Sequence-Based Deep Learning DNA Methylation Prediction and Whole-Genome Epigenetic Aging Prediction

Sophie Parsa and Marc Huo
Department of Computer Science
Stanford University
Stanford, CA 94305
sjparsa@stanford.edu, marchuo@stanford.edu

1 Abstract

DNA methylation is as an epigenetic regulator of gene expression and has been implicated in cell differentiation, cancer progression, and gene regulation. Recently, methylation of CpG regions was shown to occur de novo in regions related to key developmental genes, implicating its role in the biological process of aging. However, current predictive models of DNA methylation state and aging index suffer from the profound shortage of cost-inefficient manually annotated genomic sequences and a priori defined features. Thus, we developed a proof-of-concept deep learning model to circumvent such bottlenecks, allowing for imputation of missing methylation state by DNA sequence and prediction of an epigenetic age index. Our models leveraged the use of a convolutional autoencoder to learn the latent representations of genomic sequences as well as a series of fully-connected stacked convolutional layers and a Bidirectional LSTM. Overall all supervised learning models achieved accuracy above 80% with future improvements possible by addressing class imbalance. We also trained a random forest regression model for age prediction that was able to learn a high correlation between complete methylation profile and chronological age and outperformed our deep learning model.

2 Introduction

Reliable predictors of biological age can play important roles in early disease prediction and prevention. Currently much research is being done to develop accurate prediction models of cellular age based on molecular factors. Commonly investigated factors include telomere length, metabolomic and transcriptomic markers, as well as proteomic variations. However, the most promising molecular mark is DNA methylation, the addition of methyl groups to nucleotides.

DNA methylation has been implicated in a wide range of biological processes, including "chromosome instability, X-chromosome inactivation, cell differentiation, cancer progression and gene regulation" (Jones, 2012). Overall, DNA methylation acts as a regulatory mechanism: methylated regions are associated with condensed chromatin and can act as gene suppressors if located near a promoter, and can also downregulate if located near a repressor. It was recently shown that DNA methylation can occur de novo in regions related to key developmental genes, which implicates its role in the biological process of aging (Rakyan et al., 2010). Furthermore, DNA methylation alterations have been shown to contribute to the progression of cancer due to hypermethylation of tumor suppressing genes and hypomethylation of tumor enhancing genes (Rakyan et al., 2010). In previous studies, DNA methylation age acceleration has also been associated with many neurodegenerative diseases such as Alzheimer's, Parkinson Disease, Huntington's, and Amyotrophic Lateral Sclerosis (Yokoyama et al., 2017).

Methylation of DNA cytosine residues are often found in the sequence context CpG (Laird, 2010), where unmethylated CpGs are predominantly clustered in large groups called CpG Islands (Laird, 2010). For cellular age determination, methylation of cytosines at CpG dinucleotides are of particular interest and have been demonstrated to effectively estimate age state as described by Hannum and Horvath biological clocks. By looking at a few hundred base pairs we can estimate cell age with incredible accuracy and by changing the epigenome of old cells, we may fundamentally alter and "reverse" aging processes. This has many applications in disease treatment and prevention. Currently, researchers quantify DNA methylation levels in cells by single-cell methylation analyses including genome-wide bisulfite sequencing (scBS-seq) or reduced representation protocols (scRRBS-seq) (Smallwood et al., 2014).

However, these methods are limited by the small amounts of DNA material and lack of CpG coverage. Moreover, there exists a profound shortage of publicly available TAB-seq or oxBS-seq datasets due to significant cost and skilled labor for high-throughput analysis. In specific experimental setups, measurement of methylation levels for all cell/tissue types across every developmental stage and physiological state is unfeasible due to the perturbation of environment by sequencing based assays. This requires an in silico solution for imputation of datapoints. Furthermore, DNA sequencing based protocols have amplification and fragment selection setups, which result in biased sampling procedures that predispose certain cytosines in a genome to be either overrepresented or underrepresented (Gu,H., et al. 2011). RRBS-seq, which is one of the most widely used assays for measuring DNA methylation, is especially prone to imbalanced weighting of cytosines as only a very small fraction of cytosines have reliable coverage for querying methylation (Gu, H., et al., 2011). Inherent stochasticity of sampling methods may also result in erroneous measurements of methylation levels.

As such, there is a growing demand for predictive models to impute missing methylation data points and predict the methylation state of DNA sequences. In this paper we develop a latent autoencoder to learn sequence representations as well as four supervised learning models combining LSTMs and convolutional layers. This aids in predicting missing methylation states for genome wide association studies (GWAS) and circumvents current bottlenecks in methylation predictions.

Deep learning has emerged as an effective modeling technique for biomedical applications with artificial neural networks. The increasing accessibility of large datasets, availability of graphics processing units (GPUs), and recent rapid progression in model architectures have made deep learning approaches a prime candidate for applications in genomic data (Levy et al., 2020). Thus, by leveraging the unsupervised nature of deep learning models, we aim to identify meaningful relationships between CpG region methylation and age index predictions. Our project is bi-faceted: 1) Prediction of methylation state by DNA sequence for imputation of missing methylation state data 2) Prediction of age index based on complete genome methylation profiles using machine learning.

3 Related Works

Early models for DNA methylation prediction employed Support Vector Machines (SVMs) and decision trees; it used structure and sequence derived features (Bhasin et al., 2005) for classification. A study by Bhasin et al. (2005) utilized an SVM-based model with a feed-forward back-propagation network with a single hidden layer for DNA methylation prediction. The model achieved an AUC value of 0.82 and ACC and MCC of 75% and 0.504, respectively. Das et al. (2006) devised a two-class radial basis kernel SVM-based model with recursive feature elimination (RFE) for feature selection with 84% accuracy.

A study by Pavlovic et al. in 2017 devised a supervised integrative learning framework to perform whole-genome methylation and hydroxymethylation predictions in CpG dinucleotides, along with imputation of missing data in existing datasets. The motivation for this model was to circumvent the current bottleneck of lack of publicly available TAB-seq or oxBS-seq datasets due to significant expenditure and skilled labor as well as the invasive and destructive nature of sequencing based assays, which may be unfeasible in certain experimental setups. Thus, the study developed the Discriminative IntegRative whole Epigenome Classification at single nucleotide resoluTION (DIRECTION) model, which can be trained on shotgun sequencing-based methylation datasets and predicts methylation levels based on genomic sequence-based traits as predictor variables. DIRECTION is a scalable ensemble-learning framework based on a decision tree architecture with a biologically driven topology

- it utilized separate Support Vector Machine and Radial Basis Function Random Forest based predictive model, chosen based on dataset as SVMs are effective with small datasets and RFs are resistant to outliers. DIRECTion predicted methylation status with 82% accuracy across the whole genome. However, these methods are limited by their requirement of annotated datasets and explicit labeling of methylation status for training. Therefore, later models were created to overcome the necessity of labor-intensive and costly labeling tasks and to devise a framework that need only be semi-supervised or unsupervised.

DeepCpG (Angermueller et al., 2017) has been developed to accurately predict methylation states in single cells using deep neural networks. The model architecture is broken down into three separate modules, the DNA module, the CpG module, and the joint module. The DNA module uses two convolutional and pooling layers to identify motifs in the input sequences (sequences of 1000 bp centered around the CpG site) and a fully connected layer to capture interactions between the motifs. The CpG module uses a bidirectional gated recurrent network (GRU) to obtain features from the CpG areas of cells. The joint module combines the previous two modules to learn interactions between higher level features and make cellular level predictions. The DNA module of this study is especially relevant to our study as it detects informative sequence patterns for prediction of DNA methylation. The module scans sequence motifs from large DNA sequence windows using stacked convolution pooling and filtering layers to learn high-level interactions between sequence motifs.

A standard that many age regression models compare against is the accuracy and performance of the biological clock defined by Hannum and Horvath (Horvath, 2013), which utilizes elastic net penalized regression to identify sets of CpGs strongly correlated with aging to provide accurate age estimation. The epigenetic clocks identified "sparse but accurate estimators, with utility in predicting phenotypic outcomes" (Bell et al., 2019). However, there is a lack of understanding of the number of and manner with which features confer aging. Similarly, the cause of residual between chronological age and methylation age, which are tightly associated with disease risk and mortality, is currently unclear.

Another crucial deep learning model for DNA methylation prediction tasks is the MethylNet model, which can be used to perform classification, regression, and multi-output regression tasks. The primary contribution of MethylNet was the application of a variational autoencoder (VAE) to identify latent representations of DNA sequences and apply a downstream predictive architecture to identify meaningful relationships between the latent space and associated phenotypic factors like smoking status, cancer subtype classification, and age regression. The use of variational autoencoders (VAEs) "embed the methylation profiles in a way that represents the original data with high fidelity while revealing nuances" (Way & Greene, 2017). Many autoencoder approaches represent the data using an encoder but utilize a non-neural network such as a SVM to finalize the predictions. In contrast, MethylNet executed “end-to-end training approach that both extracts biologically meaningful features through latent encoding and performs predictions using the derived features” (Levy et al., 2020). The authors found excellent concordance with the predicted age and traditional age estimation measures like Hannum and Horvath clocks.

Recently, Galkin et al. 2021 engineered the DeepMAge model, which frames the age prediction task as a regression problem using 353 regression coefficients and DNAm data. The model input has high dimensionality (originally 24,538 features) so feature selection was applied to reduce the number of features before training the model. In addition, the age labels were transformed using a log transform based on the Hannum and Horvath clocks and then the output from the model is reverse transformed. The model itself uses a feed forward neural network with more than three hidden layers to enable high dependencies in the data fitting.

4 Data

For our first aim of methylation status prediction, the data we will use comes from a Kaggle dataset of 50,000 CpG sites with stable methylation states. The data is split into 30,000 training samples and 20,000 testing samples. The data is stored as a dictionary - for each sample, we have the chromosome number, base location on the chromosome, CpG island information, CpG position relative to genes, CpG position relative to CpG islands, CpG position relative to regulatory elements, 120 bases of the surrounding forward DNA sequence, 2000 bases of the other surrounding DNA sequence, and the beta value indicating 0 for unmethylated and 1 for methylated. The 20,000 test datapoints are unlabeled and cannot be used for supervised learning. We will instead use them to train a generative convolution autoencoder to learn latent representations for genomic sequences. The remaining 30,000 labeled data points are split into training, validation and test sets according to a 70/20/10 percent split respectively.

For the second aim of age index prediction based on methylation states, we will work with tabulated data from a Gene Expression Ombinus (GEO) repository. It contains analysis of 476,366 sites throughout the genome of white blood cells from 431 people ranging in age between 14 and 94 years old. The raw data for each sample is available in the form of an IDAT file, which contains microarray analysis metrics. The age labels for each sample follow a semi-normal distribution with slightly more samples at the lower extreme of ages relative to the upper extreme (Johansson, 2013). The samples were split into training, validation, and test set according to a 70/20/10 split, resulting in a training set of 524 people, validation set of 132 people, and test set of 73 people.

5 Approach

By leveraging a recurrent neural network, we seek to circumvent the bottlenecks of past models, which led to an inability to account for intra-cellular CpG region differentiation and leverage linkages between neighboring CpG sites and DNA sequence patterns. Our proposed model similarly does not require explicit annotations and labeling and instead predicts DNA methylation with deep neural networks.

For the first binary prediction task of predicting methylation state based on nucleotide sequence, we trained a deep learning model using the 120 base pairs surrounding stable CpG sites. We centered all samples around the CpG site and used the 60 nucleotides before and after this site as the training sample. We then one-hot-encoded the nucleotides such that every example for the model was represented as a 120x4 matrix. A very small amount of the samples (less than 5 $\%$) contained a CpG site of the form CT or CA rather than CG. Since this was not a large enough percentage of the training set to meaningfully inform the model during training, we removed these samples from the dataset. To train our deep learning model, we used a batch size of 512 training samples and 128 validation examples.

For our first approach to the methylation prediction task, we utilized a convolutional autoencoder to first learn the latent representations of unlabeled 120-bp sequences as a means of introducing a model that could make use of unlabeled datasets, which are vastly available and more common than annotated datasets. We hoped to leverage the fact that unlabeled genomic sequences are more readily available than their counterparts in order to effectively and feasibly make future predictions on methylation state. The autoencoder learns a compressed representation of the DNA sequences, resulting in a bottleneck layer of the most important features used to reconstruct the original input. By compressing the input data, the model captures intrinsic relationships between data variables and allows for accurate downstream analyses (Pratella et al., 2021). The convolution autoencoder consists of two partitions, the encoder, which learns a latent representation for the genomic sequences, and the decoder, which reconstructs the genomic sequence from the latent representation. The encoder has an architecture of a series of 1D Convolutional layers followed by 1D Max Pooling layers and a final Flatten and Dense layer which represent the latent space. The decoder has an architecture of a series of 1D Convolutional layers followed by 1D Upsampling layers to generate the original genomic sequence, which mirror that of the encoder structure, thus resulting in a "U-net-like" architecture. We first trained the autoencoder on the $n = 20,000$ unlabeled genomic sequences. Upon tuning hyperparameters, kernel size, and achieving effective generative efficacy of the genomic sequences, we partitioned the encoder and formulated a supervised model that consisted of the trained encoder with model weights as well as one final dense layer with a sigmoid activation to make a final prediction on methylation state. Figure 1 summarizes the architecture of the autoencoder and deep learning model.

After testing the efficacy of the autoencoder-based supervised model, we then formulated more supervised models for the methylation prediction task. Since the task predicts on DNA sequences, we devised several models with different combinations of LSTMs and 1D convolutions. The first model consists of two stacked 1D Convolution Layers with ReLu activation. The first layer has 32 filters of size 5 and the second layer has 32 filters of size 3. Each convolutional layer is followed by a max pooling layer of size 4 and 20$\%$ dropout. This is followed by a bidirectional LSTM with 256 hidden

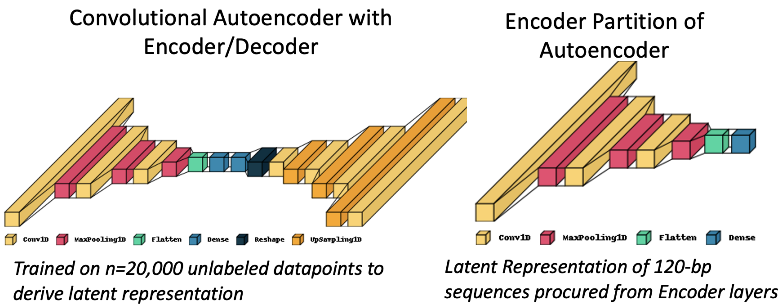

Convolutional Autoencoder with Encoder/Decoder
Encoder Partition of Autoencoder
Conv1D
MaxPooling1D
Flatten
Dense
Reshape
UpSampling1D
Trained on n=20,000 unlabeled datapoints to derive latent representation
Conv1D
MaxPooling1D
Flatten
Dense
Latent Representation of 120-bp sequences procured from Encoder layers

Figure 1: Unsupervised Latent Autoencoder and deep learning model for methylation prediction.

units. Finally, there is a fully connected layer of 32 units with ReLu activation and one Dense layer for the final sigmoid activation. The second architecture consists of 2 stacked convolutional layers as described above, followed by a 100 neuron fully connected layer and a final one unit sigmoid activation. The third model consists of 3 stacked bidirectional LSTMs with hidden units increasing from 64 to 128 to 256, followed by a 100 neuron fully connected layer and a final one unit sigmoid activation. The fourth model consists of one bidirectional LSTM with 256 hidden units followed by 2 convolutional layers as described earlier. The model also ends with a 32 neuron fully connected layer and a final one neuron sigmoid activation layer. Figure 2 summarizes the architecture of the four supervised learning architectures.

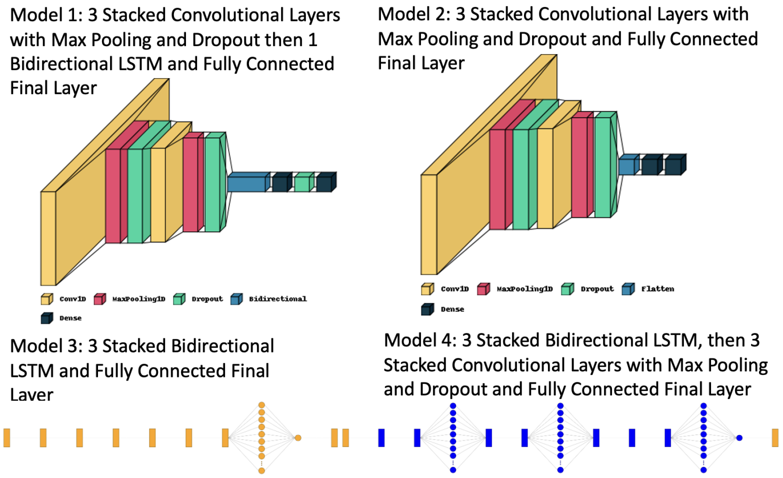

Model 1: 3 Stacked Convolutional Layers with Max Pooling and Dropout then 1 Bidirectional LSTM and Fully Connected Final Layer
Model 2: 3 Stacked Convolutional Layers with Max Pooling and Dropout and Fully Connected Final Layer
Conv1D
MaxPooling1D
Dropout
Bidirectional
Dense
Conv1D
MaxPooling1D
Dropout
Flatten
Dense
Model 3: 3 Stacked Bidirectional LSTM and Fully Connected Final Laver
Model 4: 3 Stacked Bidirectional LSTM, then 3 Stacked Convolutional Layers with Max Pooling and Dropout and Fully Connected Final Layer

Figure 2: 4 Deep Learning Model Architectures for supervised methylation prediction.

For the age prediction task, we compared a deep learning architecture consisting of 5 fully connected layers with ELU activation with a Random Forest regressor using 100 decision trees. The Random Forest regressor expands all nodes until the leaves are pure or the leaves contain fewer than two samples. Figure 3 summarizes the architecture for the age prediction models.

6 Experiments

Our first approach utilized a convolutional autoencoder to first learn the latent representation of unlabeled 120-bp sequences ($n = 20,000$) followed by a supervised methylation state prediction model consisting of the trained encoder with a sigmoid activation layer for state prediction. Following fine tuning of our convolutional layer kernel and window size, we trained the generative autoencoder to reveal latent representations of the sequences (Fig. 4A). Following pre-training of the autoencoder, the learned encoder partition was combined with a Dense and sigmoid activation layer for methylation

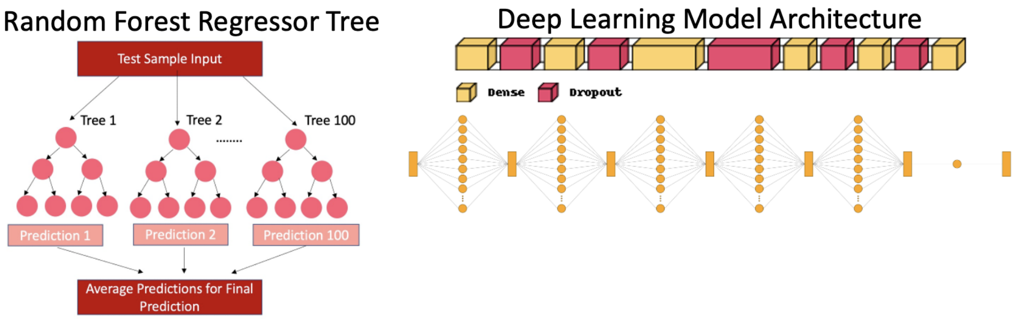
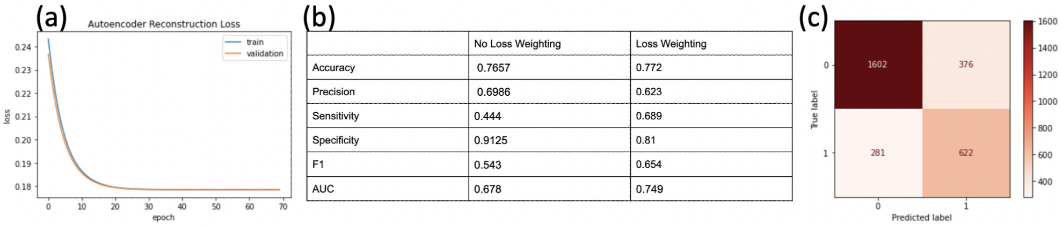

Random Forest Regressor Tree
Test Sample Input
Tree 1
Tree 2
Tree 100
....
Prediction 1
Prediction 2
Prediction 100
Average Predictions for Final Prediction
Deep Learning Model Architecture
Dense
Dropout

Figure 3: Age Regression Deep Learning Model and Random Forest Regressor.

state prediction. We noted an accuracy of 77% and 78% on predictions for our no loss weighted and loss weighted test set, respectively (Fig. 4B,C).



	No Loss Weighting	Loss Weighting
Accuracy	0.7657	0.772
Precision	0.6986	0.623
Sensitivity	0.444	0.689
Specificity	0.9125	0.81
F1	0.543	0.654
AUC	0.678	0.749

Figure 4: Loss curve for generative autoencoder and metrics for supervised methylation prediction encoder model

For our second approach, we fine tuned hyper parameters for the LSTMs and convolutional layers of the supervised learning models. Specifically, the kernel size and number of filters were tuned for convolutional layers and the number of hidden units were tuned for the LSTM layers. Initial experiments with the architectures lead to inefficient learning where learning plateaued for the first 20-30 epochs and then increased in a non smooth layer. After tuning the kernel size and number of convolutional filters, we achieved improved learning curve shapes and increased accuracy. Figure 5 shows the learning curves for the stacked convolutional model and bidirectional model followed by stacked convolutional layers.

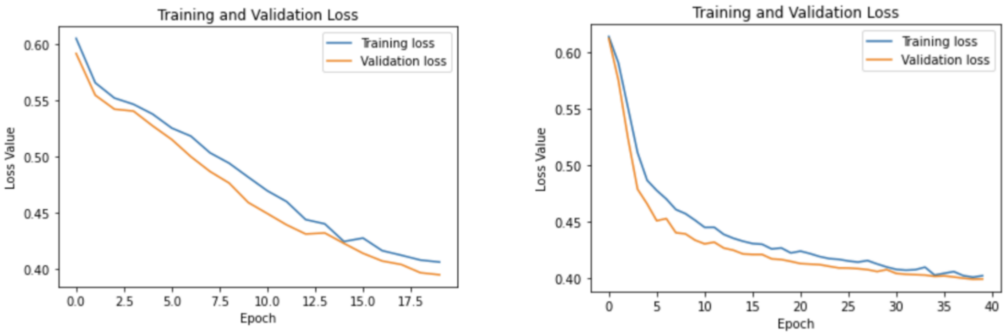
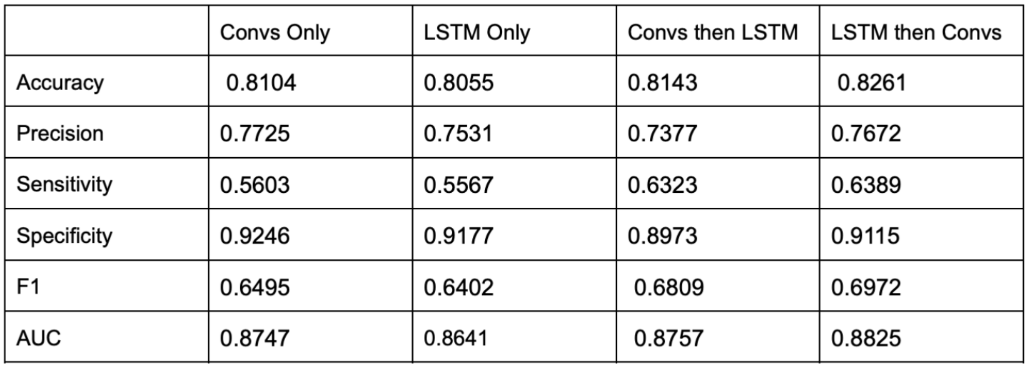


Figure 5: Training and Validation loss for deep learning models.

Initial LSTM models consisting of 3 stacked Bidirectional LSTMs with 256 hidden units were unable to learn and simply predicted the negative class output for every sample. Further fine tuning of the number of hidden states allowed the model to learn effectively and achieve accuracy comparable with the other architectures.

For the model consisting of bidirectional LSTMs followed by stacked convolutions, we experimented with having multiple stacked LSTMs. This did not improve model accuracy while drastically increasing the training time due to the many parameters that must be trained for a bidirectional LSTM. As a result, our final architecture uses just one bidirectional LSTM to extract biologically relevant information from the sequence before the convolutional layers.

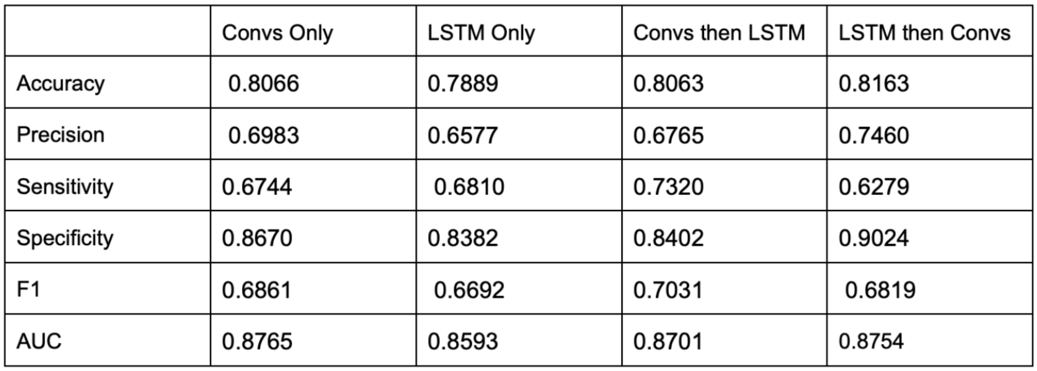
Overall the four architectures all performed similarly on the validation set with the model consisting of a bidirectional LSTM followed by stacked convolutions performing slightly better on the validation set than the others. Figure 6 shows the results after evaluating each model on the test set.



	Convs Only	LSTM Only	Convs then LSTM	LSTM then Convs
Accuracy	0.8104	0.8055	0.8143	0.8261
Precision	0.7725	0.7531	0.7377	0.7672
Sensitivity	0.5603	0.5567	0.6323	0.6389
Specificity	0.9246	0.9177	0.8973	0.9115
F1	0.6495	0.6402	0.6809	0.6972
AUC	0.8747	0.8641	0.8757	0.8825

Figure 6: Test set evaluation metrics for four methylation prediction architectures.

From these results we see that all architectures disproportionately have better specificity than sensitivity. Thus, these models are quite effective at predicting sequences that are unmethylated but poor at predicting methylated sequences. We hypothesized that this was due to class imbalance in the training set where only 30 % of the data were associated with the positive class and the remaining 70 % belong to the negative class. To address this, we re-trained each model using class weighting with a ratio of 1:1.5 for negative to positive (approximately the ratio between the classes reversed) and achieved comparable accuracy with improvements in the positive predictive power. The aforementioned last two models could maintain comparable accuracy with a 1:2 ratio and improved in positive predictive power. The results on the weighted models are summarized in Figure 7. Note that the sensitivity of the convolutions only model, for example, increases from 0.5603 with no class weighting to 0.6744 with class weighting which marks a drastic improvement.



	Convs Only	LSTM Only	Convs then LSTM	LSTM then Convs
Accuracy	0.8066	0.7889	0.8063	0.8163
Precision	0.6983	0.6577	0.6765	0.7460
Sensitivity	0.6744	0.6810	0.7320	0.6279
Specificity	0.8670	0.8382	0.8402	0.9024
F1	0.6861	0.6692	0.7031	0.6819
AUC	0.8765	0.8593	0.8701	0.8754

Figure 7: Training and Validation loss for deep learning models with class weighting to address class imbalance.

For the second prediction task of age prediction based on complete genome methylation profile, we used the MethylPrep package to download and process all the raw IDAT files and extract the beta values for each location in the genome. In total, after removing NA values, we are left with 323,466 locations that represent features for the age prediction. We then used PCA to reduce the dimensionality of the data to allow training of a deep learning neural network. With 441 principal components we were able to preserve 95 % of the variation amongst the methylation profiles for different ages. We ran a Standard Scalar normalization on the methylation data before and after the PCA and we used max scaling on the age labels to reduce the range to a smaller scale.

Initial performance of the deep learning model did not show markedly significant correlation. We experimented with three different activation functions: ReLu, Elu, and SeLu. The choice of activation function did not change the model correlation, explain variance between models, or error, so we used ELU in our final model as the previous DeepMAge had employed the same function. To optimize our model, we employed a grid search over the number of neurons in each network for a five layer fully connected network. The optimal neuron numbers were 512, 512, 64, 128, 256 for layers one through five and the model ends with a single neuron layer with Elu activation for the age regression prediction. Even after optimizing the model with Grid-Search, we noted suboptimal correlation. Thereafter, we trained a Random Forest Regressor on the data with 100 estimators to compare the two models and had noted improved r squared of 0.6039 on validation set compared to 0.3644 for the deep learning model. We then visualized which features of the Random Forest were the most important in informing the model predictions. Since the RF model is trained on PCA data these features correspond to the principal components of the original data that are highly indicative of age. These features are shown in Figure 8.

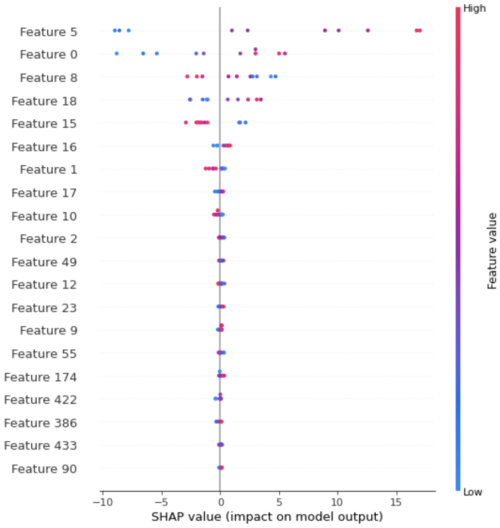
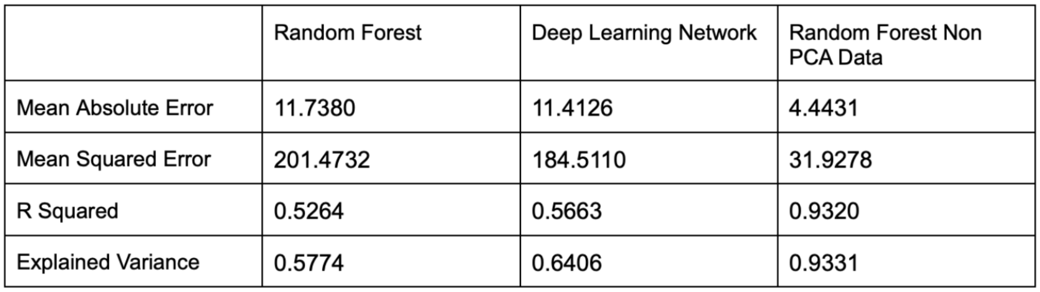


Figure 8: Shapley values for the random forest regressor.

Given the markedly improved performance of the Random Forest Regressor on the dimensionality reduced data, we then tested the model's efficacy on the original non-PCA pre-processed data as well. We noted an extremely strong correlation of 0.9180 on validation and 0.9500 on the test set. Figure 9 summarizes the regression metrics for the three models and figure 10 shows scatterplots of the predictions vs actual ages across the three models. Future work could be done to investigate the informative features for this model, which may correlate directly to CpG regions in the genome. However, we were unable to complete this feature extraction in the scope of this project as our previous method of visualizing the most important features could not run in reasonable time (due to the 300k+ informative features).

Pavolic et al. previously trained a random forest regressor for age prediction based on methylation using a different dataset and achieved an r squared of 0.82, which our model outperforms. Our random forest regressor performs with comparable metrics to deep learning models for age prediction based on methylation profile. The DeepMAge model can estimate human age with a MedAE of 2.77 years, which aligns closely with our model's mean absolute error of 3.8 years.

The MethylNet model is trained on the same dataset as our age prediction model and achieved an R^2 of 0.96 and mean absolute error 3.0. Our random forest model achieves a comparable R^2 value of 0.92 and mean absolute error 4.6, indicating comparable model performance to MethylNet.



	Random Forest	Deep Learning Network	Random Forest Non PCA Data
Mean Absolute Error	11.7380	11.4126	4.4431
Mean Squared Error	201.4732	184.5110	31.9278
R Squared	0.5264	0.5663	0.9320
Explained Variance	0.5774	0.6406	0.9331

Figure 9: Regression metrics for deep learning model compared to machine learning Random Forest Regressor shows the Random Forest Regressor greatly out performing deep learning.

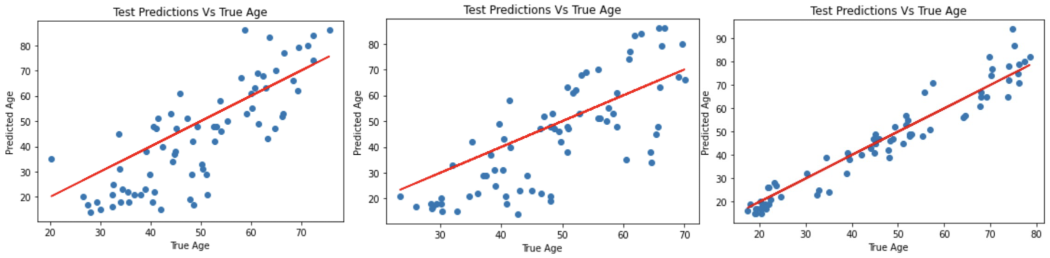


Figure 10: Figure 5: Predicted vs actual ages for deep learning model on the left, random forest on pca data in the center, and random forest on the full dimensionality data on the right.

7 Conclusion

Overall, our methylation status prediction models performed excellently, all achieving accuracy above 0.8. Likely due to class imbalance in the training data, our model had higher specificity than sensitivity. We mitigated this with a weighted loss function but future work would focus more on this issue with model finetuning and preprocessing steps. Of note, in the future we plan to use longer sequences (1200 base pairs) as opposed to the shorter 120 base pair sequences we used to train our model. A previous model using a convolutional deep learning architecture on the 1,200-bp sequences was able to achieve 97 percent accuracy of unmethylated sites and 85 percent accuracy of methylated sites.

Our random forest regression model performed with greater efficacy than the deep learning model on the age prediction task. Previous literature has shown that deep learning may be very effective at learning age from methylation profile, but we would likely need to use a more complex form of dimensionality reduction or feature encoding in comparison to PCA. For example, the MethylNet architecture works very well and uses a variational auto encoder to encode the complete methylation profile. This is further supported by the significant performance of a Random Forest Regressor on the complete data as this indicates there is a learnable, meaningful relationship between the methylation profile and age in our data set. Experimenting with future methods for this problem are a direction we are interested in exploring further.

Finally, we would like to create a joint module that combines the methylation prediction deep learning model with the age prediction model. As shown in DeepCpG, a joint module can be formulated by creating a DNA methylation prediction module that informs an age prediction module. Furthermore, we would like to expand on our model's ability to predict methylation state and broaden the scope of our project to predict on smoking status, cancer subtype propogation, and disease progression, as was done in previous joint models.

8 Author Contributions

Mahdi Moqri helped us with coming up with the project idea and pointed us towards the datasets that we used. Sophie and Marc worked together on the introduction and writing code to preprocess the DNA sequences and one hot encode them.

Sophie worked on writing the previous deep learning models section of the related works, the section about the datasets, and the conclusion and future directions. She also wrote and trained the 4 supervised learning models for methylation status prediction, fine tuned them, and wrote code to asses performance metrics. She also retrained all of them with class weighting and reassessed performance. Sophie also researched ways to process the IDAT files for the age prediction and downloaded and processed them all. She then worked on training the deep learning model for age prediction including grid search for hyper parameters, trained the random forest, and did all the data preprocessing and dimensionality reducuction using pca for the data. She also wrote code to evaluate metrics and plot the results. She also wrote up all the parts of the paper about these topics.

Marc worked on writing the introduction section specifically about DNA methylation and current bottlenecks in DNAm prediction, the related works section about early models for DNA methylation prediction, the autoencoder-related section of the Approach column, and the experimental results section for the autoencoder generative and autoencoder supervised training. He also wrote and trained the generative autoencoder for DNA sequence reconstruction and the supervised model with the encoder for methylation state prediction. Marc also worked on generating SHAPley feature values for model assessment as well as creating the model architecture figures for visual assessment.

9 References

Angermueller, C., Lee, H. J., Reik, W., amp; Stegle, O. (2017). DeepCpG: Accurate prediction of single-cell DNA methylation states using Deep Learning. Genome Biology, 18(1). https://doi.org/10.1186/s13059-017-1189-z

Bell, C. G., Lowe, R., Adams, P. D., Baccarelli, A. A., Beck, S., Bell, J. T., Christensen, B. C., Gladyshev, V. N., Heijmans, B. T., Horvath, S., Ideker, T., Issa, J.-P. J., Kelsey, K. T., Marioni, R. E., Reik, W., Relton, C. L., Schalkwyk, L. C., Teschendorff, A. E., Wagner, W., … Rakyan, V. K. (2019). DNA methylation aging clocks: Challenges and recommendations. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1824-y

Bhasin, M., Zhang, H., Reinherz, E. L., amp; Reche, P. A. (2005). Prediction of methylated cpgs in DNA sequences using a support vector machine. FEBS Letters, 579(20), 4302–4308. https://doi.org/10.1016/j.febslet.2005.07.002

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H., amp; Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences. Proceedings of the National Academy of Sciences, 103(28), 10713–10716. https://doi.org/10.1073/pnas.0602949103

Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A., amp; Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nature Protocols, 6(4), 468–481. https://doi.org/10.1038/nprot.2010.190

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biology, 14(10). https://doi.org/10.1186/gb-2013-14-10-r115

Johansson, Å., Enroth, S., amp; Gyllensten, U. (2013). Continuous aging of the human DNA methylome throughout the human lifespan. PLoS ONE, 8(6). https://doi.org/10.1371/journal.pone.0067378

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. Nature Reviews Genetics, 13(7), 484–492. https://doi.org/10.1038/nrg3230

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics, 11(3), 191–203. https://doi.org/10.1038/nrg2732

Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., amp; Christensen, B. C. (2019). MethylNet: An automated and modular deep learning approach for DNA methylation analysis. https://doi.org/10.1101/692665

Pavlovic, M., Ray, P., Pavlovic, K., Kotamarti, A., Chen, M., amp; Zhang, M. Q. (2017). Direction: A machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. Bioinformatics, 33(19), 2986–2994. https://doi.org/10.1093/bioinformatics/btx316

Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T. P., Beyan, H., Whittaker, P., McCann, O. T., Finer, S., Valdes, A. M., Leslie, R. D., Deloukas, P., amp; Spector, T. D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Research, 20(4), 434–439. https://doi.org/10.1101/gr.103101.109

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., amp; Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods, 11(8), 817–820. https://doi.org/10.1038/nmeth.3035

Way, G. P., amp; Greene, C. S. (2017). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. https://doi.org/10.1101/174474