Explainable AI to predict a complex multifactorial outcome, childhood obesity: Application to clinical epidemiology

Fuling Chen [*], Phillip E Melton [†], Kevin Vinsen [‡], Trevor Mori [§], Lawrence Beilin [¶], Rae-Chi Huang [‖]

July 9, 2025

Abstract

Childhood obesity is a complex and multifactorial condition influenced by genetic predisposition, environmental exposures, and early-life anthropometrics. While machine learning (ML) models have shown promise in predicting obesity trajectories, their adoption in clinical settings is limited due to a lack of interpretability. In this study, we apply Kolmogorov-Arnold Networks (KAN), an explainable deep learning framework, to predict body mass index (BMI) as an obesity risk indicator at 8 years old, using early-life epidemiological factors and polygenic risk scores (PGS) from the Raine Study Gen2 cohort. KAN's formularization mechanism enables mathematical representation of predictive relationships, allowing for improved model transparency. Our results demonstrate that KAN outperforms traditional

[*]Fuling Chen is with the International Centre for Radio Astronomy Research at University of Western Australia, 35 Stirling Hwy, WA, 6009 Australia (e-mail: fuling.chen@uwa.edu.au).

[†]Phillip E Melton is with Menzies Institute for Medical Research at University of Tasmania, Churchill Ave, Hobart, TAS, 7005 Australia (e-mail: phillip.melton@utas.edu.au).

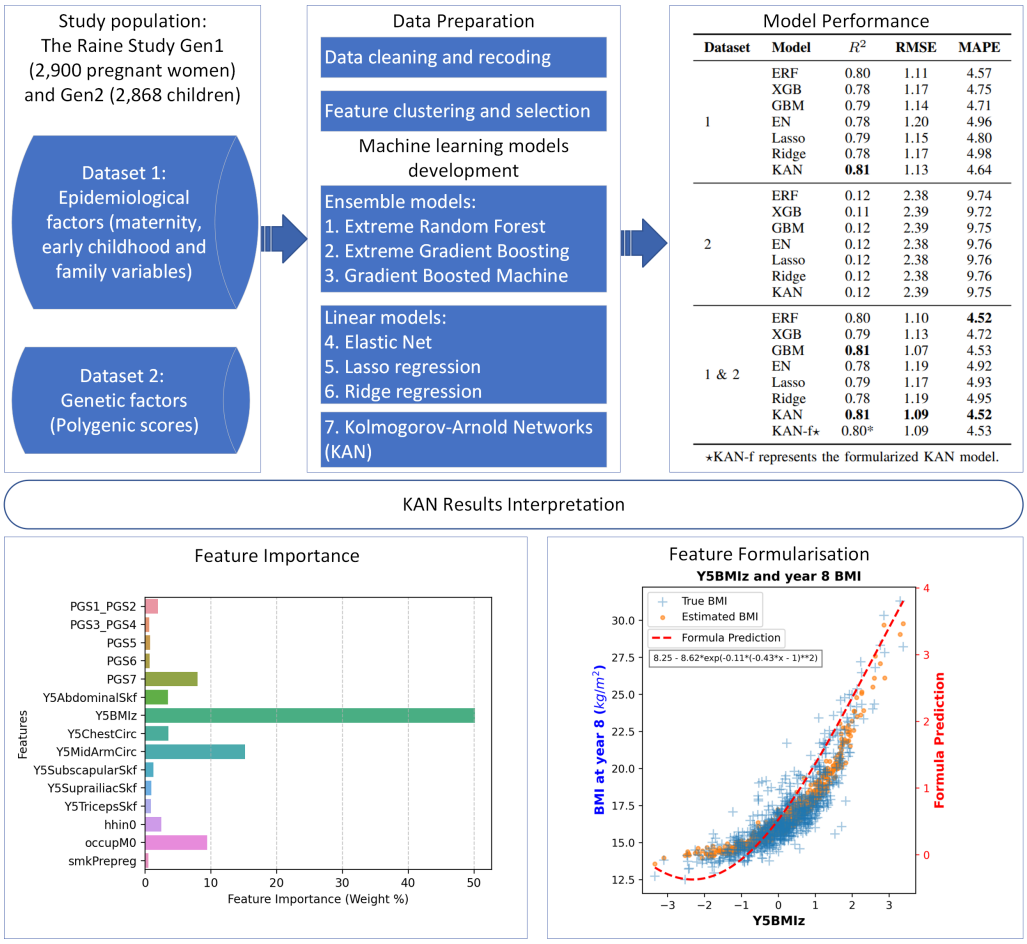
[‡]Kevin Vinsen is with the International Centre for Radio Astronomy Research at University of Western Australia, 35 Stirling Hwy, WA, 6009 Australia (e-mail: kevin.vinsen@uwa.edu.au).

[§]Trevor Mori is with the Medical School at the University of Western Australia, 35 Stirling Hwy, WA, 6009 Australia (e-mail: trevor.mori@uwa.edu.au)

[¶]Lawrence Beilin is with the Medical School at the University of Western Australia, 35 Stirling Hwy, WA, 6009 Australia (e-mail: lawrie.beilin@emeriti.uwa.edu.au)

[‖]Rae-chi Huang is with the Nutrition & Health Innovation Research Institute at Edith Cowan University, 2 Bradford Street, Mount Lawley, WA, 6050 Australia (e-mail: r.huang@ecu.edu.au).

ML models—including Extreme Random Forest, Extreme Gradient Boost, and Lasso regression—achieving the highest R^2 score (0.81) when integrating polygenic scores and epidemiological data. Feature importance analysis identifies Body Mass Index (BMI) z-score at Year 5, triceps and suprailiac skinfold thickness, and polygenic scores as key contributors. These findings highlight the potential of explainable deep learning in personalized obesity prevention, providing a transparent and interpretable AI-driven tool for early intervention strategies.



Dataset	Model	R^2	RMSE	MAPE
1	ERF	0.80	1.11	4.57
	XGB	0.78	1.17	4.75
	GBM	0.79	1.14	4.71
	EN	0.78	1.20	4.96
	Lasso	0.79	1.15	4.80
	Ridge	0.78	1.17	4.98
	KAN	**0.81**	1.13	4.64
2	ERF	0.12	2.38	9.74
	XGB	0.11	2.39	9.72
	GBM	0.12	2.39	9.75
	EN	0.12	2.38	9.76
	Lasso	0.12	2.38	9.76
	Ridge	0.12	2.38	9.76
	KAN	0.12	2.39	9.75
1 & 2	ERF	0.80	1.10	**4.52**
	XGB	0.79	1.13	4.72
	GBM	**0.81**	1.07	4.53
	EN	0.78	1.19	4.92
	Lasso	0.79	1.17	4.93
	Ridge	0.78	1.19	4.95
	KAN	**0.81**	**1.09**	**4.52**
	KAN-f⋆	0.80*	1.09	4.53

⋆KAN-f represents the formularized KAN model.

Figure 1: KAN model predicts childhood obesity (BMI at age 8), showcasing key features, top performance, and accurate formularised results with epidemiological and genetic factors.

Keywords: *Machine learning, Explainable AI, Obesity, Epidemiology, Body Mass Index*

1 Introduction

During the past decade, artificial intelligence, particularly machine learning (ML) models, has played a transformative role in making sense of complex datasets. Obesity, a multifaceted disease driven by both environmental and genetic factors, has been widely studied using body mass index (BMI) as a key surrogate measure of health risk at the population level [1]. Reviews focusing on ML applications in this field [2, 3] highlight that numerous models have been developed to predict BMI. Commonly utilized models include Lasso [4], Logistic Regression (LR) [5], Random Forests (RF) [6], Support Vector Machines (SVM) [7], Extreme Gradient Boosting (XGB) [8], Artificial Neural Networks (ANN) [9, 10], and Long Short-Term Memory network [11].

Both the aforementioned reviews [2, 3] underscore the advantages of ML models, particularly their superior predictive accuracy—ensemble methods and deep learning models often outperform traditional statistical techniques. ML models also capture complexity and nonlinear relationships among obesity-related variables. Nonetheless, significant challenges remain. Many ML models, especially deep learning, are criticized for their lack of interpretability, often functioning as "black boxes." SHapley Additive exPlanations (SHAP) has emerged as a gold-standard tool for explaining ML models lacking intrinsic explainability [12]. Yet its variants can produce inconsistent feature importance values and struggles to clarify decision-making in deep models, questioning the reliability of its explanations [13]. Moreover, high-performing models frequently require large, high-quality datasets, which can limit their application in settings where such data is unavailable.

Kolmogorov-Arnold Networks (KAN) represent a recently developed deep learning model [14]. By eliminating linear weights entirely, replacing them with univariate functions parametrized as splines, this innovative architecture is believed to enhance the accuracy of the model compared to MultiLayer Perceptrons (MLP), claiming superior performance, requiring smaller and shallower architectures to achieve faster neural scaling laws, thus improving computational efficiency.

Critically, KAN provides intuitive visualizations, formularization and explicitly presents the weights of variables across layers, offering insights into the model's decision-making process and formularization [14]. These features position KAN as a robust and explainable alternative to traditional deep learning approaches.

Increased interpretability is highly valued in scenarios requiring transparency and explanation [15, 16]. This is particularly pertinent to clinical scenarios. When clinical decisions are being made, potentially affecting an individual's course of care, understanding of the logic is reassuring, if not

mandatory, rather than devolving decision making to a "black box".

We aimed to demonstrate the application of KAN, which allows artificial intelligence to be applied to complex biological and environmental data with transparency of the processes to attain the solution. This approach has been applied to the clinical challenge of predicting obesity trajectories in early life.

The transition from childhood to adulthood represents a critical period for obesity prevention. Childhood obesity significantly increases the likelihood of remaining obese into adulthood. Longitudinal studies indicate that approximately 55% of obese children remain obese in adolescence, 80% of obese adolescents continue to be obese as adults, and 70% maintain obesity over age 30 [17]. These findings underscore the long-term health implications of childhood obesity and the importance of early intervention.

Unlike previous studies, which often focus on isolated factors or lack comprehensive interpretability, we proposed a novel ML model that balances predictive performance with robust interpretability.

The aim of this study was to compare the application of explainable AI (KAN) with traditional supervised machine learning to predict the future risk of obesity at 8 years of age. The prediction factors were all collected prior to school entry age (including a broad range of perinatal and early life factors, and relatively easily measured biometrics, such as anthropometrics (skin folds) and genetic profile).

2 Materials and methods

2.1 Study population

This study utilized data from the Raine Study Gen2[1]. The Raine Study is a large, well-established longitudinal cohort designed to track health and developmental outcomes from pregnancy into adulthood [19]. The Raine Study initially recruited 2,900 pregnant women (Generation 1, Gen1) and followed 2,868 children (Generation 2, Gen2). Recruitment was primarily conducted at King Edward Memorial Hospital in Perth, Western Australia, making the study population representative of the regional population. Dataset 1 collection time points were at 18, 24, 34 weeks' gestation during pregnancy, birth, 1, and 5 years old. The outcome measure data collection occurred at 8 years old.

[1]The variables in the Rain Study are mostly harmonized with the LifeCycle Project-EU Child Cohort Network [18]

2.2 Dataset 1 – Epidemiological factors (maternity, early childhood and family variables)

In this study, Dataset 1 was created to explore the influence of maternal, early childhood, and family factors on BMI development over time.

This dataset initially included 201 raw variables (Step 1) spanning various domains, such as maternal characteristics (e.g., weight gain and smoking status during pregnancy), early childhood physical measures (e.g., weight and height at birth, year 1 and year 5, breastfeeding duration and child care status), and family demographics (e.g. parental education, occupation, and household income). The details of the variables are provided in the Supplementary Material.

The processing of dataset 1 involved obtaining the raw data, excluding the anomalies, data cleansing, encoding variables, feature selection and clustering. Figure 2 depicts the data flow and the corresponding sample size and number of variables at each step.

Data preparation involved rigorous cleaning and filtering stages to enhance data quality and minimize bias. We began by excluding cases with significant congenital anomalies or multiple births to ensure a focus on children with typical developmental trajectories, which helps in drawing more generalizable conclusions regarding BMI development (Step 2).

Subsequent filtering focused on addressing the high level of missing data among certain variables, a common challenge in cohort studies. To address this, we removed variables where more than 50% of the values were missing (Step 3), which aligns with best practices for ensuring sufficient data density [20]. Then, we removed cases without target BMI data (Step 4) and retained cases with key BMI-related data, such as weight and height at birth, and ages at one and five years (Step 5). These physical measurements were essential for calculating BMI z-scores, which provide a standardized approach for assessing BMI changes relative to age and gender. To ensure accuracy across different ages, we used the CDC Extended BMI-for-age growth charts for very high BMI values at age five, given the chart's improved reliability in capturing elevated BMIs above the 95th percentile [21]. For younger age points (birth and year one), we applied the World Health Organization's growth standards, which are widely recognized as appropriate for early developmental stages [22] (Step 6).

Dataset 1 also included categorical variables for parental race and ethnicity, represented using one-hot encoding (Step 6), which converts each category into a binary vector for compatibility with machine learning algorithms. This encoding captured demographic diversity with five categories—Caucasian, Aboriginal, Polynesian, Asian, and Other—enabling a

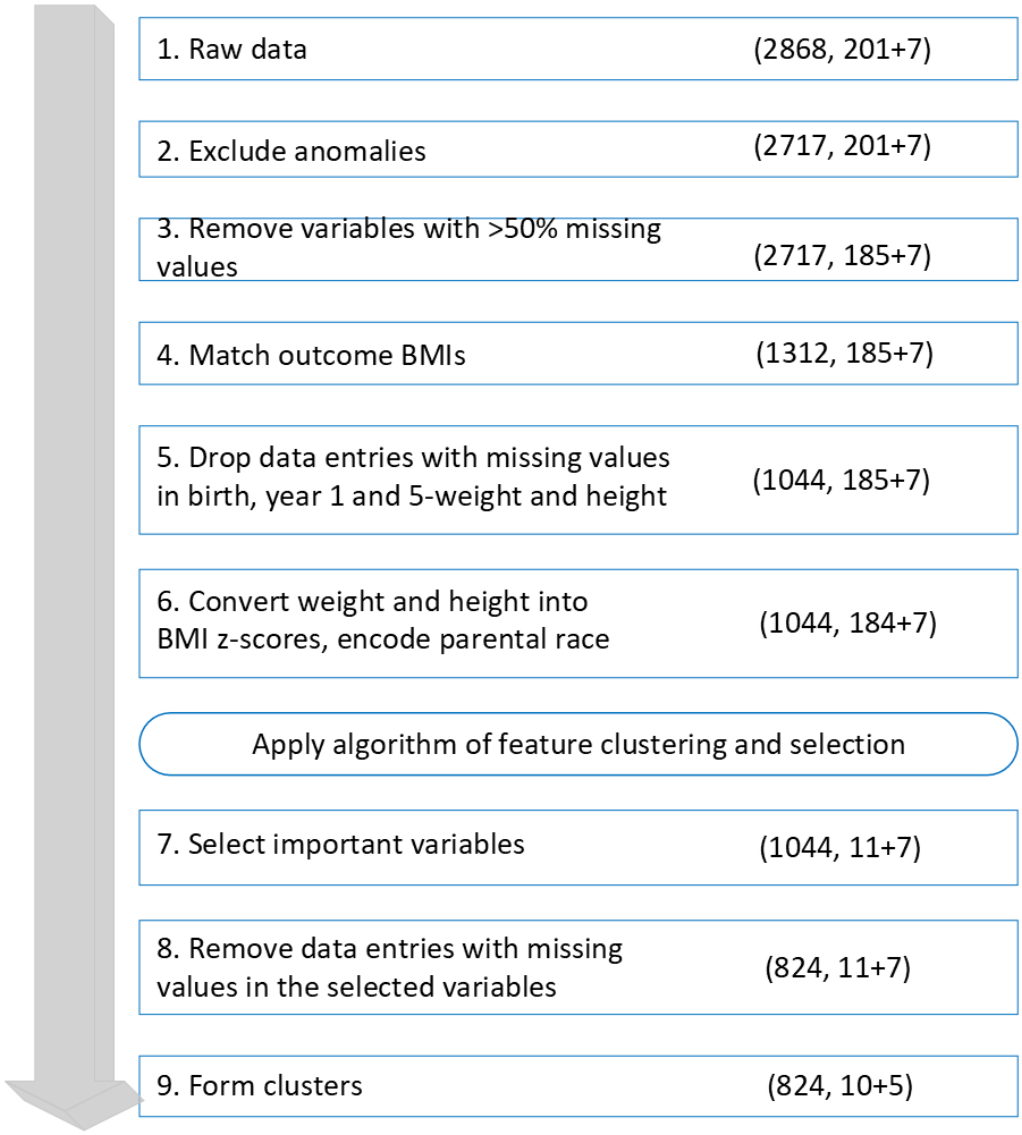


Figure 2: Datasets preprocessing workflow. The form of $(i, j + k)$ represents the number of samples (i), the number of dataset 1 variables (j), and the number of polygenic scores of dataset 2 (k).

more nuanced understanding of BMI development across ethnic backgrounds, as genetic and cultural factors may influence childhood BMI trajectories [23, 24].

Despite these steps, several variables were found to be strongly correlated, and some variables continued to exhibit missing data. To form independent

variables and maximize data retention while minimizing redundancy, we applied a correlation-based clustering algorithm, which groups highly correlated variables, facilitating the identification of core features that contribute independently to BMI prediction. Clustering was followed by importance-based variable selection, allowing us to prioritize variables with the greatest predictive value for BMI outcomes (more details in Appendix A). This combination of clustering and selection has been shown to enhance model performance by retaining relevant variables while reducing overfitting due to extraneous data [25]. We obtained the cluster patterns and selected the inclusive variables (Step 7). Finally, we excluded cases without data on selected key variables (Step 8) and formed clusters (Step 9), resulting in a dataset optimized for our subsequent analyses.

After preprocessing, the final dataset includes 824 samples and 10 variables. The baseline characteristics and the descriptive statistics of the selected variables are shown in Table 1. Among these variables, smoking before pregnancy (*smkPrepreg*) and cigarettes/day before pregnancy (*cigPrepreg*) were combined into a single factor, hereafter referred to as *smkPrepreg*.

2.3 Dataset 2 - Genetic factors

In addition to Dataset 1, we incorporated seven published BMI polygenic scores (PGS), as outlined in Table 2. These scores were generated using the pgsc_calc pipeline [26], which computes PGS by matching genetic variants from scoring files published in the Polygenic Score Catalog with those in the target dataset.

Preprocessing occurred using Dataset 1, Dataset 2, and a combination of both datasets for predicting the outcome of BMI at 17 years old. When merging Dataset 1 and Dataset 2, the seven PGS were treated as extra variables and integrated with those in Dataset 1. After clustering, these seven PGS were grouped into 5 independent variables with PGS1 and PGS2 as one cluster, and PGS3 and PGS4 as one cluster. The combined data workflow is shown in Figure 2, with the final number of clusters being 15. The metadata of the seven PGS are shown in Table 1.

2.4 Outcome measure

For this study, weight and height measurements were collected at 8 years old. BMI was calculated by weight divided by height squared, serving as the primary outcome variable. The year 8 BMI is presented in Table 1.

Variable Name	Description	Data Type	Units	Mean ± SD	Min	Max
		Baseline characteristics				
sex	Child sex	Binary	N/A	443 males, 381 females		
raceM	Race of mother	Categorical	N/A	792 Caucasian, 32 others		
raceF	Race of father	Categorical	N/A	805 Caucasian, 19 others		
		Critical variables				
hhin0	Household income in the first year [a]	Categorical	AUD	1.99 ± 1.02	1	4
occupM0	Mother occupation category at age <1 year [b]	Categorical	N/A	5.76 ± 2.92	0	9
smkPrepreg	Smoking before pregnancy	Binary	N/A	0.54 ± 0.50	0	1
cigPrepreg	Cigarettes/day before pregnancy	Numeric	Count	0.86 ± 0.87	0	2
Y5ChestCirc	Chest circumference at year 5	Numeric	cm	58.32 ± 3.81	50.0	78.9
Y5MidArmCirc	Mid-arm circumference at year 5	Numeric	cm	17.93 ± 1.68	12.9	26.8
Y5TricepsSkf	Triceps skinfold at year 5	Numeric	mm	10.09 ± 2.80	4.1	24.5
Y5SubscapularSkf	Subscapular skinfold at year 5	Numeric	mm	6.01 ± 2.35	3.0	22.0
Y5SuprailiacSkf	Suprailiac skinfold at year 5	Numeric	mm	8.70 ± 4.02	3.5	29.4
Y5AbdominalSkf	Abdominal skinfold at year 5	Numeric	mm	7.54 ± 3.61	3.0	29.6
Y5BMIz	BMI z-score at year 5	Numeric	N/A	0.18 ± 0.95	-3.34	3.38
PGS1	Polygenic Score PGS002313	Numeric	N/A	-0.96 ± 0.35	-2.10	0.16
PGS2	Polygenic Score PGS002161	Numeric	N/A	0.02 ± 0.06	-0.20	0.23
PGS3	Polygenic Score PGS000027	Numeric	N/A	38.84 ± 0.15	38.42	39.34
PGS4	Polygenic Score PGS004150	Numeric	N/A	-1.50 ± 1.62	-5.99	4.40
PGS5	Polygenic Score PGS003884	Numeric	N/A	0.63 ± 0.36	-0.40	1.68
PGS6	Polygenic Score PGS002853	Numeric	N/A	0.24 ± 0.31	-0.77	1.16
PGS7	Polygenic Score PGS000921	Numeric	N/A	-4.56 ± 0.49	-6.10	-3.08
Y8BMI	BMI at year 8	Numeric	kg/m^2	16.86 ± 2.47	12.51	31.31

[a] Household income: 1 = 4th quartile (highest), 2 = 3rd quartile, 3 = 2nd quartile, 4 = 1st quartile.
[b] Occupation categories: 1 = Legislators, senior officials, managers; 2 = Professionals; 3 = Technicians and associate professionals; 4 = Clerks; 5 = Service workers, shop, and market sales workers; 6 = Skilled agricultural and fishery workers; 7 = Craft and related trades workers; 8 = Plant and machine operators and assemblers; 9 = Elementary occupations; 0 = Armed forces occupations.

Table 1: Baseline characteristics and descriptive statistics of critical variables used in the study.

PGS ID	Reference	PGS Pub. ID	No. Variants	Development Method	Score Development		
					No. Sample	Ancestry	Cohort
PGS1	[27]	PGP000332	1,109,311	BOLT-LMM	NR	NR	NR
PGS2	[28]	PGP000263	990,022	LDpred2 (bigsnpr)	391,124	European	UKB
PGS3	[29]	PGP000017	2,100,302	LDpred	119,951	European	UKB
PGS4	[30]	PGP000517	1,039,042	UKBB-EUR.MultiPRS.CV	359,913	European	UKB
PGS5	[31]	PGP000501	1,067,771	PRS-CSx	NR	NR	NR
PGS6	[32]	PGP000393	7,446,664	DBSLMM	23,381	European	MGI
PGS7	[33]	PGP000243	1,947,711	LDpred	1,000	European (Danish)	Inter99

Table 2: Summary of polygenic scores (PGS) used in this study. "NR" indicates not reported.

2.5 Machine learning models

This study used six widely used machine learning models and the KAN [14] to analyse the data. The conventional models included three ensemble methods – Extreme Random Forest (ERF) [34], Extreme Gradient Boosting (XGB) [35], and Gradient Boosted Machine (GBM) [36] – and three linear models: Elastic Net (EN) [37], Lasso [38], and Ridge [39] regression. These models were selected for their strong performance in regression tasks and high interpretability, making them well-suited for understanding the relationships within the data. The conventional models were implemented using the Python package scikit-learn (version 1.0.1). At the same time, the KANs were adapted to the target dataset using source code from the creator's publicly available GitHub repository[2].

2.5.1 Ensemble models

The ensemble models employed (ERF, XGB and GBM) were selected for their ability to capture complex relationships in data, offering strong predictive performance and interpretability [40–42]. ERF extends the traditional Random Forest algorithm by building numerous unpruned decision trees on random subsets of data, with predictions made by averaging outputs across trees. This approach reduces variance and overfitting while effectively capturing nonlinear relationships with minimal parameter tuning [34].

XGB and GBM, both employing boosting techniques, construct trees sequentially, each tree addressing errors from its predecessors. XGB, an optimized gradient boosting method, is renowned for its speed, accuracy, and scalability. It incorporates features such as parallel computations, regularization to mitigate overfitting, and efficient handling of missing data [35]. Traditional GBM, while less computationally efficient, uses gradients to build trees iteratively and adjusts each tree's impact with a learning rate, ensuring precision in predictions [36]. Together, these models leverage complementary strengths, bolstering the robustness and interpretability of BMI predictions.

2.5.2 Linear models

The three linear models (EN, Lasso, and Ridge regression) were chosen for their effectiveness in handling multicollinearity and their ability to deliver interpretable results in regression tasks. Lasso regression employs an $L1$ penalty, which drives certain coefficients to zero, effectively performing feature selection by retaining only the most significant predictors. This approach

[2]https://github.com/KindXiaoming/pykan.git

simplifies the model and enhances interpretability by reducing the number of included variables [38].

Ridge regression, in contrast, applies an $L2$ penalty that shrinks all coefficients toward zero without eliminating them [39]. This regularization technique enhances model stability, particularly when predictors are highly correlated, but does not perform feature selection. The EN combines the strengths of Lasso and Ridge by incorporating both $L1$ and $L2$ penalties, striking a balance between feature selection and coefficient stability. This makes EN particularly effective in scenarios involving strong multicollinearity or when the number of predictors exceeds the number of observations [37]. Together, these models offer a robust and adaptable framework for managing complexity and interpretability in regression analysis.

2.5.3 Kolmogorov-Arnold Networks

The KAN, a deep learning model [14], is inspired by the Kolmogorov-Arnold Representation Theorem, which states that any multivariate continuous function $f(x_1, x_2, ..., x_n)$ can be decomposed into a sum of univariate functions:

$$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$

where $\phi_{q,p}(x_p)$ are learnable univariate function. Unlike traditional neural networks, KANs represent activation functions as spline-based transformations on edges rather than fixed functions on nodes. The general deep KAN model is defined as:

$$KAN(x) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)(x)$$

where each layer transforms inputs via learnable splines.

Each edge function in KANs is represented by a B-spline:

$$\phi(x) = w_b b(x) + w_s \sum_i c_i B_i(x)$$

where:

- $b(x) = silu(x) = \frac{x}{1+e^{-x}}$ is a residual basis function.
- $B_i(x)$ are B-spline basis functions with trainable coefficients c_i.
- w_b and w_s control activation magnitudes.

The tuning involves grid extension, which defines the B-spline resolution, and regularization of lambda, $L1$ and entropy penalties to induce sparsity. In the current study, we optimised the hyperparameters by searching grid extension between 3 and 5, lambda from 0.01, 0.05, 0.1, and 0.2, $L1$ from 1, 2, and 5, and entropy from 1, 2, and 5.

KANs use sparsification and pruning to reduce complexity and improve interpretability by removing the nodes that are less important. The importance score of a node at the l-th layer i-th node is computed as

$$I_{l,i} = \max_k |\phi_{l-1,i,k}|_1, O_{l,i} = \max_j |\phi_{l+1,j,i}|_1$$

where k and j are the indexes of the nodes at the layer below and above the current layer l. Nodes are pruned if:

$$I_{l,i} < \theta \text{ or } O_{l,i} < \theta$$

where θ is a predefined cut-off threshold. In this study, we pruned the model by setting the cut-off value as 0.01 for the node and 0.03 for the edge, which indicates that if the attribution score of a node or an edge is below the cut-off value, it was considered dead and was set to zero. This removes weak connections while maintaining model performance.

After pruning, we further symbolized the expressions by constraining edge functions to known forms from the function library, which includes most typically used functions such as linear, polynomial, reciprocal, root, exponential, logarithmic, trigonometric, hyperbolic, Gaussian, a constant zero function and other activation functions. A learned function $\varphi(x)$ can be fitted to a known symbolic function from this library via:

$$y \approx cf(ax + b) + d$$

where a, b, c and d are trainable affine parameters. The final symbolised formula is a compact mathematical expression containing all the input variables x_i.

In this study, we demonstrated the use of a shallow KAN model with a single hidden layer between the input and output layers. Consequently, the importance of each input variable is determined by multiplying the activation weights along the paths from the input to the output.

$$I_p = \prod_{l=1}^{L} \phi_{l,i,j}$$

where the importance score I_p reflects how strongly an input influences the output, making KANs more interpretable than MLPs. To facilitate interpretation, the final weights of each input variable were normalized to percentages, representing their relative contributions.

2.6 Model training and evaluation metrics

All seven models were deployed across Dataset 1, Dataset 2, and their combined datasets 1 and 2. A 5-fold cross-validation strategy with varying randomization was implemented to split the data into training and testing sets. Each model was trained on the training set and evaluated on the testing set, with predictions collected from all five testing folds for further analysis. This approach ensured robust evaluation by mitigating the risk of overfitting and providing reliable performance estimates.

For each model, hyperparameters, including the number of clusters and the r_{cutoff} value, were finely tuned. For formalization, we utilized the default formula library to derive formulas for calculating the final predictions. This approach assumes no prior knowledge of the input variables' distributions or their relationships to the final outcomes.

Model performance was evaluated using three key metrics: the coefficient of determination (R^2), root mean squared error (RMSE), and Mean Absolute Percentage Error (MAPE), to assess explained variance and prediction accuracy. The best-performing model and dataset were selected based on these metrics for subsequent results presentation and analysis.

- R^2 Score (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ is the mean of the actual values. An R^2 score closer to 1 indicates better model performance.

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Both RMSE and MAPE closer to 0 represent better performance.

The performance of the KAN models with formalization was compared to the original KAN models and seven conventional models by examining R^2 scores, RMSE and MAPE. Additionally, the estimates of the KAN models, both with and without formalization, were compared against the true BMI values. We further analysed key factors and their derived formulas identified by the KAN models —- such as year 5 BMI z-scores (Y5BMIz) and the PGS7 -— for their role in predicting the year 8 BMI.

3 Results

3.1 Models and datasets comparison

Table 3 presents the R^2, RMSE and MAPE values of seven models trained on Dataset 1, Dataset 2, and their combination. When comparing the impact of datasets on BMI estimation, the combined use of Dataset 1 (early life epidemiology) and Dataset 2 (genetics) yielded the best results, slightly better than using Dataset 1 (early life epidemiology) alone across the seven models. In contrast, using Dataset 2 (genetics) in isolation provided little value in predicting BMI.

The KAN model trained on the combined datasets yielded the best result. Therefore, we performed the following analysis based on the KAN model outputs.

3.2 Feature importance

The feature importance in the decision-making process of the KAN model was determined by multiplying the weights of each activation function from the leaf nodes (input variables) to the output (year 8 BMI value). Figure 3 shows the aggregated feature importance of the selected factors. The year 5 BMI z-score dominates the estimation of the year 8 BMI, followed by mid-arm circumference at year 5, the occupation of the mother in the first year, and the PGS7.

In this study, we pruned the model by setting the cut-off value as 0.01 for the node and 0.03 for the edge, as described in Section 2.5.3. Figure 4 displays the original and pruned tree plots of input variable importance for BMI predictions.

Table 3: Performance comparison of different models across datasets using R^2, RMSE, and MAPE.

Dataset	**Model**	R^2	**RMSE**	**MAPE**
1	ERF	0.80	1.11	4.57
	XGB	0.78	1.17	4.75
	GBM	0.79	1.14	4.71
	EN	0.78	1.20	4.96
	Lasso	0.79	1.15	4.80
	Ridge	0.78	1.17	4.98
	KAN	**0.81**	1.13	4.64
2	ERF	0.12	2.38	9.74
	XGB	0.11	2.39	9.72
	GBM	0.12	2.39	9.75
	EN	0.12	2.38	9.76
	Lasso	0.12	2.38	9.76
	Ridge	0.12	2.38	9.76
	KAN	0.12	2.39	9.75
1 & 2	ERF	0.80	1.10	**4.52**
	XGB	0.79	1.13	4.72
	GBM	**0.81**	1.07	4.53
	EN	0.78	1.19	4.92
	Lasso	0.79	1.17	4.93
	Ridge	0.78	1.19	4.95
	KAN	**0.81**	**1.09**	**4.52**
	KAN-f⋆	0.80*	1.09	4.53

⋆KAN-f represents the formularized KAN model.

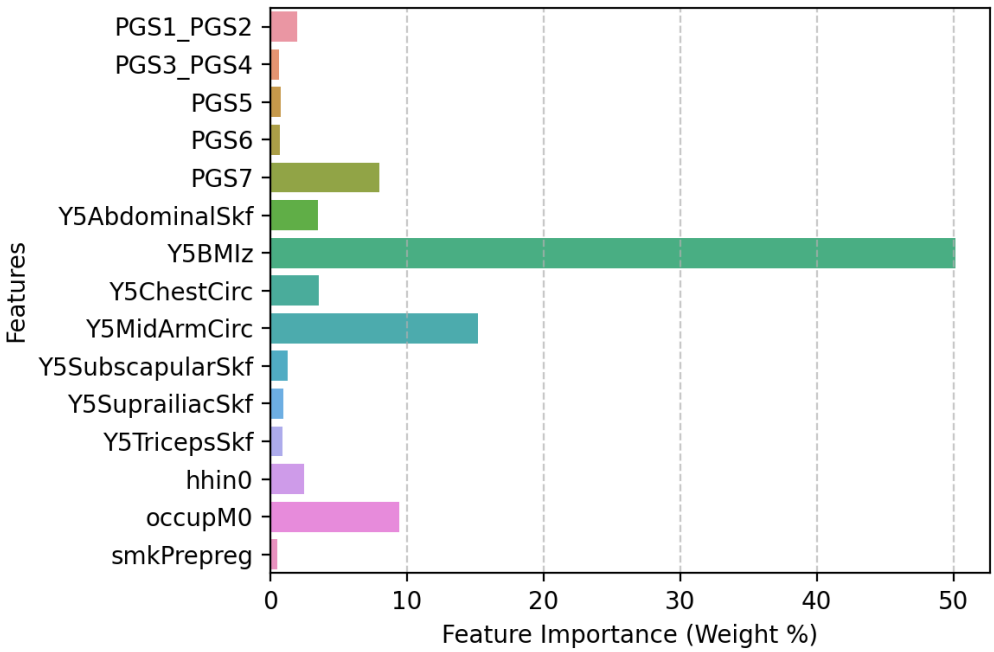


Figure 3: Feature importance

3.3 Formularization

To mathematically explain the decision-making process of the KAN models, we derived the formulas used by the models to estimate the target outcomes. Without making any prior assumptions about the functional form or distribution of the variables, we employed the activation functions recommended by the KAN model package. These functions were chosen as the optimal fit from the available function library, as described in section 2.5.3.

We fixed the activation functions and re-evaluated the formularized models on the testing data to obtain the estimated BMI values. Figure 5 compares the performance of the models before and after formularization against the ground truth. The results clearly show that the model maintains comparable performance after formularization. This consistency allowed us to further explore the mathematical functions associated with variables of interest and hypothesize their distributions in predicting target BMI values.

Diving deeper, we analyzed the formulas of the critical features and their relationships with the final outcome, the year 8 BMI. To illustrate the results, we selected Y5BMIz (the BMI z-score at year 5) and PGS7 to demonstrate the mathematical interpretation provided by KAN. The plots for the other important features (Y5MidArmCirc - mid-arm circumference at year

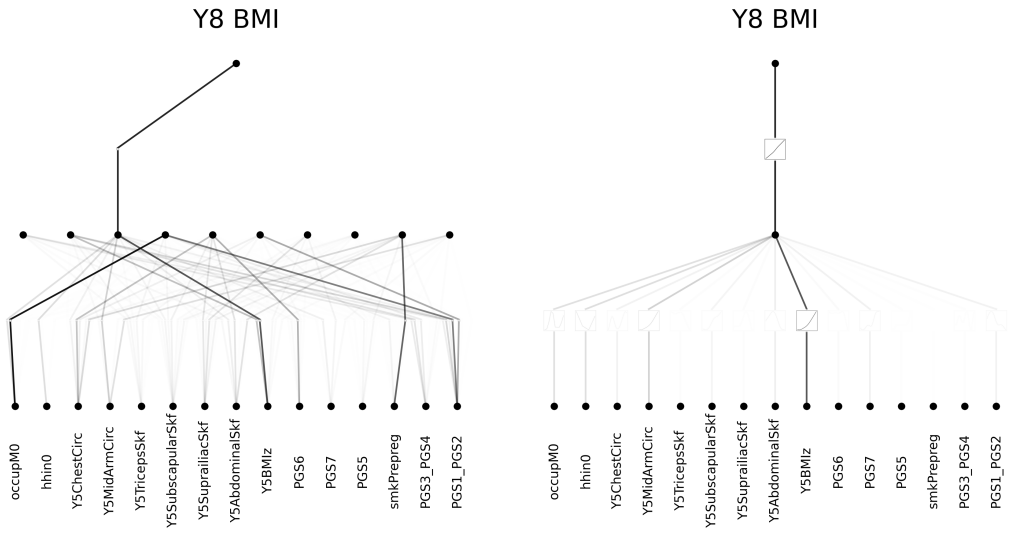


Figure 4: Original (left) and pruned (right) tree plots of KAN. The bottom black points represent the selected factors used in model development, the middle layer points represent the first-layer nodes of the KAN models, and the top point corresponds to the year 8 BMI value. The colour intensity of the connecting lines reflects the impact of the lower node on the upper node. For the pruned model, the nail plots in the middle of the connecting lines represent the activation functions used in the models.

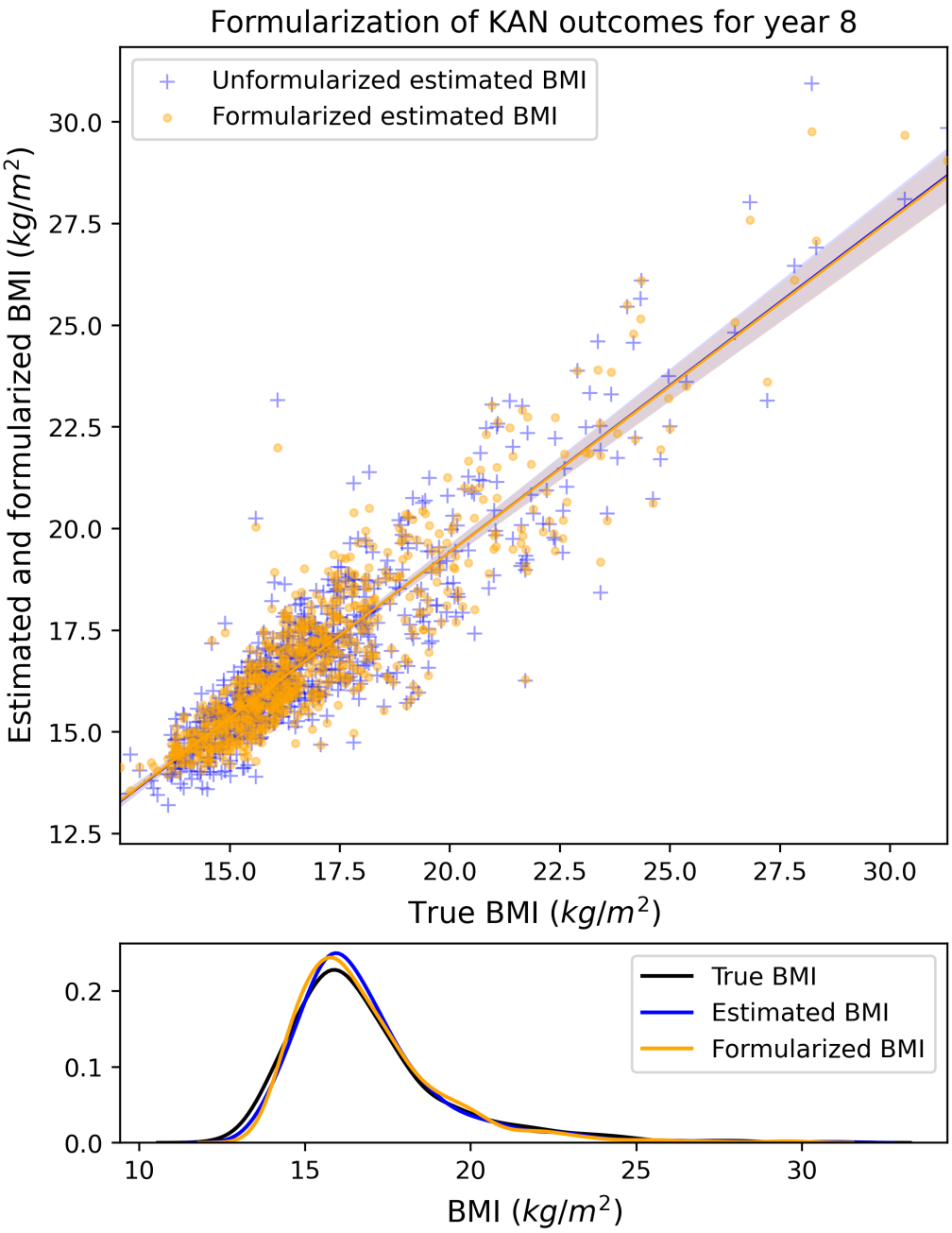


Figure 5: Scatter and distribution plots of the estimated and real BMI values of year 8, before and after formularization

5; and occupM0 - occupation of the mother at the birth of the child) are provided in the Supplementary Materials. Our findings reveal that the true and estimated BMI values exhibit an exponential or first-quadrant sine-like functional relationship with Y5BMIz, as shown in Fig. 6. In contrast, PGS7 does not follow a distinct functional pattern in relation to the target BMI, as observed in Fig. 7. However, a positive correlation is evident between PGS7 and BMI at 8 years old.

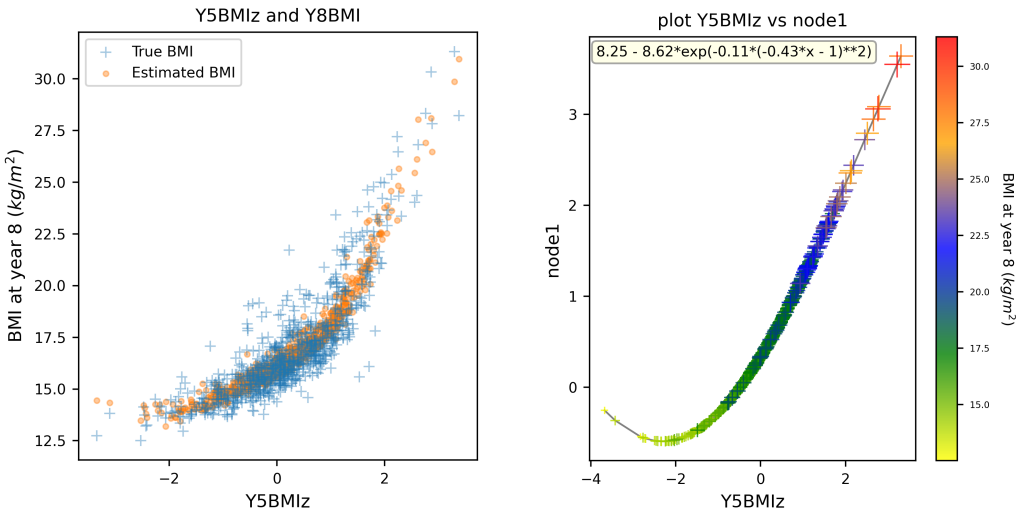


Figure 6: Left - the relationship between Y5BMIz on the x-axis and the true BMI (blue crosses) alongside the model's estimated BMI values (orange dots) at year 8 on the y-axis. Right - the KAN node function for Y5BMIz, where the data points (crosses) are colour-coded and size-scaled to represent BMI values. The functional form after standardization is provided in the top-left corner of the right panel.

Finally, we aggregated the formularization results, reversed the standardization of the input variables, and derived the overall features formula for predicting BMI at 8 years old, which is:

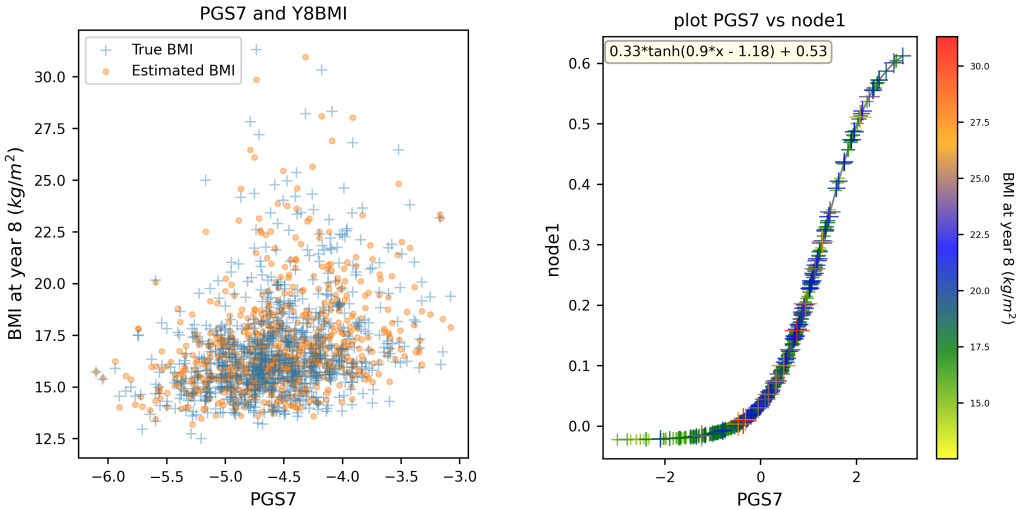


Figure 7: Left - the relationship between PGS7 on the x-axis and the true BMI (blue crosses) alongside the model's estimated BMI values (orange dots) at year 8 on the y-axis. Right - the KAN node function for PGS7, where the data points (crosses) are colour-coded and size-scaled to represent BMI values. The functional form after standardization is provided in the top-left corner of the right panel.

$$
\begin{aligned}
\text{Y8BMI} = 29.16 \times \tan^{-1} \Big[& \qquad\qquad\qquad\qquad\qquad\qquad 0.00022 \times (1 - 0.66 \times \text{sm} \\
& + 0.00493 \times \sin(17.29 \times \text{PGS6} - 9.44) \\
& - 0.01491 \times \sin(0.41 \times \text{occupM0} - 7.49) \\
& + 0.00302 \times \sin(0.08 \times \text{Y5SuprailiacSkf} - 3.04) \\
& + 0.00008 \times \tan(4.83 \times \text{hhin0} - 9.75) \\
& + 0.00004 \times \cosh(0.76 \times \text{PGS3_PGS4} - 0.11) \\
& + 0.06 \times \tanh(3.65 \times \text{PGS7} + 15.45) \\
& - 0.0165 \times \tanh(0.22 \times \text{Y5TricepsSkf} - 6.40) \\
& - 0.00118 \times |29.24 \times \text{PGS1_PGS2} + 9.42| \\
& + 0.15 \times \tan^{-1}(0.0095 \times \text{Y5AbdominalSkf} - 0.07) \\
& + 0.06 \times \tan^{-1}(0.0587 \times \text{Y5ChestCirc} - 5.90) \\
& - 1.68 \times \exp\left(-0.09 \times (-0.51 \times \text{Y5BMIz} - 1)^2\right) \\
& - 0.022 \times \exp\left(-0.28 \times (1 - 0.12 \times \text{Y5SubscapularSkf})^2\right) \\
& + 0.56 \times \exp\left(-7.35 \times (1 - 0.03 \times \text{Y5MidArmCirc})^2\right) \\
& + 0.0078 \times \exp\left(-113.53 \times (0.66 - \text{PGS5})^2\right) \\
& + 1.13 \Big] + 27.32
\end{aligned}
$$

These findings highlight that the PGS variables alone (Dataset 2) do not enable the model to achieve strong performance, as the model struggles to identify a clear functional relationship between the PGS variables and BMI. However, when combined with Dataset 1, the correlations between the PGS variables and BMI contribute positively to the training process, enhancing the model's performance by leveraging the complementary information from both datasets.

4 Discussion

This study demonstrates that the Kolmogorov-Arnold Network (KAN) model outperforms traditional machine learning models in predicting childhood obesity at age 8, achieving an R^2 of 0.81 and RMSE of 1.09 when combining epidemiological and genetic data. The KAN's interpretability revealed year 5 BMI z-score as the dominant predictor, followed by mid-arm circumference, maternal occupation, and polygenic score PGS7, offering actionable insights into obesity risk factors.

In biological research, conventional statistical models, such as linear regression or generalized linear models, have long been the cornerstone for analysing relationships between input variables and outcomes. These models are widely favoured for their simplicity, transparency, and ease of implementation. However, they inherently assume linearity in the relationships between input variables, among input variables and between inputs and the target outcome. While such assumptions work well for data with straightforward, linearly distributed relationships, they fail to handle complex datasets. For instance, input variables in biological studies often exhibit intricate interdependencies, nonlinear distributions, and hierarchical relationships, which are difficult to capture with conventional models. Additionally, relying on a single overarching model to fit all variables, irrespective of their distributional properties, may lead to suboptimal performance and inaccurate interpretations.

Deep learning models have emerged as powerful tools to address these limitations. Their capacity to model highly nonlinear and complex relationships has led to significant advancements in various fields, including biology and medicine. However, despite their impressive predictive capabilities, deep learning models often lack interpretability—a critical aspect in fields like biomedical research where understanding the underlying biological mechanisms is as important as achieving high predictive accuracy. The opaque nature of deep learning models has hindered their widespread adoption in scenarios requiring clear and actionable insights.

The KAN model presents a promising alternative to bridge this gap. KAN achieves a balance between predictive performance and interpretability, leveraging its innovative architecture to provide actionable insights. KAN incorporates a formularization mechanism enabled by fixing the activation functions, allowing it to model complex, nonlinear relationships while preserving interpretability. This capability is particularly advantageous for tasks involving heterogeneous input variables with varied distributions and intricate interrelationships.

The results of this study underscore the utility of KAN in BMI prediction. By outperforming traditional models and maintaining a high degree of interpretability, KAN demonstrates its potential as a robust tool for analysing complex biomedical data. The incorporation of formularization functions provides researchers and clinicians with not only accurate predictions but also valuable insights into the underlying factors driving those predictions.

It is important to acknowledge the limitations of the KAN model. The Kolmogorov-Arnold Representation theorem, on which KAN is based, assumes that any multivariate continuous function on a bounded domain can be expressed as a finite composition of continuous functions of a single vari-

able and the binary operation of addition. This assumption poses challenges when dealing with binary variables, categorical variables with limited bins, or entirely non-continuous discrete variables, as these may not fit naturally into the model's formularization framework. Future studies should explore strategies to address this limitation, such as integrating piecewise functions into the model architecture. In this study, we observed that certain variables, such as ethnicity and smoking status, are non-continuous. To enable the model to account for these factors, we applied one-hot encoding where necessary. For categorical variables with ordered bins, such as household income and education levels, we treated them as continuous variables, as the number of bins approximates continuity.

In summary, this study demonstrates the potential of explainable advanced machine learning models, such as KAN, to address and explain the complexities of BMI prediction. By providing both high predictive accuracy and interpretability, KAN offers a potential tool for utilising the multifaceted influences on BMI in real-world predictions.

A Algorithm of variables clustering and selection

To optimize data cleansing while preserving as many entries as possible, we developed a variable clustering and selection algorithm, applied after Step 6 as described in Section 2.2 and Figure 2: "Convert weight and height into BMI z-scores, encode parental race". This algorithm consists of three main stages: Step A ("Deep Clean"), Step B ("Cluster Variables"), and Step C ("Select Important Clusters"). Detailed descriptions of these steps are provided in Figure 8. The input data for this algorithm corresponds to the dataset obtained from Step 6 in Figure 2.

In Step A ("Deep Clean"), we refined the dataset by removing variables with more than 20% missing values, imputing categorical variables with over 80% similarity, and dropping data entries with missing values in the remaining variables. Additionally, constant variables were eliminated. As shown in Figure 8, this process significantly reduced the number of data entries. While such a reduction could adversely impact machine learning model development if used directly for training, we utilized this "clean" data solely for clustering and selection to identify important factors and eliminate redundant variables.

The implementation of Step B ("Cluster Variables") is illustrated in Figure 9. First, we computed the correlation matrix of the input variables. For

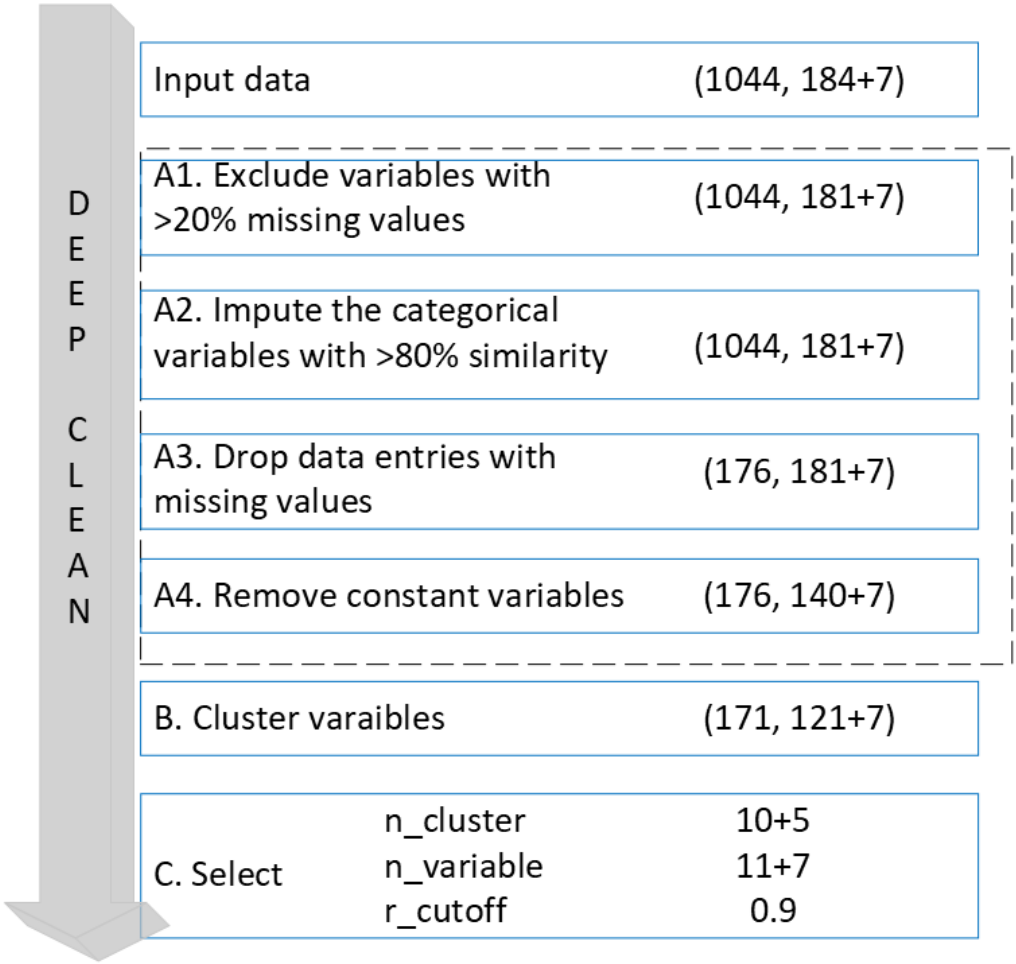


Figure 8: Workflow of variables clustering and selection of Dataset 1 and 2. The form of $(i, j + k)$ represents the number of samples (i), the number of Dataset 1 variables (j) and the number of Dataset 2 PGS (k)

any pair of items with a correlation coefficient ($r(A, B)$) exceeding the threshold value (r_{cutoff}), we replaced them with a single component (X) derived through Principal Component Analysis (PCA). Each item could represent either a single variable or a cluster of multiple variables. The r_{cutoff} value was tuned during model training and selected from options of 0.6, 0.7, 0.8, and 0.9.

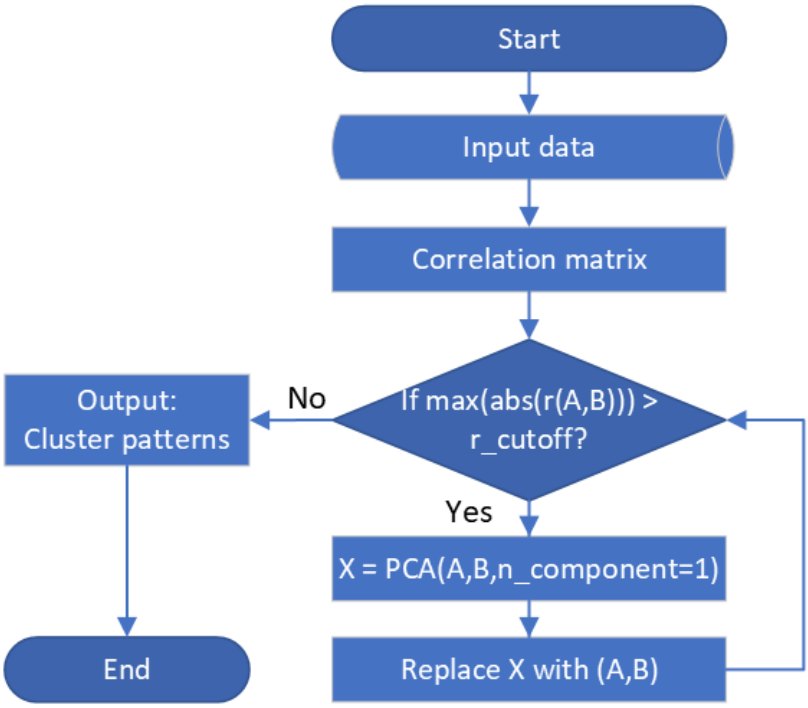


Figure 9: Correlated-based clustering algorithm.

After clustering, we obtained both the cluster patterns and the clustered dataset for Step C ("Select $n_{cluster}$, $n_{variable}$, r_{cutoff}") and subsequent analyses. These clusters were treated as independent units, exhibiting minimal dependency on one another.

In Step C ("Selection"), we employed Recursive Feature Elimination (RFE) integrated with ERF to identify potential predictors. The optimal number of predictors was determined by tuning among 10, 20, 30, and 40.

Acknowledgment

We gratefully acknowledge all Raine Study participants and their families for their continued participation in the study, as well as the Raine Study team

for study co-ordination and data collection. We also thank the NHMRC and the Raine Medical Research Foundation for their support. The core management of the Raine Study is funded by The University of Western Australia, Curtin University, The Kids Research Institute Australia, Women and Infants Research Foundation, Edith Cowan University, Murdoch University, The University of Notre Dame Australia and the Western Australian Future Health Research and Innovation Fund (2023-2024; Grant ID WACSOSP2023-2024). The Pawsey Supercomputing Centre provided computation resources to carry out analyses required with funding from the Australian Government and the Government of Western Australia. The data collection of the Raine Study Gen1- and Gen2-1, 2, 5, 8 year follow-ups was funded by NHMRC Grant and The Raine Medical Research Foundation.

References

References

[1] F. Rubino, D. E. Cummings, R. H. Eckel, R. V. Cohen, J. P. Wilding, W. A. Brown, F. C. Stanford, R. L. Batterham, I. S. Farooqi, N. J. Farpour-Lambert *et al.*, "Definition and diagnostic criteria of clinical obesity," *The Lancet Diabetes & Endocrinology*, 2025.

[2] G. Colmenarejo, "Machine learning models to predict childhood and adolescent obesity: a review," *Nutrients*, vol. 12, no. 8, p. 2466, 2020.

[3] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," *Computers in biology and medicine*, vol. 136, p. 104754, 2021.

[4] M. J. Gerl, C. Klose, M. A. Surma, C. Fernandez, O. Melander, S. Männistö, K. Borodulin, A. S. Havulinna, V. Salomaa, E. Ikonen *et al.*, "Machine learning of human plasma lipidomes for obesity estimation in a large population cohort," *PLoS biology*, vol. 17, no. 10, p. e3000443, 2019.

[5] D. Scheinker, A. Valencia, and F. Rodriguez, "Identification of factors associated with variation in us county-level obesity prevalence rates using epidemiologic vs machine learning models," *JAMA Network open*, vol. 2, no. 4, pp. e192 884–e192 884, 2019.

[6] K. Jindal, N. Baliyan, and P. S. Rana, "Obesity prediction using ensemble machine learning approaches," in *Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 2*. Springer, 2018, pp. 355–362.

[7] C. A. C. Montañez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, J. Hind, and N. Radi, "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2743–2750.

[8] X. Pang, C. B. Forrest, F. Le-Scherban, and A. J. Masino, "Understanding early childhood obesity via interpretation of machine learning model predictions," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 1438–1443.

[9] B. Singh and H. Tawfik, "A machine learning approach for predicting weight gain risks in young adults," in *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE, 2019, pp. 231–234.

[10] P. Fergus, A. Hussain, J. Hearty, S. Fairclough, L. Boddy, K. A. Mackintosh, G. Stratton, N. D. Ridgers, and N. Radi, "A machine learning approach to measure and monitor physical activity in children to help fight overweight and obesity," in *Intelligent Computing Theories and Methodologies: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015, Proceedings, Part II 11*. Springer, 2015, pp. 676–688.

[11] M. Gupta, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti, "Obesity prediction with ehr data: A deep learning approach with interpretable elements," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 3, pp. 1–19, 2022.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[13] J. Bajorath, "From scientific theory to duality of predictive artificial intelligence models," *Cell Reports Physical Science*, vol. 6, no. 4, 2025.

[14] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[15] C. He, M. Ma, and P. Wang, "Extract interpretability-accuracy balanced rules from artificial neural networks: A review," *Neurocomputing*, vol. 387, pp. 346–358, 2020.

[16] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 6, pp. 741–760, 2021.

[17] M. Simmonds, A. Llewellyn, C. G. Owen, and N. Woolacott, "Predicting adult obesity from childhood obesity: a systematic review and meta-analysis," *Obesity reviews*, vol. 17, no. 2, pp. 95–107, 2016.

[18] V. W. Jaddoe, J. F. Felix, A.-M. N. Andersen, M.-A. Charles, L. Chatzi, E. Corpeleijn, N. Donner, A. Elhakeem, J. G. Eriksson, R. Foong *et al.*, "The lifecycle project-eu child cohort network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents," *European journal of epidemiology*, vol. 35, pp. 709–724, 2020.

[19] L. Straker, J. Mountain, A. Jacques, S. White, A. Smith, L. Landau, F. Stanley, J. Newnham, C. Pennell, and P. Eastwood, "Cohort profile: the western australian pregnancy cohort (raine) study–generation 2," *International journal of epidemiology*, vol. 46, no. 5, pp. 1384–1385j, 2017.

[20] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *Test*, vol. 18, no. 1, pp. 1–43, 2009.

[21] R. Wei, C. L. Ogden, V. L. Parsons, D. S. Freedman, and C. M. Hales, "A method for calculating bmi z-scores and percentiles above the 95th percentile of the cdc growth charts," *Annals of human biology*, vol. 47, no. 6, pp. 514–521, 2020.

[22] W. H. Organization *et al.*, "Who child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development," in *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*, 2006.

[23] J. Min, X. Wen, H. Xue, and Y. Wang, "Ethnic disparities in childhood bmi trajectories and obesity and potential causes among 29,250 us children: Findings from the early childhood longitudinal study-birth and kindergarten cohorts," *International journal of obesity*, vol. 42, no. 9, pp. 1661–1670, 2018.

[24] A. D. Guerrero, C. Mao, B. Fuller, M. Bridges, T. Franke, and A. A. Kuo, "Racial and ethnic disparities in early childhood obesity: growth trajectories in body mass index," *Journal of racial and ethnic health disparities*, vol. 3, pp. 129–137, 2016.

[25] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[26] S. A. Lambert, B. Wingfield, J. T. Gibson, L. Gil, S. Ramachandran, F. Yvon, S. Saverimuttu, E. Tinsley, E. Lewis, S. C. Ritchie *et al.*, "Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization," *Nature Genetics*, vol. 56, no. 10, pp. 1989–1994, 2024.

[27] O. Weissbrod, M. Kanai, H. Shi, S. Gazal, W. J. Peyrot, A. V. Khera, Y. Okada, A. R. Martin, H. K. Finucane *et al.*, "Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores," *Nature Genetics*, vol. 54, no. 4, pp. 450–458, 2022.

[28] F. Privé, H. Aschard, S. Carmi, L. Folkersen, C. Hoggart, P. F. O'Reilly, and B. J. Vilhjálmsson, "Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort," *The American Journal of Human Genetics*, vol. 109, no. 1, pp. 12–23, 2022.

[29] A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick *et al.*, "Polygenic prediction of weight and obesity trajectories from birth to adulthood," *Cell*, vol. 177, no. 3, pp. 587–596, 2019.

[30] R. Monti, L. Eick, G. Hudjashov, K. Läll, S. Kanoni, B. N. Wolford, B. Wingfield, O. Pain, S. Wharrie, B. Jermy *et al.*, "Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning," *The American Journal of Human Genetics*, vol. 111, no. 7, pp. 1431–1447, 2024.

[31] I. Shim, H. Kuwahara, N. Chen, M. O. Hashem, L. AlAbdi, M. Abouelhoda, H.-H. Won, P. Natarajan, P. T. Ellinor, A. V. Khera *et al.*, "Clinical utility of polygenic scores for cardiometabolic disease in arabs," *Nature Communications*, vol. 14, no. 1, p. 6535, 2023.

[32] Y. Ma, S. Patil, X. Zhou, B. Mukherjee, and L. G. Fritsche, "Exprsweb: An online repository with polygenic risk scores for common health-related exposures," *The American Journal of Human Genetics*, vol. 109, no. 10, pp. 1742–1760, 2022.

[33] D. Borisevich, T. M. Schnurr, L. Engelbrechtsen, A. Rakitko, L. Ängquist, V. Ilinsky, M. Aadahl, N. Grarup, O. Pedersen, T. I. Sørensen *et al.*, "Non-linear interaction between physical activity and polygenic risk score of body mass index in danish and russian populations," *Plos one*, vol. 16, no. 10, p. e0258748, 2021.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.

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

[36] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[37] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[39] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[40] A. V. Konstantinov and L. V. Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowledge-Based Systems*, vol. 222, p. 106993, 2021.

[41] A. Palkar, C. C. Dias, K. Chadaga, and N. Sampathila, "Empowering glioma prognosis with transparent machine learning and interpretative insights using explainable ai," *IEEE access*, vol. 12, pp. 31 697–31 718, 2024.

[42] H. Du, Q. Yang, A. Ge, C. Zhao, Y. Ma, and S. Wang, "Explainable machine learning models for early gastric cancer diagnosis," *Scientific Reports*, vol. 14, no. 1, p. 17457, 2024.