scientific reports

Check for updates

OPEN

Non-invasive acoustic classification of adult asthma using an XGBoost model with vocal biomarkers

Yi Lyu[1,2,6], Quan-Cheng Jiang[1,6], Shuai Yuan[3,6], Jing Hong[1,2], Chun-Feng Chen[4], Hai-Mei Wu[4], Yi-Qin Wang[1,2], Yu-Jing Shi[5], Hai-Xia Yan[1,2✉] & Jin Xu[1,2✉]

Traditional diagnostic methods for asthma, a widespread chronic respiratory illness, are often limited by factors such as patient cooperation with spirometry. Non-invasive acoustic analysis using machine learning offers a promising alternative for objective diagnosis by analyzing vocal characteristics. This study aimed to develop and validate a robust classification model for adult asthma using acoustic features from the vocalized /ɑː/ sound. In a case-control study, voice recordings of the /ɑː/ sound were collected from a primary cohort of 214 adults and an independent external validation cohort of 200 adults. This study extracted features using a modified extended Geneva Minimalistic Acoustic Parameter Set and compared seven machine learning models. The top-performing model, Extreme Gradient Boosting, was further assessed through ten-fold cross-validation, external validation, and feature analysis using SHapley Additive exPlanations and Local Interpretable Model-Agnostic Explanations. The Extreme Gradient Boosting classifier achieved the highest performance on the test set, with an accuracy of 0.8514, an Area Under the Curve of 0.9130, a recall of 0.8804, a precision of 0.8387, an F1-score of 0.8567, a Kappa coefficient of 0.7018, and a Matthews Correlation Coefficient of 0.7071. On the external validation set, the model maintained strong performance with an accuracy of 0.8100, AUC of 0.8755, recall of 0.8300, precision of 0.7981, F1-score of 0.8137, Kappa of 0.6200, and Matthews Correlation Coefficient of 0.6205. Interpretability analysis identified formant frequencies as the most significant acoustic predictors. An Extreme Gradient Boosting model utilizing features from the extended Geneva Minimalistic Acoustic Parameter Set is an accurate and viable non-invasive method for classifying adult asthma, holding significant potential for developing accessible tools for early diagnosis, remote monitoring, and improved asthma management.

Abbreviations

Acc	Accuracy
AdaBoost	Adaptive boosting
AI	Artificial intelligence
AUC	Area under curve
BART	Bayesian additive regression trees
CLSA	Computerized lung sound analysis
CNN	Convolutional neural network
ComParE	Computational paralinguistics challenge
COPD	Chronic obstructive pulmonary disease
CQCCs	Constant-Q cepstral coefficients
DL	Deep learning
DT	Decision tree
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set

[1]School of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, People's Republic of China. [2]Shanghai Key Laboratory of Health Identification and Assessment, Shanghai 201203, People's Republic of China. [3]Yangzhou Municipal Center for Disease Control and Prevention, Yangzhou 225007, People's Republic of China. [4]Shanghai Lingyun Community Health Service Center, Shanghai 200237, People's Republic of China. [5]Affiliated hospital of Nanjing University of Chinese Medicine, Nanjing 210029, People's Republic of China. [6]Yi Lyu, Quan-Cheng Jiang, and Shuai Yuan contributed equally to this work. ✉email: hjy2012ok@163.com; xujin3264@hotmail.com

nature portfolio

1

ET	Extra trees
GeMAPS	Geneva minimalistic acoustic parameter set
GMM	Gaussian mixture models
GWAS	Genome-wide association studies
HNR	Harmonics-to-noise ratio
ISCB	INTERSPEECH 2013 computational paralinguistics challenge baseline
InterSp12	Interspeech 2012 speaker trait set
KNN	K-nearest neighbors
KS	Kolmogorov–Smirnov
LightGBM	Light gradient boosting machine
LLDs	Low-level descriptors
MCC	Matthews correlation coefficient
MFCC	Mel-frequency cepstral coefficients
ML	Machine learning
NB	Naive Bayesian
PEF	Peak expiratory flow
Prec.	Precision
PSO	Particle swarm optimization
RF	Random forest
RFR	Random forest regressor
RNN	Recurrent neural network
ROC	Receiver operating characteristic (implied by "ROC curves")
SHAP	SHapley Additive explanation
SpO_2	Peripheral capillary oxygen saturation
Std	Standard deviation
SVM	Support vector machine
TCA	Total classification accuracy
TT	Total computation time
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

Asthma, a prevalent and heterogeneous respiratory disease, affects over 300 million individuals globally[1]. Clinical observations indicate that symptom manifestation can occur at any age[2,3]. However, studies suggest that a majority of patients diagnosed with asthma exhibit airway abnormalities during childhood[4,5], establishing it as the most common chronic disease in children[1]. Although asthma's prevalence is not specific to age or race, evidence indicates a higher susceptibility among individuals from lower socioeconomic backgrounds[6,7]. As a chronic condition, asthma significantly impacts patients' physical health and exerts a substantial influence on societal, financial, and healthcare systems in numerous countries[2,8,9]. In view of the pervasive impact of asthma and the effect of pathological changes on voice, researchers have initiated the exploration of non-invasive diagnostic methods based on patients' vocal characteristics. By analysing voice signals such as wheezing and coughing, these methods can identify airway abnormalities, providing new ways to intervene at an early stage.

As a common condition caused by chronic inflammation of the lower respiratory tract, it is associated with allergy and features chronic airway inflammation, airway hyperresponsiveness, variable reversible airway limitation, and airway remodeling as the disease progresses[10]. Clinically, asthma presents with recurrent episodes of wheezing, coughing, chest tightness, and shortness of breath. These symptoms result from airway narrowing due to persistent inflammation, which in turn exacerbates airway hyperresponsiveness[11,12]. While the precise pathogenesis of asthma remains incompletely understood, it is acknowledged that gene-environment interactions are significant risk factors for its onset[3,11]. Environmental factors, including but not limited to air pollution, microorganisms[13,14], dietary changes[15], obesity[16] and others, can increase the risk of asthma attack. In recent years, genome-wide association studies (GWAS) have identified the influence of several genes on the heritability of asthma[11,17,18], yet these studies have not fully elucidated the mechanisms of asthma development. It has been hypothesized that the pathogenesis of asthma is attributable to the disruption of immune-inflammatory pathway balance[13]. Nonetheless, the potential inflammation may exist in the absence of symptoms, so inflammation is the core of asthma diagnosis and management[11].

Lung function tests, such as spirometry and peak flow meters, are widely regarded as the gold standard for asthma diagnosis. However, spirometry accuracy depends on patient cooperation and physician technical proficiency, which can be challenging to achieve, particularly in children and the elderly who may exhibit poor compliance. Furthermore, peak flow meters fails to measure flow rate in the minor airways which is affected by asthma[19]. Respiratory diseases result in pathological changes within the lungs, which can be detected by acoustic diagnostic methods[20]. These changes include localized or generalized airway obstruction or narrowing, abnormal airway opening, and secretions that produce distinct breathing sounds. The most prevalent clinical abnormal sound is a decrease in intensity, with wheeze and coarse crackles being commonly encountered in obstructive airway diseases[20,21], especially asthma, which frequency is usually below 400 Hz. However, many pathologic breathing sounds can be confused with wheeze, and there is an overlap of symptoms between asthma and chronic obstructive pulmonary disease (COPD)[22–24], so it is necessary to choose other more representative sounds for the diagnosis and classification of asthma. It is possible to develop asthma detection technology that does not require subjects to produce unnatural cough and wheeze by understanding the encoding forms of asthma-related acoustic features in different verbal sounds.

This study elected to utilize vocalized /ɑ:/ sounds, the feasibility of which has been confirmed[25], as the material for analysis. The objective study of acoustic diagnosis is facilitated by research into lung sound acquisition and Computerized Lung Sound Analysis (CLSA)[26,27]. The specific application of objective acoustic diagnosis in disease prediction, diagnosis and monitoring has been proven by numerous studies[12,28,29]. The subjective nature of traditional lung auscultation has been demonstrated to result in diagnostic delays and differences. Nevertheless, the capacity of CLSA to identify alterations in acoustic signals associated with airway inflammation has led to the conclusion that it can reduce the risk of asthma exacerbations through the provision of early warning and improve patient outcomes. CLSA quantifies voice characteristics by digitizing voices, achieving objective analysis of digitally recorded voice data using signal processing technology. The integration of voice diagnosis technology with machine learning (ML) and deep learning (DL) methodologies has the capability to propel it towards a state of automation[30,31]. The introduction of artificial intelligence (AI), ML, and DL[32] has significantly improved the accuracy and objectivity of asthma diagnosis by enabling deep pattern recognition and automated analysis of sound signals[30]. This approach effectively addresses the subjective nature of traditional auscultation, while also capturing faint sounds inaudible to the human ear. Furthermore, it optimises the poor correlation between doctors' subjective assessments of asthma severity and lung function. Advancements in technology have catalysed the evolution of auscultation, encompassing the transition from traditional methods to digital stethoscopes, and culminating in the advent of CLSA, a technological advancement underpinned by artificial intelligence and ML. These innovations have continuously enhanced the objectivity and automation of diagnostic analysis, thereby providing a technological foundation for remote monitoring and early warning systems[10], and reduce the demand for offline medical visits. Honorata Hafke-Dys et al. have confirmed that AI trained on 1,043 cases of auscultation data significantly outperforms traditional auscultation in distinguishing normal and abnormal breath sounds in asthma patients, and can achieve home asthma symptom monitoring[33]. The capacity of CLSA to detect subclinical inflammation or occult asthma in asymptomatic patients remains to be studied and validated.

ML and DL, subsets of AI, have demonstrated significant potential in the automatic detection and classification of abnormal breathing sounds and/or the diagnosis of lung pathologies[20]. Traditional asthma diagnostic methods rely on stethoscope auscultation and pulmonary function tests, which are subjective and time-consuming. ML algorithms offer a more objective and data-driven diagnostic approach for this prevalent disease. In recent years, ML has emerged as a powerful tool for assisting in asthma diagnosis, risk assessment, and treatment. As a popular tool for diagnosing pulmonary diseases, several ML models—such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Extra Trees (ET), and Extreme Gradient Boosting (XGBoost)—have proven more effective than traditional examination methods in distinguishing between normal and abnormal breathing sounds in asthma. While acoustic analysis is a promising alternative, existing approaches often rely on inconsistent biomarkers like coughs or wheezes, which can overlap with other respiratory conditions. Furthermore, previous studies utilizing more stable vocalizations have often employed traditional machine learning models that may not fully harness the predictive power of the acoustic data. This reveals a clear research gap: a need for a novel method that combines a stable, easily acquired vocal biomarker with a state-of-the-art classifier to improve diagnostic accuracy for adult asthma. Therefore, this study aims to address this gap by developing and validating an XGBoost model using features extracted from the sustained vowel /ɑ:/.

Objectives

The primary objectives of this study are as follows:

- To develop and validate an XGBoost-based classification model for the non-invasive diagnosis of adult asthma.
- To utilize acoustic features extracted from the vocalized /ɑ:/ sound as the primary input for the diagnostic model.
- To employ the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for feature extraction and assess its efficacy in distinguishing between asthmatic and healthy individuals.
- To compare the performance of the XGBoost model against six other machine learning algorithms to identify the optimal classifier for this task.
- To validate the final model's performance and generalizability using both ten-fold cross-validation and an independent external validation cohort.
- To identify the most significant acoustic features for asthma classification using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to better understand the acoustic changes associated with the condition.

Contributions

The main contributions of this study are as follows:

- This study developed a novel, non-invasive classification framework for adult asthma by uniquely combining acoustic analysis of a sustained vowel sound (/ɑ:/), the standardized extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for feature extraction, and an Extreme Gradient Boosting (XGBoost) model was developed.
- The model's performance and generalizability, not only through ten-fold cross-validation but also on an independent external cohort, where it maintained strong performance with an accuracy of 0.81 and an Area Under the Curve (AUC) of 0.8755 were rigorously validated.

- A quantitative analysis of the most significant acoustic biomarkers for adult asthma using SHAP and LIME was provided. The results identified formant frequencies as top predictors, offering deeper insight into the vocal changes associated with the condition.
- A systematic comparison of seven machine learning algorithms, empirically demonstrating the superior performance of the XGBoost classifier for this specific diagnostic task was conducted.

Paper structure

The remainder of this paper is organized as follows. The "Methods" section details the study participants, data collection protocol, feature extraction process, and the machine learning models employed. The "Results" section presents the comparative performance of the models, the validation results for the optimal XGBoost model, and the feature interpretability analysis. The "Discussion" section examines the implications of findings, analyzes the model's interpretability, addresses the study's limitations and threats to validity, and discusses directions for future research. Finally, the "Conclusion" summarizes the key findings and contributions of this work.

Related works

ML and Explainable Artificial Intelligence (XAI) are being applied extensively in a range of medical disciplines[34–36]. Objective classification studies employing ML based on the acoustic characteristics of different nonverbal vocalizations (e.g., breathing sounds, coughs) currently play a significant role in diagnosing lung diseases. In 2015, Ulukaya et al.[37] used frequency analysis and time analysis for feature extraction to distinguish monophonic and polyphonic wheezes, and classified features by SVM, KNN, and NB. The findings indicated that the optimal outcomes were attained when the features calculated by the two analysis methods were integrated using KNN, which achieved the highest accuracy of 0.7578. In 2021, Kim et al.[38] recorded 1,918 lung sounds from 871 patients suffering from various respiratory diseases. In this study, the Mel-spectrograms were utilized for the processing of lung sounds, with the convolutional neural network (CNN) being employed for the purpose of classification. This study achieved an accuracy of 0.865 and an AUC of 0.93 in detecting abnormal sounds, with an overall accuracy of 0.857 and a mean AUC of 0.92 in classifying abnormal lung sounds. In 2023, Emeryk et al.[29] trained the random forest regressor (RFR) model to compare the parameter performances of various lung sound data (wheezes, rhonchi, and coarse and fine crackles intensity; respiratory and heart rate; and the parameter performances of inspiration-expiration ratio) obtained with artificial intelligence-assisted stethoscope and peripheral capillary oxygen saturation (SpO_2) and peak expiratory flow (PEF). The results showed that wheezes achieved an AUC of 0.838 in children and rhonchi achieved 0.813 in older children respectively, performing better than other parameter indicators. In 2024, AhmadianKalati et al.[39] developed a predictive model using Bayesian additive regression trees (BART). The performance of this model was then compared with that of adaptive boosting, logistic regression, DT, neural network, RF, and SVM in predicting the diagnosis of childhood asthma. BART demonstrated the optimal predictive performance. The accuracy rate was 0.88, the F1-score was 0.36, and the AUC was 0.75.

Beyond these common machine learning models, XGBoost, a popular tree-structured classifier, has been employed in various medical and other contexts[40–44] particularly in diagnosing respiratory disorders in conjunction with acoustic features. Purkovic et al.[45] enhanced a two-layer framework combining CNN with XGBoost and adaptive boosting (AdaBoost) models using particle swarm optimization (PSO) metaheuristics, achieving a classification accuracy of 0.9814 for determining whether a patient has a respiratory disease. Mouawad et al.[46] employed cough and sustained vowel 'ah' as research data. The classification performance of COVID-19 was evaluated using five classifiers: DT, SVM, KNN, RF and XGBoost. The results showed that XGBoost had the best performance in classifying cough and sustained vowel 'ah', with an accuracy of 0.97 and 0.99, respectively, and a mean F1-score of 0.91 and 0.89 after optimization. Bijoy et al.[47] employed a range of ML algorithms, including RF, ET, GraidentBoost and XGBoost, to classify the sounds of various respiratory conditions (including COPD, Bronchiectasis, Pneumonia and asthma, etc.), and the results showed that XGBoost obtained 0.9863 accuracy, 0.9863 recall, 0.9867 precision and 0.9862 F1-score (slightly inferior to ET, with 0.9947 accuracy). In a similar vein, Varun et al.[48] conducted a classification model training study on an audio database of various respiratory disorders and demonstrated that SVM accuracy led the way with 0.9275, followed by XGBoost's 0.9103.

In contrast to the aforementioned studies on respiratory sound classification, some researchers have chosen verbal vocalizations as their research material. Balamurali et al.[25] studied cough and vocalised /ɑː/ sounds which had not been covered in previous studies. In this study, MFCC and constant-Q cepstral coefficients (CQCCs) were utilised to extract the features of cough and vocalised /ɑː/ sounds, and GMM-UBM model was employed for classification. The model trained using cough demonstrated higher accuracy, sensitivity, and specificity compared to the model trained using vocalised /ɑː/ sounds and the fused model with both cough and vocalised /ɑː/ sounds, 0.973, 0.956, and 0.95, respectively, and the accuracy of the vocalised /ɑː/ model is 0.722. The author hypothesises that this discrepancy can be attributed to the significantly smaller number of available /ɑː/ sounds compared to cough sounds. As evidenced in[25] nonverbal sounds (cough, wheeze) are more reliably indicative of asthma than verbal sounds when classifying subjects with asthma and healthy subjects[19]. In 2018, Yadav et al.[49] utilised sustained phonetic sounds (/ɑː/, /iː/, /uː/, /eɪ/, /oʊ/) as the research object and an SVM as the classifier. The results revealed that the classification accuracy of the sustained vowel /iː/ attained 0.8079, which was higher than cough and lower than wheeze. In 2020, the same research team[19] augmented the original set of phonetic sounds, namely /s/ and /z/, utilising the MFCC as the baseline feature and considering the INTERSPEECH 2013 Computational Paralinguistics Challenge baseline (ISCB) acoustic features for the classification task. The research results indicate that nonverbal stimulus are superior to all verbal stimulus, and /oʊ/ performed best among all verbal stimulus and second best among all verbal and nonverbal stimulus. In addition, the total

Study	Disease	Classifier (best)	Results	Advantages	Disadvantages
Ulukaya et al.[37]	Asthma	KNN	Acc: 0.7578	The evaluation focuses on the classification performance of both time-domain and frequency-domain features	The impact of ML models on classification performance has been weakened
Kim et al.[38]	Various respiratory diseases	CNN	Acc: 0.865 AUC: 0.93	In the research, the transfer learning method was employed, which is simple, fast and capable of using multiple features	The few studies on the classification of mixed breathing sounds have the challenge of noise filtering
Emeryk et al.[29]	Asthma	RFR	AUC: 0.838	Compared with SpO_2 and PEF, the advantages of non-invasive breathing sound classification are proved	The researchers employed reference standards that were established on the basis of doctors' personal experience and subjective decisions
AhmadianKalati et al.[39]	Asthma	BART	Acc: 0.88 AUC: 0.75 F1-score: 0.36	BART emphasized key predictors, including wheezing, coughing and medication, etc	It is evident from the results that even the most effective BART exhibited suboptimal performance
Mouawad et al.[46]	COVID-19	XGBoost	Acc (coughs): 0.99 F1-score (vowel): 0.91	They presented a novel method of informative undersampling using an information theoretic measure of information rate	This model failed to consider its performance in relation to disease diagnosis
Bijoy et al.[47]	Various respiratory diseases	ET	Acc: 0.9947	To tackle class imbalance, they implemented two methods: data augmentation and a hybrid oversampling-undersampling approach	
Varun et al.[48]	Various respiratory diseases	SVM	Acc: 0.9275	Research has demonstrated the superiority of certain classifiers in classifying certain diseases	While some classifiers have an advantage in classifying certain diseases, they may misclassify others
Balamurali et al.[25]	Asthma	GMM-UBM	Acc (coughs): 0.953		The author believes that the reason for the poor performance of vowel is that the sample size is almost ten times smaller than that of cough
Yadav et al.[49]	Asthma	SVM	Acc (wheeze): 0.9025 Acc (/i:/): 0.8079 Acc (coughs): 0.7872	Research has demonstrated that the classification performance of vowels is less robust than that of wheezing, yet superior to that of coughing	
Yadav et al.[19]	Asthma	SVM	Acc (exhale): 0.778 Acc (/oʊ/): 0.754	While adding voice sounds, the ISCB feature set was selected as the classification, which proved to be superior to MFCC	

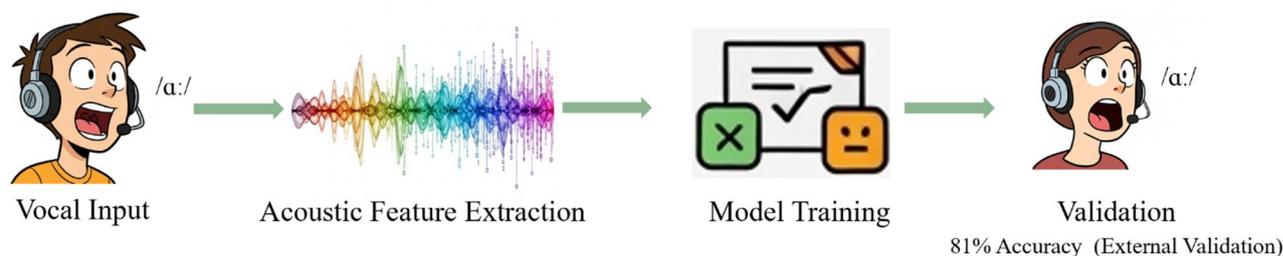**Table 1**. Comparison with papers on related methods that have been published.

Fig. 1. Schematic overview of the research methodology and key result.

classification accuracy (TCA) for vowels with ISCB features showed significant improvement, and /s/ performs best among the fricatives.

As demonstrated in the relevant research (Table 1), ML has been widely applied in the diagnosis of asthma as a technique for classifying respiratory sounds. However, the selected ML methods have not achieved encouraging results. As an emerging ML tool, XGBoost has been applied to the classification and diagnosis of respiratory diseases. However, its performance in identifying mixed diseases is often inferior to that of alternative models, and to the best of the paper's knowledge, no research has specifically employed XGBoost to identify asthma patients. Historically, coughing, wheezing and other such conditions have frequently been employed as sound biomarkers for the classification of disease. However, the research by Yadav et al.[19,49] has demonstrated that the classification performance of vowel sounds has surpassed that of cough and is approaching that of wheezing. It is evident that the ML model employed by the researchers was not fully capable of capitalizing on the advantages of vowel sounds. Therefore, this study is motivated by the clear need to investigate the efficacy of combining a state-of-the-art classifier (XGBoost) with a stable vocal biomarker (/ɑː/) to achieve a higher level of classification performance for adult asthma.

Figure 1 depicts the methodological pipeline of this study. The workflow commences with the acquisition of vocal inputs, followed by the extraction of relevant acoustic features. Subsequently, these features facilitate a model training phase, which involves iterative evaluation and optimization. The final model demonstrated a robust performance, achieving an accuracy of 81% on an external validation set. These findings offer significant

insights for researchers in the fields of machine learning, voice-based diagnostics, and specifically, the diagnosis of adult asthma.

Methods

The study protocol was approved by the Institutional Review Board of Shanghai University of Traditional Chinese Medicine (Approval Number: 2023-3-10-08-08) and the study complied with the Declaration of Helsinki. Clinical trial number: not applicable.

Study participants and datasets

The model development was based on a case-control study involving a primary cohort of 214 participants. This cohort comprised 107 healthy controls and 107 patients diagnosed with asthma. Participants were recruited between September 2019 and December 2021 from two institutions affiliated with the Shanghai University of Traditional Chinese Medicine: the Shanghai Municipal Traditional Chinese Medicine Hospital and Yueyang Integrated Traditional Chinese Medicine and Western Medicine Hospital.

To assess the model's generalizability, external validation using an independent cohort was conducted. This validation dataset was sourced from the Shanghai Qiangsheng Employees' Hospital, with data collected from January 2017 to December 2018. It consisted of 200 participants, evenly split between 100 healthy controls and 100 patients with asthma.

Inclusion criteria

The inclusion criteria were as follows:

① History of typical variable respiratory symptoms[50].
② Confirmed variable expiratory airflow[50].
③ Participants in the healthy group were required to be free from any respiratory disease.
④ All participants were required to complete both the collection and the signing of the informed consent form.
⑤ Participants must be over 18 years old.

Exclusion criteria

The exclusion criteria were as follows:

① Participants with allergic rhinitis, bronchiectasis, and chronic obstructive pulmonary disease.
② Participants with severe endocrine, blood, metabolic system diseases, severe gastrointestinal disease or kidney diseases.
③ Participants with malignant tumors.
④ Participants who refused to participate.
⑤ Severe incomplete clinical data.

Voice data collection

This study utilized a Logitech H340 headset microphone. Voice information required for this research was collected using the voice diagnosis module of the SmartTCM-1 Traditional Chinese Medicine Life Information Analysis System, which was jointly developed by Shanghai University of Traditional Chinese Medicine and Shanghai Asia-Pacific Computer Information System Co., Ltd. In an environment where the ambient noise was ensured to be less than 45 dB, subjects were equipped with the headset microphone. The microphone boom was adjusted so that the microphone was positioned approximately 2 cm directly in front of the subject's lips. The earphone volume was calibrated to ensure that subjects could clearly receive auditory prompt commands. Subjects were instructed to naturally vocalize the vowel of /ɑː/ according to the prompt commands heard through the earphones. Recordings were made at a sampling frequency of 16 kHz and saved as audio files in WAV format with a 16-bit sampling depth.

Feature extraction and feature selection

This study employed the eGeMAPS for feature extraction. eGeMAPS is recognized as a standard baseline feature set for a variety of acoustic analysis tasks. In contrast to larger, more comprehensive parameter sets, eGeMAPS offers a compact set of voice parameters, representing an expansion of the foundational Geneva Minimalistic Acoustic Parameter Set (GeMAPS).

The GeMAPS set comprises 62 acoustic features. These are generated by applying statistical functionals to 18 Low-Level Descriptors (LLDs)—encompassing 6 frequency-related, 3 energy/amplitude-related, and 9 spectral (balance) parameters—and subsequently incorporating 6 temporal features. The eGeMAPS builds upon GeMAPS by introducing 7 additional LLDs: 5 spectral (balance/shape/dynamics) parameters and 2 frequency-related parameters, as detailed in Table 2. The application of statistical functionals to these 7 supplementary LLDs yields 25 further acoustic parameters. These, in conjunction with the 62 features from GeMAPS and the equivalent sound level, constitute the standard 88-feature eGeMAPS[51].

A key modification in the present study was the exclusion of the standard deviation functional during the statistical calculations performed on all 25 LLDs (the 18 original to GeMAPS and the 7 additional ones for eGeMAPS), to reduce dimensionality. Consequently, the feature set utilized in this research contains a reduced number of features compared to the standard eGeMAPS configuration.

This research employed the openSMILE open-source toolkit[52] for the analysis and processing of the acquired audio, from which eGeMAPS parameters were extracted.

Parameter	LLD
Frequency related parameters	Pitch, logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)
	Jitter, deviations in individual consecutive F0 period lengths
	Formant 1, 2, and 3 frequency, centre frequency of first, second, and third formant
	Formant 1, bandwidth of first formant
Energy/amplitude related parameters	Shimmer, difference of the peak amplitudes of consecutive F0 periods Loudness, estimate of perceived signal intensity from an auditory spectrum Harmonics-to-noise ratio (HNR), relation of energy in harmonic components to energy in noise-like components
Spectral (balance) parameters	Alpha Ratio, ratio of the summed energy from 50–1000 Hz and 1–5 kHz
	Hammarberg Index, ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
	Spectral Slope 0–500 Hz and 500–1500 Hz, linear regression slope of the logarithmic power spectrum within the two given bands
	Formant 1, 2, and 3 relative energy, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F_0
	Harmonic difference H1–H2, ratio of energy of the first F_0 harmonic (H1) to the energy of the second F_0 harmonic (H2)
	Harmonic difference H1–A3, ratio of energy of the first F_0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)
Temporal features	The rate of loudness peaks, i.e., the number of loudness peaks per second
	The mean length and the standard deviation of continuously voiced regions ($F_0 > 0$)
	The mean length and the standard deviation of unvoiced regions ($F_0 = 0$; approximating pauses)
	The number of continuous voiced regions per second (pseudo syllable rate)
Spectral (balance/shape/dynamics) parameters	MFCC 1–4 Mel-Frequency Cepstral Coefficients 1–4
	Spectral flux difference of the spectra of two consecutive frames
Frequency related parameters	Formant 2–3 bandwidth added for completeness of Formant 1–3 parameters

Table 2. 25 LLD and 6 Temporal features of eGeMAPS[51].

Machine learning classification

Traditional classification algorithms often suffer from limited generalization ability and are not well-suited for handling large-scale, high-dimensional datasets. The advent of big data has necessitated the development of more sophisticated classification techniques. ML, which leverages statistical methods, database systems, and computer algorithms, has emerged as a promising paradigm for addressing these challenges. Classification problems, as a core component of ML, have been extensively studied[53–55]. ML classification applications span a wide range of fields, from healthcare and finance to retail, e-commerce, image recognition, and natural language processing, playing a significant role in each.

In this study, the dataset of 214 participants was randomly divided into train and test sets using a 70/30 split. The best-performing model underwent internal validation using ten-fold cross-validation and external validation using an independent external dataset.

The versions used in this study: Python (3.8.19), scikit-learn (1.2.2), PyCaret (3.2.0), openSMILE (3.0.2), SHAP (0.44.1), and LIME (0.2.0.1). The random_state = 123.

Performance evaluation metrics

The ML classification evaluation metrics used in this paper mainly include Accuracy, Area under curve (AUC), Recall, Precision, F1-score, Kappa Coefficient, and Matthews Correlation Coefficient (MCC). Superior model performance is reflected by higher numerical values in all evaluation metrics.

Explainability method

SHAP is a game-theoretic approach to explain the output of any ML model. SHAP values measure how much each feature contributes, both positively and negatively, to the model's prediction for a given instance. In simpler terms, SHAP helps us understand why a ML model makes a particular prediction.

LIME helps understand why a complex "black-box" ML model made a specific prediction for a single instance. It does this by creating a simple, local, and interpretable model (like a linear model) that approximates the black-box model's behavior around that particular prediction.

As illustrated in Fig. 2, methodological pipeline commences with the acquisition of /ɑː/ vowel sounds to establish two distinct cohorts: a primary cohort ($n = 214$) designated for model development and an independent cohort ($n = 200$) reserved for external validation. Following feature extraction from both cohorts, the model development phase was initiated. This phase involved training a suite of seven distinct ML models on the primary cohort's data, which was partitioned into a 70/30 train/test split, utilizing a rigorous 10-fold cross-validation protocol. Subsequently, the developed models were subjected to a final evaluation against the independent validation cohort for the selection of the optimal model. In the final step, this optimal model underwent a post-hoc interpretability analysis using SHAP and LIME to elucidate its predictive mechanisms and ensure transparency.
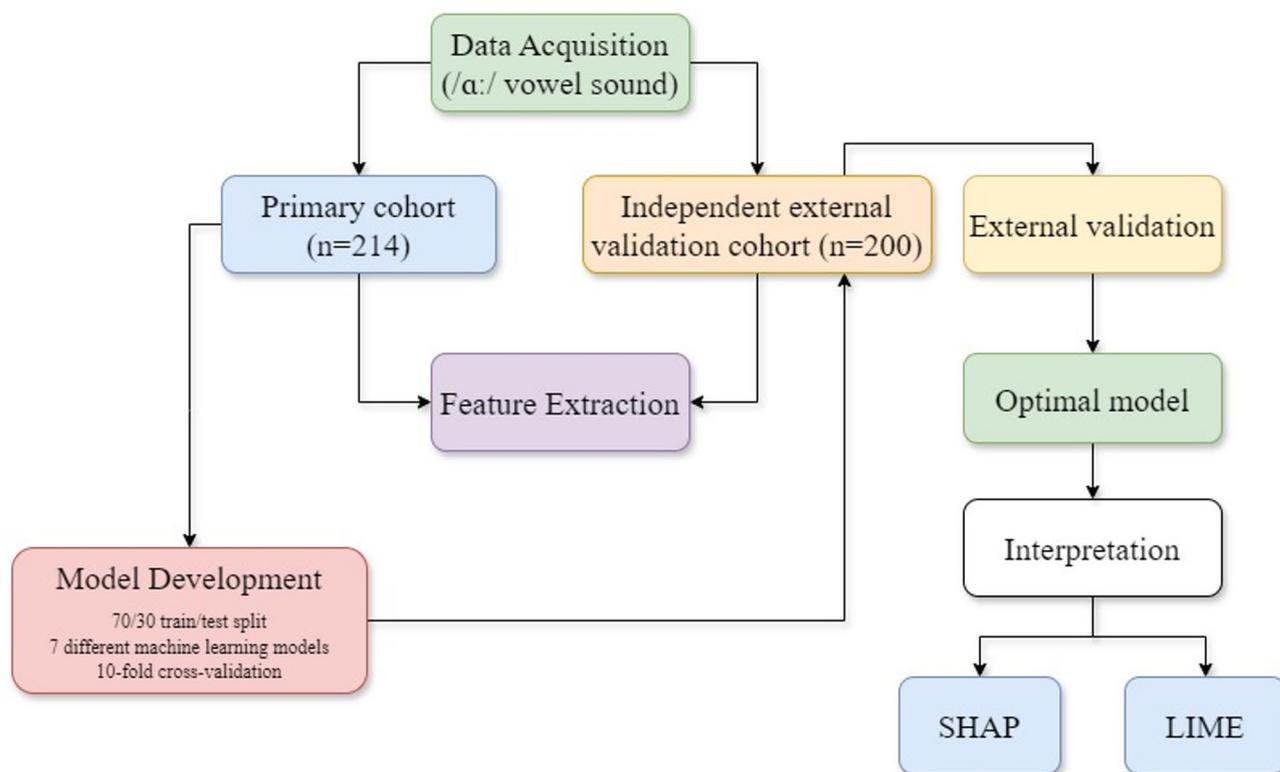
Fig. 2. Methodological workflow of the study.

Characteristic	Total ($N = 214$)	Healthy control group ($n = 107$)	Asthma group ($n = 107$)	P-value
Age, years, mean ± standard deviation	28.0 ± 3.4	28.1 ± 3.2	27.8 ± 3.6	0.545[a]
Sex, n (%)				0.100[b]
Male	98 (45.8)	43 (40.2)	55 (51.4)	
Female	116 (54.2)	64 (59.8)	52 (48.6)	
Education, n				0.131[c]
High school or less	71	33	38	
College	95	44	51	
Graduate degree or higher	48	30	18	

Table 3. Baseline characteristics of the study participants. Data are presented as mean ± SD or n (%). [a]P-value was calculated using the independent samples t-test. [b]P-value was calculated using the chi-square test. [c]P-value was calculated using the Mann–Whitney U test.

Results

The study enrolled 214 participants, who were divided equally into the asthma group ($n = 107$) and the healthy control group ($n = 107$). A detailed summary of the baseline characteristics is provided in Table 3.

Overall, the mean age of the cohort was 28.0 ± 3.4 years. The comparison between groups showed no significant difference in mean age (28.1 ± 3.2 vs. 27.8 ± 3.6, $P = 0.545$). The study included 98 (45.8%) males and 116 (54.2%) females, and the sex distribution was not significantly different between the healthy control and asthma groups ($P = 0.100$). Similarly, there was no significant difference in the distribution of education levels between the two groups ($P = 0.131$).

Figure 3 illustrates the baseline characteristics of the healthy control ($n = 107$) and asthma ($n = 107$) groups. (A) Bar chart of the mean age with standard deviation error bars. (B) Bar chart showing the gender distribution by count. (C) Bar chart showing the distribution of education levels by count. There were no statistically significant differences between the two groups in age ($P = 0.545$), gender ($P = 0.100$), or education level ($P = 0.131$).

The seven ML algorithms previously mentioned were employed for the construction of diagnostic models. Table 4 presents the performance metrics of these seven ML models using their default hyperparameters. Ten-fold cross-validation, stratified to maintain class proportions across folds, was conducted on the training data. The final evaluation focused on assessing the generalization capability of the models on test set. The XGBoost classifier demonstrated superior performance, achieving the highest values for accuracy, recall, F1-score, and
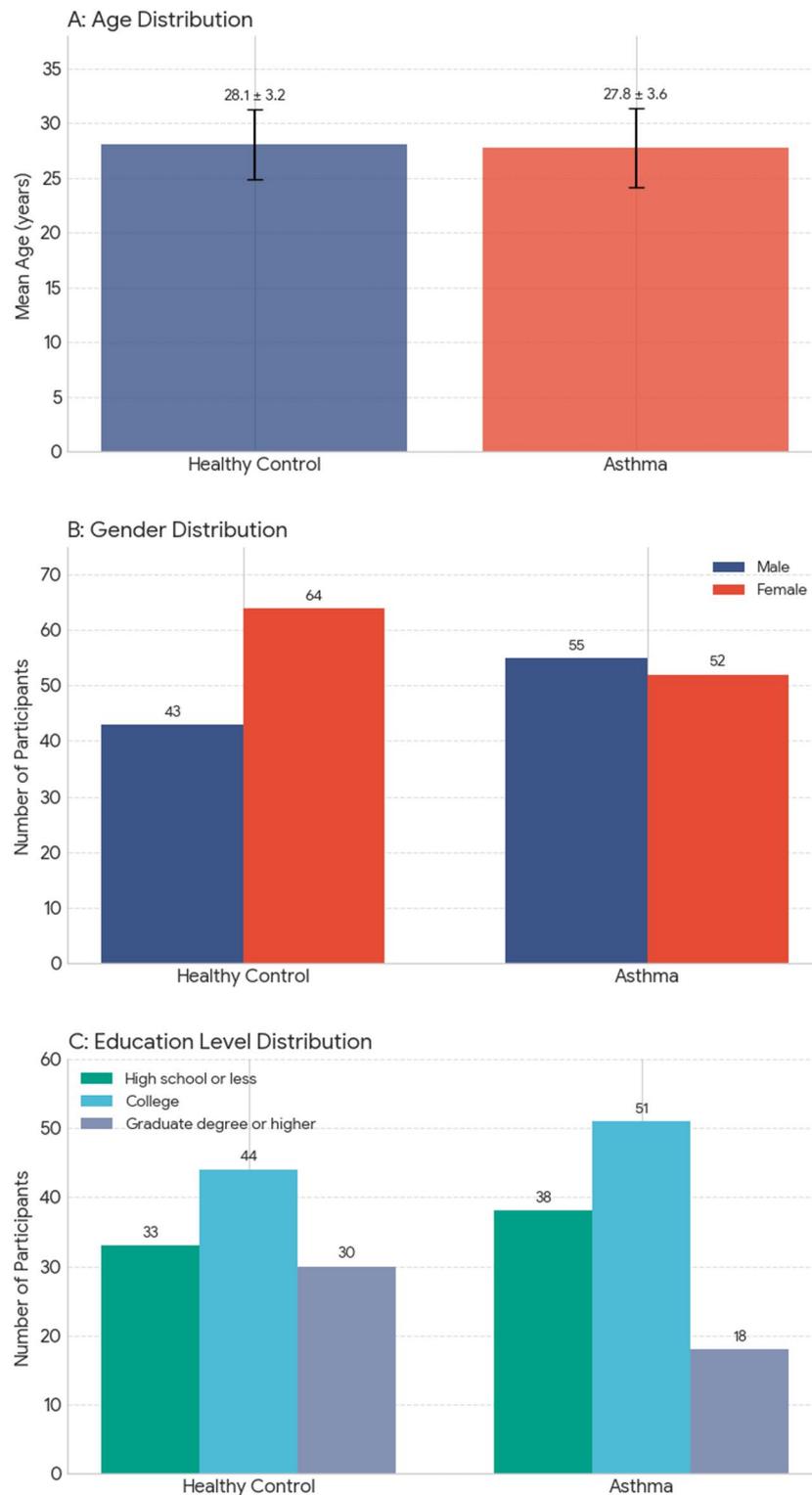
Fig. 3. Demographic distribution of study participants.

Kappa Coefficient. Considering a comprehensive assessment that included total computation time (TT), the XGBoost classifier was selected as the optimal ML model for subsequent investigation.

From Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 and 23, the ROC curves, confusion matrix, threshold plot, and feature importance for all 7 ML models on test set are shown.

Table 5 details the results of a ten-fold cross-validation for an XGBoostClassifier, evaluating its performance across various metrics. The XGBoostClassifier generally demonstrates strong and relatively consistent performance across the ten folds, as indicated by the mean values of the evaluation metrics. The XGBoostClassifier

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (s)
Extreme gradient boosting	0.8514±0.0622	0.9130±0.0774	0.8804±0.0758	0.8387±0.0757	0.8567±0.0594	0.7018±0.1254	0.7071±0.1275	0.015
Extra trees classifier	0.8510±0.0936	0.9125±0.0604	0.8536±0.1605	0.8556±0.0881	0.8467±0.1146	0.7012±0.1878	0.7107±0.1843	0.028
CatBoost classifier	0.8505±0.1055	0.9286±0.0565	0.8500±0.1971	0.8498±0.1045	0.8378±0.1524	0.7009±0.2107	0.7130±0.2028	1.179
Random forest classifier	0.8443±0.0946	0.9157±0.0655	0.8268±0.1547	0.8645±0.0983	0.8374±0.1145	0.6882±0.1898	0.6982±0.1866	0.023
Light gradient boosting machine	0.8371±0.1294	0.9010±0.0821	0.8268±0.1547	0.8481±0.1323	0.8345±0.1396	0.6730±0.2602	0.6762±0.2602	0.016
Decision tree classifier	0.8257±0.0675	0.8250±0.0662	0.8125±0.1036	0.8506±0.1066	0.8236±0.0663	0.6504±0.1336	0.6625±0.1385	0.004
Naive Bayes	0.7776±0.0773	0.8311±0.1033	0.7821±0.1143	0.7794±0.0816	0.7761±0.0817	0.5519±0.1557	0.5582±0.1552	0.004

Table 4. The performance comparison of different ML models.
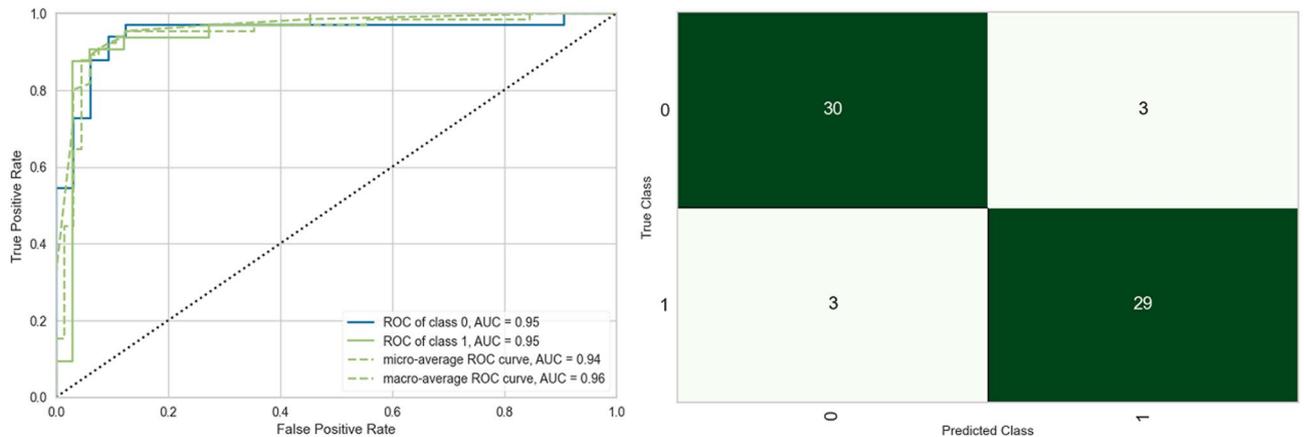
Fig. 4. ROC curves and confusion matrix for XGBoostClassifier on test set.

appears to be a robust model for this diagnostic task, with high average performance in distinguishing classes and overall accuracy.

The performance of the XGBoostClassifier was further evaluated on an independent external validation set, as presented in Table 6. On this external dataset, the model achieved an accuracy of 0.81 and an AUC of 0.8757, indicating strong discriminative capability on previously unseen data. The model demonstrated a recall of 0.83 and a precision of 0.7981, leading to a balanced F1-score of 0.8137. Furthermore, the Kappa coefficient was 0.62 and the MCC was 0.6205, suggesting a robust and reliable classification performance beyond chance agreement when applied to this new dataset.

Figure 24 shows that the XGBoost model performs well on external validation task. It demonstrates strong discriminative power (AUC=0.8755) and achieves good accuracy (81%) and recall (83%). Although a certain number of false positives and false negatives exist, these metrics are acceptable in many application scenarios.

Figure 25 shows that the curve maintains a very high precision (close to 1.0) at lower recall rates, and even when the recall rate reaches around 0.8, the precision can still be maintained above 0.75. This shows that the model can identify most of the positive examples while ensuring high accuracy.

Figure 26 shows that the probability distributions of the two classes have very little overlap. This strongly suggests that the model can effectively distinguish between positive and negative classes.

Figure 27 shows that when selecting the top 20% of the population with the highest predicted scores, the lift value is close to 1.9, which means that the model's selection efficiency is almost twice that of random selection. Even when selecting the top 60% of the population, the lift value is still much greater than 1. Selecting the top 40% of the population by score captures approximately 80% of the positive cases, selecting the top 60% captures about 90% of the positive cases. This is much more efficient than random selection (dashed line), indicating that the model is very effective at ranking positive cases at the top of the scores.

Figure 28 shows the degree and direction of each feature's impact on the model output. The features with the most significant impact on the model's predictions are: F2frequency_sma3nz_amean, F1frequency_sma3nz_amean, and F3frequency_sma3nz_amean.

F2frequency_sma3nz_amean: There is a clear inverse relationship. High values (red dots) consistently have negative SHAP values, meaning they decrease the model's output. Low values (blue dots) consistently have positive SHAP values, increasing the model's output.

F1frequency_sma3nz_amean: There is a clear positive relationship. High values (red dots) have positive SHAP values, increasing the model's output. Low values (blue dots) have negative SHAP values, decreasing the model's output.

F3frequency_sma3nz_amean: Shows a similar, though perhaps slightly less distinct, positive relationship to F1: higher values tend to increase the output.
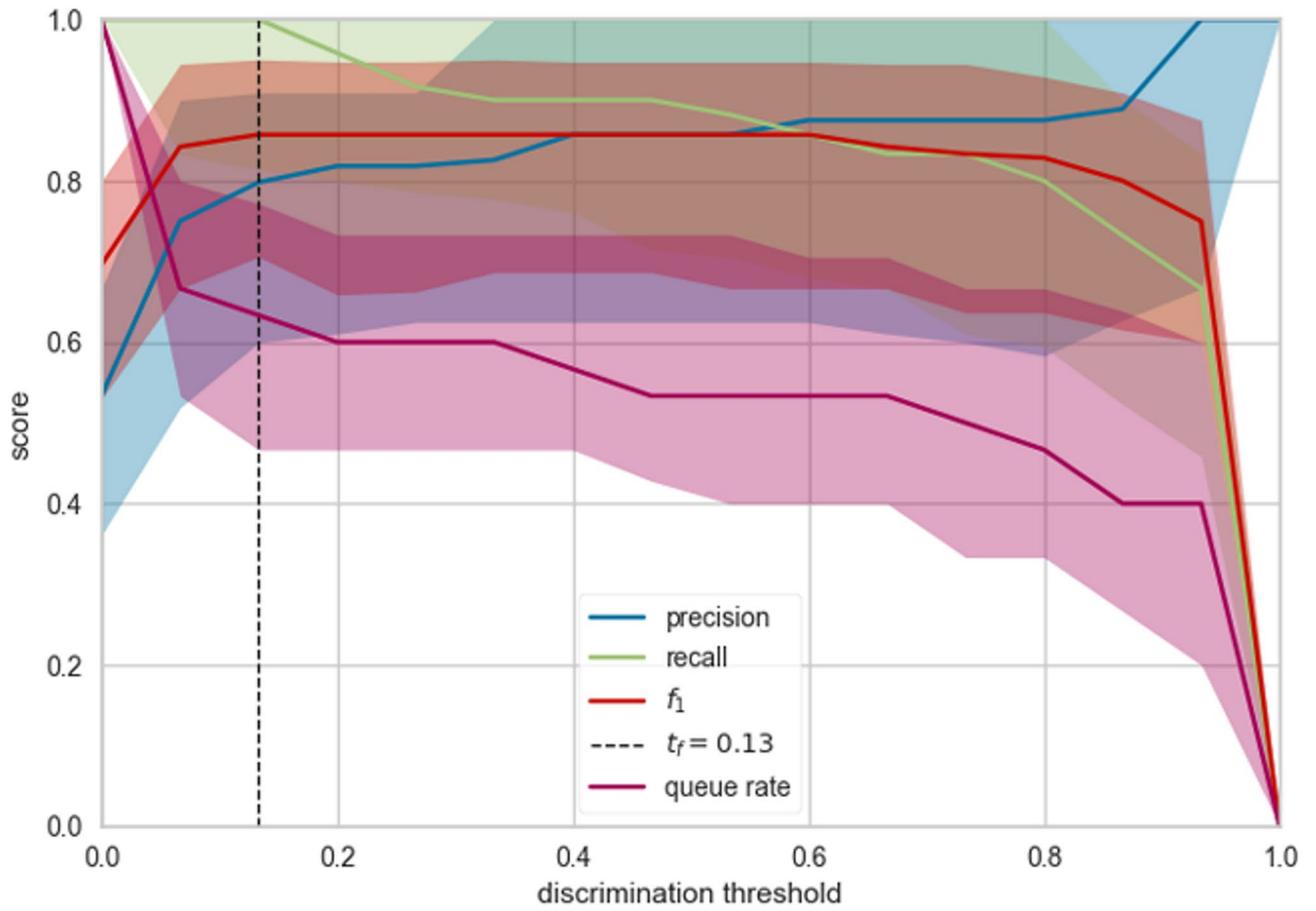
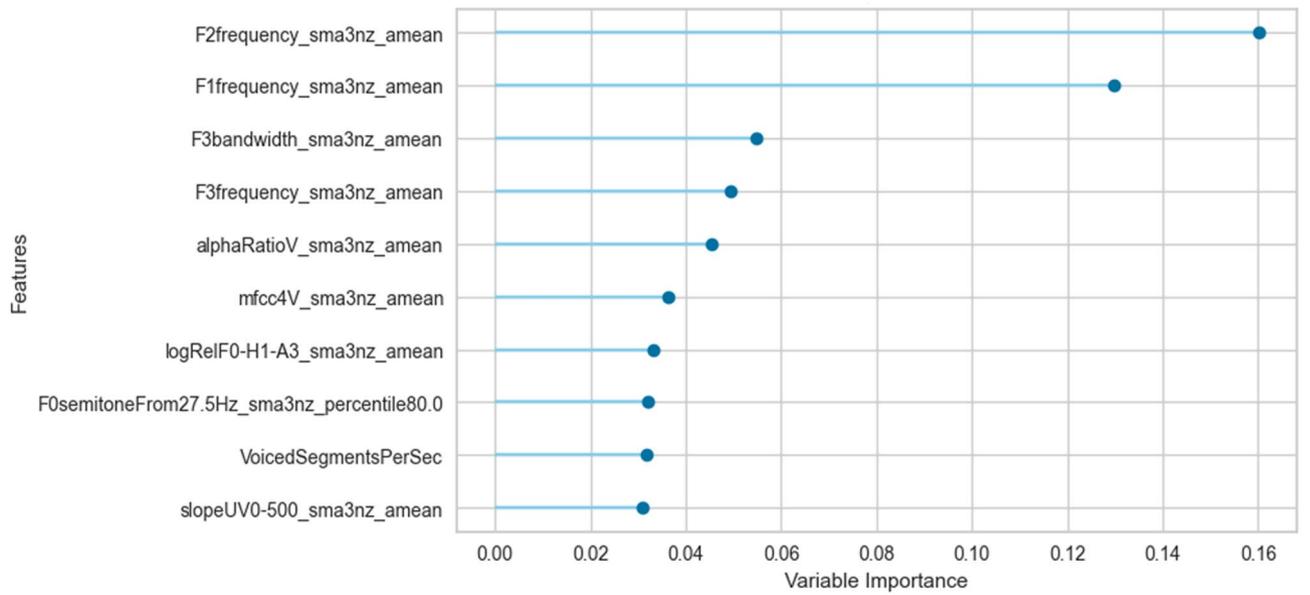**Fig. 5**. Threshold Plot for XGBoostClassifier on test set.

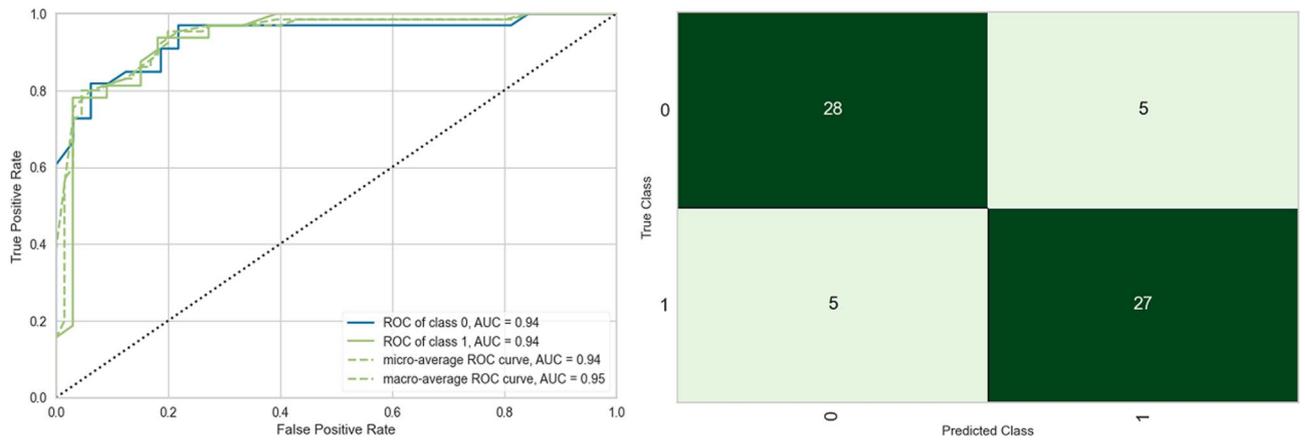**Fig. 6**. Feature importance of XGBoostClassifier on test set.

Fig. 7. ROC curves and confusion matrix for ExtraTreesClissifier on test set.
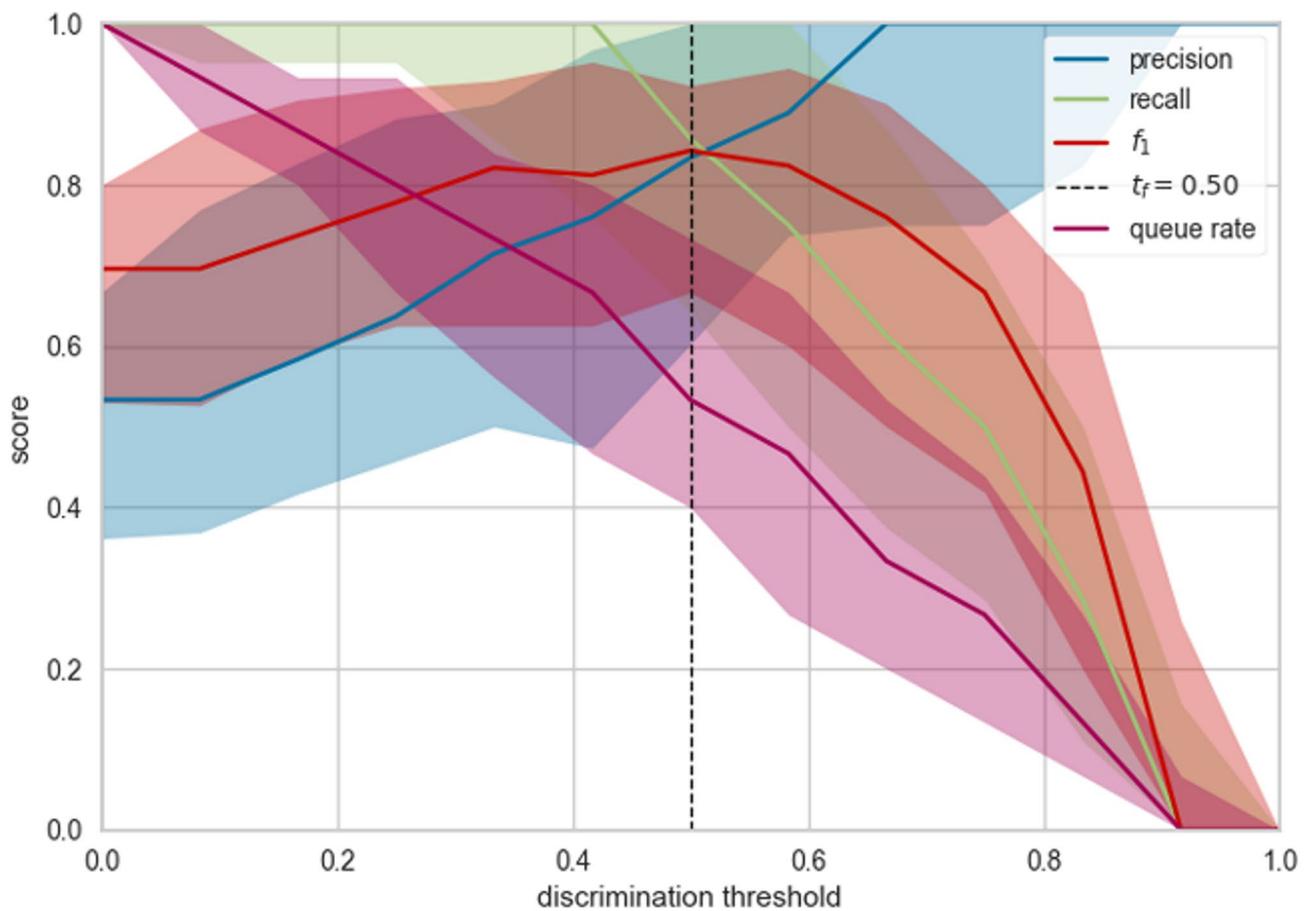

Fig. 8. Threshold Plot for ExtraTreesClassifier on test set.

Figure 29 displays a LIME explanation for why a specific instance (Instance 0) was predicted as class 0 by the model.

- X-axis: Represents the magnitude and direction of each feature's contribution to the prediction. Positive values (green bars) indicate that the feature's presence or condition meeting the criteria increases the likelihood of predicting class 0. Negative values (red bars) indicate that the feature's presence or condition meeting the criteria decreases the likelihood of predicting class 0. The larger the absolute value, the stronger the influence.
- Y-axis: Lists the features and their conditions that contributed most significantly to the current prediction.

12
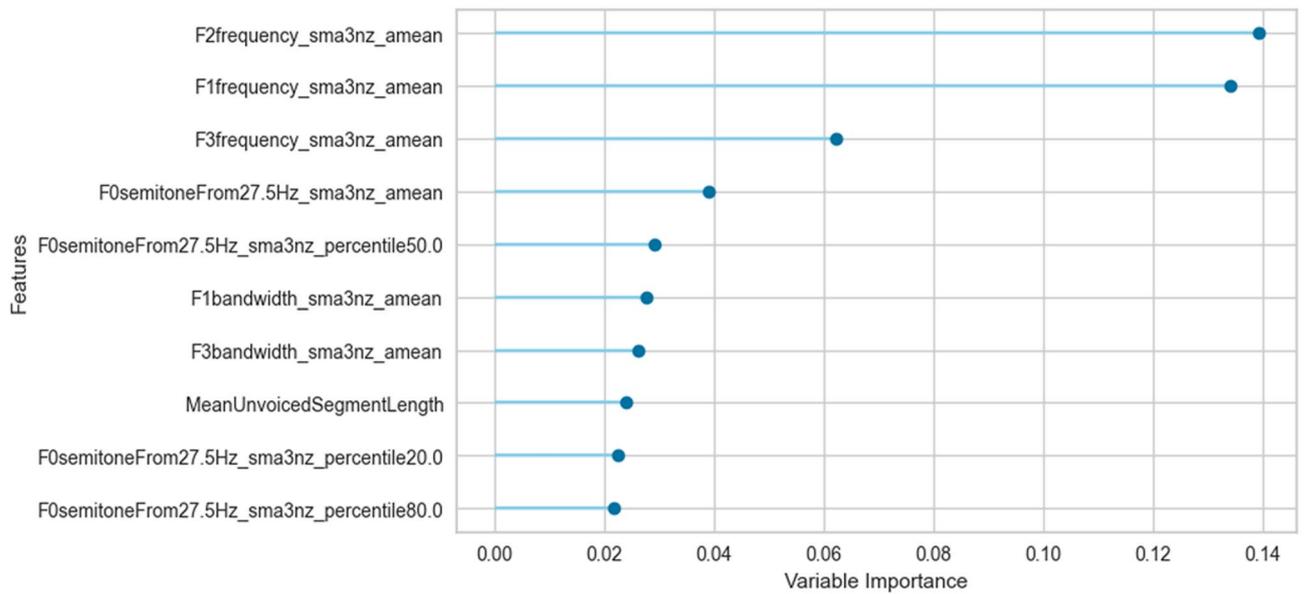
Fig. 9. Feature importance of ExtraTreesClassifier on test set.
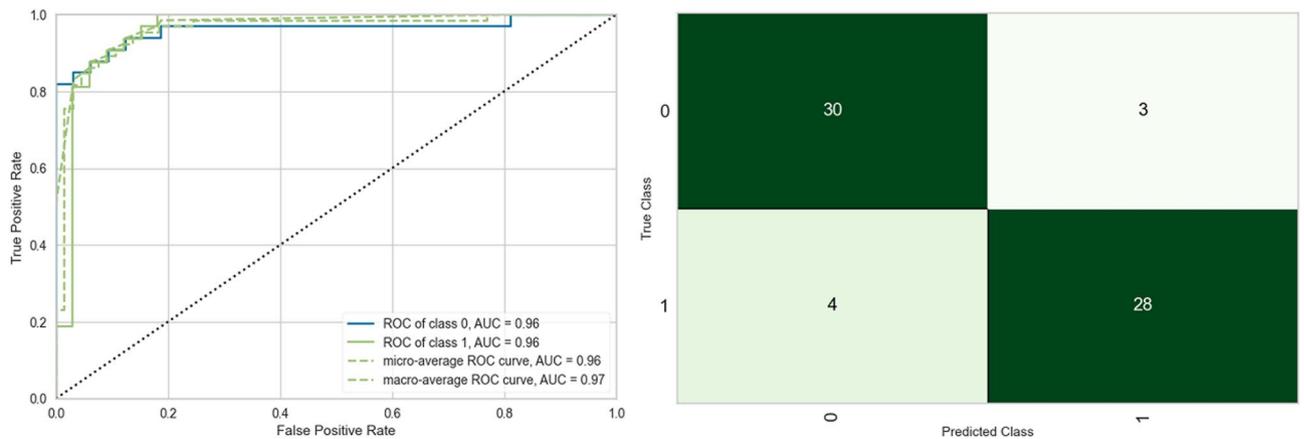
Fig. 10. ROC curves and confusion matrix for CatBoostClassifier on test set.

- Green Bars: These feature conditions (e.g., "F2frequency_sma3nz_amean < = 1420.28") are the main support-ing evidence for the model's prediction of class 0 for this instance. "F2frequency_sma3nz_amean < = 1420.28" has the largest positive contribution, followed by "F3frequency_sma3nz_amean < = 2510.42" and "F1frequen-cy_sma3nz_amean < = 501.55".
- Red Bars: These feature conditions (e.g., "F2bandwidth_sma3nz_amean > 1203.65") are the main counter-ev-idence or hindering factors for the model's prediction of class 0. "F2bandwidth_sma3nz_amean > 1203.65" is the largest negative contributor, meaning its presence makes the model less likely to predict class 0.

For this particular instance, the model predicted it as class 0 primarily because its F2, F3, and F1 frequency means were below certain thresholds. Although some factors (like F2 bandwidth mean being above a threshold) somewhat opposed this prediction, the positive influence of the frequency features (shown by green bars) significantly outweighed the negative impacts, leading to the final class 0 prediction.

Discussion

This study investigated the viability of ML techniques as an auxiliary tool for diagnosing asthma in adults. These findings demonstrate that ML can provide accurate asthma diagnoses. Seven computational models were developed, and as shown in Table 4, XGBoost exhibited the best classification performance, achieving an accuracy of 0.8514, an AUC of 0.913, a recall of 0.8804, a precision of 0.8387, an F1-score of 0.8567, a Kappa of 0.7018, and an MCC of 0.7071. The top 10 most important features are: F2frequency_sma3nz_amean, F1frequency_ sma3nz_amean, F3bandwidth_sma3nz_amean, F3frequency_sma3nz_amean, alphaRatioV_sma3nz_amean, mfcc4V_sma3nz_amean, logRelF0-H1-A3_sma3nz_amean, F0semitoneFrom27.5Hz_sma3nz_percentile80.0,
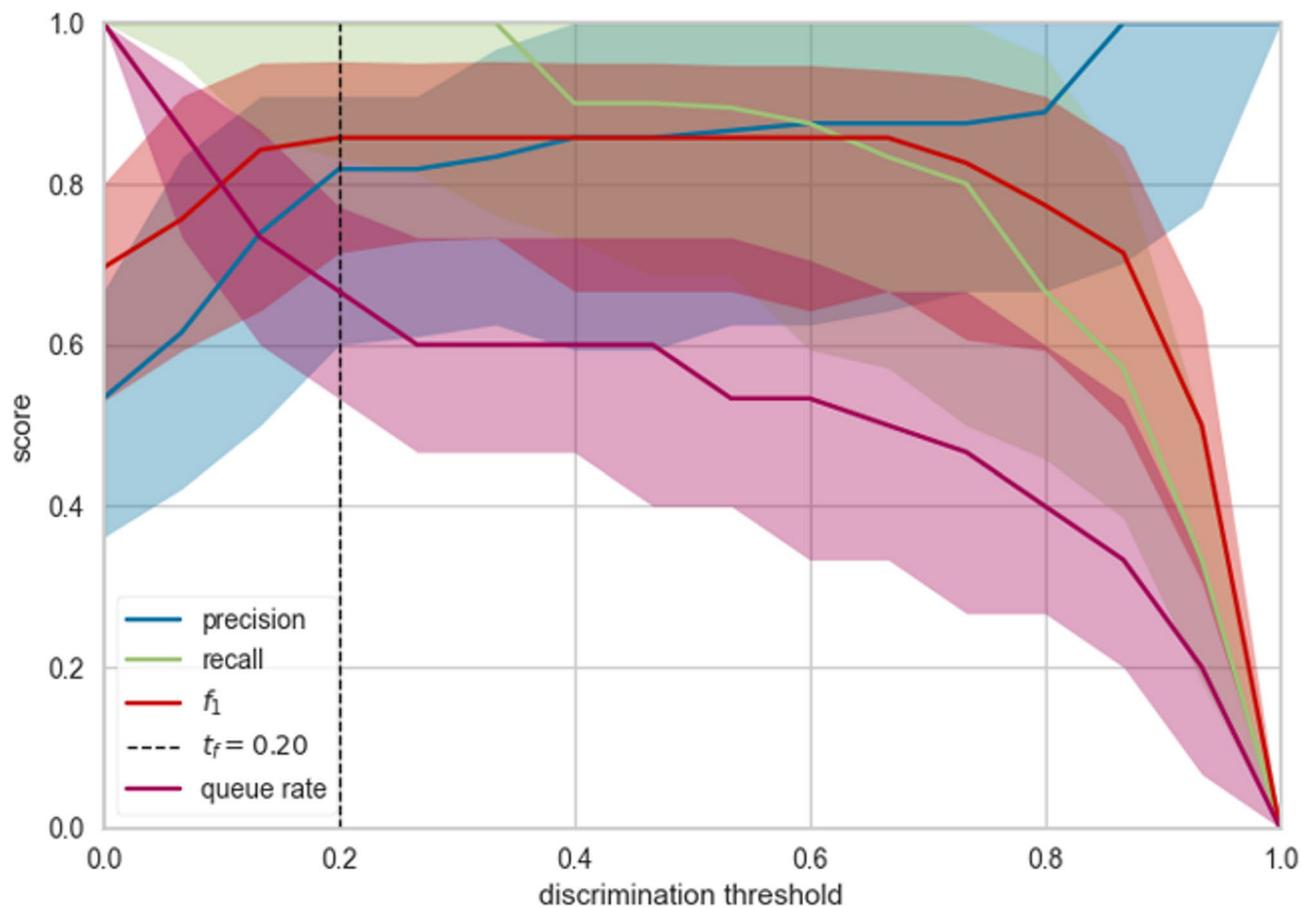
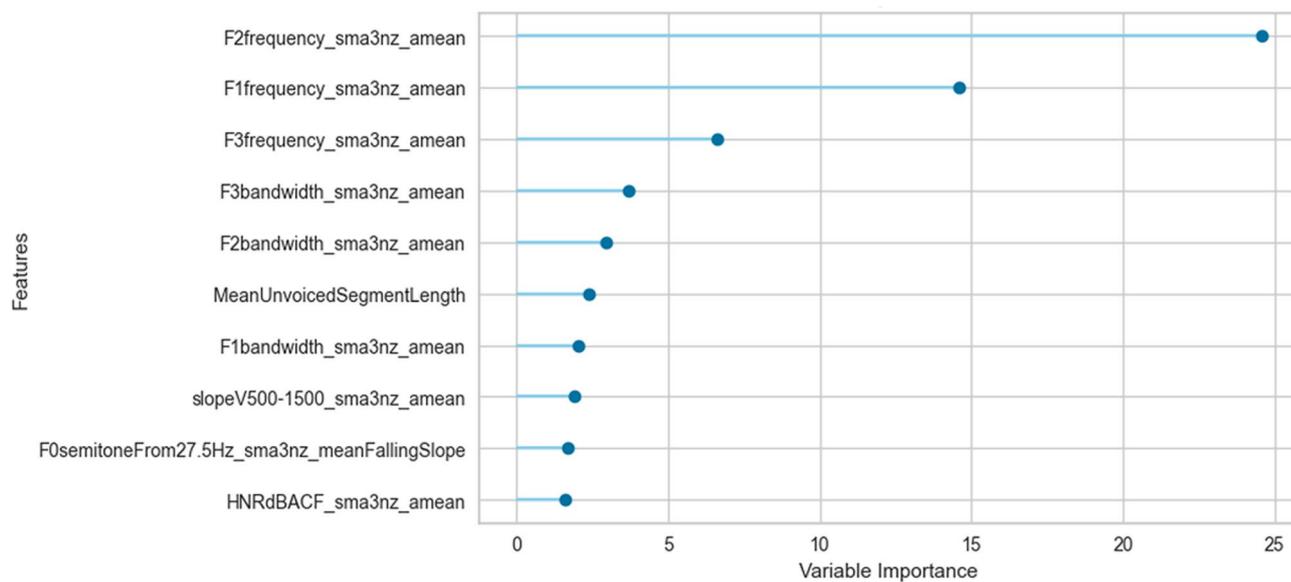**Fig. 11**. Threshold Plot for CatBoostClassifier on test set.

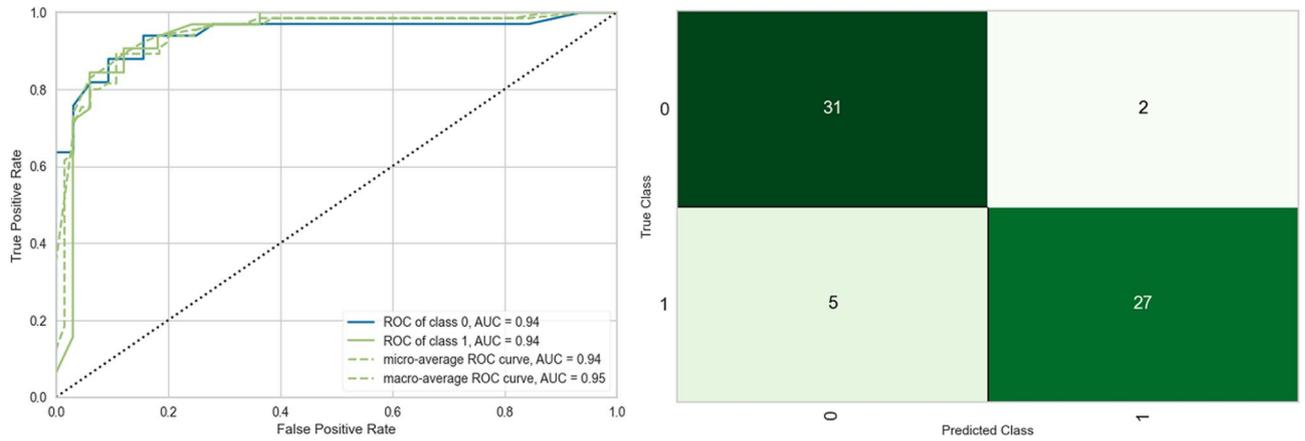**Fig. 12**. Feature importance of CatBoostClassifier on test set.

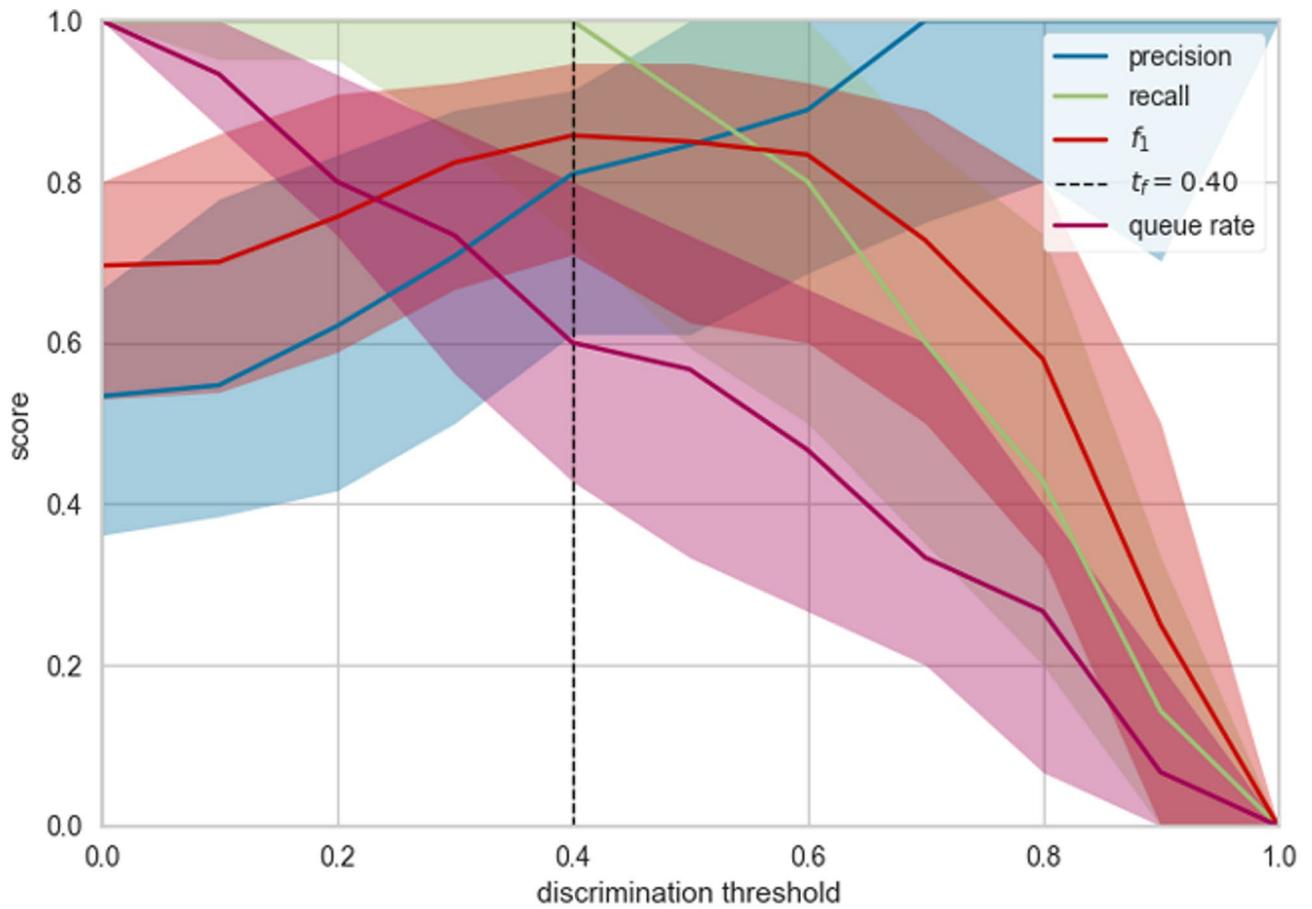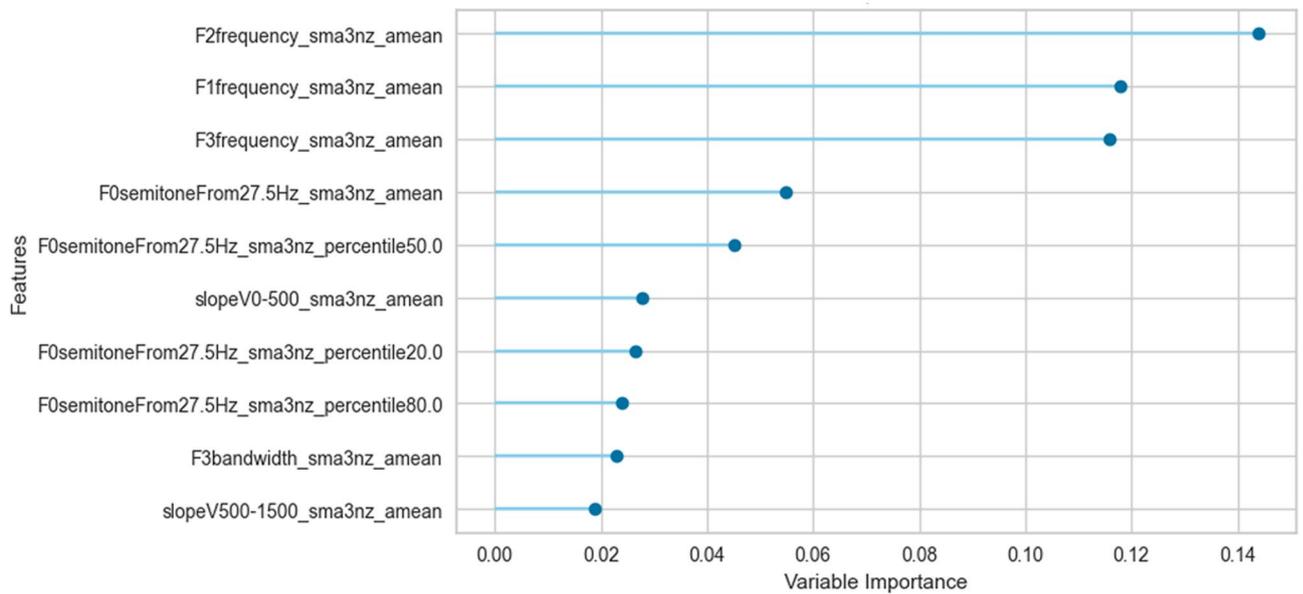**Fig. 13**. ROC curves and confusion matrix for RandomFoerstClassifier on test set.

Fig. 14. Threshold plot for CatBoostClassifier on test set.

VoicedSegmentsPerSec, and slopeUV0-500_sma3nz_amean. On the external validation set (Table 6), XGBoost maintained strong performance with an accuracy of 0.81 and an AUC of 0.8755.

Analysis of model interpretability results

To enhance the interpretability of XGBoost model, the global feature analysis from SHAP with a local explanation for an individual prediction using LIME was supplemented. This dual approach allows us to both understand general trends across the dataset and dissect the reasoning behind a specific outcome. For this analysis, "Instance 0," which the model correctly predicted as "Class 0" (Healthy).

Fig. 15. Feature importance of RandomFoerstClassifier on test set.
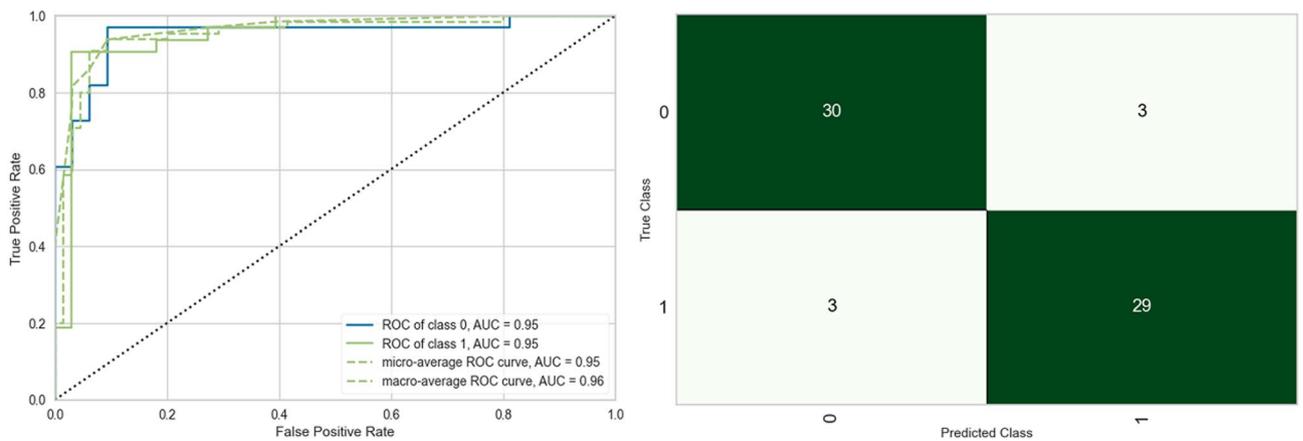
Fig. 16. ROC curves and confusion matrix for LightGBMClassifier on test set.

At a high level, both methods agree on the central importance of formant frequencies in distinguishing between healthy and asthmatic subjects.

- The SHAP summary plot clearly ranks F2frequency_sma3nz_amean, F1frequency_sma3nz_amean, and F3frequency_sma3nz_amean as the top three most impactful features globally.
- Similarly, the LIME explanation for Instance 0 shows these same three features providing the largest contributions (both positive and negative) to this specific prediction.

This consistency confirms that the model's predictions are stably driven by acoustically significant features.
For several features, the LIME explanation for Instance 0 serves as a perfect local example identified by SHAP.

- F1 and F3 Frequencies: LIME shows that a low F1 frequency ($<= 501.55$) and a low F3 frequency ($<= 2510.42$) are strong contributors to the "Healthy" prediction for this instance. This aligns perfectly with the SHAP plot, where low values for F1 and F3 (blue dots) consistently have negative SHAP values, pushing the model's output towards the "Healthy" class.

The most insightful finding comes from an apparent contradiction in the analysis of the top feature, F2frequency_sma3nz_amean.

- SHAP: The SHAP plot shows a clear and strong trend: low values of F2 frequency (blue dots) have positive SHAP values, meaning they are powerful predictors of the "Asthma" class.
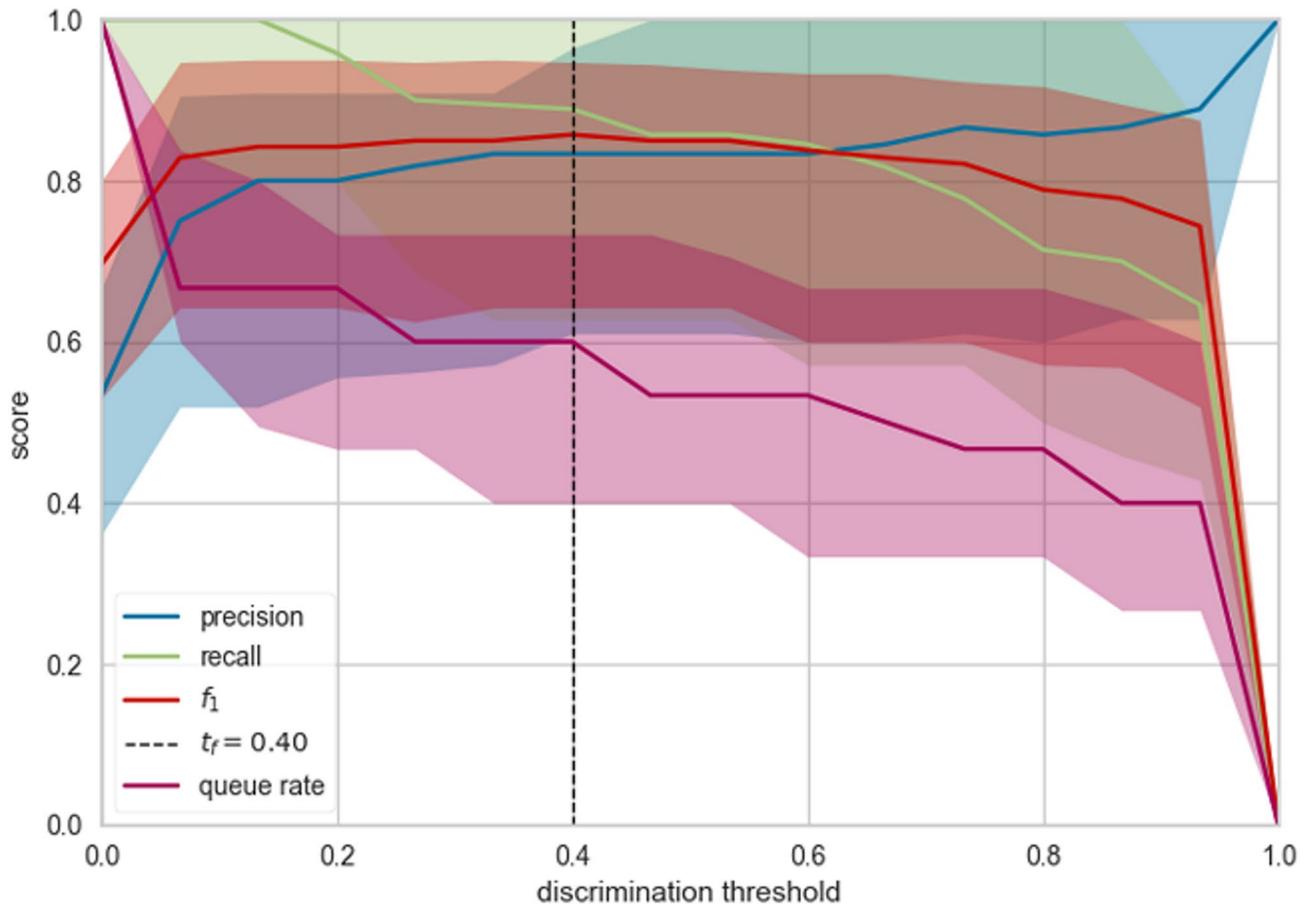
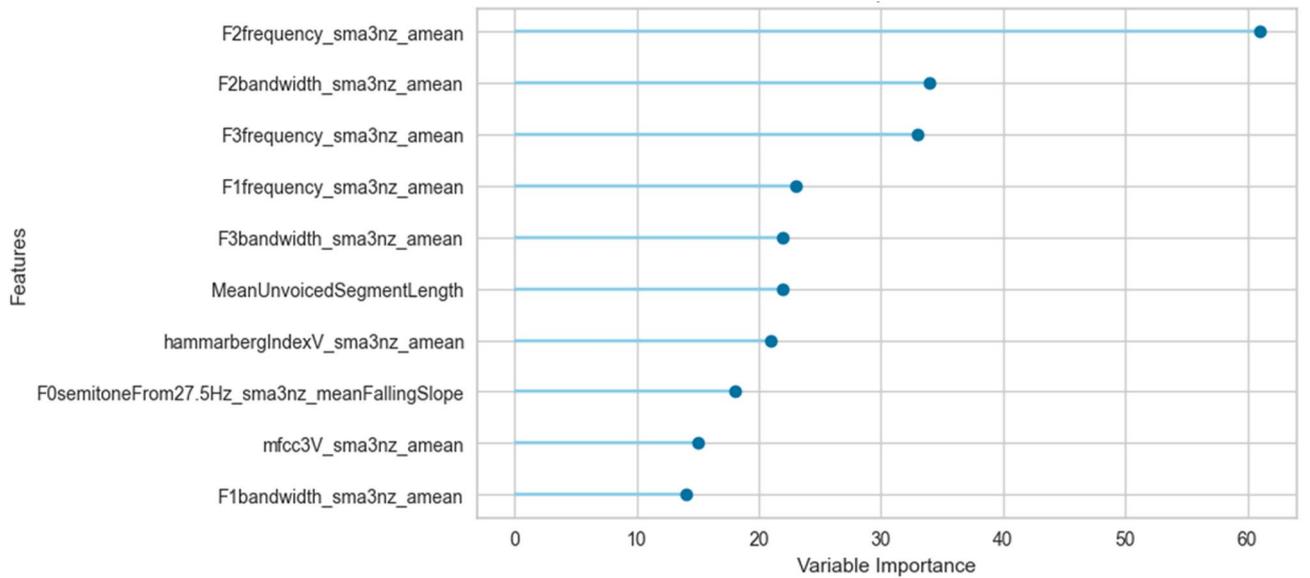**Fig. 17**. Threshold plot for LightGBMClassifier on test set.

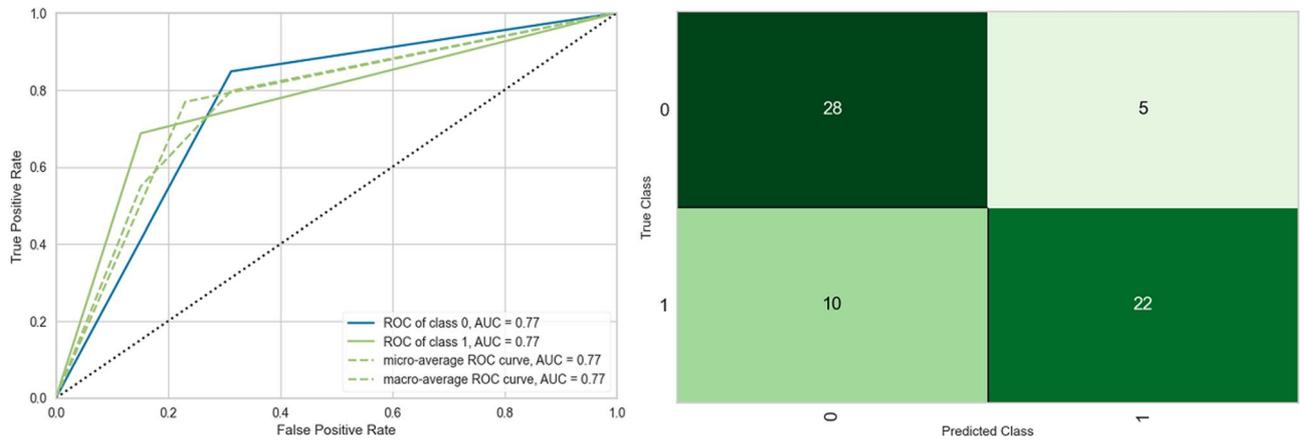**Fig. 18**. Feature importance of LightGBMClassifier on test set.

Fig. 19. ROC curves and confusion matrix for DecisionTreeClassifier on test set.
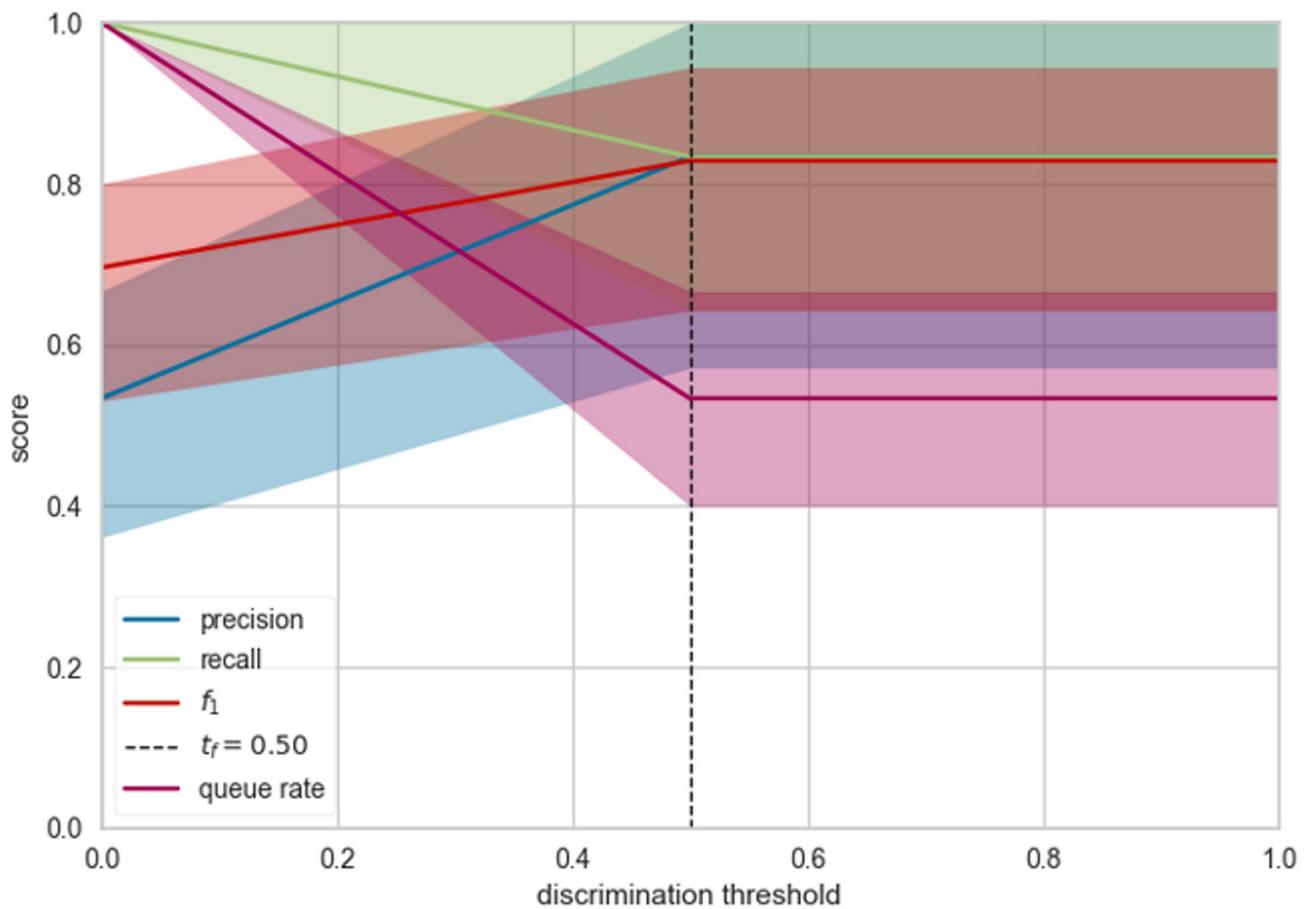

Fig. 20. Threshold plot for DecisionTreeClassifier on test set.

- LIME: However, for this specific "Healthy" instance, LIME identifies a low F2 frequency as the single most powerful feature supporting the "Healthy" prediction ($<= 1420.28$).

This apparent discrepancy is not a flaw in the model or the methods. Instead, it reveals a crucial aspect of the model's sophistication:

- Non-Linearity and Feature Interaction: This divergence strongly suggests that the relationship between F2 frequency and an asthma diagnosis is not simply linear. The effect of F2 is likely dependent on the values of other features. For Instance 0, its specific combination of very low F1 and F3 frequencies may have created a
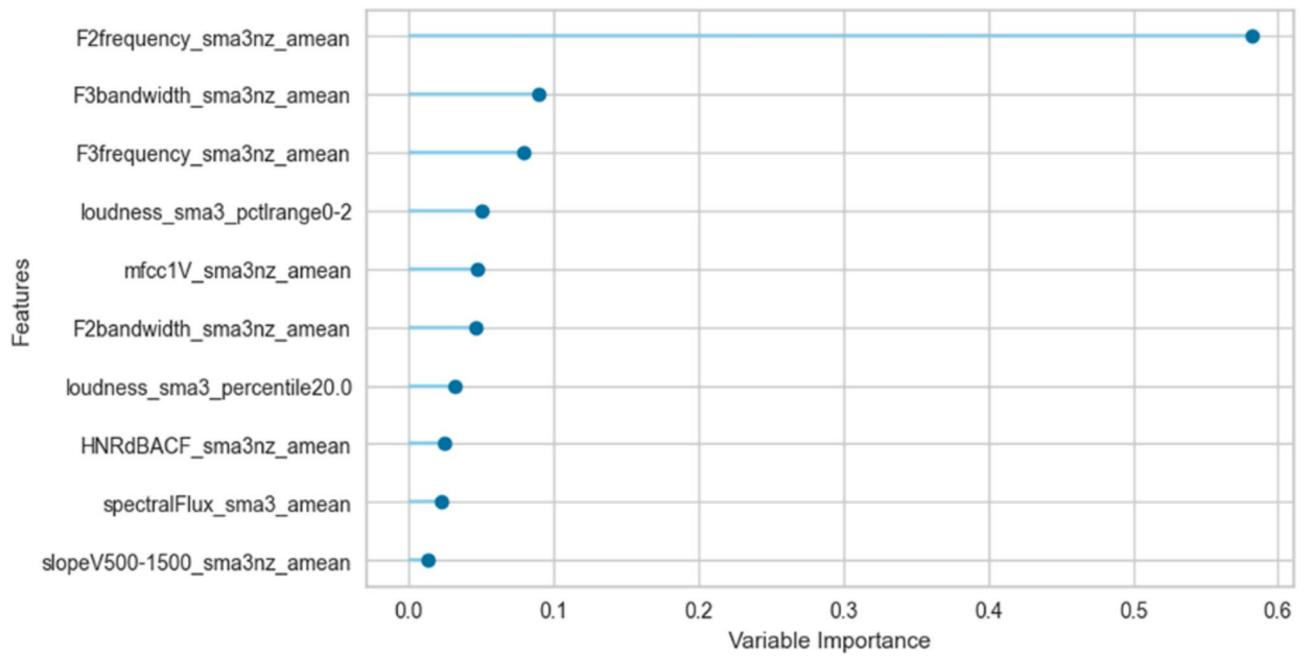
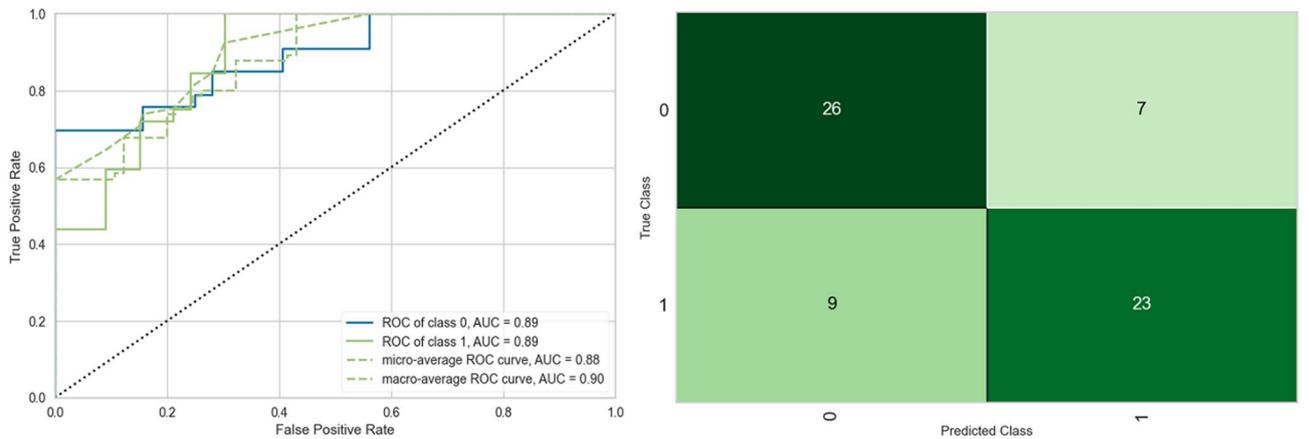**Fig. 21**. Feature importance of DecisionTreeClassifier on test set.

Fig. 22. ROC curves and confusion matrix for NaiveBayes on test set.

context where a low F2 frequency contributes to a "Healthy" prediction, even though that is contrary to the general rule. LIME, by creating a simple linear model in a tiny local neighborhood, captures this localized behavior. SHAP, while also providing local explanations, presents a global summary that averages these effects and reveals the more dominant, overarching trend.

In summary, combining SHAP and LIME provides a more complete and nuanced picture. While SHAP defines robust, global rules the model has learned, LIME can probe specific instances to reveal local exceptions and highlight the complex, non-linear interactions that allow the model to achieve high accuracy. This dual analysis demonstrates that the model is not a simple "black box" but a complex system whose decisions, while intricate, can be systematically explored and understood.

Comparative analysis and clinical context

A critical evaluation of this study's performance involves comparing it to existing non-invasive asthma detection techniques, including the clinical gold standard, spirometry, and other audio-based classifiers. While spirometry is the standard, its efficacy is often compromised by its dependence on patient cooperation and technical proficiency, and its limited sensitivity to issues in minor airways. Our method circumvents these issues by using an effortless vocalization. When compared quantitatively to modern, AI-assisted spirometry-related methods, our model demonstrates a distinct advantage. For instance, the study by Emeryk et al.[29]. reported an AUC of
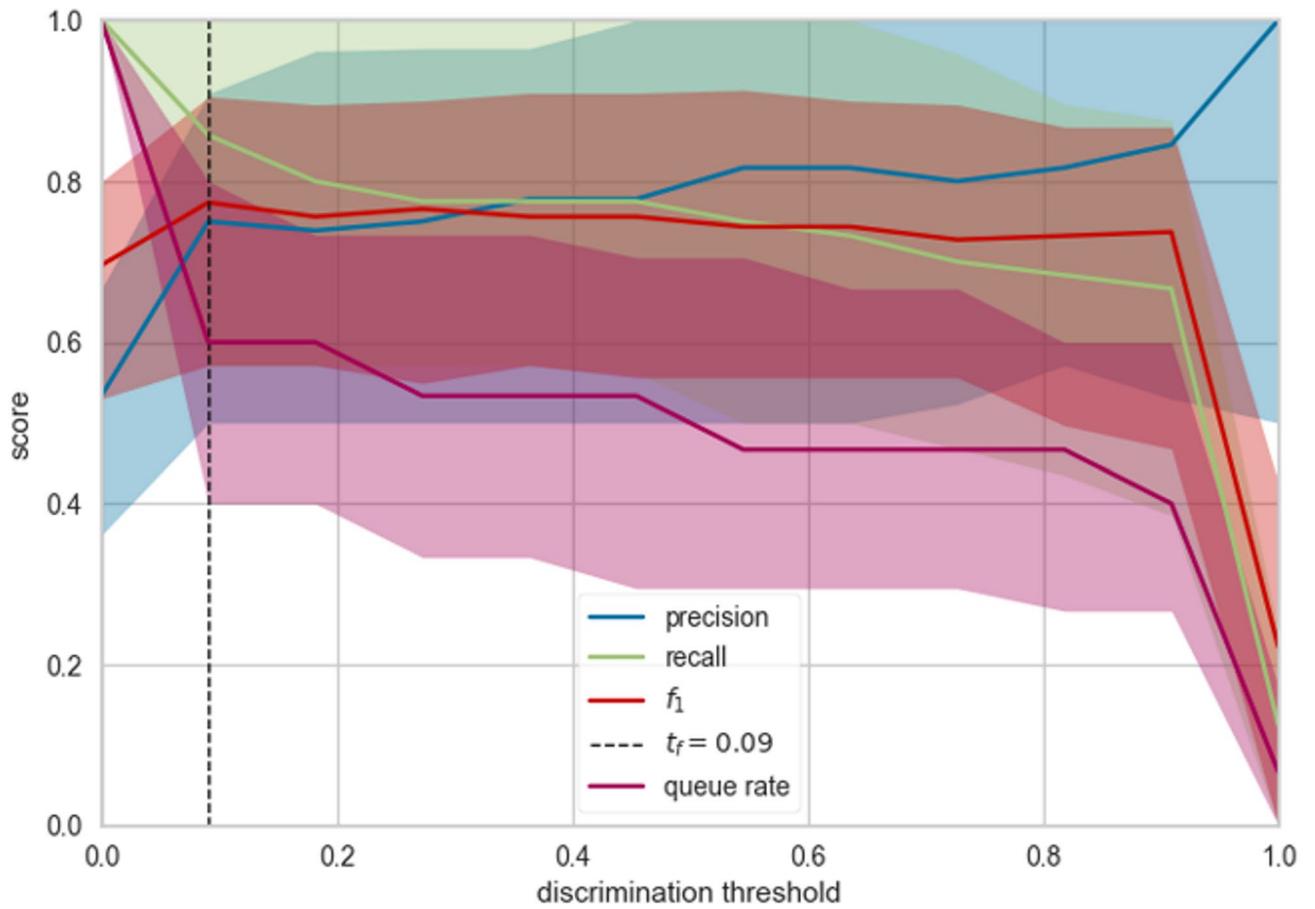
Fig. 23. Threshold plot for NaiveBayes on test set.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8667	0.9821	1.0000	0.7778	0.8750	0.7368	0.7638
1	0.9333	1.0000	1.0000	0.8750	0.9333	0.8673	0.8750
2	0.8667	0.8750	0.8571	0.8571	0.8571	0.7321	0.7321
3	0.8667	0.9286	0.8571	0.8571	0.8571	0.7321	0.7321
4	0.9333	1.0000	0.8750	1.0000	0.9333	0.8673	0.8750
5	0.8667	0.9107	0.8750	0.875	0.8750	0.7321	0.7321
6	0.8000	0.9286	0.8750	0.7778	0.8235	0.5946	0.6001
7	0.8667	0.9464	0.8750	0.875	0.8750	0.7321	0.7321
8	0.8000	0.8036	0.8750	0.7778	0.8235	0.5946	0.6001
9	0.7143	0.7551	0.7143	0.7143	0.7143	0.4286	0.4286
Mean	0.8514	0.9130	0.8804	0.8387	0.8567	0.7018	0.7071
Std	0.0622	0.0774	0.0758	0.0757	0.0594	0.1254	0.1275

Table 5. Tenfold cross-validation result of XGBoostClassifier.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Extreme gradient boosting	0.8100	0.8755	0.8300	0.7981	0.8137	0.6200	0.6205

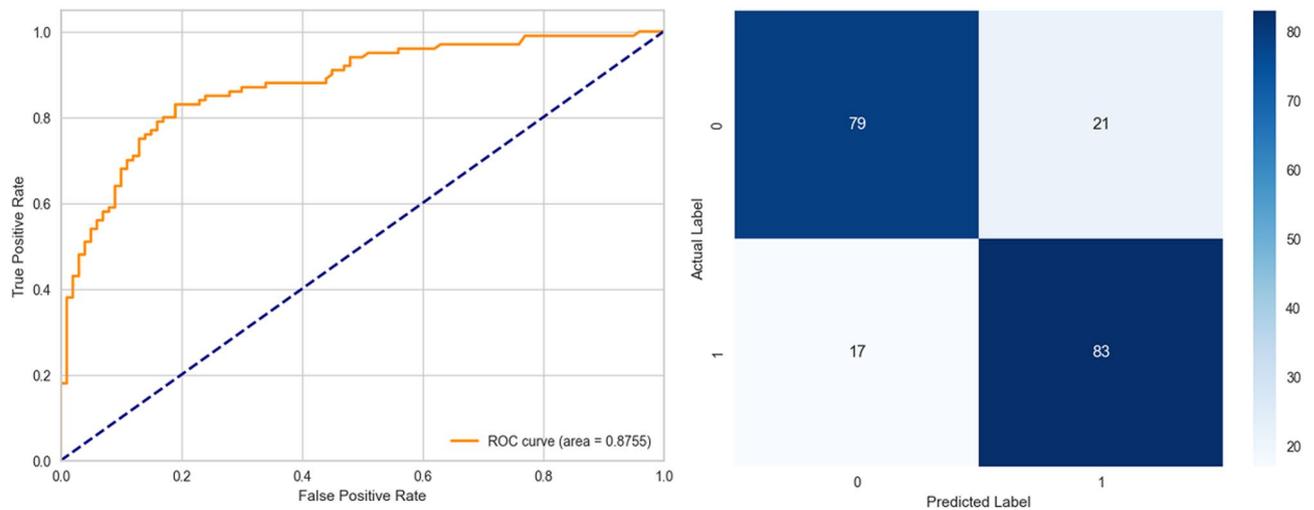**Table 6**. External validation set result of XGBoostClassifier.

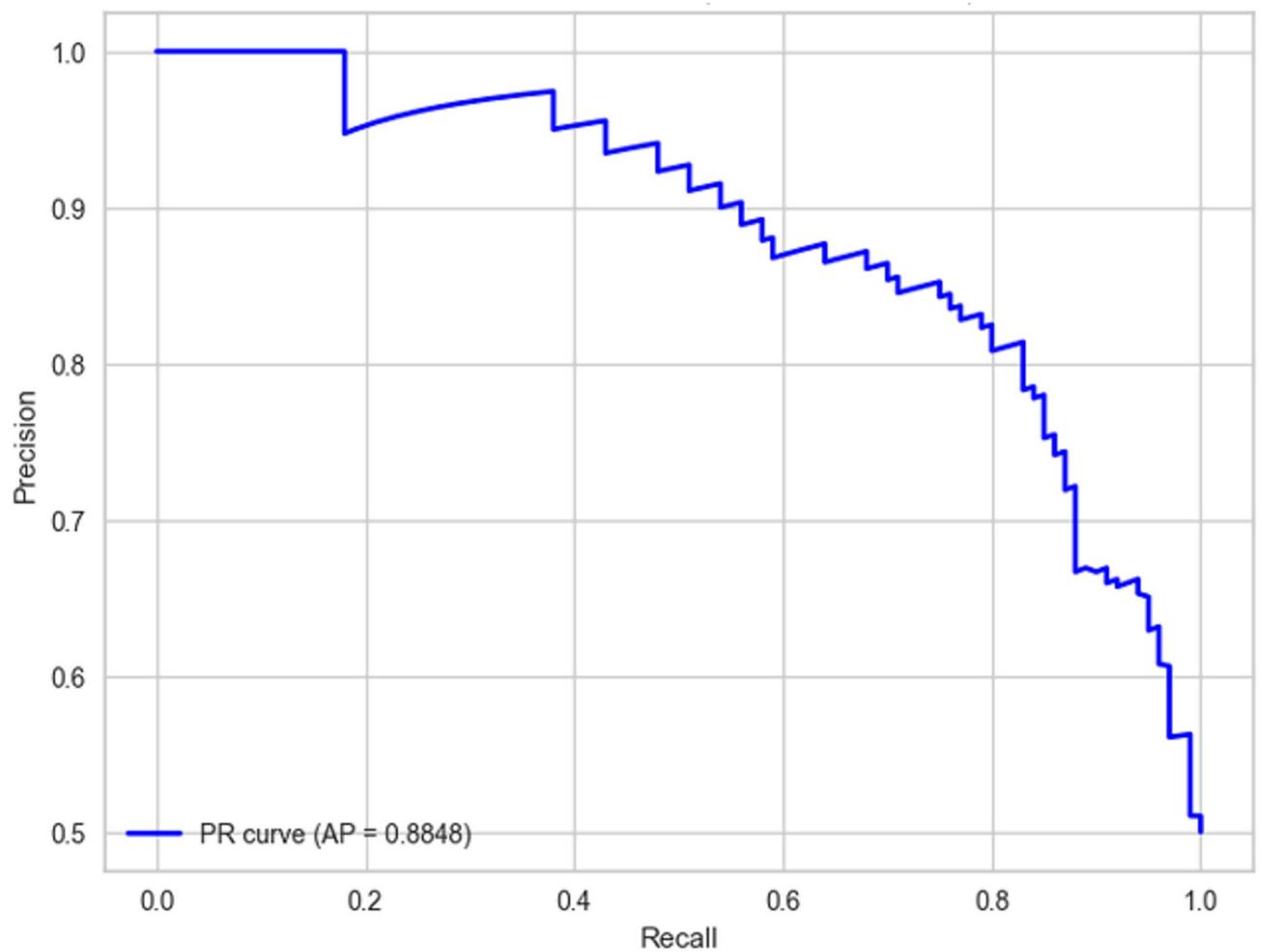**Fig. 24**. ROC curve and confusion matrix for XGBoostClassifier on external validation set.

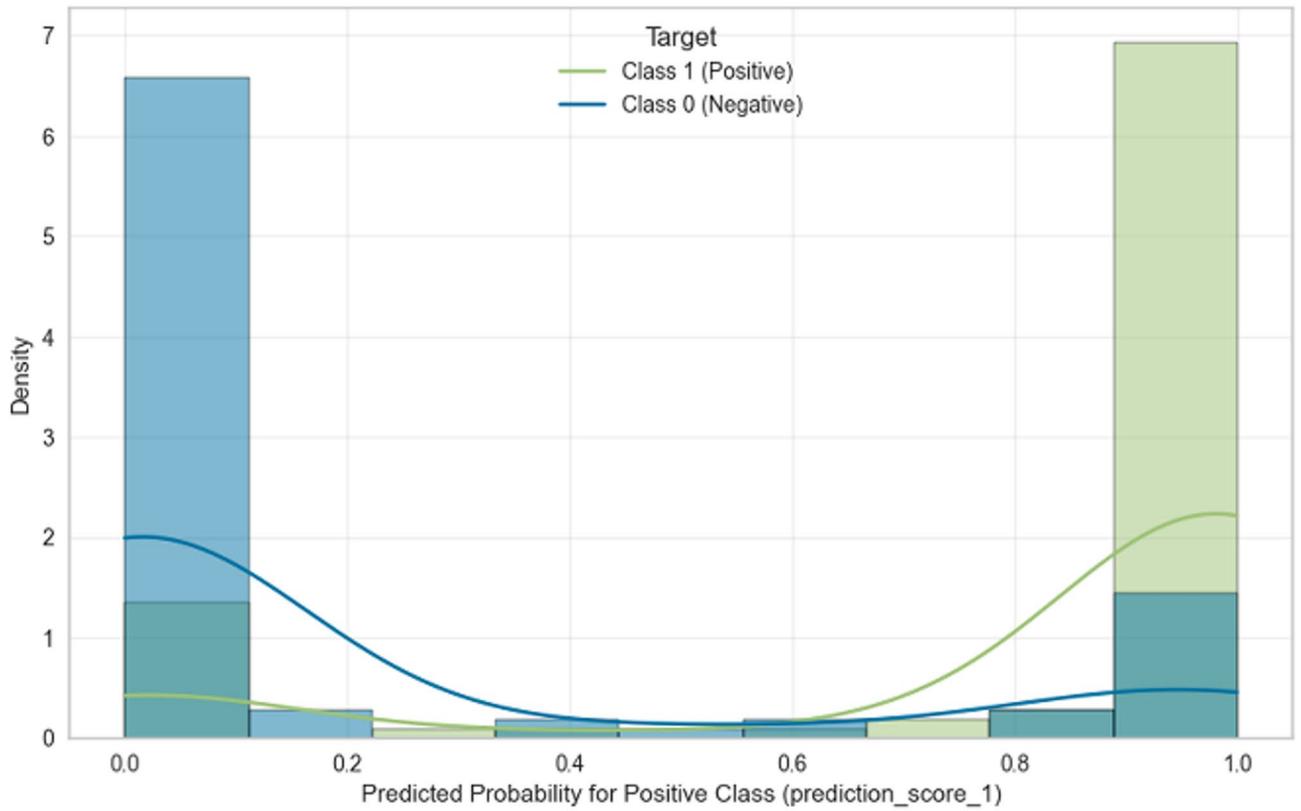**Fig. 25**. Precision-recall curve for XGBoostClassifier on external validation set.

Fig. 26. Prediction score distribution (by True Class "Target") for XGBoostClassifier on external validation set.
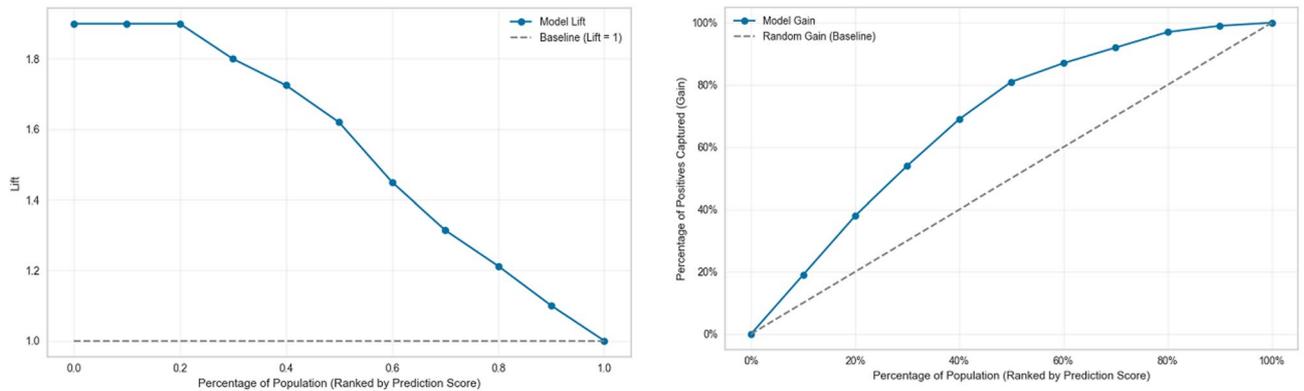

Fig. 27. Lift chart and gain chart for XGBoostClassifier on external validation set.

0.715 for PEF analysis in adults, which is substantially lower than the 0.8755 AUC our model achieved on its independent external validation cohort.

In the domain of audio-only classifiers, cough sounds are a common biomarker. However, our model's external validation accuracy of 0.81 surpasses the 0.7872 accuracy reported for cough-based classification by Yadav et al.[49] Notably, some previous research, such as the work by Balamurali et al.[25] found that vowel-based models performed worse than cough-based models. Our study's superior results suggest that the key to success is not just the choice of biomarker, but the powerful combination of a stable vocalization (the /ɑː/ vowel) with a highly effective, empirically selected classifier (XGBoost) and a robust feature set (eGeMAPS).

In summary, the proposed framework is not only more practical and convenient than effort-dependent tests like spirometry but also demonstrates a superior or highly competitive classification performance compared to other audio-based non-invasive methods. This positions it as an advancement with strong potential for clinical applications, especially in telemedicine and remote monitoring scenarios.
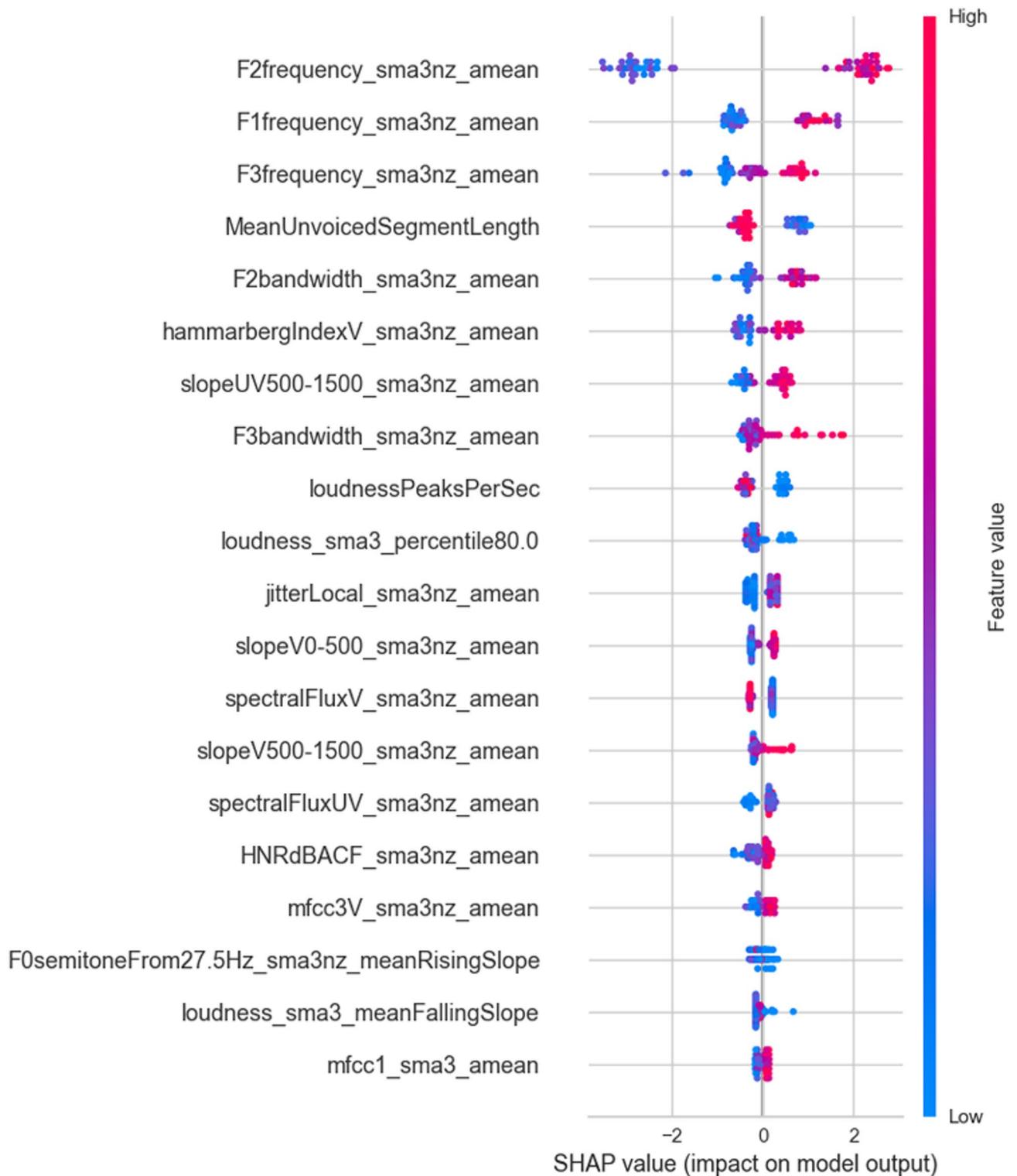
Fig. 28. SHAP value of features for XGBoostClassifier.

Threats to validity

The generalizability of findings to broader populations and settings may be constrained by several factors:

- Population specificity: This study population was comprised exclusively of adults recruited from specific hospitals in Shanghai. Therefore, the model's performance on pediatric patients, who represent a significant portion of the asthma population, remains unevaluated. Furthermore, its applicability to populations with
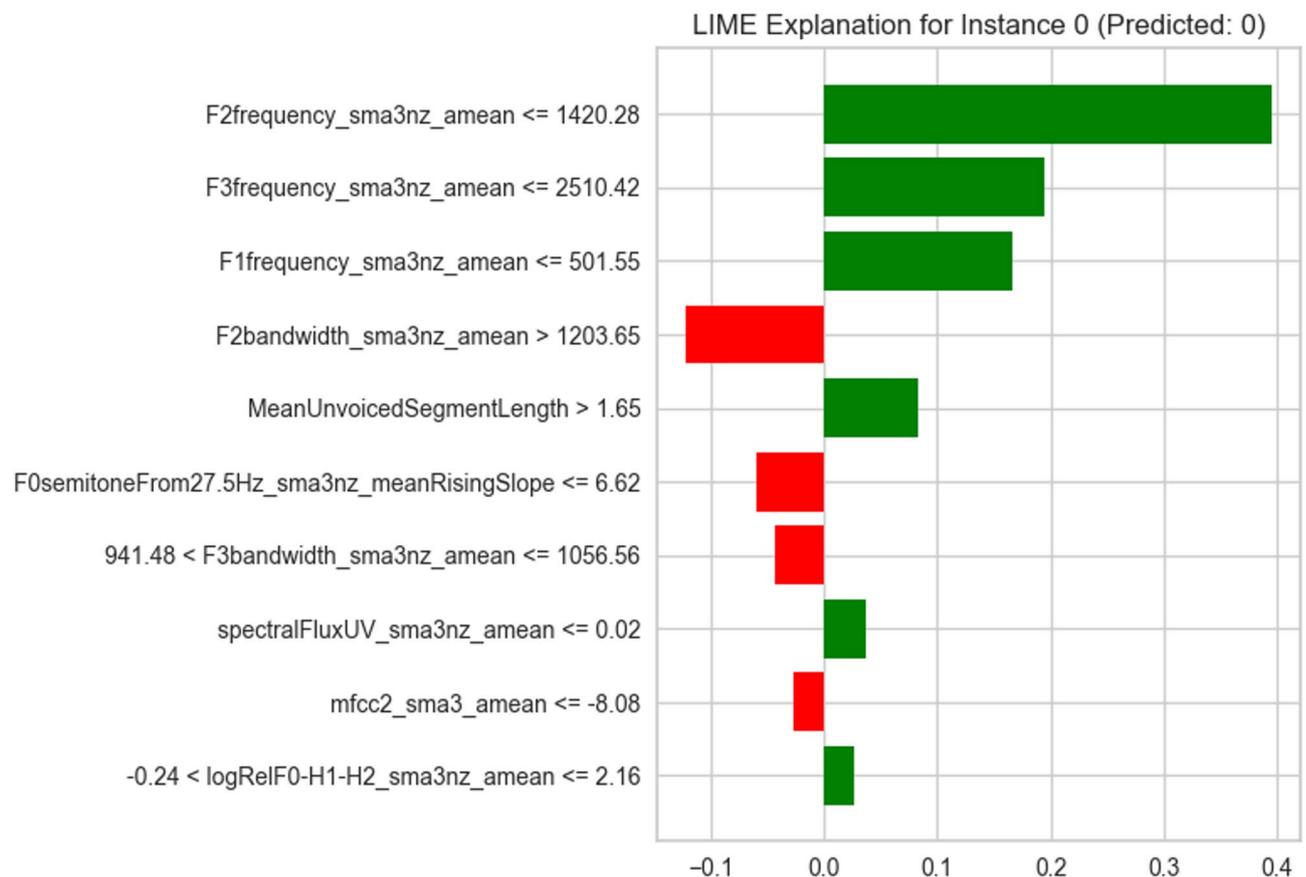
Fig. 29. LIME explanation.

different genetic backgrounds, environmental exposures, or language accents outside of this specific demographic is unknown.

- Data acquisition hardware: The voice data was collected using a single model of headset microphone (Logitech H340) in a controlled, low-noise environment. The model's robustness and performance may vary when used with different recording equipment, such as built-in smartphone microphones, which are common in telemedicine settings.
- Limited acoustic stimuli: This research focused solely on the sustained vowel /ɑː/. While this ensures consistency, it is a potential threat to generalizability as the model was not trained or tested on other phonemes, coughs, or wheezing sounds, which may also contain relevant pathological information.

Construct validity concerns whether the intended underlying concept is truly being measured—in this case, the acoustic manifestation of asthma.

- Completeness of the Pathological Construct: While the model achieved high accuracy, using only the /ɑː/ sound may not capture the full multidimensional acoustic changes caused by asthma. Pathological changes in the airways could manifest differently across a wider range of vocalizations. Future research incorporating a more diverse set of sounds could build a more comprehensive and robust diagnostic construct.

This category relates to the statistical power and robustness of conclusions.

- Sample Size: Although this study included a primary cohort of 214 participants and an external validation cohort of 200, larger datasets are essential in machine learning to further enhance model robustness, improve generalization, and minimize the risk of overfitting. These findings, while strong, should be further validated on larger and more diverse cohorts.
- Lack of Prospective Clinical Validation: A primary limitation of this study is the retrospective nature of its validation. Although the model demonstrated strong performance on an external dataset, it has not undergone prospective clinical validation. This means its performance has not been tested in a real-time clinical workflow, and its outputs have not been used by clinicians to aid in actual patient diagnosis. The true clinical utility and impact of the model on diagnostic accuracy, efficiency, and patient outcomes can only be determined through prospective trials, which represents a crucial next step for this research.

Speech is a good indicator of pathological conditions, especially respiratory diseases[56]. The GeMAPS is based on an automated extraction system that reliably extracts acoustic parameter sets from clean audio waveforms without the need for manual interaction or correction. GeMAPS is a minimal set of speech parameters with 62 features. Research has demonstrated that GeMAPS or the eGeMAPS outperforms the two largest parameter sets, the Computational Paralinguistics Challenge (ComParE) and the Interspeech 2012 speaker trait set (InterSp12), in terms of arousal[51]. GeMAPS is a widely utilised tool for the extraction and identification of emotion-distinguishing parameters[57]. But in respiratory diseases, there are also changes in acoustic characteristics caused by emotional changes. Consequently, it is theoretically feasible for GeMAPS to be used to identify respiratory diseases, a hypothesis that has been confirmed by studies. Jing Han et al.[58] analyzed the voice records of patients infected with COVID-19, employing two data sets from eGeMAPS and ComParE to extract features and used SVMs as the classifier. The average accuracy of both datasets was 0.69 in estimating disease severity; however, the advantage of eGeMAPS was that it consisted of only 88 features from 25 LLDs. Despotovic et al.[59] used GeMAPS, eGeMAPS and ComParE feature sets to extract cough signals from COVID-19 patients, but the author only employed the ComParE set for feature selection. The findings demonstrated that GeMAPS and eGeMAPS exhibited inferior performance in comparison to the ComParE set across all six classification models. Nevertheless, they also exhibited the capacity to discern intrinsic features from coughs.

Although XGBoost and eGeMAPS has gained popularity in studies classifying respiratory diseases, there are very few studies that specifically use it to classify adult asthma based on acoustic features. This study specifically focuses on asthma, in contrast to previous studies that used coughing and wheezing as sound stimulus for this classification task. This paper obtained pathological or physiological audio materials exclusively through the pronunciation of the monosyllabic vowel /ɑː/. Acoustic features were extracted from the acquired audio using eGeMAPS, and the classification performance of various ML models was evaluated based on tenfold cross-validation. XGBoost emerged as the top performer among all ML classifiers. In terms of data acquisition, this paper only used single-syllable vowel as raw material, which is a type of verbal sound. The active collection of vowel phonation data ensures the stability and reproducibility of audio acquisition. Furthermore, single-syllable phonation exhibits high consistency, whilst coughs or respiratory sounds are susceptible to interference from environmental noise, resulting in elevated data heterogeneity. Compared to having patients speak a long string of words as audio data, obtaining single-syllable vowel is more convenient and can reduce patient discomfort to some extent. Most importantly, it can mitigate pronunciation differences caused by dialects, thereby facilitating the standardisation of the features obtained. Although the model's performance did not exceed that reported for XGBoost models in other respiratory disease contexts, this study utilized a wider range of evaluation metrics (including accuracy, AUC, recall, precision, F1-score, Kappa, and MCC) to enhance the reliability of the results. Compared to studies that also used vowel sounds as research data, this model exhibits considerably higher levels of accuracy, precision, recall, and AUC. Traditional non-invasive asthma diagnostic tests encompass spirometry and cough, among others. The AUC of the inspiration-to-expiration ratio and PEF obtained by Emeryk et al. through training ML models was only 0.621 and 0.715 in adults, respectively[29]. Yadav et al.[49] trained an SVM model to compare the classification efficacy of cough, wheezing, and several vowel pronunciations. The findings indicated that the classification accuracy of coughs was 0.7872, which was lower than that of vowel sounds in the aforementioned study. It is evident from the results that the classification performance of trained model is significantly superior to that of traditional non-invasive asthma detection techniques.

Asthma is a respiratory disease that affects a considerable proportion of the global population, particularly children. Consequently, clinicians are increasingly focusing on the convenience, rapidness and high accuracy of asthma diagnosis and monitoring. Pulmonary auscultation remains a pivotal component of the physical examination, surpassing the efficacy of alternative respiratory examination methods. No other clinical procedure can provide clinical information with such expediency and accessibility[21]. In recent years, there has been increasing interest in using verbal occurrence and non-verbal vocalization as acoustic markers of asthma, particularly in the context of the continuing improvement in passive measurement capabilities provided by wearable digital devices. A number of studies have attempted to develop wearable devices that can detect the acoustic characteristics of the lungs and apply them to diagnosis and treatment, making a significant contribution to the medical industry[60–62]. The development of portable devices, exemplified by the integration of an electronic stethoscope with a recorder, and the subsequent storage and analysis of data collected by both the stethoscope and recorder with a smartphone application that employs ML classifiers and DL models, has rendered lung auscultation a more expeditious process[21]. In consideration of technologies such as ML and wireless transmission, acoustic detection has the potential to facilitate the implementation of telemedicine for conditions such as asthma[20]supporting doctors to remotely assess patients' conditions. This approach is particularly well-suited to the provision of follow-up for chronic asthma and contactless diagnosis and treatment during a pandemic. As research into COVID-19 intensifies, the GeMAPS and eGeMAPS technologies are gradually being valued in the study of respiratory diseases. However, their application in the study of the identification and diagnosis of asthma remains insufficient, more studies are needed to prove that they have more superiority compared with other feature extractions. XGBoost, combined with noise reduction principles, has been proven to be helpful in training DL models to select reliable features and use them for the diagnosis of respiratory diseases[48]. In comparison with ML, DL models (e.g. CNN and RNN) demonstrate superior performance in the processing of high-dimensional, unstructured acoustic data. This enhanced capability enables the extraction of a greater number of layers of features, thereby leading to an improvement in classification accuracy. In future studies, DL models can be introduced to explore their diagnostic and classification performance for asthma.

The selection of an optimal machine learning classifier is highly dependent on the specific characteristics of the dataset, a concept formalized by Wolpert's "no free lunch" theorem[63,64]which posits that no single algorithm is inherently superior for all problems. While several powerful gradient boosting algorithms such as LightGBM and CatBoost have demonstrated state-of-the-art performance in various domains, the "no free lunch" theorem

cautions against assuming their superiority for specific task of acoustic asthma classification without direct empirical evidence.

Therefore, rather than pre-selecting a classifier based on its general popularity, an empirical, data-driven approach was adopted. A comprehensive comparative analysis of seven different machine learning models, including multiple advanced gradient boosting machines (XGBoost, LightGBM, CatBoost) was conducted. Results, presented in Table 3, provide a clear justification for choice. On specific dataset, the XGBoost classifier demonstrated a superior balance of performance across multiple key metrics, achieving the highest accuracy (0.8514), F1-score (0.8567), and Kappa Coefficient (0.7018), while also maintaining a highly efficient computation time. While LightGBM and CatBoost are highly optimized for speed, XGBoost's built-in L1 and L2 regularization, which helps control over-fitting, may have provided a performance advantage on high-dimensional acoustic feature set. The findings underscore the importance of empirical validation for model selection in novel application domains.

Conclusion
This study successfully developed and validated an XGBoost model for the non-invasive classification of adult asthma using acoustic features from the sustained vowel /ɑː/. The model demonstrated robust performance and generalizability, achieving an accuracy of 0.81 and an AUC of 0.8755 on an independent external validation cohort. The results confirm that this machine learning approach to vocal biomarker analysis is an accurate and viable method, holding significant potential for the development of accessible tools for early diagnosis and remote asthma management.

Future works
To accomplish this long-term objective, several potential challenges must be addressed: (1) As demonstrated in the study by Balamurali et al.[25] the smaller sample size of the /ɑː/ sound model compared to the cough sound model resulted in inferior outcomes. In subsequent research phases, collecting more audio data from asthma patients will be necessary to enrich the database and refine the model. (2) This study focused exclusively on the monosyllabic vowel /ɑː/ and did not include other vowels (e.g., /iː/, /oʊ/), thereby constraining the model's capacity to capture multidimensional pathological features. Future research should combine a wider range of acoustic parameters and multiple types of sound data (e.g., introducing more vowel sounds, as /oʊ/ has shown strong performance among verbal stimuli[19] to improve classification specificity. (3) This study only included adult asthma patients and excluded children, limiting the model's applicability. Future studies should verify the model's robustness by including a broader range of subjects.

Data availability
The datasets generated and/or analyzed during the current study are not publicly available due to ethical concern but are available from the corresponding author on reasonable request.

References
1. Asher, I. & Pearce, N. Global burden of asthma among children. *Int. J. Tuberc. Lung Dis.* **18**, 1269–1278. https://doi.org/10.5588/ijtld.14.0170 (2014).
2. Enilari, O. & Sinha, S. The global impact of asthma in adult populations. *Ann. Glob. Health*. **85** https://doi.org/10.5334/aogh.2412 (2019).
3. Dharmage, S. C., Perret, J. L. & Custovic, A. Epidemiology of asthma in children and adults. *Front. Pediatr.* **7**, 246. https://doi.org/10.3389/fped.2019.00246 (2019).
4. Melén, E. et al. Asthma inception: epidemiologic risk factors and natural history across the life course. *Am. J. Respir. Crit. Care Med.* **210**, 737–754. https://doi.org/10.1164/rccm.202312-2249SO (2024).
5. Pijnenburg, M. W., Frey, U., De Jongste, J. C. & Saglani, S. Childhood asthma: pathogenesis and phenotypes. *Eur. Respir. J.* **59** (2022).
6. Rusconi, F. et al. Paracetamol and antibiotics in childhood and subsequent development of wheezing/asthma: association or causation? *Int. J. Epidemiol.* **40**, 662–667. https://doi.org/10.1093/ije/dyq263 (2011).
7. Antova, T. et al. Exposure to indoor mould and children's respiratory health in the PATY study. *J. Epidemiol. Community Health*. **62**, 708–714. https://doi.org/10.1136/jech.2007.065896 (2008).
8. Al Efraij, K. & FitzGerald, J. M. Current and emerging treatments for severe asthma. *J. Thorac. Dis.* **7**, E522–525. https://doi.org/10.3978/j.issn.2072-1439.2015.10.73 (2015).
9. Gauvreau, G. M. et al. Sounding the alarmins—the role of alarmin cytokines in asthma. *Allergy* **78**, 402–417 (2023).
10. Semmad, A. & Bahoura, M. Comparative study of respiratory sounds classification methods based on cepstral analysis and artificial neural networks. *Comput. Biol. Med.* **171**, 108190 (2024).
11. Mims, J. W. Asthma: definitions and pathophysiology. *Int. Forum Allergy Rhinol.* **5** (Suppl 1), S2–S6. https://doi.org/10.1002/alr.21609 (2015).
12. Shimoda, T., Nagasaka, Y., Obase, Y., Kishikawa, R. & Iwanaga, T. Prediction of airway inflammation in patients with asymptomatic asthma by using lung sound analysis. *J. Allergy Clin. Immunol. Pract.* **2**, 727–732. https://doi.org/10.1016/j.jaip.2014.06.017 (2014).
13. Tang, H. H. F., Sly, P. D., Holt, P. G., Holt, K. E. & Inouye, M. Systems biology and big data in asthma and allergy: recent discoveries and emerging challenges. *Eur. Respir. J.* **55** https://doi.org/10.1183/13993003.00844-2019 (2020).
14. Rosas-Salazar, C. et al. Respiratory syncytial virus infection during infancy and asthma during childhood in the USA (INSPIRE): a population-based, prospective birth cohort study. *Lancet* **401**, 1669–1680 (2023).
15. Yang, W. et al. Dietary factors and risk for asthma: A Mendelian randomization analysis. *Front. Immunol.* **14**, 1126457. https://doi.org/10.3389/fimmu.2023.1126457 (2023).
16. Peters, U., Dixon, A. E. & Forno, E. Obesity and asthma. *J. Allergy Clin. Immunol.* **141**, 1169–1179. https://doi.org/10.1016/j.jaci.2018.02.004 (2018).

17. Ntontsi, P., Photiades, A., Zervas, E., Xanthou, G. & Samitas, K. Genetics and epigenetics in asthma. *Int. J. Mol. Sci.* **22** https://doi.org/10.3390/ijms22052412 (2021).
18. Kabesch, M. & Tost, J. Recent findings in the genetics and epigenetics of asthma and allergy. *Semin. Immunopathol.* **42**, 43–60. https://doi.org/10.1007/s00281-019-00777-w (2020).
19. Yadav, S., Keerthana, M., Gope, D. & Ghosh, P. K. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6789–6793 (IEEE).
20. Rao, A., Huynh, E., Royston, T. J., Kornblith, A. & Roy, S. Acoustic methods for pulmonary diagnosis. *IEEE Rev. Biomed. Eng.* **12**, 221–239. https://doi.org/10.1109/rbme.2018.2874353 (2019).
21. Bohadana, A., Izbicki, G. & Kraman, S. S. Fundamentals of lung auscultation. *N. Engl. J. Med.* **370**, 744–751. https://doi.org/10.1056/NEJMra1302901 (2014).
22. Guntupalli, K. K. et al. Evaluation of obstructive lung disease with vibration response imaging. *J. Asthma*. **45**, 923–930. https://doi.org/10.1080/02770900802395496 (2008).
23. Sen, I., Saraclar, M. & Kahya, Y. P. Differential diagnosis of asthma and COPD based on multivariate pulmonary sounds analysis. *IEEE Trans. Biomed. Eng.* **68**, 1601–1610. https://doi.org/10.1109/tbme.2021.3049288 (2021).
24. Yan, Y., van Bemmel, L., Franssen, F. M., Simons, S. O. & Urovi, V. Developing a multi-feature fusion model for exacerbation classification in asthma and COPD. *Comput. Methods Progr. Biomed.* 108796 (2025).
25. B, T. B. et al. Asthmatic versus healthy child classification based on cough and vocalised /ɑ/ sounds. *J. Acoust. Soc. Am.* **148**, El253. https://doi.org/10.1121/10.0001933 (2020).
26. Gurung, A., Scrafford, C. G., Tielsch, J. M., Levine, O. S. & Checkley, W. Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis. *Respir. Med.* **105**, 1396–1403. https://doi.org/10.1016/j.rmed.2011.05.007 (2011).
27. Emmanouilidou, D., McCollum, E. D., Park, D. E. & Elhilali, M. Computerized lung sound screening for pediatric auscultation in noisy field environments. *IEEE Trans. Biomed. Eng.* **65**, 1564–1574. https://doi.org/10.1109/tbme.2017.2717280 (2018).
28. Sharan, R. V., Abeyratne, U. R., Swarnkar, V. R. & Porter, P. Automatic croup diagnosis using cough sound recognition. *IEEE Trans. Biomed. Eng.* **66**, 485–495. https://doi.org/10.1109/tbme.2018.2849502 (2019).
29. Emeryk, A. et al. Home monitoring of asthma exacerbations in children and adults with use of an AI-Aided stethoscope. *Ann. Fam Med.* **21**, 517–525. https://doi.org/10.1370/afm.3039 (2023).
30. Huang, D. M. et al. Deep learning-based lung sound analysis for intelligent stethoscope. *Military Med. Res.* **10**, 44 (2023).
31. Park, J. S., Park, S. Y., Moon, J. W., Kim, K. & Suh, D. I. Artificial intelligence models for pediatric lung sound analysis: systematic review and Meta-Analysis. *J. Med. Internet Res.* **27**, e66491. https://doi.org/10.2196/66491 (2025).
32. Ruchonnet-Métrailler, I., Siebert, J. N., Hartley, M. A. & Lacroix, L. Automated interpretation of lung sounds by deep learning in children with asthma: scoping review and strengths, weaknesses, opportunities, and threats analysis. *J. Med. Internet. Res.* **26**, e53662 (2024).
33. Hafke-Dys, H. et al. Artificial intelligence approach to the monitoring of respiratory sounds in asthmatic patients. *Front. Physiol.* **12**, 745635. https://doi.org/10.3389/fphys.2021.745635 (2021).
34. Palkar, A., Dias, C. C., Chadaga, K. & Sampathila, N. Empowering glioma prognosis with transparent machine learning and interpretative insights using explainable AI. *IEEE Access.* **12**, 31697–31718 (2024).
35. Chadaga, K., Prabhu, S., Sampathila, N. & Chadaga, R. A machine learning and explainable artificial intelligence approach for predicting the efficacy of hematopoietic stem cell transplant in pediatric patients. *Healthc. Anal.* **3**, 100170 (2023).
36. Goenka, N. et al. A regularized volumetric ConvNet based alzheimer detection using T1-weighted MRI images. *Cogent Eng.* **11**, 2314872 (2024).
37. Ulukaya, S., Sen, I. & Kahya, Y. P. Feature extraction using time-frequency analysis for monophonic-polyphonic wheeze discrimination. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2015**, 5412–5415. https://doi.org/10.1109/embc.2015.7319615 (2015).
38. Kim, Y. et al. Respiratory sound classification for crackles, wheezes, and Rhonchi in the clinical field using deep learning. *Sci. Rep.* **11**, 17186. https://doi.org/10.1038/s41598-021-96724-7 (2021).
39. Ahmadiankalati, M., Boury, H., Subbarao, P., Lou, W. & Lu, Z. Bayesian additive regression trees for predicting childhood asthma in the CHILD cohort study. *BMC Med. Res. Methodol.* **24**, 262. https://doi.org/10.1186/s12874-024-02376-2 (2024).
40. Tasic, A. et al. Towards sustainable societies: convolutional neural networks optimized by modified crayfish optimization algorithm aided by adaboost and XGBoost for waste classification tasks. *Appl. Soft Comput.* **175**, 113086 (2025).
41. Villoth, J. P. et al. Two-tier deep and machine learning approach optimized by adaptive multi-population firefly algorithm for software defects prediction. *Neurocomputing.* **630**, 129695 (2025).
42. Bacanin, N. et al. The explainable potential of coupling hybridized metaheuristics, xgboost, and SHAP in revealing toluene behavior in the atmosphere. *Sci. Total Environ.* **929**, 172195. https://doi.org/10.1016/j.scitotenv.2024.172195 (2024).
43. Zivkovic, T., Nikolic, B., Simic, V., Pamucar, D. & Bacanin, N. Software defects prediction by metaheuristics tuned extreme gradient boosting and analysis based on Shapley additive explanations. *Appl. Soft Comput.* **146**, 110659 (2023).
44. Salb, M. et al. Enhancing internet of things network security using hybrid Cnn and Xgboost model tuned via modified reptile search algorithm. *Appl. Sci.* **13**, 12687 (2023).
45. Purkovic, S. et al. Audio analysis with convolutional neural networks and boosting algorithms tuned by metaheuristics for respiratory condition classification. *J. King Saud Univ.-Comput. Inform. Sci.* **36**, 102261 (2024).
46. Mouawad, P., Dubnov, T. & Dubnov, S. Robust detection of COVID-19 in cough sounds: using recurrence dynamics and variable Markov model. *SN Comput. Sci.* **2**, 34 (2021).
47. Bijoy, M. H., Mondal, P. K., Plabon, M. F. & Das, D. In *2nd International Conference on Information and Communication Technology (ICICT).*, 95–99 (IEEE, 2024).
48. Varun, B. S., Jonnalagadda, I., Vardhan, C. V., Srinivas, M. & Ullas, S. in *15th International Conference on Computing Communication and Networking Technologies (ICCCNT).*, 1–7 (IEEE, 2024).
49. Yadav, S., Kausthubha, N., Gope, D., Krishnaswamy, U. M. & Ghosh, P. K. in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).*, 1400–1403 (IEEE, 2018).
50. Asthma., G. I. f. *Global Strategy for Asthma Management and Prevention* (2025).
51. Eyben, F. et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **7**, 190–202 (2015).
52. Eyben, F., Wöllmer, M. & Schuller, B. In *Proceedings of the 18th ACM international conference on Multimedia* 1459–1462 (Association for Computing Machinery, 2010).
53. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F. & Campbell, J. P. Introduction to machine learning, neural networks, and deep learning. *Transl. Vis. Sci. Technol.* **9**, 14. https://doi.org/10.1167/tvst.9.2.14 (2020).
54. Deo, R. C. Machine learning in medicine. *Circulation.* **132**, 1920–1930. https://doi.org/10.1161/circulationaha.115.001593 (2015).
55. Sun, X., Yin, Y., Yang, Q. & Huo, T. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. *Eur. J. Med. Res.* **28**, 242. https://doi.org/10.1186/s40001-023-01065-y (2023).
56. Dibazar, A. A., Narayanan, S. & Berger, T. W. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology.* 182–183 (IEEE).
57. De Boer, J. et al. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychol. Med.* **53**, 1302–1312 (2023).

58. Han, J. et al. An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety. arXiv preprint arXiv:2005.00096 (2020).
59. Despotovic, V., Ismael, M., Cornil, M., Mc Call, R. & Fagherazzi, G. Detection of COVID-19 from voice, cough and breathing patterns: dataset and preliminary results. *Comput. Biol. Med.* **138**, 104944 (2021).
60. Yu, C., Tsai, T. H., Huang, S. I. & Lin, C. W. Soft stethoscope for detecting asthma wheeze in young children. *Sens. (Basel).* **13**, 7399–7413. https://doi.org/10.3390/s130607399 (2013).
61. Sang, B. et al. An Accelerometer-Based wearable patch for robust respiratory rate and wheeze detection using deep learning. *Biosens. (Basel).* https://doi.org/10.3390/bios14030118 (2024).
62. Oletic, D., Arsenali, B. & Bilas, V. Low-power wearable respiratory sound sensing. *Sens. (Basel).* **14**, 6535–6566. https://doi.org/10.3390/s140406535 (2014).
63. Wolpert, D. The lack of A priori distinctions between learning algorithms. *Neural Comput.* **8**, 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341 (1996).
64. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82. https://doi.org/10.1109/4235.585893 (1997).

Author contributions

Y.L.: Conceptualization, data curation, methodology, software, validation, visualization, and writing—original draft. Q.-C.J.: Conceptualization, formal analysis, and writing—original draft. S.Y.: Methodology, validation, and visualization. J.H.: Data curation, investigation, and resources. C.-F.C.: Formal analysis and investigation. H.-M.W.: Formal analysis and validation. Y.-Q.W.: Data curation, funding acquisition, project administration, and resources. Y.-J.S.: Methodology. H.-X.Y.: Data curation, project administration, and resources. J.X.: Data curation, funding acquisition, supervision, and writing—review and editing.

Funding

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Institutional Review Board of Shanghai University of Traditional Chinese Medicine and the study complied with the Declaration of Helsinki. All participants received informed consent form and signed.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.-X.Y. or J.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.