Pooja and Pushpalatha, J Clin Respir Dis Care
2019, 5:1

Journal of
Clinical Respiratory Diseases and Care

Research Article Open Access

Machine Learning Approaches for Risk Stratification and Predictive Modeling of Asthma

Pooja MR* and Pushpalatha MP

Department of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

***Corresponding author:** Pooja MR, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India, Tel: +91-9886819448; E-mail: pooja.mr@vvce.ac.in

Received date: Feb 16, 2019; **Accepted date:** Aug 06, 2019; **Published date:** Aug 13, 2019

Abstract

Chronic respiratory diseases like Asthma and Chronic Obstructive Pulmonary Diseases (COPD) have attracted research interest in the area of risk stratification and many machine learning techniques have been the subject of interest in prediction systems involving risk stratification that perform early identification of the risk factors for the disease. Identification of patient populations at high risk is an important intervention in the early detection and clinical assessment of chronic diseases like asthma, as it can lead to targeted and personalized therapies. Here, we propose and deploy different machine learning approaches for the risk stratification under different study settings. All the approaches primarily predict the disease outcome or identify the severity/control level by recognizing the key risk factors for the disease depending on the nature of the data made available.

Keywords: Risk stratification; ISAAC; Machine learning; Spirometric; Predictive framework; Tiffeneau-Pinelli index

Introduction

Asthma is one of the chronic respiratory diseases that are basically reversible in nature. Early detection of the disease by identifying the associated risk factors leads to prevention of critical consequences such as asthma exacerbations which are detrimental to the lives of the subjects [1,2]. Asthma is heterogeneous encompassing a variety of symptoms and treatment of the targeted symptom identified at an early phase can be one of the major concerns in the effective treatment of the disease. Phenotype may be defined as the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment. The phenotype of an asthmatic is identified by their clinical presentation rather than by a single blood test [3]. Machine learning approaches play a vital role in the process of predicting and identifying the risk factors that contribute to the disease phenotyping that lead to effective targeted therapy [4,5]. A lot of emphasis has been laid on developing suitable tools that are easy to use and monitor the disease progression as well as sensitize the disease patterns which are an important intervention for the early identification of the disease [1,4,6]. Identification of the risk factors and comorbidities at a very early stage is even more important as the disease is progressive in nature [7,8]. A predictive model based on Principal Component Analysis and Least Square Support Vector Machine Classifier was used in to perform asthma outcome prediction [9]. A simple and robust tool for predicting asthma at school age in preschool children with wheeze or cough was deployed in using demographic and perinatal data, eczema, upper and lower respiratory tract symptoms, and family history of atopy [10-12]. The predictive model constructed using K-Nearest Neighbor and Support Vector Machine which identified asthma severity indicators showed good predictive performance [7]. Mining the most interesting patterns from multiple phenotypes medical data poses a great challenge in these days. The prevalence of airway diseases is underestimated. Although these diseases present several common characteristics, they have different clinical outcomes. The differentiation between asthma and chronic pulmonary disease in the early stage of disease is extremely important for the adoption of therapeutic measures. However because of the high prevalence of these diseases and the common pathophysiological pathways, some patients with different diseases may present with similar symptoms. Hence, there exists a strong requirement to delineate similarities and differences between these diseases in terms of risk factors, symptoms, diagnosis and treatment [6,13]. Risk stratification approaches aim at deploying different learning techniques for devising models that either predict the risk factors or persons at risk, while performing evaluation of the same, which would better define asthma phenotypes that may improve the understanding of the underlying pathobiology of the phenotypes and lead to targeted therapies for individual phenotypes[13,14].

Here we discuss a computerized hybrid decision support system developed by us that effectively predict the asthma outcome given the symptoms and comorbidities that contribute to the disease. Also we have developed a predictive framework that efficiently predicts the asthma control level in a clinical trial study that involves measured levels of various cytokines and counts of critical cell types present in the BAL fluids [8]. The asthma control levels are used to distinguish healthy, controlled and uncontrolled asthmatic subjects. The risk factors identified uniquely in each of the systems have significantly contributed to the precision of the systems. While feature clustering is employed to identify the features in the first technique, the second approach uses combined feature scoring technique to do the same [1,15]. Further, we also present a neural network approach for the risk assessment of the disease through predicting Tiffeneau-Pinelli index, a strong indicator of the asthma in a setting that involves spirometer tests.

Common binary classifiers generally adopted for two class classification problems include Logistic Regression, Support Vector Machine (SVM), K Nearest Neighbor (KNN) classifier, Naïve Bayes and Ensemble models have seen to exhibit comparatively good

performance when compared to the traditional classifiers and hence we opted to deploy the variants of ensemble techniques namely sequential and parallel techniques [7,15]. Further, one of the hybrid decision support system developed by us was effectively used to predict the outcome when binary attributes were used. A neural network approach was used in to predict the Tiffeneau-Pinelli index using Forced Expiratory Volume (FEV1) and Forced Vital Capacity (FVC) to assess the outcome of asthma which obtained desirable values for the Mean Absolute Error (MAE) [16]. A transformed feature set containing predicted lung function parameters using the raw features such as age, sex and height was used to train the model. Many cytokines are involved in the development of the atopic state and of the chronic inflammatory process of asthma, ultimately contributing to airway remodeling, bronchoconstriction, and bronchial hyper responsiveness. The potential role of each cytokine in these processes can be evaluated by studying their properties [17]. It was evident from the results in, that the cytokines and chemokines namely IL-8, IL-16 and IL1-RA as well as neutrophil percentage play a significant role in the process of distinguishing controlled and uncontrolled asthmatics [18].

The rest of the paper is organized as follows:

In Section 2, we discuss the various machine learning approaches deployed along with the description of the data on which they are validated. Section 3 presents the results and the related discussions of all the approaches. Section 4 provides concluding remarks.

Materials and Methods

Hybrid decision support system for the identification of asthmatics

The system identified the comorbid features that correlate significantly to the asthma via feature clustering technique which follows the principle of Fuzzy C Means clustering that incorporates a correlation based objective function. The feature subset is constituted by extracting only those features that coexist along with the asthma feature in the cluster containing them. The features included in the subset were presence/absence of permanent wheeze, the occurrence of 4 or more attacks of wheeze in the past 12 months, Sleep disturbance from wheeze, 1 or more nights a week in the past 12 months, Speech limited by wheeze in the past 12 months, Nose and eye symptoms in the past 12 months and symptoms of nose in the months of January, February, March, April, June and November.

The other clusters and the attributes included in them are ignored. Further prediction of asthma is performed by using only the features in the feature subset as independent attributes while the asthma attribute is used as the dependent attribute [19]. We clustered the available subjects into two clusters within which Classification and Regression Tree (CART) was employed for the purpose of predicting the asthma outcome. The technique was hybrid in nature as it incorporated the fusion of unsupervised and supervised techniques. This decision support system predicted the asthma outcome with a precision and recall of 1.0. Some of the commonly deployed binary classifiers which were analyzed for their performance in were deployed for a comparative study against the hybrid approach used by us [6,11]. The results obtained by the hybrid model were further compared with the traditional binary classifiers as well as ensemble methods including parallel and sequential ensembles and are projected in Table 1.

Method	F1	Precision	Recall
Gradient boost	0.787	0.818	0.817
Random Forest Learner	0.774	0.803	0.803
Naive Bayes	0.783	0.797	0.789
SVM* Learner	0.773	0.813	0.81
Logistic Regression	0.708	0.733	0.732
KNN*	0.717	0.761	0.761
SVM: Support Vector Machine; KNN: K Nearest neighbor			

Table 1: Comparative results for various classifiers via stratified 10-fold cross validation.

Random forest was used for the parallel ensemble technique while Stochastic Gradient Descent (SGD) was used for the sequential technique.

Dataset

The Neyveli asthma dataset available as study data gathered through questionnaires by ISAAC (International Studies of Asthma and Allergies in Childhood) was used for the purpose of validating the model. The attributes were binary in nature incorporating only the presence/absence of the disease which were encoded as 0 and 1 respectively and indicated symptoms including wheeze, speech limitation arising from wheeze, nose irritations, presence of hay fever, rashes, breathing problems following exercise and gender, shortness of breath, sleep disturbance and frequency of recurrence in symptoms [11]. A balanced dataset including an approximately equal number of subjects with and without asthma were used to train and test the model. The problem of prediction was reduced to a two – class prediction with a positive class containing 78 samples and a negative class including 64 samples.

Common metrics for performance evaluation of classifiers include precision, recall and F1 measure. Hence we present the results for the classifiers deployed in terms of these metrics. A 10-fold stratified cross validation scheme is used to test the performance of the system wherein the data is divided into 10 groups and at any point of time while one of the group called fold is used in the test set, the rest of them will be used for training. Further, in the test on test data scheme, a separate dataset that is unused for training will be deployed to test the performance of the model.

A total of 78 subjects were identified to have asthma while a total 3203 were identified to not have asthma. This distribution indicates a clear problem of imbalance. Thus, we opted to choose a collection of about 64 samples which constitute nearly 2% of the total samples drawn from the set of subjects with asthma outcome as 2. This leads to a balanced class problem where we finally have a positive class with 78 samples and a negative class with 64 samples. We have used ANOVA, Chi2, Relief and FCBF feature scoring techniques and finally apply combined feature scoring technique to find the most relevant features that are applicable for the problem of classification. This gives rise to a reduced feature set that characterizes asthma morbidity and is used in the successive stages for the prediction of asthma outcome. The following features constituting about 25% of the features from the

original feature space which were identified as risk factors are used in reduced feature set.

whezev- Wheeze ever

whez12- Wheeze in the past 12 months

nwhez12- 4 or more attacks of wheeze in the past 12 months

awake12-Sleep disturbance from wheeze, 1 or more nights a week in the past 12 months

speech12-Speech limited by wheeze in the past 12 months

ieyes12-Nose and eye symptoms in the past 12 months

pnosejan, pnosefeb, pnosemar, pnoseapr, pnosejun, pnosenov-Nose symptoms in the respective months

Predictive framework for the assessment of asthma control level

The hybrid decision support system however yielded optimal results with the ISAAC dataset wherein, the features were binary in nature. In the case of the second dataset discussed below, which included a combination of nominal and binary features, the regression techniques coupled with regularization was applied to predict the asthma outcomes which were categorized as healthy, controlled and uncontrolled. Of all the regularization techniques, elastic net worked well with the regression to result in least MSE, RMSE and MAE which projects the good performance of the system. Further, a binary classification problem to classify the controlled and uncontrolled subjects was formulated, the results for which are projected in Table 2 [8].

Method	F1	Precision	Recall
SGD*	1	1	1
Logistic Regression	0.571	0.16	0.4
Naive Bayes	0.667	0.85	0.8
Random Forest Learner	0.667	0.85	0.8
SVM* Learner	1	1	1
SGD: Stochastic Gradient Descent; SVM: Support Vector Machine			

Table 2: Classification results using test on test data.

The dataset was divided into two sets namely training and test set. The classification process was performed by considering only two classes namely uncontrolled and controlled asthma. 80% of the total data constituting 20 instances were used in the train set while the test data consisting of 5 instances, covering 3 instances with controlled asthma and 2 instances with uncontrolled asthma. The test set was used to evaluate the performance of the model and the results obtained were considerably fair.

Dataset

In yet another study involving asthma data taken from Department of Asthma, Allergy and Lung Biology, King's College London School of Medicine, U.K. primarily made available on the Dryad repository, Stochastic gradient Descent(SGD) ensemble was used to classify controlled and uncontrolled asthmatic subjects [17]. The dataset included measured levels of various cytokines and chemokines in the Bronchoalveolar Lavage (BAL) fluid along with the counts of various cell types and the usage of medications including ICS, SABA, LABA and ICS and their dosage [18]. Other binary classifiers including random forest, Naïve Bayes, SVM learner, Logistic Regression and KNN techniques were deployed to perform the same task. The SGD technique outperformed the other binary classifiers by scoring optimal values in terms of F1, precision and recall.

About 25% of the total features were extracted from the feature set as risk factors as depicted in Table 3.

Features	Information Gain	Gain Ratio	Gini	ReliefF	FCBF
ICS (dose µg/day)	0.970950594	0.551374456	0.48	0.4624	0.707961792
% predict FEV1	0.728212946	0.364106473	0.36	0.226222222	5.10E-05
LABA	0.716642278	0.767227049	0.387692308	0.72	0.720990999
ICS (use)	0.609986547	0.609986547	0.32	0.564	6.62E-05
Age	0.590468571	0.297394066	0.3	0.314064516	4.05E-05
IL-1RA	0.367248898	0.183624449	0.2	-0.034420969	2.63E-05
IL-8	0.367248898	0.183624449	0.2	0.069567307	2.63E-05
% neutrophil	0.272242172	0.137116877	0.146666667	0.044809524	1.99E-05
IL-16	0.242737649	0.121368824	0.12	0.006743686	1.79E-05
M-CSF	0.242737649	0.121368824	0.12	-0.0125732	1.79E-05
total cells (x106)	0.242737649	0.121368824	0.12	-0.017178571	1.79E-05
% mo0/Mac	0.190468571	0.09593097	0.121666667	-0.031826923	1.32E-05

Table 3: Feature subset obtained by combined feature scoring.

The factors included Inhaled corticosteroids (ICS) (dose µg/day), percentage of predicted FEV1 (Forced Expiratory Volume in the first second), Long-acting beta-agonists (LABA), use of Inhaled corticosteroids Age, Interleukin (IL)-1α, IL-1β, IL-1 receptor antagonist(IL-1RA), IL-8,percentage of neutrophils, IL-16, Macrophage Colony-Stimulating Factor(M-CSF), total cells, percentage of macrophages were finally chosen to perform the prediction using combined feature scoring technique. The different feature scoring techniques involved in weighted feature averaging includes Analysis of Variance (ANOVA), Chi2, ReliefF and Fast Correlation Based Feature Selection (FCBF) techniques.

Performance metrics for problems involving regression include Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE) are presented in Table 4 below for the variants of regression used in the prediction framework.

Method	MSE	RMSE	MAE
Linear Regression -Lasso	0.273	0.523	0.465
Linear Regression-Elastic Net	0.158	0.398	0.358
Linear Regression- Ridge	2.665	1.633	1.444
Linear Regression-No Regularization	2.675	1.636	1.447

Table 4: Prediction results using test on test data.

Also the predicted values *vs* actual values encoded for asthma outcome using elastic regression is shown in Table 5.

	Asthma Control Level	Predicted Continuous Value	After Round Off
Elastic net	0	0.031675699	0
	1	0.945590652	1
	1	1.043728739	1
	2	1.911049088	2
	2	2.036280123	2
	0	0.031675699	0
	0	0.031675699	0

Table 5: Actual and predicted outcomes test on test data.

Neural network approach for risk assessment of asthma

A neural network approach for risk assessment of asthma was deployed to identify patients at risk by predicting a significant pulmonary function parameter called Tiffeneau-Pinelli index by using the factors such as age, gender height along with FEV1 and FVC. The Tiffeneau-Pinelli index is a strong indicator of the asthma disease as a drop in the index below 0.70 implies a diagnosis of asthma. Various neural network types were used to train the model on data obtained using spirometry tests in a clinical setting [16]. The transformed input dataset with respect to the two attributes FEV1P and FVCP is used to train the different neural networks for the prediction process to predict the target parameter which is obtained by computing the ratio of Forced Expiratory Volume in the first second (FEV1) and Forced Vital Capacity (FVC), i.e. FEV1/ FVCP, (FEV1 and FVC, as obtained from the tests carried out on the patients). The FEV1P, i.e. FEV1 predicted and FVCP i.e, FVC predicted is computed using the formulae published by Association for Respiratory Technology and Physiology for males and female. Of the different neural network types, radial basis function network was most preferred as it obtained optimal results. The trained model is now used to predict the target index at any point of time using age, gender and height as input data using the already trained model to predict the index. Tables 6 and 7 present the actual vs. predicted index values and the mean absolute error for the sample male and female subjects respectively.

Actual	Predicted
0.8875	0.866
0.89	0.8677
0.8954	0.8385
0.8624	0.8596
0.8801	0.8656
MAE: 0.0236	
MAE: Mean Absolute Error	

Table 6: Actual and predicted vectors for Tiffeneau-Pinelli index for sample male Subject.

Actual	Predicted
0.8515	0.8307
0.827	0.8329
0.8167	0.8318
0.814	0.8276
0.8207	0.8237
0.8194	0.8476
0.8022	0.8454
0.741	0.8297
MAE: 0.0273; MAE: Mean Absolute Error	

Table 7: Actual and predicted vectors for Tiffeneau-Pinelli index for sample female subject.

Dataset

The data is taken from the SPIROLA dataset. The dataset basically contains longitudinal data with respect to individual patients recorded over time; however the data is not strictly periodic in nature though closer to being called time series data [15]. The dataset contains the following features: age, sex, race, pulmonary function parameters such as Forced Expiratory Volume in 1 second (FEV1), Forced Vital capacity(FVC) are included as the primary attributes along with optional attributes such as second best FEV1, second best FVC. For the sex attribute, male is encoded as 1 in our dataset and female as 2.

Results and Discussion

The results obtained by the hybrid model were further compared with the traditional binary classifiers as well as ensemble methods including parallel and sequential ensembles and are projected in Table 1. Random forest was used for the parallel ensemble technique while Stochastic Gradient Descent (SGD) was used for the sequential technique. A confusion matrix is employed as shown in Table 8 to reflect the results of hybrid decision support system wherein true positives, true negatives, false positives and false negatives are represented.

Predicted	Actual		Predicted	Actual	
	Positive	Negative	Positive	Negative	
Positive	61	0	Positive	17	0
Negative	0	15	Negative	0	49
Cluster 1			Cluster 2		

Table 8: Confusion matrix obtained using hybrid model.

The subjects available for study were first clustered into two clusters namely Cluster1 and Cluster2 and the prediction of asthma outcome was performed within the individual clusters. The total number of false positives and false negatives were zero in both clusters indicating the optimal performance of the system used.

Table 4 depicts the performance of the various regularization techniques for the predictive framework. The predicted outcome values obtained against the actual outcome values using elastic regression are presented in Table 5.

Table 5 shows the risk factors along with their scores as discussed earlier and Table 2 shows the two – class classification results using various classifiers when tested on test data.

Tables 6 and 7 show actual and predicted indices using radial basis function network for male and female subjects respectively for the neural network approach discussed earlier.

The Mean Absolute Error (MAE) is shown in both the cases, which is considerably a good value indicating least error between actual and predicted Tiffeneau-Pinelli vectors.

Overall the results of the different approaches can be summarized as below: The first scenario involved asthma data that involved observable characteristics (symptoms) that were recorded as part of written questionnaire that was conducted by ISAAC. The system generated the binary prediction with a precision and recall of 1.0 as the number of false positives and false negatives were zero. The second method was applied on data that was obtained as part of clinical study involving a panel of cytokines and cell types present in the BAL fluids. The combined feature scoring technique used to identify the risk factors that constitute the reduced feature set has proved to be fruitful as it has led to optimal levels of precision and accuracy in prediction. The prediction framework involving elastic net regression was able to predict all the three levels of asthma control level with a high degree of precision. Further the SGD ensemble adopted for binary classification between controlled and uncontrolled asthma also yielded very accurate results by scoring a value of 1.0 for all the classification performance metrics under both the testing schemes, namely test on test data and random sampling. Finally, a third approach involving neural network technique that can be adopted in a spirometric setting was successfully used to assess the healthy subjects at risk by predicting the Tiffeneau-Pinelli which implies the level of asthma severity.

Conclusion

The different approaches have effectively identified the risk factors which are used to predict the outcome of asthma with a high degree of precision. The results of all the approaches are systematically validated using suitable performance metrics. The performance evaluation of the approaches suggests that all the methods are quite promising, because of which they could be considered as strong models for deployment in clinical settings under various tests. They can be suitably incorporated into the clinical pathways that assess patients at risk and identify the different degrees of severity of the disease.

References

1. Sly PD, Boner AL, Björksten B, Bush A, Custovic A, et al. (2008) Early Identification of Atopy in the Prediction of Persistent Asthma in Children. Lancet 372: 1100-1106.
2. Chang TS, Lemanske RF, Guilbert TW, Gern JE, Coen MH, et al. (2013) Evaluation of the Modified Asthma Predictive Index in High-Risk Preschool Children. J Allergy Clin Immunol 1: 152-156.
3. Shivade C, Raghavan P, Fosler- Lussier E, Embi PJ, Elhadad N, et al. (2013) A Review of Approaches to Identifying Patient Phenotype Cohorts using Electronic Health Records. J Am Med Inform Assoc 21: 221-230.
4. Prasadl BDCN, Prasad Krishna PESN, Sagar Y (2011) An Approach to develop Expert Systems in Medical Diagnosis using Machine Learning Algorithms (Asthma) and a Performance Study. Int J Soft Comp 2: 26-33.
5. Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE (2014) Predicting Phenotypes of Asthma and Eczema with Machine Learning. BMC Med Genomics 7: S7.
6. Simpson A, Tan VY, Winn J, Svensen M, Bishop CM, et al. (2010) Beyond Atopy: Multiple Patterns of Sensitization in Relation to Asthma in a Birth Cohort Study. Am J Res Crit Care Med 181: 1200-1206.
7. Pushpalatha MP and Pooja MR (2017) A Predictive Model for the Effective Prognosis of Asthma using Asthma Severity Indicators. Computer Communication and Informatics (ICCCI), 2017 International Conference on IEEE.
8. Pooja MR and Pushpalatha MP (2019) A Predictive Framework for the Assessment of Asthma Control Level. Int J Eng Adv Tech 8: 239-245.
9. Chatzimichail E, Paraskakis E, Sitzimi M, Rigas A, et al. (2013) An Intelligent System Approach for Asthma Prediction in Symptomatic Preschool Children. Comput Math Methods Med 2013.
10. Pescatore AM, Dogaru CM, Duembgen L, Silverman M, Gaillard EA, et al. (2013) A Simple Asthma Prediction Tool for Preschool Children with Wheeze or Cough. J Allergy Clin Immunol 133: 111-118.
11. Asher MI, Keil U, Anderson HR, Beasley R, Crane J, et al. (1995) International Study of Asthma and Allergies in Childhood (ISAAC): Rationale and methods. Eur Respir J 8: 483-491.
12. Sanchez-Morillo D, Fernandez- Granero MA, Leon- Jimenez (2016) A Use of Predictive Algorithms in-Home Monitoring of Chronic Obstructive Pulmonary Disease and Asthma: A Systematic Review. Chron Respir Dis 13: 264-283.
13. Howard R, Rattray M, Prosperi M, Custovic A (2015) Distinguishing Asthma Phenotypes using Machine Learning Approaches. Curr Allergy Asthma Rep 15: 38.
14. Smit HA, Pinart M, Anto JM, Keil T, Bousquet J, et al. (2015) Childhood Asthma Prediction Models: A Systematic Review. The Lancet Respir Med 3: 973-984.
15. Pooja MR and Pushpalatha MP (2017) An Empirical Analysis of Machine Learning Classifiers for Clinical Decision Making in Asthma.

International Conference on Cognitive Computing and Information Processing Springer 2017.

16. Pooja MR and Pushpalatha MP (2018) A Neural Network Approach for Risk Assessment of Asthma Disease. J Health Inform Manag 2:1.
17. Hosoki K, Ying S, Corrigan C, Qi H, Kurosky A, et al. (2015) Analysis of a Panel of 48 Cytokines in BAL Fluids Specifically Identifies IL-8 Levels as the only Cytokine that Distinguishes Controlled Asthma from Uncontrolled Asthma, and Correlates Inversely With FEV1. PloS One 10: e0126035.
18. Pooja MR, Pushpalatha MP (2019) Analysis of A Panel Of Cytokines in BAL Fluids to Differentiate Controlled and Uncontrolled Asthmatics Using Machine Learning Model. Journal of Respiratory Research 5: 142-145.
19. Pooja MR and Pushpalatha MP (2015) A hybrid decision support system for the identification of asthmatic subjects in a cross-sectional study. Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on IEEE, 2015.