2425MED1006 Developing a Machine Learning Approach to Identify "Digital Conterparts" Using Targeted Clinical Genomic Sequencing Data

Student Name:	Lee Cheuk Hang Jacob	Project Mentor:	Professor Wong, Jason Wing Hon
Major:	Bioinformatics	Department:	Medicine

INTRODUCTION

The aim of this project was to develop a deep learning-based system capable of identifying "digital counterparts" — historical cancer patients with similar molecular and clinical profiles — to support treatment decisions for new patients. This would involve matching targeted sequencing data to cases with known treatment outcomes using a deep learning framework inspired by a recent research paper.[1] However, due to time constraints, it was not feasible to perform full treatment matching or outcome prediction. Instead, the project focused on understanding and replicating the preprocessing pipeline described in the paper, engineering the necessary molecular and clinical features, and preparing the dataset to be compatible with the model's input requirements, and retrain the model based on the features we could extract from our data. This work serves as a foundational step towards creating the full system and highlights the practical challenges of applying machine learning in a clinical genomics setting.

METHODOLOGY

Data Collection

There was a total of 1604 patients' data being used in this project. These patients had either blood or tissue biopsies performed, and the blood was collected for ctDNA extraction; Meanwhile, formalin-fixed, paraffin-embedded tumor tissue was collected from primary or metastatic sites. Blood and tumor TP-NGS were performed using the FDA-approved FoundationOne Liquid CDx (F1LCDx) and FoundationOne CDx (F1CDx) assays.

Model Architecture

The model is called GDD-ENS, which is an ensemble model of 10 separate models. The training dataset was first split into 10 sets using the StratifiedShuffleSplit function in scikit-learn, then each is used to train a fully connected feed-forward neural network. The model was trained using Adam, with a batch size of 32 for 200 epochs. The hyperparameters were evaluated using a gaussian process, using gp_hedge as an acquisition function. The training process also has a measure against overfitting, early-stopping model developments after either 5 model updates or 500 calls to the gp_minimize function used in the training process. [1]
In the modified training process, the missing features in our dataset were replaced with 0.

Data Processing

Our data processing pipeline integrated multiple genomic and clinical data sources through a comprehensive series of computational steps. We began by combining variant summaries, SNV data, and mutation hotspot information into a unified dataset. The feature engineering phase involved creating binary indicators for gene mutations and truncations, while also generating mutation signatures. We extracted mutational signatures using COSMIC reference signatures and the MutationalPattern package, complemented by detailed trinucleotide context information processing.

To ensure data quality, we implemented thorough cleaning procedures including median imputation for continuous variables and conversion of categorical variables to binary indicators. The pipeline standardized column names and data formats while removing redundant features. Throughout the process, we maintained strict quality control measures by validating data completeness, ensuring consistent feature encoding, and verifying the absence of missing values.

However, due to limited time and the complexity of features related to CNV, there are still some features not replicated in our dataset.

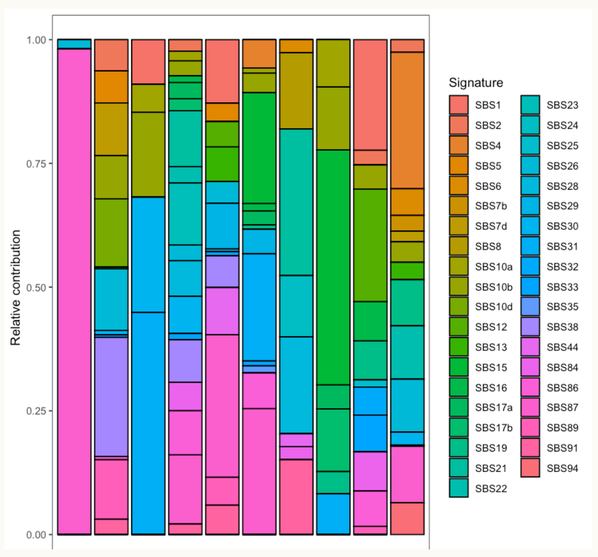

Figure 1. Plot of Contribution of Each Mutational Signatures for 10 Tumor Samples

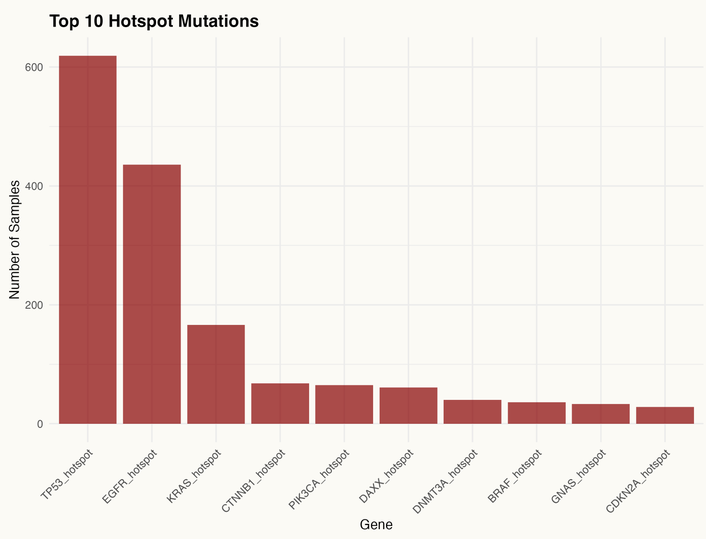

Figure 2. Plot of Top 10 Most Common Hotspot Mutations among the 1604 Samples

RESULTS

The first few layers of the model training returned sub-par performance with the current features, as some features could not be replicated either due to dataset limitations or time constraints.

```
training it: 0 0.6433685646096148
training it: 1 0.6464370951244459
training it: 2 0.647800886464371
```

CONCLUSION

This project lays the groundwork for leveraging deep learning in personalized cancer treatment by adapting key preprocessing steps to real-world data. With more time, the approach has a strong potential to guide clinical decisions through identifying patient similarity.

References:

1. Darmofal M, Suman S, Atwal G, Toomey M, Chen J-F, Chang JC, Efsevia Vakiani, Varghese AM, Anoop Balakrishnan Rema, Syed A, et al. 2024 Feb 27. Deep Learning Model for Tumor Type Prediction using Targeted Clinical Genomic Sequencing Data. Cancer discovery. doi:https://doi.org/10.1158/2159-8290.cd-23-0996.