

# LUNGVISION: Seeing the Future of Asthma Risk

Deepta Susan Beji, Ericsson A. David, Fabricio Montero Zúñiga  
 Institute of Bioengineering, University of Tartu, Tartu, Estonia



UNIVERSITY OF TARTU

Institute of Bioengineering

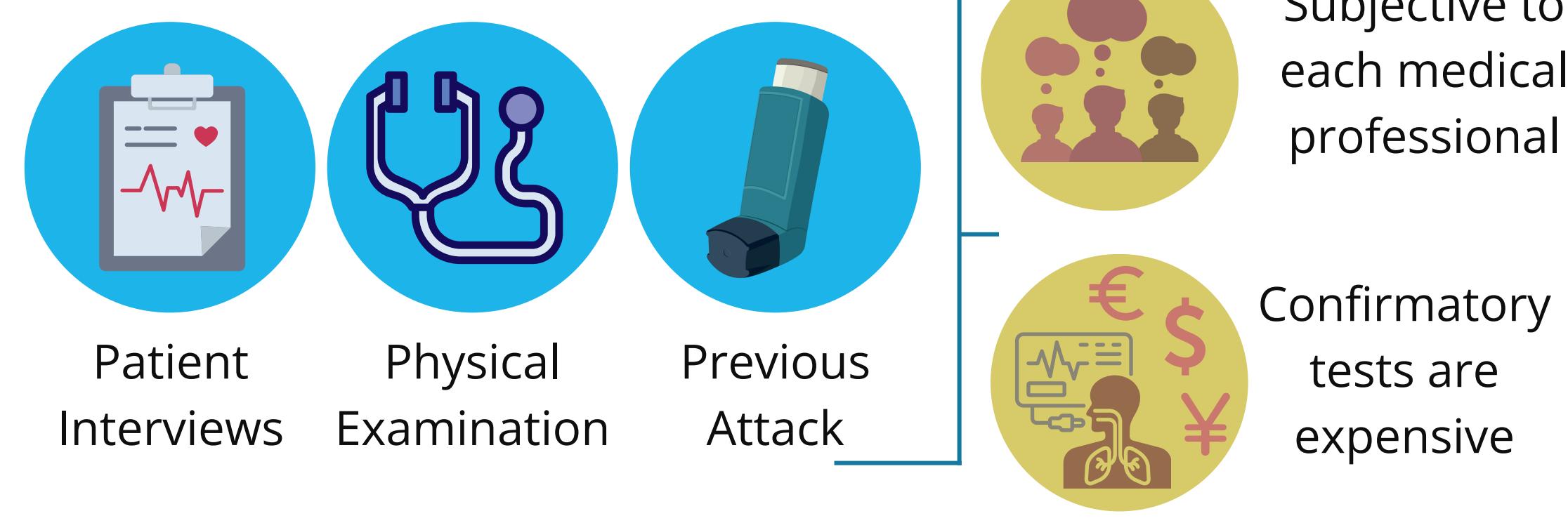
## ABSTRACT

Asthma is a major pulmonary condition affecting numerous individuals worldwide; thus a better systematic diagnostic approach is needed. In this study, the CDC's National Health and Nutrition Examination Survey (NHANES) data was used to identify predictors that may be used to diagnose the prevalence of asthma and its onset. Top predictors were shortlisted using Pearson correlation and chi-square statistical tests. Six models were tested and Logistic Regression had the best Accuracy (69%) and AUC (0.63) for asthma risk prediction. For onset prediction, Logistic Regression achieved optimal performance (Accuracy = 76%, AUC = 0.82), outperforming XGBoost, Random Forest, and ExtraTrees models. SHapley Additive exPlanations (SHAP) feature importance analysis identified four key predictors of childhood-onset asthma: BMI, respiratory symptoms, cholesterol levels, and hypertension status. The weak statistical significance of the predictors highlights limitations in the current data and underscores the need for focused data collection in asthma prevention and management strategies.

## INTRODUCTION

Asthma is a chronic inflammatory disorder of the airways resulting in variable airflow bronchial obstruction which is potentially reversible<sup>1</sup>.

### CURRENT FIRST-LINE DIAGNOSIS FOR ASTHMA



Certain Biomarkers are available that have correlation to asthma but it fails to differentiate between childhood or adult-onset which require different approaches for treatment or care<sup>2</sup>.

### PROPOSED PREDICTOR-BASED PROTOCOL



## METHODOLOGY

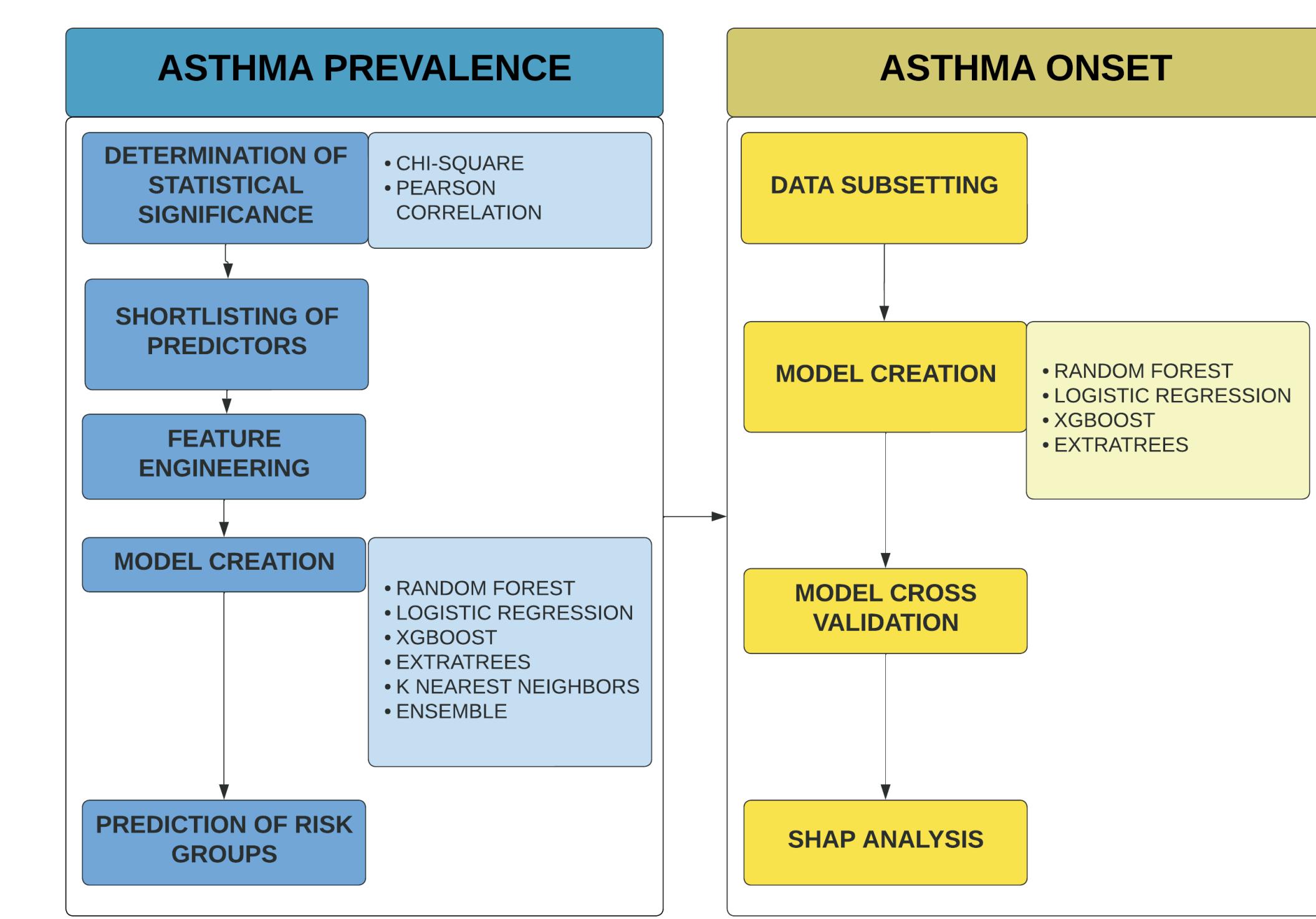
### DATASET INFORMATION

The CDC conducts the National Health and Nutrition Examination Survey (NHANES) which is the only national health survey that includes health exams, laboratory tests, and dietary interviews for participants of all ages in the United States. The survey is done yearly with around 5000 participants. In this study the 2011-2014 datasets were utilized.

(<https://www.cdc.gov/nchs/nhanes/default.aspx>)



### PROJECT WORKFLOW



### REFERENCES

- [1] World Health Organization. (2013). *Diagnosis and Management of Asthma: Primary health care service Delivery Iraq 2013 guidelines*. Iraq: World Health Organization.
- [2] Buhl, R., Korn, S., Menzies-Gow, A., Aubier, M., Chapman, K. R., Canonica, G. W., Picado, C., Donica, M., Kuhlbusch, K., Korom, S., & Hanania, N. A. (2020). Prospective, single-arm, Longitudinal Study of biomarkers in real-world patients with severe asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(8). <https://doi.org/10.1016/j.jaip.2020.03.038>

## RESULTS AND DISCUSSION

### ASTHMA RISK CLASSIFICATION MODEL

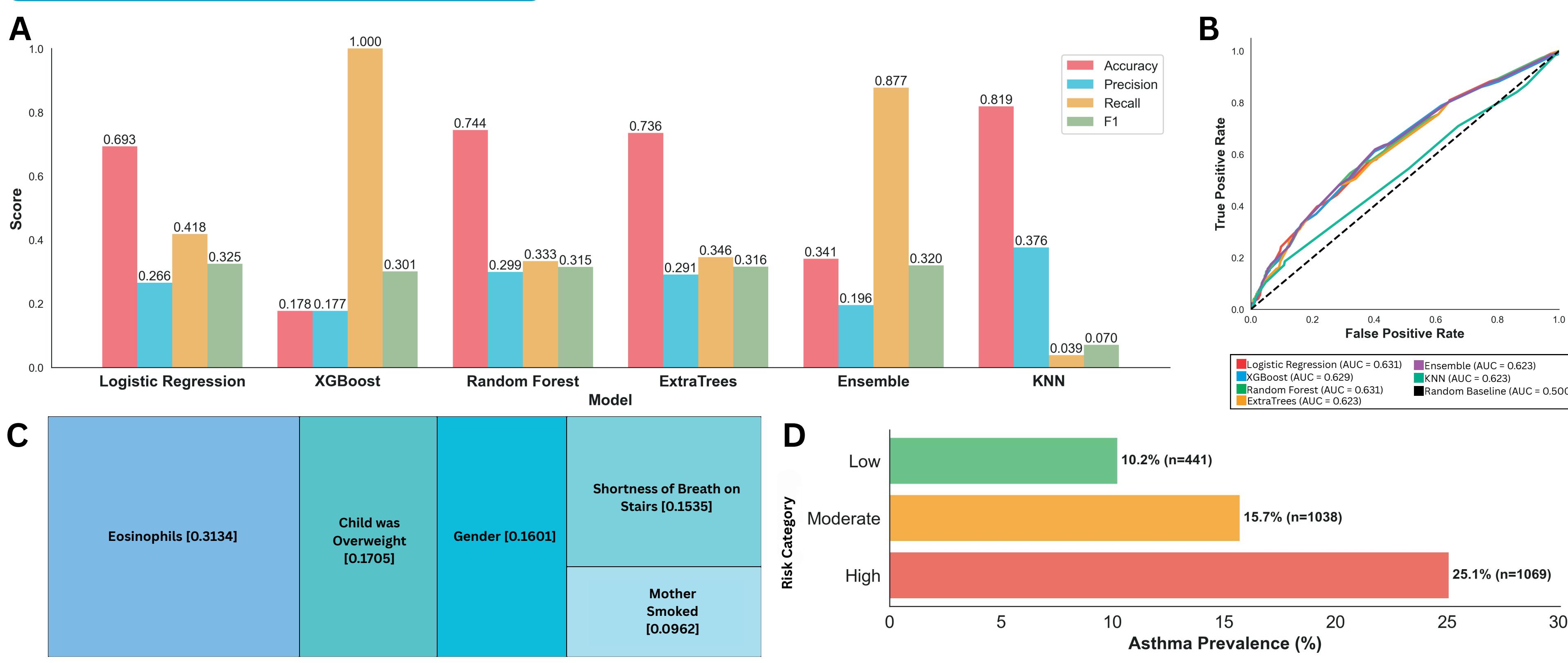


Figure 1. (A) Performance metrics for asthma risk classification models. (B) Model AUC comparison. (C) Top predictors from Logistic Regression. (D) Asthma prevalence across risk categories.

Several models were trained and tested. Among them, XGBoost (Fig. 1A) achieved a remarkable recall score of 1.000, while KNN exhibited a respectable accuracy of 0.819. However, Logistic Regression had the best overall performance with an accuracy of 69% and AUC of 0.63 (Fig. 1B) for asthma risk prediction. Moreover, top predictors of asthma risk were identified (Fig. 1C). In addition, the model was able to classify individuals into three different risk groups (Fig. 1D).

### ASTHMA-ONSET MODEL

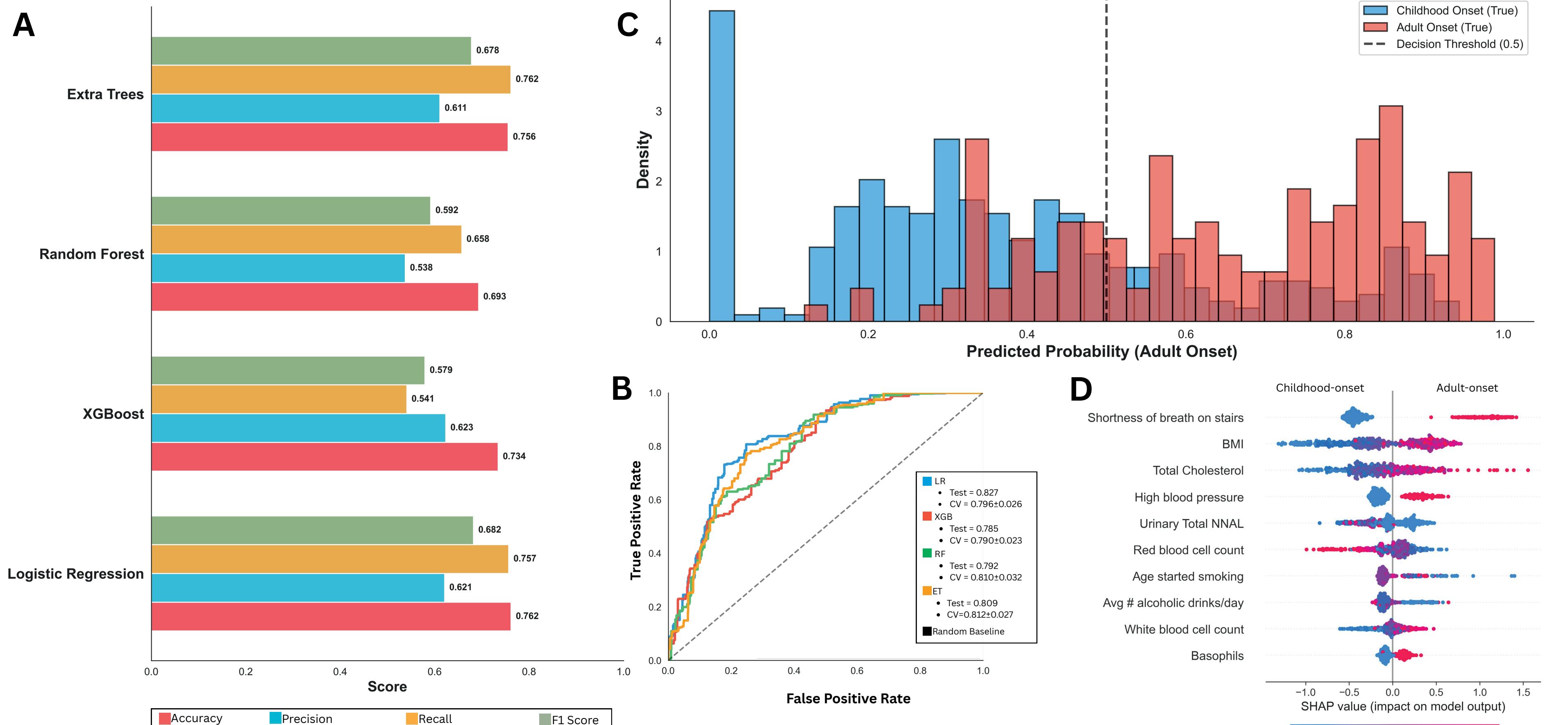


Figure 2. (A) Performance metrics for asthma onset classification models. (B) Model AUC comparison. (C) Probability distribution of predictions. (D) Key onset predictors from SHAP analysis.

Among the models tested for onset prediction, Logistic Regression (Fig. 2A) achieved the highest accuracy (76.2%) and AUC as verified in the ROC curve (Fig. 2B). Also, as shown in Fig. 2C, model predictions show strong confidence with distinct peaks: childhood cases near 0 and adult cases near 1. SHAP analysis reveals basophils and high blood pressure separating the different onset cases (Fig. 2D).

## CONCLUSION AND RECOMMENDATIONS

- Among all the different trained models, Logistic Regression provided the most reliable performance for both asthma risk prediction and onset classification on the NHANES data. Additionally, SHAP analysis identified BMI, respiratory symptoms, cholesterol levels, and hypertension status as key predictors of childhood-onset asthma.
- Weak statistical significance across predictors suggests that the current dataset may be insufficient for high-precision diagnostics. Hence, for future projects asthma-specific data must be collected.

### ACKNOWLEDGEMENT

UNIVERSITY OF TARTU  
 Institute of Computer Science

