NATIONAL UNIVERSITY OF SINGAPORE
DSA4262: SENSE-MAKING CASE ANALYSIS HEALTH AND MEDICINE

FINAL REPORT

Name:

Andrea Cho (A0205675B)

Carine Ho (A0205168H)

Noorus Suhaina (A0205289Y)

Sanraj Mitra (A0200075X)

Shreya Sriram (A0205079E)

Group: genericteam

To access our github repository, please click here.

Problem Statement

m6A which is known as N6- methyladenosine is when a methyl group can be added to an adenosine molecule at N6 position. Such a post-transcriptional modification is vital in regulating the different biological processes such as protein synthesis (Khan et al., 2022) and as such helps to maintain the normal physiological function of cells as well as organisms. However, any mutations in the m6A sites are closely linked to life-threatening diseases like cancer (Hendra et al., 2021). As such, it is important to identify these m6A modification sites to identify and overcome these diseases.

Using the nanopore direct RNA sequencing technology which allows for direct, real-time analysis of RNA fragments, we are able to capture information like the direct RNA current for each RNA molecule and the dwelling time of the molecule as it passes through a nanopore. Using such features, we developed a supervised machine learning model to predict the m6A modification sites.

Feature Engineering

Numerical and categorical features were created:

1. Mean and variance: we studied the distributions in terms of the central tendency (mean) and variability (variance). Due to the large number of reads, the values follow a gaussian distribution approximately by the Central Limit Theorem.
2. Purpose of using left, centre and right positions: the distribution of the neighbouring 1-flanking position may provide information in predictions for the current transcript position.
3. Nucleotide bases count: the frequency of A, T, G and C in the nucleotide sequence (positions 1 through 7, except positions 4 and 5 since they are always ‘AC’) is counted. The frequency count does not depend on the order of nucleotide bases hence it may be useful.
4. Relative transcript position: since transcript position may vary significantly across cell lines which results in large ranges, the relative position was considered instead. To obtain the relative position, we grouped by the transcript ID, and ranked the transcript position.
5. One-hot encoding of nucleotide sequences: from the given DRACH motif (middle 5-mers) with the neighbouring 1-flanking position, one-hot encoding was performed for the nucleotides from positions 1 through 7, except positions 4 and 5 which are ‘A’ and ‘C’.

Exploratory Data Analysis

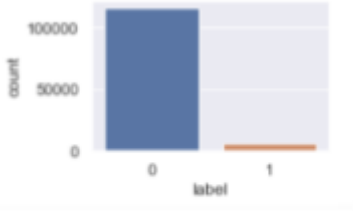


Figure 1: Class Imbalance in Training Data

There is a class imbalance in our dataset with the majority class being unmodified m6A sites while the minority class being modified m6A sites with a ratio of roughly 23 to 1. We will account for this class imbalance in our hyperparameter tuning of our model using scale_pos_weight.

Distribution

We explore differences in feature distributions between the modified sites and unmodified sites to better understand how each feature contributes to the prediction of m6A sites.

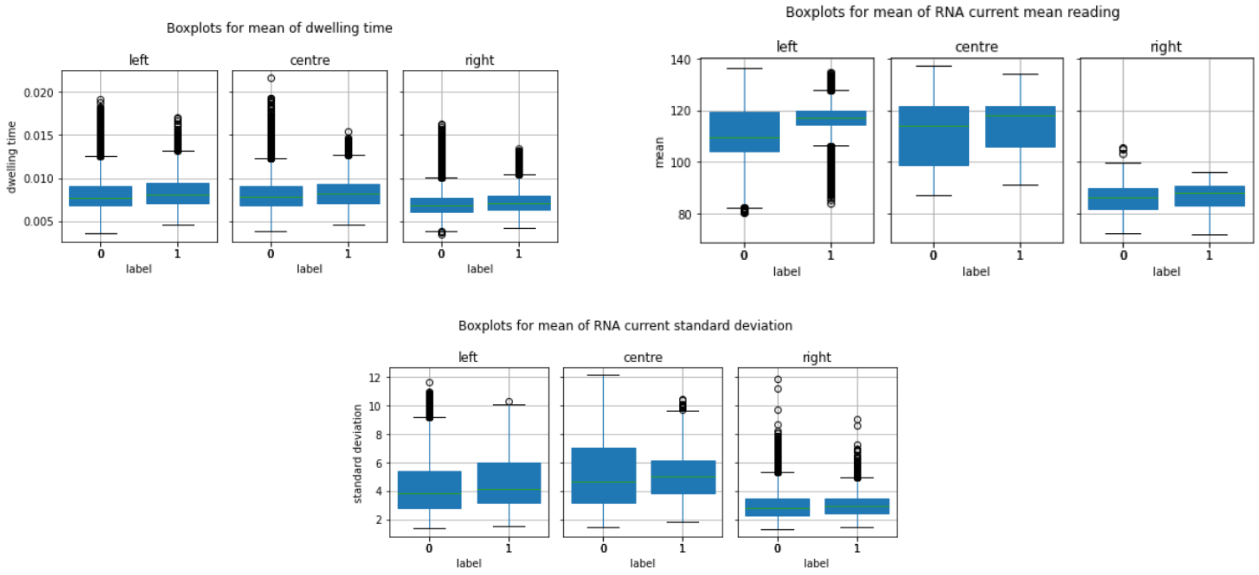


Figure 2: Comparing the mean readings of dwelling time and direct RNA current (its mean and standard deviation)

Mean direct RNA current for the left position and the centre position shows the strongest difference between modified and unmodified sites. However, we observe that the mean dwelling time for the left, centre and right position shows the smallest difference when comparing between modified and unmodified sites. As such, this indicates a possibility that mean dwelling time of the 3 different positions may not be as informative in predicting the m6A sites.

Methods: SG-NEx Dataset

The SG-NEx dataset consists of Nanopore RNA-Seq of 5 cancer cell lines drawn from human samples:

- A549: lung cancer cell line (2 replicates)
- Hct116: colon cancer cell line (3 replicates)
- HepG2: liver cancer cell line (2 replicates)
- K562: bone marrow leukemia cell line (3 replicates)
- MC57: breast cancer cell line (2 replicates, both replicates are the exact same)

To obtain the 12 datasets, the *wget* command was used to download the files from the SG-NEx server. The SG-NEx data was accessed on 21/10/2022 at registry.opendata.aws/sg-nex-data.

The number of replicates indicates the number of repeated samples taken of a single cancer tissue. The replicates were obtained from growing the cell cultures in different days or different dishes, which results in variation of information extracted from the Nanopore Sequencing

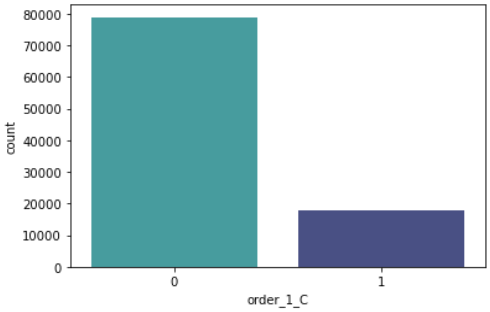
Technology. Our objective, going forward, is building a model that predicts m6A modification sites within these 12 datasets.

Feature Selection

We performed a 80-20 split using gene id's of the dataset and performed the following feature selection techniques on the training dataset. This is to combat any form of data leakage.

Baseline Feature Selection: Variance Threshold

Preliminarily, categorical variables with low variance - namely those that only had one value in 80% of samples in our training dataset were removed based on the Variance Threshold feature selector. This was performed as such features provide limited new information for modelling.



'order_1_C 'is an indicator of whether the first base of the left 1-flanking position was a 'C'. As seen in Figure 3 on the left, it is the only feature exhibiting extremely low variance, indicating that 'C' appeared less than 1 in 5 times for that position. Hence, this feature was removed.

Figure 3: Bar Chart of Order_1_C feature distribution

Recursive Feature Selection

Extreme Gradient Boosting Model (XGBoost) was our model of choice. To limit overfitting, an optimised approach to feature selection was vital. We used Recursive Feature Elimination (RFE), which progressively reduces model complexity from its original 39 features, by removing features one by one until the model is optimised based on XGBoost's feature importance score.

This process involved:

1. Varying the number of features retained in RFE's parameter 'n_features_to_select' between 20 (51% of features retained) and 39 (all features retained).
2. For each set of features chosen, manual cross validation using gene_id's with 5 folds was performed to aggregate performances across different training and validation sets.

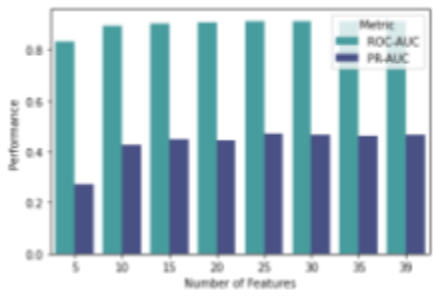


Figure 4: Model Performance Across Different Numbers of Features on ROC-AUC and ROC-AUC

The model's performance on both ROC-AUC and PR-AUC was optimal when 25 features (64.1% of features) were used. As seen in figure 4 on the left, the model performance slightly decreased from using 39 features to 25 features, and then significantly decreased after using less than 25 features. As we want to make our model less complex, we chose to use 25 features that gave a relatively high performance while still remaining fairly uncomplex. Mean dwelling time for the 3-flanking

positions were dropped which is congruent with our exploratory data analysis which showed the least difference between the modified and unmodified sites. (Figure 2)

Manual Implementation of Cross Validation to Eliminate Data Leakage

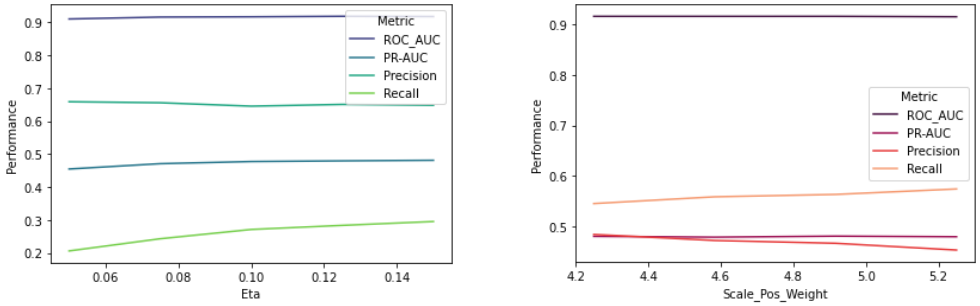
Cross validation was manually implemented using pre-set segmentations of 5 folds, **where no two folds had overlapping gene IDs**. This prevents data leakage - making predictions on a gene ID that was trained on could lead to gene-specific data within training data being "seen" by the test set. Our model would overestimate its performance, which is a problem when in practice, it would be used for predictions on **entirely new** genes from different species/tissues.

Hyperparameter Tuning

Certain critical hyperparameters from our XGBoost model were selected for hyperparameter tuning. Namely, max_depth, n_estimators, eta, and scale_pos_weight:

1. Max_depth: maximum distance between root and leaf node
2. N_estimators: controls number of weak learner trees
3. Eta: Analogous to learning rate
4. **Scale_pos_weight**: deals with imbalance in class labels. Higher value leads to more emphasis placed on minority class (likely with higher recall, but lower precision)

Hyperparameter tuning was conducted in a sequential manner, where eta was tuned first. After it was optimised, scale_pos_weight was tuned with eta kept constant at 0.15.



Figures 5 & 6: Effect of changing eta (left) and scale_pos_weight (right) across key metrics

The same was performed for max_depth and n_estimators with prior hyperparameters fixed. Based on Figure 7, small variations in max_depth heavily impact precision and recall. A good tradeoff was critical.

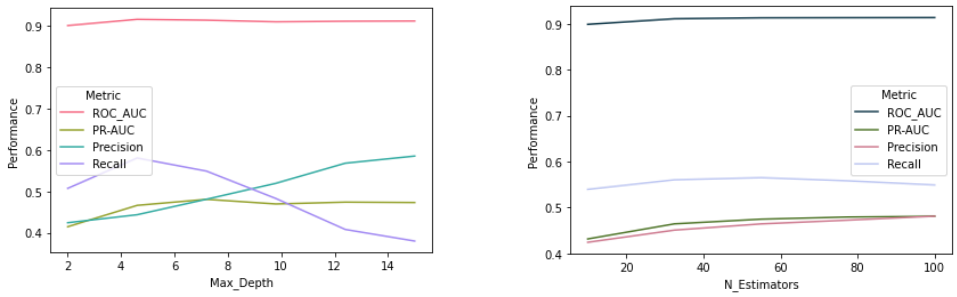


Figure 7 & 8: Effect of varying min_data_in_leaf and scale_pos_weight across key metrics

Varying hyperparameters independently of each other (sequentially) was imperative for **clear visualisation** of the effects of changing individual hyperparameters on **multiple** metrics.

A manual implementation of cross validation was once again utilised to prevent data leakage. The XGBoost Classifier's performance was optimised with max_depth = 7, eta = 0.15, n_estimators = 85, scale_pos_weight = 4.9.

Results and Discussion (Task 1)

The performance of the models are based on the test set generated from the 80-20 split.

Baseline Model Performance

The base model is a Logistic Regression model with no hyperparameter tuning performed and only features provided by the original dataset were used. Training data was scaled using the MinMaxScaler from sklearn and applied to the test data.

Features used are:

- Mean of left, centre and right of dwelling time, standard deviation, and mean of reads per group of Gene ID, transcript ID & Position Number.
- One-hot encoded nucleotide sequences.

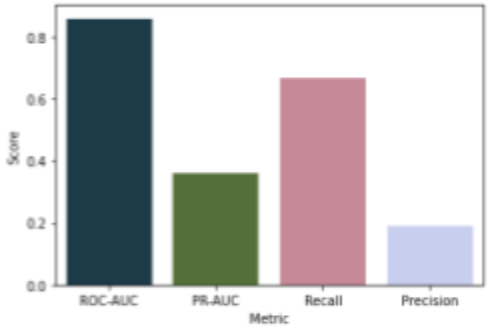


Figure 9: Base Model Performance Across Key Metrics

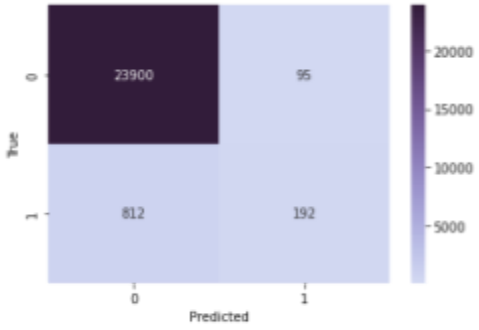
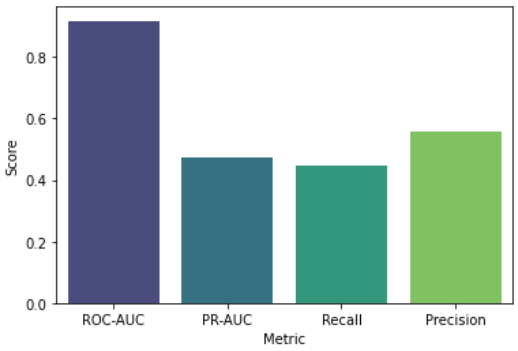


Figure 10: Confusion Matrix of Base Model Performance

Final Model Performance

The optimised XGBoost Model is trained on the 25 chosen features concluded from EDA and Feature Extraction and the hypertuned values described in the section above.



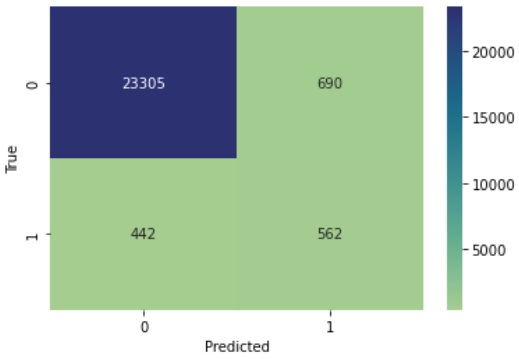


Figure 11: XGB model performance across key metrics

Figure 12: Confusion Matrix of XGB Model Performance

ROC-AUC

The XGBoost achieved an ROC-AUC Score of 0.918, indicative of extremely strong performance in distinguishing both classes along various thresholds. It performs significantly better than the LR baseline model whose ROC-AUC Score is only 0.859. However, given the class imbalance, PR-AUC Score may provide a better indicator of the model's ability to correctly identify 1's (presence of m6A modification), since it does not account for true negatives (correctly identifying lack of m6A modification).

PR-AUC Score

The model achieved a PR-AUC Score of 0.475. This is a significant improvement over the baseline model, whose PR-AUC is only 0.361. It is also considerably better than a random classifier, which would achieve a mere PR-AUC score of 0.04, given the heavy class imbalance.

Precision & Recall

Based on Figures 9 & 10, the XGBoost model's precision of 0.449 indicates that 562 of 1252 (44.9%) test points predicted positive for m6A modification actually had an m6A modification. Likewise, a recall of 0.560 indicates that 562 of 1004 (56.0%) of all m6A modification points in the test set were successfully identified.

Results and Discussion (Task 2)

Prediction of m6A sites on SG-NEx Samples

Our model made predictions on the 12 SG-NEx cancer cell lines, as described below. Figure 13 aggregates the mean performance of replicates of the same cancer cell line. The transcript is classified as m6A modification when the probability is greater than or equal to 0.5.

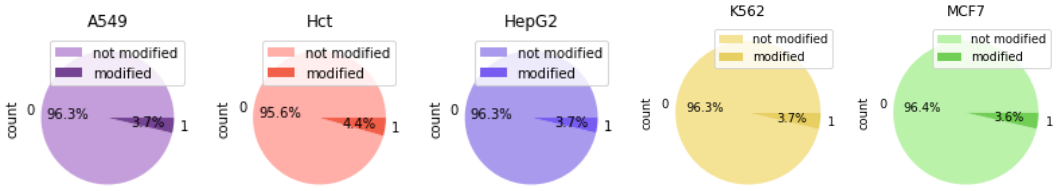


Figure 13: Pie Chart of Aggregated m6A Modification Rate Across 5 Cell Lines

The model identifies approximately similar rates of m6A modification (hovering around 96%) amongst all 5 tissues. It is interesting that the predicted rate of modification on the tissue it was trained on (Hct116), is approximately similar to that of other tissues. One possibility is that there are many overlapping transcript IDs across tissues, leading to similar rates of m6A modifications. As such, we did further analysis on transcript IDs.

Variance across Replicates: Hct116

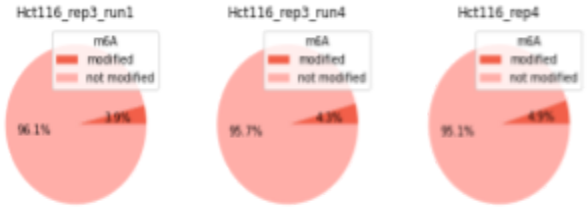


Figure 14: Pie Chart of m6A modifications in Hct116 across replicates

Between replicates, only the Hct116 cell line had a notable difference in rate of modifications in Hct116 across replicates, with a 1% difference between the highest and lowest, perhaps because between replicates, certain features were hugely different. There was a huge difference in the number of reads & distinct transcripts IDs between the two. This may be caused by noise and experimental differences in the environment when the cell line was grown.

Macroscopic Transcript Analysis: Transcript ID and Model Performance

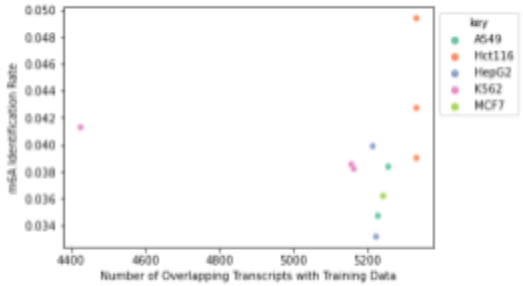


Figure 16: Scatter Plot of Overlapping Transcripts & m6A Modification Rate

"Number of Overlapping transcripts with Training Data" refers to the number of distinct transcript IDs shared by our training data (Hct116) and the current predicted cell line (e.g. A549 replicate 1). m6A identification rate refers to the percentage of data points in a predicted cell line (e.g. A549 replicate 1) that contain m6A modifications. We found a notable positive correlation between the number of overlapping transcripts with training data and m6a identification rate. This correlation is 0.55 excluding an outlier replicate for cell line K562, which could have experimental errors. Since the model was trained on the Hct116 dataset, this high correlation may account for Hct116 having a higher m6A identification rate than any other cell line, having a high number of overlapping transcripts with its training data. Now knowing that transcript ID is critical, we performed an in-depth analysis of specific transcript IDs' effects on m6A modification rate in subsequent sections.

Microscopic Transcript Analysis: Identifying Key Transcripts with High Modification

Modification rate = # of positions with modification present in transcript X / total # of positions in transcript X. Each transcript is associated with 12 data points in the box plot below.

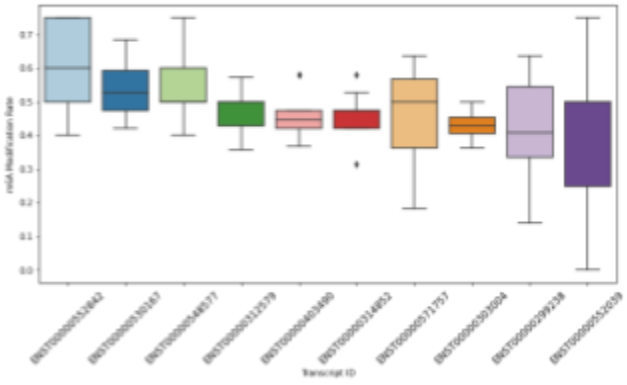


Figure 17: Boxplot of transcript IDs with highest m6A Modification Rate

Figure 17 shows the distribution of the top 10 transcripts present across all 12 cell lines with the highest m6A modification rates. We found the number of modifications for each of these transcripts within each replicate divided by the total number of times it appeared in that replicate (specifically - number of unique positions per transcript). We generated a box plot describing the distribution of the modification rates across the 12 replicates.

Transcript ENST00000552842 has the highest mean m6A modification rate across 12 replicates. Based on the box plot, it also has a visibly high variance compared to the others. In the following section, we examine ENTS00000552842 more closely.

Breakdown by Cell Lines and Replicates: Transcript ID 'ENST00000552842'

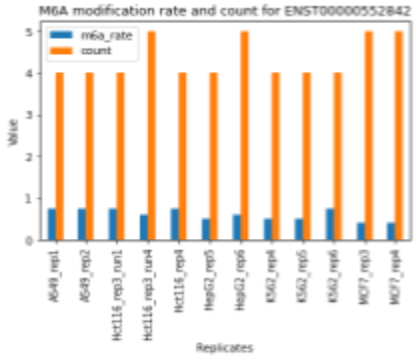


Figure 18: Cell Line and Replicate Breakdown for ENST00000552842

The transcript with the highest modification rate and count of 52 is 'ENST00000552842' across the 12 replicates. It had a modification rate of 0.596 across all replicates, and looking at the figure 18, we see that the lowest modification rate occurs on MCF7 replicates while the highest occurs on A549, HCT116_rep3_run1 and K562_rep6. ENST00000552842 is found in RNASEK gene (Protein Atlas), which is linked to prostate cancer in humans (National Center for Biotechnology Information). There may be a link between the 5 cancer cell lines (lung, colon, liver, bone marrow, breast) and prostate cancer.

Breakdown by Cell Lines and Replicates

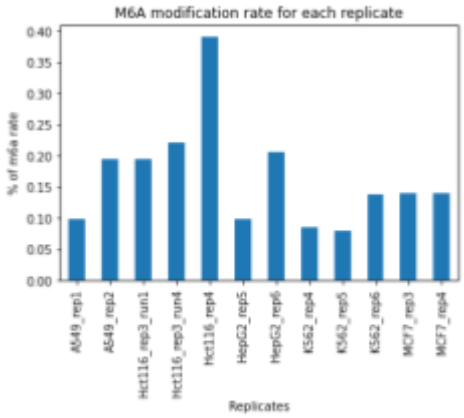


Figure 19: Cell Line and Replicate Breakdown for all Transcripts

For each replicate, we found the number of unique transcripts that have a modification rate of at least 0.3 and has a count of more than 20 in that replicate for a reliable assessment. From Figure 19, Hct116_rep4 has the highest modification rate of about 0.4%, whereas K562_rep5 has the lowest rate of 0.07%. This is an interesting observation as we would expect modification rates between the same cell lines to be similar, but that is not the case here. The only consistent rate is found in MCF7 datasets. This is an interesting observation to conduct further research in.

Limitations

One of the limitations faced was potentially high false negatives present in the training labels, caused by errors (such as human error and random error) in m6A detection during experimentation. Hence, the XGBoost model may pick up noise from the training dataset. Since training and testing was done on the human colon cancer dataset, extrapolating on other tissues and species may not be accurate and our findings may be inaccurate.

Future Works

Other modelling approaches such as the deep Multiple Instance Learning can be explored, since it is a powerful deep neural network which helps in scenarios without fully annotated data. Neural networks are capable of independently performing feature engineering and feature extraction which may deliver promising results.

References

1) Khan, A., Rehman, H. U., Habib, U., & Ijaz, U. (2022). M6A-finder: Detecting M6A methylation sites from RNA transcriptomes using physical and statistical properties based features. *Computational Biology and Chemistry*, *97*, 107640. https://doi.org/10.1016/j.compbiolchem.2022.107640
2) Hendra, C., Pratanwanich, P. N., Wan, Y. K., Goh, W. S. S., Thiery, A., & Göke, J. (2021). Detection of M6A from direct RNA sequencing using a multiple instance learning framework. https://doi.org/10.1101/2021.09.20.461055
3) Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
4) Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
5) The Human Protein Atlas. (n.d.). *RNASEK (HGNC Symbol)*. RNASEK protein expression summary - the human protein atlas. Retrieved November 6, 2022, from https://www.proteinatlas.org/ENSG00000219200-RNASEK
6) National Library of Medicine. (n.d.). RNASEK ribonuclease K [homo sapiens (human)] - gene - NCBI. National Center for Biotechnology Information. Retrieved November 6, 2022, from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=440400