Nonperiodic pathologic voice signals classification using Mel-Spectrogram and VGGish

Joana Filipa Teixeira Fernandes[1,2][0000-0002-0618-4627], João Viana Pinto[3][0000-0002-6485-6786],Carla Pinto Moura[4, 5 [0000-0001-7291-6163], Helena Vilarinho[5, 6][0000-0001-7339-8409], Felipe Teixeira[1,7][0000-0002-3439-826X], Diamantino Freitas[2][0000-0003-4260-9677] and João Paulo Teixeira[1][0000-0002-6679-5702]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), - Instituto Politécnico de Bragança (IPB), Bragança 5300, Portugal; joana.fernandes@ipb.pt, joaopt@ipb.pt

[2] Faculty of Engineering of University of Porto (FEUP), Porto, 4200-465, Portugal; dfreitas@fe.up.pt

[3] Otorhinolaryngology Department, University Hospital Centre of São João, Unit of Otorhinolaryngology, Department of Surgery and Physiology, University of Oporto Faculty of Medicine, Alameda Professor Hernâni Monteiro, 4200- 319, Oporto, Portugal, Centre for Health Technology and Services Research (CINTESIS), Rua Dr. Plácido da Costa, 4200-450, Oporto, Portugal, joaopvfpinto@gmail.com

[4] Genetics, Department of Pathology, Faculty of Medicine, University of Porto, Porto, Portugal.

[5] Department of Otorhinolaryngology, University Hospital Centre of São João, Porto, Portugal; cmoura@med.up.pt, vilarinhocastro@gmail.com

[6] School of Health Sciences (ESSUA), University of Aveiro, Aveiro, Portugal

[7] Applied Management Research Unit (UNIAG) - Instituto Politécnico de Bragança (IPB), Bragança 5300, Portugal; Engineering Department, School of Sciences and Technology, University of Trás-os-Montes and Alto Douro (UTAD), Quinta de Prados, Vila Real, 5000-801, Portugal, felipe.lage@ipb.pt

Abstract. In this work and the literature, voice signals can be classified as periodic (type 1) or either some periodicity (type 2) and chaos (type 3). This work aims to classify signs into types 1, 2 or 3 to be subsequently applied in a classification system for pathological/control signs. The original dataset is composed of 466 type 1 individuals, 900 type 2 individuals, and 84 type 3 individuals classified by an otolaryngologist. 15% of the data was used for testing and the remaining 85% was used for training and validation. A data augmentation technique was applied to balance the data in training set. Therefore, for the test set, 3380 sounds were used, 1020 type 1, 1280 type 2 and 1080 type 3. Of these, 80% were used for training and 20% for validation. The Mel spectrograms of the signals were used in the input of a VGGish to retrain the model in classifying the 3 types of signals. Regarding test accuracy, this network obtained 71.2%.

Keywords: Disordered Signals Types, Mel Spectrogram, Convolutional Neural Networks (CNN), VGGish.

1 1. Introduction

There are various reasons to analyse voice acoustic signals. The human voice has been demonstrated to include a wealth of information regarding a person's general health and well-being from a health science standpoint [1].

To properly determine voice signals for disturbance analysis, Titze, 1994 [1] recommends a pre-processing step, which classifies the voice signal into three types, considering the degree of periodicity [2].

Titze, 1994 [1] defined type 1 voices as quasi-periodic, type 2 signals as containing intermittency, strong subharmonics or modulations, and finally, type 3 signals as chaotic or random. This type of classification is helpful as it prevents voice pathologists from using analysis tools that may be unreliable for inappropriate signals, which can lead to misleading values that do not carry any physiological meaning [2].

The most indicated uses for analysing type 2 and 3 voices are spectrograms, perceptive analysis, and non-linear dynamic analysis. Meanwhile, type 1 signals are adequate for perturbation measures such as jitter and shimmer[3].

Considering this analysis, several authors started to use this classification scheme to determine the suitability of voice signals for disturbance analysis [4]–[8].

Sprecher et al., 2010 [3] added a fourth voice type to Titze's 1994 scheme. In this new scheme for classifying voice signals, types 1 and 2 do not change their definition, while type 3 was divided into two, becoming types 3 and type 4. In this classification, type 3 includes only voices that present a chaotic behaviour; in contrast, type 4 voices comprise signals with a stochastic solid component. Both voice types 3 and 4 need an apparent periodic structure. In this way, the problem of distinguishing between type 3 and type 4 voices, similar to the issue of differentiating a chaotic (deterministic) dynamic from a stochastic (non-deterministic) one [9]. Thus, we will use Titze's 1994 classification scheme in this article.

Both Titze, 1994 [1] and Sprecher et al., 2010[3] consider spectrograms the most appropriate tool for classifying voice signals into different types, regardless of whether the scheme of 3 or 4 types of signals is applied. To complement the visual information obtained by the spectrograms, the physicians must hear the voice recordings to detect noise or essential changes in tone, and the evaluation of the periodicity of the signal must also be combined [10]. In a non-invasive way, the auditory acoustic analysis intends to quantify and characterise the sound signal. Still, it needs to be more objective and depends on the experience of the physician who performs the evaluation[11].

These factors have led to the need to develop systems to assist clinicians in perceptive tasks and use a pattern recognition approach that automatically uses objective measures to classify signals [12]. The development of automatic acoustic analysis techniques saves both patient and specialist time and can improve the accuracy of assessments [12]–[14].

Based on the above, the objective of this work is to automatically detect whether a signal corresponds to type 1, 2 or 3 using a convolutional network, which has mel spectrograms in its input to be later used as a parameter in the classification of pathological/control signals.

This article is organised as follows: section 1 presents the introduction. Next follows the materials and methods section. Section 3 presents the results and discussion; finally, section 4 presents the conclusion.

1.1 State of the Art

Although the importance of classifying signals is recognised in the scientific community, the methods are complex and include subjective assessment, with most of the work described in the literature [15]–[18] only describing quantitative measures to determine the type of signal. Signal and the classification step were only addressed in the work of Lee, 2016 [19] and Miramont et al. 2022[2].

In his research, Lee, 2016 [19] classified audio signals using metrics such as the bicoherence value (BV) based on the linear predictive coding (LPC) residual, the coefficient of normalised skewness variation (CSV), and the coefficient of normalised kurtosis variation (CKV). The acoustic, CSV, CKV, and BV parameter performances were estimated using the classification and regression tree (CART) and the LPC residual. When jitter, shimmer, and SNR were examined as acoustic characteristics, the best tree produced by jitter alone had an average accuracy of 78.6%. The average accuracy of the decision tree created using the acoustic, CSV, CKV, and BV parameters was 82.1%.

In their work, Miramont et al. 2022 [2] propose a pattern recognition method for automatic voice signal typing based on a multi-class linear Support Vector Machine. They achieve accuracies of 82.71% by combining nonlinear dynamics measures with relatively well-known parameters like Jitter, Shimmer, Harmonic-to-Noise Ratio, and Cepstral Prominence Peak.

2 Materials and Methods

This section describes the database used, the type 1, 2 or 3 signals, the features used and their extraction, the classification model used and the database augmentation.

2.1 Database

The German Saarbrucken Voice Database (SVD) was used. This database is available online from the Institute of Phonetics at the University of Saarland [20].

The database consists of voice signals from over 2000 subjects with several diseases and control subjects. Each person has the recording of phonemes /a/, /i/ and /u/ in the low, average and high tones, swipe-along tones, and the German phrase "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The size of the sound files is between 1 and 3 seconds, has a resolution of 16 bits, and has a sampling frequency of 50 kHz.

All signals below than 800 ms were excluded from this analysis, as Miramont et al. 2022[2]. The specialist then classified the signal using the signal spectrogram, where 1450 sound files were classified, 466 type 1 signals, 900 type 2 signals and 84 type 3

signals. These files were from pathological subjects, distributed across various diseases and by vowels /a/, /i/ and /u/ in 3 tones: high, low and normal.

2.2 Signal Typing

A bifurcation is a term used in nonlinear dynamics to describe a qualitative change in the behaviour of a dynamical system. As a rule, it occurs when there is a gradual change in some parameters of the vibratory system, such as pulmonary pressure, tension in the vocal cords, or asymmetry between the vocal cords [2].

The following classification scheme is adopted to recognise the nature of bifurcations in speech signals. Classification is crucial to all other considerations in voice acoustic analysis. It follows the general principles of nonlinear coupled oscillator dynamics [1].

- Type 1 signals – almost periodic signals that do not present qualitative changes in the analysis segment. If modulation frequencies or sub-harmonics are present, the energy of these signals is an order of magnitude below the power of the fundamental frequency.
- Type 2 Signals - Signals with qualitative variations (bifurcations) in the analysis segment, or signals with subharmonic frequencies or modulating frequencies whose energies approach the energy of the fundamental frequency; there is, therefore, no single precise fundamental frequency across the entire segment.
- Type 3 signals – signals that do not have a periodic structure.

2.3 Mel-spectrogram

In sound analysis, spectrograms play a crucial role as they offer essential information about the frequency of energy distribution over time. The spectrogram is a widely used as visual aid for audio data where one wants to describe the energy distribution in various frequency bands [21]. However, Newman et al. [22] invented the Mel scale, which provides a non-linear transition from the traditional linear frequency scale to better match the perception of the human ear. The transformation to the Mel spectrum is more effective in collecting acoustic elements, highlighting the percentage of relevant information within the audio stream, and applying a bank of Mel filters in the signal's frequency domain, creating the Mel spectrum. Each filter in the Mel filter bank has a central point in the frequency domain and comprises triangular window functions. When you divide the signal into multiple frequency bands, these filters can imitate the human ear picking out frequencies. The corresponding triangular window functions of the Mel filter bank are applied using the squared magnitude of the discrete Fourier transform (DFT) coefficients [23].

In short, the Mel spectrum provides a perceptually relevant representation of the audio data, as it is formed from the Mel scale and the Mel filter bank. Our understanding of the acoustic properties of various audio signals is enhanced by their use in sound analysis, speech recognition, and other audio-related activities, which allow for a more accurate, human-centred interpretation of frequency energy distribution.

2.4 Convolutional Neural Networks (CNN)

Most images are composed of a vast pixel density, which for a fully connected neural network would require many weights in the first hidden layer [24], [25].

In this way, convolutional neural networks are better, as they only have one of their neurons connected to a specific region of the layer next to it. In this way, computers detect and extract data from these images and compare them, pixel by pixel, with a reference image. When comparing images, the algorithm detects the image, section by section, which are called features [24], [25].

In a CNN, the first layer is the Convolution layer, where the image features are extracted through the movement of a filter of a specific size over the image, which will allow the creation of the feature map. This map is a matrix of positive and negative pixel values created by mathematical operations when moving the filter. The next layer is the rectified linear activation function (ReLU), which receives the previous layer's output as input and creates its input as output if it is positive. If the input pixel value is negative, the ReLU output becomes zero. Next, the Pooling layer tries to reduce the dimension of the resource map, thus reducing the calculation time and making the process faster. Lastly, the CNN layers are fully connected. The input vector for the fully connected layer is subjected to an activation function, which allows the calculation of the probabilities. A higher likelihood would indicate the presence of the required image feature, thus indicating successful image detection [24], [25].

Transfer Learning

Training large datasets through deep neural networks requires several days to weeks. To avoid this situation, pre-trained models are used. Using pre-trained models reduces errors and the time needed for training. These pre-trained models use weights reused across layers to adapt to the new problem. This way, a part of the pre-trained or entire model can be joined to the latest neural network model [26].

Several pre-trained models can be used; however, the VGGish model will be used in this work.

Architecture CNN – VGGish

The VGGish model, commonly called the VGGish network, comprises a popular convolutional neural network (CNN) architecture, which Google researchers trained for feature extraction from audio content. It is a pre-trained network with several notable improvements over previous models, resulting in better performance, ease of use and reduced training time [27].

Using smaller convolutional filters, such as 3×3 , allows VGGish to achieve better generalisation and reduce the likelihood of overfitting during training, which is one of the standout features. Furthermore, considering the information collected from the previous layer, each layer of VGGish's hierarchical structure can gradually acquire a more complex representation of the input image [27].

After receiving the input image, VGGish's first layer generally scales it to a fixed dimension. Convolutional layers are used to make up the model, and its function is to

use various filters to extract information from the input image. VGGish adds five max-pooling layers after each convolutional layer to down sample the feature maps, thus decreasing the spatial dimensions and increasing the network's resilience.

Due to its exceptional performance and adaptability, the VGGish model is used in several areas, including drone recognition [28] and audio signal classification [27]. It is an effective tool for image recognition and classification in various medical and machine vision applications due to its versatility and resilience.

2.5 Data augmentation

Since the data is unbalanced and this negatively impacts the training model, it was necessary to try to balance the data and thus have more samples using data augmentation.

In this process, 15% of data was selected for testing, excluded from the data that would be augmented and used for training and validation.

Therefore, from the 1450 signals, 218 were saved for testing, and the remaining 1232 signals were augmented to balance the classes.

The technique used to augment the audio data was random sequential augmentations, which defines augmentation as a sequence of augmentations applied probabilistically. The order in which the augmentations are applied is always the same.

The following conditions were applied for the stretch time parameter: probability of applying time stretch – 0.5, range of time stretch speedup factor – [0.8, 1.2], applying time stretch – true, time stretch speedup factor – 0.8. The following conditions were applied to the shift pitch parameter: probability of applying pitch shift – 0.5, range of pitch shift – [-2, 2], apply pitch shift – true, pitch shift – (-3). To control volume, the probability of applying volume control – 0.5, the range of volume gain (dB) – [-3, 3], apply volume gain – true, and volume gain (dB) – (-3) were applied. In adding noise, the probability of applying noise addition – 0, range of noise addition SNR (dB) – [0, 10], applying noise addition – true and noise addition SNR (dB) – 5. Lastly, in shift time, used the probability of applying time shift – 0.5, range of time shift (s) – $[-5e^{-3}, 5e^{3}]$, apply time shift – true and time shift – $5e^{-3}$.

2.6 Feature Extraction with VGGish

The audio input signal is split up into time blocks of 975 milliseconds and has an overlap percentage of 50% of the blocks.

A 96x64 frequency-time spectrum appropriate for the convolutional layers' input is created from each block.

The convolutional and pooling layers successively extract the frequency-time spectrum's characteristics.

The final linear mapping layer compresses the retrieved data from previous layers into a 128-dimensional vector of audio features.

This 128-dimensional vector effectively describes the input audio material and can be applied to sound event detection, classification, and other audio processing tasks. To extract meaningful features from voice, music, ambient sounds, and other audio content, the VGGish model has been trained on a large dataset of different audio content.

3 Results and Discussion

This work aims to classify the three types of signals and use the mel spectrogram of the signals, which is extracted and used as an image for input into the CNN. The VGGish model was used, where 15% of the data was initially removed for testing. Data augmentation was carried out for the remaining 85%, and then files with a duration of less than 975 milliseconds were eliminated. Therefore, were left with 3380 sounds, 1020 type 1, 1280 type 2 and 1080 type 3. Of these, 80% were used for training and 20% for validation. For the training process, 6 epochs were used, with a validation frequency of 20; for the training function, the 'adam' function was used, and the data was shuffled before each training epoch; as a learning rate schedule, the function "piecewise", the patience of validation stopping 5, output network is 'best – validation'.

3.1 Results

In Figure 1, it is possible to observe the results obtained when applying the test data. These correspond to the 15% initially withdrawn. There would be 218 signals. However, it was necessary to eliminate those that were shorter than 975 milliseconds, giving a total of 187 signals.

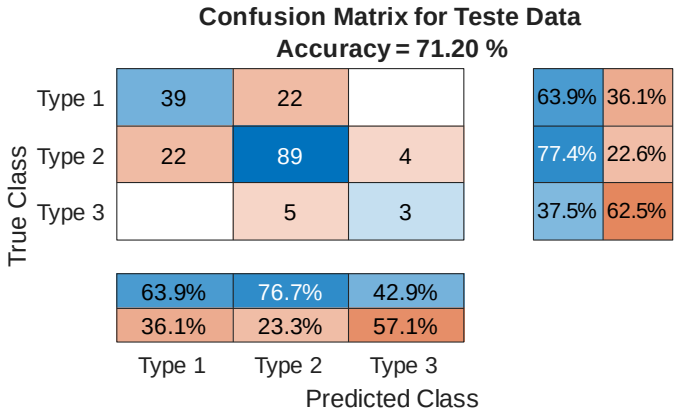

Fig. 1. Confusion Matrix

Fig.1 shows that the network's performance in the test set obtained an accuracy of 71.2%.

3.2 Discussion

This work aimed to classify type 1, type 2 and type 3 signals. However, as can be seen from the results, using mel-spectrogram images, it was not possible to obtain better results than those existing in the literature. The use of mel-spectrogram images does

not reach the accuracy of 82.71% achieved by Miramont et al. 2022 [2] when combining the parameters Jitter, Shimmer, Harmonic-Noise Ratio, and Cepstral Prominence Peak.
Theoretically, it makes no sense for jitter and shimmer, which measure periodicity, to be used in type 3 signals since they are chaotic. Therefore, they do not have periodicity. However, the algorithms used to measure jitter and shimmer developed by Teixeira and Gonçalves 2016 [29] work on type 3 signals since when measuring jitter and shimmer in type 3 signals, these are very high, demonstrating a non-periodicity of the signal.

Therefore, the idea that jitter and shimmer cannot be used in non-periodic signals is untrue. What will allow us to use these parameters to classify pathological/control signals.

4 Conclusion

This work aims to classify type 1, 2 or 3 signals using the Mel spectrogram.

So, this analysis started with an expert classifying 1,450 signals into three types. 466 type 1 signs, 900 type 2 signs and 84 as type 3 were classified.

The spectrogram of the signals that served as input parameters was extracted using the VGGish pre-process and then served as input to a pre-trained VGGish model, achieving an accuracy of 71.2% on the test set.

The results are lower than those that used jitter shimmer Harmonic-Noise Ratio, and Cepstral Prominence Peak. We can use the jitter and shimmer values to further aid in classifying pathological/control signals since the algorithms used for subsequent classification in these parameters obtain high values in type 3 signals.

Acknowledgements

This work was supported by national funds through FCT/MCTES (PIDDAC): CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020) and 2021.04729.BD (DOI: 10.54499/2021.04729.BD).

References

[1] I. R. Titze, “Workshop on Acoustic Voice Analysis,” *Natl. Cent. Voice Speech, Am.*, pp. 1–36, 1994, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Workshop+on+Acoustic+Voice+Analysis#2

[2] J. M. Miramont, J. F. Restrepo, J. Codino, C. Jackson-Menaldi, and G. Schlotthauer, “Voice Signal Typing Using a Pattern Recognition Approach,” *J. Voice*, vol. 36, no. 1, pp. 34–42, Jan. 2022, doi: 10.1016/J.JVOICE.2020.03.006.

[3] A. Sprecher, A. Olszewski, J. J. Jiang, and Y. Zhang, “Updating signal typing in voice: Addition of type 4 signals,” *J. Acoust. Soc. Am.*, vol. 127, no. 6, pp. 3710–3716, Jun.

2010, doi: 10.1121/1.3397477.

[4] S. H. Choi, Y. Zhang, J. J. Jiang, D. M. Bless, and N. V. Welham, "Nonlinear Dynamic-Based Analysis of Severe Dysphonia in Patients With Vocal Fold Scar and Sulcus Vocalis," *J. Voice*, vol. 26, no. 5, pp. 566–576, Sep. 2012, doi: 10.1016/J.JVOICE.2011.09.006.

[5] C. Fabris, W. De Colle, and G. Sparacino, "Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 920–926, Nov. 2013, doi: 10.1016/J.BSPC.2013.08.010.

[6] D. Stone *et al.*, "Voice outcomes after transoral laser microsurgery for early glottic cancer - Considering signal type and smoothed cepstral peak prominence," *J. Voice*, vol. 29, no. 3, pp. 370–381, 2015, doi: 10.1016/j.jvoice.2014.08.018.

[7] B. . H. Barsties U.; Maryn, Y., "Spektrografische Stimmtypenklassifizierung zur Beurteilung der Stimmqualität TT - The Evaluation of Voice Quality via Signal Typing in Voice using Narrowband Spectrograms," *Laryngorhinootologie*, vol. 95, no. 02, pp. 105–111, 2016, doi: 10.1055/s-0035-1559678.

[8] S. Vaz Freitas, P. Melo Pestana, V. Almeida, and A. Ferreira, "Integrating Voice Evaluation: Correlation Between Acoustic and Audio-Perceptual Measures," *J. Voice*, vol. 29, no. 3, pp. 390.e1-390.e7, May 2015, doi: 10.1016/J.JVOICE.2014.08.007.

[9] J. P. . Teixeira, D. . Freitas, D. . Braga, M. J. . Barros, and V. Latsch, "Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB," in *Proceedings of Eurospeech'01 – International Conference on Spoken Language Processing*, 2001, pp. 1707–1710. doi: 8790834100, 978-879083410-4.

[10] J. A. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, G. Castellanos-Dominguez, and J. I. Godino-Llorente, "Emulating the perceptual capabilities of a human evaluator to map the GRB scale for the assessment of voice disorders," *Eng. Appl. Artif. Intell.*, vol. 82, pp. 236–251, Jun. 2019, doi: 10.1016/J.ENGAPPAI.2019.03.027.

[11] J. P. Teixeira, J. Fernandes, F. Teixeira, and P. O. Fernandes, "Acoustic analysis of chronic laryngitis statistical analysis of sustained speech parameters," in *BIOSIGNALS 2018 - 11th International Conference on Bio-Inspired Systems and Signal Processing, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018*, 2018, vol. 4, pp. 168–175. doi: 10.5220/0006586301680175.

[12] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders," *J. Voice*, vol. 33, no. 6, pp. 947.e11-947.e33, Nov. 2019, doi: 10.1016/J.JVOICE.2018.07.014.

[13] V. Guedes, A. Junior, J. Fernandes, F. Teixeira, and J. P. Teixeira, "Long Short Term Memory on Chronic Laryngitis Classification," *Procedia Comput. Sci.*, vol. 138, pp. 250–257, Jan. 2018, doi: 10.1016/J.PROCS.2018.10.036.

[14] J. P. . Teixeira and D. Freitas, "Segmental Durations Predicted With a Neural Network," in *Proceedings of Eurospeech'03 – International Conference on Spoken Language Processing*, 2003, pp. 169–172.

[15] P. N. Carding, I. N. Steen, A. Webb, K. Mackenzie, I. J. Deary, and J. A. Wilson, "The reliability and sensitivity to change of acoustic measures of voice quality," *Clin. Otolaryngol. Allied Sci.*, vol. 29, no. 5, pp. 538–544, 2004, doi: 10.1111/j.1365-

2273.2004.00846.x.

[16] L. D'Alatri, F. Bussu, E. Scarano, G. Paludetti, and M. R. Marchese, "Objective and subjective assessment of tracheoesophageal prosthesis voice outcome," *J. Voice*, vol. 26, no. 5, pp. 607–613, 2012, doi: 10.1016/j.jvoice.2011.08.013.

[17] J. J. Houlton *et al.*, "Voice outcomes following adult cricotracheal resection," *Laryngoscope*, vol. 121, no. 9, pp. 1910–1914, 2011, doi: 10.1002/lary.21915.

[18] L. M. Kopf *et al.*, "Pitch Strength as an Outcome Measure for Treatment of Dysphonia," *J. Voice*, vol. 31, no. 6, pp. 691–696, 2017, doi: 10.1016/j.jvoice.2017.01.016.

[19] J. Y. Lee, "Parameter estimations for signal type classification of Korean disordered voices," *Int. J. Eng. Technol.*, vol. 7, no. 6, pp. 1977–1988, 2016.

[20] M. Pützer and W. J. Barry, "Saarbruecken Voice Database," *Institute of Phonetics at the University of Saarland*, 2007. http://www.stimmdatenbank.coli.uni-saarland.de (accessed Nov. 05, 2021).

[21] A. Maity, A. Pathak, and G. Saha, "Transfer learning based heart valve disease classification from Phonocardiogram signal," *Biomed. Signal Process. Control*, vol. 85, p. 104805, Aug. 2023, doi: 10.1016/J.BSPC.2023.104805.

[22] S. S. S. J. V. E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *J. Acoust. Soc. Am.*, vol. 8, pp. 185–190, 1937, doi: https://doi.org/10.1121/1.1915893.

[23] J. Fernandes, L. Silva, F. Teixeira, V. Guedes, J. Santos, and J. P. Teixeira, "Parameters for Vocal Acoustic Analysis - Cured Database," *Procedia Comput. Sci.*, vol. 164, pp. 654–661, Jan. 2019, doi: 10.1016/J.PROCS.2019.12.232.

[24] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification," *Proc. IEEE Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl. CENTCON 2021*, vol. 1, pp. 96–99, 2021, doi: 10.1109/CENTCON52345.2021.9687944.

[25] S. Khaleghian, H. Ullah, T. Kræmer, N. Hughes, T. Eltoft, and A. Marinoni, "Sea ice classification of sar imagery based on convolution neural networks," *Remote Sens.*, vol. 13, no. 9, pp. 1–20, 2021, doi: 10.3390/rs13091734.

[26] H. C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.

[27] M. Diwakar and A. B. Gupta, "The robust feature extraction of audio signal by using VGGish model," *Internatioinal J. Comput. Sci. Inf. Secur.*, vol. 21, no. 6, pp. 1–18, 2023.

[28] H. Samma, S. A. Suandi, N. A. Ismail, S. Sulaiman, and L. L. Ping, "Evolving Pre-Trained CNN Using Two-Layers Optimizer for Road Damage Detection from Drone Images," *IEEE Access*, vol. 9, pp. 158215–158226, 2021, doi: 10.1109/ACCESS.2021.3131231.

[29] J. P. Teixeira and A. Gonçalves, "Algorithm for Jitter and Shimmer Measurement in Pathologic Voices," *Procedia Comput. Sci.*, vol. 100, pp. 271–279, Jan. 2016, doi: 10.1016/J.PROCS.2016.09.155.