



A scoping review of asthma and machine learning

Ulfat A. Khanam, Zhiwei Gao, Darryl Adamko, Anthony Kusalik, Donna C. Rennie, Donna Goodridge, Luan Chu & Joshua A. Lawson

To cite this article: Ulfat A. Khanam, Zhiwei Gao, Darryl Adamko, Anthony Kusalik, Donna C. Rennie, Donna Goodridge, Luan Chu & Joshua A. Lawson (2022): A scoping review of asthma and machine learning, Journal of Asthma, DOI: [10.1080/02770903.2022.2043364](https://doi.org/10.1080/02770903.2022.2043364)

To link to this article: <https://doi.org/10.1080/02770903.2022.2043364>



View supplementary material [↗](#)



Published online: 02 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 88



View related articles [↗](#)



View Crossmark data [↗](#)



A scoping review of asthma and machine learning

Ulfat A. Khanam, MSc, MBBS^a, Zhiwei Gao, PhD^b , Darryl Adamko, MD^c, Anthony Kusalik, PhD^d, Donna C. Rennie, PhD^e, Donna Goodridge, PhD^f, Luan Chu, PhD^g and Joshua A. Lawson, PhD^h

^aHealth Sciences Program, College of Medicine, Canadian Centre for Health and Safety in Agriculture, Respiratory Research Centre, University of Saskatchewan, Saskatoon, SK, Canada; ^bDepartment of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada; ^cDepartment of Paediatrics, College of Medicine, University of Saskatchewan, Saskatoon, SK, Canada; ^dDepartment of Computer Science, College of Arts and Science, University of Saskatchewan, Saskatoon, SK, Canada; ^eCollege of Nursing and Canadian Centre for Health and Safety in Agriculture, University of Saskatchewan, Saskatoon, SK, Canada; ^fDepartment of Medicine, College of Medicine, University of Saskatchewan, Saskatoon, SK, Canada; ^gProvincial Research Data Services, Alberta Health Service, Calgary, AB, Canada; ^hDepartment of Medicine, Canadian Centre for Health and Safety in Agriculture, and Respiratory Research Centre, College of Medicine, University of Saskatchewan, Saskatoon, SK, Canada

ABSTRACT

Objective: The objective of this study was to determine the extent of machine learning (ML) application in asthma research and to identify research gaps while mapping the existing literature.

Data Sources: We conducted a scoping review. PubMed, ProQuest, and Embase Scopus databases were searched with an end date of September 18, 2020.

Study Selection: DistillerSR was used for data management. Inclusion criteria were an asthma focus, human participants, ML techniques, and written in English. Exclusion criteria were abstract only, simulation-based, not human based, or were reviews or commentaries. Descriptive statistics were presented.

Results: A total of 6,317 potential articles were found. After removing duplicates, and reviewing the titles and abstracts, 102 articles were included for the full text analysis. Asthma episode prediction (24.5%), asthma phenotype classification (16.7%), and genetic profiling of asthma (12.7%) were the top three study topics. Cohort (52.9%), cross-sectional (20.6%), and case-control studies (11.8%) were the study designs most frequently used. Regarding the ML techniques, 34.3% of the studies used more than one technique. Neural networks, clustering, and random forests were the most common ML techniques used where they were used in 20.6%, 18.6%, and 17.6% of studies, respectively. Very few studies considered location of residence (i.e. urban or rural status).

Conclusions: The use of ML in asthma studies has been increasing with most of this focused on the three major topics (>50%). Future research using ML could focus on gaps such as a broader range of study topics and focus on its use in additional populations (e.g. location of residence).

ARTICLE HISTORY

Received 28 October 2021

Revised 8 February 2022

Accepted 12 February 2022

KEYWORDS

Asthma; epidemiology; machine learning

Introduction

Asthma is a chronic health condition affecting all age groups globally. It is a heterogeneous inflammatory disease of the lung characterized by variable symptoms such as cough, wheeze, shortness of breath, chest tightness, and variable airflow limitation (1). Worldwide, more than 262 million people suffered from asthma in 2019 with 461,000 people dying due to asthma (2). Asthma is a multifactorial disease. Many studies have been conducted to assess the risk factors, natural history, severity, underlying biology, and treatment of the disease (3–6).

Despite all these data, the goal of achieving asthma prevention remains elusive. Many environmental and genetic factors have been identified which contribute to the onset of asthma but the exact causal mechanisms are not fully understood (3–7). Part of the reason could be the complexity of asthma. It is commonly believed that many environmental factors, genes, and interactions including gene*gene and gene*environment interactions play an important role in the etiology of the diseases. Highly dimensional data is difficult to comprehend by humans, which may lead to difficulties in assessing patterns or poor decisions (8,9). To better understand this condition,

CONTACT Ulfat A. Khanam  Khanamulk292@usask.ca  Canadian Centre for Health and Safety in Agriculture, Respiratory Research Centre, University of Saskatchewan, 1231 E Wing, 104 Clinic Place, Saskatoon, SKS7N 5E5, Canada.

 Supplemental data for this article can be accessed at [publisher's website](#).

© 2022 Taylor & Francis Group, LLC

it may be beneficial to use novel technologies and techniques.

Machine learning is a relatively new method of data analysis. It is an area of artificial intelligence suited to recognizing patterns in large datasets (10). The focus of conventional statistical analysis is to explain data using human expertise to formulate and discover cause-effect relationships driven by a set of pre-determined assumptions. There could be human error in this approach. In contrast, machine learning (ML) is focused on data and hypothesis generation by building predictive models using algorithms (11). Use of machine learning allows for higher order interaction analysis (12).

Given the complexity of asthma, machine learning could be a useful approach to understand asthma. We know that machine learning techniques have been successfully applied to a variety of studies such as those investigating asthma phenotypes (13,14), wheeze exacerbation trajectories (15), asthma prediction (16,17), asthma diagnosis (18), asthma hospitalization (19,20), asthma control (21,22), and asthma severity (23). However, we do not know the extent to which ML has been used in asthma research. We also do not know which areas of asthma research would benefit most from machine learning techniques. A better understanding of what ML approaches have been used in asthma is needed if we hope to improve asthma understanding and management. Scoping reviews are an important method to help map existing literature and detect research gaps (24). The objective of our current review was to examine the nature of ML activity related to asthma research and identify gaps in the existing literature.

Methods

The methodology of this study used the framework developed by Arksey and O'Malley (24), which outlines the work into six stages: (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; (5) collating, summarizing, and reporting the results; and (6) consultation. In addition, the PRISMA checklist that included an extension for scoping reviews (25) was followed.

Our first step was to identify the research question. We followed the suggestion by Daudt et al., 2013 which states that "Scoping studies aim to map the literature on a particular topic or research area and provide an opportunity to identify key concepts; gaps in the research; and types and sources of evidence to inform practice, policymaking, and research" (26).

Following this principle, our review was guided by the following questions:

What are the most common study areas of asthma research where machine learning is used? In which groups (e.g., population, locations) are machine learning methods most commonly used in asthma research? Which machine learning methods are most commonly used in asthma research? In what ways can machine learning be used in asthma studies to aid in identifying etiology and clinical practice?

We performed a comprehensive literature search between December 2019 and February 2020 using three electronic databases: PubMed, ProQuest, and Embase Scopus databases. We searched again in September 2020 to obtain more recent publications. We considered all literature from the databases until September 2020. PubMed covers the area of medicine, nursing, health care systems, preclinical sciences, and biomedical and life sciences. ProQuest covers the area of public health information and research. Embase covers a wide range of core and niche scientific, medical, and healthcare disciplines. Keywords such as 'asthma', 'wheeze', 'machine learning', 'human', 'cluster', 'hierarchical', 'artificial intelligence', 'naive Bayes', 'k nearest neighbor', 'k means clustering', 'decision tree', 'support vector machine', 'principal component analysis', 'hidden Markov model', 'Bayesian network', 'random forest', 'gradient boosting model', 'regression', 'artificial neural network', and 'deep learning' were used to search for literature in all three databases. All citations were imported into Mendeley reference manager to create a single Research Information Systems (RIS) file. The RIS file was then imported into the Web-based systematic review software DistillerSR (Evidence Partners, Ottawa, ON) to remove duplicates and help screen titles and abstracts.

Eligible studies were included if they described any use of machine learning techniques in relation to asthma. We determined any study as machine learning if the study explicitly mentioned machine learning algorithms or any techniques mentioned above as supervised or unsupervised. Inclusion criteria were an asthma focus, human participants, and use of machine learning methods in the analysis. Studies were excluded if they were published in any language other than English, did not include humans, included only an abstract, were simulation or data mining based, or were reviews or commentaries.

After completing the screening of titles and abstracts, full text articles of all citations considered relevant for the full text review were extracted and imported to Mendeley for an in-depth review. Studies were excluded in this phase if they did not meet the

inclusion criteria. The data were compiled and imported into Microsoft Excel for validation and coding.

For each paper, the following information was recorded: location/country where the study was conducted, year of publication, publication type (journal/conference), study topic, study design, machine learning methods used in the study, sample size, study population (children/adolescents/adults), and location of residence (urban/rural).

Operational definitions

We grouped the studies by their clinical research focus. Groupings are described below.

Asthma phenotype

Asthma phenotypes are the identifiable clusters of demographic, clinical and/or pathophysiological characteristics (7,27,28). Asthma phenotypes include clinical, molecular, and atopy phenotype.

Asthma episode prediction

Asthma episodes are the acute conditions of worsening of symptoms. It could be the initial presentation of asthma in some cases (1). For our purposes, this category includes wheeze, shortness of breath, chest tightness and cough, variable expiratory airflow limitation, prediction of exacerbation, hospitalization, and emergency department visits.

Asthma diagnosis prediction

Diagnosis of asthma is based on identifying a characteristic pattern of respiratory symptoms (i.e. wheezing, shortness of breath, chest tightness or cough) and variable expiratory airflow limitation (29). This category includes studies describing the diagnostic value of spirometry and impulse oscillometry, diagnostic support for pulmonary diseases, and assessment of key factors in asthma diagnosis.

Allergic sensitization

Allergic sensitization is a process by which the body's immune system develops an abnormal reaction (i.e. producing an antibody) to an allergen (i.e. certain foods, pollen, mold, or medications) (30). Depending on the individual, allergic sensitization response can range from minor to potentially life-threatening. This category includes studies focused on allergic sensitization, allergen, atopic endotype, and determinants of allergic rhinitis among children with asthma.

Cough sound

Cough sounds are coughing sounds from patients with and without pathological respiratory conditions (31). This category includes studies looking at the development and analysis of algorithms analyzing coughing sounds, respiratory sounds, and breath sounds.

Genetic profile of asthma

This category includes studies related to genetic profiles of asthma patients. These include genetic profiles related to phenotype, genetic factors, gene-environment interaction, genome-by-environment, fetal lung transcriptomes, and vitamin-D related genes.

Metabolomics

This category includes studies related to asthma and -omics, specifically metabolites. These include urine metabolomics, metabolomic prediction of asthma, exhaled breath condensate metabolome clusters, and metabolomics of asthma control.

Miscellaneous

We categorized a study as miscellaneous if any one of the following topics were the focus of a study: asthma and absenteeism, mixed longitudinal outcomes related to the asthma state, neonatal lung function, developmental profiles, text-based asthma status based on mobile phone technology, plasma biomarkers, data-driven order set generation, airway obstruction, exhaled biomarkers for asthma, bronchodilator response, exhaled air temperature, infant wheeze, and computer-based assessment of asthma strategies. There was only one study conducted for each of these topics.

Machine learning techniques used

We also grouped the studies by their machine learning approach. The types of machine learning we looked at were clustering, Bayesian network, random forest, neural network, regression, support vector machine, decision tree, hidden Markov model, and gradient boosting model. Specific details of each technique are listed in [Appendix B](#). In addition, we classified the study as using one vs. multiple machine learning types.

Analysis

We completed a number of descriptive analyses. Initially, we looked at frequencies of studies per year (1993–2020), study designs, study topics, and machine learning techniques. Subsequently, we looked at

various cross-tabulations between these variables. Simple descriptive statistics such as frequencies and graphs were created in Excel and SPSS.

Results

The study inclusion flow chart is located in [Figure 1](#). A total of 6,317 articles were found and imported into the citation managing software DistillerSR. After removing 908 duplicates, 5,409 articles were included for title and abstract review. From this, 280 articles were considered for full text review. Articles related to simulation, non-human study, and abstracts only were excluded ($n=178$) from the study. Finally, 102 articles were included for the full text review. The data charting of the articles including location/country, year of publication, study topic, design, M/L method, sample size, study population, and location of residence are presented in the [Supplemental Table 1](#).

Based on our review, we found there was no asthma related study using machine learning before 1993. In 1993, artificial neural networks were first used to assess the relationship between pulmonary function of children with asthma and air pollution (32). However, since 2000, the use of machine learning has increased over time in

asthma related studies, reaching its peak in 2017–2020 ([Figure 2](#)). Forty-one percent of studies used a small sample size of fewer than 300 people.

There was wide variability in the study topics investigated. The three topics most commonly considered were asthma episode prediction, asthma phenotyping, and genetic profile of asthma ([Figure 3](#)). The most frequently used study designs were cohort ($n=54$), cross-sectional ($n=21$), and case-control ($n=12$) ([Figure 4](#)). Machine learning was used in randomized controlled trial/experimental, time series, case study, aggregate panel, prospective panel, or retrospective chart studies very rarely ([Figure 4](#)).

Most of the studies related to asthma episode prediction used a cohort design while a smaller portion used a cross-sectional design ([Table 1](#)). The majority of studies related to asthma phenotyping used a cohort and cross-sectional design, while fewer studies used a case-control or randomized controlled trial design ([Table 1](#)). Finally, most of the studies related to genetic profiles of asthma used a case-control or randomized controlled trial design, while only a few used a cohort or cross-sectional study design ([Table 1](#)).

With regard to the type of machine learning techniques used, most of the studies used a single machine

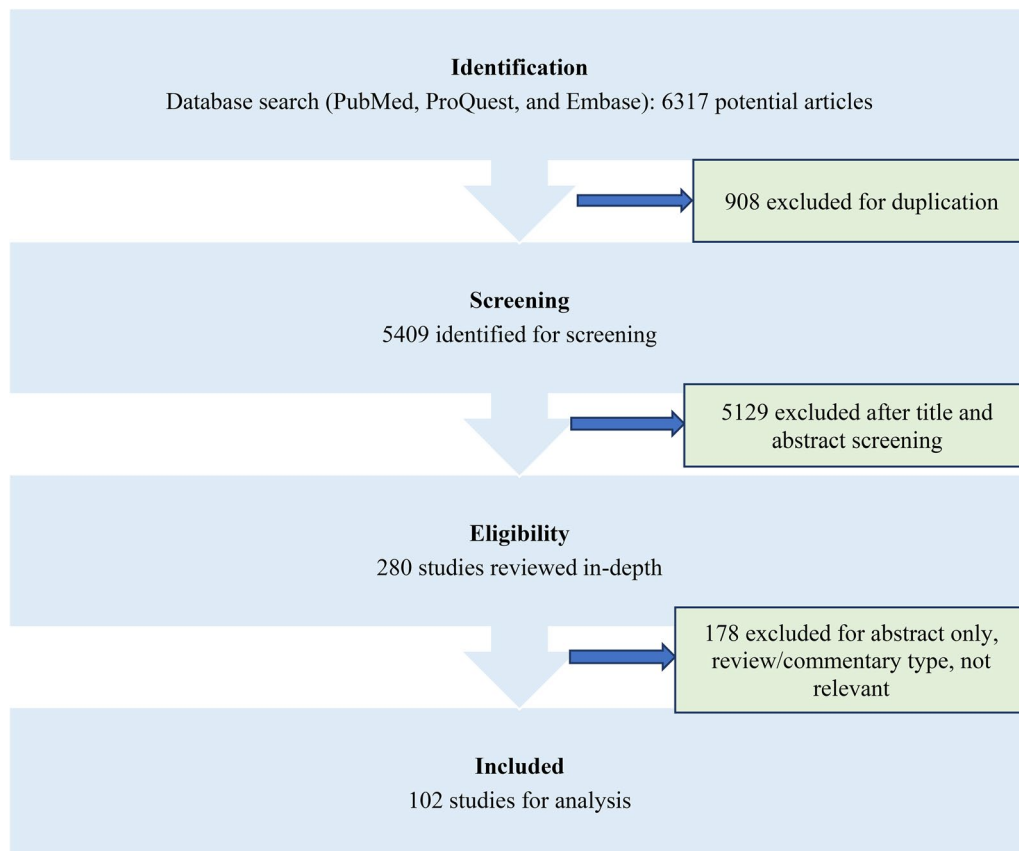
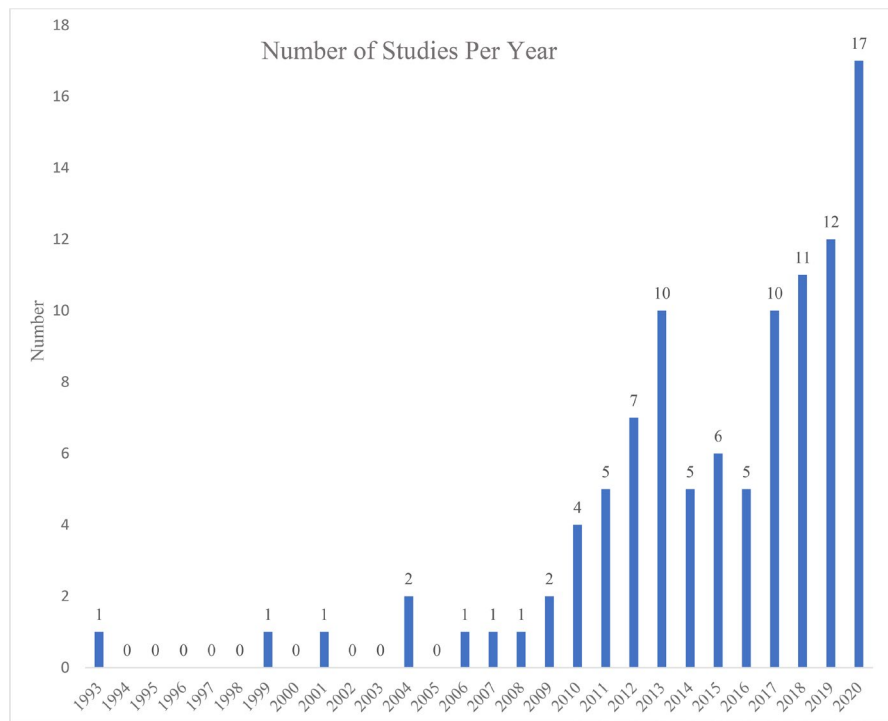


Figure 1. Flow chart of the study selection procedure.

Table 1. Proportion of studies using a specific study design by study topic.

Topic	Sample size (n)	Study design (row %)									
		Cohort	Cross-sectional	Case-control	Randomized controlled trial	Time series	Retrospective chart study	Aggregate panel study	Prospective panel	Case study	Unknown
Asthma phenotype	(n=17)	47.1	41.2	5.9	5.9	–	–	–	–	–	–
Asthma prediction											
Asthma episode prediction	(n=25)	80.0	12.0	–	4.0	–	4.0	–	–	–	–
Asthma diagnosis prediction	(n=7)	42.9	57.1	–	–	–	–	–	–	–	–
Genetic profiles	(n=13)	15.4	15.4	30.8	23.1	–	–	–	7.7	7.7	–
Allergic sensitization	(n=9)	66.7	11.1	22.2	–	–	–	–	–	–	–
Cough sound	(n=6)	50.0	33.3	16.7	–	–	–	–	–	–	–
Air pollution	(n=5)	20.0	–	20.0	–	40.0	–	20.0	–	–	–
Metabolomics	(n=4)	25.0	–	25.0	50.0	–	–	–	–	–	–
Asthma control	(n=3)	100.0	–	–	–	–	–	–	–	–	–
Miscellaneous	(n=13)	53.8	15.4	15.4	7.7	–	–	–	–	–	7.7

**Figure 2.** Number of studies published on asthma and machine learning by year.

learning method ($n=67$). The top three machine learning techniques were neural networks, clustering, and random forests (Figure 5). Among the studies looking at asthma episode prediction, the top three machine learning techniques were decision tree, naïve Bayes, and support vector machine. Among those studies looking at asthma phenotype, the top three machine learning techniques were clustering, hidden Markov model, and boosting. Among the studies looking at genetic profile of asthma, the top three machine learning techniques used were principal component analysis, Bayesian network, and nearest neighbor (Table 2).

Overall, 67% of studies included only children and adolescents (age ≤ 18 years) while 21% included only adults. The remaining studies included both children and adults. Consistently for studies related to asthma episodes, phenotype, and genetic profile, children and adolescents were most often included (Table 3).

The majority of machine learning and asthma studies were performed in the USA and Europe with only two studies conducted in Canada, both by Farion et al. (33,34). Very few studies mentioned location of dwelling (urban or rural status) while describing or

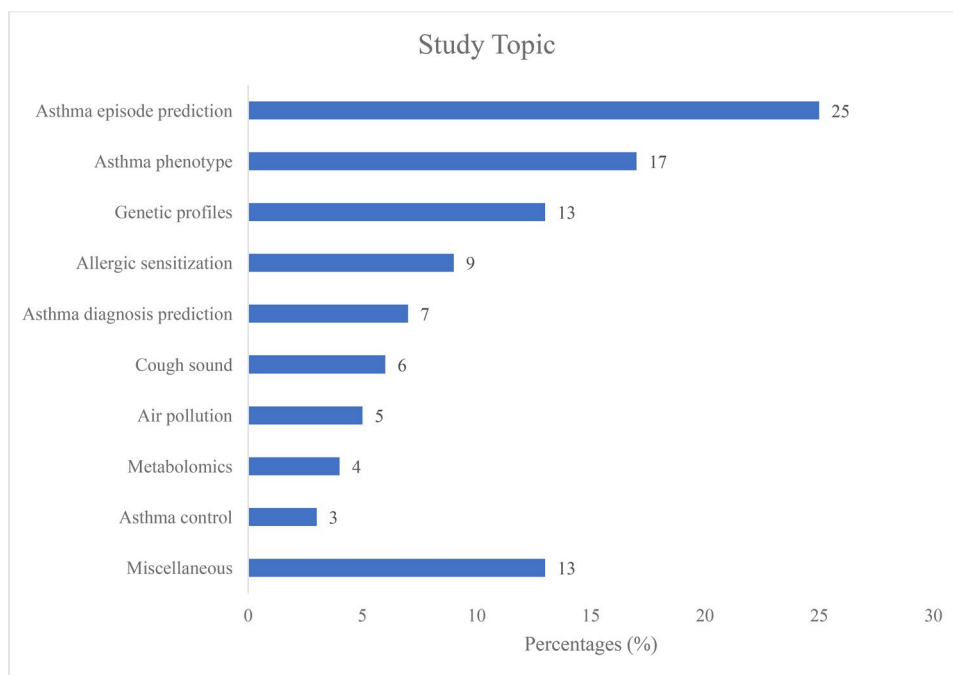


Figure 3. Percentages of published studies focused on asthma that used machine learning techniques by study topic (N=102).

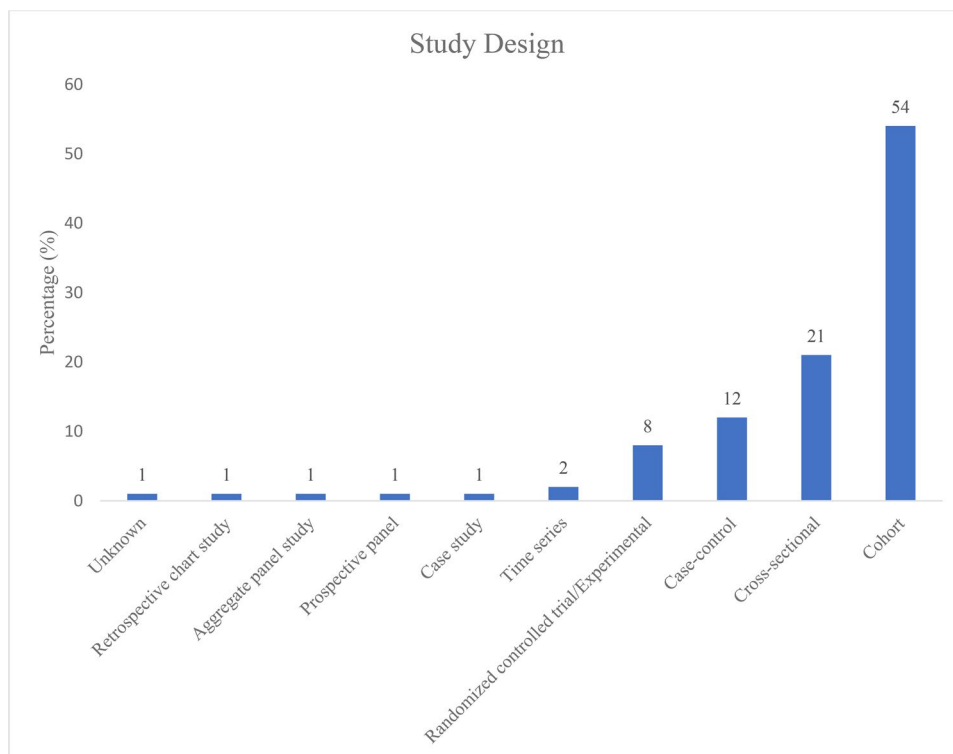


Figure 4. Percentages of published studies focused on asthma that used machine learning techniques by study design (N=102).

analyzing the study population. Only three studies (14,35,36) mentioned including both urban and rural populations in the analysis. The research topics for these three studies were air pollution, asthma episode prediction, and genetics. Support vector machine, k-means clustering, regression, and principal

component analysis were used in these urban-rural studies. Seven studies (19,32,37–41) described location of residence and considered an urban only population. The research topics from these seven studies were asthma phenotype, asthma episode prediction, genetic profile of asthma, allergic sensitization, and neonatal

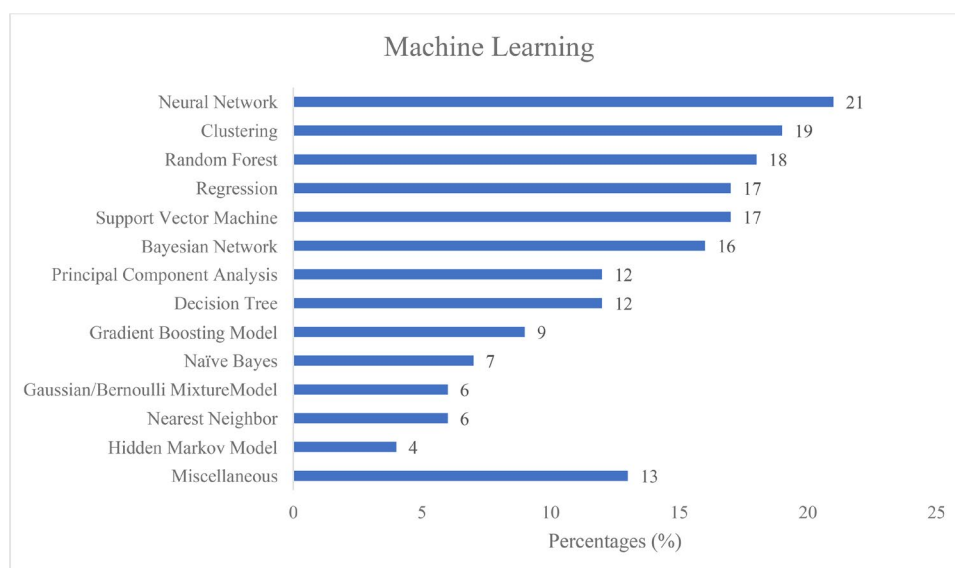


Figure 5. Percentage of studies using specific machine learning techniques used in published studies focused on asthma. Note: The total number of studies does not add up to 102 as many studies included multiple machine learning techniques.

Table 2. Study topic vs. machine learning technique used.

Topic	Sample size (n)	Machine learning technique (row %)													
		Cluster	BN	Neural network	Regression	PCA	RF	SVM	DT	Boosting	GBMM	NB	NN	HMM	Misc
Asthma phenotype	(n = 17)	58.8	11.8	5.9	23.5	17.6	11.8	5.9	11.8	17.6	5.9	5.9	–	11.8	17.6
Asthma prediction															
Asthma episode prediction	(n = 25)	12.0	20.0	24.0	12.0	8.0	16.0	28.0	24.0	12.0	–	12.0	8.0	–	4.0
Asthma diagnosis prediction	(n = 7)	14.3	–	28.6	14.3	–	28.6	28.6	28.6	28.6	–	14.3	14.3	–	28.6
Genetic profiles	(n = 13)	23.1	23.1	7.7	15.4	23.1	15.4	7.7	7.7	–	–	7.7	7.7	–	15.4
Allergic sensitization	(n = 9)	11.1	11.1	11.1	22.2	11.1	22.2	–	11.1	–	33.3	–	–	11.1	22.2
Cough sound	(n = 6)	–	–	33.3	33.3	–	–	16.7	–	–	33.3	–	–	16.7	–
Air pollution	(n = 5)	–	40.0	40.0	–	–	–	20.0	–	–	–	–	–	–	–
Metabolomics	(n = 4)	–	50.0	–	25.0	–	50.0	–	–	–	–	–	–	–	–
Asthma control	(n = 3)	–	–	66.7	–	33.3	66.7	33.3	–	33.3	–	33.3	33.3	–	–
Miscellaneous	(n = 13)	7.7	7.7	30.8	15.4	15.4	15.4	23.1	–	–	–	–	7.7	–	23.1

Cluster = Clustering; BN = Bayesian networks; PCA = Principal component analysis; RF = Random forest; SVM = Support vector machine; DT = Decision tree; GBMM = Gaussian/Bernoulli mixture model; NB = Naïve Bayes; NN = Nearest neighbor; HMM = Hidden Markov model; Misc = Miscellaneous.

Note: The total row% does not add up to 100% as some studies included multiple machine learning techniques.

lung function. Random forests, decision trees, regression, principal component analysis, Bayesian networks, and k-means clustering were used in these urban only studies. There were no rural-only studies mentioned.

Discussion

Our scoping review presents an overview of the studies which used machine learning on asthma-related data. We followed the methodology of the five-step scoping review protocol (25) and PRISMA. Our scoping review provides a clear map on the use of machine learning in the field of asthma. Among the studies included in this review, asthma episode prediction, asthma phenotype, and genetic profiles were identified

as the most widely studied topic in studies of asthma where machine learning was used. A reason behind the use of machine learning around these topics is that it can reduce some forms of human bias, although human error is still possible through mechanisms such as the use of inappropriate techniques, coding errors, or selection and information biases in data collection. ML uses specific algorithms to analyze data in an unbiased way (e.g. clustering, decision tree) although the data itself may be biased due to selection and classification issues.

Among the asthma episode prediction, asthma phenotype, and genetic profile topics, most of the studies used cohort, cross-sectional, and case-control study designs. A cohort design has the advantage of aiding

Table 3. Characteristics of the study populations by study topic.

Topic	Sample size (n)	Study population (row %)					
		Children	Children/ adolescent	Children/ adolescent/ adult	Adult	Fetus	Unknown
Asthma phenotype	(n = 17)	29.4	29.4	5.9	23.5	–	11.8
Asthma prediction							
Asthma episode prediction	(n = 25)	8.0	56.0	8.0	28.0	–	–
Asthma diagnosis prediction	(n = 7)	14.3	14.3	–	57.1	–	14.3
Genetic profiles	(n = 13)	15.4	46.2	–	7.7	7.7	23.1
Allergic sensitization	(n = 9)	55.6	11.1	–	22.2	–	11.1
Cough sound	(n = 6)	50.0	33.3	–	16.7	–	–
Air pollution	(n = 5)	–	40.0	20.0	20.0	–	20.0
Metabolomics	(n = 4)	50.0	50.0	–	–	–	–
Asthma control	(n = 3)	–	100.0	–	–	–	–
Miscellaneous	(n = 13)	46.2	38.5	7.7	7.7	–	–

in the assessment of causality between exposure and outcome of the disease due to its ability to consider temporality. Cohort study designs are the preferred observational design although it is also important in epidemiology to use other study designs to help confirm study results. The use of case-control and randomized controlled trial designs were less frequently observed in asthma phenotype and asthma episode prediction topics while, cohort and case-control designs in genetic profiles were less frequently used. The randomized controlled trial study design could be used in testing an intervention (42). However, in studies looking at etiology, for example, with an outcome of asthma prediction or asthma phenotype, randomized controlled trial study designs are usually not possible and so observational study designs are used. For studies of asthma episode prediction, none of the studies we found used a case-control design.

Regarding machine learning techniques, most of the studies concerning asthma phenotype used a clustering method. The reason for using a clustering approach would be to identify or classify phenotypes within the data based on features or attributes. Classification of asthma based on phenotype is important because accurate categorization may be useful for treating patients appropriately as well as for etiological understanding. Clustering may not be an appropriate technique where classification or categorization is not required, for example: predicting presence of asthma, asthma episode or diagnosis. Therefore, very few studies related to episode prediction used a clustering algorithm. Clustering could be used in genetic profiles and allergic sensitization where categorization is required; however, the use of clustering in genetic profiles and allergic sensitization was minimal. For the asthma episode prediction topic, most of the studies used a decision tree or support vector machine algorithm. Decision tree is an appropriate method for making predictions and accurate

classification in asthma diagnosis and medical decision making (34). All other study topics used a variety of machine learning algorithms.

It is difficult to suggest best/potential M/L methods by study topic. Because the application of machine learning depends on not only the study topic, but also on the nature of the specific research question and how big the sample size. It may be most appropriate to employ multiple methods to determine the best fit. For example, if the intent is to determine phenotypes and with an accompanying research question to identify any asthma phenotypes based on the data, an unsupervised technique (e.g. clustering) could apply. However, if the research question is to find predictors of an established phenotype (e.g. atopic asthma) then a variety of supervised machine learning techniques might be applicable (e.g. naïve Bayes, decision tree, SVM, Bayesian networks). Given the sample size, decision tree generally allows us to detect a higher order interaction. Each machine learning method has specific advantages and disadvantages, the use of multiple techniques would allow for the most appropriate selection. Therefore, an “optimal” approach based on the performance of the several M/L techniques to be used consecutively instead of reporting a “best” approach. Based on our scoping review, we found that 34% of the studies used more than one M/L technique.

The majority of studies included children and adolescents. Asthma is more common in children (43) which could be a reason for inclusion of children and adolescents as study participants in most studies. However, asthma among adults could be different compared to children in various ways such as asthma presentation, potential phenotypes, determinants or risk factors (44). Therefore, application of machine learning in asthma studies including adults as study participants could be helpful for understanding the disease.

A significant gap was identified when considering location of residence, such as urban and rural. Among the included studies, most studies did not specify location of residence. Location of residence may have a significant impact on the results as there is evidence of urban-rural variation in asthma prevalence and related outcomes (45–49). The prevalence of asthma is lower among the rural population compared to the urban population (45,46,50–52). Clustering could be used in phenotype analysis considering an urban-rural gradient. However, a prediction modeling (e.g. via decision tree) would be appropriate and has not been used in the prediction of an urban-rural gradient of asthma. New technologies will help us to identify why there is an urban-rural variation.

Recently, two additional reviews looking at machine learning and asthma were published (53,54). These came after the search and collection of articles for our review. Our study differs from these previous reviews and adds to the knowledge around the topic (53,54). The previous reviews focused on artificial intelligence (53,54), which may include machine learning; however, there has not been a single scoping review specifically on machine learning and asthma. The objective of the first prior review was to review asthma screening and diagnosis, patient classification, asthma management and treatment for identifying trends and potential gaps (53) while the objective of the second prior review was to review fundamental analytical approaches conducted in literature and critical analysis of these approach and their usefulness in medicine (54). In contrast to the previous studies, we followed a specific framework for scoping reviews developed by Arksey and O'Malley (24) and PRISMA (25). The first study (53) used a systematic review focused on asthma while the second study (54) used a bibliometric analysis (quantitative statistical analysis to analyze published studies in a given research area) focused on big data analytics of asthma. Regarding methodology, the first prior review (53) included articles from two sources (DBLP and PubMed), the second review (54) included studies from the Scopus database, whereas our source of studies was from PubMed, ProQuest, and Embase. After selecting studies, researchers of the first review (53) categorized studies into four broad types including asthma screening and diagnosis, patient classification, asthma management and monitoring, and asthma treatment. The researchers of the second review (54) categorized articles into three categories based on machine learning technique including multivariate statistics, machine learning/artificial intelligence, and Bayesian statistics. However, we divided

our studies into a larger range of topics (e.g. asthma phenotypes, asthma episode prediction) and considered analysis type. This included a descriptive cross-tabulation looking at study type by analysis type which has not been conducted by previous studies. As a scoping review, we also describe the study designs and populations used in the included studies more thoroughly as well as their cross-tabulations by study type and machine learning type. We further described the study populations included with regard to age (e.g. children, adults) and location including urban-rural status.

There are strengths and limitations of this scoping review. A strength is that it is based on the approach of knowledge synthesis which is effective for mapping gaps in the existing literature. However, scoping reviews lack quality assessment, which could be a problem for precise interpretation from the articles although it is not a requirement for conducting a scoping review. The conclusions are based on the existence of the selected literature, not on their fundamental qualities. Issues could have occurred from reporting or publication biases as well. We included literature from electronic databases and key journals. However, we did not include any unpublished literature and so our results may have a potential publication bias. We carefully searched and selected the studies for inclusion to overcome 'researcher bias' through the process. We included articles in the English language only which might lead to potential 'language bias' (24).

Conclusions

Different types of machine learning techniques were used while investigating asthma. This is appropriate given the variety of study topics. Our study found several gaps in the application of machine learning to studies of asthma such as minimal use of some study designs, limited involvement of some study populations (e.g. adult participants), and limited research by location of residence (i.e. Canada, urban-rural). Future work should focus on addressing these gaps.

Acknowledgements

Health Sciences Program, College of Medicine, University of Saskatchewan, Canadian Center for Health and Safety in Agriculture (CCHSA).

Declaration of interest

The authors report no conflicts of interest.

Financial

This research has been supported by personal scholarships from the University of Saskatchewan University Graduate Scholarship (UGS), Canadian Centre for Health and Safety in Agriculture Founding Chairs Fellowships, and Entrance Scholarship from the University of Saskatchewan.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Zhiwei Gao  <http://orcid.org/0000-0002-5313-4884>

References

- Global Initiative for Asthma. Global strategy for asthma management and prevention updated 2020 [Internet]. 2020. Available from: www.ginasthma.org.
- Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study. 2019. *Lancet*. 2020;396(10258):1204–1222. Available from: <https://www.thelancet.com/gbd/summaries>
- Beasley R, Semprini A, Mitchell EA. Risk factors for asthma: is prevention possible? *Lancet*. 2015;386(9998):1075–1085. doi:10.1016/S0140-6736(15)00156-7.
- Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *Lancet Respir Med*. 2019;7(6):509–522. doi:10.1016/S2213-2600(19)30055-4.
- D'Amato G, Liccardi G, D'Amato M, Holgate S. Environmental risk factors and allergic bronchial asthma. *Clin Exp Allergy*. 2005;35(9):1113–1124. doi:10.1111/j.1365-2222.2005.02328.x.
- Beasley R, Crane J, Lai CKW, Pearce N. Prevalence and etiology of asthma. *J Allergy Clin Immunol*. 2000;105(2 II):466–472.
- Bel EH. Clinical phenotypes of asthma. *Curr Opin Pulm Med*. 2004;10(1):44–50.
- Johnstone IM, Titterton DM. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4237–4253. doi:10.1098/rsta.2009.0159.
- Clarke R, Ransom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*. 2008;8(1):37–49. doi:10.1038/nrc2294.
- Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260. doi:10.1126/science.aaa8415.
- Chowriappa P, Dua S, Todorov Y. Introduction to machine learning in healthcare informatics. Springer; 2014. p. 1–23.
- Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365.
- Ilmarinen P, Tuomisto LE, Niemelä O, Tommola M, Haanpää J, Kankaanranta H. Cluster analysis on longitudinal data of patients with adult-onset asthma. *J Allergy Clin Immunol Pract*. 2017;5(4):967–978.e3. doi:10.1016/j.jaip.2017.01.027.
- Fernández D, Sram RJ, Dostal M, Pastorkova A, Gmuender H, Choi H. Modeling unobserved heterogeneity in susceptibility to ambient benzo[a]pyrene concentration among children with allergic asthma using an unsupervised learning algorithm. *Int J Environ Res Public Health*. 2018;15(1):106. doi:10.3390/ijerph15010106.
- Deliu M, Fontanella S, Haider S, Sperrin M, Geifman N, Murray C, Simpson A, Custovic A. Longitudinal trajectories of severe wheeze exacerbations from infancy to school age and their association with early-life risk factors and late asthma outcomes. *Clin Exp Allergy*. 2020;50(3):315–324. doi:10.1111/cea.13553.
- Messinger AI, Bui N, Wagner BD, Szefer SJ, Vu T, Deterding RR. Novel pediatric-automated respiratory score using physiologic data and machine learning in asthma. *Pediatr Pulmonol*. 2019;54(8):1149–1155. doi:10.1002/ppul.24342.
- Spyroglou II, Spöck G, Rigas AG, Paraskakis EN. Evaluation of Bayesian classifiers in asthma exacerbation prediction after medication discontinuation. *BMC Res Notes*. 2018;11(1):522. doi:10.1186/s13104-018-3621-1.
- Mozaffarinya M, Reza Shahriyari A, Karim Bahadori M, Ghazvini A, Shamsadin Athari S, Vahedi G. A data-mining algorithm to assess key factors in asthma diagnosis. *Rev Fr Allergol*. 2019;59(7):487–492. doi:10.1016/j.reval.2019.01.013.
- Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. *Acad Emerg Med*. 2018;25(12):1463–1470. doi:10.1111/acem.13655.
- Das LT, Abramson EL, Stone AE, Kondrich JE, Kern LM, Grinspan ZM. Predicting frequent emergency department visits among children with asthma using EHR data. *Pediatr Pulmonol*. 2017;52(7):880–890. doi:10.1002/ppul.23735.
- Luo G, Stone BL, Fassl B, Maloney CG, Gesteland PH, Yerram SR, Nkoy FL. Predicting asthma control deterioration in children. *BMC Med Inform Decis Mak*. 2015;15(1):84. doi:10.1186/s12911-015-0208-9.
- Pifferi M, Bush A, Pioggia G, Di Cicco M, Chinellato I, Bodini A, Macchia P, Boner AL. Monitoring asthma control in children with allergies by soft computing of lung function and exhaled nitric oxide. *Chest*. 2011;139(2):319–327. doi:10.1378/chest.10-0992.
- Swarnkar V, Abeyratne U, Tan J, Ng TW, Brisbane JM, Choveaux J, Porter P. Stratifying asthma severity in children using cough sound analytic technology. *J Asthma*. 2021;58(2):160–169. doi:10.1080/02770903.2019.1684516.
- Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol Theory*

- Pract. 2005;8(1):19–32. doi:[10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616).
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467–473. doi:[10.7326/M18-0850](https://doi.org/10.7326/M18-0850).
 26. Daudt HML, Van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Med Res Methodol*. 2013;13(1):48–49. doi:[10.1186/1471-2288-13-48](https://doi.org/10.1186/1471-2288-13-48).
 27. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med*. 2012;18(5):716–725. doi:[10.1038/nm.2678](https://doi.org/10.1038/nm.2678).
 28. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R, Castro M, Curran-Everett D, Fitzpatrick AM, National Heart, Lung, and Blood Institute's Severe Asthma Research Program, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med*. 2010;181(4):315–323. doi:[10.1164/rccm.200906-0896OC](https://doi.org/10.1164/rccm.200906-0896OC).
 29. Levy ML, Quanjer PH, Booker R, Cooper BG, Holmes S, Small IR, General Practice Airways Group. Diagnostic spirometry in primary care: proposed standards for general practice compliant with American Thoracic Society and European Respiratory Society recommendations: a General Practice Airways Group (GPIAG)1 document, in association with the Association for Respiratory Technology & Physiology (ARTP)2 and Education for Health3 1 www.gpiag.org 2 www.artp.org 3 www.educationforhealth.org.uk. *Prim Care Respir J*. 2009;18(3):130–147. doi:[10.4104/pcrj.2009.00054](https://doi.org/10.4104/pcrj.2009.00054).
 30. Van Ree R, Hummelshøj L, Plantinga M, Poulsen LK, Swindle E. Allergic sensitization: host-immune factors. *Clin Transl Allergy*. 2014;4(1):1–9.
 31. Hee HI, Balamurali BT, Karunakaran A, Herremans D, Teoh OH, Lee KP, et al. Development of machine learning for asthmatic and healthy voluntary cough sounds: a proof of concept study. *Appl Sci*. 2019;9(14):2833.
 32. Moseholm L, Taudorf E, Frøsig A. Pulmonary function changes in asthmatics associated with low-level SO₂ and NO₂ air pollution, weather, and medicine intake. An 8-month prospective study analyzed by neural networks. *Allergy*. 1993;48(5):334–344. doi:[10.1111/j.1398-9995.1993.tb02402.x](https://doi.org/10.1111/j.1398-9995.1993.tb02402.x).
 33. Farion KJ, Wilk S, Michalowski W, O'Sullivan D, Sayyad-Shirabad J. Comparing predictions made by a prediction model, clinical score, and physicians: pediatric asthma exacerbations in the emergency department. *Appl Clin Inform*. 2013;4(3):376–391. doi:[10.4338/ACI-2013-04-RA-0029](https://doi.org/10.4338/ACI-2013-04-RA-0029).
 34. Farion K, Michalowski W, Wilk S, O'Sullivan D, Matwin S. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. *J Med Syst*. 2010;34(4):551–562. doi:[10.1007/s10916-009-9268-7](https://doi.org/10.1007/s10916-009-9268-7).
 35. Choi H, Song W-M, Wang M, Sram RJ, Zhang B. Benzo[a]pyrene is associated with dysregulated myelo-lymphoid hematopoiesis in asthmatic children. *Environ Int*. 2019;128:218–232. doi:[10.1016/j.envint.2019.04.052](https://doi.org/10.1016/j.envint.2019.04.052).
 36. Chatzimichail E, Paraskakis E, Sitzimi M, Rigas A. An intelligent system approach for asthma prediction in symptomatic preschool children. *Comput Math Methods Med*. 2013;2013:240182–240186. doi:[10.1155/2013/240182](https://doi.org/10.1155/2013/240182).
 37. Lary M-A, Allsopp L, Lary DJ, Sterling DA. Using machine learning to examine the relationship between asthma and absenteeism. *Environ Monit Assess*. 2019;191(Suppl 2):332. doi:[10.1007/s10661-019-7423-2](https://doi.org/10.1007/s10661-019-7423-2).
 38. Schoos AMM, Chawes BL, Rasmussen MA, Bloch J, Bønnelykke K, Bisgaard H. Atopic endotype in childhood. *J Allergy Clin Immunol*. 2016;137(3):844–851.e4. doi:[10.1016/j.jaci.2015.10.004](https://doi.org/10.1016/j.jaci.2015.10.004).
 39. Chawes BLK, Stokholm J, Bønnelykke K, Brix S, Bisgaard H. Neonates with reduced neonatal lung function have systemic low-grade inflammation. *J Allergy Clin Immunol*. 2015;135(6):1450–1456.e1. doi:[10.1016/j.jaci.2014.11.020](https://doi.org/10.1016/j.jaci.2014.11.020).
 40. Sharma S, Poon A, Himes BE, Lasky-Su J, Sordillo JE, Belanger K, Milton DK, Bracken MB, Triche EW, Leaderer BP, et al. Association of variants in innate immune genes with asthma and eczema. *Pediatr Allergy Immunol*. 2012;23(4):315–323. doi:[10.1111/j.1399-3038.2011.01243.x](https://doi.org/10.1111/j.1399-3038.2011.01243.x).
 41. Benton AS, Wang Z, Lerner J, Foerster M, Teach SJ, Freishtat RJ. Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort. *J Asthma*. 2010;47(7):728–734. doi:[10.3109/02770903.2010.491142](https://doi.org/10.3109/02770903.2010.491142).
 42. Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am J Prev Med*. 2007;33(2):155–161. doi:[10.1016/j.amepre.2007.04.007](https://doi.org/10.1016/j.amepre.2007.04.007).
 43. CDC.gov. CDC - Asthma [Internet]. 2019. Available from: <https://www.cdc.gov/asthma/default.htm>.
 44. Trivedi M, Denton E. Asthma in children and adults—what are the differences and what can they tell us about asthma? *Front Pediatr*. 2019;7:256.
 45. Lawson JA, Rennie DC, Cockcroft DW, Dyck R, Afanasieva A, Oluwole O, Afsana J. Childhood asthma, asthma severity indicators, and related conditions along an urban-rural gradient: a cross-sectional study. *BMC Pulm Med*. 2017;17(1):4. doi:[10.1186/s12890-016-0355-5](https://doi.org/10.1186/s12890-016-0355-5).
 46. Lawson JA, Janssen I, Bruner MW, Madani K, Pickett W. Urban-rural differences in asthma prevalence among young people in Canada: the roles of health behaviors and obesity. *Ann Allergy Asthma Immunol*. 2011;107(3):220–228. doi:[10.1016/j.anai.2011.06.014](https://doi.org/10.1016/j.anai.2011.06.014).
 47. Rodriguez A, Vaca MG, Chico ME, Rodrigues LC, Barreto ML, Cooper PJ. Lifestyle domains as determinants of wheeze prevalence in urban and rural school-children in Ecuador: cross sectional analysis. *Environ Heal A Glob Access Sci Source*. 2015;14:15.
 48. Xu F, Yan S, Li F, Cai M, Chai W, Wu M, Fu C, Zhao Z, Kan H, Kang K, et al. Prevalence of childhood atopic dermatitis: an urban and rural community-based study in Shanghai, China. *PLoS One*. 2012;7(5):e36174. doi:[10.1371/journal.pone.0036174](https://doi.org/10.1371/journal.pone.0036174).
 49. Lawson JA, Rennie DC, Dosman JA, Cammer AL, Senthilselvan A. Obesity, diet, and activity in relation

- to asthma and wheeze among rural dwelling children and adolescents. *J Obes.* 2013;2013:315096–315099. doi:[10.1155/2013/315096](https://doi.org/10.1155/2013/315096).
50. Chrischilles E, Ahrens R, Kuehl A, Kelly K, Thorne P, Burmeister L, Merchant J. Asthma prevalence and morbidity among rural Iowa schoolchildren. *J Allergy Clin Immunol.* 2004;113(1):66–71. doi:[10.1016/j.jaci.2003.09.037](https://doi.org/10.1016/j.jaci.2003.09.037).
 51. Solis Soto MT, Patiño A, Nowak D, Radon K. Prevalence of asthma, rhinitis and eczema symptoms in rural and urban school-aged children from Oropeza Province - Bolivia: a cross-sectional study. *BMC Pulm Med.* 2014;14(1):16–21. doi:[10.1186/1471-2466-14-40](https://doi.org/10.1186/1471-2466-14-40).
 52. Chan-Yeung M, Zhan LX, Tu DH, Li B, He GX, Kauppinen R, Nieminen M, Enarson DA. The prevalence of asthma and asthma-like symptoms among adults in rural Beijing, China. *Eur Respir J.* 2002;19(5):853–858. doi:[10.1183/09031936.02.00250602](https://doi.org/10.1183/09031936.02.00250602).
 53. Exarchos KP, Beltsiou M, Votti C-A, Kostikas K. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. *Eur Respir J.* 2020;56(3):2000521. doi:[10.1183/13993003.00521-2020](https://doi.org/10.1183/13993003.00521-2020).
 54. Fontanella S, Cucco A, Custovic A. Machine learning in asthma research: moving toward a more integrated approach. *Expert Rev Respir Med.* 2021;15(5):609–621. doi:[10.1080/17476348.2021.1894133](https://doi.org/10.1080/17476348.2021.1894133).

Appendix A

Search protocol (September 18, 2020)

PubMed

((((((((((((((((((“machine learning”[MeSH Terms] OR (“machine”[All Fields] AND “learning”[All Fields]) OR “machine learning”[All Fields]) OR (“Proc IEEE Int Conf Clust Comput”[Journal] OR “cluster”[All Fields]) OR hierarchical[All Fields]) OR (“boidae”[MeSH Terms] OR “boidae”[All Fields] OR “python”[All Fields]) OR (“artificial intelligence”[MeSH Terms] OR (“artificial”[All Fields] AND “intelligence”[All Fields]) OR “artificial intelligence”[All Fields]) OR (naive[All Fields] AND bayes[All Fields]) OR (k[All Fields] AND nearest[All Fields] AND neighbor[All Fields]) OR (k[All Fields] AND means[All Fields] AND (“cluster analysis”[MeSH Terms] OR (“cluster”[All Fields] AND “analysis”[All Fields]) OR “cluster analysis”[All Fields] OR “clustering”[All Fields]) OR (“decision trees”[MeSH Terms] OR (“decision”[All Fields] AND “trees”[All Fields]) OR “decision trees”[All Fields] OR (“decision”[All Fields] AND “tree”[All Fields]) OR “decision tree”[All Fields]) OR (“support vector machine”[MeSH Terms] OR (“support”[All Fields] AND “vector”[All Fields] AND “machine”[All Fields]) OR “support vector machine”[All Fields]) OR (“random allocation”[MeSH Terms] OR (“random”[All Fields] AND “allocation”[All Fields]) OR “random allocation”[All Fields] OR “random”[All Fields] AND (“forests”[MeSH Terms] OR “forests”[All Fields] OR “forest”[All Fields]) OR (“principal component analysis”[MeSH Terms] OR (“principal”[All Fields] AND “component”[All Fields] AND “analysis”[All Fields]) OR “principal component analysis”[All Fields]) OR (hidden[All Fields] AND markov[All Fields] AND model[All Fields]) OR (bayesian[All Fields] AND (“Network”[Journal] OR “Netw Res Triangle Park N C”[Journal] OR “IEEE Netw”[Journal] OR “network”[All Fields]))))

OR (artificial[All Fields] AND (“neural networks, computer”[MeSH Terms] OR (“neural”[All Fields] AND “networks”[All Fields] AND “computer”[All Fields]) OR “computer neural networks”[All Fields] OR (“neural”[All Fields] AND “network”[All Fields]) OR “neural network”[All Fields])) OR (“deep learning”[MeSH Terms] OR (“deep”[All Fields] AND “learning”[All Fields]) OR “deep learning”[All Fields]) AND ((“asthma”[MeSH Terms] OR “asthma”[All Fields]) OR (“respiratory sounds”[MeSH Terms] OR (“respiratory”[All Fields] AND “sounds”[All Fields]) OR “respiratory sounds”[All Fields] OR “wheeze”[All Fields])) AND (hasabstract[Text] AND “humans”[MeSH Terms] AND English[lang]))

Total articles found: 1839

ProQuest

noft(((machine learning) OR cluster OR hierarchical OR python OR (artificial intelligence) OR (naive bayes) OR (k nearest neighbor) OR (k means clustering) OR (decision tree) OR (support vector machine) OR (principal component analysis) OR (hidden markov model) OR (random forest) OR (bayesian network) OR (artificial neural network) OR (deep learning))) AND noft(Asthma* or wheeze)

Total articles found: 112

Embase

Keywords search for ‘machine learning’, ‘cluster’, ‘hierarchical’, ‘python’, ‘artificial intelligence’, ‘naive Bayes’, ‘k nearest neighbor’, ‘k means clustering’, ‘decision tree’, ‘support vector machine’, ‘principal component analysis’, ‘hidden Markov model’, ‘Bayesian network’, ‘artificial neural network’, ‘deep learning’. Then combined these keywords by ‘OR’.

After that, Keywords search for ‘asthma’, ‘Wheeze and then combined these two by ‘OR’. Finally, combined machine learning keywords with asthma or wheeze by ‘AND’.

Total articles found: 4366

Appendix B

Machine learning categories (grouped)*

Clustering	Bayesian Network (BN)	Neural network	Regression	Principal Component Analysis (PCA)	Gaussian/Bernoulli Mixture model (GBMM)	Random Forest (RF)	Nearest Neighbor (NN)	Naïve Bayes (NB)	Support Vector Machine (SVM)	Decision Tree (DT)	Hidden Markov Model (HMM)	Gradient boosting model	Misc.
k-means clustering	Bayesian networks	Artificial neural network	Logistic regression	Principal component analysis	Gaussian Mixture model	Random Forest	k-nearest neighbor	Naïve Bayes	Support vector machine	Decision tree	Hidden markov model	Gradient boosting	Predictor pursuit
Unsupervised clustering	Bayesian hierarchical model	Deep learning	Linear regression		Bernoulli Mixture model		Nearest neighbor	Gaussian naïve Bayes	Support vector regression	CART (classification and regression tree)	Markov Chain Monte Carlo (MCMC)	Boosting	Unsupervised multidimensional factor analysis technique (NNS-PARAFAC model)
Supervised clustering	Gaussian Bayesian networks	Perceptron	LASSO				KNN		CIBERSORT	Rule-learning algorithm		XGBoost	Dynamic latent trait model with mixed hidden Markov extension
Hierarchical clustering	Adaptive Bayesian network		Elastic net							C4.5 classifier		Adaptive boosting	Modified nonnegative matrix factorization model
			Partial least square discriminant analysis									Multiboost with decision stumps	Fuzzy rule-based model
			Linear discriminant analysis										Bayesian machine learning modeling framework with an expectation propagation algorithm
			Sparse partial least squares discriminant analysis (sPLS-DA)										Natural language processing (NLP)
													Exploratory factor analysis (EFA)
													Gaussian Process (GP)
													PEAK algorithm
													Mahalanobis–Taguchi system (MTS)
													Bagging

***Clustering:** Clustering is a machine learning technique that involves dividing the population or data points into a number of groups.

Bayesian network (BN): A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph (DAG).

(Continued)

Appendix B: Continued.

***Neural network:** A neural network is a highly flexible method that works like a network of neurons in the human brain and nervous system. A neural network consists of three parts: input layer, hidden layer, and output layer. The hidden layer can be one, two, three or more layers. With a single hidden layer, the network is called an artificial neural network; with multiple hidden layers, it is called a deep learning network.

***Regression:** Regression analysis is a method that allows a person to examine the relationship between two or more variables of interest.

***Support vector machine (SVM):** The algorithm of a support vector machine creates a line or a hyperplane which separates the data into classes.

***Decision tree:** A decision tree is a structure like a flowchart in which each internal node indicates a test on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label.

***Hidden Markov model (HMM):** A Hidden Markov model is a probabilistic framework where the observed data are modeled as a series of outputs generated by one of several (hidden) internal states.

***Gradient boosting model:** A gradient boosting model is a supervised machine learning technique for regression and classification problems that aggregates an ensemble of weak individual models to obtain a more accurate final model.

***Naive Bayes (NB):** Naive Bayes is a set of supervised learning algorithms based on the application of Bayes' theorem with a "naïve" assumption (i.e. features of a measurement are independent of each other) of conditional independence but they are also Bayesian in nature as the outcome is based on a conditional probability computed on the input feature(s).

***Nearest Neighbor (NN):** K-nearest neighbor is a non-parametric learning algorithm. The principle of KNN is to find a user defined (k) number of training samples that are closest in distance to a point and make predictions from these. ***Random Forest (RF):** Random Forest is a supervised machine learning algorithm that can be used for both classification and regression tasks. The 'forest' is an ensemble of decision trees where multiple decision trees are built and combined to get more precise and stable predictions using a random selection of features at each node to determine the split.

***Principal Component Analysis (PCA):** Principal component analysis is a technique to reduce the dimensionality and maximize the variance of a dataset which has a large number of interrelated variables. In this process, the original variables are transformed to a set of new variables (principal component) which are uncorrelated and ordered where the largest variation comes first.

***Gaussian/Bernoulli mixture model (GBMM):** Gaussian mixture model is a probabilistic model where all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Gaussian distribution (normal distribution) is a bell-shaped curve consisting of symmetrically distributed data points around a mean value. Gaussian mixture models are fitted by the expectation-maximization algorithm. A specialized model for binary data is called a Bernoulli mixture model.

***Miscellaneous (Misc.):** This category includes techniques not specified elsewhere. There is only one machine learning technique used for each specific type in this category.