Supplementary material for the article “Artificial Intelligence techniques in Asthma: A systematic review and critical appraisal of the existing literature”

Table of contents

Introduction to Artificial Intelligence

The aim of this supplement is to serve as an introduction in Artificial Intelligence. First, we provide a table of abbreviations where the reader may refer, in order to facilitate the comprehension of the manuscript. In the next section “AI/Ml flavors” we describe the Data Mining process and present a rough categorization of AI/ML techniques based on the learning process. Next, we briefly describe the most commonly employed classification algorithms, especially the ones that are frequently used in medicine oriented problems.

Table of Abbreviations

In this section (**Table S1**) we provide a list of the most commonly used abbreviations pertaining to Artificial Intelligence that are frequently used throughout the manuscript.

Table S1 contains a list of the abbreviations used throughout the main body of this manuscript.

AI	Artificial Intelligence
ML	Machine Learning
DM	Data mining
ANN	Artificial Neural Networks
RF	Random Forest
DT	Decision Tree
PCA	Principal Component Analysis
SVM	Support Vector Machines
LR	Logistic Regression
BN	Bayesian Network
HMM	Hidden Markov Model
k-NN	k Nearest Neighbors
SOM	Self Organizing Map
GMM	Gaussian Mixture Model
NB	Naive Bayes
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Se	Sensitivity
TPR	True Positive Rate

Sp	Specificity
TNR	True Negative Rate
Acc	Accuracy
ROC	Receiver Operating Characteristic
AUC	Area Under ROC Curve
PPV	Positive Predictive Value
NPV	Negative Predictive Value
LOOCV	Leave One Out Cross Validation

Table S1: Table of the most frequently used abbreviations in this section.

AI/ML “flavors”

Artificial Intelligence (AI) refers to the software that is able to make a machine intelligent such that it performs human tasks, i.e. process, learn and respond to information gained from data; whereas Machine Learning (ML) is the process followed in order to make a machine learn how to perform a specific task, and in a similar manner as a human to perform better as the experience increases. Both AI and ML are data driven processes whereby the computer or the algorithm is presented with input data and the desired output and subsequently “learns” the inherent relations that lead from the input to the output. This is a completely different approach compared to a traditional computer programme where input data are fed and based on a set of extremely precise predefined instructions the computer returns a specific outcome. Similarly with AI and ML, Data Mining (DM) involves the computational and programming steps in order to “mine” large amounts of complex data for meaningful patterns and consequently knowledge. **Figure S1** depicts the steps of the DM process. There are roughly two basic phases within the DM process: i) during the **training phase,** the ML algorithm is fed with input data based on which a model is trained that captures the relations and inherent patterns within the data. During the training phase the raw input data are subject to a series of preprocessing steps aiming to increase the quality of the data, identify the set of more informative features and omit potentially redundant or irrelevant information. Inherent to the training phase is the process of model evaluation where the parameters of the trained model are further fine-tuned in order to procure a well-trained model. ii) In the **predicting phase** new instances of unknown data are fed as input to the previously trained model and the respective labels are predicted.

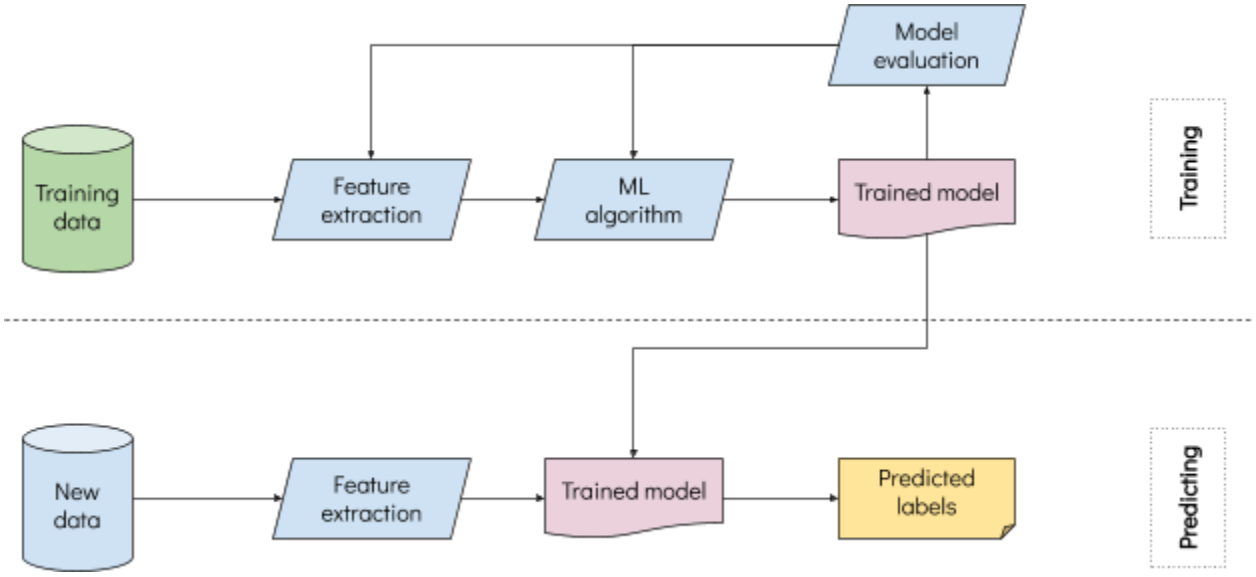

Figure S1: Flowchart of the Data Mining (DM) process.

The learning procedure of ML algorithms is divided into two broad categories, i.e. supervised and unsupervised learning, based on whether the output values (class) of the input samples are fed to the algorithm as prior knowledge or not (**Figure S2**). In the latter case the algorithm is expected to identify the underlying classes in the provided data.

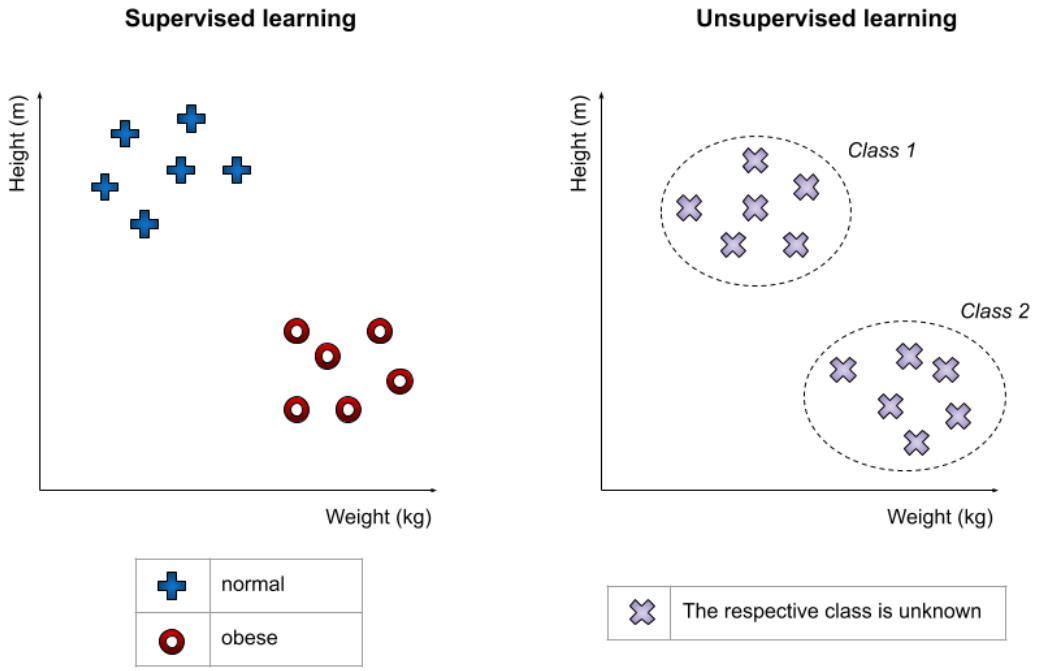

Figure S2: In supervised learning, the classes are already known and the algorithms aims to formulate a boundary that separates the given classes; in unsupervised learning the classes are unknown and the algorithm aims to “understand” the data and find inherent patterns or groupings.

Besides supervised and unsupervised learning there is another hybrid technique called semi-supervised learning which is often used when the unlabeled input data in a dataset are far more than the labeled ones. In semi-supervised learning the small amount of labeled input data is used as a starting point for training the algorithm, which is further trained with large amounts of unlabeled data. Supervised learning has two main branches, classification and regression; within a classification task the output values are a finite number of classes, whereas in the case of a regression problem the output variable is continuous. Unsupervised learning is largely represented by clustering where the algorithm aims to identify a set of clusters that are inherent to the input data (**Figure S3**).

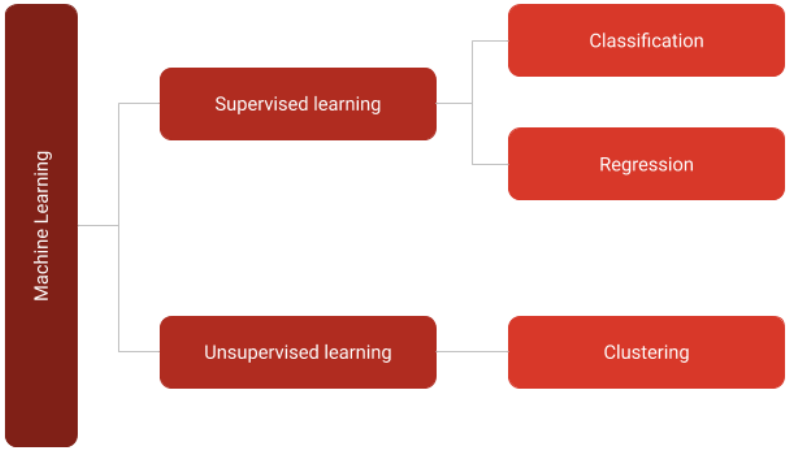

Figure S3: Supervised and unsupervised learning.

Overview of ML techniques

Over the past decades several machine learning algorithms have been presented in the literature, which differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve. Below, we will describe briefly the most popular machine learning algorithms: Bayesian networks, Naive Bayes, Artificial Neural Networks, Decision Trees, Random forests and Support Vector Machines.

Bayesian Networks

A Bayesian network (belief network, directed acyclic graph model) is a model that is built based on the observed probabilistic relationship among a set of variables (e.g. symptoms and diseases); therefore its output is rather a probability than a prediction. Bayesian networks have been widely used in series of ML problems, including medical applications since they are able to provide reasoning for the reported outcomes as well as assign a probability representing confidence for each decision. As shown in **Figure S4** below, each node of the network is accompanied by a table of probabilities defined by the values of the variables it is connected to, i.e. the ones that affect its outcome. In the case that all employed variables are "naively" considered independent, the resulting algorithm is called Naive Bayes.

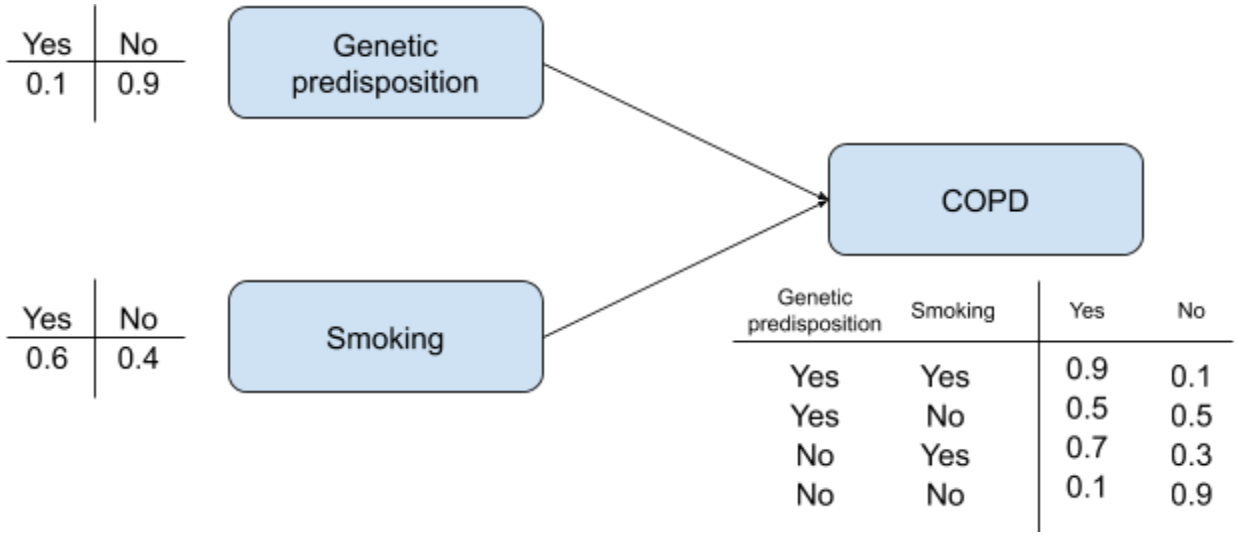

Figure S4: A provisional Bayesian network for COPD.

Artificial Neural Networks

Artificial Neural Networks are vaguely inspired by the notion and function of biological neural networks where neurons are interconnected by synapses and are trained to perform a specific task when activated. Artificial Neural Networks have proven quite useful in a series of tasks from various fields since they often perform very well. Due to their layered and often largely interconnected structure (**Figure S5**) the training process is quite time consuming and more importantly reasoning is almost impossible, therefore, they are often regarded as "black-boxes". Especially in medically oriented tasks this lack of explanation for the reported decision has attracted much criticism. Another concept that should be mentioned here is deep learning, that constitutes a subset of machine learning whereby the model resembles the layered approach of problem solving carried out by the human brain. Deep learning employs ANNs and a typical model often has at least three layers, where information is passed onto the next layer.

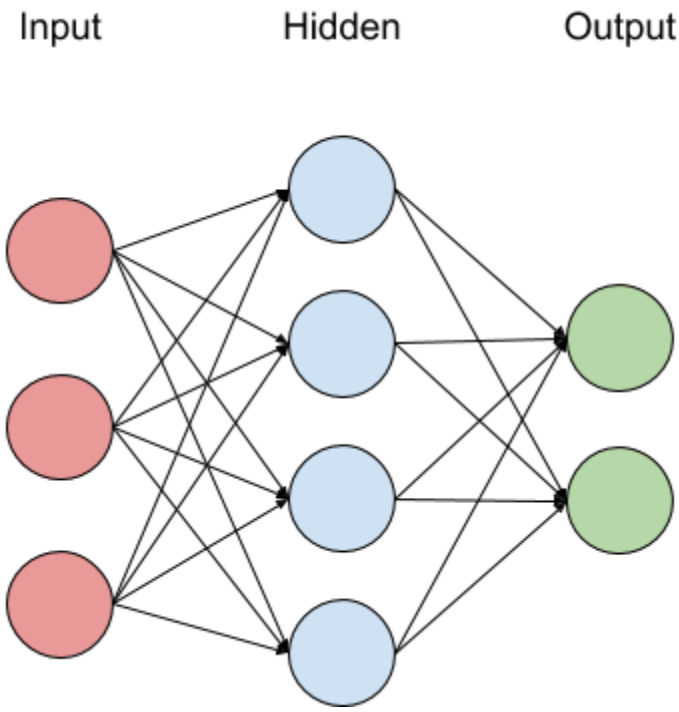

Figure S5: Architecture of an Artificial Neural Network with one hidden layer.

Decision Trees

Decision Trees constitute tree-structured classifiers where each node represents a variable and the leaves correspond to decision outcomes. The branches represent conjunctions of features that lead to the outcomes; by traversing the tree given the features values of a new sample, we are able to conjecture about its outcome. During the training phase where the tree architecture is formulated, the C4.5 algorithm is employed which often performs quite fast. The resulting architecture besides its simplicity, is also quite intuitive and transparent allowing for justified decisions. Specifically, each decision is based on a human-readable rule which provides adequate reasoning and subsequently makes Decision Trees a quite appealing solution for medical problems where transparency and reasoning are often prerequisites. **Figure S6** depicts a provisional architecture of a Decision Tree.

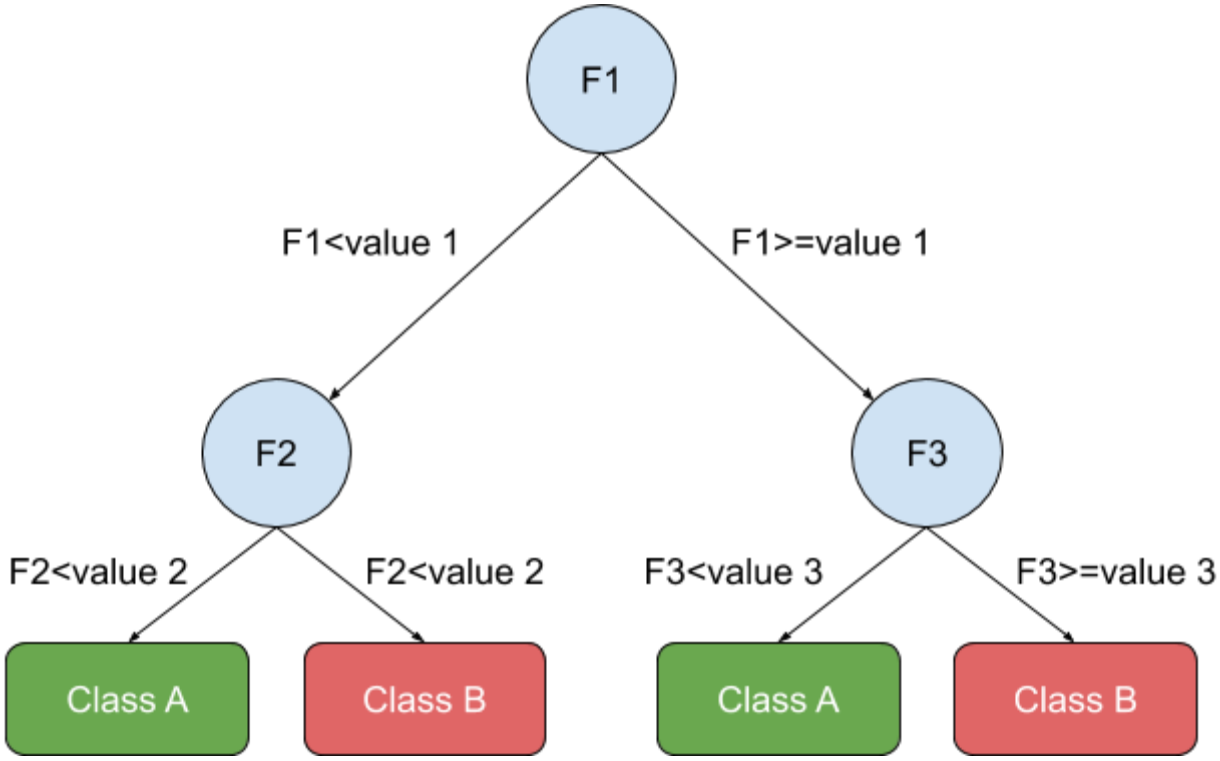

Figure S6: Provisional architecture of a Decision Tree classifier.

Random Forests

Random Forests or Random Decision Forests constitute an ensemble classifier that operates by constructing multiple Decision Trees in data subsets and assigning the output value by performing majority voting across the individual Decision Trees. **Figure S7** shows an exemplar Random Forest architecture.

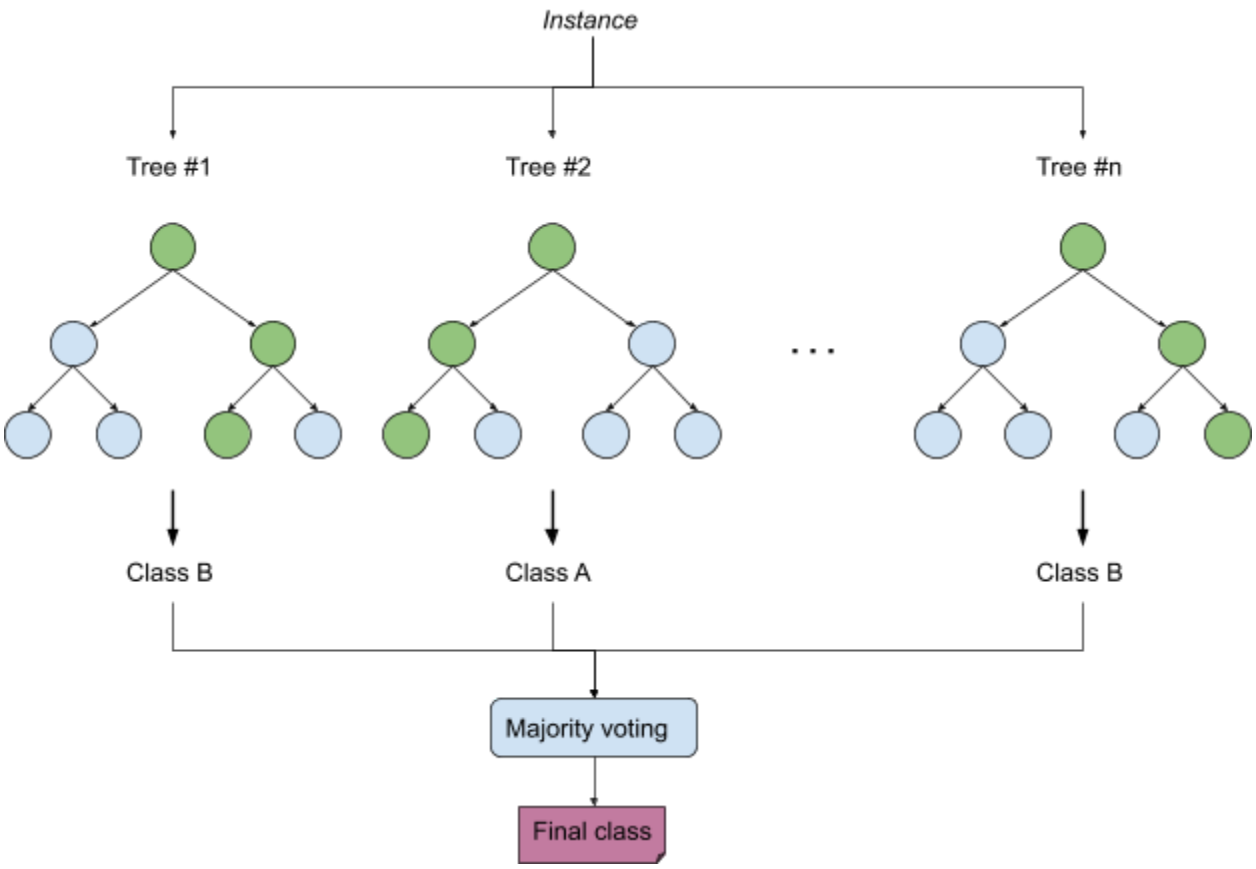

Figure S7: Architecture of a Random Forest algorithm.

Support Vector Machines

Support Vector Machines are one of the latest machine learning algorithms that has also been used extensively in medical and non-medical applications, due to the good performance and the generalization capability they often achieve. These two qualities are owed to the inherent process of training; specifically, Support Vector Machines map the initial input vector to a feature space of higher dimensionality where the samples can be separated with a linear hyperplane (kernel “trick”). Next, the algorithm searches across all possible hyperplanes that separate the samples in order to identify the one that maximizes the distance between the decision hyperplane and the most dubious instances (**Figure S8**).

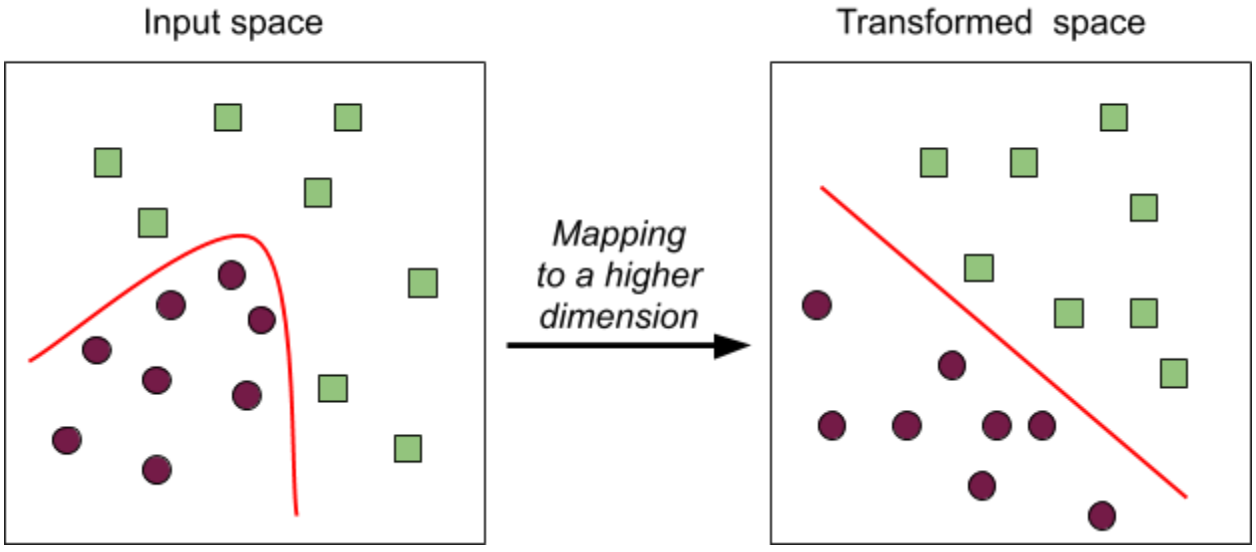

Figure S8: The kernel trick performed by the Support Vector Machines involves mapping the input vector to a higher dimensionality where the instances can be discriminated with a linear hyperplane.

Besides, the aforementioned algorithms, there are plenty of other machine learning algorithms, as well as variations of those algorithms with their respective strengths and limitations that heals towards deciding the most appropriate one for each task under consideration.

AI/ML validation

Within all “flavors” of AI or ML there are certain issues that need to be dealt with, that pertain to the fact that AI is essentially data-driven. When a model is trained with very limited data, these samples are memorized by the algorithm and the performance is nearly optimal for the specific dataset but very poor for other samples. This is much like a human that learns by heart a very specific task and is unable to perform well in other tasks. In a similar manner, an algorithm that is expected to discriminate between two classes and has been trained with an unbalanced dataset where one class is largely underrepresented, its performance towards disciminating that class will be relatively poor. This resembles a child that can recognize a basic set of common colors but when presented with one that has seen only a few times, it will most likely not recognize it.

All the aforementioned aspects regarding the performance of the algorithm are assessed quantitatively during the validation of the algorithm. For validation purposes the dataset is divided into two subsets, namely training and testing set where the latter is used in order to assess the performance of the trained model with new and previously unseen input data. Based on the size of the initial dataset, the testing set often contains 20%-40% of the input data. Another popular technique that is frequently used for validation purposes is *n*-fold cross

validation, whereby the initial dataset is partitioned in *n* equal subsets (or folds) from which *n-1* are used for training and the remaining one is used for testing; this process is repeated *n* times until all the folds have been used once for testing and the respective results are averaged in order to assess the overall performance of the model. A variation of *n*-fold cross validation is called Leave One Out Cross Validation (LOOCV) where *n* equals the total number of samples in the dataset. LOOCV is often indicated for limited datasets but is rather computationally intensive.

As for evaluation metrics, several ones have been described depending on the purpose of the machine learning algorithm, e.g. classification, regression, etc. The most widely used evaluation metrics are presented in **Table S2**.

Table S2: Most common metrics used for assessing the performance of ML algorithms.

Metric	**Formula**	**Description**
Sensitivity (Se) or True Positive Rate (TPR)	$TP/(TP+FN)$	Fraction of positive examples, predicted correctly by the model
Specificity (Sp) or True Negative Rate (TNR)	$TN/(TN+FP)$	Fraction of negative examples, predicted correctly by the model
Accuracy (Acc)	$TP+TN/(TP+FP+TN+FN)$	Overall correctness of the model, the ratio of correctly predicted outcomes and total number of examples
Receiver Operating Characteristic (ROC)	-	Graphical plot displaying the trade-off between the true positive rate and the false positive rate
Area Under ROC curve (AUC)	-	The two-dimensional area underneath the entire ROC curve
Positive predictive value (PPV)	$TP/(TP+FP)$	The proportion of positive results in the true positive results
Negative predictive value (NPV)	$TN/(TN+FN)$	The proportion of negative results in the true negative results
F1 score	$2*TP/(2*TP+FP+FN)$	The harmonic mean of PPV and Se

Kappa statistic	$[Pr(A) - Pr(E)]/[1 - Pr(E)]$ Pr(A): the percentage of observed agreement between the predictions and actual values Pr(E): the percentage of chance agreement between the predictions and actual values.	The agreement between the predicted results obtained by the model and the actual values

True Positive (TP): an outcome where the model correctly predicts the positive class.
True Negative (TN): an outcome where the model correctly predicts the negative class.
False Positive (FP): an outcome where the model incorrectly predicts the positive class.
False Negative (FN): an outcome where the model incorrectly predicts the negative class.

Artificial Intelligence and Asthma

As noted in the section 'Literature Review' of the main manuscript, the retrieved publications are divided into four categories, namely: (1) Asthma screening and diagnosis, (2) Patient classification, (3) Asthma management and monitoring, and (4) Asthma treatment. The articles from each category are summarized in a separate table where the respective studies can be compared by a set of qualitative and quantitative criteria or characteristics. In the first column ('*Ref*') we provide the reference for each study, the second column ('*ML algorithm*') shows the ML algorithm that was employed in the study. In cases where the study explored the performance of several ML algorithms, the best performing algorithm is reported. The third column ('*Sample size*') shows the total number of subjects or samples used in each study. The fourth column ('*Evaluation method*') shows the technique used for evaluating the performance of the proposed classification scheme; the fifth column ('*Performance*') contains a set of the most important reported metrics assessing the performance of the proposed work. In the last column ('*Important features*') we present the features reported in each study as being most important and informative.

Table S3, **Table S4** and **Table S5** contain studies related to 'Asthma screening and diagnosis', 'Patient classification' and 'Asthma management and monitoring', respectively.

Asthma screening and diagnosis

Table S3: Publications relevant to 'Asthma screening and diagnosis'.

Ref	ML algorithm	Sample size	Input features	Evaluation method	Performance	Important features
[14]	SVM	73	Capnography	LOOCV	Acc=94.52%, Se=97.67%, Sp=90%	Upward expiration (AR1), downward inspiration (AR2), sum of AR1 and AR2
[15]	SVM	60	Clinical (lung sound recordings)	LOOCV	Acc=93.3%	Exchange time of the instantaneous frequency
[16]	SVM	254	Clinical (medical record)	10-fold CV	Acc=98.59%,Se=98.59%,Sp=98.61%	
[17]	ANN & Fuzzy logic	780	Clinical (Portable spirometer)		Acc=97.32%	
[18]	HMM	16	Clinical (respiratory sounds)		Acc=94.91, Se=89.34%, Sp=96.28%	
[19]	k-NN	75	Forced oscillation technique parameters	N-fold CV, LOOCV	Se=82.9%, Sp=86.1%, AUC=0.91	Cross products of the FOT parameters: fr2, Xm.Cdyn [fr=resonance frequency, Xm=Mean respiratory reactance, Cdyn=Respiratory system dynamic compliance]
[20]	SVM	16	Clinical (phonopneumogr	LOOCV	Reliability (TPR*TNR)=97.36%	

			ams-respiratory sounds)			
[21]	ANN	112	Clinical (questionnaire, history)	10-fold CV	Acc=96.77%, Se=96.15%, Sp=100%	Wheezing episodes until 5th year, wheezing episodes between 3rd and 5th year, wheezing episodes until 3rd year, weight, waist's perimeter, seasonal symptoms, FEF25/75, number of family members, ICS
[22]	Fuzzy rules	278	Clinical		Se=88%, Sp=100%	
[23]	ANN		Clinical, epidemiological		AUC=0.903	
[24]	SVM	150 discharge summaries	Clinical (EMR)	10-fold CV	Acc=82%	
[25]	ANN	350	Clinical	CV		
[26]	LR	514	Clinical (EHR)	Training-Testing	Se=86%, Sp=98%	History of allergic rhinitis, eczema, family history of asthma, maternal history of smoking during pregnancy
[27]	ANN	254	Clinical	Training-Testing (70-30)	Acc=100%	Cough, symptoms of exercise induced asthma, humidity levels at home, emotional reactions, air pollution, wheeze, respiratory distress, hospitalization before 3 years of age, response to irritants, response to allergens, phlegm, allergies (both parents), pursiness
[28]	Fusion algorithm	170	Clinical (questionnaires)	10-fold CV	Se=98%, AUC=1	
[29]	SVM	30	Clinical (respiratory sounds)	Training-Testing	Acc=94.6%	
[30]	RF	132	Clinical	Training-Testing (80-20)	Precision=83%	Inhaler, MEF2575, Age, Smoker, Wheeze and Breath Shortness
[31]	Fuzzy rules & ANN	455	Clinical (spirometry, impulse oscillometry)	Independent test set	Acc=99%, Se=99%	
[32]	ANN	58	Clinical (breath sounds)	Independent test set	Se=94.6, Sp=100%	
[33]	ANN	48	Clinical (breath sounds)	Training-Testing (80-20)	Acc=92.8%	
[34]	ANN	827	Genomic (IgE reactivity)	Training-Testing (60-40)	Acc=78%	Allergens: Penicillin, Derm. Farinae, Kiwi, Timothy grass, Alpha amylase, Ph1 p1, Derp 1
[35]	ANN	51	Electronic nose, FeNO, and lung function testing	Training-Testing	Acc=95.8%	Electronic nose and FeNO
[36]	ANN	82	Genomic (SNPs)	5-fold CV	Acc=78%	
[37]	ANN	2832	Clinical (questionnaire)	Independent test set	PPV=100%	

[38]	ANN	10	Clinical (respiration sounds)	4-fold CV	Acc=80%	
[39]	ANN	180	Clinical (questionnaire)	Independent test set	Spearman rank order correlation coefficient=0.66	
[40]	SOM	32	Clinical (lung sounds)		Acc=78%, Se=52%	
[41]	DT	968	Clinical (lung function testing)	10-fold CV	PPV=66%, TPR=82%	
[42]	DT	12512	Clinical (spirometry, history, questionnaire, medication)	10-fold CV, Independent test set	Se=79%	
[43]	DT	26 signals	Clinical (lung sounds)	LOOCV	Acc=92%	
[44]	RF	554	Genetic (SNPs) and clinical	Bootstrapping	Acc=87%, AUC=0.84	Allergen sensitization, lung function markers
[45]	RF	461	Genetic and clinical	Training-Testing (80-20)	Se=97%, Sp=34%, AUC=0.82	Dust mite, pollens, pet allergens
[46]	GMM	24	Clinical (lung sounds)	LOOCV	Se=97.2%, Sp=94.2%, AUC=0.974	
[47]	LR	190	Genetic (nasal RNA)	Independent test set	AUC=0.994	
[48]	SVM	95 recordings	Clinical (respiratory sounds)		Acc=84%, Se=71.4%, Sp=88.9%	
[49]	DT	5032	Clinical (patient record)	5-fold CV	Definite asthma cases: PPV=66%,Se=98%,Sp =95%; Definite and probable asthma cases: PPV=82%, Se=96%, Sp=90%; Definite-probable and doubtful asthma cases: PPV=57%, Se=95%, Sp=67%	
[50]	SVM	283	Genetic (gene expression)	10-fold CV	Acc=95%	
[51]	RF	109	Exhaled breath condensate	Independent test set	Se=80%, Sp=75%	
[52]	RF	79	Genetic (micro RNA)	LOOCV	AUC=0.974	miR-125b, miR-16, miR-299-5p, miR-126, miR-206, miR-133b
[53]	ANN		Clinical	Independent test set	Acc=93%, Se=81%, Sp=100%	
[54]	k-NN	10	Clinical (lung sounds)	1-fold CV	Acc=77%	
[55]	ANN	60	Clinical	Training-Testing	Acc=43%	
[56]	JDINAC	461	Clinical	10-fold CV	Acc=86%, Se=84%, Sp=87%, AUC=0.94	Component-specific IgEs
[57]	LR & RF	177	Genomic (serum miRNA)	10-fold CV	Se=89%, Sp=77%, AUC=0.86	

[58]	NB	322	Clinical, patients history	10-fold CV	Acc=70.7%	
[59]	ANN		Capnogram		Acc=95.65%	
[60]	DT	1104	Clinical	10-fold CV	Se=93%, Sp=85%	Ever had asthma, current asthma, shortness of breath, atopy and wheezing, breathless but no family history
[61]	ANN & Fuzzy expert system	908	Genomic (SNPs)	Independent test set	Acc=94%	MS4A2 Glu237Gly, IL4Ra Glu375Ala

SVM: Support Vector Machine; ANN: Artificial Neural Networks; HMM: Hidden Markov Models; k-NN: k Nearest Neighbors; LR: Logistic Regression; RF: Random Forests; SOM: Self-organizing Maps; DT: Decision Trees; GMM: Gaussian Mixture Models; JDINAC: Joint density-based non-parametric differential interaction network analysis and classification; NB: Naive Bayes

Patient classification

Table S4: Publications relevant to 'Patient classification'.

Ref	ML algorithm	Sample size	Input features	Evaluation method	Performance	Important features
[67]	ANN	344	Genomic	5-fold CV	Acc=74.4%	
[68]	RF	96	Clinical	Training-Testing	Acc=70%, Se=81%, Sp=67%, AUC=0.86	15 VOCs
[54]	k-NN	10	Clinical (lung sounds)	1-fold CV	Acc=77%	
[55]	ANN	60	Clinical	Training-Testing	Acc=43%	
[56]	JDINAC	461	Clinical	10-fold CV	Acc=86%, Se=84%, Sp=87%, AUC=0.94	Component-specific IgEs
[57]	LR & RF	177	Genomic (serum miRNA)	10-fold CV	Se=89%, Sp=77%, AUC=0.86	
[58]	NB	322	Clinical, patients history	10-fold CV	Acc=70.7%	
[69]	Ensemble classifier	55	Clinical	LOOCV	PPV=95%	Tracheal wheeze sounds
[70]	Fuzzy Rules	28	Clinical (combination of 10 asthma severity scores)		Kappa coefficient=1	
[71]	DT	341	Clinical	10-fold CV	Se=84%, Sp=71%, AUC=0.83	
[72]	LASSO & stochastic gradient boosting	260	Clinical, Genomic	LOOCV	AUC=0.81	PKN2, PTK2, ALPP
[73]	SVM	346	Clinical	10-fold CV	Acc=81%, Se=62%, Sp=87%	-

[74]	DT	107	Clinical	10-fold CV	Acc=82.4%	Th2-mediated inflammation, corticosteroid insensitivity
[75]	GMM	1642	Clinical	CV	-	IL-13, IL-5
[76]	SVM	378	Clinical	LOOCV	Acc=93%	Age of asthma onset, quality of life, symptoms, medications, health care use
[77]	HMM	2255	Clinical	10-fold CV		Patterns of IgE responses over time
[78]	LR	1048	Clinical	10-fold CV	Acc=85%	
[79]	RF	348	Genomic	-	Misclassification rate=44%	ADAM33
[80]	DT	205	Genomic, Clinical	-	Acc=78%	Gene expression, clinical covariates, indicators of health outcomes
[81]	DT	3160	Clinical	Independent test set	AUC=0.72	Change in PEF, hospitalization for asthma, initial oxygen saturation on room air, initial PEF, risk stratification, emergency care of acute asthma
[59]	ANN		Capnogram		Acc=95.65%	
[60]	DT	1104	Clinical	10-fold CV	Se=93%, Sp=85%	Ever had asthma, current asthma, shortness of breath, atopy and wheezing, breathless but no family history
[82]	Fuzzy expert system	42	Clinical	-	Cohen kappa coefficient=1	
[83]	ANN	128	Clinical	10-fold CV	Acc=80%	
[84]	ANN	486	Clinical	Training-Testing	Acc=98.7%, Se=97.63%, Sp=97.83%	FEF25-75%
[85]	DT	872	Clinical	Independent test set	Cluster 1: Se=84.1%, Sp=96.3%; Cluster 2: Se=94.1%, Sp=99.5%, Cluster 3: Se=90.1%, Sp=99.3%; Cluster 4: Se=91.6%, Sp=91.9%	Comorbidities, adherence, cognitive dysfunction, depression
[86]	LR	12792	Patient records	Independent test set	AUC=0.67	Age, BMI, race, smoking history
[87]	BN	9801	Clinical	Independent test set	Average posterior probability=0.833	Eczema, wheeze, rhinitis
[88]	LR & SVM	1019	Clinical	5-fold CV	Short-term prediction=0.86; Long-term prediction=0.66	Obesity, allergy
[61]	ANN & Fuzzy expert system	908	Genomic (SNPs)	Independent test set	Acc=94%	MS4A2 Glu237Gly, IL4Ra Glu375Ala

SVM: Support Vector Machine; ANN: Artificial Neural Networks; JDINAC: Joint density-based non-parametric differential; interaction network analysis and classification; HMM: Hidden Markov Models; k-NN: k Nearest Neighbors; LR: Logistic Regression; RF: Random Forests; DT: Decision Trees; GMM: Gaussian Mixture Models; BN: Bayesian Networks

Asthma management and monitoring

Table S5: Publications relevant to 'Asthma management and monitoring'.

Ref	ML algorithm	Sample size	Input features	Evaluation method	Performance	Important features
[89]	Fuzzy expert system	25	Clinical (exacerbations)			
[90]	Ensemble classifier	96	Clinical, Patients record	5-fold CV	Acc=91.66%	Out of 140 initial variables,35 clinical variables were chosen
[91]	RF	42	Genomic	LOOCV	Acc=74%	20 features out of 30
[92]	RF	2	Clinical	10-fold CV	Acc=80.10%	FEV1, PEF,dust density, heart rate
[93]	Association rule mining	20959 ED visits	Environmental data, Patients records	Training-Testing	FDR=13%	SO2, NO, NO2, PM
[94]	Multiboost & Decision stumps	180	Clinical	10-fold CV	Acc=71.8%, Se=73.8%, Sp=71.4 %, AUC=0.757	
[95]	ANN & DT		Social media, Environmental data	10-fold CV	Precision=70%	asthma tweets, CO, NO2 and PM2.5
[96]	PCA & SVM	112	Clinical	10-fold CV	Se=95.54%	18 features
[97]	Pattern Based Decision Tree (PBDT) and Pattern Based Class-Association Rule (PBCAR)	33	Patient records, Clinical, Environmental data	Training-Testing (70-30)	PBCAR Acc=86.89%, Recall=84.12%; PBDT Acc=87.52%, Recall=85.59%	
[98]	ANN		Patients records, Clinical	CV	Acc=84%	
[99]	SVM	162	Clinical (cough signals)	-	Probability of correct classification=90%	-
[100]	RF	3206	Clinical, Patients records	Lasso penalization, out-of-bag estimation, CV, Ridge penalization	Critical care prediction: C-statistics=0.80, Se=79%; Hospitalization prediction: C-statistics=0.83, Se=75%	Advanced age, vital signs, arrival mode, comorbidities

[101]	RF	16	Clinical	LOOCV	Acc=87.4%, Se=47.2%, Sp=96.3%	Heart rate, respiratory parameters
[102]	ANN		Meteorological, Air pollution	CV	Acc=81%	
[103]	ANN	3602	Clinical, Meteorological, Air pollution	R2, Index of Agreement (IA), Root Mean Square Error (RMSE), Mean Bias Error (MBE)	0–4 years: R2=0.567; 5–14 years: R2= 0.207; 0–14 years: R2=0.528	
[104]	ANN	42	Clinical			
[105]	ANN		Clinical, Pollution data	Training-Testing	Acc=53%	Air pollution levels (NOx)
[106]	ANN	27	Clinical, Environmental data	CV		SO2, NO2, temperature, intake of medicine, relative humidity
[107]	Gradient boosting models	29354	Clinical, Patients records, Environmental, Air pollution, Neighborhood characteristics, Community viral load	3-fold CV	AUC=0.85	Oxygen saturation, pulse rate, respiratory rate, weight, age, triage acuity, weather variables
[108]	DT	200	Clinical	CV	Se=80%, Sp=89%	Dyspnea, accessory muscle use, wheezing
[109]	SVM	26	Clinical, Patient records (daily asthma diary)	Training-Testing	Acc=80%, Se=84%, Sp=80%	
[110]	RF	417	Clinical, Genomic	Independent test set	160-320 SNPs: AUC=0.66; 10 SNPs: AUC=0.57; Clinical traits: AUC=0.54	
[111]	Gradient boosting models	4548	Clinical, Environmental data	5-fold CV	AUC=0.78	Previous year bronchitic symptoms
[112]	XGBoost	7503	Air pollution, Meteorological data, Historical data	CV	AUC=0.832	Air pollution data, weather data, historical admissions data
[113]	LR	2691	Patients records	Training-Testing	Se=23%, PPV=56%, AUC=0.86	Number of ED visits in year 1, type of Insurance
[114]	BN	7001	Clinical	Training-Testing	Acc=100%, Se=100%, Sp=100%	63 variables out of 147 attributes
[115]	HMM		Clinical (respiration sounds)	CV	Se=85.7%	Cough
[116]	ANN & PCA	130	Clinical	3-fold CV	Se=100%, Sp=79.6%	FeNO, FEV1, FVC, FEV1/FVC, FEF25-75%

[67]	ANN	344	Genomic	5-fold CV	Acc=74.4%	
[68]	RF	96	Clinical	Training-Testing	Acc=70%, Se=81%, Sp=67%, AUC=0.86	15 VOCs
[58]	NB	322	Clinical, patients history	10-fold CV	Acc=70.7%	
[82]	Fuzzy expert system	42	Clinical	-	Cohen kappa coefficient=1	
[83]	ANN	128	Clinical	10-fold CV	Acc=80%	
[84]	ANN	486	Clinical	Training-Testing	Acc=98.7%, Se=97.63%, Sp=97.83%	FEF25-75%
[85]	DT	872	Clinical	Independent test set	Cluster 1: Se=84.1%, Sp=96.3%; Cluster 2: Se=94.1%, Sp=99.5%, Cluster 3: Se=90.1%, Sp=99.3%; Cluster 4: Se=91.6%, Sp=91.9%	Comorbidities, adherence, cognitive dysfunction, depression
[86]	LR	12792	Patient records	Independent test set	AUC=0.67	Age, BMI, race, smoking history
[87]	BN	9801	Clinical	Independent test set	Average posterior probability=0.833	Eczema, wheeze, rhinitis
[88]	LR & SVM	1019	Clinical	5-fold CV	Short-term prediction=0.86; Long-term prediction=0.66	Obesity, allergy

RF: Random Forests; ANN: Artificial Neural Networks; DT: Decision Tree; PCA: Principal Component Analysis; SVM: Support Vector Machine; LR: Logistic Regression; BN: Bayesian Network; HMM: Hidden Markov Model; NB: Naive Bayes