A Web Portal for Genetic Mutation Analysis: Predicting AD and Breast Cancer Risk Using ROSMAP and BRCA Datasets

Joel Chong, Evan Qiu (Kaiyang), Yuxuan Wang, Xiaoyu Zhu
School of Biomedical Engineering, The University of New South Wales, Sydney, Australia

***Abstract*—A highly interactive and user-friendly web portal was developed for cancer prediction, where users can input genetic mutation data and assess the likelihood of various cancer types. Beyond prediction, the portal offers features for exploring historical model performance to visualise connections between different genetic mutations and cancer types. The project was successfully implemented using Multi-Omics Graph cOnvolutional NETworks (MOGONET), with all relevant functional modules thoroughly verified.**

***Index Terms*—cancer prediction, genetic mutation data, multi-omics, MOGONET, web portal, interactive, biomarkers, trend analysis, predictive modeling**

I. Introduction

Cancer is one of the leading causes of mortality worldwide, and early accurate diagnosis is critical for improving patient outcomes [1]. High-throughput biomedical technologies make the collection of multi-omics data possible. Multi-omic data can offer a more comprehensive understanding of biological processes and improve disease prediction accuracy. While many existing methods focus on unsupervised integration, there is growing interest in supervised approaches that consider interactions across omics types [2].

Previous research has highlighted the importance of integrating multi-omics. For instance, [3] demonstrated that combining transcriptomic and epigenomic data improved the classification accuracy of cancer subtypes. Similarly, the use of GCNs has shown high performance in capturing non-linear relationships within complex datasets [4].

In our study, the Multi-Omics Graph Convolutional NETworks (MOGONET) will be used on our web server for cancer prediction. MOGONET integrates omics-specific learning and cross-omics correlation learning for the accurate classification of multi-omics data. It has demonstrated superior performance over other leading methods in biomedical classification, utilizing mRNA expression, DNA methylation, and microRNA expression data. In addition, it can identify significant biomarkers from different omics data types relevant to the specific biomedical issue quantitatively [5]. Therefore, we will be using MOGONET as the backbone model. The datasets ROSMAP for AD (Alzheimer's disease) and BRCA for breast cancer were used to demonstrate the model performance.

II. Method

A. Machine Learning

Two multi-omics datasets were used in this study: ROSMAP (related to Alzheimer's disease) and BRCA (breast cancer data). These datasets contain multiple omics views, including gene expression, DNA methylation, and miRNA expression. Each view corresponds to a different omics layer, providing complementary insights into the biological mechanisms of the respective diseases.

Data for each view was loaded using Python's numpy and pandas libraries. Labels for training (labels_tr.csv) and testing (labels_te.csv) were preprocessed to ensure class balance. Omics features were split into training and testing datasets, ensuring the data formats were compatible with PyTorch tensors. The number of samples for training and testing was recorded to maintain reproducibility. Missing values, if any, were imputed, and all features were normalized to ensure consistency across omics layers.

For graph reconstruction, pairwise similarity matrices were constructed for samples within each omics view using cosine similarity. Adjacency matrices were generated using a predefined adjacency parameter (adj_parameter), which determines the sparsity of the graph. For ROSMAP, adj_parameter was set to 2, while for BRCA, it was set to 10, reflecting differences in dataset complexity. Model Architecture The core of the model architecture is based on MOGONET, which integrates multiple omics views through the following components:

1) Graph Convolutional Networks (GCN): Each omics view was processed independently by a GCN, which extracted high-level feature representations by aggregating information from neighboring nodes in the graph. The hidden dimensions of the GCN layers were set as follows:

For ROSMAP: [200, 200, 100]

For BRCA: [400, 400, 200]

2) Vector-based Convolutional Deep Network (VCDN): The outputs from the GCNs were concatenated and passed through a VCDN layer for multi-class classification. The VCDN layer modeled the interdependencies between omics layers, enabling the model to make accurate predictions.

3) Training Procedures: The training process was divided into three stages: Pretraining GCNs, MOGONET Training and testing and After Trained VCDN Classification

Each GCN was pretrained independently using only its respective omics view. The optimizer used was Adam, with a learning rate of lr_e_pretrain = 0.001. The pretraining process spanned 500 epochs, with loss values computed using the cross-entropy loss function weighted by sample importance. Joint Fine-Tuning:

After pretraining, the model was fine-tuned end-to-end, jointly training all GCNs and the VCDN layer. The learning rates for this stage were set as lr_e = 0.0005 (for GCNs) and lr_c = 0.001 (for the VCDN classifier). This phase spanned 2500 epochs to ensure convergence. Metrics such as accuracy, F1-score, precision, and AUC were logged every 50 epochs to monitor performance.

4) *Evaluation Metrics:* To evaluate model performance, the following metrics were used:

Accuracy: The proportion of correctly classified samples. F1-Score: Evaluated for both weighted and macro averages to account for imbalanced datasets.

Precision: Evaluated for each class and averaged (macro).

AUC (Area Under the Curve): Calculated using one-vs-rest (multi-class) or binary approaches to assess model discriminability.

Feature Importance Analysis: To enhance interpretability, feature importance scores were calculated for each omics view. Features were perturbed individually (set to zero), and the impact on model performance was measured. The change in F1-score after perturbation was used as a proxy for feature importance. Important features were ranked, and the top-ranked features were saved to a CSV file for downstream biological interpretation.

Output Formatting: The outputs were saved in standardized formats for ease of downstream analysis:

Performance Logs: Metrics (accuracy, F1-score, precision, AUC) were saved to text files (training_performance_50_epoch.txt, testing_performance_50_epoch.txt) for easy visualization.

Feature Importance: Ranked feature importance scores were stored in separate CSV files for ROSMAP (important_features_ROSMAP.csv) and BRCA (important_features_BRCA.csv).

5) *Outcome Visualization:* The chalk function of E-Charts was applied in a Vue 3-based implementation to create bar and line charts. The bar chart was utilized to display machine learning results, where the X-axis represented feature names, and the Y-axis corresponded to feature importance values. The line charts below depicted performance metrics during machine learning training and testing, such as accuracy (ACC), F1-weighted, F1-macro, AUC, precision, and loss.

The dynamic interaction between the chart and the user was realized by enabling responsive and interactive features within the chart. The integration between machine learning results and readable diagnostic conclusions was addressed by converting complex data into comprehensible content. A web interface was designed for user input, allowing users to assign level values to the top 30 feature names based on preliminary training results. These values were used to calculate the likelihood of a patient developing a specific cancer. The probability was computed as (1):

$$\text{Probability} = \frac{\sum(l \cdot i)}{\sum(k \cdot i)} \tag{1}$$

Where l corresponds to the expression level of an omic feature, i corresponds to the feature importance, k corresponds to the maximum value of feature importance. To predict this probability, 30 expression levels and their corresponding feature importances will be input into the formula for weighted calculation. In the web implementation, k was set to 100.

B. Front-End Integration

The website is named CancerInsight. Users are required to register an account, log into the website and linked as an oncologist or researcher that is registered in the back-end database (Figure 1).

Verified users can create patient profiles and attach the mutation data file to the patient. The file type accepted is the VCF (Variant Call Format) file, which is commonly used in bio-informatics to store gene mutation information. Figure 2 shows an example webpage for the patient profiles.

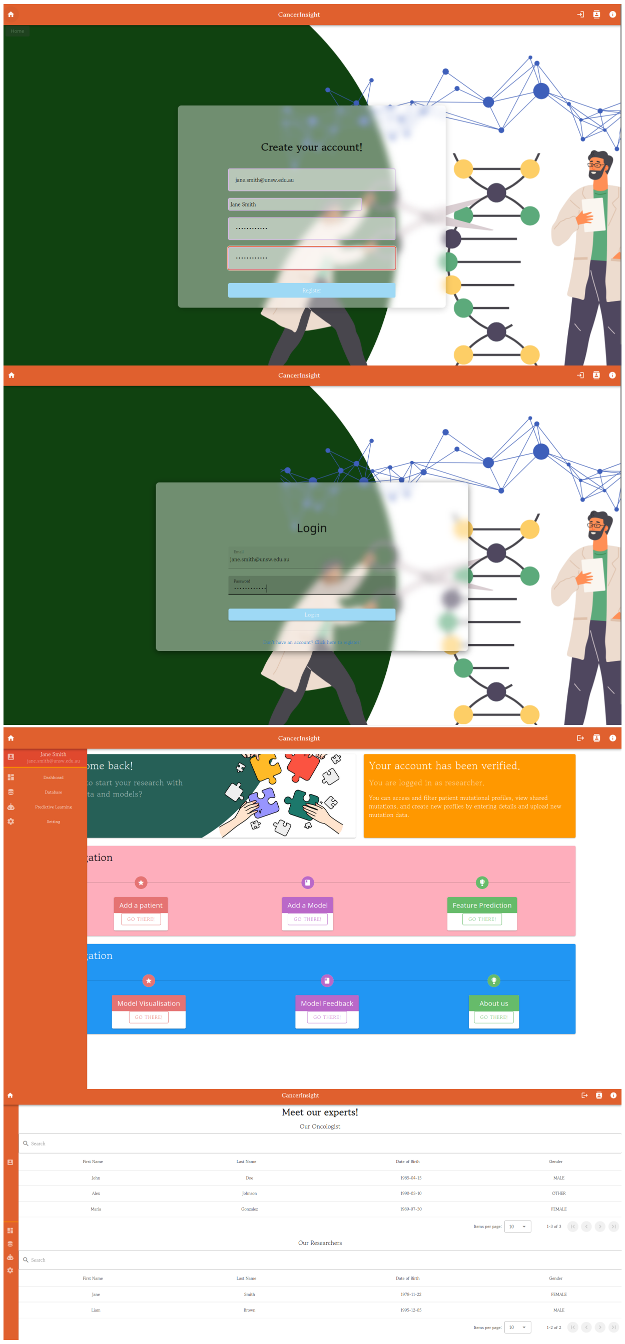
The front-end web application enables users to perform machine learning tasks and utilize trained models for various visualizations and predictive analyses (Figure 3). Users can configure parameters such as the number of training epochs and pre-training epochs, as shown in Figure 4. Once the model is trained, users can predict the probability of diseases, such as Alzheimer's disease (AD) or breast cancer, by inputting the feature names along with their corresponding weights on a scale from 0 to 100 (Figure 5). Additionally, the performance of different features is displayed on the front-end interface and will be further analyzed in the results section. Users can also provide feedback on the trained model, which will be documented in the back-end for future improvement in training performance (Figure 6).

III. Results and Analysis

The training results for the ROSMAP are presented in Figure 7. On the front end, this is displayed as a scrollable bar chart showing the top 30 features with the highest importance for the specific disease. The model's performance during training, including metrics such as accuracy, F1-weighted, F1-macro, precision, and loss, is shown in Figure 8 (with 2500 training epochs and 500 pre-training epochs). Similarly, the testing performance is illustrated in Figure 9.

IV. Conclusion

In this project a web portal for genetic mutation analysis was successfully developed. All components including the back-end database, machine learning models, front-end web interactions, and communication between various parts have been verified. The machine learning model, implemented in Python, was fully integrated into the back-end database using the Flask framework. It can make predictions based on user-submitted data, and the results are displayed on the front-end web page. Further improvements include fine-tuning the



Create your account!

jane.smith@unsw.edu.au

Jane Smith

············

············

Login

Email
jane.smith@unsw.edu.au

Password
············

Don't have an account? Click here to register!

Jane Smith
jane.smith@unsw.edu.au

ome back!

o start your research with
ta and models?

Your account has been verified.

You are logged in as researcher.

You can access and filter patient mutational profiles, view shared mutations, and create new profiles by entering details and upload new mutation data.

ation

ation

Meet our experts!

Our Oncologist

Search

First Name	Last Name	Date of Birth	Gender
John	Doe	1985-04-15	MALE
Alex	Johnson	1990-03-10	OTHER
Maria	Gonzalez	1989-07-30	FEMALE

Items per page: 10 1-3 of 3

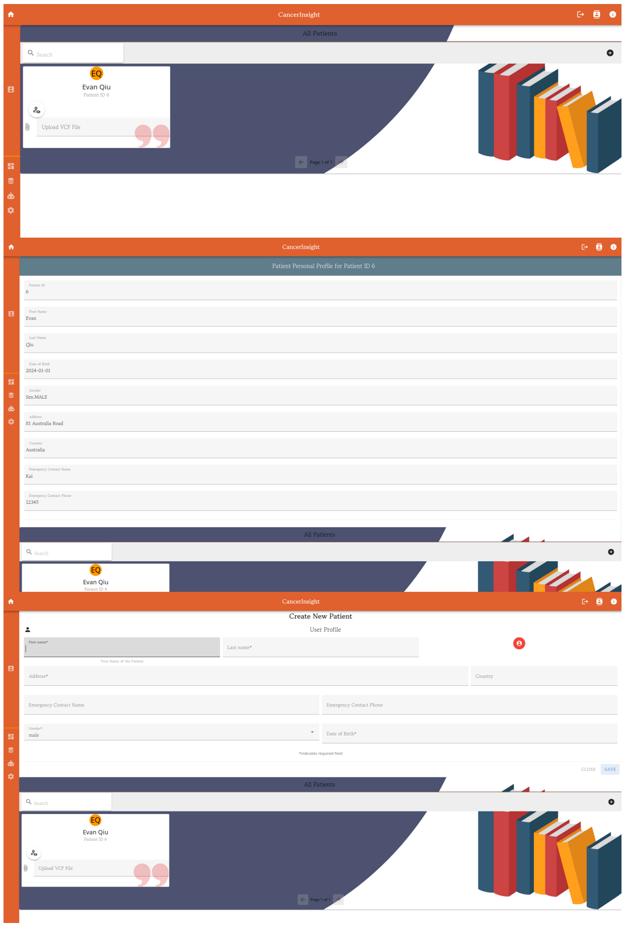
Our Researchers

Search

First Name	Last Name	Date of Birth	Gender
Jane	Smith	1978-11-22	FEMALE
Liam	Brown	1995-12-05	MALE

Items per page: 10 1-2 of 2

Fig. 1. Example Setup for the User Account



Patient Personal Profile for Patient ID 6

Patient ID
6

First Name
Evan

Last Name
Qiu

Date of Birth
2024-01-01

Gender
Sex.MALE

Address
81 Australia Road

Country
Australia

Emergency Contact Name
Kai

Emergency Contact Phone
12345

Create New Patient

User Profile

First name*

First Name of the Patient

Last name*

Address*

Country

Emergency Contact Name

Emergency Contact Phone

Gender*
male

Date of Birth*

*indicates required field

Fig. 2. Example Patient Profile Page

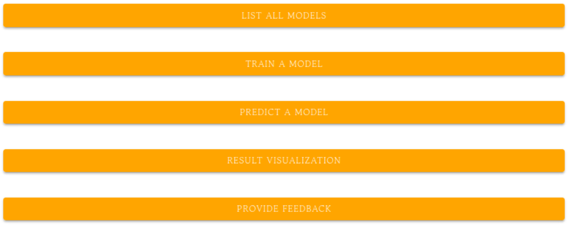
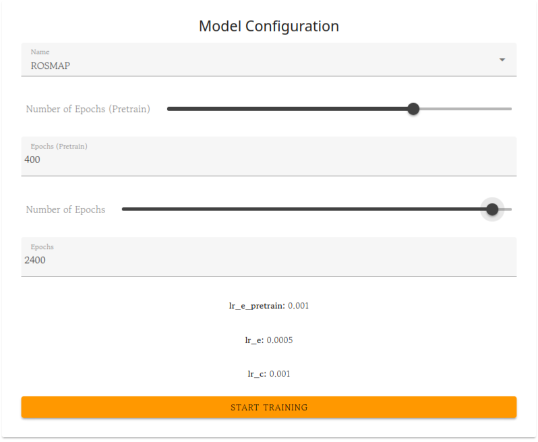


Fig. 3. Homepage for machine learning

model in real-time and trying datasets for other types of cancer prediction.



Model Configuration

Name
ROSMAP

Number of Epochs (Pretrain)

Epochs (Pretrain)
400

Number of Epochs

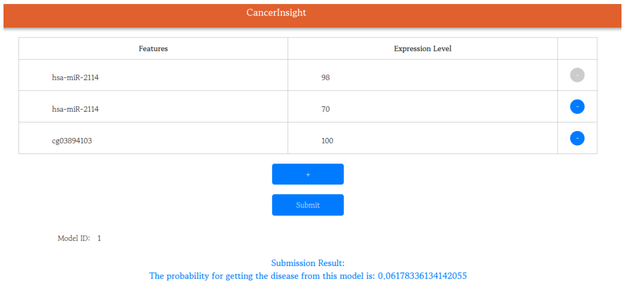
Epochs
2400

lr_e_pretrain: 0.001

lr_e: 0.0005

lr_c: 0.001

Fig. 4. Configurable training parameters



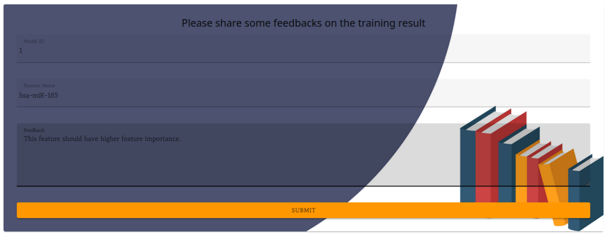
CancerInsight

Features	Expression Level	
hsa-miR-2114	98	-
hsa-miR-2114	70	-
cg03894103	100	-

Model ID: 1

Submission Result:
The probability for getting the disease from this model is: 0.06178336134142055

Fig. 5. Prediction of cancer probability based on weighting



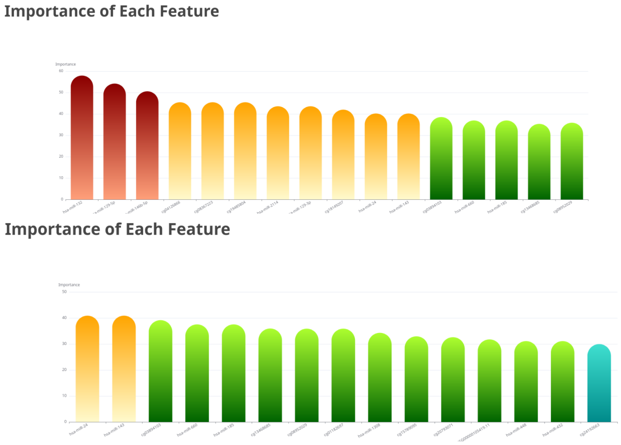
Please share some feedbacks on the training result

Model ID
1

Feature Name
hsa-miR-185

Feedback
This feature should have higher feature importance.

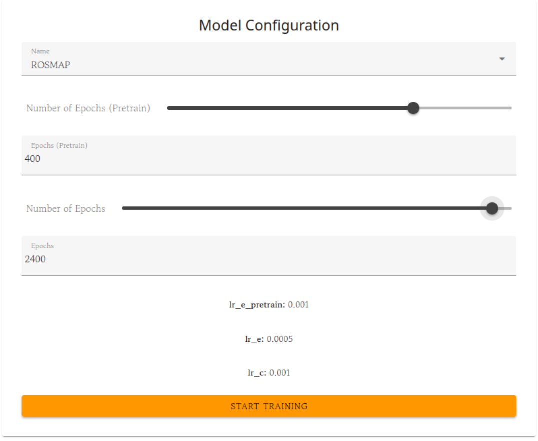
Fig. 6. Example user feedback for a trained model



Importance of Each Feature

Importance of Each Feature

Fig. 7. Top features that contribute to the disease (AD)



Model Configuration

Name
ROSMAP

Number of Epochs (Pretrain)

Epochs (Pretrain)
400

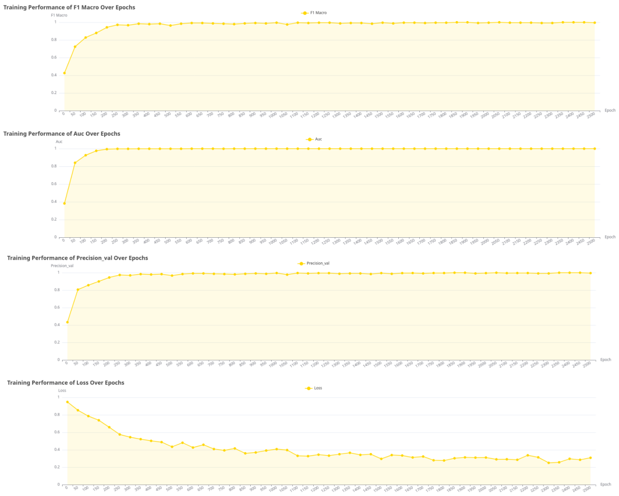
Number of Epochs

Epochs
2400

lr_e_pretrain: 0.001

lr_e: 0.0005

lr_c: 0.001



Training Performance of F1 Macro Over Epochs

Training Performance of Auc Over Epochs

Training Performance of Precision_val Over Epochs

Training Performance of Loss Over Epochs

Fig. 8. Performance during training

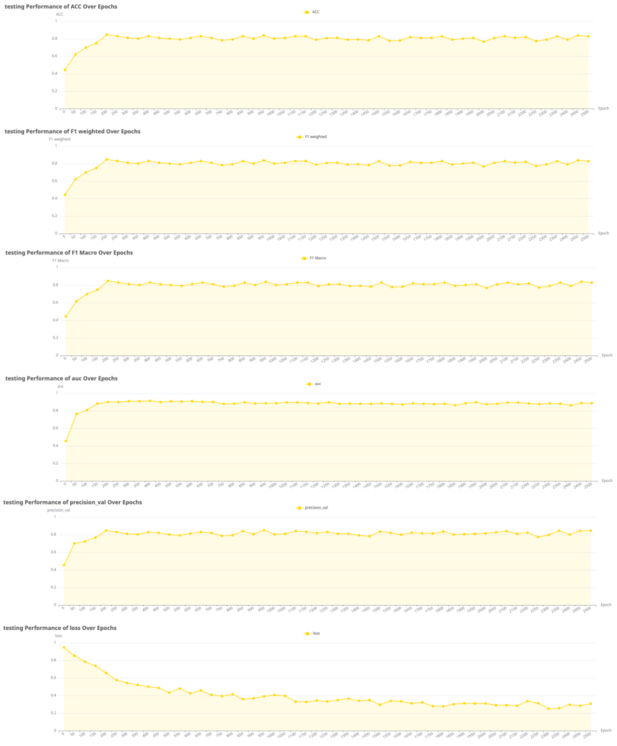


Fig. 9. Performance during testing

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018, ISSN: 0007-9235. DOI: https://doi.org/10.3322/caac.21492. [Online]. Available: https://doi.org/10.3322/caac.21492.

[2] D. Kim, J. G. Joung, K. A. Sohn, *et al.*, "Knowledge boosting: A graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," *J Am Med Inform Assoc*, vol. 22, no. 1, pp. 109–20, 2015, ISSN: 1067-5027 (Print) 1067-5027. DOI: 10.1136/amiajnl-2013-002481.

[3] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Front Genet*, vol. 8, p. 84, 2017, ISSN: 1664-8021 (Print) 1664-8021. DOI: 10.3389/fgene.2017.00084.

[4] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016, [Online]. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv160902907K.

[5] T. Wang, W. Shao, Z. Huang, *et al.*, "Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Communications*, vol. 12, no. 1, p. 3445, 2021, ISSN: 2041-1723. DOI: 10.1038/s41467-021-23774-w. [Online]. Available: https://doi.org/10.1038/s41467-021-23774-w.