A Machine Learning Framework for Sleep Apnoea Detection Based on Imbalanced Pulse and Oximetry Data

Dongjin Yang,[1] Jingqiong Zhang,[1] Zhenglin Li,[2] Heather Elphick,[3] Eishaan Bhargava,[4] and Lyudmila Mihaylova[1,5]

[1]*School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, UK*

[2]*Institute of Artificial Intelligence, Shanghai University, Shanghai 200444, China*

[3]*Department of Paediatric Respiratory and Sleep Medicine, Sheffield Children's Hospital, Sheffield, UK*

[4]*Department of Paediatric Otolaryngology, Sheffield Children's Hospital, Sheffield, UK*

[5]*Insigneo Institute The Pam Liversidge Building Sir Frederick Mappin Building The University of Sheffield Mappin Street Sheffield, S1 3JD*

Abstract

Sleep apnoea, a disorder impacting both children and adults, typically requires costly and time-intensive diagnostics. This paper introduces a novel framework that uses the wavelet transform to extract features from sleep signals and the RUSBoost algorithm to address the challenge of imbalanced data in detecting sleep apnoea, which enables home self-monitoring. Patient data features short apnoea epochs and long periods of normal breathing, creating imbalances that challenge classification algorithms. The framework was tested on three public datasets with varying imbalance ratios. Significantly, the Childhood Adenotonsillectomy Trial dataset (CHAT), with an 'apnoea' to 'normal' period ratio of 1:15, effectively reflects actual sleep apnoea signals from children. The proposed framework with the CHAT dataset achieved a maximum accuracy of 91.54%, sensitivity of 72.06%, specificity of 92.39%, and an AUC of 0.923, surpassing state-of-the-art home screening models. For the classification task, this study compared several machine learning techniques, including support vector machine, K-nearest neighbour, and Dirichlet process Gaussian mixture model algorithms. It is found that the RUSBoost algorithm provides the most accurate results when the ratio of the 'apnoea' to the 'normal' period reaches an imbalance of 1:3 or greater.

Keywords: Sleep apnoea, Children, Classification, Machine Learning, Imbalanced Data, Wavelet Transform
DOI: 10.31526/JMLFS.2025.552

1. INTRODUCTION

1.1. Background

Humans sleep about one-third of their lives; therefore, diagnosing and managing sleep-disordered breathing (SDB) is important. These conditions can greatly affect the quality of life [1]. People with sleep apnoea experience pauses in breathing during sleep, called apnoea, or significant decreases in breathing, called hypopnoea. These disruptions pose a substantial risk to people's health because they can lead to various health problems. [2]. Sleep apnoea affects approximately 1% to 6% of adults and 1.2% to 5.7% of children [3]. Given its widespread occurrence worldwide and its lasting impact on sleep, it is necessary to detect and treat sleep apnoea [4]. Sleep apnoea includes three primary types of respiratory events, each distinguished by distinct pathophysiological character- istics [5]. Central sleep apnoea (CSA) is a condition in which the central nervous system fails to send signals to the respiratory muscles, causing breathing difficulties. The patients usually stop breathing involuntarily for 10 to 30 seconds. Hypopnoea involves a substantial reduction in airflow, with decreases ranging from 30% to 90% of normal levels [6]. This phenomenon usually lasts for more than ten seconds and leads to hypoventilation and resulting hypoxaemia. Obstructive sleep apnoea (OSA) [7] occurs when the upper airway is partially or completely blocked despite the patient's persistent breathing efforts. This obstruction blocks airflow into the lungs, leading to severe oxygen desaturation and subsequent cardiovascular stress. Polysomnography (PSG) [8], widely recognized as a noninvasive diagnostic test for analysing sleep patterns and identifying various SDBs, is commonly used as a sleep study. This test is typically conducted by sleep technologists and is performed in hos- pitals, independent medical facilities, or specialized sleep clinics. However, these methods of collecting data require monitoring of patients in a hospital setting equipped with a lot of specialized equipment. The cost and limited availability of such a facility restrict its wider use. Additionally, the overnight data collected needs to be analysed by qualified medical personnel, such as physicians or clinical technologists, in compliance with the guidelines outlined in the American Academy of Sleep Medicine (AASM) manual [6]. Due to the intricate nature of the scoring process and the limited availability of trained professionals, there are often considerable delays in confirming diagnoses of SDB. Furthermore, PSG examinations are always an uncomfortable experience due to the large number of sensors attached to the patient during data acquisition. In response to this challenge, there has been a significant increase in research focused on the development of automated systems for detecting SDB. Research in this area has predominantly concentrated on enhancing detection capabilities through fewer signal channels or the choice of portable devices. Specifically, methods that utilise peripheral

haemoglobin oxygen saturation (SPO2), electrocardiogram (ECG), and acoustic signals have been extensively explored [9]. Moreover, these detection techniques have been shown to have higher accuracy than those relying on acoustic signals. Given that SPO2 signal measurements can be readily ob- tained through nighttime pulse oximeters, frameworks that leverage these metrics are particularly advantageous for facilitating sleep health monitoring at home [10]. In addition, it is also considered that combining multiple signals for detection may obtain better detection results. Given the capability of pulse oximeters to concurrently measure multiple parameters, including pulse and blood oxygen levels, combining these data from the same sensor sources to detect sleep apnoea is feasible. Figure 1 illustrates conventional and future frameworks for the detection of sleep apnoea-hypopnoea. Conventional practice (Figure 1(a)) involves monitoring patients overnight in the hospital, attaching multiple sensors to their bodies to collect data throughout the night. Be- cause patients do not remain still during sleep, the data is often heavily distorted by artefacts, which can affect the diagnoses of clinicians. The clinician will analyse each epoch (30 seconds) to formulate a diagnosis, which is an excessively time-consuming pro- cess. Machine learning techniques can streamline this process for direct diagnostics. Figure 1(b) illustrates that individuals seeking to self-diagnose apnoea can monitor their health by utilising a home oximeter during sleep. The oximeter measures blood oxygen levels and pulse rate, subsequently uploading this data to cloud storage. Once the computer downloads the data from the cloud, it employs the wavelet transform to extract the characteristics of each segment. Then, machine learning techniques will be used to classify each segment and provide the clinician with feedback on the judgement results. In sleep apnoea diagnosis, the Apnoea-Hypopnoea Index (AHI) quantifies the severity of the condition. The AHI represents the average number of apnoea and hypopnoea occurring per hour of sleep [11]. In this study, data from adults with AHI 5 are considered as 'apnoea'; otherwise, they are defined as 'normal'. For children, however, since the primary manifestation of apnoea is behavioural characteristics [3], the threshold of AHI should be adjusted to AHI 2. The technology for automatically detecting sleep apnoea based on biological signals is developing rapidly, but there are still challenges in the task of automatic real-time detection. First, medical signal data is often highly unbalanced; that is, the number of diseased segments is significantly less than the number of normal segments. This phenomenon will continue to be amplified in paediatric patients. This imbalance may cause the training results to be biased towards normal segments, thereby reducing the model's sensitivity to diseased states. Second, PSG recording is time-consuming and laborious, and the sleeping environment can also cause errors in their test results. Third, there are endless devices for recording human biological signals, and researchers hope to use simple equipment to achieve more accurate detection.

1.2. contributions

To tackle these challenges, we propose a machine learning framework for the autonomous detection of sleep apnoea that leverages wavelet transform-based features from blood oxygen and pulse data. The contributions of this paper can be summarised as follows:

- This paper describes an approach for detecting abnormal conditions using blood oxygen and pulse data that can be acquired through a single, easy-to-use device such as a pulse oximeter. This approach reduces detection costs and allows patients to monitor their sleep signals conveniently at home, enhancing the practicality and accessibility of apnoea diagnosis.
- A new framework is proposed to automatically detect sleep apnoea using the Daubechies wavelet [12] transform for effective feature extraction combined with the machine learning method RUSBoost [13]. It is found that features extracted by wavelet Daubechies 1 demonstrate good performance with SPO2 data. Wavelet Daubechies 3 is effective when using single pulse data and fusion with SPO2 data. All classification results perform well with a 30-second signal overlap (1-minute segment).
- The results of using different machine learning algorithms for state-of-the-art binary classification approaches, such as Support Vector Machines (SVM) [14], K-nearest Neighbour (KNN) [15], Dirichlet Process Gaussian Mixture Model (DPGMM) [16] approaches, and neural networks, were compared on three balanced and unbalanced datasets. It can be found that when the imbalance ratio of the 'apnoea' to the 'normal' period exceeds 1:3, RUSBoost is more suitable for classi- fication with the imbalanced data.
- Unlike prior studies that utilised balanced data from adults, our research incorporates both commonly used adult data (AHI 5) and data of children from the Childhood Adenotonsillectomy Trial (CHAT) [17] (AHI 2), which is highly imbal- anced. Our approach has achieved remarkable results with the CHAT dataset, with an accuracy of 91.54%, a sensitivity of 72.06%, a specificity of 92.39%, and an AUC of 0.923, which outperforms the state-of-the-art home screening models validated on this dataset. Furthermore, by comparing dual-signal feature fusion with single-signal processing, we found that the accuracy of fused results was 3% higher than that of single-signal outcomes. The paper is organised as follows. Recent studies on sleep apnoea are introduced in Section 2. Section 3 introduces the methodology of the framework, including feature extraction based on wavelet transform, and the two main classification algorithms used here, namely RUSBoost and DPGMM. The main results of this work are summarised in Section 4.

2. RELATED WORK

The identification of sleep apnoea has consistently been a significant subject in sleep analysis. Various automated techniques for detecting sleep apnoea have been developed in the literature. In the field of sleep apnoea detection, the electroencephalogram (EEG) is a commonly used signal. Zhao et al. [18] introduced a method for classifying OSA and CSA by analysing the properties of sample entropy and variance within two subbands of EEG signals. The classification was performed using several machine learning techniques, including SVM, random forests (RF) [19], and KNN, and reached high accuracy, which is impressive. However,

collecting EEG data is quite complicated. In addition to EEG, ECG signals are also widely used for disease detection. Li et al. [20] employed a single-lead ECG to detect sleep apnoea, benefiting from the minimal setup of only two sensors, which reduces discomfort significantly during sleep. Pinho et al. [21] extract features from the ECG signal to detect sleep apnoea. The best accuracy was 82.12%, with a sensitivity and specificity of 88.41% and 72.29%, respectively. Although the performance evaluation is not high, their feature selection provides inspiration for this research. However, given that our experiment includes data from children, we aim to minimise the number of sensors even further. In comparison, the blood oxygen and pulse signals selected in this experiment are relatively more straightforward to obtain, which is similar to what Gutierrez-Toba [7] and his colleague used. They used home oximeters to test the severity of OSA. They achieved 89% accuracy and 58% specificity in distinguishing AHI 5. This indicates a high misdiagnosis rate. However, for AHI 15 and AHI 30, the specificity was significantly improved to over 85%. This demonstrates the feasibility of their model on adult data and the feasibility of home oximeter data for detecting apnoea. Chutinan et al. [22] employed sleep sounds, SPO2, and pulse rate to detect sleep apnoea. A notable innovation introduced in their methodology was the fusion of SPO2 and pulse rate data into a singular signal for the training phase, resulting in an accuracy of 79%. While this accuracy is lower compared to previous studies on apnoea detection, it is worth noting that the dataset used here was imbalanced with only 10% positive cases. This imbalance significantly influenced the training process, skewing the results towards negative predictions. The studies mentioned above all focused on analysing the characteristics of signals based on time information. While the information about frequency, as recommended by the AASM [6], is also very useful for detection, using this frequency information well can help improve the model. In addition, airflow and SPO2 have been shown to change significantly during apnoea [23]. Vero´nica [24] uses the wavelet transform to analyse an overnight airflow signal from children and gets 77.9% accuracy on OSA detection for AHI 2, while ours is 91.54%. He suggested using the order of the appropriate frequency band for feature selection, which is an excellent way to reduce information loss. In addition, he compared the accuracy of different classification models and found that AdaBoost.M2 [25][26] performed best among other classifiers, which inspired this paper. Li et al. [27] use SPO2 as test data and achieve high accuracy. The paper proposes a new framework based on a clustering method, the DPGMM algorithm, to detect sleep apnoea. The framework used the wavelet transform to generate features in the fre- quency domain, and two datasets were compared. One dataset achieves an accuracy of 97.01%, making it a very effective classifier. However, the performance with the other dataset is slightly inferior under this model. The difference between the two datasets stems from the balance of the data, which is a common issue in medical testing. This is also the issue that is discussed in this experiment. Regarding training with imbalanced data, Hassan and Haque [28] give a RUSBoost application on sleep apnoea iden- tification that is similar to this paper. He uses ECG data and its spectral data to extract features. The accuracy of his experiments is 85.37%, which is the highest in his experiments. In further research, he added the TQWT technology [29], which is very useful in decomposing ECG signals. Compared with the test before, the accuracy increased to 91.94%. However, although they claim that their work focuses on unbalanced data, the ratio of 'apnoea' to 'normal' cycles used is only 1:3. Additionally, ECG recording necessitates that patients attach electrodes to their chest, legs, and arms, which is quite inconvenient. Our study used even more unbalanced data from children, which is more representative. Moreover, we chose SPO2 and pulse, making it more convenient for patients to record signals. In a recent publication, Manish Sharma [30] and colleagues explored multisignal detection of sleep apnoea, comparing the effects of balanced and unbalanced data using the RUSBoost algorithm. They achieved an accuracy 89.39% in the optimal data combination and reported an AUC of 0.905. However, the study did not compare the performance across different algorithms with the same dataset, limiting a deeper understanding of how data imbalance affects detection efficacy. Additionally, while the researchers balanced the data in their experiments, they did not give a specific balance and imbalance ratio. Furthermore, if the data were merely downsampled, this could result in the loss of critical information from the balanced data, potentially skewing the accuracy of the results. To more accurately match the recording and detection of the data from children, this experiment uses the CHAT dataset, with the degree of imbalance detailed in Section 4.1. Fernando and colleagues [31] previously used this dataset for sleep apnoea classification by convolutional neural networks (CNNs), analysing the SPO2 signal over 20-minute segments to classify apnoea severity. Despite the advantages of CNNs, they reported an accuracy of 77.6% and a sensitivity of 71.2%, likely constrained by the data's imbalance. Our study employed the RUSBoost algorithm to overcome this limitation and achieved significantly improved results: 89.97% accuracy and 74.79% sensitivity for individual SPO2 signals using 1-minute segments for real-time detection. Furthermore, the results of the feature fusion method discussed in Section 4 further surpass these numbers, demonstrating the effectiveness of the enhancement. Regarding multisignal feature fusion technology, Michelsanti et al. [32] summarised several simple feature fusion methods and listed the usage methods in detail, which is of great help to the feature fusion part later in this paper.

3. METHODOLOGY

In the proposed framework, the signals are preprocessed to remove the artefact first (introduced in section 4.1) and decomposed using the wavelet transform. The detail coefficients are used to compute wavelet features. In addition, statistical features based on raw signals are also used in this work. All extracted features are fed into a classifier to detect sleep apnoea. Figure 1 illustrates the proposed approach, which is further detailed in this section.

3.1. Feature Extraction

3.1.1. Wavelet features

Wavelet transform (WT) has become a new mathematical tool for the multiresolution decomposition of time series signals and has potential applications in computer vision [1]. The pulse and oxygen signals are nonstationary and originate from a nonlinear system, and WT can identify subtle morphological changes in nonstationary signals [2]. The WT is divided into discrete wavelet

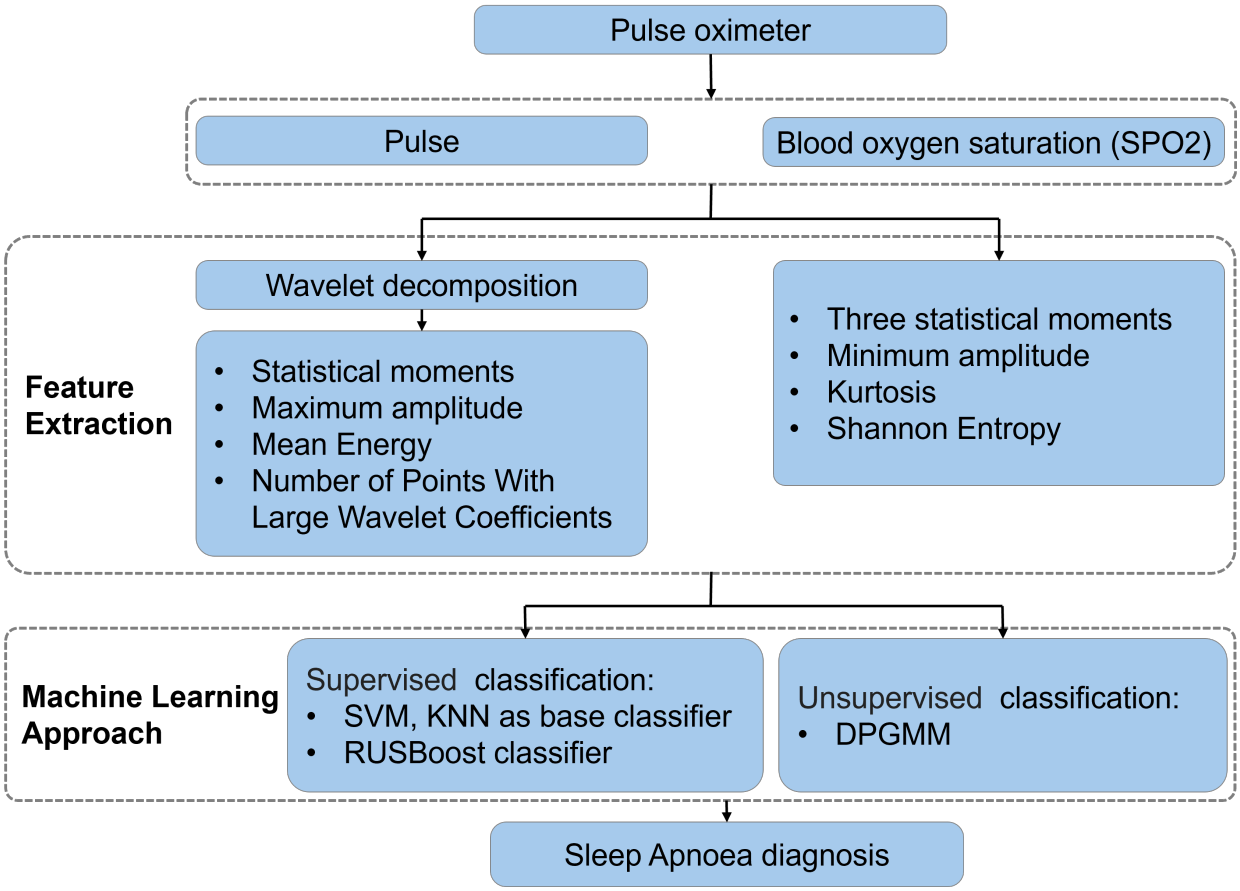

FIGURE 1: Diagram for the proposed framework

transform (DWT) and continuous wavelet transform (CWT). Compared to the CWT, the DWT can reduce computational complexity. In addition, it captures frequency and time location information more effectively than the discrete Fourier transform (DFT). Furthermore, the DWT has been used successfully in previous studies on apnoea, described in Section 2.

Since the invention of the wavelet transform, many experts have created many different types of wavelets. This experiment mainly compares Daubechies wavelets (db) families [3] 1 to 4 (db1 to db4). 1 to 4 here means the number of zero moments or vanishing moments [4], and db1 is the well-known Haar wavelet [5]. Figure 2 shows Daubechies wavelets db1 to db4 used in this research. These plots help justify the wavelet family selected for our study. For example, db4 has oscillatory and stationary parts, making it well suited for extracting transient and local features from nonstationary signals such as sleep signals. Different wavelets apply to different signals due to their different shapes, which will be confirmed in Section 4. In this study, the signal

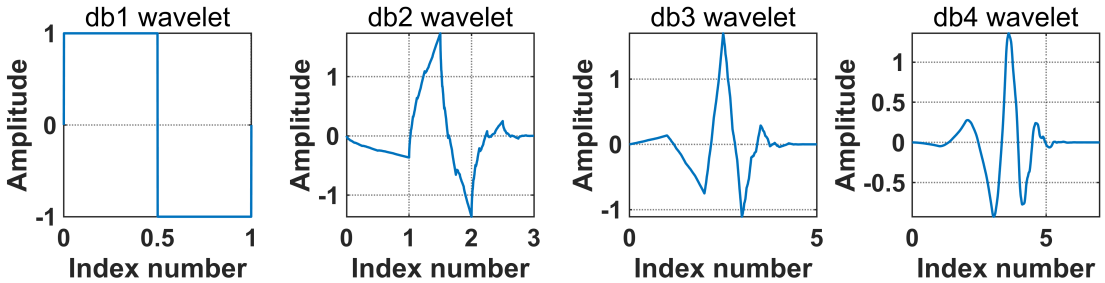

FIGURE 2: Plots of the different wavelets

was divided into multiple 1-minute windows. Since each signal has a different sampling frequency, the number of values used for calculation is also different. Additionally, typical hyperpnoea (also referred to as apnoea in this study) following apnoea episodes occurs between 0.784 and 0.890 Hz events, which increase the respiratory rate. Children who tested positive for apnoea exhibited greater variability in the Power Spectral Density (PSD) range of 0.35–0.43 Hz [6]. Therefore, experiments will select levels that contain frequencies within this range. For instance, Figure 3 shows an example of an SPO2 signal whose sample rate is 10 Hz and whose level 4 and 5 coefficients are used. After obtaining the required detail coefficients, the following features are extracted to

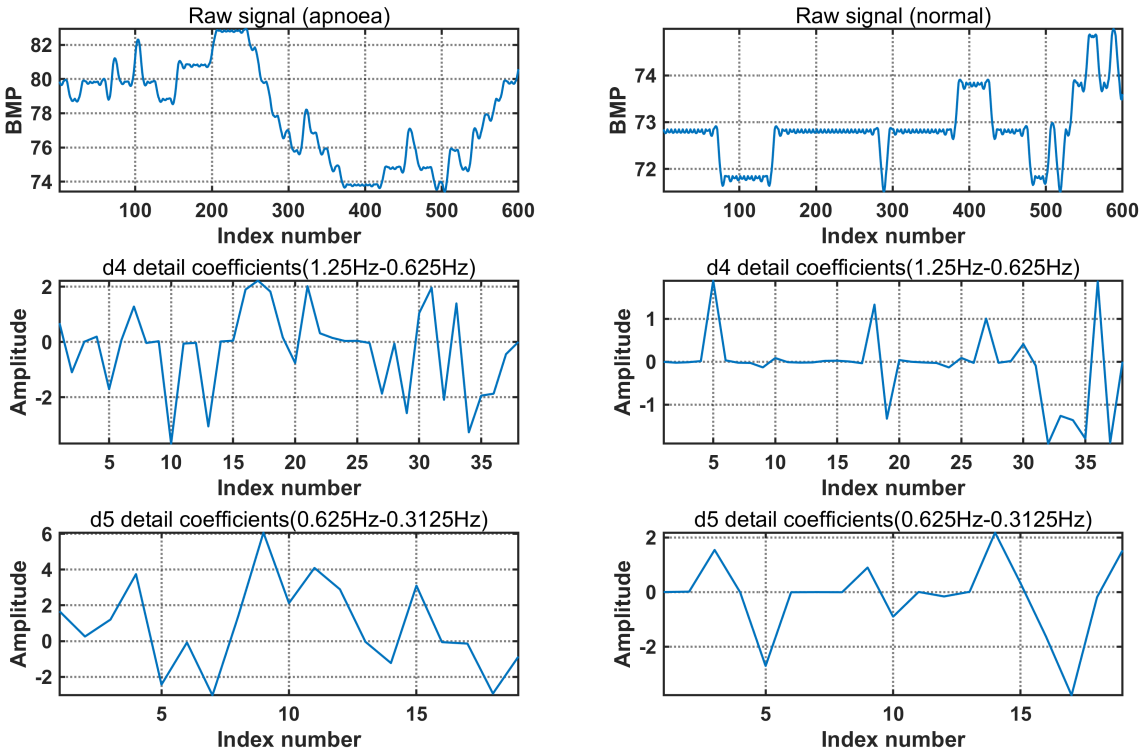

FIGURE 3: Decomposition process of a signal using DWT

measure the information they contain:

- Mean Energy [7]: the average energy found in the wavelet detail coefficients varies between the 'apnoea' and the 'normal' segments, making it a useful feature for differentiation. The average energy E of a fragment is calculated as follows:

$$E = \frac{1}{N}\Sigma_i^N d(i)^2,$$

  where $d(i)$ represents the i-th element of the wavelet coefficients for a segment, and N denotes the total number of coefficients.

- Statistical features are computed, including range, variance, and maximum value.

- Number of the large coefficients [8]: the count of points exceeding the threshold can serve as a feature to differentiate the 'apnoea' segment from the 'normal' segment in signal measurements. However, this feature is uncertain. The original work utilised this feature to achieve high accuracy, but the threshold selection is subjective. Additionally, if this feature is used, the signal needs to be resampled. To avoid this, the Adaptive Cumulative Sum (CUSUM) [9] method is used to judge different thresholds $h = 4 * \frac{\sigma}{\sqrt{f_s}}$, where f_s is the sample frequency of the signal. More details can be found in [9]. Section 4 will provide corresponding results comparisons.

3.1.2. Time-domain features

Feature extraction mainly targets signal and wavelet transform detail coefficients, which are introduced in Section 3.1.1. The specific features for signal values are as follows:

- Variance: since the variance of apnoea subsequences is usually higher than the variance presented in normal subsequences [10], variance can be used as a classification feature for the classifier.

- Range: the range is determined by the difference between the highest and lowest signal values within a segment. Typically, SPO2 values decrease after apnoea events and remain relatively stable during normal subsequences. In addition, pulse values usually oscillate more during apnoea and less during normal subsequences. Therefore, the range can be selected as a feature to detect apnoea.

- Average: SPO2 values usually drop after an apnoea event, the average value of the SPO2 segment may change after an apnoea. By observing the changes in the pulse signal, it can be found that the pulse signal has evident oscillations during the apnoea event. If it occurs for a short period, it appears as a falling signal with a very high slope.
- Minimum: SPO2 values typically decrease significantly following apnoea events, as previously mentioned, which is not a prevalent occurrence in normal sleep. Consequently, it is advantageous to evaluate the minimal SPO2 value within a segment. Similarly, the minimum pulse value typically decreases substantially within the apnoea event segment as a result of the large oscillations.
- Kurtosis [11]: the peakedness of a signal can be ascertained using kurtosis. A dataset with a kurtosis greater than 0 is characterised by a higher peakedness and more extreme values. A flatter distribution is indicated by a kurtosis that is less than 0. The stability of both SPO2 and pulse signals is altered following an apnoea event. Compared to the normal segment, the kurtosis value of the abnormal segment increases.
- Shannon Entropy [12]: Shannon entropy is a metric that quantifies the amount of information in a system using the probabilities of various scenarios within the system. It quantifies the changes in energy distribution during the decomposition process, illuminating the fundamental dynamic behaviour and signal irregularities. The Shannon entropy escalates with a rise in the uncertainty of an event or entity and diminishes with a drop in uncertainty.

3.2. Machine-Learning Approaches

In traditional machine learning binary classification, SVM is the most representative supervised learning method. Generally, it performs well in binary classifiers on balanced datasets. However, biomedical tests often contain imbalanced data, as shown in Table 1. The skewness of classes in an imbalanced dataset pushes the hyperplane toward the minority class [13], which may lead to inaccurate prediction results of SVM. Therefore, this study uses RUSBoost, a classifier that can be trained on imbalanced data. In addition, this study also compares the feasibility of unsupervised classification algorithms in this application.

3.2.1. RUSBoost Algorithm for Unbalanced Data

The RUSBoost [14] classification algorithm combines two methods: random undersampling (RUS) and boosting techniques. This method is mainly used to solve the class imbalance problem in machine learning tasks. This happens when the data categories are unevenly skewed or biased. For example, in this study, it is necessary to train the classification of the 'apnoea' category. However, this category is a minority in the overall data, which may cause bias in the training model and negatively affect its performance. The RUSBoost method is used to solve this imbalanced problem. In RUSBoost, weak learners [15] refer to simple or low-complexity classification models that perform slightly better than random guessing. The weak learner cannot achieve high accuracy, especially when dealing with complex problems. However, when multiple weak learners are combined in a boosting framework, they can collectively form a strong classifier. Each weak learner is trained on a randomly undersampled subset of the data to balance the class distribution, which is a key part of RUSBoost to handle the imbalanced classes. This section mainly introduces the mathematical theory of RUSBoost [16].

First, assume all the features $X = \{x_i\}$ and its corresponding labels $Y = \{y_i\}$, so the dataset can be express as $S = (x_i, y_i)$, where $y_i \in \{0, 1\}$ and $i = 1, 2, 3..., m$. In this study, 0 is defined as a 'normal' class, and 1 is a 'apnoea' class. 'Normal' is the category with a larger proportion. Set D_t as the weight of each weak learner and $D_1(i) = \frac{1}{m}$, which means that the weight of each instance is equal at the beginning of the model training. t here is the iteration value and $t = 1, 2, 3, ..., T$. Then, the 'normal' samples are randomly eliminated to modify the class distribution in the training set until the dataset reaches the necessary balance between classes, thus obtaining a temporary training dataset $S_t^{'}$ and new weight $D_t^{'}$. This is the meaning of RUS. Next, the temporary dataset $S_t^{'}$ and its corresponding weights $D_t^{'}$ are fed into a base learning model (WeakLearn), which trains a weak hypothesis $h_t(x_i) = \{p_{y=0}(i), p_{y=1}(i)\}$. p is the predicted probabilities of x_i from weak learner. After this, the pseudo-loss ϵ_t which measures the error of h_t weighted by D_t can be calculated as:

$$\begin{aligned} \epsilon_t &= \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y)), \\ &= \sum_{(i,y):y_i \neq y} D_t(i)(1 - p_{y=0}(i) + p_{y=1}(i)) \end{aligned} \tag{1}$$

This error can be used to compute the weight update factor $\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$, which adjusts the influence of h_t in the final model. Subsequently, weights D_{t+1} are updated.

$$D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+p_{y=0}(i)+p_{y=1}(i))} \tag{2}$$

Equation (2) shows that the weights assigned to misclassified instances increase, while those assigned to correctly classified instances decrease. This adjustment allows the model to focus on more challenging cases in future iterations. At the beginning of the next iteration, the weights D_{t+1} will be normalised as

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}, \tag{3}$$

where $Z_t = \sum_i D_{t+1}(i)$. After T iterations, the algorithm forms the final hypothesis, denoted by $H(x)$, which is constructed by combining the T weak hypotheses through weighted voting. The weight of each vote corresponds to the accuracy of the hypothesis, which is inversely proportional to α_t. The expression of $H(x)$ is given by the equation

$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} h_t(x, y) log \frac{1}{\alpha_t}. \tag{4}$$

By repeatedly concentrating on fixing its prior errors, this approach ensures that the machine learning algorithm gradually enhances its capacity to categorize increasingly challenging, frequently minority, samples, efficiently managing the imbalance in the dataset.

3.2.2. Variational inference for Dirichlet process mixture model

Blei and Jordan originally proposed a variational Dirichlet process mixture model (DPMM) [17] and subsequently optimised it by Kenichi Kurihara et al. [18]. Building on these foundational methods, Li et al. [8] developed a clustering-based classification framework specifically used for sleep apnoea detection. The model in this framework uses the log-likelihood ratio (LLR) for data classification.

The Dirichlet Process Gaussian mixture model in the stick-breaking [19] representation involves an infinite number of components, where each component is associated with a set of parameters drawn from a base distribution H. The mixing proportions are determined via a stick-breaking process governed by a parameter α. The specific process is as follows:

$$\eta_i \sim H, \tag{5}$$

$$v_i \sim Beta(1, \alpha), \tag{6}$$

$$\pi_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \tag{7}$$

where $\{\pi_i\}_{i=1}^{\infty}$ is the mixing weight. $v_i \in [0, 1]$ which is an infinite collection of 'stick lengths' $V = \{v_i\}_{i=1}^{\infty}$. In this research, features are extracted and combined from each segment of the SPO2 and pulse signal, categorized as either 'apnoea' or 'normal.' It is posited that the distribution of features extracted from apnoea and normal segments differs. Consequently, decisions can be made by comparing the probabilities of each test segment under the 'apnoea' and 'normal' models. Features from 'apnoea' and 'normal' segments can be represented using two Gaussian mixture models (GMM), as a GMM can accurately approximate any distribution when configured with the appropriate number of components and parameter adjustments. The likelihood of a data point x in a model is expressed as follows:

$$p(x|\{\pi_i, \eta_i\}_{i=1}^{\infty}) = \sum_{i=1}^{\infty} \pi_i \mathcal{N}(x; \eta_i), \tag{8}$$

where $\mathcal{N}(\cdot)$ is the Gaussian distribution. Denote the training features $X = \{x_n\}_{n=1}^{N}$, and $Z = \{z_n\}_{n=1}^{N}$ is the set of all labels. The main problem is to compute the posterior $p(z_n|X, \theta)$ over the labels and the predictive density $p(x|X, \theta)$, whose expression is:

$$p(x|X, \theta) = \int_{H,V} p(x|H, V) \int_{Z} p(W|X, \theta), \tag{9}$$

where $\theta = \{\alpha, \lambda\}$ is the hyperparameters from prior. $W = \{H, V, Z\}$ is the set of all latent variables of the DP mixture. Since $p(W|X, \theta)$ is difficult to compute analytically, a parametric family of variational distribution $q(W; \phi)$[18] is utilised to approximate the posterior.

$$q(W; \phi) = \prod_{i=1}^{L} [q_{v_i}(v_i; \phi_i^v) q_{\eta_i}(\eta_i; \phi_i^{\eta})] \prod_{i=1}^{L} q_{z_n}(z_n), \tag{10}$$

where $q_{v_i}(v_i; \phi_i^v) = p_v(v_i|\alpha)$ and $q_{\eta_i}(\eta_i; \phi_i^{\eta}) = p_{\eta}(\eta_i|\lambda)$, and $q_{z_n}(z_n)$ are discrete distributions of the component labels, so that the minimised free energy reads:

$$q_{z_n}(z_n = i) = \frac{\exp(S_{n,i})}{\sum_{j=1}^{\infty} \exp(S_{n,i})}, \tag{11}$$

where

$$S_{n,i} = E_{q_V}[\log p_z(z_n = i|V)] + E_{q_{\eta_i}}[\log p_x(x_n|\eta_i)], \tag{12}$$

$$E_{q_{v_i}}[\log v_i] = \Psi(\phi_{i,1}^v) - \Psi(\phi_{i,1}^v + \phi_{i,2}^v), \tag{13}$$

$$E_{q_{v_i}}[\log(1 - v_j)] = \Psi(\phi_{i,2}^v) - \Psi(\phi_{i,1}^v + \phi_{i,2}^v), \tag{14}$$

$$E_{q_{\eta_i}}[\log p_x(x_n|\eta_i)] = E_{q_{\eta_i}}[\eta_i]^T x_n - E_{q_{\eta_i}}[a(\eta_i)] \tag{15}$$

Then, the probability $p(x|X,\theta)$ can be approximated by:

$$p(x|X,\theta) = \sum_{i=1} E_{q_V}[\pi_i(V)]E_{q_{\eta_i}}[p_x(x|\eta_i)] + [1 - \sum_{i=1} E_{q_V}[\pi_i(V)]]E_{p_\eta}[p_x(x|\eta)] \tag{16}$$

This research example uses two Gaussian mixture models to model the feature distributions of the 'apnoea' and 'normal' segments. Therefore, two sets of features are given: $X^1 = \{x_i^1\}_{i=1}^{N^1}$ as 'apnoea' and $X^0 = \{x_i^0\}_{i=1}^{N^0}$ as 'normal.' The LLR is then computed for each segment by taking the difference between the log-likelihoods of the data under the 'apnoea' model and the 'normal' model based on the probability computation by equation (16). Then the decision [8] will be shown as

$$log\frac{p(x'|X^1)}{p(x'|X^0)} \geq c, \tag{17}$$

where c is a threshold that affects how well sensitivity and specificity are balanced. The whole framework is shown in Figure 4.

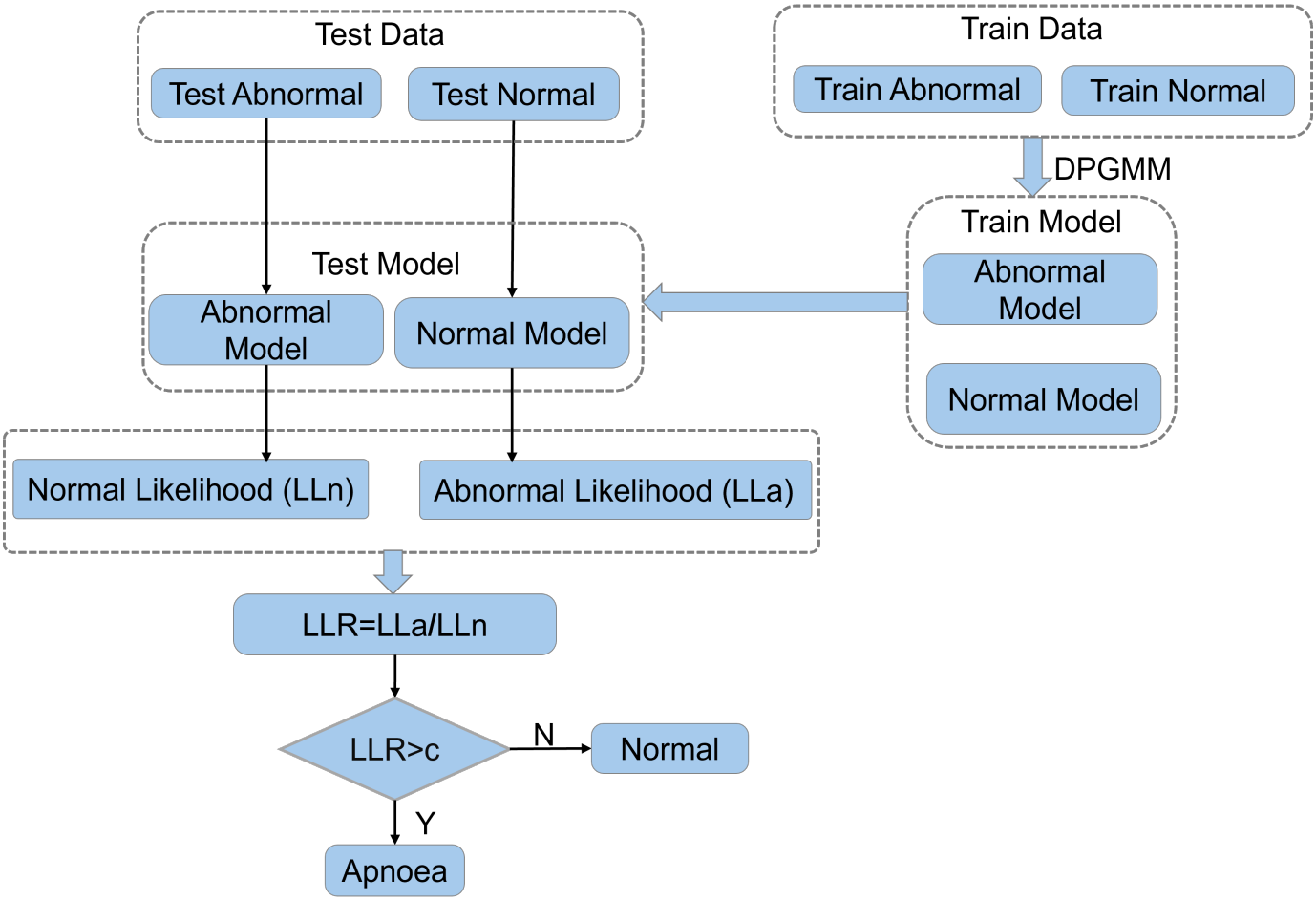

FIGURE 4: Framework for DPGMM to detect sleep apnoea

3.3. Fusion Techniques

Given that this study encompasses two types of signals, it is considered to consider integrate the signal features before inputting them into the classifier. In terms of the sequence of fusion and prediction, feature fusion is categorized into early fusion and late fusion [20]. Early fusion involves combining the features before they are input into the classifier, while late fusion entails inputting the features into the classifier independently and then fusing the prediction scores. Given the vast quantity of data and the need to manage computational complexity, this study selected concatenation fusion methods. Figure 5 shows the details of concatenation fusion used. Initially, all features are normalised to ensure uniformity. In Concatenation-Based Fusion, the extremities of two disparate signal features are linked to formulate a novel integrated feature. All the results will be compared and discussed in Section 4.

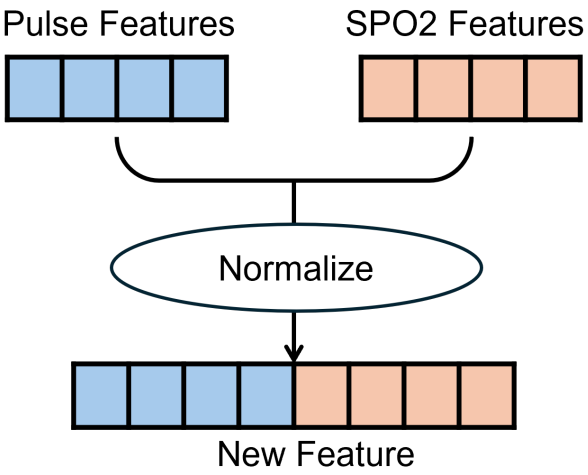

FIGURE 5: Concatenation fusion.

4. PERFORMANCE VALIDATION AND EVALUATION

4.1. Datasets

To enable near-real-time detection of SDB events, the SPO2 and pulse signal are systematically partitioned into overlapping subsequences, each of which has the same length. Multiple informative features are extracted from each subsequence for advanced analysis. The following several subsections will clarify the source of the experimental datasets and the specific methodologies applied for their preprocessing. Moreover, an explanation will be given about the features used in the classification process, which are defined within both the time and wavelet domains.

4.1.1. Sleep apnoea database

This study compared the detection results of three different public datasets. The first is the CHAT from the National Sleep Research Resource (NSRR) [21], and the second is from St. Vincent's University Hospital (St.Vincent) [22]. The third is the Apnoea-ECG Database (Apnoea-ECG) [23]. The whole dataset and the data used for the experiment are shown in Table 1. In the names on the left side of the table, '60' refers to a 60-second segment, while '10,' '20,' and '30' represent different overlaps between the segments, respectively. Here $ovlp$ denotes the overlap of segments. 'Label' is the label of the segments based on different overlaps. 0 is 'normal' and 1 is 'apnoea'. 'Count' means the number of 'normal' or 'apnoea' segments. Percentage refers to the proportion of the number of 'normal' or 'apnoea' segments to the experimental data under different overlap conditions.

	CHAT			St.Vincent			Apnoea-ECG		
	Label	Count	Percentage	Label	Count	Percentage	Label	Count	Percentage
60_10_ovlp	0	83826	93.41%	0	9094	75.94%	0	2044	60.89%
	1	5915	6.59%	1	2882	24.06%	1	1313	39.11%
60_20_ovlp	0	253720	94.47%	0	11674	78.02%	0	2036	60.99%
	1	14862	5.53%	1	3289	21.98%	1	1302	39.01%
60_30_ovlp	0	144223	95.83%	0	16137	80.86%	0	2030	61.29%
	1	6281	4.17%	1	3820	19.14%	1	1282	38.71%

TABLE 1: Three different datasets (CHAT, St Vincent and Apnoea-ECG) used in the performance validation

In these datasets, there are the St.Vincent dataset samples at 8Hz and the Apnoea-ECG at 100Hz, while the CHAT dataset has a variable sampling frequency. Figure 6 provides a statistical chart showing the number of files corresponding to each sampling frequency.

4.1.2. Data Preprocessing

Figure 6 illustrates the nonuniform sampling frequencies in the CHAT dataset, with a high count of files at 1 Hz and 2 Hz. Low sampling rates can lead to data being classified as noise, impacting further analyses, while very high rates may result in redundant information. Consequently, we excluded files with low (1 Hz and 2 Hz) and high (512 Hz) sampling frequencies from the CHAT dataset.

All data will be segmented into overlapping one-minute intervals for analysis to detect sleep disordered breathing (SDB) events. Segments containing SDB events are classified as 'apnoea', while those without respiratory disturbances are labelled 'normal'. A refined classification is utilised when an apnoea event extends across two consecutive segments. If respiratory disturbances occur but last less than five seconds in any segment, it is labelled 'normal', as the brief disturbance does not significantly alter the

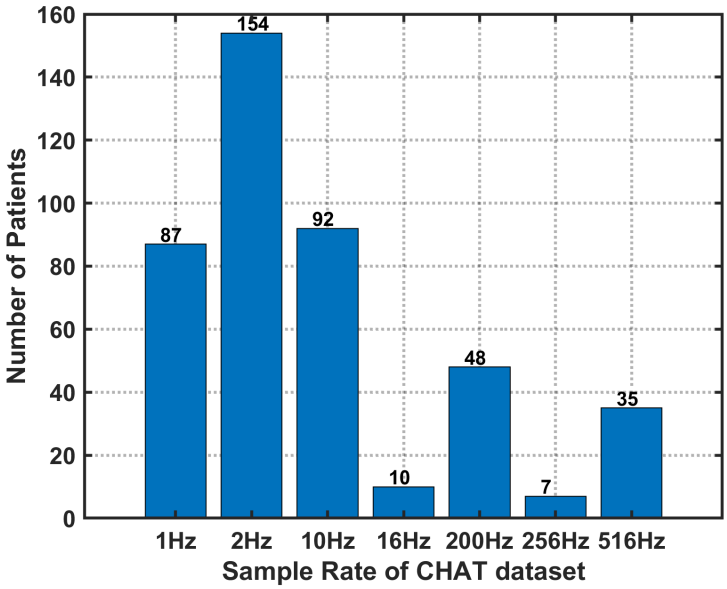

FIGURE 6: Histogram of the sampling frequency of patient data from CHAT dataset

overall respiratory pattern. Conversely, if disturbances exceed this duration in any segment, it is marked as 'apnoea', indicating a significant disruption in respiratory activity. This method ensures precise categorization of each minute based on the severity and duration of disruptions.

Insufficient contact with the pulse oximeter, often due to body movement at night, can generate artefacts. Artefacts are defined as pulse readings above 300 and below 50 and SPO2 below 50. Segments with these values are excluded from training and testing to ensure the reliability of the data.

4.2. Performance Criteria

For the RUSBoost model training setting, 'weak learners' is 100, 'number of ensemble learning cycles' is 1000 and the learn rate is 0.1. All the settings of DPGMM are default. The two publicly available databases are trained and tested independently with 10-fold cross-validation. In binary classification, the confusion matrix categorizes outcomes into four types [24]: true positives (TP), which are instances accurately identified as positive; false positives (FP), which denote negative instances mistakenly labelled as positive; true negatives (TN), which represent negative instances correctly classified as such; and false negatives (FN), which pertain to positive instances incorrectly classified as negative.

In the comparative analysis, accuracy (ACC), sensitivity (SEN), specificity (SPE), F_1 score, and Cohen's κ were utilised. The area under the Receiver Operating Characteristic (ROC) curve, which is the area under curve (AUC), is also used. These performance measures are defined as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}, \tag{18}$$

$$SEN = \frac{TP}{TP + FN}, \tag{19}$$

$$SPE = \frac{TN}{TN + FP}. \tag{20}$$

Equations (18) ,(19) ,(20) calculate accuracy, sensitivity, and specificity, which are critical metrics for assessing classifier performance. However, these metrics can be misleading in cases where there are significant class imbalances.

The F_1 score [25], a more robust metric, is defined as a weighted average of precision and recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{21}$$

$$Precision = \frac{TP}{TP + FP}, \tag{22}$$

$$Recall = \frac{TP}{TP + FN}. \tag{23}$$

Because sensitivity and recall are mathematically equivalent, the F_1 score in this paper utilises the term 'recall' to discuss the classifier's ability to correctly identify actual positive cases, which are apnoea, as well as actual negative cases, which are normal.

This difference is very important for the next steps of figuring out the F_1 score, because recall for 'apnoea' is the same as sensitivity and recall for 'normal' is the correct identification of cases that are not apnoea.

Cohen's κ [26] is a statistical coefficient designed to measure the consistency of two categorizing observers. It estimates the simple consistency between the model prediction and the actual label and corrects for chance consistency, making it more accurate and dependable than other metrics. When the model prediction and label match, the κ value is high, indicating consistency. The κ expression is as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{24}$$

$$p_o = ACC, \tag{25}$$

$$p_e = \sum_i p_{i+} * p_{+i}, \tag{26}$$

where p_{i+} and p_{+i} represent the i^{th} row probabilities and i^{th} column probabilities.

AUC, or Area under the curve [24], is ideal for assessing binary classification models on imbalanced datasets. It measures the area beneath the ROC curve, which plots the true positive rate against the false positive rate at various thresholds. AUC evaluates performance without bias from class distribution, making it crucial for datasets where one class predominates.

4.3. Results and Discussion

This subsection mainly presents the test results of single signal and feature fusion. Since there are many features and wavelets to compare, only some important results are shown. Because there is no pulse signal in the ECG database, the feature fusion results only list the results of the CHAT and St.Vincent databases.

4.3.1. Features visualisation

To evaluate the classification ability of the 18 extracted features, we generated 18 normalised histograms (see Figure 7) and box plots (see Figure 8) for the normal class and the apnoea class, respectively. Histograms show varying degrees of classification ability regarding shape, central tendency, and skewness [27]. The box plot further reveals the median, interquartile range (IQR) and outliers of the features in different categories from a statistical perspective [28]. This section takes SPO2 in the St. Vincent dataset as an example to show the distribution of each feature in the two categories. Table 2 gives the specific meaning of each feature name. The histograms from Figure 7 show that several features exhibit highly asymmetric distributions, including 'varad',

varad	Variation of level 4 detail coefficients
varbd	Variation of level 5 detail coefficients
rangead	Range of level 4 detail coefficients
rangebd	Range of level 5 detail coefficients
Powerad	Mean Energy of level 4 detail coefficients
Powerbd	Mean Energy of level 5 detail coefficients
maxad	Maximum level 4 detail coefficients
maxbd	Maximum level 5 detail coefficients
kursig	Kurtosis of signal
meansig	Mean of signal
varsig	Variation of signal
rangesig	Signal Range
CTM	Central Tendency measure of signal
minsig	Minimum signal
shannonEn	Shannon entropy of signal
TsallisEn	Tsallis entropy of signal
WavL2Num	Number of the large coefficients (level 4 detail coefficients)
WavL3Num	Number of the large coefficients (level 5 detail coefficients)

TABLE 2: Explanation of feature name

'varbd', 'rangebd', and 'Powerbd'. There is a significant shift in the peak positions between the two classes. For example, the normal samples exhibit a sharp peak close to zero, while the apnoea samples exhibit a wider right-skewed distribution, indicating that the variability and power-related measurements in the apnoea event are higher, which means that the SPO2 signal fluctuates more violently in the apnoea state, and abnormal oscillations or signal disturbances occur. Several features, including 'CTM', 'minsig', 'maxad', and 'rangesig', exhibit a significantly skewed distribution in the normal class with a single peak. In contrast, the distribution is wider and has two peaks in the apnoea class. Besides, they are distributed on both sides with less overlap, indicating that these features can effectively distinguish between normal and apnoea. In addition to practical distinguishing features, some features with similar distributions, such as 'ShannonEn', 'TsallisEn', 'WavL2Num', and 'WavL3Num', have large overlapping areas,

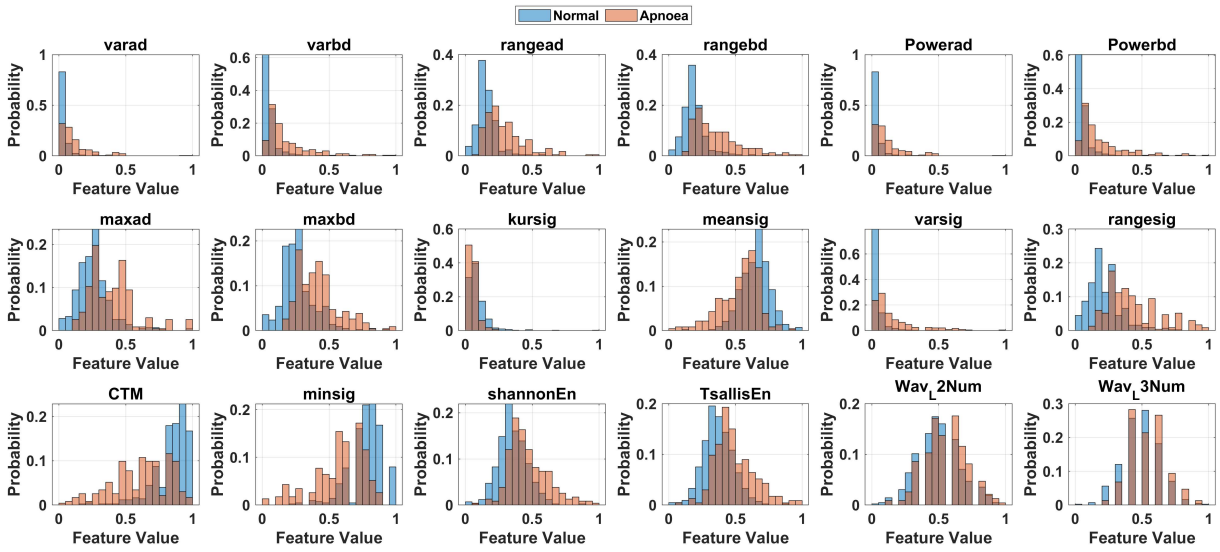

FIGURE 7: Features visualisation by histogram

indicating insufficient classification ability. However, the differences in the peak positions of the distributions allow these features to serve as auxiliary features for classification.

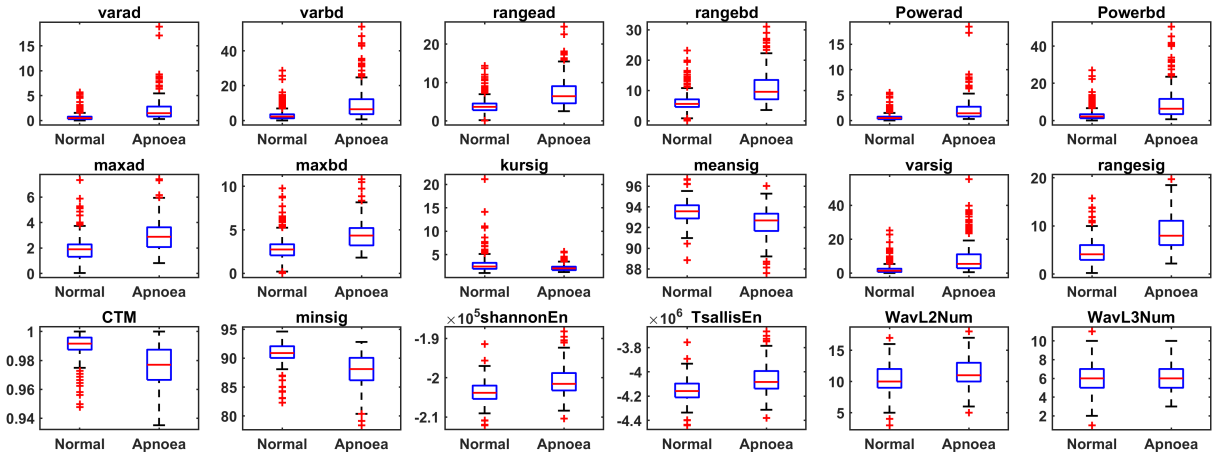

FIGURE 8: Features visualisation by boxplot

The histogram illustrates the distribution of features, while the box plot visually indicates the suitability of features for classification. The blue box in Figure 8 is the IQR. Features such as 'Powerbd', 'rangebd', and 'rangesig' show nonoverlapping IQR between classes, while the apnoea group consistently shows higher medians and wider distributions. 'CTM' and 'Meansig' also show a significant downward shift in the apnoea class, reflecting the loss of signal consistency during apnoea episodes. Non-overlapping IQR mean that the feature has very strong classification ability. In contrast, features such as 'WavL2Num' and 'WavL3Num' have overlapping medians and IQR, which means limited classification power when used alone. This is the same conclusion as the histogram.

As observed from Figures 7 and 8, features 'rangebd', 'powerbd', 'maxad', 'rangesig', 'CTM', and 'minsig' demonstrate strong discriminative capability. However, this does not imply that the remaining features are uninformative. Table 3 reports the classification performance obtained using either the whole feature set or only the aforementioned six features. The fact that feature selection does not improve classification accuracy suggests that the excluded features contribute complementary information, enhancing overall model performance.

	acc	sens	spec	F_1-N	F_1-A	κ
After feature select	74.76%	47.91%	81.14%	0.839	0.422	0.263
All feature select	84.32%	72.07%	87.12%	0.90	0.67	0.57

TABLE 3: Results of St. Vincent Database (SPO2) based on different features selected

4.3.2. Single signal results

This experiment evaluated and contrasted multiple algorithms using different training sets to examine the effects of data imbalance on training outcomes and to offer algorithmic recommendations for the medical field. Tables 4, 5, and 6 present the results of several machine learning techniques employing the identical wavelet (db1), segment duration (30-second overlap), and features. In this context, 'F_1-N' represents the F_1 score for the normal category classification, while 'F_1-A' signifies the F_1 score for the apnoea category classification. RUSBoost performs well on the St. Vincent dataset, attaining an acceptable F_1-A score and high sensitivity. The significant imbalance in the CHAT dataset leads to an F_1-A score of less than 0.5. The algorithms demonstrate exceptional performance on the ECG dataset, with a maximum accuracy of 97%. In conclusion, if the ratio of the 'apnoea' moments to the 'normal' moment is less than 1:3, the RUSBoost algorithm is advisable. For ratios over 1:3, the performance of different algorithms remains consistent, with SVM or Gaussian Naive Bayes being advisable.

db1_30_ovlp	acc	sens	spec	F_1-N	F_1-A	κ
Fine Tree	95.16%	95.86%	94.71%	0.960	0.939	0.899
Gaussian Naive Bayes	**97.31%**	**96.55%**	**97.80%**	**0.978**	**0.966**	**0.943**
Linear SVM	97.31%	95.17%	98.68%	0.978	0.965	0.943
Fine KNN	95.97%	93.79%	97.36%	0.967	0.948	0.915
Boosted Tree	96.77%	95.17%	97.80%	0.974	0.958	0.932
Bagged Tree	96.24%	93.79%	97.80%	0.969	0.951	0.920
Medium Neural Network	94.89%	91.72%	96.92%	0.959	0.933	0.892
RUSBoost	88.95%	75.47%	89.54%	0.94	0.36	0.32
DPGMM	94.20%	97.90%	91.78%	0.95	0.93	0.88

TABLE 4: Results of ECG Database (SPO2) based on different machine learning methods with the same wavelet, segment length, and features

db1_30_ovlp	acc	sens	spec	F_1-N	F_1-A	κ
Fine Tree	86.87%	58.38%	93.61%	0.920	0.630	0.551
Gaussian Naive Bayes	86.89%	52.62%	94.99%	0.921	0.605	0.529
Linear SVM	87.42%	46.07%	97.21%	0.926	0.584	0.516
Fine KNN	82.21%	54.97%	88.65%	0.890	0.542	0.432
Boosted Tree	87.77%	56.02%	95.29%	0.926	0.637	0.565
Bagged Tree	87.37%	56.02%	94.79%	0.924	0.629	0.555
Medium Neural Network	87.37%	56.02%	94.79%	0.924	0.629	0.555
RUSBoost	**85.86%**	**80.37%**	**87.17%**	**0.91**	**0.69**	**0.60**
DPGMM	84.51%	42.15%	94.54%	0.91	0.51	0.42

TABLE 5: Results of St.Vicent Database (SPO2) based on different machine learning methods with the same wavelet, segment length, and features

db1_30_ovlp	acc	sens	spec	F_1-N	F_1-A	κ
Fine Tree	95.90%	9.34%	99.69%	0.979	0.160	0.150
Gaussian Naive Bayes	92.30%	28.18%	95.10%	0.959	0.235	0.195
Linear SVM	95.81%	0.00%	100.00%	0.979	NaN	0.000
Fine KNN	93.78%	24.95	96.79	0.968	0.252	0.219
Boosted Tree	95.89%	8.09%	99.73%	0.979	0.142	0.133
Bagged Tree	95.90%	14.98%	99.44%	0.979	0.234	0.220
Medium Neural Network	95.96%	12.27%	99.62%	0.979	0.203	0.191
RUSBoost	**88.95%**	**75.47%**	**89.54%**	**0.94**	**0.36**	**0.32**
DPGMM	61.61%	85.42%	60.54%	0.75	0.16	0.09

TABLE 6: Results of CHAT Database (SPO2) based on different machine learning methods with the same wavelet, segment length, and features

Section 3.1.1 introduces an innovative feature termed 'Number of Large Coefficients'. This study aimed to validate the utility of this feature in model training. Employing the pulse signal from the St. Vincent dataset as a case study, models were systematically trained under various conditions: with and without resampling and incorporating the new feature. The results presented in Table 7 confirm the initial hypothesis that models trained without resampling attained greater accuracy. The accuracy of models trained without resampling was 2% to 3% higher than that of models trained with resampling. Furthermore, the accuracy of the model in disease identification was improved by approximately 2% as a result of the new feature's integration, which also increased training sensitivity.

db4_30_ovlp	acc	sens	spec	F_1-N	F_1-A	κ
no sample	73.93%	62.83%	76.57%	0.83	0.48	0.32
sample	71.24%	47.38%	76.91%	0.81	0.39	0.21
no resample, add new feature	**74.24%**	**64.92%**	**76.44%**	**0.83**	**0.49**	**0.33**
resample, add new feature	71.19%	47.38%	76.85%	0.81	0.39	0.21

TABLE 7: Results of St. Vincent's university hospital sleep apnoea database based on different features in same wavelet and segment length

In addition to comparing algorithms and datasets, this experiment further assesses the efficacy of different wavelet features in training models. The training results, employing the CHAT dataset for illustration, are detailed in Tables 8 and 9. The left column of each table specifies the wavelet type and segment length; for example, 'db1_10_ovlp' denotes the use of wavelet db1 with a 10-second overlap. According to the results, pulse data performs better with db3, whereas SPO2 data is best handled among the four wavelets under investigation with db1. These presentations emphasise the need to select specific wavelet features for different types of data in order to improve the efficacy of the model. Moreover, the best results are always obtained with a 30-second overlap when using identical wavelet settings.

	acc	sens	spec	F_1-N	F_1-A	κ
db1_10_ovlp	67.69%	69.15%	67.58%	0.80	0.22	0.12
db1_20_ovlp	69.64%	72.14%	69.50%	0.81	0.21	0.13
db1_30_ovlp	72.34%	73.50%	72.29%	0.83	0.18	0.12
db2_10_ovlp	68.18%	73.90%	67.78%	0.80	0.23	0.14
db2_20_ovlp	70.38%	72.29%	70.27%	0.82	0.21	0.13
db2_30_ovlp	72.19%	75.21%	72.05%	0.83	0.18	0.12
db3_10_ovlp	67.94%	72.84%	67.59%	0.80	0.23	0.13
db3_20_ovlp	70.58%	72.87%	70.45%	0.82	0.21	0.13
db3_30_ovlp	**72.41%**	**74.36%**	**72.33%**	**0.83**	**0.18**	**0.12**
db4_10_ovlp	67.94%	73.14%	67.57%	0.80	0.23	0.13
db4_20_ovlp	70.77%	72.43%	70.67%	0.82	0.21	0.13
db4_30_ovlp	72.49%	74.93%	72.38%	0.83	0.19	0.12

TABLE 8: Results of CHAT database (Pulse) based on different wavelet and segment length

	acc	sens	spec	F_1-N	F_1-A	κ
db1_10_ovlp	83.67%	65.65%	84.95%	0.91	0.35	0.28
db1_20_ovlp	84.93%	69.23%	85.85%	0.91	0.34	0.28
db1_30_ovlp	**89.97%**	**74.79%**	**90.63%**	**0.95**	**0.38**	**0.34**
db2_10_ovlp	83.98%	64.19%	85.37%	0.91	0.34	0.28
db2_20_ovlp	85.16%	66.13%	86.25%	0.92	0.33	0.27
db2_30_ovlp	89.82%	75.50%	90.44%	0.94	0.38	0.34
db3_10_ovlp	84.21%	64.19%	85.62%	0.91	0.35	0.28
db3_20_ovlp	85.17%	66.86%	86.23%	0.92	0.33	0.27
db3_30_ovlp	89.88%	75.07%	90.53%	0.94	0.38	0.34
db4_10_ovlp	84.04%	64.64%	85.40%	0.91	0.35	0.28
db4_20_ovlp	85.15%	66.86%	86.20%	0.92	0.33	0.27
db4_30_ovlp	89.89%	75.36%	90.53%	0.94	0.38	0.34

TABLE 9: Results of CHAT database (SPO2) based on different wavelet and wegment length

4.3.3. Feature fusion results

In this experiment, three types of early feature fusion, as described in Section 3.3, were employed. Using the CHAT database as an example for concatenation fusion, the performance results of the RUSBoost algorithm are presented in Table 10. From the table, wavelet db3 and 30-second overlap give the best performance. Among the assessed wavelets, db3 exhibited the best efficacy throughout the investigation. Table 11 compares the training results obtained from the St. Vincent and CHAT databases using the same feature processing method. The CHAT database demonstrates superior overall accuracy, whereas the St.Vincent database exhibits elevated F_1 scores and κ values. This is because the different imbalance ratios of the two data sets can impact model fit significantly. Figure 9 illustrates the associated ROC curve, demonstrating an AUC value of 0.923.

	acc	sens	spec	F_1-N	F_1-A	κ
db1_10_ovlp	88.25%	58.51%	90.35%	0.93	0.40	0.34
db1_20_ovlp	86.64%	68.69%	87.69%	0.93	0.36	0.31
db1_30_ovlp	91.90%	69.91%	92.86%	0.96	0.42	0.38
db2_10_ovlp	87.72%	58.57%	89.77%	0.93	0.39	0.33
db2_20_ovlp	88.75%	60.12%	90.39%	0.94	0.37	0.32
db2_30_ovlp	91.78%	70.20%	92.72%	0.96	0.42	0.38
db3_10_ovlp	87.54%	57.36%	89.66%	0.93	0.38	0.32
db3_20_ovlp	88.82%	61.00%	90.42%	0.94	0.37	0.32
db3_30_ovlp	**91.54%**	**72.06%**	**92.39%**	**0.95**	**0.42**	**0.38**
db4_10_ovlp	87.62%	55.99%	89.84%	0.93	0.37	0.31
db4_20_ovlp	89.33%	60.56%	90.98%	0.94	0.38	0.33
db4_30_ovlp	91.39%	72.06%	92.23%	0.95	0.41	0.37

TABLE 10: Results of CHAT database in concatenation features

	acc	sens	spec	F_1-N	F_1-A	κ
CHAT	91.54%	72.06%	92.39%	0.95	0.42	0.38
St. Vincent	86.72%	74.87%	89.52%	0.92	0.68	0.60

TABLE 11: Results of different database with the processing features

		acc	sens	spec	F_1-N	F_1-A	κ
CHAT	SPO2	86.32%	69.04%	87.33%	0.92	0.35	0.30
	Pulse	70.21%	**73.06%**	70.04%	0.82	0.21	0.13
	Fusion	**89.27%**	63.75%	**90.78%**	**0.94**	**0.39**	**0.34**
St Vincent	SPO2	84.32%	72.07%	87.12%	0.90	0.67	0.57
	Pulse	72.79%	60.47%	76.27%	0.81	0.49	0.31
	Fusion	**85.80%**	**74.45%**	**89.61%**	**0.91**	**0.69**	**0.60**

TABLE 12: Mean performance based on 157 patients of single signal and concatenation feature fusion

The average performance measures for each training dataset were calculated to assess the impact of feature fusion relative to nonfused features. The CHAT database exhibits an average accuracy of 86.3% for single signal classification, whereas fusion classification demonstrates an enhancement of 89.3% accuracy. In the St. Vincent database, single signal detection achieves 84.3% accuracy, while feature fusion classification enhances accuracy to 85.8%. This means that, even in instances of mild data imbalance, specificity, the metric for accurately identifying disease-free segments, is enhanced by 1.5%. In cases of substantial data imbalance, RUSBoost accuracy with integrated features is enhanced by 3% compared to single-signal features. Furthermore, data from Table 12 indicates that training data post-feature fusion fits more effectively with the model, as demonstrated by elevated F_1 scores and κ values.

5. CONCLUSION AND FUTURE WORK

This work proposes a machine learning framework for detecting sleep apnoea, focusing on addressing the classification of imbalanced medical data, which is especially well-suited for sleep apnoea diagnostics of children. We investigate several machine learning techniques across three differently imbalanced datasets. The results of the experiments show that statistical features based on Daubechies wavelets work well, and the RUSBoost algorithm performs better than other algorithms when the data category is imbalanced. Besides, with the CHAT dataset, the fusion of two single-signal features (pulse and oximetry data) resulted in a further improvement of 3% in average accuracy, underscoring the advantage of fused features over single-signal features. Moreover, this framework has achieved an accuracy of 91.54%, a sensitivity of 72.06%, and a specificity of 92.39% for noninvasive home screening, surpassing existing approaches.

Future work will explore other feature extraction techniques to enhance diagnostic abilities across various types of apnoea events and multiple types of data. Additionally, it can also be beneficial to investigate multiple classification models to deliver more detailed and tailored medical diagnosis solutions for various patient groups.

CONFLICTS OF INTEREST

The author declares that there are no conflicts of interest regarding the publication of this paper.

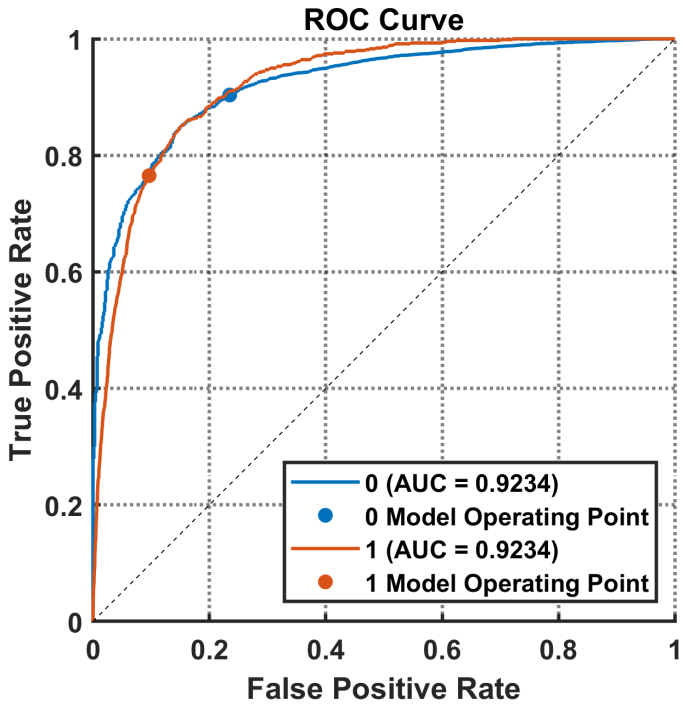

FIGURE 9: RUSBoost ROC curve of CHAT Database based on db3 and 30s overlap and fusion technique

ACKNOWLEDGEMENTS

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134,and UL1 RR024989). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). We express our gratitude to the UK Engineering and Physical Sciences Research Council (EPSRC) for their support through the following projects: EP/V026747/1 (Trustworthy Autonomous Systems Node in Resilience) and EP/T013265/1 (NSF-EPSRC: ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven Systems). We also acknowledge the support of the National Science Foundation (NSF) under Grant No. ECCS 1903466 for the ShiRAS project. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript (AAM) version arising from this work. We are grateful to the anonymous reviewers and to the Associate Editor for the constructive suggestions helping us to improve this work.

References

[1] A. N. Akansu and R. A. Haddad, "Chapter 6 - Wavelet Transform," in *Multiresolution Signal Decomposition*, 2nd ed., A. N. Akansu and R. A. Haddad, Eds. San Diego: Academic Press, 2001, pp. 391–442.

[2] K. Rajesh, R. Dhuli, and T. S. Kumar, "Obstructive sleep apnea detection using discrete wavelet transform-based statistical features," *Computers in Biology and Medicine*, vol. 130, p. 104199, 2021.

[3] C. Vonesch, T. Blu, and M. Unser, "Generalized Daubechies wavelet families," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4415–4429, 2007.

[4] G. Strang and V. Strela, "Orthogonal multiwavelets with vanishing moments," *Optical Engineering*, vol. 33, no. 7, pp. 2104 – 2107, 1994.

[5] R. S. Stanković and B. J. Falkowski, "The haar wavelet transform: its status and achievements," *Computers & Electrical Engineering*, vol. 29, no. 1, pp. 25–44, 2003.

[6] G. C. Gutiérrez-Tobal, M. L. Alonso-Álvarez, D. Álvarez, F. del Campo, J. Terán-Santos, and R. Hornero, "Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home," *Biomedical Signal Processing and Control*, vol. 18, pp. 401–407, 2015.

[7] A. Rihaczek, "Signal energy distribution in time and frequency," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 369–374, 1968.

[8] Z. Li, M. Arvaneh, H. E. Elphick, R. N. Kingshott, and L. S. Mihaylova, "A Dirichlet process mixture model for autonomous sleep apnea detection using oxygen saturation data," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–8.

[9] C. Alippi and M. Roveri, "An adaptive CUSUM-based test for signal change detection," in *2006 IEEE International Symposium on Circuits and Systems*, 2006, pp. 5752–5755.

[10] D. Álvarez, R. Hornero, J. Victor Marcos, and F. del Campo, "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2816–2824, 2010.

[11] R. A. Groeneveld and G. Meeden, "Measuring Skewness and Kurtosis," *Journal of the Royal Statistical Society Series D: The Statistician*, vol. 33, no. 4, pp. 391–399, 12 2018.

[12] J. Gao, J. Hu, and W.-w. Tung, "Entropy measures for biological signal analyses," *Nonlinear Dynamics*, vol. 68, pp. 431–444, 2012.

[13] J.-B. Wang, C.-A. Zou, and G.-H. Fu, "AWSMOTE: An SVM-based adaptive weighted SMOTE for class-imbalance learning," *Scientific Programming*, vol. 2021, pp. 1–18, 2021.

[14] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.

[15] C. Gao, N. Sang, and Q. Tang, "On selection and combination of weak learners in adaboost," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 991–1001, 2010.

[16] N. S. E. Mohd Noor, H. Ibrahim, M. H. Che Lah, and J. M. Abdullah, "Prediction of recovery from traumatic brain injury with EEG power spectrum in combination of independent component analysis and rusboost model," *BioMedInformatics*, vol. 2, no. 1, pp. 106–123, 2022.

[17] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121 – 143, 2006.

[18] K. Kurihara, M. Welling, and N. Vlassis, "Accelerated variational dirichlet process mixtures," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006.

[19] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[20] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[21] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 05 2018.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[23] T. Penzel, G. Moody, R. Mark, A. Goldberger, and J. Peter, "The apnea-ecg database," in *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, 2000, pp. 255–258.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition.

[25] P. Christen, D. J. Hand, and N. Kirielle, "A review of the F-Measure: Its history, properties, criticism, and alternatives," *ACM Comput. Surv.*, vol. 56, no. 3, Oct 2023.

[26] T. Wan, H. Jun, H. Zhang, W. Pan, and H. Hua, "Kappa coefficient: a popular measure of rater agreement," *Shanghai archives of psychiatry*, vol. 27, no. 1, p. 62, 2015.

[27] D. W. Scott, "Histogram," *WIREs Computational Statistics*, vol. 2, no. 1, pp. 44–48, 2010.

[28] L. Bessler, *Distributions, Histograms, Box Plots, and Alternative Tools*. Berkeley, CA: Apress, 2023, pp. 465–502.