Physiological
Measurement

IPEM
Institute of Physics and Engineering in Medicine

PAPER

An intelligent evaluation model for spirometry quality using wavelet and deep time-series methods

To cite this article: Haibo Yang *et al* 2026 *Physiol. Meas.* **47** 105006

View the article online for updates and enhancements.

You may also like

- A convergent electron beam approach for small-field irradiation with reduced bremsstrahlung
Yuma Hayashi and Atsushi Myojoyama
- Hysteretic behaviour of bulk (RE)BCO Bitter-like HTS electromagnets under current cycling
Ross Taylor, Hubertus W Weijers, Mark D Ainslie et al.
- A tri-modal, offline, and explainable framework for 12-lead electrocardiogram interpretation: fusing deterministic rules, a local residual network, and an agentic vision-language model
Mohammad Mahbubur Rahman Khan Mamun

This content was downloaded from IP address 64.186.255.168 on 08/10/2026 at 15:29

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

Check for updates

PAPER

RECEIVED
23 February 2026

REVISED
10 May 2026

ACCEPTED FOR PUBLICATION
11 June 2026

PUBLISHED
8 October 2026

An intelligent evaluation model for spirometry quality using wavelet and deep time-series methods

Haibo Yang[1], Xuanfeng Li[1], Ruibin Liu[1], Zhufeng Wang[3], Tao Qian[2], Jinping Zheng[3] and Chitin Hon[1,*]

[1] Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, People's Republic of China
[2] Macao Center for Mathematical Sciences, Macau University of Science and Technology, Macau, People's Republic of China
[3] Guangzhou Institute of Respiratory Health, First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: cthon@must.edu.mo

Keywords: wavelet transform; CNN-LSTM; quality control; spirometry evaluation

Abstract

Objective: Quality control of spirometry is a critical yet challenging prerequisite for accurate pulmonary function tests, especially in primary care settings where skilled personnel are often unavailable. This study aims to develop an intelligent evaluation model to automatically identify anomaly types in spirometry reports, ensuring the reliability of diagnostic data. *Approach:* We proposed a hybrid model, which integrates an artificial neural networks with wavelet analysis and integration of convolutional neural network and long short-term memory models. The model was trained and validated using over 3000 curves collected from the First Affiliated Hospital of Guangzhou Medical University. *Main results:* The proposed model achieved an average accuracy of 96.73% on the validation set and 95.84% agreement on the test set. Comparative analysis shows that our method outperforms existing benchmarks in Precision (0.965), Recall (0.971), and F1 score (0.968). The model successfully identifies a wide range of anomalies, including cough, glottic closure, and initial leakage, with high consistency. *Significance:* This approach provides a high-precision, near real-time diagnostic tool for spirometry quality assessment. By simplifying data management and standardizing evaluation processes, the model supports primary healthcare institutions in enhancing testing quality and diagnostic reliability, particularly in resource-limited environments.

1. Introduction

Spirometry is the most frequently used test in pulmonary function tests (PFTs), serving as a critical examination for evaluating respiratory system function, assessing thoracic surgery risks, and providing early diagnosis and monitoring of respiratory diseases Celli *et al* (2000), Giri *et al* (2021). In recent decades, advancements in portable spirometers and operator training have made spirometry more accessible across all healthcare levels. High-quality measurement results are crucial for effective patient management, drug eligibility trials, and clinical studies Pérez-Padilla *et al* (2008).

The necessity of quality control (QC) is multifaceted. Firstly, QC ensures consistency across different personnel and equipment, enhancing test standardization and data reliability. Secondly, accurate results are vital for clinicians in formulating treatment plans, reducing the risk of misdiagnosis. Thirdly, many regions have specific standards for PFTs Verwey *et al* (2025); QC ensures compliance, improving the professionalism of medical institutions. Finally, rigorous QC measures boost patient confidence in primary care, facilitating better clinical communication.

Currently, operators must guide patients according to the Chinese Thoracic Association CMA (2014) and ATS/ERS guidelines Graham *et al* (2019), Miller *et al* (2023). Primary care physicians manually interpret the flow–volume ($F-V$) and volume–time ($V-T$) curves based on these standards. However,

© 2026 Institute of Physics and Engineering in Medicine. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

interpretation guidelines are constantly updated as new technologies emerge, and many practitioners lack access to the latest criteria, impeding accuracy De Vos *et al* (2023). Furthermore, the diverse range of spirometers results in varying report formats, complicating manual evaluation. Therefore, developing computer software for rapid, high-quality spirometry evaluation is urgent Lopes (2025). Studies show that integrating machine learning (ML)-based quality analysis Palamuthusingam *et al* (2022) or risk detection systems Mukhtar *et al* (2024) into clinical workflows can significantly improve patient safety and alleviate institutional burdens.

Early approaches used complex mathematical models like continuity analysis Velickovski *et al* (2017), but were limited by computational power. Modern developments focus on ML and deep learning (DL), specifically convolutional neural networks (CNNs) Li *et al* (2025). Previous studies explored ML to assist operators: Luo *et al* (2017) used AdaBoost for error classification; Solińsk *et al* (2020) utilized neural networks to identify coughs with 91% success; and Wang *et al* (2022) used computer vision to analyze four curve anomalies. Software by ArtiQ.QC achieved high accuracy in both pediatric and adult testing Topole *et al* (2021). Additionally, Das *et al* (2020) and Carla Martins *et al* (2025) utilized CNN-based methods like VGG-16 to classify respiratory curves, achieving accuracies above 90%.

Despite these advancements, a critical scientific gap remains. Most existing models treat spirometry data as static images, failing to account for the intrinsic non-stationary and time-varying nature of respiratory signals. Traditional Fourier transforms lack time localization, making them unsuitable for such biomedical signals. Wavelet transforms, however, provide excellent time resolution at high frequencies and frequency resolution at low frequencies, making them invaluable for diagnostics. While wavelets have been applied to lung CT images Bairagi *et al* (2022), sound signals Rizal *et al* (2022), and EMG signals Bairagi *et al* (2020), or used for early warnings in vital capacity Wang *et al* (2013) and error identification Bonthada *et al* (2024), their integration with hybrid DL architectures for real-time QC remains limited.

To address this, we developed a hybrid model (artificial neural network model based on wavelet analysis (WANN)-CNNLSTM). The methodological rationale for this approach is to synergize the multi-scale analysis of Wavelets with the dual-feature extraction capabilities of CNN-long short-term memory (LSTM). Specifically, the CNN component is employed to extract robust spatial/morphological features from the curve shapes, while the LSTM network is uniquely suited to capture the sequential dependencies and temporal irregularities inherent in a forced expiratory maneuver. By combining these, the model can identify subtle, time-localized anomalies that static classifiers might overlook. The contributions are twofold: (1) Our model is the first to integrate wavelet analysis with CNN-LSTM for spirometry quality grading; (2) The proposed model identifies curve issues within a very short time, achieving higher classification accuracy and providing a powerful tool for general practitioners in primary care settings.

2. Methods

2.1. Basic knowledge of spirometry reports

Spirometry is an examination that uses knowledge of respiratory physiology and modern diagnostic techniques to assess the functional status of the human respiratory system. In the process of testing, we will come across some of the following terms:

The spirometry report mainly includes two parts: the numerical indicators section (abbreviations explained in table 1) and the curve image section. Clinicians first assess the quality level of the report by combining these two parts (see table 7 of reference Graham *et al* (2019) and table 2). Then, they analyze the types of ventilation function impairments and specific disease diagnoses for reports with a quality level of B or above. Note that, for convenience of expression, we refer to 'Expiratory rise time $\geqslant$ 150 ms' as poor effort and 'Extrapolated volume $\leqslant$ 5% FVC or 0.100 l, whichever is greater' as hesitation.

The main purpose of this article is to establish a rapid assessment model for the quality of PFTs. This model aims to assist clinicians in guiding patients through better clinical examinations based on the types of anomalies flagged by the model. Additionally, it helps in screening the quality of existing test reports, significantly reducing the time clinicians spend interpreting the reports.

2.2. Spirometry and data inclusion criteria

At the subject screening level, this study included only adult subjects; to ensure data independence and eliminate biological bias, only the first examination record was retained for patients with multiple spirometry reports. At the report integrity level, all included reports were required to have complete flow–volume ($F-V$) curve images and corresponding physiological parameter tables. At the digital

Table 1. Explanation of main abbreviations.

Abbreviation	Meaning
FVC	Forced Vital Capacity
FIVC	Forced Inspiratory Vital Capacity
FEV1	Forced Expiratory Volume in the First Second
FEV1/FVC	Ratio of FEV1 to FVC
VC MAX	Maximum Vital Capacity
PEF	Peak Expiratory Flow
FEF25%	Forced Expiratory Flow at 25% Vital Capacity
Vbex	Residual Volume After Expiration
MMEF75/25	Average flow rate from 25% to 75% of FVC during expiration
MVV	Maximum Voluntary Ventilation
FET	Forced Expiratory Time

Table 2. Quality grade for pulmonary function test (age $\geqslant 6$).

Grade	Requirement	Repeatability
A	$\geqslant 3$ acceptable	Best and second-best values within 0.150 l
B	2 acceptable	Best and second-best values within 0.150 l
C	$\geqslant 2$ acceptable	Best and second-best values within 0.200 l
D	$\geqslant 2$ acceptable	Best and second-best values within 0.250 l
E	$\geqslant 2$ or 1 acceptable	Best and second-best values > 0.250 l
U	0 acceptable and $\geqslant 1$ usable	Not applicable
F	0 acceptable and 0 usable	Not applicable

extraction level, raw reports were processed via a 'Graph-to-data' program and a PDF conversion program; a report was only included if at least two sets of valid curve sampling point data were successfully extracted. Furthermore, the study implemented a stringent expert-level QC process: two trained personnel evaluated the $F-V$ curves following the 2019 ATS/ERS standardized spirometry guidelines Graham *et al* (2019) and the 2014 CTS spirometry guidelines CMA (2014). Any labeling discrepancies were adjudicated by a senior pulmonary function expert with over 10 years of clinical experience, thereby excluding curves with extremely poor waveform quality or severe recognition distortion.

The dataset for this study consists of a training set, a test set, and a validation set, all of which were delivered after anonymization. To ensure high consistency in physical characteristics and sampling frequency during the signal acquisition phase, all raw data were collected using Jaeger Master Screen spirometers (Jaeger, WÜrzburg, Germany). The training and test sets are composed of multicenter data integrated by the First Affiliated Hospital of Guangzhou Medical University, covering 128 primary medical units across 12 prefecture-level cities in Guangdong Province, ensuring broad diversity in data sources. This collection includes 946 reports, of which 429 were rated FVC Grade B or above, and 266 were rated Grade E or below. Following digital conversion, a total of 2741 curves were obtained, including 2301 normal or single-anomaly curves and 440 curves containing 2–4 types of composite anomalies. To strengthen the model's ability to recognize complex clinical signs, composite anomaly curves were redundantly counted into each anomaly category they covered, ultimately forming a training and testing data pool of 3243 samples. The validation set data originated from 274 reports collected by the First Affiliated Hospital of Guangzhou Medical University, including 130 Grade B or above reports and 7 Grade E or below cases. After digital conversion and the exclusion of 21 poorly recognized curves, 606 curves containing only a single type of anomaly were included for algorithm validation. See figure 1.

During the construction of the model training and evaluation framework, we ensured that multiple test curves from the same subject were assigned exclusively to either the training set or the test set. Regarding the specific allocation logic, normal samples and each subdivided anomaly category were randomly partitioned into training and test sets according to a 8:2 ratio. This stratified random partitioning ensures a consistent distribution of categorical features across different subsets, significantly enhancing the robustness and scientific reproducibility of the experimental results.

2.3. Data processing

Raw volumetric flow data consisted of exhalation volumes (ml) sampled at 20 ms intervals. To standardize the input for the neural network and ensure the efficacy of feature extraction, a resampling technique was employed to obtain coordinates for 512 data points per flow–volume ($F-V$) curve. The choice of

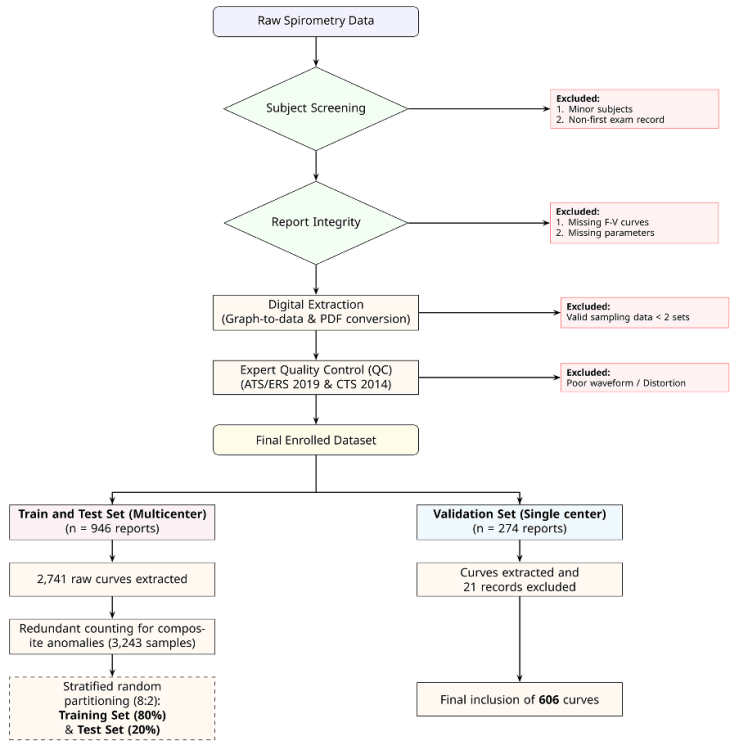


Figure 1. Data processing and allocation.

512 points was primarily based on the original data density (typically 300–500 points); thus, it preserve the signal's morphological integrity without excessive oversampling. While Haar wavelet decomposition does not strictly mandate a signal length of a power of two, adopting this length facilitates a symmetrical dyadic decomposition into low-frequency and high-frequency components across multiple levels without the need for complex boundary handling. If the division interval was not an integer, a minimal number of zero-flow points were appended to the end of the $F-V$ curve. This zero-padding has a negligible impact on the decomposition process, as these zero-valued trailing segments do not contribute significantly to the wavelet coefficients or introduce artificial frequency artifacts that would bias the subsequent analysis.

A four-level wavelet decomposition was subsequently performed. This level was selected to achieve an optimal balance between frequency resolution and computational efficiency, effectively isolating essential physiological characteristics from high-frequency noise. For data security, all data were anonymized; thus, extrinsic variables such as age or specific pulmonary impairments were not considered.

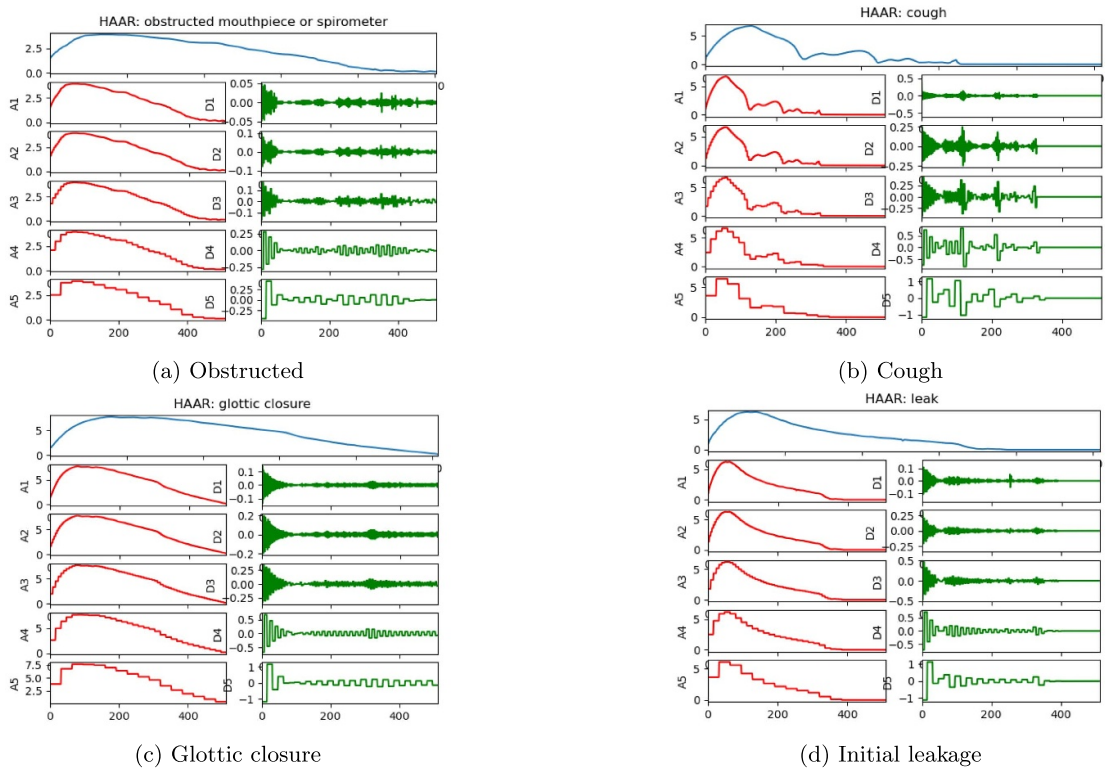
The standardized signal vectors were processed using a CNN-LSTM hybrid model. The input signals were first fed into a CNN layer, where wide convolutional kernels were utilized for adaptive feature extraction. These features underwent max-pooling to reduce dimensionality while retaining primary diagnostic information. The resulting feature maps served as sequential inputs for the LSTM layer, which automatically learned the temporal dependencies within the $F-V$ curve data. The model was optimized using the Adam algorithm for backpropagation and a Softmax activation function for the final multi-class classification.

2.4. Model evaluation

To ensure the robustness and scientific validity of the model, the following protocols were implemented:

(1) Performance metrics: Model efficacy was evaluated using Accuracy (The proportion of total correct predictions among the total number of cases), Precision (The ratio of correctly predicted positive observations to the total predicted positives), Recall (The ratio of correctly predicted positive observations to all actual positives), and the F1-score (The harmonic mean of Precision and Recall).

(2) Validation strategy: A five-fold cross-validation was employed, with the dataset partitioned into training and testing sets at an 8:2 ratio. As explicit uncertainty estimation methods were not utilized, results are reported as the mean value with sample standard deviation.



(a) Obstructed (b) Cough

(c) Glottic closure (d) Initial leakage

Figure 2. Decomposition effect examples of several errors.

(3) Statistical verification: To confirm the absence of significant bias between the training and validation subsets and demonstrate model robustness, Levene's test for homogeneity of variance and the Mann–Whitney U test were performed to ensure no significant differences existed between the results of the two sets.

2.5. Compact supported wavelet

Wavelet transform possesses both spatial and frequency localization, which enables effective extraction of information from signals. A function or signal can be analyzed through multi-scale refinement using scaling and shifting operations. In order to better utilize our model for practical applications in detection, we convert the $F-V$ curve in the report to a one-dimensional respiratory flow signal and utilize various wavelets for analysis. In our preliminary study, a comprehensive comparative analysis was conducted using various wavelet families, including Daubechies (db-1,2,4), Symlets (sym-2,4,8), Coiflets (coif-1,2,4), and Biorthogonal wavelets. These wavelets were selected as candidates due to their diverse properties in terms of vanishing moments, symmetry, and regularity, which are typically advantageous for denoising and feature extraction in complex physiological signals. See sections 3 table 4.

Thus, we selected the Daubechies-1 wavelet (i.e. Haar wavelet) to decompose the respiratory signals. It is defined as follows:

$$\Psi(t) = \begin{cases} 1, & 0 \leqslant t < \frac{1}{2}, \\ -1, & \frac{1}{2} \leqslant t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The sampled data of the curve are taken as input signal and decomposed into different levels by using Haar wavelet. This decomposition results in the acquisition of approximation coefficients and detail coefficients at each level. By analyzing these coefficients, we could identify the characteristics of different curve anomalies initially. Figure 2 provides examples of the decomposition of several common curve anomalies and their visualization results. Based on these results, we can construct feature engineering for classification models based on the approximate coefficients and detail coefficients of each layer. For each decomposition chart, the topmost blue curve represents the exhalation part of the original report FV curve. The red curve represents the approximate part of the original curve decomposed

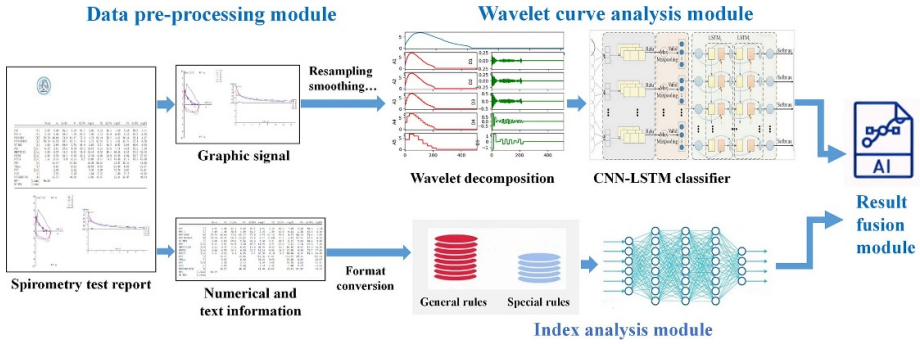


Figure 3. Algorithm flow chart.

layer by layer, while the green curve represents the detail part of the original curve decomposed layer by layer.

2.6. Model design

The model consists of four modules: data preprocessing, index analysis, wavelet analysis and result fusion. See figure 3.

1. The data preprocessing module first uses a graph-to-data program Li *et al* (2025) to extract the exhalation parts of all curves from the $F-V$ graphs and numerical information in the original PDF reports. Next, we performed quadratic spline interpolation and resampling on the curve data, and added some zero point at the end of the data to ensure that all the curve data were of the same length, facilitating wavelet analysis.
2. The metric analysis module first checks for anomalies in several numerical indicators, including Vbex values, FIVC-FVC values, and exhalation time. All numerical indicators are then input as features into the artificial neural network (ANN) model for analysis. We set up five hidden layers [64, 128, 128, 64, 32]. To prevent overfitting, we randomly drop 20% of the neurons.
3. The wavelet analysis module begins by decomposing the preprocessed, resampled data into different levels. It then identifies features in the approximation and detail parts of the decomposition results for each level and marks any potential anomalies. The decomposition results from different levels are then concatenated and fed into the CNN-LSTM model (can be referred to Buratto *et al* (2024)) as long-sequence data for analysis.
4. By combining the results of the index analysis module and the wavelet analysis module, the result fusion module outputs the types of abnormalities for each curve in the report and provides guidance accordingly.

3. Experiment

3.1. Implementation details

To ensure the reproducibility of our findings, all data processing, wavelet decomposition, model training, and statistical analyses were implemented in a Python 3.13 environment. Specifically, wavelet decomposition was performed using the PyWavelets (v1.9.0) package, while statistical significance tests, including Levene's and Mann–Whitney U tests, were conducted using SciPy (v1.3.1.1). Performance metrics (Precision, Recall, and F1-score) were calculated using scikit-learn (v1.5.1).

The WANN-CNNLSTM architecture was developed using PyTorch (v2.10.0+cu126) and torchvision (v0.25.0). We utilized a pre-trained CNN backbone, modified by replacing the final layer with a custom fully connected layer to derive weight matrices specific to lung function curve data. A batch normalization layer was subsequently integrated to mitigate potential data drift between the CNN and LSTM modules.

The image-to-data program based on OpenCV (opencv-python v4.13). This pipeline employed: (1) Template matching: To localize the curve images and parameter tables within the PDF; (2) Hough transform and HSV color segmentation: To isolate the specific spirometry curves from the background and

Table 3. The number of curves for each case in the data set.

Classes	Validation	Train and test
Qualified	89	1195(500)
Cough	78	161
Poor effort	90	210
No expiratory plateau	105	485
Initial leakage	34	184
Obstructed mouthpiece or spirometer	39	202
Hesitation	25	366
Glottic closure	48	146
FIVC-FVC $<$ 0.100 l or 5% of FVC	98	294
Total	606	3243(2548)

Table 4. Different wavelets with CNN model on train and test set.

Wavelet type	Accuracy	Wavelet type	Accuracy
Daubechies-1	**0.907**	Coiflet-1	0.898
Daubechies-2	0.899	Coiflet-2	0.902
Daubechies-4	0.901	Coiflet-4	0.903
Symlets-2	0.826	Biorthogonal-2.2	0.904
Symlets-4	0.891	Biorthogonal-3.3	0.901
Symlets-8	0.886	Biorthogonal-6.8	0.896

grid lines; (3) Quadratic spline interpolation: To reconstruct and densify the extracted curve points into a standardized format compatible with the model input.

The dataset utilized in this study originated from two distinct sources, necessitating different processing workflows: Training and Test Sets comprised raw digital data exported directly from spirometry devices, requiring no additional preprocessing, thus preserving the integrity of the original physiological signals. To evaluate the model's compatibility across different clinical settings, validation data were extracted from existing hospital PDF diagnostic reports.

3.2. Results

Table 3 presents the characteristics of the training dataset.

To select the most suitable wavelet decomposition function, we tested 12 different wavelets on the dataset. The results showed that the Daubechies-1 wavelet achieved the highest classification accuracy. See table 4. Therefore, the Daubechies-1 wavelet was consistently adopted as the decomposition method in subsequent training and validation.

The rationale is centered on the specific nature of respiratory flow anomalies: clinical artifacts such as coughing or mouthpiece obstruction do not manifest as gradual morphological shifts, but instead appear as sudden and sharp impulses in the flow signal, making the Haar wavelet's step-function nature uniquely suited to capturing these abrupt transitions. Furthermore, due to its minimal compact support, the Haar wavelet provides the highest degree of temporal localization, allowing the model to precisely pinpoint the exact moment an anomaly occurs. Additionally, as the features of FV signals do not require overly complex or high-order mathematical representations, the Haar wavelet offers a computationally efficient solution that increases the algorithm's processing speed without sacrificing classification accuracy.

Based on the available data, the classification accuracy for each category is presented in table 5. To evaluate model performance more rigorously, this study introduces confidence intervals (CIs) and sample standard deviations to quantify the uncertainty of the point estimates. The proposed method achieved an average accuracy of 96.73% in the validation set and 95.84% in the test set, and the results demonstrate high statistical consistency between the 95% CIs of the test set and the validation set. While the model maintains high overall accuracy, categories such as Initial leakage, Hesitation, and Glottic closure exhibited relatively larger fluctuations in accuracy between the test and validation sets. These are primarily attributable to the statistical stochasticity inherent in small sample sizes rather than issues regarding the stability of the model architecture. Consequently, the observed variances are more likely

Table 5. Accuracy results with sample standard deviation and significance test.

Classes	Test Acc. (%)	Validation Acc. (%)
Qualified	97.47(1.14)	97.53(0.84)
Cough	94.29(1.14)	95.13(0.89)
Poor effort	97.62(0.89)	98.22(0.55)
No expiratory plateau	97.69(0.84)	98.48(0.55)
Initial leakage	87.40(1.92)	91.76(1.64)
Obstructed mouthpiece	88.12(1.87)	90.26(2.12)
Hesitation	97.65(0.71)	95.20(0.84)
Glottic closure	93.56(1.34)	95.83(1.22)
FIVC-FVC $<$ 0.100 l or 5%	99.05(0.55)	99.18(0.45)
Average Accuracy	**95.84**	**96.73**
95% CI of *t*-distribution	[95.55, 96.13]	[96.47, 97.00]
Significance test	Parameter	Significance
Levene's homogeneity	P-value $= 0.778$	No
Mann–Whitney U test	U Statistic $= 42.000$	No

Table 6. Comparison with other results.

Index	WANN-CNNLSTM	Reference 14	Reference 15	Reference 31
Precision	0.965	0.86	0.94	0.91
Recall	0.971	0.91	0.89	0.83
F1 score	0.968	0.88	0.91	0.86

a manifestation of stochastic noise inherent to small cohorts rather than a lack of model stability or generalizability. Furthermore, the consistency between the test and validation results was rigorously evaluated through non-parametric and variance-based statistical tests. The Levene's homogeneity of variance test yielded a p-value of 0.778, indicating that the variance across the different datasets is homogeneous. Additionally, the Mann–Whitney U test ($U = 42.000$) showed no significant difference ($p > 0.05$) between the test and validation distributions. These statistical outcomes provide robust evidence that the model's predictive performance remains stable across different data splits, and the minor numerical discrepancies in specific categories do not undermine the overall reliability of the model.

Then, we compare the outcomes of our model: reference Walag *et al* (2023) utilized the recalibrated CNN model, reference Topole *et al* (2021) employed the standard ArtiQ.QC software, and reference Smith *et al* (2021) used various common classifiers, such as logistic regression, feed-forward artificial neural networks, and support vector machines, to detect one type of anomaly: cough. From table 6, our method achieved the highest precision rate, recall rate, and F1 score, despite having a slightly lower recall rate compared to the result in reference López-Canay *et al* (2026). Our model achieved an average accuracy of 96.73% in abnormality classification on the test set, surpassing the results reported by Wang *et al* Wang *et al* (2022).

4. Discussion

The primary contribution of this study is the development of the WANN-CNNLSTM model, which achieves automated anomaly detection in spirometry waveforms by combining wavelet analysis with deep time-series networks. By utilizing wavelet coefficients to capture multi-scale morphological features of respiratory curves, the model not only significantly reduces detection time but also demonstrates superior classification performance on the existing dataset. Since a single respiratory curve in clinical practice may simultaneously contain multiple operational errors (e.g. 'cough' accompanied by 'early termination'), we adopted a redundant counting approach to assign a sample to all its corresponding anomaly categories. This approach realistically simulates complex clinical conditions and enables the model to identify overlapping error features, ensuring that the reported performance metrics reflect the capture precision across independent anomaly dimensions, thereby enhancing the practical utility of the system.

Despite the encouraging results, this study has several limitations. First, the dataset integrated by the First Affiliated Hospital of Guangzhou Medical University, covering 128 primary medical units across 12

cities in Guangdong Province. While this ensures broad diversity in data sources, geographical diversity across different provinces and national borders remains to be strengthened to verify the model's universality across various ethnicities and living environments. To address the issue of equipment consistency, advanced data smoothing and standardized resampling techniques could be introduced in the future to conduct external validation across different brands. Second, regarding the scope of the study subjects, all samples in this study were restricted to adults under the age of 65. Furthermore, due to data privacy constraints, variables such as patients' basic physiological status, past medical history, and living environment were not included, which to some extent limits the model's in-depth analysis of complex clinical cases.

To further enhance the system's auxiliary decision-making efficacy in complex clinical environments, subsequent research should consider incorporating auxiliary input variables such as height, weight, gender, occupational history, smoking history, and existing clinical diagnoses (e.g. chronic obstructive pulmonary disease (COPD) or asthma). This background information is crucial for distinguishing 'physiological respiratory limitation' from 'technical operational errors'. For instance, by integrating occupational history or pulmonary disease status, the model can more intelligently adjust discrimination thresholds, thereby reducing false positives caused by pathological characteristics. Additionally, to address the challenge of periodically updated clinical guidelines, the modular architecture of this model features a flexible multi-layer adaptation mechanism: for changes in quantitative physical rules, only the parameters of the indicator analysis module need to be modified; for significant qualitative changes, the modular nature of the WANN-CNNLSTM allows for targeted fine-tuning of the connection and output layers. This enables model iteration using a small sample of new standards without retraining the entire network, greatly improving maintenance efficiency under evolving guidelines.

To enhance clinical trust, this model utilizes wavelet analysis to localize signal variations in the time–frequency domain, aligning its feature extraction with clinical morphological logic. For instance, the model effectively identifies continuous fluctuations in glottic closure, intense amplitude spikes in coughs, and persistent vibrations from obstructed mouthpieces. Furthermore, the model's internal weighting mirrors clinical expertise by prioritizing key diagnostic parameters such as FEV1/FVC, PEF, and FEV1. This alignment between DL logic and medical knowledge provides the physical interpretability necessary to guide patient maneuvers in real-time, significantly reducing misdiagnosis and unnecessary repeat testing. Future work will implement visualization techniques to generate heatmaps on flow–volume curves, translating 'black-box' logic into familiar morphological descriptions. Ultimately, integrating this model into various spirometer systems will shorten preprocessing time and support primary healthcare settings with limited expertise, effectively optimizing the diagnostic workflow.

5. Conclusions

In this work, we developed a model based on wavelet analysis, ML, and a deep time-series framework, in accordance with the latest guidelines for spirometry test standardization. The proposed WANN-CNNLSTM model supports primary healthcare institutions in building a standardized spirometry information system by enabling high-precision, near real-time quality assessment of spirometry tests. This capability helps general practitioners ensure high-quality testing and enhances diagnostic reliability, particularly in resource-limited settings. Labeled anomalies simplify data management and analysis, thereby improving overall service quality. Integrating the model into clinical workflows may further facilitate the construction of standardized information systems and support continuous quality improvement in spirometry.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Funding

This work was supported in part by Macau Science and Technology Development Fund (FDCT) Grants 0020/2023/RIB1, 0002/2024/RDP, 0128/2022/A; Specialized Grant for Medical High Ground Development, Guangzhou Municipal Health Commission, Grant 2022-3398.

Author contributions

Haibo Yang 0000-0001-7756-0119
Formal analysis (equal), Methodology (equal), Writing – original draft (equal)

Xuanfeng Li 0009-0002-6022-1000
Visualization (equal)

Ruibin Liu 0000-0002-8252-5585
Software (equal)

Zhufeng Wang 0000-0002-8771-5118
Data curation (equal), Resources (equal)

Tao Qian 0000-0002-8780-9958
Methodology (equal), Project administration (equal)

Jinping Zheng
Resources (equal)

Chitin Hon 0000-0002-5658-3503
Writing – review & editing (equal)

References

Bairagi V and Kanwade A B 2020 Classification of chronic obstructive pulmonary disease (COPD) using electromyography *Sadhana* **45** 132

Bodasingi N, Balaji N and Jammu B R 2022 Automatic diagnosis of pneumonia using backward elimination method based SVM and its hardware implementation *Int. J. Imaging Syst. Technol.* **32** 1000–14

Bonthada S, Perumal S P, Naik P P, Padukudru M A and Rajan J 2024 An automated deep learning pipeline for detecting user errors in spirometry test *Biomed. Signal Process. Control* **90** 105845

Buratto W G *et al* 2024 Wavelet CNN-LSTM time series forecasting of electricity power generation considering biomass thermal systems *IET Gener. Transm. Distrib.* **18** 3437–51

Celli B R 2000 The importance of spirometry in COPD and asthma: effect on approach to management *Chest* **117** 15S–19S

Chinese Medical Association Respiratory Branch Pulmonary Function Professional Group 2014 Guidelines for pulmonary function testing (part 2) - spirometry *Chin. J. Tuberc. Respir. Dis.* **37** 481–6

Das N *et al* 2020 Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria *Eur. Respir. J.* **56** 2000603

De Vos M *et al* 2023 Insights into spirometry usage and the potential for AI-based decision support software in Belgian primary care *Gen. Pract. Prim. Care* **62** OA764

Giri P C *et al* 2021 Application of machine learning in pulmonary function assessment: where are we now and where are we going? *Front. Physiol.* **12** 678540

Graham B L *et al* 2019 Standardization of spirometry 2019 update an official American Thoracic Society and European Respiratory Society technical statement *Am. J. Respir. Crit. Care Med.* **200** e70–e88

Li X *et al* 2025 Enhancing the interpretation of spirometry: joint utilization of n-order adaptive Fourier decomposition and deep learning techniques *IEEE Trans. Instrum. Meas.* **74** 1–14

Lopes A 2025 Pulmonary function tests: an integrated approach to interpreting results in the search for treatable traits *Expert Rev. Respir. Med.* **19** 121–43

López-Canay J *et al* 2026 Development of an intelligent clinical decision support system for spirometry quality control *Diagnostics* **16** 213

Luo A Z, Whitmire E, Stout J W, Martenson D and Patel S 2017 Automatic characterization of user errors in spirometry *2017 39th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* pp 4239–42

Martins C, Barros H and Moreira A 2025 Transfer learning in spirometry: CNN models for automated flow-volume curve quality control in paediatric populations *Comput. Biol. Med.* **184** 109341

Miller M R, Stanojevic S and Graham B L 2023 Do American Thoracic Society and European Respiratory Society Journals have a role in promoting official American Thoracic Society and European Respiratory Society standards? *Ann. Am. Thorac. Soc.* **20** 1694–6

Mukhtar S A *et al* 2024 Predictive analytics for early detection of hospital-acquired complications: an artificial intelligence approach *Health Inf. Manag. J.* **54** 109–20

Palamuthusingam D *et al* 2022 Evaluating data quality in the australian and New Zealand dialysis and transplant registry using administrative hospital admission datasets and data-linkage *Health Inf. Manag. J.* **52** 212–20

Pérez-Padilla R, Vázquez-García J C, Márquez M N and Menezes A M B 2008 Spirometry quality-control strategies in a multinational study of the prevalence of chronic obstructive pulmonary disease *Respir. Care* **53** 1019–26
Rizal A and Puspitasari A 2022 Lung sound classification using wavelet transform and entropy to detect lung abnormality *Serb. J. Electr. Eng.* **19** 79–98
Smith E *et al* 2021 ERS International Congress 2020 virtual: highlights from the allied respiratory professionals assembly *ERJ Open Res.* **7** 00808–2020
Solińsk M, Ł epek M and Ukasz Koł towski Ł 2020 Automatic cough detection based on airflow signals for portable spirometry system *Inform. Med. Unlocked* **18** 100313
Topole E, Biondaro S and Montagna I 2021 Artificial intelligence assists in quality assessment of spirometry in clinical trials *Am. J. Respir. Crit. Care Med.* **203** A4606
Velickovski F *et al* 2017 Automated spirometry quality assurance: supervised learning from multiple experts *IEEE J. Biomed. Health Inform.* **22** 276–84
Verwey C *et al* 2025 Pulmonary function testing in healthy infants in rural Bangladesh: feasibility study *Pediatr. Pulmonol.* **60** e27461
Walag D *et al* 2023 Deep learning algorithm for visual quality assessment of the spirograms *Physiol. Meas.* **44** 085004
Wang W, Finkelstein S M and Hertz M I 2013 Automatic event detection in lung transplant recipients based on home monitoring of spirometry and symptoms *Telemed. e-Health* **19** 658–63
Wang Y *et al* 2022 Deep learning for spirometry quality assurance with spirometric indices and curves *Respir. res.* **23** 98