applied sciences

MDPI

Article

An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection

Ji-Na Lee [1] and Ji-Yeoun Lee [2,*]

[1] Division of Global Business Languages, Seokyeong University, Seogyeong-ro Seongbuk-gu, Seoul 02173, Republic of Korea
[2] Department of Bigdata Medical Convergence, Eulji University, 553 Sanseong-daero, Sujeong-gu, Seongnam-si 13135, Republic of Korea
* Correspondence: jylee@eulji.ac.kr

Abstract: The Saarbruecken Voice Database (SVD) is a public database used by voice pathology detection systems. However, the distributions of the pathological and normal voice samples show a clear class imbalance. This study aims to develop a system for the classification of pathological and normal voices that uses efficient deep learning models based on various oversampling methods, such as the adaptive synthetic sampling (ADASYN), synthetic minority oversampling technique (SMOTE), and Borderline-SMOTE directly applied to feature parameters. The suggested combinations of oversampled linear predictive coefficients (LPCs), mel-frequency cepstral coefficients (MFCCs), and deep learning methods can efficiently classify pathological and normal voices. The balanced datasets from ADASYN, SMOTE, and Borderline-SMOTE are used to validate and evaluate the various deep learning models. The experiments are conducted using model evaluation metrics such as the recall, specificity, G, and F1 value. The experimental results suggest that the proposed voice pathology detection (VPD) system integrating the LPCs oversampled by the SMOTE and a convolutional neural network (CNN) can effectively yield the highest accuracy at 98.89% when classifying pathological and normal voices. Finally, the performances of oversampling algorithms such as the ADASYN, SMOTE, and Borderline-SMOTE are discussed. Furthermore, the performance of SMOTE is superior to conventional imbalanced data oversampling algorithms, and it can be used to diagnose pathological signals in real-world applications.

Keywords: pathological voice; disordered voice; imbalanced learning; voice pathology classification; SMOTE; ADASYN; Borderline-SMOTE; deep learning; intelligent medical diagnosis system

Citation: Lee, J.-N.; Lee, J.-Y. An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Appl. Sci.* **2023**, *13*, 3571. https://doi.org/10.3390/app13063571

Academic Editors: Jongweon Kim and Yongseok Lee

Received: 2 January 2023
Revised: 20 February 2023
Accepted: 21 February 2023
Published: 10 March 2023

1. Introduction

Artificial intelligence has been used successfully in medical applications in recent years [1–10]. Pathological voice detection systems based on deep learning algorithms and characteristic parameters have become popular research topics. In biomedical signal processing, different parameters are extracted from databases to create voice pathology detection (VPD) systems that effectively classify pathological and normal voices. Most studies have experimented with the Saarbruecken Voice Database (SVD) and Massachusetts Eye and Ear Infirmary (MEEI) database [11], which are widely used in VPD systems [12,13]. However, in previous studies on VPD, many researchers failed to consider the imbalanced distributions of the voice samples when using SVD and MEEI databases.

The MEEI Voice Disorders database was developed by the Kay Elemetrics Corp and was released in 1994 [11]. The recordings consist of sustained phonations of the vowel /ah/ (657 pathological and 53 normal) and rainbow passage sentences (657 pathological and 53 normal). The SVD includes 2000 voice samples and electroglottography (EGG) signal sets. It contains the records of 687 healthy people (428 women and 259 men) and 1356 people (727 women and 629 men) with different pathologies [12]. The recordings include the

/i, a, u/ vowels produced by a normal voice, the /i, a, u/ vowels with high and low pitches, and the German sentence "Guten Morgen, wiegeht es Ihnen?". This dataset is pretty recent and is useful for research related to voice disorders [14–18]. In conclusion, the distributions of pathological and normal voices are disproportionate in these two databases [1–10,13–18]. Therefore, considering the accuracy metric as the evaluation result of a classifier, the performance of the constructed classification model for pathological voice may be better than the actual performance. This is because, when trained on imbalanced datasets, it is generally a biased classifier that achieves better performances for majority classes and poorer performances for minority classes [19–21]. In real-world applications, imbalanced class data are caused by insufficient and disproportionate samples in the utilized disordered voice database. This also prevents algorithms that detect various types of pathological speech from performing well in the existing VPD systems. However, collecting balanced class datasets in experiments on medical diagnosis and rehabilitation is very difficult because signal acquisition through the patient's laryngoscope and labeling through manual supervision are required. Given its importance, the diagnosis of pathological voices using disproportionate data has attracted researchers' attention [22,23]. Therefore, when using a class-imbalanced pathological speech database, imbalanced learning methods must be implemented using oversampling to model an effective VPD system.

Few studies have applied oversampling methods to the SVD and MEEI [2,22,24–26]; however, here, we introduce two representative papers. Recent work has demonstrated the development of a system to automatically classify pathological and normal speech on an imbalanced class dataset, utilizing a fuzzy clustering synthetic minority oversampling technique algorithm (FC-SMOTE) [24]. The accuracy (90%) obtained from FC-SMOTE with a convolutional neural network (CNN) was superior to that of the traditional oversampling algorithms using imbalanced data [24]. Kwok et al. proposed CGAN-IFCM combining a conditional generative adversarial network (CGAN) and an improved fuzzy c-means clustering (IFCM) algorithm [22]. The classification ability of the CGAN-IFCM exceeded that of existing models, with an accuracy of 95.15%.

The synthetic minority oversampling technique (SMOTE) is the oversampling method most commonly used to manage imbalanced data [27]. The SMOTE improves the random oversampling (ROS) method to reduce the risk of overfitting, but it can be a source of overgeneralization [28]. The proposed SMOTE-based methods include Borderline-SMOTE [29], adaptive synthetic sampling (ADASYN) [30], and Random-SMOTE [31]. However, most of these approaches are highly complex or mitigate only some of the drawbacks of SMOTE [32]. In imbalanced deep learning scenarios, overlapping between the classes and unclear boundaries affects the performance of oversampling techniques. Eventually, noisy samples without an information value are generated. In particular, the k-nearest neighbors algorithm can be useful for a pathological voice dataset with ambiguous boundaries and overlapping between classes [33]. Thus, we use this advantage of the k-nearest neighbors method in imbalanced learning as an integrated technique for oversampling tasks. Furthermore, related studies [32] mainly use the k-nearest neighbors algorithm. Various oversampling methods based on the k-nearest neighbors algorithm are used in this study; these methods can overcome the influences of unclear points and noise and the simultaneously imbalanced distributions within and between classes. Finally, representative minority samples with as many discriminating features as possible should be generated to increase the accuracy of pathological voice classification models.

Most pathological voice detection and intelligent diagnostic studies typically use MEEI and SVD as standard databases for VPD systems, but they do not consider imbalances in the number of pathological and normal samples in each database [1–10,14–18]. Accuracy is not suitable for an imbalanced class dataset because it makes model performances inaccurate as a measure of model evaluation. To address these problems, this work proposes a VPD system that combines various oversampling methods based on SMOTE, ADASYN, and Borderline-SMOTE and deep learning models such as CNN and feedforward neural network (FNN) to classify between normal and pathological voices with the imbalanced

SVD. In addition, this study proposes oversampled linear predictive coefficients (LPCs) and mel-frequency cepstral coefficients (MFCCs) as the input parameters of the model. Then, the balanced class datasets with the oversampled MFCC or LPC inputs are utilized to validate and evaluate the combinations of deep learning models in the modeling phase of the VPD system. Finally, evaluation metrics [34] such as recall, specificity, G, and F1 are presented for the classifier model. The contributions of this study can be summarized as follows:

- This paper introduces efficient deep learning models based on various oversampling methods, such as the SMOTE, Borderline-SMOTE, and ADASYN, and directly applies them to feature parameters for VPD.
- The suggested combinations of the oversampled MFCCs, LPCs, and deep learning methods can efficiently classify pathological and normal voices.
- Several experiments are conducted to verify the usefulness of the developed VPD system using the SVD.
- The results highlight the excellence of the proposed classification system, which integrates a CNN and LPCs based on the SMOTE in terms of monitoring voice disorders; it is an effective and reliable system.

2. Materials and Methods

2.1. Database

The SVD, which was developed by the Phonetics Research Institute at Saarland University, Germany [12], contains more than 2000 voiced samples of sustained /a/, /i/, and /u/ vowels and the sentence "Guten Morgen, wie geht es Ihnen?". All recordings are sampled at 50 kHz, and their resolutions are 16 bits. The lengths of the speech files with sustained vowels are between 1 and 3 s. However, the dataset consists of recordings of the /a/, /i/, and /u/ vowels from 687 normal voices and 1354 pathological voices with one or more of 71 different pathologies. Table 1 shows the dataset information regarding the SVD utilized in this study. Although the numbers normal and pathological voice samples are imbalanced, they are balanced after oversampling proposed in this paper. Then, the pathological samples are the same for the two datasets, and the 687 normal samples appear also in the balanced set.

Table 1. Number of samples in the experimental dataset.

	Imbalanced Class	Balanced Class
Number of normal voices	687	1354
Number of pathological voices	1354	1354

2.2. Overview of the Framework

Figure 1 shows the overall structure proposed in this study. The framework includes four steps, feature extraction, oversampling algorithm processing, model validation, and evaluation, all of which are used to address the class imbalance of the SVD and develop a VPD classification system. First, the MFCCs or LPCs extracted from the input signals are oversampled to create a balanced dataset. An important point in this work is to make new minority samples and obtain balanced datasets using various oversampling techniques for feature parameters such as MFCCs and LPCs. Then, various combinations of deep learning models are evaluated and validated using the balanced dataset as the final task.

2.3. Feature Extraction

The theoretical background of the MFCC [35] and LPC [35,36] feature parameters is briefly described in this section. The MFCC approach is the most widely known feature extraction method for automated VPD systems. Its steps consist of preprocessing, fast Fourier transform, mel filtering, log power determination, and discrete cosine transform (DCT). Firstly, the voice signal is pre-emphasized, framed, and windowed. Then, the

magnitude spectrum is extracted from a short-time Fourier analysis. The magnitude spectrum is obtained into a mel spectrum with an equal center frequency distribution using 24 overlapping triangular windows for the mel scale windows [24]. The square of the mel spectrum—that is, the log power of each filter bank output—is calculated [24]. Finally, the 20th-order MFCCs are extracted via the log power by applying the DCT [35,36].

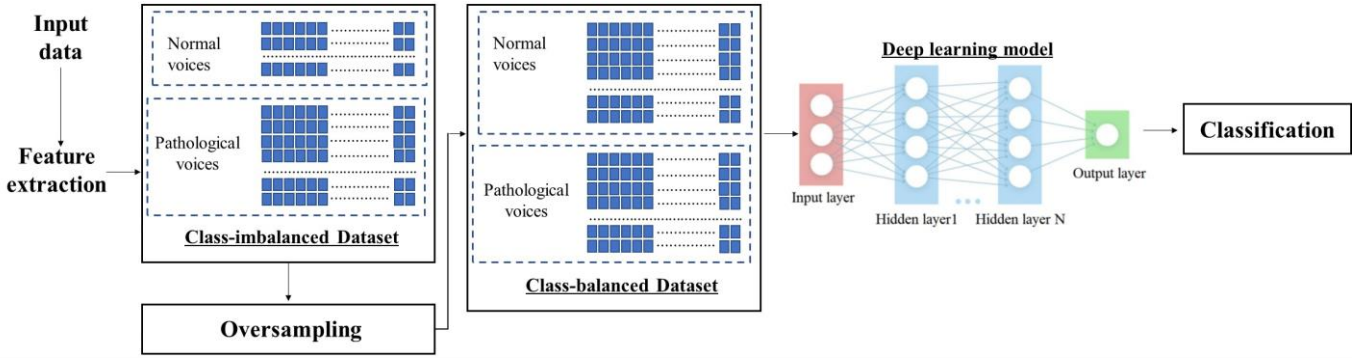


Figure 1. Structure of the proposed algorithm.

Vocal tract information of a given speech is successfully extracted through linear predictive analysis. That is, LPCs represent source behaviors that are periodic and steady. Linear prediction (LP) is the analysis approach most used in speech signal processing. LP is based on the theory that a speech sample can be approximated as a linear combination of previous samples [35]. Then, a unique set of prediction coefficients is established by minimizing the sum of the squared differences between the actual speech and the linearly predicted samples during a finite interval [35,36]. Two independent components of LP analysis, vocal tract parameters and a glottal excitation, are obtained. They are referred to as LP coefficients and LP residuals, respectively. It is assumed that speech is generated by the excitation of an impulse train and random noise for voiced and unvoiced speech, respectively [36,37]. In this research, the 20th MFCC and LPC parameters for each speech frame (window size = 40 ms; frame overlapping rate = 50%) are computed during the feature extraction stage [38].

2.4. Oversampling Methods

The SMOTE [27] generates synthetic samples for the minority class to balance the dataset. This approach can solve the overfitting problem due to random oversampling. First, the SMOTE creates synthetic data samples using the k-nearest neighbors method. It starts by randomly selecting data from the minority class and sets up the k-nearest neighbors algorithm for the data. Synthetic data are then created between the random data and the randomly selected k-nearest neighbors data [27]. Figure 2 shows the working procedure of the SMOTE. The SMOTE does not consider the data locations of the adjacent majority class while generating the synthetic data of the minority class. Therefore, the classes can overlap or create noise, so this method is inefficient for the classification of high-dimensional data. However, in this experiment, the SMOTE is considered effective because it is a binary classification technique that classifies pathological and normal voice samples.

ADASYN [30] is a generalized form of the SMOTE algorithm. This algorithm also intends to oversample data of the minority class by creating synthetic data. However, ADASYN differs from the SMOTE in that it considers the density distribution to determine the number of synthetic data. The advantage of this algorithm is that it can adaptively change its decision boundaries based on challenging samples. Figure 3 shows the ADASYN procedure.

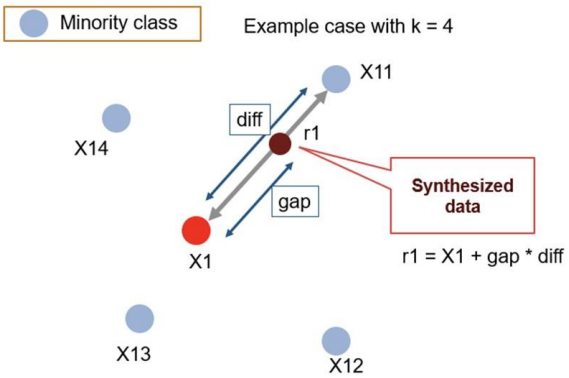


Figure 2. Working procedure of SMOTE. Source: https://github.com/minoue-xx/Oversampling-Imbalanced-Data (accessed on 20 February 2023).

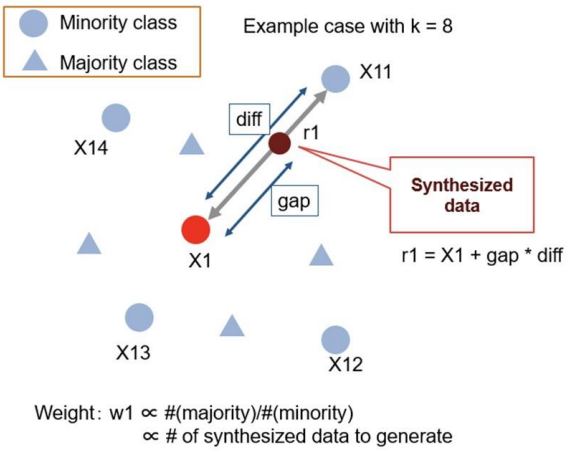


Figure 3. Working procedure of ADASYN. Source: https://github.com/minoue-xx/Oversampling-Imbalanced-Data (accessed on 20 February 2023).

If observations that appear in multiple classes away from the minority classes are present, they create multiple classes and line bridges, causing problems for the SMOTE. Borderline-SMOTE [29] addresses the above issue. Figure 4 shows the Borderline-SMOTE procedure. The starting point of this algorithm is to classify the minority class observations. If the observations are disregarded while generating synthetic data and the majority class include all neighbors, any minority observations are classified as noise points. The algorithm also classifies minority points as boundary points with both multiple and minority layers as neighbors and resamples them completely.

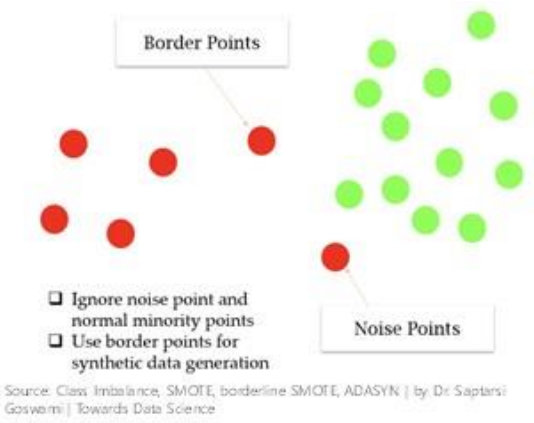


Figure 4. Working procedure of Borderline-SMOTE.

3. Results

3.1. Experimental Setup

The 20-dimensional MFCC and LPC parameters are obtained from each speech file of the SVD. The modules used to extract the MFCC and LPC are set to a frame size with 40 ms and a frame overlapping rate of 50%. The SMOTE, ADASYN, and Borderline-SMOTE algorithms with the four nearest neighbors are used (k = 4), with a total of 42 random states. We experimented SMOTE, ADASYN, and Borderline-SMOTE algorithms with k = 3, 4, and 5. When k was equal to 4, it was confirmed that the best performance was shown. This study focuses on the effectiveness of using the proposed method to generate minority class samples rather than controlling the hyperparameters of the deep learning classifier to achieve optimal classification performance. Table 2 lists the main parameters of each model. First, we study the classification between normal and pathological voices using an FNN with two hidden layers. A rectified linear unit (ReLU) is activated after the first layer, and the activation of softmax function occurs after the last layer [38]. The main parameter values are shown in Table 2. This study also uses a CNN with four consecutively convolutional layers, in which the convolutional mask has a kernel of size 3 × 3 and ReLU activation functions with 64, 64, 32, and 32 convolutional masks for each layer. The CNN also has four max pooling layers with sizes of 2 × 2, one dense layer with 512 nodes where each node has an ReLU activation function, and one softmax output layer with four neurons. The details are presented in Table 2. Experimental results are obtained through 10-fold cross-validation to ensure that each fold of the training (70%) and testing (30%) data contains at least one sample from the minority class. The models are implemented using Python 3.7 with the scikit-learn, imbalanced-learn, and PyTorch libraries.

Table 2. The main parameters of each model.

Main Hyperparameters	CNN	FNN
Activation function	ReLU	ReLU
Kernel size	(3, 3)	·
Optimizer	SGD + momentum	SGD + momentum
Number of epochs	100	100
Loss function	Cross-entropy	Cross-entropy
Dropout	0.3	·
Pooling window	Max pooling (2,2)	·
Neurons in the dense layer	512	·
Learning rate	0.001	0.00001

3.2. Model Evaluation Measures

Accuracy is one of the most common tools used to measure classifier performance. However, general measurements of classifier performance are inadequate for imbalanced datasets because the results can be biased for one class. Selecting the wrong metric to evaluate the model can lead to the selection of a defective model or incorrect recognition of the expected performance of the model in the worst-case scenario. Therefore, this study uses several evaluation metrics, such as recall, specificity, G, and F1, to determine the performance of the model using an imbalanced class dataset. Recall indicates the accuracy over the minority class. Specificity measures the accuracy over majority class. Additionally, the G and F1 values consider the classification performance achieved for the minority class [24]. In short, these evaluation metrics are generally considered "unbiased" because they are unaffected by the numbers of samples obtained from different classes and can be used in the database in which the classes are unbalanced [24]. They are calculated according to Equations (1)–(5). Table 3 contains the values of TP, TN, FP, and FN. In this work, TP refers to actual normal samples predicted as normal samples and is the number of true-positive samples. FN refers to actual normal samples predicted as pathological samples and is the number of false-negative samples. FP refers to the actual pathological samples predicted as normal samples and is the number of false-positive samples. TN

means the actual pathological samples predicted as pathological samples and is the number of true-negative samples [24].

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$G = \sqrt{Recall * Specificity} \quad (4)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

Table 3. The confusion matrix.

Actual Class	Prediction Results	
	Positive Class	Negative Class
Positive class	TP	FN
Negative class	FP	TN

3.3. Oversampling Method Comparison

Figure 5 shows the fifth and fifteenth MFCC distributions of the normal voice samples obtained before and after using various oversampling methods, including the ADASYN, SMOTE, and Borderline-SMOTE on the SVD database. Figure 5a,b show the fifth and fifteenth MFCCs of the original samples, respectively. Overall, when comparing the oversampled waveform with the original waveform, the amplitude is slightly higher, and the samples are slightly tighter in the oversampled MFCCs. The different components are partially marked with circles. In Figure 5c, the samples oversampled by the SMOTE are observed in the second and third circles. This waveform is more densely oversampled than the Borderline-SMOTE waveform of the original sample, as shown in Figure 5a. In Figure 5e, it is possible to observe the aspects oversampled by ADASYN in the last circle. When oversampled with Borderline-SMOTE, the dense waveforms can be observed in the first circle of Figure 5g. When oversampling the fifth MFCCs extracted from the normal voices using the three methods discussed, the SMOTE best interpolates the samples while maintaining the characteristics of the original MFCC of the fifth waveform, and the fifth MFCC oversampled by ADASYN appears to be the most similar to the original MFCC of the fifth waveform. In the case of the fifteenth MFCCs, in Figure 5d, the samples oversampled by the SMOTE are clearly observed in the first and second circles. When oversampling the fifteenth MFCCs extracted from normal voices using the three methods mentioned, the SMOTE best interpolates the samples, allowing for tight waveforms. The fifteenth MFCC oversampled by Borderline-SMOTE appears to be most similar to the original MFCC of the fifteenth waveform.

Figure 6 shows the fifth and fifteenth LPC distributions obtained for the normal voice samples before and after using various oversampling methods, including the ADASYN, SMOTE, and Borderline-SMOTE, on the SVD database. Figure 6a,b show the original LPCs of the fifth and fifteenth samples, respectively. As is true of the aspects observed in the MFCCs, when comparing the oversampled waveform with the original waveform, the amplitude is slightly higher, and the samples are slightly tighter in the oversampled LPCs. The different components are partially marked with circles. In the LPCs oversampled by Borderline-SMOTE in Figure 6g,h, significant differences are observed in all circles.

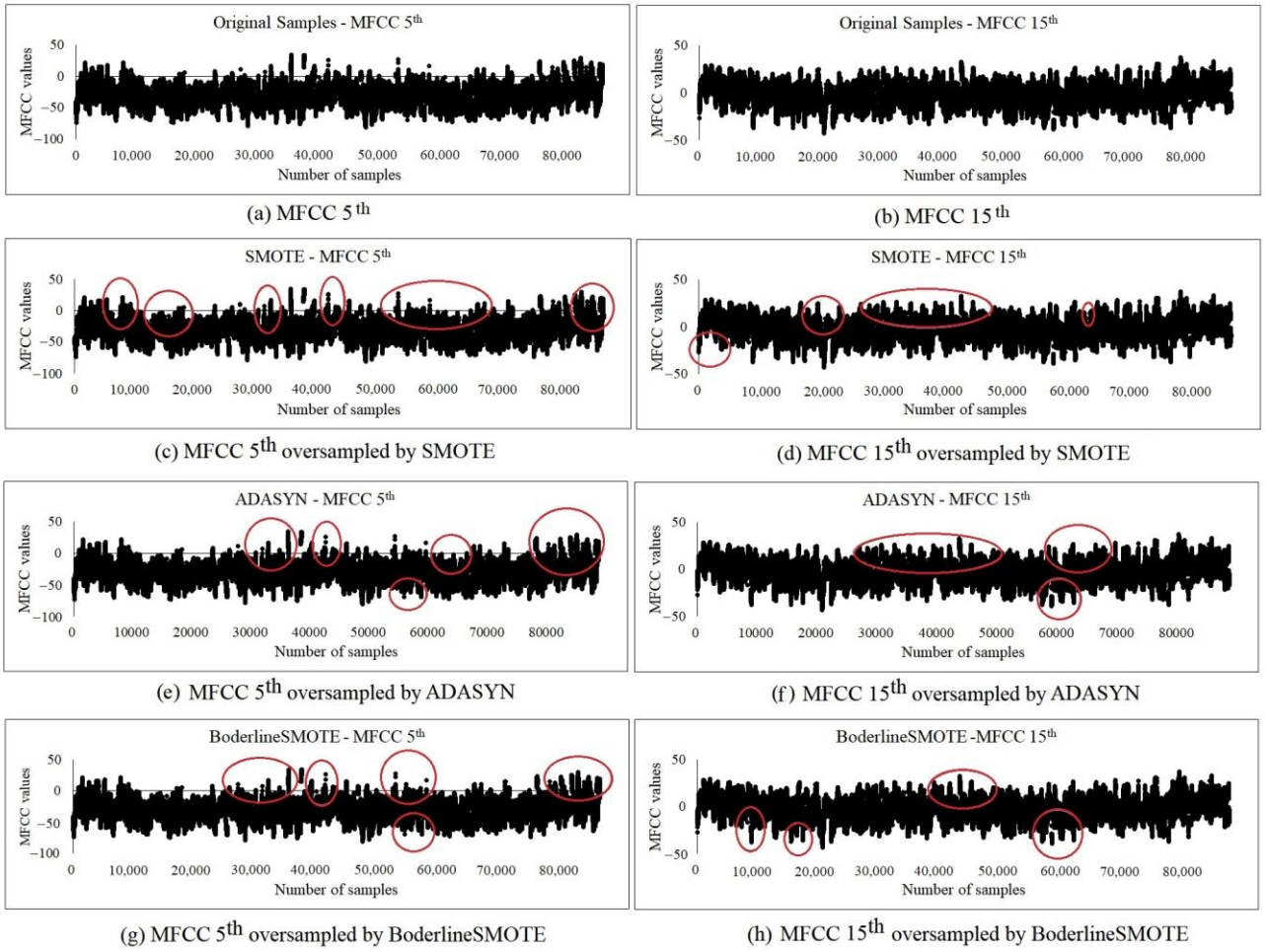


Figure 5. MFCCs of the fifth and fifteenth normal voice samples oversampled using various oversampling methods.

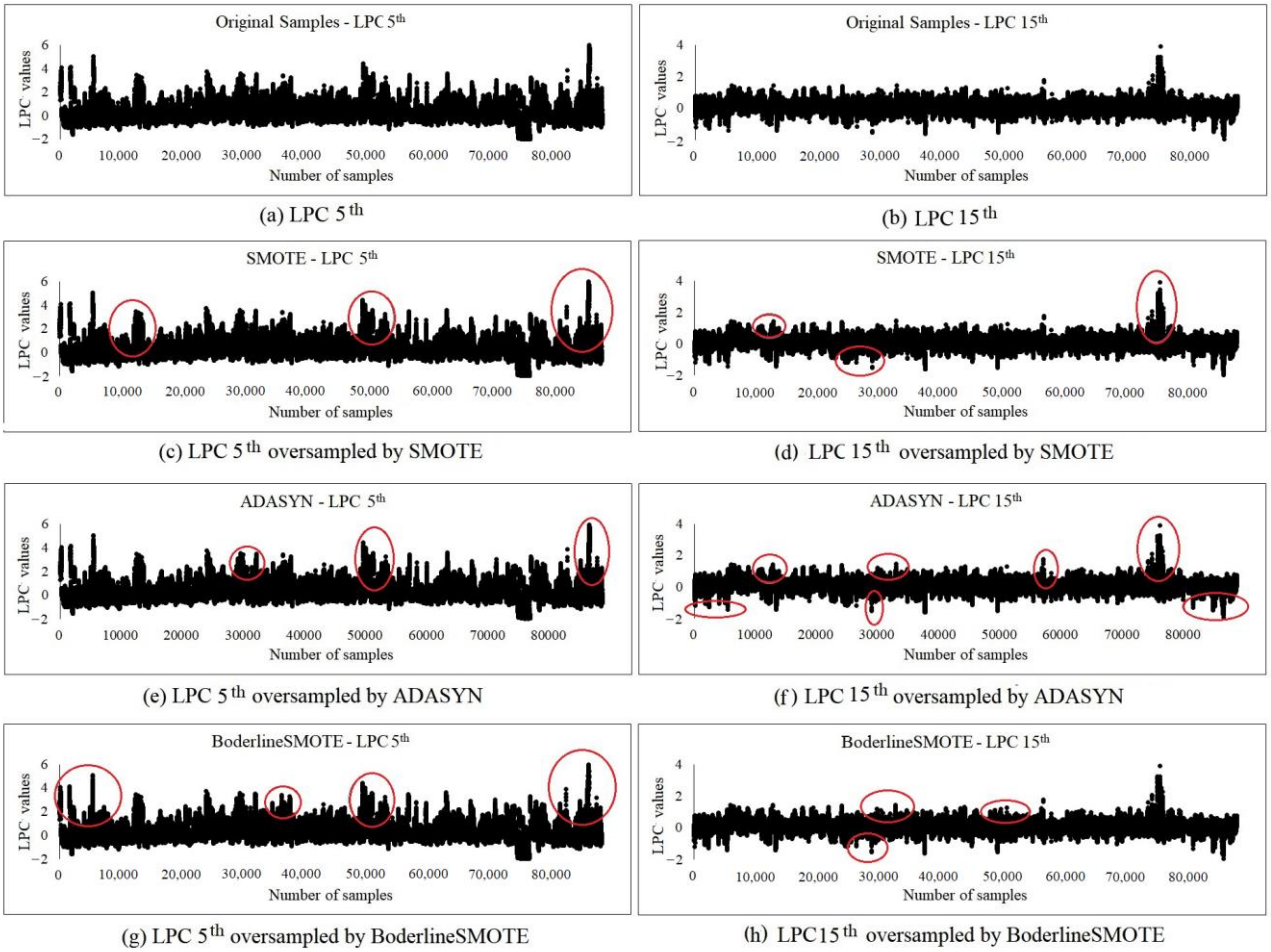
Similarly, based on the results of other oversampling methods, such as ADASYN and the SMOTE, prominent differences are observed in the various circles shown in Figure 6c–f. Therefore, among the three methods, Borderline-SMOTE best interpolates the samples while maintaining the characteristics of the original LPC waveform.

The fifth and fifteenth percentiles of the MFCCs and LPCs are randomly selected to show the oversampling phenomenon visually. Overall, as shown in Figures 5 and 6, the samples oversampled for the LPC tend to be better interpolated than those of the MFCC. In both cases, the samples oversampled using the ADASYN method tend to be similar to the original waveform. Consequently, an equal number of normal and pathological voice samples are produced. Although the data lengths in (c) to (h) appear to match those in (a) and (b) in Figures 5 and 6, this is because the oversampled samples overlap at the same time and are plotted together. In reality, the normal voice data samples are as numerous as the pathological voice data samples.

3.4. Experimental Results and Analysis

The model evaluation measures are obtained through 10-fold cross-validation with the specificity, recall, G, and F1 value. Table 4 presents the results measured by the FNN and CNN models before using oversampling methods. The performances of all models are poor. In SVD, most model evaluation metrics are lower than 0.6 because the number of class distribution in the sample is unbalanced. Figure 7 shows the confusion matrix produced by each deep learning model using feature parameters to classify pathological and normal voices in the initial imbalanced class dataset. As shown in Table 4 and Figure 7,

the recognition rate of classifiers is biased toward samples of the majority class composed of pathological voices, whereas the accuracy for samples of the minority class composed of normal voice is insufficient. Therefore, the model misleads the overall accuracy results. From Figures 7–10, 72.96% accuracy is shown for the combination of MFCCs and the FNN. Additionally, accuracies of 72.31% and 71.34% are obtained with a combination of LPCs and MFCCs using the FNN and CNN, respectively. The lowest performance of 66.94% is obtained using the LPCs and CNN, as shown in Figure 10a.



(a) LPC 5 th (b) LPC 15 th

(c) LPC 5 th oversampled by SMOTE (d) LPC 15 th oversampled by SMOTE

(e) LPC 5 th oversampled by ADASYN (f) LPC 15 th oversampled by ADASYN

(g) LPC 5 th oversampled by BoderlineSMOTE (h) LPC15 th oversampled by BoderlineSMOTE

Figure 6. LPCs of the fifth and fifteenth normal voice samples oversampled by various oversampling methods.

Table 5 shows the model evaluation matrices obtained from each classifier after performing imbalanced handling using various oversampling algorithms. The results in terms of binary class confusion matrices are shown in Figures 8 and 9. Model evaluations of FNN and CNN classifiers described in Table 5 show that VPD systems with SMOTE have better specificity, recall, G value, and F1 value. Compared to various deep learning models and the feature parameters, the combination of the CNN and LPCs oversampled by the SMOTE has the highest accuracy (98.89%), and the best overall performance is achieved when evaluating the model metrics with the implementation of the SMOTE algorithm. The optimal classifier, such as the CNN with LPCs oversampled by the SMOTE, increases the recall, the specificity, the G value, and the F1 value to 1.0, 0.97, 0.98, and 0.99, respectively. These studies are conducted on an imbalanced dataset, with significant deviations ranging from 0.20 to 0.73 between the recall and specificity. Our proposed method, which combines the CNN and LPCs oversampled by the SMOTE, improves the recall and specificity from 0.01 to 0.73 and 0.00 to 0.28, respectively, compared to the performance of other conventional methods.

Table 4. Evaluation measures obtained by each model on the class-imbalanced SVD.

FNN		
MFCC	Recall	0.43
	Specificity	0.88
	G value	0.62
	F1 value	0.51
LPC	Recall	0.35
	Specificity	0.90
	G value	0.56
	F1 value	0.45
CNN		
MFCC	Recall	0.44
	Specificity	0.87
	G value	0.62
	F1 value	0.51
LPC	Recall	0.27
	Specificity	0.95
	G value	0.51
	F1 value	0.40

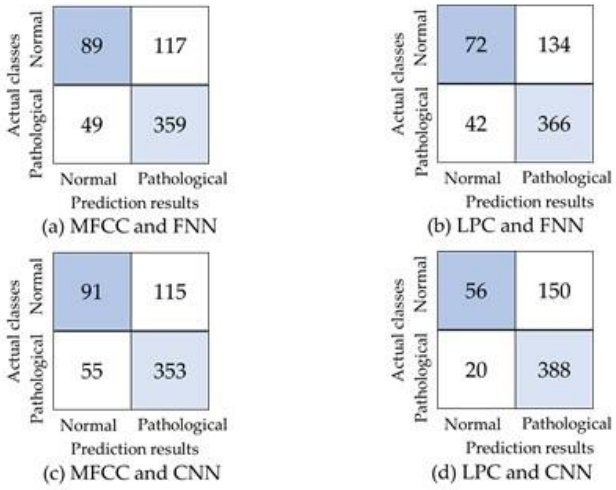


Figure 7. The confusion matrices produced for the MFCCs and LPCs by the FNN and CNN on the initial class-balanced dataset.

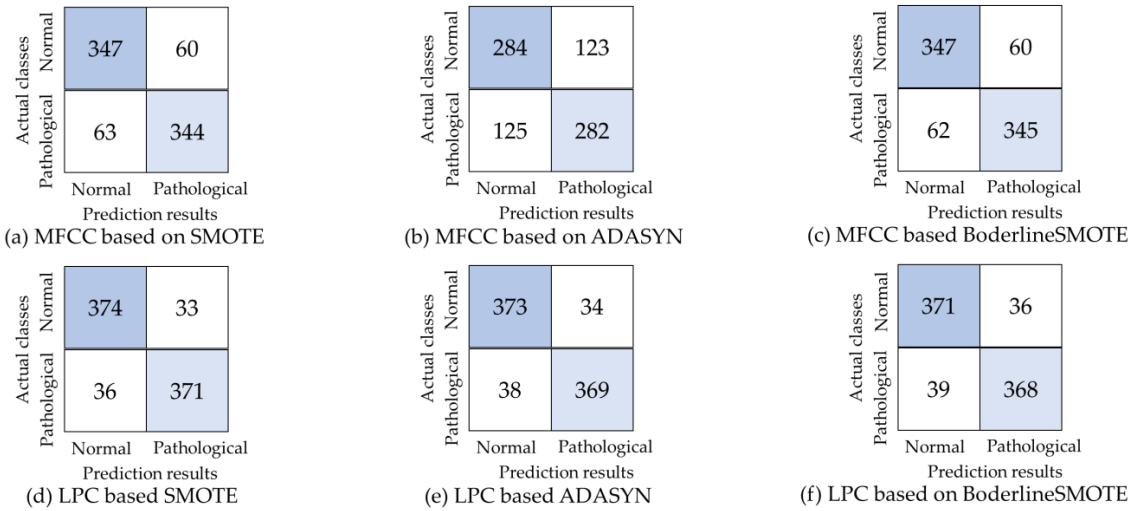


Figure 8. The confusion matrices obtained by the FNN on the class-balanced dataset under various conditions.

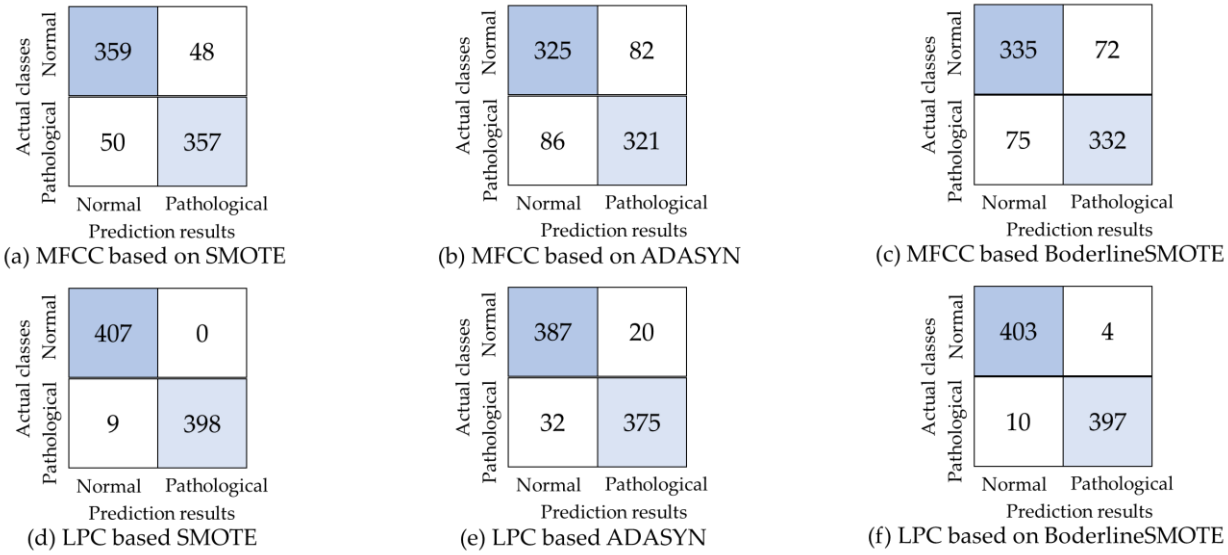


Figure 9. The confusion matrices produced by the CNN on the class-balanced dataset under various conditions.

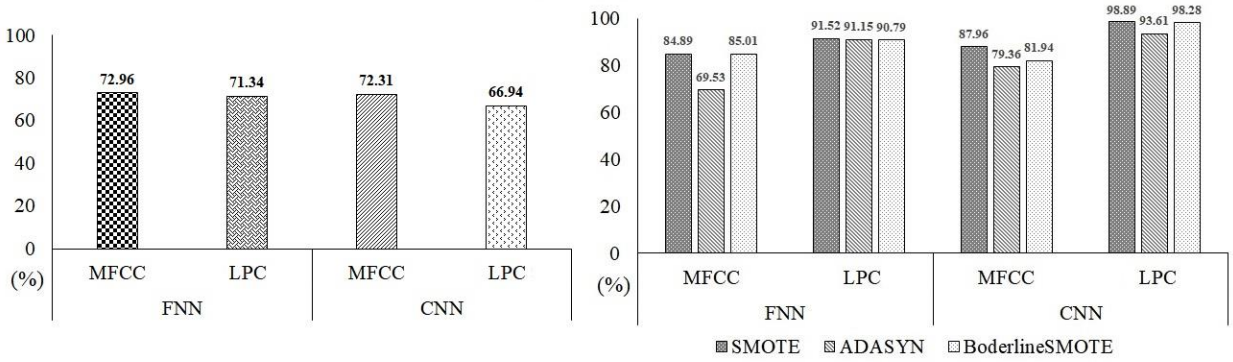


Figure 10. Performance comparison among various combinations of methods on the class-imbalanced and class-balanced dataset.

Table 5. Evaluation measures produced by each model on the class-balanced SVD.

FNN		SMOTE	ADASYN	Borderline-SMOTE
MFCC	Recall	0.85	0.7	0.85
	Specificity	0.85	0.69	0.84
	G value	0.85	0.69	0.85
	F1 value	0.85	0.69	0.85
LPC	Recall	0.92	0.92	0.91
	Specificity	0.91	0.91	0.9
	G value	0.91	0.91	0.9
	F1 value	0.91	0.91	0.9
CNN				
MFCC	Recall	0.88	0.8	0.82
	Specificity	0.88	0.78	0.81
	G value	0.88	0.79	0.81
	F1 value	0.88	0.8	0.82
LPC	Recall	**1.0**	0.95	0.99
	Specificity	**0.97**	0.92	0.98
	G value	**0.98**	0.94	0.98
	F1 value	**0.99**	0.93	0.98

Figures 8 and 9 show the confusion matrices of each deep learning model using feature parameters to classify normal and pathological voices in the balanced dataset. Comparing the distributions of the confusion matrices in Figures 7–9, we find that the classification performance for majority class samples composed of pathological voices is lower for the VPD system using the FC-SMOTE algorithm, but it signiFIcantly improves its ability to classify minority class samples composed of normal voices.

Figure 10b demonstrates that utilizing the CNN classifier with the LPCs oversampled by the SMOTE yields the highest accuracy at 98.89% compared to combinations of the other deep learning classifiers and feature parameters. Additionally, the next-highest performances (98.28% and 93.61%) are obtained via the LPCs oversampled by Borderline-SMOTE and ADASYN, respectively, with the CNN classifier. In the FNN classifier, the best accuracy (91.52%) is achieved with the LPCs oversampled by the SMOTE. The combination of the MFCCs oversampled by ADASYN and the FNN classifier yields the lowest performance of 69.53%.

Overall, as a deep learning model, the CNN outperforms the single deep learning classifier for disordered voice detection, which has also been featured in the most recent published paper on this topic [6]. Additionally, the proposed method, such as the combination of the LPCs oversampled by the SMOTE and the CNN in the binary class confusion matrix of two models, is generally superior in terms of the exact prediction for each normal and pathological voice. The experimental results indicate that the SMOTE is a useful approach for building a binary classification model between pathological and normal voices. It also confirms that our suggested VPD algorithm can train minority classes better and can achieve an improved binary classification performance.

Because the MFCC is widely used in speech signal processing, the VPD system can also achieve good classification and detection performances on imbalanced datasets, as demonstrated in a recent study [39,40]. Therefore, the performances obtained from the MFCCs are better than those of the LPCs when using the two deep learning models in the class-imbalanced binary classification, as shown in Figure 10a. The performance of the FNN is approximately 6% better than that of the CNN. However, regarding the accuracies obtained on the balanced class dataset, all the models represent good predictive ability in the results of binary classification for normal and pathological voices, with LPCs oversampled by various methods. The deep learning model using the CNN achieves the best performance overall. Experimentally, the LPC appears to be more sensitive to oversampling methods than the MFCC. In conclusion, the CNN (as a deep learning model) and LPCs oversampled by the SMOTE as feature parameters obtain the highest performance in classification between pathological and normal voice using the SVD. Notably, the classification rate of the VPD system configured using the SMOTE method demonstrates considerable improvement compared to the results of the non-sampling. Therefore, we conclude that the proposed method, such as the combination of the CNN and the LPCs oversampled by the SMOTE, can efficiently increase the performance for classification between the pathological and normal voices.

In summary, our VPD system uses the SMOTE, ADASYN, and Boderline-SMOTE algorithms to generate the binary imbalanced class data in the SVD, and the accuracy of the algorithm is verified using a set of deep learning classifiers such as FNN and CNN. To classify normal and pathological voices, when compared with a VPD system without various oversampling algorithms, all performances of our VPD system with the SMOTE method with the recall, specificity, G value, and F1 value are higher than those of the former. These results confirm that our proposed method is a good strategy for achieving successful classification between pathological and normal voices. It also justifies that our methods for solving class imbalances in a limited pathological speech database can be applied to pathological speech detection in the field of biomedical engineering.

3.5. Comparison with Existing Techniques

In the previous subsections, the performance of our method is compared with that of deep learning methods using vocal tract-based cepstral features and their combinations as references. The results in Figure 10b and Table 6 show that the LPC combination based on the SMOTE + CNN yields the best overall detection accuracy. In this subsection, this optimal combination is compared with existing methodologies and deep learning techniques. Many studies have developed different VPD techniques over the past few decades; in our work, we select four studies with databases and deep learning methods similar to those used in our research, as shown in Table 6.

Table 6. Comparative analysis between related works tested on various datasets.

Work	Feature	Database	Methodology	Accuracy
[2]	MFCC	SVD	BPGAN and GAN	87.60%
[24]	MFCC	MEEI	FC-SMOTE and RF	100%
	MFCC	SVD	FC-SMOTE and CNN	90%
[22]	·	SVD	IFCM and CGAN	95.15%
[26]	Spectrogram	Spanish Parkinson's Disease Dataset (SPDD)	Semi supervised GAN	96.63%
Proposed method	MFCC and LPC	SVD	LPC based on the SMOTE and CNN	98.89%

Table 6 presents the databases, methodologies, features, deep learning methods, and performances of competing approaches under the binary detection model scenario for voice disorder detection. In [22], the CGAN-IFCM algorithm achieves an accuracy of 95.15%, with 869 normal and 1356 pathological voice data in the SVD. Although our study also uses all the SVD data, we are not sure why the number of points differs from the total amount of data in this study. The authors of [24] prove that the proposed FC-SMOTE method outperforms the other oversampling methods by 100% and 90% in terms of the accuracy of the CNN model with the MEEI and SVD, respectively. Then, for the SVD data, 687 normal and 194 pathological voice data are used, and pathological voice data are augmented to 687 via oversampling. Although the approach in [26] demonstrated good performance, that study used the SPDD. In a comparison with [2,22,24] using the SVD, the proposed combination containing LPCs based on the SMOTE and the CNN improved the resulting accuracy by 3.74–11.29%. Compared to [26], using the SPDD, our proposed method increased the performance by 2.26%. All existing works [2,22,24,26] solved imbalanced class issues with various features, databases, and methodologies. In conclusion, our method (LPCs based on the SMOTE and the CNN) has improved accuracy by up to 11.29% compared to papers using the same parameters (MFCC) and database (SVD) as our experiments.

4. Conclusions

This paper recommends a VPD system combined with various imbalanced learning algorithms. By analyzing the distributions between and within the sample classes, various oversampling methods are applied to minority class samples. This study also proposes oversampled MFCCs and LPCs as the input parameters of the deep learning model.

First, we perform empirical and numerical experiments on SVD. We evaluate and validate deep learning model sets using the resulting balanced class dataset as input and select a reasonable metric set as the model evaluation scale. The experimental results are evaluated using several model evaluation metrics such as recall, specificity, G, and F1, which are reasonable measures for imbalanced class learning. The experimental results suggest that the VPD system without the various oversampling algorithms has a very poor performance in classifying minority and majority classes and struggles to distinguish pathological voices from normal voices. The VPD system with the CNN and LPCs oversampled by the SMOTE is a useful method for building the deep learning model for pathological voice classification, which means that the VPD model can train minority classes better

and achieve improved performance. Vocal tract information of pathological and normal voices is successfully extracted through linear predictive analysis. Finally, the performances of oversampling algorithms such as the SMOTE, ADASYN, and Borderline-SMOTE are discussed though waveform analysis in Figures 5 and 6. In pathological voice detection, the SMOTE achieves better performance than traditional oversampling methods.

In the future, our method may help clinicians diagnose pathological voices, solving problems related to insufficient medical resources and shortening the diagnosis time required for patients. In addition, the SMOTE can learn more pathological voices to form an effective VPD system for the classification of a wider range of class-imbalanced pathological voices encountered in multi-classification problems. In addition, we will undertake a study to classify disease types and grade, roughness, breathiness, asthenia, strain (GRBAS) scales of pathological voices.

Author Contributions: Conceptualization, J.-Y.L.; methodology, J.-Y.L.; software, J.-Y.L. and J.-N.L.; validation, J.-Y.L. and J.-N.L.; writing—original draft preparation, J.-Y.L. and J.-N.L.; writing—review and editing, J.-Y.L.; visualization, J.-Y.L.; funding acquisition, J.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2022R1H1A209259811). The sponsor had no involvement in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miliaresi, I.; Poutos, K.; Pikrakis, A. Combining acoustic features and medical data in deep learning networks for voice pathology classification. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 1190–1194.
2. Jinyang, Q.; Denghuang, Z.; Ziqi, F.; Di, W.; Yishen, X.; Zhi, T. Pathological Voice Feature Generation Using Generative Adversarial Network. In Proceedings of the 2021 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD), Nanjing, China, 21–23 October 2021.
3. Khan, M.A.; Kim, J. Toward Developing Efficient Conv-AE-Based Intrusion Detection System Using Heterogeneous Dataset. *Electronics* **2020**, *9*, 1771. [CrossRef]
4. Al-Dhief, F.T.; Latiff, N.M.A.; Malik, N.N.N.A.; Salim, N.S.; Baki, M.M.; Albadr, M.A.A.; Mohammed, M.A. A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access* **2020**, *8*, 64514–64533. [CrossRef]
5. Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A.; Khanapi Abd Ghani, M.; Maashi, M.S.; Garcia-Zapirain, B.; Oleagordia, I.; Alhakami, H.; AL-Dhief, F.T. Voice Pathology Detection and Classification Using Convolutional Neural Network Model. *Appl. Sci.* **2020**, *10*, 3723. [CrossRef]
6. Verde, L.; de Pietro, G.; Ghoneim, A.; Alrashoud, M.; Al-Mutib, K.N.; Sannino, G. Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 Through Speech and Voice Analysis. *IEEE Access* **2021**, *9*, 65750–65757. [CrossRef]
7. Zhang, T.; Wu, Y.; Shao, Y.; Shi, M.; Geng, Y.; Liu, G. A Pathological Multi-Vowels Recognition Algorithm Based on LSP Feature. *IEEE Access* **2019**, *7*, 58866–58875. [CrossRef]
8. Alhusseimn, M.; Muhammad, G. Automatic Voice Pathology Monitoring Using Parallel Deep Models for Smart Healthcare. *IEEE Access* **2019**, *7*, 46474–46479. [CrossRef]
9. Verde, L.; de Pietro, G.; Alrashoud, M.; Ghoneim, A.; Al-Mutib, K.N.; Sannino, G. Leveraging Artificial Intelligence to Improve Voice Disorder Identification Through the Use of a Reliable Mobile App. *IEEE Access* **2019**, *7*, 124048–124054. [CrossRef]
10. Verde, L.; de Pietro, G.; Sannino, G. Voice Disorder Identification by Using Machine Learning Techniques. *IEEE Access* **2018**, *6*, 16246–16255. [CrossRef]
11. Eye, M.; Infirmary, E. *Voice Disorders Database, Version. 1.03 (cd-rom)*; Kay Elemetrics Corporation: Lincoln Park, NJ, USA, 1994.

12. William, J.B.; Manfred, P. *Saarbrucken Voice Database*; Institute of Phonetics, Univ. of Saarland: Saarbrücken, Germany, 2007. Available online: http://www.stimmdatenbank.coli.uni-saarland.de/2007 (accessed on 29 December 2022).
13. Islam, R.; Tarique, M.; Abdel-Raheem, E.A. Survey on Signal Processing Based Pathological Voice Detection Techniques. *IEEE Access* **2020**, *8*, 66749–66776. [CrossRef]
14. Reddy, M.K.; Alku, P.A. Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation. *IEEE Access* **2021**, *9*, 135953–135963. [CrossRef]
15. Hemmerling, D.; Skalski, A.; Gajda, J. Voice data mining for laryngeal pathology assessment. *Comput. Biol. Med.* **2016**, *9*, 270–276. [CrossRef] [PubMed]
16. Naranjo, L.; Perez, C.J.; Martin, J.; Campos-Roca, Y.A. two-stage variable selection and classification approach for Parkin-son's disease detection by using voice recording replications. *Comput. Methods Prog. Biomed.* **2021**, *142*, 147–156. [CrossRef]
17. Wu, Y.; Zhou, C.; Fan, Z.; Wu, D.; Zhang, X.; Tao, Z. Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection. *IEEE Access* **2021**, *9*, 30–44. [CrossRef]
18. Tuncer, T.; Dogan, S.; Özyurt, F.; Belhaouari, S.B.; Bensmail, H. Novel Multi Center and Threshold Ternary Pattern Based Method for Disease Detection Method Using Voice. *IEEE Access* **2020**, *8*, 84532–84540. [CrossRef]
19. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
20. Vandewiele, G.; Dehaene, I.; Kovács, G.; Sterckx, L.; Janssens, O.; Ongenae, F.; de Backere, F.; de Turck, F.; Roelens, K.; Decruyenaere, J.; et al. Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artif. Intell. Med.* **2020**, *111*, 101987–102003. [CrossRef] [PubMed]
21. Jing, X.Y.; Zhang, X.; Zhu, X.; Wu, F.; You, X.; Gao, Y.; Shan, S.; Yang, J.Y. Multiset feature learning for highly imbalanced data classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 139–156. [CrossRef]
22. Chui, K.T.; Lytras, M.; Vasant, P. Combined Generative Adversarial Network and Fuzzy C-Means Clustering for MultiClass Voice Disorder Detection with an Imbalanced Dataset. *Appl. Sci.* **2020**, *10*, 4571. [CrossRef]
23. Fan, Z.; Qian, J.; Sun, B.; Wu, D.; Xu, Y.; Tao, Z. Modeling Voice Pathology Detection Using Imbalanced Learning. In Proceedings of the 2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD), Xi'an, China, 15–17 October 2020; pp. 330–334.
24. Fan, Z.; Wu, Y.; Zhou, C.; Zhang, X.; Tao, Z. Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method. *Appl. Sci.* **2021**, *10*, 3450. [CrossRef]
25. Esmaeilpour, M.; Cardinal, P.; Koerich, A.L. Unsupervised Feature Learning for Environmental Sound Classification Using Weighted Cycle-Consistent Generative Adversarial Network. *Appl. Soft Comput.* **2020**, *86*, 105912–105943. [CrossRef]
26. Trinh, N.H.; O'Brien, D. Semi-Supervised Learning with Generative Adversarial Networks for Pathological Speech Classification. In Proceedings of the 2020 31st Irish Signals and Systems Conference (ISSC), Letterkenny, Ireland, 11–12 June 2020; pp. 1–5.
27. Chawla, N.V.; Bowyer, K.W.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
28. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106–121. [CrossRef] [PubMed]
29. Hui, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 330–334.
30. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
31. Dong, Y.; Wang, X. A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Dalian, China, 10–12 August 2013; pp. 10–12.
32. Douzas, G.; Bacaoa, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [CrossRef]
33. Wei-Chao, L.; Chih-Fong, T.; Ya-Han, H.; Jing-Shang, J. Clustering-based undersampling in-class imbalanced data. *Inf. Sci.* **2017**, *409–410*, 17–26.
34. Gautheron, L.; Habrard, A.; Morvant, E.; Sebban, M. Metric Learning from Imbalanced Data with Generalization Guarantees. *Pattern Recognit. Lett.* **2020**, *133*, 298–304. [CrossRef]
35. Alim, S.A.; Alang Rashid, N.K. Some Commonly Used Speech Feature Extraction Algorithms. Available online: https://www.intechopen.com/chapters/63970 (accessed on 29 December 2022).
36. Makhuol, J. Linear Prediction: A Tutorial Review. *Proc. IEEE* **1975**, *63*, 561–580. [CrossRef]
37. Kumar, S. Real-time implementation and performance evaluation of speech classifiers in speech analysis-synthesis. *ETRI J.* **2020**, *43*, 82–94. [CrossRef]
38. Lee, J.Y. Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database. *Appl. Sci.* **2021**, *11*, 7149. [CrossRef]

39. Kadiri, S.R.; Alku, P. Analysis and Detection of Pathological Voice Using Glottal Source Features. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 367–379. [CrossRef]
40. Amami, R.; Smiti, A. An incremental method combining density clustering and support vector machines for voice pathology detection. *Comput. Electr. Eng.* **2017**, *57*, 257–265. [CrossRef]