INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1–7
Manuscript received February 10, 2025; revised June, 2025. doi: 10.24425/ijet.2025.153621

Utilizing CNN architectures for non-invasive diagnosis of speech disorders – further experiments and insights

Filip Ratajczak, Mikołaj Najda, and Kamil Szyc

***Abstract*—This research investigated the application of deep neural networks for diagnosing diseases that affect the voice and speech mechanisms through the non-invasive analysis of vowel sound recordings. Using the Saarbruecken Voice Database, the voice recordings were converted to spectrograms to train the models, specifically focusing on the vowels /a/, /u/, and /i/. The study used Explainable Artificial Intelligence (XAI) methodologies to identify essential features within these spectrograms for pathology identification, with the aim of providing medical professionals with enhanced insight into how diseases manifest in sound production. The F1 Score performance evaluation showed that the DenseNet model scored 0.70 ± 0.03 with a top of 0.74. The findings indicated that neither vowel selection nor data augmentation strategies significantly improved model performance. Additionally, the research highlighted that signal splitting was ineffective in enhancing the models' ability to extract features. This study builds on our previous research [1], offering a more comprehensive understanding of the topic.** [1]

***Keywords*—Voice Disorder Diagnosis; Vowel Sound Analysis; Convolutional Neural Networks (CNNs); Explainable Artificial Intelligence (XAI)**

I. Introduction

WITHIN the realm of medical diagnostics, there is an increasing focus on creating non-invasive and easily accessible tests that aim to improve patient quality of life while also lowering hospital expenses. The emergence of machine learning (ML) and artificial intelligence (AI) is paving the way for innovative approaches that hold the potential to improve diagnostics by providing solutions that are minimally intrusive, economically viable, and broadly applicable [2], [3].

The early detection of diseases represents a significant advancement in modern medicine. It is well established that prompt diagnosis of conditions can significantly improve patient outcomes. Using extensive patient data, deep learning (DL) models are used to analyze large amounts of information, revealing patterns that traditional diagnostic methods cannot detect. These innovations can help with disease detection, facilitating more timely and accessible interventions [4] [5], [6].

Recent strides in speech analysis via machine learning have facilitated the detection of diseases by scrutinizing vocal attributes like pitch, tone, rhythm, and breath control. The core premise of this approach is that certain diseases can be pinpointed by noting subtle changes in a patient's voice or speech patterns, providing an auxiliary tool for healthcare professionals during examinations [7]. Such observations can sometimes be detected before more apparent symptoms develop, enhancing speech analysis's diagnostic capability. There is already substantiation that speech can function as a biomarker for conditions such as neurological [8], mental [9], voice [10], and respiratory disorders [11], with continuing research seeking to expand its use to a wider array of diseases, including heart failure [12].

This research seeks to showcase the potential of analyzing speech data in healthcare technologies by differentiating between people with normal voices and those suffering from voice disorders. The analysis will focus on the frequency patterns of /a/, /u/, and /i/ sounds. Building on our earlier research utilizing the Saarbruecken Voice Database to explore the use of CNNs pathological voices [1], this study incorporates new models - DenseNet, EfficientNet, RegNet, and DeiT - to extend our previous analysis. The process involves converting voice recordings into frequency domain representations using spectrograms bolstered by diverse audio augmentation techniques to expand the dataset. Models were evaluated through various configurations, particularly with respect to vowel selection and augmentation strategies, employing the F1-score as the main metric of performance. Furthermore, the study uses XAI methods to pinpoint critical features in the spectrograms that distinguish healthy from pathological voice samples. Our approach provides a comprehensive methodology for disease detection and can offer clinicians improved insights into the acoustic features associated with voice disorders, thereby facilitating a better understanding and aiding the diagnostic process.

This paper is organized as follows: Section II discusses prior research on voice analysis, spectrogram generation, deep learning models, XAI, and pathological voice classification. Section III outlines the data preparation steps, model evalua-

F. Ratajczak and K. Szyc are with the Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Wrocław, Poland (e-mail: frata-jczak124@gmail.com, kamil.szyc@pwr.edu.pl).

M. Najda is with the Institute of Data Science, Maastricht University, The Netherlands (e-mail: mikolaj.najda@maastrichtuniversity.nl).

[1] All results are fully reproducible, the source code is available at https://github.com/Tesla2000/DepCoS2024/

tion processes, and experimental outcomes. Lastly, Section IV recaps the main findings and contributions of this study.

II. Related works

Voice analysis methods are typically grouped into three main categories: static analysis, time-series analysis, and signal-to-image conversion [13]–[15]. Static analysis focuses on extracting distinctive features from voice recordings, such as pitch, loudness, jitter, and shimmer [16]. Time-series techniques consider the voice as a series of data points, using Long Short-Term Memory (LSTM) models to detect time-related patterns and variations [17], [18]. The utilisation of image-based techniques allows for the application of methodologies such as CNNs or Vision Transformers (ViTs) trained with spectrograms to recognize unique patterns in this frequency representation of the signal [19], [20].

Our study focuses on the latter method, where spectrograms are utilized to capture the specific details regarding the frequency and temporal aspects of voice signals. These spectrograms afford a thorough signal representation by depicting energy distribution across time and frequencies, making them ideal for CNN/ViT evaluation [21]. Other methods compared to spectrograms include the short-time Fourier transform (STFT) using both linear and Mel scales, the constant-Q transform (CQT), and the continuous wavelet transform (CWT). Each method provides a distinct insight into the signal's time-frequency features. For example, CWT excels at capturing fine temporal details with different resolutions [22], while CQT offers a logarithmic frequency scale, better fitting human auditory perception [23]. Nonetheless, spectrograms continue to be widely favored due to their simple graphical depiction and compatibility with conventional image processing techniques, thus serving as an excellent choice for deep learning tasks in voice analysis.

CNNs [24] have played a pivotal role in the advancement of image recognition tasks, with significant contributions from networks such as VGG [25], ResNet [26], and more recently, EfficientNet [27]. These architectures have demonstrated remarkable performance, reaching notable milestones in the field and are providing a robust foundation for the analysis of complex features within voice sample spectrograms. They enable the precise extraction of time-frequency representations and effectively support the classification of various speech characteristics.

Recently, transformer-based models have emerged, enhancing the ability to process sequential data such as speech. Originally developed for natural language processing tasks, transformers excel at capturing complex patterns but often require more data to be trained effectively compared to CNNs [28]. Models like the Vision Transformer (ViT) [29] and the Spectrogram Transformer [30] achieve strong performance in benchmark spectrogram analyses by utilizing self-attention mechanisms. These results position transformers as a powerful complementary or alternative solution to traditional CNN-based methods in speech analysis.

By utilizing straightforward voice recordings, particularly vowel sounds, our methodology is grounded in the observation that specific illnesses cause a noteworthy shift in vocal characteristics , which may be observed in vowel pronunciation. The utilisation of the Saarbruecken Voice Database [31], which spans a diverse set of voice disorders, illustrates the capacity of voice recordings to discern and categorise pathological conditions.

A number of research teams have conducted studies investigating the potential of CNNs for the classification of pathological voice recordings, frequently utilising the Saarbruecken Voice Database as a benchmark. For example, the authors [32] leveraged deep learning by modifying the VGG architecture, specifically targeting organic dysphonia disorders. Their method consisted of training ensemble models on diverse vowel subsets proved an effective approach, with an 82% accuracy rate in identifying pathological speech. This highlights the efficacy of employing model combinations and transfer learning to address the limitations of restricted dataset sizes in classification of disordered speech.

In another study, a group of researchers explored a new algorithm, OSELM, to distinguish and categorise the various types of voice disorders [33]. The conjunction of vowel sounds and an uninterrupted flow of speech spoken at various pitches achieved an accuracy rate exceeding 87% across multiple metrics, indicating the potential of the algorithm for developing clinically applicable software that can be used in real time.

Similarly, the work presented in [34] introduced CS-PVC, a system designed to classify disordered voices. Mel-frequency cepstral coefficients (MFCCs) were used as features, which were fed into a DCA-ResNet architecture that included attention modules to emphasize relevant features. This resulted in an accuracy rate of 81.6% in the Saarbruecken Voice Database.

The highest accuracy of 82.69% was achieved by another work [35] using a CNN classifier with linear prediction cepstrum coefficients (LPCCs) features for the vowel in male voice samples. LPCCs were extracted from 40-ms windows with a 20-ms frame shift, providing a detailed representation of the vocal tract characteristics. This proved crucial for the differentiation between pathological and healthy voices.

It is noteworthy that the prior mentioned research, as well as ours, employ a variety of voice sample subsets for the training and assessment of their voice pathology classification models.

CNNs leveraged in the detection of diseases through voice analysis facilitate the use of XAI techniques [36]. Techniques such as Grad-CAM [37], Score-CAM [38], and Ablation-CAM [39] offer visual clarity on how CNN models arrive at decisions by emphasizing the most pertinent areas in spectrograms. Grad-CAM employs gradient information to create heatmaps indicating the most significant regions, while Score-CAM enhances this by weighting each activation map based on actual output scores. Ablation-CAM, on the other hand, systematically removes input components to observe their effects on outcomes. These approaches offer a means to comprehend the reasoning behind model predictions, aiding in the interpretation and verification of automated analyses, which in turn supports better-informed clinical decisions.

III. Experiments

The Saarbruecken Voice Database is a well-regarded resource for scholars specializing in speech analysis [31]. This research investigates audio signal from recorded articulation of the vowels /a/, /i/, and /u/, gathered from both individuals who do not have speech disorders and those who have been diagnosed with speech impairments.

The purpose of this study is to differentiate between samples from individuals in good health and those with different medical conditions, comprising 2031 recordings from healthy subjects and 2289 from those with voice disorders evenly spread across the three vowels. The dataset covers a spectrum of disorders [40], such as Dysphonia, Functional Dysphonia, Hyperfunctional Dysphonia, Laryngitis, and Recurrent Nerve Paralysis. Our methodology involves several stages, including data preparation, augmentation, and model evaluation.

1) Data Preparation: The study examined whether the generation of spectrograms through the application of signal slicing with 400 ms windows and a 100 ms stride could facilitate improved model performance in comparison with the resizing of the frequency representation to 224x224 pixels. This contrasts with the approaches used in other studies. The hypothesis suggests that when mel-spectrograms of differing durations are resized to a common size, there is a loss of information. The mel-spectrograms were generated using these parameters: 512 samples separated successive frames, and each window contained a specific number of 2048 samples, and the number of mel filter banks was 128. These mel-spectrograms are illustrated in Figure 1.

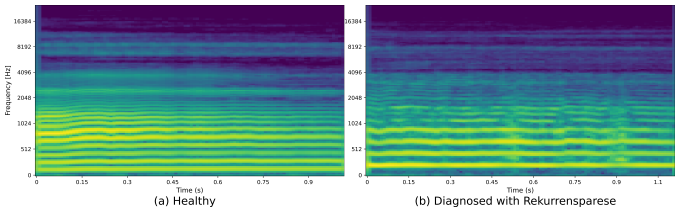

Fig. 1. The figure demonstrates a comparison between a spectrogram of a person without speech impairment (a) with that of a person with Rekurrensparese, in the dataset (b).

Furthermore, mel-spectrogram data augmentation techniques [41] have been investigated, encompassing time masking, frequency masking, a combination of these methods, and audio augmentation through noise - see Figure 2.

2) Model Evaluation: The training of each model was carried out using the Adam optimizer, incorporating various data enhancement techniques, including a baseline without augmentation. A one-fold stratified cross-validation approach was employed, with a 50% split for training and testing. Three of the configurations entailed training discrete models on individual vowels, whereas the fourth combined samples of /a/, /u/, and /i/ into a unified set. In the fifth approach, all available vowel samples for each subject (up to three) were combined into a single three-channel unit, creating an approach called the ”multichannel” approach. Early stopping was used to prevent overfitting. Models trained with segmented

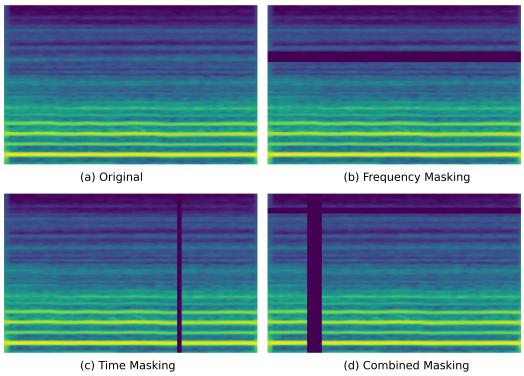

Fig. 2. The figure illustrates the initial spectrogram (a) and its alterations via frequency masking (b), time masking (c), and the application of both techniques together (d).

signals (slicing) rather than continuous signals were evaluated with noise-augmented and non-augmented audio data.

We investigated seven models: VGG-19 [25], ResNet-18, ResNet-101 [26], DenseNet-121 [42], EfficientNet-B2 [27], RegNet-X32GF [43] and DeiT [44].

3) Results: The best models obtained over the course of the experiments are presented in Table I. Notably, DenseNet achieved an F1 score of 0.74 for Multichannel and Slicing in both scenarios: without augmentation and with added noise. For most of the models, the highest scores were achieved without augmentation. The use of augmentation was not statistically significant for any specific model or across all results combined (p-value=0.38). Results regarding the use of augmentation in terms of the F1 score are presented in Figure 3.

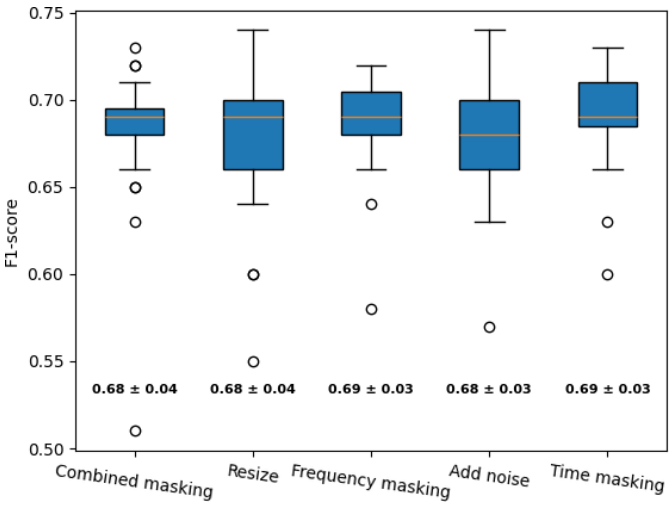

Fig. 3. F1 scores of specific models and augmentations

Dividing signal into windows corresponded to a significant loss of efficiency in DenseNet, EfficientNet, DeiT, and the results of all models combined. Augmentations, on the other hand, did not cause the results to change significantly in any model.

TABLE I
BEST RESULTS FOR EACH MODEL AND AUGMENTATION TYPE

Model	Slicing	Augmentation	Vowels	F1
VGG	**No**	**No**	**/a/ /u/ /i/**	**0.73**
	No	Add Noise	/a/	0.72
	Yes	Pad Zeros	/a/	0.72
	No	Add Noise	multichannel	0.72
	Yes	Pad Zeros	multichannel	0.72
ResNet-18	**No**	**Frequency Masking**	**/a/**	**0.72**
	Yes	**Pad Zeros**	**/a/**	**0.72**
	No	**Frequency Masking**	**/a/ /u/ /i/**	**0.72**
	No	Frequency Masking	multichannel	0.72
	Yes	Add Noise And Pad	/a/	0.71
ResNet-101	**No**	**Time Masking**	**/u/**	**0.71**
	No	Time Masking	/a/	0.69
	No	Frequency Masking	/a/	0.69
	Yes	Add Noise And Pad	/i/	0.69
	No	Combined Masking	/i/	0.69
DenseNet	**No**	**No**	**multichannel**	**0.74**
	No	**Add Noise**	**multichannel**	**0.74**
	No	Combined Masking	/a/	0.73
	No	Time Masking	/i/	0.73
	No	No	/i/	0.72
EfficientNet	**No**	**No**	**/a/**	**0.73**
	No	Frequency Masking	/a/	0.71
	No	Time Masking	/a/	0.71
	No	No	multichannel	0.70
	No	Combined Masking	/a/	0.70
RegNet	**No**	**Time Masking**	**/a/**	**0.72**
	No	Frequency Masking	/a/	0.71
	No	No	/a/	0.71
	No	Combined Masking	/a/	0.71
	No	Frequency Masking	/u/	0.71
DeiT	**No**	**Add Noise**	**multichannel**	**0.70**
	No	**Frequency Masking**	**/a/ /u/ /i/**	**0.70**
	No	**Time Masking**	**/a/ /u/ /i/**	**0.70**
	No	No	multichannel	0.69
	No	Combined Masking	/a/	0.69

The use of specific vowels or a combination of them significantly influenced the performance of VGG, DeiT, ResNet-101, and RegNet. Table II shows which vowel or group of vowels (Superior vowel) outperformed (Inferior vowel) with statistical significances as well as used statistical tests and calculated p-values.

The use of vowels combined into a multichannel signal resulted in statistically worse performance on ResNet-101. Across all other models and results combined, the difference was not statistically significant.

Results of the model, both in terms of their F1 score and statistical significance, are presented in Figures 4 and Table 5.

The Grad-CAM, Score-CAM, and Ablation-CAM techniques were applied to highlight the distinctions between normal and pathological states in mel spectrograms. Unlike the approach in [45], which showed these methods on a limited set of samples, we opted to compute averages from 50 randomly selected samples (Figures 6 and 7). These heatmaps visualize the areas of the input image or feature map that the specific model focuses on most while making predictions; the red areas indicate high importance, while the blue areas signify lower importance. The graphical depiction of the findings shows that the model weights are more prominently activated in the case of subjects with illnesses. Pathological samples demonstrate greater variability and intensity in color, both in multi-channel and single-channel data, highlighting the regions of the image

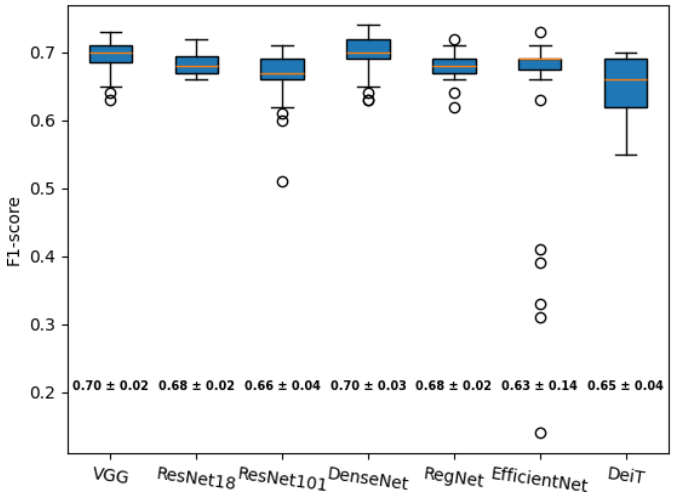

Fig. 4. F1 scores of specific models and augmentations

TABLE II
SIGNIFICANT DIFFERENCES BETWEEN VOWELS DIVIDED TO MODELS

Model	Superior Vowel	Inferior Vowel	p-value	test
VGG	/a/ /u/ /i/	/i/	0.014	t-test
	/a/ /u/ /i/	/u/	0.009	t-test
	/a/	/i/	0.039	Wilcoxon
	/a/	/u/	0.023	Wilcoxon
	multichannel	/u/	0.040	Wilcoxon
ResNet-101	/a/	multichannel	0.008	Wilcoxon
	/a/ /u/ /i/	multichannel	0.022	Wilcoxon
	/u/	multichannel	0.044	t-test
RegNet	/a/	/i/	0.001	t-test
	/a/	/a/ /u/ /i/	0.005	t-test
	multichannel	/i/	0.006	t-test
	multichannel	/a/ /u/ /i/	0.041	t-test
	/u/	/i/	0.001	t-test
DeiT	/a/ /u/ /i/	/i/	0.019	t-test
	/a/ /u/ /i/	multichannel	0.012	t-test
	/a/ /u/ /i/	/u/	0.021	t-test
	/a/ /u/ /i/	/a/	0.008	Wilcoxon
	/i/	/u/	0.042	t-test
	/i/	/a/	0.013	Wilcoxon
	multichannel	/a/	0.008	Wilcoxon
	/u/	/a/	0.039	Wilcoxon

that the model identified as critical for diagnosing health issues. Conversely, healthy subjects present a more consistent and subdued color pattern, indicating fewer problematic areas.

IV. SUMMARY

This research project examined the potential of deep learning architectures for diagnosing speech disorders based on vowel recordings. The study employed spectrograms from the Saarbruecken Voice Database, with a particular focus on the analysis of vowels /a/, /u/, and /i/. A number of CNN models were trained on the spectrograms, including VGG, ResNet-18, ResNet-101, DenseNet, EfficientNet, RegNet, and DeiT. The objective was to classify pathological and healthy voice recordings. Among the models tested, DenseNet demonstrated the highest performance, with an F1-score of 0.74, both when evaluated with multi-channel data and when noise was introduced to the audio samples.

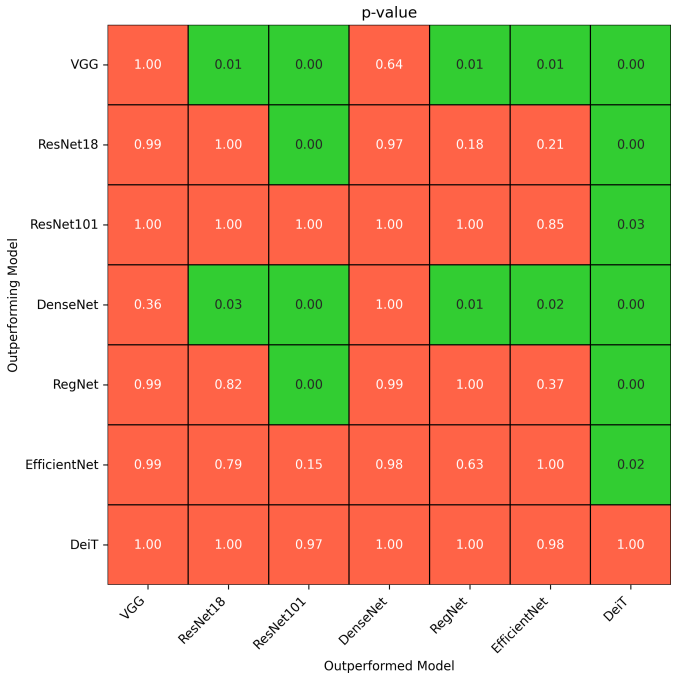

Fig. 5. Significant differences between models

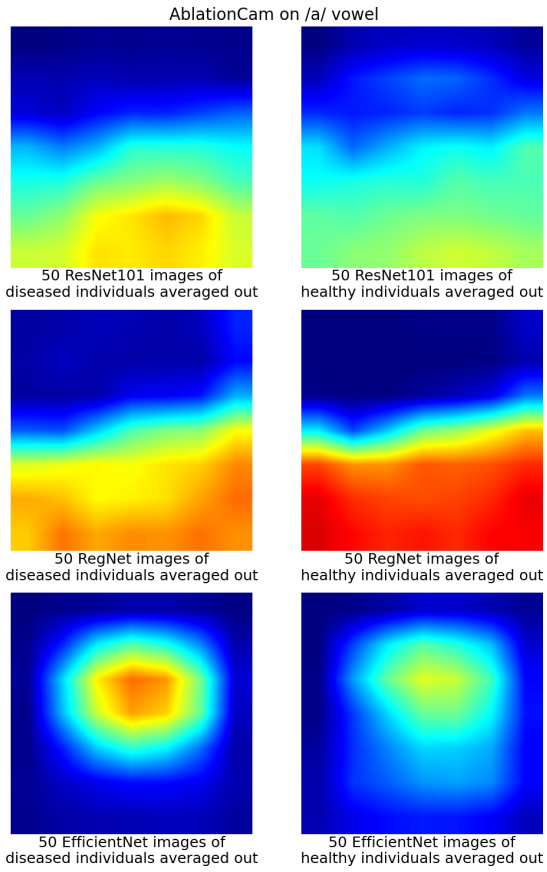

Fig. 6. Comparison of Ablation Cam results of /a/ vowel across different models.

Despite the success of DenseNet, statistical analysis revealed that the selection of vowels had a significant influence on the performance of the model in certain cases. To illustrate, the VGG model attained an F1-score of 0.73 when all vowel sounds were combined (/a/, /u/, /i/) but statistical tests (Wilcoxon and t-tests) demonstrated that vowel /i/ exhibited a significantly inferior performance compared to vowels /a/ and /u/ (p-values < 0.05). Furthermore, ResNet-101, which demonstrated moderate performance with a maximum F1-score of 0.71, also indicated that the combination of vowel sounds into multichannel signals resulted in statistically inferior outcomes compared to the use of single vowels ($p < 0.01$).

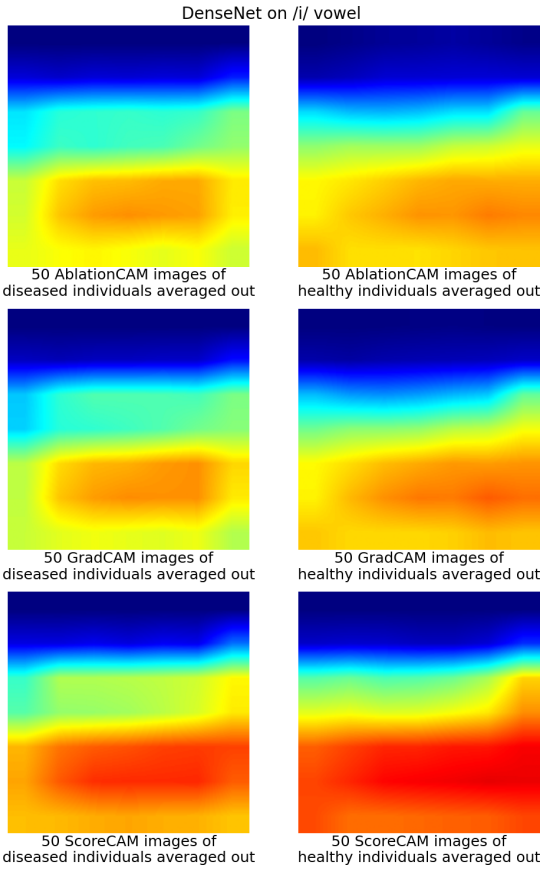

Fig. 7. Comparison of DenseNet results of /i/ vowel across different XAI visualization techniques.

The statistical analysis revealed significant discrepancies between the models in their performance with specific vowel groups. For instance, the RegNet model demonstrated superior performance with the vowel /a/, exhibiting an F1-score of 0.72, as compared to the combined vowel set ($p = 0.005$). Conversely, the DeiT model exhibited a statistically significant decline in performance with the combined vowels, as compared to single vowels /a/ and /u/, when evaluated on the combined vowel set ($p < 0.05$).

With regard to data augmentation, techniques such as frequency masking, time masking, and the addition of noise were employed, yet they did not result in notable enhancements in model performance. The partitioning of audio signals into smaller windows for training also failed to enhance model generalization.

XAI techniques were employed to provide visual explanations for the model's predictions. By averaging the specific XAU results across 50 samples, it was observed that for all models except EfficientNet, vital information came from lower frequency bands with a higher or lesser focus on the beginning and end of the recording. Different XAI visualization techniques provided similar, in terms of frequency and time results, which differed in intensity. Ablation-CAM is characterized by the most mild activation, and Score-CAm is the most intense one.

In conclusion, while DenseNet demonstrated strong performance, the study highlights the importance of refining vowel selection and signal processing techniques. Statistical analyses confirmed that certain vowels when used in isolation, outperform combined vowel sets in specific models.

REFERENCES

[1] F. Ratajczak, M. Najda, and K. Szyc, *Utilizing CNN Architectures for Non-invasive Diagnosis of Speech Disorders*. Springer Nature Switzerland, 2024, p. 218–226. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-61857-4_21

[2] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, and J. Wang, "Applications of deep learning to mri images: A survey," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 1–18, 2018.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[4] R. Shrivastav, D. A. Eddins, and S. Anand, "Pitch strength of normal and dysphonic voices," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2261–2269, 2012.

[5] R. Deepa, S. Arunkumar, V. Jayaraj, and A. Sivasamy, "Healthcare's new frontier: Ai-driven early cancer detection for improved well-being," *AIP Advances*, vol. 13, no. 11, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1063/5.0177640

[6] O. Obulesu, N. Venkateswarulu, M. Sri Vidya, S. Manasa, K. Pranavi, and C. Brahmani, *Early Prediction of Healthcare Diseases Using Machine Learning and Deep Learning Techniques*. Springer Nature Singapore, 2023, p. 323–338. [Online]. Available: http://dx.doi.org/10.1007/978-981-99-1588-0_29

[7] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in digital health*, vol. 4, p. 886615, 2022.

[8] D. Hemmerling, M. Wójcik-Pdziwiatr, P. Jaciów, B. Ziółko, and M. Igras-Cybulska, "Monitoring of parkinson's disease progression based on speech signal," in *2023 6th International Conference on Information and Computer Technologies (ICICT)*, 2023, pp. 132–137.

[9] M. L. Joshi and N. Kanoongo, "Depression detection using emotional artificial intelligence and machine learning: A closer review," *Materials Today: Proceedings*, vol. 58, pp. 217–226, 2022.

[10] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947–e11, 2019.

[11] L. van Bemmel, W. Harmsen, C. Cucchiarini, and H. Strik, *Automatic Selection of the Most Characterizing Features for Detecting COPD in Speech*. Springer International Publishing, 2021, p. 737–748. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-87802-3_66

[12] M. K. Reddy, P. Helkkula, Y. M. Keerthana, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku, "The automatic detection of heart failure using speech signals," *Computer Speech & Language*, vol. 69, p. 101205, 2021.

[13] R. Monir, D. Kostrzewa, and D. Mrozek, "Singing voice detection: a survey," *Entropy*, vol. 24, no. 1, p. 114, 2022.

[14] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[15] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[16] E. Keller, "The analysis of voice quality in speech processing," *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pp. 54–73, 2004.

[17] J. C. B. Gamboa, "Deep learning for time-series analysis," *arXiv preprint arXiv:1701.01887*, 2017.

[18] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," in *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2015, pp. 187–191.

[19] R. V. Sharan, H. Xiong, and S. Berkovsky, "Benchmarking audio signal representation techniques for classification with convolutional neural networks," *Sensors*, vol. 21, no. 10, p. 3434, 2021.

[20] S. Seo, C. Kim, and J.-H. Kim, "Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices," *Journal of Web Engineering*, pp. 497–522, 2022.

[21] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.

[22] T. Bartosch, D. Seidl *et al.*, "Spectrogram analysis of selected tremor signals using short-time fourier transform and continuous wavelet transform," 1999.

[23] P. Abdzadeh and H. Veisi, "A comparison of cqt spectrogram with stft-based acoustic features in deep learning-based synthetic speech detection," *Journal of AI and Data Mining*, vol. 11, no. 1, pp. 119–129, 2023.

[24] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[30] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021. [Online]. Available: https://arxiv.org/abs/2104.01778

[31] S. University, "Saarbruecken voice database," database of voice recordings for speech and voice disorders research. [Online]. Available: https://stimmdb.coli.uni-saarland.de/help_en.php4

[32] L. Vavrek, M. Hires, D. Kumar, and P. Drotár, "Deep convolutional neural network for detection of pathological speech," in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2021, pp. 000 245–000 250.

[33] F. T. Al-Dhief, M. M. Baki, N. M. A. Latiff, N. N. N. A. Malik, N. S. Salim, M. A. A. Albader, N. M. Mahyuddin, and M. A. Mohammed, "Voice pathology detection and classification by adopting online sequential extreme learning machine," *IEEE Access*, vol. 9, pp. 77 293–77 306, 2021.

[34] H. Ding, Z. Gu, P. Dai, Z. Zhou, L. Wang, and X. Wu, "Deep connected attention (dca) resnet for robust voice pathology detection and classification," *Biomedical Signal Processing and Control*, vol. 70, p. 102973, 2021.

[35] J.-Y. Lee, "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database," *Applied Sciences*, vol. 11, no. 15, p. 7149, Aug. 2021. [Online]. Available: http://dx.doi.org/10.3390/app11157149

[36] R.-K. Sheu and M. S. Pardeshi, "A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system," *Sensors*, vol. 22, no. 20, p. 8068, 2022.

[37] 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.

[38] H. Wang, M. Du, F. Yang, and Z. Zhang, "Score-cam: Improved visual explanations via score-weighted class activation mapping," *CoRR*, vol. abs/1910.01279, 2019. [Online]. Available: http://arxiv.org/abs/1910.01279

[39] H. G. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 983–991.

[40] Y. Naqvi and V. Gupta, *Functional Voice Disorders*. StatPearls Publishing, Treasure Island (FL), 2023. [Online]. Available: http://europepmc.org/books/NBK563182

[41] Y. Hwang, H. Cho, H. Yang, D.-O. Won, I. Oh, and S.-W. Lee, "Mel-spectrogram augmentation for sequence to sequence voice conversion," *arXiv preprint arXiv:2001.01401*, 2020. [Online]. Available: https://arxiv.org/abs/2001.01401

[42] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[43] J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, "Regnet: Self-regulated network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 9562–9567, 2022.

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers distillation through attention," 2021. [Online]. Available: https://arxiv.org/abs/2012.12877

[45] R. Jegan and R. Jayagowri, "Voice pathology detection using optimized convolutional neural networks and explainable artificial intelligence-based analysis," *Computer Methods in Biomechanics and Biomedical Engineering*, pp. 1–17, 2023.