RESEARCH
O&G Forum 2024; 34 – 2s: 225 - 229

ENHANCED MACHINE LEARNING FOR SPIROMETRY CURVE ASSESSMENT IN CHRONIC RESPIRATORY DISEASES

Sumit Kumar Roy[1], Saurabh Gupta[2], Dibakar Sahu[3], Hemlata Sinha[4*]
[1]Department of Biomedical Engineering National Institute of Technology Raipur, India
skroy.phd2022.bme@nitrr.ac.in
[2]Department of Biomedical Engineering National Institute of Technology Raipur, India
sgupta.bme@nitrr.ac.in
[3]Department of Pulmonary Medicine AIIMS Raipur, India
dibakarsahu@aiimsraipur.edu.in
[4]Department of ETC, SSIPMT Raipur, India
sinha.hemlata552@gmail.com

Abstract
Chronic respiratory diseases, including Chronic Obstructive Pulmonary Disease (COPD) and asthma, pose a significant global health challenge with far-reaching economic implications. This study introduces a machine learning-based approach for predicting exacerbation risks in respiratory dis- eases, specifically focusing on COPD and asthma. The frame- work employs an advanced machine learning architecture to enable real-time and precise detection of respiratory events. spirometry standards mandate that correct maneuvers shouldbedevoidofcoughartifacts,especiallyintheinitialseconds of forced exhalation. Achieving this standard becomes inherently challengingwhenpatientsmanifestincreasedcoughingtendencies duringtheexamination.Thisstudyaimtofacilitateunsupervised or minimally supervised spirometry measurements, expanding accessibility in diverse settings, including home monitoring and general practitioner oversight. By prioritizing specificity in the machinelearningtechniques,themodelachievesacomprehensive understanding and accurate classification of symptoms, thereby contributing to personalized risk assessment.

Index Terms—Chronic Obstructive Pulmonary Disease, Asthma, Machine Learning, Spirometry, Random Forest, KMean

Abstract—

I. INTRODUCTION

In the era of burgeoning big data, opportunities and challenges abound in understanding, analyzing, and extracting knowledge. Traditional statistical analyses, once adequate, may now fall short in the face of pervasive measurability. Amidst this backdrop, the healthcare sector stands to gain immensely from identifying shared patterns across diverse patient datasets. This research focus on two prevalent respiratoryconditions: asthma and CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) [1].

Asthma, characterized by chronic airway inflammation, triggers bronchospasm and airway constriction upon exposure to environmental stimuli. While a genetic predisposition may exist, symptoms typically manifest following exposure to triggerviralinfectionandallergy.COPDencompasseschronic bronchitis and emphysema, both contributing to persistent air- way obstruction and daily breathlessness, with smoking being a primary contributor. Globally, asthma affects approximately 300 million patients, leading to 250,000 annual deaths, while COPD afflicts 330 million patients, causing around 3 million deaths each year. The urgent need to develop tools for the early prediction and diagnosis of these respiratory diseases is evident [2].This study employs machine learning techniques to scrutinize the factors underpinning the diagnosis of asthma and COPD. Aligned with existing research, our methodology exploresboth linear and non-linear models, placing emphasis on the importance of complexity and interpretability [3].

Inthislandscape,effortstodetectcoughsusingdiversesensors such as ECG sensors, thermistors, chest belts, and oximeters havebeenexplored.Wearablecoughdetectionsystems,including those integrated into smartwatches for ambulatory monitoring, signify recent innovations in this domain. Intriguingly, cough detection has extended beyond human patients to encompass veterinary monitoring of farm animals, underscoring the versatility and applicability of such algorithms [1] [2].The predominant focus in existing literature on cough detectionhasrevolvedaroundaudio(sound)signalsoraccelerometer recordings. However, this study diverges from this trendby

examining airflow signals passing through the spirometer. Analyzing airflow signals presents distinct advantages, notably the reduction of troublesome environmental noise influences that often plague sound-based approaches. This departure aligns with the overarching objective of developing a robust and accurate cough detection algorithm, particularly for spirometry applications [1] [3].

Some studies concentrated on monitoring cough overtime eandcounting occurrences, others delved into assessing theseverity of various respiratory disorders, including cysticfibrosis, cold, tuberculosis, and COPD.

Problem statement: Current spirometry standards mandate thatcorrectmaneuversshouldbedevoidofcoughartifacts,especiallyi n the initial seconds of forced exhalation. Achieving this standard becomes inherently challenging when patients manifest increased coughing tendencies during the examination. Our work aims to facilitate unsupervised or minimally supervised spirometry measurements, expanding accessibility in diverse settings, including home monitoring and general practitioner oversight.

The subsequent sections of this article are structured as follows: Methodology' details the data sources and tools employed.

'Results' unveils statistical and machine learning outcomes, alongside an exploration of variablesignificance. Finally,' Discussion and Conclusion' presents closingremarks, delineateslimitationsandoutlinespotentialdirectionsforfuture research.

II. Methodology

The cough reflex serves as a critical physiological response to airway irritations, characterized by sudden, forceful expiratory efforts. Initiated by the stimulation of cough receptors within the respiratory tract, this reflex plays apivotal role in clearing irritants and maintaining airway integrity. In the context of pulmonary health, the cough reflex assumes heightened significance, often reflecting respiratory irritation or underlying pathology. Conditions such as asthma, cystic fibrosis, and chronic obstructive pulmonary disease (COPD) frequently present with cough as an early symptom, necessitating precise monitoring and diagnostic approaches[1] [4].

Spirometry, awidely employedmethodfordiagnosing and monitoring pulmonary functions, introduces a notable challenge in the context of patients with chronic pulmonary diseases. The procedure involves a forced expiratory maneuver, requiring considerable effort from the patient. However, individuals afflicted with conditions like COPD commonlyexperienceheightenedlevelsofbreathlessness and cough, particularly during exacerbations. The need for accurate spirometry measurements is paramount for assessing disease progression and treatment efficacy [5].

This work elucidates a solution for automatic cough detection developed specifically Comprising a portable spirometer, a computer and an internet cloud for data storage. This system aims to provide accurate and robust cough detection invariousenvironments,includingclinicalandin-homesettings. The algorithm is designed to analyze the airflow signal transmitted from the spirometer to the computer application during measurements, ensuring the reliability of clinically important parameters such as forced vital capacity (FVC), forced expiratory volume in the first second (FEV1), theirratio (FEV1/FVC), peak expiratory flow (PEF), among others [6].

To facilitate algorithm development, a large dataset of spirometry airflow recordings was utilized for training and validation. Preprocessing methods were applied to enhancethequalityofthedataset,emphasizingtheimportance of meticulous data preparation in ensuring algorithmic accuracy.Theanalyticalmethodsadaptedforconstructing the cough detection algorithm were driven by machine learning approaches, leveraging the capabilities of advanced mathematical techniques [7].

The machine learning algorithms employed were systematically presented, accompanied by an in-depth exploration of their results. The emphasis on high specificity in the algorithm's design was underscored, acknowledgingthe potential consequences of misclassification, such as the unjustifiedrepetitionofmeasurementsoruserdiscouragement.

As spirometry continues to be a cornerstone in respiratory diagnostics, the presented algorithm holds promise for enhancingtheaccuracyandefficiencyofpulmonaryfunction

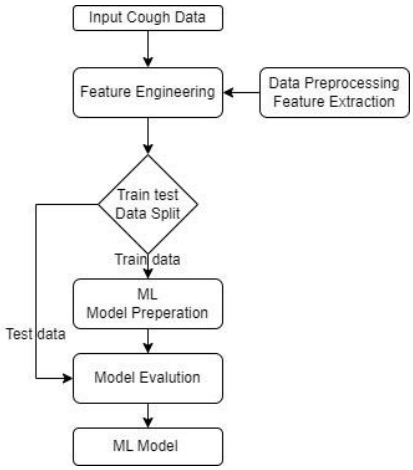

Fig.1.Processflowofthework

assessments in diverse clinical and non-clinical settings. The ongoing evolution of such algorithms signifies a critical step toward advancing respiratory health monitoring, offering potentialbenefitsforbothhealthcareprovidersandpatientsalike [6][7].Thefigure1showstheprocessflowoftheprocess.

A. Dataaccusation

ResearchdataoriginatedfromtheNHANESdatabase,man- aged by the American National Center for Health Statistics[8]. This free resource includes raw spirometry curve data aligned with American Thoracic Society guidelines. Utilizing specificdatabasesubsets[9],diversepatientsunderwentinitial and, if criteria were met, subsequent spirometry examinations with β2-adrenergic bronchodilator inhalation. The dataset en- compasses raw signals from multiple individual spirometry curves, categorized into subsets, including those of acceptable quality, with large time to peak flow or non-repeatable peak flow, without plateau, or with cough. Experienced experts curated two classes: (a) ATS-acceptable curves and (b) other errorcurvesversuscoughcurves.Figure2showgraphsof the Correct

ATS-Acceptable Maneuver Subset, Clear Cough Occurrence Subset and Less Manifested Cough Occurrence Subset respectively.

B. DataPreprocessing

Preprocessing of raw spirometry data is an indispensable precursor to feature extraction, involving systematic proceduresfor standardization and noise mitigation. This scientific approach encompasses several sequential operations:

1) Segmentation of Forced Exhale Signal: Initialization involves the segmentation of the raw curve todelineate the forced exhale signal, a critical maneuver in spirometryassessments.Thissegmentationisparamount to isolate the primary respiratory effort from ancillary componentssuchasinhalesprecedingorsucceedingthe main maneuver.
2) Length Standardization: A subsequent step addressesthe standardization of the forced exhale signal's length.

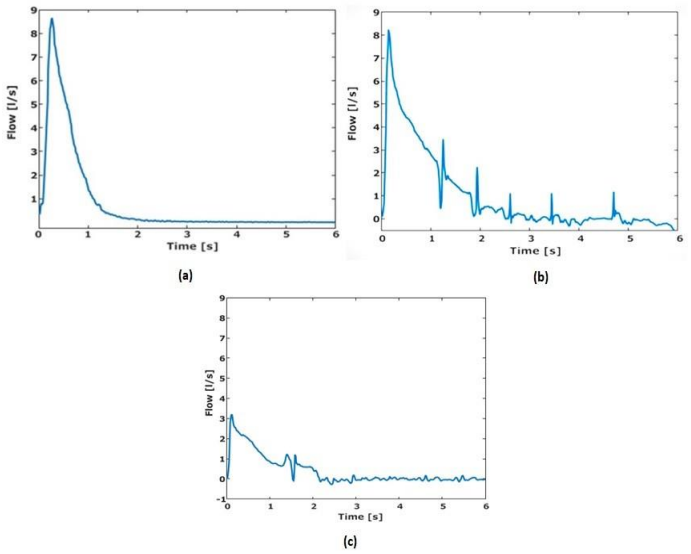

Fig. 2.(a): Correct ATS-Acceptable Maneuver Subset Graph, (b): ClearCough Occurrence Subset Graph and (c): Less Manifested Cough OccurrenceSubset Graph

To established spirometry standards, truncation is done, with particular emphasis on the first second, a temporal window crucial for assessing parameters such as Forced Expiratory Volume in the first second (FEV1).

3) Zeroing Negative Values: Negative values within the signalaresystematicallyzeroed.Whilethisoperation is agnostic to the subsequent feature extraction process, it serves the purpose of diminishing the flow span ofthe data and obliterating residual inhale fragments that might persist from the initial segmentation.
4) Steady-Flow Detection Preprocessing: A specialized preprocessing step is dedicated to discerning the nature of the signal, specifically identifying whether it corresponds to a steady-flow characteristic, typical in scenarios such as Flow/Volume generator. This discernment reliesonameticulousevaluationofdissimilaritymetrics between signal samples and the signal median.

The culmination of these preprocessing steps orchestrates the refinementandhomogenizationofrawspirometrycurves,systematically priming them for subsequent analytical processes. Thisrigorousscientificapproachensuresthestandardizationof signals and the amelioration of potential confounding factors, thereby fortifying the robustness and reliability of ensuing analyses, notably in the context of cough detection within spirometry evaluations [1] [2].

C. FeatureExtraction

Thestudyemploysarigorousfeatureextractionprotocol to distill essential characteristics indicative of cough within individual spirometry curves. The algorithm input is defined by computationally efficient features, meticulously tailored to encapsulate distinctive cough-related attributes. The features are as follows:

1) Number of PEF-Related Local Maxima (Spikes): This feature quantifies local maxima (spikes) occurring post-peak expiratory flow (PEF). The chosen threshold distinguishescough-relevantpeaksfromshorter,potentially inconsequential fluctuations.
2) Number of PEF-Related Local Maxima with Right- Slope Amplitude: The right-slope amplitude is defined as the amplitude between peak maximum and the point where the first derivative undergoes a sign change, indicating the initiation of an amplitude increase.
3) Number of Crossings at Various Percentages of PEF: Deriving four distinct features, this metric counts signal crossings with horizontal lines positioned at 15%, 25%, 50%, and 75% of PEF.

   Correlation analysis is instrumental in gauging the relevance and interdependence of features. The preliminary feature selection process involved assembling a set of potentially useful features,withexclusioncriteriabasedonhighcorrelationwith others [1] [2].

D. MachineLearning

The implementation of machine learning models took place withinthePython-3environment.Acomprehensivesuiteof algorithms underwent systematic training and evaluation, seekingtheoptimalperformeramonglogisticregression(LR), feed-forward artificial neural network (ANN), support vector machines (SVM), and random forest (RF). The F1 score served as the paramount metric for model optimization and performance comparison.

TheANNmodelwascharacterizedbyasigmoidfunction as the transfer function in the neuron model. Employing k- fold cross-validation (k = 10) for validation. LR, integrated withk-foldcross-validation,utilizedthemeansquareerror as the loss function. SVM, employing a radial-type kernel, underwentparametertuningthroughgridsearch.RF,grounded in Breiman's algorithm for classification and regression [1] [10].

The architectural configuration of the final artificial neural network is visually articulated. This methodological rigor in algorithmic selection, parameter tuning, and model validation ensurestherobustnessandreliabilityofthechosenma-chinelearningapproachforcoughdetectionwithinspirometry datasets.

E. ProposedMachinelearningApproach

Employing clustering (unsupervised machine learning technique)and similar techniques aids in consolidating logging responses for classification, offering valuable insights into diverse data types and pattern identification. Analyzing the correlation between logging responses and blood flow enables the understanding of system behavior, issue detection, and overall system enhancement. Utilizing random-forest with KMeanclustering(KMRF)formodelingandforecastingproves robustforpredictingcoughwithinspirometrydatasets.KMean clusteringcategorizesdata,addressingthechallengeofaweak loggingresponseofcoughrelationshipbygroupingsimilardata points. The subsequent application of RF algorithms to each category enhances cough prediction accuracy based on distinct group features within spirometry datasets [10].

F. PerformanceMetrices

Table 1 outlines various metrics employed to assess the performance of classifiers in this study [10].

TABLEI
PERFORMANCEMETRICSOFOURPREDICTIVEMODE

Metrices	Methods
Accuracy	$\frac{tP + tN}{tP + tN + fP + fN}$
Sensitivity	$\frac{tP}{tP + fN}$
F1-score	$2\ X\ (Recall\ X\ Pressi$ / $Recall + Pressior$
Specificity	$\frac{tN}{tN + fP}$
ClassificationError	1-Accuracy

III. RESULTSANDDISCUSSION

In the study, proposed algorithm under scrutiny exhibits noteworthy proficiency in mitigating the influence of cough events and artifacts on spirometric curves, thereby adeptly preservingpertinentinformationduringtheinterstitialperiods. Furthermore, its discerning capacity extends to the extraction of meaningful data from curves traditionally deemed aberrant, thereby overcoming historical impediments in data mining from such atypical sources. The comparative evaluation of diverse machine learning algorithms incorporates meticulous consideration of established metrics encompassing sensitivity, specificity, F1 score, and accuracy. These metrics collectively furnishacomprehensiveandquantitativeassessmentofthealgorithm' sefficacyinspirometrydataanalysis.Thealgorithm's commendable performance, as elucidated through these metrics, underscores its potential significance in the realm of medical diagnostics, particularly in scenarios characterized by irregularities and perturbations in spirometric data.

In the study, cough events were classified from the NHANES database,andtheperformanceofeachMLmodelwasassessed by conducting a series of studies.

Theclassificationprocess,illustratedinFigure1,involvesfour key steps. Firstly, the feature engineering phase generates a fully featured dataset through data preprocessing and feature extraction. Subsequently, 10-Fold Cross-Validation partitions the dataset into ten unique subsets for systematic evaluationof classifiers. Machine learning algorithms are trained using a selection dataset to develop a cough event detection model, emphasizingminimalredundancyandmaximumrelevance infeatureselection.OurstudyexploresLR,ANN,SVM, RF, and proposed KMRF, evaluating each model's efficacy. Classifier performance relies on data quality, with inherent advantages and limitations. Performance metrics, including accuracy, sensitivity, specificity, F1-score, and classification error,assessMLefficiency,asdepictedinTable2.Figure 3 illustrates a comparison of classifier performance based on specificity and sensitivity.

TABLE II
PERFORMANCECOMPARISONOFDIFFERENTML TECHNIQUESFORIDENTIFYING COUGH EVENTS

ML Techniques	**PerformanceMetrices**		
	Accuracy	***F1-Score***	***Classification Error***
LR	0.913	0.812	0.087
ANN	0.912	0.916	0.088
SVM	0.902	0.832	0.098
RF	0.921	0.843	0.079
ProposedKMRF	0.96	0.92	0.04

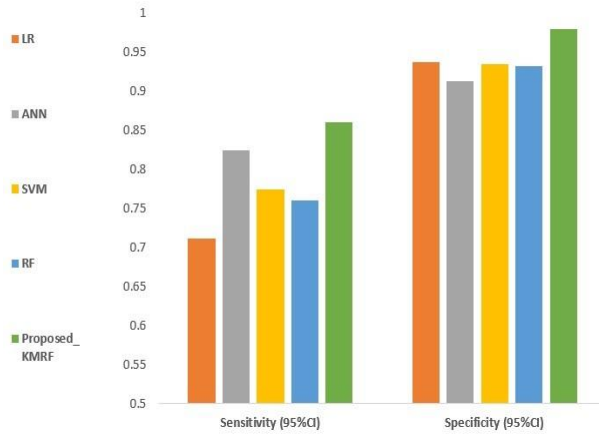

Fig. 3.Performance comparison of the classification model using specificityand sensitivity

This study focuses on detecting cough events in spirometry curves using airflow signals, addressing the complexity of distinguishing subtle cough manifestations from other disturbances. The research incorporates a diverse training dataset from the NHANES database.

The proposed algorithm demonstrates superiority over a previously described approach. Despite some misclassifications, particularly in cases of small cough disturbances or similar disturbances,thealgorithmprovesrobustforclinicalandhome monitoring in spirometry.

Thestudyevaluatestheperformanceofvariousclassifiers in

detecting cough events within spirometry data. Classifier accuracies are reported, with consistent misclassifications observed across models, suggesting feature selection's primary influence. Specific misclassification patterns involve spirometry maneuvers lacking forced exhalation and coughs with spikes only at the end.

The Proposed KMRF algorithms excel in handling high-dimensional datasets and capturing intricate data patterns, enhancing cough event classification. The integration of KM clustering with RF algorithms enables precise feature ex- traction and improved classification. KM clustering reduces feature space dimensionality, facilitating algorithmic identificationofrelevantfeaturesanddatarelationships.Fromtable2a ndfigure3ProposedKMRFoutperformsothermodels in metrics like accuracy, f1-score, specificity, and sensitivity, showcasingeffectivedataclusteringforenhancedprecision andpredictionaccuracy.Thesealgorithmsmitigateoverfitting, boost model generalization, and improve predictive accuracy, achieving a remarkable 96% accuracy in cough event pre-diction. The study emphasizes the importance of algorithm selection and the benefits of employing feature extraction techniques like KM for enhanced prediction accuracy.

Comparisonswithpreviousworkrevealhigherspecificityand overall F1 score, emphasizing the algorithm's effectiveness in minimizing false positives. A preference for specificity aligns with real-world application needs, preventing unnecessary repeated maneuvers.

Balancingthedatasetadverselyaffectsperformance,attributed tothelossofinformationabouteventfrequencies.Despitethis, the model demonstrates reasonable generalization, indicating avoidance of over-training.

The study underscores the significance of feature choice, the balancebetweensensitivityandspecificity,andconsiderations for real-world usability in developing an effective cough detection algorithm for spirometry.Despite these occasional challenges, the overall performance andcapabilitiesofouralgorithmrepresentaconsiderable step forward in the accurate detection of cough events during spirometry. As technology continues to evolve, our research contributestotheongoingeffortstorefinealgorithms,making them even more adept at discerning subtle respiratory patterns and further enhancing the utility of spirometry in both clinical and home settings.

The study acknowledges limitations and proposes future work to enhance the algorithm's performance and reliability. Additionally, the suggestion of modifying the algorithm to provide temporal information on cough occurrences, such as duringthe first second of exhalation, highlights potential avenues for improvement and broader applicability in spirometry quality assessment.

IV. Conclusion

Detectingcougheventsinspirometrycurvesthroughairflow signals poses challenges due to diverse cough manifestations, ranging from clear disruptions to subtle flow disturbances. Leveraging a comprehensive training dataset comprising the NHANES database, the study produced a robust classification algorithm surpassing prior methodologies. The ProposedKMRFrepresentsanotableadvancementinspirometry monitoring for clinical and home settings with 96% accuracy. Despite occasional misclassifications in maneuvers featuring minor cough disturbances, our algorithm's features contribute to its accuracy. Comparative analysis with a competing algorithm demonstrates our superior overall F1 score of 92%, showcasing enhanced precision and recall. The balanced performancesuggestsouralgorithm'sreliabilityforprecisecough eventdetectionduringspirometry,highlightingitspotentialfor refined respiratory health monitoring.

The study acknowledges limitations and proposes future work to enhance the algorithm's performance and reliability.

ACKNOWLEDGEMENT

We wish to thanks the Biomedical Engineering Department of National Institute of Technology, Raipur for assistance and encouragement in preparing this paper.

References

1. *Solin´ski,M.,etal.Automaticcoughdetectionbasedonairflowsignalsfor portable spirometry system. Informatics in medicine unlocked, 18,(2020), 100313.*
2. *SmithJ,WoodcockA.CoughanditsimportanceinCOPD.IntJChronicObstrPulm Dis 2006;1(3):305–14.*
3. *Standardization of Spirometry 2019 Update. An official AmericanthoracicsocietyandEuropeanrespiratorysocietytechnicalstatement. Am J Respir Crit Care Med October 15, 2019;200(8).https://doi.org/10.1164/rccm.201908-1590ST.*
4. *Hofman A, Kupczyk M, Kuna P, et al. Application of the AioCaresystem in monitoring exacerbations of bronchial asthma. Pol J AllergolSpec Issues 2018;(1): A.8. in Polish.*
5. *GoelM,SabaE,StiberM,etal.Spirocall:measuringlungfunctionovera phone call. In: Proceedings of the 2016 CHI Conference on humanFactors in computing systems. ACM; 2016. p. 5675–85.*
6. *LarsonEC,etal.Accurateandprivacypreservingcoughsensingusing a low-cost microphone. In: Proceedings of the 13th internationalconference on ubiquitous computing. ACM; 2011. p. 375–84.*
7. *Di Perna L, et al. An automated and unobtrusive system for coughdetection. In: Life sciences conference (LSC). IEEE; 2017. p. 190–3.*
8. *National Center for Health Statistics. National Health and nutritionexamination Survey. website, https://wwwn.cdc.gov/Nchs/Nhanes/ (lastaccess 11 January 2019).*
9. *National Center for Health Statistics. NHANES data links,https://wwwn.cdc.gov/Nchs/Nhanes/2007- − 2008/SPXRAWE.htm;11 January 2019. https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/SPXRAW F.htm. https://wwwn.cdc.gov/Nchs/Nh anes/2011-2012/SPXRAWG.htm.*
10. *Jain P, Gupta S. Multi-exposure Laser Speckle Contrast Imaging(MECI)-Based Prediction of Blood Flow Using Random Forest (RF)With K-Means (KM). Cureus. 2023 Jun 12;15(6).*