Analysing DNA Elements to Understand Transcriptional Regulation

Introduction

Accurate classification of cancer samples – differentiating between 'Primary Tumour' and 'Solid Tissue Normal' is a challenge in cancer diagnostics, with significant implications for patient outcomes and treatment strategies. Advances in high-throughput molecular techniques, such as RNA sequencing (RNA-seq) and DNA methylation profiling, offer new avenues for improving the precision of these classification (Cantini et al., 2021; Wojewodzic & Lavender, 2024). However, it remains unclear which molecular data type provides predictive power for classifying cancer samples. RNA-seq measures gene expression levels, while DNA methylation captures epigenetic modifications that regulate gene expression. Understanding which data type yields superior performance in cancer classification could enhance the development of personalised medicine approaches in oncology.

This study aims to investigate the comparative predictive power of DNA methylation and RNA-seq data in classifying cancer samples. Specifically, we seek to answer the question: which data type performs better in distinguishing between 'Primary Tumour' and 'Solid Tissue Normal' samples? Leveraging datasets from The Cancer Genome Atlas (TCGA), which includes both gene expression and DNA methylation beta values across a range of tissue types – including breast, colon, kidney, liver, and lung cancers – two machine learning models will be developed. One model will be trained on RNA-seq data and the other on DNA methylation data, with their performance evaluated to determine which is more effective in terms of classification accuracy. Additionally, the models will be tested on a shared 'mystery' dataset, which lacks specific tissue information, to assess the model's ability to generalise across different tissue types.

RNA-seq and DNA methylation provide distinct insights into cancer mechanisms. RNA-seq offers a direct measure of gene expression, while DNA methylation represents an epigenetic mechanism that can either silence or activate gene expression by modifying DNA accessibility. These epigenetic changes are known to play a crucial role in tumorigenesis, making DNA methylation a potential early indicator of cancer (Wojewodzic & Lavender, 2024). Meanwhile, RNA-seq provides immediate transcriptional information, reflecting active cellular processes. This distinction highlights the need to compare the performance of these data types in classification tasks.

Previous research has demonstrated the utility of both data types in cancer research. For example, Cantini et al. (2021) showed that combining muti-omics data, including RNA-seq and DNA methylation, significantly improved cancer classification accuracy by capturing complementary biological information. Wojewodzic and Lavender (2024) also found that DNA methylation data, when used with machine learning models, offers highly accurate cancer classifications. These findings underscore the potential of both data types in diagnostic applications.

In this study, machine learning classifiers, such as neural networks, will be applied to both DNA methylation and RNA-seq datasets. Model performance will be evaluated using metrics like accuracy, sensitivity, and specificity to determine which data type offers better classification outcomes. Feature selection and normalisation techniques will also be explored to optimise the models' predictive power by applying these models to both labelled datasets and the 'mystery' dataset, this research will provide insights into the effectiveness of each data type in cancer classification.

The findings aim to contribute to the broader discussion of molecular data's utility in cancer diagnostics. By comparing DNA methylation and RNA-seq data, this research may help clarify which biomarkers offer the most reliable information for classifying cancer samples across various tissue types (Cantini et al., 2021; Wojewodzic & Lavender, 2024). This work could inform future biomarker selection and improve machine learning models for cancer diagnostics, benefitting both research and clinical practice.

Data Set and Method

The dataset used in this analysis consist of gene expression and DNA methylation beta values from specific tissues. These datasets classify samples as "Primary Tumour" or "Solid Tissue Normal" and were sourced from the ACGT muti-omic benchmark datasets. The data includes features such as gene names and CpG probes from the Illumina HumanMethylation450 BeadChip.

Data Scaling and Normalisation

After loading the datasets, data cleaning was performed including handling missing values. Columns where 90% or more values were zero were removed to reduce dimensionality of the data, ensuring only informative features were retained. To ensure consistent scaling, Min-Max scaling was applied to normalise the features between 0 and 1. This step is crucial for models that rely on distance-based and gradient-based computations, such as neural networks, as it prevents bias caused by large numeric ranges and ensures that all features contribute equally to the machine learning process.

Feature Selection and Validation

Feature Selection was performed using the Boruta algorithm, with a Random Forest Classifier serving as the random model to evaluate feature importance. This approach identified the most relevant features for classification while reducing the dimensionality of the datasets.

Model Training – Neural Network

A feedforward neural network (NN) was trained on both the DNA methylation and RNA-seq datasets. Each model consisted of a single hidden layer with ReLU activation functions. For the DNA methylation dataset, Stochastic Gradient Descent (SGD) optimisation was used with 3 hidden units and a learning rate of 0.01. The RNA-seq model employed the Adam optimiser, with 5 hidden units and a learning rate of 0.000005. Both models were trained for up to 100,000 epochs, with early stopping implemented based on validation loss to prevent overfitting. Cross-validation was also applied to fine-tune hyperparameters, ensuring effective performance. This careful design allowed the models to capture the underlying patterns in the data while adapting to the specific characteristics of the DNA methylation and RNA-sequence datasets.

Model Testing – Accuracy, Precision and Recall

Model performance was evaluated using metrics such as accuracy, precision and recall. These metrics were calculated based on predictions made by the trained neural network on a reserved test set, ensuring the model's generalisability.

Model Testing – Liver Dataset

Both the DNA methylation and gene expression datasets were tested independently using the trained neural network. The performance on the liver dataset was compared to determine which data type (methylation or gene expression) provided better predictive power in classification of "Primary Tumour" versus "Solid Tissue Normal".

Model Testing – Mystery Dataset

The mystery dataset, which did not specify the tissue of origin, was used for final testing. The neural network models trained on liver data were applied to this dataset, with predictions made for the binary classification task. The performance metrics from this phase were crucial for evaluating the effectiveness and flexibility of the trained models.

Results

DNA methylation data (5,002 columns) had 0% of columns with >90% zeros, while RNA gene expression data (20,533 columns) had 8.6% of columns with >90% zeros. Following filtering, RNA expression columns were reduced to 18,760, with no columns exceeding 90% zeros, and DNA methylation remained unchanged (see Appendix 1 – Data Cleaning). Min-Max scaling preserved the relative distribution of DNA methylation, while RNA expression showed reduced skewness and a spread across 0-1 range, as seen in Figure 1. A subset of features is added to illustrate the effects of Min-Max scaling for both DNA methylation and RNA sequencing. (see Appendix 2 – Standardisation and Normalisation)

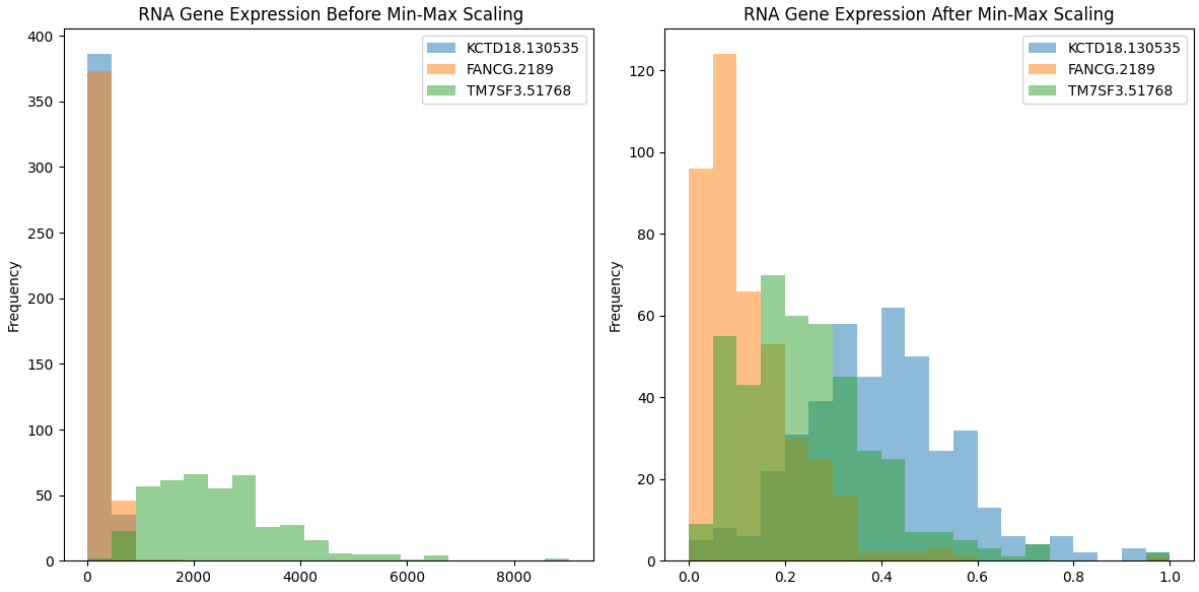

Figure 1: RNA expression distributions (KCTD18.130535, FANCG.2189, TM7SF3.51768) show reduced skewness and spread across 0-1 after min-max scaling.

Figure 3 shows the performance of the two classification models when tested against an unseen subset of the training dataset (test data), and separate unseen "mystery" dataset. Appendix 3 includes accuracy, recall and precision metrics for both models when tested on the training data, test data, and a mystery dataset independent of the training data. Accuracy is used as a rough metric to assess how often the models make correct classifications. Recall is also a useful metric for performance because the cost of a false positive classification in cancer diagnosis is high, so recall should be maximised to prevent unnecessary costly and potentially harmful further treatment. Precision is also relevant, as a cancer screening tool with a high rate of false negatives will prevent cancer patients from receiving further treatment until a later stage when prognosis will likely decline.

Both the DNA methylation and RNA-seq models appear to perform moderately well on the reserved test dataset, with accuracy >97%. However, the DNA methylation model has lower precision (89.5%) but perfect recall, whilst the RNA-seq model is the opposite with lower recall (75%) but perfect precision. This result can be understood by looking at the confusion matrices to see that in Figures 1A and 1B, the DNA methylation model generates false positives whilst the RNA-seq model generates false negatives, though both models' false classifications form less than 5% of the total dataset. A training and testing dataset containing more Primary Tumour samples would help to see the performance more clearly, as the dataset used contains far more healthy samples than tumour samples.

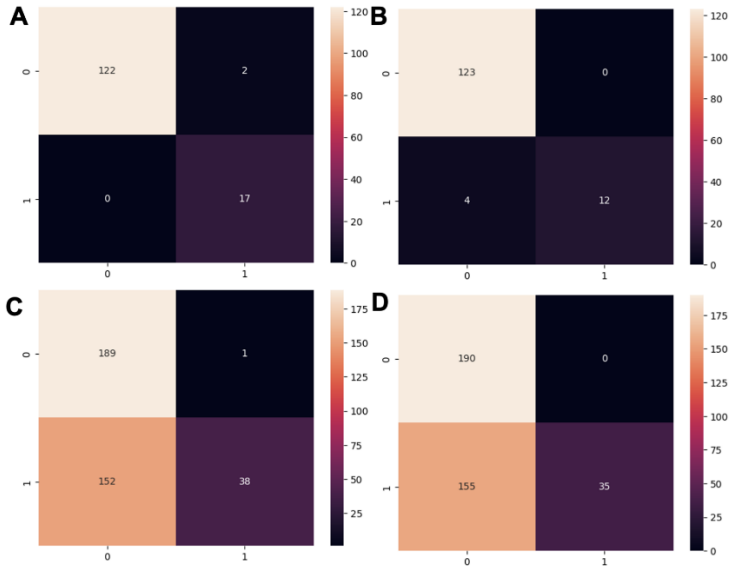

Figure 3: Confusion matrices from testing of the DNA methylation model (A and C) and RNA-seq model (B and C) on data from an unseen partition of the training dataset (A and B) and a separate unseen "mystery" dataset (C and D). Vertical axis represents predicted label, horizontal axis represents true label, with 1 = 'Primary Tumour' 0 = 'False Solid Tissue Normal'.

Whilst there are differences in performance when tested on the test data, which comes from the same dataset as the training data, both models perform similarly poorly on the truly independent "mystery" dataset, with accuracy of approximately 60%. Both models have recall values of approximately 20% that indicate poor discernment of positive results leading to many false negatives, seen in Figures 1C and 1D. Interestingly, the models have high precision (>97%), generating few or no false positives.

Both models appear to be overfitting to the training data, as performance is perfect when tested on the test data, but imperfect on test data that comes from the same dataset as the training data (see Appendix 3). This overfitting may explain the poor performance of the models on the independent "mystery" dataset, as the overfitted models have learnt the details of the training dataset rather than learning the more general patterns that would allow the models to generalise to new data effectively.

Overall, the two models appear to both perform well on the independent "mystery" testing data when considering the low number of false positives, but also perform poorly in that they produce many false negatives. Alongside the strong suggestion of overfitting, the results suggest the training data may be missing some key features or patterns that are present in some cases of liver cancer seen in the "mystery" test dataset. This may also be caused or exacerbated by flawed feature selection; perhaps there are biologically significant genes or methylation regions that often differ in liver cancer tissue compared to healthy tissue but aren't present in the training dataset used here.

Discussion

After conducting feature selection using computational methods, it is essential to validate the selected features through biological relevance, particularly in the context of gene expression, DNA methylation and liver cancer mechanisms. Based on relevant literatures, we identified a set of 47 genes that are consistently reported as being either over-expressed or downregulated in LIHC. These genes represent key molecular association to the progression and pathology of LIHC. To further assess the robustness of the features filtered through Boruta, we compared these 47 biologically relevant genes with the list of features retained by Boruta. This comparison revealed that 23 genes from the Boruta output overlapped with our gene set.

This finding highlights the importance of integrating biological validation with computational approaches, as many computationally selected features may not align with known biological markers. For DNA methylation, Xu et al. (2017) use similar computational approach to develop a diagnostic predictive model, which the feature selection involved LASSO and random-forest, and markers overlapped are selected. Although the 10 markers chosen from Xu et al. (2017) are not present in our features selection, this may be due to the different feature selection method used. High stringency in Boruta select features based on their statistical significance compared to random features, even if LASSO and random forest both found it significant, it could still be discarded as it did not show statical significance compared to noise. The limited overlap between the two sets underscores the need for further investigation to determine whether the remaining features from Boruta may have novel, yet undiscovered biological significance or if additional refinement of the feature selection process is required.

By comparing the two types of data, DNA methylation model slightly performs better than the RNA-seq model, which RNA-seq model has a relatively lower recall than DNA methylation model. This may be due to the biological difference between DNA methylation and RNA-seq data. Methylation patterns are usually more stable over time, while RNA-seq measures the gene expression levels, which is more dynamic and can be affected by environmental stimuli, cell signalling or disease progression. The dynamic features of RNA-seq data result in a noisier dataset, which makes the model harder to avoid false negatives, leading to lower recall.

However, the results above also suggested that the neural network models are overfitting, which the model performs well on the training set, while performs poorly in the mystery data set. Overfitting occurs when the model closely adapts to the training model, learning even noise and irrelevant features, resulting in not generalizing in new data set. One significant reason for overfitting is that feature selection removed biological significant features which are not as statistically significant as other features, resulting in missing of important features that the model cannot learn from. As above mentioned, features remain after features selection does not strongly correlate with the published literatures, we suspect that the faulty feature selection results from Boruta is one of the reasons for overfitting. To address overfitting, regularization and ensemble methods can be utilized to reduce noise in the data and reduce variance within the dataset respectively. Moreover, we can use two feature selection methods and take the overlapping results, which can avoid bias faulty results.

In conclusion, the DNA methylation model slightly outperformed the RNA-seq model, which might be due to biological difference between methylation and RNA-seq data. Yet, both models are overfitting as they perform too well on the training set but fail to perform similarly in the mystery data, which might be due to the feature selection process.

References

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, *12*(1), 124. https://doi.org/10.1038/s41467-020-20430-7

Wojewodzic, M. W., & Lavender, J. P. (2024). Diagnostic classification based on DNA methylation profiles using sequential machine learning approaches. *PLOS ONE*, *19*(9), e0307912. https://doi.org/10.1371/journal.pone.0307912

Xu, R. H., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., Yi, S., Shi, W., Quan, Q., Li, K., Zheng, L., Zhang, H., Caughey, B. A., Zhao, Q., Hou, J., Zhang, R., Xu, Y., Cai, H., Li, G., Hou, R., … Zhang, K. (2017). Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature materials*, *16*(11), 1155–1161. https://doi.org/10.1038/nmat4997

Appendix

Appendix 1: Data Cleaning

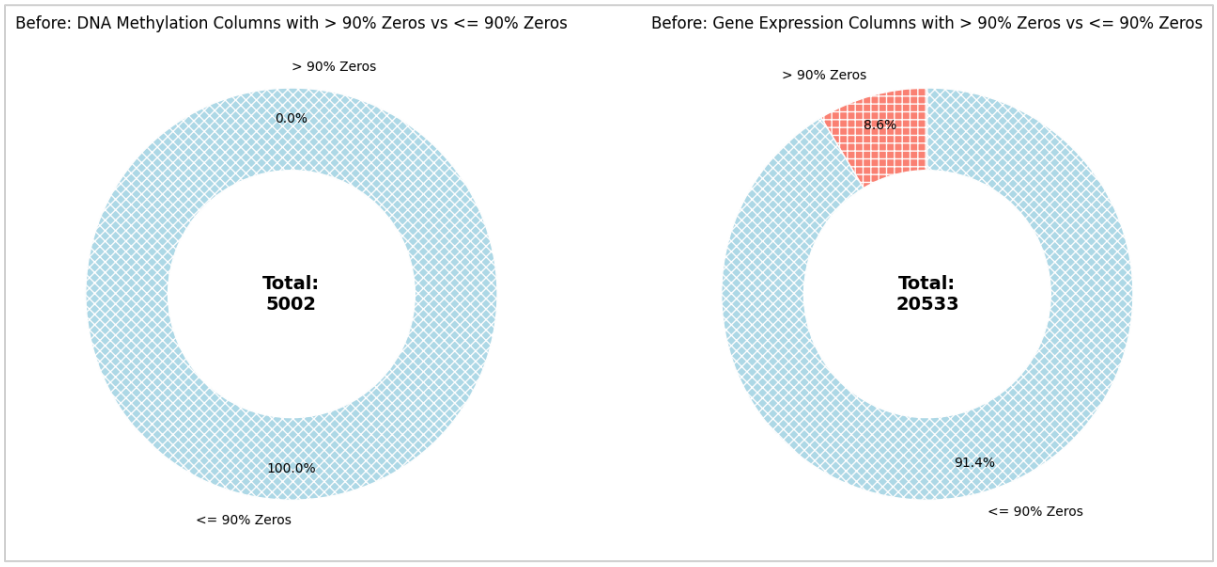

Figure 1A and Figure 1B: Before filtering, all DNA methylation columns (5,002) had ≤90% zeros, while 91.4% of RNA expression columns (20,533) had ≤90% zeros and 8.6% had >90% zeros.

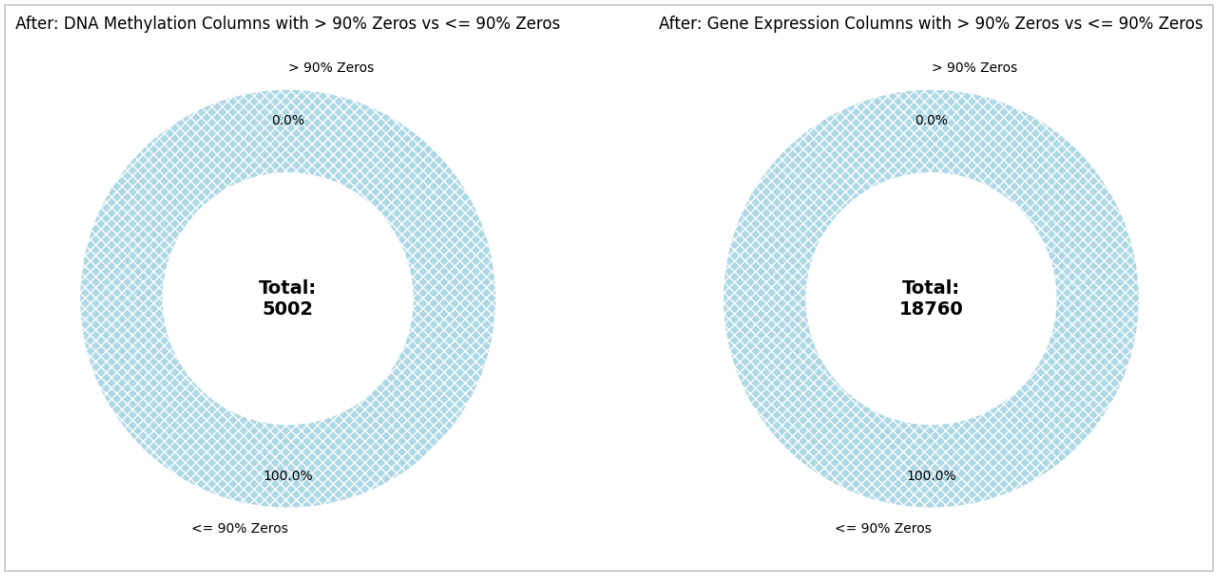

Figure 2A and 2B: After filtering, all DNA methylation (5,002) and RNA expression columns (18,760) had ≤90% zeros, with no columns >90% zeros.

Appendix 2: Standardisation and Normalisation

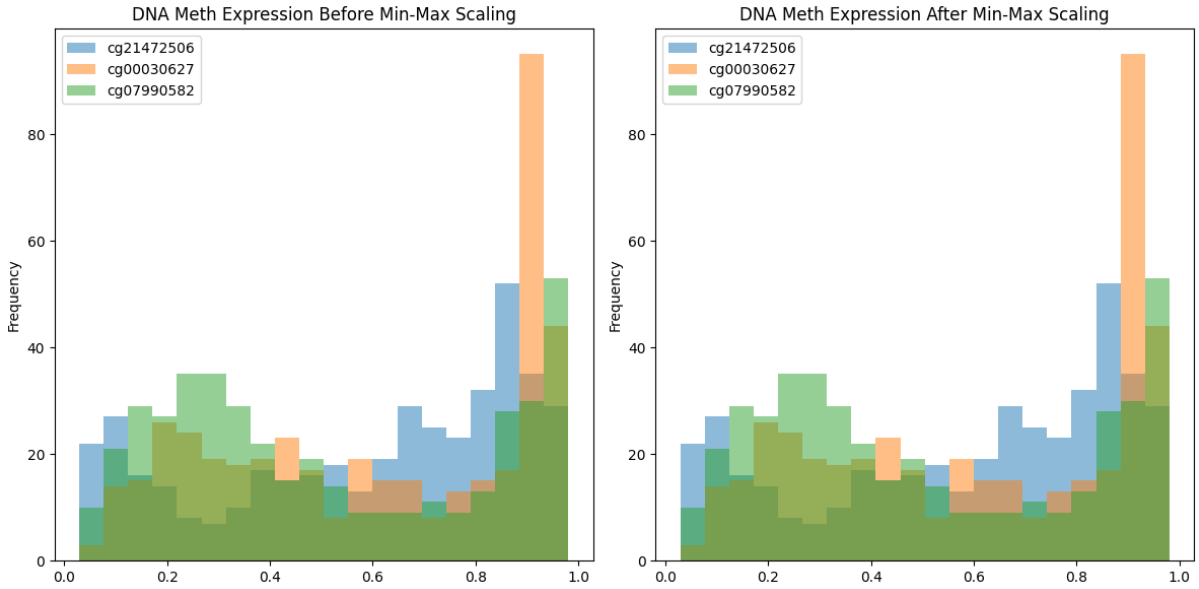

Figure 3: DNA methylation site distributions (cg21472506, cg00030627, cg07990582) before and after min-max scaling show minor shifts.

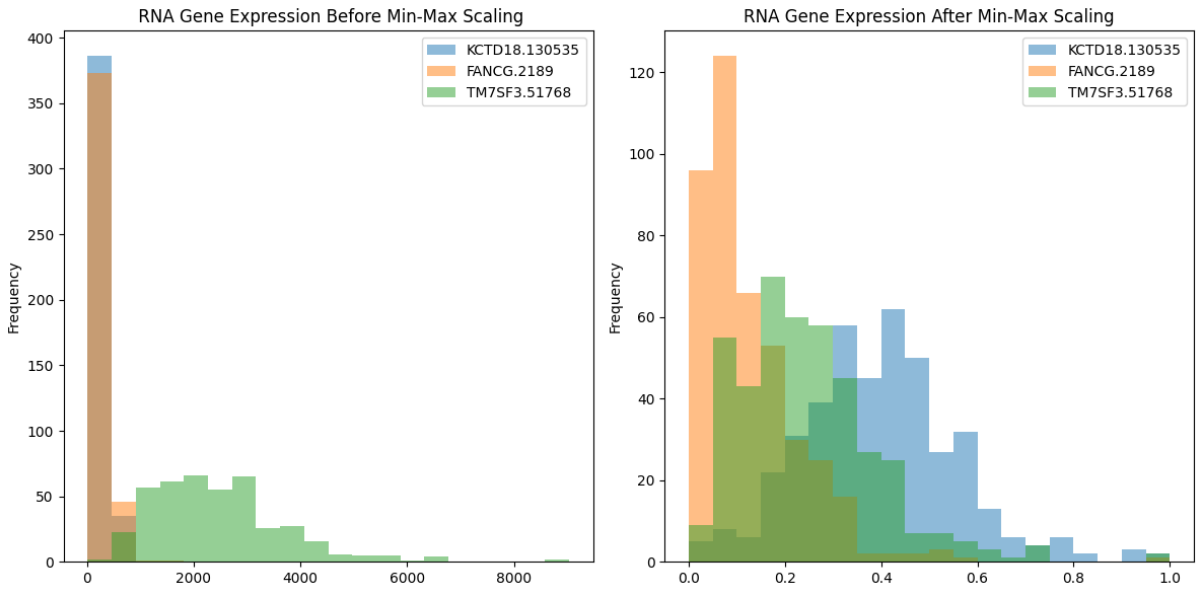

Figure 4: RNA expression distributions (KCTD18.130535, FANCG.2189, TM7SF3.51768) show reduced skewness and spread across 0-1 after min-max scaling.

Appendix 3: Performance Metrics

Table 1: Performance metrics for the DNA-methylation and RNA-seq classification models, tested against training, test, and "mystery" datasets

Model	Testing dataset	Accuracy	Recall	Precision
DNA-meth	Train	100%	100%	100%
	Test	98.6%	100%	89.5%
	Mystery	59.7%	20.0%	97.4%
RNA-seq	Train	100%	100%	100%
	Test	97.1%	75.0%	100%
	Mystery	59.5%	18.9%	100%