Research Article

Health Informatics Journal

Health Informatics Journal
1–26

DOI: 10.1177/14604582241283968
journals.sagepub.com/home/jhi

Machine learning approaches for asthma disease prediction among adults in Sri Lanka

JRNA Gunawardana
Institute for Health Policy, Sri Lanka and Robert Gordon University, UK

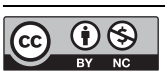
SD Viswakula
Department of Statistics, University of Colombo, Sri Lanka

Ravindra P Rannan-Eliya and Nilmini Wijemunige
Institute for Health Policy, Sri Lanka

Abstract

Objectives: Addressing the challenge of cost-effective asthma diagnosis amidst diverse symptom patterns among patients, this study aims to develop a machine learning-based asthma prediction tool for self-detection of asthma. **Methods:** Data from 6,665 participants in the Sri Lanka Health and Ageing Study (2018-2019) are used for this research. Thirteen machine learning algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbors, Gradient Boost, XGBoost, AdaBoost, CatBoost, LightGBM, Multi-Layer Perceptron, and Probabilistic Neural Network, are employed. **Results:** A hybrid version of Logistic Regression and LightGBM outperformed other models, achieving an AUC of 0.9062 and 79.85% sensitivity. Key predictive features for asthma include wheezing, breathlessness with wheezing, shortness of breath attacks, coughing attacks, chest tightness, nasal allergies, physical activity, passive smoking, ethnicity, and residential sector. **Conclusion:** Combining Logistic Regression and LightGBM models can effectively predict adult asthma based on self-reported symptoms and demographic and behavioural characteristics. The proposed expert system assists clinicians and patients in diagnosing potential asthma cases.

Corresponding author:
J R N A Gunawardana, Institute for Health Policy, 72, Park Street, Colombo 02, Western Province, Sri Lanka.
Email: nishaniamalka@gmail.com



Keywords
asthma, classification, disease prediction, machine learning, LightGBM, logistic regression

Introduction

Asthma is a chronic, non-communicable disease characterized by inflammation and narrowing of the bronchial tubes, resulting in reduced airflow in and out of the lungs.[1] It affected approximately 260 million individuals worldwide in 2019, and it was the 34th leading cause of disease burden, measured by Disability-Adjusted Life Years (DALYs).[2]

Asthma can be relieved and controlled by avoiding triggers and lifestyle changes, and it can be effectively controlled with medication. Control requires diagnosis and continuing management with medication, both of which are best done at the primary care level.[3] Whilst the burden of disease from asthma is mainly through its symptoms, it can be fatal if poorly controlled. When asthma is poorly controlled, people may require admission and the risk of mortality increases substantially. For this reason, asthma hospitalizations and deaths are often regarded as "avoidable", and admission and mortality rates for asthma have been used as indicators of primary care quality.[4,5]

Sri Lanka is a low-middle income developing country, which the evidence indicates suffers from a high burden of asthma. Asthma was the fourth leading cause of death in Sri Lanka in 2019 according to global estimates.[6] Although national estimates are not available for Sri Lanka, a study covering 7 out of 9 provinces found that wheezing (an indicator of asthma) was reported by 24% of adults, 80% of whom had at least one other symptom of asthma.[7] However, only 12% reported a diagnosis of asthma, indicating low levels of diagnosis.[7] The conclusion of low levels of diagnosis and inadequate treatment is reinforced by Sri Lanka's very high asthma hospital admission rates of 895 per 100,000 population,[8] fifteen times more than the average rate in Organization for Economic Co-operation and Development (OECD) countries.[9] Additionally, Sri Lanka had one of the highest reported asthma mortality rates in the world (1.3 deaths per 100,000), nearly two to five times the mortality rates reported in Europe and high-income countries.[10] Death rates were also higher in poorer socioeconomic quintiles and highest in estate areas of the central hill country,[10] suggesting that disparities in access to healthcare may be contributing.

Clinical diagnosis of asthma requires spirometry with testing of bronchodilator reversibility in a healthcare setting. There is limited capacity for this in the Sri Lankan healthcare system, even at the hospital level, and limited availability of equipment and supplies often results in spirometry testing not being done in patients with possible asthma.[11] Having lower-cost non-clinical options to identify people with a high risk of having asthma could improve diagnosis and treatment rates by prioritizing those at higher risk for clinical testing, and by increasing treatment rates to reduce morbidity and mortality.

One option for improving screening at the community level might be to provide the public with access to reliable screeners that they can self-use to identify if they might have asthma and should seek proper assessment by a physician. This would require a screening tool adapted to the Sri Lankan population that uses non-clinical information to identify individuals with a high likelihood of having asthma.

One approach to doing this is to use self-reported data along with Machine Learning (ML) and Artificial Intelligence (AI) to streamline the process of identifying individuals who may require further diagnostic assessments or treatment for asthma. By integrating AI algorithms, which analyze patient-reported symptoms and triggers such as coughing or wheezing, with ML models that leverage both current data and historical records, it becomes possible to predict the presence or

exacerbation of asthma with a personalized risk assessment.[12–15] These AI-driven decision support systems can assist healthcare providers in interpreting patient-reported symptoms and characteristics, potentially accelerating the diagnosis and treatment process.[13]

Moreover, a model that is developed using self-reported data could be used by individuals, or clinicians at primary or secondary care settings, particularly in countries without the resources for expensive laboratory tests. For a more comprehensive exploration of how ML and AI can be applied in this context, please refer to the "Related Work" section. Thus, we aimed to develop a predictive model incorporating adult asthma features and risk factors which can be deployed as a web application.

Related Work

Many existing clinical prediction models for asthma are designed for use within healthcare settings and require data that is typically only available through medical facilities. These models often involve clinical populations and use detailed clinical data, which is not always accessible to the general public, particularly in resource-limited settings like Sri Lanka. In contrast, the goal in the Sri Lankan context is to develop a self-use screener that can be utilized by ordinary people outside of healthcare facilities. This aligns with the need for accessible, population-based solutions that leverage available data to facilitate early detection and prompt medical consultation.

Tomita and colleagues[16] developed a model using Logistic Regression, Support Vector Machine (SVM), and Deep Neural Network (DNN) to predict adult asthma diagnoses. Their study used clinical features such as symptoms, physical signs, biochemical findings, lung function tests, and bronchial challenge test results from 566 adult outpatients at Kinki University Hospital, achieving a DNN model accuracy of 98%. Similarly, researchers from Ionian University, Greece, developed an asthma predictive model[17] using demographic, medical, and lung measurements, habits, and symptoms data from 132 patients, finding that Random Forest outperformed other models like Naïve Bayes, Logistic Regression, and SVM. These studies highlight the effectiveness of ML models but underscore the need for clinical data, which limits their applicability in non-clinical settings.

Philippine researchers integrated ML models to predict asthma using genetic information,[18] specifically Single-Nucleotide Polymorphism (SNP) data. They used Random Forest and Recursive Feature Elimination (RFE) algorithms to identify significant SNPs, with the integrated RF-SVM model achieving an accuracy of 62.5%. Priya and Priyadharshini[19] developed a Convolutional Neural Network (CNN) model that used a broad range of clinical features to achieve 98.36% accuracy. In another study, Chinese computer scientists[20] used classifiers such as Naïve Bayes and Random Forest on a dataset from a hospital in Pakistan, with Naïve Bayes attaining an accuracy of 98.75%. Additionally, a collaboration between South Korean and US researchers implemented a mobile health application using the Internet of Things (IoT).[21] They developed a CNN model using Peak Expiratory Flow Rates (PEFR), indoor particulate matter data, and weather data. This model, implemented as a mobile app on smartphones, demonstrated lower error rates compared to other benchmark techniques. These models, though highly accurate, rely on detailed clinical and genetic data.

In contrast, our approach focuses on developing a population-based asthma screener that can be used by individuals to self-assess their need for medical consultation. This approach leverages readily available data, making it more accessible and practical for use outside healthcare facilities. For instance, the National Health and Nutrition Examination Surveys (NHANES) data was used to predict asthma attacks in a cross-sectional United States (US) population, achieving an accuracy of

73.7% with an XGBoost model.[22] Additionally, a study combined ML with AI,[23] using data from social media and doctor visits to create a Decision Tree model with 87% accuracy, deployed as an Android app for user self-assessment.

Several real-world applications have successfully implemented ML and AI to assist with asthma management, highlighting the potential for self-use tools. For example, Smart Asthma: Forecast Asthma uses historical and real-time data to provide personalized forecasts and insights into potential triggers, enabling users to take preventive measures.[24,25] Similarly, Propeller Health uses sensors attached to inhalers to track medication use and environmental data, providing users with health forecasts based on local conditions and reducing emergency visits by 57%.[26–28] Hailie[29] and KagenAir[30] also offer personalized insights and forecasts to help users manage their asthma based on real-time data and usage patterns.

Most of the previously published approaches rely on data input by physicians and information from laboratory investigations and testing procedures. In Sri Lanka, where the priority is to triage people for clinical assessment, there is an urgent need for alternative tools that can be used by the public at the community level using information readily available to individuals. In Sri Lanka, where medical testing resources are limited, the widespread access to mobile phones with internet connectivity[31–33] provides an opportunity to develop a relevant technology solution. A public-facing web tool based on a population-based screener can help individuals self-assess their asthma risk and determine the need for medical consultation. This approach leverages accessible data and technology, making it a practical and scalable solution for early detection and management of asthma in Sri Lanka.

AI-based systems have been successfully used to screen for undiagnosed conditions in other health areas. For example, Singapore's AI system SELENA + addresses diabetic retinopathy by analyzing retinal images for early signs of eye diseases, significantly scaling up screening capabilities.[34] Singapore's healthcare system also uses AI for early detection, treatment targeting, and resource optimization, demonstrating the broad potential of AI in improving healthcare delivery and efficiency.[34] These examples illustrate the feasibility and benefits of implementing AI-based screening tools in various health contexts, and such tools could likely be useful for asthma detection in resource-limited settings like Sri Lanka.

Methodology

Overall framework

The study followed a framework of six phases:

- Phase 1: Data extraction
- Phase 2: Data preprocessing and exploratory data analysis
- Phase 3: Feature engineering
- Phase 4: Model fitting
- Phase 5: Model evaluation
- Phase 6: Model deployment

Figure 1 provides a visual representation of the comprehensive data analysis workflow executed within each of these phases.

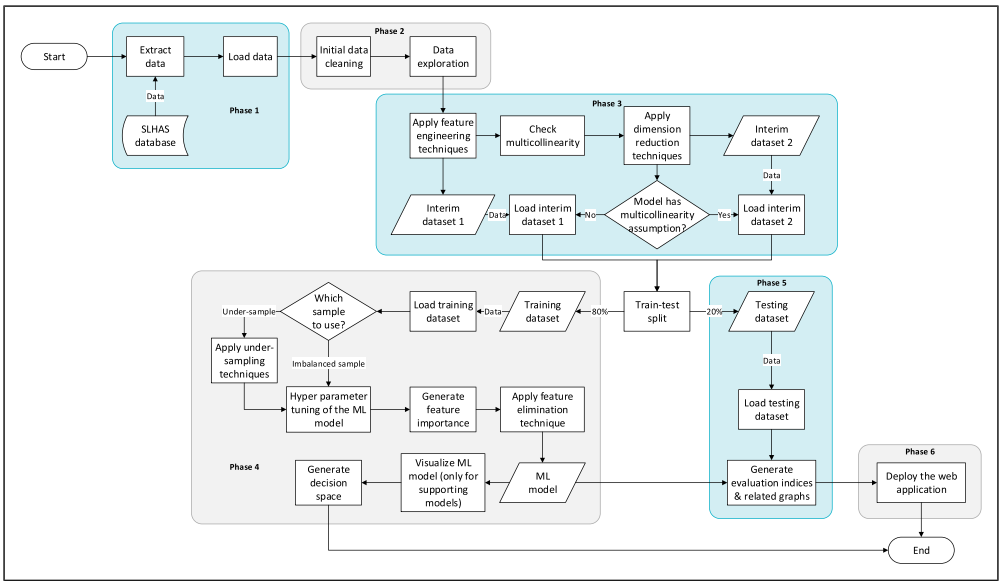


Figure 1. Overall framework.

Study design and participants

We utilized data from the inaugural wave of the Sri Lanka Health and Ageing Study (SLHAS), a nationally representative longitudinal cohort study, conducted between mid-November 2018 and mid-November 2019.

The SLHAS employed a stratified, multistage probability sampling strategy to procure a representative sample of the non-institutionalized adult population (≥18 years) across Sri Lanka.[35,36] This methodological framework treated all 14,014 Grama Niladhari Divisions (GNDs), the smallest administrative units in Sri Lanka, as primary sampling units (PSUs).[35,36] These divisions were stratified according to district, residential sector, and Area Socio-Economic Status (ASES), with the latter quantified through Principal Component Analysis (PCA) of socio-economic indicators derived from the 2012 national census data furnished by the Department of Census and Statistics (DCS).[35,36]

To achieve a robust sample, two or more PSUs were selected from each stratum via probability-proportionate-to-size sampling. Within each PSU, households were systematically sampled, and one eligible adult meeting inclusion criteria: resident, aged ≥18 years, not pregnant, and able to provide informed consent; was randomly selected, with age weighting applied to ensure demographic representativeness. Households were excluded if the individual selected could not participate.[35,36]

Fieldwork was rotated across provinces to mitigate seasonal bias. Participants attended local field clinics, where they underwent comprehensive interviews to gather data on chronic diseases and symptoms, risk factors, and socio-economic variables. Each participant also underwent anthropometric measurements and blood tests. For participants with mobility limitations, interviews and abbreviated assessments were conducted in their homes. Data was collected using a Computer-Assisted Personal Interviewing (CAPI) application deployed on tablet devices.[35,36]

Of the 10,689 households sampled, 10,062 agreed to participate. Among these, 6,624 adults attended field clinics, with an additional 41 individuals completing home interviews, yielding an effective response rate of 65.0%. Response rates varied, with higher participation observed among women (69%), adults aged ≥45 years (74%), rural residents (70%), and individuals with known diabetic conditions (73%).[35,36]

Data collection

The SLHAS extensively leveraged electronic data capture methods, prominently featuring CAPI for streamlined data acquisition. The integration of CAPI offered several advantages:

- Accelerated and enhanced precision in data acquisition within field settings compared to traditional methods.
- Mitigation of costs and errors associated with manual data entry from paper-based forms, including the need for subsequent data cleaning.
- Facilitation of intricate data collection designs, including randomized questionnaire modules, Global Positioning System (GPS) data recording, and coding of complex responses.
- Drastically reduced turnaround time from data acquisition to data availability for analysis.
- Agile adaptation of data collection tools in response to survey experience or project adjustments.

To facilitate electronic data capture, the SLHAS utilized iFormBuilder, a cloud-based mobile data collection platform provided by Zerion Software.[37] Customized data collection forms were employed, enabling data entry through typing, point-and-click methods, barcode scanning, and GPS signal capture via tablet PCs (Personal Computers). Furthermore, the iForm collection tools incorporated data validation checks to minimize erroneous entries, while the question design minimized reliance on free-text responses. More importantly, all Personally Identifiable Information (PII) was safeguarded through Public-Key Cryptography (PKC), ensuring robust data security and patient privacy.

Questionnaire tools

All the questionnaires used in this study are validated, translated, back-translated, and pilot-tested in the field prior to the actual commencement of the survey. Questions from standard international questionnaires were used. The European Community Respiratory Health Survey (ECRHS) II[38] was used to determine respiratory symptoms. The World Health Organization's (WHO) STEPwise approach to noncommunicable disease risk factor surveillance (STEPS)[39,40] was used to determine alcohol usage and physical activity, while questions from the National Health and Nutrition Examination Survey (NHANES) 2015[41,42] was used to determine smoking habits.

Household features such as garbage disposal and cooking were determined using questions from the Sri Lankan Household Income & Expenditure Survey (HIES) 2016[43] and Demographic & Health Survey (DHS) 2016,[44,45] while demographic details were based on questions from the Sri Lankan Census of Population & Housing (CPH) 2011.[46,47] These tools and questionnaires are available for public use, such as those from NHANES, HIES, DHS, and CPH, whilst the others are available for use with appropriate citation.

Variables considered for models

We followed an evidence-based reasoning approach through literature for the initial selection of input features. A detailed description of the variables used in this study is given in Table 1 below.

Preprocessing

Initial SLHAS datasets in Stata format were retrieved and then cleansed and explored graphically. Feature engineering techniques were applied, including binning, outlier detection, one-hot encoding, and imputation of missing data. We imputed missing data columns using the Multivariate Imputation by Chained Equations (MICE) algorithm when the missing data percentage was less than 40%.[52]

Two variables are said to be highly correlated if the magnitude of the correlation coefficient is higher than 0.7.[53] Some ML models are sensitive to multicollinearity. Therefore, correlation coefficient calculation techniques like Cramer's V (for categorical features), Pearson's r (for numerical features), and correlation ratio (for mixed data) were employed to detect multicollinearity.

In our study, we employed three different dimension reduction techniques to handle correlated features for our ML models. We utilized Principal Component Analysis (PCA) for numerical features, Multiple Correspondence Analysis (MCA) for categorical features, and Factor Analysis of Mixed Data (FAMD) for mixed data. These techniques were chosen to address multicollinearity assumptions in our models by combining correlated features.

Data sampling

The model-building process of this study started with an 80% versus 20% train-test split with stratification. In tuning model hyperparameters, we adopted the grid search method with 10-fold cross-validation. Moreover, the 10-fold cross-validation technique allowed for RFE and performance score calculations.

Due to the unequal distribution between asthmatics and non-asthmatics in the dataset (known as the imbalance property), we used two categories of ML models: models built on an under-sampled dataset (mentioned as data level approach models in Table 3) and models that incorporated class weights to overcome the imbalance in the dataset, referred to as balanced models (also mentioned as algorithmic level approach models in Table 3).

Machine learning models

Logistic regression. Logistic Regression (Univariate Logistic Regression) is a statistical model that models the relationship between a dichotomous outcome variable and one or more categorical or continuous response variables, resulting in an equation to predict the outcome.[54]

Support vector machine. Support Vector Machine (SVM) finds an optimal hyperplane in the dimensional space which separates observations belonging to one class (group) from another.[55] SVMs can handle both linear and nonlinear classifications.

Decision tree. A Decision Tree follows the divide-and-conquer approach to perform classification.[56] It recursively partitions the dataset using information gain metrics until each partition consists totally or mainly of cases from only a single class.[57]

Table 1. Overview of variables.

Category	Variable	Detailed description
Outcome	*asthma_status*	Presence of asthma based on respondent self-report, doctor diagnosis, or the use of prescribed asthma medications
Respiratory symptoms[48,49]	*wheez*	The respondent experienced wheezing or whistling sounds in the chest at any time in the past year
	brthless_wheez	The respondent experienced breathlessness while wheezing occurred
	wheez_cold	The respondent experienced wheezing or whistling sounds even in the absence of a cold
	awaken_chest_tightness	The respondent woke up with a feeling of tightness in the chest at any point during the past year
	dyspnea_rest	The respondent had an attack of shortness of breath (SOB) while at rest during the day at least once in the past year
	dyspnea_strenuous_act	The respondent had an attack of SOB following strenuous activity at any time in the past year
	awaken_dyspnea_1yr	The respondent was awakened by an attack of SOB at any time in the past year
	awaken_dyspnea_3month	The respondent was awakened by an attack of SOB in the last 3 months
	avg_dyspnea_attk_1wk_3months	The respondent experienced attacks of SOB at least once a week on average in the last 3 months
	avg_dyspnea_attk_3month	The average number of times per week the respondent was awakened by SOB in the last 3 months
	awaken_cough	The respondent was awakened by an episode of coughing at any time in the past year
	nasal_allergy	The respondent has experienced nasal allergies, including hay fever
Behavioural risk factors	*smoker*	The respondent is currently a smoker of cigarettes, beedees, hand-rolled cigarettes, shisha, cigars, pipes, or e-cigarettes containing nicotine, or they have smoked them in the past
	passivesmoke	The respondent has been exposed to indirect smoke at home or workplace in the past 30 days
	alcohol	The respondent has consumed an alcoholic drink within the past year
	MET_PA	The respondent's total physical activity level, measured in metabolic equivalent-minutes (MET-minutes) per week, calculated using the global physical activity questionnaire (GPAQ) analysis guide.[50]

(continued)

Table 1. (continued)

Category	Variable	Detailed description
Household features	*garbage*	The respondent's household contributes to air pollution by burning the garbage they dispose of (The classification is taken from research conducted by Subramanian et al[51])
	light	The respondent's household adds to air pollution by primarily using kerosene or generators as their main source of lighting (The classification is taken from research conducted by Subramanian et al[51])
	cookfuel	The type of cooking fuel used by the respondent's household contributes to air pollution, categorized as high, medium, or low polluting (The classification is taken from research conducted by Subramanian et al[51])
	cookplace	The location where the respondent's household cooks: Within the house, in a separate building, or outdoors
Demographic details	*ageyrs*	The age (in years) of the respondent at the time of the survey, calculated using their date of birth
	sex	The gender of the respondent
	ethnicity	The ethnic group of the respondent
	educat	The highest level of education completed by the respondent
	sector	The residential sector of the respondent, categorized as urban, rural, or estate
	rec_lat, *rec_alt*	The precise latitude and altitude of the respondent's household location, collected using tablet PCs during household recruitment visits
	hhincome	Monthly income of the household
	rechhsize	Number of members who usually live in the household
Anthropometric measurements	*weight*	The weight of the respondent, measured using an OMRON BF511 body composition monitor to the nearest 0.1 kg.[35]
	height	The height of the respondent, measured using a seca 240 cm height measure (seca, Hamburg, Germany) to the nearest 0.1 cm.[35]
	waist	Waist circumference measurement of the respondent, collected using a seca 200 cm tape measure at the level of the natural indent of the trunk during expiration.[35]
	bmi	The body mass index (BMI) of the respondent, calculated as weight (kg) divided by the square of height (m^2) (i.e., weight/height^2).[35]
	WHR	Waist-to-hip ratio (WHR) of the respondent, calculated as waist circumference (cm) divided by hip circumference (cm) (i.e., waist/hip)
	WHtR	Waist-to-height ratio (WHtR) of the respondent, calculated as waist circumference (cm) divided by height (cm) (i.e., waist/height)

Random forest. Random Forest evolved from Decision Trees and is a combination of multiple decision trees. Each Decision Tree results in a predicted class, and the Random Forest selects the most voted class as the final prediction.[58]

Naïve bayes. The Naïve Bayes algorithm is a supervised learning probabilistic classifier that utilizes Bayes' theorem together with feature independence.[59]

K-nearest neighbour. K-Nearest Neighbour (KNN) employs distance metrics to determine a group of K-similar samples. Then the class of an unknown case is obtained using the class attributes of the nearest K neighbour.[60]

Gradient boost. Gradient Boost is an effective classifier designed to gradually convert a weak learner to a strong learner by minimizing the loss function. This loss function serves as a metric for measuring the prediction error. In essence, the objective function guides the gradient descent optimization process for this systematic reduction in the loss function.[61]

Extreme gradient boost. The Extreme Gradient Boost (XGBoost) algorithm is an efficient and flexible form of the Gradient Boost algorithm.[62] The XGBoost differs from the Gradient Boost because it introduces a regularization term to the objective, apart from the loss function.

Adaptive boost. Adaptive Boost (AdaBoost) is a boosting algorithm that converts a weak learner into a strong one. It does this by adjusting weights with no prior knowledge of the learning ability of the learner.[63]

Categorical boosting. Categorical Boosting (CatBoost) is a recently developed Gradient Boost algorithm that uses binary decision trees as the base predictor. It works exceptionally well with categorical features, resulting in the lowest information loss. Compared to other Gradient Boost algorithms, CatBoost is diverse due to its utilization of ordered boosting, its ability to be used even on small datasets, and its automatic handling of categorical features.[64]

Light gradient boosting machine. Light Gradient Boosting Machine (LightGBM) is another algorithm that adopts the gradient boosting framework. It is designed to improve computational efficiency and is thus suitable for large datasets. LightGBM differs from other tree-based models by growing trees leaf-wise instead of level-wise. This prioritizes nodes with the most significant impact on reducing loss, leading to faster training times and frequently superior predictive performance, especially with large datasets. Therefore, LightGBM is a better model when compared to other tree-based models.[65]

Multi-layer perceptron. Multi-Layer Perceptron (MLP) is a DNN with single or multiple hidden layers between input and output layers.[66] MLPs also have the characteristic of fully connected layers.

Probabilistic neural network. The Probabilistic Neural Network (PNN) belongs to the ANN group and is based on Bayes' theory. It estimates the probability density function of each class. A PNN consists of four layers: input layer, pattern layer, summary layer, and decision layer.[67]

Hybrid machine learning model. A Hybrid model combines two or more different ML modeling techniques to create a single, more powerful model.[68]

Model building

We applied multiple ML classifiers identified in the literature to identify the most appropriate model for asthma prediction. The models we selected were Logistic Regression, SVM, Decision Tree, Random Forest, Naïve Bayes, KNN, Gradient Boost, XGBoost, AdaBoost, CatBoost, LightGBM, MLP, PNN, and a hybrid version of Logistic Regression and LightGBM models (referred to as the Hybrid Model). Of the constructed models, Logistic Regression, Naïve Bayes, MLP, PNN, and Hybrid Model utilized dimension reduction techniques to address the multicollinearity assumption.

Depending on the ML classifier, we employed several feature-importance techniques in this study: coefficients as feature importance (Logistic Regression and SVM), decision tree feature importance (Decision Tree, Random Forest, Gradient Boost, XGBoost, CatBoost, and LightGBM), and permutation feature importance (Naïve Bayes, KNN, AdaBoost, MLP, and PNN).

Model evaluation

The two main model evaluation techniques incorporated in this study were the confusion matrix and the Receiver Operating Characteristic (ROC) curve. Thus, accuracy, precision, recall (sensitivity), specificity, and the F1 score were calculated for the confusion matrix, and the Area Under Curve (AUC) was calculated for the ROC curve.

Finally, the decision boundaries (decision spaces) of each classifier were visualized in a two-dimensional space.

Model deployment

We utilized a Python-based web application framework called Flask to make the model available for end-users, enabling them to use it for practical decision-making.

Results

Characteristics of participants

We excluded 171 participants with missing data for self-reported or doctor-diagnosed asthma status or lacking data to conclude asthmatic medication intake, leaving 6,494 (97.43%) participants for analysis. Their mean age was 50.05 with a 95% Confidence Interval (CI) of (49.63–50.47), with 3,313 (51.02%) being female. A more detailed view of the characteristics of the asthma sample is given in Table 2.

A significant number of individuals with asthma (~71%) in the study displayed wheezing and other respiratory symptoms. Asthmatics demonstrated lower physical activity levels, evidenced by a lower mean *MET_PA* compared to non-asthmatics (4,892.62 vs 7,018.20). Demographic characteristics, environmental exposures, and anthropometric measurements showed similarities between asthmatics and non-asthmatics.

Model comparison

The evaluation indices of all fitted models are listed in Table 3. It highlights that many ML models, those developed on imbalanced datasets (mentioned as algorithmic level approach models in Table 3) outperformed those built on under-sampled datasets (mentioned as data-level approach

Table 2. Characteristics of SLHAS asthma sample.

Variable	All (N = 6,494)	Asthmatics (N = 596)	Non-asthmatics (N = 5,898)
Respiratory symptoms			
wheez, yes	799 (12.30%)	424 (71.14%)	375 (6.36%)
brthless_wheez, yes	606 (9.33%)	364 (61.07%)	242 (4.10%)
wheez_cold, yes	382 (5.88%)	255 (42.79%)	127 (2.15%)
awaken_chest_tightness, yes	435 (6.70%)	211 (35.40%)	224 (3.80%)
dyspnea_rest, yes	349 (5.37%)	185 (31.04%)	164 (2.78%)
dyspnea_strenuous_act, yes	902 (13.89%)	341 (57.21%)	561 (9.51%)
awaken_dyspnea_1yr, yes	419 (6.45%)	200 (33.56%)	219 (3.71%)
awaken_dyspnea_3month, yes	248 (3.82%)	136 (22.82%)	112 (1.90%)
avg_dyspnea_attk_1wk_3months, yes	138 (2.13%)	93 (15.60%)	45 (0.76%)
avg_dyspnea_attk_3month	0.09 (0.05–0.13)	0.76 (0.38–1.14)	0.02 (0.01–0.03)
awaken_cough, yes	908 (13.98%)	280 (46.98%)	628 (10.65%)
nasal_allergy, yes	1,214 (18.69%)	272 (45.64%)	942 (15.97%)
Behavioural risk factors			
smoker, yes	1,433 (22.07%)	132 (22.15%)	1,301 (22.06%)
passivesmoke, yes	1,617 (24.90%)	131 (21.98%)	1,486 (25.19%)
alcohol, yes	1,700 (26.18%)	110 (18.46%)	1,590 (26.96%)
MET_PA	6,827.57 (6,623.67–7,031.47)	4,892.62 (4,424.27–5,360.97)	7,018.20 (6,799.47–7,236.93)
Household features			
garbage			
Polluting	3,001 (46.21%)	266 (44.63%)	2,735 (46.37%)
Non-polluting	3,405 (52.43%)	326 (54.70%)	3,079 (52.20%)
light			
Polluting	58 (0.89%)	7 (1.17%)	51 (0.86%)
Non-polluting	6,346 (97.72%)	585 (98.15%)	5,761 (97.68%)
cookfuel			
High-polluting	4,441 (68.39%)	385 (64.60%)	4,056 (68.77%)
Medium-polluting	70 (1.08%)	5 (0.84%)	65 (1.10%)
Low-polluting	1,887 (29.06%)	199 (33.39%)	1,688 (28.62%)
Non-polluting	5 (0.08%)	2 (0.34%)	3 (0.05%)
cookplace			
Inhouse	5,802 (89.34%)	529 (88.76%)	5,273 (89.40%)
Separate building	527 (8.12%)	51 (8.56%)	476 (8.07%)
Outdoor	140 (2.16%)	15 (2.52%)	125 (2.12%)
Demographic details			
ageyrs	50.05 (49.63–50.47)	53.95 (52.52–55.38)	49.66 (49.22–50.10)
sex, female	3,313 (51.02%)	348 (58.39%)	2,965 (50.27%)
ethnicity			
Sinhala	4,569 (70.36%)	414 (69.46%)	4,155 (70.45%)
SL Tamil	1,268 (19.53%)	123 (20.64%)	1,145 (19.41%)
Muslim	413 (6.36%)	34 (5.70%)	379 (6.43%)

(continued)

Table 2. (continued)

Variable	All (N = 6,494)	Asthmatics (N = 596)	Non-asthmatics (N = 5,898)
Indian Tamil	200 (3.08%)	23 (3.86%)	177 (3.00%)
Other	43 (0.66%)	2 (0.34%)	41 (0.70%)
educat			
No schooling	251 (3.87%)	39 (6.54%)	212 (3.59%)
Gr 1–5	908 (13.98%)	97 (16.28%)	811 (13.75%)
Gr 6-12	2,357 (36.30%)	203 (34.06%)	2,154 (36.52%)
Passed O-level	1,599 (24.62%)	142 (23.83%)	1,457 (24.70%)
Passed A-level	1,102 (16.97%)	85 (14.26%)	1,017 (17.24%)
Degree and above	265 (4.08%)	26 (4.36%)	239 (4.05%)
sector			
Urban	1,971 (30.35%)	200 (33.56%)	1,771 (30.03%)
Rural	3,567 (54.93%)	314 (52.68%)	3,253 (55.15%)
Rural/Estate	791 (12.18%)	65 (10.91%)	726 (12.31%)
Estate	165 (2.54%)	17 (2.85%)	148 (2.51%)
rechhsize	2.97 (2.94–3.00)	3.00 (2.88–3.12)	2.97 (2.93–3.01)
hhincome	47,777 (45,254–50,300)	39,223 (35,677–42,769)	48,554 (45,810–51,298)
Anthropometric measurements			
weight	59.75 (59.43–60.07)	58.17 (57.17–59.17)	59.91 (59.58–60.24)
height	158.41 (158.18–158.64)	156.32 (155.62–157.02)	158.62 (158.38–158.86)
waist	85.66 (85.36–85.96)	86.43 (85.42–87.44)	85.58 (85.27–85.89)
bmi	23.77 (23.66–23.88)	23.8 (23.41–24.19)	23.77 (23.65–23.89)
WHR	0.94 (0.94–0.94)	0.95 (0.94–0.96)	0.94 (0.94–0.94)
WHtR	0.54 (0.54–0.54)	0.55 (0.54–0.56)	0.54 (0.54–0.54)

models in Table 3). The balanced LightGBM model and the balanced logistic regression model returned the first and second-highest sensitivity of all individual models (excluding the Hybrid model) (as denoted by the [a] symbol in Table 3).

This implies that balanced LightGBM and balanced Logistic Regression are the first and second-best individual models for accurately identifying actual asthma patients (Table 3). This led us to develop a model combining the balanced LightGBM and the balanced Logistic Regression, which we called the Hybrid model. The Hybrid model recorded the highest value of sensitivity (recall) and AUC of all the models we had constructed and was selected as the best model, excelling in terms of sensitivity, AUC, and the number of features utilized (as elaborated in the section "Optimum Features").

Considering the number of models we experimented with, only those demonstrating high evaluation scores are presented in the forthcoming sections.

Feature importance

Figure 2 illustrates the importance of individual features in a few selected models. In general, all ML models indicate that input variables related to wheezing are noticeable in their ability to predict

Table 3. Evaluation indices of fitted ML models.

Model	Accuracy	Precision	Recall	F1 score	AUC
Data level approach					
Logistic regression	93.53	71.95	52.88	59.39	0.8841
SVM	93.23	66.40	53.71	58.81	0.8945
Decision tree	92.15	55.39	76.36	63.99	0.8505
Random forest	93.07	67.25	46.14	54.16	0.8997
Naive bayes	89.53	46.82	69.02	55.17	0.8786
KNN	90.30	50.36	69.77	57.74	0.8954
Gradient boost	92.99	64.01	55.38	59.05	0.8988
XGBoost	93.07	66.08	50.53	56.95	0.9040
AdaBoost	93.23	64.31	59.70	61.62	0.8726
CatBoost	93.07	66.43	49.62	56.39	0.9028
LightGBM	92.61	64.59	46.29	53.51	0.9008
MLP	92.99	66.97	49.55	55.68	0.8665
PNN	84.91	31.89	54.62	39.91	0.7666
Algorithmic level approach					
Logistic regression (balanced)[a]	90.38	48.37	77.35	59.42	0.8890
SVM (balanced)	92.15	55.39	76.36	63.99	0.8505
Decision tree (balanced)	92.15	55.39	76.36	63.99	0.8505
Random forest (balanced)	92.15	55.39	76.36	63.99	0.8505
Gradient boost (balanced)	93.46	67.19	56.21	60.76	0.9013
XGBoost (balanced)	91.99	54.75	76.36	63.60	0.8546
AdaBoost (balanced)	93.23	67.03	54.62	59.57	0.8979
CatBoost (balanced)	92.15	55.39	76.36	63.99	0.8505
LightGBM (balanced)[a]	91.76	53.99	78.11	63.62	0.9060
Hybrid model[b]	90.15	48.76	79.85	60.32	0.9062

[a]Signifies individual models with high recall and AUC values.
[b]Denotes the model with the highest recall and AUC values among all models.

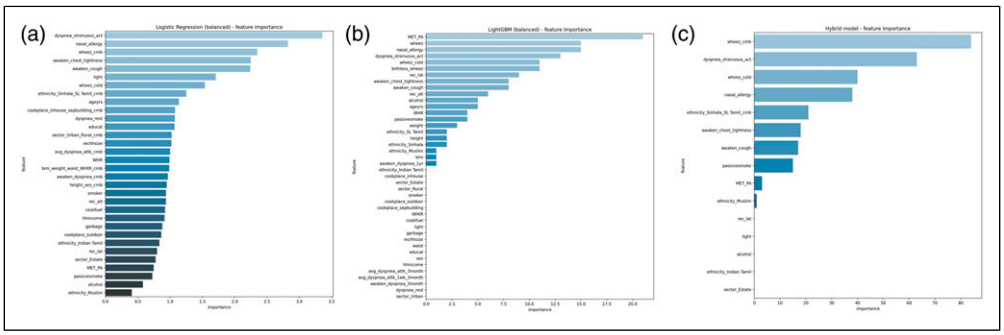


Figure 2. Feature importance of selected models: (a) balanced Logistic Regression model; (b) balanced LightGBM model; (c) Hybrid model.

asthma. In conjunction with wheezing, the occurrence of SOB episodes following strenuous activity, and the presence of nasal allergies such as hay fever, emerge as factors related to having asthma, as indicated by feature importance scores from models exhibiting high accuracy rates. Notably, among behavioural characteristics, engaging in less physical activity appears to be associated with asthma.

Optimum features

By utilizing feature importance and cross-validated recall, the RFE algorithm provided the optimal number of input features and identified which ones should be utilized. Figure 3 presents the number of input features required to achieve the highest cross-validated recall in a set of selected models. We considered the initial set of features that yield the highest cross-validated recall as the optimal number for each ML model. Consequently, the balanced Logistic Regression model yielded the best outcomes with the first 15 features, the balanced LightGBM model with the first 30 features, and the hybrid model with the first 12 features. Thus, among the models showcasing the highest accuracy levels, the Hybrid model stood out for requiring the fewest optimal features.

ROC curves

Figure 4 displays the ROC curves of the models, showcasing high AUC values. These curves visually depict the sensitivity and specificity of the models. Points closer to the top-left corner of the curve indicate better suitability and a closer approximation to the model's ideal state. All three models demonstrated AUC values near 0.90. However, the Hybrid model outperformed the others by achieving the highest AUC while utilizing the least number of features.

Decision space

The decision space, depicted in Figure 5, visually illustrates how a set of fitted models delineate the feature space into different classes, accompanied by the AUC metric in the bottom right corner. Each model's prediction process and division of the feature space are showcased. The X and Y axes represent input features, reduced to two dimensions through dimension reduction algorithms. Red dots denote asthmatic patients, while blue dots denote non-asthmatic patients. The background colours, shades of red for asthmatics and shades of blue for non-asthmatics, highlight the respective feature spaces. The decision boundary, delineated by contour lines, marks regions where the classifier assigns class labels based on input features. Notably, the Hybrid model and balanced

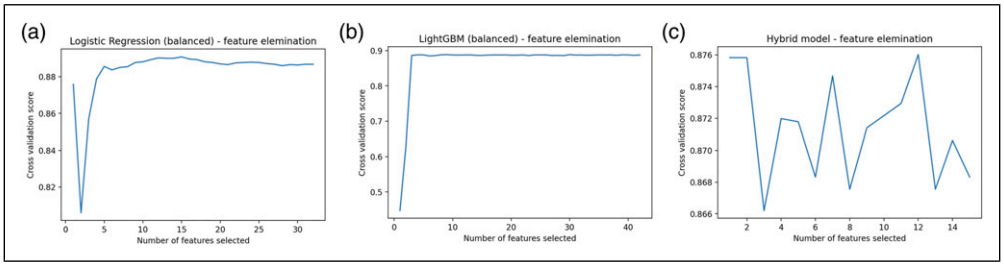


Figure 3. The optimum number of features in selected models: (a) balanced Logistic Regression model; (b) balanced LightGBM model; (c) Hybrid model.

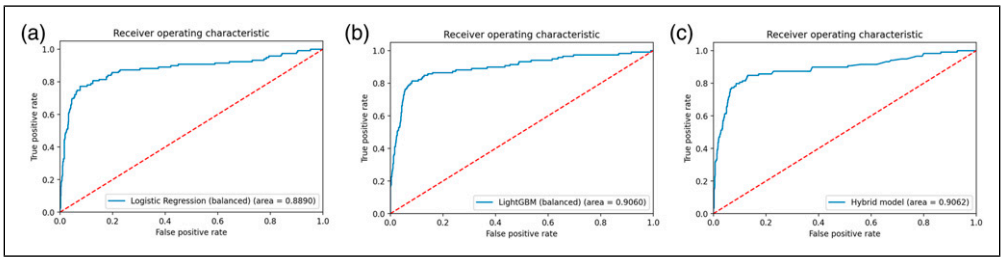


Figure 4. ROC curves of selected models: (a) balanced Logistic Regression model; (b) balanced LightGBM model; (c) Hybrid model.

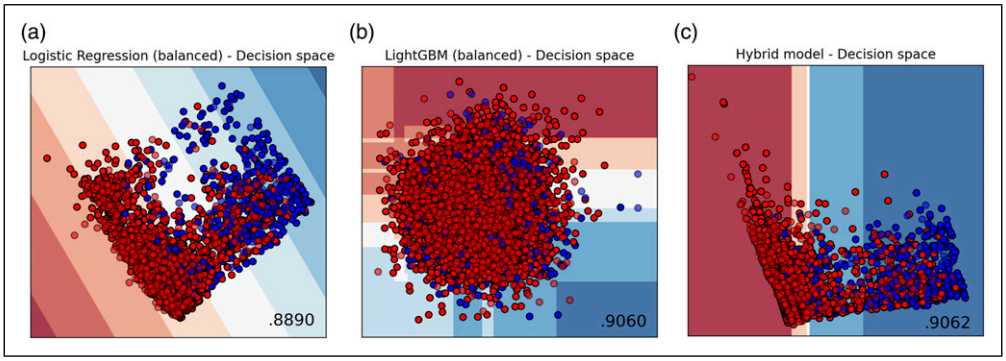


Figure 5. Decision space of selected models: (a) balanced Logistic Regression model; (b) balanced LightGBM model; (c) Hybrid model.

Logistic Regression model demonstrate clear separations between asthmatic and non-asthmatic classes, contrasting with the balanced LightGBM model, which lacks such a distinct separation.

Discussion

Discussion on fitted models

This study commenced by identifying the problem of accurate early diagnosis of asthma at affordable costs. It was crucial to avoid misdiagnosis due to its potential consequences. Hence, it became evident that a precise methodology was needed to predict asthma. As a resolution, a group of ML and statistical classifiers, along with a web application, were developed for asthma prediction.

In this paper, a total of 23 ML models were utilized to predict asthma among adults in Sri Lanka, employing a representative database collected through digital means. Different sampling techniques were employed to enhance the ML models' performance. Thirteen models were tested on under-sampled data, while ten were tested on imbalanced data. The study found that ML models trained on imbalanced data with class weights performed better in predicting asthma compared to those trained on under-sampled data. These classifiers include Logistic Regression, SVM, Decision Tree, Random Forest, Naïve Bayes, KNN, Gradient Boost, XGBoost, AdaBoost, CatBoost, LightGBM, MLP, and PNN.

To determine the best model, two evaluation metrics were utilized: recall (sensitivity) and AUC. Given the low prevalence of asthma, high accuracy alone is not sufficient since correctly identifying individuals with asthma is crucial. Therefore, recall was prioritized as the primary measure to evaluate all models. Subsequently, AUC was used to assess overall model performance. This approach facilitated accounting for variations in dataset balance, including both balanced and imbalanced datasets.

Following this evaluation approach, the balanced LightGBM model emerged as the top-performing individual model, with a recall of 78.11% and an AUC of 0.9060. The second-best individual model was the balanced Logistic Regression model, with a recall of 77.35% and an AUC of 0.8890.

To further enhance model performance, a Hybrid approach was adopted. This involved using the balanced Logistic Regression model to identify the most significant features, leveraging its interpretability and effectiveness in feature selection. These selected features were then used to fit the balanced LightGBM model, combining the strengths of both models. This Hybrid model approach enabled leveraging the recall and AUC benefits of the LightGBM model while ensuring the best set of features of the Logistic Regression model were used for prediction. This resulted in the highest level of accuracy and reliability in identifying asthma cases.

This hybrid model utilized *wheez*, *brthless_wheez*, *wheez_cold*, *awaken_chest_tightness*, *dyspnea_strenuous_act*, *awaken_cough*, *nasal_allergy*, *passivesmoke*, *MET_PA*, *ethnicity*, and *sector* as input variables. These are six asthma symptom variables (wheeze, breathlessness with wheeze, wheeze in the absence of a cold, awaking with chest tightness or cough, and shortness of breath with strenuous activity), two risk factors (nasal allergies and passive smoking), two demographic factors (ethnicity and sector of residence), and one behavioural factor (total physical activity).

The Hybrid model reported an accuracy of 90.15%, indicating that the proportion of those correctly categorized by the model is 90.15%. Since the Hybrid model recorded a 79.85% sensitivity, it can correctly identify actual asthma patients with a probability of 79.85%. A specificity of 91.18% indicates that the model can accurately identify non-asthmatics with a probability of 91.18%. The AUC of 0.9062 of the model suggests that the likelihood of correctly distinguishing an asthmatic from a non-asthmatic is 90.62%.

Model deployment

The ultimate solution integrated into this research is a Flask app where users can answer a specified set of questions and obtain a prediction of having asthma. Once the app is launched, users can view the front-end question set. As users provide answers to the questionnaire presented, their responses are fed into the developed model as features to generate the prediction outcome. Figure 6 provides the complete questionnaire.

After all fields in the questionnaire are successfully filled, the user clicks the "submit" button and is shown the prediction result. The values inputted by the user are extracted and fetched to the hybrid model, which is saved as a Pickle object, to perform calculations and provide the output. A positive prediction is indicated by displaying the result in red, while a negative prediction is shown in green (Figure 7). Additionally, for greater transparency, the AUC and predicted probability are presented to the users with a brief description.

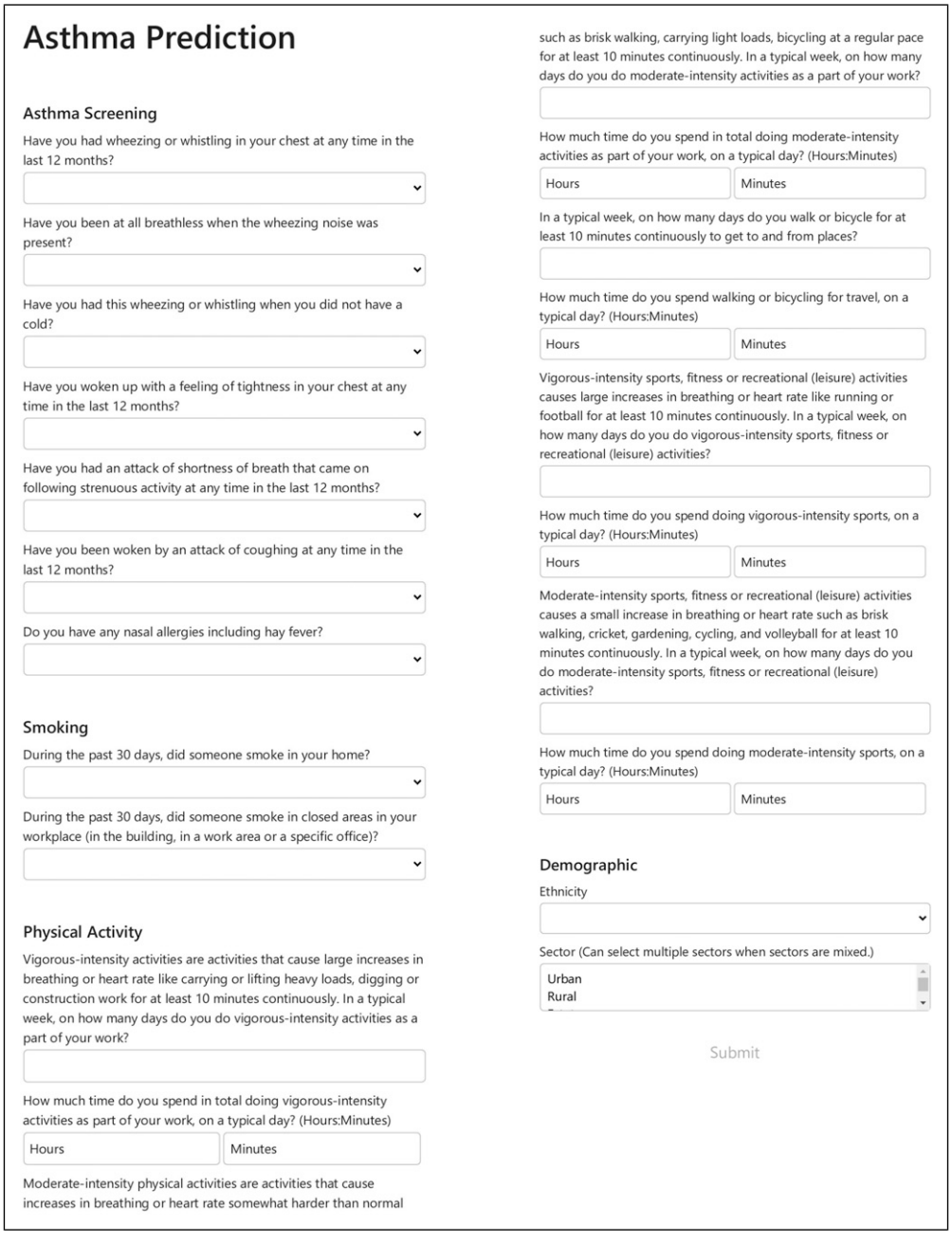


Figure 6. Front-end questionnaire for data entry by users of the Flask app.

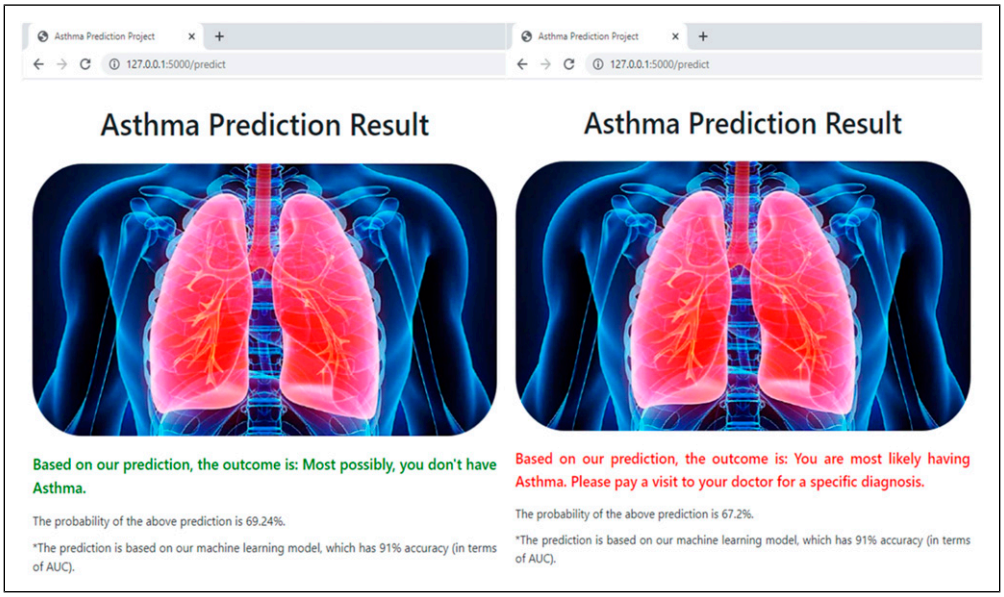


Figure 7. Prediction outcome of the Flask app.

Strengths

This research uses data from a nationally representative adult health survey, enabling a robust analysis of various factors impacting asthma. By examining multiple variables such as respiratory symptoms, behavioural risk factors, and demographic details, the study offers a comprehensive understanding of the disease rather than focusing solely on isolated factors. This approach enhances the local validity of the study results, facilitating informed decision-making.

The study differentiates itself by evaluating a large number of ML models (thirteen) on undersampled and imbalanced data to find the best predictor for asthma outcomes. A rigorous evaluation was undertaken, utilizing a variety of ML and statistical models to enhance predictive accuracy.

Unlike previous studies that often relied on biochemical findings or lung function,[16–18,20,69–72] this study prioritizes variables that do not require laboratory or lung function testing. This is important for several reasons. Firstly, it broadens the use of ML-based apps to settings with limited resources. Sri Lanka, like other low- and middle-income countries, has limited resources for testing. The recent economic crisis in Sri Lanka has further strained the health system, leading senior clinicians to advocate for greater reliance on "clinical judgment" and reduced dependence on laboratory investigations.[73] Our web application can supplement such clinical judgment as a low-cost tool that relies on self-reported symptoms and background characteristics and potentially triages patients for further investigation.

Secondly, by excluding spirometry data and anthropometric measurements in the final model, the study enhances self-detection by individuals in the community. The application allows individuals to actively participate in the prediction process by easily inputting their data and receiving personalized prediction outcomes. With an intuitive interface and easy-to-navigate forms, the application encourages users to interact with the tool, promoting proactive asthma management.

With ever-increasing access to the internet and the ready availability of mobile phone devices, such a web-based application is becoming increasingly accessible to the general public in Sri Lanka. For instance, in 2022, 52.6% of Sri Lankans had internet access and there were 1.5 mobile connections per person on average, indicating a high penetration of mobile devices.[31] In low- and middle-income countries, as many as 70% of the population were mobile users[32] and 55% had internet access in 2023.[33] This trend demonstrates the rapid growth of connectivity and the potential for digital health interventions in these regions. Furthermore, the individual-centric approach could increase individual awareness, foster engagement, and facilitate health-seeking for asthma, particularly in the Sri Lankan setting, where the prevalence of asthma is likely to be high (estimated at 23%, which is far higher than in other South Asian countries), but diagnosis and treatment rates are likely to be low, due to social stigmas associated with asthma and underdiagnosis by the healthcare system.[7]

Sri Lanka boasts over 1,000 Healthy Lifestyle Centres (HLCs) that actively screen for diabetes, hypertension, and cardiovascular disease,[74] reaching approximately 0.5 million adults aged 35 and over annually.[75] These efforts align closely with the WHO Package of Essential Noncommunicable (PEN) Disease Interventions for Primary Health Care,[76] although Sri Lanka focuses predominantly on cardiovascular disease, diabetes, and hypertension. While PEN includes guidelines for treating asthma symptoms, and Sri Lanka has established asthma treatment protocols in primary care,[77] there is no national policy for widespread asthma screening, unlike the systematic screening for cardiovascular disease, hypertension, and diabetes conducted at HLCs. Additionally, Sri Lanka benefits from frequent doctor-patient interactions and a robust public health infrastructure.[78] Public health midwives routinely visit homes with newborns, and public health inspectors conduct various community health initiatives. These existing settings and resources provide opportunities to disseminate information about this app to the public. The app could be promoted at screening centers for individuals to use independently, or even potentially be adapted into existing screening programs at resource-limited primary or secondary facilities by healthcare staff.

Limitations of the study

Using secondary data presented certain limitations, primarily due to the unavailability of crucial explanatory variables and a high rate of missing data on vital variables. Key parameters such as outdoor air pollution, history of allergic diseases beyond nasal allergies and hay fever, family history of allergic diseases, climate conditions, and exposure to pets were absent from the dataset. Additionally, due to a significant proportion of missing values, variables related to occupational exposures (~53%) and family history of asthma (~49%) had to be excluded.

Another limitation was the lack of data on children for inclusion in model development. The data was from the SLHAS survey, which focused exclusively on the health and ageing aspects of adults, and did not survey children.

The study encountered challenges related to the computational intensity and lengthy training times required for ML models. Despite achieving high AUC values, prioritizing recall over precision resulted in many models not attaining high precision values. This trade-off between recall and precision represents another limitation of the study.

Furthermore, the predictive model developed in the study has inherent limitations in predicting the outcome of physician-diagnosed asthma. The outcome variable relied on self-reported asthma or medical records/prescriptions related to asthma rather than a predetermined set of diagnostic criteria by a clinician. In some cases, this reliance on participants' recall of an asthma diagnosis or their

medical records may have led to misclassification. Nonetheless, the model predicts the outcome variable of physician-diagnosed asthma based on participants' reports or medical records.

Suggestions for future work

In future iterations, enhancing the best-fitted model could involve incorporating variables that were unavailable in the current study. Additionally, a variation could explore the integration of biomarkers and clinical measurements into the model, specifically for clinical settings.

Another avenue for improvement lies in integrating wearable sensor data from devices such as IoT devices or wearable sensors. This data could capture real-time physical activity levels, environmental exposures, and climate data, thereby enriching the model's accuracy and predictive capabilities.

It is recommended that this app be trialed in the community, with user feedback collected from both clinical professionals and the general public on various parameters. Key aspects to evaluate should include ease of use, likelihood of continued use, and user satisfaction. Additionally, it is important to assess the app's impact on health outcomes, specifically looking at parameters such as an increase in new asthma diagnoses, improved adherence to medication, and overall management of asthma symptoms. Collecting and analyzing this data will provide valuable insights into the app's effectiveness and areas for further improvement.

While this study explored many single-type ML algorithms, it primarily focused on one mixed ML model, the Hybrid model. Future work could involve comparing this Hybrid model with other mixed models to determine the most effective model in terms of prediction power.

To further enhance the user experience, steps can be taken to make the web application even more user-friendly than its current interface. Additionally, exploring options such as web hosting and the development of a mobile application could increase the accessibility and usability of the tool in future iterations.

Conclusion

A combination of Logistic Regression and LightGBM classifiers (the Hybrid model) demonstrates a 90.62% accuracy rate in predicting asthma, as measured by AUC. The study identifies wheezing or whistling sounds in the chest, breathlessness accompanied by wheezing, attacks of SOB, coughing attacks, chest tightness, nasal allergies, physical activity level, exposure to passive smoking, ethnicity, and sector of residence as the most significant factors associated with asthma. Additionally, the web application provides the opportunity to deploy the model for real-time predictions at home or in clinical settings.

We anticipate that this study will prove valuable to healthcare providers, IT service providers in healthcare, and patients within the Sri Lankan healthcare system. It may encourage possible asthmatics to seek appropriate healthcare, as well as aid in the initial screening of individuals in lower-level healthcare facilities, facilitating further specialist evaluation where necessary.

Acknowledgements

The authors thank their colleagues in the SLHAS consortium, consisting of the Institute for Health Policy, University of Colombo, University of Ruhuna and University of Peradeniya, for their input to the design of the survey tools and support of data collection, in particular Dr Renuka Jayatissa (Medical Research Institute) whose staff provided training in anthropometric measurement; colleagues in the Ministry of Health, who

facilitated the SLHAS, especially Dr S Sridharan, Deputy Director General (Planning); and Dr Anuji Gamage (Kotelawala Defence University).

Author contributions

The SLHAS survey was conceived, designed, and managed by RRE, NW, and other SLHAS collaborators. JRNAG designed and implemented the framework for the asthma prediction model. This includes data extraction, preprocessing, exploratory data analysis, feature engineering, model fitting, evaluation, and deployment. SDV provided technical advice and supervision for model implementation. JRNAG prepared the manuscript and it was reviewed and improved by NW and SDV. All authors have read and approved the final version of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Swiss Agency for Development Cooperation (SDC) and the Swiss National Science Foundation (SNSF) through the Swiss Programme for Research on Global Issues for Development (r4d programme) by the grant "Inclusive Social Protection for Chronic Health Problems" (Grant number 400640_160374), and the Institute for Health Policy Public Interest Research Fund (Grant number PIRF-2018-02).

Ethical statement

Ethical approval

This study involved human participants and was approved by the Sri Lanka Medical Association Ethical Review Committee (ERC/18-022). Participants gave informed written consent to participate in the study before taking part.

ORCID iDs

J R N A Gunawardana https://orcid.org/0000-0002-5641-9332
Ravindra P Rannan-Eliya https://orcid.org/0000-0002-5013-2816

Data availability statement

The data supporting the findings of this study are available from the SLHAS Consortium, but restrictions apply to their availability. These data were used under license for the current study and are not publicly accessible. The SLHAS Consortium, which has implemented an Open Data policy, will grant access to SLHAS Wave 1 data starting in 2024 upon application to the Consortium by interested researchers. The specific data file utilized for this paper can be acquired from the corresponding author upon reasonable request and with permission from the SLHAS Consortium.

Supplemental Material

Supplemental material for this article is available online.

References

1. National Heart, Lung and BI. *Asthma - what is asthma?* Maryland: National Heart, Lung, and Blood Institute 2022. https://www.nhlbi.nih.gov/health/asthma
2. Global Asthma Network. The global asthma report 2022. *Int J Tubercul Lung Dis*. 2022; 102: 7.
3. World Health Organization. *Asthma* 2020. https://www.who.int/news-room/q-a-detail/asthma
4. OECD. Health at a glance 2011. Paris: OECD, 2011.
5. OECD and World Health Organization. *Health at a glance: asia/pacific 2012*. Geneva: OECD and World Health Organization, 2012.
6. GBD 2019 Diseases and Injuries Collaborators. *Global burden of disease study 2019: country profile - Sri Lanka*. Washington: Institute for Health Metrics and Evaluation (IHME), 2020. https://www.healthdata.org/sri-lanka
7. Gunasekera KD, Amarasiri WADL, Undugodage UCM, et al. Prevalence of asthma and its symptoms in Sri Lankan adults. *BMC Publ Health* 2022; 22: 2330.
8. Perera C, Rannan-eliya RP, Senanayake S, et al. Public hospital inpatient discharge survey 2005. *IHP Heal Stat Reports Ser Number 1* 2009; 162: 1-209. https://www.ihp.lk/publications/docs/HSR0901.pdf
9. OECD. Health at a glance 2007. *Vasc Health Risk Manag* 2022; 18: 915–925. DOI: 10.1787/9789264105133-ko.
10. Rannan-Eliya RP, Anuranga C, Brearley L, et al. An assessment of the burden , Issues and policy options in curative care services delivery and non-communicable diseases in Sri Lanka. https://ihp.lk/publications/docs/NCD.pdf
11. Rannan-Eliya RP, Wijemanne N, Liyanage IK, et al. Quality of inpatient care in public and private hospitals in Sri Lanka. *Health Pol Plann* 2015; 30(Suppl 1): i46–i58.
12. Tsang KCH, Pinnock H, Wilson AM, et al. Application of machine learning algorithms for asthma management with mHealth : a clinical review. *J Asthma Allergy* 2022; 15: 855–873.
13. Exarchos KP, Beltsiou M, Votti C, et al. Artificial intelligence techniques in asthma : a systematic review and critical appraisal of the existing literature. *Eur Respir J* 2020; 56: 2000521. DOI: 10.1183/13993003.00521-2020.
14. Xiong S, Chen W, Jia X, et al. Machine learning for prediction of asthma exacerbations among asthmatic patients : a systematic review and meta - analysis. *BMC Pulm Med* 2023; 23: 278–315.
15. Molfino NA, Turcatel G and Riskin D. Machine learning approaches to predict asthma exacerbations : a narrative review. *Adv Ther* 2024; 41: 534–552.
16. Tomita K, Nagao R, Touge H, et al. Deep learning facilitates the diagnosis of adult asthma. *Allergol Int* 2019; 68: 456–461.
17. Spathis D and Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Inf J* 2017; 25: 811–827.
18. Gaudillo J, Rodriguez JJR, Nazareno A, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLoS One* 2019; 14: e0225574–e0225612.
19. Priya L and Priyadharshini B. The analysis of adult asthma using convolutional neural Network. *Int Res J Eng Technol* 2020; 7: 2914–2918.
20. Akbar W, Wu W-P, Faheem M, et al. Machine learning classifiers for asthma disease prediction: a practical illustration. In: *16th International Computer Conference on Wavelet Active Media Technology and Information Processing*. Chengdu, China: IEEE, 2019, pp. 143–148.
21. Bhat GS, Shankar N, Kim D, et al. Machine learning-based asthma risk prediction using IoT and smartphone applications. *IEEE Access* 2021; 9: 118708–118715.
22. Huang AA and Huang SY. Use of feature importance statistics to accurately predict asthma attacks using machine learning: a cross-sectional cohort study of the US population. *PLoS One* 2023; 18: e0288903.

23. Murad SA, Adhikary A, Muzahid AJM, et al. AI powered asthma prediction towards treatment formulation: an android app approach. *Intell Autom Soft Comput* 2022; 34: 87–103.
24. Smart Respiratory Products Ltd. Smart Asthma: Forecast Asthma - Apps on Google Play. https://play.google.com/store/apps/details?id=com.srp.spf&hl=en&gl=US
25. Smart Respiratory Products Ltd. Smart asthma: forecast asthma on the app store. https://apps.apple.com/id/app/smart-asthma-forecast-asthma/id1409692428
26. Propeller Health. Digital therapeutic platform. *Propeller Health*. https://propellerhealth.com/our-platform/
27. Amazon Web Services. Propeller health case study. https://aws.amazon.com/solutions/case-studies/propeller-health-case-study/
28. Propeller Health. *Outcomes | 150+ peer-reviewed*. California: Publications | Propeller Health. https://propellerhealth.com/outcomes/
29. Adherium Limited. Learn how the Hailie™ solution works | asthma and COPD. https://www.hailie.com/pages/hailie-app
30. KagenAir LLC. KagenAir™ - discover how weather and your environment affect your health. https://www.kagenair.com/
31. Kepios Pte. Ltd. Digital 2022: Sri Lanka — DataReportal – global digital insights. https://datareportal.com/reports/digital-2022-sri-lanka
32. International Telecommunication Union. Facts and figures 2023 - mobile phone ownership. https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-mobile-phone-ownership/
33. Statista. Internet penetration rate by country income level 2023. https://www.statista.com/statistics/1361735/internet-penetration-rate-by-country-income-level/
34. Ta AWA, Goh HL, Ang C, et al. Two Singapore public healthcare AI applications for national screening programs and other examples. *Heal Care Sci* 2022; 1: 41–57.
35. Rannan- RP, Kapuge Y, Gunawardana N, et al. Prevalence of diabetes and pre- diabetes in Sri Lanka : a new global hotspot – estimates from the Sri Lanka Health and Ageing Survey 2018/2019. *BMJ Open Diabetes Res Care*; 11: e003160. DOI: 10.1136/bmjdrc-2022-003160.
36. Rannan-Eliya RP, Wijemunige N, Perera P, et al. Prevalence and associations of hypertension in Sri Lankan adults: estimates from the SLHAS 2018–19 survey using JNC7 and ACC/AHA 2017 guidelines. *Glob Heart*; 17: 50. DOI: 10.5334/gh.1135.
37. Zerion Software. iFormBuilder | Zerion Software 2024. https://www.zerionsoftware.com/iformbuilder/
38. Burney P and Jarvis D. The European Community Respiratory Health Survey II Ecrhs II Main Questionnaire ECRHS II-interviewer administered questionnaire. https://www.ecrhs.org/_files/ugd/4aa474_815bac5d61e94be08eb8306d946a29ad.pdf
39. World Health Organization. Standard STEPS instrument. https://www.who.int/publications/m/item/standard-steps-instrument
40. World Health Organization. WHO STEPS instrument question-by-question guide (core and expanded). https://cdn.who.int/media/docs/default-source/ncds/ncd-surveillance/steps/part5.pdf?sfvrsn=c7be3ad6_5
41. Centers for Disease Control and Prevention/National Center for Health Statistics. *NHANES 2015-2016 Overview* 2015. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2015
42. Centers for Disease Control and Prevention. *SMOKING and tobacco USE-SMQ target group: SPs 0-11 years and 18+*. Atlanta: Centers for Disease Control and Prevention.
43. Department of Census and Statistics. Household income and expenditure survey 2016. https://www.statistics.gov.lk/IncomeAndExpenditure/StaticalInformation/HouseholdIncomeandExpenditureSurvey2016FinalReport
44. Department of Census and Statistics. *Demographic and Health Survey* 2016. https://www.statistics.gov.lk/Health/StaticalInformation/DHS

45. Department of Census and Statistics. Demographic and Health Survey 2016. https://www.aidsdatahub.org/sites/default/files/resource/srilanka-dhs-2016.pdf
46. Department of Census and Statistics. Census of population & housing 2011. https://www.statistics.gov.lk/Population/StaticalInformation/CPH2011
47. Department of Census and Statistics. Census of population & housing 2011. https://catalog.ihsn.org//catalog/4270/download/56158
48. Grassi M, Rezzani C, Biino G, et al. Asthma-like symptoms assessment through ECRHS screening questionnaire scoring. *J Clin Epidemiol* 2003; 56: 238–247.
49. The European community respiratory health survey II steering committee. The European community respiratory health survey II. *Eur Respir J* 2002; 20: 1071–1079.
50. World Health Organization. Global physical activity questionnaire (GPAQ) - analysis guide 2002. https://www.who.int/ncds/surveillance/steps/resources/GPAQ_Analysis_Guide.pdf
51. Subramanian SV, Ackerson LK, Subramanyam MA, et al. Domestic violence is associated with adult and childhood asthma prevalence in India. *Int J Epidemiol* 2007; 36: 569–579.
52. Azur MJ, Stuart EA, Frangakis C, et al. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; 20: 40–49.
53. Mukaka MM. Statistics Corner : a guide to appropriate use of Correlation coefficient in medical research. *Malawi Med J* 2012; 24: 69–71.
54. Digangi EA and Hefner JT. Ancestry estimation. In: DiGangi EA and Moore MK (eds). *Research Methods in Human Skeletal Biology*. Cambridge: Academic Press, 2012, pp. 117–149.
55. Pisner DA and Schnyer DM. Support vector machine. In: Mechelli A and Vieira S (eds). *Machine Learning*. Cambridge: Academic Press, 2020, pp. 101–121.
56. Myles AJ, Feudale RN, Liu Y, et al. An introduction to decision tree modeling. *J Chemom* 2004; 18: 275–285.
57. Omurca Sİ, Ekinci E, Çakmak B, et al. Using machine learning approaches for prediction of the types of asthmatic allergy across the Turkey. *Data Sci Appl* 2019; 2: 8–12.
58. Mao W and Wang F-Y. Cultural modeling for behavior analysis and prediction. In: Mao W and Wang F-Y (eds). *New Advances in Intelligence and Security Informatics*. Cambridge: Academic Press, 2012, pp. 91–102.
59. Anastasiou A, Kocsis O and Moustakas K. Exploring machine learning for monitoring and predicting severe asthma exacerbations. In: *10th Hellenic Conference on Artificial Intelligence*. New York: Association for Computing Machinery, 2018, pp. 1–6.
60. Noi PT and Kappas M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors* 2018; 18: 1–20.
61. Boehmke B and Greenwell B. Gradient boosting. *Hands-On Machine Learning with R*. Boca Raton: CRC Press, 2019. DOI: 10.1201/9780367816377.
62. Patrous ZS. *Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors*. Stockholm: KTH Royal Institute of Technology, 2018. https://www.diva-portal.org/smash/get/diva2:1240595/FULLTEXT01.pdf
63. Chengsheng TU, Huacheng LIU and Bing XU. AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*. France: EDP Sciences, 2018, pp. 1–6.
64. Ben S, Gharib C, Mefteh-wali S, et al. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technol Forecast Soc Change* 2021; 166: 1–19.
65. Liang W, Luo S, Zhao G, et al. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics* 2020; 8: 765–817.

66. Carlsson L. *Using multilayer perceptrons as means to predict the end-point temperature in an electric arc furnace*. Stockholm: KTH Royal Institute of Technology, 2015. https://www.diva-portal.org/smash/get/diva2:904219/FULLTEXT01.pdf.
67. Rehman ZU, Mirza MT, Khan A, et al. Predicting G-protein-coupled receptors families using different physiochemical properties and pseudo amino acid composition. In: Conn PM (ed). *Methods in Enzymology*. Cambridge: Academic Press, 2021, pp. 61–79.
68. Zhou F, Fan H, Liu Y, et al. Hybrid model of machine learning method and empirical method for rate of penetration prediction based on data similarity. *Appl Sci*; 13: 5870. DOI: 10.3390/app13105870.
69. Emanet N, Öz HR, Bayram N, et al. A comparative analysis of machine learning methods for classification type decision problems in healthcare. *Decis Anal* 2014; 1: 6–20.
70. Tomita K, Sano H, Chiba Y, et al. A scoring algorithm for predicting the presence of adult asthma: a prospective derivation study. *Prim Care Respir J* 2013; 22: 51–58.
71. Chatzimichail E, Paraskakis E and Rigas A. Predicting asthma outcome using partial least square regression and artificial neural networks. *Adv Artif Intell* 2013; 2013: 1–7. DOI: 10.1155/2013/435321.
72. Princy JC and Sivaranjani K. Survey on asthma prediction using classification technique. *Int J Comput Sci Mobile Comput* 2016; 5: 515–518.
73. Matthias AT and Jayasinghe S. Worsening economic crisis in Sri Lanka: impacts on health. *Lancet Global Health* 2022; 10: e959.
74. Mallawaarachchi DSV, Wickremasinghe SC, Somatunga LC, et al. Healthy Lifestyle Centres: a service for screening noncommunicable diseases through primary health-care institutions in Sri Lanka. *WHO South-East Asia J public Heal* 2016; 5: 89–95.
75. Medical Statistics Unit, Ministry of Health. *2020 annual health bulletin Ministry of healthh Sri Lanka*. Sri Lanka: Medical Statistics Unit, Ministry of Health. 2020; 1–270.
76. World Health Organization. WHO package of essential noncommunicable (PEN) disease interventions for primary health care. 2020. https://iris.who.int/handle/10665/334186
77. Non Communicable Disease Unit, Ministry of Health N and IM. Management of common non communicable chronic respiratory diseases guidelines for primary health care providers non communicable disease unit Ministry of health, nutrition and indigenous medicine. 2018. https://www.health.gov.lk/
78. National UHC dynamics card Sri Lanka | P4H Network, https://web.archive.org/web/20221223153517/https://p4h.world/en/national_uhc_dynamics_card_sri-lanka