WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2026 Volume 11, Issue 5
Advances in Precision Medicine
ISSN: 2424-9106 (Online)
ISSN: 2424-8592 (Print)

Development and Validation of an Interpretable Machine Learning Model for Predicting Sleep Disorder Risk in Asthma Patients: A Nationwide Retrospective Cohort Study

Wenhao Zhao[1], Shuang Li[2], Ziyue Cui[1], Mengfan Zhang[1], Zhizheng Zhao[1], Shaobo Lyu[1]*, Lei Wu[1]*

[1]Hebei Key Laboratory of Mental Health and Brain Science, School of Psychology and Mental Health, North China University of Science and Technology, Tang'shan 063210, Hebei, China

[2]School of Public Health, North China University of Science and Technology, Tang'shan 063210, Hebei, China

Author to whom correspondence should be addressed.

Abstract: *Background:* Sleep disorders are prevalent among patients with asthma and negatively impact disease control, quality of life, and clinical outcomes. This study aimed to develop and validate a machine learning-based model for predicting sleep disorder risk in this population using routinely accessible clinical and physiological biomarkers. *Methods:* Data from the NHANES 2005–2010 cohort were analyzed. LASSO regression was applied to identify key predictors. Seven machine learning algorithms were subsequently developed and compared. Model performance was evaluated based on discrimination, calibration, and net clinical utility. SHAP analysis was used to interpret the optimal model and provide individualized risk assessments. *Results:* A total of 1,059 asthma patients were included and randomly divided into training and validation sets. LASSO regression identified seven predictors: hypertension status, congestive heart failure, asthma attack in the past year, serum iron, BMI, HDL-C, and eGFR. In the validation set, the optimal model, random forest, demonstrated the highest AUC of 0.800 (95% CI: 0.772–0.828), with satisfactory calibration (Brier score: 0.164) and positive net benefit across a range of threshold probabilities in decision curve analysis. SHAP analysis revealed complex non-linear and threshold effects for metabolic and renal predictors. *Conclusion:* We developed and validated an interpretable, machine learning-based prediction model for sleep disorder risk in asthma patients using seven accessible clinical and laboratory parameters, offering a practical tool for early identification and targeted interventions in clinical practice.

Keywords: Sleep disorder; Machine learning; Prediction model; Asthma

Online publication: May 26, 2026

1. Introduction

Asthma is a chronic respiratory disease characterized by persistent airway inflammation, airway hyperresponsiveness, and reversible airflow limitation. Asthma constitutes a formidable global public health challenge, with a prevalence exceeding 300 million individuals worldwide. Quantified by disability-adjusted life years (DALYs), the disease accounts for approximately 1.1% of the aggregate global disease burden [1]. This large patient population poses significant challenges

in medical resource allocation and disease management, requiring substantial human and material investments for asthma diagnosis, treatment, and rehabilitation.

Sleep disorders are highly prevalent among patients with asthma, constituting a critical yet frequently overlooked component of the disease burden. Studies indicated that the prevalence of clinically significant sleep disorders ranges around 35% in asthma patients, a rate substantially higher than in the general population [2]. Moreover, the presence of sleep disorders is strongly associated with worse asthma control, increased frequency of asthma exacerbations, greater healthcare utilization, impaired lung function variability, and reduced responsiveness to standard pharmacotherapy [3–7]. Importantly, this relationship is bidirectional; uncontrolled asthma worsens sleep, while poor sleep quality may exacerbate airway inflammation and symptom perception, creating a detrimental cycle [8]. Consequently, early identification of asthma patients with high risk for sleep disorders and the development of integrated, personalized management strategies are essential for breaking this cycle, optimizing asthma outcomes, and improving long-term patient well-being.

Despite its clinical importance, systematic screening for sleep disorders in respiratory outpatient clinics remains suboptimal, partly owing to limited time and expertise during routine visits [9]. Accordingly, this study investigates risk factors for sleep disorders in patients with asthma using large-scale data from the United States National Health and Nutrition Examination Survey (NHANES). It also employs readily available clinical and laboratory indicators to build and validate a multivariate prediction model for sleep disorders in patients with asthma. This predictive approach would enable clinicians to identify high-risk individuals promptly and implement targeted interventions to improve sleep quality and potentially reduce the downstream health impacts associated with poor sleep in asthma.

2. Methods

2.1. Data sources

Data for this cross-sectional study were obtained from three continuous cycles of NHANES spanning 2005–2010. The NCHS Research Ethics Review Board approved all study protocols, and written informed consent was obtained from all participants. Detailed methodologies and datasets are publicly available on the NCHS website (http://www.cdc.gov/nchs/nhanes/index.htm).

2.2. Study design and sample screening

The study population consisted of adult participants (aged ≥ 20 years) diagnosed with asthma. Asthma was determined by affirmative responses to two sequential questions: “Has a doctor or other health professional ever told you that you have asthma?” and “Do you still have asthma?” [10]. Strict exclusion criteria were applied as follows. The detailed flowchart of participant selection is illustrated in **Figure 1**.

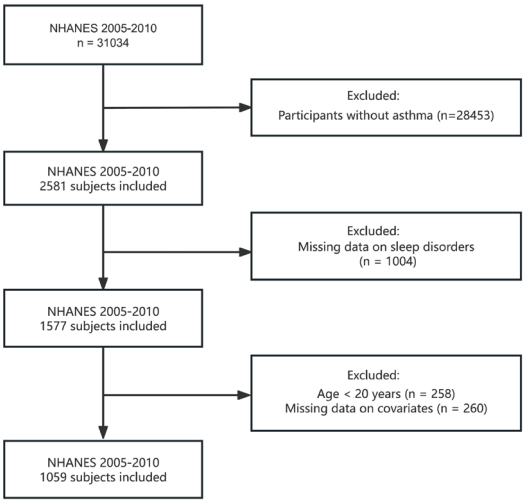

Figure 1. Flowchart of participant selection from NHANES 2005–2010.

2.3. Diagnosis of sleep disorders

The primary outcome of this study was the presence of sleep disorders, assessed using the standardized Sleep Disorders Questionnaire module. Specifically, sleep disorder status was determined based on the variable SLQ060, which asks: "Has a doctor or other health professional ever told you that you have a sleep disorder?" Consistent with previous methodology, participants who responded "Yes" were classified as the positive group, while those answering "No" were classified as controls [11].

2.4. Candidate prediction variables

A comprehensive set of candidate predictors was extracted from the NHANES database, encompassing sociodemographic characteristics, lifestyle factors, clinical comorbidities, and laboratory indices.

2.5. Statistical analysis

All statistical analyses were conducted in accordance with the analytic guidelines of the CDC. All analyses conformed to the established standards.

To assess model generalizability and prevent overfitting, the analytical sample was randomly partitioned into a training set (70%) and a validation set (30%) [12]. To optimize feature selection and reduce data dimensionality, we employed least absolute shrinkage and selection operator (LASSO) logistic regression via the glmnet package (version 4.1.9) in R.

Following the identification of the optimal feature subset, seven machine learning classifiers were developed to predict the risk of sleep disorders in asthma patients. All models were trained exclusively on the training set. To ensure robust performance and derive optimal model configurations, hyperparameter tuning was systematically executed using grid search coupled with 5-fold cross-validation. The area under the receiver operating characteristic curve (ROC) was designated as the primary optimization metric during this tuning process.

To comprehensively evaluate the generalization performance of the trained models, their discrimination, calibration, and clinical utility were rigorously assessed on the independent validation set. Recognizing that the inherent "black-box" nature of advanced machine learning algorithms often limits the interpretability of how individual risk variables influence predictions, we sought to elucidate the underlying prediction mechanism of the optimal model. To this end, a game-theoretic approach based on SHAP analysis was performed using the Python shap library (version 0.47.0).

3. Results

3.1. Baseline characteristics of the study population

Baseline sociodemographic and clinical characteristics of the 1,059 adult asthma patients included in this analysis are summarized in **Table 1**. The entire analytic cohort (N = 1,059) was randomly partitioned into a training set (70%, n = 742) and an independent validation set (30%, n = 317).

Table 1. Baseline information included in this study

Variable	**Overall (n = 1059)**	**Non sleep disorders (n = 880)**	**Sleep disorders (n = 179)**	***p***
Gender (%)				0.508
male	394 (37.20)	323 (36.70)	71 (39.66)	
female	665 (62.80)	557 (63.30)	108 (60.34)	
Race (%)				0.557
Mexican American	91 (8.59)	77 (8.75)	14 (7.82)	
Other Hispanic	84 (7.93)	67 (7.61)	17 (9.50)	

Non-Hispanic White	584 (55.15)	494 (56.14)	90 (50.28)	
Non-Hispanic Black	253 (23.89)	204 (23.18)	49 (27.37)	
Other Race	47 (4.44)	38 (4.32)	9 (5.03)	
Hypertension.Status (%)				< 0.001
No	556 (52.50)	492 (55.91)	64 (35.75)	
Yes	503 (47.50)	388 (44.09)	115 (64.25)	
Congestive heart failure (%)				< 0.001
No	989 (93.39)	838 (95.23)	151 (84.36)	
Yes	70 (6.61)	42 (4.77)	28 (15.64)	
asthma.attack.last.year (%)				0.001
No	542 (51.18)	472 (53.64)	70 (39.11)	
Yes	517 (48.82)	408 (46.36)	109 (60.89)	
Age (median [IQR])	49.00 [34.00, 63.00]	48.00 [33.00, 63.25]	52.00 [42.50, 62.00]	0.028
BMI (median [IQR])	30.33 [25.54, 35.27]	29.63 [25.07, 34.44]	34.40 [29.16, 40.42]	< 0.001
HbA1c (median [IQR])	5.50 [5.20, 5.90]	5.50 [5.20, 5.90]	5.60 [5.30, 6.20]	< 0.001
HDL.C (median [IQR])	51.00 [42.00, 62.00]	52.00 [42.75, 63.00]	47.00 [37.00, 58.00]	< 0.001
eGFR (median [IQR])	94.42 [77.31, 109.60]	95.12 [77.51, 110.67]	90.52 [75.19, 104.95]	0.028

3.2. Prediction variables screening

To identify the optimal subset of predictors for model development, LASSO logistic regression incorporating 10-fold cross-validation was performed on the training cohort. At the optimal tuning parameter (λ_min) that minimized the binomial deviance, seven variables retained non-zero coefficients and were consequently selected as the final input features **(Figure 2 A, B)**. These critical predictors comprised a history of hypertension, congestive heart failure, recent asthma attacks, BMI, serum iron, HDL-C, and eGFR.

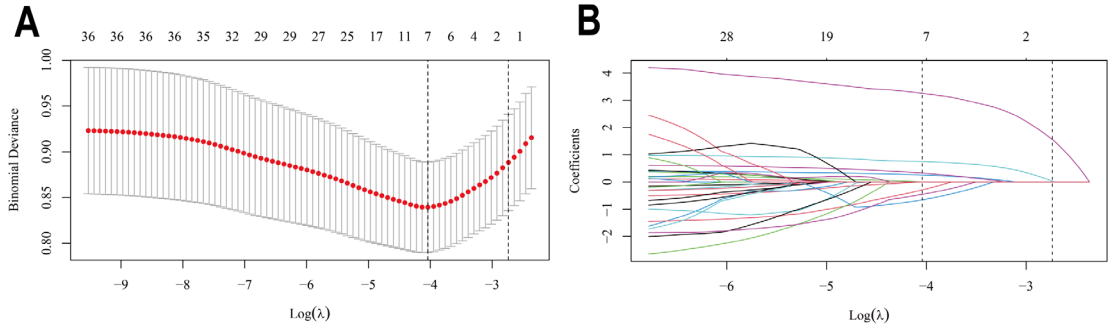

Figure 2. Feature selection using LASSO logistic regression.

3.3. Prediction model development and performance comparison

Utilizing the optimal subset of seven predictor variables, we developed predictive models on the training set using seven machine learning algorithms. Model performance within the training cohort was illustrated in **Figure 3A-C**, which respectively depict the ROC curves, Calibration plots, and DCA.

To rigorously assess model generalizability, the predictive performance of the seven machine learning classifiers was further evaluated using the independent validation set. The detailed performance metrics for each model were summarized in **Table 2**. Among these models, the Random Forest model demonstrated superior discriminative ability, achieving the highest area under the receiver operating characteristic curve (AUC = 0.800, 95% CI: 0.772–0.828), along with a balanced F1 score (0.638), sensitivity (55.8%), specificity (89.9%), and a favorable Youden's index (0.457). Although LightGBM

and ANN exhibited comparable discriminative performance, the Random Forest model showed better calibration (Brier score: 0.164) and yielded a greater net clinical benefit across a wider range of threshold probabilities. Consequently, considering its robust discrimination, excellent calibration, and superior clinical utility, the Random Forest algorithm was selected as the optimal prediction model. Its ROC curve, calibration plot, and decision curve analysis derived from the validation set are presented in **Figure 3D- F**.

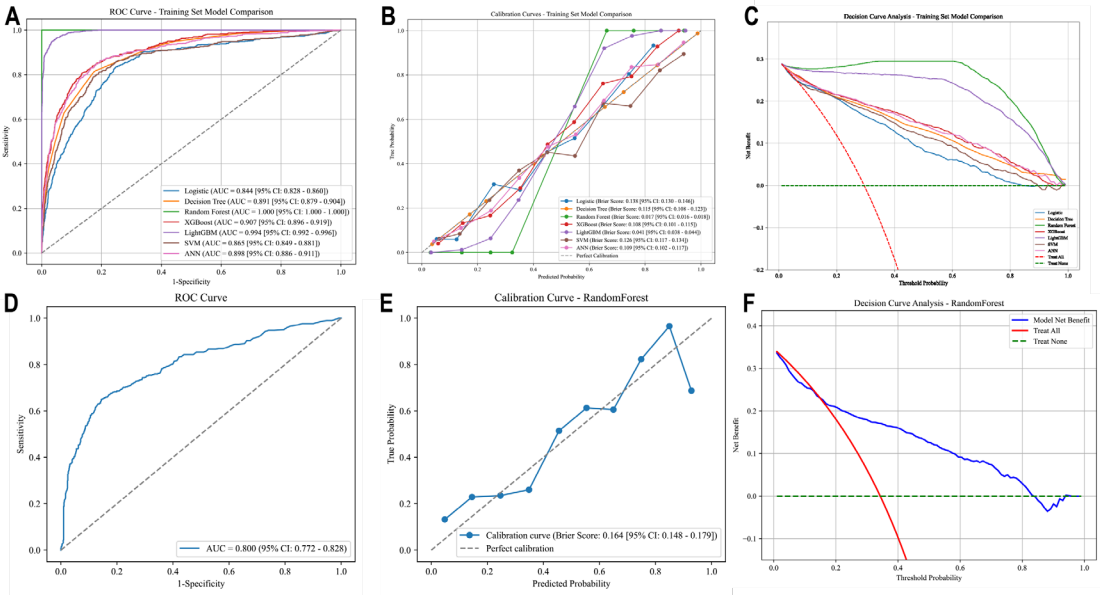

Figure 3. Performance evaluation of the developed machine learning models.

Table 2. Performance of classifier models under 5-fold cross-validation on the validation set

Model	Logistic	Decision tree	Random forest	XGBoost	LightGBM	SVM	ANN
AUC (95%CI)	0.7760 (0.7455, 0.8067)	0.7838 (0.7556, 0.8113)	0.8004 (0.7719, 0.8280)	0.7814 (0.7531, 0.8103)	0.7893 (0.7598, 0.8180)	0.7754 (0.7440, 0.8058)	0.7956 (0.7672, 0.8230)
Accuracy	0.7527	0.7736	0.7808	0.7455	0.7790	0.7582	0.7736
Precision	0.6608	0.7661	0.7448	0.7244	0.7285	0.6727	0.7185
Sensitivity	0.5864	0.4974	0.5576	0.4267	0.5759	0.5864	0.5681
Specificity	0.8407	0.9197	0.8989	0.9141	0.8864	0.8490	0.8823
F1 Score	0.6214	0.6032	0.6377	0.5371	0.6433	0.6266	0.6345
Kappa	0.4387	0.4546	0.4852	0.3773	0.4863	0.4490	0.4737
Youden's J	0.4271	0.4170	0.4565	0.3408	0.4623	0.4354	0.4503
PPV	0.6608	0.7661	0.7448	0.7244	0.7285	0.6727	0.7185
NPV	0.7935	0.7757	0.7934	0.7509	0.7980	0.7951	0.7943

3.4. SHAP analysis

To interpret the optimal Random Forest model, we applied SHAP analysis to quantify the contribution of each predictor. As shown in **Figure 4A**, BMI exhibited the highest mean absolute SHAP value (0.11), indicating that it was the most influential feature driving the model's predictions. The SHAP summary beeswarm plot **(Figure 4B)** further revealed the directionality of these effects. Furthermore, **Figure 4C** presents the SHAP dependence plots to illustrate the complex trajectories of individual features. BMI exhibited a clear threshold effect, yielding consistently positive SHAP contributions when values exceeded approximately 35 kg/m^2. HDL-C demonstrated a monotonic negative association, where higher levels corresponded to lower

predicted risk, while eGFR showed modest positive SHAP attributions at low levels (< 60 mL/min/1.73 m^2) that gradually shifted to negative values with improving renal function. To translate the global model into actionable clinical insights, SHAP waterfall and force plots were employed for individualized risk assessment **(Figure 4D, E)**.

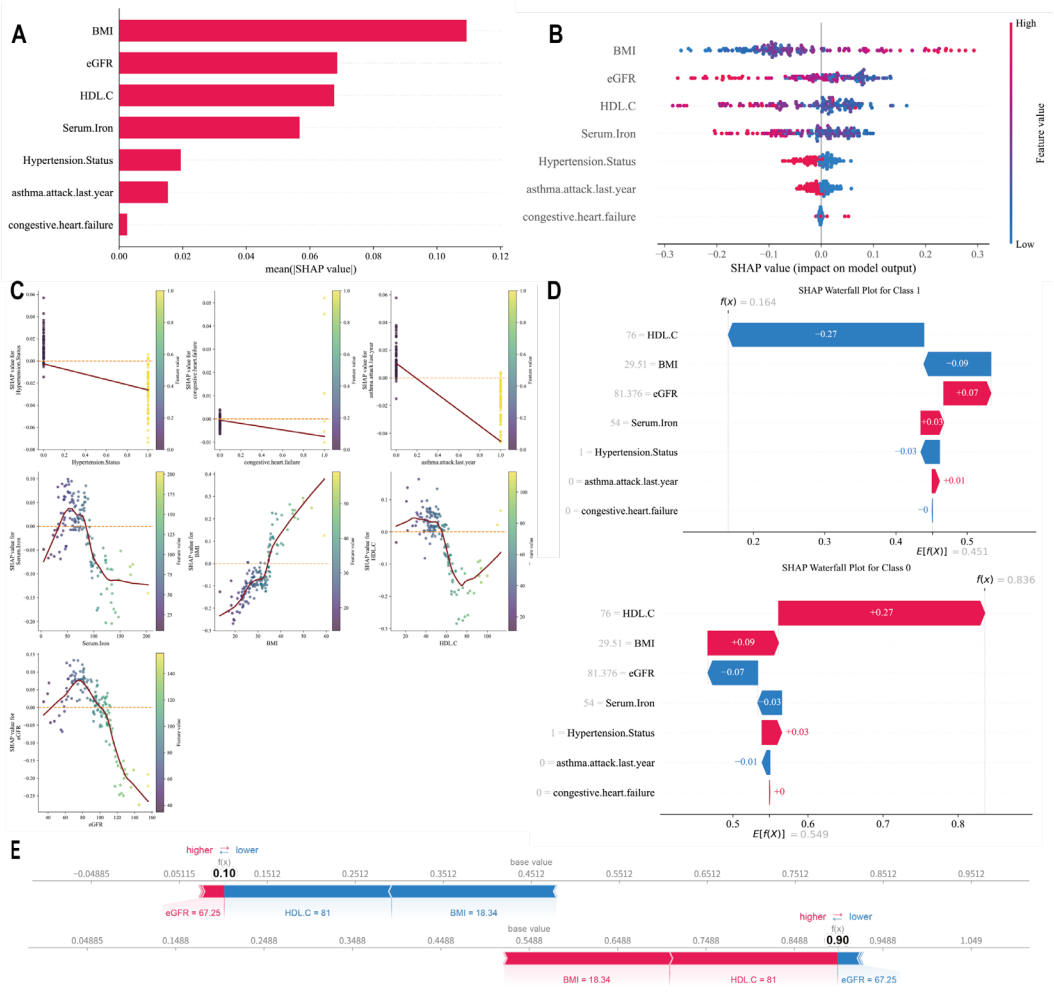

Figure 4. SHAP-based feature attribution for the optimal RandomForest model. (A) Bar plot of mean absolute SHAP values ranking global feature importance. (B) SHAP summary beeswarm plot. (C) SHAP dependence plots. (D) SHAP waterfall plots. (E) SHAP force plots.

4. Discussion

There is a complex bidirectional relationship between asthma and sleep disorders. Nocturnal asthma symptoms—including cough, wheezing, and dyspnea, directly disrupt sleep architecture by triggering frequent awakenings and reducing slow-wave sleep [13]. Conversely, sleep fragmentation and deprivation can alter autonomic nervous system function, promote systemic inflammation, and increase airway hyperresponsiveness, thereby exacerbating asthma control [14]. Despite growing mechanistic insights, the translation of this knowledge into routine clinical practice remains limited. In clinical settings, patients with comorbid asthma and sleep disorders frequently remain unrecognized and untreated. Hence, in this study, we introduced a transparent, machine learning-based prediction model for sleep disorder risk in asthma patients, incorporating seven routinely available clinical and laboratory indicators. This model may assist clinicians in the early identification of at-risk patients, facilitate timely and targeted interventions, improve patient sleep quality and asthma control, and ultimately enhance overall health outcomes.

Nevertheless, several limitations warrant consideration. First, although our model demonstrated robust internal

validity, it was derived solely from the NHANES database. The absence of external validation using independent, geographically diverse datasets may limit the generalizability of our findings to broader global populations. Second, the study population was restricted to adults aged 20 years and older, and children with asthma were not included, which may introduce selection bias and limit the applicability of our model to pediatric populations. Third, due to the cross-sectional nature of NHANES, temporal or causal relationships cannot be established, and reverse causation remains a possibility.

5. Conclusion

This study established and validated a practical and highly interpretable clinical prediction model for assessing sleep disorder risk in patients with asthma. By incorporating seven routinely collected clinical and laboratory parameters, this model enables the prompt identification of individuals at elevated risk. Designed to be straightforward, transparent, and clinically accessible, it holds significant promise for reducing the under-recognition of sleep disorders among asthmatic patients and facilitating the prioritization of personalized, targeted interventions.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Oh J, Kim S, Kim M, et al., 2025, Global, Regional, and National Burden of Asthma and Atopic Dermatitis, 1990–2021, and Projections to 2050: A Systematic Analysis of the Global Burden of Disease Study 2021. The Lancet Respiratory Medicine, 13: 425–446.

[2] Liu T, Qin H, 2025, Association of Obstructive Sleep Apnea Risk with Allergic Asthma: A Systematic Review and Meta-Analysis. Medicine, 104: e41918.

[3] Luyster F, Shi X, Baniak L, et al., 2020, Associations of Sleep Duration with Patient-Reported Outcomes and Health Care Use in US Adults with Asthma. Annals of Allergy, Asthma & Immunology, 125: 319–324.

[4] Mindus S, Gislason T, Benediktsdottir B, et al., 2024, Respiratory Symptoms, Exacerbations and Sleep Disturbances Are More Common Among Participants with Asthma and Chronic Airflow Limitation: An Epidemiological Study in Estonia, Iceland and Sweden. BMJ Open Respiratory Research, 11: e002063.

[5] Kumar P, Goyal J, Singh K, 2022, Association of Symptoms of Sleep-Related Breathing Disorders with Asthma Control in Indian Children. Indian Journal of Pediatrics, 89: 31–36.

[6] Capanoglu M, Ginis T, Buyuktiryaki B, et al., 2024, The Frequency of Sleep-Disordered Breathing in Preschool Children with Asthma and Its Effects on Control of Asthma. Allergologia et Immunopathologia, 52: 38–43.

[7] Xu X, Tran T, Golam S, et al., 2025, Systemic Corticosteroid Dose–Response Effects in Asthma: An Observational Cohort Study. ERJ Open Research, 11: 00172-2024.

[8] Yau E, Cheung J, Saini B, et al., 2025, Understanding How Sleep Disorders Are Managed in People with Asthma: A Scoping Review of the Literature. Journal of Asthma, 62: 1266–1279.

[9] Selsick H, Heidbreder A, Ellis J, et al., 2024, Assessment and Management of Chronic Insomnia Disorder: An Algorithm for Primary Care Physicians. BMC Primary Care, 25: 138.

[10] Yang M, Qian X, Zhu L, et al., 2025, Associations of Serum Neurofilament and Asthma: Evidence from the NHANES 2013–2014. Frontiers in Neurology, 16: 1567158.

[11] Wang H, Zhang C, Chen B, et al., 2025, Development, Validation, and Visualization of a Machine Learning-Based Predictive Model for Depression Risk in Sleep Disorder Patients. BMC Psychiatry, 26: 91.

[12] Xu Y, Goodacre R, 2018, On Splitting Training and Validation Set: A Comparative Study of Cross Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing, 2: 249–262.

[13] Brockmann P, Sanchez T, Castro-Rodriguez J, 2016, Sleep-Disordered Breathing in Children with Asthma: A Systematic Review on the Impact of Treatment. Journal of Asthma and Allergy, 2016: 83.

[14] Yau E, Cheung J, Saini B, et al., 2025, Understanding How Sleep Disorders Are Managed in People with Asthma: A Scoping Review of the Literature. Journal of Asthma, 62: 1266–1279.