Current existing systems

In a research carried out by, jGaudilo (2019) explored the use of machine learning algorithms for predicting individual susceptibility to asthma using single nucleotide polymorphism (SNP) data. The study employed random forest (RF) and recursive feature elimination (RFE) algorithms for feature selection and K-nearest neighbor (kNN) and support vector machine (SVM) algorithms for classification. The results indicated that RF outperformed RFE in selecting relevant features associated with asthma, and the RF-SVM model achieved the highest accuracy (62.5%), precision (65.3%), and sensitivity (69%) compared to other models. The RF-SVM and RF-kNN models also achieved AUC scores of 0.64 and 0.62, respectively. The study faced challenges in obtaining a sufficient sample size and identifying biological variables associated with complex diseases. SNP data was obtained from public repositories, and the RF algorithm was used for feature selection due to its robustness and ability to measure the predictive importance of each feature. Overall, the study demonstrates the potential of machine learning approaches for SNP-based prediction of asthma susceptibility.

Joseph Finkelstein (2019) aimed to explore the use of telemonitoring data to build machine learning algorithms for predicting asthma exacerbations before they occur. The study involved the preparation of stratified training datasets, predictive feature selection, and the evaluation of resulting classifiers. The study utilized a naive Bayesian classifier (NB) with a sensitivity of 0.80, specificity of 0.77, and accuracy of 0.778, an adaptive Bayesian network with a sensitivity of 1.00, specificity of 1.00, and accuracy of 1.00, and support vector machines (SVM) with a sensitivity of 0.84, specificity of 0.80, and accuracy of 0.80. The adaptive Bayesian classifier was based on Bayesian networks and the models were evaluated using the MDL principle. The SVM algorithm used a subset of training data as support vectors that provided the greatest separation between classes, which made them useful in instances with nonlinear class boundaries and avoided the problem of overfitting. The study concluded that machine learning has significant potential in developing personalized decision support for chronic disease telemonitoring systems. However, the dataset’s distribution was highly skewed, presenting a challenge in machine learning with a two-class problem. To address this issue, the dataset was rearranged for three experiments, resulting in better prediction results. Additionally, the small sample size and class imbalance problem were addressed by using two stratified samples with approximately equal target class distribution.

A study by Dilini M. Kothalawala (2021) aimed to develop machine learning-based prediction models for childhood asthma. The study identified that it is difficult to predict which children will develop asthma in the future, and machine learning approaches may provide better predictive performance and generalisability than existing models. Two models were developed: Childhood Asthma Prediction in Early life (CAPE Model) and Childhood Asthma Prediction at Preschool age (CAAP Model). Recursive Feature Elimination (RFE) was used to identify the optimal subset of features predictive of school-age asthma for each model. Seven state-of-the-art machine learning classification algorithms were employed to develop prognostic models. The performance of the models was evaluated on a test set using fivefold cross-validation, imputation, and resampling. The models were further externally validated in the Manchester Asthma and Allergy Study (MAAS) cohort. The study found that the SVM algorithm provided the best performance for both the CAPE and CAAP models. The models demonstrated good generalisability across multiple time-points and excellent sensitivity to predict a subgroup of individuals with persistent wheeze. For the CAPE model, the area under the receiver operating characteristic curve (AUC) was 0.71 in the test set, and the model demonstrated good generalizability in the Manchester Asthma and Allergy Study (MAAS) cohort, with AUCs of 0.71 at 8 years and 0.71 at 11 years.For the CAPP model, the AUC was 0.82 in the test set, and the model demonstrated good generalizability in the MAAS cohort, with AUCs of 0.83 at 8 years and 0.79 at 11 years.Performance measures in the IOWBC test set and MAAS were evaluated at thresholds of 0.42 (CAPE model) and 0.73 (CAPP model).Overall, the study reported that the machine learning models offered improved predictive performance over current regression-based methods and demonstrated good generalizability to predict school-age asthma across multiple time points without degrading the predictive power to rule in asthma.

As reported by Dohyeong Kim (2020) who aimed to explore the feasibility of predicting asthma attacks by analyzing the effects of indoor PM concentrations on the peak expiratory flow rates (PEFR) of asthmatic children. The study involved collecting self-reported PEFR results from 16 pediatric asthma patients, measured twice a day between September 2019 and August 2022 The study used a continuous measurement device with low-cost monitoring sensors installed in each patient's residence during the same period when the PEFR data were collected. The matched data was analyzed to obtain guidance for predicting modeling. The study used statistical association between PM concentrations and PEFR classifications over time to predict the probability of each risk category. The study used Long Short-Term Memory (LSTM) deep learning network, which is a variant of recurrent neural network (RNN) that is designed to overcome the error back-flow (vanishing gradient) problems, through the use of memory cells and several gates, with each of these components being associated with a particular aspect of learning. The Python scikit-learn package was used to implement MNL regression, while TensorFlow was used for the LSTM modeling. The study used 67% of the matched data as a training dataset for model training, and the remaining 33% of the matched dataset was used as a test database.

The study conducted by Seyed Vahid (2021) aimed to identify asthma-prone areas in Tehran, Iran by considering environmental and spatial factors. The study used a machine learning model, particularly a random forest (RF) model, to model and validate asthma-prone areas. They used 70% of the locations of children with asthma for modeling and 30% for validation. The study found that the distance to parks and streets, as well as PM 2.5 and PM 10, had the greatest impact on asthma occurrence in the study area. The RF model showed good accuracy with an area under the curve of 0.987 and 0.921 for training and testing data, respectively, based on the receiver operating characteristic results.

As observed in the case study by Tsang et al. (2020) he employed several well-known supervised learning algorithms for classification, including decision trees, logistic regression, naive Bayes, and support vector machine (SVM), to differentiate stable and unstable periods in asthma self-management. For feature selection, they used the least absolute shrinkage and selection operator (LASSO) method to rank and select a handful of input features. The regularisation performed by LASSO not only helped to avoid over-fitting but also kept ranking input features based on their predictive power.The decision trees algorithm had an accuracy of 0.716, sensitivity of 0.683, and specificity of 0.766. The SVM algorithm had an accuracy of 0.638, sensitivity of 0.591, and specificity of 0.620. The logistic regression classifier had an accuracy of 0.748, sensitivity of 0.737, and specificity of 0.762. The naive Bayes-based classifier provided high accuracy (AUC > 0.87), sensitivity of 0.866, and specificity of 0.725.To compare the performance of the different classifiers, they used the area under the receiver operating characteristic curves (AUC-ROC), which is a standard comparison metric for binary classification. The ROC curve reflects the obtainable balance between sensitivity (true positive rate (TPR), the proportion of unstable periods correctly classified) and specificity (true negative rate (TNR), the proportion of stable periods correctly classified). The ROC curve was dependent on the training-test set segmentation for cross-validation. The SVM model varied most over the different validation sets. The median performance figures of 500 evaluations.Overall, their study demonstrated the potential of machine learning algorithms to provide real-time feedback and tailored advice to patients based on their monitoring to enhance asthma self-management. The high accuracy of the logistic regression and naive Bayes-based classifiers suggests that they may be useful in developing early warning algorithms for unstable asthma events.

This study examined the use of machine learning techniques in the diagnosis of asthma and chronic obstructive pulmonary disease (COPD). The study used different machine learning algorithms for classification, including auto-associative memory neural networks (AMNN), Bayesian networks (BN), ID3, C4.5, support vector machine (SVM), Gaussian process, random forest, naive Bayes, and logistic regression.In one study, logistic regression was used to predict the development of asthma in a sample of 1226 children, with an AUC score of 0.74. Another study compared different classifiers, including artificial neural network (ANN), SVM, Gaussian process, and expert-based BN, in the diagnosis of asthma and COPD. The expert-based BN achieved the highest accuracy, while the classifiers performed slightly worse.The data used in another study were obtained from the UK General Practice Research Database, which included patients diagnosed with COPD (n=2,699) and asthma (n=7,931). Random forest achieved the highest precision for both asthma (80.3%) and COPD (97.7%) using 10-fold cross-validation. Naive Bayes (73.4%) and logistic regression (72.7%) were the next best performers for asthma using 10-fold cross-validation, while naive Bayes (96.9%) and logistic regression (96.2%) were the next best performers for COPD using 10-fold cross-validation.In both cases, the 10-fold cross-validation technique performed better, except for logistic regression in asthma and neural network in COPD, where the 5-fold cross-validation performed better. Overall, the findings of these studies suggest that machine learning algorithms can be used to diagnose asthma and COPD with high accuracy, and that different algorithms may perform better for different types of data and classification tasks.

In another research conducted by Eskezeia et al(2022) concerning the development and validation of asthma risk prediction model using co-expression gene modules and machine learning methods, Their study was based on multiple genes that may affect asthma, they found out that by identifying differently co-expressed genes followed by functional annotation can inform their understanding of the molecular mechanisms in asthma pathogeniesis. In their study they used the airway epithelial cells(AECs) and nasal epithelial cells (NECs) dataset and they implemented weighted gene co-expression network analysis(WGCNA) and machine learning (ML) techniques to develop asthma classification and predicative models. The model was evaluated using external bronchial epithelial cells(BECs), airway smooth muscle(ASM) and whole blood(WB) dataset. AECs area under curve achieved (AUC=0.90) and NECs(AUC = 0.99). They further validated AECs derived DEGs in BECs (AUC= 0.96), ASM (AUC= 0.72) and WB (AUC= 0.67). Similarly, NECs derived DEGs in BECs (AUC= 0.88), ASM (AUC= 0.87) and WB (AUC= 0.68). Both Both AECs and NECs based gene-signatures showed a strong diagnostic performance with high sensitivity and specificity.They used four machine leaning algorithm namely: Random Forest achieving best perfomance AUCcriteria(AUC=0.90), recursive Feature Elimination (RFE), least absolute shrinkage and selection operator(Lasso) and Boruta.In addition, the diagnostic performance of WGRF-RF method achieved MCC = 0.66 and F-score = 0.86. In the evaluation, using BECs, NECs, ASM and WB datasets, then diagnostic model derived from AECs exhibited a performance with F-measure of 0.93, 0.83, 0.82 and 0.43 respectively.

In another study,deep learning approaches for sleep disorder and prediction in an asthma cohort carried out by Dinh-van et al(2020). They examined samples included in the Taiwan National Health Insurance Research Database(NHIRD). The disease histories of the asthma patients were transferred to sequences and matrices for the prediction of sleep disorder by applying machine learning(ML) algorithms, including K-Nearest Neighbors(KNN), Support Vector Machine (SVM), and Random Forest(RF), and deep learning(DL) models including Recurrent Neural Network(RNN), Long Short-Term Memory (LSTM), Gated Recurrent Units(GRU), and convolution Neural Network(CNN). From the new asthma subjects: The KNN, SVM and RF algorithms were demonstrated to be successful sleep disorder prediction models, with accuracies of 0.798, 0.793, and 0.813 (AUC: 0.737, 0.690, AND 0.719, respectively). The results of the DL models showed the accuracies of the RNN, LSTM, GRU, and CNN to be 0.744,0.815,0.782, and 0.951.The results showed that CNN model had the best performance for sleep disorder prediction in the asthma cohort.

| Author and Year | Dataset Used | ML Used and Feature Selection | Strengths | Challenges |
| --- | --- | --- | --- | --- |
| Dilini M. Kothalawala, 2021 | IOWBC, MAAS | SVM with RFE, 7 classification algorithms | Timely predictions, improved performance, good generalisability, sensitivity to subgroup | Predominantly Caucasian populations, feature selection potentially biased results, no genomic data considered |
| Dohyeong et al. 2020 | 16 pediatric asthma patients | LSTM deep learning network, MNL regression | Feasibility study, low-cost monitoring sensors, comparison between LSTM and MNL approaches | Small sample size, limited dataset used, limited features considered |
| Gaudilo et al.(2019) | OpenSNP database, 23andMe, deCODEme, and FamilyTreeDNA | K-nearest neighbor (kNN) and support vector machine (SVM), and Recursive Feature Elimination (RFE) and Random Forest (RF) for feature selection | RF outperformed other feature selection methods in terms of robustness | Obtaining a sufficient sample size to develop predictors and equipping predictors with biological variables associated with complex disease |
| Seyed Vahid, 2021 | Location data of children with asthma | Random Forest, Spatial Auto-correlation, Distance Criteria | Accurate prediction of asthma-prone areas in Tehran, Iran considering environmental and spatial factors | Lack of data on individual exposure to air pollution, Lack of consideration of indoor air pollution and other risk factors, Limited generalization beyond the study area |
| Kevin C. H. Tsang, 2020 | Asthma Mobile Health Study (AMHS) | Decision Trees, SVM, Logistic Regression, Naive Bayes | High accuracy (AUC > 0.87) in predicting stable and unstable periods in asthma patients using personalized feedback | Identification of unstable events with a weekly resolution,  Limited representativeness of the dataset due to device restriction,  Limited validation of the algorithm in real world setting,  potential bias in the data due to device restriction. |
| Pascual, J. et al. | UK General Practice Research Database | Artificial neural network (ANN), Support vector machine (SVM), Gaussian process and expert-based Bayesian networks (BN) | The study compared several classifiers and found that the expert-based BN achieved the highest accuracy for both COPD and asthma. Random Forest achieved the highest precision in predicting asthma and COPD, respectively. The study also used both 5-fold and 10-fold cross-validation techniques to evaluate the models. | The sample size was relatively large, but the study acknowledged that the performance of the classifiers could still be improved with a larger dataset. The study also highlighted the challenge of over-fitting in conventional Decision Trees, which was addressed by using Random Forest. |
| Eskezeia et al(2022) | publicly available NCBI Gene Expression Omnibus (GEO) database. | Random Forest, recursive Feature Elimination,least absolute shrinkage and selection operator, SVM | From classification performance, the random forest-based model, WGRF, was identified as robust method to select potential gene  features to improve the diagnostic performance. | However most models perform better, they predicted poorly in external validation dataset,duet to weak extrapolation possiblity resulted from overfitting problem. |
| Dinh-van et al(2020) | National Health Insurance Research Database(NHIRD) | K-Nearest Neighbors(KNN), Support Vector Machine (SVM), and Random Forest(RF), and deep learning(DL) RNN, LSTM, GRU CNN | CNN model had the best performance for sleep disorder prediction in the asthma cohort | Underfitting, Insufficient model complexity and potential bias in the data |

Challenges Facing Machine Learning Algorithms from the current systems.

Despite the potential benefits of machine learning algorithms, there are several challenges that are currently being faced in predicting Asthma.

Over fitting: Over fitting is a common problem in machine learning when a model is trained too well on the training data and performs poorly on new, unseen data. This can lead to inaccurate predictions on asthma.

Model complexity: The complexity of the machine learning model used can also affect the accuracy of predictions. Simple models may not capture all the relevant factors that contribute to student dropouts, while complex models may be difficult to interpret and may not generalize well to other contexts.

As per their study they acknowledge that the size of the sample is not considered big for them to be able to generalise their predictions. However the variety of the attributes/ features allowed them to describe each patient of their dataset thoroughly. kevin C. H. Tsang (2020)

Another challenge is the implementation of predictive technologies that do not require invasive medicine or expensive devices; spirometry in their case.kevin C. H. Tsang (2020)

A limitation in the data was the identification of the unstable event, using the weekly survey, with the resolution of a n unstable event being a week. Also the app was limited to Apple device users, who are not representative of the whole US population, thus introducing potential bias in the data.( kevin C. H. Tsang,2020)