Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study

Álvaro Torres-Martos [a,b,c,1], Augusto Anguita-Ruiz [c,d,1], Mireia Bustos-Aibar [a,c,e], Alberto Ramírez-Mena [f], María Arteaga [g], Gloria Bueno [c,e,h], Rosaura Leis [c,i], Concepción M. Aguilera [a,b,c,*], Rafael Alcalá [g], Jesús Alcalá-Fdez [g]

[a] *Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain*

[b] *Instituto de investigación Biosanitaria ibs.GRANADA, Granada, 18012, Spain*

[c] *CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, 28029, Spain*

[d] *Barcelona Institute for Global Health, ISGlobal, Barcelona, 08003, Spain*

[e] *Growth, Exercise, Nutrition and Development (GENUD) Research Group, Institute for Health Research Aragón (IIS Aragón), Zaragoza, 50009, Spain*

[f] *Bioinformatics Unit, Centre for Genomics and Oncological Research, GENYO Pfizer/University of Granada/Andalusian Regional Government, PTS, Granada, 18016, Spain*

[g] *Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain*

[h] *Pediatric Endocrinology Unit, Facultad de Medicina, Clinic University Hospital Lozano Blesa, University of Zaragoza, Zaragoza, 50009, Spain*

[i] *Unit of Pediatric Gastroenterology, Hepatology and Nutrition, Pediatric Service, Hospital Clínico Universitario de Santiago. Unit of Investigation in Nutrition, Growth and Human Development of Galicia-USC, Pediatric Nutrition Research Group-Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, 15706, Spain*

ARTICLE INFO

Keywords:
Pediatric obesity
Insulin resistance
Epigenomics
Multiomics
Machine Learning
Explainable Artificial Intelligence

ABSTRACT

Pediatric obesity can drastically heighten the risk of cardiometabolic alterations later in life, with insulin resistance standing as the cornerstone linking adiposity to the increased cardiovascular risk. Puberty has been pointed out as a critical stage after which obesity-associated insulin resistance is more difficult to revert. Timely prediction of insulin resistance in pediatric obesity is therefore vital for mitigating the risk of its associated comorbidities. The construction of effective and robust predictive systems for a complex health outcome like insulin resistance during the early stages of life demands the adoption of longitudinal designs for more causal inferences, and the integration of factors of varying nature involved in its onset. In this work, we propose an eXplainable Artificial Intelligence-based decision support pipeline for early diagnosis of insulin resistance in a longitudinal cohort of 90 children. For that, we leverage multi-omics (genomics and epigenomics) and clinical data from the pre-pubertal stage. Different data layers combinations, pre-processing techniques (missing values, feature selection, class imbalance, etc.), algorithms, training procedures were considered following good practices for Machine Learning. SHapley Additive exPlanations were provided for specialists to understand both the decision-making mechanisms of the system and the impact of the features on each automatic decision, an essential issue in high-risk areas such as this one where system decisions may affect people's lives. The system showed a relevant predictive ability (AUC and G-mean of 0.92). A deep exploration, both at the global and the local level, revealed promising biomarkers of insulin resistance in our population, highlighting classical markers, such as Body Mass Index z-score or leptin/adiponectin ratio, and novel ones such as methylation patterns of relevant genes, such as *HDAC4*, *PTPRN2*, *MATN2*, *RASGRF1* and *EBF1*. Our findings highlight the importance of integrating multi-omics data and following eXplainable Artificial Intelligence trends when building decision support systems.

[*] Corresponding author at: Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain.

E-mail addresses: alvarotorres@ugr.es (Á. Torres-Martos), augusto.anguita@isglobal.org (A. Anguita-Ruiz), mbustos@iisaragon.es (M. Bustos-Aibar), alberto.ramirez@genyo.es (A. Ramírez-Mena), mariaartj@correo.ugr.es (M. Arteaga), mgbuenol@unizar.es (G. Bueno), mariarosaura.leis@usc.es (R. Leis), caguiler@ugr.es (C.M. Aguilera), alcala@decsai.ugr.es (R. Alcalá), jalcala@decsai.ugr.es (J. Alcalá-Fdez).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.artmed.2024.102962
Received 27 February 2024; Received in revised form 31 July 2024; Accepted 16 August 2024
Available online 20 August 2024

1. Introduction

According to the World Obesity Atlas (WOA) 2023, published by the World Obesity Federation, more than half of the global population will be living with overweight or obesity by 2035 if the current trend persists. In children and adolescents, the situation is even worse, with obesity rates rising faster than in adults. Unless significant action is taken, by 2035, WOA estimates obesity cases to reach 1.5 billion among adults and nearly 400 million in children. Aside from the devastating population health impact, it is estimated that the total cost of treating obesity-related illnesses will amount to $4 trillion per year, representing almost 3% of worldwide gross domestic product

Obesity is associated with an increased risk of mortality, especially if originated from the early stages of life [1,2]. Premature mortality in people with obesity is mainly caused by the appearance of cardio-metabolic disturbances including cardiovascular diseases, Type II Diabetes (T2D), and Metabolic Syndrome (MetS) [3–5]. Insulin Resistance (IR), defined as a pathological condition in which cells become less responsive to the effects of insulin on a systemic level, is the metabolic comorbidity of obesity that shows the earliest appearance in patients and represents a cornerstone linking adiposity to the rest of cardiometabolic complications [6,7].

The onset of IR in patients with obesity usually occurs already from the very early stages of life (∼10 years old) and can get worse with the occurrence of key developmental events such as puberty [8]. During puberty, a range of dynamic physiological changes take place (e.g., secretion of sex steroids and accumulation of fat and lean mass) that are related to distinct prognostics of IR, highlighting the importance of this developmental stage for long-term health. Nevertheless, pubertal alterations appear to impact individuals differently [9]. In healthy normal-weight children, there is a physiological decrease in insulin sensitivity during mid-puberty, which typically recovers by the end of the pubertal period. However, evidence suggests that IR persists in children with obesity as they enter puberty, leading to higher cardiometabolic risk [6]. Consequently, puberty has been pointed out as a critical stage upon which obesity-associated adverse cardiometabolic disturbances are more difficult to revert [10]. In this regard, the early childhood appears a magnificent window of opportunity for the implementation of preventive actions against obesity-associated IR worsening and appearance [2,11].

The prediction of which pre-pubertal children will develop pubertal IR, and will subsequently exhibit adverse cardiometabolic trajectories during adulthood is a challenging task. Nevertheless, translational findings would improve the capacity for preventive care through the prioritization of nutritional or lifestyle interventions for high-risk pre-pubertal children [12]. Indeed, there is strong evidence that not all children with obesity develop chronic IR after puberty, maintaining a healthy metabolic status throughout their life (a condition known as metabolically healthy obese) [13,14]. The totality of factors conditioning the worsening of metabolic health and IR in children with obesity are not fully understood yet, possibly involving complex interactions between environmental and molecular factors. Consequently, the metabolic evaluation of children at high risk currently relies on classical biomarkers with limited predictive ability [15]. Therefore, there is an emerging need to identify new biomarkers to incorporate into predictive systems to help pediatricians accurately diagnose the developing IR at an early stage. [16]. These clinical Decision Support Systems (DSS) would enable pediatricians to estimate each child's metabolic risk and provide more effective and personalized treatments in primary care [17].

Thanks to current technological advances and the increased use of high-throughput molecular screening systems, large amounts of omic data have increasingly become available for biological and clinical research (e.g., genomics, epigenomics, transcriptomics, proteomics, metabolomics), identifying novel and promising predictive biomarkers for many diseases, including IR [16,18]. The heterogeneous and complex nature of the different types of omic biomarkers available demands on the other hand the employment of advanced analysis techniques able to jointly integrate all these multi-modal data if we want to build upon them reliable and robust clinical DSS [19]. Likewise, the adoption of longitudinal study designs is mandatory for this task [20,21], which allows for building better systems by considering the intrinsic across-time huge variation of this type of data in humans [6,9,22].

In this midst of this need, Artificial Intelligence (AI), and particularly Machine Learning (ML) techniques, have been successfully applied to predict the IR due to their ability to automatically integrate data from different information sources in order to obtain descriptive or predictive models that enable us to develop the inference engine of a clinical DSS, helping pediatricians to detect and diagnose diseases in an earlier and more accurate way. The authors of previous works [23,24] used ML techniques to identify children and adolescents with or without pre-diabetes from clinical or single-layer omic data belonging to cross-sectional studies. However, these studies provide static predictions without considering the dynamic physiological changes that occur during puberty or the information available from other omics. In addition, experts often do not trust the latest technical and methodological approaches (e.g., Deep Learning), despite their high accuracy, as they provide models for which it is not possible to explain in a human-understandable way how they make their predictions [25]. Pediatricians do not trust the decisions generated by these models unless they are accompanied by exhaustive and easily understandable explanations since in many cases these models should be considered as clinical DSS, being in many cases more important to understand "how the decision was made" than the decision itself. A number of studies have attempted to unravel the inner workings of complex systems and offer explanations regarding their decisions, either by understanding how the systems perform or by explaining their decisions. This new trend is called eXplainable Artificial Intelligence (XAI) [26], which recommends the use of transparent systems that by their nature are self-explanatory and post-hoc explainability techniques that aim to provide understandable information about how a complex system makes its predictions for any given input. This transparency of the systems is especially essential in high-risk areas, such as healthcare, in which the output of predictive systems has an impact on the patient's lives, so experts will only deploy and use them if they can be trusted [27]. Due to that, the European Commission has published the "Ethics Guidelines for Trustworthy AI" and has recently accepted the first draft of an AI law that promotes legal, lawful and robust AI.[2]

The use of these predictive tools from early ages could improve the healthcare and knowledge of children having a high risk to develop cardiometabolic alterations during adulthood. Accordingly, this paper utilizes an XAI-based pipeline to generate an accurate and understandable DSS that enables the prediction of pubertal IR in children from their pre-pubertal multi-omic information, to identify new molecular mechanisms of IR in pediatric population and to report the most relevant features to determine their potential future integration into clinical practice. To this end, we employ multi-omic information (three layers: genetics variants, DNA methylation measures, and anthropometry, biochemistry measurements and protein biomarkers) from the pre-pubertal stage of a longitudinal cohort of 90 children. A methodology based on ML good practices is followed, analyzing different resampling techniques for class balancing and testing the results obtained by various ML algorithms for the early diagnosis of IR from the information derived through different ways of joining data layers. To improve comprehension of the working of the DSS, SHapley Additive exPlanations (SHAP) [28] were used to examine the impact of each omic layer on longitudinal predictions using global explanations (feature contributions for the whole system) and local explanations (feature contributions for each specific instance). Both types of explanation are

[2] https://www.europarl.europa.eu.

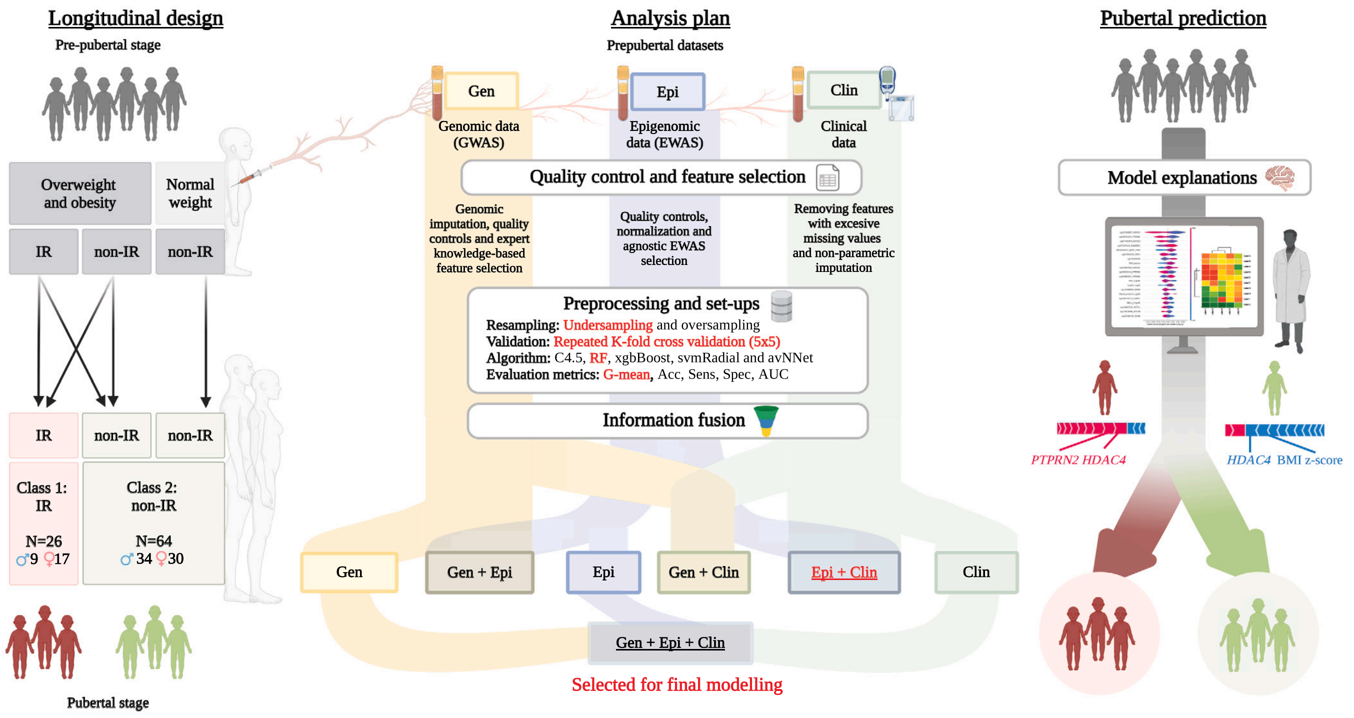
Longitudinal design
Pre-pubertal stage
Overweight and obesity
Normal weight
IR
non-IR
non-IR
IR
non-IR
non-IR
Class 1: IR
N=26
♂9 ♀17
Class 2: non-IR
N=64
♂34 ♀30
Pubertal stage
Analysis plan
Prepubertal datasets
Gen
Epi
Clin
Genomic data (GWAS)
Epigenomic data (EWAS)
Clinical data
Quality control and feature selection
Genomic imputation, quality controls and expert knowledge-based feature selection
Quality controls, normalization and agnostic EWAS selection
Removing features with excesive missing values and non-parametric imputation
Preprocessing and set-ups
Resampling: Undersampling and oversampling
Validation: Repeated K-fold cross validation (5x5)
Algorithm: C4.5, RF, xgbBoost, svmRadial and avNNet
Evaluation metrics: G-mean, Acc, Sens, Spec, AUC
Information fusion
Gen
Gen + Epi
Epi
Gen + Clin
Epi + Clin
Clin
Gen + Epi + Clin
Selected for final modelling
Pubertal prediction
Model explanations
PTPRN2 HDAC4
HDAC4 BMI z-score

Fig. 1. Summary of the experimental design. The longitudinal study consisted of pre-pubertal children who were followed into puberty three years later. The pre-pubertal information was used as input to generate the classifiers and the output was the pubertal IR status. The analysis plan utilizes genomic (*Gen*), epigenomic (*Epi*), and clinical (*Clin*) data from pre-pubertal children. The chosen data combination, algorithm, and resampling method are highlighted in red. Subsequently, we made pubertal predictions and analyzed the final classifier's behavior using post-hoc explainer. Abbreviations: Acc, Accuracy; BMI, Body Mass Index; EWAS, Epigenome-Wide Association Study; GWAS, Genome-Wide Association Study; Sens, Sensitivity; Spe, Specificity; AUC, Area Under the ROC curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

effective to understand the performance of a wide range of ML systems by supplying an importance value to each input feature for every prediction, enhancing the reliability and transparency of the systems through a deeper comprehension of the fundamental causes influencing each prediction [28].

The proposed pipeline makes novel contributions, including the importance of pre-processing in omic and clinical data, the evaluation of different data layer fusion in combination with diverse resampling methods and classification algorithms, a deep understanding of clinical specialists of the DSS at the global level, and a better comprehension of how the DSS stratifies children depending on clinical and omic characteristics at the local level. With the aim of assessing the effectiveness of our pipeline, we compared our proposal with other ML approaches selected among the most accurate and the most interpretable or understandable ones from a comparative study [29], employing a stratified repeated 5-fold cross-validation approach to assess our system's performance. Several metrics and nonparametric statistical test were considered to analyze the performance of the analyzed algorithms. Supplementary material from this study has been included in a web page associated with this article (i.e., https://sci2s.ugr.es/MultiOmics_IR_Pred).

The rest of this paper is structured as follows. Section 2 introduces the datasets used and goes in-depth with technical details of the longitudinal multi-omic analysis performed. Section 3 presents the main reported results from the longitudinal predictive analysis. Section 4 discusses and highlights some promising omic biomarkers in the prediction of IR in pediatric population. Finally, Section 5 concludes the paper.

2. Material and methods

2.1. Overview of the analysis plan

The present work aimed to evaluate the ability of a group of systems to predict the risk of developing pubertal IR in a longitudinal cohort of Spanish children. For this purpose, we used three-layer input molecular data (*Gen*, *Epi* and *Clin* data) derived from the pre-pubertal stage, see Fig. 1. The analytic pipeline followed in this study included data pre-processing, feature selection, construction of predictive systems and interpretation of the best candidate. In order to carry out this analysis, we had to face the inherent problems of omic data (biological heterogeneity, background noise, missing values, high dimensionality, etc.) [30,31]. During the construction of predictive systems, the different layers of input predictors were used independently and combined as described in Fig. 1.

The use of complex ML systems usually presents difficulties in understanding the ethical implications of how decision-making could affect patients' lives. Consequently, health professionals prefer to rely on interpretable systems instead of accurate ones [26]. Several innovative techniques currently offer the possibility of opening the black box and understanding its hidden mechanism. In this contribution, we employ interpretable and complex systems that offer in the latter case visualization and feature importance based on SHAP to clarify the drivers of the final system and its predictive contribution [28]. A summary of the whole experimental design and approaches can be found in Fig. 1.

2.2. Study population

The study population analyzed here comes from the "Puberty and metabolic risk in obese children. Epigenetic alterations and pathophysiological and diagnostic implications" (PUBMEP) project. This project is a longitudinal study based on the follow-up of a cohort of Spanish children who previously participated in the "Association between genetic variants, biomarkers of oxidative stress, inflammation and cardiovascular risk in obese children" (GENOBOX) project [6]. The main objective of the project was to unveil the molecular mechanisms behind the appearance of obesity and cardiometabolic complications such as IR from the early stages of life. In the PUBMEP study, pre-pubertal boys and girls initially enrolled in the GENOBOX study who had already initiated puberty were invited to participate. During the PUBMEP study (2015–2018), children underwent regular medical check-ups by the same pediatricians. For this work, a sub-population derived from the PUBMEP study composed of 90 children (43 males and 47 females) was selected. Children were allocated into two classes according to their IR status (26 IR and 64 non-IR) after the onset of puberty. The class was defined by an IR measure frequently used by pediatricians called Homeostatic Model Assessment - IR (HOMA-IR) [32]. The HOMA-IR cut points were published in previous studies [33]. Input data sources comprised three different molecular data layers (*Gen*, *Epi* and *Clin* data). *Gen* and *Epi* data layers contained information about single nucleotide polymorphisms (SNPs) and DNA methylation of CpG sites, respectively. The *Clin* data layer comprised anthropometry, biochemistry, clinical measures, cardiovascular/inflammatory protein biomarkers and adipokines measured in blood samples.

2.3. Data extraction and quality control

To predict pubertal IR, we used *Gen*, *Epi* and *Clin* data collected at the pre-pubertal stage. More details on each data layer can be found below.

2.3.1. Genomic data

Firstly, all children were genotyped for ~654.000 SNPs using Infinium Global Screening Array-24 v3.0 Kit (Infinium HTS Assay platform) for the GWAS analysis. The genotype calls for all children were obtained using the GenomeStudio generating the standard format files (.ped and .map) in the GRCh38/hg38 reference genome. We encoded these files to binary formats (.bed, .bim and .fam) to save space and speed up the subsequent analysis [34].

In the present work, we matched our SNPs (GRCh37/hg19 genomic annotation) with the latest reference panel of the Haplotype Reference Consortium to apply a genotype imputation using the Minimic 4 algorithm through the cloud-based interface of the Michigan Imputation Server. An automated quality control analysis was performed prior to genomic imputation following the default settings of Michigan Imputation Server.

Once the genomic imputation was performed, several commonly used standard quality control filters were applied: (1) filter SNPs that have a low minor allele frequency ($MAF > 0.01$), (2) discard the SNPs that were not in the Hardy–Weinberg equilibrium ($HWE < 10^{-6}$) and (3) remove SNPs with poor imputation quality ($R^2 > 0.9$). Data were transformed to dosage format (.raw) according to the additive genetic model. Consequently, 5,894,726 SNPs remained in the *Gen* dataset. The quality control protocol, management and encoding of *Gen* data were performed using bcftools and PLINK 1.9 command-line programs [34,35].

Table 1
The summary table shows the number of features before and after quality control and feature selection.

	Genomic data	Epigenomic data	Clinical data
Initial features	651,563	866,091	48
Filtered features based on quality control & missing values	512,937	834,371	34
Feature selection method	Expert knowledge-based	Data-driven (agnostic)	–
Final features after feature selection	151	267	34

2.3.2. Epigenomic data

Epi data comprised DNA methylation values for ~850.000 CpG sites which were measured across the whole genome in buffy coat with the Infinium Methylation EPIC (Illumina platform) generating the raw data (IDAT files) for the EWAS analysis. We loaded the raw data into the R environment utilizing the minfi package. We employed Beta-Mixture Quantile (BMIQ) intra-array normalization to eliminate undesirable variability across and within samples [36]. Moreover, low-performing probes were excluded based on established criteria: probes with a detection p-value greater than 0.01 in over 10% of samples (230 probes), probes affected by SNPs (30,432 probes), cross-reactive probes mapping to several locations (25,570 probes), and probes situated on the Y chromosome (246 probes). A total of 834.371 probes remained in the *Epi* dataset. To determine the methylation at each CpG site Beta and M values were calculated. For the purposes of this paper, M values are used due to their statistical robustness [37–39].

2.3.3. Clinical data

In addition to anthropometrical, biochemical and clinical measures, *Clin* data layer included cardiovascular/inflammatory protein biomarkers and adipokines data. Cardiovascular/inflammatory biomarkers and adipokines were measured in blood samples through XMap technology on the Luminex Corporation platform, utilizing human monoclonal antibodies (Miliplex Map Kit). These data were measured as published elsewhere [6,33].

Due to the presence of missing values, we performed an exploratory analysis of the missing values patterns in the *Clin* dataset to check the random structure of the missing values. It is important to consider that the percentage of missing values in the *Clin* dataset was less than 1% after removing the features that contained 5 or more missing values. The final number of features in this layer was 34. Therefore, we decided to use a non-parametric method imputation known as missForest. It has advantages over other imputation methods in that it only generates a unique imputed dataset and also its performance is quite robust, as it does not require a tuning parameter stage [40].

2.4. Feature selection

Omic data are inherently complex due to their high dimensionality. This complexity often leads to challenges in ML, as algorithms may struggle with high-dimensional datasets, producing inaccurate and unrobust models of lower quality. The issue, known as the "curse of dimensionality", involves an increased likelihood of finding spurious statistical associations in large datasets, impacting the quality of the models developed. This is aggravated by the small sample sizes that are common in molecular epidemiological and longitudinal studies, where data is limited and complex to collect over time. To address this issue, a feature selection process is necessary to reduce the number of variables by choosing an optimal subset of variables [30,31]. This process is an important pre-processing step in ML, enabling more accurate and robust feature selection and more trustworthy application of ML techniques [41].

In this context, numerous methods for feature selection have been suggested, including data-driven approaches based on the use of specialized analytical tools, as well as a priori strategies based on expert knowledge. In our study, we employed different approaches depending on the data layer, in order to ensure the most optimal selection of markers likely to be involved in the outcome of our study, particularly in our population. For the *Gen* dataset, we relied on expert knowledge informed by existing literature, a technique endorsed in molecular epidemiology as a credible method [42,43]. Conversely, for the *Epi* data, given its unique characteristics as elaborated subsequently, we implemented a data-driven approach. The particularities considered in the feature selection of each data layer included in this study are described in the following subsections [38,44]. Table 1 presents the number of features in each dataset before and after quality control and feature selection.

2.4.1. Genomic data

A search of literature and databases (GWAS and PGS catalog) [45,46] focusing on IR studies within large European populations enabled the selection of three articles [47–49]. We selected a subset of SNPs tested in studies with a large sample size guaranteeing substantial statistical power to identify the small size effects that a SNP may influences have on a phenotype. Among the 258 SNPs associated with IR from previous studies, 151 were present in our *Gen* dataset.

2.4.2. Epigenomic data

Regarding the *Epi* dataset, we performed a data-driven selection where CpG sites differentially methylated and linked to IR were identified across the genome without prior hypotheses. This process was performed in longitudinal and cross-sectional approaches from an independent population study, which is part of the PUBMEP project. In this case, we considered that agnostic selection was a better choice due to environment-dependent epigenetic variability, in contrast to feature selection on *Gen* data. From this strategy, we selected 267 CpG sites. More details regarding the selection of candidate CpGs can be found elsewhere [18]. The choice to perform data-driven selection on the phenotype of interest (IR) rather than selecting features based on literature findings, as in the case of the *Gen* dataset, was motivated by the fact that epigenetic findings are strongly related to population-specific environmental exposures. In this regard, choosing CpG sites from the same sample used to build the system, sharing characteristics with the current study cohort, proved to be a more advantageous option than selecting CpG sites based on other European studies, where research on children is limited [50].

The associated web page contains comprehensive information on the variables used in each of the datasets (see Table S1). The distribution of clinical features are shown in Table S2.

2.5. Set-ups and imbalance considerations

As we have introduced, an ideal investigation would employ a joint approach in which omic data could be integrated. To elucidate the predictive information attributed to each omic layer, we propose to understand and evaluate the predictive information of omic data both separately and together [51]. This strategy generates many layers and combinations of information data: (1) *Gen* dataset generated from GWAS analysis, (2) *Epi* dataset generated from EWAS analysis, (3) *Clin* dataset, (4) fusion of Gen and Epi datasets (*Gen+Epi*), (5) fusion of Gen and Clin datasets (*Gen+Clin*), (6) fusion of Epi and Clin datasets (*Epi+Clin*) and fusion of all datasets (*Gen+Epi+Clin = All*). See Fig. 1, which shows the different approaches followed in this study [30,31].

A stratified 5-fold cross-validation, repeated 5 times for a total of 25 executions, was chosen to assess the predictive ability of each approach. This approach is appropriate for scenarios in which the population size is limited, decreasing estimation errors, achieving a good balance between bias and variance, and minimizing the influence of the chosen seed by dividing the population into the training and test sets [30,31,52].

Since the population derived from the PUBMEP study is significantly imbalanced (26 IR and 64 non-IR), we analyzed the performance of different sampling methods (oversampling and undersampling) to avoid biasing the learning algorithms towards the majority class (non-IR), which would result in a higher misclassification rate for the minority class (IR) [53]. In this study, we have considered 6 resampling methods available in the R Themis package: SMOTE, SMOTE-NC, ADASYN, ROSE, NearMiss and TomeK. Note that these resampling methods have only been applied to the training set of each fold, so the test sets were not affected and maintained their original values and proportions. Additional information on resampling methods can be found on the associated website. In this paper, we will analyze the results obtained with the NearMiss class balancing method, whose method generated the classifiers with the best performance. The results obtained with the rest of resampling methods are available in the Table S3-S7 (see the supplementary material in the associated web page) [54].

The classification algorithms analyzed have been selected among the most accurate and the most interpretable or understandable from the comparative study presented in [29], in which the performance of 179 classifiers from 17 families was evaluated. From these, we have analyzed: a decision tree algorithm (C4.5 [55]), two ensembles algorithms (Random Forest (RF [56,57]), eXtreme Gradient Boosting (xgBoost [58]), a support vector machine (svmRadialCost [59,60]), and a neural network (avNNet [61]). Further information about algorithms can be found on the associated webpage. All of these algorithms are available in the R Caret package and their parameters have been set to their default values, following the recommendations indicated by their authors when they were published in order to facilitate comparisons and take advantage of the use of configurations that work well in most cases [62].

To assess the performance of the classifiers, we used some classification metrics extensively described in the literature (Accuracy, Sensitivity, Specificity, AUC, and G-mean). Some of these metrics, such as sensitivity and specificity, allow us to analyze their predictive ability on a specific class. Other metrics, such as G-mean, have been designed to combine the predictive ability of the algorithm on both classes searching for a balance between the majority and minority classes. The G-mean measure represents the geometric mean between sensitivity and specificity. Thus, in our population, a low performance in predicting non-IR cases will imply a low value for the G-mean metric even if all cases with IR are correctly classified. These metrics help us to avoid overfitting the majority class and underfitting the minority group. Another interesting analysis is to perform an in-depth study of the performance of the methods in multiple groups of predicted risk, or in groups of true-positive rate or false-positive rate, when the population size allows it [63]. More details about the classification metrics can be found in the supplementary material on the associated web page [64,65].

2.6. Model explanations

In this paper, we have used both interpretable ML algorithms and others that fall into the black box category, in which it is difficult to understand the decision-making mechanisms given its complexity. In order to understand the mechanism underlying complex systems, we use SHAP explainers to calculate feature contributions for each prediction.

SHAP is an algorithm based on the cooperative game theory concept of Shapley values [28]. This approach allows for the explanation of predictions by assigning a contribution value to each feature for a specific prediction. We have used SHAP to calculate the attribution of each predictor, which allows experts to understand the mechanisms behind each of those predictions. Our systems employed the inputs to

Table 2
Classification metrics for classification algorithms.

Fus.	Methods	Classification metrics				
		Accuracy	Sensitivity	Specificity	AUC	G-mean
Gen	RF	0.55 (0.09)	0.51 (0.17)	0.56 (0.12)	0.54 (0.10)	0.53 (0.10)
	xgBoost	0.51 (0.11)	0.48 (0.23)	0.53 (0.14)	0.50 (0.12)	0.48 (0.12)
	C4.5	0.49 (0.09)	0.43 (0.27)	0.51 (0.12)	0.47 (0.12)	0.42 (0.19)
	svmRadial	0.48 (0.16)	0.55 (0.25)	0.45 (0.23)	0.50 (0.14)	0.45 (0.17)
	avNNet	0.58 (0.11)	0.56 (0.22)	0.59 (0.14)	0.57 (0.13)	0.55 (0.16)
Epi	RF	0.57 (0.12)	0.57 (0.20)	0.57 (0.14)	0.57 (0.12)	0.56 (0.13)
	xgBoost	0.57 (0.09)	0.61 (0.21)	0.55 (0.13)	0.58 (0.10)	0.56 (0.11)
	C4.5	0.56 (0.13)	0.50 (0.22)	0.58 (0.17)	0.54 (0.13)	0.52 (0.14)
	svmRadial	0.50 (0.16)	0.51 (0.21)	0.50 (0.21)	0.51 (0.16)	0.44 (0.23)
	avNNet	0.55 (0.24)	0.78 (0.29)	0.46 (0.42)	0.62 (0.14)	0.38 (0.36)
Clin	RF	0.61 (0.08)	0.78 (0.16)	0.54 (0.14)	0.66 (0.07)	0.64 (0.08)
	xgBoost	0.61 (0.10)	0.71 (0.23)	0.57 (0.14)	0.64 (0.10)	0.62 (0.11)
	C4.5	0.59 (0.12)	0.60 (0.20)	0.58 (0.16)	0.59 (0.11)	0.57 (0.13)
	svmRadial	0.53 (0.13)	0.70 (0.23)	0.46 (0.17)	0.58 (0.14)	0.55 (0.14)
	avNNet	0.35 (0.14)	0.69 (0.37)	0.21 (0.30)	0.45 (0.10)	0.13 (0.21)
Gen + *Epi*	RF	0.60 (0.10)	0.62 (0.16)	0.60 (0.14)	0.61 (0.10)	0.60 (0.10)
	xgBoost	0.62 (0.07)	0.61 (0.17)	0.62 (0.11)	0.62 (0.08)	0.61 (0.08)
	C4.5	0.53 (0.08)	0.48 (0.20)	0.55 (0.11)	0.51 (0.10)	0.48 (0.17)
	svmRadial	0.55 (0.15)	0.56 (0.24)	0.55 (0.17)	0.56 (0.16)	0.54 (0.17)
	avNNet	0.68 (0.08)	0.06 (0.14)	0.93 (0.15)	0.02 (0.07)	0.08 (0.17)
Gen + *Clin*	RF	0.61 (0.10)	0.70 (0.20)	0.58 (0.14)	0.64 (0.10)	0.62 (0.11)
	xgBoost	0.61 (0.11)	0.70 (0.23)	0.58 (0.15)	0.64 (0.12)	0.62 (0.13)
	C4.5	0.58 (0.14)	0.62 (0.20)	0.57 (0.18)	0.59 (0.13)	0.58 (0.13)
	svmRadial	0.48 (0.17)	0.57 (0.26)	0.45 (0.45)	0.51 (0.15)	0.44 (0.20)
	avNNet	0.44 (0.17)	0.58 (0.43)	0.39 (0.40)	0.49 (0.07)	0.17 (0.21)
Epi + *Clin*	RF	**0.70 (0.11)**	**0.73 (0.23)**	**0.69 (0.14)**	**0.71 (0.12)**	**0.69 (0.13)**
	xgBoost	0.64 (0.09)	0.71 (0.19)	0.61 (0.12)	0.66 (0.10)	0.64 (0.10)
	C4.5	0.61 (0.12)	0.67 (0.21)	0.59 (0.16)	0.63 (0.13)	0.61 (0.13)
	svmRadial	0.52 (0.18)	0.59 (0.27)	0.49 (0.23)	0.54 (0.18)	0.50 (0.21)
	avNNet	0.56 (0.17)	0.37 (0.28)	0.64 (0.32)	0.51 (0.09)	0.33 (0.24)
All	RF	0.67 (0.10)	0.71 (0.20)	0.66 (0.41)	0.68 (0.10)	0.67 (0.10)
	xgBoost	0.64 (0.10)	0.67 (0.22)	0.62 (0.13)	0.64 (0.11)	0.63 (0.12)
	C4.5	0.58 (0.11)	0.61 (0.20)	0.57 (0.12)	0.59 (0.12)	0.58 (0.12)
	svmRadial	0.49 (0.17)	0.59 (0.27)	0.45 (0.23)	0.52 (0.17)	0.46 (0.22)
	avNNet	0.58 (0.14)	0.33 (0.30)	0.68 (0.27)	0.51 (0.10)	0.33 (0.25)

generate a binary output between 0 (non-IR) and 1 (IR). Thanks to using SHAP, we can compute the individual variable effect or SHAP value for each feature in each prediction (instance-based or local explanations). Likewise, SHAP values can be understood as the contribution of each predictor in the final decision of each prediction. This innovative method can provide reliable explanations to researchers or physicians in each case of study, making them feel more comfortable with the decision-making process [28]. In this work, we calculate the SHAP values, as shown in Eq. (1), where $\phi_{(j)}$ is the SHAP value for feature j, S represents a subset of features excluding feature j, F encompasses the entire set of features, $f_{S\cup\{j\}}$ denotes the system trained with the inclusion of feature j, f_S indicates the system trained without feature j, and x_S reflects the input feature values within subset S.

The overall importance of each feature was calculated as the mean of the SHAP values for each feature across all samples associated with a specified dataset. As illustrated in Eq. (2), I_j is the importance for feature j, where n is the sample size and $\phi_j^{(i)}$ is the SHAP value for sample i and feature j [28,66]. In other words, the SHAP value of a feature is calculated as the difference in prediction with and without that feature in each child. In order to do this, the system is re-trained with all possible subsets of features from the complete set of features [67].

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\cdot(|F|-|S|-1)!}{|F|!}[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)] \tag{1}$$

$$I_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_j^{(i)}| \tag{2}$$

3. Results

Predictive ability for all assessed systems and combinations of data layers are shown in Table 2; this table shows the average classification metrics over test folds with the highest values highlighted in bold. Overall, the algorithms that obtain the lowest classification metrics are those corresponding to the *Gen* data layer. The *Epi* data layer provides more predictive capacity to the algorithms than the *Gen* data layer but less than the *Clin* data layer. The results obtained are better when the algorithms extract integrated information from different layers of information. Of all the data combinations, the fusion of the *Epi* data layer and *Clin* provides the classifiers with the best predictive performance; the *Epi* + *Clin* combination enabled 3 of the 5 algorithms to obtain their best results. Fig. 2 shows an overview of the results obtained.

For this reason, we analyze the differences in the results of the classifiers generated with the *Epi*+*Clin* data fusion by means of the Friedman test. The Friedman ranking non-parametric test was applied to each pair of classifiers to compare their overall performance. The results showed that RF was the best classifier according to 4 of the 5 metrics; see the top of Table 3 where the algorithm that achieved the highest position is highlighted in bold.

Then, we considered conducting pairwise comparisons between classifiers and calculated the adjusted p-values. The numerical outputs for the comparisons are shown at the bottom of Table 3. We rejected the equality hypothesis with greater than 95% confidence in most measures. For measures where RF is the best performing method, significant differences of at least 0.05 are found with svmRadial and avNNet in 4 of 5 measures and with C4.5 in 2 of 5 measures. As for the sensitivity measure, where xgBoost is the best-ranked method

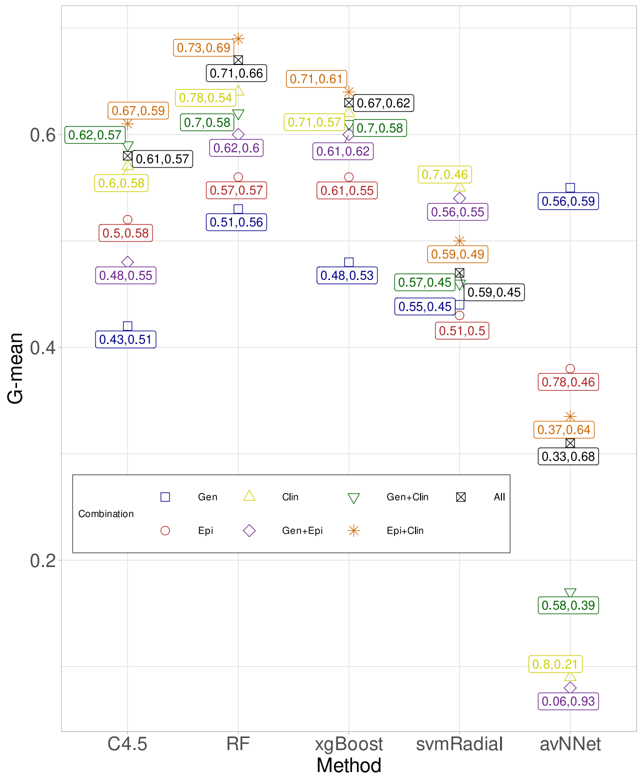

Fig. 2. G-mean is shown in the *y*-axis and sensitivity and specificity are shown in the label for the employed classifiers across the combinations of data layers. According to the predictive metrics, *Epi+Clin* data fusion is the most accurate data combination in 3 of the 5 algorithms.

Table 3
Comparison of classification algorithms.

Methods	Friedman ranking				
	Accuracy	Sensitivity	Specificity	AUC	G-mean
RF	**2.02**	2.48	**2.43**	**1.98**	**1.98**
xgBoost	2.66	**2.42**	3.02	2.56	2.36
C4.5	3.22	2.66	3.26	3.10	2.88
svmRadial	3.70	3.22	3.72	3.54	3.32
avNNet	3.40	4.22	2.66	3.82	4.46
Methods	Adjusted p-values by Holm's procedure				
	Accuracy	Sensitivity	Specificity	AUC	G-mean
RF	–	1.18	–	–	–
xgBoost	0.15	–	0.25	0.19	0.39
C4.5	0.01	1.18	0.11	0.02	0.08
svmRadial	<0.01	0.22	<0.01	<0.01	<0.01
avNNet	<0.01	<0.01	0.47	<0.01	<0.01

according to Friedman procedure, we found no significant differences between them, revealing that both methods show similar behavior in the minority class (IR). On the other hand, no significant statistical differences are observed between RF and xgBoost, but we stated that the metrics of the RF method were the highest in this data combination. Based on the statistical results obtained, we can see how RF is the ML technique that obtains the models with the best statistical results on the test sets.

With the aim of interpreting RF and extracting useful biological insights from it, RF was trained with the whole population (N = 90), combined with an undersampling strategy. Table 4 shows the values obtained by the model learned on the whole population for the best data combination following the Nearmiss undersampling. It is noteworthy that RF obtained values of 0.90, 1, and 0.92 in the Accuracy, Sensitivity, and G-mean metrics with the whole dataset, respectively. Sensitivity is defined as the ratio of true positives to false negatives. A sensitivity value of 1 indicates that the system has correctly predicted all true positives, without any false negatives (see Performance measures in the supplementary material on the associated web page).

Table 4
Predictive ability of the final classifier.

Method	Classification metrics				
	Accuracy	Sensitivity	Specificity	AUC	G-mean
RF	0.90	1.00	0.85	0.92	0.92

Next, we uncovered the hidden mechanism behind our final classifier using the SHAP values. These values, properly visualized, are really useful for the identification of the most important features used by the system as predictors and the directionality of the association (e.g. higher values of a certain variable incline the system to classify an individual as part of the IR class). Interestingly, these values can be extracted at the level of the whole study population (global explanations), which gives us an idea of the overall structure of the system, as well as at the level of groups of individuals (local explanations), which is very useful for identifying whether the system is using different features to predict class in different subgroups. For these reasons, SHAP values have been postulated as a promising tool to open black box systems such as RF with clinical applications [28,66].

Fig. 3 illustrates a visual representation of the feature importance of our top 20 predictors based on their predictive importance in our final system. The features are arranged according to their overall contribution. Each dot signifies the contribution of a predictor to the classifier's prediction for an individual child. The color of the dot corresponds to the value of the feature, with pink indicating high values and blue indicating low values. We generated a dot plot by class to study the impact of the features in the final output. The graph on the left shows the more relevant features for discriminating the negative class (Non-IR), while the graph on the right shows those corresponding to the positive class (IR). Figures S2a-S2b (see supplementary material on the associated web page) show a dot plot and a violin plot with the contribution of each feature to all samples, without distinguishing between different classes.

In general, the contributions of each feature had a small effect in each example; see Tables S8-10 (see supplementary material on the associated web page). The complementariness of information between both data layers seems to have a strong influence on the strength of the prediction. We found that DNA methylation patterns along with adipokines, anthropometric, and biochemical measures were implicated in the predictive ability of the classifier. Interestingly, the analysis of the SHAP values shows that the variables with the greatest contribution to the predictive work belong to the *Epi* data layer. It is quite relevant that the patterns used by the system to differentiate between the positive and negative classes are slightly different.

4. Discussion

In this work we built a that, using different types of omic data from pre-pubertal children aged to 6–12 years old, is able to predict the future IR status of children when they reach the pubertal stage (3 years later). This approach is an unprecedented work that shows how *Epi* and *Clin* information contains significant predictive power for predicting longitudinal trajectories of metabolic diseases. As a main conclusion, our work demonstrates that to achieve this it is necessary to combine the different data layers, which justifies their integration in this type of system beyond their use as individual factors.

To develop this work we fixed an ambitious objective and analyzed heterogeneous omic data from several platforms and technologies. Thus, we show how it is possible to perform a single-omic and multi-omic analysis with ML pipeline to evaluate the predictive ability of

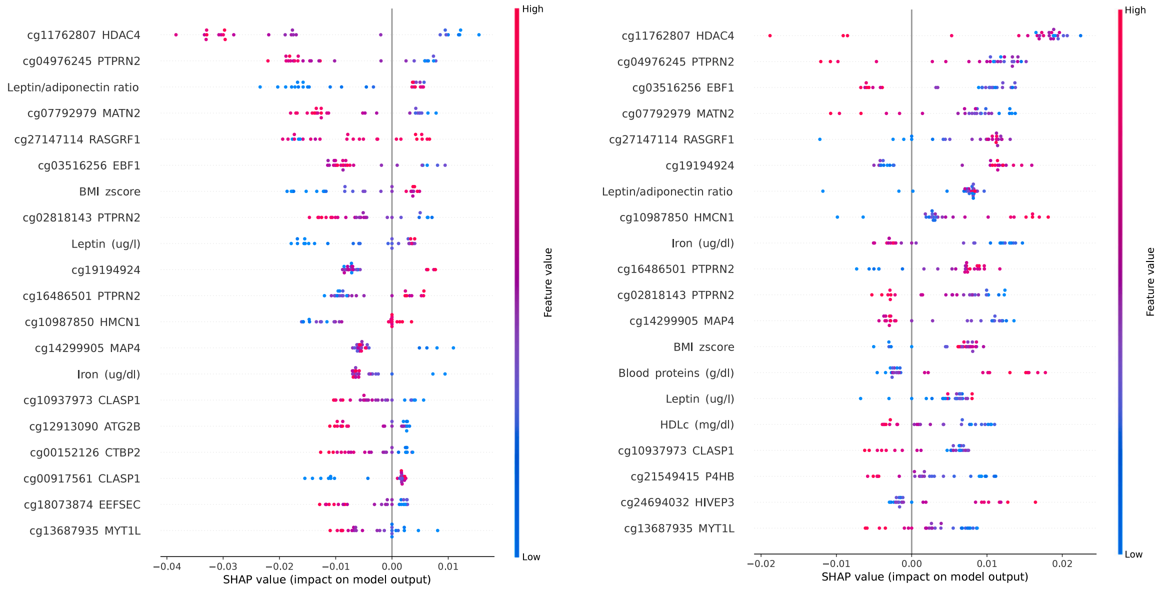

Fig. 3. The SHAP analysis was conducted for our final system to provide global explanations. The top 20 features, ranked by their contributions, are presented separately for Non-IR (left) and IR children (right). Each point represents the contribution of a specific child and feature to the system. The color of the point indicates the value of the feature, with pink representing high values and blue representing low values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

each dataset. It is readily apparent from this multi-omic study that the genetic data layer does not provide the algorithms with extra predictive information. The invariant nature of the SNPs could explain the result obtained, as they have a small cumulative effect size and need to be transformed into genetic risk scores to predict obesity risk more successfully [68]. It was not surprising that the *Clin* dataset provided the most predictive information individually because it contains anthropometric, biochemical, and adipokine variables, which are commonly used markers of cardiometabolic risk. The *Epi* dataset, which reflects a molecular signature of environmental exposures or disease changes, ranks in an intermediate position in terms of predictive utility [69].

The chosen strategy consists of analyzing each data layer separately and fusing the layers to interpret the results, demonstrating the need to develop intelligent data integration methods [70–72]. In general, we can highlight that data fusion has generated classifiers with better predictive performance than individual data layers, especially if *Gen* data is excluded. Particularly, systems use the complementarity between molecular information from Epi data and Clin data to achieve better predictive performance. For this reason, we hypothesized that the *Clin* data layer could provide a real-time metabolic signature, e.g., adipokines concentration (leptin and adiponectin) correlated with adiposity levels, BMI z-score and other biochemical blood parameters such as HDL (High Density Lipoprotein) and iron levels, which provide valuable information to the classifier regarding the children's metabolic status at the precise moment of risk estimation. Meanwhile, the *Epi* data layer offered a molecular signature that is a consequence of the long-term and medium-term lifestyle such as diet and physiological conditions of the children. Thus, these layers of omic data could provide distinct temporary molecular information that explains the predictive effectiveness of their combination.

The algorithm with the best predictive performance (0.90, 1 and 0.92 as Accuracy, Sensitivity and G-mean values, respectively), summarized by G-mean values (see Table 4), in the best data fusion was RF, an algorithm widely used in the area of bioinformatics due to its robustness and its great capacity to work with datasets with class imbalance or high dimensionality problems [73]. However, it is an ensemble method that has a major drawback for research and clinical use, as it is a black box system [26]. To overcome this limitation of RF, post-hoc explainers have been proposed to interpret and understand the classifier by visualizing the impact of each variable on decision-making. The SHAP values stand out because they provide researchers with global explanations about the influence of each feature on the whole system, and with local explanations of the effect that each feature has in classifying a specific child. This approach makes it possible to study of the impact of all variables on the overall prediction of the system. The utilization of SHAP values enables us, like classical approaches, to determine the ranking of the most significant variables. However, it also allows us to identify the directionality of the statistical association and the complementariness between features in the decision-making process. Due to their characteristics, SHAP values might prove to have great prospects in the medical area with respect to the diagnosis, monitoring, prognosis and treatment of various pathologies [28,66].

4.1. Global explanations

We analyzed SHAP values to have an idea of the general behavior of the final system, identiying some potential biomarkers for IR risk estimation. Leveraging averaged SHAP values for each feature across individuals, we built a ranking of feature importance. The top important feature of the classifier was the DNA methylation of a CpG site (cg11762807) annotated in the *HDAC4* (Histone Deacetylase 4) gene and several CpG sites related to *PTPRN2* (Protein Tyrosine Phosphatase Receptor Type N2) gene along with leptin/adiponectin ratio and BMI z-score, among others features, see Fig. 3. Surprisingly, the DNA methylation levels of the *HDAC4* and *PTPRN2* genes might be a robust biomarker from the pre-pubertal stage to predict the pubertal IR due to homogeneous behavior in all individuals independently of the adipokine levels or BMI z-score. However, it is important to note that each variable's contribution had a small effect. In other words, the system's prediction for each child is the sum of all variable contributions.

HDAC4 is a crucial participant in a complex regulatory network of gene expression in several tissues related to IR such as the adipose tissue or pancreas. Importantly, the researchers observed a *HDAC4* hypermethylation in adipose tissue samples in response to a six-month exercise intervention, corresponding to reduced *HDAC4* gene expression after exercise [74]. Furthermore, *HDAC4* was differentially methylated

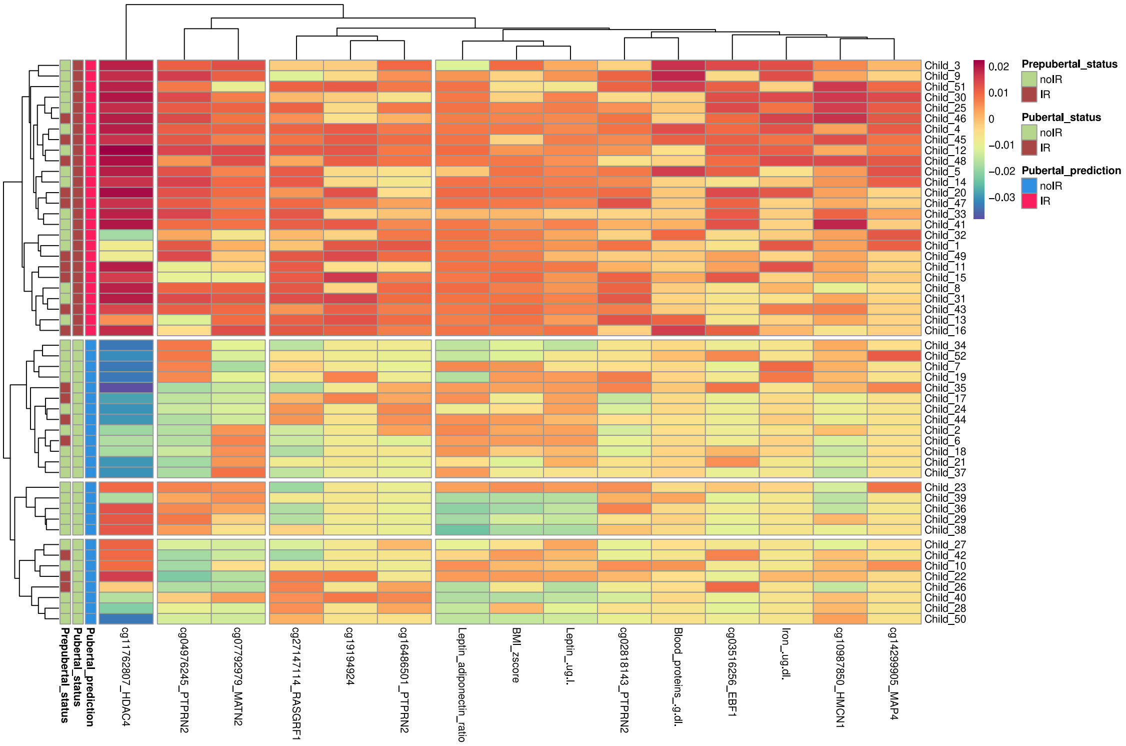

Fig. 4. The heatmap displays clusters of children (rows) and variables (columns) based on their SHAP values. Clustering was only performed on the children used to generate the final classifier (RF) and the fifteen variables that contributed the most to the system based on their SHAP values. The legend shows that red and green represent children who were in IR and non-IR in pre-pubertal and pubertal states, respectively. Blue and red represent the non-IR and IR predictions of RF, respectively. The identification number of each child is displayed on the right-hand side. The visualization displays four clusters of children and variables. The initial cluster of variables comprises solely *HDAC4* methylation, while the following two clusters consist of methylation of the main genes and the last cluster of variables comprises clinical variables such as BMI z-score and leptin/adiponectin ratio together with other methylation patterns. The first, second, third, and fourth clusters of children are composed of the following individuals with IDs from child 3 to child 16, from child 34 to child 37, from child 23 to child 38 and from child 27 to child 50, respectively It is worth noting that the first and second clusters of children can be distinguished by their *HDAC4* methylation pattern, while the last two clusters are characterized by their heterogeneity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in both visceral and subcutaneous adipose tissue samples before and after gastric bypass surgery and was also associated with weight loss. The study also identified strong correlations between *HDAC4* methylation in subcutaneous adipose tissue and fasting glucose levels [75]. According to studies conducted on mice and culture media, *HDAC4* appears to play a key role in the differentiation of insulin-producing beta cells in the pancreas [76]. Similar to our findings, differential methylation of the *HDAC4* gene in white blood cells was found between controls and children with obesity from a cohort of Spanish children aged around 10 years [77]. Subsequently, another study confirmed the relationship between childhood obesity and DNA methylation patterns in other CpG sites of the *HDAC4* gene in peripheral blood leukocytes [78]. Recently, a meta-analysis of four European cohorts studied the association between DNA methylation in peripheral blood and T2D, once again highlighting a CpG site of the *HDAC4* gene, which was also related to gender, age, glucose tolerance and the C-reactive protein [79]. On the other hand, a proteomic study in peripheral blood mononuclear cells found that *HDAC4* was down-regulated in individuals with obesity and induced by physical exercise. *HDAC4* levels were positively correlated with maximum oxygen consumption and negatively correlated with BMI and the inflammatory chemokine RANTES. For these reasons, *HDAC4* has been proposed as a therapeutic target for the control and management of excess adiposity and IR due to its protective role against obesity [80].

It should be noted that this study casts a new light on the predictive importance of DNA methylation patterns of the *PTPRN2* gene as there are 3 Cpg sites of this gene that are among the top 20 in our final system. Previous studies have already demonstrated the association between *PTPRN2* and Insulinoma and T2D. Also, it is known that this gene encodes the major auto-antigen in Type I Diabetes and is involved in pathways related to the Immune System and *PAK* signaling [81,82]. Differential methylation in *PTPRN2* CpG sites in blood samples were reported in previous studies dealing with childhood obesity [83], childhood adiposity [84], newborn adiposity [85] and gestational diabetes [84,86,87]. Other highlighted CpG sites are annotated in genes, such as *MATN2* (Matrilin 2), *RASGRF1* (Ras Protein Specific Guanine Nucleotide Releasing Factor 1) and *EBF1* (Early B Cell Factor 1), which seem to be associated functionally with the pubertal IR. Additional information regarding the molecular pathways of the most significant genes at the predictive level is available on the associated web page.

The study of SHAP values and their relationship to the most important variables shed some light on how RF makes predictions (see interactive Figure S3-S4 of the supplementary material on the associated web page). For instance, the system distinguishes individuals with hypermethylated or hypomethylated *HDAC4* gene by using a cut-off point of 5.15. Children with hypomethylated *HDAC4* (below 5.15) have positive SHAP values that classify them as positive (IR), while those with hypermethylated *HDAC4* (above 5.15) have negative SHAP values. Another interesting insight from the use of SHAP values in our system was the identification of a cut-off point of 0.8 in the leptin/adiponecting ratio domain, which could have some clinical utility. Thus, the classifier assigns positive SHAP values to children with a leptin/adiponectin ratio higher than 0.8, placing them in the positive class (IR). Conversely, individuals with a leptin/adiponectin ratio below 0.8 are assigned negative SHAP values, pushing them into the negative class (non-IR). This ratio of adipokines is a well-known marker of IR and cardiometabolic risk [88].

Also, low values of Iron and HDL blood levels were associated with the positive class (IR) in both cases. This result made biological sense because a decrease in HDL is linked to a worse metabolic state leading

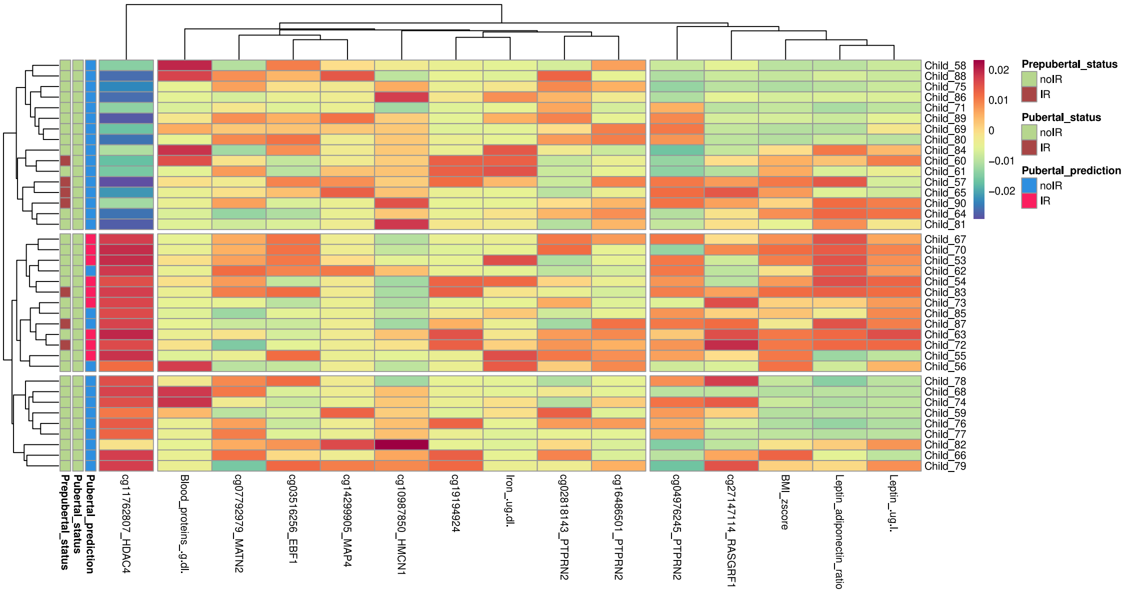

Fig. 5. The heatmap displays clusters of children (rows) and variables (columns) based on their SHAP values. Clustering was only performed on the excluded children during the undersampling procedure and the fifteen variables that contributed the most to the system based on their SHAP values. The legend shows that red and green represent children who were in IR and non-IR in pre-pubertal and pubertal states, respectively. Blue and red represent the non-IR and IR predictions of RF, respectively. The visualization displays three clusters of children and variables. The first cluster of variables is formed only by *HDAC4* methylation, as in Fig. 4. The first, second and third clusters of children are composed of the following individuals with IDs from child 58 to child 81, from child 67 to child 56 and from child 78 to child 79, respectively. Thus, the children highlighted in pink are examples where RF failed in its prediction. It is noteworthy that all of these children belong to the second cluster, which is characterized by positive SHAP values for the first and third clusters of variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to IR and MetS, while reduced iron levels have been associated with obesity over recent years. Significantly, there is a complex bidirectional relationship between iron metabolism and body fat, glucose and lipid metabolism. Alterations in iron status altered the distribution and deposition of body fat and metabolic risk. Similarly, systemic glucose, lipid and insulin are involved in the iron-regulatory pathways [89]. On the other hand, high levels of blood proteins appear to be associated with the positive class (IR), as indicated by positive SHAP values for this variable. This variable might provide information about the functionality of adipose tissue, because when the adipose tissue is hypertrophied and dysfunctional a multitude of adipokines and inflammatory factors of protein nature are released into the bloodstream [9]. Additional information on how the values of different features contribute to the final system with positive or negative SHAP values can be found interactively in Figure S3-S4 of the supplementary material on the associated web page.

Finally, the predictive ability of the final system depends on several predictors included in both datasets. Although RF uses certain variables more frequently than others to make predictions, it is important to note that the sum of small SHAP values of all variables determines the prediction towards one class or another. Therefore, the final prediction is the sum of the small contributions of several features. Further details regarding the influence of each variable on the final prediction can be found in Figures S3-S4 of the supplementary material on the associated web page.

4.2. Local explanations

Clustering was performed to group the children and the variables based on their SHAP values. The clustering was initially performed only in the children used to generate the final system (RF). Clustering on SHAP values allowed us to identify subgroups of both variables and individuals. Children clustering enables the grouping of children who are predicted similarly, with comparable SHAP value patterns. Variable clustering helps to identify which variables have similar patterns of contributions to predicting whether a child is IR or non-IR. Examining Fig. 4, four clusters of children are identifiable. Further, it appears that the first two children clusters can be distinguished by the SHAP values corresponding to *HDAC4* methylation. The first cluster pertains to IR children, while the second cluster includes non-IR children. The second cluster emphasizes the predictive significance of *HDAC4* methylation and other methylation patterns by improving the accuracy of the final classifier (RF). These methylation patterns improve the accuracy of the prediction by refining it; because relying solely on anthropometric (BMI z-score) and adipokine information (leptin/adiponectin ratio) would not result in accurate predictions in the children belonging to the second cluster. The predictive enhancement of the methylation patterns is due to the presence of children with a high BMI z-score and a high leptin/adiponectin ratio. Consequently, the children in the second cluster exhibit positive SHAP values, for their anthropometric (BMI z-score) and adipokine information (leptin/adiponectin ratio), despite belonging to the negative class (non-IR). The behavior of the next two minority and heterogeneous clusters is challenging to explain.

During the undersampling process to generate the final classifier, certain negative class examples were excluded. As a result, the final classifier failed to predict nine out of the thirty-eight children that were set aside during undersampling. To study RF behavior in children who did not participate in the training process, the SHAP values were clustered similarly in these children to observe similarities among them. Fig. 5 shows the results, which reveal three distinct clusters of children. Intriguingly, the second and third clusters of children in Fig. 4 appear to be similar to the first and third clusters of children in Fig. 5. These nine children, who were predicted incorrectly by RF and belong to the second cluster of children, are characterized by positive SHAP values for the first and third clusters of variables. The first cluster of variables is formed by *HDAC4* methylation, while the second one is formed by leptin/adiponectin ratio, BMI z-score, leptin, and *RASGRF1* and *PTPRN2* methylation.

To comprehend why RF was unsuccessful in predicting the pubertal IR status of these children, who have positive SHAP values on the mentioned variables, we analyzed the phenotypic traits of these children. Curiously, these nine children had a high BMI z-score and unhealthy adipokines profile, some of them were even IR, in their pre-pubertal stage. They experienced an observable weight loss in their

BMI z-scores during puberty, thus improving their IR parameters, see Fig. 6. However, these children present HOMA-IR index values that are close to the established cut-off point for IR. Therefore, they are individuals with a high metabolic risk and they might likely develop IR over time if they remain in the same metabolic condition, see Figure S5-S8 of the supplementary material on the associated web page. To explore our hypothesis more deeply, we generated local explanations for one of the nine children. The local explanations revealed that the variables contributing significantly to the system's prediction were the leptin/adiponectin ratio and BMI z-score, see Fig. 7.

With these examples, we show how using SHAP values enables us to gain a better understanding of our final system's behavior and why it makes errors in specific cases. The identification of misclassified children as at-risk individuals underscores the final classifier's usefulness. In future works, it would be beneficial to explore the information provided by SHAP explanations in conjunction with other explanations, such as counterfactual. These explanations may help human experts to become familiar with unknown processes by understanding the hypothetical input conditions under which the model's prediction for a patient changes [90,91].

The principal strengths of our study lie in our ability to generate a system with a relevant predictive ability and in our comprehensive exploration of both global and local levels, which has led to the identification of promising biomarkers of IR in our population. The system highlights the predictive importance of classical markers, such as BMI z-score or leptin/adiponectin ratio, and novel ones such as methylation patterns of IR-relevant genes, such as *HDAC4*, *PTPRN2*, *MATN2*, *RASGRF1* and *EBF1*. Our findings highlight the importance of integrating multi-omics data and following XAI trends when building clinical DSS [92].

As a main limitation, we have been unable to use automatic feature selection methods, which has been a methodological limitation argued in our limited sample size. It is also necessary to validate the predictive capacity of the system used through a validation study in an independent population. Furthermore, our system uses omic features that are not easily quantifiable in the clinical setting due to their high cost (e.g., DNA methylation), which hinder its straightforward implementation as a clinical DSS. Noteworthy, this work supports the predictive utility of using epigenetic features in clinical settings. In the future, the development of high-throughput technology may result in lower costs and greater clinical accessibility to epigenetic data.

5. Conclusions

There is an urgent need to implement early life prediction programs that include the use of DSS to address childhood obesity early on to prevent the worsening of health status, before pubertal IR occurs. In this sense, this study presents an accurate and understandable system derived from a longitudinal cohort of 90 children to predict pubertal IR according to multi-omic and clinical data from the pre-pubertal stage. A comprehensive analysis of global and local explanations produced from the SHAP values has allowed us to identify the variables with the greatest influence on the system's predictions, to determine the direction of association of each of them, and to differentiate subgroups of children according to the risk factors they present. This analysis has highlighted the relevance of contrasted markers such as BMI z-score and leptin/adiponectin ratio along with emerging epigenetic biomarkers such as *HDAC4* and *PTPRN2* methylation. In the future, it would be of great interest to identify the exposures responible for these methylation patterns, by adding a new layer of environmental data into the classifier. Likewise, future research should investigate the predictive utility of promising transcriptomic, proteomic or metabolomic signatures individually or in combination with other omics for the prediction of pubertal IR.

Our approach denotes the potential of XAI techniques to address the challenges of multi-omic analysis, improving the understanding of disease pathophysiology by enabling predictive analysis using different layers of information. This scientific and technological advancement into expected to lead to the integration of omic platforms in clinical practice, along with the use of clinical DSS co-managed by physicians. In the long term future, the responsible use of clinical DSS that integrates multiple data layers could become a standard monitoring tool for the clinical management of childhood obesity, approaching the everyday clinical practice to a more precise medicine.

CRediT authorship contribution statement

Álvaro Torres-Martos: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Augusto Anguita-Ruiz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mireia Bustos-Aibar:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alberto Ramírez-Mena:** Methodology, Investigation, Formal analysis. **María Arteaga:** Formal analysis. **Gloria Bueno:** Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Rosaura Leis:** Project administration, Methodology, Investigation, Funding acquisition. **Concepción M. Aguilera:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Rafael Alcalá:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Jesús Alcalá-Fdez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Instituto de Salud Carlos III co-funded by the European Union and ERDF A way of making Europe (grant numbers PI20/00711, PI20/00563, PI20/00924, P20/00988, PI23/00028, PI23/00129, PI23/01032, PI23/00165 and also PI23/00191), and by the European Union through the Horizon Europe Framework Programme (eprObes project, grant number GA 101080219). The authors also acknowledge Instituto de Salud Carlos III for personal funding of A.A.R, A.T.M and M.B.A.: i-PFIS and PFIS contracts: IIS doctorates - company in health sciences and technologies of the Strategic Health Action (IFI17/00048, IFI22/00013 and FI23/00042). We also thank the support from the grant FJC2021-046952-I by Ministerio de Ciencia, Innovación y Universidades y Agencia Estatal de Investigación.. Funding for open access charge: Universidad de Granada/CBUA.

Code availability

To promote open access and reproducibility, the code is publicly available on Github: https://github.com/AlvaroTorresMartos/IR_prediction.

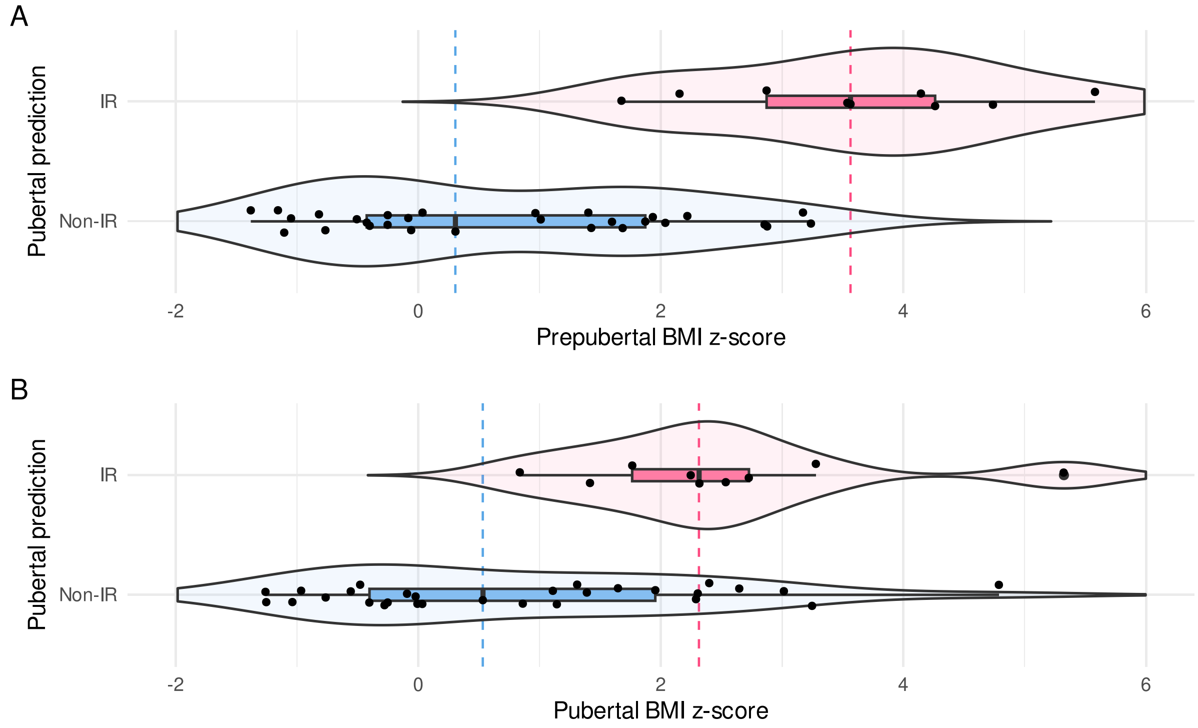

Fig. 6. The violin plots display BMI z-score along the x-axis and pubertal prediction along the y-axis, where blue indicates non-IR and pink denotes IR. As a result, nine misclassified children are highlighted in pink. Fig. 6A and B differentiate BMI z-scores for pre-pubertal and pubertal stages, correspondingly. The difference between the two pink lines, which indicates the average pre-pubertal and pubertal BMI z-score, suggests weight loss as the average BMI z-score decreases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

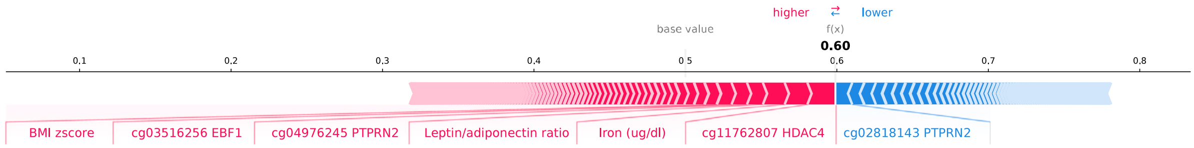

Fig. 7. Force plot of a child belonging to the nine children misclassified as IR (local explanation). The most influential variables in the final prediction are labeled, such as leptin/adiponectin ratio and BMI z-score among others.

References

[1] Jayedi A, Soltani S, Zargar MS, Khan TA, Shab-Bidar S. Central fatness and risk of all cause mortality: systematic review and dose-response meta-analysis of 72 prospective cohort studies. BMJ 2020;23(370):m3324. http://dx.doi.org/10.1136/bmj.m3324.

[2] Lister NB, Baur LA, Felix JF, Hill AJ, Marcus C, Reinehr T, Summerbell C, Wabitsch M. Child and adolescent obesity. Nat Rev Dis Prim 2023;9(1). http://dx.doi.org/10.1038/s41572-023-00435-4.

[3] Kim MS, Kim WJ, Khera AV, Kim JY, Yon DK, Lee SW, Shin JI, Won H-H. Association between adiposity and cardiovascular outcomes: an umbrella review and meta-analysis of observational and Mendelian randomization studies. Eur Heart J 2021;42(34):3388–403. http://dx.doi.org/10.1093/eurheartj/ehab454.

[4] Jacobs DR, Woo JG, Sinaiko AR, Daniels SR, Ikonen J, Juonala M, Kartiosuo N, Lehtimäki T, Magnussen CG, Viikari JS, et al. Childhood cardiovascular risk factors and adult cardiovascular events. New Engl J Med 2022;386(20):1877–88. http://dx.doi.org/10.1056/nejmoa2109191.

[5] Magnussen C, Ojeda FM, Leong DP, Alegre-Diaz J, Amouyel P, Aviles-Santa L, De Bacquer D, Ballantyne CM, Bernabé-Ortiz A, Bobak M, et al. Global effect of modifiable risk factors on cardiovascular disease and mortality. New Engl J Med 2023;389(14):1273–85. http://dx.doi.org/10.1056/nejmoa2206916.

[6] Lamas C, Kalen A, Anguita-Ruiz A, Perez-Ferreiros A, Picans-Leis R, Flores K, Moreno L, Bueno G, Gil A, Gil-Campos M, et al. Progression of metabolic syndrome and associated cardiometabolic risk factors from prepuberty to puberty in children: The PUBMEP study. Front Endocrinol 2022;13:1082684. http://dx.doi.org/10.3389/fendo.2022.1082684.

[7] James DE, Stöckli J, Birnbaum MJ. The aetiology and molecular landscape of insulin resistance. Nat Rev Mol Cell Biol 2021;22(11):751–71. http://dx.doi.org/10.1038/s41580-021-00390-6.

[8] Hannon TS, Janosky J, Arslanian SA. Longitudinal study of physiologic insulin resistance and metabolic changes of puberty. Pediatr Res 2006;60(6):759–63. http://dx.doi.org/10.1203/01.pdr.0000246097.73031.27.

[9] González-Gil EM, Anguita-Ruiz A, Kalén A, De Las Lamas Perez C, Rupérez AI, Vázquez-Cobela R, Flores K, Gil A, Gil-Campos M, Bueno G, et al. Longitudinal associations between cardiovascular biomarkers and metabolic syndrome during puberty: the PUBMEP study. Eur J Pediatr 2023;182(1):419–29. http://dx.doi.org/10.1007/s00431-022-04702-6.

[10] Reinehr T, Roth CL. Is there a causal relationship between obesity and puberty? Lancet Child Adolesc Health 2019;3(1):44–54. http://dx.doi.org/10.1016/S2352-4642(18)30306-7.

[11] Fernandez-Jimenez R, Al-Kazaz M, Jaslow R, Carvajal I, Fuster V. Children present a window of opportunity for promoting health: JACC review topic of the week. J Am Coll Cardiol 2018;72(25):3310–9. http://dx.doi.org/10.1016/j.jacc.2018.10.031.

[12] Hannon TS, Arslanian SA. Obesity in adolescents. New Engl J Med 2023;389(3):251–61. http://dx.doi.org/10.1056/nejmcp2102062.

[13] Tsatsoulis A, Paschou SA. Metabolically healthy obesity: Criteria, epidemiology, controversies, and consequences. Curr Obes Rep 2020;9(2):109–20. http://dx.doi.org/10.1007/s13679-020-00375-0.

[14] April-Sanders AK, Rodriguez CJ. Metabolically healthy obesity redefined. JAMA Netw Open 2021;4(5):e218860. http://dx.doi.org/10.1001/jamanetworkopen.2021.8860.

[15] Hampl SE, Hassink SG, Skinner AC, Armstrong SC, Barlow SE, Bolling CF, Avila Edwards KC, Eneli I, Hamre R, Joseph MM, et al. Clinical practice guideline for the evaluation and treatment of children and adolescents with obesity. Pediatrics 2023;151(2). http://dx.doi.org/10.1542/peds.2022-060640.

[16] Aleksandrova K, Egea Rodrigues C, Floegel A, Ahrens W. Omics biomarkers in obesity: Novel etiological insights and targets for precision prevention. Curr Obes Rep 2020;9(3):219–30. http://dx.doi.org/10.1007/s13679-020-00393-y.

[17] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17. http://dx.doi.org/10.1038/s41746-020-0221-y.

[18] Anguita-Ruiz A, Torres-Martos Á, Ruiz-Ojeda F, Alcalá-Fdez J, Bueno G, Gil-Campos M, Roa-Rivas J, Moreno L, Gil A, Leis R, et al. Integrative analysis of blood cells DNA methylation, transcriptomics and genomics identifies novel epigenetic regulatory mechanisms of insulin resistance during puberty in children with obesity. 2022, http://dx.doi.org/10.1101/2022.12.13.22283415, medRxiv. URL https://www.medrxiv.org/content/early/2022/12/16/2022.12.13.22283415.

[19] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New Engl. J. Med. 2019;380(14):1347–58. http://dx.doi.org/10.1056/nejmra1814259.

[20] Schüssler-Fiorenza Rose SM, Contrepois K, Moneghetti KJ, Zhou W, Mishra T, Mataraso S, Dagan-Rosenfeld O, Ganz AB, Dunn J, Hornburg D, et al. A longitudinal big data approach for precision health. Nat. Med. 2019;25(5):792–804. http://dx.doi.org/10.1038/s41591-019-0414-6.

[21] Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, Zhang MJ, Rao V, Avina M, Mishra T, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature 2019;569(7758):663–71. http://dx.doi.org/10.1038/s41586-019-1236-x.

[22] Llorente-Cantarero FJ, Aguilar-Gómez FJ, Anguita-Ruiz A, Rupérez AI, Vázquez-Cobela R, Flores-Rojas K, Aguilera CM, Gonzalez-Gil EM, Gil-Campos M, Bueno-Lozano G, et al. Changes in physical activity patterns from childhood to adolescence: Genobox longitudinal study. Int. J. Environ. Res. Public Health 2020;17(19):7227. http://dx.doi.org/10.3390/ijerph17197227.

[23] Kushwaha S, Srivastava R, Jain R, Sagar V, Aggarwal AK, Bhadada SK, Khanna P. Harnessing machine learning models for non-invasive pre-diabetes screening in children and adolescents. Comput Methods Programs Biomed 2022;226:107180. http://dx.doi.org/10.1016/j.cmpb.2022.107180.

[24] Khan MS, Cuda S, Karere GM, Cox LA, Bishop AC. Breath biomarkers of insulin resistance in pre-diabetic hispanic adolescents with obesity. Sci Rep 2022;12(1). http://dx.doi.org/10.1038/s41598-021-04072-3.

[25] Castelvecchi D. Can we open the black box of AI? Nature 2016;538(7623):20–3. http://dx.doi.org/10.1038/538020a.

[26] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[27] Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. New Engl. J. Med. 2023;388(13):1201–8. http://dx.doi.org/10.1056/nejmra2302038.

[28] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2(10):749–60. http://dx.doi.org/10.1038/s41551-018-0304-0.

[29] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15(90):3133–81. http://dx.doi.org/10.5555/2627435.2697065.

[30] Torres-Martos Á, Anguita-Ruiz A, Bustos-Aibar M, Cámara-Sánchez S, Alcalá R, Aguilera CM, Alcalá-Fdez J. Human multi-omics data pre-processing for predictive purposes using machine learning: A case study in childhood obesity. In: Bioinformatics and biomedical engineering. Lecture notes in computer science, Springer International Publishing; 2022, p. 359–74. http://dx.doi.org/10.1007/978-3-031-07802-6_31.

[31] Torres-Martos Á, Bustos-Aibar M, Ramírez-Mena A, Cámara-Sánchez S, Anguita-Ruiz A, Alcalá R, Aguilera CM, Alcalá-Fdez J. Omics data preprocessing for machine learning: A case study in childhood obesity. Genes 2023;14(2):248. http://dx.doi.org/10.3390/genes14020248.

[32] Levy-Marchal C, Arslanian S, Cutfield W, Sinaiko A, Druet C, Marcovecchio ML, Chiarelli F. Insulin resistance in children: Consensus, perspective, and future directions. J Clin Endocrinol Metab 2010;95(12):5189–98. http://dx.doi.org/10.1210/jc.2010-1047.

[33] Anguita-Ruiz A, Mendez-Gutierrez A, Ruperez AI, Leis R, Bueno G, Gil-Campos M, Tofe I, Gomez-Llorente C, Moreno LA, Gil Á, et al. The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity. Metabolism 2020;105:154187. http://dx.doi.org/10.1016/j.metabol.2020.154187.

[34] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559–75. http://dx.doi.org/10.1086/519795.

[35] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10(2):giab008. http://dx.doi.org/10.1093/gigascience/giab008.

[36] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. Bioinformatics 2013;29(2):189–96. http://dx.doi.org/10.1093/bioinformatics/bts680.

[37] Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinform 2010;11(1):587. http://dx.doi.org/10.1186/1471-2105-11-587.

[38] Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the illumina infinium platform. BMC Bioinform 2012;13(1):59. http://dx.doi.org/10.1186/1471-2105-13-59.

[39] Xie C, Leung Y-K, Chen A, Long D-X, Hoyo C, Ho S-M. Differential methylation values in differential methylation analysis. Bioinformatics 2019;35(7):1094–7. http://dx.doi.org/10.1093/bioinformatics/bty778.

[40] Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8. http://dx.doi.org/10.1093/bioinformatics/btr597.

[41] Pfeifer B, Holzinger A, Schimek MG. Robust random forest-based all-relevant feature ranks for trustworthy AI. Stud Health Technol Inform 2022;294:137—138. http://dx.doi.org/10.3233/shti220418.

[42] Maitre L, Guimbaud J-B, Warembourg C, Güil-Oumrait N, Petrone PM, Chadeau-Hyam M, Vrijheid M, Basagaña X, Gonzalez JR. State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. Environ Int 2022;168:107422. http://dx.doi.org/10.1016/j.envint.2022.107422.

[43] Bustos-Aibar M, Aguilera CM, Alcalá-Fdez J, Ruiz-Ojeda FJ, Plaza-Díaz J, Plaza-Florido A, Tofe I, Gil-Campos M, Gacto MJ, Anguita-Ruiz A. Shared gene expression signatures between visceral adipose and skeletal muscle tissues are associated with cardiometabolic traits in children with obesity. Comput Biol Med 2023;163:107085. http://dx.doi.org/10.1016/j.compbiomed.2023.107085.

[44] Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL. DNA methylation-based predictors of health: applications and statistical considerations. Nat Rev Genet 2022;23(6):369–83. http://dx.doi.org/10.1038/s41576-022-00465-w.

[45] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucl Acids Res 2019;47(D1):D1005–12. http://dx.doi.org/10.1093/nar/gky1120.

[46] Battram T, Yousefi P, Crawford G, Prince C, Sheikhali Babaei M, Sharp G, Hatcher C, Vega-Salas MJ, Khodabakhsh S, Whitehurst O, et al. The EWAS catalog: a database of epigenome-wide association studies. Wellcome Open Res 2022;7:41. http://dx.doi.org/10.12688/wellcomeopenres.17598.2.

[47] Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42(2):105–16. http://dx.doi.org/10.1038/ng.520.

[48] Lotta LA, Gulati P, Day FR, Payne F, Ongen H, van de Bunt M, Gaulton KJ, Eicher JD, Sharp SJ, Luan J, et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. Nat Genet 2017;49(1):17–26. http://dx.doi.org/10.1038/ng.3714.

[49] Kotnik P, Knapič E, Kokošar J, Kovač J, Jerala R, Battelino T, Horvat S. Identification of novel alleles associated with insulin resistance in childhood obesity using pooled-DNA genome-wide association study approach. Int J Obes (London) 2018;42(4):686–95. http://dx.doi.org/10.1038/ijo.2017.293.

[50] González-Martín JM, Torres-Mata LB, Cazorla-Rivero S, Fernández-Santana C, Gómez-Bentolila E, Clavo B, Rodríguez-Esparragón F. An artificial intelligence prediction model of insulin sensitivity, insulin resistance, and diabetes using genes obtained through differential expression. Genes (Basel) 2023;14(12):2119. http://dx.doi.org/10.3390/genes14122119.

[51] Verdonck T, Baesens B, Óskarsdóttir M, vanden Broucke S. Special issue on feature engineering editorial. Mach Learn 2021;113(7):3917–28. http://dx.doi.org/10.1007/s10994-021-06042-2.

[52] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365. http://dx.doi.org/10.1371/journal.pone.0224365.

[53] López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inform Sci 2013;250:113–41. http://dx.doi.org/10.1016/j.ins.2013.07.007.

[54] Hvitfeldt E. themis: Extra recipes steps for dealing with unbalanced data. 2023, URL https://CRAN.R-project.org/package=themis. R package version 1.0.2.

[55] Salzberg SL. C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn 1994;16(3):235–40. http://dx.doi.org/10.1007/BF00993309.

[56] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. 2022, URL https://CRAN.R-project.org/package=e1071. R package version 1.7-11.

[57] Liaw A, Wiener M. Classification and regression by randomforest. R News 2002;2(3):18–22, URL https://CRAN.R-project.org/doc/Rnews/.

[58] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, et al. xgboost: Extreme gradient boosting. 2022, URL https://CRAN.R-project.org/package=xgboost. r package version 1.6.0.1..

[59] Karatzoglou A, Smola A, Hornik K. kernlab: Kernel-based machine learning lab. 2022, URL https://CRAN.R-project.org/package=kernlab. R package version 0.9-31.

[60] Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 package for kernel methods in R. J Stat Softw 2004;11(9):1–20. http://dx.doi.org/10.18637/jss.v011.i09.

[61] Venables WN, Ripley BD. Modern applied statistics with S. 4th ed.. New York: Springer; 2002, ISBN 0-387-95457-0.

[62] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26. http://dx.doi.org/10.18637/jss.v028.i05.

[63] Carrington AM, Manuel DG, Fieguth PW, Ramsay T, Osmani V, Wernly B, Bennett C, Hawken S, Magwood O, Sheikh Y, McInnes M, Holzinger A. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Trans Pattern Anal Mach Intell 2023;45(1):329–41. http://dx.doi.org/10.1109/tpami.2022.3145392.

[64] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23(1):40–55. http://dx.doi.org/10.1038/s41580-021-00407-0.

[65] Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nature Rev Genet 2022;23(3):169–81. http://dx.doi.org/10.1038/s41576-021-00434-9.

[66] Ramírez-Mena A, Andrés-León E, Alvarez-Cubero MJ, Anguita-Ruiz A, Martinez-Gonzalez LJ, Alcala-Fdez J. Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. Comput Methods Programs Biomed 2023;240:107719. http://dx.doi.org/10.1016/j.cmpb.2023.107719.

[67] Lombardi A, Arezzo F, Di Sciascio E, Ardito C, Mongelli M, Di Lillo N, Fascilla FD, Silvestris E, Kardhashi A, Putino C, et al. A human-interpretable machine learning pipeline based on ultrasound to support leiomyosarcoma diagnosis. Artif Intell Med 2023;146:102697. http://dx.doi.org/10.1016/j.artmed.2023.102697.

[68] Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, Distefano M, Senol-Cosar O, Haas ME, Bick A, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. Cell 2019;177(3):587–96. http://dx.doi.org/10.1016/j.cell.2019.03.028.

[69] Ling C, Rönn T. Epigenetics in human obesity and type 2 diabetes. Cell Metab 2019;29(5):1028–44. http://dx.doi.org/10.1016/j.cmet.2019.03.009.

[70] Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735–46. http://dx.doi.org/10.1016/j.csbj.2021.06.030.

[71] Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform 2020;21(2):541–52. http://dx.doi.org/10.1093/bib/bbz015.

[72] Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature Commun 2021;12(1):124. http://dx.doi.org/10.1038/s41467-020-20430-7.

[73] Herrmann M, Probst P, Hornung R, Jurinovic V, Boulesteix A-L. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform 2021;22(3):bbaa167. http://dx.doi.org/10.1093/bib/bbaa167.

[74] Rönn T, Volkov P, Davegardh C, Dayeh T, Hall E, Olsson AH, Nilsson E, Tornberg A, Dekker Nitert M, Eriksson K-F, et al. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLoS Genet. 2013;9(6):e1003572. http://dx.doi.org/10.1371/journal.pgen.1003572.

[75] Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, Griffiths L, Hoffman EP, Stubbs RS, Macartney-Coxson D. An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. Genome Biol 2015;16(1):8. http://dx.doi.org/10.1186/s13059-014-0569-x.

[76] Lenoir O, Flosseau K, Ma FX, Blondeau B, Mai A, Bassel-Duby R, Ravassard P, Olson EN, Haumaitre C, Scharfmann R. Specific control of pancreatic endocrine β- and δ-cell mass by class IIa histone deacetylases HDAC4, HDAC5, and HDAC9. Diabetes 2011;60(11):2861–71. http://dx.doi.org/10.2337/db11-0440.

[77] Samblas M, Milagro FI, Mansego ML, Marti A, Martinez JA, GENOI members. PTPRS and PER3 methylation levels are associated with childhood obesity: results from a genome-wide methylation analysis. Pediatr Obes 2018;13(3):149–58. http://dx.doi.org/10.1111/ijpo.12224.

[78] Li Y, Zhou Y, Zhu L, Liu G, Wang X, Wang J, Wang J, You L, Ji C, Guo X, et al. Genome-wide analysis reveals that altered methylation in specific CpG loci is associated with childhood obesity. J Cell Biochem 2018;119(9):7490–7. http://dx.doi.org/10.1002/jcb.27059.

[79] Juvinao-Quintero DL, Marioni RE, Ochoa-Rosales C, Russ TC, Deary IJ, van Meurs JBJ, Voortman T, Hivert M-F, Sharp GC, Relton CL, et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin. Epigenet. 2021;13(1):40. http://dx.doi.org/10.1186/s13148-021-01027-3.

[80] Abu-Farha M, Tiss A, Abubaker J, Khadir A, Al-Ghimlas F, Al-Khairi I, Baturcam E, Cherian P, Elkum N, Hammad M, et al. Proteomics analysis of human obesity reveals the epigenetic factor HDAC4 as a potential target for obesity. PLoS One 2013;8(9):e75342. http://dx.doi.org/10.1371/journal.pone.0075342.

[81] Lan MS, Wasserfall C, Maclaren NK, Notkins AL. IA-2, a transmembrane protein of the protein tyrosine phosphatase family, is a major autoantigen in insulin-dependent diabetes mellitus. Proc Natl Acad Sci USA 1996;93(13):6367–70. http://dx.doi.org/10.1073/pnas.93.13.6367.

[82] Lu J, Li Q, Xie H, Chen ZJ, Borovitskaya AE, Maclaren NK, Notkins AL, Lan MS. Identification of a second transmembrane protein tyrosine phosphatase, IA-2beta, as an autoantigen in insulin-dependent diabetes mellitus: precursor of the 37-kDa tryptic fragment. Proc Natl Acad Sci USA 1996;93(6):2307–11. http://dx.doi.org/10.1073/pnas.93.6.2307.

[83] Lee S. The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. Sci Rep 2019;9(4855). http://dx.doi.org/10.1038/s41598-019-40486-w.

[84] Yang IV, Zhang W, Davidson EJ, Fingerlin TE, Kechris K, Dabelea D. Epigenetic marks of in utero exposure to gestational diabetes and childhood adiposity outcomes: the EPOCH study. Diabet. Med. 2018;35(5):612–20. http://dx.doi.org/10.1111/dme.13604.

[85] Sasaki A, Murphy KE, Briollais L, McGowan PO, Matthews SG. DNA methylation profiles in the blood of newborn term infants born to mothers with obesity. PLoS One 2022;17(5):e0267946. http://dx.doi.org/10.1371/journal.pone.0267946.

[86] Weng X, Liu F, Zhang H, Kan M, Wang T, Dong M, Liu Y. Genome-wide DNA methylation profiling in infants born to gestational diabetes mellitus. Diabetes Res. Clin. Pract. 2018;142:10–8. http://dx.doi.org/10.1016/j.diabres.2018.03.016.

[87] Awamleh Z, Butcher DT, Hanley A, Retnakaran R, Haertle L, Haaf T, Hamilton J, Weksberg R. Exposure to gestational diabetes mellitus (GDM) alters DNA methylation in placenta and fetal cord blood. Diabetes Res. Clin. Pract. 2021;174:108690. http://dx.doi.org/10.1016/j.diabres.2021.108690.

[88] Frithioff-Bø jsøe C, Lund MA, Lausten-Thomsen U, Hedley PL, Pedersen O, Christiansen M, Baker JL, Hansen T, Holm J-C. Leptin, adiponectin, and their ratio as markers of insulin resistance and cardiometabolic risk in childhood obesity. Pediatr. Diabetes 2020;21(2):194–202. http://dx.doi.org/10.1111/pedi.12964.

[89] Hilton C, Sabaratnam R, Drakesmith H, Karpe F. Iron, glucose and fat metabolism and obesity: an intertwined relationship. Int. J. Obes. (London) 2023;47(7):554–63. http://dx.doi.org/10.1038/s41366-023-01299-0.

[90] Del Ser J, Barredo-Arrieta A, Díaz-Rodríguez N, Herrera F, Saranti A, Holzinger A. On generating trustworthy counterfactual explanations. Inform Sci 2024;655:119898. http://dx.doi.org/10.1016/j.ins.2023.119898.

[91] Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. Artif Intell Med 2024;150:102819. http://dx.doi.org/10.1016/j.artmed.2024.102819.

[92] Göndöcs D, Dörfler V. AI in medical diagnosis: AI prediction and human judgment. Artif Intell Med 2024;149:102769. http://dx.doi.org/10.1016/j.artmed.2024.102769.