International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 358-369 March 2026-Sept 2026 www.csjournals.com

Predicting the respondents Age by extracting a subset of features from large NHANES

Komal, Dr. Anupam Bhatia, Dr. Naveen

MCA Student[1], Associate. Professor[2], Asst. Prof.[3]
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Demography-based ageing is a serious issue in health informatics because of the growing importance of estimating chronological age based on routinely collected physiological and biochemical data. The National Health and Nutrition Examination Survey (NHANES) is a large but high-dimensional data source for such studies, and it can be difficult to model with the full set of variables, which may lead to overfitting and obscurity. The purpose of this paper is to present an empirical analysis that estimates age of respondents based on a small set of health measures from NHANES. From seven possible biomarkers, two independent feature ranking methods (univariate F-regression scoring (SelectKBest) and Random Forest impurity importance) were employed and four biomarkers (triglycerides, insulin, body mass index, and fasting glucose) were retained. Using the 80:20 split and 5-fold cross validation, six regression algorithms were trained and benchmarked with two different configurations: (1) All features, (2) Reduced 4 features.An all-feature configuration and a reduced 4-feature configuration were used to train and benchmark six regression algorithms, Linear Regression, Decision Tree, Random Forest, Support Vector Regression, K-Nearest Neighbors and XGBoost. With the full feature set the best performance was achieved with the Random Forest model (R2 = 0.3470, MAE = 13.51 years), and with the reduced feature set, Support Vector Regression was the best performing model (R2 = 0.3454, MAE = 13.35 years). The close correspondence of these results suggests that a small number of interpretable biomarkers may provide an approximation to the predictive power of the full set of variables and provide a less burdensome and more transparent alternative to age-estimation in survey-based health analytics. Mathematical models—Random Forest, Support Vector Regression, Mathematical models for chronological age prediction, Mathematical models for feature selection, NHANES, Healthcare data analytics,

Mathematical models for machine learning regression; Mathematical models for support vector regression; Mathematical models for random forest.

I.INTRODUCTION

The ability to estimate a person's chronological age by health examination measurements is in the middle of statistical learning and biomedical informatics. Usually age is considered a fixed demographic variable, but the biochemical and anthropometric signatures of ageing (e.g., changing glucose regulation, accumulation of lipids, changes to insulin sensitivity) allow a well-trained model to approximate the same age value solely from physiological signals. This ability can be useful wherever written age is not available or is unreliable, and it can also be an indirect test of the ability of certain biomarkers to reflect the aging process.

The National Center for Health Statistics (NCHS), part of the U.S. Centers for Disease Control and Prevention (CDC), conducts the National Health and Nutrition Examination Survey (NHANES) to assess the health and nutritional status of the U.S. population. NHANES is a multi-modal survey that includes structured interviews, physical exams, and laboratory assays, yielding a data set that includes anthropometric, biochemical, dietary, and demographic information for a large and diverse sample size.

This is also NHANES's main analytic challenge. A survey instrument with hundreds of variables that is administered repeatedly over time results in a dimensionality that is much larger than the one required by the prediction task at hand. When an unfiltered variable set is fed into a learning algorithm, it often causes poor generalization, slow training, and makes it hard to understand how each prediction is made, which is known as the curse of dimensionality. Systematic feature selection is thus not something that can be done as a fine-tuning, but rather a necessary step to create a defensible, interpretable model with such data.

Studies on metabolic aging have been conducted repeatedly in the past and have consistently demonstrated correlations between metabolic markers, including fasting glucose, serum triglycerides concentration, insulin concentration and body mass index, and the age-related deterioration in metabolism. These four are low-cost, routinely tested, and appealing options to include in a model suitable for clinical use that will estimate age. The present study follows the idea and adopts a principled feature-selection procedure in conjunction with a comparison of the performance of six regression techniques twice, using the full seven-variable predictor set and the four-variable predictor set selected from that set.

This paper is organized as follows: Next, in Section II, the literature related to machine learning applications to NHANES data and biological age estimation is surveyed. Section III provides a focus for the specific gap that this study will address. The research objectives are given in Section IV. The proposed methodology, system architecture, algorithms, and implementation details are explained in Sections V-VIII. The experimental results are reported and discussed in Section IX and the conclusions, directions for future work in Section X and XI are provided.

II. LITERATURE REVIEW

The machine learning methods used to analyze NHANES data have been predominantly focused on binary or multi-class classification of diseases, which is in contrast to the focus on continuous outcome prediction. Abbas et al. [2] used Adaptive Boosting (AdaBoost) with Support Vector Machines to identify diabetes and cardiovascular disease risk in NHANES participants, obtaining an ensemble accuracy of 98.54%, and determining that age, body mass index, and glucose concentration were the features most discriminatory for the problem. Their discovery that these three variables are most important in models of disease risk also indirectly supports the use of the same biomarkers as candidate predictors if the direction of the relationship is reversed, namely for age as the quantity to be predicted.

Vangeepuram et al. [3] examined the ability of the existing guidelines for pediatric and adolescent screening to identify prediabetes and diabetes risk in NHANES records, only 43.1% of the rule-based guideline was sensitive, and 67.6% was specific.

More relevant to the current study, Deng et al. [4] developed the Nanchang biological-age (NC-BA) model using LASSO feature selection and a generalized linear model and deep generalized

linear model and tested the model's transferability in an independent NHANES cohort. Their research showed that a survey of biomarkers could be used to measure chronological age very accurately and help identify a risk of non-alcoholic fatty liver disease (NAFLD), highlighting the clinical utility of age-prediction pipelines based on a data-driven biomarker panel.

Fu et al. [5] created a model for osteoarthritis risk in NHANES individuals age 45 or older using several tree-based learners, and found that a reduced set of predictors performed best in terms of discrimination, with age, sex, and BMI ranking as the top three contributors to the model. Similarly, Tachie et al. [6] evaluated K-Nearest Neighbors, Decision Trees, Support Vector Machines and Artificial Neural Networks for predicting fatty-acid composition in U.S. snack food products collected from NHANES, and determined that the ensemble tree method Random Forest was the best performer of the algorithms tested, in terms of low prediction error and high coefficient of determination.

Chen et al. [7] also created a prediction model for erectile dysfunction based on XGBoost when using NHANES cycles 2001–2004, which had an AUC of approximately 0.89, and identified, once again, age and adiposity measures as important predictors. As a whole, the data exhibits two common trends: gradient-boosted and bagged tree ensembles frequently outperform simpler linear baselines across a variety of NHANES-related tasks, and a small number of metabolic and anthropometric variables - glucose, triglycerides, insulin, and BMI - consistently show up as high-value predictors across a wide range of outcome variables, a pattern that leads to their selection as the reduced feature panel analyzed in this study.

TABLE I. SUMMARY OF REVIEWED STUDIES

Ref.	Focus	Method	Key Finding
[2]	Diabetes/CVD risk	AdaBoost+SVM	98.54% acc.; age, BMI, glucose top
[3]	Youth diabetes screen	ML vs. guideline	Guideline sens. 43.1%; ML better
[4]	Biological age/NAFLD	LASSO+GLM	Transferred to NHANES; age-NAFLD link
[5]	Osteoarthritis risk	CatBoost	Best AUC; age, sex, BMI top
[6]	Fatty-acid class	KNN/DT/SVM/ANN/RF	RF lowest error, highest R2
[7]	Erectile dysfunction	XGBoost	AUC≈0.89; age, adiposity key

III. RESEARCH GAP

When reading the above studies and the literature on NHANES in general, four gaps are identified that can be resolved by this study. First, chronological age is almost always used as an

input variable for predicting some other clinical outcome rather than as the outcome to be predicted, and the latter is relatively unexplored. Third, the use of a wide range of algorithms (linear, tree-based, kernel-based, instance-based, and boosted) and a full vs. reduced feature set on the same NHANES extract is rare; most papers report one of the best-performing models, but not a side-by-side experimental comparison. Fourth, the specific combination of triglycerides, insulin, BMI, and glucose, assessed together for the specific purpose of estimating age, has not been investigated in the literature reviewed as a whole. To address these gaps, this paper focuses on age as the prediction objective, uses two independent feature-ranking algorithms to justify the decreased feature set, and compares six regression algorithms on both the feature set size, under a consistent experimental protocol.

IV. OBJECTIVES

The following objectives will guide the research:

1)To review the available methods of age estimation using physiological and health survey information and the limitations of the methods.

2) To develop a process that can be replicated for the NHANES extract to perform duplicate screening, remove leakage-safe columns, correct for skewness, and scale features.

3) To find the most informative predictor subset for age estimation by applying two different predictor subset selection methods – SelectKBest using F-regression and Random Forest importance ranking.

4.) Train and test six regression models: Linear Regression, Decision Tree, Random Forest, SVR, KNN and XGBoost on the original and shortened predictor sets.

5) To evaluate model performance with respect to MAE, RMSE and R2 with five fold cross validation for generalizing the model.

6) To compare the predictive accuracy of the four biologically interpretable biomarkers-only configuration with the seven-variable configuration.

V. PROPOSED METHODOLOGY

The overall workflow is similar to a typical supervised-regression pipeline, but adjusted to the particular type of NHANES extract utilized for this study.

A.Dataset

The dataset used for this project is called the working dataset and includes 2,278 NHANES respondent records from the ages of 12 to 80 years, which are available in the public domain via the CDC/NCHS NHANES data repository in addition to the UCI machine learning repository. The 10 fields are present at the beginning: the identifier for the respondent (SEQN), a derived age group classification (AGE10), the variable of interest (RIDAGEYR – age in years), and seven candidate predictors: gender (RIAGENDR), physical activity participation (PAQ605), body mass index (BMXBMI), fasting glucose (LBXGLU), diabetes diagnosis status (DIQ010), triglycerides (LBXGLT), and insulin (LBXIN). There were no missing ages, and the mean age was 41.80 years, with a standard deviation of 20.16 years.

B. Preprocessing

Duplicate rows were checked and none were found. The SEQN identifier was not predictive and was dropped, and the age_group field was dropped due to it being a direct categorical derivative of the target variable, which attempts to label leakage. This resulted in 7 predictor columns and 1 target column. Glucose, triglycerides and insulin were highly right-skewed, (raw skewness of 7.19, 2.88, 2.78 respectively), and were subsequently log1p transformed which reduced their skewness to 2.76, 0.62, 0.31 respectively. Each of the predictor features was then scaled to be centered to the mean of the training partition of the data only, for the test partition to be scaled, but not seen during the scaling, and a fixed random seed was used for the split so the test set is reproducible, with 1,822 training and 456 test records.

C. Feature Selection

Two independent, complementary techniques were used to rank the seven candidate predictors. SelectKBest with F-regression scoring and RF importance measures the strength of linear association between each predictor and the target, and the non-linear and interactions effects that a linear score fails to capture, respectively, across the trees of a fitted ensemble. It is worth to note that both rankings resulted in the same top four predictors: triglycerides, insulin, BMI and glucose, with a cumulative importance of about 95% on Random Forest. This joint use of a statistical and a model-based method resulted in a second, simplified experimental configuration, which was used solely with the four variables.

VI. SYSTEM ARCHITECTURE

The entire system is run as a five-layer pipeline, where each layer uses the output of the previous layer as its input, and thus all downstream models receive a representation that is exactly the same.

A.Data Input Layer

Reads the raw NHANES CSV file and checks the shape and data types and initializes a check for missing values and duplicate rows.

B. Preprocessing Layer

Drops columns with data that are irrelevant or repeated and corrects skewed data by using log1p transformation, as well as standardizing all numerical predictors.

C. Feature Engineering Layer

Performs SelectKBest (F-regression) and Random forest (RF) importance ranking in parallel and creates two experimental feature sets, one with 7 features, and one with 4 features

D. Modeling Layer

Trains each of the six regression algorithms independently on each of the two feature sets with an identical 80:20 train-test split and thus any performance difference is due to the algorithm and configuration of the features and not due to a difference in how the data is split.

E. Evaluation and Output Layer

Calculates MAE, RMSE and R2 for each model, does 5-fold cross-validation on the best model found, compares the results between models and configurations, and saves the winning model with joblib to be reused later.

VII. ALGORITHMS / MODELS USED

Six supervised regression algorithms were chosen from four families: linear, tree-based, kernel-based, instance-based, and boosted-ensemble, to provide a wide range of comparative results on the age-prediction task, thus reflecting the different inductive biases used by each algorithm.

A. Linear Regression

Is the default model, which fits a linear function of the predictors by minimizing the sum of squared residuals under ordinary least squares. The upper limit of its performance suggests the linearity of the age-biomarker relationship.

C. Support Vector Regressor

Recursively divides the feature space into regions and estimates the mean target value on each region. The depth was limited to 5 and the minimum split size was 10 and the minimum leaf size was 5, to prevent overfitting.

C. Random Forest Regressor

This is an ensemble of 200 bagged decision trees (with maximum depth of 6 and minimum split of 15 and minimum leaf of 5) whose prediction are averaged to reduce variance and improve generalization and also provided the importance scores used for feature selection.

D. Support Vector Regressor (SVR)

Fits an epsilon-insensitive hyperplane with default RBF kernel (which is appropriate for non-linear age-biomarker relationship after features are standardized).

E. K-Nearest Neighbors Regressor (KNN)

Regresses the age of responses to the target age of one of its neighbors that is closest in Euclidean distance when the distance between the feature vectors is computed in the scaled feature space.

F. XGBoost Regressor

A sequential ensemble of 300 shallow trees (max depth of 4, learning rate of 0.05, subsample of 0.8, column-subsample of 0.8) with gradient boosting that minimizes the residual errors of previous trees, while regularizing against overfitting**.**

VIII. IMPLEMENTATION

The pipeline has been realized in Python in a Jupyter Notebook on Windows 11. Data loading, cleaning and numerical operations were performed using Pandas and NumPy, preprocessing utilities (StandardScaler, SelectKBest) and numerical classifiers (Linear Regression, Decision Tree, Random Forest, SVR, KNN) were provided by scikit-learn, while the gradient-boosted regressor was provided by XGBoost library, the exploratory visualizations (histograms, scatter plots, correlation heatmaps) were obtained using Matplotlib and Seaborn and the selected model was serialized using joblib.

The eight steps of the implementation were the same for both configurations: (1) construct scaled feature matrix and target vector for given configuration; (2) split data into two partitions for training and testing (80:20); (3) create instances of each model, using hyperparameters described in Section VII, and fit the models on the training partition; (4) predict on held out test partition; (5) calculate MAE, RMSE and R2 values as difference between prediction and actual data; (6) perform five-fold cross validation on Random Forest to get more conservative estimation of performance; (7) check for overfitting by comparing training and test scores; and (8) save the learned Random Forest model to disk.

The same split, scale and evaluation process was applied to all algorithm and feature configurations, so any differences seen in performance are due to the algorithm's inductive bias, or due to the differences in how the features were reduced, and not due to inconsistencies in data handling.

IX. RESULTS AND DISCUSSION

The outcomes of the feature selection process.

A.Feature Selection Outcomes

The F-scores for the seven candidate predictors are reported in Table II and the importance scores in Table III using the SelectKBest and Random Forest methods, respectively. The rankings are the same for both, with triglycerides, glucose, BMI, and insulin having the highest scores, and gender, physical activity and diabetes diagnosis having relatively low scores. This agreement between the two methods, one based on statistics and the other on correlation, and the second based on reduction in impurity, lends confidence that the four biomarkers that were selected are not artifacts of any one scoring criterion.

TABLE II. SELECTKBEST F-REGRESSION SCORES

Rank	Feature	F-Score
1	LBXGLT (Triglycerides)	291.31
2	LBXGLU (Glucose)	202.93
3	BMXBMI (BMI)	50.38
4	LBXIN (Insulin)	22.83
5	DIQ010 (Diabetes)	5.70
6	PAQ605 (Activity)	1.54

Rank	Feature	F-Score
7	RIAGENDR (Gender)	0.09

TABLE III. RANDOM FOREST FEATURE IMPORTANCE

Rank	Feature	Importance
1	LBXGLT (Triglycerides)	0.2760
2	LBXIN (Insulin)	0.2580
3	BMXBMI (BMI)	0.2396
4	LBXGLU (Glucose)	0.1764
5	RIAGENDR (Gender)	0.0228
6	PAQ605 (Activity)	0.0211
7	DIQ010 (Diabetes)	0.0061

B. Experiment 1 — All Seven Features

Interruption of the Fourteen Triadic Patterns

Table IV shows the results of the Modelfit, for the case of including all predictors. Random Forest achieved the lowest error and highest fit among the six algorithms (MAE = 13.51 years, RMSE = 16.50 years, R2 = 0.3470), with SVR close behind (R2 = 0.3419) and KNN in third place (R2 = 0.3307). The other two models did not perform well and both were around R2 = 0.22-0.23, suggesting that neither a linear model nor a shallow decision tree is able to capture the interaction effects within the metabolic predictors. A 5-fold cross validation for the Random Forest model gave the score of 0.2724, 0.3248, 0.3498, 0.3123, and 0.2959, with an average score of 0.3111, and training and test scores of 0.4499 and 0.3470, respectively, indicating mild overfitting but still good generalization.

TABLE IV. MODEL PERFORMANCE — ALL FEATURES (7 PREDICTORS)

Model	MAE	RMSE	R2
Linear Regression	15.00	17.95	0.2279
Decision Tree	14.34	17.99	0.2243
Random Forest	13.51	16.50	0.3470
SVR	13.52	16.57	0.3419
KNN	13.54	16.71	0.3307
XGBoost	≈13.6	≈16.8	≈0.330

C. Experiment 2 — Selected Four Features

Selected Four Features.

Performance for the reduced four feature configuration is reported in Table V. Here SVR takes the lead (MAE = 13.35 years, RMSE = 16.53 years, R2 = 0.3454), narrowly ahead of

Random Forest (R2 = 0.3398). The biggest drop in this configuration, KNN (R2 = 0.2339) compared to KNN with all features (R2 = 0.3307), occurred when the three variables (gender, activity level and diabetes status) that were removed were not important on their own. Random Forest cross-validation in this configuration produced fold scores of 0.2668, 0.3203, 0.3473, 0.3042, and 0.2900 (mean 0.3057), with a training score of 0.4435 against a test score of 0.3398.

TABLE V. MODEL PERFORMANCE — SELECTED FEATURES (4 PREDICTORS)

Model	MAE	RMSE	R2
Linear Regression	15.01	17.86	0.2354
Decision Tree	14.34	17.99	0.2243
Random Forest	13.56	16.60	0.3398
SVR	13.35	16.53	0.3454
KNN	14.16	17.88	0.2339
XGBoost	≈13.7	≈16.9	≈0.328

D. Comparative Analysis

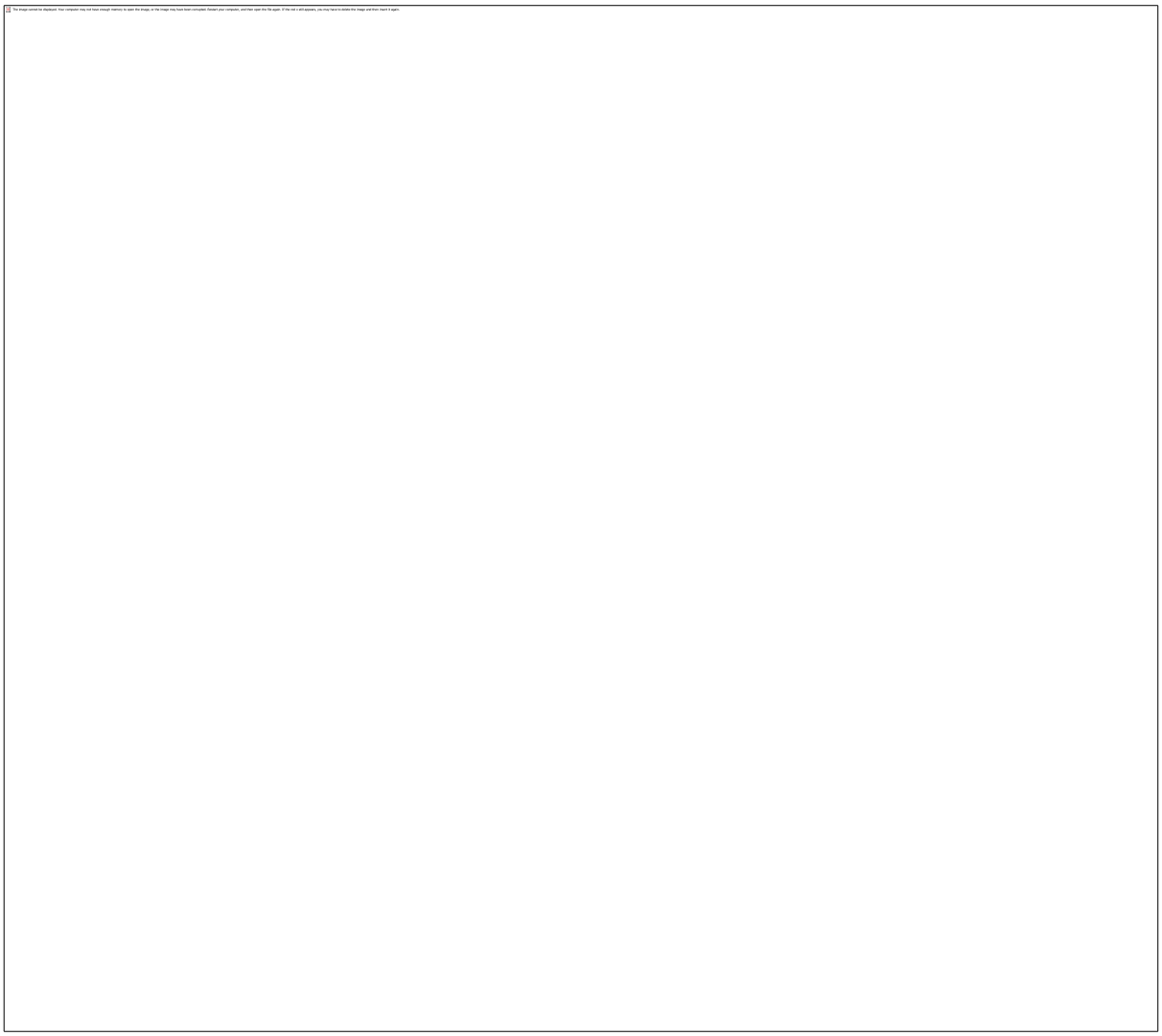
Table VI makes it easy to compare the R2 values between the two configurations. The change for 4 of the 6 algorithms is fairly small (about 0.01 R2), which shows that the overall loss in predictive accuracy of dropping three of the least important variables is minimal. SVR and Linear Regression benefit slightly from the smaller set – likely due to the addition of mild noise and/or collinearity in the larger set, which is less useful to these models that are sensitive to feature geometry; Random Forest and XGBoost lose a little accuracy – consistent with tree ensembles being useful even on weakly informative variables, when available; and the Decision Tree is not affected as it is shallow enough to rely on the same dominant features in both sets. The obvious exception is KNN whose sharp drop suggests that feature importance measures computed by linear or tree-based methods do not necessarily carry over to distance-based learners.

TABLE VI. COMPARATIVE R2 ACROSS FEATURE CONFIGURATIONS

Model	R2(All)	R2(Sel)	ΔR2	Best
Linear Regression	0.2279	0.2354	+0.0075	Selected
Decision Tree	0.2243	0.2243	0.0000	Equal
Random Forest	0.3470	0.3398	−0.0072	All Features
SVR	0.3419	0.3454	+0.0035	Selected

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 358-369 March 2026-Sept 2026 www.csjournals.com

Model	R2(All)	R2(Sel)	ΔR2	Best
KNN	0.3307	0.2339	−0.0968	All Features
XGBoost	≈0.330	≈0.328	≈−0.002	All Features



E. Discussion

There are several implications from these findings. The R2 ceiling that was achieved by all the algorithms is not due to any particular algorithm, but is instead due to the task itself; chronological age is influenced by many genetic, behavioral, and environmental factors that are not captured in these 7 metabolic and demographic variables and is therefore not surprising that a moderate R2 was achieved. This margin by ensemble methods (Random Forest, XGBoost) outperforming linear regression reinforces that the ensemble models are capturing the relationship between these biomarkers and age well, and therefore have interaction effects that a linear coefficient per biomarker can't capture. This is somewhat of

an interesting result, as it shows that SVR performs better than Random Forest when the feature space is limited to four variables, suggesting that its RBF kernel is especially well suited to modelling a reduced, less noisy biomarker space, which is a practically useful result as it suggests that a light weight, four-input model is a possible candidate for deployment. Finally, the simple difference between the single-split test scores and the more conservative averages of the five-fold cross-validated scores for the full feature set (0.3470 versus 0.3111) and reduced set (0.3398 versus 0.3057) remind us that the single train-test split can be slightly optimistic and that the cross-validated estimates have to be viewed as the more reasonable expectation of how the model will perform on unseen respondents.

X. CONCLUSION

The purpose of this study was to assess if a small, clinically meaningful subset of NHANES health measures could be used to predict chronological age similarly accurately as the larger set of measures. The two independent feature selection methods, a statistical one and a model-based method, agreed upon the four biomarkers, which gained confidence in the relevance of these biomarkers for the task. In all of the six regression algorithms tested, using all seven features (R2 = 0.3470, MAE = 13.51 years) versus four features (R2 = 0.3454, MAE = 13.35 years) had little difference for all except SVR; for that algorithm, it was the strongest when using all seven features compared to the four. The results validate the main hypothesis of this study: a small number of easily interpretable, biological-based biomarkers can match the predictive capacity of a larger, less interpretable feature set, while providing a more tractable and computationally efficient alternative for age-estimation problems in survey-based health data.

XI. FUTURE SCOPE

Based on this study, the following directions for extension were identified:

Adding more complex neural network models like deep learning networks (DLN's), like multi-layer perceptrons or attention-based models, to capture more complex non-linear relationships between biomarkers.

Utilizing other NHANES variables for the predictor set such as dietary intake, blood pressure, complete blood count variables, to increase potential R2.

Combining multiple cycles of NHANES to increase sample size and enhance the representativeness of populations by subgroups.

Using the target as a biological age, not chronological age, to allow for deviations between the target and actual age to identify accelerated ageing or higher health risk.

Partial tuning of hyper-parameters for the six algorithms (as done in this study) instead of performing a systematic search (grid, random or bayesian optimization).

Incorporation of explainability features like SHAP to highlight local and global feature contribution would support the argument for clinical usage.

Harvesting external clinical validation of reduced-feature model with healthcare practitioners.

Raising the proportion of adults and senior respondents by either resampling or class-weighted training.

REFERENCES

[1] National Center for Health Statistics, "About NHANES," Centers for Disease Control and Prevention, Dec. 2024. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about/

[2] I. Abbas, S. Israil, M. Bayu, and R. W. Sembiring, "Data-driven approach to age prediction on patients' diabetes and cardiovascular diseases using machine learning: National Health and Nutrition Health Survey (NHANES)," Research Square preprint, 2023, doi: 10.21203/rs.3.rs-3764619/v1.

[3] N. Vangeepuram, B. Liu, P. Chiu, L. Wang, and G. Pandey, "Predicting youth diabetes risk using NHANES data and machine learning," Sci. Rep., vol. 11, art. 11212, 2021, doi: 10.1038/s41598-021-90406-0.

[4] L. Deng, J. Huang, H. Yuan, Q. Liu, W. Lou, P. Yu, X. Xie, X. Chen, Y. Yang, L. Song, and L. Deng, "Biological age prediction and NAFLD risk assessment: a machine learning model based on a multicenter population in Nanchang, Jiangxi, China," BMC Gastroenterol., vol. 25, art. 172, 2025, doi: 10.1186/s12876-025-03752-y.

[5] Y. Fu, Y. Yu, and W. Chen, "Constructing machine learning-based risk prediction model for osteoarthritis in population aged 45 and above: NHANES 2011–2018," Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-99411-z.

[6] C. Y. E. Tachie, D. Obiri-Ananey, N. A. Tawiah, N. Attoh-Okine, and A. N. A. Aryee, "Machine learning approaches for predicting fatty acid classes in popular US snacks using NHANES data," Nutrients, vol. 15, no. 15, art. 3310, 2023, doi: 10.3390/nu15153310.

[7] X.-Y. Chen, W.-T. Lu, D. Zhang, M.-Y. Tan, and X. Qin, "Development and validation of a prediction model for ED using machine learning: according to NHANES 2001–2004," Sci. Rep., vol. 14, art. 27279, 2024, doi: 10.1038/s41598-024-78797-2.

[8] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16), 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[12] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in Advances in Neural Information Processing Systems 9 (NIPS 1996), 1997, pp. 155–161.

[13] J. D. Hunter, "Matplotlib: A 2D graphics environment," Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.