Respiratory Artefact Removal in Forced Oscillation Measurements: A Machine Learning Approach

Thuy T. Pham, *Student, IEEE,* Cindy Thamrin, *Member, IEEE,* Paul D. Robinson, Alistair L. McEwan, *Member, IEEE,* and Philip H.W. Leong, *Senior Member, IEEE*

***Abstract*—Respiratory artefact removal for the forced oscillation technique can be treated as an anomaly detection problem. Manual removal is currently considered the gold standard but this approach is laborious and subjective. Most existing automated techniques used simple statistics and/or rejected anomalous data points. Unfortunately, simple statistics are insensitive to numerous artefacts, leading to low reproducibility of results. Furthermore, rejecting anomalous data points causes an imbalance between the inspiratory and expiratory contributions. From a machine learning perspective, such methods are unsupervised and can be considered simple feature extraction. We hypothesize that supervised techniques can be used to find improved features that are more discriminative and more highly correlated with the desired output. Features thus found are then used for anomaly detection by applying quartile thresholding which rejects complete breaths if one of its features is out of range. The thresholds are determined by both saliency and performance metrics rather than qualitative assumptions as in previous works. Feature ranking indicates that our new *landmark* features are among the highest scoring candidates regardless of age across saliency criteria. F1-scores, receiver operating characteristic, and variability of the mean resistance metrics show that the proposed scheme outperforms previous simple feature extraction approaches. Our subject-independent detector, 1IQR-SU, demonstrated approval rates of** 80.6% **for adults and** 98% **for children, higher than existing methods.**

***Index Terms*—Respiratory artefacts, lung, machine learning.**

I. Introduction

THE forced oscillation technique (FOT) [1] is a lung function test that can provide useful information from short duration recordings, and only requires passive cooperation from the subject [2]. FOT assesses breathing mechanics by superimposing small external pressure signals to the spontaneous breathing of the subject. A total respiratory mechanical impedance (*Zrs*) which includes airway resistance together with elastic and inertive behavior of the lungs and the chest wall is then measured at one oscillation frequency (mono-frequency oscillations) or several (multi-frequency). *Zrs* is described as a complex number with *real* and *imaginary* components, called the resistance (*Rrs*) and reactance (*Xrs*) respectively. A primary reason hindering its widespread adoption lies in difficulties associated with removing artefacts. This results in lower reproducibility than the most common pulmonary function test, spirometry. Manual removal of artefacts, called the *human-based* method, is currently considered the gold standard. This, however, is typically done in an ad-hoc manner which is laborious, and subjective.

Several recent studies have explored the role of FOT in telemedicine [3–5]. The ability to automatically and objectively detect artefacts, enabling quality control on a patient's unsupervised self-measurements, would greatly facilitate the use of FOT in such applications.

To detect artefacts, several automated refinements include detecting low (e.g., transducer noise) and high frequency artefacts (e.g., light coughing, mouth piece leak, swallowing, glottic closure and tongue occlusion). According to the quality control guidelines [6], low frequency noise removal rejects low magnitude-squared coherence values of pressure and flow [7]. Several transient artefacts are removed by identifying deviations from the norm.

To exclude respiratory artefacts, two different strategies are point rejection [8–10] and complete-breath rejection [7, 11]. For example, the *3SD* point-based method [8] introduced a statistical filter that rejected any impedance points greater than three standard deviations (SD) from the mean *Rrs* or *Xrs* value. Alternatively, the *B-3SD complete-breath* approach rejects entire breaths as defined by the starting and ending points of breath cycles in which at least one data point is out of the *3SD* range [7]. Complete-breath rejection has been reported to be more accurate than the point approach as it can avoid an imbalance between the inspiratory and expiratory contributions to each breath [7]. Nonetheless, these automated techniques still miss numerous artefacts.

From a machine learning perspective, each breath is represented by a vector of features. Features are then classified by a model (*detector*) constructed from domain-knowledge and/or human annotations (*labels*). The aforementioned methods are unsupervised techniques in which simple feature extraction is used and threshold values are chosen as a number of standard deviations away from the mean of a single measurement. We hypothesize that more sophisticated two-dimensional (2D) features may alleviate the above limitations of current automated methods. The relevancy of features can be confirmed quantitatively by supervised techniques (*feature selection*), e.g., correlation of feature candidates with artefacts can be measured by mutual information (Shannon's information theory [12]).

The clusterability of a candidate [13] indicates the efficiency of using threshold values to detect artefacts. Two typical ways to assess clusterability are the variance ratio of clusters [13]

T. Pham*, A. McEwan and P. Leong are with the Dept. of Electrical and Information Engineering, The University of Sydney, NSW, Australia. *:correspondence thuy.pham@sydney.edu.au. C. Thamrin is with Woolcock Institute of Medical Research and Paul D. Robinson is with The Children's Hospital at Westmead, NSW, Australia. Funding resources: Endeavour/Prime Minister's Australia Scholarship and the Faculty Research Cluster Program at The University of Sydney.

and the separability as calculated by Euclidean distances from an instance to a *near-hit* and *near-miss* [14].

Given an exploratory feature pool, by selecting the k highest ranking candidates (e.g., $k = 10$ often used in literature of feature selection), we can construct a more accurate anomaly detector as non-salient features which cause overfitting are discarded. Several challenging factors should be noted. One is the time-dependency of lung function (e.g., lung elasticity [15]). The others are clinical aspects of FOT; e.g., *Rrs* and Xrs are dependent on body size and possibly racial/ethnic differences [2]. To avoid dependency, feature ranking scores should be accumulated in a recording-wise scheme. We also noticed that *Rrs* within a recording can be non-Gaussian, with a strong kurtosis. Hence, when applying threshold values, we do not assume a particular distribution. Instead we use quartile percentages, called *quartile thresholding*. In contrast to earlier works, the deviation threshold is also not assumed, rather it is determined from the receiver operating characteristic (ROC) and other performance metrics obtained from training data.

The main contributions of this work are:

- This is the first reported application of supervised machine learning to respiratory artifact detection in FOT.
- We identify new features that are more relevant and more discriminative than those previously proposed.
- We propose an anomaly detector which, to the best of our knowledge, achieves the best reported performance for FOT data regardless of participants' age.

II. Methods

A. Data Collection

Subjects and Protocol: We collated data from two different age groups (*Paediatrics* and *Adults*, Table I). The paediatric dataset comprised a random sample of 9 subjects (total 69 FOT runs) for development and 5 subjects (total 31 runs) for out-of-sample tests. These were taken from a much larger ongoing epidemiological study, which has been described in detail elsewhere (Ultrafine Particles from Traffic Emissions and Children's Health, UPTECH) [16][17]. The epidemiological study collected FOT data, as part of its respiratory function assessment, in eight- to eleven-year-old children recruited from 25 different public primary schools in Queensland, Australia. FOT was performed at least 30 minutes after supervised medication administration and with at least 10 minutes rest prior to recording. Zrs was measured at 6 Hz, using an in-house built FOT device (transducer Sursense DCAL-4, Honeywell Sensing & Control; more details as in [18]) and modification to comply with recent recommendations [19]. Children were encouraged to breathe in a regular manner, avoid swallowing and maintain a tight mouthpiece seal. Children had multiple recordings in a single session as part of the study protocol.

For the adult group, 9 healthy participants and 10 asthmatic patients were recruited from staff and patients of the Royal North Shore Hospital, St Leonards, Australia and the Woolcock Institute of Medical Research volunteer database (Glebe, Australia) [20]. Healthy participants were non-smokers with no known respiratory disease. Asthmatic adults had a physician diagnosis of asthma (clinically stable as defined by GINA guidelines [21]) and had no reported diagnoses of any other cardiac or pulmonary disease. The asthmatic and control subjects had three recordings over seven days within a 10-day period at the Respiratory Investigation Unit at Royal North Shore Hospital [20]. To ensure clinical stability, asthmatic patients continued to take their usual medications and were reviewed by a specialist physician at each visit for any changes in their usual symptoms. All recordings were performed at the same schedule to avoid any diurnal variation effects. Zrs was measured at an oscillation frequency of 6 Hz from a FOT device of similar general design and specifications as the children dataset [22]. Three separate consecutive recordings were collected with subjects breathing tidally for 60 second at each session (day). The participants put their nose clip on and placed their hands on cheeks to reduce the upper airway shunt. Recordings were assessed from visual inspection by a technician if tidal volume and breathing frequency appeared stable. Artefact labels were made by the operator using recommendations in [7] (more details in [20]). All subjects gave written, informed consent and the study was approved by The Human Research Ethics Committee of Northern Sydney Central Coast Health (protocol no. 0903-050M). For children, the study was approved by the Queensland University of Technology Human Research Ethic Committee.

TABLE I
Data Sets used in this work.

Dataset	Subjects	Recordings	Breaths	Description
$Ds1$	9	69	1110	Development, children
$Ds2$	9	261	3067	Development, adults
$Ds3$	5	31	580	Test, children
$Ds4$	10	285	3947	Test, adults

Data Pre-processing: Flow and pressure signals were digitally sampled at 396 Hz and band-pass filtered with a bandwidth of ±2 Hz centered around 6 Hz. *Rrs* and *Xrs* were calculated at $0.1s$ intervals using a standard frequency-domain method. To ensure a balance between the inspiratory and expiratory contributions to each breath [7], incomplete or partial breaths at the beginning/end of the recording were removed. Since the "not accepted" annotations included non-eligible physiological breaths which are commonly known to be rejected by the standard FOT quality guidelines [6], we discard these artefacts in pre-processing steps and report separately in later comparisons. First, we remove breaths that contain negative *Rrs* which are non-physiological. Then, we discard breaths that have magnitude-squared coherence values of pressure and flow less than 0.9 [7]. Unusually high amplitude observations were successfully caught by the *B-3SD* approach in [7] and discarded. Finally, we apply *3IQR* (i.e., $3\ IQR$ away from the median) to *Rrs*, *Xrs*, *Volume*, *Pressure*, and *Flow*.

B. Performance Metrics

Since *Rrs* is one of the main outcomes of FOT in clinical and research usage, we consider *variability* (i.e., the standard deviation divided by mean) of the average *Rrs* for each patient to be critical metric. To quantify this, we aim for an equivalent average *Rrs*, and lower or equal SD compared with the human-based approach. However, if we only consider variability, we may not account for the number of valid breaths that remain, e.g, we may discard most valid breaths together with invalid ones to achieve low variability. Therefore, when comparing techniques, we should strive for an equivalent preservation level *and* lower variability.

The *preservation* after removal can be summarised by standard accuracy metrics (e.g., sensitivity, specificity, and F1-score [23]) and our new metrics: *throughput* and *approval rate*. In confusion matrices for accuracy calculation, we consider *groundtruth* to be the human labels (or "manual"), and *positives* to be artefacts.

True Positives (TP) are breaths which were marked as "artefacts" by both a test algorithm and the annotation. False Positives (FP) are breaths we labeled as artefacts but did not agree with the ground truth. Breaths that we failed to label as artefacts but were annotated as such, are defined as False Negatives (FN). When the test method and the human agree a breath was not an artefact, it is counted as a True Negative (TN). Sensitivity and specificity are $\frac{TP}{TP+FN}$ and $\frac{TN}{TN+FP}$, respectively. F1-score, which is the harmonic mean of precision and sensitivity, has best value at 1 and worst at 0 [23], is calculated as $\frac{2TP}{(2TP+FP+FN)}$. *Throughput* is the ratio of breath numbers in the output to input, $\frac{TN+FN}{total\ input}$. *Approval rate* of the filtered data (i.e., the breaths remaining after removal) is the ratio of breaths that are "accepted" by the human to the total output breaths, $\frac{TN}{TN+FN}$.

C. Feature Selection

We construct a pool of exploratory features. After ranking candidates, we select the most salient subset of features for further detection steps. The pool consists of 111 candidates (Table II), of which 11 have been previously reported. Others are our new features: *landmark* information, *resampling* values, and other domains.

1) Feature Extraction: *Landmark* features are extracted from *landmark* points of a breath cycle. Intuitively, we want to capture the boundary information of normal cycles to detect anomalies. For example, in Fig. 1.a, A, B, CL, CR, D, E, F, Z are seven *landmark* points of which distances contain information for artefact detection (called *7-point extraction*). Specifically, points B and Z are at the zero flow value and the higher and lower *Rrs* values, respectively. Points A and D are at the maximum and minimum of Flow, respectively. Points CR, CL and E are at the maximum (right: positive Flow area and left: negative Flow area) and minimum of *Rrs*, respectively.

Resampling features are extracted from one dimensional input with a fixed number of points for a cycle to alleviate varied durations of breaths. We noticed that the minimum length of all breaths in the training data sets is larger than 30 points. For generalization, we consider 20 points per cycle and assume this is sufficient to describe the fundamental shape information for a breath curve. Thus, we re-sample *Rrs, Flow, Xrs, Volume* at a fixed rate of 20 points/cycle (called *20-point*).

Other new candidates in the pool are from different domains. For example, we obtain the changes of polar coordinates over time for each breath (using the mapping from Cartesian coordinates to their polar ones). We also explore the wavelet decomposition analysis, DWT, (three level decomposition) with the Daubechies method [24] of the above 20-point resampling vectors. For the spectral coherence computation, we use 0.1667-second windows (as our frequency of interest is $6Hz$) and ensemble-average every three windows with 50% overlap. This and the impedance are then re-sampled at 10 Hz to effectively get the same number of coherence points as the number of impedance values. The existing are minima and maxima of *Rrs*, and DWT of *pressure* (e.g., in [7–11]).

TABLE II
LIST OF FEATURES EXAMINED IN THIS WORK.

Measurement	Domain	Function Description	New?	ID
Pressure	Frequency	Maximum value of first level DWT	No	1
Xrs	Time	Maximum, minimum, range	No	2-4
Xrs	Time	20-point re-sampled	Yes	8-27
Volume	Time	Maximum, minimum, range	No	5-7
Volume	Time	20-point re-sampled	Yes	28-47
Rrs	Time	Peaks, minimum, *Cr, Cl, E*	No	54, 56, 61
Rrs	Time	20-point re-sampled	Yes	65-84
Flow	Time	Minimum	No	62
Flow	Time 2D	Landmark *Cr,Cl, E*	Yes	55, 57, 63-64
Flow	Time	20-point re-sampled	Yes	85-104
Rrs, Flow	2D	Landmark *Z, B, A*	Yes	48-53
Rrs, Flow	2D	Landmark *Z* and *D*	Yes	58-60
Rrs, Flow	2D	Mean and std of polar coordinators from *20-point Rrs*, Flow	Yes	105-108
Rrs, Flow	2D Frequency	Maximum of full DWT from *20-point Rrs*, Flow	Yes	109, 110
Rrs, Flow	Frequency	Maximum spectral coherence *Rrs* and Flow	Yes	111

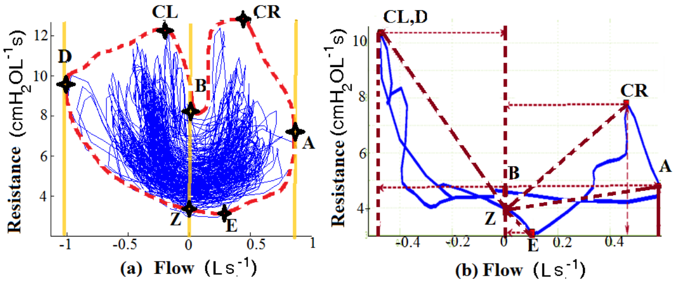


Fig. 1. (a): All clinical accepted breaths (*Rrs* against Flow) of one child (several recordings) and *7 points* proposed to determine *boundary landmarks* (dotted) for curves. (b): Example features extracted by landmarks for one breath from a child (dotted: Euclidean distances between points).

2) Saliency Criteria: We examine the relevance and clusterability of feature candidates using mutual information scores

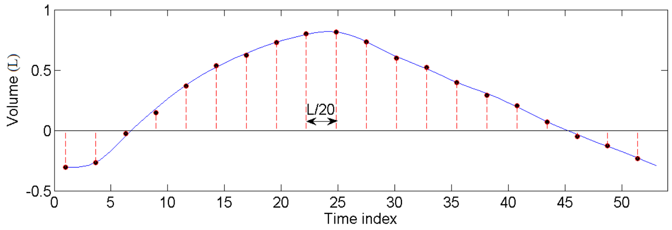


Fig. 2. Example of unified *20-point* resampling for a breath (Volume, time).

(between a feature candidate and the label) and the inter-class distances. Specifically, let X be a discrete random variable with alphabet $\mathbb{X}$ and probability mass function $p(x) = P\{X = x\}$, $x \in \mathbb{X}$. The *entropy* $H_b(X)$ of X measures its uncertainty. $H_b(X)$ is defined by $H_b(X) \overset{\text{def}}{=} -\sum_{x\in\mathbb{X}} p(x)\log_b p(x)$, where b is the base of the logarithm. In this work, we take $b = 2$, and hence entropy will be measured in bits. Let C be a target variable ($c \in \mathbb{C}$, class label set). The mutual information between X and C, $I(X;C)$ measures the relevance of X to C [12].

$$I(X;C) \overset{\text{def}}{=} H(X) - H(X|C) = \sum_{x\in\mathbb{X}}\sum_{c\in\mathbb{C}} p(xc)\log\frac{p(xc)}{p(x)p(c)} \tag{1}$$

As a measure of redundancy, symmetrical uncertainty between X and C, $\mathrm{SU}(X,C)$ [25], is a weighted average of two uncertainty coefficients: $C_{XY} \overset{\text{def}}{=} \frac{I(X;Y)}{H(Y)}$ and $C_{YX} \overset{\text{def}}{=} \frac{I(X;Y)}{H(X)}$. $\mathrm{SU}(X,C)$ is often referred to *correlation*. Thus, the relevance of a feature is computed by $\mathrm{SU}(X,C) = 2\frac{I(X;C)}{H(X)+H(C)}$ (namely the *SU* score)

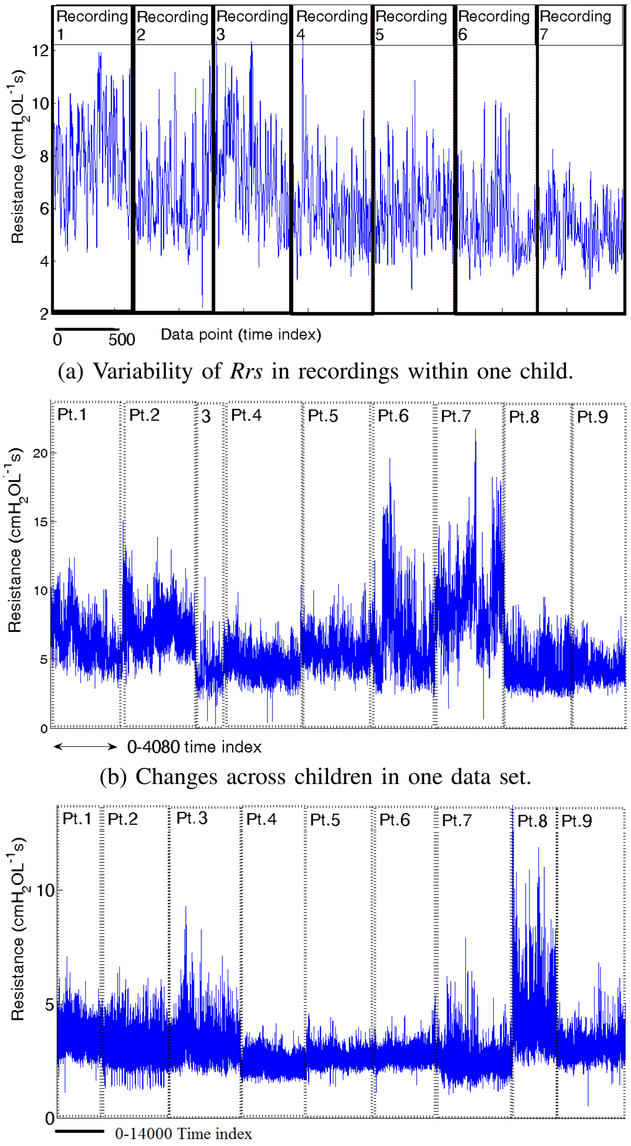
For clusterability, we use the *RELIEF* algorithm to calculate Euclidean distances between features and a *near-hit* or *near-miss* instance [14],[26]. The variance ratio of a feature X is the ratio of the between-cluster variance ($B_C(X)$) to the within-cluster variance ($W_C(X)$), $V(X) \overset{\text{def}}{=} \frac{B_C(X)}{W_C(X)}$. A higher $V(X)$ implies that it is easier to cluster X [13], therefore the feature is more desirable.

3) Challenging Factors: Fig. 3 illustrates the time dependence of samples within and between recordings (and between different age groups). These variations and artefacts are contained partly in the scaling information of the samples. This may introduce bias into ranking scores of features which are extracted from amplitude values across recordings. To reduce the bias, we accumulate scores for each feature candidate in a recording-wise manner.

4) Other Criteria: Apart from saliency ranking, we select a relevant and efficient feature set based on performance metrics. We investigated ROC, F1-score, *throughput*, and *approval rate*. For a clinical interest, we have quantified the reduction in artefactual activity and selected features by the variability of the average *Rrs*.

D. Artefact Detector

In thresholding filters, a breath is marked as an artifact and discarded if one of its features exceeds a given upper bound or is less than a lower one. Since the normality hypothesis of *Rrs* in a recording is rejected with a significance level of 0.05 (the p-values were very close to zero; 0 to 1.27×10^{-17}) by the Lilliefors test [27] and the KS test [28], we do not assume a specific data distribution. Instead we use the ROC plots to determine the threshold parameters. We refer to this detector as a quartile thresholding filter.



(a) Variability of *Rrs* in recordings within one child.

(b) Changes across children in one data set.

(c) Changes across adults in one data set.

Fig. 3. Examples of challenges in learning contaminated *Rrs* (after preprocessing) within a participant (a) and between participants (b), (c).

Let Q_1, Q_3, and IQR denote the 25^{th}, 75^{th} percentiles and the interquartile range of a variable respectively. Let n_{IQR} be a number of IQR intervals away from the Q_1 and Q_3. The lower bound θ_L is defined by n_{IQR} interquartile intervals less than Q_1. The upper bound θ_H is n_{IQR} intervals greater than Q_3, i.e.:

$$\theta_L = Q_1 - n_{IQR} \times \mathrm{IQR} \tag{2}$$

$$\theta_H = Q_3 + n_{IQR} \times \text{IQR} \quad (3)$$

To simplify parameter settings, we apply the same n_{IQR} to all features and categorize subjects into two age groups (i.e., Paediatrics and Adults). We split each age group into two data sets: one for development and the other for test. For example, the set of (development, test) for children is ($Ds1,\ Ds3$) and for adults is ($Ds2,\ Ds4$), details of data sets as in Table I.

We compare our detector with *B-3SD* [7] and the wavelet in a complete-breath rejection fashion [10] (namely *B-wavelet*). These two methods were recently proposed as the best automated ones in the literature. Note that, the work in [10] used a point rejection approach and asked the participant to intentionally introduce artefacts while *B-wavelet* uses complete-breath rejection, and was tested with our real-life artefacts. We performed *B-wavelet* with three levels of DWT coefficients (cd1,cd2,cd3) and the *db5* method for *pressure*, and then used the three recommended thresholds in [10] (i.e., $cd1^2 = 0.004$; $cd2^2 = 0.023$; $cd3^2 = 0.07$).

III. Results

A. Saliency Ranking

Ranking scores (i.e., RELIEF, SU, and Variance-Ratio) for each feature candidate are presented in Fig. 4. Features with circle markers have been currently used in the literature (Table II) while the others are our new candidates. Group "I" illustrates results for children while "II" depicts adults. We sorted scores in the entire pool from high to low by each saliency criterion. Ranking order for the pool in Fig. 4.a is from 1 to 111 (high to low); the higher saliency score indicates the higher ranking order. Fig. 4.b shows the first ten identifications of features (IDs, detailed in Table II) that have the highest scores.

As can be seen in Fig. 4.a, only three previous features (the minimum and peaks of *Rrs*) are in the top ten highest ranking candidates from the children group. For adult cases, these *Rrs* features have moderate variance ratio and very low RELIEF scores. Our novel *landmark* features dominate not only in both children and adult groups but also across all three saliency criteria. Specifically, in Fig. 4.b, they are *landmark* features ID 48, 49, 50, 54 56, 64, description for these features as in Table II. In next steps, given a detector of interest, we continue the selection by performance criteria.

B. Detection Performance and Parameter Settings

Using a quartile thresholding detector, against a wide range of deviation threshold parameters, we compared the ROC, F1-scores, *throughput*, and *approval rate* curves (Fig. 5) and the variability (Fig. 6) of three selection schemes with the case of no selection. Apart from examining effects of introducing the selection schemes, we use the above curves to determine the optimized n_{IQR} settings.

We explored n_{IQR} in a range $0 \rightarrow 10$ (incremental steps of 0.25). For $n_{IQR} > 4$, curves did not vary significantly. Hence, we depict these curves only for $n_{IQR} \leq 4$. Four criteria (*Relief*, *SU*, *Variance Ratio* or *No-sel* (i.e., no selection)) are presented with different markers. Fig. 5(a,e) presents their F1-scores. ROCs are shown in Fig. 5(b,f) with solid lines for sensitivity, and dotted for specificity. The *throughput* and clinical *approval* rate of the removal are illustrated in Fig. 5(c,g) and (d,h), respectively. The effects of n_{IQR} and the feature selection on the variability are demonstrated in Fig. 6.

We observed that characteristic curves were different between age groups. When no feature selection algorithm is used, the optimized empirical n_{IQR} is 3 for children and 2 for adults. If a feature selection algorithm is involved, the optimized empirical n_{IQR} reduced to around 1.5 for children or 1 for adults. F1-score and *throughput* are also improved significantly.

One important parameter setting is $n_{IQR} = 1$. Across three feature selection algorithms, this setting can work in a subject-independent manner with a high sensitivity (around 80%) and specificity (about 70%) regardless of participant age (i.e., children or adults). Although curves of the three saliency criteria were quite comparable, in *approval rate* and variability, the SU selection is better. Hence, we proposed a final model for the quartile thresholding detector that uses the SU selection technique and settings of $n_{IQR} = 1$, called *1IQR-SU*. In the next section, we do out-of-sample test with this model and compare with the aforementioned existing artefact removal methods.

C. Out-of-sample Tests

We used unseen test sets ($Ds3$ for children and $Ds4$ for adults) to validate the proposed detector (*1IQR-SU*). Table III compared *1IQR-SU* with existing complete-breath based methods: *B-3SD* [7] and *B-wavelet* [10]. *Manual* is the reference value calculated from removals by a human expert. Paired t-tests (two-tailed) for the variability (the test minus the operator, degrees of freedom of four ($Ds3$) or nine ($Ds4$)) are also reported in Table III. In terms of sensitivity, approval rate by operator (i.e., of the output are breaths "accepted" by the clinician), and the variability, *1IQR-SU* is the best detector. For example, in adults, although the mean *Rrs* of the *1IQR-SU* had a comparable average value with the operator, the standard deviation is lower (only 0.40 while the operator was 0.44 with p value is 0.06).

Rrs in our study ranged from $1.7 \rightarrow 8\ \mathrm{cmH_2OL^{-1}s}$ in adults (Table III), i.e. a mild to medium range of obstruction. To investigate the potential influence of obstruction on our detector, Figure 7a shows one performance metric, i.e. the approval rate, plotted against the median resistance for each recording, while Fig. 7b shows the distribution of the approval rate. It can be seen that while there is a large range, it remains mostly high regardless of the median *Rrs*. Similarly in children, Fig. 7c and 7d show that, with the exception of three recordings, approval rates remains high regardless of median *Rrs*, albeit within a smaller range of resistances. We also quantified the independence of the approval rate and *Rrs* using the distance correlation [29], and obtained 0.12 for adults and 0.27 for children, with complete independence indicated by 0.

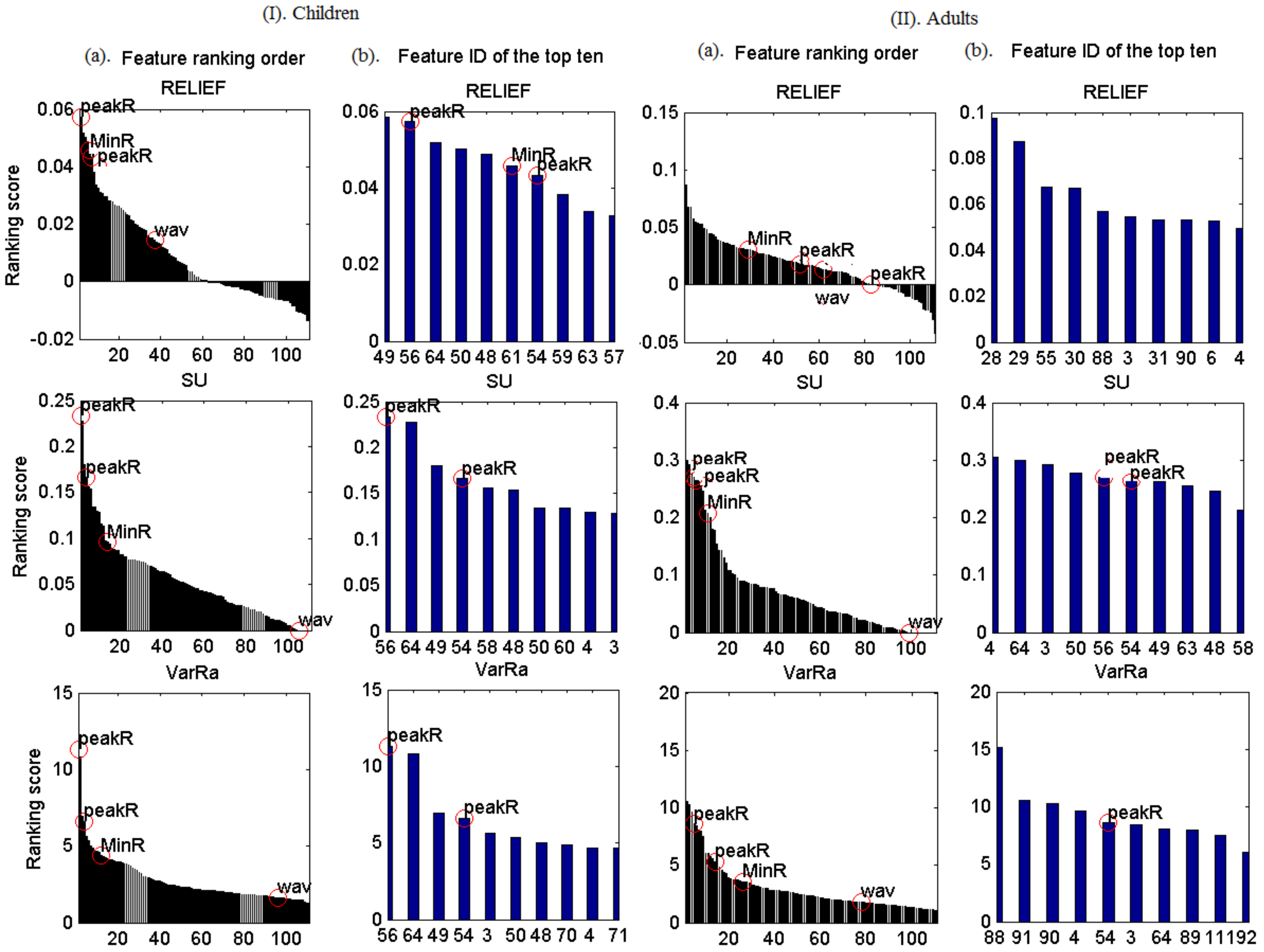


Fig. 4. Ranking scores for the feature pool (a) and the ten highest-score candidates (b). Vertical axes: scores calculated by three saliency criteria. Horizontal axes in (a): ranking order (highest =1, lowest=111); (b): feature identification (ID) in the pool. Features with circle markers were existing ones in literature.

Fig. 5. Effects of n_{IQR} and feature selection for paediatrics (top) and adult (bottom). Markers are for different feature selection algorithms. (a,e) are for F1-scores. (b,f) are for ROC curves (Solid lines: sensitivity; the dotted: specificity). (c,g) are for *throughput* curves. (d,h) are for *approval rate*.

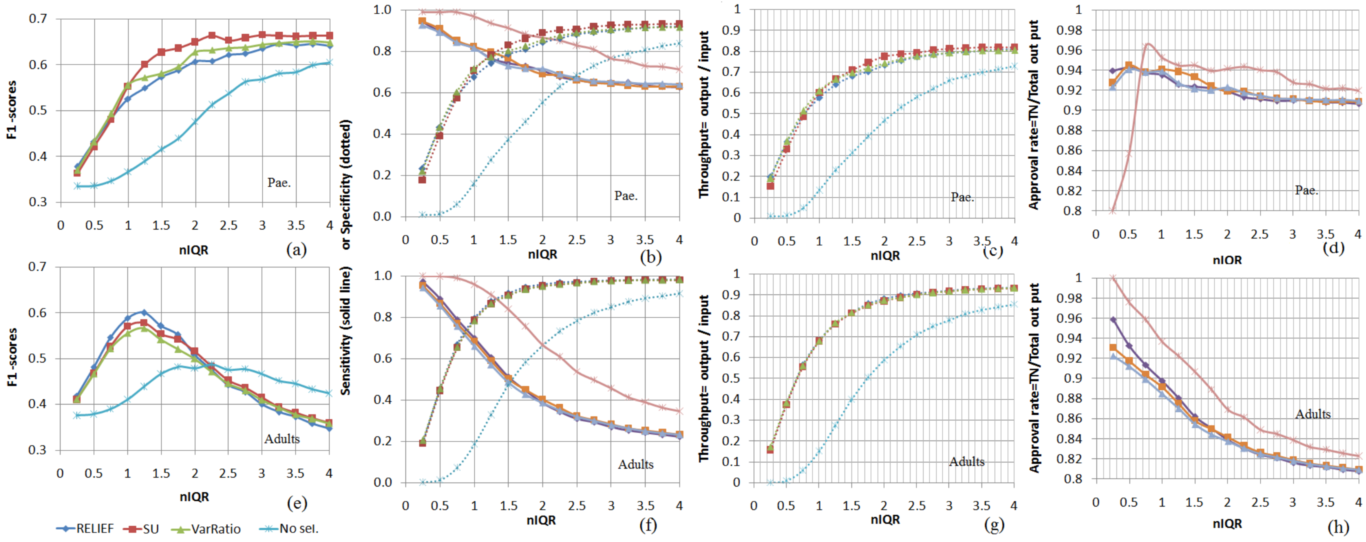


Fig. 6. Effects of n_{IQR} and feature selection on the variability of the average *Rrs* (standard deviation over the mean across patients). Markers are for different selection algorithms. (a) is for children and (b) is for adults.

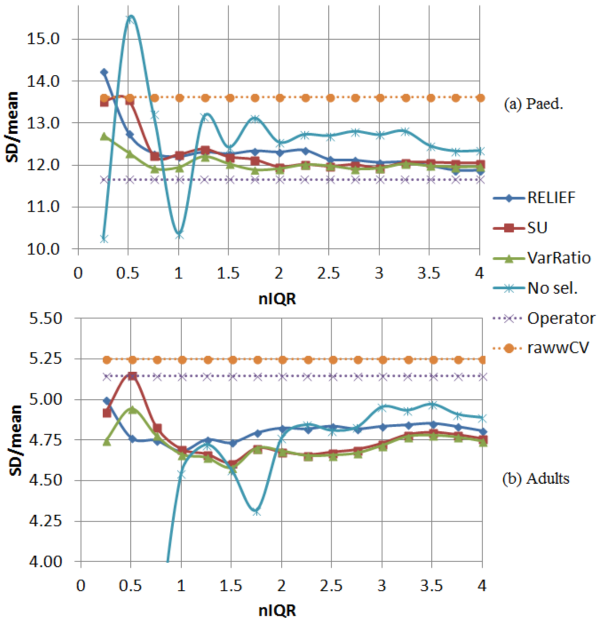


TABLE III
OUT-OF-SAMPLE TEST PHASE. *1IQR-SU*[a] IS THE PROPOSED. OTHERS ARE THE EXISTING. P VALUES [b] ARE FROM PAIRED T-TESTS (TWO-TAILED, $N = 5$ ($Ds3$) OR 10 ($Ds4$)). POSITIVES ARE ARTEFACTS [c]

Paediatrics (test $Ds3$)				
	B-3SD [7]	*B-wavelet* [10]	*1IQR-SU*[a]	Manual
F1-score[c] [%]	46.8	21.6	41.2	-
Approval[c] [%]	95.4	94.8	98.0	-
Throughput[c] [%]	82.7	30.0	67.1	84.1
Mean(±SD) *Rrs* $[cmH_2OsL^{-1}]$	3.72 (± 0.18)	3.74 (± 0.20)	3.70 (± 0.17)	3.75 (± 0.16)
P-value[b] *Rrs*	0.08	0.84	0.03	-
Mean(±SD) SD*Rrs* $[cmH_2OsL^{-1}]$	0.29 (± 0.13)	0.38 (± 0.12)	0.32 (± 0.11)	0.32 (± 0.13)
P-value[b] SD*Rrs*	0.23	0.25	0.82	-
Adults (test $Ds4$)				
F1-score [c] [%]	50.6	49.3	54.7	-
Approval[c] [%]	77.4	78.1	80.6	-
Throughput[c] [%]	85.4	55.9	63.4	68
Mean(±SD) *Rrs* $[cmH_2OsL^{-1}]$	3.69 (± 0.97)	3.69 (± 0.98)	3.66 (± 0.94)	3.67 (± 0.95)
P-value[b] *Rrs*	0.34	0.58	0.14	-
Mean(±SD) SD*Rrs* $[cmH_2OsL^{-1}]$	0.40 (± 0.23)	0.41 (± 0.27)	0.40 (± 0.24)	0.44 (± 0.21)
P-value[b] SD*Rrs*	0.05	0.86	0.03	-

[a] The detector used one interquartile range as a subject-independent parameter with the top ten salient features selected by the SU technique.

[b] compared to *Manual*, significant if $P < 0.05$.

[c] Removals by a specialist is considered ground truth. F1-score is the harmonic mean of precision and sensitivity. *Throughput* is the ratio of breath numbers in the output to input. *Approval rate* of the breaths remained after removal is the ratio of breaths that are “accepted” by the human to the total output breaths. Details of equations as in Section II-B.

IV. DISCUSSION

Our experiments were executed on recordings collected from adults and eight- to eleven-year-old children in Queensland and New South Wales, Australia. These recordings will be made available for public use on the UCI machine learning repository (http://archive.ics.uci.edu/).

For the feature extraction, we suggest to obtain *landmark* features of the two dimensional resistance-against-flow curves. This feature group is highly ranked by supervised learning techniques using saliency scores (RELIEF, SU, variance ratio). The SU score measures the correlation (mutual information) between one feature candidate of a breath and its label of abnormality. Meanwhie, RELIEF and variance ratio scores depict the clusterability of a feature candidate.

Although selecting the ten highest score candidates is common practice in the literature of feature learning, we acknowledge that an investigation for the stability of these feature preferences should be undertaken. Nevertheless, our results are consistent with more than one well-known feature selection algorithm with four separate data sets. As can be seen, scores that come after the top ten were significantly lower than the highest level. Thus, $k = 10$ satisfied our requirements. In practice, one may choose the entire landmark group and the resulting detector will perform comparably to the approach of this paper. This is because the majority of the top ten percentage are actually landmark features and the performance curves varied negligibly among selection algorithms.

While we demonstrated a reasonable degree of independence between the accuracy of our detector and levels of obstruction, further work is required to determine if the detector can be applicable to recordings from severely obstructed patients or those experiencing an exacerbation. Also of note is that in our datasets of healthy and asthmatic subjects, *Rrs* features ranked consistently high, whereas the features associated with *Xrs* did not rank highly for inclusion in the detector. This may be different in other diseases, and remains to be tested.

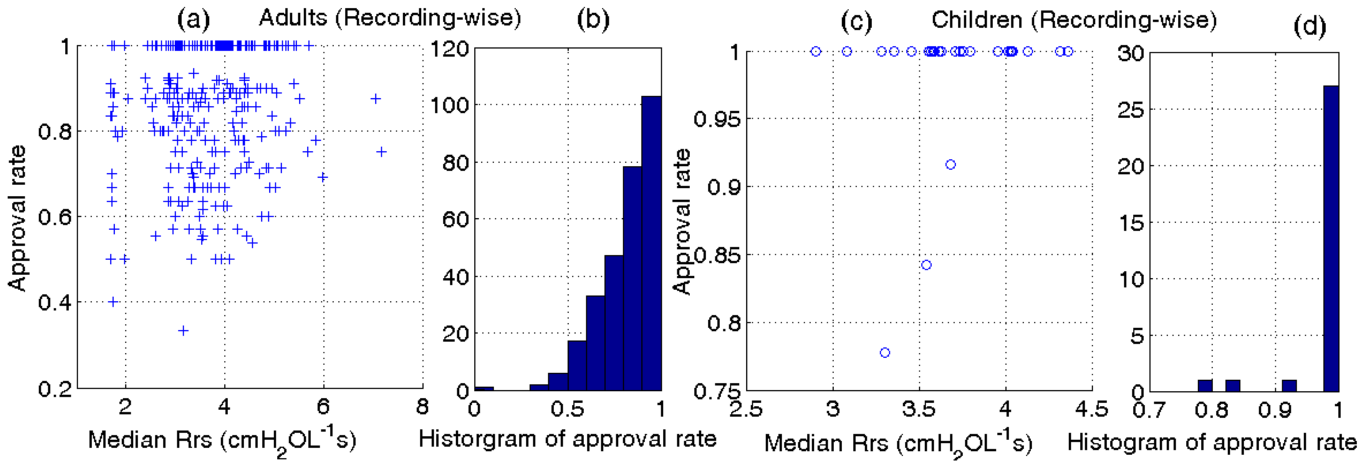
Finally, we have limited our analysis to a single frequency closest to what is commonly reported in the literature (5 Hz). However, our detector could also be applied to multi-frequency systems which are commonly used, using a similar set of features for each component frequency.

V. CONCLUSION

Recordings in lung function tests (FOT data) are naturally subject to artefacts. In this work, we studied two important problems related to of artefact removal. One is finding new features that are more relevant to respiratory artefact characteristics. The other is searching for an efficient scheme to detect and remove artefacts. Our proposed method is objective and has an equivalent reliability to the manual method (the current gold standard).

Our *landmark* features were found among the ten percentile of candidates in both children and adult groups, across the

Fig. 7. Approval rate plotted against median *Rrs* and histogram of approval rates for all recordings for the adults (a, b) and children (c, d) testing datasets.



three saliency criteria. To remove artefacts, we implemented a *thresholding* detector. The detector operates on complete-breaths to guarantee a balance between the inspiratory and the expiratory. We also proposed performance metrics (e.g., ROC and the variability) to determine threshold parameters.

During development, the performance curves (i.e., F1-scores, ROC, and the variability) against the parameter n_{IQR} showed that the top ten features outperformed the case of no feature selection. The three saliency scores yield nearly similar performance curves. The *1IQR-SU* configuration was found to give good results in a subject-independent setting, *regardless of age*. In out-of-sample tests, our detector performed similar to the *gold standard*, as assessed by through paired t-tests (two-tailed) for variability.

Our findings are an important first step towards objective and automated quality control of FOT measurements, as FOT moves beyond its long-standing role in the respiratory research realm, becomes more available in commercial systems and is increasingly adopted in clinical and home telemonitoring settings.

Acknowledgment

We would like to thank the UPTECH study team for the provision of the paediatric dataset. The UPTECH study is a collaborative research project between Queensland University of Technology (QUT), the Woolcock Institute of Medical Research (WIMR), and Queensland Institute of Medical Research (QIMR). The study was funded through a Linkage Grant LP0990134 by the Australian Research Council, QLD Department of Transport and Main Roads (DTMR) and QLD Department of Education, Training and Employment (DETE). We would like to thank Drs. Sophie Timmins and Greg King for the provision of the adult dataset. We also like to thank Dr. Diep N. Nguyen for technical discussions and suggestions to our work. There is no conflict of interest.

References

[1] A. B. DuBois *et al.*, “Oscillation mechanics of lungs and chest in man,” *Journal of Applied Physiology*, vol. 8, no. 6, pp. 587–594, 1956.

[2] H. Smith *et al.*, “Forced oscillation technique and impulse oscillometry,” *European Respiratory Monograph*, vol. 31, p. 72, 2005.

[3] A. Gobbi *et al.*, “A new telemedicine system for the home monitoring of lung function in patients with obstructive respiratory diseases,” in *eHealth, Telemedicine, and Social Medicine, 2009. eTELEMED’09. International Conference on*, IEEE, 2009, pp. 117–122.

[4] R. L. Dellacà *et al.*, “Home monitoring of within-breath respiratory mechanics by a simple and automatic forced oscillation technique device,” *Physiological measurement*, vol. 31, no. 4, N11, 2010.

[5] S. C. Timmins *et al.*, “The feasibility of home monitoring of impedance with the forced oscillation technique in chronic obstructive pulmonary disease subjects,” *Physiological measurement*, vol. 34, no. 1, p. 67, 2012.

[6] H Lorino *et al.*, “Influence of signal processing on estimation of respiratory impedance,” *Journal of Applied Physiology*, vol. 74, no. 1, pp. 215–223, 1993.

[7] P. D. Robinson *et al.*, “Procedures to improve the repeatability of forced oscillation measurements in school-aged children,” *Respiratory physiology & neurobiology*, vol. 177, no. 2, pp. 199–206, 2011.

[8] C. Schweitzer *et al.*, “Influence of data filtering on reliability of respiratory impedance and derived parameters in children,” *Pediatric pulmonology*, vol. 36, no. 6, pp. 502–508, 2003.

[9] N. J. Brown *et al.*, “A comparison of two methods for measuring airway distensibility: nitrogen washout and the forced oscillation technique,” *Physiological measurement*, vol. 25, no. 4, p. 1067, 2004.

[10] S. A. Bhatawadekar *et al.*, “A study of artifacts and their removal during forced oscillation of the respiratory system,” *Annals of biomedical engineering*, vol. 41, no. 5, pp. 990–1002, 2013.

[11] C.-L. Que *et al.*, "Homeokinesis and short-term variability of human airway caliber," *Journal of Applied Physiology*, vol. 91, no. 3, pp. 1131–1141, 2001.

[12] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal, The*, vol. 27, no. 3, pp. 379–423, 1948.

[13] M. Ackerman and S. Ben-David, "Clusterability: a theoretical study," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 1–8.

[14] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Proceedings of the tenth national conference on Artificial intelligence*.

[15] J. F. Nunn, *Applied respiratory physiology*. Butterworth-Heinemann, 2013.

[16] W. N. Ezz *et al.*, "Ultrafine particles from traffic emissions and childrens health (uptech) in brisbane, queensland (australia): study design and implementation," *International journal of environmental research and public health*, vol. 12, no. 2, pp. 1687–1702, 2015.

[17] M. Mazaheri *et al.*, "School childrens personal exposure to ultrafine particles in the urban environment," *Environmental science & technology*, vol. 48, no. 1, pp. 113–120, 2013.

[18] C. W. Thorpe *et al.*, "Modeling airway resistance dynamics after tidal and deep inspirations," *Journal of Applied Physiology*, vol. 97, no. 5, pp. 1643–1653, 2004.

[19] N. Beydon *et al.*, "An official american thoracic society/european respiratory society statement: pulmonary function testing in preschool children," *American journal of respiratory and critical care medicine*, vol. 175, no. 12, pp. 1304–1345, 2007.

[20] S. C. Timmins *et al.*, "Day-to-day variability of oscillatory impedance and spirometry in asthma and copd," *Respiratory physiology & neurobiology*, vol. 185, no. 2, pp. 416–424, 2013.

[21] E. Bateman *et al.*, "Global strategy for asthma management and prevention: gina executive summary," *European Respiratory Journal*, vol. 31, no. 1, pp. 143–178, 2008.

[22] E Oostveen *et al.*, "The forced oscillation technique in clinical practice: methodology, recommendations and future developments," *European Respiratory Journal*, vol. 22, no. 6, pp. 1026–1041, 2003.

[23] C. J. V. Rijsbergen, *Information Retrieval*, 2nd. Newton, MA, USA: Butterworth-Heinemann, 1979.

[24] I. Daubechies and B. J. Bates, "Ten lectures on wavelets," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1671–1671, 1993.

[25] W. H. Press *et al.*, "Numerical recipes in c," *Cambridge University Press*, vol. 1, p. 3, 1988.

[26] G. Brown *et al.*, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27–66, 2012.

[27] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[28] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[29] G. J. Székely *et al.*, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.