

Sprint 5 Tasks

Task	Status	Priority	Assignee	Type
Create user sign up and login page	Done	Medium	Pranali Chipkar	Development
Store user authentication details in SF (User management schema)	Done	High	Pranali Chipkar	Development
Setup feedback page	Done	Medium	Siddharth Pawar	Development
Store feedback data in a separate table	Done	High	Siddharth Pawar	Development
Create profile page for user including main buttons	Done	Medium	Aditi Deshmukh	Development
Create profile page for user including the product analytics button	Done	Medium	Aditi Deshmukh	Development
Quick prediction logic for risk	Done	Medium	Pranali Chipkar	Development
Full AI analysis logic for risk prediction	Done	High	Pranali Chipkar	Development
Integrating in-house embeddings to enhance AI analysis for risk prediction	Done	High	Aditi Deshmukh	Development
Setup UI for risk prediction page	Done	Medium	Siddharth Pawar	Development
Setup logic for risk categories	Done	High	Aditi Deshmukh	Development
Map agent for high risk patients	Done	High	Siddharth Pawar	Development
Setup network integrations for google maps api	Done	High	Pranali Chipkar	Setup

Task	Status	Priority	Assignee	Type
Maps logic based on location	Done	High	Siddharth Pawar	Development

Challenges Faced:

- Network integrations for Google Maps API setup and configuration
- Integrating and fine-tuning embedding-based context for Full AI Analysis in risk prediction
- Streamlit session state management for user authentication
- UI responsiveness issues when loading large datasets

🤖 Sprint 6: Research Assistant - Multi-Agent RAG System

Duration: Nov 22 - Dec 5 (2 weeks)
Status: ✅ Completed
Goal: Build intelligent research assistant with multi-agent architecture

Sprint 6 Tasks

Task	Status	Priority	Assignee	Type
Work on RAG agent logic for research assistant	Done	High	Aditi Deshmukh	Development
Work on Arxiv agent logic for research assistant	Done	High	Siddharth Pawar	Development
Store serp api key into config table	Done	High	Pranali Chipkar	Setup
Work on websearch agent logic for research agent	Done	High	Aditi Deshmukh	Development
Work on logic for saving research notes	Done	High	Siddharth Pawar	Development

Task	Status	Priority	Assignee	Type
Add translation, summarization features to the chat	Done	Medium	Pranali Chipkar	Development
Decide on threshold to call the 3 agents as per response confidence	Done	High	Aditi Deshmukh	Development
Create networks for arxiv and websearch access integrations	Done	High	Siddharth Pawar	Setup
UI Setup for research assistant page	Done	High	Pranali Chipkar	Development
Making UI uniform and clean across all pages	Done	Medium	Aditi Deshmukh	Development

Challenges Faced:

- Network integrations for SERP API access and authentication
- Arxiv library access setup and integration
- Agent orchestration and determining optimal confidence thresholds
- Balancing response time with agent accuracy

📈 Sprint 7: Product Analytics & Final Documentation

Duration: Dec 6 - Dec 12 (1 week)
Status: ✅ Completed
Goal: Implement Kafka-based analytics pipeline and complete all documentation

Sprint 7 Tasks

Task	Status	Priority	Assignee	Type
Setup Confluent Cloud	Done	Medium	Aditi Deshmukh	Setup
Setup topic and event structure	Done	High	Aditi Deshmukh	Development

Task	Status	Priority	Assignee	Type
Snowflake sink connector to Confluent Cloud	Done	High	Siddharth Pawar	Development
Store api keys, bootstrap url and rest api url into config table	Done	High	Pranali Chipkar	Setup
Trigger Kafka events on user_authentication success and failures	Done	High	Aditi Deshmukh	Development
Trigger Kafka events on main buttons clicks	Done	High	Siddharth Pawar	Development
Create flattened view of the event data received from Kafka	Done	High	Pranali Chipkar	Development
Setup network access to Confluent Cloud	Done	High	Aditi Deshmukh	Setup
Create UI for analytics query chatbot	Done	Medium	Siddharth Pawar	Development
Work on logic for fetching text to sql function	Done	High	Pranali Chipkar	Development
Integrate LLM to choose relevant tables for querying data	Done	High	Aditi Deshmukh	Development
Including AI insights and choice of visualization on analytics page	Done	High	Siddharth Pawar	Development
Deploy streamlit code to github	Done	High	Pranali Chipkar	Deployment
Upload snowflake ddls to github	Done	High	Aditi Deshmukh	Deployment
Documentation: poster	Done	High	Siddharth Pawar	Documentation
Documentation: github readme	Done	High	Pranali Chipkar	Documentation
Documentation: ppt file	Done	Low	Aditi Deshmukh	Documentation

Task	Status	Priority	Assignee	Type
Documentation: architecture diagram	Done	High	Siddharth Pawar	Documentation
Updating Notion for the work logs	Done	Medium	Aditi Deshmukh	Documentation
Documentation: YouTube video	Done	Medium	Aditi Deshmukh	Documentation

Challenges Faced:

- Connection of Confluent Cloud to Snowflake sink connector
- Decisiveness in event structure for product analytics
- Network access configuration between services
- Text-to-SQL function accuracy improvements
- Time management for comprehensive documentation

📈 Overall Project Summary

Sprint	Duration	# Tasks	Status
Sprint 1: Foundation & Environment Setup	Sept 13 - Sept 26	5	Done
Sprint 2: Data Pipeline Development & Deployment	Sept 27 - Oct 10	6	Done
Sprint 3: DBT Data Modeling & Transformation	Oct 11 - Oct 24	5	Done
Sprint 4: Midterm & Subtype Classification	Oct 25 - Nov 7	6	Done
Sprint 5: User Management & Risk Prediction	Nov 8 - Nov 21	14	Done
Sprint 6: Research Assistant Multi-Agent System	Nov 22 - Dec 5	10	Done
Sprint 7: Product Analytics & Documentation	Dec 6 - Dec 12	20	Done

Sprint Overview

Task Distribution by Member

- **Aditi Deshmukh:** 26 tasks
- **Siddharth Pawar:** 28 tasks
- **Pranali Chipkar:** 26 tasks

Task Distribution by Type

- **Development:** 54 tasks
- **Setup:** 12 tasks
- **Documentation:** 9 tasks
- **Research:** 4 tasks
- **Deployment:** 1 task

Completion Metrics

- **Total Tasks:** 80
- **Completed:** 80 ✅
- **Completion Rate:** 100%
- **Project Duration:** 13 weeks (Sept 13 - Dec 12)

🎯 Key Achievements

✅ Built multimodal RAG system with text and image embeddings
✅ Implemented 7 specialized AI agents
✅ Developed comprehensive risk prediction and subtype classification
✅ Created modern BI analytics pipeline with Kafka
✅ Delivered production-ready Streamlit application
✅ Completed all documentation and demo materials

🛠️ Technical Stack

Data Engineering: Airflow, DBT, Snowflake, AWS S3
ML/AI: Snowflake Cortex, Claude Code
Application: Streamlit, Python
Analytics: Kafka, Confluent Cloud
DevOps: Docker, GitHub, VS Code
APIs: PubMed, ArXiv, SerpAPI, Google Maps API