SCIENCE FAIR LOGBOOK

(2024-2025)

ERIN D’SOUZA

Grade: 9

Innovation Project: AI-Driven Early Detection of Breast Cancer

I’m estimating a Science Fair project would take a minimum of 100 hours. Working backwards, I will have to commit at least 2 hours every weekend, and about 4 hours per week on PD & other school holidays.

Date	Activity
January 6, 2024	Talked to parents about checking with current school about participating in Science Fair. Found out its too late for this year, but also school hasn’t participated before.
April 13, 2024	Visited CYSF with parents, and very excited to participate next year.
May 17, 2024	Spoke to Science teacher, but he is transferring to elementary next year and not teaching Science in Junior High.
June 5, 2024	Parents talked to Vice Principal and next year’s Science teacher about school registration with CYSF. They will let us know in September.
June 26, 2024	School’s out! Started exploring topics in Bioinformatics and …… Explored biotechnology areas to research, Python - (and learning about miRNA biology, their role in gene regulation, and their application as biomarkers in breast cancer detection. ecancer. (2023, August 16). *MicroRNA biomarker can help predict which breast cancer patients are more likely to see their cancer come back*. ecancer.org. https://ecancer.org/en/news/22364-microrna-biomarker-can-help-predict-which-breast-cancer-patients-are-more-likely-to-see-their-cancer-come-back
July 25, 2024	Started a course this summer, in Python to add to my existing knowledge of Python programming with aim to build ML (machine learning) models that I could possibly implement in DNA or RNA sequencing .
March 2024	Explored biotechnology areas to research, Python - (mention someething specific)

	Explored AI and its fundamental concepts and how it can be used for biotech research
April	Read and research genome sequencing, personlaized medicine and possible areas of research
June	Did a course in Python to add to my existing knowledge of Python programming with aim to build ML (machine learning) models that I could possibly implement in DNA or RNA sequencing
	Researched mRNA (not miRNA) and its practical applications
July	Studied mathematics/ statistics related to ML
August	understood ML in depth - studied various models - SVM, Logistic Regression, Lasso, Brushed up on Python Libraries - SkiLearn
August	Researched on miRNA and read research papers on ML and miRNA expression data research - Noticed most of the research done is recent and lot of early experimental citation (research papers)
October	Seems there isn't enough interest from other students and lack of teacher resources. Still no response from school yet about participating in CYSF. Talked to parents about alternative option.
November	Researched AI agents if I can build one based on current LLM models like Open Ai. Found it be to fairly complex, requires deep domain knowledge. Read research papers on ML modelling. l that was developed by…… Typically requires a team and intensive work. Beyond the scope of my project
December	Data selection. GDI data API. Tried downloading manually, failed multiple times. Finally realised that the data is meant to work with ML programs like python. Studied python coding required to build API. Finally after multiple tries succeed in accessing and downloading the dataset that I could use. This process was much harded than anticipated but once the data was downloaded it was very high quality data with almost no errors or cleaning required. It was a dream to work with
Nov	Researched various feature selection methods. Read on white papers on what other similar study have done . Studied miRNA pathways, miRNA expression data - Regulated, disregulated. Studies breast cancer - subtypes and molecular science behind it
Dec	Researched on various ML models - Focussed on SVM method
Jan	Researched on possible implementation of Deep Learning with ML methods - Did a deep dive in to CNN, RNN and other methods. Found out there are several shortcomings with miRNA - not with the method itself but with the data available. Good and reliable datasets are very limited (alteast the ones publicly available). Tried to understand why miRNA data is so sparse. Realised that Deep Learning typically required enormous amounts of data like LLM models (openAI). If the dataset problem could be solved it will be nothing short of revolutionary in the field of new way of data diagnostic , early cancer detection being just one of them. Studied current diagnostic standards and shortcomings mostly on mammography. Came up with Problem.
Dec	Deep dive on miRNA and its potential for disease diagnostics. Found series of research papers from the same publisher
Dec	Decided on the project of - AI models for miRNA early cancer detection

Jand	Wrote, tested and debugged code
Jan 15, 2025	Parents have applied to register me as an independent entry, yay!
Feb 3, 2025	Received access to CYSF portal, and started to review the platform requirements, videos, tips, and the judging tally sheet. Mom has connected with ‘Science Is’ to purchase the tri-fold.
Feb 6, 2025	Finalised and completed the basic project info, ethics and due care on the platform and submitted to CYSF for review
Feb 7, 2025	Approval from CYSF received.
Feb 8, 2025	Wrote tested and debugged code, Tested and fine tuned ML model
Feb 22, 2025	Working on citations.
March 1, 2025	Began work on the Presentation slides, not a lot of images online… but found a couple I like for the project image & project banner.
March 8, 2025	Looking at tri-fold measurements and how I can create posters from the slides and other graphics. (31x60)mademyposters30x60.Made themhalf(15x30)andtwoforeach
	Stickingpostersontrifold,andpracticingpitchforschoolfair.Finished compiling citations
March 18	Completed citations
November	Researched ML model Training, Evaluation and Performance
December 22 – 24, 2024	Code base: A after loading all the modules for feature selectin at once, setting up, training and validation. Read up on best practice to load the feature selection one by one to better test for overfitting. Redid the whole process found better metrics and it made logical sense too
	Feature selection working great today. All the parameters too good to be true. Ok, now I’m suspicious…
	Code: After the ML Model was constantly breaking had to reconfigure each module multiple times for proper filtering logic. Found out the modules were reverting to wrong filter parameters. Centralized the data filtering system and added comprehensive tests for edge cases. Introducing a centralized data loader and stricter validation. Changed cross validation parameters for better metrics.
December 27 – 30, 2024	Improve model testing with centralized data loading and updated testing framework using pytest and joblib.
	Performance metrics too good to be true. Possible overfitting. Readjusting parameters
	Import error, almost every single day. somehow analysis outpout file got deleted or disappeared!
	Lots of failures in the Test suite possibly because of change in the dependencies.
	Continue fixing the remaining test failures (specifically in feature selection and model robustness modules. Roll back some features maybe try a different approach.
	Continue fixing remaining test failures also in feature selection and model robustness modules
January 2 – 5, 2025	Accidently deleted some core modules. Restored form backup but some parameter are off. The stability analysis test is failing because of a key mismatch mean stability vs stability score. Updated test functions to use assertions instead of

	returning values to fix the warnings seen in the test output. Read stage 0 and 1 breast cancer specific study papers and which feature selection methods were used to understand ML modelling better. Watched youtube video on Python code
	Added better logging and debugging information.
	1. Fixed all test failures by properly handling SVM model variations
February	Ran all the test and executed analysis all out parameters looks good some minor adjustments. Performance threshold needs to be adjusted.
	fixed data loader errors. addressed class imbalance. confirmed minimal overfitting for now Conflicts with couple of visual tools. checking for version compatibility. statistical_tests.py module not working properly- fixing it. Feature selection pipeline not calling load_filtered_expression_data() , Training Module giving errors. lots of other bugs to fix. okay - the model seems to be broken nothing is working properly. Fixed! - Implemented centralized data loader with strict early-stage filtering. everything works great. Needs parameters fine tuning and some non critical features needs debugging otherwise everything looking good. Trying to optimize the model - discovered Significant Redundancy, Scattered functionality etc. Refactoring
March 1 – 2, 2025	Started write up for “problem” and “method”
March 8 – 9, 2025	Working on the “analysis” and “conclusion” this weekend.
March 15 – 16, 2025	Finalizing project write up. Reviewing citations and acknowledgements.