scientific reports

Check for updates

OPEN

Exploring the correlation between DNA methylation and biological age using an interpretable machine learning framework

Sheng Zhou[1], Jing Chen[2], Shanshan Wei[1], Chengxing Zhou[3], Die Wang[4], Xiaofan Yan[5]✉, Xun He[5]✉ & Pengcheng Yan[6]✉

DNA methylation plays a significant role in regulating transcription and exhibits a systematic change with age. These changes can be used to predict an individual's age. First, to identify methylation sites associated with biological age; second, to construct a biological age prediction model and preliminarily explore the biological significance of methylation-associated genes using machine learning. A biological age prediction model was constructed using human methylation data through data preprocessing, feature selection procedures, statistical analysis, and machine learning techniques. Subsequently, 15 methylation data sets were subjected to in-depth analysis using SHAP, GO enrichment, and KEGG analysis. XGBoost, LightGBM, and CatBoost identified 15 groups of methylation sites associated with biological age. The cg23995914 locus was identified as the most significant contributor to predicting biological age by calculating SHAP values. Furthermore, GO enrichment and KEGG analyses were employed to initially explore the methylated loci's biological significance.

Keywords DNA methylation, Biological age, GO enrichment analysis, XGBoost, Interpretable machine learning, Shapley Additive exPlanations

Background of the study

DNA methylation is a widespread epigenetic phenomenon, a functional modification of the genomic nucleic acid sequence that can affect gene expression without altering the DNA sequence. It is characterized by adding a methyl group to the DNA molecule, significantly impacting gene expression and cellular function[1]. DNA methylation occurs mainly at the CpG site, between adjacent C and G bases in the deoxyribonucleotide sequence[2]. The process is catalyzed by DNA methyltransferases, which achieve the selective addition of methyl groups to specific bases[3].DNA methylation is crucial in regulating transcriptional programs and shows systematic changes with age[4]. These changes begin at the onset of embryonic development and continue throughout the life cycle, with implications for chromatin conformation, lineage differentiation, gene expression, genome stability, and stem cell self-renewal[5].

Some changes in methylation are strongly associated with age and provide markers for biological aging. The genome continues to undergo programmed changes in methylation after birth in response to environmental inputs, acting as a memory device that may influence aging and susceptibility to various metabolic, autoimmune, and neurological diseases[6]. Related studies have shown that DNA methylation patterns in the genome are disrupted with age and that these changes can be used to predict age through the epigenetic clock[7] statistically. Many sources of evidence suggest that some CpG sites may have age-related methylation changes[8]; exploring the methylation of different CpG sites and their levels is essential to reveal the association between DNA methylation and biological age, and biological age prediction provides a fundamental basis for preventive and healthcare efforts against age-related diseases.

[1]Department of Public Health and Health, Guizhou Medical University, Guizhou Province, China. [2]Guizhou Provincial Drug Administration inspection center, Guiyang, Guizhou Province, China. [3]School of Biology&Engineering(School of Health Medicine Modern Industry), Guizhou Medical University, Guiyang, Guizhou, China. [4]College of Anesthesia, Guizhou Medical University, Guizhou Province, China. [5]School of Medicine and Health Management, Guizhou Medical University, Guizhou Province, China. [6]School of Clinical Medicine, Guizhou Medical University, Guizhou Province, China. ✉email: 137859829@qq.com; 2812878586@qq.com; 1029473656@qq.com

Experimental methods for DNA methylation detection cover restriction endonuclease-based techniques, affinity enrichment-based strategies, and bisulfite conversion-based means. Computational analysis of DNA methylation sequencing data obtained by various experimental methods may be challenging[9], partly because the methylation data obtained by experimental assays are extensive and may contain numerous CpG sites. It isn't easy to compute and analyze these massive and high-dimensional data. On the other hand, machine learning (ML) development provides an effective method for mining and parsing massive data, especially for integrated phenotypes in extensive and high-dimensional data. Machine learning can compute many covariates even in high-dimensional data and complex interactions. Compared to standard statistical methods, machine learning may have advantages regardless of their performance in terms of yield and nature[10,11]. In the field of biological age prediction, machine learning(ML) can integrate diverse data, including images (e.g., brain magnetic resonance imaging, chest radiology, retinal or facial photography)[12–14], physical activity data, as well as blood biomarkers, gut microbiome, or genomic data[15]. Epigenetic clocks, which measure changes at hundreds of specific CpG loci, can accurately predict the number of solid years in various species, including humans, and these clocks are currently the best biomarkers for predicting human mortality[16]. Dmitry Zubakov et al. (2016) explored novel age-associated mRNA and DNA methylation markers in the blood of young and old individuals of the same age using microarray technology. Validation was also carried out in independent samples covering a wide age range by alternative techniques and previously proposed DNA methylation, stress, and telomere length markers. In age prediction, the results showed that DNA methylation markers were more accurate than mRNA, sjTREC, and telomere length[17]. Jiansheng Zhang et al. (2021) constructed two prediction models using health and disease data by analyzing DNA methylation data in blood tissues. The R^2 value obtained from gradient-boosted regression for the health data was 0.86 with a mean squared absolute deviation (MAD) of 3.90, while in the disease dataset, the R^2 value was 0.89 with a MAD of 3.11[18]. David Bernard et al. (2023) created an interpretable ML framework for determining a patient's health and nutritional status using a broad population-based dataset from the National Health and Nutrition Examination Survey (NHANES) study and selecting the XGBoost algorithm as the predictor. XGBoost algorithm as a predictor to create an interpretable ML framework to determine personalized physiological age (PPA) and calculated an accurate quantitative correlation metric explaining physiological (i.e., accelerated or delayed) deviations from age-specific normative data using SHAP for each variable[19]. Through experimental and technological methods, the researchers mentioned above investigated the possibilities of DNA methylation, machine-learning approaches, and interpretable machine-learning frameworks in biological age prediction. The findings indicate that DNA methylation offers a considerable advantage in biological age prediction and has much potential when combined with machine learning methods.

Purpose of the study

This paper uses open-source DNA methylation data to explore the relationship between DNA methylation and biological age. Through data exploration, feature engineering, statistical analysis, machine learning, and the application of an interpretable machine learning framework to a large DNA methylation dataset, we seek to reveal the relationships: the statistical associations between biological age and DNA methylation, the effects of different DNA methylation sites and levels on biological age; the construction of a biological age prediction model through machine learning techniques; and the Assessing the extent to which different gene methylations contribute to biological age prediction using the interpretable machine learning framework SHAP. It provides a reference for biology and medicine to explore the mechanism and effects of methylation at different DNA loci and methylation levels on aging.

Research significance

Abnormal increases or decreases in DNA methylation lead to or are markers of cancer formation and tumor progression and DNA methylation abnormalities have also been associated with neurological disorders, immune disorders, atherosclerosis, and osteoporosis[20]. Exploring the association between DNA methylation and biological age from large DNA methylation datasets through an interpretable machine learning framework provides guidelines for the prevention and healthcare of age-related diseases; moreover, investigating the differences in the contribution of different CpG island methylation degrees and methylation levels to the prediction of biological age through an interpretable machine learning approach can help to improve the interpretability of machine learning models. Finally, analyzing methylation differences at different gene loci can help to reveal the association between cell and tissue aging and methylation from a biological perspective.

Results

DNA methylation data

The DNA methylation data for this study contained a total of 10,296 samples, of which 7,833 were healthy samples, and 2,463 were diseased samples; of these samples, 8,233 samples contained biological age data and were defined as training samples for constructing machine learning models. Each training sample contains 50,000 methylation site data, 1 item of gender data, and 1 item of biological age data, and DNA methylation data is measured by the methylation 450 K platform. Of the 8233 samples used to construct the machine learning model, 6266 were healthy samples, and 1967 were diseased samples; in terms of gender distribution, 4409 (53.55%) samples were male samples, and 3824 (46.45%) samples were female samples. Biological age had a minimum value of 0 years, a maximum value of 114 years, a median value of 56 years, a mean value of 53.6597 years, and a standard deviation of 25.8249 years.

Data preprocessing results

First, the statistics showed that 23,688,484 methylation sites data were missing, accounting for 5.7544% of the total data volume (total data volume was 411658233), and all the vacant data were filled with 0. Secondly, we

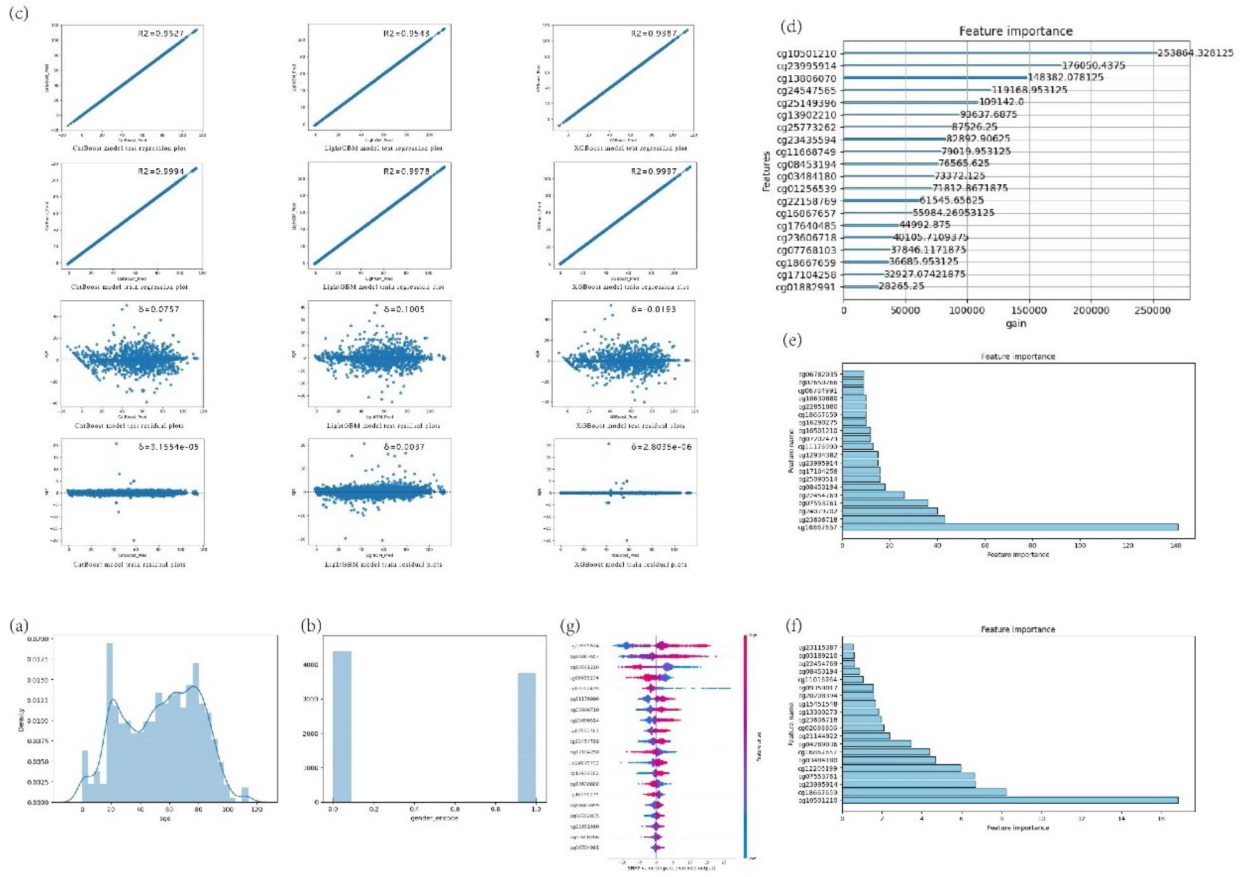

Fig. 1. (**a**) Density estimate of biological age distribution; (**b**) Histogram of gender distribution; (**c**) Regression plots, residual plots for training and testing of feature selection models XGBoost, LightGBM, and CatBoost models; (**d**) Plot of importance of top 20 features of XGBoost model; (**e**) Plot of importance of top 20 features of LightGBM model; (**f**) CatBoost model top 20 feature importance plot; (**g**) Best model (LithtGBM) top 20 feature SHAP values.

Model	Training MAE	Test MAE
XGBoost	0.0539	3.6546
LightGBM	0.6815	3.3041
CatBoost	0.3313	4.0189

Table 1. Model training MAE and test MAE.

performed data coding on the gender data: 4409 male category data were coded as 0, and 3824 female category data were coded as 1. Subsequently, we performed a normality test on the 50,000 DNA methylation site data. The results showed that none of the 50,000 DNA methylation site data fulfilled the normal distribution (all P-values were less than 1%). Finally, we plotted the distribution histograms and kernel density estimates for the sex and biological age data. See Fig. 1a and b.

Feature selection results

The XGBoost, LightGBM, and CatBoost models were used for the wraparound strategy for feature selection, and 50,000 methylation sites and gender data were used as feature data. 80% (6586 samples) of the data were used as training sets, and 20% (1647 samples) were used as test sets. The model training MAE and testing MAE are shown in Table 1.

Table 1 shows that the XGBoost model has the lowest training MAE (0.0539), the LightGBM model has the highest training MAE(0.6815), the LightGBM model has the lowest testing MAE(3.3041), and the CatBoost model has the highest testing MAE(4.0189). When analyzing the test MAE, LightGBM obtained the lowest MAE.

The regression and residual plots of the predicted biological age versus actual age are shown in Fig. 1c. As can be seen in Fig. 1c, in training set, the XGBoost model predicted results with the actual biological age R2 value of 0.9997, which is the best performance, and the LightGBM model predicted results with the actual biological age R2 of 0.9978, which is the worst performance; in the test dataset, the LightGBM model predicted results with the actual results R2 of 0.9543, while the CatBoost model predicted results with actual biological age R2 of 0.9327, which is relatively poor performance. Regarding the prediction results' stability, the XGBoost model has the smallest average residuals of 2.8035e-06 in the training dataset. In contrast, the LightGBM model has the most significant average residuals of 0.0037. In the test dataset, the XGBoost model had the smallest average residual of -0.0193, while the LightGBM model had the most significant average residual of 0.1005.

Considering the model training MAE, testing MAE, training dataset prediction result vs. real biological age R2, testing dataset prediction result vs. real biological age R2, training dataset prediction result vs. actual biological age average residuals, and testing dataset prediction result vs. actual biological age average residuals together, the top 20 methylation sites screened by the LightGBM model were selected as the feature data for constructing the biological age prediction model. Figure 1d and e, and 1f show the top 20 methylated sites output by the XGBoost, LightGBM, and CatBoost models based on feature importance, respectively, and Fig. 1g shows the top 20 methylated sites output by the LightGBM model based on feature importance SHAP.

Statistical analysis results

The 20 methylation data obtained by feature selection were first tested for normality using the Scipy library regular test[21] function, and the test results are shown in Table 1 of the supplementary document. Figure s1 shows the histogram of data distribution and kernel density estimation for the 20 methylation sites. Based on the analysis of normality tests, histograms, and kernel density estimation plots, not all methylation data conformed to a normal distribution. Therefore, the Spearman correlation coefficient was used to assess the correlation between the 20 methylation data and biological age. The results of the correlation analysis are shown in Table 2, and Fig. 2a demonstrates the heat map of the Spearman correlation coefficient.

Table 2; Fig. 2a correlate the 20 methylation sites and biological age. We selected the features with an absolute correlation coefficient value greater than 0.45 as the final features for constructing the machine learning model, and 15 groups of methylation data were obtained. To understand the distribution of individual methylation data as well as to detect further and process abnormal data to improve data quality, we calculated the descriptive statistical features (mean, standard deviation, minimum, first quartile, median, third quartile, and maximum) of the 15 groups of methylation data (Table 2 of the supplementary document) and plotted box plots (Fig. 2b) and scatter plots (Figure s2) of the 15 groups of methylation data.

From Table 2 of the supplementary document, Figure s2, and Fig. 2b, it can be seen that the distribution consistency of the 15 groups of methylation data is relatively poor, and there are outliers and outliers in most methylation data. Further observation of the box plots shows that there are generally outliers in the methylation data of the 15 groups, a result that matches the conclusions of the descriptive statistical analysis and the regression scatterplot; meanwhile, the box plots show that there is a significant difference in the distribution range of the methylation data of the 15 groups. To ensure the data quality and accelerate the convergence of the machine learning model, we first replaced the outliers and anomalies with the median according to the regression scatter

DNA methylation site	Spearman	*P*
cg23995914	0.724931	0.000000e+00
cg11176990	0.697864	0.000000e+00
cg25090514	0.654536	0.000000e+00
cg07553761	0.630504	0.000000e+00
cg18667659	0.566513	0.000000e+00
cg24079702	0.546958	0.000000e+00
cg23606718	0.541610	0.000000e+00
cg02650266	0.517605	0.000000e+00
cg16867657	0.504503	0.000000e+00
cg22454769	0.473058	0.000000e+00
cg06782035	0.409698	0.000000e+00
cg06784991	0.363914	2.940928e-256
cg12934382	0.310387	2.458023e-183
cg17104258	-0.131589	3.999642e-33
cg22851880	-0.205071	6.913939e-79
cg10630880	-0.477016	0.000000e+00
cg07202479	-0.536098	0.000000e+00
cg16290275	-0.556945	0.000000e+00
cg08453194	-0.602760	0.000000e+00
cg10501210	-0.645215	0.000000e+00

Table 2. 20 Spearman correlation coefficients of methylation data with biological age.

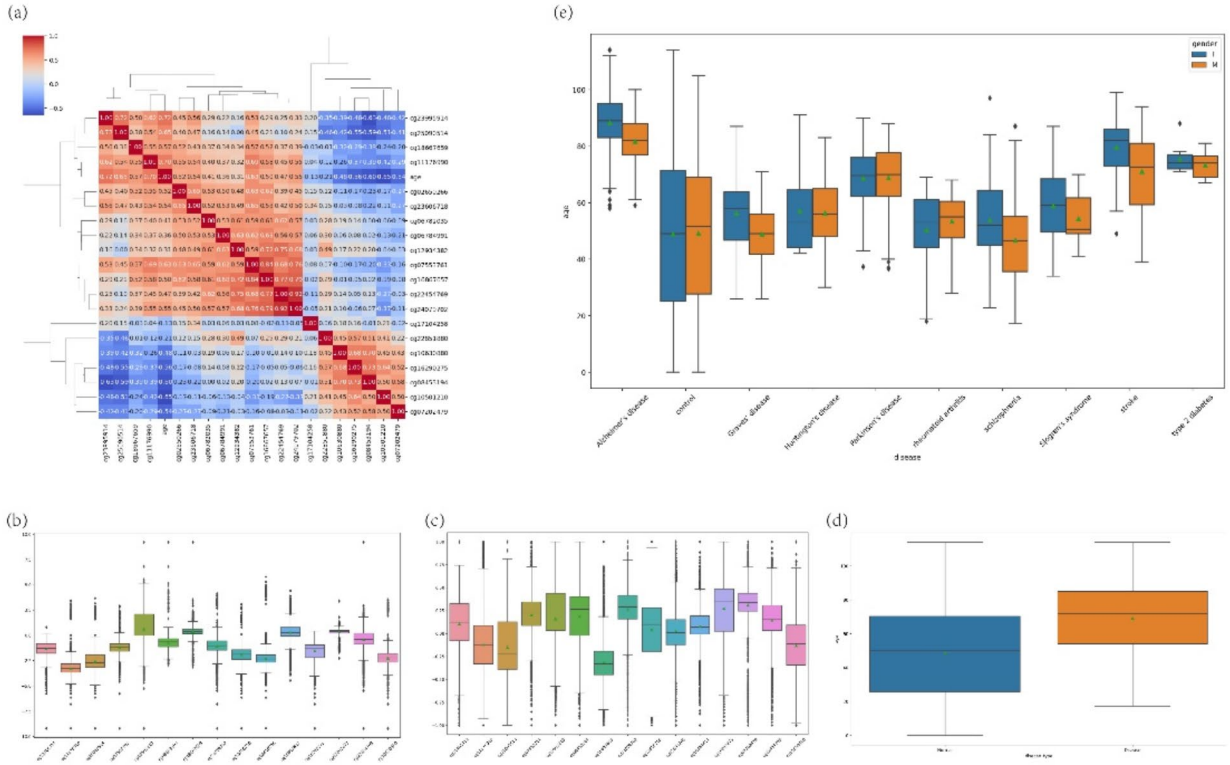

Fig. 2. (**a**) Heat map of Spearman's correlation coefficient; (**b**) Box plot of 15 methylation data; (**c**) Box plot of methylated data after data normalization and scaling; (**d**) Box plots of biological age in healthy and diseased groups; (**e**) Box plots of biological age in the diseased group.

plot (Fig. s1 of the supplementary document) and the box plot (Fig. 2b), and then standardized and normalized the data by using the StandardScaler[22] and MinMaxScaler[23] functions of the Sklearn library to scale the data between −1 and +1. The feature scatter plots and box plots after anomaly data replacement, normalization, and scaling are shown in Figure s3 and Fig. 2c.

The results of the Mann-Whitney U test (statistic value: 3327968.5, P value: 0.00) and the Kolmogorov-Smirnov test (statistic value: 0.325113, P value: 0.00) demonstrated a statistically significant discrepancy in biological age between the healthy and diseased groups. Figure 2b illustrates the distribution of biological age between the healthy and diseased groups. The Kruskal-Wallis H-test (statistic: 1751.851222, p-value: 0.00) demonstrated that there was also a statistically significant difference in biological age between the various disease groups. Figure 2e depicts the distribution of biological age among the diseases.

Machine learning model training results

Following the processing of anomalous data, which included normalization and scaling, we selected XGBoost, LightGBM, CatBoost models, and deep neural networks as machine learning models for training purposes. The models mentioned above were trained using 10-fold cross-validation. XGBoost, LightGBM, and CatBoost models are all models built up based on tree structure, and the feature importance can be output after training to understand the degree of contribution of 15 sets of methylation data to biological age prediction[24–26]; the methylation sites that contribute more to biological age prediction can be used as a reference for the biological and medical neighborhoods to study DNA methylation and cellular and tissue aging; deep neural networks have a solid nonlinear fitting ability, which can fully explore the nonlinear relationship between the data[27]. The fully connected neural network was constructed using TensorFlow, using MAE as the loss function, initialized learning rate of 0.001, early stop and checkpoint techniques were applied to monitor the loss function and MAE changes in real-time during the training process, and the performance scheduling strategy was used to adjust the learning rate automatically. Different training rounds using the above parameters are shown in Figure s4. Testing 2000, 1000, and 800 rounds of training found that the MAE curves all showed a rising trend in the late stage of training, while when 300 rounds of training were chosen, the MAE curves did not rise in the late stage of training. Therefore, the number of training rounds for the deep neural network was finally determined to be 300.

Table 3 shows the results of machine learning model training using 15 sets of methylation data and 1 item of gender data, and the deep neural network training MAE curve is shown in Fig. 3a.

ML model	Train MAE	Validate MAE	Test MAE
XGBoost	0.0189	3.8027	3.6412
LightGBM	2.4831	4.2630	4.1559
CatBoost	0.8253	4.1477	4.1104
deep neural network	7.2615	5.4537	5.4531
XGBoost(GridSearchCV)	0.7689	3.6953	3.6089

Table 3. Machine learning model training results. XGBoost, LightGBM, and CatBoost training MAE and validation MAE were obtained using 10-fold cross-validation training. The average MAE and test MAE were calculated for the test set. The deep neural network validation MAE is obtained by dividing the validation set and setting the TensorFlow model fit method validation_data parameter. The test MAE is calculated using the TensorFlow model to evaluate the method.

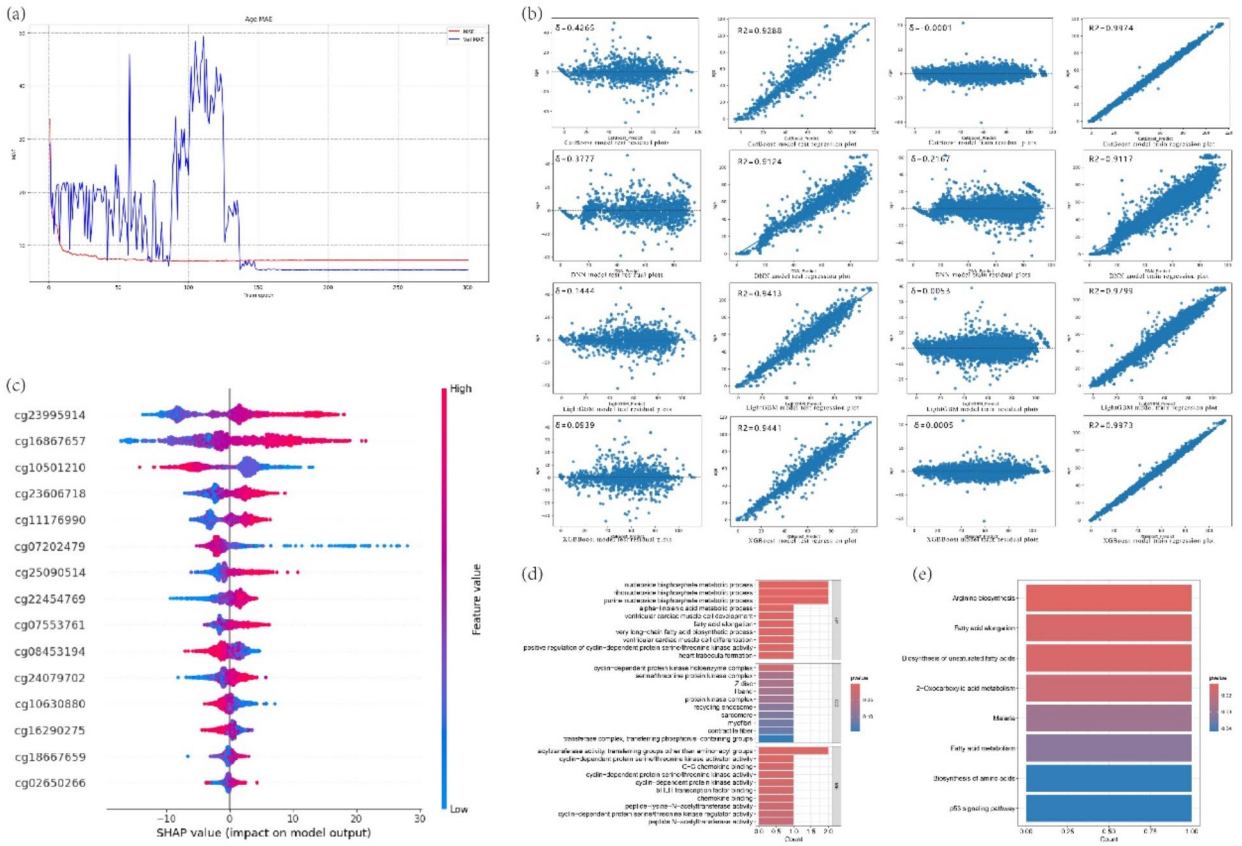

Fig. 3. (**a**) Deep neural network training MAE plot; (**b**) Machine learning model and deep neural network training and testing regression plots, residual plots; (**c**) SHAP plot of methylated features of XGBoost model; (**d**) GO enrichment analysis plot; e: KEGG analysis plot, KEGG data were obtained using the KEGG website enrichment[28].

From Table 3, it can be seen that the XGBoost model has the lowest training MAE in the training dataset (0.0189), and the deep neural network has the highest training MAE (7.2615); in the testing dataset, the XGBoost model has the lowest testing MAE (3.6412), and the deep neural network has the highest testing MAE (5.4531). Therefore, the XGBoost model performs the best biological age prediction performance, and the deep learning model could perform better.

The scatterplots and residual plots of the regression of the model prediction results with the actual biological age are shown in Fig. 3b. As can be seen from Fig. 3b: in the training dataset, the CatBoost model prediction results have a high degree of agreement with biological age (R2:0.9974), the profound neural network prediction results have a significant difference with biological age (R2:0.9117), and the average residuals for the CatBoost model are the lowest (-0.0001), and the average residuals for the deep neural network are the highest (0.2167).

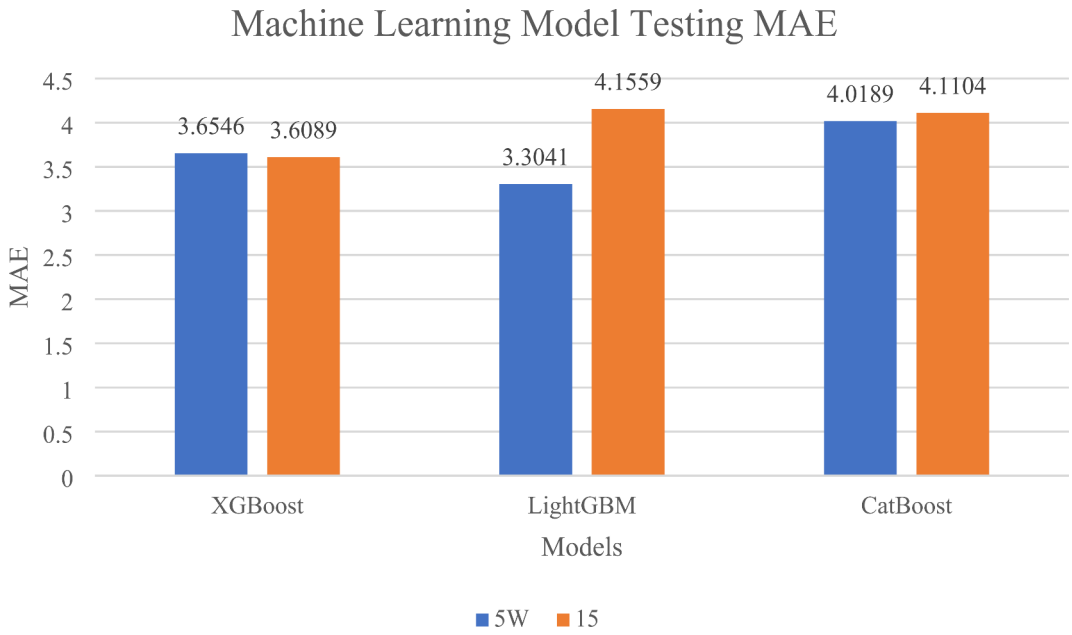

Fig. 4. Model testing MAE for different numbers of methylated sites.

In the test dataset, the XGBoost model predicted relatively good agreement with biological age (R2: 0.9441), the deep neural network predicted poor agreement with biological age (R2:0.9124), the XGBoost model had the lowest average residual (0.0939), and the deep neural network had the highest average residual (0.3777). Considering the R2 values and average residuals of the model predictions with biological age, the XGBoost model showed high accuracy compared to the relatively poor performance of the deep neural network. Ultimately, the XGBoost model was parameterized to identify the optimal hyperparameters, utilizing a grid search with 10-fold cross-validation. After the grid search, the XGBoost model's training MAE was 0.7689, the validation MAE was 3.6953, and the test MAE was 3.6089.

The SHAP value integrates the effect of a given biological variable by itself and the impact of the variable's interaction with other biological parameters. For a given individual (local interpretation), the sum of the SHAP values of all model variables represents the individual's deviation from the mean of the actual age predicted by the entire data set. Finally, we used SHAP to calculate the SHAP values for the 15 methylation data sets in the XGBoost model. We plotted SHAP summary and bar stacked plots to visualize how much the 15 methylation data sets contributed to the biological age prediction, as in Fig. 3c. To ascertain the biological significance of the genes associated with the 15 groups of methylation sites, we initially queried the associated genes by methylation site CG number in the EWAS database (Table 3 of the Supplementary file). Following the database query, the final 12 groups of genes were subjected to a GO enrichment analysis (Fig. 3d, Figure s5, Figure s6) and a KEGG analysis (Fig. 3e, Figure s7) to gain insight into the cellular components, biological processes, molecular functions and pathways of the related genes.

KEGG and GO analyses were implemented using clusterProfiler v4.10.1 in an R4.3.3 environment, and the following 12 proteins were ultimately used for KEGG and GO analyses: ARHGEF33、
ABHD14A-ACY1、CCND3、AMER3、PCBP4、ELOVL2、ZNF518B、FHL2、ABHD14B
、TRIM59、ABHD14A、ACKR1. A KEGG analysis revealed that the genes associated with 15 groups of methylation sites were primarily involved in arginine biosynthesis, fatty acid elongation, unsaturated fatty acid biosynthesis, 2-Oxocarboxylic acid metabolism, malaria, fatty acid metabolism, amino acid biosynthesis, and the p53 signaling pathway, alpha – linolenic acid metabolic process. A Gene Ontology (GO) enrichment analysis revealed that two of the 15 related genes were involved in nucleoside diphosphate metabolism, ribonucleoside bisphosphate metabolism, purine nucleoside diphosphate metabolism, alpha-linolenic acid metabolism, and ventricular cardiomyocyte differentiation. The genes were found to be involved in several processes, including cardiomyocyte development, fatty acid elongation, very long-chain fatty acid biosynthesis, ventricular cardiomyocyte differentiation, positive regulation of the activity of the cell-cycle-dependent protein serine/threonine kinases, and cardiac trabeculae formation. Each of the genes was found to be involved in one of these processes.

Discussion

Screening sites critical for cell and tissue aging from 50,000 methylation sites using traditional statistical and medical experimental methods is challenging and a huge workload. However, feature selection with the help of machine learning methods is expected to reduce the workload and fulfill the potential of screening key methylation sites. Follow-up studies will explore the effectiveness of methylation sites screened by feature selection.

Figure 4 illustrates the mean absolute error (MAE) of the test set following the training of the machine learning model with 50,000 sets of methylated data during the feature selection phase and subsequent training of the model with 15 sets of methylated data following feature selection. As illustrated in Fig. 3, the MAE of the test set for the XGBoost model demonstrates a reduction following the reduction of the methylated data from 50,000 to 15 sets (0.03% of the total number of features). The LightGBM model test set MAE increased by 0.0457, while the CatBoost model test set MAE increased by 0.0915. The MAE of the LightGBM model test set increased the most following the feature selection, which reduced the number of features by 99.97%. In contrast, the MAE of the XGBoost model trained with 15 sets of methylation sites is lower than that of the model trained with 5 W sets, and the potential explanations for this phenomenon are twofold. Firstly, the wraparound feature selection and the initial correlation filtering facilitate the retention of practical features, removing interfering features and reducing data dimensionality. Secondly, the detection and replacement of abnormal data, coupled with data normalization and scaling, enhance the data quality, thereby facilitating the learning of correlations between data points. Finally, a 10-fold cross-validated grid search for different hyperparameter combinations helped the XGBoost model to find better hyperparameters. It can be observed that the application of feature selection is a practical approach for the identification of methylation sites and is capable of accurately discerning those that are predictive of biological age.

The results obtained by our machine learning approach are compared with the results of the most recent studies as follows: the test MAE obtained using the XGBoost model for Personalized Physiological Age (PPA) in the results of Emmanuel Doumard et al. was 7.89 with an R2 of 0.75, the test MAE obtained by MLP was 7.23 with an R2 of 0.75[29], and the test set MAE obtained by us using the 15 sets of methylated loci to train the XGBoost model obtained a test set MAE of 0.3089 with an R2 of 0.9441. the remaining machine learning studies using data similar to ours for training are relatively few. In a 2018 study, Morgan E. Levine and colleagues developed phenotypic age estimates using NHANES III data combined with a proportional risk penalty regression model using nine biomarkers and actual age. They subsequently validated this in NHANES IV. Subsequently, they developed DNAm PhenoAge estimates using InCHIANTI data to regress phenotypic age on blood DNA methylation data[30]. This resulted in the development of the DNAm PhenoAge, which significantly outperforms previous metrics in predicting outcomes of interest, including all-cause mortality, cancer, healthy lifespan, physical function, and Alzheimer's disease. Ake T. Lu et al. employed training and test data from the Framingham Heart Study to define and validate an alternative DNAm-based smoking packet model. They also validated proxy DNAm-based biomarkers of smoking pack-years and plasma protein levels. They constructed a DNAm GrimAge using elastic net regression modeling to automatically select covariates from the actual age, sex, and DNAm-based biomarkers of smoking pack years, and 12 plasma protein levels were evaluated. The results demonstrated significant improvements in the associations with age-related diseases, clinical biomarkers, and PhenoAge, which introduces biomarkers to develop estimates of 'phenotypic age' and effectively predicts aging outcomes[31]. In contrast, GrimAge introduces smoking pack-years and other biological data to construct epigenetic age and performs well in predicting age-related diseases. Performance. It is, therefore, evident that biomarkers and behavioral lifestyles should be considered when constructing epigenetic-based biological ages. In the future, we intend to collect and incorporate biomarker and behavioral lifestyle data into our study to better understand the association between biological age and DNA methylation.

Table 3 of the supplementary document shows the results obtained by consulting the Probes & Genes[32] and STRING[33] databases. The genes and proteins corresponding to the 15 sets of methylation sites were obtained, and the methylation site order in Table 3 of the supplementary document used the SHAP data in Fig. 3c. A comparison of the SHAP value with Spearman's correlation coefficient revealed that both identified cg23995914 as the locus with the highest contribution to the prediction, with the corresponding gene being ZNF518B. The Genecard[34] database indicates that ZNF518B is a protein-coding gene. The gene ontology (GO) annotations related to this gene include DNA-binding transcription factor activity and RNA polymerase II specificity. ZNF518B was identified as a protein-coding gene, and the gene ontology (GO) annotations related to this gene include DNA-binding transcription factor activity and RNA polymerase II specificity. Furthermore, a discrepancy was observed between the SHAP and Spearman correlation coefficients for the cg25090514 locus, which is positioned seventh in the SHAP calculation, and the Spearman correlation coefficient. The value is 0.654536, which is located in the third position. However, the related gene is not identified in the Probes & Genes or STRING databases. Additionally, no associated genes or proteins were identified in these databases. A query of the relevant databases revealed that some methylation sites are related to multiple genes (cg16867657, cg11176990, cg This demonstrates the intricate nature of the methylation process and its correlation with cellular aging. The Spearman correlation coefficients of cg25090514 and cg02650266 were more significant than 0.5 (0.654536 for cg25090514 and 0.517605 for cg02650266), which may indicate a potential link between the two methylation sites and cell and tissue aging. This may suggest that the two methylation sites are associated with cellular and tissue aging.

The KEGG and GO enrichment analyses revealed that the methylation-related genes were predominantly associated with fatty acid elongation, long-chain fatty acid biosynthesis, cell cycle protein-dependent protein serine/threonine kinase activity, and α-linolenic acid metabolism. The cell cycle protein-dependent protein serine/threonine kinases were identified as playing a role in the biological processes, molecular functions, and cellular components[35–37]. Cell cycle protein-dependent kinases (CDKs) constitute a group of serine/threonine kinases that are pivotal in regulating cell cycle progression. The activity of these kinases is induced by cell cycle proteins[38]. The evidence is mounting that CDKs and cell cycle proteins play an active role in regulating stem cell transcription, epigenetic mechanisms, metabolic processes, and self-renewal capacity[39]. Alpha-linolenic acid (ALA) is an essential omega-3 fatty acid for human health. Essential fatty acids are thought to profoundly influence various metabolic processes, including regulating energy supply, enzyme activity, and gene expression[40]. ALA has been demonstrated to possess many biological functions, including cardiovascular

protection, neuroprotection, anticancer, anti-osteoporosis, anti-inflammatory, and antioxidant effects[41]. In a study conducted by Wenyuan Huang and colleagues to investigate the inhibitory effects of ALA on the fatty acid synthesis pathway and apoptosis in breast cancer cells, it was found that ALA could inhibit the invasion and metastasis of tumor cells by inhibiting the fatty acid synthase-induced apoptosis[42] The results of the KEGG and GO enrichment analyses indicated that some of the 15 groups of methylation site-related genes may be involved in the regulation of cell cycle expression. Further investigation at the biological and cellular levels will be conducted to elucidate the biological significance of these 15 groups of methylation-related genes.

Finally, our research results have the following advantages: (1) 15 groups of methylation sites were screened from 50,000 methylation sites using the wraparound strategy using XGBoost, LightGBM, and CatBoost models as feature screening models, which substantially reduced the workload, time cost, and workforce and cost consumption of methylation site screening compared with the traditional statistical and biological medical methods. (2) A machine learning model was constructed for the 15 groups of methylation sites for biological age prediction based on the feature selection strategy, which provides a valuable reference for age-related disease prevention and health care. (3) Based on the 15 groups of methylated sites screened by the wraparound feature selection strategy, we found the related genes and proteins in the relevant databases, which can provide a valuable reference for studying the aging of cellular nucleus tissues in biology and medicine. Furthermore, KEGG and GO enrichment analyses were employed to elucidate the biological significance of the genes associated with the 15 identified methylation sites. This provides a reference point for the study of cell cycle expression regulation. The present study is also subject to the following limitations: 1. The limited methylation data (5 W) used did not comprehensively cover all methylation sites, which may have omitted some sites associated with cellular and tissue senescence. Furthermore, the absence of clinical, biological indicators, and behavioral lifestyle data in our dataset prevented us from constructing biological ages that could reflect the effects of behavioral lifestyle on senescence. Consequently, the screening of methylation sites based on the tagged biological age in the dataset is subject to certain limitations.3. Despite the KEGG and GO enrichment analyses initially exploring the biological processes in which the genes associated with the 15 groups of methylation sites might be involved, they failed to provide specific explanations at the level of cellular and molecular mechanisms. Consequently, further improvements are required in studying the molecular mechanisms of the 15 methylation sites associated with cell and tissue aging.

We successfully constructed a biological age prediction model by applying data preprocessing, wraparound feature selection, statistical analysis, and training machine learning model methods on human methylation data obtained from the first World Science Intelligence Contest: Life Science Track - Biological Age Evaluation and Age-Related Disease Risk Prediction held on AliCloud's Tianchi platform. We constructed a biological age prediction model and queried relevant databases to obtain relevant genes and proteins. XGBoost, LightGBM, CatBoost, and deep neural network models were applied to build the biological age prediction model. The XGBoost model obtained the best performance compared with other models, with an MAE of 0.7689 in the training dataset, an R2 value of 0.9973 for the prediction results versus the biological age, and an average residual difference of 0.0005. In the test data set, the MAE is 3.6089, the R2 value of the prediction result with biological age is 0.9441, and the average residual is 0.0939. The biological age prediction model we constructed provides an essential reference for preventing and providing health care for age-related diseases.

The SHAP of the 15 groups of methylation data was calculated to show that the cg23995914 locus had the highest contribution value in biological age prediction. In addition, the information on related genes and proteins corresponding to the 15 groups of methylation sites was obtained by querying the Probes & Genes database and STRING database (Table 3 of the supplementary document), which provides a reference for the study of cellular and tissue aging in biological and medical neighborhoods.

Materials and methods

Data sources

We obtained the human DNA methylation data through the first World Science Intelligence Contest: Life Science Track - Biological Age Evaluation and Age-Related Disease Risk Prediction held on the AliCloud Tianchi platform DNA methylation data; the dataset contains 10,296 samples, of which 7,833 are healthy samples 2,463 are diseased samples; The DNA methylation data were quantified using methylation 450k microarrays, and the specific details can be retrieved from the EWAS[43] database. Each sample provides methylation data, age, and disease at 485,512 loci. 8233 (79.96%) of the samples from the human DNA methylation data were defined as training set data (containing biological age data), and 2063 samples were defined as test samples (not containing biological age data). In this study, we selected the 8,233 samples containing biological age data to construct the prediction model. The data processing and machine learning process is shown in Fig. 5. The first stage of the process involved data pre-processing, which included the following: counting and processing blank data, detecting and replacing anomalous data, converting data types and encoding data.2. A parenthesized feature selector was constructed to search for methylated features using the XGBoost, LightGBM and CatBoostmo models.3. Statistical analyses were performed, which included tests for normality of the methylated data, correlation calculation, data normalization and scaling. Differences in biological age between healthy and diseased groups and between diseases were also analysed.4. Machine learning models were constructed to fit the DNA methylation and biological age data, and SHAP was used to calculate the contribution of each methylation site to the model prediction. GO enrichment, and KEGG analyses explored the biological significance of genes associated with methylation sites.

Our study was approved by the Ethics Committee of Guizhou Medical University under the approval number 2024 Lunar Review No. (159). It was a retrospective, non-interventional study using data from the Aliyun Tianchi database without direct patient contact. We applied for a waiver of informed consent from our host

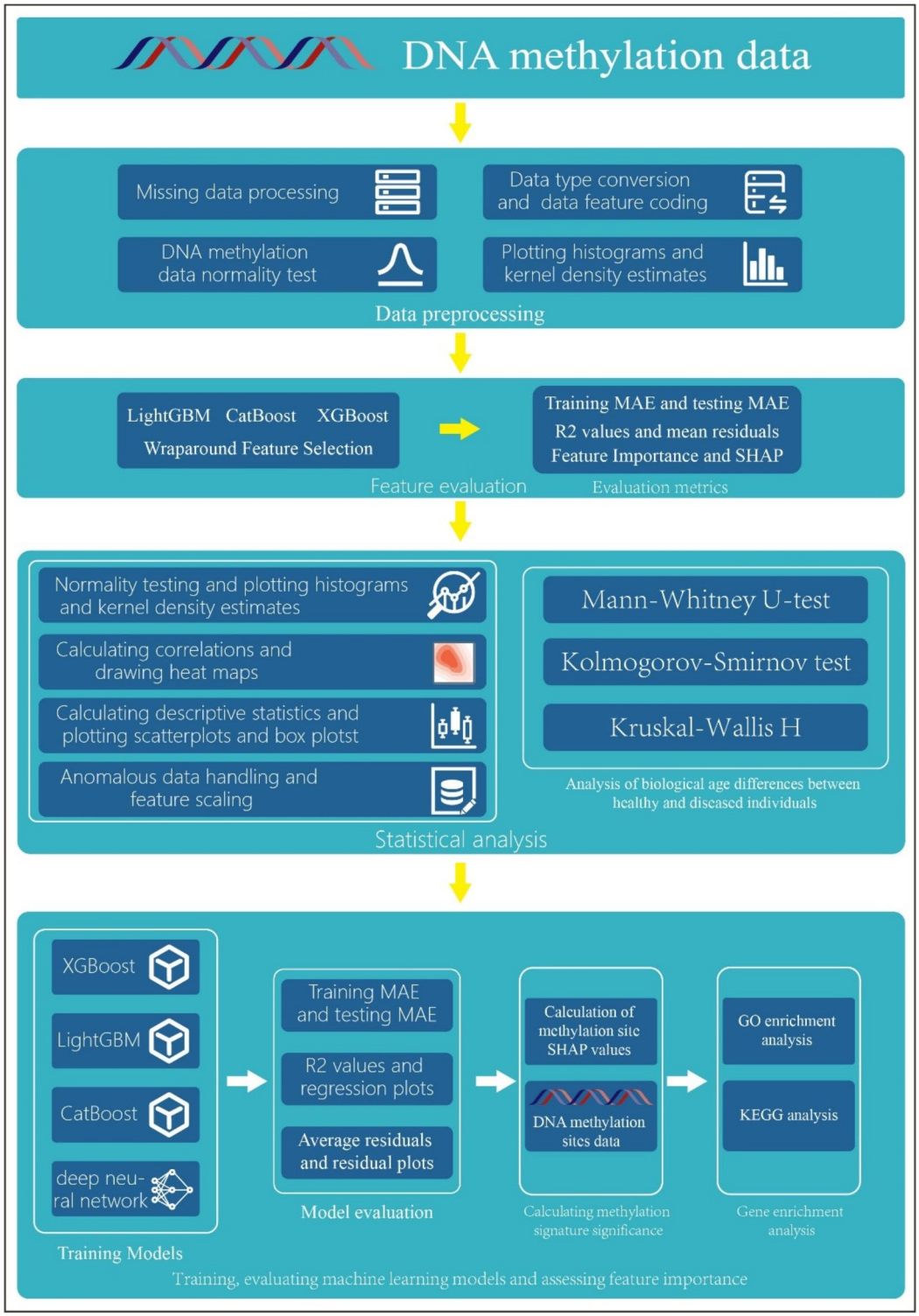

Fig. 5. Data processing and machine learning flowchart.

institution and received approval from the Ethics Committee. The Declaration of Helsinki of the World Medical Association conducted the study.

Data pre-processing

First, we performed statistical calculations and filled in the missing data in the dataset to achieve data integrity; subsequently, we performed data type conversion and data feature coding and used the 0 ~ of ~ n-1 method to code the fixed-class data into n values ranging from 0 to n-1 (n depends on the total number of fixed-

class data categories in the coding work) for the subsequent analysis work. After missing data processing and transformation of the fixed-class data, the methylation data were examined for average distribution properties. Finally, histograms of biological age and sex data and kernel density estimates were plotted to assess the distribution of biological age data and sex data.

Feature selection

Given the large amount of methylation data contained in each sample in the dataset, which is about 485,512 sites, a comprehensive analysis of all methylation sites is burdensome, and it is difficult to determine which sites contribute more to biological age prediction. Therefore, performing feature engineering to filter key features and reduce feature dimensionality is especially critical. We adopted a wraparound strategy for feature engineering, in which XGBoost, LightGBM, and CatBoost are used as feature selection models. We run the model and output the training and testing MAE, R2 value, and average residual of the prediction result and the actual result, and the feature importance of each model; we comprehensively evaluate the training and testing MAE, R2 value, and average residual of the prediction result and the natural result of each model and select the optimal model and the feature importance of the output of the model as the result of feature engineering. Finally, SHAP was used to calculate the contribution of methylation data to biological age after feature selection.

Statistical analysis

Firstly, we tested the normality of the data after feature selection and plotted data distribution histograms and kernel density estimation plots to understand the data distribution of the data; according to the results of the normality test to calculate the correlation between the methylation data and the biological age, if the normality test was passed, then we selected the Pearson correlation coefficient. Otherwise, we selected the Spearman correlation coefficient and, through the correlation coefficient heatmap and the P-value heatmap, Visualize the correlation between methylation data and biological age. Select the features with a correlation greater than 0.45 as the final feature data for machine learning, calculate the descriptive statistics of the final feature data (mean, standard deviation, minimum, first quartile, median, third quartile, and maximum), and draw box plots and regression scatter plots to visualize the distribution of the feature data; Data processing and feature scaling of anomalous data based on descriptive statistics results, box plots and regression scatter plots to improve data quality and accelerate machine learning model convergence.

Morgan E. Levine et al., in a study exploring epigenetic biomarkers of aging using an integrated clinical measure that combines phenotypic age, argued that differences in the rate of aging would have implications for a wide range of diseases and conditions[28]. Accordingly, the Mann-Whitney U and Kolmogorov-Smirnov tests were employed to investigate the discrepancies in biological age between the healthy and diseased groups. Additionally, histograms were constructed to illustrate the distribution of biological age between the two groups. The Kruskal-Wallis H-test was employed to investigate the variance in biological age across different disease groups. At the same time, ridge and box plots were utilized to visualize the distribution of biological age within these groups.

Machine learning model training and evaluation

Following data preprocessing, feature selection, and statistical analysis, we employed XGBoost, LightGBM, CatBoost, and deep neural networks as machine learning models for training. This was conducted using 10-fold cross-validation and grid search for hyper-parameter tuning of the optimal models. The metrics employed to assess model training efficacy included the training MAE and the testing MAE. The consistency of model predictions was evaluated using R2 values for the predicted biological age and the actual biological age. The stability of model predictions was gauged by examining the mean residuals of the predicted biological age and the actual biological age. The optimal model was selected through a comparative analysis of the metrics above.

To enhance the model's interpretability and comprehend the impact of methylation at disparate sites on biological age prediction, we employed SHAP to ascertain the SHAP values of methylated sites. We generated SHAP summary plots and bar-stacked plots to illustrate the SHAP of the methylation data. Additionally, we conducted GO and KEGG analyses to investigate the biological significance of genes associated with methylated sites.

Data availability

DNA methylation data is available at https://tianchi.aliyun.com/competition/entrance/532114/introduction?spm=a2c22.12281925.0.0.22bc7137BkwykY, and the code and images covered in this article are available at https://github.com/Kosonora/DNA-methylation-and-biological-age-study.git.

Received: 22 May 2024; Accepted: 7 October 2024
Published online: 15 October 2024

References

1. Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective[J]. *Trends Genetics*. **38** (7), 676–707 (2022).
2. K H L et al. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res.* **44** (14), 6693–6706 (2016).
3. Olya, R. J. E. & ,Mathieu, Y. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin*. **10** (1), 23 (2017).
4. Jones, M. J., Goodman, S. J. & Kobor, M. S. DNA methylation and healthy human aging. *Aging Cell.* **14** (6), 924–932. https://doi.org/10.1111/acel.12349 (2015). Epub 2015 Apr 25. PMID: 25913071; PMCID: PMC4693469.
5. Benayoun, B. A., Pollina, E. A. & Brunet, A. Epigenetic regulation of aging: linking environmental inputs to genomic stability. *Nat. Rev. Mol. Cell. Biol.* **16** (10), 593–610. https://doi.org/10.1038/nrm4048 (2015). Epub 2015 Sep 16. PMID: 26373265; PMCID: PMC4736728.

6. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine[J]. *Lancet.* **392** (10149), 777–786 (2018).
7. Higham, J., Kerr, L., Zhang, Q. et al. Local CpG density affects the trajectory and variance of age-associated DNA methylation changes[J]. *Genome Biol.* **23** (1), 216 (2022).
8. Goel, N., Karir, P. & Garg, V. K. Role of DNA methylation in human age prediction[J]. *Mech. Ageing Dev.* **166**, 33–41 (2017).
9. Ieva, R., Finn, D. & Beck, M. R. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clin. Epigenetics*. **11** (1), 193 (2019).
10. Bi, Q., Goodman, K. E., Kaminsky, J. & Lessler, J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. ;188(12):2222–2239. doi: (2019). https://doi.org/10.1093/aje/kwz189. PMID: 31509183.
11. Shin, S. et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail.* **8** (1), 106–115 (2021). Epub 2020 Nov 17. PMID: 33205591; PMCID: PMC7835549.
12. Lombardi, A. et al. Explainable deep learning for personalized age prediction with brain morphology. *Front. Neurosci.* **15**, 674055. https://doi.org/10.3389/fnins.2021.674055 (2021). PMID: 34122000; PMCID: PMC8192966.
13. Nusinovici, S. et al. Retinal photograph-based deep learning predicts biological age and stratifies morbidity and mortality risk. *Age Ageing*. **51** (4), afac065. https://doi.org/10.1093/aging/afac065 (2022). PMID: 35363255; PMCID: PMC8973000.
14. Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. J. W. L. & Lu, M. T. Deep learning to Estimate Biological Age from chest radiographs. *JACC Cardiovasc. Imaging*. **14** (11), 2226–2236. https://doi.org/10.1016/j.jcmg.2021.01.008 (2021). Epub 2021 Mar 17. PMID: 33744131.
15. Galkin, F. et al. A methylation aging clock developed with deep learning. *Aging Dis.* **12** (5), 1252–1262. https://doi.org/10.14336/AD.2020.1202 (2021). PMID: 34341706; PMCID: PMC8279523.
16. Unnikrishnan, A., Freeman, W. M., Jackson, J. et al. The role of DNA methylation in epigenetics of aging[J]. *Pharmacol. Ther.* **195**, 172–185 (2019).
17. Zubakov, D., Liu, F., Kokmeijer, I. et al. Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length[J]. *Forensic Sci. Int. Genet.* **24**, 33–43 (2016).
18. Jiansheng, Z., Hongli, F., Yan, X. & Genes Age Prediction of Human Based on DNA Methylation by Blood Tissues. 12(6):870–870. (2021).
19. Bernard, D., Doumard, E., Ader, I. et al. Explainable machine learning framework to predict personalized physiological aging[J]. *Aging Cell.* **22** (8), e13872 (2023).
20. Ehrlich, M. DNA hypermethylation in disease: mechanisms and clinical relevance. *Epigenetics*. **14** (12), 1141–1163 (2019).
21. script. Stats. normal test — SciPy v1.12.0 Manual.
22. sklearn Preprocessing.StandardScaler — sci-kit-learn 1.4.0 documentation.
23. sklearn Preprocessing.MinMaxScaler — sci-kit-learn 1.4.0 documentation.
24. Marbaniang, I. A., Choudhury, N. A. & Moulik, S. Cardiovascular disease (CVD) prediction using machine learning algorithms[C]//2020 IEEE 17th India Council International Conference (INDICON). IEEE, : 1–6. (2020).
25. Bian, L. et al. Application, interpretability, and prediction of machine learning method combined with LSTM and LightGBM-a case study for runoff simulation in an arid area. *J. Hydrol.* **625**, 130091 (2023).
26. Lyu, G. & Nakayama, M. Prediction of respiratory failure risk in patients with pneumonia in the ICU using ensemble learning models. *Plos One*. **18** (9), e0291711 (2023).
27. Liang, S. & Srikant, R. Why deep neural networks for function approximation? arXiv preprint arXiv:1610.04161, (2016).
28. https://www.kegg.jp/kegg/kegg1.html
29. Bernard, D. et al. Explainable machine learning framework to predict personalized physiological aging. *Aging cell.* **22** (8), e13872 (2023).
30. Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*. **10** (4), 573 (2018).
31. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)*. **11** (2), 303 (2019).
32. Probes & Genes (cncb.ac.cn)
33. STRING. functional protein association networks (string-db.org).
34. GeneCards. – the human gene database (www.genecards.org).
35. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** (1), 27–30 (2000).
36. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28** (11), 1947–1951 (2019).
37. Kanehisa, M. et al. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51** (D1), D587–D592 (2023).
38. Ghafouri-Fard, S. et al. A review on the role of cyclin-dependent kinases in cancers. *Cancer Cell Int.* **22** (1), 325 (2022).
39. Lim, S. & Kaldis, P. Cdks, cyclins, and CKIs: roles beyond cell cycle regulation. *Development*. **140** (15), 3079–3093 (2013).
40. Kaur, N., Chugh, V. & Gupta, A. K. Essential fatty acids as functional components of foods-a review. *J. Food Sci. Technol.* **51**, 2289–2303 (2014).
41. Wiggins, A. K. A., Mason, J. K. & Thompson, L. U. Growth and gene expression differ over time in alpha-linolenic acid treated breast cancer cells. *Exp. Cell Res.* **333** (1), 147–154 (2015).
42. Huang, W. et al. α-Linolenic acid induces apoptosis, inhibits the invasion and metastasis, and arrests the cell cycle in human breast cancer cells by inhibiting fatty acid synthase. *J. Funct. Foods*. **92**, 105041 (2022).
43. EWAS Datahub (cncb.ac.cn).

Acknowledgements

We thank the World Science Intelligence Contest: Life Science Track - Biological Age Evaluation and Age-Related Disease Risk Prediction Contest held by Aliyun Tianchi and the CFFF Intelligent Computing Platform of Fudan University for providing data support.

Author contributions

Conceptualization: Xun He, Xiaofan Yan, Pengcheng Yan; Methodology: Sheng Zhou, Shanshan Wei; Software: Sheng Zhou, Shanshan Wei, Chengxing Zhou; Validation: Xun He, Xiaofan Yan, Pengcheng Yan, Sheng Zhou, Chengxing Zhou, Jing Chen, Die Wang; Formal analysis: Sheng Zhou, Jing Chen, Shanshan Wei; Investigation: Die Wang, Shanshan Wei, Pengcheng Yan; Resources: Sheng Zhou, Xiaofan Yan; Data Curation: Sheng Zhou, Chengxing Zhou, Xiaofan Yan; Writing - Original Draft: Sheng Zhou, Shansha Wei; Writing - Review & Editing: Shanshan Wei, XiaoFan Yan, Pengcheng Yan; Visualization Sheng Zhou, Shanshan Wei, Chengxing Zhou; Supervision: Shanshan Wei, Die Wang; Project administration: Xun He, Xiaofan Yan, Pengcheng Yan; Funding acquisition: Xun He, Xiaofan Yan.

Funding

This work was supported by the Research Center for Pharmaceutical Economics and Management of Guizhou Medical University under Grant No. [GMUMEM2022-B17].

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-75586-9.

Correspondence and requests for materials should be addressed to X.Y., X.H. or P.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.