iDT-diet: Toward Personalized Health Forecasting-An Intelligent Digital Twin Model for Diet-Influenced Biomarker Trajectories (Student Abstract)

Ashikur Nobel[1], **Jacob Matos**[2], **Honggang Wang**[1], **Hua Fang**[1]

[1]Yeshiva University, New York, NY 10016 USA
[2]University of Massachusetts Dartmouth, North Dartmouth, MA 02747 USA

Abstract

We present iDT-diet, an intelligent digital twin prototype designed to model the long-term influence of diet quality on health biomarkers and chronic conditions. The system integrates three novel components: (i) a random forest learning model enhanced with Choquet LASSO feature selection for capturing complex, nonlinear interactions in temporal health data; (ii) a translation module that converts predictive outputs into natural language narratives of physical and biomarker states; and (iii) a generative 3D visualization engine that produces dynamic, personalized digital twins reflecting evolving health trajectories. This integration uniquely links advanced machine learning, interpretable communication, and immersive visualization within a single framework. While the current implementation focuses on retrospective digital twin generation, the system architecture supports real-time data integration, enabling continuous monitoring, predictive simulation, and personalized recommendation delivery for diet and lifestyle management.

Introduction

In this work, we introduce iDT-diet, an intelligent digital twin prototype designed to model and visualize the long-term influence of diet quality on health outcomes and chronic conditions. The system integrates three tightly coupled components: (i) a Random Forest model enhanced with Choquet LASSO feature selection for capturing nonlinear, high-dimensional interactions in temporal health data; (ii) a translation module that converts predictive outcomes into natural language descriptions of biomarkers and physical states; and (iii) a 3D visualization engine that generates dynamic, personalized avatars to represent evolving health trajectories.

Our approach builds on established evidence linking diet quality, as measured by indices such as the Alternate Healthy Eating Index (AHEI-2010), with biomarkers and chronic diseases including obesity, diabetes, and cardiovascular conditions (Fallaize et al. 2018; Xu et al. 2020; Lynch et al. 2024). Traditional feature selection methods often fail to capture higher-order interactions in health data (Iranzad and Liu 2024), while fuzzy measures such as the Choquet integral provide a systematic but computationally intensive alternative (Murofushi, Sugeno et al. 2000; Sugeno 1974). By integrating Choquet-based modeling with LASSO regularization, iDT-diet achieves efficient feature and interaction selection, improving predictive performance while maintaining interpretability compared to similar works (Wang et al. 2007; Bresson et al. 2020; Fang et al. 2010).

Digital Twin visualization technologies have shown great promise in healthcare. However, most current implementations remain static or organ-specific (Erol, Mendi, and Doğan 2020). For example, (Viola et al. 2023) modeled cardiac geometry for real-time 3D heart visualization, enhancing doctor-patient communication. Similarly, (Gkouskou et al. 2020) demonstrated that patient-specific VR consultations improved understanding and satisfaction in metabolic and gut-microbial nutrition analysis. In contrast, iDT-diet extends beyond these organ- or system-focused applications and learns about patients from longitudinal dietary behaviors and their relationships with physical biomarkers and chronic disease progression. It provides interpretable, visually guided feedback to support understanding and decision-making.

iDT-diet: Intelligent Digital Twin System Prototype

We propose a digital twin framework for diet and biomarker-based prediction. First, Choquet LASSO Regression is applied to select features and random forest to complete prediction collectively named Choquet-LASSO-RF.

Let $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ be the feature vector.

$$C_v(f(x)) = \sum_{i=1}^{n} \left(a_{(i)}x_{(i)} - a_{(i-1)}x_{(i-1)}\right) \cdot v(A_{(i)}), \quad (1)$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ with respective coefficient $a_{(i)}$ and cumulative set $A_{(i)} = \{(i), (i+1), \ldots, (n)\}$. The model output is defined as $\hat{y} = C_v(f(x))$

For regression tasks, the learning objective minimizes mean squared error (MSE) with LASSO regularization on the linear coefficients:

$$\mathcal{L}_{\text{total}} = \frac{1}{m}\sum_{k=1}^{m}\left(y^{(k)} - \hat{y}^{(k)}\right)^2 + \lambda_{\text{LASSO}}\sum_{i=1}^{n}|a_i|, \quad (2)$$

where λ_{LASSO} is a tuned hyper-parameter for lasso regularization.

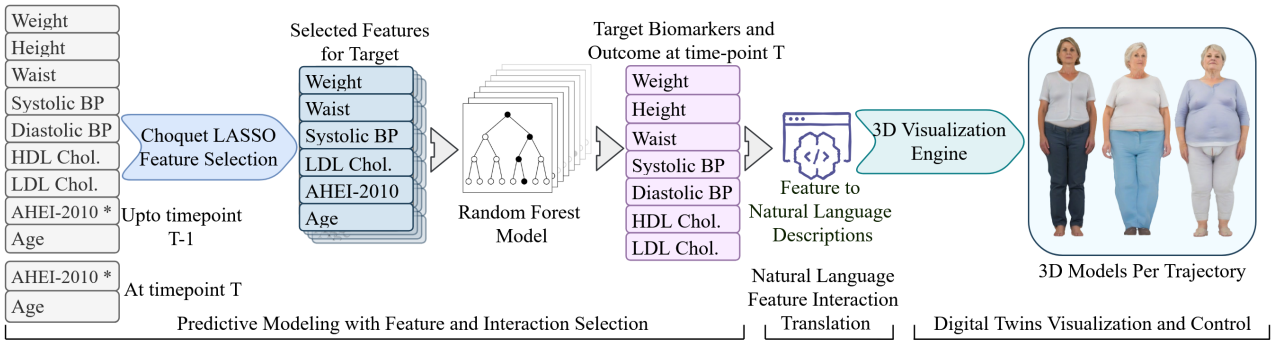

Figure 1: Overview of iDT-Diet: Intelligent Digital Twin Prototype Diet-Influenced Biomarker Trajectories

Target	MSE				MAE			
	Choquet-LASSO-RF	XGBoost	LightGBM	CatBoost	Choquet-LASSO-RF	XGBoost	LightGBM	CatBoost
Height	35.72	56.25	**34.09**	41.62	**1.48**	1.80	1.79	2.98
Weight	**69.69**	87.88	85.33	127.41	**4.46**	5.09	4.98	6.97
Waist	**90.58**	100.19	101.76	121.97	**5.53**	5.90	5.98	7.09
Syst. BP	**208.75**	258.96	233.22	240.21	**10.04**	11.65	10.83	11.20
Dias. BP	**81.26**	86.46	84.13	86.34	**6.90**	7.32	7.10	7.31
HDL Chol.	87.16	96.09	**84.32**	123.15	7.21	7.49	**7.08**	8.48
LDL Chol.	**994.98**	1166.77	1016.70	1129.68	24.75	26.48	**24.61**	26.43

Table 1: Model performance evaluation of Choquet-LASSO RF across biomarkers at 6 years from baseline

Pairwise interactions using Sugeno λ-measure (Murofushi, Sugeno et al. 2000) are computed as:

$$v(\{A_{(i)}, j\}) = v(\{A_{(i)}\}) + v(\{j\}) + I_{i,j}, \qquad (3)$$

where λ_{sugeno} is a hyper-parameter controlling the interaction strength and $I_{i,j} = \lambda_{\text{sugeno}} \cdot v(\{A_{(i)}\}) \cdot v(\{j\})$ is the pairwise feature interaction value from which we pick the top-k features for a random forest pipeline: with biomarker prediction at each timepoint using data upto the target timepoint and diabetes outcome probabilities from the base biomarkers and predicted biomarkers selected over all times.

Natural Language Translation Predicted biomarkers are transformed into descriptive sentences to enable natural language interpretation. The structured template used is:

> *A [age] year old [race] [gender] with a(n) [height] height, a(n) [weight] build, and a(n) [waist] waist.*

Subject details such as age, race, and gender are directly inserted, while height, weight, and waist circumference are categorized into descriptive buckets derived from national averages (Fryar et al. 2021; NHLBI-Obesity 2000).

3D Visualization Engine The visualization system operates in two phases. First, generative AI converts the descriptive sentences into 3D digital twin models, with multiple generations produced to ensure reliability where the models were generated using Meshy.ai (Meshy 2025) from predicted biomarkers. Second, the models are exported as `.fbx` files and imported in a visualization environment, where lighting and scene adjustments enhance clarity with tests in Unity Graphical Engine(Unity 2025).

iDT-Diet Numerical Analysis

iDT-Diet was evaluated on a selected longitudinal dataset of 402 participants with repeatedly measured biomarkers and diet quality measurements at baseline, year 1, year 3, and year 6. Inputs included all biomarkers and diet quality components till target timepoints, while targets were biomarkers and diabetes outcomes subsequent to the input. Data were normalized, with 80% for training and 20% for testing and Choquet-LASSO-RF applied from feature selection to target prediction. Models per target were tuned and trained with cross-validation, and results averaged across 5 random seeds to ensure replicability while retaining only 50-70% of candidate features based on target. Performance was assessed using MSE and MAE for biomarker regression, and accuracy for diabetes classification. Compared to state-of-the-art models LightGBM (Ke et al. 2017), CatBoost (Prokhorenkova et al. 2018), XGBoost (Chen and Guestrin 2016)), iDT-Diet generally achieved lower prediction errors across multiple targets(Table-1), with diabetes prediction accuracy around 97%. Finally, we present defined 3D visualizations to communicate physical progression based on these predictions over the years.

Discussion and Conclusion

The iDT-Diet framework was developed to provide an interpretable, data-driven approach for linking longitudinal diet quality with biomarkers with better interpretability and trajectory visualization. These capabilities support personalized nutrition strategies and public health planning, with potential for future work on expanding population diversity and integrating additional lifestyle and chronic condition interactions.

Acknowledgements

This research was partly supported by NIH 1R56DK114514-01A1 and NIH R01DK129432 to Dr. Fang, and partly supported by NSF-IIS 2140729 to Drs. Fang and Wang.

Ethics Statement

This study was conducted in accordance with U.S. federal regulations for the protection of human subjects (45 CFR 46, the Common Rule) and NIH policies on ethical research. All human data used were either fully de-identified or collected with informed consent from participants, and the study protocol was reviewed and approved by the Institutional Review Board of University of Massachusetts Dartmouth. Data handling, storage, and analysis complied with federal and NIH guidelines to ensure participant privacy, confidentiality, and responsible use. The development and evaluation of the diet-quality based digital twin models were performed with attention to minimizing risk, avoiding harm, and ensuring that model outputs are used ethically in research and potential practical applications.

References

Bresson, R.; Cohen, J.; Hüllermeier, E.; Labreuche, C.; and Sebag, M. 2020. Neural representation and learning of hierarchical 2-additive Choquet integrals. In *IJCAI-PRICAI-20-Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*, 1984–1991. International Joint Conferences on Artificial Intelligence Organization.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Erol, T.; Mendi, A. F.; and Doğan, D. 2020. The digital twin revolution in healthcare. In *2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, 1–7. IEEE.

Fallaize, R.; Livingstone, K. M.; Celis-Morales, C.; Macready, A. L.; San-Cristobal, R.; Navas-Carretero, S.; Marsaux, C. F.; O'Donovan, C. B.; Kolossa, S.; Moschonis, G.; et al. 2018. Association between Diet-Quality Scores, Adiposity, Total Cholesterol and Markers of Nutritional Status in European Adults: Findings from the Food4Me Study. *Nutrients*, 10(1): 49.

Fang, H.; Rizzo, M. L.; Wang, H.; Espy, K. A.; and Wang, Z. 2010. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. *Pattern recognition*, 43(4): 1393–1401.

Fryar, C.; Carroll, M.; Gu, Q.; Afful, J.; and Ogden, C. 2021. Anthropometric Reference Data for Children and Adults: United States, 2015-2018. *Vital & health statistics. Series 3, Analytical and epidemiological studies*, 1–44.

Gkouskou, K.; Vlastos, I.; Karkalousos, P.; Chaniotis, D.; Sanoudou, D.; and Eliopoulos, A. G. 2020. The "virtual digital twins" concept in precision nutrition. *Advances in Nutrition*, 11(6): 1405–1413.

Iranzad, R.; and Liu, X. 2024. A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*, 1–15.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Lynch, S.; Killeen, S. L.; O'Brien, E.; Mullane, K.; Hokey, E.; Mealy, G.; and McAuliffe, F. M. 2024. Diet quality and blood pressure among pregnant women with overweight or obesity: A secondary analysis of two randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, 103(6): 1073–1082.

Meshy. 2025. Meshy AI - The #1 AI 3D Model Generator for Creators. Accessed on July 29, 2025.

Murofushi, T.; Sugeno, M.; et al. 2000. Fuzzy measures and fuzzy integrals. *Fuzzy measures and integrals: theory and applications*, 2000: 3–41.

NHLBI-Obesity. 2000. *The practical guide: identification, evaluation, and treatment of overweight and obesity in adults*. National Institutes of Health, National Heart, Lung, and Blood Institute ....

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Sugeno, M. 1974. Theory of fuzzy integrals and its applications. *Doctoral Thesis, Tokyo Institute of Technology*.

Unity. 2025. Unity Real-Time Development Platform 3D, 2D, VR & AR Engine. Accessed on July 29, 2025.

Viola, F.; Del Corso, G.; De Paulis, R.; and Verzicco, R. 2023. GPU accelerated digital twins of the human heart open new routes for cardiovascular research. *Scientific reports*, 13(1): 8230.

Wang, H.; Fang, H.; Sharif, H.; and Wang, Z. 2007. Nonlinear classification by genetic algorithm with signed fuzzy measure. In *2007 IEEE International Fuzzy Systems Conference*, 1–6. IEEE.

Xu, Z.; Steffen, L. M.; Selvin, E.; and Rebholz, C. M. 2020. Diet quality, change in diet quality and risk of incident CVD and diabetes. *Public health nutrition*, 23(2): 329–338.