Machine Learning-Based DNA Methylation Score for Fetal Exposure to Maternal Smoking: Development and Validation in Samples Collected from Adolescents and Adults

Introduction:

The paper aims at finding an effective way to predict if a person is exposed under maternal smoking using DNA methylation data. Datasets used for training , testing and validating are three datasets with sample number smaller than 1000. Before finding suitable models, they do data pre-selecting by statistical method, which make the computation less expensive without harming the performance. To solve the unbalance dataset problem, they use SMOTE to create ‘artificial’ data in the minority class. After pre-processing, they try about ten network structures and fine tune four of them with better performance. Their result shows that elastic net perform best on this problem, and their model achieve better sensitivity and specificity than existing models.

Background:

1. T-value, p-value and significant difference

P-value is a statistical index that can be used to determine whether the class result induce significant difference on a feature. To understand how p-value is calculated, we should first introduce t-value. The equation for t-value is provided below:

In brief, t-value is the difference of a feature between two classes of objects normalized with respect to their standard deviation. Larger t-value implies that difference of the feature between to classes are more significant. P-value is simply the probability the variable is larger than t-value under Student's t-distribution. Lower p-value implies that it is more likely there are differences between two set. We can set a threshold p-value (normally 0.05). When a feature has lower p-value than the threshold, we say that there is significant difference and select the feature.

1. Bonferroni correction

Using p-value to select hypothesis gives a upperbound on type 1 error (wrongly rejecting null hypothesis). However, when there are multiple hypothesis, the type 1 error rate increase. Bonferroni correction is a conservative way to protect the model from such effect. If there are n hypothesis, the threshold p-value is divided by n (harder to show significant difference). However, the method is not perfect. Because the correction is too conservative on type 1 error, it makes the model vulnerable to type 2 error.

1. SMOTE

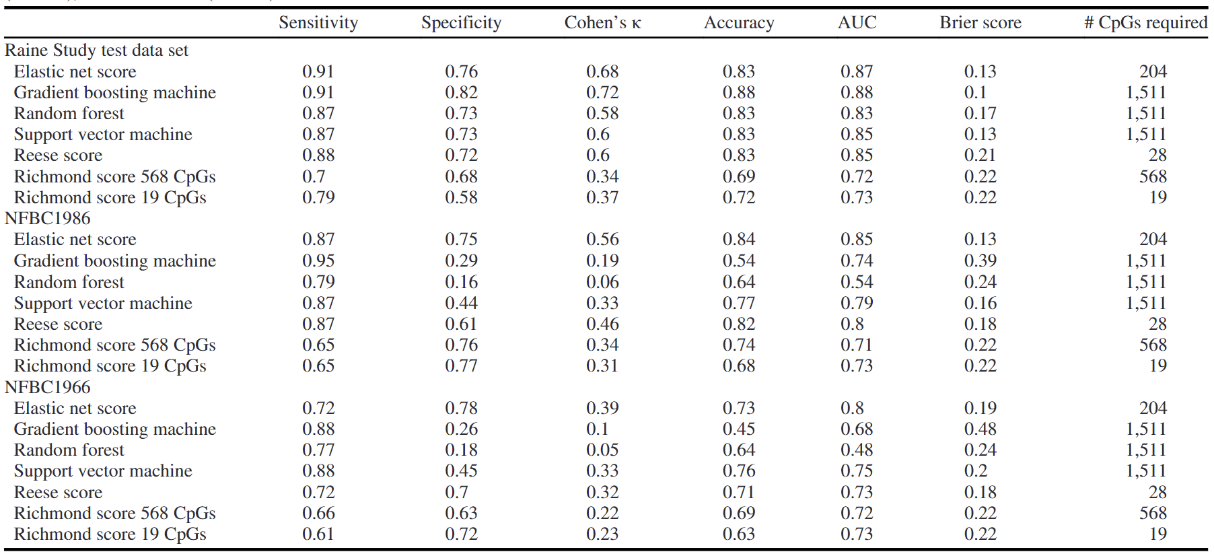
Smote is a method to enlarge the minority data to deal with unbalance dataset. First for a minority data point, find its neighbors. Then, randomly choose one from the neighbor set and create an artificial data point between them. This make the cost of making mistake on minority class higher.

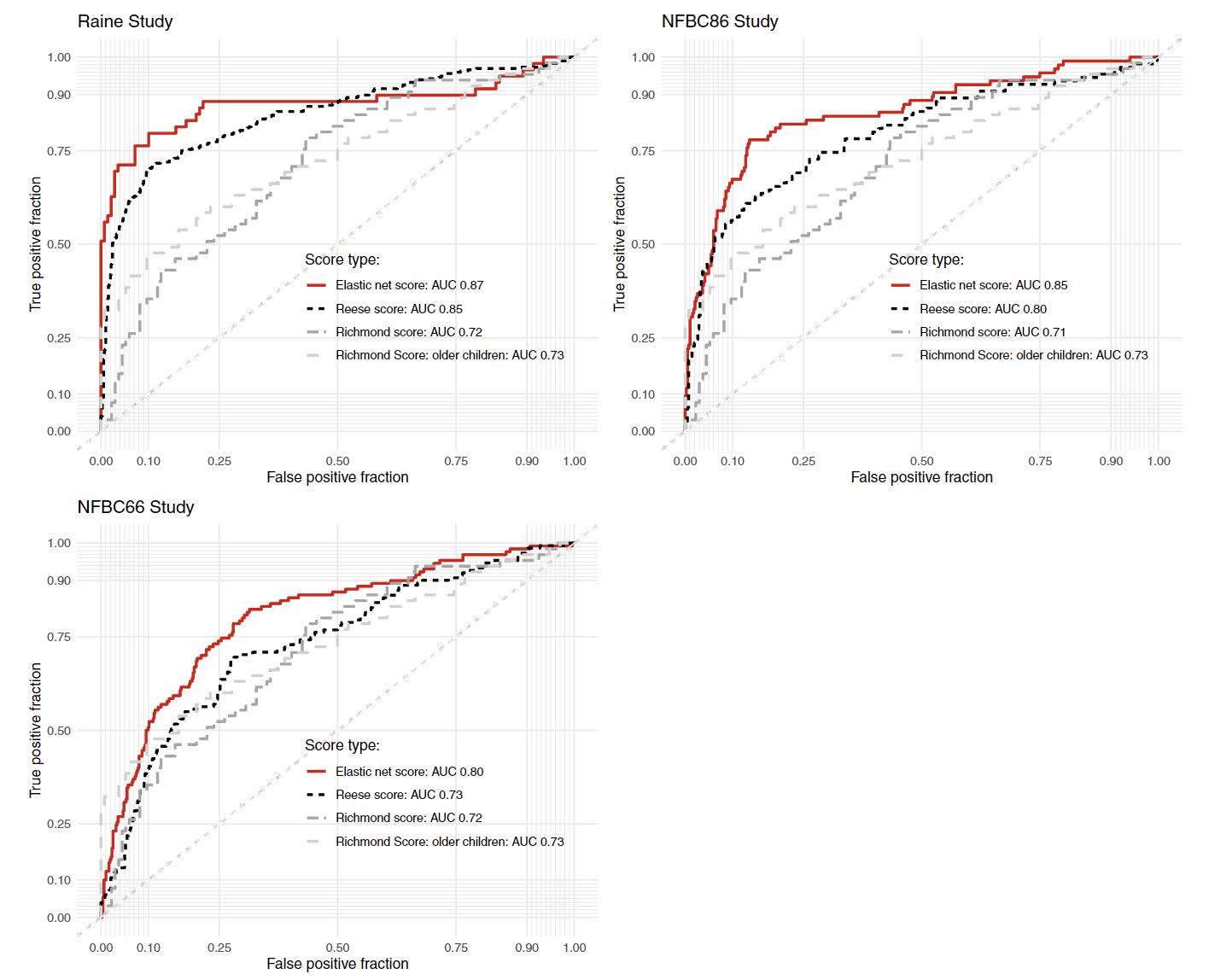
Discussion:

This work does not contain any fancy network structure. All models they used are standard one provided in suites. The interesting part are the data pre-processing. Because of the high dimension of DNA methylation data, feature selection becomes an important process. Using p-value helps selecting data truly causes difference, which makes the computational cost affordable and avoid overfitting. For the unbalance data problem, SMOTE generate more data by hand, which increase the cost of misclassifying minority class samples.

Result & Comparison:

* Result show that, after fine tune, elastic net regression had the best overall performance on the task.
* Reese score: On Raine set, the performance shows no significant difference between elastic net and Reese score both on the ROC curve and statistical metrics. While on NFBC1986 and NFBC1966, the model provided outperform on nearly all metrics.
* Richmond: On Raine set, the performance is much better than Richmond score, which can be observed in the ROC curve. While in NFBC1986, the specificity is slightly lower, while other metrics shows significant advantages. For NFBC1966, elastic net outperforms Richmond in all metrics.





However, there are still some parts ambiguous in these processes. First, for the pre-selecting algorithm, they introduce Bonferroni correction to control type 0 error. While at the end, the level they choose for significant difference are between the value with and without correction. Since the given threshold has no statistical significance, I don’t think introducing Bonferroni correction is needed. I think the threshold value is simply a fine-tuned result. Next, I doubt that if using SMOTE to generate artificial data in the minority class is always feasible. This method strongly relies on how data points are distributed in the space. If data points are not linear separable, this method might cause trouble since a point between two in same class does not necessary belongs to the same set.

**A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking**

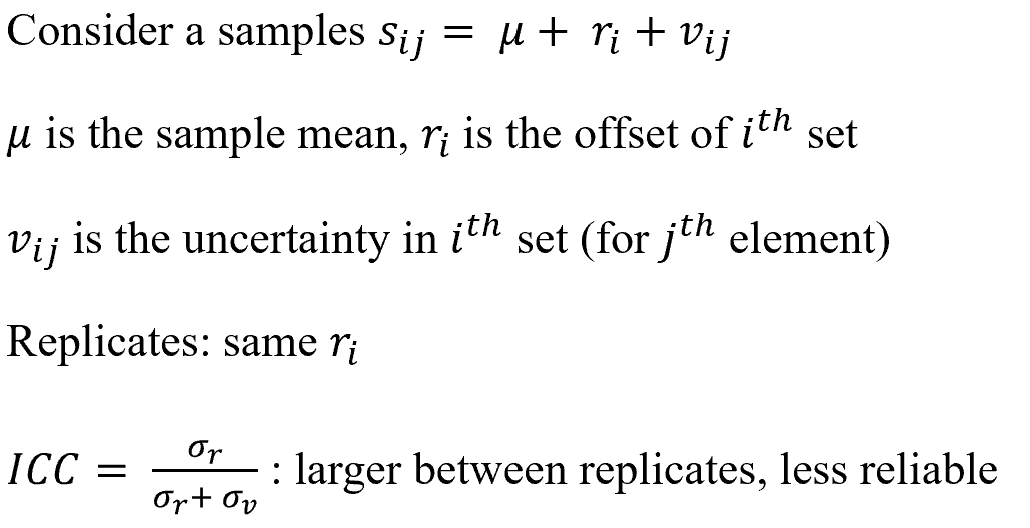
Introduction:

Epigenetic clock can be helpful on clinical diagnosis. However, existing epigenetic clock has limited application because of a serious problem: reliability. It is found that existing epigenetic clock give predictions with large variance between technical replicates, which should be, ideally, exactly the same. Thus, enhancing reliability become an important issue. In the work, it is found that simply filtering feature using ICC does not lead to good solution, and perform weighting still suffer from the noise. Moreover, the authors found that base on statistical analysis, there are much more gene related with aging than those are considered in existing clocks. Thus, they choose to apply PCA on the data given to include more information and reduce the noise effect. Their result shows that principal components indeed have better reliability than original CpGs, and using CpGs for training can reduce the sample size and dimension needed to achieve convergence.

Background:

ICC:

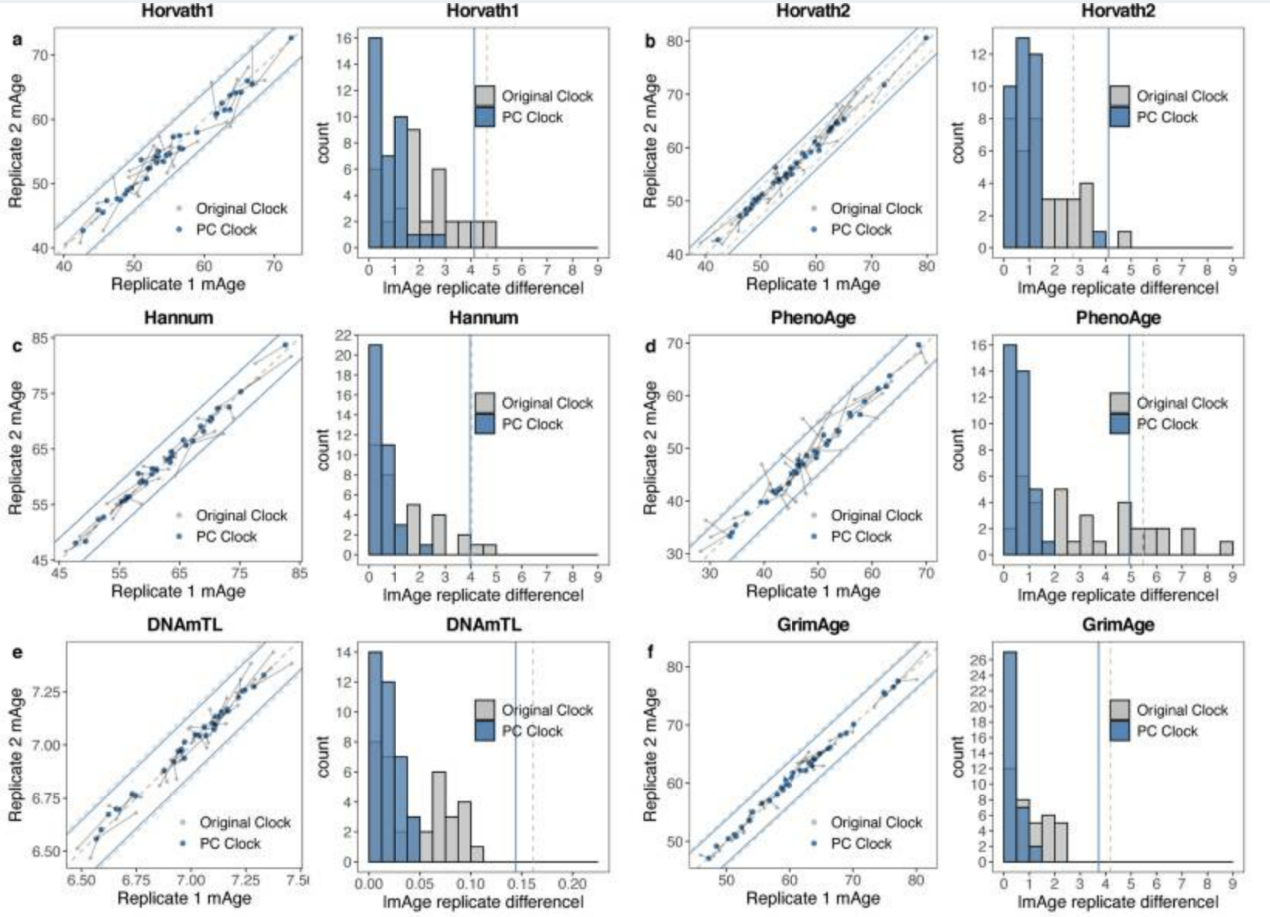
In brief, interclass correlation coefficient (ICC) is a number between 0 and 1, used to represent how similar members in a small group are, compare with the entire dataset.

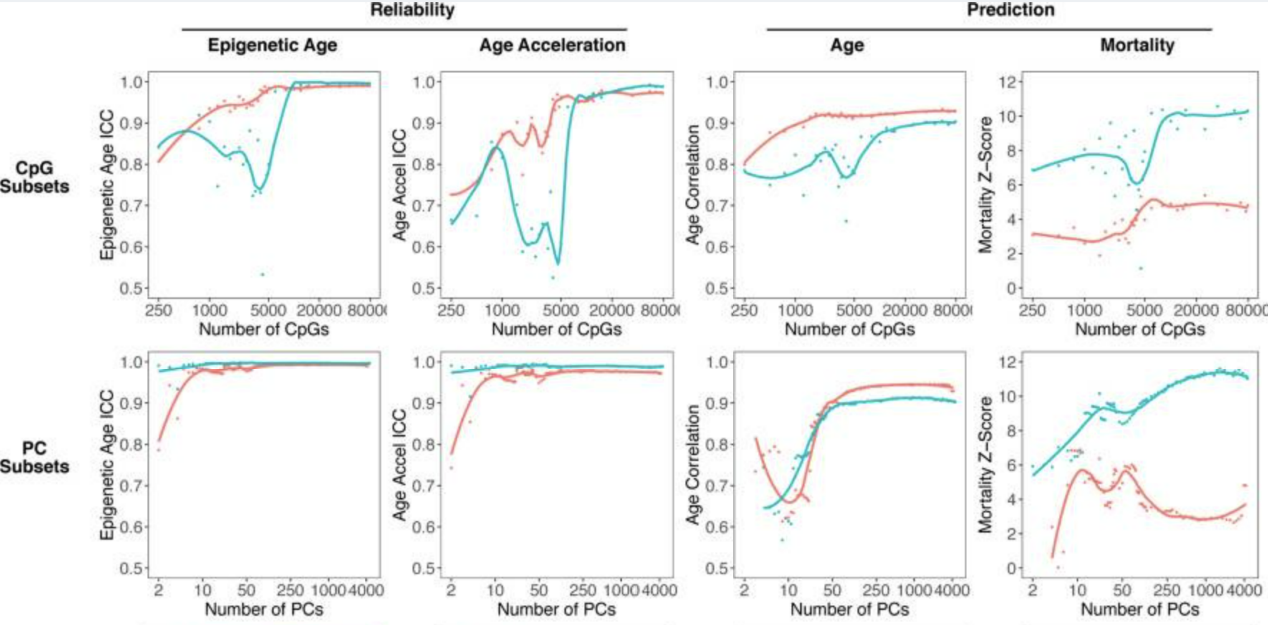
In brief, the coefficient is the variance of interest divided by total variance. For example, if we want to evaluate how reliable a test is, we can repeat the test on same sample to create replicates. The variance of interest is the variance between different samples, and the total variance is variance of interest plus variance among replicates.

Comment:

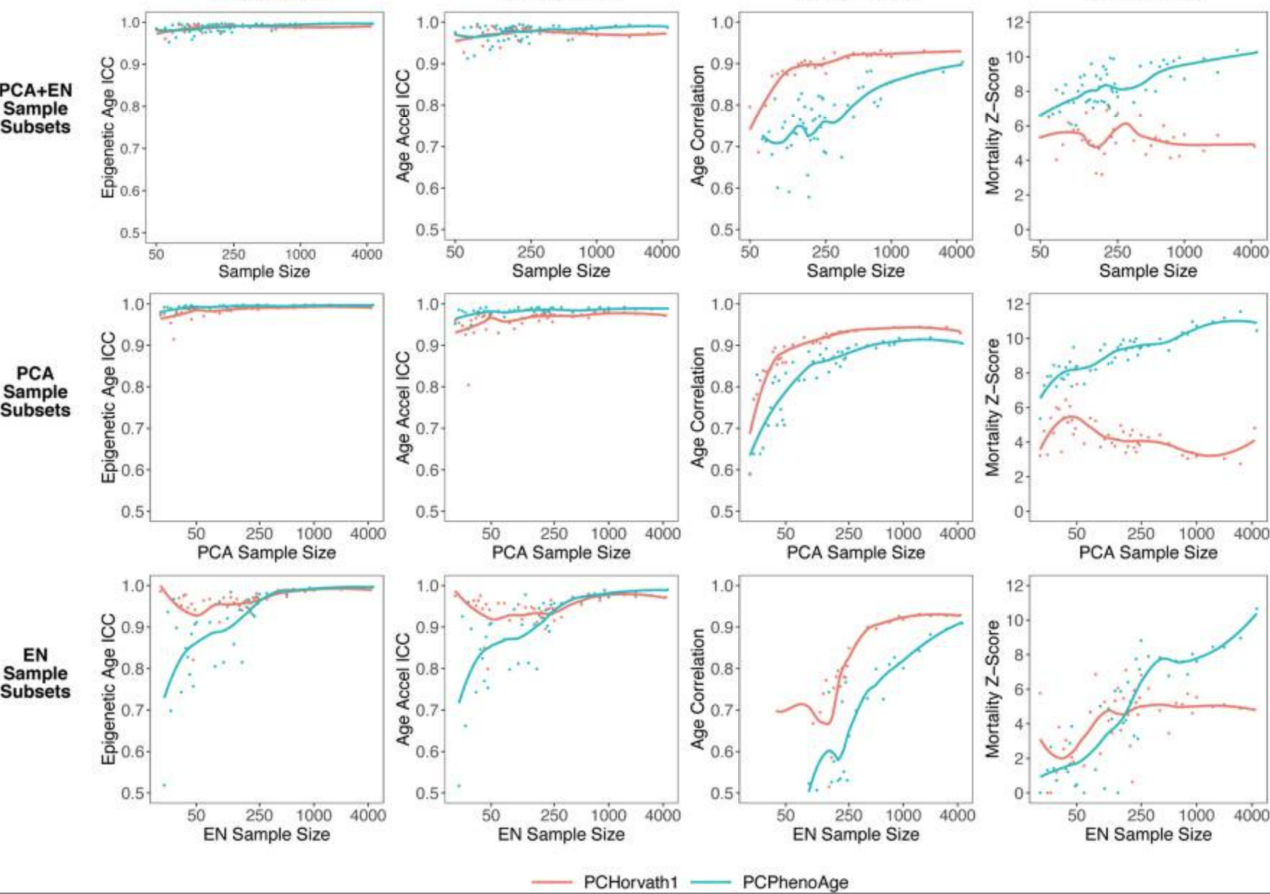
This work shows that sometimes simple structure can lead to good results. PCA allows the model to include more information (since they observe that there are much more CpGs related with aging than those already included in the existing clock) while keeping the dimension the same. Moreover, the noise is reduced effectively and the reliability of the clock has increased significantly. To be mentioned, their data shows that sometimes PC with low variance contain important information. Thus, after doing PCA, they didn’t filter out PCs by hand. Instead, they apply elastic net regression, using the regularization term to select feature automatically. They claim that this method can get rid of terms representing noise effectively.

Result & Comparison:

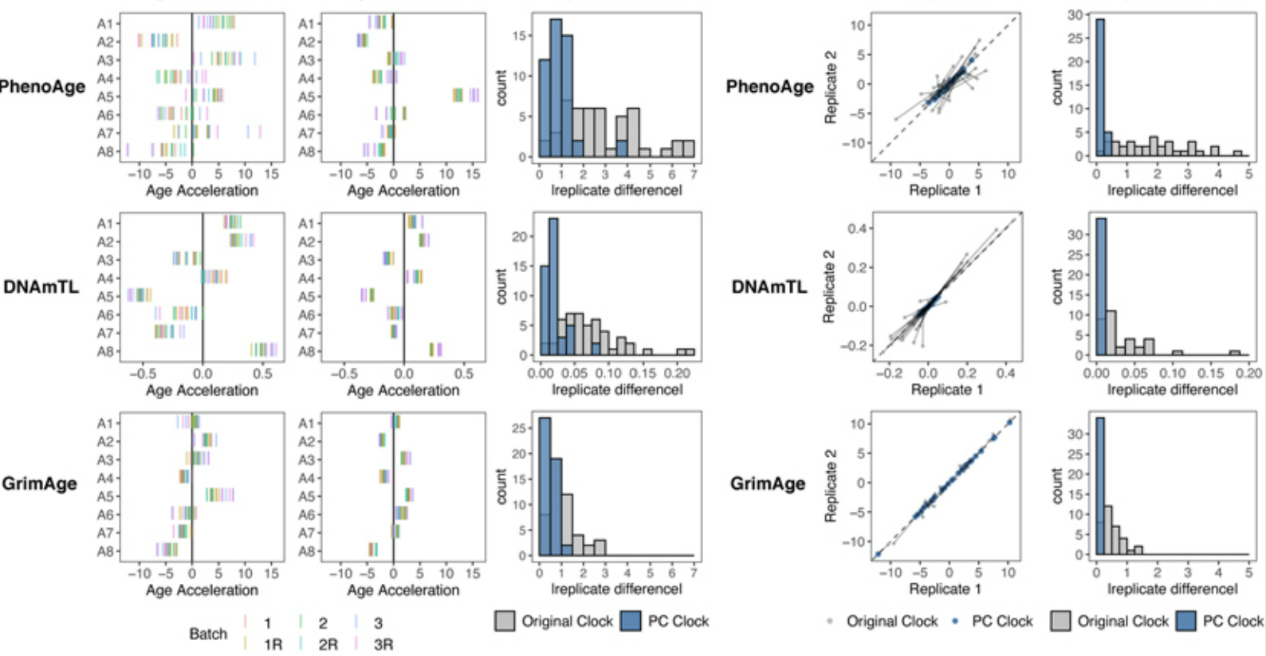
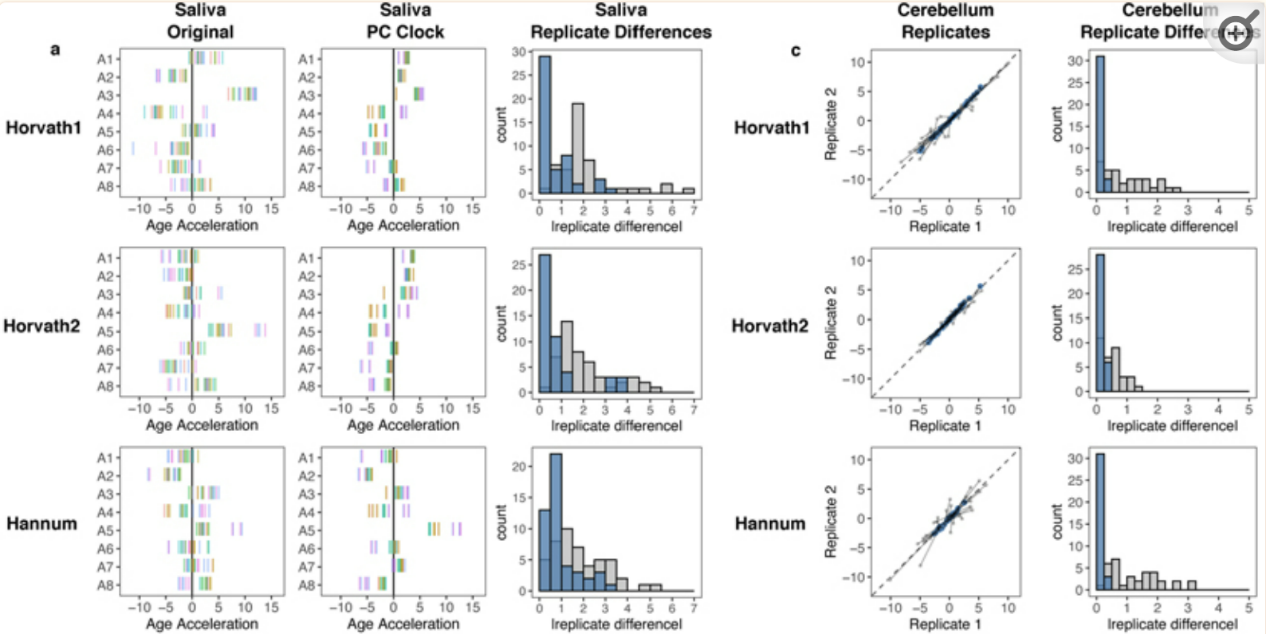
* The graph gives the relation between technical replicates. It is shown that using PCA, the ICC of features (CpGs for original clocks, PC for new method) has increased significantly, which implies that the new clock should have better reliability.
* Using PCA, less dimension is needed for achieving same performance.



* Probably because of the higher reliability of features, PC clock need less sample to achieve same accuracy. Notice that the number of samples cannot be too low, since the number of principal components can be extracted is upper bounded by sample number.



* Though the clock is trained using blood-samples, the team claim that, to some extent, their result can be applied to other tissue such as saliva and brain. The graph shows that PC clock still achieve lower variance between technical replicates when applied on those tissue.



Other Issue:

When observing that weighting doesn’t help the reliability much, the team claim that it is because the noise is amplified simultaneously. We think this is not the true reason. Weighting features with higher ICC do increase the SNR mathematically. We think that the true reason for limited performance improvement needs further investigation.

Drug Response Prediction Based on 1D Convolutional Neural Network and Attention Mechanism

Introduction:

In modern clinical medicine, many seemingly identical symptoms respond very differently to drugs. To apply most suitable treatment at the first timing, predicting the drug response on a patient using genetic information become an important topic. This work uses a dataset containing expression, methylation, copy number variation to develop a drug response prediction model. They first integrate the three-dimensional data into biological pathways, then apply simulated annealing to do feature selection. Last, they use 1D convolutional network with attention to train the predictor. They claim that the method performs better than random forest approach.

Background:

Attention:

Attention is initially applied on language related task.

Simulated Annealing:

Discussion:

The team treat the prediction problem as a binary classification problem. The response profile is separated into two class, i.e. responders (including complete and partial response) and nonresponders (including stable disease and progressive disease). After defining the problem, this work spent lots of time on data processing. They first use gene synthesis to score the training set. The score is then mapped onto KEGG pathway to obtain the weight of subapthways on each training data. Now, the training data are integrated into functional blocks. Then they apply simulated annealing to do feature selection. The advantage of such process is that the result is somehow explainable since the features used for training are functional blocks. After pre-processing, they got 17x3 features on expression, methylation, copy number variation. They then apply CNN, pooling, dropout, and softmax function to produce the final prediction.

Result & Comparison

I don’t think the result given by this work is solid. The team only provide their training curve on testing and validation, and comparison between random forest and the proposed structure. No data from other related works are provided.

Other Issue:

Although the attention mechanism is mentioned in the article, the team provide no information about how they integrate attention module to their model.