Predicting Cancer Reocurrence With AI

Team sdmay24-10: Chris Tague, Eric Schmitt, Mark Hanson,

Thriambak Giriprakash, Bishal Ghataney, Norfinn Norius

Faculty Advisor / Client: Ashraf Gaffar

FINAL REPORT

Introduction / Background

Problem statement

Killing over half a million people every year, Cancer is one of the leading causes of death in the United States. Efforts to find cures for cancer have been going on for decades, and no cure has been found. Many medical professionals work tirelessly to analyze cancerous cells in the hopes to find some correlation between them and the chances of them coming back (reoccurrence). Naturally this generates a large amount of data which is impossible for a relatively small number of doctors to process and study. Thankfully, artificial intelligence provides the computing power to handle this type of data. Our task is to leverage AI in order to find a correlation between cutting edge cancer spectral data and cancer reocurrence.

Intended users and uses

This project is intended to be used by patients who have had cancer and their doctors. It is used by uploading CSV files containing data spectrums of cancer cells to a secure website where a trained AI model is hosted. The AI model will use the uploaded data to return the predicted amount of time until cancer recurrence. The website requires users to make an account in order to login and use the model.

Context

A number of large studies have been done recently to assess the usefulness of AI in the realm of oncology. The resulting conclusion is that AI has strong potential in predicting and diagnosing cancer using pathology profiles and images studies (Zhang et al., 2023). The University of Pittsburgh in 2020 created a very accurate machine learning technique that diagnoses prostate cancer with a specificity of 98% and sensitivity of 98% (Zhang et al., 2023). Another AI technique that has been used recently was based on a Google DeepMind algorithm and was used to predict breast cancer more accurately than human specialists, also in 2020 (McKinney et al., 2020)

Oncology imaging studies using AI have an advantage in that training AI on images is relatively straightforward with huge results, such as the above mentioned case where breast cancer prediction was more accurate than a human specialist in that area. There are several disadvantages of using imaging for cancer research. One is that in some cases it is heavily biased, such as in detecting skin cancer the accuracy varies depending on the color of skin (Wen et al., 2021). Another study done showed that the AI could tell which institution had supplied the images and ended up lumping patients together by institution when training itself on the data which could lead to results based off of the institution rather than individual biology (Wood, 2021).

For AI training based on pathology data there is an issue of procuring good data. Training a model requires massive datasets to create accurate profiles, and this is tricky in the healthcare industry due to issues such as patient privacy, lack of data shared between institutions, and availability of data in general (Khan et al., 2023).

We have trained our AI model on cancer spectrum data pulled from images of cancerous and non-cancerous cells in the form of csv files. Each file is one image and column A gives a position x coordinate and column B is the corresponding value which together can be read as a vector. The advantage to our approach is that we are specifically training just on the cells, so we can identify any kind of cancer since all cancer cells look the same. We also do not run into any bias such as encountered in skin cancer image studies.
The corresponding disadvantage is we cannot differentiate what kind of cancer it is.

Khan, B. *et al.* (2023) *Drawbacks of artificial intelligence and their potential solutions in the healthcare sector*, *Biomedical materials & devices (New York, N.Y.)*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9908503/ (Accessed: 22 October 2023).

McKinney, S.M. *et al.* (2020) *International Evaluation of an AI system for breast cancer screening*, *Nature News*. Available at: https://www.nature.com/articles/s41586-019-1799-6 (Accessed: 22 October 2023).

Wen, D. (2021) *Characteristics of publicly available skin cancer image datasets: A ...*, *The Lancet Digital Health*. Available at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext (Accessed: 22 October 2023).

Wood, M. (2021) *Artificial intelligence models to analyze cancer images can take shortcuts that introduce bias for minority patients*, *UChicago Medicine*. Available at: https://www.uchicagomedicine.org/forefront/research-and-discoveries-articles/artificial-intelligence-models-to-analyze-cancer-images-can-take-shortcuts-that-introduce-bias-for-minority-patients (Accessed: 22 October 2023).

Zhang, B., Shi, H. and Wang, H. (2023) *Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach*, *Journal of multidisciplinary healthcare*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10312208/#:~:text=Machine%20learning%20(ML)%2C%20a,in%20predicting%20cancer%20than%20clinicians (Accessed: 22 October 2023).

Revised Design

Requirements (functional and non-functional)

Learning AI principles and tools for cancer recurrence prediction

- Familiarize ourselves with AI models suitable for predicting cancer recurrence based on pathology data
- Understand the principles of machine learning algorithms applicable to this problem, such as linear regression, neural networks, and convolutional neural networks

Construct an AI model for cancer recurrence prediction

- Identify and select appropriate AI models capable of processing the provided pathology dataset
- Design a logical pipeline for data flow through the selected models, ensuring a clear path from input to output

Data Preprocessing

- Utilize data preprocessing techniques to clean, normalize, and transform the raw pathology data into a format suitable for model training
- Handle missing values, outliers, and inconsistencies in the dataset to ensure data quality and integrity

Model Training and Validation

- Employ techniques such as cross-validation, regularization, and hyperparameter tuning to train the AI model using the preprocessed pathology data
- Utilize learning rate scheduling, early stopping, and other optimization techniques to improve model performance and prevent overfitting
- Validate the trained model using appropriate evaluation metrics and techniques, such as mean absolute error (MAE) and t-statistic tests

Web Application Development

- Design and develop a user-friendly web application that allows doctors to upload pathology data and view cancer recurrence predictions
- Implement a secure and efficient backend system to process uploaded data and generate predictions using the trained AI model
- Ensure the web application is HIPAA-compliant and adheres to relevant data privacy and security regulations
- Optimize the web application for responsiveness and compatibility across various devices and browsers to enhance user experience

Model Deployment and Integration

- Deploy the trained AI model on a scalable and reliable cloud platform, such as AWS EC2 instances or serverless architectures
- Integrate the deployed model with the web application backend to enable seamless data processing and prediction generation
- Implement necessary API endpoints and data transfer protocols to facilitate communication between the web application and the deployed model

Engineering standards

General Software Standards

- Clear and concise comments on code
- Properly named variables
- Files named properly
- Files must be placed and go through our shared Git repository (Minimal local work)
- Code must be extensively reviewed

Security concerns and countermeasures

Our project requires use of patient data. All dealings with medical information must be dealt with in a manner that is HIPPA compliant. To protect against medical information leaking, our website has a secure login page where every user must make an account. The user account information is stored using Supabase, which takes special precautions to be HIPAA compliant.

Evolution since 491 Implementation details

Initial Planning and Requirements Gathering

- Identified the goal of predicting cancer recurrence using spectral data.
- Gathered requirements for data collection, preprocessing, model building, and evaluation.

Implementation Phase Challenges

- Faced challenges understanding the spectral data due to the nature of the project.

Initial Model Architecture

- Started with a simple neural network architecture consisting of one dense layer with 1,000 nodes that feeds into another dense layer with 1 node with Stochastic Gradient Descent(SGD) optimizer.
- Encountered high Mean Absolute Error (MAE), indicating poor model accuracy.

Exploration of Advanced Architectures

- Experimented with more complex architectures including Conv1D, MaxPooling1D, and Flatten layers.
- Observed worsening of MAE, leading to abandonment of advanced architectures.

Attempted Ensemble Bagging Approach

- Tried ensemble bagging and boosting approach with the original architecture to improve MAE.
- Minimal improvement observed in MAE, prompting abandonment of ensemble approach.

Adoption of Early Stopping and Optimizer Change

- Implemented early stopping mechanism to prevent overfitting during model training.
- Switched from Stochastic Gradient Descent (SGD) optimizer to Adam optimizer for better convergence.
- Noticed gradual decrease in MAE from 61 to 45.

Introduction of Dropout, ReLU Activation and Learning Rate Scheduler

- Added dropout layers after each dense layer for regularization to prevent overfitting.
- Employed Rectified Linear Unit (ReLU) activation function for better performance.

- Utilized learning rate scheduler based on validation loss to dynamically adjust the learning rate during training for better convergence.
- Achieved significant reduction in MAE to 37.46.

Exploration of Transfer Learning:

- Explored transfer learning by using pre-trained models to enhance model accuracy.
- Encountered system crashes and errors during implementation, leading to discontinuation of this approach.

Detailed design

Data Preparation

- Read data from an Excel file containing samples and their corresponding recurrence data
- Load spectral data for each sample from CSV files.
- Preprocess the data by filtering out missing samples and normalizing the input features using StandardScaler.

Model Architecture

- Utilize a neural network model consisting of:
    - StandardScaling applied to input features
    - Input layer with BatchNormalization
    - Dense layers with ReLU activation
    - Dropout layers following dense layers
- We tried many different approaches to improve the evaluation of our model but failed on most of them due to higher MAE
    - Advanced architecture: Conv1D, MaxPooling1D, and Flatten
    - LSTM was used with no meaningful improvement to results

Training Process:

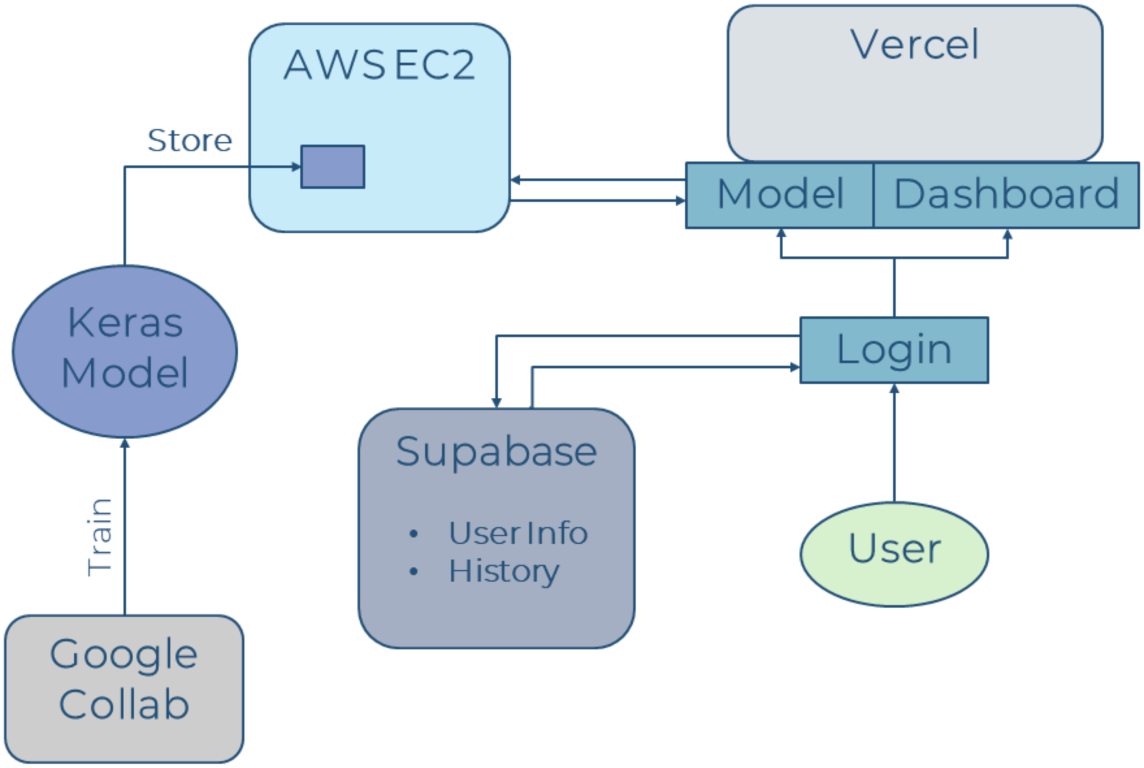
- Split the dataset into training and testing sets using train_test_split.
- Train the neural network model with Adam optimizer and Mean Absolute Error (MAE) loss function.
- Employ callbacks for early stopping and learning rate reduction during training.

Evaluation:

- Evaluate the trained model on the test set to assess its performance.
- Calculate the Mean Absolute Error (MAE) metric to quantify the model's accuracy.

Website

- Keras Model hosted on AWS EC2
- Supabase to store and verify login information
- Vercel frontend that contains the webpages



Description of functionality

- The code aims to predict the time until the recurrence happens based on spectral data.
- It pre-processes the input data, normalizing it to ensure consistent scaling.
- A neural network model is constructed and trained on the preprocessed data.

- The model is trained to minimize the Mean Absolute Error (MAE) loss function, optimizing its ability to predict time until recurrence.
- During training, callbacks are used to monitor and control the training process, including early stopping to prevent overfitting and learning rate reduction for better convergence.
- Finally, the trained model is evaluated on a separate test set, and the Mean Absolute Error (MAE) is calculated to assess its performance.
- The web page is hosted on Vercel
- The webpage hands the CSV file the user uploads to the model that is hosted on EC2 AWS, after which the model returns a reoccurrance prediction.
- The web application uses Supabase to store necessary login information

Notes on implementation

- Ensure that the input data is preprocessed and normalized before feeding it into the neural network model.
- Experiment with different architectures, hyperparameters, and optimization techniques to improve model performance.
- Monitor the training process using callbacks to prevent overfitting and achieve better convergence.
- Evaluate the model's performance using appropriate metrics, such as Mean Absolute Error (MAE), to gauge its accuracy on unseen data.
- Save important artifacts such as the scaler object for preprocessing and the trained model for future use or deployment.

Testing & Implementation

Process

Data Preparation

- Read and preprocess the data.
- Normalize the input data using StandardScaler.

Model Evaluation

- Split the dataset into training and testing sets.
- Train the neural network model on the training data.
- Use callbacks like EarlyStopping and ReduceLROnPlateau for better training control.
- Evaluate the trained model on the test set to assess its performance.

Save Scaler

- This scaler will be used to normalize new input data during deployment.

Mean Absolute Value (MAE) Model Testing:

- Evaluate the model’s performance using MAE metric.
- MAE measures the average absolute difference between the predicted and true values.
- Lower MAEs indicates better model accuracy.
- The best MAE we were able to get for the model was 37.4.
- The model with the best MAE that we got is based on testing and retraining the model with different other optimizers and by constantly adjusting the learning rate.

Front end/middle end Testing:

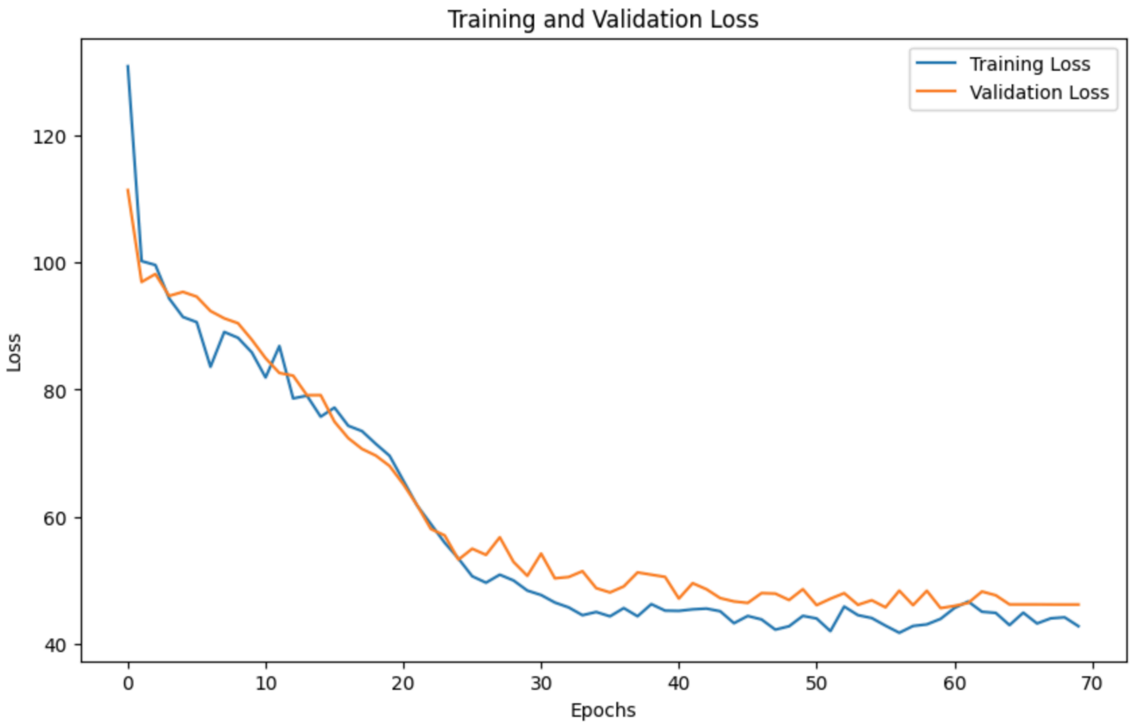
- Tested for proper user creation and login.
- Tested for proper page navigation upon user input.
- Tested for improper user creation and login.
- Tested for direct URL access prevention.
- Tested flask post requests to send csv to model hosted on EC2.
- Tested proper storage of previous runs to a specific user.

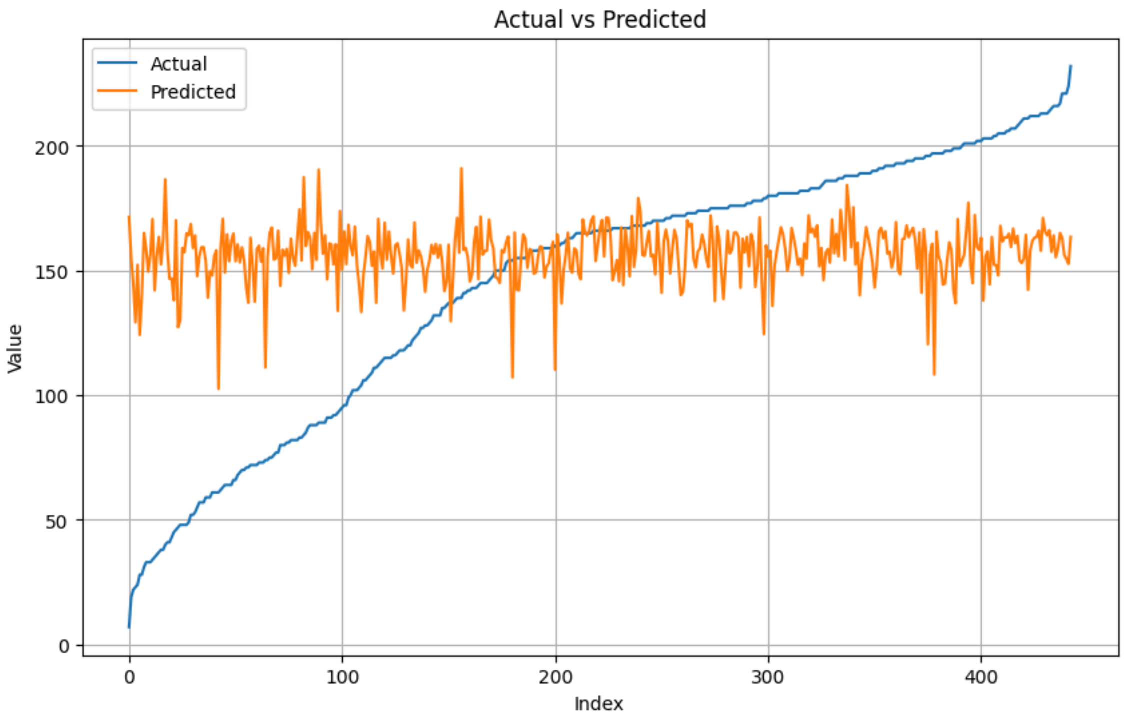
Final Evaluation

- Assess the model’s MAE on the test set.
- Compare the test MAE with the training MAE to gauge generalization performance.
- Ensure the model meets performance requirements for deployment.
- Ensured the model is correctly implemented through the webpage

Results

- Achieved MAE of ~37%.

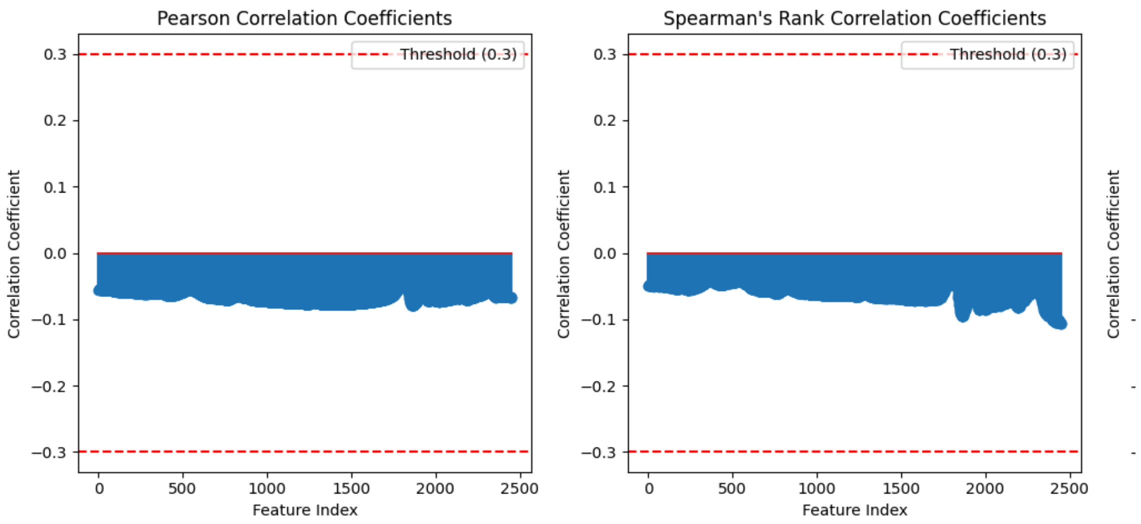




Note: The above chart was created by sorting our data. A model with less error would roughly follow the blue line.

- Found a lack of correlation between the spectra and the recurrence time.

- Based on the above chart, we can see the model appears to be guessing between about 125 to 175 months for all data entries, suggesting the model only learned to guess the average of its training data.



The two statistical analyses above all show a very slight negative correlation between the input fields (spectrum and reoccurrence). Pearson correlation coefficients measure linear correlations, whereas Spearman's Rank correlation measures can be used to measure any sort of correlation using monotonic measurements.

Broader Context

Societal Needs Addressed

- While the project aimed to address the need for predicting cancer recurrence, the findings suggest that the spectral data used in this study is not a reliable predictor. This highlights the importance of continued research into alternative methods and biomarkers for cancer recurrence prediction to better support patients, their families, and medical professionals in making informed decisions about ongoing care and monitoring.

Public Health, Safety, and Welfare

- The project's findings emphasize the need for caution when developing and deploying AI-based tools in healthcare. Reliance on ineffective predictors could lead to false reassurance or unnecessary stress for patients and misdirect healthcare resources.

- The medical field should prioritize the development of evidence-based, clinically validated tools for cancer recurrence prediction. Clear communication about the limitations and uncertainties of any such tools is essential to ensure they are used appropriately in clinical decision-making.

Global, Cultural, and Social

- This project shows that there needs to be more support in accurate cancer prediction. There is still much to learn in this space and continued support of cancer research will drive more projects like this which one day, will find a correlation between a set of data and cancer reoccurrence, which will save lives.

Environmental

- The project's findings suggest that the development and deployment of AI-based tools for cancer recurrence prediction should be approached with caution. The environmental costs associated with large-scale data processing and cloud computing for ineffective tools, is unnecessary.

Economic

- The project's outcomes highlight the need for careful consideration of the long-term cost-effectiveness of AI-based tools for cancer recurrence prediction. Investing in the development and deployment of tools based on unreliable predictors could lead to significant waste of healthcare resources.
- The medical field should prioritize the development of cancer recurrence prediction methods that are both clinically effective and economically sustainable. This will require robust cost-effectiveness analyses and the development of strategies to ensure equitable access to such tools.

Conclusions

Review progress

Many different methods of inference were developed and tested against the given dataset. Based on their results, based on the models attempted and the analysis run, we were unable to find a statistically significant correlation between the spectral data and months until cancer reoccurrence. In the process of developing our model we gained experience working with tensorflow, hosting models in the cloud, and creating user facing applications.

Value of current design

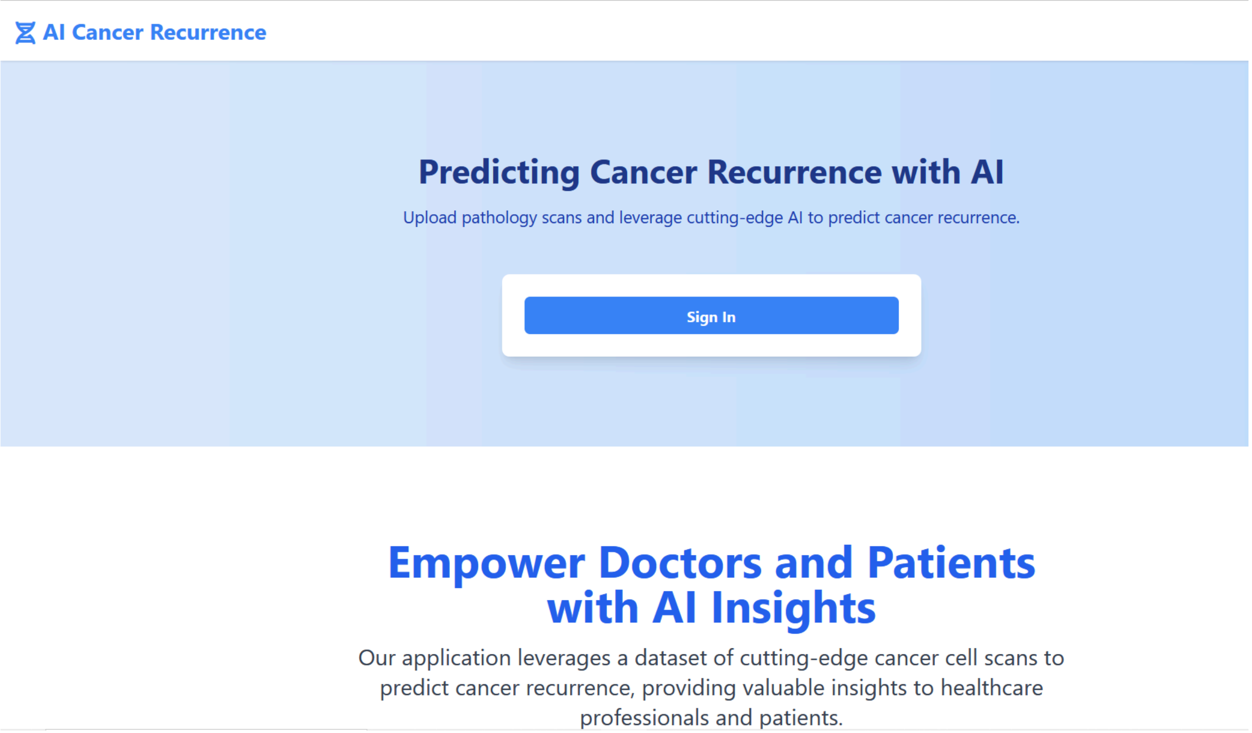
Right now, our project is not very valuable to the healthcare industry. Our best AI model does not have a statistically significant improvement over guessing the average at each instance of data upload. However, should a better model be created, it could be uploaded to EC2, and the existing framework that was created would already be in place to then use the better model. The current design has security features, history, and sharing results already implemented.

Potential future steps

As time goes on, the hope is that more data will be collected from cancer patients and that there are new developments in AI technology that can produce a statistically significant model for cancer predictions. This model would then be uploaded to EC2 and work with the existing website to provide healthcare professionals and patients with an online cancer prediction tool.

Appendix 1 – Operation Manual

1. Navigate to the website through the link https://sdmay24-10.vercel.app/.
2. Click the blue Sign In button.



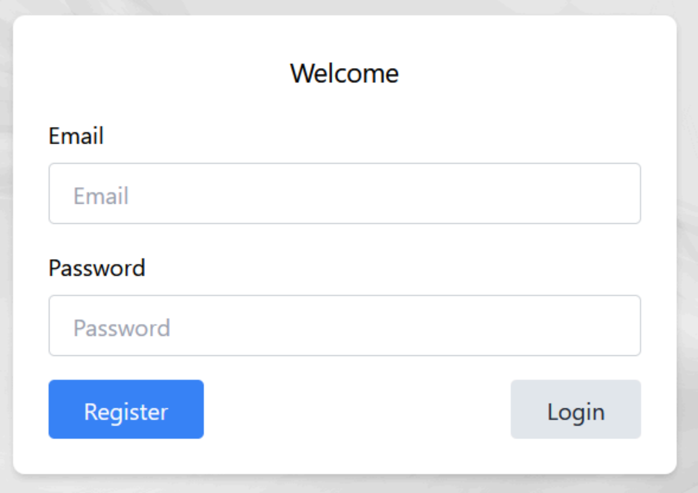
Predicting Cancer Recurrence with AI

Upload pathology scans and leverage cutting-edge AI to predict cancer recurrence.

Empower Doctors and Patients with AI Insights

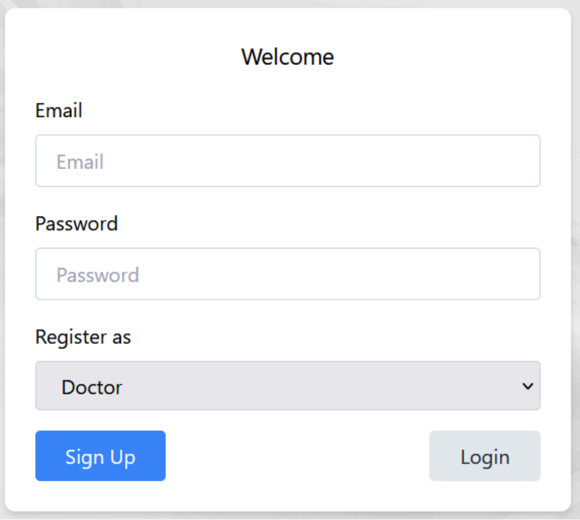
Our application leverages a dataset of cutting-edge cancer cell scans to predict cancer recurrence, providing valuable insights to healthcare professionals and patients.

3. Enter email and password into the designated email and password fields and click the login button.



4. If the user does not have an account, they can create an account by clicking on the blue text “Register”.

a. Enter email address and a password
b. Select doctor or patient
c. Hit the blue Sign Up button.

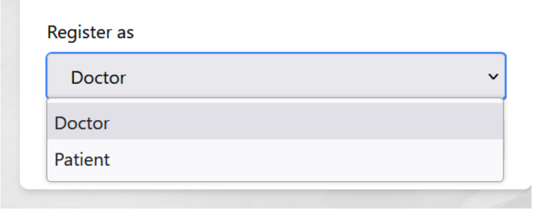


Welcome

Email: Email

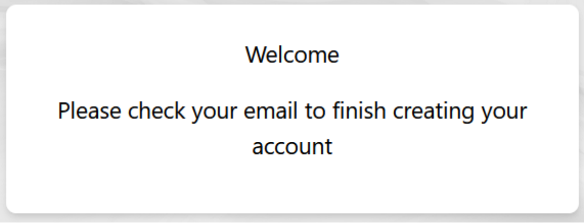
Password: Password

Register as: Doctor



Register as: Doctor

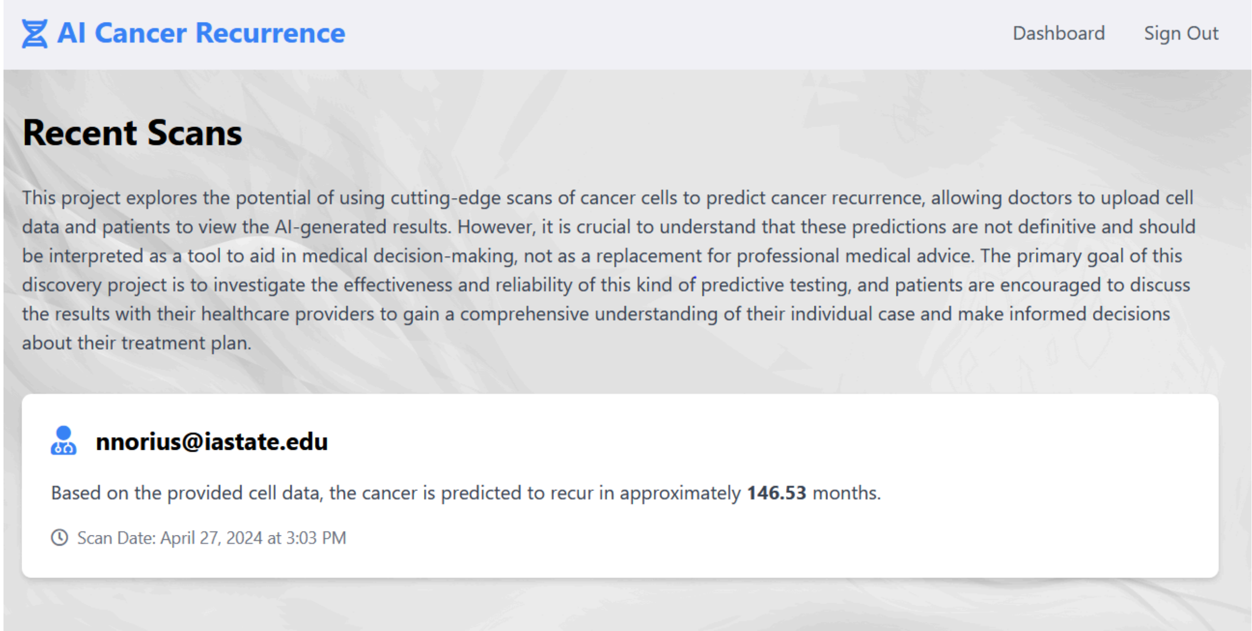
5. They will then be told to verify their account before logging in. Once verified, the user can login as above.



Welcome

Please check your email to finish creating your account

6. The user will then be taken to their dashboard. If the user is a patient, they can see their past results from the scans, what doctor performed the scan for them, and the time of the scan.



AI Cancer Recurrence

Recent Scans

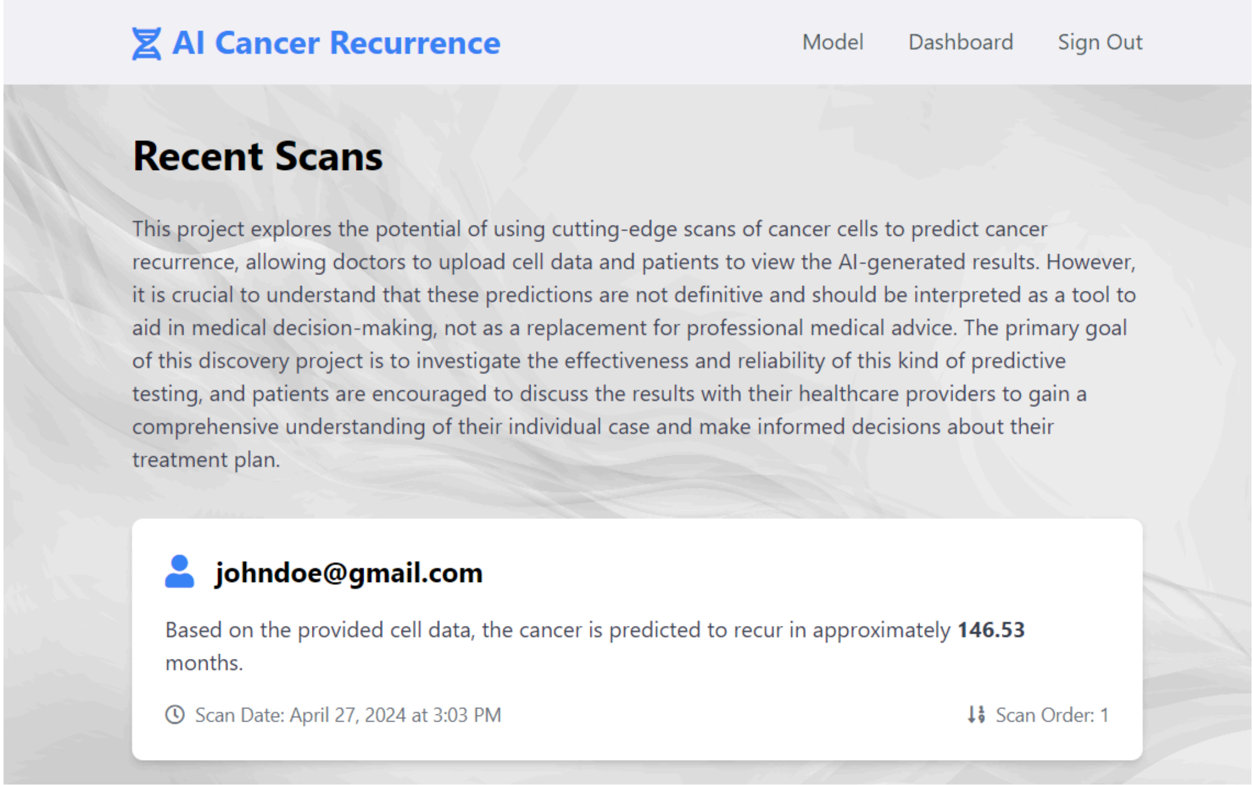
This project explores the potential of using cutting-edge scans of cancer cells to predict cancer recurrence, allowing doctors to upload cell data and patients to view the AI-generated results. However, it is crucial to understand that these predictions are not definitive and should be interpreted as a tool to aid in medical decision-making, not as a replacement for professional medical advice. The primary goal of this discovery project is to investigate the effectiveness and reliability of this kind of predictive testing, and patients are encouraged to discuss the results with their healthcare providers to gain a comprehensive understanding of their individual case and make informed decisions about their treatment plan.

nnorius@iastate.edu

Based on the provided cell data, the cancer is predicted to recur in approximately **146.53** months.

Scan Date: April 27, 2024 at 3:03 PM

7. If the user is a doctor, the doctor will see all their patient results and time of each scan on their dashboard.



AI Cancer Recurrence

Recent Scans

This project explores the potential of using cutting-edge scans of cancer cells to predict cancer recurrence, allowing doctors to upload cell data and patients to view the AI-generated results. However, it is crucial to understand that these predictions are not definitive and should be interpreted as a tool to aid in medical decision-making, not as a replacement for professional medical advice. The primary goal of this discovery project is to investigate the effectiveness and reliability of this kind of predictive testing, and patients are encouraged to discuss the results with their healthcare providers to gain a comprehensive understanding of their individual case and make informed decisions about their treatment plan.

johndoe@gmail.com

Based on the provided cell data, the cancer is predicted to recur in approximately **146.53** months.

Scan Date: April 27, 2024 at 3:03 PM

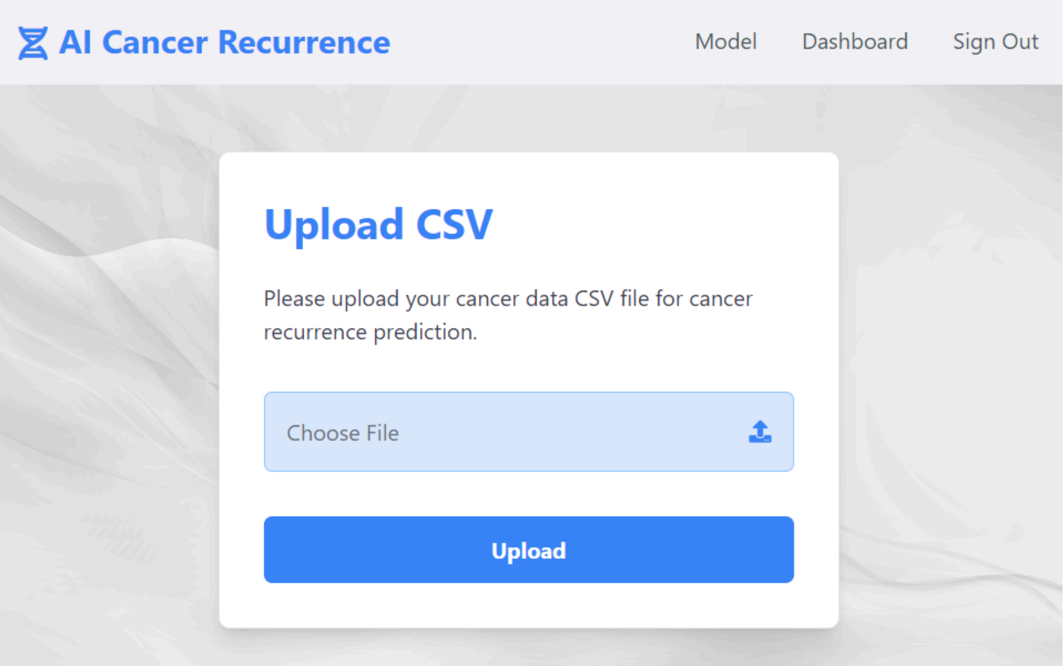
Scan Order: 1

8. If the user is a doctor, they also have access to the AI model, and can upload scans to the website.

a. In the upper right hand corner, the user can click on Model to be taken to the model page.

b. Click on the choose file and navigate in the file explorer to the data they wish to upload. Or, they can also drag and drop the file into the upload section to upload the data.

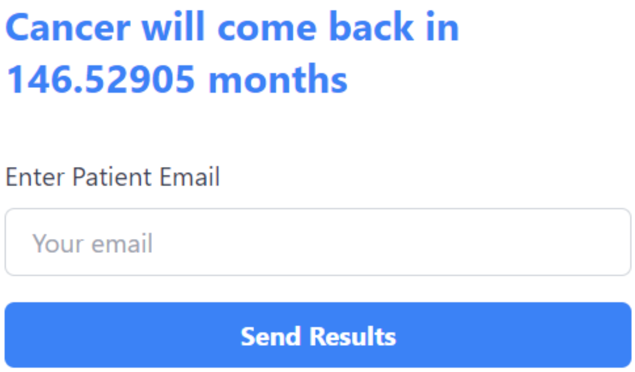
c. Once the data has been uploaded click the Upload button.



Upload CSV

Please upload your cancer data CSV file for cancer recurrence prediction.

9. The page will then return a prediction of cancer recurrence, and allow the user to send the information in an email to the patient. To do this, the user will type in the patient email and click Send Results.



Cancer will come back in 146.52905 months

Enter Patient Email

10. Click on Sign Out in the upper left hand corner to logout

Step-by-step instructions on how to setup/demo/test the system

1. Visit https://sdmay24-10.vercel.app/
2. Create account or sign in through the button on the top right
3. Select the ‘Model’ button on the top right of the screen
4. Find and upload a spectrum file to where it says upload
5. Enter the patient's email to store the data on the user’s dashboard.

Appendix 2 – Alternative/initial version of design

Versions considered before client’s specifications have changed

None

Versions considered before learning more about the project

None

Versions that resulted in failure to achieve specifications, etc.

The first idea to host the model on AWS was to use Lambda. This would allow effortless scalability, and provide all the benefits of serverless.

Along with lambda to host the model, a cloudfront hosting strategy was also attempted. Both failed due to particularities in our build file, and node_modules which could not be ported over to an AWS hosting strategy in a functioning manner. Since the issue was the build phase of website deployment, a new technology was leveraged which we had not considered before, that being AWS Amplify. Amplify offers an all in one solution to hosting a web page. By providing a git repository to read from, Amplify provisions a lambda, codebuild, and cloudfront resource in order to compile, build, and host a webpage. Unfortunately, the same issue with node_modules resurfaced so another hosting strategy had to be considered. This is when we landed on Vercel. Like Amplify, Vercel compiles, builds, and hosts a web page from a git repo. Fortunately, Vercel is well suited to host svelte libraries, so that was the hosting service moving forward.

Reason for revision

The initial problem with Lambda was the size restriction. Lambda has a limit of 250 Mb, but Tensorflow is over 500 Mb. So it was decided to use Tensorflow Lite. But a new problem was encountered: numpy was not importing. The reason for this was that numpy uses binary code to stay fast, and our architecture was different from that of the Lambda. The problem could have been fixed by using docker to build the zip file or by building the zip file on an EC2 instance running with Amazon Linux. However, it was easier and less complicated to just run the model on an EC2 server.

Appendix 3 – Other considerations

Any miscellany you deem important, what you learned, anything funny, anecdotes from your project experience

We learned a lot about Machine Learning. Most of us were not familiar with Machine Learning, and had to teach ourselves about AI through online resources. We familiarized ourselves with tensorflow and keras libraries. We took what we learned and built our first models on Google Colab. From there we made adjustments to the data and algorithms to try to find a more accurate model, our most accurate model had an accuracy of 37%. Upon statistical analysis, we found that the data seemed to show no correlation, we would suggest that the methods of gathering the spectrum dataset should be altered for the purpose of cancer reoccurrance prediction.