Analysis of Frequency Range Effect on the Detection of Voice Disorder using Convolutional Neural Networks Trained on Spectogram Images

José Alberto Souza Paulino[1], Herman Martins Gomes[1], Leonardo Vidal Batista[2], Leonardo Wanderley Lopes[3]
[1]Department of Systems and Computing, Federal University of Campina Grande (UFCG)
{[2]Department of Computer Systems, [3]Integrated Voice Studies Laboratory}, Federal University of Paraiba (UFPB)
Email: josealberto@copin.ufcg.edu.br, hmg@dsc.ufcg.edu.br, leonardo@ci.ufpb.br, lwlopes@lievlab.com

***Abstract*—Considering the current advancements in signal processing and machine learning (ML), non-invasive techniques for assessing vocal quality have become increasingly popular, especially with the use of spectrograms in acoustic analysis, which typically do not evaluate patterns in regions above 5kHz, either through visual inspection or using ML algorithms. This study aims to assess the relevance of different frequency ranges in classifying healthy and disordered voices using convolutional neural networks (CNN), as well as to investigate whether the combination of frequency ranges can improve classification results. To achieve this goal, spectrogram subsets were generated from 16 frequency ranges in two datasets, obtained through a bank of band-pass filters, and trained in CNN models with transfer learning. The study was conducted by first evaluating the relevance of each frequency range individually. Then, the results of the 65,536 possible combinations obtained with the 16 frequency ranges were assessed. This analysis revealed that it is possible to characterize voice pathology patterns in frequency regions above 5kHz, but the interval between 1 to 1,462 Hz is substantially better in terms of descriptive capacity in spectrograms. Additionally, it was observed that high-frequency regions, when combined with other frequency ranges, produce better classification results, improving the test accuracy from 80.53% to 82.10% in the SVD dataset and from 78.11% to 82.12% in the AVFAD dataset.**

I. Introduction

Acoustic analysis has been employed in clinical practice as a complementary strategy to auditory-perceptual analysis, and due to its non-invasive nature, it has become an increasingly prominent topic in voice research. In this type of analysis, the spectrogram is a very useful tool for evaluating changes in the voice production mechanism. As a result of severals advances in signal processing methods, notably in the field of machine learning (ML) and computer vision, there are numerous efforts published in the literature aimed at classifying voice signals using ML algorithms, particularly making use of Convolutional Neural Networks (CNN) for spectrogram classification, as in [1]–[3].

It is noteworthy that due to human limitations in acoustic analysis to identifying patterns in higher frequency ranges, regions corresponding to low frequency ranges are commonly adopted in the analysis or classification of spectrograms, as described in [4]. There is no consensus about a threshold to characterize what constitutes high frequency in voice signals, so this nomenclature is often associated with different frequency ranges in the literature. In [5], various studies analyzing the influence of frequency ranges on the perception and production of human voice sounds were examined, and based on their findings, the authors refer to high frequencies as those above 8 kHz.

When it comes to the clinical analysis of spectrograms to evaluate voice characteristics, regions above 5 kHz are rarely considered. Some authors remove regions above 5 kHz in the spectrogram, as observed in [6], where this cutoff was made without clear scientific evidence. Historically, energy characteristics in frequencies above 5 kHz in the spectrum have been neglected in various voice studies because the portion corresponding to the low-frequency regions of the speech spectrum was considered sufficient from a perceptual standpoint, as stated in [4]. Furthermore, the inspection of a typical voice spectrogram reveals that acoustic energy in frequencies above 5 kHz tends to decrease sharply as the frequency range increases. However, it is important to highlight that frequency characteristics of the voice signal may vary as a result of other factors, such as age, language and gender, as discussed in [7] and [8].

Factors like these reduce the interest in expending efforts to investigate the relevance of high-frequency regions. However, evidences of the existence of relevant energy patterns in these regions of the spectrum have been reported in some studies. For example, the findings described in [9], related to the analysis of four types of voice disorders in high-frequency regions, have shown that the octave centred at 8 kHz (between 5,657-11,314 kHz) could be more relevant for distinguishing the studied disorders. Similarly in [10], it was identified that when comparing healthy and dysphonic voices, there is a higher concentration of energy in the ranges between 6 kHz and 16 kHz for voices with dysphonia.

These empirical evidence [9], [10] suggest that, for certain voice disorders, discriminative patterns can be found in higher frequencies, and highlight a gap in the scientific literature regarding the investigation of the importance of these high frequencies. Thus, this study aims to evaluate the capacity of different ranges of the frequency spectrogram in the task of classifying healthy voices from voices with organic disorders, witch special interest in the ranges above 5 kHz. Additionally, we investigate whether if it is possible to improve classification

results by combining different frequency ranges.

II. Related Works

Through a quantitative analysis of different methods developed for the classification of voices with disorders, conducted in [11], it is possible to observe a variety of approaches proposed to solve this classification problem in studies published between 2012 and 2022. This list of approaches comprises different classification algorithms, various databases, different partitioning strategies, and distinct procedures for subsampling signals. The wide variety of existing approaches also reflects in the discrepancy between the published results, with reported accuracies ranging from 67% to 100%.

Despite this diversity, there is a common characteristic among these studies, which is the lack of methodology description, in order to allow reproducibility and to guarantee reliability in the reproduction and comparison many of results. This deficiency is corroborated in [12], who made attempts to reproduce the methodologies described in published studies, but achieved results falling short of those reported by the authors.

Although there are several studies focused on the classification of voice disorders using machine learning algorithms, no efforts have been identified up to date specifically aimed at evaluating the relevance of frequency ranges for the classification of voice disorders in spectrograms. However, it is important to note that some studies provide evidence that supports the hypothesis that high-frequency regions contain descriptive potential in characterizing patterns of disordered voices. Among these, the following stand out:

1) The approach for classifying voice disorders in spectrograms using a hybrid classifier, proposed in [13], employs pre-trained CNN layers and an SVM classifier. Despite the limitations of that study regarding the number of samples and the data augmentation strategy, the authors perform a relevant analysis of the classified spectrograms using the GradCam algorithm. In the analysis, it is evident that the CNN attends to different regions of the spectrogram to characterize voice signal patterns, particularly the high-frequency regions for identifying patterns of disordered voices;
2) In [14], the influence of different frequency regions of voice signals is investigated to differentiate between two types of pathologies using correlation functions. Although the paper does not focus on spectrograms and does not use machine learning algorithms, the study contributed by evaluating the relevance of frequency ranges in the classification of voice signals. The approach [14] considered frequency ranges on a logarithmic scale, in octaves. Furthermore, the aforementioned study provided some evidence that the most relevant frequency range for the classification of disordered voice signals corresponds to the interval between 1 kHz and 8 kHz.

III. Methodology

The proposed methodology was divided into two stages. Initially, to evaluate the relevance of the frequency ranges, classifiers were trained with spectrograms generated in each of the target frequencies ranges, and the performance of these classifiers was compared. Next, it was tested whether combining the results of the classifiers could yield better outcomes than those obtained by the individual classifiers.

The experiments were conducted on two distinct datasets. Additionally, the sample selection and dataset partitioning processes were described in detail to ensure reproducibility, differently from the work reported in [6], [13], and [15]. This Section also describes the parameters adopted for generating the spectrograms, the rationale for choosing the frequency ranges, and the selection of the CNN architecture.

A. Voice Datasets

1) **SVD** [16]: of German origin, it contains recordings of 1,853 individuals, 851 are classified as controls (423 men and 428 women) and 1,002 are classified as having voice disorders (454 men and 548 women), grouped into 71 types of pathologies. The recordings for each individual include the sustained vowel emissions /a/, /i/, and /u/ at low, normal, and high intonations, along with the short phrase "Guten Morgen, wie geht es Ihnen?". The audio and electroglottography (EGG) recordings are available at a sampling rate of 50 kHz and a resolution of 16 bits.
2) **AVFAD** [17]: of European Portuguese origin, contains recordings of 709 individuals, 363 are classified as controls (113 men and 250 women) and 346 are classified as having voice disorders (97 men and 249 women). The recordings for each individual include the sustained vowel emissions /a/, /i/, and /u/, with three repetitions in the same recording. Additionally, each individual recorded the reading of six sentences from the CAPE-V (Consensus auditory-perceptual evaluation of voice), which is a scale for voice assessment. The audio are available at a sampling rate of 44 kHz and a resolution of 16 bits.

When performing a binary classification to distinguish voices from the control group and the disorder group, it is essential to consider that some types of pathologies have very small number of samples, especially in the SVD database. This can directly impact the characterization of patterns in the disorder class and create biases in this class. For this reason, this study followed the strategy outlined in [15] for selecting voice samples, in which only organic pathologies related to structural alterations in the larynx were considered. These pathologies in this database include: laryngitis, leukoplakia, Reinke's edema, recurrent laryngeal nerve paralysis, vocal fold carcinoma, and vocal fold polyps. Additionally, only recordings corresponding to the sustained vowel /a/ at normal intonation were considered.

Using only the organic pathologies in the SVD database and the sustained vowel /a/, the disorder group consisted

of 482 samples. To balance the dataset, the control group included the first 482 samples from the database (considering an alphabetical ordering by file name). After sample selection, the SVD database contained 964 recording entries, of which 20% were used for the test set through the *train_test_split* function[1] with a seed equal to 42. From the remaining samples, a new partitioning was performed using the same function to generate the training and validation sets in an 80%-20% ratio. Only a single emission of the sustained vowel per individual was considered.

Similarly, in the AVFAD database, only the sustained vowel /a/ was selected. For the samples in this database, 20% were used for the test set, and the remaining samples were partitioned in an 80%-20% ratio to generate the training and validation sets. It is important to note that during the process of data split, after sorting the files in ascending order by name, stratification was performed respecting the class and gender proportions for the new sets generated. Additionally, since the AVFAD database contains three emissions of the same sustained vowel in each recording, as a data augmentation strategy (only in the training set), these emissions were used separately as individual signals. In the validation and test sets, to avoid data leakage, only a single emission of the sustained vowel per individual was considered.

B. Data Pre-processing

The spectrogram used in this study was inspired by the spectral representation proposed in [15]. After the voice signal was resampled to 25 kHz, the Short-Time Fourier Transform (STFT) was applied to the time-domain signal to obtain its spectral-domain representation. The following parameters were adopted for the STFT: a Hamming window of size 1,024, corresponding to 40.96 ms, and a window spacing of 4 ms. Next, the magnitude of the complex-valued spectrogram is calculated to represent how energy is distributed across various frequencies over time. Finally, the spectrogram values were normalized for the entire interval between 0 and 255. The Python programming language and the Librosa[2] library were used for implementing the preprocessing described in this Section.

C. Definition of frequencies ranges

The spectral representations generated in [15], according to the aforementioned parameters, have a fixed height of 513 points and a variable width depending on the signal duration. However, the authors use only a small portion of this data, corresponding to the first 60 lines of the spectrogram, justifying that there wouldn't be relevant information in the upper regions beyond this threshold. By adopting only the first 60 lines, the authors are using a frequency range between 1 Hz and 1,462.60 Hz, value that will be rounded to 1,462 Hz. This calculation is based on the Nyquist-Shannon theorem, which establishes that the maximum representable frequency in a digital signal is half the sampling rate. Thus, using a sampling rate of 25 kHz and generating a representation with 513 lines, this spectrogram can contain frequencies up to 12.5 kHz. Therefore, it can be determined that the first 60 lines correspond to a range up to 1,462 Hz.

This approach of selecting a region of the spectrogram served as inspiration for creating a segmentation of the frequency in linear ranges to be evaluated within the possible range of up to 12.5 kHz. For this, fixed segments with 1,462 Hz were used, with 50% overlap, defined by the arithmetic progression $f_n = f_{\text{initial}} + d(n-1)$, where f_n is the n-th frequency range, the constant f_{initial} corresponds to the value of the first frequency range, that is, the first term of the progression ($f_{\text{initial}} = 1,462$ Hz), the constant d corresponds to the difference between consecutive terms ($d = 731$ Hz), and n is the frequency number (term of the progression). This segmentation approach generated 16 frequency ranges, as shown in Table I.

TABLE I
FREQUENCY RANGES FOR EVALUATION

Range	Start Frequency	End Frequency
01	1	1,462
02	731	2,193
03	1,463	2,924
04	2,193	3,655
05	2,924	4,386
06	3,655	5,117
07	4,386	5,848
08	5,117	6,579
09	5,848	7,310
10	6,579	8,041
11	7,310	8,772
12	8,041	9,503
13	8,772	10,234
14	9,503	10,965
15	10,234	11,696
16	10,965	12,427

Since all the frequency ranges used have a variation of 1,462 Hz, the generated spectrogram will have a standard height of 60, to any chosen frequency range. Finally, to standardize the width of the spectrogram, a duration of 1 second of the voice signal was used, corresponding to a width of 251 points or columns. The 1 second width was chosen due to the characteristics of the SVD database, in which the median signal duration is only 1.25 seconds. This is sufficient since organic voice disorders exhibit distinct characteristics even in short segments. A single spectogram has been used per voice signal. Furthermore, silence regions of the signal were discarded using the split_on_silence function, from the Pydub[3] library, considering the silence threshold equals to -40dB.

D. Spectrogram Adjustments

The complete spectrogram, with frequencies covering the range up to 12.5 kHz, is illustrated in Figure 1 (a). By applying a bank of band-pass elliptic filters (order 5) , from the ellip[4] function, to the original voice signal, spectrograms for each of

[1]A function of the Scikit-learn library of Python language

[2]Librosa: library package for music and audio analysis of Python language

[3]Pydub: library for audio processing of Python language

[4]ellip: function available in the module scipy.signal of Python language

the ranges described in Table I were generated, as illustrated in Figure 1 (b), these spectrograms have the dimensions (60, 251). The transformations shown in Figure 1 (c) are necessary for the spectrogram to fit within the input dimensions required for using pre-trained CNN architectures. In this way, the spectrogram initially obtained with dimensions (60, 251) was resized to (224, 224). Then, the spectrogram was transformed to adopt the three-channel standard, simulating RGB channels. To achieve this, the (224, 224) representation was replicated across the other channels, forming three channels with the same spectrogram representation, resulting in the dimensions (224, 224, 3).

Utilizing visual representations from a different domain, such as ImageNet, is effective for spectrogram analysis, as highlighted in [18]. Voice databases with limited samples often lead to overfitted CNN models, and transfer learning can mitigate this issue. The initial layers of a CNN trained on general images typically include low-level filters that detect patterns such as contours and textures. These filters are also useful for spectrogram analysis, which rely on similar low-level features.

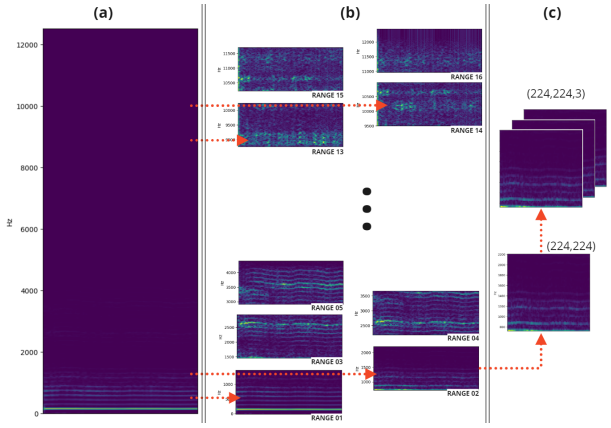

Fig. 1. Transformations applied to the spectrogram: (a) input spectrogram; (b) spectrograms obtained after band-pass filter bank; (c) resizing and replication into 3 channels.

E. CNN Architecture

For the spectrogram classification experiments, a MobileNet V3 Small convolutional neural network, as proposed in [18], pre-trained on the ImageNet database, was adopted. The choice of this architecture occurred after comparative experiments with the results obtained by other convolutional networks, namely: VGG-16, VGG-19, MobileNet V2, MobileNet V3, MobileNet V3 Large, ResNet50, ResNet101, and the architecture proposed by [15]. This comparative evaluation was not detailed as it is outside the scope of this paper. Considering the pre-trained MobileNet V3 Small network, only its convolutional layers with frozen weights were used. The following layers were then added for classification purposes: a flatten layer, a dropout layer with a 30% rate, a fully connected dense layer with 1024 neurons, another dropout layer with a 30% rate, and finally, the output layer with 2 neurons, one for each class. The described architecture is illustrated in Figure 2. These hyperparameters were obtained empirically.

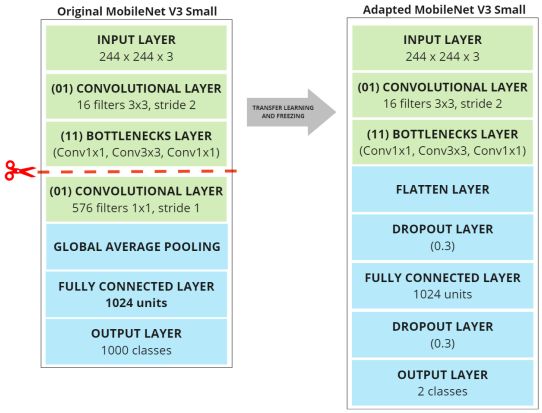

Fig. 2. CNN architecture with transfer learning

F. Experiment design

The experiment was designed to evaluate the capability of characterizing voice patterns in different frequency ranges of the signal. The process workflow is detailed in Figure 3. The first step involves applying a bank of band-pass filters, as described in Section III-C, to generate a dataset of spectrograms obtained in each frequency range. For each dataset (SVD and AVFAD), sixteen spectrogram subsets corresponding to the sixteen evaluated frequency ranges were created. Adjustments were then made to the spectrograms, as described in Section III-D, sixteen CNN models were trained using the architecture described in Section III-E, one model for each spectrogram dataset. Finally, the models are evaluated to obtain their respective performance metrics.

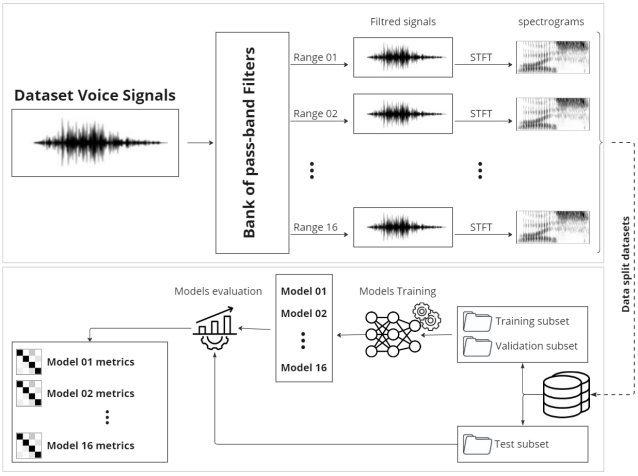

Fig. 3. Steps of the frequency range evaluation process

For the experiments, the hyperparameters were obtained empirically and used for training all models, including: a learning rate of 0.001, the Stochastic Gradient Descent (SGD) optimizer and Binary Cross-Entropy as the loss function. Additionally, to deal with potential plateaus, we used the Keras

function ReduceLROnPlateau. A factor of 1 has been applied to the function if no error reduction occurred for 30 consecutive epochs. An early stopping criterion associated with the validation set loss function was also employed. Whenever there was no improvement error reduction for 50 epochs, the training was halted, and the weights from the epoch with the lowest validation error were considered. Finally, 600 epochs were adopted for the training of all models and a batch size of 128. To evaluate the performance of the models, we used the followings metrics: balanced accuracy, F1 score, sensitivity, specificity, and precision.

After evaluating the performance metrics of the classifiers for different frequency ranges, the next stage of the experiment involves assessing whether combining the results obtained from these classifiers could improve the classification outcomes. The number of possible combinations is given by $\sum_{k=1}^{16} C(n,k)$=65,536, where n=16 is the number of frequency ranges, and k are the combination sizes.

For each combination, a simple majority criterion is adopted, thus each spectrogram is evaluated individually and is classified according to the votes of the combination models. Thus, to determine the classification result of a combination, the outputs generated by each model used in the combination are evaluated, and the majority number of votes determines the classification of a given input. In case of a tie, the vote the model applied to the lowest frequency is considered. For example: when combining frequencies 03, 04, 05, and 06, if the sample is classified as voice disorder in frequencies 03 and 06, and as healthy in frequencies 04 and 05, the classification of frequency 03 will be considered, and the classification in this combination will be voice disorder.

IV. Results and Discussion

Based on the balanced accuracy (ACC) metric, it is possible to make a direct comparison between the results obtained by the classifiers trained on each of the frequency ranges. These results are presented in Figure 4.

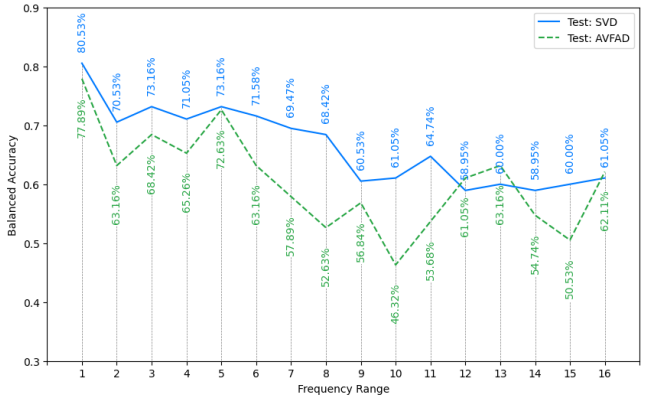

Fig. 4. Comparison of balanced accuracy results obtained from the SVD and AVFAD subset for each frequency range

It is observed that the first frequency range, which spans from 1 Hz to 1.462 Hz, demonstrates a greater descriptive capacity. For the other frequency intervals, even with good performance in various range in some classifiers, a decrease is perceptible, though not linear, in the performance of the other classifiers, particularly on the AVFAD dataset. On the other hand, the last frequency range, between 10.965 Hz and 12.472 Hz, performs slightly better than the four preceding ranges.

After observing the individual performance of the classifiers, the evaluation of the results from the combination of these classifiers was conducted. Given that the experiment generated 65,536 results, it is impractical to present all the metrics for each combination. Therefore, only the combination with the best result is presented.

Table II shows a comparison between the metrics obtained on the SVD dataset for each of the analyzed frequencies, the results obtained from the model proposed by [15], and the results from the frequency combination that achieved the best result. Considering the results on the SVD dataset, it is observed that combining the frequencies 01, 03, 05, 07, 11, 13, 15, and 16, provided a better balanced accuracy result, increasing by 1.7 percentage points compared to frequency range 01. In this scenario, the performance improved from 80.50% to 82.10%. Comparing the accuracy results of the combined frequencies with the model proposed in [15], which achieved 77% using the same dataset and spectrogram pattern, the classification performance is better.

TABLE II
Performance metrics in SVD dataset

Range of Frequency	ACC	F1	SP	SN	P
01 [1 - 1,463] Hz	0.805	0.820	0.726	0.884	0.763
02 [732 - 2,193] Hz	0.705	0.726	0.632	0.779	0.678
03 [1,463 - 2,924] Hz	0.732	0.756	0.632	0.832	0.693
04 [2,193 - 3,655] Hz	0.711	0.724	0.663	0.758	0.692
05 [2,924 - 4,386] Hz	0.732	0.736	0.716	0.747	0.724
06 [3,655 - 5,117] Hz	0.716	0.713	0.726	0.705	0.720
07 [4,386 - 5,848] Hz	0.695	0.704	0.663	0.726	0.683
08 [5,117 - 6,579] Hz	0.684	0.709	0.600	0.768	0.657
09 [5,848 - 7,310] Hz	0.605	0.607	0.600	0.611	0.604
10 [6,579 - 8,041] Hz	0.610	0.602	0.632	0.589	0.615
11 [7,310 - 8,772] Hz	0.647	0.638	0.674	0.621	0.655
12 [8,041 - 9,503] Hz	0.590	0.602	0.558	0.621	0.584
13 [8,772 - 10,234]Hz	0.600	0.591	0.621	0.579	0.604
14 [9,503 - 10,965]Hz	0.590	0.602	0.558	0.621	0.584
15 [10,234 - 11,696]Hz	0.600	0.591	0.621	0.579	0.604
16 [10,965 - 12,427]Hz	0.611	0.641	0.526	0.695	0.594
Model proposed in [15]	0.770	0.780	0.790	0.760	0.810
Range Combination (01/03/05/07/11/13/15/16)	**0.8210**	**0.835**	**0.905**	**0.737**	**0.774**

Table III shows a comparison of the performance for each of the frequencies analyzed in the AVFAD database, as well as the performance for the combination with the best result. For the results obtained in the AVFAD database, the combination of ranges 01, 04, 05, 06, 08, 16, provided substantially better results than those observed in the frequency range 01. The balanced accuracy from the range combination reached 82.12% compared to 78.10% for range 01, representing an improvement of over 4 percentage points. Although no studies employing the same methodology for spectrogram generation and sample selection in this database were identified in the

literature, precluding a direct comparasion, results ranging from 77.3% (for voices with organofunctional pathologies) to 82.8% (for voices with organic pathologies) are reported in [19]. Even though, the results reported in [19] used a different sample of that study were using different sample selection and partitioning strategies, it is possible to perceive some similarities between their accuracies and the ones obtained by combining frequency ranges in the present study.

TABLE III
PERFORMANCE METRICS IN AVFAD DATASET

Range of Frequency	ACC	F1	SP	SN	P
01 [1, 1462] Hz	0.781	0.769	0.848	0.714	0.833
02 [732, 2193] Hz	0.631	0.646	0.609	0.653	0.640
03 [1463, 2924] Hz	0.681	0.722	0.565	0.796	0.661
04 [2193, 3655] Hz	0.653	0.653	0.674	0.633	0.674
05 [2924, 4386] Hz	0.725	0.745	0.674	0.776	0.717
06 [3655, 5117] Hz	0.630	0.653	0.587	0.673	0.635
07 [4386, 5848] Hz	0.579	0.592	0.565	0.592	0.592
08 [5117, 6579] Hz	0.524	0.563	0.457	0.592	0.537
09 [5848, 7310] Hz	0.566	0.602	0.500	0.633	0.574
10 [6579, 8041] Hz	0.462	0.485	0.435	0.490	0.480
11 [7310, 8772] Hz	0.536	0.560	0.500	0.571	0.549
12 [8041, 9503] Hz	0.608	0.648	0.522	0.694	0.607
13 [8772, 10234]Hz	0.630	0.660	0.565	0.694	0.630
14 [9503, 10965]Hz	0.543	0.606	0.413	0.673	0.550
15 [10234, 11696]Hz	0.502	0.552	0.413	0.592	0.518
16 [10965, 12427]Hz	0.619	0.647	0.565	0.673	0.623
Range Combination (01/04/05/06/08/16)	**0.8212**	**0.825**	**0.816**	**0.826**	**0.833**

V. CONCLUSION

This study aimed to investigate the descriptive capability of different frequency ranges for voice classification in disorders. It was observed that high-frequency regions, defined as frequencies above 5 kHz, despite showing modest performance in the classification of spectrograms (pathology vs healthy), demonstrate the presence of useful patterns, especially when the results from classifications using different frequency ranges are combined.

The main contribution of this study is to provide evidence that there is relevant information in high-frequency regions, which contradicts many studies that classify spectrograms and assume the absence of relevant information in these regions, as cited in [15]. Another contribution of this study is the proposed method for spectrogram classification by combining results obtained from different frequency ranges. As results indicated, the proposed method showed potential to improve classification accuracy.

As future investigations, we propose to develop a CNN architecture that extracts embeddings from different frequency ranges and then merges them into a single feature vector for combined classification. Additionally, training/testing on other databases and/or using merged databases is expected (like MEEI of Massachusetts Eye and Ear Infirmary or AVPD of Arabic voice pathology database). We is also intend to apply explainability frameworks in visual patterns for better understanding of the results, such as Grad-CAM in [20] or Occlusion Maps in [21].

REFERENCES

[1] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Analysis and Applications*, pp. 423–431, 2021.

[2] C. Graham, "L1 identification from l2 speech using neural spectrogram analysis." in *Interspeech*, 2021, pp. 3959–3963.

[3] N. Vrebčević, I. Mijić, and D. Petrinović, "Emotion classification based on convolutional neural network using speech data," in *2019. 42nd MIPRO*. IEEE, 2019, pp. 1007–1012.

[4] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in psychology*, p. 91153, 2014.

[5] E. Jacewicz, J. M. Alexander, and R. A. Fox, "Introduction to the special issue on perception and production of sounds in the high-frequency range of human speecha)," *The Journal of the Acoustical Society of America*, no. 5, pp. 3168–3172, 11 2023.

[6] D. R. Leite *et al.*, "Desenvolvimento de um modelo de classificação da tipologia dos sinais vocais com base no deep learning," Ph.D. dissertation, Universidade Federal da Paraíba, 2022.

[7] A. Bürki, "Variation in the speech signal as a window into the cognitive architecture of language production," *Psychonomic Bulletin & Review*, no. 6, pp. 1973–2004, 2018.

[8] N. Lavan, A. M. Burton, S. K. Scott, and C. McGettigan, "Flexible voices: Identity perception from variable vocal signals," *Psychonomic bulletin & review*, pp. 90–102, 2019.

[9] J. M. Liss, S. LeGendre, and A. J. Lotto, "Discriminating dysarthria type from envelope modulation spectra," *Journal of Speech, Language, and Hearing Research*, 2010.

[10] N. V. Naranjo, E. M. Lara, I. M. Rodríguez, and G. C. García, "High-frequency components of normal and dysphonic voices," *Journal of Voice*, no. 2, pp. 157–162, 1994.

[11] J. B. Lee and H. G. Lee, "Quantitative analysis of automatic voice disorder detection studies for hybrid feature and classifier selection," *Biomedical Signal Processing and Control*, p. 106014, 2024.

[12] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the saarbrücken voice database: Effects of pathology subset and audio materials," in *Proceedings of the Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 4850–4854.

[13] X. Peng, H. Xu, J. Liu, J. Wang, and C. He, "Voice disorder classification using convolutional neural network based on deep transfer learning," *Scientific Reports*, no. 1, p. 7264, 2023.

[14] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *Journal of Voice*, no. 1, pp. 3–15, 2017.

[15] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "Convolutional neural networks for pathological voice detection," in *2018 40th annual international conference of the ieee engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 1–4.

[16] B. Woldert-Jokisz, "Saarbruecken voice database," in *Saarbruecken Voice Database*. Institut für Phonetik, Universität des Saarlandes, 2007.

[17] L. M. Jesus, I. Belo, J. Machado, and A. Hall, "The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research," in *Advances in Speech-language Pathology*. IntechOpen, 2017.

[18] S. Qian, C. Ning, and Y. Hu, "Mobilenetv3 for image classification," in *2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. IEEE, 2021, pp. 490–497.

[19] S. Moura *et al.*, "Detecção e classificação de categorias de disfonias com redes neurais convolucionais," Master's thesis, Universidade Tecnológica Federal do Paraná, 2023.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[21] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *International Journal of Computer Vision*, pp. 736–760, 2021.