On the Problem of Data Availability in Automatic Voice Disorder Detection

Dayana Ribas, Antonio Miguel, Alfonso Ortega and Eduardo Lleida
ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

Keywords: Voice Disorder, Saarbruecken Voice Database (SVD), Advanced Voice Function Assessment Database (AVFAD), Opensmile, SVM, SMOTE.

Abstract: In order to support medical doctors in having more versatile health assistance, automatic voice disorder detection systems enable the remote diagnosis, treatment, and monitoring of voice pathologies. The main problem for developing the related technology is the availability of audio data of healthy and pathological voices manually labeled by experts. Saarbruecken Voice Database (SVD) was created in 1997, with a collection of more than 5 hours of healthy and pathologica audio data. This database has been widely used for developing voice disorder detection systems. However, it has some issues in the distribution of data and the labeling that makes it difficult to conduct conclusive studies. This paper evaluates an Automatic Voice Disorder Detection (AVDD) system using the recent Advanced Voice Function Assessment Database (AVFAD) with almost 40 hours of audio data and SVD as a reference. The system consists of a representation using spectral, prosody, and voice quality parameters followed by an SVM classifier that can obtain up to 88% accuracy in phrases and 86% in sustained vowel *a*. Data augmentation strategy is assessed for handling the problem of data imbalance with the SMOTE method which improves the performance of male, female, and gender-independent models without decreasing the results for scenarios with data balance. Finally, we release the system implementation for voice disorder detection including the list of train-test partitions for both databases.

1 INTRODUCTION

Nowadays, voice disorders have an impressive prevalence in the population. Previous studies (Bhattacharyya, 2014) reported an important affectation due to voice problems among the population, especially for professionals that use the voice as a primary tool, such as teachers, telemarketers, TV presenters, singers, etc (Roy et al., 2004). There are many reasons why affected people do not go to the doctor in time neglecting the problem while getting worse. However, with the recent development of remote health services, there is an opportunity to use smart solutions to assist doctors in remotely screening patients contributing to early diagnosis and continuous monitoring of patient evolution without the need for a hospital visit.

Figure 1 shows a scheme of the scenario of application for health remote services, where the patient and the specialist interact with the AVDD system from their own side. Behind these services, the AVDD system (Ali et al., 2017b; Verde et al., 2019) consists of a traditional machine learning scheme where the system learns to identify the pathological cues from speech audio samples. There are many previous related approaches for studying suitable representations and classifiers for distinguishing between healthy and pathological speech (Al-nasheri et al., 2016; Ali et al., 2017a; Harár et al., 2017; Mohammed et al., 2020). However, a principal problem

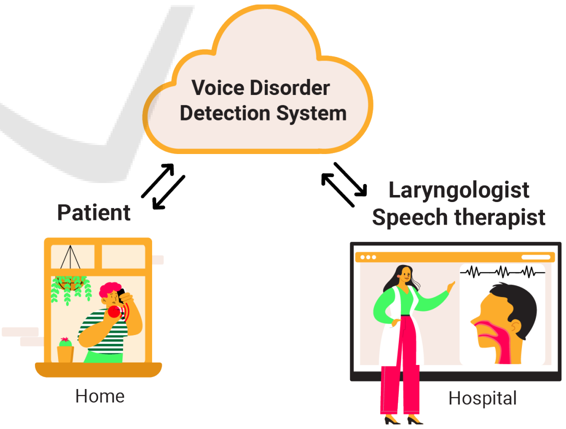
Voice Disorder Detection System
Patient
Laryngologist Speech therapist
Home
Hospital

Figure 1: Illustration of the scenario of application for an Automatic Voice Disorder Detection system.

Ribas, D., Miguel, A., Ortega, A. and Lleida, E.
On the Problem of Data Availability in Automatic Voice Disorder Detection.
DOI: 10.5220/0011669300003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 330-337
ISBN: 978-989-758-631-6; ISSN: 2184-4305

is the availability of audio data of healthy and pathological voices manually labeled by experts for developing the technology related to the AVDD systems. Usually, these data result from research projects with medical institutions, where the condition is to keep data private. Therefore, the availability of datasets for developing AVDD systems is quite limited.

Saarbruecken Voice Database (SVD) (Pützer and Koreman, 1997) is one of the few freely available databases for this task. Considering the number of studies in the related literature based on this database, we would not hesitate to say that SVD is almost a standard for developing AVDD systems (Sarika Hegde and Dodderi, 2019). SVD contains many recordings and speakers with labeled healthy and pathological speech, including more than 5 hours of speech recordings of the vowels *a, i, u* and short phrases. However, there is great inequality in the distribution of individual pathologies. Some pathologies only have one audio sample, and they can end up only in the testing set, which is a problem for training AVDD systems. Also, there are some issues with labeling. For instance, there are many audios labeled with more than one pathology, as well as many diagnoses that are not really voice pathologies, such as Cordectomy or Down's Disease. All these issues difficult the interpretation of the AVDD system's performance results over SVD. Typically, these issues are avoided by selecting those pathologies with enough audio samples for training and test set (Al-nasheri et al., 2016; Ali et al., 2017a; Verde et al., 2019; Mohammed et al., 2020). However, this scenario doesn't reflect the realistic case of use, where the system would process a wide variability of pathologies out of a fixed selected set.

In this paper, we introduce the AVFAD for the AVDD task. It is a recent dataset with healthy and pathological speech (Jesus et al., 2017). This corpus is well documented with an extended study of its acoustic characteristics. However, because of its recent release, it has not been previously used for voice disorders classification tasks, so its performance results for AVDD tasks are a novel contribution. In the following, we evaluate the performance of an AVDD system using SVD and AVFAD with male, female, and gender-independent models. Then, we address the problem of class imbalance using SVD as an example of an imbalanced dataset and AVFAD as an example of a balanced one. For this aim, we evaluate a data augmentation strategy in the feature domain known as Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). This method generates new samples from existing samples in the minority class to increase the information for training the model. SMOTE has been previously used in different classification tasks, including speech processing (He and Ma, 2013; **?**). For voice disorder detection, there are some previous works studying the behavior of SMOTE algorithm in this context (Fan et al., 2021; Chui et al., 2020). Although their systems and experimental setup differ from our proposal, obtained results are encouraging and support our motivation to use SMOTE for balancing the dataset in this framework. Finally, we release the system implementation to contribute to reproducibility and further developments.

Contributions of this paper are:

- Performance evaluation of an automatic voice disorder detection system for assisting on the diagnosis of voice pathologies using AVFAD and SVD corpus, conformed by a representation using spectral, prosody and voice quality parameters and an SVM classifier.
- Evaluation of SMOTE algorithm to deal with the problem of class imbalance in the voice disorder detection framework.
- Free available implementation [1] of the voice disorder detection system based on a machine learning approach for binary and multi-class classification, including the list of train-test partitions in AVFAD and SVD databases.

In the following, Section 2 comments on the SMOTE method for handling the problem of class imbalance. Section 3 presents the materials used in this study, including the databases and the performance metrics. Then Section 4 describes the AVDD system. Section 5 describes the experimental evaluation and discusses obtained results. Finally, Section 6 concludes the paper.

2 DATA AUGMENTATION FOR CLASS IMBALANCE: SMOTE

The main weakness in developing an AVDD system relies on the data. Therefore, the data accommodation stage is the most delicate part of the AVDD development process. A frequent data problem in health-related applications is the availability of an unequal amount of samples for healthy and pathology classes. However, the usual low availability of data dismisses the simple solution of discarding some samples to balance the training set. Thus, training a machine learning model with an imbalanced dataset produces poor performance on the minority class, although in some

[1] https://github.com/dayanavivolab/voicedisorder

Table 1: Distribution of audio samples by fold for each model gender divided by training and evaluation sets: Male (Train—Test), Female (Train—Test), Both (Train—Test).

	Dataset: SVD			**Dataset: AVFAD**		
Model	Male	Female	Both	Male	Female	Both
Audio-type: Sustained vowel *a*						
Fold 1	715—171	916—239	1631—410	168—42	398—100	566—142
Fold 2	700—186	915—240	1615—426	168—42	398—100	566—142
Fold 3	715—171	934—221	1649—392	168—42	398—100	566—142
Fold 4	711—175	921—234	1632—409	168—42	399—99	567—141
Fold 5	703—183	934—221	1637—404	168—42	399—99	567—141
Audio-type: Phrases						
Fold 1	709—171	881—227	1590—398	1008—252	2389—598	3397—850
Fold 2	695—185	877—231	1572—416	1008—252	2389—598	3397—850
Fold 3	711—169	895—213	1606—382	1008—252	2390—597	3398—849
Fold 4	707—173	884—224	1591—397	1008—252	2390—597	3398—849
Fold 5	698—182	895—213	1593—395	1008—252	2390—597	3398—849

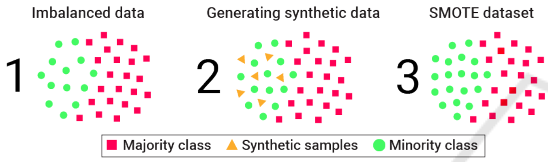

Figure 2: SMOTE algorithm samples generation.

cases it is the performance of the minority class that is most important (He and Ma, 2013).

SMOTE is a type of data augmentation for the minority class (Chawla et al., 2002). It addresses the imbalanced dataset problem by oversampling the minority class. SMOTE consists of synthesizing new samples from the existing examples of the minority class, increasing the information to train the model. In order to be less tied to the application, it works in the feature space rather than in the sample space. The process consists of selecting a random sample from the minority class and its $k = 5$ nearest neighbors samples. Then a line between the pair of samples is drawn, and a new feature in a random middle point is generated. The synthetic examples drive the model to create larger and less specific decision regions, bringing more generalization for the minority class. Figure 2 shows an illustration of the process.

SMOTE over-sampling is then combined with under-sampling such that the classifier learns on the dataset perturbed by over-sampling the minority class and under-sampling the majority class. It consists of under-sampling the majority class by randomly removing samples until the minority class becomes some specified percentage of the majority class to produce a higher presence of the minority class in the training set. It forces the model to experience different degrees of under-sampling, so the initial bias of the learner towards the majority class is inverted in favor of the minority class.

Some approaches of the SMOTE method have been developed by focusing on the selectivity of the minority class examples used for generating new synthetic observations. For instance, Borderline SMOTE selects misclassified examples of the minority class according to a k-nearest neighbor classification model. This way, it provides robustness to the sampling process oversampling only the more difficult instances. Subsequently, (Nguyen et al., 2011) proposes an alternative that uses an SVM method to identify the misclassified examples on the decision boundary. This approach also includes those regions with fewer density of observations belonging to the minority class and attempts to extrapolate towards the class boundary. Related to this proposal, Adaptive Synthetic Sampling (ADASyn) SMOTE (He et al., 2008) also attempts to generate more synthetic examples in those areas where the density of minority examples is low, while fewer, where the density is high.

3 EXPERIMENTAL SETUP

3.1 Databases

This study is based on the following databases:

- The Saarbruecken Voice Database (SVD) (Pützer and Koreman, 1997) which is an open access dataset including the healthy and pathology-labeled speech of 71 voice disorders. It has been broadly used in previous works for studying automatic detection and assessment of voice pathologies (Sarika Hegde and Dodderi, 2019). It contains around 5 hours of voice recordings of 687 healthy persons and 1356 patients.
- The Advanced Voice Function Assessment

Database (AVFAD) (Jesus et al., 2017) which is also an open-access dataset in the Portuguese language. It includes almost 40 hours of recordings for 363 persons with no vocal alterations and 346 clinically diagnosed within 26 vocal pathologies.

In this study, we use the sustained vowel *a* and the phrases included in the recording sessions of SVD and AVFAD. Experiments are carried out in a cross-validation framework, where audio data are in five folds, each with train and evaluation sets. Note that the dataset distribution of SVD is approximately one healthy by two pathological samples, such that classes in the evaluation set are imbalanced. On the other side, the sample distribution in AVFAD provided a better balance between healthy and pathology classes, such that the evaluation set are mostly balanced. Table 1 presents the information on audio samples distribution for each fold.

3.2 Performance Metrics

The system performance is evaluated in terms of the Accuracy

$$ACC = \frac{TN+TP}{TN+TP+FN+FP}, \qquad (1)$$

and Unweighted Average Recall

$$UAR = 0.5 \cdot \frac{TP}{TP+FN} + 0.5 \cdot \frac{TN}{FP+TN}. \qquad (2)$$

Both metrics are computed from the true and false positive and negative rates (TP, FP, TN, FN). The distributions of ACC and UAR are quite similar for balanced classes. However, if this is not the case, UAR considers each class by itself, while ACC provides a more general metric. There is also computed the Recall and F1-Score

$$Recall = \frac{TP}{TP+FN}, \qquad (3)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \qquad (4)$$

where $Precision = \frac{TP}{TP+FP}$.

4 SYSTEM

The AVDD system consists of the machine learning classification system depicted in Fig. 3. The following subsections describe the main modules of the system.

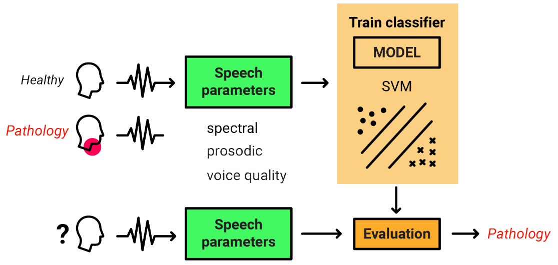

Figure 3: System for Automatic Voice Disorder Detection.

4.1 Speech Parameters

The representation module uses a set of parameters designed for automatic recognition of paralinguistic issues during the Computational Paralinguistics ChallengE (ComParE)[2] in 2013. It was implemented in the openSMILE toolkit[3] for extracting suprasegmental audio feature sets in real time. The ComParE acoustic feature set includes spectral, cepstral, prosodic, and voice quality parameters of speech signals obtained by applying a large set of statistical functionals to acoustic low-level descriptors. These descriptors cover a broad set of parameters usually employed for representing speech signals, such as Mel Frequency Cepstral Coefficients (MFCC) and RASTA. As well as sound quality descriptors frequently used for voice pathology analysis such as jitter, shimmer, and Harmonic to Noise Ratio (HNR) (Teixeira et al., 2013). See all parameters in Table 2. On top of this, several statistical functions are applied to low-level descriptors to obtain better representations, including mean, variance, kurtosis, skewness, percentiles, etc. The final dimension of the ComParE feature set is 6373 parameters.

In (Huckvale and Buciuleac, 2021) authors reported results with ComParE for phrases in the SVD dataset with an accuracy of 80.71% (taken from table 2 in (Huckvale and Buciuleac, 2021)) similar to the accuracy of 82.8% reported in (Barche et al., 2020) for /aiu/ concatenated vowels using SVD.

4.2 Model

Support Vector Machine (SVM) is selected for the classification module (Schölkopf and Smola, 2002). SVM has been widely used for speech pathology analysis in previous works (Sarika Hegde and Dodderi, 2019). This method attempts to find the optimal hyperplane to establish the boundary between the samples of different classes in the training set. To

[2] http://www.compare.openaudio.eu/

[3] https://github.com/audeering/opensmile

Table 2: Speech parameters in the ComParE audio set used as frontend (Weninger et al., 2013).

Spectral and cepstral parameters
RASTA spectrum (bands 1-26, frecuency 0-8 kHz)
Spectral energy 250-650 Hz (1-4 kHz)
Spectral roll-offf point (0.25, 0.50, 0.75, 0.90)
Spectral Flux, Centroid, Entropy, Slope
Psychoacoustic Sharpness, Harmonicity
Spectral Variance, Skewness, Kurtosis
MFCC (coefficients: 1-14)
Prosodic parameters
Sum of the auditory spectrum (loudness)
RASTA spectrum (energy)
RMS Energy, Zero-Crossing Rate
F0 (SHS and Viterbi smoothing)
Voice quality parameters
Probability of voicing
Log. HNR, Jitter (local, delta), Shimmer (local)

choose the suitable configuration of SVM, we conducted several auxiliary experiments to test the kernel and the hyper-parameters C for error control. Finally, we selected the linear kernel with $C = 1$.

4.3 Implementation

The AVDD system is implemented in python and released for reproducibility. We used open and standard python libraries such as OpenSMILE for the representation, Sklearn for the classifier, and Imbalanced-learn for the data augmentation. The free available implementation is located in the following link[4].

5 EXPERIMENTS AND RESULTS

The following subsections present the experiments carried out for assessing the performance of the automatic voice disorder detection system in a binary format, namely healthy vs. pathology. Experiments are evaluated in two data scenarios related to the balance between the number of samples of the classes healthy and pathology, which is represented by the datasets:

1. SVD: Data imbalanced 30% health vs. 70% path.
2. AVFAD: Data balanced 50% health vs. 50% path.

Then, several experiments are designed for handling the class imbalance using the SMOTE data augmentation strategy.

5.1 Voice Disorder Detection

This section presents the results of the performance of the AVDD system for the SVD and the AVFAD corpora in terms of classification accuracy, specifying the recall for healthy and the pathology classes, F1-Score, and UAR to assess the imbalance between classes. There are results for three different types of models by gender (male, female, and gender-independent), where the training and evaluation sets include audio samples of the specific gender.

Results in the first columns of Table 3 correspond to the models trained with audios of the sustained vowel *a*. Considering that spectral characteristics between males and females are usually remarkable (Priya et al., 2022), a better performance could be expected for the gender-dependent models (male and female). However, results show that comparing gender-dependent and gender-independent models there is not a great difference in SVD, while in AVFAD we can see better performance for those models, including females samples. Note that the best F1 for experiments with vowel *a* corresponds to female models in AVFAD ($F1 = 87.81$). This could be related to females' audio samples being almost double of males' audio samples in AVFAD (table1), so the female model is expected to be more robust.

The right columns in Table 3 present the results for phrases. Compared to the models with the vowel *a* the system performance increases for all gender-dependent models, though this is more remarkable for SVD dataset. It is an expected result, considering that phrases have larger and more diverse audio material than a single vowel. Therefore beyond the sustained sound, there is information in the transition among sentence phonemes that would be useful for disorder detection. Again gender-independent models do not present better performance than gender-dependent models, especially for the SVD. In AVFAD, there is also better performance for models with female audio samples, supporting the results obtained for the vowel *a*. So, looking at the trade-off between data and performance, we could conclude that training gender-independent models is more convenient, especially when the availability of training audio is reduced.

The system works better for the AVFAD than the SVD dataset, mainly in experiments with vowel *a*. This result reflects that AVFAD has more hours of speech data than SVD, even though the amount of speakers is less for AVFAD than for SVD. In both datasets, $Recall_{Healthy}$ is lower than $Recall_{Path}$ for all models, indicating the difficulty of detecting healthy samples over pathological ones. This behavior could

[4] https://github.com/dayanavivolab/voicedisorder

Table 3: Performance metrics for binary voice disorder classification: healthy vs. pathology.

SVD	**Audio type: Vowel *a***					**Audio type: Phrases**				
Gender	ACC	Recall_Health	Recall_Path	F1	UAR	ACC	Recall_Health	Recall_Path	F1	UAR
Male	72.81	42.94	85.42	81.62	64.18	82.64	63.90	90.02	88.04	76.96
Female	72.18	55.49	82.14	78.74	68.81	83.38	73.82	88.48	87.43	81.15
Indep.	71.94	46.58	84.82	80.03	65.70	83.27	71.26	88.86	87.83	80.06

AVFAD	**Audio type: Vowel *a***					**Audio type: Phrases**				
Gender	ACC	Recall_Health	Recall_Path	F1	UAR	ACC	Recall_Health	Recall_Path	F1	UAR
Male	78.57	72.53	85.80	78.33	79.17	85.32	79.66	91.95	85.26	85.80
Female	86.75	77.94	95.62	87.81	86.78	88.15	79.30	96.99	89.12	88.15
Indep.	86.44	78.76	94.63	87.18	86.70	86.84	78.49	95.57	87.65	87.03

be related to including pathological samples with low severity in the pathology class of the training set, such that they could be closer to the healthy than the pathological class. For instance, a voice sample of a patient with a low level of dysphonia could sound similar to a healthy voice. Thus, there will be misses close to the SVM boundary, inducing some uncertainty during model training.

In order to see the possible improvement margin for accuracy-related metrics, Table 4 shows the oracle reference[5] for the models with gender-independent models. Note that the best $Recall_{Healthy}$ is again worst than the best $Recall_{Path}$. It confirms that detecting healthy samples is more difficult for system classification than a pathological sample.

Table 4: Oracle reference of gender-independent models in SVD and AVFAD datasets.

Dataset	ACC	Recall *Healthy*	Recall *Path*	F1	UAR
	Audio type: Vowel a				
SVD	86.86	72.41	94.19	90.48	83.30
AVFAD	91.63	84.04	99.56	92.08	91.80
	Audio type: Phrase				
SVD	92.19	84.50	95.79	94.35	90.14
AVFAD	90.03	82.17	98.25	90.60	90.21

5.2 Classes Imbalance: SMOTE

In order to handle the class imbalance problem, the following experiments assess the system performance when applying the data augmentation method SMOTE. To illustrate the data distribution, Fig. 4 shows a 2-dimensional visualization of the features using TSNE before and after class balancing for gender-independent models of phrases. In both datasets, the feature distribution in SVD is considerably overlapped compared to AVFAD.

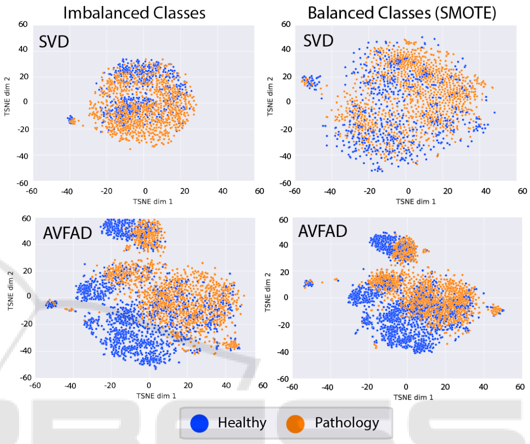

Figure 4: Visualization of features after a dimension reduction using TSNE for gender-independent models of phrases.

Table 5: Oracle reference with SMOTE for gender-independent models in SVD and AVFAD datasets.

Dataset	ACC	Recall *Healthy*	Recall *Path*	F1	UAR
	Audio type: Vowel a				
SVD	89.25	95.27	83.23	88.55	89.25
AVFAD	91.52	83.45	99.59	94.64	91.52
	Audio type: Phrase				
SVD	94.80	97.73	91.87	92.15	94.80
AVFAD	90.22	81.75	98.69	90.98	90.22

Then, Table 5 shows results for the oracle reference when using SMOTE balancing method for gender-independent models. Comparing between oracle reference of SMOTE with the original data (Table 4), the system performance improves indicating that the data augmentation provides some amount of new information by means of oversampling. Note that the improvement is noticeable for SVD (UAR=89.25 and 94.80 in Table 4 with respect to UAR=83.30 and 90.14 in Table 5), where the SMOTE really increases the samples because as AVFAD's folds are already balanced, the data augmentation is slight or none.

Fig. 5 shows the system results with gender-

[5] Oracle reference means the system trained and evaluated with the same training partition.

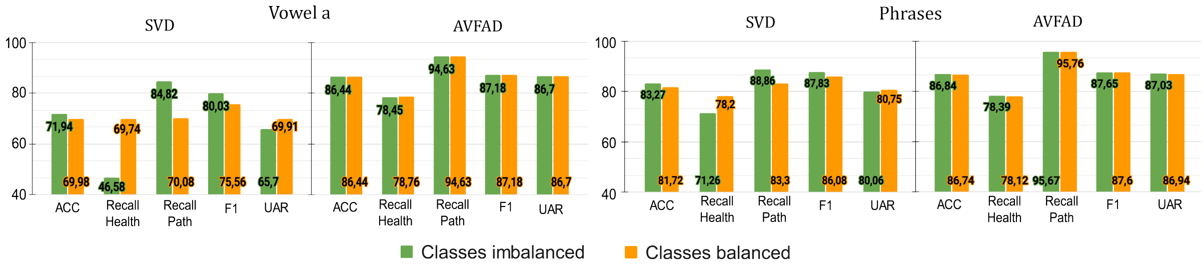

Figure 5: Results of ACC, $Recall_{Healthy}$, $Recall_{Path}$, F1-Score, and UAR for the original data distribution in SVD and AVFAD (green-left bars) and the SMOTE method for balancing classes (orange-right bars) for the gender-independent model.

independent models for the original imbalanced classes along with classes after balancing. First, note that adjacent result bars in AVFAD charts are almost equal. This shows no improvement with the data augmentation in AVFAD experiments because classes in each fold are originally balanced. This achievement is consistent with results in Table 3, where ACC and UAR are similar in AVFAD experiments, indicating that these models are well-balanced for classes.

Applying data augmentation in SVD's experiments there is a clear improvement in the recall of healthy class. Even though the recall of the pathology class decreases, the trade-off between these metrics produces a final UAR improvement. For the original system (green bars), the difference between ACC and UAR is 6.24% for vowel *a* and 3.21% for phrases. Then, when applying the balance compensation method, it reduces to 0.07% for vowel *a* and 0.97% for phrases. In both cases, SMOTE makes ACC and UAR approach, which indicates that the models are more robust for the class imbalance scenario. Concluding the data augmentation technique successfully handles the class imbalance in the SVD, while at the same time, it does not harm the system performance for the already balanced AVFAD dataset.

Further experiments were carried out for comparing the original SMOTE with the extensions SVM-SMOTE and ADASyn. However, obtained results do not show a remarkable performance difference among the methods. This could be related to the high overlap among the intrinsic distribution of these datasets (Fig. 4), which does not allow finding isolated areas of the minority class for taking advantage of the data augmentation.

6 CONCLUSIONS

This work studied the problem of data availability for detecting voice disorders using two different scenarios determined by the dataset. On the one hand, the SVD, which is almost a standard for evaluating AVDD systems, is an imbalanced dataset with more pathological samples than healthy ones. On the other hand, the AVFAD is a recently released dataset with almost balanced healthy and pathological audio samples. This dataset has not been previously evaluated for classification tasks. The AVDD system employed here is a state-of-the-art system that uses spectral, prosody, and voice quality parameters for representing speech, followed by an SVM model for classifying between healthy and pathology samples. Experimental results include the performance for gender-dependent and gender-independent models and employing phrases and the sustained vowel a. Comparing the results among models, we conclude that it is more convenient to use gender-independent models to not be forced of training two models with fewer data each, especially when there is low availability of data.

Then we evaluated the performance of data augmentation by means of the SMOTE method for handling the class imbalance. The results show that the AVDD system augmented with the SMOTE method increases the recall of the healthy class and makes ACC and UAR closer together in SVD, considering that SVD- is a class-imbalanced dataset. While at the same time, it do not decrease the performance of the system for AVFAD that is originally balanced. Further experiments with SMOTE method extensions did not show performance improvement on top of the basic SMOTE method in the datasets evaluated.

In the future, we plan to extend the study on the performance of the SMOTE algorithm to multiclass classification, considering the wide range of pathologies in the datasets. First, we will study how to establish an organization of the labeled diagnosis on this corpus by speech characteristics related to the pathology condition. Furthermore, we plan to study the behavior of the system performance when defining different operation points related to the application use case of use.

ACKNOWLEDGEMENTS

This work was supported by the EU Horizon 2020 program MSCA-101007666; by MCIN/AEI/10.13039/501100011033 and by the NextGeneration-EU/PRTR under Grants PDC2021-120846-C41 & PID2021-126061OB-C44, and by the Government of Aragon (Grant Group T36_20R).

REFERENCES

Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., Malki, K. H., and Bencherif, M. A. (2016). An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification. *Journal of Voice*, 31(1):113.e9–113.e18.

Aldraimli, M., Soria, D., Parkinson, J., Thomas, E., Bell, J., Dwek, M., and Chaussalet, T. (2020). Machine learning prediction of susceptibility to visceral fat associated diseases. *Health and Technology*, 10:925–944.

Ali, Z., Alsulaiman, M., Elamvazuthi, G. M. I., Al-Nasheri, A., Mesallam, T., Farahat, M., and Malki, K. (2017a). Intra- and Inter-Database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology? *Journal of Voice*, 31:386.e1–386.e8.

Ali, Z., Muhammad, G., and Alhamid, M. F. (2017b). An automatic health monitoring system for patients suffering from voice complications in smart cities. *IEEE Access*, 5:3900–3908.

Barche, P., Gurugubelli, K., and Vuppala, A. K. (2020). Towards automatic assessment of voice disorders: A clinical approach. In *Proc. Interspeech 2018*, pages 2537–2541.

Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the United States. *The Laryngoscope*, 124:2359–2362.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Chui, K. T., Lytras, M. D., and Vasant, P. (2020). Combined generative adversarial network and fuzzy c-means clustering for multi-class voice disorder detection with an imbalanced dataset. *Applied Sciences*, 10(13).

Fan, Z., Wu, Y., Zhou, C., Zhang, X., and Tao, Z. (2021). Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. *Applied Sciences*, 11(8).

Harár, P., Alonso, J., Mekyska, J., Galáž, Z., Burget, R., and Smekal, Z. (2017). Voice pathology detection using deep learning: a preliminary study. pages 1–4.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

He, H. and Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press.

Huckvale, M. and Buciuleac, C. (2021). Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials. In *Interspeech*, pages 1399–1403.

Jesus, L. M., Belo, I., Machado, J., and Hall, A. (2017). The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research. In Fernandes, F. D. M., editor, *Advances in Speech-language Pathology*, chapter 14. IntechOpen, Rijeka.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., and AL-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, 10(11).

Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 3(1):4–21.

Priya, E., S, J. P., Reshma, P. S., and S, S. (2022). Temporal and spectral features based gender recognition from audio signals. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 1–5.

Pützer, M. and Koreman, J. (1997). A German database of pathological vocal fold vibration. pages 143–153.

Roy, N., Merrill, R. M., Thibeault, S., Parsa, R. A., Gray, S. D., and Smith, E. M. (2004). Prevalence of Voice Disorders in Teachers and the General Population. *Journal of Speech Language and Hearing Research*, 47:281–293.

Sarika Hegde, Surendra Shetty, S. R. and Dodderi, T. (2019). A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *Journal of Voice*, 33(6):947.e11–947.e33.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press.

Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal acoustic analysis – jitter, shimmer and hnr parameters. *Procedia Technology*, 9(Complete):1112–1122.

Verde, L., Pietro, G. D., Alrashoud, M., Ghoneim, A., Al-Mutib, K. N., and Sannino, G. (2019). Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app. *IEEE Access*, 7:124048–124054.

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., and Scherer, K. R. (2013). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, 4(292). ID: unige:97889.