Heliyon 9 (2023) e19444

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Machine learning based analysis for intellectual disability in Down syndrome

Federico Baldo [a,1], Allison Piovesan [b,1], Marijana Rakvin [a], Giuseppe Ramacieri [c], Chiara Locatelli [d], Silvia Lanfranchi [e], Sara Onnivello [e], Francesca Pulina [e], Maria Caracausi [b], Francesca Antonaros [b], Michele Lombardi [a,**], Maria Chiara Pelleri [b,*]

[a] Department of Computer Science and Engineering, University of Bologna, Viale Risorgimento 2, 40136, Bologna, BO, Italy
[b] Department of Biomedical and Neuromotor Sciences (DIBINEM), University of Bologna, Via Massarenti 9, 40138, Bologna, BO, Italy
[c] Department of Medical and Surgical Sciences (DIMEC), University of Bologna, Via Massarenti 9, 40138, Bologna, BO, Italy
[d] Neonatology Unit, IRCCS University General Hospital Sant'Orsola Polyclinic, Via Massarenti 9, 40138, Bologna, BO, Italy
[e] Department of Developmental Psychology and Socialisation, University of Padova, Via Venezia 8, 35131, Padua, PD, Italy

ARTICLE INFO

Keywords:
Down syndrome
Intellectual disability
Data mining
Machine learning

ABSTRACT

Down syndrome (DS) or trisomy 21 is the most common genetic cause of intellectual disability (ID), but a pathogenic mechanism has not been identified yet. Studying a complex and not monogenic condition such as DS, a clear correlation between cause and effect might be difficult to find through classical analysis methods, thus different approaches need to be used. The increased availability of big data has made the use of artificial intelligence (AI) and in particular machine learning (ML) in the medical field possible.

The purpose of this work is the application of ML techniques to provide an analysis of clinical records obtained from subjects with DS and study their association with ID.

We have applied two tree-based ML models (random forest and gradient boosting machine) to the research question: how to identify key features likely associated with ID in DS. We analyzed 109 features (or variables) in 106 DS subjects. The outcome of the analysis was the age equivalent (AE) score as indicator of intellectual functioning, impaired in ID. We applied several methods to configure the models: feature selection through Boruta framework to minimize random correlation; data augmentation to overcome the issue of a small dataset; age effect mitigation to take into account the chronological age of the subjects.

The results show that ML algorithms can be applied with good accuracy to identify variables likely involved in cognitive impairment in DS. In particular, we show how random forest and gradient boosting machine produce results with low error (MSE <0.12) and an acceptable R^2 (0.70 and 0.93). Interestingly, the ranking of the variables point to several features of interest related to hearing, gastrointestinal alterations, thyroid state, immune system and vitamin B12 that can be considered with particular attention for improving care pathways for people with DS.

* Corresponding author.Department of Biomedical and Neuromotor Sciences (DIBINEM), University of Bologna, via Massarenti, 9, 40138, Bologna, BO, Italy.
** Corresponding author. Department of Computer Science and Engineering, University of Bologna, Viale Risorgimento 2, 40136, Bologna, BO, Italy.
E-mail addresses: michele.lombardi2@unibo.it (M. Lombardi), mariachiara.pelleri2@unibo.it (M.C. Pelleri).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2023.e19444
Received 2 February 2023; Received in revised form 19 July 2023; Accepted 23 August 2023
Available online 27 August 2023

In conclusion, ML-based model may assist researchers in identifying key features likely correlated with ID in DS, and ultimately, may improve research efforts focused on the identification of possible therapeutic targets and new care pathways. We believe this study can be the basis for further testing/validating of our algorithms with multiple and larger datasets.

1. Introduction

Down syndrome (DS) or trisomy 21 [1] is the most common genetic alteration being observed in 1 in 1000–1100 live births around the world [2]. It is caused by the presence of an extra full or partial copy of chromosome 21 [3]. The most constant and typical features of DS are intellectual disability (ID) and craniofacial dysmorphisms, together with other variable signs and symptoms, such as cardiac malformations and growth delay [4–9]. In particular, ID is present to some degree of severity in virtually all individuals with DS [2,10] and represents the most relevant clinical symptom for which a pathogenic mechanism has not been identified yet [11].

A great effort has been made to investigate the hypothesis of DS as a metabolic disease, firstly proposed by Lejeune [12]. Many studies identified specific metabolic alterations in DS plasma [13–15] reviewed in Ref. [16], also reporting a cognitive development influence by homocysteine [17]. However, a recent study failed to confirm a specific association of the DS metabolic profile with cognitive aspects [18]. In fact, studying a complex and not monogenic disease such as DS, a clear correlation between cause and effect might be difficult to find through classical analysis methods, thus different approaches need to be used.

In this context, the increasing availability of big data and of electronic health records has paved the way for Artificial Intelligence (AI) in the medical field. Most notably, Machine learning (ML) has been one of the leading approaches to bring new perspectives. The reason behind the use of ML-based approaches, instead of more classical methods such as Linear Regressions and Logistic Regressions, is that they can capture non-linear correlations in the data. This can lead to uncovering unexpected information that would not be observable by human analysts or more canonical statistical approaches. This technique is increasingly employed as it has the potential both to offer significant advantages to health professionals, through effective tools for data representation, and to assist medical doctors in order to diagnose and accurately predict the risk of disease [19].

Accurate analyses through ML techniques can give significant contribution in the medical field and have also been applied in the context of DS. In particular, several studies have focused with promising results on computerised diagnosis by analysing facial images, taking advantage of craniofacial dysmorphisms typical of DS [20–22]. Among others, Koivu and coll [23]. evaluated the use of ML algorithms for improved risk assessment for DS in prenatal screening also considering first trimester biochemical measurements. Regarding the cognitive profile, other studies have tried to build intelligent tutoring systems with the aim of assessing the skills of DS children and helping to improve their progress and learning skills [24–26]. Finally, several works focused on DS using the great quantity of data available from mouse models, predicting locomotor activity and identifying critical proteins from protein profiles generated from mice cortex through different AI approaches [27–29]. Although the results obtained were accurate and promising, the lack of homology of relevant chromosome 21 sequences and the impossibility to measure human superior cognitive functions in mice needs to be taken into account in the investigation of a pathogenic mechanism of ID in DS.

The purpose of this work is the application of ML techniques in order to find a model to analyze clinical data of DS subjects and study their association with ID. In particular, our goal is to identify features that are highly (and possibly non-linearly) correlated with ID scores. By doing so, we may assist researchers in identifying key features likely associated with ID in DS, and ultimately, may improve research efforts focused on the identification of possible therapeutic targets and new care pathways.

In particular, we have applied two tree-based ML models (Random Forest and XGBoost) to the research question: how to identify key features likely associated with ID in DS. We analyzed 109 features (or variables) in 106 DS subjects. The outcome of the analysis was the age equivalent (AE) score as indicator of intellectual functioning, impaired in ID. We applied several methods to configure the models: feature selection through Boruta framework to minimize random correlation; data augmentation to overcome the issue of a small dataset; age effect mitigation to take into account the chronological age of the subjects.

2. Materials and Methods

2.1. Dataset and preprocessing

The dataset used in this study was obtained during routine follow up visits provided for DS at the Unit of Neonatology of IRCCS University General Hospital Sant'Orsola Polyclinic in Bologna, Italy from February 2014 to July 2019. Inclusion criteria were diagnosis of DS with homogeneous or mosaic trisomy 21 and age >2 years (yrs). The Independent Ethics Committee of the IRCCS University General Hospital Sant'Orsola Polyclinic of Bologna Italy has granted the ethical approval for this study (number: 39/2013/U/Tess). For all participants involved in the present study, written informed consent was obtained from the subjects themselves if over 18 years of age or parents and/or legal guardians if under 18 years of age, according to the approved protocol for the collection of urine and blood samples and clinical data. All procedures were carried out in accordance with the Ethical Principles for Medical Research Involving Human Subjects of the Helsinki Declaration.

The DS clinical dataset here reported in Supplementary Table 1 is composed of a collection of anonymized personal records regarding genetic, diagnostic, clinical, and auxological information. For details regarding the recorded data, we refer to previous works where subjects are indicated with the same DS subject code [11,18,30,31].

Table 1
Dataset variables descriptions, names and types.

Description	Variable	Type
Blood test data	Leukocytes (10^3/mmc)	Continuous
	Erythrocytes (10^6/mmc)	Continuous
	HGB-Hemoglobin (g/dL)	Continuous
	Hematocrit (%)	Continuous
	MCV-Mean corpuscular volume (fL)	Continuous
	MCH-Mean corpuscular hemoglobin (pg)	Continuous
	MCHC-Mean corpuscular hemoglobin concentration (g/dL)	Continuous
	RDW-Red blood cells distribution width (%)	Continuous
	HDW-Hemoglobin distribution width (g/dL)	Continuous
	Neutrophils (10^3/mmc)	Continuous
	Lymphocytes (10^3/mmc)	Continuous
	Monocytes (10^3/mmc)	Continuous
	Eosinophils (10^3/mmc)	Continuous
	Basophils (10^3/mmc)	Continuous
	Platelet count (10^3/microL)	Continuous
	MPV-Mean platelet volume (fL)	Continuous
	$CD3^+$ (PAN T) (%)	Continuous
	$CD3^+CD4^+$ (%)	Continuous
	$CD3^+CD4^+$ (Helper) (mmc)	Continuous
	$CD3^+CD8^+$ (%)	Continuous
	CD4+/CD8+	Continuous
	$CD56^+CD16^+CD3^-$ (NK) (%)	Continuous
	$CD19^+$ (PAN B) (%)	Continuous
	Glucose (mg/dL)	Continuous
	HbA1c glycated hemoglobin (mmol/mol)	Continuous
	Fructosamine (micromol/L)	Continuous
	Urea (mg/dl)	Continuous
	Creatinine (mg/dl)	Continuous
	Uric acid (mg/dL)	Continuous
	Total cholesterol (mg/dL)	Continuous
	Triglycerides (mg/dL)	Continuous
	Cholesterol HDL (mg/dl)	Continuous
	Sodium (mmol/L)	Continuous
	Potassium (mmol/L)	Continuous
	Chloride (mmol/L)	Continuous
	Zinc (micromol/L)	Continuous
	Magnesium (mg/dL)	Continuous
	Total protein (g/dL)	Continuous
	Albumin	Continuous
	Direct bilirubin (mg/dL)	Continuous
	Indirect bilirubin (mg/dL)	Continuous
	AST-aspartate aminotransferase (GOT-glutamic-oxaloacetic transaminase) (U/L)	Continuous
	ALT-alanine Aminotransferase (GPT-glutamic-pyruvic transaminase) (U/L)	Continuous
	Iron (microgr/dL)	Continuous
	Transferrin (mg/dL)	Continuous
	Ferritin (ng/mL)	Continuous
	Folic acid (ng/mL)	Continuous
	Vitamin B12 (pg/ml)	Continuous
	Immunoglobuline G (mg/dl)	Continuous
	Immunoglobuline A (mg/dl)	Continuous
	Immunoglobuline M (mg/dl)	Continuous
	Anti-insulin antibodies	Continuous
	Human tissue transglutaminase IgA antibodies (U/mL)	Continuous
	Thyrotropin (microIU/ml)	Continuous
	Free triiodothyonine (FT3) (pg/mL)	Continuous
	Free thyroxine (FT4) (pg/mL)	Continuous
	Anti-thyroglobulin Antibody (AbTg) (UI/mL)	Continuous
	TPO-Thyroid peroxidase antibodies (UI/mL)	Continuous
	Albumin (%)	Continuous
	Alpha 1 globulin (%)	Continuous
	Alpha 2 globulin (%)	Continuous
	Beta 1 globulin (%)	Continuous
	Beta 1 globulin (%).1	Continuous
	Gamma globulin (%)	Continuous
	A/G albumin/globulin ratio (1)	Continuous
	Deaminated gliadin peptide IgG Antibodies (U/mL)	Continuous
	Homocysteine (μmol/L)	Continuous

(*continued on next page*)

Table 1 (*continued*)

Description	Variable	Type
Genetic data	MTHFR Genotype	Categorical
Data at birth	Mother Age yrs	Continuous
	Father Age yrs	Continuous
	Gestational Age day	Continuous
	Birth Weight kg	Continuous
	Birth Length cm	Continuous
	Birth OFC cm	Continuous
	Apgar Score 1	Discrete
	Sex	Categorical
Development	Started Sitting at month	Continuous
	Started Babbling at month	Continuous
	Started Walking at month	Continuous
	Sphincter Control at month	Continuous
Data at examination	Age (months)	Continuous
	Weight kg	Continuous
	Height cm	Continuous
	OFC cm	Continuous
	BMI	Continuous
	Short Stature	Categorical
	Flat Occiput [Plagiocephaly]	Categorical
	Flat Facial Profile	Categorical
	Small Ears	Categorical
	Hearing Loss	Categorical
	Strabismus	Categorical
	Myopia	Categorical
	Heart Surgery	Categorical
	Separation of the Abdominal Muscle	Categorical
	Umbilical Hernia	Categorical
	Duodenal Atresia	Categorical
	Imperforate Anus	Categorical
	Hirschprung Disease	Categorical
	Obstructive Sleep Apnea	Categorical
	Seizures	Categorical
	Hypothyroidism	Categorical
	TSH μU per mL	Continuous
	Alopecia	Categorical
	Celiac Disease	Categorical
	Small Genitalia	Categorical
	Masticatory Disfunction	Categorical
	Constipation	Categorical
	Jackson's Signs (number)	Discrete
	Diarrhea	Categorical
Indicator of intellectual disability	Age equivalent (AE) score	Analysis output

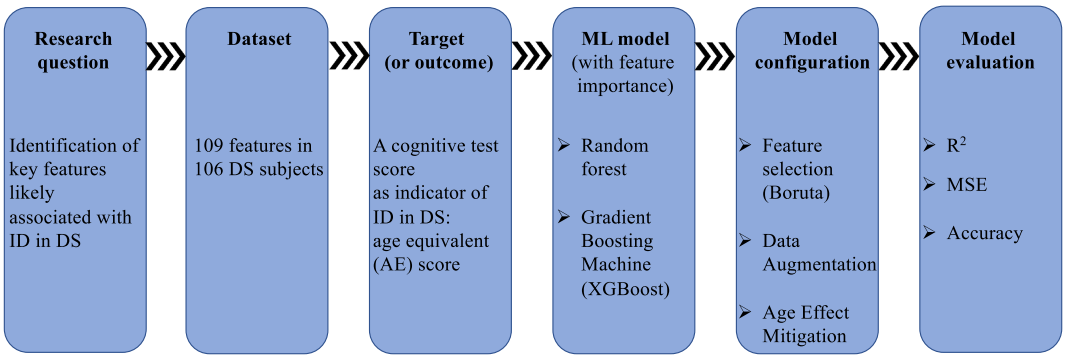

Fig. 1. Overview of machine learning analysis workflow. We have applied two tree-based machine learning (ML) models (Random Forest and XGBoost) to the research question: how to identify key features likely associated with intellectual disability (ID) in Down syndrome (DS). We analyzed 109 features (or variables) in 106 DS subjects. The target (or outcome) of the analysis was the age equivalent (AE) score as indicator of intellectual functioning, impaired in ID. We applied several methods to configure the models: feature selection thought Boruta framework to minimize random correlation; data augmentation to overcome the issue of a small dataset; age effect mitigation to take into account the chronological age of the subjects.

To perform a correct analysis of the data sample we preprocessed the data. This phase was composed of the following steps: (i) starting from Supplementary Table 1, we removed 72 variables, which were deemed unimportant or not informative due to missing values; (ii) we codified categorical attributes introducing a numerical encoding of the values signifying different classes (e.g., YES/NO becomes 0/1); (iii) attributes with signs or string not digestible by an ML algorithm (for instance, numerical attribute containing ">" symbols) were processed and cast to data type conform to the input values required by the model. After preprocessing steps, the dataset has 5 sets of variables on 106 subjects. Table 1 lists all final 109 variables in the dataset, and the associated datatype for each of the attributes. In particular, the acquisition of the development skill milestones was recorded with the month at which the subject started babbling ("Babbling"), sitting without support ("Sitting"), walking without assistance ("Walking") and controlling sphincter and urination ("Sphincter control"), as reported by parents.

A general overview of our ML analysis workflow is shown in Fig. 1.

In our analyses the predictive target (or outcome) is a cognitive test score as indicator of ID in DS. Following a procedure broadly used in the field [18,30,32] in order to assess cognitive functioning in individuals with DS, tests more appropriate for the expected mental age instead of chronological age were used. This allowed us to avoid the floor effect (and the consequent lack of information) that is often present when tests appropriate for chronological age are used. In particular, the cognitive score was calculated starting from raw scores obtained from the Griffiths-III test [33] for DS subjects between 3 years and 6 years and 11 months and the WPPSI-III test [34] for 7-year-old or older DS subjects. Raw scores were transformed into AE scores according to normative data. In order to consider comparable test scales between the two groups, for the scope of this analysis we considered only scale A ("foundations of learning") Griffiths-III test (column "A AE G" in Supplementary Table 1) and "total" WPPSI-III test AE scores (column "Tot AE W" in Supplementary Table 1). However, to take into account the subject's chronological age in the prediction of the outcome, the age effect mitigation was applied as explained below.

2.2. Machine learning models and model building

In the present study, we have applied two decision tree-based ML models (Random Forest and Gradient Boosting Machine) to identify key features likely associated with ID in DS.

Random Forest (RF) is based on training of multiple trees on random subsets of variables. The final prediction is the result of a voting system, where each tree returns a predictive outcome for an input instance and the majority output will be the overall returning value [35].

Gradient Boosting Machine (GBM) is another tree-based method. If RF builds decision trees in parallel, approximating different views (or subsamples) of the dataset, GBM builds models sequentially, gradually minimizing the loss function and consequently the overall training error. The final output is computed incrementally, i.e., the prediction of each decision tree is summed to obtain the overall outcome. Among the most popular GBM models available we choose XGBoost, which in recent years has been widely used to solve a variety of problems with a high level of accuracy [36].

2.3. Feature importance

Feature importance was used to quantify how important each input attribute is for reaching a prediction. In our case it was used in both RF and XGBoost methods to identify which attribute is more correlated with the ID score, so as to guide future investigations. Whenever we create a split while we train a tree model (both RF and GBM), this will be associated to an increase of the performance indicator. By keeping track of the cumulative increase, it is possible to estimate their impact on the tree behavior indicating feature importance scores.

Feature importance scores are the key instrument for our analysis. For this reason, it is important to understand that, since most ML models (including the ones we use) operate on the basis of correlations in the data, a high feature importance is not an indicator of cause-effect relation. In the experimental results, we will display the feature importance as a bar plot, showing the ranking of the first ten variables with higher relevance based on their score.

It is also important to observe that the reliability of feature importance score is limited by the performance of the ML model itself. In other words, the importance score extracted from an inaccurate model may be misleading.

2.4. Data manipulation

In this section, we will present three methods for tackling different issues relative to the dataset used in our analysis (Fig. 1). Each of these methods refers to a specific problem that arose during the analysis.

2.4.1. Feature selection

Feature selection allows the model to focus on the most important features or on features whose importance is of statistical significance. By doing so, we reduce the risk of overfitting, and we obtain more reliable importance scores as a side effect.

We applied Boruta, a wrapper-based feature selection method to filter out a set of features that are relevant to the target variable [37]). It is implemented through the following steps: 1) The features are randomly shuffled and then stitch together with the actual feature matrix to form a new feature matrix. 2) The importance of the shuffled and actual features is obtained by inputting the new feature matrix into the random forest. 3) The actual features with importance greater than the maximum importance of the shuffled features are retained. By iterating the above steps several times, the important features are identified by Boruta.

2.4.2. Data augmentation

To increase the ML models accuracy starting from a small sample size, we used data augmentation techniques. The number of data augments quadratically, going from n sample to $\sum_i n\text{-}i$ by comparing pairs of input instances, where each new data entry is computed as the ratio between the variables pertaining to two subjects. With this step we are changing the question our model is designed to answer: rather than asking "given information about subjects what is the estimated AE score?", we now ask "given information about ratios of input attributes for a pair of subjects, what is the ratio of their AE score?". In scientific terms, the new question is weaker than the previous one, in fact given an oracle for the former it is possible to answer the latter while the opposite is not possible.

That said, the new question is still highly informative from the point of view of our main goal: if the ratio between (say) vitamin B12 for two subjects is predictive of the ratio between their AE scores, there is intuitively a chance the feature is involved in ID.

The question can be further simplified by focusing on the detection of the higher AE score in the pair of subjects. In this case, we ask "given information about ratios of input attributes for a pair of subjects, which subject has the higher AE score?". This method allows the formulation of the problem as a classification task, where the target is formulated as a category. More precisely, if we consider two subjects q_i and q_j, and the category associated with their comparison, c_k:

- $c_k = 1$ if $q_i > q_j$.
- $c_k = 2$ if $q_i \leq q_j$.

2.4.3. Age effect mitigation

In our analysis the outcome is a cognitive test score as indicator of ID in DS. In particular, we used the AE score of children with DS. It is important to note that the chronological age of the subject is highly related to the target of the model and might distort the real importance of other features. Moreover, this argument can be made for each variable that is correlated to the chronological age of the children. In this situation, features strongly correlated with chronological age would naturally have a higher importance score in our model.

To address this issue we have applied an age effect mitigation: in this way we avoid to overfit about the chronological age using age information to compute normalization factors (estimated using ML models, in particular artificial neural networks). The chronological age of the subject was used as a predictor of any other variable in our sample. In this case, we might face two scenarios: (i) the age of the subject is a good predictor of the variable. Therefore, all its values will be normalized to approximately one, meaning it will not be as informative during the training process; (ii) the age of the subject is not a good predictor of the variable, then the values of the variables will be re-scaled according to the predicted value.

2.5. Model evaluation

In order to have a reliable interpretation of the results, we needed to evaluate the ML models used to approximate the samples, since scarce performances would not allow for a trustworthy analysis. To this end, we relied on commonly used performance metrics, such as: R^2, to measure the goodness of the fit, and Mean Squared Error along with the accuracy score, to measure the error associated to the models.

To avoid incurring in overfitting problems, we relied of a K-fold cross validation approach ($K = 10$); thus all the results provided regarding the metrics are specified as mean and standard deviation among the different validation splits.

3. Results

The clinical dataset includes a total of 106 DS subjects (41 females; 65 males) for which cognitive data are available. The mean age of cognitive assessment is 8.88 years with a standard deviation (SD) of 3.96 years. Anonymized personal, genetic, diagnostic, clinical, and auxological information are available in the Supplementary Table 1.

In this section, we are presenting the results of the analysis on the complete dataset, except for the features removed during the preprocessing phase. All the main findings will be presented outlining advantages and disadvantages of each approach motivating the use of the methodologies presented in Materials and Methods section. The investigation was performed using both RF and GBM (i.e. XGBoost), whereas the findings are presented using feature importance. Each model will be evaluated providing the metrics introduced in the paragraph 2.5 of Materials and Methods section which are summarized in Table 2.

The computer code has been implemented with Python programming language.

Table 2
Evaluation of the models. R^2: coefficient of determination; MSE: mean squared error; RF: random forest; FS: feature selection; DA: data augmentation; AEM: age effect mitigation; XGB: XGBoost.

Model	Regression		Classification	
	R^2	MSE	R^2	Accuracy
RF + FS	$0.54 \pm 8.19 \cdot 10^{-5}$	154.31 ± 72.93		
RF + FS + DA	0.69 ± 0.00011	0.13 ± 0.047	$0.86 \pm 4.3 \cdot 10^{-5}$	0.66 ± 0.070
RF + FS + DA + AEM	0.70 ± 0.00019	0.12 ± 0.043	$0.86 \pm 0.17 \cdot 10^{-5}$	0.67 ± 0.076
XGB + FS + DA + AEM	$0.93 \pm 5.83 \cdot 10^{-5}$	0.11 ± 0.043	$0.96 \pm 8.61 \cdot 10^{-5}$	0.67 ± 0.094

The results shown below are a partial view of the analysis, focused on motivating the methodological process followed during this study, a complete report of the analysis is available in the Supplementary Figures.

3.1. Random forest combined with feature selection

Due to the scarcity of data samples and the high number of variables, we deployed a feature selection method aimed at reducing the number of random correlations in the dataset and removing the unimportant variables with reference to the predictive target (as described in paragraph 2.4.1 in Materials and Methods section and shown in Fig. 1).

The application of the feature selection algorithm reduced the set of variables to a few candidates, where the predictor with higher importance is the chronological age of the subject (Supplementary Fig. 1).

However, as reported in Table 2, the metrics reveal a bad performance of the model.

3.2. Random forest combined with feature selection and data augmentation

To counteract the absence of data points, we resorted to a form of data augmentation, as described in paragraph 2.4.2 of Materials and Methods section and shown in Fig. 1. We are basically changing the feature space to obtain more samples and produce more accurate models.

This approach allows us to formulate the original learning problem in two forms: a regression task and a classification task. In the regression task, the predictive target is a continuous variable representing the AE score of the child. As we can see in Supplementary Fig. 2 the predictor with higher importance is APGAR score at 1 min after birth, a measure of the physical condition of a newborn infant. The classification model has a discrete outcome, namely, a discrete numerical value obtained by the AE comparison in a pair of subjects, as described in paragraph 2.4.2 of Materials and Methods section. As shown in Supplementary Fig. 3 the categorical variables (e.g. hearing loss and duodenal atresia) gain higher importance.

Table 2 shows better performance metrics for the model (RF combined with feature selection and data augmentation) compared to the previous one (RF combined with feature selection). Even though we now have more reliable results, the quality of the information we retrieved through the analysis is debatable since the chronological age of the subject is highly related to the target of the model (AE) and might distort the real importance of other features (Fig. 2). Moreover, this argument can be made for each variable that is known to be correlated to the chronological age of the children (e.g., creatinine level, homocysteine level, development milestones, height, occipital frontal circumference). To avoid this issue, we can mitigate the effect of age on the predictive model by applying the age effect mitigation method as described in paragraph 2.4.3 of Material and Methods section and shown in Fig. 1.

3.3. Random forest combined with feature selection, data augmentation and age effect mitigation

In order to obtain an analysis not influenced by the chronological age of the subject, we resorted to the normalization method presented in paragraph 2.4.3 of Material and Methods section. The regression task (Fig. 3A) confirms the importance of the APGAR score at 1 min after birth and of almost all the other features compared to the previous model without age effect mitigation (Supplementary Fig. 2). The classification task (Fig. 3B) highlights another categorical variable related to gastrointestinal alteration (Hirschsprung disease). Both classification and regression tasks show the importance of hearing loss, language (the month at which the subject started babbling), magnesium, immune system ($CD19^{+}$ and immunoglobulins M) and vitamin B12. The evaluation score shows no significant variations demonstrating the robustness of the model (Table 2).

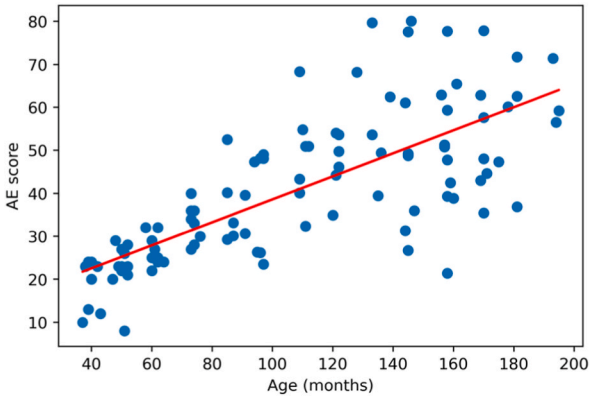

Fig. 2. Correlation with chronological age. Chronological age of the subjects is strongly correlated to the outcome of our analysis (age equivalent or AE score). Correlation coefficient r = 0.747. The data show that chronological age could be considered as a trivial predictor of the outcome and can influence the results related to variables strongly dependent on chronological age.

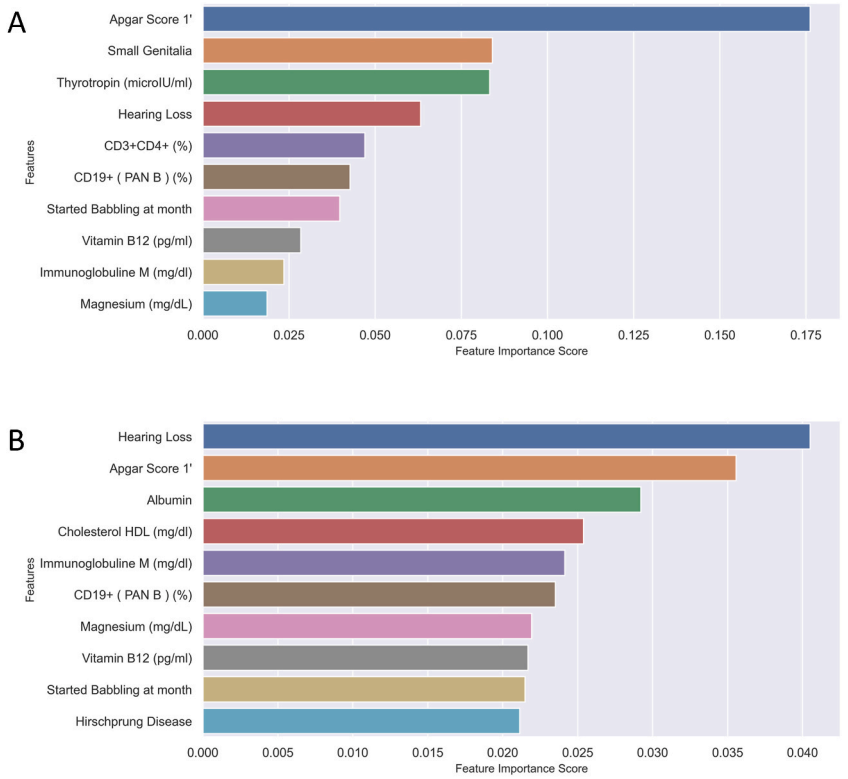

Fig. 3. Random Forest combined with Feature Selection, Data augmentation and Age Effect Mitigation. Feature importance obtained by random forest combined with feature selection, data augmentation and age effect mitigation. **3A.** Regression task. **3B.** Classification task.

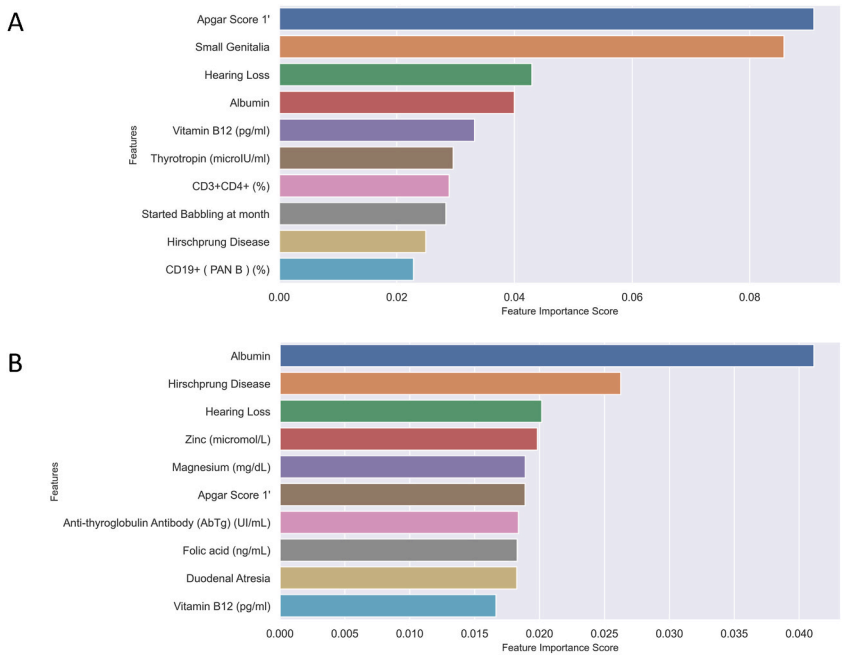

Fig. 4. Gradient boosting machine combined with Feature Selection, Data augmentation and Age Effect Mitigation. Feature importance obtained by gradient boosting machine combined with feature selection, data augmentation and age effect mitigation. **4A.** Regression task. **4B.** Classification task.

3.4. Gradient boosting machine combined with feature selection, data augmentation and age effect mitigation

XGBoost combined with Feature Selection and regression and classification tasks of XGBoost combined with Feature Selection and Data Augmentation model results are shown in Supplementary Figs 4, 5 and 6, respectively.

The XGBoost combined with all data manipulation steps confirms most of the results obtained by RF method. As shown in Fig. 4 A (regression task) the importance of language (the month at which the subject started babbling and hearing loss), small genitalia, thyrotropin, immune system ($CD3^+CD4^+$ and $CD19^+$), and APGAR score at 1 min after birth are confirmed with a higher R^2 ($R^2 = 0.93$, Table 2). Fig. 4B (classification task) shows again categorical variables related to gastrointestinal alterations (duodenal atresia, Hirschsprung disease).

4. Discussion

The purpose of this study is to use a dataset comprising personal, genetic, diagnostic, clinical, and auxological data of children with DS to build a model useful to better understand their association with ID. In particular, models built using ML algorithms may assist researchers in identifying key features likely associated with ID in DS, and ultimately, may improve research efforts focused on the identification of possible therapeutic targets.

ML approaches can be useful to deal with two issues related to complex conditions such as DS: first, the collection of complex and mixed data types; second, ML methods have the ability to identify non-linear relationships between a variety of data types and can be used to generate a model in a hypothesis-free manner in order to identify new pathways involved in cognitive impairment in DS.

In the present study, we decided to apply RF and GBM approaches to our dataset: these tree-based methods can indicate which features are most strongly associated with the output of interest (or target). In particular, we decided to use a cognitive test score as target for our analysis, with the purpose of identifying features likely involved in cognitive delay. The application of Boruta framework in feature selection helps in removing random correlations from the analysis. Then, feature importance application can summarize the impact of individual features (or variables) on response variable (or target) and they are referred to as "variables of importance". The response variable (or target) is a cognitive test score (AE score) as indicator of cognitive performance in children with DS, thus the identification of variables of importance leads to the building of models useful for the identification of features likely associated with the cognitive development in DS. In the present study, we extract the variables of importance in an effort to demonstrate the feasibility of a ML based model to identify possible new altered pathways associated to ID in DS.

It is important to note that we applied an "age effect mitigation" in order to mitigate the effect of the chronological age of the children on the target: the variable "age" and all the other variables highly related to the age might distort the real importance of other features in predicting the outcome. Moreover, to address the issue of low number of data, we applied a form of data enrichment (or augmentation), considering the comparison between couples of subjects, where each new data entry is computed as the ratio between the variables relative to the comparison between two children. This allows us to augment the number of data quadratically; moreover, this method allows the formulation of the problem as a classification problem, where the target (i.e., AE score) is formulated as a category.

We described all the main features important for the application of a supervised ML method; in particular, (i) the label for prediction (cognitive AE score as an indicator of ID in DS) and (ii) candidate features listed in the dataset and described in Material and Methods section and in Table 1.

(i) As previously described [18,30], the cognitive score was calculated starting from raw scores obtained from the A scale of the Griffiths-III test [33] and the total score of the WPPSI-III test [34]: raw scores were transformed into AE scores. In this way we have applied tests that were more appropriate for expected mental age rather than chronological age to assess cognitive functioning in individuals with DS [32]. Moreover, we applied the age effect mitigation in order to take into account the subject's chronological age in the prediction of the outcome.

(ii) As regards the candidate features (Table 1), it is important to note that they are diverse, potentially unharmonized data and the application of ML methods allows to handle numerical and non-numerical data and to analyze a huge amount of mixed data. In particular, we considered both categorical (clinical data) and continuous variables (lab measurements).

The results show that ML algorithms can be applied with good accuracy to identify variables likely involved in cognitive delay in DS. In this context it is very important to take into account the effect of chronological age on the outcome. Considering all the steps described above (tree-based models and Boruta framework, feature importance tasks, age effect mitigation, data enrichment) we have shown two well performing models (RF and XGBoost both combined with feature selection, data augmentation and age effect mitigation) with low error (MSE <0.12) and an acceptable R^2 (0.70 and 0.93) (Table 2). XGBoost is providing the best performance for both the regression ($R^2 = 0.93$, MSE $= 0.11$) and the classification task ($R^2 = 0.96$, Accuracy $= 0.67$).

Interestingly, the variables of importance show several features that can be considered with particular attention during the follow up of DS patients. In particular, in the regression tasks, the results show two groups of variables of importance (Figs. 3A and 4A). First, different characteristics might be grouped in perinatal and neonatal status and development. The APGAR score at 1 min after birth is a measure of the physical condition of a newborn infant that can influence the neonatal development [38]. Hearing loss has been demonstrated to affect language and cognitive skills development [39]. This might slow down the achievement of developmental milestones, such as babbling (Fig. 3A, B and 4A), that could be predictors of later motor, cognitive skills and language [31]. The ML models are consistent with the idea that early developmental milestones can be important variables to consider in planning early

phenotype-informed interventions that have the potential to influence positive trajectories in community living and participation across an individual's lifespan. Finally, thyrotropin concentration can be related to thyroid disorders that can affect the nervous system and play a role in cognitive delay [40].

The second group of variables of importance is related to the immune system (e.g. $CD19^{+}$ and immunoglobulins M) underlining the documented immune dysfunction typical of DS [41].

In the classification tasks (Fig. 4A–B), among the top ten variables of importance we find variables related to gastrointestinal alterations (e.g. Hirschsprung disease and duodenal atresia). Variables related to gastrointestinal disorders can be related to microbiome alterations. Since there are increasing evidence that the composition of the resident bacteria within the gastrointestinal tract can influence cognitive functions [42], treatments for gastrointestinal disorders and modulation of the microbiota should be deeper investigated. Moreover, the presence of congenital gastrointestinal alterations can lead to a surgical operation in the first stages of life, and this may influence child's development in his newborn life. Lastly, results obtained through both RF and XGB models show the importance of vitamin B12: a concentration threshold of vitamin B12 has been highlighted to be important in the DS population which might have greater vitamin requirements [30].

It is known that cognitive delay may be exacerbated by the presence of chronic health conditions [43]. Individuals with DS and co-occurring chronic disorders may benefit from early interventions to mitigate their risk for adverse cognitive outcomes: for this reason, it is very important to identify important variables that may affect cognitive development and to use these data to improve care pathways, in particular perinatal care.

In conclusion, in the context of the routine follow up provided for DS, the collection of a large quantity of data is possible. Complex conditions as DS can be better understood if we integrate different information (genetic, diagnostic, clinical, and auxological data). The ML pipeline developed here can be applied to identify important variables likely involved in cognitive development in DS. Further studies will be necessary to better understand how these variables affect cognitive delay in DS and their predictive value. This approach may improve research efforts focused on the identification of possible therapeutic targets and new care pathways. Moreover, from a future perspective, this study can be the basis for further testing/validating of our algorithms with multiple and larger datasets with the opportunity to increase the power of the analysis. With this purpose, we deposited the associated computer code in the public domain on the GitHub platform (https://github.com/lompabo/down-syndrome-feature-screening.git) in the hopes of contributing to an open-source community.

Funding

This work has been supported by Fondazione Cassa di Risparmio in Bologna. The fellowship for G.R. has been funded by donations from the Fondazione Umano Progresso. The fellowship for F.P. has been funded by donations from the Fondazione Rosa Pristina. The fellowship for F.A. has been funded by donations from the Fondazione Umano Progresso and Matteo and Elisa Mele and by European Union - NextGenerationEU - Alma Idea 2022 – Alma Mater Studiorum Università di Bologna.

Author contribution statement

Federico Baldo, Allison Piovesan, Marijana Rakvin: Performed the experiments. Giuseppe Ramacieri, Chiara Locatelli, Silvia Lanfranchi, Sara Onnivello, Francesca Pulina: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data. Maria Caracausi, Francesca Antonaros: Analyzed and interpreted the data. Michele Lombardi, Maria Chiara Pelleri: Conceived and designed the experiments; Wrote the paper.

Data availability statement

Data are available at https://github.com/lompabo/down-syndrome-feature-screening.git.

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.

Acknowledgements

The authors heartily thank all the children and their families who participated in the study. We are very grateful to Donatella Pascai and Giuseppina Dibenedetto for their expert organization and performance of blood sampling.

We are especially grateful to the "Fondazione Umano Progresso", Milano, and to "Fondazione Rosa Prístina", Pisa, for their valuable support to our research on trisomy 21.

We wish to sincerely thank for their support to our research Matteo and Elisa Mele; the dad and mom of Gabriele; Veronica Sandroni "for those who can make progress for the good of all"; Samuel and Simona; Associazione "Il Sorriso - das Lächeln APS" - Associazione Genitori e Amici di Persone con Sindrome di Down, Bolzano; Associazione "Sette Per Te Ventuno - Non abbiate paura - Onlus", Gorgonzola (MI); Associazione "Amicorum" and Associazione "+ di 21" (piudi21) ODV, Cassano Magnago (VA); "Comitato Arzdore di Dozza" (BO) and the Costa family; Associazione "Vola con Martin oltre il 21" ODV, Mandello del Lario (LC); Sichim Alfa Srl;

Guido Marangoni; Studio Ballotta, Sghirlanzoni & Associati; Maurizio Funazzi; Eleonora Riboldi and all the readers of her book; Giulia Nicoletti, and her relatives and friends, in occasion of her birthday; Simona Galletta and Antonio Lazzari.

Very special thanks to "Progetto Pulcino Onlus" - Reggio Emilia, to its Founder and Vice-President Dr. Cristiana Magnani, and to Dr. Massimiliano Iori for the valuable donation of the −80 °C freezer hosting our biobank and of a refrigerated centrifuge.

Our heartfelt thanks for their initiatives in support of our study, as well all the participants in these events, go to: "Parrocchie di Pieve Corleto e Basiago", and "Gruppo Dcuore", in Faenza; Luciano Perondi and Paolo Ciaroni for having created a series of car meetings called "Top Down", devoted to spider and cabriolet cars; Barchetta Social Club.

We are very grateful to all the other people that have very kindly contributed through individual donations to support this research, in particular: Biologi Officina Trasfusionale Romagna; Matteo Brivio; Giovanni Bubani; Massimiliano Buonamici; Ornella Carciani; Carla and Guido, grandparents of Pietro, Milan; Marcello Colombo and the group of whisky tasting; Enrico Donà; in memory of Sergio Ferro; Frallicciardi Family; Dott. Giulia Gramellini; Leonesio and Tanno Families; Cecilia Martinelli; the Martinz family: Elena, Enrico, Alfredo and Concetta; Patrizia Masello and Associazione MPV-CAV, Campodarsego (PD); Massimo Montanari; Marcella Monti; Irene Panzavolta; Orlando and Ramirez Pennica; Francesco Pugliarello; Paola Rancati; Elena Rossi and Davide Finelli; Gabriele Rossi; Beniamino and Diletta Sincich; Cristina Todaro; Anselmo Zaniboni.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e19444.

References

[1] J. Lejeune, M. Gauthier, R. Turpin, [Human chromosomes in tissue cultures], Comptes rendus hebdomadaires des seances de l'Academie des sciences 248 (4) (1959) 602–603.
[2] A.F. Lukowski, H.M. Milojevich, L. Eales, Cognitive functioning in children with down syndrome: current knowledge and future directions, Adv. Child Dev. Behav. 56 (2019) 257–289.
[3] M.C. Pelleri, E. Cicchini, M.B. Petersen, L. Tranebjaerg, T. Mattina, P. Magini, et al., Partial trisomy 21 map: ten cases further supporting the highly restricted Down syndrome critical region (HR-DSCR) on human chromosome 21, Mol. Genet. Genomic Med. 7 (2019) e797.
[4] G. Anneren, K.H. Gustavson, V.R. Sara, T. Tuvemo, Growth retardation in Down syndrome in relation to insulin-like growth factors and growth hormone, Am. J. Med. Genet. Suppl. 7 (1990) 59–62.
[5] K. Gardiner, Y. Herault, I.T. Lott, S.E. Antonarakis, R.H. Reeves, M. Dierssen, Down syndrome: from understanding the neurobiology to therapy, J. Neurosci. 30 (45) (2010) 14943–14945.
[6] N.J. Roizen, D. Patterson, Down's syndrome, Lancet 361 (9365) (2003) 1281–1289.
[7] P. Strippoli, M.C. Pelleri, M. Caracausi, L. Vitale, A. Piovesan, C. Locatelli, et al., An integrated route to identifying new pathogenesis-based therapeutic approaches for trisomy 21 (Down Syndrome) following the thought of Jérôme Lejeune, Science Postprint 1 (1) (2013), e00010.
[8] M.C. Pelleri, E. Gennari, C. Locatelli, A. Piovesan, M. Caracausi, F. Antonaros, et al., Genotype-phenotype correlation for congenital heart disease in Down syndrome through analysis of partial trisomy 21 cases, Genomics 109 (5–6) (2017) 391–400.
[9] P. Strippoli, M.C. Pelleri, A. Piovesan, M. Caracausi, F. Antonaros, L. Vitale, Genetics and genomics of Down syndrome, State of the Art of Research on Down Syndrome 56 (2019) 1–39. Academic Press.
[10] S. Onnivello, F. Pulina, C. Locatelli, C. Marcolin, G. Ramacieri, F. Antonaros, et al., Cognitive profiles in children and adolescents with Down syndrome, Sci. Rep. 12 (1) (2022) 1936.
[11] C. Locatelli, S. Onnivello, C. Gori, G. Ramacieri, F. Pulina, C. Marcolin, et al., A reassessment of Jackson's checklist and identification of two Down syndrome sub-phenotypes, Sci. Rep. 12 (1) (2022) 3104.
[12] J. Lejeune, On the mechanism of mental deficiency in chromosomal diseases, Hereditas (Lund) 86 (1) (1977) 9–14.
[13] M. Pogribna, S. Melnyk, I. Pogribny, A. Chango, P. Yi, S.J. James, Homocysteine metabolism in children with Down syndrome: in vitro modulation, Am. J. Hum. Genet. 69 (1) (2001) 88–95.
[14] R. Obeid, K. Hartmuth, W. Herrmann, L. Gortner, T.R. Rohrer, J. Geisel, et al., Blood biomarkers of methylation in Down syndrome and metabolic simulations using a mathematical model, Mol. Nutr. Food Res. 56 (10) (2012) 1582–1589.
[15] M. Caracausi, V. Ghini, C. Locatelli, M. Mericio, A. Piovesan, F. Antonaros, et al., Plasma and urinary metabolomic profiles of Down syndrome correlate with alteration of mitochondrial metabolism, Sci. Rep. 8 (1) (2018) 2977.
[16] M. Dierssen, M. Fructuoso, M. Martínez de Lagrán, M. Perluigi, E. Barone, Down syndrome is a metabolic disease: altered insulin signaling mediates peripheral and brain dysfunctions, Front. Neurosci. 14 (670) (2020).
[17] J.L. Gueant, G. Anello, P. Bosco, R.M. Gueant-Rodriguez, A. Romano, C. Barone, et al., Homocysteine and related genetic polymorphisms in Down's syndrome IQ, J. Neurol. Neurosurg. Psychiatry 76 (5) (2005) 706–709.
[18] F. Antonaros, V. Ghini, F. Pulina, G. Ramacieri, E. Cicchini, E. Mannini, et al., Plasma metabolome and cognitive skills in Down syndrome, Sci. Rep. 10 (1) (2020), 10491.
[19] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (1) (2019) 64.
[20] Q. Zhao, K. Rosenbaum, K. Okada, D.J. Zand, R. Sze, M. Summar, et al., Automated Down syndrome detection using facial photographs, Annu Int Conf IEEE Eng Med Biol Soc 2013 (2013) 3670–3673.
[21] M.E. Özdemir, Z. Telatar, O. Eroğul, Y. Tunca, Classifying dysmorphic syndromes by using artificial neural network based hierarchical decision tree, Australas. Phys. Eng. Sci. Med. 41 (2) (2018) 451–461.
[22] W. Srisraluang, K. Rojnueangnit, Facial recognition accuracy in photographs of Thai neonates with Down syndrome among physicians and the Face2Gene application, Am. J. Med. Genet. A. 185 (12) (2021) 3701–3705.
[23] A. Koivu, T. Korpimäki, P. Kivelä, T. Pahikkala, M. Sairanen, Evaluation of machine learning algorithms for improved risk assessment for Down's syndrome, Comput. Biol. Med. 98 (2018) 1–7.
[24] A. Mahmoud, A.F. Belal M, Y. Helmy, Towards an intelligent tutoring system to down syndrome, Int. J. Comput. Sci. Inf. Technol. 6 (2014) 129–137.
[25] M.F. Jojoa-Acosta, S. Signo-Miguel, M.B. Garcia-Zapirain, M. Gimeno-Santos, A. Méndez-Zorrilla, C.J. Vaidya, et al., Executive functioning in adults with down syndrome: machine-learning-based prediction of inhibitory capacity, Int. J. Environ. Res. Publ. Health 18 (20) (2021).
[26] C. Gupta, P. Chandrashekar, T. Jin, C. He, S. Khullar, Q. Chang, et al., Bringing machine learning to research on intellectual and developmental disabilities: taking inspiration from neurological diseases, J. Neurodev. Disord. 14 (1) (2022) 28.

[27] C.D. Nguyen, A.C. Costa, K.J. Cios, K.J. Gardiner, Machine learning methods predict locomotor response to MK-801 in mouse models of down syndrome, J. Neurogenet. 25 (1–2) (2011) 40–51.
[28] C. Higuera, K.J. Gardiner, K.J. Cios, Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome, PLoS One 10 (6) (2015), e0129126.
[29] H. Kulan, T. Dag, In silico identification of critical proteins associated with learning process and immune system for Down syndrome, PLoS One 14 (1) (2019), e0210954.
[30] F. Antonaros, S. Lanfranchi, C. Locatelli, A. Martelli, G. Olivucci, E. Cicchini, et al., One-carbon pathway and cognitive skills in children with Down syndrome, Sci. Rep. 11 (1) (2021) 4225.
[31] C. Locatelli, S. Onnivello, F. Antonaros, A. Feliciello, S. Filoni, S. Rossi, et al., Is the age of developmental milestones a predictor for future development in down syndrome? Brain Sci. 11 (5) (2021) 655.
[32] F. Pulina, R. Vianello, S. Lanfranchi, Cognitive profiles in individuals with Down syndrome, in: S. Lanfranchi (Ed.), State of the Art of Research on Down Syndrome, vol. 56, Academic Press, 2019, pp. 67–92.
[33] E. Green, L. Stroud, S. Bloomfield, J. Cronje, C. Foxcroft, K. Hurter, et al., Griffiths Scales of Child Development, third ed., 2016 (Griffiths III).
[34] D. Wechsler, Wechsler Preschool and Primary Scale of Intelligence, third ed., 2002. WPPSI-III.
[35] M. Belgiu, L. Drăguţ, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogrammetry Remote Sens. 114 (2016) 24–31.
[36] A. Ogunleye, Q.G. Wang, XGBoost model for chronic kidney disease diagnosis, IEEE ACM Trans. Comput. Biol. Bioinf 17 (6) (2020) 2131–2140.
[37] M.B. Kursa, W.R. Rudnicki, Feature Selection with the Boruta Package, Journal of Statistical Software 36 (11) (2010) 1–13.
[38] L. Del Hoyo Soriano, T.C. Rosser, D.R. Hamilton, D.J. Harvey, L. Abbeduto, S.L. Sherman, Relationship between Apgar scores and long-term cognitive outcomes in individuals with Down syndrome, Sci. Rep. 11 (1) (2021), 12707.
[39] G. Laws, A. Hall, Early hearing loss and language abilities in children with Down syndrome, Int. J. Lang. Commun. Disord 49 (3) (2014) 333–342.
[40] H. Khaleghzadeh-Ahangar, A. Talebi, P. Mohseni-Moghaddam, Thyroid disorders and development of cognitive impairment: a review study, Neuroendocrinology (2021).
[41] R.H.J. Verstegen, M.A.A. Kusters, E.F.A. Gemen, E. De Vries, Down syndrome B-lymphocyte subpopulations, intrinsic defect or decreased T-lymphocyte help, Pediatr. Res. 67 (5) (2010) 563–569.
[42] M.G. Gareau, Cognitive function and the microbiome, Int. Rev. Neurobiol. 131 (2016) 227–246.
[43] K.C. Gandy, H.A. Castillo, L. Ouellette, J. Castillo, P.J. Lupo, L.M. Jacola, et al., The relationship between chronic health conditions and cognitive deficits in children, adolescents, and young adults with down syndrome: a systematic review, PLoS One 15 (9) (2020), e0239040.