PharmaNutrition 23 (2023) 100326

Contents lists available at ScienceDirect

PharmaNutrition

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

Predictive factors for allergy at 4–6 years of age based on machine learning: A pilot study

Kim Kamphorst [a,b,*], Alejandro Lopez-Rincon [c,f], Arine M. Vlieger [b], Johan Garssen [c,e], Esther van 't Riet [d], Ruurd M. van Elburg [a]

[a] Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Dept. Pediatrics, Amsterdam Gastroenterology, Metabolism & Nutrition, Amsterdam Reproduction & Development Amsterdam, the Netherlands
[b] St. Antonius Hospital, Dept. Pediatrics, Nieuwegein, the Netherlands
[c] Utrecht Institute for Pharmaceutical Sciences, Division Pharmacology, Utrecht University, Utrecht, the Netherlands
[d] Research Office, Department of Strategy and Policy, University Medical Center Utrecht, Utrecht, the Netherlands
[e] Danone Nutricia Research, Utrecht, the Netherlands
[f] Julius Center for Health Sciences and Primary Care, Department of Data Science, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Prediction
Feature selection
AI
Artificial intelligence
Atopic disorders
Cytokines

ABSTRACT

Background: In Europe, allergic diseases are the most common chronic childhood illnesses and the result of a complex interplay between genetics and environmental factors. A new approach for analyzing this complex data is to employ machine learning (ML) algorithms. Therefore, the aim of this pilot study was to find predictors for the presence of parental-reported allergy at 4–6 years of age by using feature selection in ML.

Methods: A recursive ensemble feature selection (REFS) was used, with a 20% step reduction and with eight different classifiers in the ensemble, and resampling given the class unbalance. Thereafter, the Receiver Operating Characteristic Curves for five different classifiers, not included in the original ensemble feature selection technique, were calculated.

Results: In total, 130 children (14 with and 116 without parental-reported allergy) and 248 features were included in the ML analyses. The REFS algorithm showed a result of 20 features and particularly, the Multi-layer Perceptron Classifier had an area under the curve (AUC) of 0.86 (SD 0.08). The features predictive for allergy were: tobacco exposure during pregnancy, atopic parents, gestational age, days of: diarrhea, cough, rash, and fever during first year of life, ever being exposed to antibiotics, Resistin, IL-27, MMP9, CXCL8, CCL13, Vimentin, IL-4, CCL22, GAL1, IL-6, LIGHT, and GMCSF.

Conclusions: This ML model shows that a combination of environmental exposures and cytokines can predict later allergy with an AUC of 0.86 despite the small sample size. In the future, our ML model still needs to be externally validated.

1. Introduction

In most European countries, allergic diseases are the most common chronic childhood illnesses [1]. Allergic diseases are the result of a complex interplay between genetics and environmental factors [2]. So far, allergic diseases have been associated with a variety of early life events such as: infections, nutrition, and exposure to medication [3]. However, most univariate/multivariate analysis techniques currently in use are not sufficient to study associations in a large dataset because of the complex interplay of all these factors [4]. A new approach for analyzing this complex data is to employ machine learning (ML) algorithms.

ML finds patterns in the data and applies these patterns to new data to make predictions [5]. Within the INCA study (INtestinal microbiota Composition after Antibiotic treatment in early life), a Dutch prospective birth cohort study, a large dataset has been built that incorporates various early life exposures/characteristics. Previously, we showed that antibiotic exposure in the first week of life was associated with allergy at 4–6 years of age [6]. The aim of this pilot study was to find predictors for the presence of parental-reported allergy at 4–6 years of age by using feature selection in ML.

* Correspondence to: Emma Children's Hospital, Amsterdam UMC, Room H8.235, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands.
E-mail address: k.kamphorst@amsterdamumc.nl (K. Kamphorst).

https://doi.org/10.1016/j.phanu.2022.100326
Received 22 November 2022; Received in revised form 6 December 2022; Accepted 10 December 2022
Available online 14 December 2022

Table 1
Data collection of variables used in the machine learning model.

Source	Time of collection	Variables
Hospital records	First week of life	Sex, birth mode, gestational age in days, birth weight, exposure to antibiotics, antibiotic duration, type of antibiotics, and hospital of birth
Parental questionnaire	At inclusion	Tobacco exposure during pregnancy, presence of siblings, birth order, presence of pets, atopy in the family, household education, parental country of birth, and living environment.
A daily checklist	During the first year of life	Start and duration of: wheezing, eczema, infantile colic, rash, fever, cough, runny nose, otalgia/ otorrhea, diarrhea, daycare, and exposure to tobacco smoke
A monthly questionnaire	During the first year of life	Weight, length, type of feeding, duration of (exclusively) breastfeeding, presence of pets, number and type of additional antibiotic courses, general practioner visits, and probiotic use
doctors' diagnoses (general practitioner) based on the International Classification of Primary Care (ICPC)	At one year of age	A14 (infantile colic), D10 (vomiting), D70 (infectious diarrhea), D73 (susceptible gastrointestinal infection), R02 (dyspnea), R03 (wheezing), R05 (cough), R96 (asthma like symptoms), R74_R74.1_R74.01 (acute respiratory tract infection (RTI)), R78 (acute bronchiolitis), R81 (pneumonia), R21 (symptoms of the throat), R25 (abnormal sputum), R27 (concern about respiratory illness), R77 (acute laryngitis), R83 (other RTI), R88 (influenza), R99 (other RT diseases), S21 (other symptoms/ complaints of the skin), S21.01 (dry skin/flaking), S87 (eczema).
Blood sample in a subgroup	At one year of age	IL1RA, IL1b, IL4, IL5, IL6, IL10, IL12, IL13, IL17, IL17F, IL18, IL21, IL22, IL25, IL26, IL27, IL31, IL33, IL37, TNFa, IFNa, IFNg, TSLP, LIGHT, APRIL, MIF, TGFb, YKL40, CCL2, CCL7, CCL8, CCL11, CCL13, CCL17, CCL18, CCL22, CCL25, CCL26, CCL27, CCL28, CXCL4, CXCL8, CXCL9, CXCL10, CXCL11, CXCL13, GMCSF, VEGF, BDNF, sICAM1, sVCAM1, sCD14, sCD19, sCD25, sCD27, sCD40L, IL1R1, IL1R2, TNFR1, TNFR2, sIL6R, sIL7Ra, sVEGFR1, Gal1, Gal3, Gal7, Gal9, Eselectin, S100A8, HSP70, SAA1, MMP9, TIMP1, Apelin, Vimentin, TPO, Adipsin, Resistin, RBP4, Adiponectin, CRP, Leptin, and Chemerin, IgE, and the radioallergosorbent test (RAST) for: dust mite, cat, dog, hence egg, cow's milk, peanut, and grass pollen mix.
Parental questionnaire	4–6 years of age	Presence of allergies (food, drug, insect venom, inhalation or contact allergy), length, weight, exposure to antibiotics, and exposure to antibiotics and number of courses before two years of age. Information on conditions related to upper respiratory tract infection, pseudocroup, and ear nose and throat (ENT) disorders
doctors' diagnoses (general practitioner) based on the ICPC	4–6 years of age	including treatments and the treating specialist were retrieved. The same doctors' diagnoses as at one year of age were requested, supplemented with: D01 (generalized abdominal pain), D06 (other localized abdominal pain), D12 (obstipation), H71 (otitis media acute), H72 (otitis media with effusion), R08 (other nose symptoms/complaints), and R90 (hypertrophy/chronic infection tonsils/adenoid).

2. Material and methods

2.1. Study design and data collection

The study design and the data collection of the prospective birth-cohort INCA study has been previously described [7]. In short, between 2012 and 2015, 436 infants born at term were recruited from maternity and neonatal wards of four teaching hospitals in the Netherlands. The method of collection and the collected variables are presented in Table 1. Around one year of age, a blood sample was obtained if parents gave additional informed consent (n = 149). The serum samples were aliquoted after centrifugation and stored at −80 °C until further use. The method of analyzing the immune markers with a multiplex immunoassay and the measurement of the IgE antibodies and the radioallergosorbent test (RAST) have been described elsewhere [8,9]. At one year of age, doctors' diagnoses were requested using the general practitioner electronic medical database based on the International Classification of Primary Care (ICPC) [10]. Between May 2018 and May 2019, at 4–6 years of age depending on the year of inclusion, parents of the included children were approached to complete an online questionnaire. In this questionnaire, information about parental-reported allergies (food, drug, insect venom, inhalation or contact allergy) was collected.

2.2. Ethics

Both parents of all participating children gave informed consent. At 4–6 years of age, doctors' diagnoses were only collected after additional informed consent. The study was approved by the ethical board of the St. Antonius Hospital in Nieuwegein, the Netherlands and was registered in the clinical trials register as NCT02536560.

2.3. Data analyses with machine learning

In order to predict allergy from early life exposures/characteristics, it is necessary to select the characteristics that can optimally distinguish between allergic children and healthy not allergic children. In this sense, popular approaches used for feature selection range from univariate statistical considerations, to iterated runs of the same classifier with an increasingly reduced number of features to assess the contribution of the features to the overall result [11,12]. Because allergy is multifactorial and therefore very complex, it is not sufficient to rely on simple statistical analyses. In addition, features extracted using an iterative method on one classifier are likely to work well only for that specific classifier. Following the idea behind *ensemble feature selection* [11], multiple algorithms should be used to obtain a more robust and general predictive performance. An ensemble approach has the advantage of obtaining features that are effective across several classifiers, with a better likelihood of being more representative of the data, and not just of the inner workings of a single classifier. For this study, a set of classifiers was

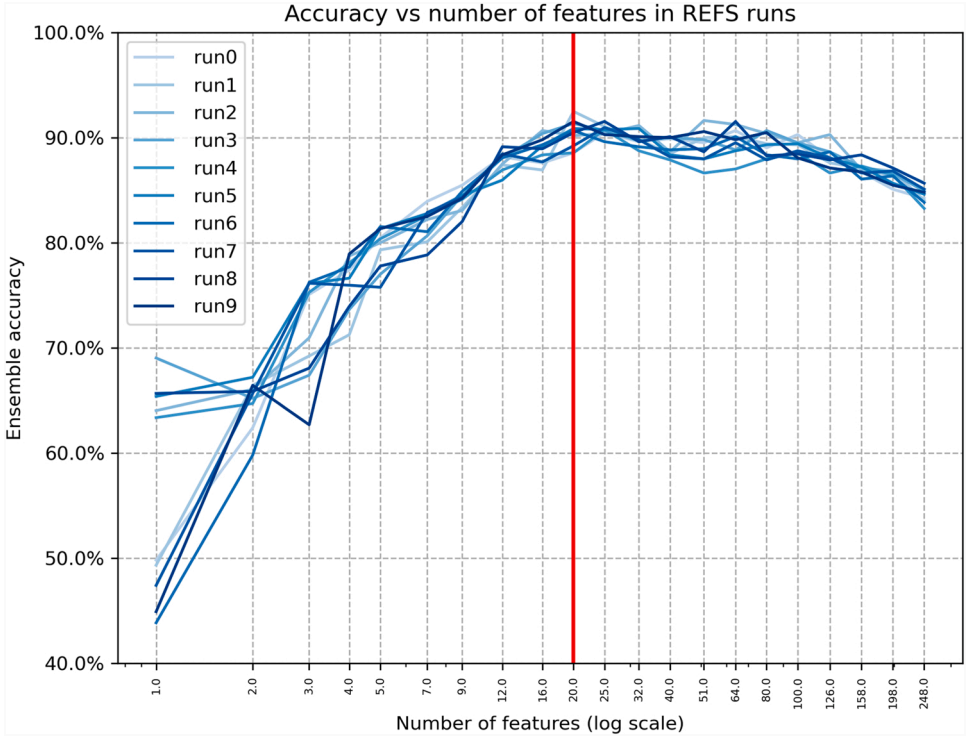

Fig. 1. 10 Runs of the recursive ensemble feature selection (REFS) algorithm. The best number of features is 20.

trained to extract a sorted list of the most relevant features from each. Features were only included if there was less than 10 % missing data. Three features had more than 10 % missing data and were therefore excluded: duration of rupture of membranes, GBS status of the mother, and exposure to antibiotics intrapartum. For the features with less than 10 % missing data, the missing data was imputated with the mean of median for the continuous variables, depending on the distribution of the data, or with the mode for nominal and ordinal variables.

To find the most important features, a recursive ensemble feature selection (REFS) was used, with a 20% step reduction and with eight different classifiers in the ensemble [13], and resampling given the class unbalance. The classifiers are from the scikit-learntoolbox [14]: Gradient Boosting [15], Random Forest [16], Logistic Regression [17], Passive Aggressive [18], Stochastic Gradient Descent [19], Support Vector Machine [20], Ridge [21] and Bagging [22]. Each classifier ranks the features independently and then the features with the highest ranking were chosen with a reduction step of 20%. Next, the best set of features was selected with the highest average accuracy in 10 independent runs in a nested cross-validation scheme [23]. In each run, the accuracy at each step is calculated in a 10-fold cross-validation. Finally, to test the results the Receiver Operating Characteristic (ROC) curve [24] was calculated in five classifiers that were not part of the ensemble to assure no overfitting, namely: Adaboost [25], Extra Trees [26], K-Neighbors [27], LASSO [28] and Multi-layer Perceptron (MLP) [29] classifier.

3. Results and discussion

From 149 children with a blood sample at one year of age, 130 (87%) children had complete follow-up until the age 4–6 years. These 130 children (14 with and 116 without parental-reported allergy) and 248 features were included in the ML analyses.

The recursive ensemble feature selection (REFS) algorithm gets a result of 20 features as shown in Fig. 1 in a stratified 10-fold cross validation scheme with oversampling. Next, the Receiver Operating Characteristic Curves (ROC) for five different classifiers, not included in the original ensemble feature selection technique, were calculated (Fig. 2). Particularly, the Multi-layer Perceptron Classifier showed an area under the curve (AUC) of 0.86 (standard deviation 0.08) which is considered as “very good” in diagnostic accuracy. The resulting features that are predictive for allergy at 4–6 years of age, and presented in Fig. 3, were: tobacco exposure during pregnancy, atopic parents, gestational age in days, days of: diarrhea, cough, rash, and fever during first year of life, and ever being exposed to antibiotics. The included immune mediators/cytokines determined at one year of age were: Resistin, Interleukin 27, Matrix metallopeptidase 9, C-X-C Motif Chemokine Ligand 8, C-C Motif Chemokine Ligand 13, Vimentin, Interleukine-4, C-C motif chemokine 22, Galactokinase 1, Interleukine 6, levels of tumor necrosis factor superfamily 14, and Granulocyte-macrophage colony-stimulating factor.

The use of ML to analyze the complex interplay between early life variables in this large prospectively collected dataset is an important strength of this study. Since ML-based models have a better predictive capacity than statistical-based models, they should have a more important role in future prediction studies, but this requires specific expertize from experts in bioinformatics. Limitations of the study were that we had only 14 children with a parental-reported allergy in our cohort. The results should therefore be interpreted as a pilot. We did not validate the model externally. Moreover, only parental-diagnosed allergy could be used as an outcome measure, because the doctor-diagnosed allergy group was even smaller.

4. Conclusion

We used a state-of-the-art technique to examine the combination of early life factors and later allergy development using ML. Our pilot study shows that an ML-model with 20 exposures/characteristics in early life can predict later allergy with an AUC of 0.86 despite the small sample size. This model emphasizes the importance of early life, since almost all included variables occur in the first year of life. In the future, our ML model still needs to be externally validated in a larger doctor diagnosed allergy dataset.

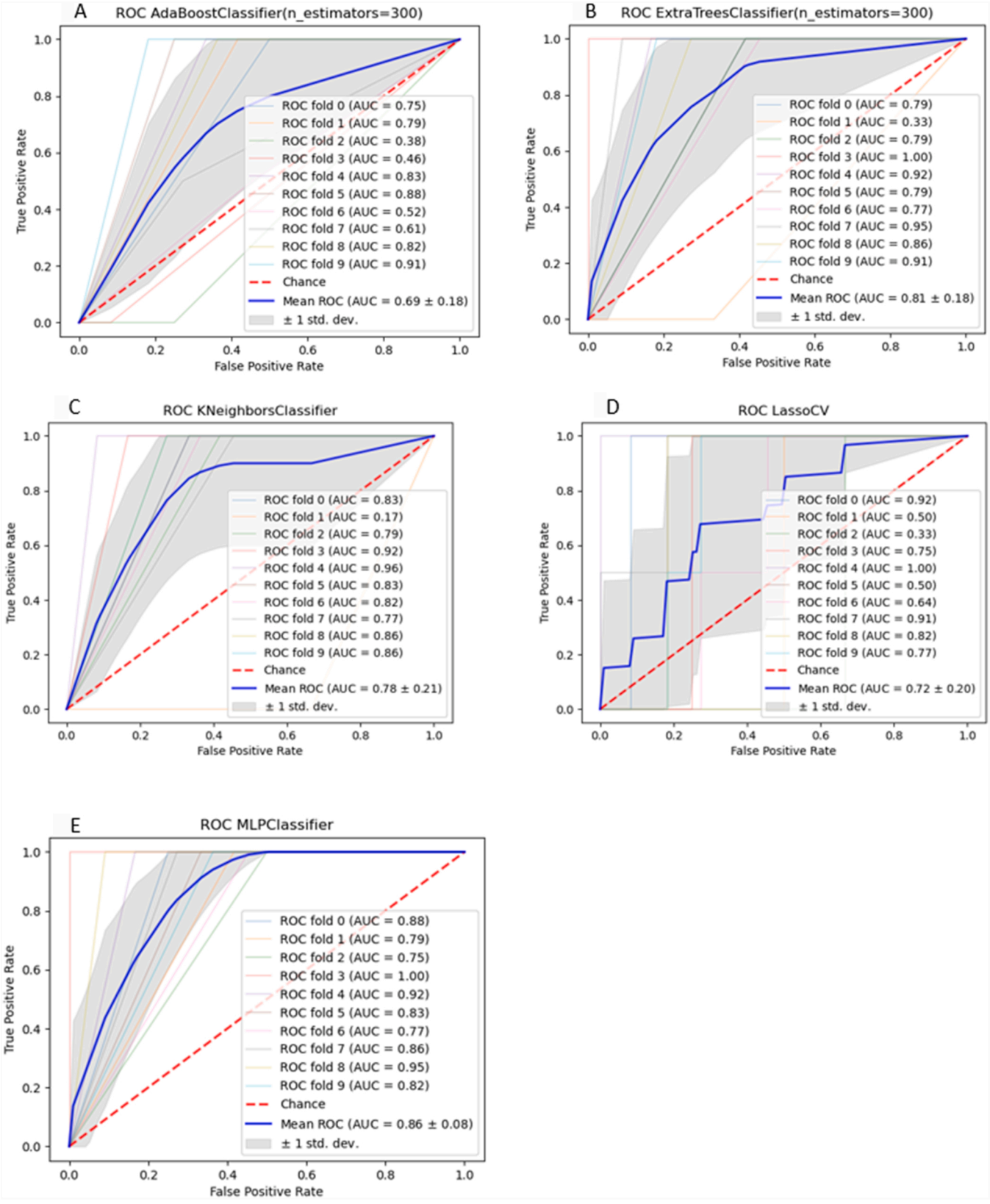

Fig. 2. The Receiver Operating Characteristics (ROC) curves for five different classifiers, that were not included in the original ensemble feature selection technique: 2A = Adaboost, 2B = Extra Trees, 2C = K-Neighbors, 2D = LASSO and 2E = Multi-layer Perceptron classifier.

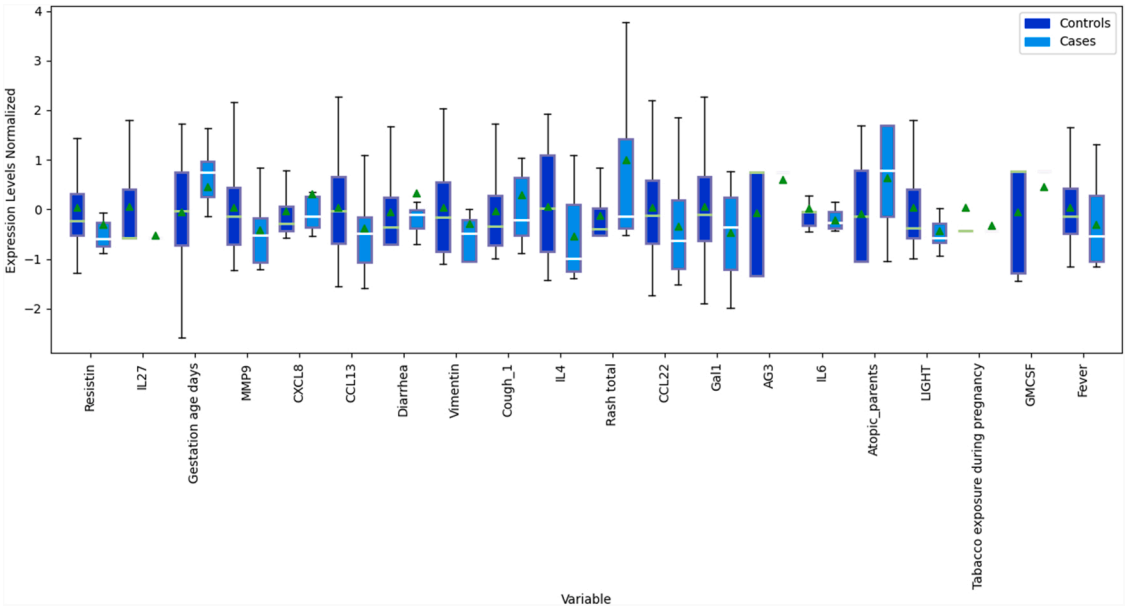

Fig. 3. Boxplot of the included features. IL-27 = Interleukin 27*, MMP9 = Matrix metallopeptidase 9*, CCL13 = C-C Motif Chemokine Ligand 13*, Diarrhea = days of diarrhea during the first year of life, Cough_1 = days of cough during the first year of life, IL-4 = Interleukine 4*, Rash total = days of rash during the first year of life, CCL22 = C-C motif chemokine 22*, Gal 1 = Galactokinase 1*, AG3 = ever being exposed to antibiotics, IL-6 = Interleukin 6*, Atopic_Parents = the presence of atopic parents, LIGHT = levels of tumor necrosis factor superfamily 14*, GMCSF = Granulocyte-macrophage colony-stimulating factor*, Fever = days of fever during the first year of life. *determined at one year of age.

CRediT authorship contribution statement

Kim Kamphorst: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft. **Alejandro Lopez-Rincon:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Arine M. Vlieger:** Conceptualization, Investigation, Writing – review & editing. **Johan Garssen:** Conceptualization, Methodology, Writing – review & editing. **Esther van 't Riet:** Methodology, Writing – review & editing. **Ruurd M. van Elburg:** Conceptualization, Methodology, Investigation, Writing – review & editing.

Declaration of interest

None.

Data Availability

Data will be made available on request.

Acknowledgments

NA.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] P. Eigenmann, M. Atanaskovic-Markovic, J.O.'B. Hourihane, G. Lack, S. Lau, P. Matricardi, A. Muraro, L. Namazova Baranova, A. Nieto, N. Papadopoulos, Testing children for allergies: why, how, who and when: an updated statement of the European Academy of Allergy and Clinical Immunology (EAACI) Section on Pediatrics and the EAACI-Clemens von Pirquet Foundation, Pediatr. Allergy Immunol. 24 (2013) 195–209.

[2] W. Loh, M.L. Tang, The epidemiology of food allergy in the global context, Int. J. Environ. Res. Public Health 15 (2018) 2043.

[3] D. Campbell, R. Boyle, C. Thornton, S. Prescott, Mechanisms of allergic disease–environmental and genetic determinants for the development of allergy, Clin. Exp. Allergy 45 (2015) 844–858.

[4] R.H. Hariri, E.M. Fredericks, K.M. Bowers, Uncertainty in big data analytics: survey, opportunities, and challenges, J. Big Data 6 (2019) 1–16.

[5] I.S. Hofer, M. Burns, S. Kendale, J.P. Wanderer, Realistically integrating machine learning into clinical practice: a road map of opportunities, challenges, and a potential future, Anesth. Analg. 130 (2020) 1115.

[6] K. Kamphorst, A.M. Vlieger, B.C. Oosterloo, S. Waarlo, R.M. van Elburg, Higher risk of allergies at 4–6 years of age after systemic antibiotics in the first week of life, Allergy 76 (2021) 2599–2602.

[7] N. Rutten, G. Rijkers, C. Meijssen, C. Crijns, J. Oudshoorn, C. Van der Ent, A. Vlieger, Intestinal microbiota composition after antibiotic treatment in early life: the INCA study, BMC Pediatr. 15 (2015) 204.

[8] B.C. Oosterloo, B. Van't Land, W. de Jager, N.B. Rutten, M. Klöpping, J. Garssen, A. M. Vlieger, R.M. van Elburg, Neonatal antibiotic treatment is associated with an altered circulating immune marker profile at 1 year of age, Front. Immunol. 10 (2020) 2939.

[9] B.C. Oosterloo, R.M. van Elburg, N.B. Rutten, C.M. Bunkers, C.E. Crijns, C. B. Meijssen, J.H. Oudshoorn, G.T. Rijkers, C.K. van der Ent, A.M. Vlieger, Wheezing and infantile colic are associated with neonatal antibiotic treatment, Pediatr. Allergy Immunol. 29 (2018) 151–158.

[10] M. Verbeke, D. Schrans, S. Deroose, J. De Maeseneer, The International Classification of Primary Care (ICPC-2): an essential tool in the EPR of the GP, Stud. Health Technol. Inform. 124 (2006) 809.

[11] Y. Saeys, T. Abeel, Y.V.d. Peer, Robust feature selection using ensemble feature selection techniques, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2008, pp. 313–25.

[12] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: homogeneous and heterogeneous approaches, Knowl.-Based Syst. 118 (2017) 124–139.

[13] A. Lopez-Rincon, M. Martinez-Archundia, G.U. Martinez-Ruiz, A. Schoenhuth, A. Tonda, Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection, BMC Bioinform. 20 (2019) 1–17.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[15] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[16] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[17] D.R. Cox, The regression analysis of binary sequences, J. R. Stat. Soc.: Ser. B (Methodol.) 20 (1958) 215–232.

[18] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer. Online passive aggressive algorithms, 2006.

[19] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Proceedings of the Twenty-first International Conference on Machine Learning, 2004, 116.

[20] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (1998) 18–28.

[21] A.N. Tikhonov, On the stability of inverse problems, Dokl. Akad. Nauk SSSR (1943) 195–198.

[22] L. Breiman, Pasting small votes for classification in large databases and on-line, Mach. Learn. 36 (1999) 85–103.

[23] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, PLoS One 14 (2019), e0224365.
[24] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (2010) 1315–1316.
[25] R.E. Schapire, Explaining adaboost. Empirical Inference, Springer, 2013, pp. 37–52.
[26] M. Ghaemi, M.-R. Feizi-Derakhshi, Feature selection using forest optimization algorithm, Pattern Recognit. 60 (2016) 121–129.
[27] V. Bala, S. Goyal, Learning from neighbours, Rev. Econ. Stud. 65 (1998) 595–621.
[28] V. Fonti, E. Belitser, Feature selection using lasso, VU Amst. Res. Pap. Bus. Anal. 30 (2017) 1–25.
[29] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron, in: Proceedings of the Third IEEE International Conference on Automatic face and gesture recognition, IEEE, 1998, pp. 454–9.