Comparative Analysis of Predictive strength of DNA methylation v/s Gene Expression data in Diagnosing/Classifying Tumor Status

1. Introduction

With the development of high-throughput sequencing technology, more and more biological data of samples from tumor tissues are being collected. These biological data, including DNA genome sequence, RNA expression, and DNA methylation levels, will help aid in the diagnosis and treatment of cancer.[1] Databases such as The Cancer Genome Atlas (TCGA), which stores real cancer data, are beginning to be established for such purposes[2].

DNA methylation studies and the detection of RNA expression of certain genes are an important aspect in the diagnosis of cancer. DNA methylation is an epigenetic modification that involves the addition of methyl groups to a DNA molecule, usually affecting gene expression without altering the underlying gene sequence. It is primarily involved in the regulation of gene expression, maintenance of genomic stability and control of cell differentiation, and is one of the key hallmarks in cancer development[3]. RNA-seq, on the other hand, provides a panoramic view of gene expression by quantifying RNA in a sample, thus revealing active biological processes within the cell. According to existing research it has been shown that there are large differences in RNA expression of certain genes between tumor and healthy tissues[4].

In this study, we worked with two datasets, which consist of DNA methylation and Gene expression information. Our aim was to compare which dataset of the two is more suitable for tumor diagnosis. We made use of the machine learning approach, with the goal to create robust models that will be able to differentiate tumor samples from normal ones. The DNA methylation dataset and the RNA-seq dataset (RNA expression) used for training were obtained from kidney cancer samples from The Cancer Genome Atlas (TCGA). All the data were either labeled as ‘solid normal tissue’ or ‘primary tumor’. Each sample in the dataset consisted of several features that are important to differentiate the two samples. However, there were also such features that were not related to tumor or in other words certain features that did not prove very relevant to the current analysis .

Therefore, as part of creating a robust machine learning model to classify tumors from normal samples, we began with a thorough analysis of the provided data. Using “python/scikit learn” we explore the data to understand the data structure, the type of data provided, the distribution of values, presence of outliers or missing values. We then moved forward with the pre-processing of the data in order to prepare the raw data for the learning process. This would include handling missing values, outliers, feature selection. Followed by processing of data, we then trained the data on one model each and compared the outcomes on factors that will be able to identify the better model for the respective data. Finally we selected the best model to test on the unlabeled data for both the datasets. This report provides a brief of the analysis and the obtained results.

2. Method

2.1 Exploratory Data analysis

A preliminary analysis of the data was performed, using functions like head(), describe(), info() to get a statistical overview of the data. Additionally we also performed a visual analysis of the data by creating histograms to understand the distribution of the data. It was

observed that the data was highly imbalanced between the two samples in both the datasets (fig1), implying that results could be biased towards the majority class.

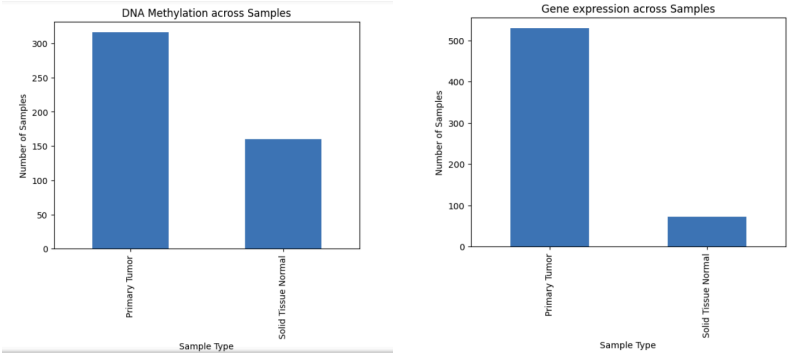


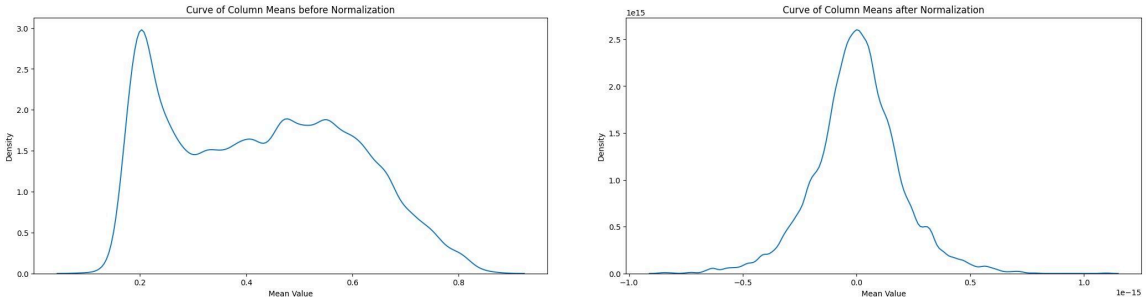
Figure.1 *(a) Distribution of DNA methylation data across tumor and normal samples (b) Distribution of Gene expression data across tumor and normal samples*

2.2 Pre-preprocessing

a) Label Encoding: Any numerical columns were converted into numerical form using Label Encoder

b) Scaling: We then scaled the data using StandarScalar method (fig2), which transforms the data such that the mean of the data becomes 0 and the std. deviations become 1.

DNA Methylation data:



Gene Expression data:

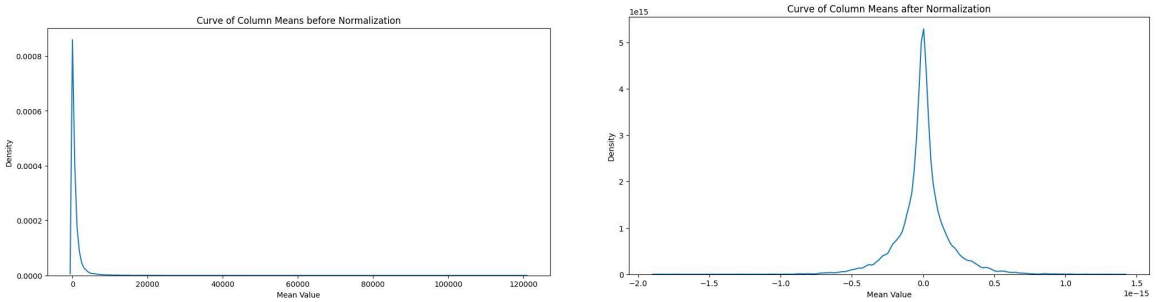


Figure2*: Data set before and after scaling*

d) One- Hot Encoding: We also performed one-hot encoding of the data to convert the Label (tumor vs normal samples) into a binary format for easy interpretation of learning models.

e) Feature selection: This was a crucial part of the process as given the vast amount of features, approximately 5000, it would be computationally complex to train the model on such big data, also causing unnecessary bias during the learning process. Boruta was used for feature selection, which works by selecting optimal features using the Random Forest (RF) algorithm[6].

f) Splitting of the data: Finally, the data was split into a training and a validation set using the train_test_split, where 33% of the data was kept aside for validation of the models.

g) Model Implementations: We implemented two models for the study, mainly Random Forest and Neural Network for the current analysis.

3. Model Selection criteria and Result

In medical diagnosis, it is essential not only to confirm the presence of a disease but also to accurately rule out its absence. This makes sensitivity and specificity crucial metrics in evaluating the performance of a machine learning model. **Sensitivity** (or true positive rate) measures the ability of a model to correctly identify patients with the disease. High sensitivity ensures that most actual cases of the disease are detected, reducing the likelihood of false negatives. A **false negative** occurs when a patient with the disease is incorrectly classified as healthy, which can have serious consequences by delaying or preventing necessary treatment. On the other hand, **specificity** (or true negative rate) evaluates the model's ability to correctly identify healthy individuals. High specificity minimizes false positives, where a healthy individual is wrongly classified as having the disease. False positives can lead to unnecessary treatments, causing undue stress and potential harm to the patient. In cases where both false positives and false negatives need to be minimized, the **F1 score** provides a balanced metric by combining precision (how many predicted positives are actually correct) and recall (sensitivity). The F1 score is particularly useful when the data is imbalanced, as it helps to balance the trade-off between sensitivity and precision, offering a single performance measure that reflects the overall accuracy of positive predictions.

Thus, we make use of the **confusion matrix** in this project to compare the performance of the model between the two datasets. By evaluating these metrics, we aim to select the most appropriate model based on the F1 score, balancing sensitivity, specificity, and precision to ensure robust and reliable predictions.

Outcome:

For the DNA Methylation dataset, we have chosen to implement the neural network model (NNM) as an architecture to create our machine learning model. In our NeuralNetwork, there are two layers, which are the input-to-hidden layer(fc1) and the hidden-to-output layer (fc2). An epoch number of 10000 was selected to maximize the accuracy while minimizing losses. By iteration 9000, the loss is able to be minimized to less than 0.0001 (*Fig3*). The outcome of our neural network model has an accuracy of ~99.4% when tested on our kidney DNA Methylation dataset. When the DNA methylation dataset from unknown tissue is tested on the established NNM, it performed with an accuracy of 60%.

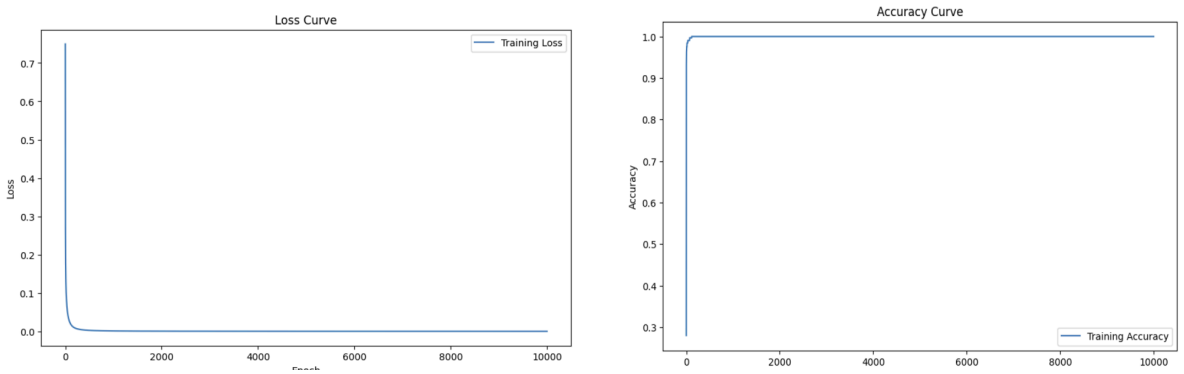


Figure 3: *Loss Curve and Accuracy Curve of NNM model training on DNA-methylation Data*

With gene expression dataset, we have chosen to use a random forest architecture to train our machine learning model. Using Boruta, it has selected 187 features from a list of 20531 features. After the model was trained on the kidney training data, it was able to achieve similarly high accuracy of 99.5% accurate. From the classification report, it scores 1 for precision, but having a lower recall at 0.96, suggesting missing some positive instances. When the gene expression data from unknown tissue was fed to our model trained on gene expression dataset, it achieved an accuracy result of 59.7%. From the classification report, a is 0.55, and recall is 1.00 (indicating all class 0 instances are correctly identified, but many false positives for class 0). For class 1: Precision is 1.00, but recall is low at 0.19, suggesting that the model struggles to identify most class 1 instances. *(Fig 4)*

	precision	recall	f1-score	support
0	0.55	1.00	0.71	190
1	1.00	0.19	0.33	190
accuracy			0.60	380
macro avg	0.78	0.60	0.52	380
weighted avg	0.78	0.60	0.52	380

Figure 4: *Table of the Classification report of RF on unknown gene expression data*

4. Conclusion/Discussion

The classification models applied to both DNA methylation and gene expression datasets uncovered critical insights into the unique challenges and strengths of each data type for cancer detection. By utilizing the Boruta method, we selected a number of key features for DNA methylation and gene expression data that emphasized the complexity of these biological markers. For the DNA Methylation data we chose the Neural Network model, which excelled achieving an accuracy of 99.4% in the known dataset. WHile for the Gene Expression dataset we decided to run the Random Forest model which also delivered near-perfect accuracy. This could imply the potent discriminatory power of the selected features. This exceptional performance raises concerns about potential overfitting, particularly with the neural network, suggesting that the models might be overly tailored to the training data. While this leads to high accuracy on the known datasets, it poses a significant risk to their ability to generalize to new, unseen data, a crucial factor in real-world cancer classification.

When challenged with the mystery datasets, both models showed a considerable drop in performance. With accuracy dropping to around 60% for the DNA methylation data, the neural network model misclassified a substantial portion of class 1 instances as class 0. These findings reveal that while the models are highly adept at identifying 'normal' samples, they struggle to detect 'tumor' samples, likely due to class imbalance in the training data or the intricate and subtle patterns associated with class 1 . In the Gene expression mystery data, the Random Forest model exhibited a clear bias towards class 0 ('Solid Tissue Normal'), perfectly identifying these instances but failing to capture class 1 ('Primary Tumor') accurately, resulting in a low recall of 0.20 for class 1. Addressing this bias is not just a technical necessity but a critical step for improving the models' reliability and generalizability, particularly in clinical applications where precise tumor detection can be life-saving.

The results of our analysis, comparing the predictive power of DNA methylation and gene expression data in classifying "Primary Tumor" versus "Solid Tissue Normal" samples in kidney cancer, suggest that both data types performed similarly. Random Forest, applied to the DNA methylation data, and the Neural Network model, used for gene expression data, yielded comparable classification accuracy across both the available training data and the mystery dataset. However, it is important to note that this similarity in performance may be due to the models chosen for each dataset. Random Forest is computationally efficient, requiring less time and resources than a Neural Network, yet the lack of significant performance difference raises questions about whether the models themselves could be masking deeper insights that might emerge from more sophisticated or varied model architectures.

A critical limitation of this study is that only one model was applied to each dataset type, which means that our findings cannot be considered conclusive. Moreover, since the models were trained solely on kidney cancer data, the features that drive classification in this organ may not generalize to other cancer types. This presents a risk of false negatives when applying the models to samples from other tissues or cancers.

To draw more robust conclusions about whether DNA methylation or gene expression is the superior predictor in cancer classification, further studies are needed. These should involve a broader range of models, such as ensemble learning methods, support vector machines, or deeper neural networks. Additionally, expanding the study to include data from multiple cancer types or organs could provide insights into the generalizability of these data types across diverse biological contexts.

Thus, while our initial findings suggest no clear advantage of either DNA methylation or gene expression data for tumor classification in this setting, a more extensive exploration of models and datasets is required to definitively answer this question.

Disclosure:
This project and the report writing was a combined effort of our team members. John was primarily responsible for the code development while Justy, Han and Peter assisted with the implementation of the codes, report building, and identifying relevant literature. Han contributed to writing the Introduction section, while Justy authored the Methodology section, and Peter was responsible for compiling and writing the Results section. The Discussion section was a joint effort by Han, Peter, Justy, and John. All contributors reviewed and approved the final version of the report.

References:

1. Neyshabouri, M. M.; Jun, S. H.; Lagergren, J. Inferring Tumor Progression in Large Datasets. *PLOS computational biology/PLoS computational biology* **2020**, *16* (10), e1008183–e1008183. https://doi.org/10.1371/journal.pcbi.1008183.
2. Hagan, E.; Proctor, S.; Krane, S. MP28-14 REPRESENTATION OF KIDNEY CANCER IN THE CANCER GENOME ATLAS KIDNEY CANCER COHORT. *The Journal of urology* **2018**, *199* (4), e361–e362. https://doi.org/10.1016/j.juro.2018.02.914.
3. Jones, P. A. DNA Methylation and Cancer. *Oncogene* **2002**, *21* (35), 5358–5360. https://doi.org/10.1038/sj.onc.1205597.
4. Chibon, F. Cancer Gene Expression Signatures – The Rise and Fall? *European journal of cancer (1990)* **2013**, *49* (8), 2000–2009. https://doi.org/10.1016/j.ejca.2013.02.021.
5. de Amorim, L. B. V.; Cavalcanti, G. D. C.; Cruz, R. M. O. The Choice of Scaling Technique Matters for Classification Performance. *Applied soft computing* **2023**, *133*, 109924. https://doi.org/10.1016/j.asoc.2022.109924.
6. Leong, L. K.; Abdullah, A. A. Prediction of Alzheimer's Disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method. *Journal of Physics: Conference Series* **2019**, *1372* (1), 12065. https://doi.org/10.1088/1742-6596/1372/1/012065.

Supplementary-

- The entire code for this project is available on Github (https://github.com/OHM314159/7000-2.git) repository .