EAACI guidelines on environmental science in allergic diseases and asthma – Leveraging artificial intelligence and machine learning to develop a causality model in exposomics

Mohamed H. Shamji[1,2] | Markus Ollert[3,4] | Ian M. Adcock[1,2] | Oscar Bennett[5] | Alberto Favaro[5] | Roudin Sarama[1,2] | Carmen Riggioni[6] | Isabella Annesi-Maesano[7] | Adnan Custovic[1,2] | Sara Fontanella[1,2] | Claudia Traidl-Hoffmann[8,9] | Kari Nadeau[10] | Lorenzo Cecchi[11] | Magdalena Zemelka-Wiacek[12] | Cezmi A. Akdis[13] | Marek Jutel[12,14] | Ioana Agache[15]

[1]National Heart and Lung Institute, Imperial College London, London, UK

[2]NIHR Imperial Biomedical Research Centre, London, UK

[3]Department of Infection and Immunity, Luxembourg Institute of Health (LIH), Esch-sur-Alzette, Luxembourg

[4]Department of Dermatology and Allergy Center, Odense Research Center for Anaphylaxis (ORCA), University of Southern Denmark, Odense, Denmark

[5]Faculty Science Limited, London, UK

[6]Pediatric Allergy and Clinical Immunology Service, Institut de Reserca Sant Joan de Deú, Barcelona, Spain

[7]Research Director and Deputy DIrector of Institut Desbrest of Epidemiology and Public Health (IDESP) French NIH (INSERM) and University of Montpellier, Montpellier, France

[8]Environmental Medicine Faculty of Medicine University of Augsburg, Augsburg, Germany

[9]CK-CARE, Christine Kühne Center for Allergy Research and Education, Davos, Switzerland

[10]Sean N. Parker Center for Allergy and Asthma Research, Stanford University School of Medicine, Stanford, California, USA

[11]SOS Allergology and Clinical Immunology, USL Toscana Centro, Prato, Italy

[12]Department of Clinical Immunology, Wroclaw Medical University, Wroclaw, Poland

[13]Swiss Institute of Allergy and Asthma Research (SIAF), University Zurich, Davos, Switzerland

[14]ALL-MED Medical Research Institute, Wroclaw, Poland

[15]Faculty of Medicine, Transylvania University, Brasov, Romania

Correspondence
Mohamed H. Shamji, National Heart and Lung Institute, Imperial College London, London, UK.
Email: m.shamji99@imperial.ac.uk

Ioana Agache, Faculty of Medicine, Transylvania University, Brasov, Romania.
Email: ibrumaru@unitbv.ro

Abstract

Allergic diseases and asthma are intrinsically linked to the environment we live in and to patterns of exposure. The integrated approach to understanding the effects of exposures on the immune system includes the ongoing collection of large-scale and complex data. This requires sophisticated methods to take full advantage of what this data can offer. Here we discuss the progress and further promise of applying artificial intelligence and machine-learning approaches to help unlock the power of complex

Marek Jutel, Department of Clinical Immunology, Wroclaw Medical University, Wroclaw, Poland.
Email: marek.jutel@all-med.wroclaw.pl

environmental data sets toward providing causality models of exposure and intervention. We discuss a range of relevant machine-learning paradigms and models including the way such models are trained and validated together with examples of machine learning applied to allergic disease in the context of specific environmental exposures as well as attempts to tie these environmental data streams to the full representative exposome. We also discuss the promise of artificial intelligence in personalized medicine and the methodological approaches to healthcare with the final AI to improve public health.

KEYWORDS

allergy, artificial intelligence, asthma, environment, exposome

1 | INTRODUCTION

Asthma and allergic diseases are prototypes of environmental-driven diseases with the important polygenic background. Natural and man-made environments, such as air, water and soil quality, together with all the physical, chemical, biological, and psychosocial features of our surroundings, have a major influence on the control and severity of allergic diseases and asthma. The environmental triggers are able to induce also epigenetic changes with a variable effect on allergic disease severity and progression. The application of environmental science to tackle the growing burden of allergic diseases and asthma has been a key priority for the European Academy of Allergy and Clinical Immunology (EAACI) Research Agenda.

An integrated approach toward environmental and health policies is needed to tackle environmental risks, based on high-quality evidence in order to implement appropriate measures. This requires high-quality tools in a form of a framework delivered by academia in the format of evidence-based guidelines. The precision medicine approach based on big data sets has the potential to unveil causality instead of associations and to promote an integrated surveillance network. It is important to note that correlation does not imply causality. Causality and association are two concepts used in research and statistics. The main difference between the two is that causality implies a causal relationship between two variables, while association simply refers to a statistical relationship or correlation between the variables. Causality suggests that a change in one variable causes a change in the other variable, whereas association means that the variables are related in some way.

The current report postulates the methodological approach based on leveraging the power of artificial intelligence (AI) through the use of machine-learning (ML) tools to tackle a range of questions about how the environment is linked to the development and exacerbation of allergic diseases and asthma (Figure 1).

2 | THE RISE OF EXPOSOMICS

During their lifetime individuals are exposed to a wide array of environmental factors, which can have both short- and long-term effects on their health status. These factors include nutrition, levels of physical activity, social and psychological stressors, exposure to toxins and pollutants, allergens, alcohol and smoking habits, and many more. Collectively, these are known as the exposome, and their study is called exposomics.[1,2] An individual's contact with external environmental factors is known as the eco-exposome and the internal effects that occur after interaction with the exposome as the endo-exposome. Three domains of exposome have been defined as general external environment (biodiversity, climate urban environment, socioeconomic factors, professional triggers at occupational exposure); specific external environment (allergens, microbes, diet, tobacco, indoor and outdoor pollutants, and other toxic substances); and host-dependent internal environment (metabolic factors, hormones, inflammation, and oxidative stress).

In their lifetime, people encounter many different exposures. Thus, collecting good-quality data sets is understandably challenging. Nevertheless, there are data collection programs underway aiming to help to understand the impact of exposomes on human health all the way down to the molecular level. The United Kingdom Biobank has collected a huge amount of data from 500,000 individuals during a lengthy study in which they collected a wide range of genetic, phenotypic, clinical, and lifestyle data.[3,4] The "All of Us" research program is ongoing and intends to recruit a million people across the United States with the plan to collect a similarly broad and powerful data set.[5] Rich data sets like these combined with the powerful analytical approaches provided by AI and ML will drive the understanding of the exposome and its impact on health in an unprecedented manner. Data-driven approaches to health risk stratification,[6] multi-omics analysis,[7,8] novel biomarker discovery,[9–11] and causal analysis[12] will invaluably contribute to the growing field of personalized medicine.

3 | WHAT IS AI?

AI is an emerging field which leverages powerful computer algorithms to carry out challenging tasks that surpass the human level intelligence to perform. An important subfield within AI known as ML involves the use of algorithms specially designed to ingest large
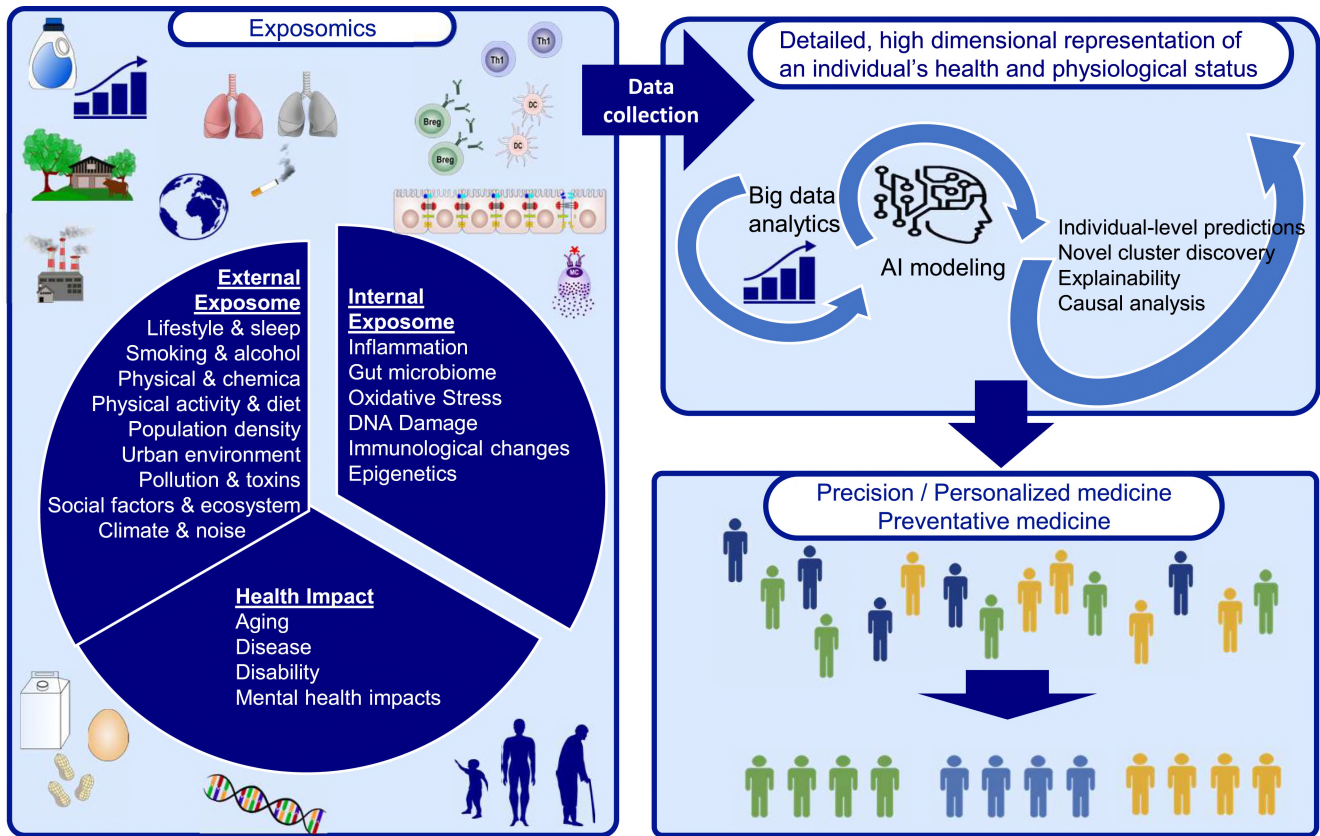
FIGURE 1 Implementation of AI and ML in medical science. The ongoing collection of large-scale and complex data is a prerequisite for implementing AI/ ML in allergy and asthma. Complex exposomics-derived data processed by AI/ML tools provide causality models for exposure and intervention. This accounts for the novel methodological approaches in personalized and preventive medicine aiming at integrated management and individualized healthcare. AI, artificial intelligence; DNA, deoxyribonucleic acid; ML, machine learning

amounts of complex data and automatically extract meaning and insights in an unbiased way. ML models "trained" on data are then able to carry out a range of potentially very useful tasks, such as estimating the risk of an outcome of interest in an individual,[6] finding natural groupings within the data,[13] or automatically extracting meaning from the content of images,[14,15] videos or text.[16] These powerful analytical approaches are changing the way complex, data-rich systems are evaluated, and the medical field is no exception. Comprehensive recent avenues of investigation have been developing involving the use of AI in almost all areas of our discipline.[17]

This report outlines how AI and ML can contribute to the understanding of how environmental exposures over a lifetime (known as the exposome) drive the inception and the severity of allergic diseases and asthma and how we can use this knowledge for primary and secondary prevention.

4 | TYPES OF ML

ML is the process of using mathematical models of data to help a computer learn without direct instruction. This enables a computer system to continue learning and improving on its own, based on experience. Broadly speaking, there are three different ML paradigms: supervised learning,[18] unsupervised learning[19] and reinforcement learning.[20]

Supervised learning involves algorithms and methods that learn the relationship between data points and associated labels. The unsupervised learning involves a problem setup, where labels are not available and the ML algorithm needs to find a useful structure within a data set without the guidance of these explicit labels. Cluster analysis is a typical example of unsupervised learning.[21] An important task closely related to unsupervised learning is dimensionality reduction.[22] This is a way of taking high-dimensional data sets (data sets with a very large number of features) and finding a lower-dimensional representation of them that retains most of the important original information. A simple and commonly used approach is principal component analysis.[23] Following on from dimensionality reduction is the area of feature selection.[24] ML algorithms which are trained to predict specific labels from a collection of features can also be used to determine which of these available features are most predictive of the label under study. Selecting out the most useful features in this way can provide a range of benefits for an ML pipeline[24] but can also provide insight for researchers into the relative importance of the different features used in the predictive task – often an important result on its own.

Another useful application of ML that can be approached in either a supervised or unsupervised way is anomaly detection.[25] This is a problem setup where a large collection of data is available, and the algorithm is trained to recognize when data diverge significantly from a learned representation of "normal."[26]

Computer vision and natural language processing (NLP) are an important group of ML methods and algorithms used to automatically extract the content and meaning contained within images and human language, respectively.[27,28] Within the medical sciences, applications of these methods are growing in importance, with computer vision being applied to solve problems and automate processes in diagnostic imaging[15] and NLP being applied to the automatic parsing of, for example, electronic health records and diagnostic reports.[16]

In the analysis of sequence-type data, ML models can be trained on the past behavior of a system in order to produce forecasts of future predicted behavior.[29,30] Another application is the analysis of data that come in sequences such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and peptide sequences.[31]

5 | ML MODELS

ML model designs can range from very simple to very complex and there are important trade-offs to consider when choosing one.[32] Linear models are usually considered the simplest group of models. These models combine available features in simple, linear ways in order to produce an output.[33] Linear regression and logistic regression are two common examples. Increasing in complexity are tree-based models.[34] They are loosely based on the idea of decision trees but with a wide range of powerful extensions to improve their power and accuracy. Random forests (RF), for example, combine many decision trees together in principled ways to produce more accurate and reliable decisions.[35] The ML principle of "boosting" is often applied to these models. Boosting involves multiple models collected in a group learning to anticipate and correct for the mistakes made by the other models in the group.[36]

Other important modeling approaches worth calling out include support vector machines (SVM)[37] and K-nearest neighbors (KNN)[38] which build up an internal representation of what they learn using specific examples of data points pulled from the training data itself. SVMs also commonly perform complex nonlinear transformations of the space the features lie within to improve the power of the model.

The most complex and potentially powerful types of ML models are artificial neural networks (ANNs).[39] The specific field of building and training ANNs with many layers of neurons is often referred to as deep learning (DN).[40] These models are powerful and have the potential to extract very complex and subtle patterns in data but also require a very large amount of data to learn from in order to be effective.[39] Thus, ANNs are a sensible choice only when working with very large data sets.

6 | VALIDATION OF ML MODELS

Validating the performance of an ML model and confirming any related insights obtained from the pattern of its predictions is needed to avoid overly optimistic conclusions. The most important thing that needs to be demonstrated is that the performance/behavior/

insights observed with the modeling approach applied to specific data will generalize to new data sets and new situations. A range of methods is designed to carry out performance assessments in ways that avoid or correct for bias: holdout test sets, cross validation,[41] leave-one-out approaches,[42] or optimism correction methods.[43] In addition, it is important to validate conclusions in other ways, following the model of trial results replication in other populations, geographical regions, etc. to demonstrate the robust nature of the conclusions being proposed.

7 | AI SAFETY CONSIDERATIONS

7.1 | Explaining ML model decisions

It is desirable to understand how an ML model makes a decision. Of note, providing the user with intelligible explanations of ML model decisions is often considered an important part of deploying ML applications in a safe and responsible way. This form of insight can range from simple to extremely complex to obtain depending on the nature of the model and the task it is performing. Explainability strategies are typically divided into two categories: intrinsic and post hoc explainability.[44] Intrinsic explainability is achieved by using a type of model that is sufficiently simple, or intrinsically understandable, that merely unpacking the internal workings of the model will provide a user with insight into how a model arrives at a decision.[44] Many ML models, particularly ANNs, are internally too complex to provide this opportunity. These complex models are referred to as "black boxes" since it is difficult to look inside and understand how decisions are reached.[45] Post hoc explainability methods like local interpretable model agnostic explanations (LIME)[46] and shapley values[47] are used in these situations.

Another distinction worth highlighting in this context is the difference between local and global explanations of an ML model's predictions.[44] A local explanation provides an explanation of how a model arrived at a specific decision for a specific input. For example, this can provide an explanation for why a model predicted that an individual would respond well to an intervention. A global explainability method would provide information more broadly about how the model makes decisions in general. In other words, what features on average push a model's decision in a certain direction. For example, an insight provided by global model explainability might tell that people who are older usually (but not necessarily always) respond better to a certain treatment.

7.2 | Robustness

ML models tend to provide more reliable predictions when they are presented with data broadly similar to what they were trained on.[48] When outside this region either unintentionally or intentionally an area known as an adversarial attack,[49] the ML model can behave in unpredictable and/or underperformance ways. Mitigation strategies to detect or avoid these scenarios are important to plan for.

7.3 | Subgroup performance variation

For certain types of data, the performance of a model might drop to unacceptable levels.[50] It is important to have validation steps in place to detect these situations and ideally also improve them, for example, with further model training.

7.4 | Fairness

Fairness is the process of understanding bias introduced by data and ensuring that this model provides equitable predictions across all demographic groups. This safety feature is particularly important if certain groups of individuals can be significantly advantaged or disadvantaged by the outcome of a model's decision.[51] Examples include ML applications in credit scoring, predictive policing, or job candidate assessments. There are usually a set of "protected characteristics" which for ethical reasons are not allowed to contribute to ML decisions. Incorporating fairness into ML model decisions usually requires careful thought and sophisticated technical methodology as the simple removal of the protected characteristics from the features presented in a model is often ineffective. Models are often able to learn about the protected characteristics indirectly from the other available features.[52]

7.5 | Privacy

Privacy is an important consideration when working with any sort of personal data. The requirements set out in the relevant legal frameworks such as the general data protection regulation (GDPR) in Europe[53] and the Health Insurance Portability and Accountability Act of 1996 (HIPPA) in the USA[54] is important to adhere to when working with personal data, but there are some additional complexities specific to ML. In particular, it is important to ensure that an ML model has not "memorized" details about specific individuals in a way that could be subsequently extracted if the model is released to users unauthorized to have such information. Differential privacy in ML[55] provides ways to avoid these problems and is an important and active area of ongoing research.

7.6 | Data integrity

Accurate research results are the central element of high-performance biomedical research. To be able to draw meaningful AI-based conclusions from a biomedical research data set, rigorous quality procedures and standardization are needed to be implemented together with an evaluation of data quality issues. This includes approaches for dealing with missing data using imputation methods. Furthermore, to ensure the integrity of data, the information technology (IT) infrastructure and workflows onto which AI algorithms are applied need stringent protection measures (e.g., data encryption, data separation, specific pseudonymization procedures, access restriction, data federation, etc.).

8 | SPECIFIC USE CASES OF AI/ML IN ENVIRONMENTAL SCIENCE FOR ALLERGIC DISEASES AND ASTHMA

8.1 | Prediction of environmental exposures

8.1.1 | Pollen count models

Accurate and up-to-date monitoring of airborne pollen grains on importance due to the dramatic global rise of pollen-induced allergy (Figure 2). Conventionally, airborne pollen samples are collected and counted under the microscope, however, this is limited by the difficulty in identifying pollen at the species level, while requiring highly specialized experts. Automatic pollen recognition becomes thus crucial and can be efficiently solved using DL. Specifically, convolutional neural networks (CNNs) have been used to increase the accuracy of identifying and counting airborne pollen. The study in question used CNNs to distinguish between low-allergenic pollen species *Urtica* and allergic species *Parietaria* of the Urticaceae family. One of the classes was >98% correctly identified, whereas the other two exhibited high error rates. This was thought to be due to a lack of sufficient variability in the training data.[56] Another CNN model was trained with DL using 122,000 pollen reference samples to overcome the laborious and costly process of pollen analysis. This model resulted in high-throughput analysis processing 10,000 pollen grains per minute, which was increased to 600 pollen grains per second following training. Two types of experiments were assessed using the optimized model. The splitting experiment used three samples of two different species for training the CNN model, and a fraction of the same samples were used for validation. A leave-one-out experiment used the same sample set, however, used single test examples and the rest of the data as training data in a sequence of experiments to produce validation predictions across a range of
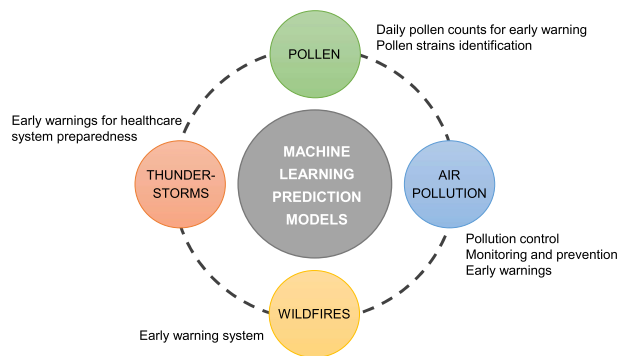

FIGURE 2 Understanding the environment. Artificial intelligence/machine-learning applications for forecasting and mitigating healthcare issues.

examples. The experiments had an overall accuracy of 0.98 and 0.41, respectively.[57]

The ability to accurately forecast the daily concentration of airborne pollen might also benefit from ML support. An ML predictive model trained with data from 24 years' worth of pollen concentration measurement in addition to atmospheric weather data robustly estimated the concentration of the airborne *Ambrosia* pollen with a correlation coefficient between the estimation by the model and actual pollen concentration of 0.82.[58]

8.1.2 | Thunderstorm asthma prediction

Thunderstorm-triggered asthma (TA) is the occurrence of acute asthma attacks immediately following a thunderstorm during the pollen season.[59] Different ML models have been evaluated for the prediction of TA occurrences. One example is a DL neural network (DLNN) model used to generate ≤15-h predictions of thunderstorm occurrences in South Texas.[60] However, not enough data are present on AI prediction models specifically for TA outbreaks, which can be crucial for alerting patients and medical providers prior to an outbreak. One study utilized social media posts on Twitter to provide early alerts for acute outbreaks of TA. The authors created a monitoring algorithm based on the relevance of the tweets and time-between events. In three cases, the algorithm detected the outbreak before the official time, and in five cases prior to news reports.[61]

8.1.3 | Wildfire risk prediction

Wildfires are increasing in frequency in different parts of the world. Identifying the probability of wildfire occurrence is important for asthmatic patients living in areas of high risk because these fires can trigger severe asthma symptoms. Sayad et al.[62] proposed a new methodology for predicting the occurrence of wildfires by utilizing big data, remote sensing, and ML models such as ANN and SVM. Both models had a high prediction accuracy: 98.32% for ANN and 97.48% for SVM. It has also been possible to use data on wildfires to improve our ability to anticipate changes in air pollution using sophisticated ML techniques.[63]

8.1.4 | Air pollution risk prediction

It is well established that long-term exposure to air pollution contributes to the development of and directly exacerbates a range of respiratory diseases. ML classifiers using electronic health records (EHR) and epidemiological data were unable to successfully predict the future risk of asthma attacks in 29,396 patients with asthma in Sweden.[64] The authors suggested that additional data on environmental exposures including weather, pollen, and air pollution levels would be needed to improve prediction models. Indeed, a recent study demonstrated that imbalanced sampling ML approaches can be used to predict the association between indoor air quality exposure and changes in peak expiratory flow rate (PEFR) in subjects with asthma.[65] The performance of these algorithms was further improved by the application of transfer learning, a DL method, which indicated the importance of particulate matter ($PM_{2.5}$) and carbon monoxide (CO) in predicting changes in PEFR. Future studies will have to link measurable asthma features with either local environmental exposure or, preferably, personalized exposure monitoring such as used recently in Delhi with the AirSpec device.[66] Spatial–temporal modeling of six air pollution parameters (CO, PM_{10} and $PM_{2.5}$, nitrogen dioxide, sulfur dioxide, and ozone) using a random forest model identified $PM_{2.5}$ and PM_{10}[67] and distance from parks[68] being associated with seasonal occurrences of asthma in children across Tehran, Iran. Finally, data from the National Air Toxics Assessment (NATA) in the USA were used to assess air pollution in the residential areas of patients with asthma who were part of the AiRway in Asthma (ARIA) cohort.[69] The ML algorithm used (Data-driven ExposurE Profile extraction, DEEP) identified 18 separate air-toxic molecules and 20 combinations of molecules including acrylic acid, ethylidene dichloride, and hydroquinone as being significantly associated with asthma outcomes.

8.1.5 | Dust storms prediction

One of the extreme weather events that are becoming more frequent and severe due to global warming and climate change is desert dust storms or sandstorms, which arise from arid regions when strong winds blow large amounts of loose sand and dirt into other areas.[70] Desert dust contributes to the poor prognosis and mortality of chronic respiratory patients.[71] In a recent study, the estimation of ambient $PM_{2.5}$ in Iraq and Kuwait from 2001 to 2018 using machine learning was reported.[72] Remote sensing random forest ML and a generalized additive mixed model to estimate daily high-resolution (1 km × 1 km) visibility over the region using satellite-based aerosol optical depth (AOD) and airport visibility data were combined. The spatially and temporally resolved visibility data were then used to estimate $PM_{2.5}$ concentrations from 2001 to 2018 by converting visibility to $PM_{2.5}$ using empirical relationships derived from available regional $PM_{2.5}$ monitoring stations.

8.2 | Prediction of clinical outcomes linked to environmental exposures

AI may also be useful to identify the deleterious effects of exposure to environmental stressors and find interventions to reduce these effects using the adverse outcome pathway analysis based on very

large publicly available chemical libraries to generate structure activity relationships.[73]

Examples of ML models that have been assessed in the literature and utilized for assessing the impact of different environmental factors are summarized in Table 1.

8.2.1 | Prediction of the severity of response in patients with atopic dermatitis or contact dermatitis exposed to chemical irritants

Skin serves as a natural barrier against harmful substances, including chemicals, UV light, and pathogens. The disruption of the protective function of the skin, for example, by detergents, is a main risk factor for the development of an epithelial barrier-linked disease such as allergy or asthma. Epidermal proteins such as filaggrin, lipids like ceramide, and tight junctions play an important role in maintaining the skin barrier and their malfunction is associated with inflammatory skin diseases such atopic dermatitis (AD).[79] However, the molecular mechanisms behind the disturbance of the skin barrier are incompletely understood, with limited research on the role of genetics for dermal chemical exposure and uptake. Studies that are more recent suggest a substantial contribution of genetics in this context. In addition, as numerous microorganisms colonize the human skin, many skin diseases (such as AD and psoriasis) are not only characterized by disrupted skin barriers, but also by imbalanced skin microbiota compositions.[80] Although the human gastrointestinal microbiota has an extensive capacity to metabolize environmental chemicals and even contribute to toxicity,[81] the role of the human skin microbiota in the context of dermal exposure to environmental chemicals is less studied and thus, may be underestimated.

Large consortium projects, like the BIOMarkers in atopic dermatitis and psoriasis (BIOMAP)[82] aiming to advance personalized medicine for AD and psoriasis by identifying biomarkers that predict therapeutic response and disease progression are urgently needed. Predictive AI algorithms for skin barrier dysfunction will have to integrate various data sets in a combinatorial approach, such as genetic, epigenetic, transcriptomic, proteomic, functional, and microbiota data. Through harmonized, huge-scale data sets such as the one generated in the BIOMAP approach, available on a secure, centralized, and access-controlled data platform, invaluable bio-resources are built for future research. Such quality- and access-controlled resources will be instrumental not only for validating novel hypotheses but also for interrogating the data for predictive biomarkers via AI-based algorithms.

The potential for novel AI-driven approaches to differentiate between allergic and irritant contact dermatitis has recently been demonstrated in the clinical patch testing model, using four contact sensitizers and two irritants with widely different physicochemical properties and high relevance to occupational exposures. Using combinatorial transcriptome analysis and AI-based ML-driven biomarker discovery, robust gene sets for the distinction between the two disease entities were identified, thus providing high-potential AI-based molecular biomarker candidates for further clinical evaluation.[83]

8.2.2 | Predicting disease sub-phenotypes in patients with food allergy

Diagnostic oral food challenges (OFC) are the gold standard to diagnose food allergies like peanut allergy (PA) and to monitor the impact of therapeutic intervention, such as oral immunotherapy (OIT). To reduce the burden of potentially harmful and unnecessary OFC a clear need exists for the identification of ex-vivo predictive biomarkers.

Deep immune profiling via high-dimensional mass cytometry was applied to provide data-driven targets for correlation with clinical outcomes during OFC in pediatric PA.[84] Comparing OFC-positive and OFC-negative patients, similar immune baseline characteristics were compared and allowed to identify immune changes in peripheral blood that was specific for allergic reactivity in peanut-allergic reactions. Using a novel unsupervised computational analysis for clustering and dimension reduction, which is adapted for huge-scale data sets of >1 billion cells enabling large overclustering (1024 clusters), a high-resolution view into immune cell populations and subpopulations can be achieved without the need for reclustering.[85] Such studies provide a comprehensive overview of temporal changes in immune signatures during OFC in PA, reflecting dynamic processes of immune cell migration and inflammation. Combined with other harmonized omics and clinical data sets, such ML-informed huge data analysis is expected to pave the way toward biomarker discovery for endotyping patients and predicting clinical outcomes in PA and other food-allergic conditions. Of note, the analysis and visualization using such novel huge-scale clustering algorithms can be performed without down-sampling, thus ensuring the detection of rare, but clinically important immune events.

8.3 | AI approaches to explore the exposome and personalized approaches

So far, research that uses ML to explore the combination of environmental factors simultaneously is scanty. A recent study used ML approaches to characterize the urban exposome predisposition to obesity.[86] The pluralistic analysis of environmental obesogens strengthened the existing evidence on the role of neighborhood socioeconomic position, urbanicity, and air pollution. In the field of asthma and allergy, the approach made it possible to assess associations between a large set of exposures and asthma outcomes in birth cohorts. However, it cannot address complex interactions (i.e., of order ≥3) or a mixture of effects.[87] Recently, ML took up the ambitious challenge of achieving precision medicine in allergy characterizing allergic endotypes, understanding allergic multimorbidity relationships, contextualizing the impact of the exposome and

TABLE 1 Examples of ML models that have been assessed in the literature and utilized for assessing the impact of different environmental factors.

Environmental factor	Model type	Prediction Target	Data source	Training data	Validation data	Performance Metric	Metric score	References
Pollen	• DNN • RF • XGB • Bayes-R	Ambrosia pollen daily concentrations	European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric weather and land surface reanalysis	ECMWF data 1987–2011	ECMWF data 2012–2017	Pearson correlation coefficient	DNN: 0.82 RF: 0.81 XGB: 0.81 Bayes-R: 0.75	58
	CNNs	Increase accuracy in counting airborne pollen and distinguish between the Urtica and Parietaria species of the Urticaceae family	Pollen grains were collected from four different plants per species. Urticaceae pollen collected from aerobiologic samples from Leiden, the Netherlands as well as from Lleida and Vielha, Catalonia, and Spain	Pollen image data set (Pollen_Projector)	Urticaceae pollen from aerobiologic samples collected from different locations in Spain and the Netherlands	Percentage successful identification	Urtica membranacea was successfully identified (>98%). Urtica and Parietaria showed high error rates (up to 44.4%)	56
	CNNs	Present an automated method for pollen analysis	Reference pollen library prepared from flower fields in the provinces of Scania and Småland, southern Sweden between 2012 and 2020	Splitting experiment: all samples Leave-one-out: two-third of the samples	Splitting: a fraction of the training data Leave-one-out: remaining samples	Prediction accuracy	Splitting experiment: 0.98 Leave-one-out experiment: 0.41	57
Thunderstorms	Deep-Learning Neural Network (DLNN)	Thunderstorm occurrence within 400km² of South Texas domains for up to 15 h in advance	North American Mesoscale Forecast System (NAM)	2004–2006 and 2009–2013 NAM data	2007–2008 NAM data	AUC	0.860	60
	Multiple ensemble predictions a poor man's ensemble of lagged deterministic forecasts, a high resolution-limited area ensemble (AromeEPS), and two global ensembles (PEARP and IFSens)	Predicting the probability of thunderstorm occurrence	Lightning data provided by the Météorage company. The model compared the raw Arome-EPS ensemble to the calibrated AromeEPS + AROLAG/AromeEPS + PEARP/PEARP + IFSens blends	One month from the period of June to August	Data from June to August 2018	Coefficients of the reliability calibration over 3 months	AromeEPS+AROLA: 0.147–0.165 AromeEPS + PEARP: 0.147–0.208 PEARP + IFSens: 0.249–0.325	74
	Four-step architecture combining natural language processing and statistical monitoring	Automated identification of reports of asthma in tweets, and using this data to detect outbreaks prior to the event	Twitter posts from the area of Melbourne, Australia between 2014 and 2016. Eighteen experimental combinations were conducted using different data set and classifier combinations	Tweet vectors data set	Eighteen experimental combinations	Thunderstorm asthma detection	Three asthma outbreak cases were detected 9 h before official reports, and five were detected before news reports	61

(Continues)

TABLE 1 (Continued)

	Model type	Prediction Target	Data source	Training data	Validation data	Performance Metric	Metric score	References
Wildfires	Big Data, Remote Sensing, and Data-Mining algorithms: • ANNs • SVM	Predicting the occurrence of wildfires	The Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), and Thermal Anomalies (TA)	70% of the total data collected	30% of the remaining data	Prediction accuracy percentage	ANN: 98.32% SVM: 97.48%	62
	• LR • RF • SVM • MLP	To identify wildfire events with a high probability of becoming a large wildfire.	• Environmental Information Network of Andalusia (REDIAM) • The Spanish Meteorological Agency (AEMET) • The Regional Forest Fire Fighting Plan of Andalusia (INFOCA)	Different data sets were compared with the different models tested. Data sets: • Original • ADASYN • RUS • SMOTE TOMEK • Smote • TOMEK LINKS	• Original • ADASYN • RUS • SMOTE TOMEK • Smote • TOMEK LINKS	Recall F1 score G-mean	LR model using SMOTE, ADASYN, SMOTE TK, and RUS data sets gave the best results and greater accuracy in predicting large wildfires. Recall: 0.87 F1: 0.53 G-means: 0.63–0.76	75
Air pollution	• LR • SVM • SNNs • RF • XGB • LightGBM	Predicting the peak demand days of cardiovascular disease admissions using the hospital admissions data, air quality data, and meteorological data in Chengdu, China from 2015 to 2017	Hospital admissions data from the Health Information Centre of Sichuan Province, China from 1 January 2015 to 31 December 2017. Meteorological data and air quality data derived from the Chengdu Meteorological Monitoring Database.	80% of data	Remaining 20% of data	AUC	LR: 0.842; SVM: 0.834; ANN: 0.890; RF: 0.926; XGB: 0.930; LightGBM: 0.940	76
						Logarithmic loss function	LR: 0.513; SVM: 0.344; ANN: 0.296; RF: 0.358; XGB: 0.227; LightGBM: 0.218	
						Accuracy	LR: 0.766; SVM: 0.748; ANN: 0.858; RF: 0.862; XGB: 0.876; LightGBM: 0.913	
						Sensitivity	LR: 0.848; SVM: 0.879; ANN: 0.333; RF: 0.909; XGB: 0.818; LightGBM: 0.758	
						Specificity	LR: 0.751; SVM: 0.724; ANN: 0.951; RF: 0.854; XGB: 0.886; LightGBM: 0.941	
						Precision	LR: 0.378; SVM: 0.362; ANN: 0.551; RF: 0.527; XGB: 0.563; LightGBM: 0.695	

TABLE 1 (Continued)

Model type	Prediction Target	Data source	Training data	Validation data	Performance Metric	Metric score	References
• LR • SVM • RF • KNN • XGB-Tree and XGB-Linear	Forecast the pattern of demand for hemorrhagic stroke healthcare services based on air quality 2016–2017	Hemorrhagic stroke events from Center for Disease Control and Prevention in the Longquanyi District of China. Air pollution data were obtained from the environmental monitoring stations.	MaxLag-N data subset		AUC	0.7971	77
	Predict the weekly number of childhood asthma admission in the greater Athens area, Greece.	Datasets included the years 2001–2004 in Greater Athens. • Hourly meteorological data from the National Observatory of Athens. • Ambient air pollution data from seven different areas. • Medical data were obtained from the three main children's hospitals.	2001–2003 period data set	2004 data set	Mean bias error (MBE) Root-mean-square deviation (RMSE) R-squared (R2) Index of Agreement (IA)	MBE: 1.4 RMSE: 6.8 R2: 0.528 IA: 0.837	78

Abbreviations: ANN, artificial neural network; AUC, area under a receiver operating characteristic curve; Bayes-R, Bayesian ridge; CNN, convolutional neural network; DNN, deep neural networks; KNN, K-nearest neighbor algorithm; LightGBM, light-gradient boosting machine; LR, logistic regression; MLP, multi-layer perception; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

ancestry/genetic risks, achieving actionable multi-omics integration, and using this information to develop adequately powered patient cohorts and refined clinical trials.[88]

The domain of unsupervised learning, and in particular the methodologies provided by cluster analysis have a range of powerful ways to discover natural groupings in high-dimensional data sets which would be challenging to surface in simpler or more manual ways.[21] The natural groupings discovered in these ways can unearth previously undiscovered subtypes of disease or subgroups within exposure patterns hidden in large and complex exposome data sets.[89] The real power of these insights is provided by the possibility of tailored diagnosis, treatment, or prevention strategies.

These AI approaches will make it possible to unlock the combined potential of clinical data and research data. Multiple data sets such as omics data (genomics, epigenetics, proteomics, etc.), deep immune profiling, and digital phenotyping (voice, mobility measurement, etc.) can be interrogated together with nutrition, metagenomics, and environmental exposure information to identify disease hubs and pathways. High-quality clinical readouts from real-world evidence (RWE) and patient-reported outcomes will further add to a better stratification of patients at risk of developing certain disease phenotypes. AI is poised to provide personalized approaches in many ways.

8.3.1 | Risk stratification

One of the core functionalities provided by AI modeling is a sophisticated and powerful way of producing predictions of the risk of certain outcomes. Clinical scoring systems estimating a risk across a population are to be replaced by personalized risk profiles based on a wide range of factors presented as features of an ML model.[6] These types of personalized risk stratifications will provide the opportunity to better target interventions for individuals where they will make the most difference.

8.3.2 | Cluster analysis

As described above, the domain of cluster analysis, which sits within the unsupervised learning paradigm, provides the opportunity to automatically identify natural groupings of individuals within datasets which share characteristics in important and potentially complex ways.[21] It is likely that the identification of such clusters within disease states or exposure patterns will provide opportunities to tailor interventions in more personalized ways for the individuals within these groups.

8.3.3 | Causal analysis

Data-driven approaches that look for the underlying signatures of cause and effect will continue to provide vital insights into how to

intervene in biological processes to promote health and to treat disease in individuals. For example, asthma and other allergic diseases are complex multifactorial syndromes. AI has identified the existence of underlying causal pathways that are distinct from their downstream pathways related to disease symptoms in other complex diseases or syndromes such as rheumatoid arthritis.[90,91] This suggests that targeting the downstream pathways may treat symptoms but will not elicit a cure unless the upstream tissue remodeling pathways are targeted, that is, the precise disease endotypes are identified and treated. These pathways, such as calcium sensing pathways in the case of asthma, may link airway smooth muscle function with airway hyperresponsiveness and remodeling.[92]

8.3.4 | Uplift modeling of randomized controlled trials (RCTs) data

When data sets are collected by investigating the causal effect of an intervention, usually during RCTs, an ML approach known as uplift modeling[93] can be used to extend conclusions beyond an average treatment effect across a population to allow predictions of treatment effects at the individual level. This has the potential to massively increase the amount we can learn from such trials if they are sufficiently powered with large enough sample sizes to support these approaches. For example, demonstrated the benefits of using uplift modeling applied retrospectively to an RCT data set from a trial looking into the effects of chemotherapy agents in patients with colon cancer.[94] They showed that the resulting AI model is able to predict treatment response accurately enough that it was possible to transform an ineffective chemotherapy agent combination into an effective one when its application was individually targeted using the trained ML model.

8.3.5 | AI for reducing healthcare costs

There is a wide range of ways that AI-based transformation in healthcare will increase efficiency and drive down the costs. There are applications of AI that can contribute to this transformation at all levels of the healthcare organizational structure, from the strategic and operational delivery of healthcare all the way down to how we make decisions about individual patients.

Perhaps, the examples most relevant to AI and the exposome would fall into the categories of personalized preventative medicine. AI will provide the opportunity to learn, on a very individual level, the way the exposures accumulated in life are linked to health problems which in turn lead to healthcare costs. These insights will provide the opportunity to understand how to target interventions in more efficient and impactful ways in order to avoid or mitigate these future health problems along with the associated costs of addressing them.

8.3.6 | AI for improving patients' quality of life

The field of personalized medicine will benefit greatly from the insights provided by AI-powered analytical approaches.[95] The ability to tailor treatments and interventions to the unique characteristics of an individual will not only make them more effective at treating disease and disability, but also make it easier to avoid side effects or unintended negative consequences of these interventions.[94] Personalized medicine, reinforced in this way will lead to better health outcomes and improved quality of life for patients.

8.4 | AI for clinical and biomedical research

Close collaboration between researchers and healthcare professionals (HCPs) is essential for the translation of research into clinical practice. This is especially importance for AI-related research since the deployment of AI-based solutions will involve more than the practicalities of distribution and deployment but also a culture change toward accepting AI as part of the way healthcare is managed.[96] Involving clinical staff in AI research and in the development journey will help to raise awareness and understanding of AI among the healthcare workforce. Additionally, as with other types of translational research, maintaining a close collaboration between researchers and HCPs will be vital to ensure the solution to the problems to deliver the best value to the clinic.

8.5 | AI for identifying novel biomarkers

AI provides the powerful pattern recognition methodology necessary to identify more complex associations between clinical states of interest and physiological, genetic, or biochemical markers.[9] It may even become possible to discover more complex "meta-biomarkers" which would consist of complex combinations of individual biomarkers considered together rather than in isolation. The ability of machine-learning approaches to identify the predictive power of complex patterns of features could make this possible.

A large pan-European consortium project (taxonomy, treatments, targets, and remission, 3TR) has recently been initiated in the field of immune diseases.[97] As its unique aspect, the 3TR precision medicine project brings several medical specialties together (respiratory medicine, rheumatology, neurology, and gastroenterology), to study disease mechanisms across seven disease entities: asthma, chronic obstructive pulmonary disease, systemic lupus erythematosus, rheumatoid arthritis, multiple sclerosis, ulcerative colitis, and Crohn's disease. Despite the fact that these autoimmune, inflammatory, and allergic diseases are highly heterogeneous conditions in their clinical phenotype, it has been shown that they can share certain genetic and epigenetic risks and several disease pathways. Consequently, individuals with one disease may share an inflammatory molecular pattern

with individuals of the other diseases, and thus may share pathways of response to treatment and disease progression. The 3TR project will generate huge amounts of molecular and clinical metadata that will be integrated and constantly analyzed. With the help of novel AI algorithms, data from one disease will be meta-analyzed across diseases, thus accelerating the identification of new biomarkers for responsiveness or nonresponsiveness to therapy or novel targets for therapeutic intervention based on molecular pathway similarity.

8.6 | AI enabling the access and use of real-world data (RWD)

RWD collected in somewhat passive ways, such as in hospital EHR systems or routine public health data capturing procedures, compared with the more targeted and deliberate data collection during clinical trials, tends to consist of much bigger datasets, often multiple orders of magnitude bigger. However, these types of data sets are less structured (e.g., lots of information locked away in free text) and there are many data quality issues to be solved (e.g., missing data fields, anomalous data points). These problems have often hampered traditional approaches to analysis, but there are ways that AI is beginning to help unlock the potential within these data sets. For example, NLP is being used to extract content and meaning from the unstructured free text in clinical reports and medical notes in automated and easily scalable ways.[16] AI models can also be used to impute missing values in smart, more effective ways,[98] as well as automate the identification of anomalous, likely inaccurate or misleading data points using the AI approaches of anomaly detection.[25]

Combining AI with other technologies, the so-called AI-driven "game changers," has the potential to deliver transformative solutions and can be harnessed to enhance current efforts to address environmental issues. These game changers can be defined by five primary features as previously published by the World Economic Forum. Some of these features include transformationally impact (i.e., ability to completely alter or disrupt current approaches), adoption potential (i.e., the effect of population size toward the approach), and systems impact (i.e., ability to shift the dial across human systems). Moreover, other features of game changers include the centrality of AI to the solution and finally a realizable enabling environment (i.e., whereby enabling environment can be identified and supported). An example of such a game changer includes the Natural Capital Project and InVEST (a computer software program helping decision-makers) to plan cities for more natural environments and help with carbon sequestration. This program focuses on understanding human dependence and impacts on nature and the deep societal transformations needed to secure people and nature.[99] The work spans fundamental research and policy-oriented initiatives to open inclusive and green development pathways. InVEST identify the locations where conservation should be a top priority because of ecological services provide a high economic value. These calculations about how preserving land and the environment provide financial benefits have played a monumental role

in benefitting both the people and the habitat of regions around the world. The InVEST AI program co-develops pragmatic approaches, engaging with governments, multilateral development banks, investors, businesses, farmers and ranchers, communities, and nongovernmental organizations (NGOs).

9 | LIMITATIONS AND BARRIERS OF APPLICATION IN ALLERGY ASTHMA RESEARCH AND MANAGEMENT

AI/ML has the potential to revolutionize the field of allergy diagnosis and treatment, but there are also limitations to these technologies. One limitation is the lack of data diversity, as most current models are trained on data from a limited demographic, resulting in potential biases and inaccuracies when applied to diverse populations. Another limitation is the need for large amounts of labeled data for training AI/ML models, which can be difficult to obtain in the case of rare allergies. Additionally, there is a lack of understanding of the underlying mechanisms of many allergies, making it difficult to develop accurate models. Finally, there is a potential for overdiagnosis and overtreatment if AI and ML are not properly validated and integrated into clinical practice. AI and ML models may not be able to account for individual differences in patients, such as genetic variations, which can lead to inaccurate diagnosis and treatment. AI and ML models are only as good as the data they are trained on, so if the data are biased or not representative of the population, the models will also be biased and not generalizable.

10 | WHAT THE FUTURE HOLDS?

The application of AI to the domain of exposomics has the potential to unlock a whole range of impactful insights from the large amount of complex data. The power of ML to investigate the impact of exposome will continue to expand as the power of hardware used to train and run ML models and the sophistication of the modeling approaches to solve problems improves exponentially.

There are some challenges for AI. AI is still a fairly nascent field and there is a degree of trust and acceptance that still needs to be achieved, especially in high-stake domains like health and biomedical research.[96] This will likely come in time as the field matures and further awareness of and research goes into the principles of AI safety.

The understanding and assessing the outcomes of AI research applied to exposomic data requires a degree of understanding of AI methodology which is not universally present in the healthcare community yet. We hope that articles like this one which introduce and explain the underlying concepts can make the field more accessible to a wider audience.

Another issue to consider is that the size and complexity of both the data and the models being used to carry out the analysis will become more difficult in practice.[100] Moving toward a research culture in which open access to both the code and data

will likely help to make this more achievable. It may even be possible for research groups to publish the trained models themselves to demonstrate the behavior they claim. However, all this will be done in a way that assures the privacy of any sensitive data involved – a complication likely to pose its own challenges down the road.

Finally, the power of AI and ML to investigate the exposome is directly related to the quality of analyzed data. It is key to drive forward scalable ways to collect large, high-quality data sets to support this type of research.[101] Centralised models of data collection like those ones demonstrated by the UK Biobank[4] and the "All of Us" research program[5] are good examples. More such programs will be vital to provide opportunities to unlock the power of data hungry AI in the future.

ORCID

Mohamed H. Shamji https://orcid.org/0000-0003-3425-3463
Markus Ollert https://orcid.org/0000-0002-8055-0103
Ian M. Adcock https://orcid.org/0000-0003-2101-8843
Oscar Bennett https://orcid.org/0000-0003-4546-1807
Carmen Riggioni https://orcid.org/0000-0002-8745-0228
Isabella Annesi-Maesano https://orcid.org/0000-0002-6340-9300
Claudia Traidl-Hoffmann https://orcid.org/0000-0001-5085-5179
Kari Nadeau https://orcid.org/0000-0002-2146-2955
Lorenzo Cecchi https://orcid.org/0000-0002-0658-2449
Magdalena Zemelka-Wiacek https://orcid.org/0000-0001-7201-8638
Cezmi A. Akdis https://orcid.org/0000-0001-8020-019X
Ioana Agache https://orcid.org/0000-0001-7994-364X

REFERENCES

1. Smith MT, de la Rosa R, Daniels SI. Using exposomics to assess cumulative risks and promote health. *Environ Mol Mutagen*. 2015;56(9):715-723. doi:10.1002/em.21985
2. Agache I, Miller R, Gern JE, et al. Emerging concepts and challenges in implementing the exposome paradigm in allergic diseases and asthma: a Practall document. *Allergy*. 2019;74(3):449-463. doi:10.1111/all.13690

3. Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

4. UK Biobank. Retrieved December 6, 2022. www.ukbiobank.ac.uk/.

5. NIH All of Us Research Program. Retrieved December 6, 2022. https://allofus.nih.gov/

6. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J*. 2017;38(23):1805-1814. doi:10.1093/eurheartj/ehw302

7. Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623-629. doi:10.1038/s41586-021-04278-5

8. Radzikowska U, Baerenfaller K, Cornejo-Garcia JA, et al. Omics technologies in allergy and asthma research: an EAACI position paper. *Allergy*. 2022;77(10):2888-2908. doi:10.1111/all.15412

9. Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*. 2012;28(13):1766-1774. doi:10.1093/bioinformatics/bts238

10. Ogulur I, Pat Y, Ardicli O, et al. Advances and highlights in biomarkers of allergic diseases. *Allergy*. 2021;76(12):3659-3686. doi:10.1111/all.15089

11. Celebi Sozener Z, Ozdel Ozturk B, Cerci P, et al. Epithelial barrier hypothesis: effect of the external exposome on the microbiome and epithelial barriers in allergic disease. *Allergy*. 2022;77(5):1418-1449. doi:10.1111/all.15240

12. Prosperi M, Guo Y, Sperrin M. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*. 2020;2:369-375. doi:10.1038/s42256-020-0197-y

13. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2

14. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6

15. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K. Deep learning for chest X-ray analysis: a survey. *Med Image Anal*. 2021;72:102125. doi:10.1016/j.media.2021.102125

16. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak*. 2021;21(1):179. doi:10.1186/s12911-021-01533-7

17. Türk M, Ertaş R, Zeydan E, et al. Identification of chronic urticaria subtypes using machine learning algorithms. *Allergy*. 2022;77(1):323-326. doi:10.1111/all.15119

18. Hastie T, Tibshirani R, Friedman J. Overview of supervised learning. *The Elements of Statistical Learning*. Springer; 2008:9-41.

19. Hastie T, Tibshirani R, Friedman J. Unsupervised learning. *The Elements of Statistical Learning*. Springer; 2008:485-585.

20. Sutton R, Barto A. *Reinforcement Learning*. MIT Press; 2018.

21. Frades I, Matthiesen R. Overview on techniques in cluster analysis. *Methods Mol Biol*. 2010;593:81-107. doi:10.1007/978-1-60327-194-3_5

22. Sorzano CO, Vargas J, Montano AP. A survey of dimensionality reduction techniques. arXiv; 2014. 10.48550/arXiv.1403.2877.

23. Abdi H, Williams L. Principal component analysis. *Comput Stat*. 2010;2(4):433-459. doi:10.1002/wics.101

24. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517. doi:10.1093/bioinformatics/btm344

25. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv*. 2009;41:1-58. doi:10.1145/1541880.1541882

26. Nakao T, Hanaoka S, Nomura Y, et al. Unsupervised deep anomaly detection in chest radiographs. *J Digit Imaging*. 2021;34(2):418-427. doi:10.1007/s10278-020-00413-2

27. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med*. 2021;4(1):5. doi:10.1038/s41746-020-00376-2

28. Chowdhary K. Natural language processing. *Fundamentals of Artificial Intelligence*. Springer; 2020:603-649.

29. Hamilton J. *Time Series Analysis*. Princeton University Press; 1994.

30. Pavlyshenko, B. M.. Linear, machine learning and probabilistic approaches for time series analysis. arXiv; 2017. doi: 10.48550/arXiv.1703.01977.

31. Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*. 2006;65(1):40-48. doi:10.1002/prot.21078

32. Osisanwo F, Akinsola J, Awodele O, et al. Supervised machine learning algorithms: classification and comparison. *Int J Comp Trends Technology*. 2017;48(3):128-138. doi:10.14445/22312803/IJCTT-V48P126

33. Maulud D, Abdulazeez A. A review on linear regression comprehensive in machine learning. *J Appl Sci Technol Trends*. 2020;1(4):140-147. doi:10.38094/jastt1457

34. Chen H, Zhang H, Si S. Robustness Verification of Tree-Based Models. 33rd Conference on Neural Information Processing Systems; 2019.

35. Louppe G. Understanding random forests: from theory to practice. arXiv; 2015. doi: 10.48550/arXiv.1407.7502.

36. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016;785–794. doi: 10.1145/2939672.2939785

37. Byvatov E, Schneider G. Support vector machine applications in bioinformatics. *Appl Bioinforma*. 2003;2(2):67-77.

38. Zhang ML, Zhou ZH. A k-nearest neighbor based algorithm for multi-label classification. *IEEE Int Conf Granular Comput*. 2005;2:718-721. doi:10.1109/GRC.2005.1547385

39. Krogh A. What are artificial neural networks? *Nat Biotechnol*. 2008;26(2):195-197. doi:10.1038/nbt1386

40. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

41. Browne MW. Cross-validation methods. *J Math Psychol*. 2000;44(1):108-132. doi:10.1006/jmps.1999.1279

42. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recogn*. 2015;48(9):2839-2846. doi:10.1016/j.patcog.2015.03.009

43. Iba K, Shinozaki T, Maruo K, Noma H. Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Med Res Methodol*. 2021;21(1):9. doi:10.1186/s12874-020-01201-w

44. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)*. 2020;23(1):18. doi:10.3390/e23010018

45. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x

46. Lee E, Braines D, Stiffler M, Hudler S, Harborne D. Developing the sensitivity of LIME for better machine learning explanation. Artificial intelligence and machine learning for multi-domain operations applications. *Proc SPIE*. 2009;1100610:1100610. doi:10.1117/12.2520149

47. Merrick L, Taly A. The explanation game: explaining machine learning models using Shapley values. *Machine Learning and Knowledge Extraction*. Springer; 2020:17-38. doi:10.48550/arXiv.1909.08128

48. Drenkow N, Sani N, Shpitser I, Unberath M. Robustness in Deep Learning for Computer Vision: Mind the Gap? arXiv; 2021. doi: 10.48550/arXiv.2112.00639.

49. Kurakin A, Goodfellow I, Bengio S. Adversarial Machine Learning at Scale. arXiv; 2017. doi: 10.48550/arXiv.1611.01236.

50. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn*. 2020;2020:151-159. doi:10.1145/3368555.3384468

51. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2022;54(6):1-35. doi:10.1145/3457607

52. Caton S, Haas C. Fairness In Machine Learning: A Survey. arXiv. doi: 10.48550/arXiv.2010.04053.

53. GDPR. Retrieved December 7, 2022. https://gdpr-info.eu/.

54. HIPAA Home. Retrieved December 7, 2022. https://www.hhs.gov/hipaa/index.html.

55. Ji Z, Lipton Z, Elkan C. Differential Privacy and Machine Learning: a Survey and Review. arXiv; 2014. doi: 10.48550/arXiv.1412.7584.

56. Polling M, Li C, Cao L, et al. Neural networks for increased accuracy of allergenic pollen monitoring. *Sci Rep*. 2021;11(1):11357. doi:10.1038/s41598-021-90433-x

57. Olsson O, Karlsson M, Persson AS, et al. Efficient, automated and robust pollen analysis using deep learning. *Methods Ecol Evol*. 2021;12(5):850-862. doi:10.1111/2041-210X.13575

58. Zewdie GK, Lary DJ, Levetin E, Garuma GF. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *Int J Environ Res Public Health*. 2019;16(11):1992. doi:10.3390/ijerph16111992

59. D'Amato G, Akdis CA. Desert dust and respiratory diseases: further insights into the epithelial barrier hypothesis. *Allergy*. 2022;77(12):3490-3492. doi:10.1111/all.15392

60. Kamangir H, Collins W, Tissot P, King A. A deep-learning model to predict thunderstorms within 400 km^2 South Texas domains. *Meteorol Appl*. 2020;27:e1905. doi:10.1002/met.1905

61. Joshi A, Sparks R, McHugh J, Karimi S, Paris C, MacIntyre CR. Harnessing tweets for early detection of an acute disease event. *Epidemiology*. 2020;31(1):90-97. doi:10.1097/EDE.0000000000001133

62. Sayad Y, Mousannif H, Moatassime H. Predictive modeling of wildfires: a new dataset and machine learning approach. *Fire Saf J*. 2019;104:130-146.

63. Li L, Girguis M, Lurmann F, et al. Ensemble-based deep learning for estimating PM$_{2.5}$ over California with multisource big data including wildfire smoke. *Environ Int*. 2020;145:106143. doi:10.1016/j.envint.2020.106143

64. Lisspers K, Ställberg B, Larsson K, et al. Developing a short-term prediction model for asthma exacerbations from Swedish primary care patients' data using machine learning – based on the ARCTIC study. *Respir Med*. 2021;185:106483. doi:10.1016/j.rmed.2021.106483

65. Bae WD, Kim S, Park CS, et al. Performance improvement of machine learning techniques predicting the association of exacerbation of peak expiratory flow ratio with short term exposure level to indoor air quality using adult asthmatics clustered data. *PLoS One*. 2021;16(1):e0244233. doi:10.1371/journal.pone.0244233

66. Mueller W, Wilkinson P, Milner J, et al. The relationship between greenspace and personal exposure to PM$_{2.5}$ during walking trips in Delhi, India. *Environ Pollut*. 2022;305:119294. doi:10.1016/j.envpol.2022.119294

67. Razavi-Termeh SV, Sadeghi-Niaraki A, Choi SM. Effects of air pollution in Spatio-temporal modeling of asthma-prone areas using a machine learning model. *Environ Res*. 2021;200:111344. doi:10.1016/j.envres.2021.111344

68. Razavi-Termeh SV, Sadeghi-Niaraki A, Choi SM. Asthma-prone areas modeling using a machine learning model. *Sci Rep*. 2021;11(1):1912. doi:10.1038/s41598-021-81147-1

69. Li YC, Hsu HL, Chun Y, et al. Machine learning-driven identification of early-life air toxic combinations associated with childhood asthma outcomes. *J Clin Invest*. 2021;131(22):e152088. doi:10.1172/JCI152088

70. D'Amato G, Annesi-Maesano I, Urrutia-Pereira M, et al. Thunderstorm allergy and asthma: state of the art. *Multidiscip Respir Med*. 2021;16(1):806. doi:10.4081/mrm.2021.806

71. Boğan M, Kul S, Al B, et al. Effect of desert dust storms and meteorological factors on respiratory diseases. *Allergy*. 2022;77(7):2243-2246. doi:10.1111/all.15298

72. Li J, Garshick E, Hart JE, et al. Estimation of ambient PM2.5 in Iraq and Kuwait from 2001 to 2018 using machine learning and remote sensing. *Environ Int*. 2021;151:106445. doi:10.1016/j.envint.2021.106445

73. Zhu H. Big data and artificial intelligence modeling for drug discovery. *Annu Rev Pharmacol Toxicol*. 2020;60:573-589. doi:10.1146/annurev-pharmtox-010919-023324

74. Bouttier F, Marchal H. Probabilistic thunderstorm forecasting by blending multiple ensembles. *Tellus*. 2020;72:1-19. doi:10.1080/16000870.2019.1696142

75. Pérez-Porras FJ, Triviño-Tarradas P, Cima-Rodríguez C, Meroño-de-Larriva JE, García-Ferrer A, Mesas-Carrascosa FJ. Machine learning methods and synthetic data generation to predict large wildfires. *Sensors (Basel)*. 2021;21(11):3694. doi:10.3390/s21113694

76. Qiu H, Luo L, Su Z, Zhou L, Wang L, Chen Y. Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. *BMC Med Inform Decis Mak*. 2020;20(1):83. doi:10.1186/s12911-020-1101-8

77. Chen J, Li H, Luo L, et al. Machine learning-based forecast of hemorrhagic stroke healthcare service demand considering air pollution. *J Healthc Eng*. 2019;2019:7463242-7463248. doi:10.1155/2019/7463242

78. Moustris KP, Douros K, Nastos PT, et al. Seven-days-ahead forecasting of childhood asthma admissions using artificial neural networks in Athens. *Greece Int J Environ Health Res*. 2012;22(2):93-104. doi:10.1080/09603123.2011.605876

79. Kim BE, Leung DYM. Significance of skin barrier dysfunction in atopic dermatitis. *Allergy, Asthma Immunol Res*. 2018;10(3):207-215. doi:10.4168/aair.2018.10.3.207

80. Nomura T, Kabashima K. Advances in atopic dermatitis in 2019-2020: Endotypes from skin barrier, ethnicity, properties of antigen, cytokine profiles, microbiome, and engagement of immune cells. *J Allergy Clin Immunol*. 2021;148(6):1451-1462. doi:10.1016/j.jaci.2021.10.022

81. Claus SP, Guillou H, Ellero-Simatos S. The gut microbiota: a major player in the toxicity of environmental pollutants? *NPJ Biofilms Microbiomes*. 2016;2:16003. doi:10.1038/npjbiofilms.2016.3 Erratum in: NPJ Biofilms Microbiomes 2017 Jun 22;3:17001.

82. Broderick C, Christian N, Apfelbacher C, et al. The BIOMarkers in atopic dermatitis and psoriasis (BIOMAP) glossary: developing a lingua franca to facilitate data harmonization and cross-cohort analyses. *Br J Dermatol*. 2021;185(5):1066-1069. doi:10.1111/bjd.20587

83. Fortino V, Wisgrill L, Werner P, et al. Machine-learning-driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc Natl Acad Sci U S A*. 2020;117(52):33474-33485. doi:10.1073/pnas.2009192117

84. Klueber J, Czolk R, Codreanu-Morel F, et al. High-dimensional immune profiles correlate with phenotypes of peanut allergy during food-allergic reactions. *Allergy*. 2023;78(4):1020-1035. doi:10.1111/all.15408. Epub ahead of print.

85. Barone SM, Paul AG, Muehling LM, et al. Unsupervised machine learning reveals key immune cell subsets in COVID-19, rhinovirus infection, and cancer therapy. *eLife*. 2021;10:e64653. doi:10.7554/eLife.64653

86. Ohanyan H, Portengen L, Huss A, et al. Machine learning approaches to characterize the obesogenic urban exposome. *Environ Int*. 2022;158:107015. doi:10.1016/j.envint.2021.107015

87. Guillien A, Cadiou S, Slama R, Siroux V. The Exposome approach to decipher the role of multiple environmental and lifestyle determinants in asthma. *Int J Environ Res Public Health*. 2021;18(3):1138. doi:10.3390/ijerph18031138

88. Proper SP, Azouz NP, Mersha TB. Achieving precision medicine in allergic disease: Progress and challenges. *Front Immunol.* 2021;12:720746. doi:10.3389/fimmu.2021.720746

89. Hu X, Walker DI, Liang Y, et al. A scalable workflow to characterize the human exposome. *Nat Commun.* 2021;12(1):5575. doi:10.1038/s41467-021-25840-9

90. Schett G, McInnes IB, Neurath MF. Reframing immune-mediated inflammatory diseases through signature cytokine hubs. *N Engl J Med.* 2021;385(7):628-639. doi:10.1056/NEJMra1909094

91. Schett G, Tanaka Y, Isaacs JD. Why remission is not enough: underlying disease mechanisms in RA that prevent cure. *Nat Rev Rheumatol.* 2021;17(3):135-144. doi:10.1038/s41584-020-00543-5

92. Riccardi D, Ward JPT, Yarova PL, et al. Topical therapy with negative allosteric modulators of the calcium-sensing receptor (calcilytics) for the management of asthma: the beginning of a new era? *Eur Respir J.* 2022;60(2):2102103. doi:10.1183/13993003.02103-2021

93. Gutierrez P, Gerardy JY. Causal inference and uplift modeling, a review of the literature. *JMLR: Workshop Conf Proc.* 2016;67:1-13.

94. Jaroszewicz S, Rzepakowski P. Uplift modeling with survival data. ACM SIGKDD Workshop on Health Informatics; 2014.

95. Peng J, Jury EC, Dönnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Front Pharmacol.* 2021;12:720694. doi:10.3389/fphar.2021.720694

96. Gille F, Jobin A, Lenca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intelligence-Based Med.* 2020;1-2:100001. doi:10.3929/ethz-b-000430039

97. 3TR Home. Retrieved November 21, 2022. https://3tr-imi.eu/.

98. Rahman MM, Davis D. Machine learning-based missing value imputation method for clinical datasets. *IAENG Transactions on Engineering Technologies. Lecture Notes in Electrical Engineering.* Vol 229. Springer; 2013:245-257. doi:10.1007/978-94-007-6190-2_19

99. Retrieved December 7, 2022. http://www.cs.put.poznan.pl/sigml/wp-content/uploads/2013/11/Uplift_Modeling_web.pdf.

100. Pineau J, Vincent-Lamarre P, Sinha K. Improving reproducibility in machine learning research. *J Mach Learn Res.* 2021;22:1-20. doi:10.48550/arXiv.2003.12206

101. Kinkorová J, Topolčan O. Biobanks in the era of big data: objectives, challenges, perspectives, and innovations for predictive, preventive, and personalised medicine. *EPMA J.* 2020;11(3):333-341. doi:10.1007/s13167-020-00213-2