Improving Non-invasive Aspiration Detection with Auxiliary Classifier Wasserstein Generative Adversarial Networks

Kechen Shu, Shitong Mao, James L. Coyle, and Ervin Sejdić, *Senior Member, IEEE*

***Abstract*— Aspiration is a serious complication of swallowing disorders. Adequate detection of aspiration is essential in dysphagia management and treatment. High-resolution cervical auscultation has been increasingly considered as a promising noninvasive swallowing screening tool and has inspired automatic diagnosis with advanced algorithms. The performance of such algorithms relies heavily on the amount of training data. However, the practical collection of cervical auscultation signal is an expensive and time-consuming process because of the clinical settings and trained experts needed for acquisition and interpretations. Furthermore, the relatively infrequent incidence of severe airway invasion during swallowing studies constrains the performance of machine learning models. Here, we produced supplementary training exemplars for desired class by capturing the underlying distribution of original cervical auscultation signal features using auxiliary classifier Wasserstein generative adversarial networks. A 10-fold subject cross-validation was conducted on 2079 sets of 36-dimensional signal features collected from 189 patients undergoing swallowing examinations. The proposed data augmentation outperforms basic data sampling, cost-sensitive learning and other generative models with significant enhancement. This demonstrates the remarkable potential of proposed network in improving classification performance using cervical auscultation signals and paves the way of developing accurate noninvasive swallowing evaluation in dysphagia care.**

***Index Terms*— Aspiration detection, cervical auscultations, data augmentation, deep learning, generative adversarial networks, swallowing accelerometry, swallowing vibrations.**

I. Introduction

SWALLOWING is a complex and coordinated biomechanical process that allows safe intake and transportation of substance while eating and drinking [1]. Neurological and other medical conditions or iatrogenic factors may lead to swallowing disorders also known as dysphagia [1], [2]. Aspiration, which defines the incursion of food or liquid into airway, is one of the most clinically significant symptoms of dysphagia, and may result in various unobservable effects to mortal consequences including airway obstruction or severe aspiration pneumonia [3]–[5]. Aspiration presence and severity is commonly measured by 8-point Penetration-Aspiration scale (PAs) that classifies possible observation of airway protection based on a videofluoroscopic swallow study (VFSS) [6]. According to its rating rules, PAs score of 1 represents the complete airway protection and score of 2 suggests mild, shallow and temporary airway invasion. PAs of scores of 3 and greater indicates deeper invasion of the airway and the presence of post-swallow airway residue [6]. The detailed description of PAs representations can be found in Appendix.

The VFSS is an x-ray imaging examination for clinical experts to visualize and evaluate swallowing physiology and airway protection [5]. However, the VFSS is not always feasible or desirable to many patients, since it exposes patients to radiation and is relatively expensive in both terms of imaging equipment and human resources

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD092239, while the data was collected under Award Number R01HD074819.

Kechen Shu and Shitong Mao are with the Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, 15260 USA (e-mail: kes247@pitt.edu; shm136@pitt.edu).

James L. Coyle is with Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, PA, 15260, USA, and also with Department of Otolaryngology, School of Medicine, University of Pittsburgh, PA, 15260, USA(e-mail: jcoyle@pitt,edu).

Ervin Sejdić is with the Edward S. Rogers Department of Electrical and Computer Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, Ontario, M5S 2E4, Canada, and also with north York General Hospital, Toronto, Ontario, M2K 1E1, Canada (e-mail: esejdic@ieee.org).

[7], [8]. Recently, high-resolution cervical auscultation (HRCA) is considered as a promising and alternative noninvasive swallowing screening tool [9]. The HRCA signals are collected by attaching a microphone and a tri-axial accelerometer to anterior neck. Previous studies have revealed the associations between HRCA characteristics with swallowing kinematic events including hyoid bone displacement, laryngeal vestibule opening, and upper esophageal sphincter opening [10]–[14]. Further deep learning algorithms established on HRCA signals have contributed to more systematic and reliable analysis comparable to human interpretations of VFSS such as swallow identification, upper esophageal sphincter opening segmentation, hyoid bone tracking, laryngeal closure duration estimation and automatic detection of aspiration [15]–[23]. Although the HRCA signals have demonstrated effectiveness in detecting a variety of swallowing kinematic events, the capability of identifying unsafe swallows has not been fully investigated.

One major challenge in developing such advanced diagnostic models is to collect sufficient HRCA signal samples accompanied by appropriate interpretations. In this study, laborious annotations on concurrent VFSS images by trained human judges were used as ground truth. However, the high financial and infrastructural requirements constrain the VFSS acquisition therefore provide an insufficient amount of exemplars of the ground truth reference data. Moreover, the need of specific training for VFSS annotations and the limited number of enrolled patients also leads to HRCA data paucity. Data augmentation, defined by the process of data oversampling by applying certain transformations to current real data samples, is an efficient technique to improve the generalization and prevent overfitting of deep learning models [24], [25]. Common data augmentation in time domain performs noise injection, temporal permutation, scaling, and cropping [24]. In feature spaces, manipulations including upsampling, perturbation, interpolation, and extrapolation also produces additional training samples. Nonetheless, such approaches may induce undesired distortion and alter the physiological representation when dealing with HRCA features [26].

Furthermore, the collected HRCA samples showed strong imbalance with limited occurrences of unsafe swallows. This class imbalance problem may also affect the diagnostic performance as most machine learning methods assumed balanced distribution by assigning equal misclassification cost for each class [27], [28]. A variety of solutions have been developed to handle imbalanced samples at both data and algorithm levels [29]. Straightforward data sampling techniques include random over-sampling and under-sampling strategies that enlarges minority class by adding duplicated samples, and decreases the size of majority class by reducing samples in random manner respectively [29], [30]. The synthetic minority over-sampling technique (SMOTE), which creates synthetic samples by exploring the features space, was considered fundamental to other synthetic sampling methods [27], [31]. However, the over-sampling techniques may cause overfitting, and under-sampling methods may discard useful information [29], [32]. Additionally, Cost-sensitive learning imposes different cost for misclassification of positive and negative instances in algorithm level but the implementation is not always feasible for all classification algorithms [31], [32].

Generative adversarial networks (GANs) are strong candidates for data augmentation and have been increasingly applied to imbalance learning [33]–[35]. GANs' potential in modeling underlying distribution of data allows production of infinite amounts of realistic samples [36]. Besides its convincing performance in synthetic image generation, recent studies attempted to synthesize sequential or high-level feature samples for various applications in medical domain [37]–[39].

In this work, we produced HRCA signal features using an auxiliary classifier Wasserstein GAN (AC-WGAN) under the hypothesis that incorporation of synthetic HRCA features will improve the performance for HRCA based aspiration detection. HRCA feature data are classified to either safe(healthy) or unsafe(abnormal) according to subjective PAs ratings during VFSS examinations. To address the issues of imbalanced class distribution and small sample size, we implement AC-WGAN to generate more HRCA feature samples for positive class. Additionally, the proposed AC-WGAN based data augmentation is compared to an ensemble of basic data augmentation and imbalance data approaches, including random over-sampling, random under-sampling, SMOTE, cost-sensitive learning, Wasserstein GAN (WGAN) and conditional WGAN (CWGAN).

II. Methods

A. Data acquisition and description

This study was under approval of Institutional Review Board of the University of Pittsburgh under the study number 19040040 and 19030185. A total number of 189 patients (115 males ages ranged between 23–94, 74 females ages ranged between 19–89) with suspected neurogenic dysphagia participated and provided informed consent. Each participant underwent swallowing assessment using VFSS conducted by speech language pathologists (SLPs) in the context of clinical standard rather than solely for research purposes.

During the VFSS, subjects were positioned laterally to a standard x-ray machine (Ultimax system, Toshiba, Tustin, CA) with a contact microphone (model C 411L, AKG, Vienna, Austria) and a tri-axial accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) attached to subjects' anterior neck as shown in Figure 1. The accelerometer was attached with surgical tape over the cricoid cartilage for optimal signal quality [42]. Its main axes were aligned parallel to the sagittal axis, longitudinal/vertical axis and frontal/horizontal axis of the neck. These axes were referred to anterior-posterior (AP), superior-inferior (SI) and medial-lateral (ML) directions respectively. The microphone was placed, slightly below the accelerometer, over the anterolateral side of larynx to avoid occlusion of swallowing mechanism in the VFSS [43], [44]. The video stream was captured by AccuStream Express HD (Foresight Imaging, Chelmsford, MA) at 30 frames per second (FPS). The audio and acceleration signals were band-pass filtered to 0.1-3000 Hz and amplified by gain of 10 using AC amplifier (model P55, Grass Technologies, Warwick, Rhode Island) before sampled to 20kHz via National Instrument 6210 DAQ. In the final stage, the video and signal recordings were acquired and synchronized using Labview Program Signal Express (National Instrument, Austin, Texas).

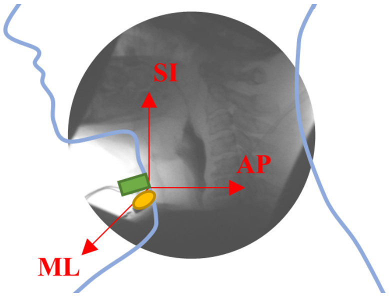


Fig. 1. Data collection setup: the video was captured on the lateral side of subject with microphone (indicated by yellow ellipse shape) and tri-axial accelerometer (green rectangle shape) attaching to the neck. Axes of accelerometer (in red arrows) align the anterior-posterior (AP), superior-inferior (SI) and medial-lateral (ML) directions of subject correspondingly.

B. Data pre-processing and feature representation

Segmentation of concurrent video and signal recordings to individual swallows was performed by a frame-by-frame analysis solely based on VFSS images. A complete swallow begins when the head of bolus reaches the ramus of mandible and ends when the hyoid returns to its lowest position after full clearance of bolus from pharynx [44]. 2079 swallows obtained from VFSS segmentation were considered in this study.

Segmented HRCA signals were firstly downsampled at 4kHz to overcome undesirable noise due to other physiological events or environmental sources while preserving most of swallowing-related information according to previous studies [45]–[47]. Inherent device noise from both accelerometer and microphone were initially characterized by fitting an autoregressive model to their zero-input responses. The autoregressive coefficients were then used to create finite impulse response filters to remove the device noise [47]. Additional low-frequency components caused by motion artifacts were eliminated from acceleration signals using fourth-order least-square spline models. Lastly, wavelet denoising was applied to reduce the effect of broadband noise through tenth order Meyer wavelet decomposition [42], [48].

A set of essential features which have been proven significant to swallowing pathology were extracted from preprocessed acceleration and audio signals [44], [46], [47]. As presented in Table I, these signal features involve analysis in time, frequency, information-theoretic, and time-frequency domains. In this study, the HRCA signal of each swallowing recording is composed of 4 channels: 3 vibrations and 1 audio. As we extracted 9 features from each channel, final dataset contains therefore 36 features.

C. Data labeling

Single segmented swallow videos were analyzed by trained judges. The presence and severity of aspiration was rated using the 8-point PAs based on the extent of airway invasion [6]. The annotations involved in swallowing segmentation and PAs rating maintained excellent intra-rater and inter-rater reliability by achieving high interclass correlation coefficients (> 0.99) on a randomly 10% of the selected swallow data. In this study, the swallows with PAs less or equal to 2 were considered safe and PAs greater or equal to 3 were defined as disordered/unsafe swallows. The aspiration detection was implemented by establishing a binary classification model to identify safe (negative) and unsafe (positive) swallows. Unsafe swallows are less frequently observed than safe swallows according to our dataset. The number of swallow samples in different PAs is presented in Table II. Examples of AP acceleration from both categories are shown in Figure 2.

D. GAN-based data augmentation

The framework of our HRCA data augmentation using AC-WGAN is illustrated in Figure 3. The synthetic HRCA features S were obtained by estimating the true distribution of original feature dataset R using proposed

TABLE I
FEATURES EXTRACTED ON BOTH VIBRATORY AND ACOUSTIC HRCA SIGNALS.

Domain	**Feature**	**Definition**
Time	Standard deviation Skewness Kurtosis	Variation of the signal around mean value Asymmetry of statistical distribution of the signal Sharpness of the peak of signal amplitude distribution
Information-theoretic	Lempel-Ziv complexity [49] Entropy rate	Regularity of the signal Randomness of the signal
Frequency	peak frequency Centroid frequency Band width	Frequency that corresponds to the maximal spectral energy Frequency that divides the spectrum into two equal parts Difference between the uppermost and lowermost frequencies of the signal spectrum
Time-frequency	Wavelet entropy [50]	Disordered/ordered behavior of the signal

TABLE II
DATA SUMMARY

PAs	1	2	3	4	5	6	7	8
# of samples	1026	675	173	60	30	51	27	37
	1701 (safe)		378 (unsafe)					

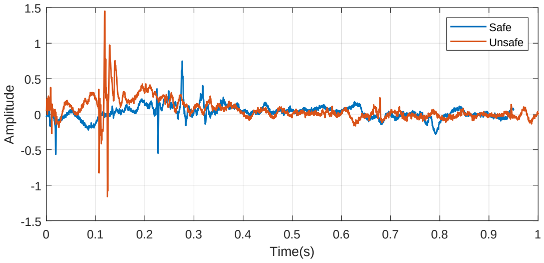


Fig. 2. HRCA signal examples from safe and unsafe instances

AC-WGAN model. By merging generated HRCA features to real ones, data augmented and balanced dataset $S \cup R$ is deployed to train the further aspiration detection model.

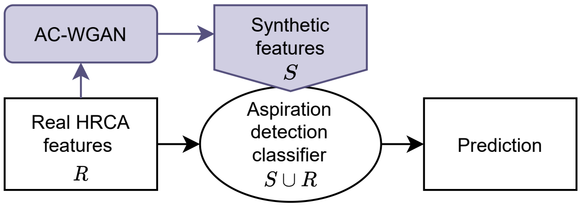


Fig. 3. Aspiration detection with AC-WGAN augmented HRCA data.

1) AC-WGAN: The fundamental structure of GAN consists of two independent components: a generator (denoted as G) and a discriminator (denoted as D) [36]. While the D attempts to distinguish real or generated samples, the G tends to fool the D by capturing the distribution of real samples (x) thus producing realistic samples ($\tilde{x}$) from noise input (z). The two networks are jointly trained to reach convergence. However, the vanilla GAN is prone to instability and mode collapse [51]. WGAN then employed Wasserstein divergence with better convergence and gradient penalty component was further applied to enforce the Lipschitz constraint of D in WGAN approach [52], [53].

To introduce the categorical information into synthetic data generation, Conditional GAN and CWGAN take supplementary class label (c) into both G and D structures [40], [54]. In our case, c refers to 0 for safe swallows and 1 for aspirated swallows.

And to further enforce production of distinguishable sample from different classes, CWGAN can be modified by reshaping D to output additional classification results without feeding the label information to it [41]. Thus, we implement auxiliary decoder(classifier) network embedded in D and constructed AC-WGAN as shown in Figure 4. In the AC-WGAN model, both label representation and noisy latent vector are fed into G to condition the generated samples. The D performs two tasks: 1. differentiate the generated data from real data; 2. classify both real and fake data into categories. Our proposed objective function of D and G can be expressed as follows:

$$\underset{\theta_D}{minL} = \underbrace{\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g,c\sim\mathbb{P}_c}\left[y|\tilde{x},c\right] - \mathbb{E}_{x\sim\mathbb{P}_r,c\sim\mathbb{P}_c}\left[y|x,c\right]}_{\text{Wasserstein loss}} \underbrace{+\lambda_{gp}\,\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}},c\sim\mathbb{P}_c}\left[\left(\left\|\nabla_{\hat{x}|c}(y|\hat{x},c)\right\|_2 - 1\right)^2\right]}_{\text{gradient penalty}} \underbrace{-\lambda_{ac}\,\mathbb{E}_{x\sim\mathbb{P}_r,c\sim\mathbb{P}_c}\left[\log(s|x,c)\right]}_{\text{auxiliary classification loss}} \tag{1}$$

$$\underset{\theta_G}{minL} = \underbrace{-\,\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g,c\sim\mathbb{P}_c}\left[y|\tilde{x},c\right]}_{\text{Wasserstein loss}} \underbrace{-\lambda_{ac}\,\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g,c\sim\mathbb{P}_c}\left[\log(s|\tilde{x},c)\right]}_{\text{auxiliary classification loss}} \tag{2}$$

Where θ_G and θ_D represent the trainable weights of G and D; $\mathbb{P}_{\hat{x}}$ stands for the distribution of sampled interpolation between real data distribution $\mathbb{P}_r$ and generated distribution $\mathbb{P}_g$.

In (1), Wasserstein loss and gradient penalty term are optimized on both real and fake samples, whereas the classification loss is optimized solely on real data samples by maximizing the log-likelihood of correct class prediction. As for (2), the auxiliary classification loss, that depends on generated samples, makes G more sensitive to the label prediction s (as indicated in Figure 4) and thus drives G to generate category-specified data samples according to the input condition c.

The constant of gradient penalty λ_{gp} is set to 10 as in the original work [53]. The constant of auxiliary classification loss λ_{ac} is updated to 30 percent of current Wasserstein loss of D for each training iteration [55].

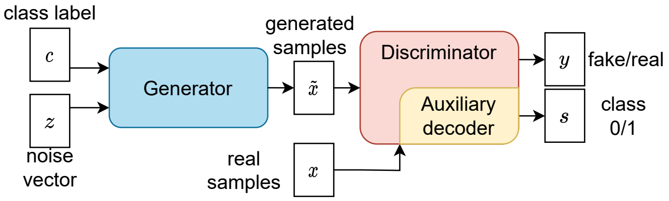


Fig. 4. Structure of AC-WGAN.

2) Implementation setup: Each HRCA sample have 36 dimensions. The outputs of both auxiliary decoder component of AC-WGAN (s in Figure 4) and aspiration classifier model (as illustrated in Figure 3) are probabilities representing either unsafe or safe swallows. G and D were both neural networks with fully connected layers. These two networks were constructed in nearly symmetric way except D divides into two branches correspond to fake/real identification and unsafe/safe classification in the output. The design of generator and discriminator have been studied with different depth of networks and various number of neurons in each layer. A final architecture of AC-WGAN generator and discriminator, as shown in Figure 5, was decided with optimal Maximum mean discrepancy (MMD) values and minimized network complexity.

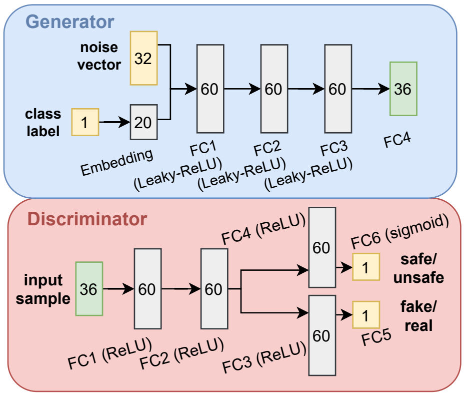


Fig. 5. Network architectures of AC-WGAN (FC stands for fully connected layers)

The generative network was optimized by RMSprop optimizer as suggested by the work [53]. The learning rate was set to 1×10^{-4} for both generator and discriminator and the weights of discriminator is updated for 8 times whereas generator is trained for 1 step during each training iteration.

E. Aspiration detection classifier

To predict aspiration from HRCA features, we implemented several machine learning algorithms including Support Vector Machine (SVM), K-means, Naive Bayes, and Artificial Neural Network (ANN) [23]. For SVM, a radial basis function was used as kernel, in which the coefficient was set to $1/\big(n_{features} * Var(x)\big)$ where the number of features equals to 36. The K-means was applied to form 2 centroids for clustering and the Naive bayes algorithm assumed Gaussian likelihood function. The ANN was composed of 3 fully connected layers with 60, 40, and 1 neurons. All layers, except the last one, were constrained by dropout (0.3 dropout rate), L1 and L2 weight regularization with a regulation values of 1×10^{-5} and 1×10^{-4} to avoid overfitting. The model was then trained to optimize binary cross entropy loss by an Adam optimizer, and the learning rate was set to 2×10^{-4}. The training process of the ANN classifier model was stopped when the improvement of training loss did not exceed 1×10^{-4} over the last 10 epochs. The hyper-parameter tuning of all the classifiers were performed according to achieve optimal classification performance on validation dataset for each data augmentation approaches.

F. Training and testing set

In this study, a total of 2079 swallow samples were collected, and a 10-fold subject cross-validation was applied. In detail, the subjects were randomly divided into 10 groups to include approximately 10% swallow samples for each group. Each group contains roughly 208 swallowing trials (range 199-233) from 18.9 (range 16-21) different subjects. For each fold of the cross-validation, One of the 10 groups was used for testing whereas the rest (approximately 1871 samples) was for training purpose. This procedure was repeated 10 times until each group has been treated as validation set for once. This subject based cross validation avoids intra-subject dependencies between training/testing sets and only performs analysis on swallow data from unseen subjects.

G. Evaluation metrics

1) MMD: The MMD that estimates the difference between two distributions from two set of samples over

Kernel Hilbert space has been widely used in GANs evaluation [39], [56]. The general value of MMD ranges between zero and one, where zero represents complete equality between two samples and one correspond to minimum similarity. The MMD is defined as follow:

$$MMD^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq 1}^{n} K\left(x_i, x_j\right) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} K\left(x_i, y_j\right) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j\neq 1}^{m} K\left(y_i, y_j\right) \tag{3}$$

We used Gaussian RBF kernel expressed by: $K\left(x_1, x_2\right) = \sum_{j=1}^{k} \exp -\alpha_j \left\|x_1 - x_2\right\|^2$ Where the bandwidth α equals the median of pairwise distance between two set of samples [56], [57].

The MMD was computed during training cycle between testing samples and generated samples of same size to evaluate synthesized data quality. We also employed MMD to compare interclass similarity between fake samples for measurement of data discriminability.

2) Classification evaluation: The classification performance of aspiration detection model is estimated by a set of metrics: accuracy, sensitivity, specificity, F1 score and Matthews correlation coefficients (MCC), of which the mathematical definitions were shown in Equation 4. Among these metrics, F1 score and MCC are particular meaningful when imbalanced dataset was considered [58]. The MCC value ranges from -1 to 1. A value of -1 corresponds to complete misclassification and +1 signifies perfect predictions [59].

$$\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \\
Sensitivity &= \frac{TP}{TP+FN} \times 100\% \\
Specificity &= \frac{TN}{TN+FP} \times 100\% \\
F_1\ score &= \frac{TP}{TP+0.5*(FP+FN)} \times 100\% \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}
\end{aligned} \tag{4}$$

Where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative predictions correspondingly.

H. Experiment baseline

To further investigate the efficacy of AC-WGAN, we conducted comparative experiments with other approaches in imbalance learning: random over-sampling, random under-sampling, SMOTE and cost-sensitive learning. The data augmentation performance of the AC-WGAN was further compared to other state-of-the-art GAN based methods including WGAN and CW-GAN as described in [54]. Two independent WGANs were established to generate safe and unsafe samples respectively while a single CWGAN and AC-WGAN produced features from both classes. All GAN models were implemented with similar structures.

III. Results

A. Performance of proposed AC-WGAN

In this section, we performed HRCA feature generation using proposed AC-WGAN framework. During the training process of AC-WGAN, the negative critic loss (negative Wasserstein divergence) and auxiliary classification loss rapidly converged towards their minimums (around 0.4 and 1.0 respectively). The MMD between testing samples and synthetic samples converged to a small value (around 0) as well. The mean values of training losses and MMD throughout 10-fold patient cross-validation were shown in Figure 6, in which the shaded area refers to the standard deviation. The optimization process indicates that the additional auxiliary decoder component did not affect the training stability.

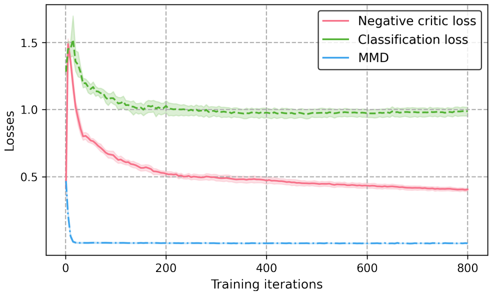


Fig. 6. Negative critic loss, auxiliary classification loss, and MMD values during AC-WGAN training.

We discovered that the AC-WGAN generated HRCA feature samples resemble the true underlying distribution of real features by comparing 200 random synthetic features with same amount of real ones and as shown in Figure 7. The synthetic HRCA features demonstrated diversity and the mode-collapse problem which is a critical issue in GAN training did not occur in our experiments. Both sets of samples showed consistency in median values and most synthetic features exhibited greater variances from larger interquartile ranges. An analysis of feature importance were applied on practical collected HRCA data and AC-WGAN generated samples using Random Forest algorithm, as shown in Figure 7. The original features presented relatively similar importance, while certain features gained more importance with AC-WGAN augmented data. These most influential features correspond to kurtosis and peak frequency of AP vibration, centroid frequency and peak frequency of SI vibration and kurtosis of ML acceleration respectively.

To visually examine the characteristics of generated samples by AC-WGAN, we plotted 200 real and synthetic HRCA features by t-SNE algorithm, as shown in Figure 8 [60]. The generated features shared similar distributions with real samples from same classes (Figure 8b and 8c). Albeit original samples suffered from overlapping distributions as shown in Figure 8a, the generated signal features showed significant distinction from different classes (Figure 8d).

The diversity and representativeness of generated samples were then further examined by computing MMD over 200 real and generated samples from safe and unsafe classes, as shown in Figure 9. WGAN generated more distinct features from different classes than AC-WGAN as greater interclass MMD values suggested. However, compared with WGAN, the intraclass MMD achieved higher values, indicating that the WGAN generated features less resembled the real ones. Since safe and unsafe samples were individually modeled by two WGANs, only a subset of training data were fed into each WGAN, which might cause WGAN synthetic data less representative than conditional GANs. In comparison to CWGAN, synthetic samples generated by AC-WGAN exhibited more realistic characteristics from lower intraclass MMD between real and fake data in both positive and negative classes (filled in green in Figure 9). In addition, greater interclass MMD between synthetic safe and unsafe swallows (filled in red) suggested more distinguishable samples were generated by AC-WGAN.

B. Effect of data augmentation on aspiration detection

Aspiration detection was carried out by 10-fold patient cross-validation with various GAN based data augmentation and other basic data sampling and learning techniques. The classification effectiveness is evaluated by accuracy, sensitivity, specificity, F1-score and MCC metrics as shown in Table III. Averaged number of training data are provided for each experiments. For GAN based methods, the number of data points were determined by the optimal classification performances. Four common types of classifiers: Naive Bayes, K-means, SVM, and ANN were employed as described in section II-E. Among the four classifiers, only SVM and ANN are applicable for cost-sensitive learning.

A set of paired t-test were conducted to compare classification performance on all classifiers using different data sampling and augmentation techniques and the results were listed in table IV. The testing hypothesis was that the first-listed method outperforms the second. According to the table IV, All techniques except AC-WGAN significantly reduced overall accuracy and specificity. The sensitivity and F1-score improved remarkably but the MCC rarely changed in these experiments. AC-WGAN is the only method that boosted sensitivity, F1-score and MCC without affecting overall accuracy and specificity notably.

Among common data imbalance methods, random over-sampling techniques outperforms the random over-sampling and SMOTE with greater accuracy, specificity and MCC despite of the decreased sensitivity. The over-sampling experiment also led to better accuracy and specificity compared to cost-sensitive learning on SVM and ANN models.

The effectiveness of AC-WGAN as an over-sampler was then compared to the random over-sampler and other GAN-based methods. AC-WGAN defeats the random over-sampling with higher sensitivity, F1-score and MCC. It also outperforms significantly WGAN in all metrics and CWGAN in all aspects other than sensitivity.

The statistical analysis was further applied on paired classification performance before and after the AC-WGAN augmentation on classifier basis as shown in last lines in Table IV. The accuracy of Naive Bayes was reduced while the sensitivity and F1-score had been boosted by AC-WGAN. K-means resulted in increased specificity and decreased sensitivity. As for the SVM and ANN classifiers, significant improvements on sensitivity, F1 score and MCC were discovered. However, the specificity has been degraded in ANN after the implementation of AC-WGAN.

IV. Discussion

In this work, we propose the AC-WGAN architecture to conduct HRCA feature augmentation and thus improved the overall aspiration classification accuracy. The original HRCA data present limited sample size and contains much more healthy instances than abnormal ones. This data imbalance issue results in strongly biased classification outcomes. The AC-WGAN, derived from CWGAN and AC-GAN, produced class specific synthetic samples and effectively enlarged training set for aspiration detection. The auxiliary classification loss component in AC-WGAN objective function enforced generation of representative and diverse synthetic data for different classes. Both generator and discriminator of AC-WGAN consist of fully connected layers and the time complexity of these networks are proportional to number of weights, number of iterations and training

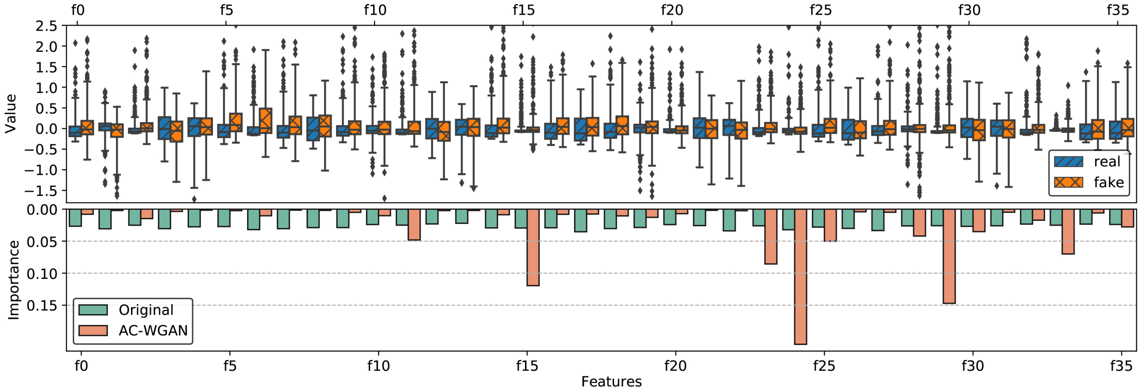
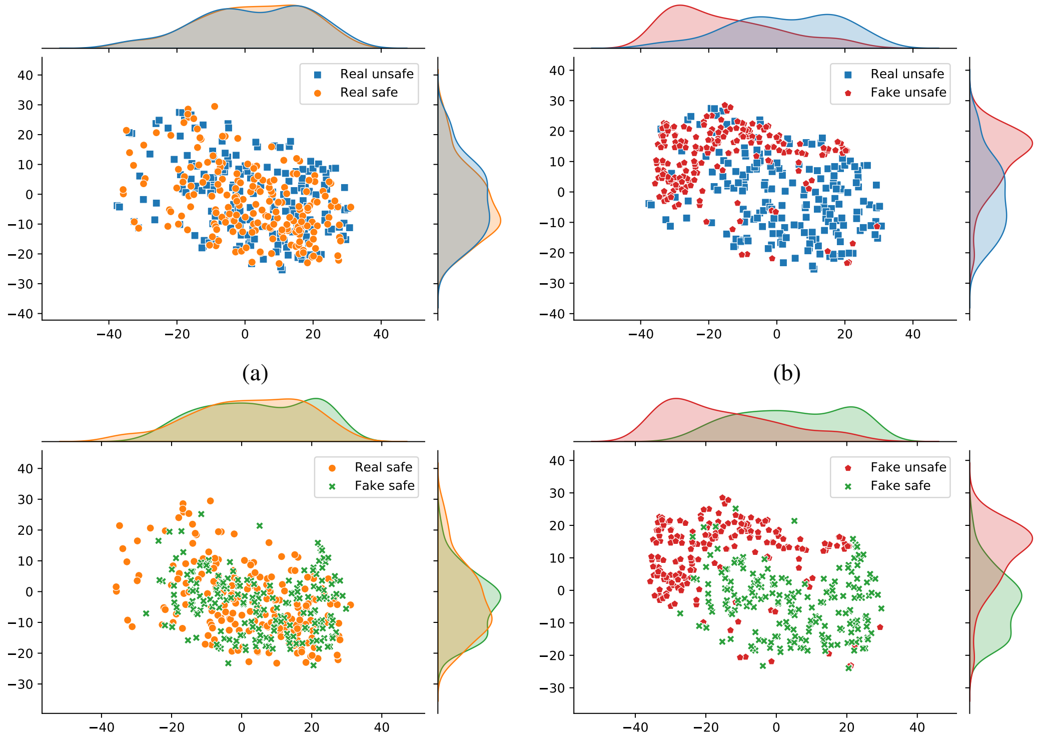


Fig. 7. Up: Box plots of real(blue) and generated(orange) features which has been previously defined in Table I; Down: Feature importance analysis using Random Forest before(green) and after(red) AC-WGAN data augmentation



(a) (b)

(c) (d)

Fig. 8. t-SNE visualization of real and AC-WGAN generated HRCA features: (a) shows safe and unsafe swallows from the original real data, (b), (c) compares intra-class samples from real and generated datasets, and (d) presents synthetic samples for different classes.

WGAN	Real safe	Fake unsafe	**CWGAN**	Real safe	Fake unsafe	**AC-WGAN**	Real safe	Fake unsafe
Real unsafe		3.42×10^{-3}	Real unsafe		3.57×10^{-3}	Real unsafe		2.81×10^{-3}
Fake safe	5.34×10^{-3}	9.10×10^{-3}	Fake safe	3.92×10^{-3}	0.88×10^{-3}	Fake safe	2.36×10^{-3}	6.89×10^{-3}

Fig. 9. Average MMD between real and synthetic samples generated by (left): WGAN, (center): CWGAN, and (right): AC-WGAN.

TABLE III
CLASSIFICATION RESULTS FOR ASPIRATED SWALLOW DETECTION.

Methods	None	Under-sampling	Over-sampling	SMOTE	Cost-sensitive	WGAN	CWGAN	AC-WGAN
# of data points	1871	680.4	3061.8	1871	1871	4×10^4	4×10^4	2×10^5
Classifier: Naive Bayes								
Accuracy (%)	74.47	64.91	71.42	33.97		64.92	67.58	66.38
Sensitivity (%)	12.88	24.79	19.26	67.92		30.26	23.95	39.03
Specificity (%)	85.39	72.34	80.37	26.03		70.80	77.27	74.60
F1-score (%)	12.79	19.70	17.89	26.62		22.38	20.53	22.02
MCC	-0.0065	-0.0129	0.0019	-0.0508		0.0158	0.0123	0.0324
Classifier: K-means								
Accuracy (%)	70.32	71.66	71.17	69.74		69.89	68.35	72.94
Sensitivity (%)	17.64	15.00	15.36	17.68		18.33	22.25	12.40
Specificity (%)	82.06	83.99	83.40	80.94		81.31	78.60	86.41
F1-score (%)	17.46	15.51	15.72	17.06		17.31	19.85	13.24
MCC	0.0035	-0.0076	-0.0093	-0.0067		-0.0011	0.0128	-0.0009
Classifier: SVM								
Accuracy (%)	77.19	53.64	64.22	59.06	61.80	69.67	68.76	75.02
Sensitivity (%)	10.17	44.02	29.72	30.04	33.02	13.32	19.91	21.71
Specificity (%)	93.21	55.67	71.82	65.56	68.17	82.28	78.46	86.84
F1-score (%)	13.13	24.72	22.72	20.63	23.03	13.12	17.76	**22.83**
MCC	0.0615	0.0315	0.0138	-0.0343	0.0077	-0.0356	-0.0133	**0.0938**
Classifier: ANN								
Accuracy (%)	69.20	51.09	70.15	67.15	69.70	68.91	60.91	71.39
Sensitivity (%)	17.07	43.92	20.87	20.5	18.93	18.59	28.83	32.84
Specificity (%)	80.67	52.13	80.9	77.70	81.00	79.97	67.51	79.78
F1-score (%)	16.46	24.12	19.52	18.04	17.78	17.82	20.31	**28.75**
MCC	-0.0207	-0.0306	0.0157	-0.0137	-0.0005	-0.0079	-0.0318	**0.1171**

TABLE IV
P VALUES OF PAIRED T-TEST ACROSS 10-FOLD VALIDATIONS BETWEEN DATA SAMPLING METHODS

Comparison	Accuracy	Sensitivity	Specificity	F1-score	MCC
Under-sampling & None	1	< 0.0001	1	0.0015	0.869
Over-sampling & None	0.9735	0.0094	0.9822	0.0215	0.5802
SMOTE & None	1	< 0.0001	1	0.0018	0.9703
Cost-sensitive & None(SVM, ANN)	0.9992	0.0003	0.9996	0.0021	0.7958
WGAN & None	0.9939	0.0122	0.9929	0.0496	0.7908
CWGAN & None	0.9998	< 0.0001	0.9999	0.0033	0.7608
AC-WGAN & None	0.8284	0.0003	0.9245	0.0043	0.0110
Over-sampling & Under-sampling	< 0.0001	1	< 0.0001	0.9874	0.0384
Over-sampling & SMOTE	< 0.0001	0.9989	< 0.0001	0.9069	0.0004
Over-sampling(SVM, ANN) & Cost-sensitive	0.0466	0.6293	0.0823	0.3113	0.2368
AC-WGAN & Oversampling	0.0624	0.01644	0.1171	0.0384	< 0.0001
AC-WGAN & WGAN	0.0029	0.0774	0.0393	0.0095	0.0002
AC-WGAN & CWGAN	0.0001	0.4417	0.0393	0.0922	0.0003
AC-WGAN(Naive Bayes) & None(Naive Bayes)	0.9822	0.0496	0.9147	0.0274	0.1729
AC-WGAN(K-means) & None(K-means)	0.0652	0.9538	0.0319	0.9087	0.6586
AC-WGAN(SVM) & None(SVM)	0.1546	0.0003	0.6248	0.0003	0.0006
AC-WGAN(ANN) & None(ANN)	0.9229	0.0021	0.9979	0.0069	0.0485

data size. Therefore, n_{epoch} training epochs of AC-WGAN has $\mathcal{O}(n_{epoch} * n_{data} * (8 * n_{\theta_D} + n_{\theta_G}))$ time complexity, where n_{data} equals to 2×10^5, n_{θ_D} and n_{θ_G} are approximately 12.5k and 13.1k respectively. Adding more neurons or layers in AC-WGAN, as well as increasing the number of generated samples into training set, would remarkably raise the time complexity of the algorithms without further improvement on detection performance.

The t-SNE visualizations and MMD evaluations, as illustrated in Figure 8 and Figure 9, suggest that AC-WGAN produced more separable and realistic HRCA samples compared to CWGAN. This implies that encouraging discriminability and diversity in GAN-based algorithms is constructive for representative generated samples [26], [41].

The effect of all data sampling methods and GAN-based augmentation experiments was analyzed by performing aspiration detection on resampled/generated training data. Our baseline experiments on real dataset showed insufficient classification outcome due to the imbalance distribution of the swallowing data [23]. When the safe data were under-sampled, both SVM and ANN classifiers showed poor accuracies, which demonstrate that the aspiration detection are sensitive to limited data size. The random over-sampling methods resulted in better classification performance than SMOTE which may indicate that the linear combination of neighboring unsafe samples does not represent validated distribution of HRCA features. Cost sensitive learning had similar outcome to over-sampling as both methods forced better prediction on minority class by assigning higher weights to these samples. Although WGAN and CWGAN oversampler succeeded in improving the sensitivity and F1 score for all classifiers, the accuracy and specificity were significantly degraded and the MCC value remains rarely increased. This may caused by the less representative and separable features generated by both models than AC-WGAN as illustrated by Figure 9.

Overall performance of the four common classifiers have been significantly improved by inducing AC-WGAN synthesized samples. The AC-WGAN outperforms other data sampling methods in almost all the evaluation metrics. While considering AC-WGAN augmentation influence on each classifier, only Naive Bayes classifier resulted in significantly decreased accuracy after data augmentation as shown in Table IV. For K-means classifier, none of the accuracy, F1 score and MCC had significantly improved. This suggests that the associations between augmented multi-dimensional HRCA signal features could be highly nonlinear and more sophisticated methods are required. Therefore, both the ANN and SVM model largely improved the classifications performances with boosted F1 score and MCC. Meanwhile, if the classifiers were trained solely on real data, the hidden complexity of the data may increase the burden of the classifiers when exploring mapping functions between inputs and outputs. In contrast, AC-WGAN helped capturing more underlying pattern of safe and unsafe feature distributions and leverages classifiers by enriching prior knowledge from original HRCA feature statistics.

The proposed AC-WGAN in data augmentation have proven considerable boost in aspiration detection performance. However, more experiments on general image or signal dataset are needed to validate the model. In current study, the sensitivity and specificity of aspiration detection with AC-WGAN data augmentation are not sufficient for practical swallow screening. Since the HRCA features were selected based on previous statistical analysis. Further studies would focus on alternative feature selection using advanced methods such as autoencoder to achieve better results. Moreover, only dysphagia suspected population was involved in this study, including more swallowing data from healthy subjects may produce more generalized samples and better explore the predictive relation between HRCA signal features and aspiration. AC-WGAN may also help to develop more robust predictive models for other swallowing related classification problems and provide accurate assistance of noninvasive swallowing assessment.

V. Conclusion

In this paper, a WGAN with auxiliary classifier (AC-WGAN) was proposed to improve the noninvasive aspiration detection on imbalanced HRCA dataset. The AC-WGAN is trained by optimizing a combination of Wasserstein losses and auxiliary classification loss, and therefore forcing more distinguishable and representative sample generation. Inducing the AC-WGAN synthetic samples to training data significantly improves the classification performance. The ANN-based model trained on augmented dataset achieved the best aspiration detection. Our findings further demonstrate HRCA's potential in dysphagia assessment and contribute to developing HRCA diagnosis adjunctive to instrumental swallowing evaluation.

Appendix

A detailed description of 8-point PAs decision is presented in Table V according to original publication.

References

[1] A. J. Miller, "The neurobiology of swallowing and dysphagia," *Developmental Disabilities Research Reviews*, vol. 14, no. 2, pp. 77–86, Jul. 2008.

[2] S. R. Achem and K. R. DeVault, "Dysphagia in aging," *Journal of Clinical Gastroenterology*, vol. 39, no. 5, pp. 357–371, 2005.

[3] S. Singh and S. Hamdy, "Dysphagia in stroke patients," *Postgraduate Medical Journal*, vol. 82, no. 968, pp. 383–391, Jun. 2006.

[4] J. Iruthayarajah, M. Saikaley, P. W. West, N. Foley, R. Martino, M. Richardson, R. Orenczuk, and R. Teasell, "Dysphagia and aspiration following stroke," in *Evidence-Based Review of Stroke Rehabilitation*, London, Ontario, Canada, 2018.

[5] K. Matsuo and J. Palmer, "Anatomy and physiology of feeding and swallowing: Normal and abnormal," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 19, no. 4, pp. 691–707, Nov. 2008.

[6] J. Rosenbek, J. Robbins, E. Roecker, J. Coyle, and J. Wood, "A penetration-aspiration scale," *Dysphagia*, vol. 11, pp. 93–8, Mar. 1996.

[7] R. D. Wilson and E. C. Howe, "A cost-effectiveness analysis of screening methods for dysphagia after stroke," *PM&R*, vol. 4, no. 4, pp. 273–282, Apr. 2012.

TABLE V
8-POINT PAS REPRESENTATION [6].

PAs	description
1	Material does not enter the airway
2	Material enters the airway, remains above the vocal folds, and is ejected from the airway
3	Material enters the airway, remains above the vocal folds, and is not ejected from the airway
4	Material enters the airway, contacts the vocal folds, and is ejected from the airway
5	Material enters the airway, contacts the vocal folds, and is not ejected from the airway
6	Material enters the airway, passes below the vocal folds, and is ejected into the larynx or out of the airway
7	Material enters the airway, passes below the vocal folds, and is not ejected from the trachea despite effort
8	Material enters the airway, passes below the vocal folds, and no effort is made to eject

[8] C. Donohue, Y. Khalifa, S. Perera, E. Sejdić, and J. L. Coyle, “How closely do machine ratings of duration of UES opening during videofluoroscopy approximate clinician ratings using temporal kinematic analyses and the MBSImP?” *Dysphagia*, Sep. 2020.

[9] A. K. Joshua M. Dudik and, J. L. Coyle, and E. Sejdić, “A statistical analysis of cervical auscultation signals from adults with unsafe airway protection,” *Journal of NeuroEngineering and Rehabilitation*, vol. 13, p. 7, Jan. 2016.

[10] Q. He, S. Perera, Y. Khalifa, Z. Zhang, A. S. Mahoney, A. Sabry, C. Donohue, J. L. Coyle, and E. Sejdić, “The association of high resolution cervical auscultation signal features with hyoid bone displacement during swallowing,” *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 27, no. 9, pp. 1810–1816, Sep. 2019.

[11] A. Kurosu, J. L. Coyle, J. M. Dudik, and E. Sejdić, “Detection of swallow kinematic events from acoustic high-resolution cervical auscultation signals in patients with stroke,” *Archives of Physical Medicine and Rehabilitation*, vol. 100, no. 3, pp. 501–508, Mar. 2019.

[12] C. Rebrion, Z. Zhang, Y. Khalifa, M. Ramadan, A. Kurosu, J. L. Coyle, S. Perera, and E. Sejdić, “High-resolution cervical auscultation signal features reflect vertical and horizontal displacements of the hyoid bone during swallowing,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–9, Dec. 2019.

[13] E. Zahnd, F. Movahedi, J. L. Coyle, E. Sejdić, and P. G. Menon, “Correlating tri-accelerometer swallowing vibrations and hyoid bone movement in patients with dysphagia,” in *Proceedings of the ASME 2016 International Mechanical Engineering Congress and Exposition. Volume 3: Biomedical and Biotechnology Engineering*, Nov. 2016.

[14] K. Shu, J. L. Coyle, S. Perera, Y. Khalifa, A. Sabry, and E. Sejdić, “Anterior-posterior distension of maximal upper esophageal sphincter opening is correlated with high-resolution cervical auscultation signal features,” *Physiological Measurement*, Feb. 2021.

[15] Y. Khalifa, J. L. Coyle, and E. Sejdić, “Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings,” *Scientific Reports*, vol. 10, p. 8704, May 2020.

[16] Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, “Upper esophageal sphincter opening segmentation with convolutional recurrent neural networks in high resolution cervical auscultation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 493–503, Feb. 2021.

[17] S. Mao, Z. Zhang, Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, “Neck sensor-supported hyoid bone movement tracking during swallowing,” *Royal Society Open Science*, vol. 6, no. 7, p. 181982, Jul. 2019.

[18] S. Mao, A. Sabry, Y. Khalifa, J. L. Coyle, and E. Sejdić, “Estimation of laryngeal closure duration during swallowing without invasive x-rays,” *Future Generation Computer Systems*, vol. 115, pp. 610–618, Feb. 2021.

[19] Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdic, “Upper esophageal sphincter opening segmentation with convolutional recurrent neural networks in high resolution cervical auscultation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 493–503, Feb. 2021. [Online]. Available: https://doi.org/10.1109/jbhi.2020.3000057

[20] Y. Khalifa, D. Mandic, and E. Sejdić, “A review of hidden markov models and recurrent neural networks for event detection and localization in biomedical signals,” *Information Fusion*, vol. 69, pp. 52–72, May 2021. [Online]. Available: https://doi.org/10.1016/j.inffus.2020.11.008

[21] E. Sejdić, C. M. Steele, and T. Chau, “Classification of penetration–aspiration versus healthy swallows using dual-axis swallowing accelerometry signals in dysphagic subjects,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1859–1866, Jul. 2013.

[22] J. Lee, C. M. Steele, and T. Chau, “Classification of healthy and abnormal swallows based on accelerometry and nasal airflow signals,” *Artificial Intelligence in Medicine*, vol. 52, no. 1, pp. 17–25, May. 2011.

[23] C. Yu, Y. Khalifa, and E. Sejdić, “Silent aspiration detection in high resolution cervical auscultations,” in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, May 2019, pp. 1–4.

[24] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, “Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks,” *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Nov. 2017.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[26] D. Kiyasseh, G. A. Tadesse, L. N. T. Nhan, L. Van Tan, L. Thwaites, T. Zhu, and D. Clifton, “PlethAugment: GAN-based PPG augmentation for medical diagnosis in low-resource settings,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3226–3235, Nov. 2020.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[28] S. Belarouci and M. Chikh, “Medical imbalanced data classification,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 116–124, Apr. 2017.

[29] Y. SUN, A. K. C. WONG, and M. S. KAMEL, “Classification of imbalanced data: A review,” *International Journal of Pattern*

Recognition and Artificial Intelligence, vol. 23, no. 04, pp. 687–719, 2009.

[30] A. Jain, S. Ratnoo, and D. Kumar, "Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach," in *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, 2017, pp. 1–8.

[31] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection," *Journal of Healthcare Engineering*, vol. 2018, May. 2018.

[32] G. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?" in *DMIN*, 2007.

[33] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv:1711.04340*, Mar. 2018.

[34] Fanny and T. W. Cenggoro, "Deep learning for imbalance data classification using class expert generative adversarial network," *Procedia Computer Science*, vol. 135, pp. 60–67, 2018, the 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

[35] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications*, vol. 91, pp. 464–471, 2018.

[36] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672—-2680.

[37] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[38] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," *arXiv:1806.01875*, Jun. 2018.

[39] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," *arXiv:1706.02633*, Dec. 2017.

[40] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, Nov. 2014.

[41] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2642–2651.

[42] J. M. Dudik, J. L. Coyle, S. Perera, and E. Sejdić, "Dysphagia screening: Contributions of cervical auscultation signals and modern signal-processing techniques," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 465–477, Aug. 2015.

[43] K. Takahashi, M. E. Groher, and K. Michi, "Methodology for detecting swallowing sounds," *Dysphagia*, vol. 9, no. 1, pp. 54–62, 1994.

[44] J. M. Dudik, I. Jestrović, B. Luan, J. L. Coyle, and E. Sejdić, "A comparative analysis of swallowing accelerometry and sounds during saliva swallows," *BioMedical Engineering Online*, vol. 14, p. 3, Jan. 2015.

[45] S. Damouras, E. Sejdić, C. M. Steele, and T. Chau, "An online swallow detection algorithm based on the quadratic variation of dual-axis accelerometry," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3352–3359, Jun. 2010.

[46] J. Lee, C. M. Steele, and T. Chau, "Time and time–frequency characterization of dual-axis swallowing accelerometry signals," *Physiological Measurement*, vol. 29, no. 9, pp. 1105–1120, Sep. 2008.

[47] E. Sejdić, V. Kosmisar, C. M. Steele, and T. Chau, "Baseline characteristics of dual-axis cervical accelerometry signals," *Annals of Biomedical Engineering*, vol. 38, no. 3, pp. 1048–1059, Mar. 2010.

[48] E. Sejdić, C. M. Steele, and T. Chau, "A procedure for denoising dual-axis swallowing accelerometry signals," *Physiological Measurement*, vol. 31, no. 1, pp. N1–9, Jan. 2010.

[49] M. Aboy, R. Hornero, D. Abasolo, and D. Alvarez, "Interpretation of the lempel-ziv complexity measure in the context of biomedical signal analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2282–2288, 2006.

[50] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Başar, "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," *Journal of Neuroscience Methods*, vol. 105, no. 1, pp. 65–75, 2001.

[51] K. E. Smith and A. O. Smith, "Conditional GAN for timeseries generation," *arXiv:2006.16477*, Jun. 2020.

[52] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223.

[53] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, ser. NIPS'17, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5769–5779.

[54] Y. Luo and B. L. Lu, "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 2535–2538.

[55] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," *Expert Systems with Applications*, vol. 174, p. 114582, Jul. 2021.

[56] A. M. Delaney, E. Brophy, and T. E. Ward, "Synthesis of realistic ecg using generative adversarial networks," *arXiv:1909.09150*, Sep. 2019.

[57] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton, "Generative models and model criticism via optimized maximum mean discrepancy," *arXiv:1611.04488*, Jan. 2021.

[58] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[59] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[61] S. Sarraf Shirazi, C. Buchel, R. Daun, L. Lenton, and Z. Moussavi, "Detection of swallows with silent aspiration using swallowing and breath sound analysis," *Medical & Biological Engineering & Computing*, vol. 50, no. 12, pp. 1261–8, Dec. 2012.