# Deployment of a Phenotypic Characterization System for Effective Identification of the Onset of Asthma Disease

Pooja M. R[1], Vinayakumar Ravi[2,*], Alanoud Al Mazroa[3] and Pradeep Ravi[4]

[1]*Department of Computer Science & Engineering Vidyavardhaka College of Engineering, Mysuru, Karnataka 57002*

[2]*Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia*

[3]*Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), P.O. Box 84428, Riyadh 11671, Saudi Arabia*

[4]*Department of Computer Science and Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India*

**Abstract:**

*Background:* Essentially, machine learning techniques help with clinical decision-making by forecasting prediction results based on recent and historical data, which are frequently found in carefully chosen clinical data repositories. In order to uncover hidden patterns in the data, machine learning applies sophisticated analytical techniques that conduct an exploratory analysis while constructing prediction models to support clinical judgment.

*Objective:* To effectively identify asthmatics in two distinct cohorts representing India's rural and urban populations by adopting a phenotypic characterization approach.

*Methods:* Cross-sectional and categorical in design, the data represent the two populations, with clinical history information emphasizing clinical symptoms and patterns defining the condition. The method adopts a hybrid approach since it uniquely blends the unsupervised and supervised learning techniques to explore the advantages of both. The clustering data emphasizing the phenotypic characteristics of asthma is input to the classifier, and the performance of the classifier was continuously monitored for significant improvement in the results.

*Results:* Asthma disease outcome predictions made by the hybrid decision support system were quite accurate, with classification accuracy reaching up to 85.1% and 95.3% for the two datasets, respectively.

*Conclusion:* Since asthma is a heterogeneous disease with multiple subtypes, employing clustering information in the form of cluster evaluation scores as an input parameter to the classifiers can effectively predict disease outcomes.

**Keywords:** Asthma, Correlation, Clustering, Hybrid clustering, Phenotypic, Healthcare.

*Address correspondence to this author at the Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia; E-mails: vravi@pmu.edu.sa, vinayakumarr77@gmail.com

## 1. INTRODUCTION

Asthma is one of the most common long-term pulmonary respiratory disorders. Itis recognized to be influenced by several variables, including genetics and environmental elements that are crucial to the disease's persistence and advancement [1]. When evaluated using data from electronic health records, clinical symptoms, their patterns, and their frequency significantly influence the course and impact of the disease [2]. The well-known

ISAAC developed core questionnaires on asthma and associated comorbidities, which have been well-received in most pilot studies in various countries. It has been suggested that a two-stage hybrid decision support system is created, with the first stage's outputs serving as a major input. [3-5]. Our study has a solid chance of serving as a significant intervention by facilitating early risk assessment and efficient illness prognosis through risk stratification. The disease may grow and ultimately become irreversible if it is not detected at an early stage, even though it is reversible in nature.

Clinicians and clinical researchers frequently use predictive models to identify and assess asthmatic patients' risks and weigh the advantages of preventive interventions against their costs and inconveniences [6, 7]. We present a weighted feature averaging technique for extracting the most relevant and prominent features concerning asthma outcome—the problem under the study's target.

Without any regard to the underlying illness, a phenotype refers to the distinguishable features of a disease, such as its clinical, morphologic, physiological, biochemical, and therapeutic response [8]. An individual's phenotype, which can change over time as a result of interactions between the environment and genes, includes having asthma. Classifying asthma can be challenging because there can be significant similarities between phenotypes. The best strategy to achieve better results would be to stratify asthma based on phenotypes or endotypes [9-11]. Below are other similar studies, and it has been widely observed and deduced that unsupervised learning utilizing cluster analysis has been one of the strategies used to undertake asthma phenotyping.

Predictive models and decision support systems currently in use perform poorly and with insufficient precision. There are ongoing efforts to enhance the same, and one such effort in this area was made by us when we adopted a system that effectively combined clustering and classification learning approaches to get the best prediction accuracies at an early stage [12]. An empirical investigation of the various standard learning classifiers described previously shows that the suggested solution outperforms the traditional classifiers for predicting the course of the disease. The system, therefore created, has a good sensitivity-to-specificity ratio, indicating that one may trust the constructed model to accurately forecast the condition. It can be incorporated into any asthma tools that provide self-management surveys by doing this.

By utilizing both unsupervised and supervised techniques, the decision support system's main goal is to accurately predict the course of the asthma illness. The system works in two steps: first, a target feature cluster is identified, which then determines the relevant features using a modified version of fuzzy C Means clustering; second, subject clusters are formed, with samples within each cluster sharing a common phenotype. Forecasting of illness outcome is achieved by adjusting a binary classification tree on the given data, which now includes categorical data indicative of identified subtypes of asthma created from the preliminary stage involving subject clustering.

The approach has successfully classified many asthma phenotypic groups with associated clinical risk factors, which is a crucial step in promoting the use of the appropriate medications early in the course of the disease to lessen its damaging effects. However, the clinical results in this work were mostly calculated based only on medical history.

## 1.1. Related Work

To explore asthma data and uncover latent patterns that add value, clinical decision support systems that use machine learning techniques to mine the useful data may be implemented. This would offer an additional source of greater understanding for decision-making regarding the identification of risk factors and the assessment of the disease's severity [13-15]. Also, it would be helpful to evaluate the degree to which it relates to other lung diseases like Chronic Obstructive Pulmonary Disease (COPD). [16-18].

Predictive models and decision assistance systems currently in use perform poorly with poor precision. As part of our continued efforts to improve the system, we adopted a method that successfully combined clustering and classification learning approaches to achieve the best prediction accuracies. An empirical investigation of the various standard learning classifiers described in the previous studies shows that the suggested solution outperforms the traditional classifiers for predicting the course of the disease. The system, therefore, has a good sensitivity-to-specificity ratio, indicating that one may trust the constructed model to accurately forecast the condition. This allows it to be incorporated into all asthma tools that provide self-management questionnaires. Incorporating supervised learning techniques into Bayesian classification approaches has produced classes that are primarily recognized as correct and have a good amount of face validity, as well as significant relationships to asthma, variations in airway reactivity, and lung function, which demonstrate content validity and provide a good amount of satisfaction [19-21]. In order to assess the risk factors, predictive analytics havebeen effectively used to combine independent data from varied sources [22]. The population of asthmatics who attend emergency rooms is described by rule-based systems that integrate expert-driven feature selection with rough sets techniques [23]. Depending on the underlying data type and the ensemble learning strategies used, a respectable level of precision was achieved in the job of distinguishing asthma from other respiratory illnesses. [24,25,21,26].

The work is centered on outlining and showcasing the efficient application of machine learning methods to support clinical judgments about the early identification and evaluation of asthma, a long-term respiratory condition that affects the respiratory system. The usefulness of clinical data-driven predictive models using machine learning for asthma prognosis due to efforts to enhance clinical procedures and attain better disease

outcomes, history, and findings has been steadily increasing [27]. This can be very advantageous for doctors and patients, thus cutting down on overall expenses.

Disease predictive modeling is carried out by utilizing instruments that are thought to aid in decision-making. While some concentrate on forecasting a clinical outcome, others categorize patients who might be susceptible to forming a specific medical condition. Including domain expertise in finding the clinical characteristics that are most pertinent to prediction models entails disease classification based on a prior understanding of the illness. Although physicians still employ a subjective method, machine learning techniques for predictive modeling of healthcare and medical data have become more popular recently. To uncover hidden relationships in clinical data, machine learning techniques primarily use data-driven approaches to explore the nature of the data. They then examine the data to identify intriguing trends that lead to a useful prognosis for a particular illness.

When it comes to providing the best level of precision, traditional machine learning techniques have their own limitations. The reasons for this include the presence of noise or outliers in the input data, the inability to normalize the data, the existence of redundant and irrelevant attributes, the lack of diversity in the training data, and the use of incorrect estimation techniques to determine the machine learning methods' accuracy level. Therefore, unless machine learning techniques are specifically designed to handle unknown data, using them as a solution is not a good idea. Instead, optimization techniques should be used.

## 2. MATERIAL AND METHODS

### 2.1. Dataset

It is observed that the current asthma prediction models have lower accuracy levels and are not trusted by medical professionals, thus, attempts are being made to make improvements [28]. Various subjective evaluation methods for clinical evaluation of asthma, include the score on the symptom tracker and the asthma control questionnaire. Recently, questionnaires have been used to gather firsthand information on the Probability of subjects having asthma at risk. The demographic and environ- mental information, the frequency and incidence of prognostic factors like wheezing episodes, parental history of the disease, comorbidities like hay fever and rhinitis, and symptoms like cough, wheezing, allergies, eczema, sleep and speech disturbances and their varied presentations are all shared among these questionnaires. Since there is no gold standard to characterize disease patterns, it is possible that different questionnaire data have some variation in the factors considered for the study of the disease. Therefore, we turn to the ISAAC data, which is standard clinical asthma questionnaire data that has been accepted and used in several pilot studies around the world.

The system performance was validated using response data from the ISAAC core questionnaires collected during the first phase of the study, which has been applied to multiple pilot studies [29]. Usually, the data collection questions represent delicate and particular disease indicators. The data are cross-sectional because they are analyzed with respect to a specific population, which is typically a sizable subset that is most representative at a given moment in time. Two age groups were used to collect the data: 6–7 years and 13–14 years. While the older children in one group completed the response independently, the parents in the other group completed the same task. The New Delhi dataset included information on 2961 individuals that was analyzed in relation to 47 variables that were suggestive of symptoms and comorbidities related to asthma. For 112 samples, the disease outcome was reported as unknown. Therefore, we eliminated all 112 samples, resulting in a dataset with 2849 samples. Since there were only 110 asthmatic subjects in the entire population, we attempted to address this imbalance by using stratified sampling, which allowed us to select 110 samples from the non-asthmatic group as well. This allowed us to create a balanced dataset, fed as the input data. Similarly, in order to counteract the class imbalance that would have produced 78 asthmatics, we attempted to select 2% of the total population, or 64 subjects, from the non-asthmatic group.

### 2.2. Clustering of Features and Phenotypic Characterization

Feature clustering is first used to select the most common features that can be recognized as important comorbidities of asthma disease. Unlike the traditional Fuzzy C Means technique, which uses an objective function based on distances, Modified Fuzzy C Means Clustering (MFCM) uses an objective function based on correlation to cluster features. Based on subtractive clustering evaluation results, we decided to divide the input feature set into four clusters. The cluster containing the asthma attribute extracts the features that coexist with the asthma attribute. As a result, the process's feature set is diminished.

Fuzzy C Mean Clustering uses an objective function that is based on correlation. Here, the distance metric includes a Pearson correlation coefficient, which shows how different two features are from one another or from a cluster. A positive or negative correlation can be indicated by the direct or inverse correlation between features. All related features will be grouped together regardless of how they are expressed. The definition of the distance metrics is $d_{ij} = 1 - \rho^2 X_i, C_j$, where $\rho X_i, C_j$ is the Pearson measure of the correlation between a feature $(x_i)$ and a cluster $(c_j)$. In the event of a high correlation—positive or negative—the correlation would be 0.

To generate silhouette scores, we employed K means clustering, and the number of clusters used to indicate the "K" parameter was adjusted from 2 to 8. With the naive data, the silhouette score was high for two clusters, and with the New Delhi data, it was high for eight clusters. In Fig. (**1a** and **b**), the distribution patterns of asthmatics and non-asthmatics in the various clusters are clearly shown. With the decreased feature set, the participants in the two datasets were grouped into two and eight clusters, respectively. After that, a classification and regression tree was used to predict people with and without asthma.
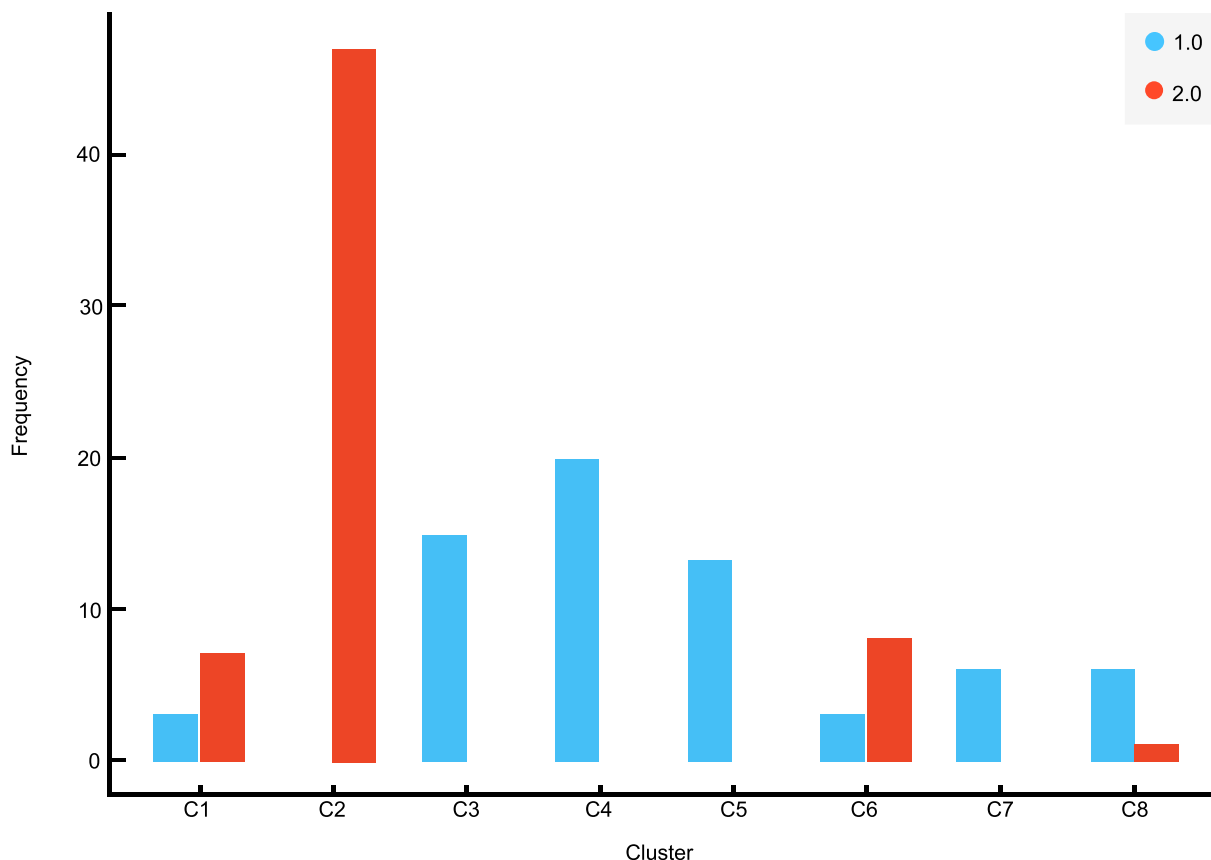
**Fig. (1).** Sample distribution of subjects within the two groups.

As shown in Tables **1** and **2**, both random initialization and K means ++ are used to initialize the cluster centroids for K-means clustering in order to establish a strong basis for the number of clusters chosen. Both initialization methods produced the same result, indicating agreement on selecting two and eight clusters in Naively and New Delhi data, respectively. The silhouette score for Naively data was 0.325 using both methods, and silhouette scores for New Delhi data were 0.384 and 0.377 (highest) using random initialization and K-means++, respectively. The hybrid decision support system predicts the course of an illness.

**Table 1. Cluster evaluation for rural asthma.**

| Initialization Method | Clusters | Silhouette Score |
|---|---|---|
| Dataset | | Naively data |
| | 2 | 0.325 |
| | 3 | 0.236 |
| | 4 | 0.217 |
| Random Initialization | 5 | 0.220 |
| | 6 | 0.194 |
| | 7 | 0.204 |
| | 8 | 0.217 |
| | 2 | 0.325 |
| | 3 | 0.236 |
| | 4 | 0.217 |
| K-Means++ | 5 | 0.216 |
| | 6 | 0.206 |
| | 7 | 0.219 |
| | 8 | 0.232 |

**Table 2. Cluster evaluation for urban asthma.**

| Dataset | New Delhi Data | |
|---|---|---|
| Random initialization | 2 | 0.384 |
| | 3 | 0.376 |
| | 4 | 0.409 |
| | 5 | 0.422 |
| | 6 | 0.442 |
| | 7 | 0.441 |
| | 8 | 0.464 |
| K-Means++ | 2 | 0.384 |
| | 3 | 0.377 |
| | 4 | 0.405 |
| | 5 | 0.414 |
| | 6 | 0.446 |
| | 7 | 0.452 |
| | 8 | 0.461 |

**Table 3. Performance evaluation of hybrid system.**

| Dataset | Method | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| Rural | 10-fold Cross Validation | 0.855 | 0.851 | 0.851 | 0.854 | 0.851 |
| Urban | | 0.977 | 0.953 | 0.953 | 0.954 | 0.953 |

Predicting the course of an illness is the hybrid decision support system's main goal. By utilizing both supervised and unsupervised techniques, asthma with a high degree of accuracy. The system is introduced in two phases, the first of which is the Target feature cluster identification, which leads to the identification of the pertinent features utilizing a modified alternative to Fuzzy C means clustering, then a subsequent stage when subject clusters are formed such that each is characterized by samples that share a common phenotype. The forecast of Fitting a binary classification tree to the input data, which now includes the illness outcome, provides category information indicating the previously known subgroups of asthma and grouping of subject-related stages.

## 3. RESULTS

Table **3** illustrates the performance of the hybrid decision system as determined by 10-fold cross-validation, stratified random sampling, and random sampling. It shows an analysis of the performance of the classifiers in comparison to other classifiers. It has been noted that the hybrid decision system outperforms the other classifiers in terms of performance.

## 4. DISCUSSION

Two testing schemes were used to assess the performance: 10-fold cross-validation and stratified random sampling, which are covered in more detail below.

By dividing the initial sample into a training set for model training and a test set for performance evaluation, the cross-validation technique is used to assess prediction models. The provided input sample is thoroughly divided into ten equal-size subsamples for 10-fold cross-validation.

Nine of the remaining "k" subsamples are used to train the model, with one sample set aside as part of the validation set for model testing. There are ten iterations in the process. One estimate can then be generated by averaging the outcomes of these ten folds. Stratified random sampling is.

In situations where the strata are assumed to be too dissimilar and too important to the problem under study and/or when investigators want to oversample a primarily minor group of interest, the chosen technique is used instead of simple random sampling. In this case, the entire dataset is divided into train and test sets.

To confirm that the findings are accurate, the procedure is repeated 100 times using 20% of the samples in the test set and 80% of the samples in the train set. According to a performance comparison between conventional classifiers and the optimized gradient boosting method, the latter method offers a higher classification accuracy, with 10-fold cross-validation yielding results of 84.5% and 91.4% for the Neyveli and New Delhi datasets, respectively.

The proposed method is seen to perform better than the conventional classifiers for an empirical investigation of the numerous traditional learning classifiers shown in the previous studies revealed the prediction of disease outcome. Thus, the system was created to ensure a strong sensitivity to specificity ratio, indicating that the model created to forecast the illness can be successfully deployed effectively. This allows it to be incorporated into any asthma tools thatprovides self-management tests or any other resources the therapist offers that are aimed at preliminary disease detection screening.

## CONCLUSION

The system proposed and developed has a good sensitivity to specificity ratio, indicating that one may trust the constructed model to accurately forecast the condition. This allows it to be incorporated into all asthma tools that provide self-management questionnaires. The performance evaluation clearly shows that the developed hybrid decision support system explores the advantages of both supervised and unsupervised learning techniques, and results obtained are quite promising for appropriate deployment in real-time scenarios. For the two datasets we examined in our study, the hybrid decision support system achieved classification accuracy as high as 85.1% and 95.3%, demonstrating its effectiveness in making accurate predictions regarding the outcome of asthma disease. Additionally, the two datasets' respective precision rates of 85.4% and 95.4% were satisfactory. Furthermore, with asthma being considered a heterogeneous disease with a wide range of symptoms, using clustering information as an input parameter to classifiers in the form of cluster evaluation scores can effectively predict disease outcomes.

## ABBREVIATIONS

COPD = Chronic Obstructive Pulmonary Disease

MFCM = Modified Fuzzy C Means Clustering

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

## HUMAN AND ANIMAL RIGHTS

Not applicable.

## CONSENT FOR PUBLICATION

Not applicable.

## STANDARDS OF REPORTING

STROBE guidelines were followed.

## AVAILABILITY OF DATA AND MATERIALS

All the data and supporting information is provided within the article.

## FUNDING

None.

## CONFLICT OF INTEREST

There is no conflict of interest for this study.

## ACKNOWLEDGEMENTS

Declared none.

## REFERENCES

[1] Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. Acad Emerg Med 2018; 25(12): 1463-70.
http://dx.doi.org/10.1111/acem.13655 PMID: 30382605

[2] Lai CKW, Beasley R, Crane J, Foliaki S, Shah J, Weiland S. Global variation in the prevalence and severity of asthma symptoms: Phase three of the international study of asthma and allergies in childhood (ISAAC). Thorax 2009; 64(6): 476-83.
http://dx.doi.org/10.1136/thx.2008.106609 PMID: 19237391

[3] Chakraborty C, Mitra T, Mukherjee A, Ray AK. CAIDSA: Computer-aided intelligent diagnostic system for bronchial asthma. Expert Syst Appl 2009; 36(3): 4958-66.
http://dx.doi.org/10.1016/j.eswa.2008.06.025

[4] Turcatel G. Machine learning models to predict asthma exacerbations. CHEST 2023; 164(4): A53.
http://dx.doi.org/10.1016/j.chest.2023.07.092

[5] Lovrić M. Predicting treatment outcomes using explainable machine learning in children with asthma. Children 2021; 8(5): 376.

[6] Rohankumar V. A smart health care ecosystem. Recent Trends in Computational Sciences. CRC Press 2024; pp. 339-45.

[7] Louis G, Schleich F, Guillaume M, et al. Development and validation of a predictive model combining patient-reported outcome measures, spirometry and exhaled nitric oxide fraction for asthma diagnosis. ERJ Open Res 2023; 9(1): 00451-2022.
http://dx.doi.org/10.1183/23120541.00451-2022 PMID: 36755965

[8] Shivade C, Raghavan P, Lussier FE, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014; 21(2): 221-30.
http://dx.doi.org/10.1136/amiajnl-2013-001935 PMID: 24201027

[9] Prosperi MCF, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med 2013; 188(11): 1303-12.
http://dx.doi.org/10.1164/rccm.201304-0694OC PMID: 24180417

[10] Lee CH, Chen JCY, Tseng VS. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. Comput Methods Programs Biomed 2011; 101(1): 44-61.
http://dx.doi.org/10.1016/j.cmpb.2010.04.016 PMID: 20554074

[11] Moore WC, Meyers DA, Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med 2010; 181(4): 315-23.
http://dx.doi.org/10.1164/rccm.200906-0896OC PMID: 19892860

[12] Tapia R, de Jesus S. Early prediction of Asthma. J Clin Med 2023; 12(16): 5404.

[13] Salau AO, Pooja MR, Hasani NF, Braide SL. Model based risk assessment to evaluate lung functionality for early prognosis of asthma using neural network approach. Math Model Eng Probl 2022; 9(4): 1053-60.
http://dx.doi.org/10.18280/mmep.090423

[14] Pooja MR, Pushpalatha MP. Cluster analysis to characterize the patterns of complementary and alternative medicines usage in asthma controls. Open Public Health J 2020; 13: 1.
http://dx.doi.org/10.2174/1874944502013010227

[15] Pooja MR, Pushpalatha MP. A predictive framework for the assessment of asthma control level. Int J Eng Adv Technol 2019; 8: 239-45.

[16] Farion K, Michalowski W, Wilk S. Developing a decision model for asthma exacerbations: Combining rough sets and expert-driven selection of clinical attributes. Rough Sets and Current Trends in Computing. Springer 2006; pp. 428-37.
http://dx.doi.org/10.1007/11908029_45

[17] Pooja MR. A predictive model for the early prognosis and characterization of asthma. J Pharm Negat Results 2022; 13(4): 1-9.
http://dx.doi.org/10.47750/pnr.2022.13.04.001

[18] Pooja MR, Pushpalatha MP. An empirical analysis of machine learning classifiers for clinical decision making in asthma. Cognitive Computing and Information Processing. Springer 2017; pp. 105-17.

[19] Wu W, Bleecker E, Moore W, et al. Unsupervised phenotyping of severe asthma research program participants using expanded lung data. J Allergy Clin Immunol 2014; 133(5): 1280-8.

http://dx.doi.org/10.1016/j.jaci.2013.11.042 PMID: 24589344

[20]  Opina MTD, Moore WC. Phenotype-driven therapeutics in severe asthma. Curr Allergy Asthma Rep 2017; 17: 1-13.

[21]  Tsang KCH, Pinnock H, Wilson AM. Application of machine learning algorithms for asthma management with mHealth: A clinical review. J Asthma Allergy 2022; 15: 855-73.

[22]  Pooja MR, Pushpalatha MP. A hybrid decision support system for the identification of asthmatic subjects in a cross-sectional study. 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). 288-93.
http://dx.doi.org/10.1109/ERECT.2015.7499028

[23]  Schmidt S, Li G, Chen Y-P P. Medical knowledge discovery from a regional asthma dataset. Advanced Intelligent Computing Theories and Applications With Aspects of Artificial Intelligence. Springer 2008; pp. 888-95.
http://dx.doi.org/10.1007/978-3-540-85984-0_107

[24]  Palmu HH, Jaakkola MS, Mäkikyrö, EMS. Subtypes of asthma and cold weather-related respiratory symptoms. Int J Environ Res Public Health 2022; 19(14): 8790.

[25]  Pushpalatha MP, Pooja MR. A predictive model for the effective use of asthma severity indicators. 2017 International Conference on Computer Communication and Informatics (ICCCI).
http://dx.doi.org/10.1109/ICCCI.2017.8117717

[26]  AlSaad R, Malluhi Q, Janahi I, Boughorbel S. Predicting emergency department utilization among children with asthma using deep learning models. Healthc Anal 2022; 2: 100050.
http://dx.doi.org/10.1016/j.health.2022.100050

[27]  Patel D, Hall GL, Broadhurst D, Smith A, Schultz A, Foong RE. Does machine learning have a role in the prediction of asthma in children? Paediatr Respir Rev 2022; 41: 51-60.
PMID: 34210588

[28]  Molfino NA, Turcatel G, Riskin D. Machine learning approaches to predict asthma exacerbations: A narrative review. Adv Ther 2023; 41(2): 534-52.
PMID: 38110652

[29]  Pushpalatha MP, Pooja MR. A predictive model for the effective prognosis of Asthma using Asthma severity indicators. 2017 International Conference on Computer Communication and Informatics (ICCCI). 1-6.
http://dx.doi.org/10.1109/ICCCI.2017.8117717