Machine Learning-Based Risk Prediction of Asthma

Arnav Kolte*1, Divija Joshi *2, Hitesh Choudhary *3, Juhi Meshram *4, Prof. Mrs. Ashwini Kukade*5

*1 UG Student, Dept. of Artificial Intelligence, G.H Raisoni College of Engineering, Nagpur, Maharashtra, India
*2 UG Student, Dept. of Artificial Intelligence, G.H Raisoni College of Engineering, Nagpur, Maharashtra, India
*3 UG Student, Dept. of Artificial Intelligence, G.H Raisoni College of Engineering, Nagpur, Maharashtra, India
*4 UG Student, Dept. of Artificial Intelligence, G.H Raisoni College of Engineering, Nagpur, Maharashtra, India
*5 Asst Professor, Dept. of Artificial Intelligence, G. H Raisoni College Of Engineering, Nagpur Maharashtra, India

---

Abstract - This paper will attempt to detect asthma early based on machine learning. This work utilizes a dataset consisting of 21 features such as family history of asthma, intake of drugs, smoking status, and quality of air and much more. The work tries to develop models with maximum precision, recall, and accuracy in asthma classification. These include several machine learning algorithms evaluated: SVM, Decision Trees, and Random Forest. This study focuses on both medical and environmental data when used in combination. The results here show that the Random Forest model performed well in achieving an overall high performance, with an accuracy of 0.926829. This has dire implications for evaluation and early diagnosis of Asthma.

Key Words: Asthma detection, Machine learning, Patient history, Environmental factors, Support Vector Machines (SVM), Decision Trees, Random Forests, Predictive models, Air quality, Smoking habits, Medication usage, Family asthma history, Ac- curacy, Precision, Recall, Early diagnosis, Healthcare, Medical data, Environmental data, Predictive power

1. INTRODUCTION

Asthma is a chronic respiratory disease that affects millions of people around the world who suffer recurring attacks of breathlessness, coughing, and wheezing. If not treated appropriately, these recurring attacks can be deadly. The disease is tricky because asthma symptoms vary with the patient, making timely diagnosis and even treatment extremely crucial. The present diagnosis of asthma is more dependent on subjective patient reporting, and clinical assessment which measures lung function. While useful, these methods are often unreliable and require many hours in time and frequent visits to the hospital in many cases, unfortunately delaying critical interventions, particularly in emergency settings. It has hope in these challenges because machine learning processes great datasets that determine patterns that clinicians might not notice by traditional methods. By opening new doors with machine learning algorithms identifying subtle correlations in most patient and environmental factors, the accuracy and speed of detecting asthma are improved. Predicting an asthma attack before it occurs will also help advance the field of preventative care and managing the disease properly. A set of strong predictive models using machine learning will be constructed that make predictions about asthma attacks based on the analysis of a comprehensive combination of both patient medical histories and environmental data like air quality, humidity, and exposure to allergens. We use the best available balanced dataset that takes into account a wide range of features such as family history of asthma, medication usage, smoking habits, and previous instances of asthma attacks. This dataset allows us to build models that not only identify the at-risk individuals but also yield early warning signs for potential exacerbations. In this paper, we present several machine learning algorithms - namely SVM, Random Forests, and Decision Trees - and de- termine the best performing model for predicting asthma. We intend to develop a system that can assist the health providers to better diagnose and manage asthma by testing the above models on key performance metrics like accuracy, precision, recall, and F1-score. This will ensure a more effective amount of burden reduction on healthcare resources while improving patient outcomes due to earlier and more accurate diagnoses.

2. LITERATURE REVIEW

With recent incorporation of machine learning in asthma prediction and management, various studies have come under emphasis, which are stating that combining patient data and bio-signals with the present environmental conditions can make predictions useful. Vincenzo et al. (2023) discussed the way in which oxidative stress plays an important role in childhood asthma. The study describes the manner in which pollutants in smoke, allergens, etc., raise oxidative stress inside children's airways; these create inflammation and put a child at higher risks for asthma. [1], suggests that early lifestyle interventions and antioxidant therapies may keep this stress under check, thereby preventing asthma from worsening with age in children. Tsang et al. (2022) have reviewed the application of machine learning within mHealth platforms for asthma management based on data from diverse sources, such as sleep quality and environmental triggers. The studies also reveal that the predictive models in mHealth can predict asthma attacks and track symptoms, owing to personalized, age-independent support of asthma patients throughout. [2], this suggests the potential for mobile health interventions in the asthma management of diverse patient populations. Another study examined the performance of machine learning models on the accuracy of determining risk for asthma by analyzing the blood sample through Rahat Ullah et al. in 2019. [3], employed some classifiers, such as ANN and SVM, RF, to be able to distinguish between an asthma patient sample and one from a healthy individual by concluding that SVM was performed the best. This study will show that non-invasive tools like combining Raman spectroscopy and machine learning could be so effective in diagnosing a patient with asthma. This approach taken by Alharbi et al. (2021) was the inclusion of bio-signal data including PEFR and FEV1 lung function measures along with environmental factors such as air quality and temperature. [4], shows that using both bio signals and environmental data makes asthma attack predictions more reliable, highlighting the benefits of including multiple sources of information to improve the accuracy of these models. Zein et al. (2021) leveraged large-scale electronic health record (EHR) data to predict asthma ex-acerbations, including non-severe episodes, emergency visits, and hospitalizations. [5], applied logistic regression, Random Forests, and Gradient Boosting models to the data with the confirmation that the flares of asthma can be well predicted with EHR. The strategy appears fruitful in clinics for decision of flares by a doctor so that timely intervention can be taken to avert dangerous repercussions. In its totality, these studies exemplify the promise of machine learning for asthma control. Researchers are getting one step closer toward creating prediction models that enable more proactive and tailor-made asthma treatment using patients' histories, bio signals, and environmental features. [6] Discusses how Machine learning techniques are increasingly being applied to the management of asthma in order to predict exacerbations thus opening doors to better care and early intervention. Recently, Jayamini et al. published a systematic review of 20 studies between the years 2010-2023 using ML-based models for predicting asthma attacks based on clinical, environmental, and socio-demographic data. Other models utilized Random Forest, XG-Boost, and Logistic Regression. The ensemble models show the best performance. Importantly, with a short prediction window like 3–7 days, the model performed better; however, several challenges and drawbacks have been discussed and are mostly related to issues such as data imbalance and lack of interpretability. Future studies should focus on explain- able AI, the optimization of hyperparameters, and inclusion of real-time data from connected devices, thus increasing predictability and clinical utility. In a systematic review, [7] concludes the predictive performance of machine learning models on asthma exacerbation. Here, the authors analyzed 23 models from 11 studies and showed the validity of logistic regression, Random Forest, and boosting in which boosting got the maximum AUROC to be at 0.84. The key predictors were steroids, emergency visit, age, and exacerbation history. Although these ML models show good predictability, there are several problems identified in heterogeneity, bias, and low generalization. The study emphasizes the point that future studies can make the clinical application better with good generalization and data integration in real time. [8] did a clinical review that focuses on the application of machine learning in asthma management using mHealth technologies. The current activities involved technology development, attack prediction, and patient clustering. Among the algorithms of ML applied were logistic regression, Random Forest, and support vector machines on data extracted from smart devices such as smartphones, smart inhalers, and wearable sensors. Most of the promising technologies had small datasets without any validation outside the group that doesn't allow gener- alization. Once again, the review itself hints at larger, real- world studies and better integration of the data sources as core aspects to more impactful management solutions for asthma.

[9] The paper is titled "Diagnosing Asthma and Chronic Obstructive Pulmonary Disease with Machine Learning." This work is better compared to other works relating to asthma and COPD diagnosis with machine learning. Here, several algorithms depending on the information retrieved from 132 patients associated with detection efficacy from diagnostic features have been applied. The best performing was the Random Forest classifier of COPD with 97.7 % precision in which smoking and age feature are the most important one. [10] deals with the issue that machine learning can be used as an intermediary to improve the detection of asthma or not? It has used the set of data regarding the hospitals located in Tehran by using a comparison of KNN, SVM, and random forest algorithms. Major emphasis was given to data preprocessing and cross-validation with the hope that better results would come out from the proposed models.

Results: The results for the KNN algorithm reflect high sensitivity, specificity, and accuracy in five neighbors, making the proposed algorithm the best asthma diagnosis model.

3. METHODOLOGY

A. Dataset

The dataset used in this work is an application containing medical and environmental variables about 21 features that describe the patient's past medical history and external sources causing asthma. The features applied for the dataset are as follows: use of medications, family history of asthma, history of asthma attacks, past nights not disturbed by sleep, smoking, environmental asthma triggers, AQI, and relative humidity. The target variable is, of course, the occurrence of an asthma attack - and this is very binary, in that 0 signifies no attack and 1 represents an attack. The model is able to capture a more realistic view of asthma triggers since it considers both clinical

and environmental variables, and this is critical because asthma is known to be influenced by intrinsic factors, including genetics, and extrinsic factors, like pollution and allergens. Diversity in patient demographics, environmental

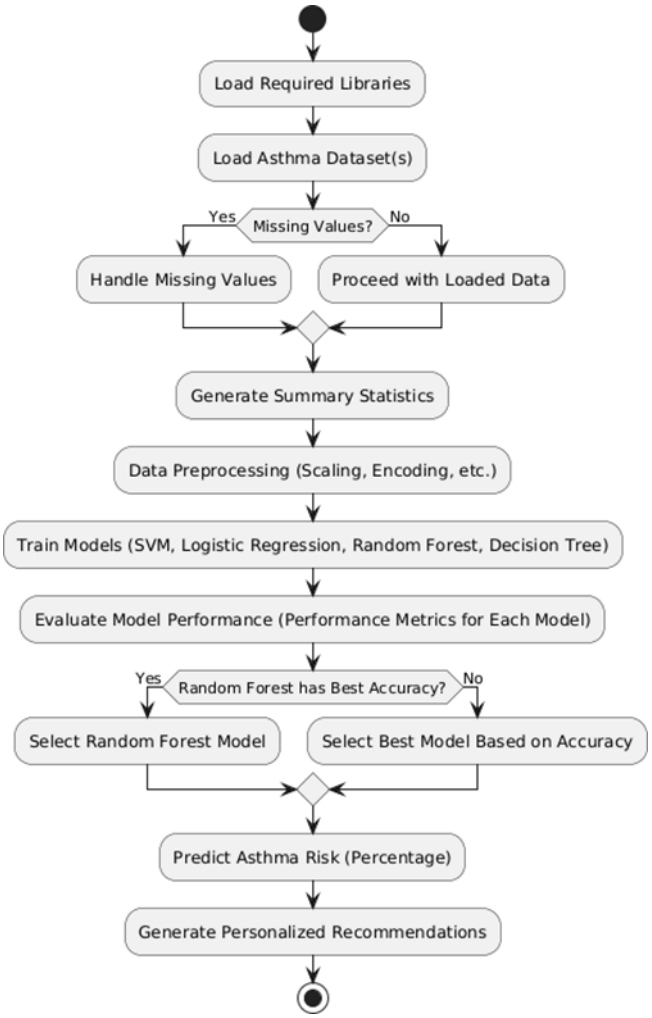

Fig. 1. Methodology.

conditions, and clinical variables was provided by sourcing the dataset from health care records and environmental monitoring systems.

B. Data Preprocessing

Preprocessing would be needed for the machine learning model preparation process. Data was presented with missing values, outliers, and categorical variables that needed to be addressed before the actual process of training the model. Handling Missing Data: Missing values are imputed appro- priately with strategies taken along the types of variables. For numeric features, missing values are replaced by the mean of the respective column. For categorical features, the mode- thus, the most frequent category-is used. This imputation helps in achieving completeness in the dataset, without inducing bias due to removal of rows with missing data. Features contributed equally to the model and improved distance-based algorithms like SVM. Numerical features were normalized by using min-max scaling: This scaling transforms all features into the range between 0 and 1, which improves convergence during training by eliminating feature magnitude effects on algorithms sensitive to these quantities. One-Hot Encoding: Some categorical features-the ”smoking status” and ”asthma medication usage”-are encoded with one-hot encoding to be numerically transformed. A type of numerical encoding that is being used for all the categorical variables is the development of new columns for each category in the actual feature. This is so the ordinal relationship that may exist between categories would not be misinterpreted by the model. Outlier Detection and Treatment: With the help of box plots and Z-scores, several outliers in the continuous variables, such as AQI and humidity were detected. Influence of extremely large points in the model performance was curtailed through clipping or replacing values larger than three standard deviations from the mean by the median. Data Splitting: This dataset was split into training and testing sets in a ratio of 80:20. Training set was utilized to fit the models while the testing set was used to assess the generalization performance of the models on unseen data. In this work, stratification splitting was used that ensures the proportion of cases of asthma attacks was preserved in both sets; it helps in avoiding bias while doing the model evaluation.

C. Feature Selection

One of the important tasks is feature selection to improve model performance. At this stage, the features that are not relevant or redundant, which may lead to overfitting, are removed. In this paper, the combination of correlation analysis and RFE was used for feature selection. Analysis of Correlations: Firstly, the correlation matrix is generated. It helps understand and view the relationship between the independent features and target variable (asthma attack). Features having very low correlation with the target are available to be removed. It also checks multicollinearity using correlation among independent features. Features that show a high correlation of 0.85 with one another are removed to decrease redundancy and improve the interpretability of the model. Recursive Feature Elimination (RFE): The RFE algorithm was applied to the dataset, systematically deleting the worst-performing features and ranking them according to how important they were to the task of prediction. This was particularly useful in identifying which features to use and eliminating the ”noise” that were not contributing much to the model’s correct outcome. By using feature selection, we were able to reduce the dataset down into a more tractable subset of features, which in addition to improving model performance also greatly reduces computer processing time.

D. Model Selection and Training

After studying several machine learning models, we chose the one that suits an asthma prediction. Here, we have some models which are trained and tested. Support Vector Machines (SVM): SVM is a robust classifier that performs very well in high-dimensional spaces. It makes it better when the features in a dataset exceed the number of observations, which makes this dataset the choice. SVM was trained with radial basis function and RBF kernel to address the non-linear interaction

between features and target variable. The hyper parameters which incorporate penalty parameter and kernel coefficient were optimized through grid search. Decision Trees: Classi-

fication problems can be well represented by decision trees, and it is very interpretable, too. The Gini impurity criterion was used to train the model. In a decision tree model, nodes are split based on the Gini impurity criterion, but that doesn't mean decision trees don't face the problem of overfitting the given data. Techniques of pruning were applied to prevent it from creating overly complex trees. Random Forest combines the forecasts from a series of decision trees to reduce over- fitting and increase accuracy. Number of decision trees in the forest and also the maximal depth of each tree was tuned by grid search cross-validation. Logistic Regression: The baseline model used was logistic regression. Because logistic regression is a linear model, even though the models we would compare against it were rather complex, this would give a good point of comparison. The regularization parameter has been tuned to avoid overfitting the model. Each model was also trained over the training set, using k-fold cross-validation, k=5; as such, we had to prevent the results from depending on one specific split of the data.

E. Evaluation Metrics

To assess the performance of the models, several evaluation metrics have been used: Accuracy: This measures the proportion of correctly predicted instances (both positive and negative) out of the total instances. Accuracy is useful but can be misleading in imbalanced datasets. Precision and Recall: Precision is defined as the ratio of true positives to all predicted positives, and recall measures the ratio of true positives to all actual positives. These are especially important metrics because of the high imbalance of this dataset with no asthma attack at a large majority class and hence may dominate the results. F1 Score: It was the harmonic mean of precision and recall, employed to give one metric that addressed the balance of precision and recall and which is more informative for datasets that are imbalanced in nature. Area under the receiver operating characteristic curve (AUC): AUC will be used to evaluate the model's ability to discriminate between positive classes and negative classes. Results from these models were compared with the metrics, and the model that produced the highest F1-score with the AUC-ROC as high as possible was regarded as the best model predicting asthma exacerbations. Based on the observations after training multiple models we found out the random forest model to be the most suitable for detecting asthma based on evaluation metrics such as accuracy, precision, recall, and F1-score. In this way the model is evaluated using these metrics that help to select a suitable model for predicting the risk of asthma.

4. RESULT

We had implemented the Random Forest algorithm in this project to analyze a given dataset that incorporates several health parameters and corresponding demographic factors in asthma. We built and trained a wide dataset incorporating age, gender, and medical history using this model. The Random Forest model, too, after intensive training and validation, has an accuracy of around 0.926829, implying that it fairly well

TABLE I
MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.843902	0.824742	0.842105	0.833333
Decision Tree	0.912195	0.896907	0.915789	0.906250
Random Forest	0.926829	0.908163	0.936842	0.922280
Support Vector Machine	0.653659	0.611111	0.694737	0.650246

predicts who should be put at risk of having an asthma attack. Other measures that affirm the efficacy of the model are precision, recall, and F1 score: they set the model as one which kills off least false positives and false negatives. Of course, the model can add up to individualized recommendations if the risk factors identified are interpreted. For instance, high-risk individuals were offered individual counseling in adhering to medication usage and lifestyle modification that can decrease their risk. The results of the application indicate that machine learning can help treat asthma patients by giving real-time risk scores and actionable information about the appropriate course of action. It also shows that integrating predictive analytics into practice would allow patients and doctors to better make decisions regarding their asthma management, hence providing better health outcomes. In the end, we also get the severity of symptoms i.e. how serious those symptoms are. Final product shows the percentage of asthma risk and gets a personalized recommendation based on the given inputs.

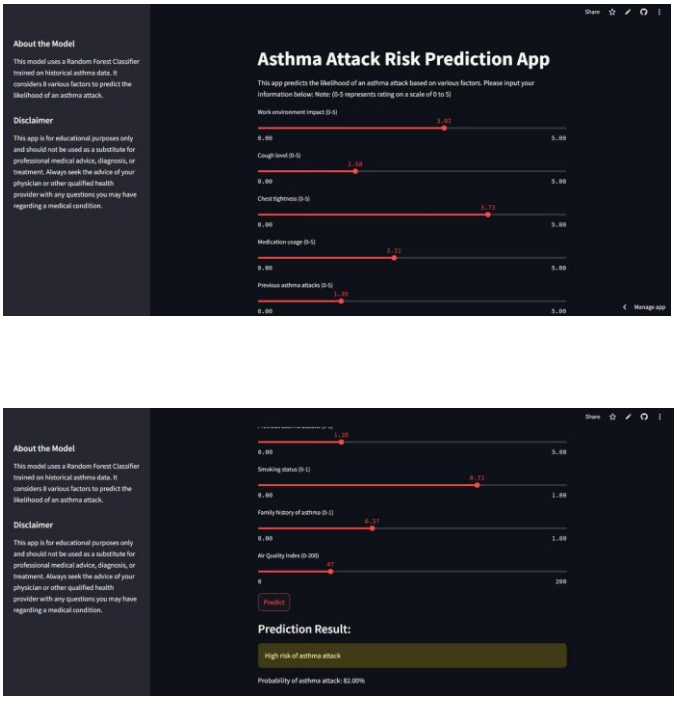

Fig. 2. Prediction 1 (High Risk).

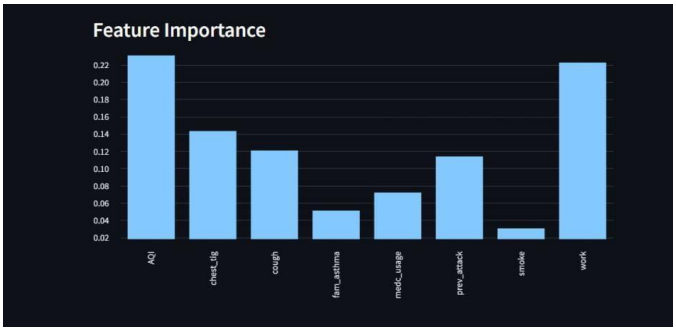

Fig. 3. Feature Importance

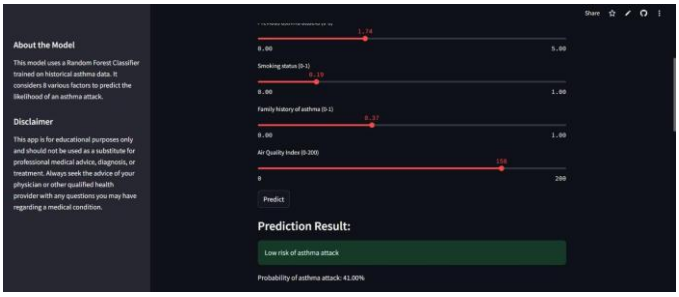

Fig. 4. Prediction 2 (Low Risk).

Fig. 5. Prediction History

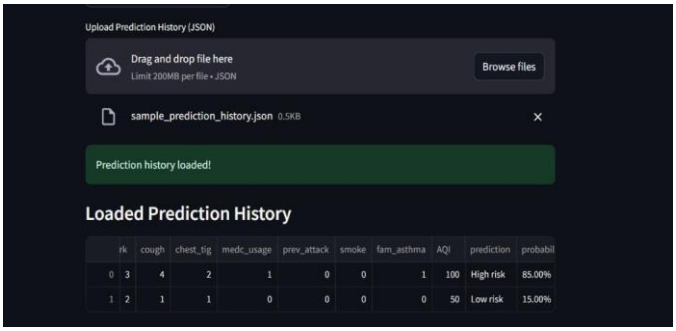

Fig. 6. Uploading past prediction history and viewing it.

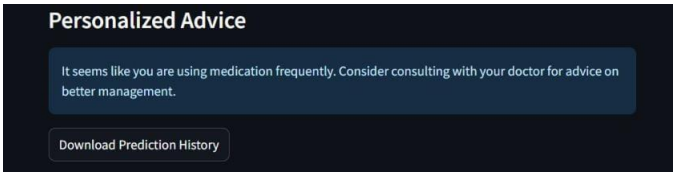

Fig. 7. Personalized Advice.

5. CONCLUSION

This study leaps forward by bringing machine learning techniques for personalized healthcare to the world by predicting asthma attack risks in this study. In short, the Random Forest model identified at-risk patients whose exposure might be of such high risks, which helped empower the patient and their caregiver to proactively take actions into better management of their asthmatic condition. It shall improve the asthma management and produce recommendations based on identified risk factors, hopefully reducing emergency visits and providing a much healthier quality of life for the patient. At the same time, the project exhibits the possibility of applying technology in the health arena as a support element for the personally tailored approach towards chronic diseases. Further work may then rely on the developed foundation, fusing the real-time data from wearable devices and mobile applications in continuous monitoring and dynamic risk assessment. Moreover, this model could relate to more diverse demographics represented in the dataset, thereby making it universally applicable and effective for wider ranges of populations. With such an approach, we seek to achieve better results for the patients and contribute to- ward the natural development of asthma management through innovative technology-based solutions toward every patient's needs.

6. REFERENCES

[1] Vincenzo, S.D.; Ferrante, G.; Ferraro, M.; Cascio, C.; Malizia, V.; Licari, A.; La Grutta, S.; Pace, E. Oxidative Stress, Environmental Pollution, and Lifestyle as Determinants of Asthma in Children. Biology 2023, 12, 133. https://doi.org/10.3390/biology12010133

[2] Tsang, K. C. H., Pinnock, H., Wilson, A. M., & Shah, S. A. (2022). Application of Machine Learning Algorithms for Asthma Management with mHealth: A Clinical Review. Journal of Asthma and Allergy, 15, 855–873. https://doi.org/10.2147/JAA.S285742.

[3] Rahat Ullah, Saranjam Khan, Hina Ali, Iqra Ishtiaq Chaudhary, Muham- mad Bilal, Iftikhar Ahmad, A comparative study of machine learn- ing classifiers for risk prediction of asthma disease, Photodiagnosis and Photodynamic Therapy, Volume 28,2019,Pages 292-296,ISSN,1572-1000,https://doi.org/10.1016/j.pdpdt.2019.10.011.

[4] Alharbi, E.T., Nadeem, F. & Cherif, A. Predictive models for person- alized asthma attacks based on patient's biosignals and environmental factors: a systematic review. BMC Med Inform Decis Mak 21, 345 (2021). https://doi.org/10.1186/s12911-021-01704-6.

[5] Joe G. Zein, Chao-Ping Wu, Amy H. Attaway, Peng Zhang, Aziz Nazha, Novel Machine Learning Can Predict Acute Asthma Exacerbation, Chest, Volume 159, Issue 5, 2021, Pages 1747-1757, ISSN 0012-3692, https://doi.org/10.1016/j.chest.2020.12.051.

[6] Darsha Jayamini, W.K., Mirza, F., Asif Naeem, M. et al. Investigating Machine Learning Techniques for Predicting Risk of Asthma Exacerba- tions: A Systematic Review. J Med Syst 48, 49 (2024).

[7] Xiong, S., Chen, W., Jia, X. et al. Machine learning for predic- tion of asthma exacerbations among asthmatic patients: a system- atic review and meta-analysis. BMC Pulm Med 23, 278 (2023). https://doi.org/10.1186/s12890-023-02570-w

[8] Ekpo, Raphael Henshaw, et al. "Machine learning classification approach for asthma prediction models in children." Health and Technology 13.1 (2023): 1-10.

[9] Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics Journal. 2019;25(3):811-827. doi:10.1177/1460458217723169

[10] Tahasamadsoltaniheris, M., Mahmoodvand, Z. and Zolnoori, M., 2013. Intelligent diagnosis of Asthma using machine learning algorithms. International Research Journal of Applied and Basic Sciences, 5(1), pp.140-145.