COMPUTATIONAL BIOLOGY

Final Project Presentation

Joshua Efraim
Registan

Goal

This project aims to train and create a machine learning model that can identify the patterns of each stage of cancer based on clinical, genomic, and demographic data, thus being able to accurately predict the cancer stage.

Using 5 different models: Logistic Regression, Random Forest, XGBoost, ANN, SVM

Independent variables: “Stage 1-3” and “Stage 4”

Tasks / Work Distribution

Efraim:

- Did background research
- Choosing the models based on the reviewed works
- Did the model training and setting the parameters for each model
- Explained the preprocessing and model training in the documentation
- Did the visualization for the code and help to write the discussion

Registan:

- Found and reviewed works related to the topic
- Created script to preprocess original dataset (Normalization, Replacing null values, etc)
- Wrote documentation for Abstract, Introduction, Discussion, and Conclusion

Task	Week	Member	Status
Topic Selection	2	Efraim and Registan	Done
Research	2-6	Efraim and Registan	Done
Dataset Finding and Preprocessing	4-7	Registan	Done
Model Selection and Training	7-11	Efraim	Done
Debugging	11-12	Efraim and Registan	Done
Final Report	10-12	Efraim and Registan	Done

Gantt Chart

Done | Partial | Not start

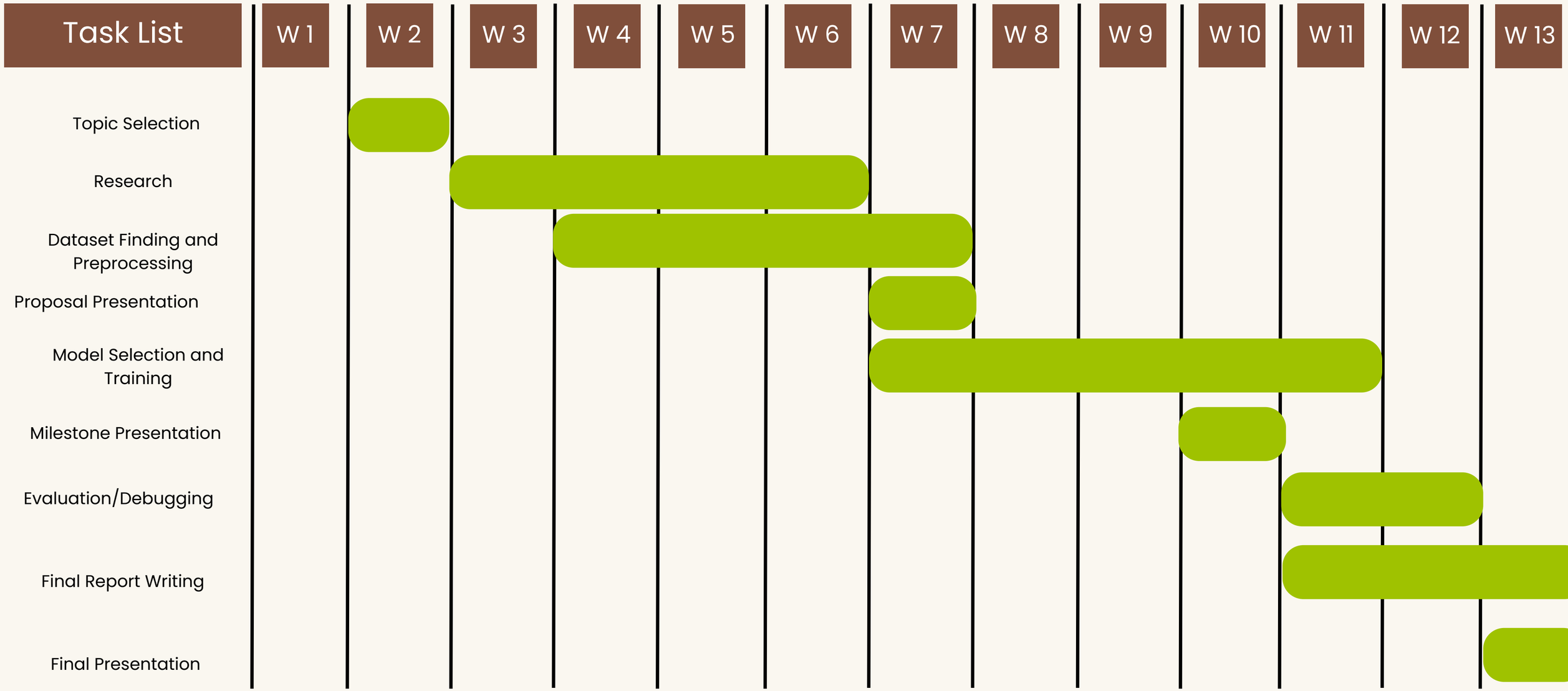
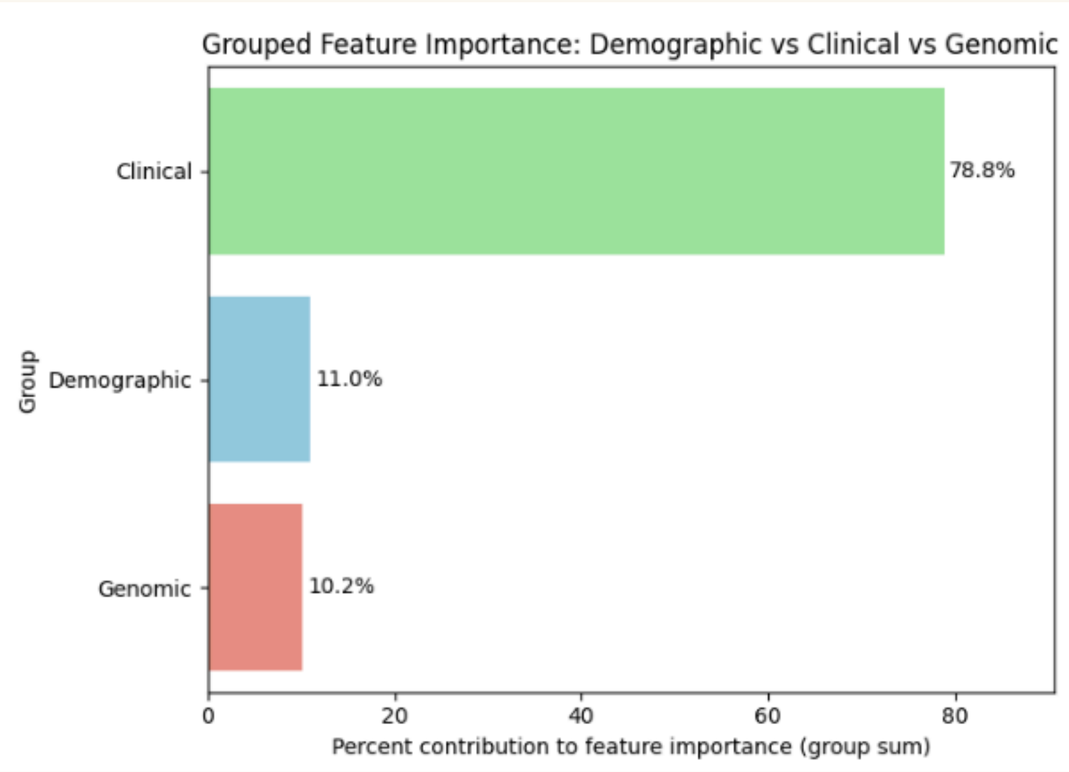


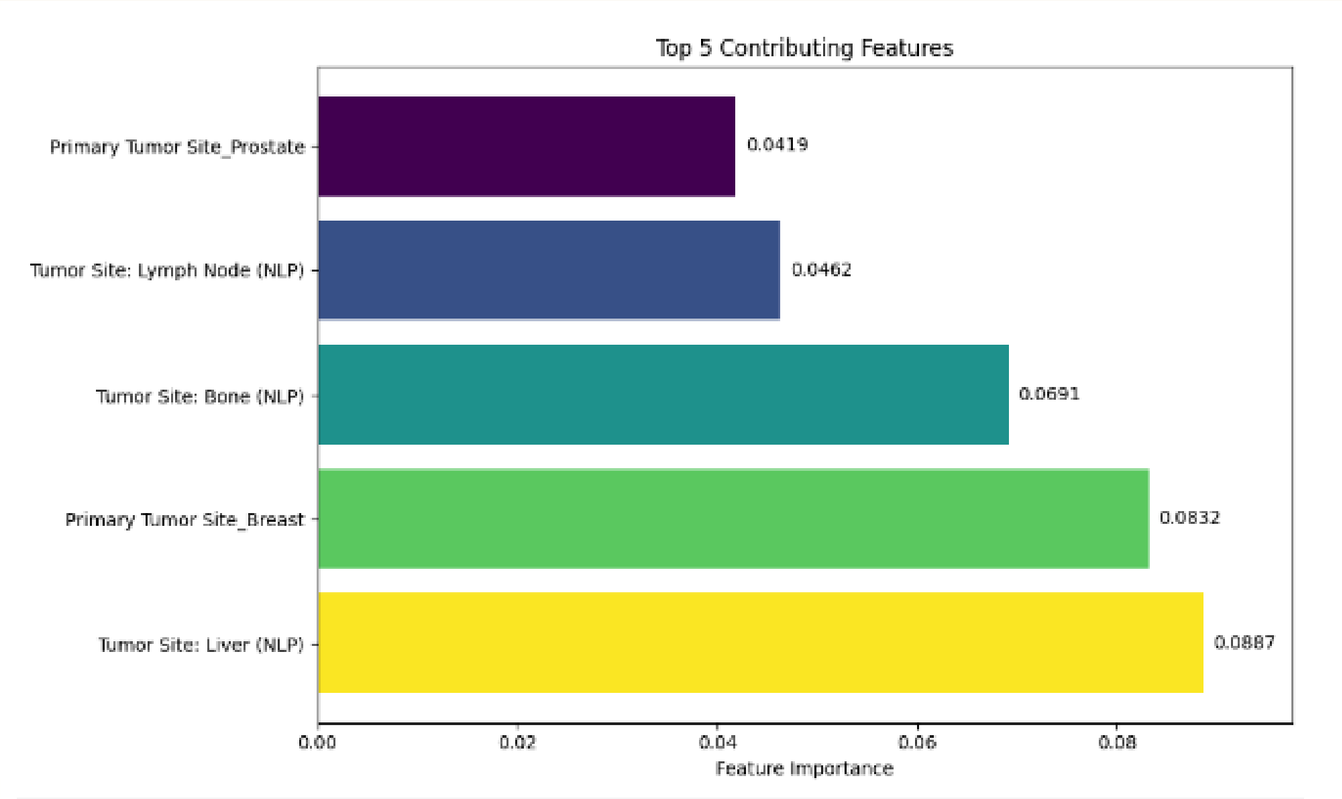
Table 1. Macro-Averaged Model Performance Comparison

Model	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.70	0.71	0.70	71.09%
Random Forest	0.73	0.73	0.73	73.46%
XGBoost	0.73	0.74	0.73	73.92%
ANN	0.72	0.72	0.72	72.54%
SVM	0.72	0.73	0.72	73.04%

By using F1-score, the model's performance is filtered through precision and recall. If either precision or recall is low, then the F1-score metric will be penalized, thus having a high F1-score means that the model is performing well and the metric is more reliable than using accuracy.

Model	Recall (Stage 4)
Random Forest	0.73
XGBoost	0.75


Grouped Feature Importance: Demographic vs Clinical vs Genomic
Clinical
Demographic
Genomic
78.8%
11.0%
10.2%
Group
0
20
40
60
80
Percent contribution to feature importance (group sum)


Top 5 Contributing Features
Primary Tumor Site_Prostate
Tumor Site: Lymph Node (NLP)
Tumor Site: Bone (NLP)
Primary Tumor Site_Breast
Tumor Site: Liver (NLP)
0.0419
0.0462
0.0691
0.0832
0.0887
0.00
0.02
0.04
0.06
0.08
Feature Importance

Limitations:

1. The model only predicts between stage 1-3 and 4, so it cannot be used to predict the specific stage of cancer. The model can only be used to help determine whether the cancer is metastatic or not.
2. The model f1-score is at 0.73 with accuracy of ~70%, this shows that the model still has almost a 30% chance of being incorrect. Thus it can only be used as a supporting decision rather than the main conviction when predicting the cancer stage.
3. There might be an imbalance in values of the features, some features may contain more of one value thus the model may be more biassed towards the certain cancer stage.

Conclusion:

The results indicate that tree-based models, specifically XGBoost and Random Forest, achieved superior macro averaged F1-scores compared to other evaluated models. While XGBoost and Random Forest had the same macro averaged F1-scores of 0.73, but XGBoost particularly had a superior sensitivity for the metastatic class with a score of 0.75. The analysis also revealed a slight variability in model performance across multiple runs highlighting the importance of model robustness when applied in medical datasets. In future work, the integration of a larger dataset could be explored to improve predictive performance and accuracy.

The End