BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(3), 2021, Article number: e137347
DOI: 10.24425/bpasts.2021.137347

SPECIAL SECTION

Deep learning vs feature engineering in the assessment of voice signals for diagnosis in Parkinson's disease

Ewelina MAJDA-ZDANCEWICZ[1*], Anna POTULSKA-CHROMIK[2], Jacek JAKUBOWSKI[1], Monika NOJSZEWSKA[2], and Anna KOSTERA-PRUSZCZYK[2]

[1] Faculty of Electronics, Military University of Technology, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland
[2] Department of Neurology, Medical University of Warsaw, ul. Banacha 1a, 02-097 Warsaw, Poland

Abstract. Voice acoustic analysis can be a valuable and objective tool supporting the diagnosis of many neurodegenerative diseases, especially in times of distant medical examination during the pandemic. The article compares the application of selected signal processing methods and machine learning algorithms for the taxonomy of acquired speech signals representing the vowel *a* with prolonged phonation in patients with Parkinson's disease and healthy subjects. The study was conducted using three different feature engineering techniques for the generation of speech signal features as well as the deep learning approach based on the processing of images involving spectrograms of different time and frequency resolutions. The research utilized real recordings acquired in the Department of Neurology at the Medical University of Warsaw, Poland. The discriminatory ability of feature vectors was evaluated using the SVM technique. The spectrograms were processed by the popular *AlexNet* convolutional neural network adopted to the binary classification task according to the strategy of transfer learning. The results of numerical experiments have shown different efficiencies of the examined approaches; however, the sensitivity of the best test based on the selected features proposed with respect to biological grounds of voice articulation reached the value of 97% with the specificity no worse than 93%. The results could be further slightly improved thanks to the combination of the selected deep learning and feature engineering algorithms in one stacked ensemble model.

Key words: voice processing; Parkinson's disease; non-linear analysis; convolutional networks.

1. Introduction

Parkinson's disease (PD) is one of the most common neurodegenerative diseases of the central nervous system. Despite many years of research, its etiopathogenesis has still not been fully studied, and symptomatic treatment is based primarily on administering medicine affecting the dopaminergic receptor. Furthermore, diagnosing this disease is still mainly based on the clinical assessment of a patient's motor status. In related works, devoted to the research and development of technical solutions supporting the diagnostics and evaluation of Parkinson's disease, the accelerometer-based data acquisition devices seem to be the most common thanks to the high sensitivity of available sensors in tremor measurements [1, 2]. However, tremor is not the only symptom of the disease and medical doctors are rather trying to identify the coexistence of several others. One of them is the deterioration in muscle activation responsible for the respiratory system, larynx, and articulation, which results in the reduced vocal loudness and clarity of speech.

2. Related work

Characteristic voice changes experienced by ca. 70 to 90% of patients already at the early stages of the disease are usually variable [3, 4]. The results of Parkinson's disease recognition system are directly related to the selection of relevant feature extraction and the classification method. The acoustic analysis of the voice signal can be conducted using different signal processing tools to extract various voice features.

The most studied parameters of the acoustic analysis in the current literature are fundamental frequency F_0 and parameters describing its variability in time (jitter). *Jitter* can be determined for its several measures and finally received parameters like *Jitta*, *PPQ5*, *RAP*. The jitter is affected mainly by the lack of control of the vibration of the cords. The next group of parameters describes the signal energy in general. *Shimmer*, which is the amplitude variation of the sound wave, is determined by *shimmer parameters* like *ShdB*, *APQ3*, *APQ5* [5–7].

Apart from the baseline features, other features were distinguished: *HNR* (harmonic-to-noise ratio), *NHR* (noise-to-harmonic ratio), *ZCR* (zero-crossing rate), and its modification *HZCRR* (high zero-crossing rate ratio). As a complement of these groups of features, several statistical measures used to describe them are also calculated [5, 8].

In addition to basic time and frequency methods used to extract features, perceptually motivated signal representations are also applied in PD recognition. Such representations, for instance, characterize the cepstral techniques. Cepstral coefficients have been rated as having the best correlation with the degree of voice pathologies in clinical research [5, 9, 10]. Also, Q-factor wavelet transform (*QWT*) was applied to vocal signals of the individuals for the diagnoses of PD [5].

The new trend in PD research is that most studies use the combination of different feature types to perform the classifi-

[*] e-mail: ewelina.majda@wat.edu.pl

Manuscript submitted 2020-10-26, revised 2021-01-07, initially accepted for publication 2021-01-15, published in June 2021

cation task rather than using separate feature types in model training. Extended feature space in these studies can be reduced via feature selection methods.

Most of the studies presented in the literature describe research that has been conducted with different databases and a limited number of recordings. Furthermore, in some articles, the sample is very small, and the algorithms have been tested on unequal patient groups. Several studies do not describe the duration and the severity of the disease.

3. Contribution

The objective of the experiments conducted by the authors in this paper is to develop a target diagnostic system which facilitates an objective taxonomy of patients' voices by means of digital signal and image processing including the use of machine learning algorithms. Such a system could be treated as technical support for physicians evaluating subjectively the voice disorder in PD and would allow them to start treatment and monitor its effectiveness during a distant medical examination. First, the paper presents the way a convolutional neural network (CNN) can be used to recognize the voice coming from PD patients. The general motivation behind the application of the CNN networks is their ability to solve many recognition tasks without using experts to find relevant features [11]. In the case of a lack of sufficient domain knowledge concerning the recognized subjects, such an approach could be of non-trivial importance. Then the study discusses the so-called feature engineering which means finding sets of parameters to be used as a base to generate feature vectors for modelling the voice of a patient including those derived from the expert knowledge of physicians. The presented material evaluates the non-linear speech signal analysis, additionally taking perceptual frequency scales into account. The vectors of selected voice descriptors were used in the classification scheme based on a classical SVM neural network and then compared to the deep learning approach utilizing images of the time-frequency representations of voice signals.

There has been an increase in interest in deep neural networks over the last few years. Up to our knowledge, deep learning architecture is a new issue in Parkinson's disease diagnosis by voice [12, 13]. However, the comparison of the results of PD recognition obtained from the deep learning approach and the traditional feature engineering is still limited. The comparison presented in this paper will lead to a better understanding of the advantages and disadvantages of each approach.

4. Data pool

The tests were conducted within the database of the patients' voice recordings collected in the Department of Neurology at the Medical University of Warsaw with the participation of the medical personnel. The test bench consisted of the Shure MX58 dynamic microphone connected to a personal computer with dedicated software via a USB adapter containing a preamplifier and ADC converter [14]. The frequency response of the set was configured especially for the voice sounds and enabled recordings over the frequency range from 50 Hz to 15 kHz. The microphone sensitivity of –54.5 dBV/Pa was as low as in other dynamic microphones, but it was sufficient enough for recording the voice at small distances from a patient's mouth. All the recordings were registered with a sampling rate of 44.1 kSa/s and a 16-bit resolution.

The acquisition scenario involved recording the vowel *a* with prolonged phonation, uttered by a patient for at least 5 seconds. The acquired voltage signals were then normalized by a factor representing their root mean square value. Exemplary acquired waveforms corresponding to both categories are depicted in Fig. 1. More problems with stable phonation resulting in an amplitude modulation can be observed in the case of a PD patient.

The patients participating in the study were selected according to the UPDRS score (Unified Parkinson's Disease Rating Scale) [4]. The material comprised of 22 patients (14 women and 8 men, aged 28–70, average: 55.5) with diagnosed Parkinson's disease. All patients were treated with L-dopy preparation for a period of at least 6 months. The disease duration ranged from one to 12 years (average: 5.3 years). The symptom severity was assessed as per the above scale – part III and amounted to 20.25 points on average. The control group was used as a reference comprising 22 persons without a diagnosed PD. They were 14 men and 8 women, aged 40 on average. The tests were conducted using the recordings articulated in two

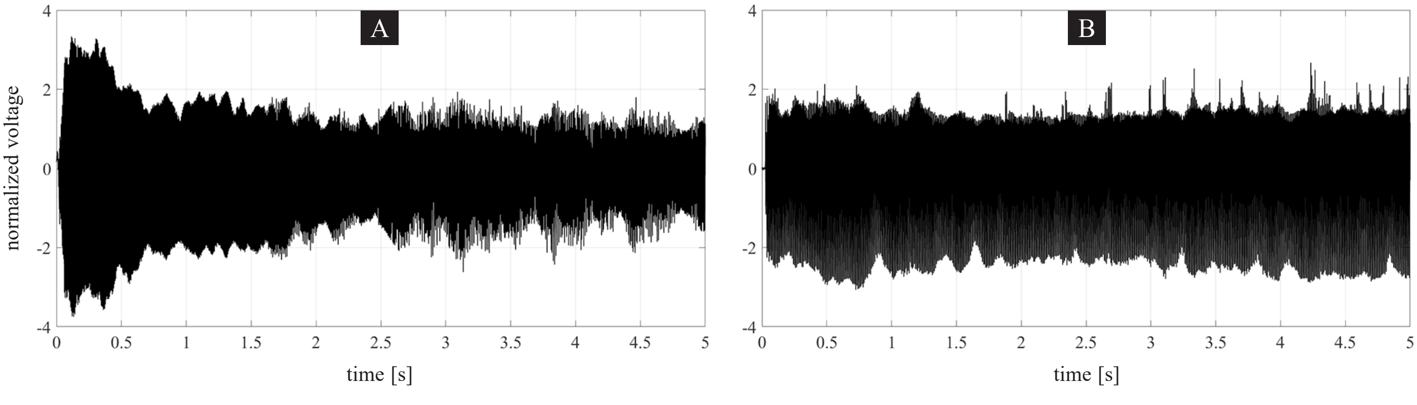


Fig. 1. Voice waveforms of the vovel *a* with 5 s pronation corresponding to a sick (A) and healthy subject (B).

successive registrations by each person. Hence, there were 88 signals acquired in total for further study, i.e. 44 voice recordings of sick people and 44 recordings of healthy people.

5. Assessment scheme

In order to reliably assess the proposed methods for recognition of PD patients, the voice database should be partitioned into subsets used for training and evaluating the performance of classifiers. The commonly used cross-validation method facilitates doing so without the risk of obtaining too optimistic and unreliable assessments, especially when the data pool is not too numerous. It involves random division of the entire data set into *N* folds of equal size. Next, a single fold is used to validate the model, while the remaining subsets are used in training. The process is repeated *N* times so that within each repetition the validating and learning sets are separated. The classification results obtained with the validation folds are then averaged. The value *N* is usually set from 5 to 10, but there are no special governing rules in this respect [15]. In this research, in order to use the testing and training corpora independently of a person, a single fold in cross-validation contained 8 recordings corresponding to all data representing 2 people from each class. In this way, the 11-fold cross-validation was preserved, and the data pool acquired from 40 people was used in training, and data from 4 people was used in the validation. At the same time, the requirement for the data of people from the validation group excluded from the learning group was satisfied.

The classification results were expressed according to the nomenclature preferred by physicians, just like sensitivity *Se* and specificity *Sp*

$$Se = \frac{TP}{TP + FN}, \tag{1}$$

$$Sp = \frac{TN}{TN + FP}. \tag{2}$$

The metrics in the above equations are taken from the concept of a confusion matrix [15]. This is a simple cross-tabulation of the actual and recognized classes and facilitates easily calculating the classifier parameters. Its diagonal cells denote the number of people *TP* correctly classified as sick and the number of people *TN* correctly classified as healthy while the off-diagonal cells contain the number of people classified in a wrong way. *FP* stands for healthy cases classified as sick and *FN* for sick ones classified as healthy. The confusion matrices may contain some additional information that is presented further in Section 4.

6. CNN – a deep learning approach

Convolutional neural networks (CNNs) are nowadays treated as versatile deep learning tools for automatic feature generation and recognition. Although primarily used in visual recognition contexts, they have been also successfully used in many one-dimensional signal processing tasks including speech, vibroacoustic, biomedical, and seismology applications [16–19]. In all of these cases, initial processing is required to obtain an image-like representation of the analyzed signal. The main advantage introduced by CNNs is that they shift the burden of the hand-crafted feature design to the system of learning along with the classification task. In other words, the nets are trained to begin from a raw input image to the final output of sufficiently labelled classes. They are successfully used in the detection, recognition, and semantic scene segmentation. Formally, the nets can be built and trained from scratch; however, they should be advanced enough to recognize objects properly and usually require a huge amount of time for training even when multiple GPUs are used. And yet, there is another alternative approach based on transfer learning that is capable of leveraging the power of a CNN. The idea is to take a pretrained network released by others and use it as an initialization for the task of interest. In this research, the results of using one of the simplest nets, called *AlexNet*, are presented [20]. The net has gained popularity thanks to its availability and numerous examples of successful use [21]. The network comprises 25 layers but only 8 of them are optimized in training: 5 convolutional layers and 3 fully connected layers. The net used for distinguishing PD patients from healthy people must be adopted to have the last layer of the same size as the number of classes in recognition. Moreover, a two-dimensional representation of a voice signal is required to feed the net. In general, voice signals are nonstationary, and their properties can be adequately described by using joint information in the time and frequency domain. The short-time Fourier transform and the square of its modulus called spectrogram is a well-known signal processing methodology to explore the instantaneous spectrum of such signals. The computations are based on discrete Fourier transform performed separately on short frames containing segments of equal length extracted from a signal. Such a procedure provides a visual representation of the signal because the Fourier spectra can be plotted as a function of time in the form of an image. Using a CNN network supported with spectrogram images is quite a common approach utilized by many researchers [16–18]. Most of them just use the distribution of energy over the time-frequency plane calculated with a fixed frame length. Some put the real and imaginary parts of the short-time Fourier transform into separate layers of the network input. Nevertheless, the length of the frame sets the frequency resolution and strongly affects the spectrogram image in which a color map is used for coding the values of the evolutionary spectrum as depicted in the upper row of Fig. 2, where three spectrograms of a different resolution described by three frame lengths are presented on a log scale. The frame lengths considered as short, medium, and long were determined by the FFT dimensions equal to 256, 1,024 and 4,096. The images shown in Fig. 2 are time-frequency representations of the PD voice signal from Fig. 1A. In contradiction to this concept, an alternative approach is proposed in this paper. It first assumes finding three spectrograms of different resolutions, expressing them as monochrome images, and then combining as three-color channels forming an RGB image. Thanks to that

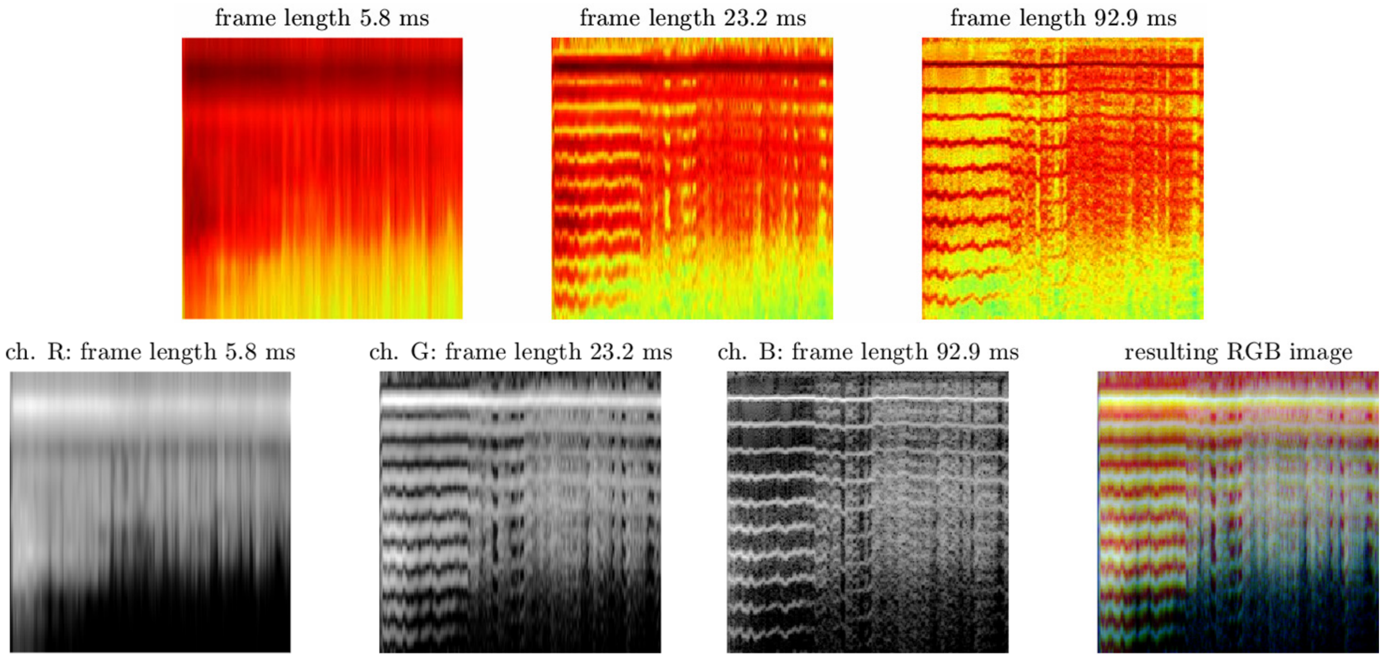


Fig. 2. Strategies of forming time-frequency image data from voice to feed a CNN network: 3 separate RGB spectrogram images of different frame lengths (upper row) and 3 channels corresponding to spectrograms of different frame lengths forming one RGB spectrogram image (lower row)

the net *AlexNet* in the transfer learning strategy was fed with images containing spectrograms of three different resolutions at the same time as depicted in the scheme presented in Fig. 2 in the lower row. The comparison of results in the form of confusion matrices obtained with respect to the assessment and cross-validation scheme proposed in Section 3 are depicted in Fig. 3.

There is some additional information in the matrices shown in Fig. 3. The column on the far right represents the percentages of all patients recognized as belonging to each class, correctly (green) and incorrectly (red) classified. Naturally, the values in green are metrics described by (1) and (2). The metrics in the lowest row show the number of patients that belong (green) and do not belong (red) to each class related to the number of cases that are correctly and incorrectly assigned to that class. The cell in the bottom right corner shows the overall accuracy and error.

As it can be seen, the concept of using triple resolution spectrogram images in the deep learning approach yielded the best results as compared to the standard concept when only one resolution was used. The overall accuracy was almost 6% better than the best result achieved when the spectrogram frame was fixed to the medium value.

7. Feature engineering approach

The results presented in the last section devoted to the deep learning approach are far from being perfect and encourage the development of an alternative processing concept based on feature engineering. Three non-linear sound processing techniques were proposed as a base for this concept aimed at finding possible hand-crafted features.

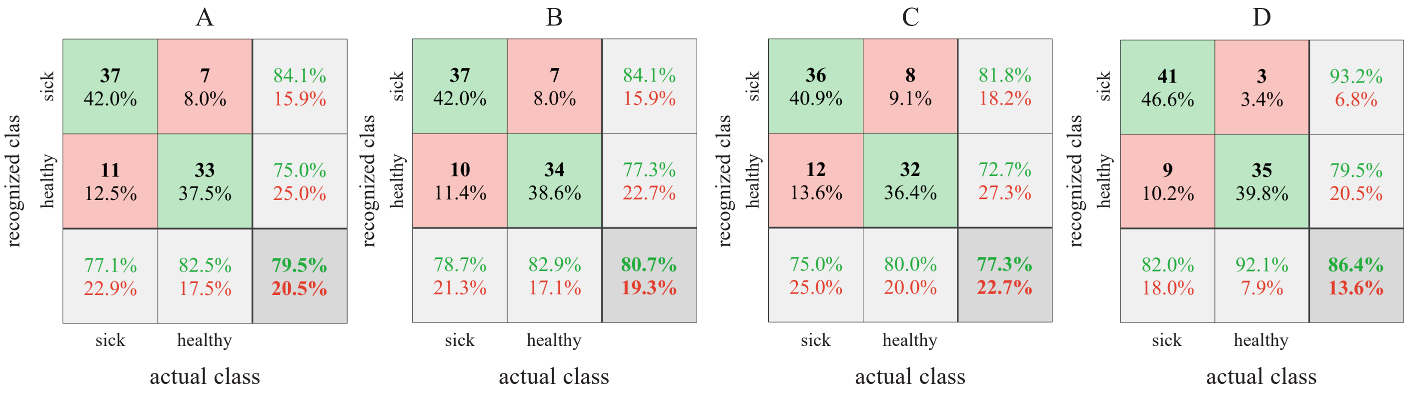


Fig. 3. Confusion matrices describing taxonomy of voice in Parkinson's disease based on CNN and spectrograms: A, B, C – single resolution spectrograms with frames 5.8 ms, 23.2 ms, and 92.9 ms, D – triple resolution spectrogram

The techniques are all outlined in the block diagram of the conducted analysis as depicted in Fig. 4. It was decided to apply 12 filter banks for each of the methods from the block diagram.

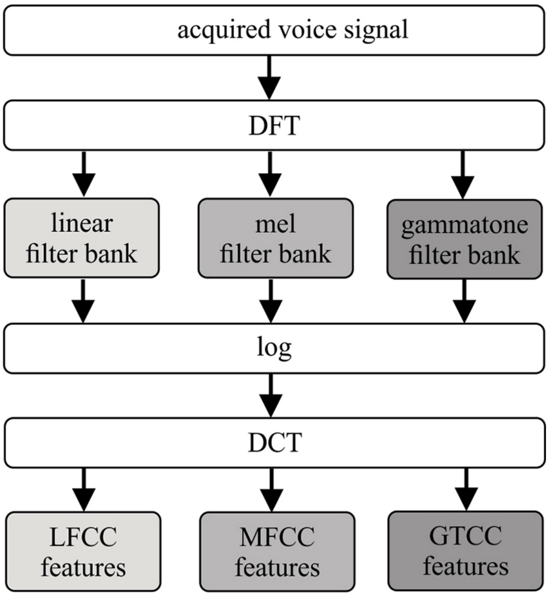


Fig. 4. Algorithms used for the acoustic analysis of voice

Consequently, three 12-dimensional feature vectors denoted further as *LFCC*, *MFCC* and *GTCC* were analyzed which were assumed to reflect the voice signal properties of the patients as suggested in [22, 23]. The vectors contain cepstral coefficients (*CC*) corresponding to frequency scales the techniques use: linear-frequency (*LF*), mel-frequency (*MF*), and gamma-tone (*GT*). The scales are depicted in Fig. 5 and a detailed discussion of the proposed features is given below with respect to the physiology of voice articulation.

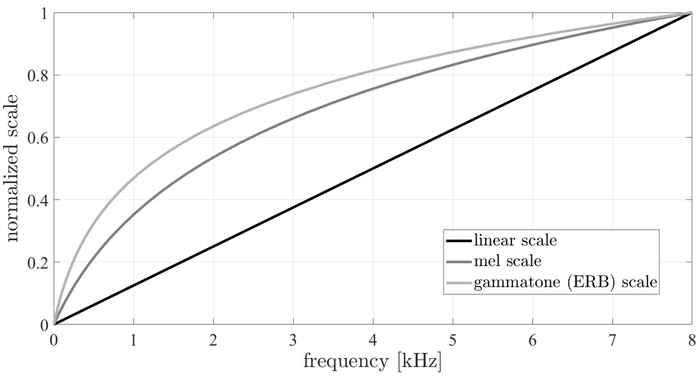


Fig. 5. Frequency scales used in non-linear voice processing

7.1. Biologically inspired features. One of the methods in voice parametrization is the cepstral analysis based on the so-called homomorphic technique. The analysis in its classic form is based on transforming a signal to the so-called pseudo-time domain using the spectral logarithm for the frequency spectrum model, called the real cepstrum [24]. In general, cepstral techniques facilitate easy deconvolution of the glottal excitation associated with the work of vocal folds and the component associated with the voice tract. Furthermore, since the human sense of hearing analyses the amplitude spectrum and very effectively intercepts voice dysfunctions, it can be assumed that it is enough to focus on the actual cepstrum, which would help us to avoid calculating the troublesome complex logarithm. The modulus of the cepstrum for the vowel *a* is shown in Fig. 6.

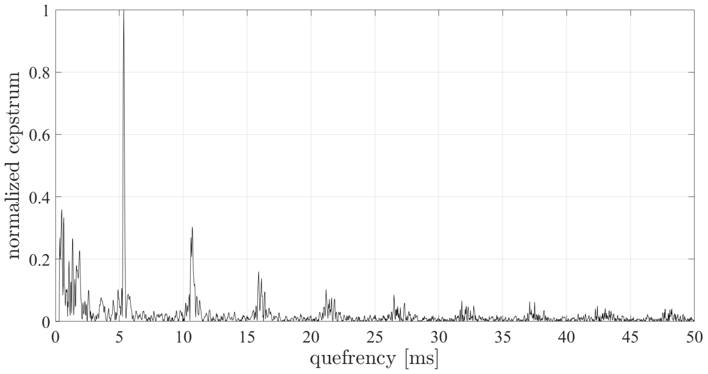


Fig. 6. The modulus of the cepstrum when the vowel *a* is analyzed

It indicates the fundamental tone period T_0 (approx. 5 ms) and pseudo-times, which characterize the vocal tract resonances (formant frequency inverses), focused around zero pseudo-time. Because the information associated with the vocal tract transmission is concentrated around zero pseudo-time, this is the area to look for concise information on what is being said. It is greatly degraded when using this technique. On the other hand, pseudo-times beyond the laryngeal sound period emphasize the concise information associated with the shape of laryngeal pulses, and since it is closely related to the anatomical structure of the larynx and glottis, it carries a lot of diagnostic information.

The cepstral method is an algorithm that raises increasing hopes for clinical and practical applications. As part of a large meta-analysis studying the usefulness of acoustic methods in diagnosing dysphonia, which covered 25 publications and 87 acoustic parameters, an international team of authors evaluated cepstral coefficients as the best correlating with a degree of dysphonia [25, 26]. A dysphonic voice is generated when any of the vocal system elements is disturbed. The breathing manner, phonation and respiratory coordination, phonation time, resonator activation, voice intensity and pitch, voice scope and average position, its timbre, and sonorousness can be abnormal. Patients with Parkinson's disease experience weakened muscles of the throat, soft palate, tongue, and mouth. A patient's speech is characterized by respiratory, articulatory, and phonation disorders resulting from the damaged subcortical nuclei (extrapyramidal system), the corpus striatum, and the dorsal pallidum in particular [27].

The outcome of the transformations is a number of linear-frequency cepstral coefficients (*LFCC*) determined by the user. In this work it was set to 12.

7.2. Other voice features. A variation of the cepstral method is the melcepstral method providing mel-frequency cepstrum coefficients (*MFCC*), which uses the non-linearity of sound perception by humans. The so-called perceptual scale used within this algorithm is characterized, in approximation, by linear mapping of low frequencies and logarithmic mapping of high frequencies, which results from a subjective connection between the frequencies of pure harmonic tones and frequencies perceived by humans [28]. The mel scale was proposed in 1937 by Stevens, Volkman, and Newman and is described by

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{\mathrm{Hz}}}{700\ \mathrm{Hz}}\right). \tag{3}$$

To put it simply, the *MFCC* method can be treated as a spectral band analysis conducted within the auditory system.

The third proposed algorithm (*GTCC*) is based, like the *MFCC* method, on the bandwidth of human hearing. The only, yet very important difference, is the application of another filter bank. The *GTCC* method involved the application of a band-pass filter designed using the gammatone function [29, 30]. Essentially, modelling involves forming the so-called filter bank, which covers the auditory frequency range. The filter bank is most usually a system of gammatone filters at intervals corresponding to the equivalent rectangular bandwidth or the *ERB*. In terms of numbers, it is equal to the bandwidth of an ideal rectangular filter with a transmittance value equal to the maximum transmittance of an auditory filter, while the power of the noise passing through this filter is equal to the power of the noise passing through the auditory filter [31]. The waveform of the dependence of an auditory filter equivalent bandwidth and the frequency is described by the subsequent relationship

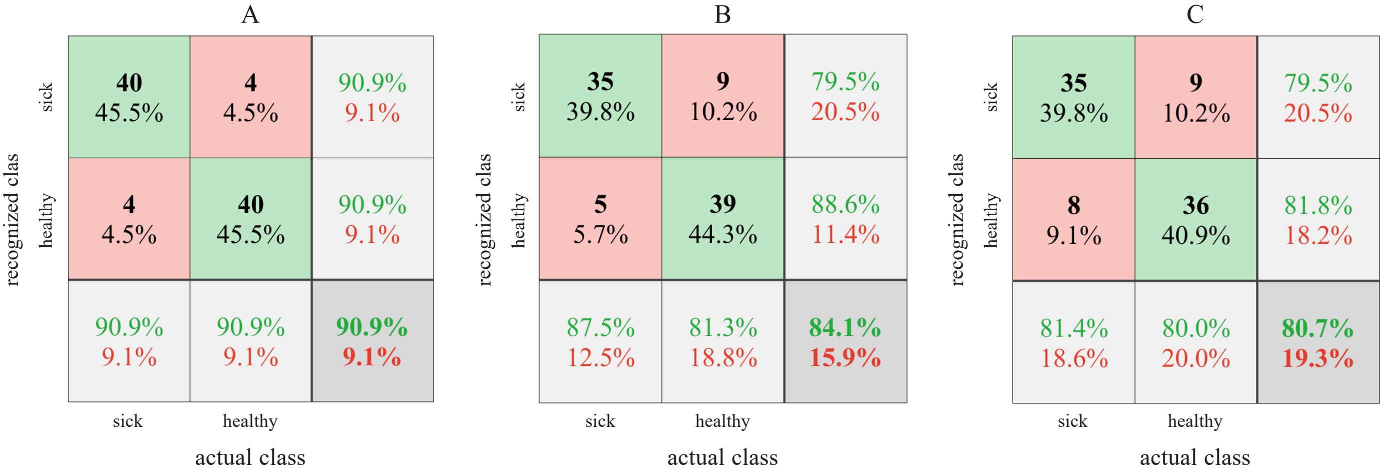
$$ERB = 24.7 \cdot (4.37F + 1), \tag{4}$$

where F is the filter mid-band frequency expressed in kHz. Therefore, in order to achieve a filter distribution of the gammatone filter bank type, one needs to define the mid-band frequencies as linearly distributed along the *ERB* scale, covering a specific frequency range with the desirable filter number.

7.3. Results of recognition. The abilities of the 12-dimensional vectors *LFCC*, *MFCC*, and *GTCC* to differentiate between healthy and PD patients were examined with the standard machine learning method finding the optimal separating hyperplane, i.e. a support vector machine SVM [32]. The conducted experiments showed that the most accurate results of recognition were achieved when the mapping to a higher dimension was performed by a cubic polynomial kernel. Because of a limited data sample, the resampling procedure based on 11-fold cross-validation was adopted again as described in Section 3. Confusion matrices found for all of the three vectors are depicted in Fig. 7. One can easily notice that this time the results are better as compared to the deep learning approach and that the vector of *LFCC* features appeared to be the best in this comparison exceeding the level 90% of the overall accuracy.

Moreover, it should be also noticed that the features in vectors were not analyzed to optimize the dimensionality of the space prior to the SVM recognition. The dimension of 12 in conjunction with 80 training samples introduces a sparse space and seems to be very close to the common requirement saying that there should be at least several samples for each dimension.

7.4. Feature selection. To decrease the dimensionality and, additionally, to take the advantage of possible synergy among the features as well the synergy of features with the SVM classifier, the algorithm of their selection based on backward sequential search strategy was applied [32]. In the strategy, the combinations of features selected by removal from a full



A

recognized clas	sick	healthy	
sick	**40** 45.5%	**4** 4.5%	90.9% 9.1%
healthy	**4** 4.5%	**40** 45.5%	90.9% 9.1%
	90.9% 9.1%	90.9% 9.1%	**90.9%** **9.1%**

actual class

B

recognized clas	sick	healthy	
sick	**35** 39.8%	**9** 10.2%	79.5% 20.5%
healthy	**5** 5.7%	**39** 44.3%	88.6% 11.4%
	87.5% 12.5%	81.3% 18.8%	**84.1%** **15.9%**

actual class

C

recognized clas	sick	healthy	
sick	**35** 39.8%	**9** 10.2%	79.5% 20.5%
healthy	**8** 9.1%	**36** 40.9%	81.8% 18.2%
	81.4% 18.6%	80.0% 20.0%	**80.7%** **19.3%**

actual class

Fig. 7. Confusion matrices assessing full, 12-dimensional vectors of various cepstral coefficients: A – *LFCC* vector, B – *MFCC* vector, and C – *GTCC* vector

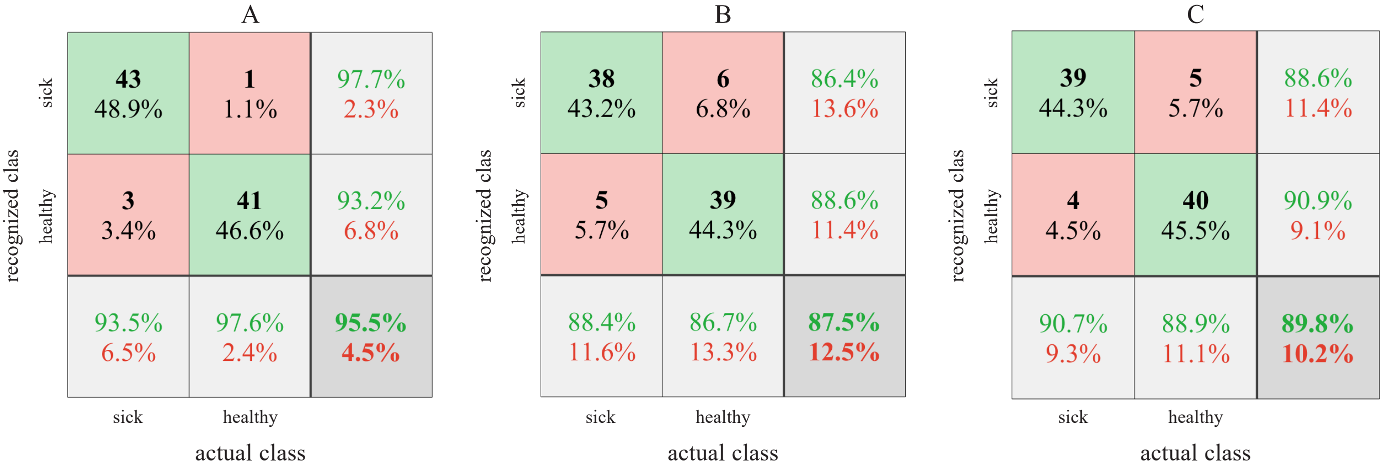


Fig. 8. Confusion matrices assessing vectors of selected cepstral coefficients: A – vector of 7 *LFCC* features, B – vector of 7 *MFCC* features, and C – vector of 8 *GTCC* features

candidate set are assessed by a class separability criterion that uses the misclassification rate given by the classifier. As a result, specific interactions between the classifier and the dataset of reduced dimensionality may be revealed and considered to improve the accuracy of recognition. The results of such a selection are depicted in Fig. 8, where confusion matrices described for new vectors of lower dimensions working with the SVM classifier are shown. In all cases, a significant improvement in overall accuracy can be observed. The highest score was found for a subset of seven features defined again by the *LFCC* vector and exceeded the level of 95%.

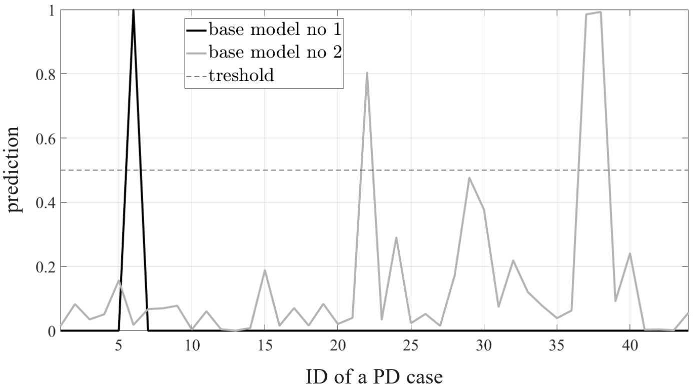


Fig. 9. Results of prediction of both base models for all PD patients

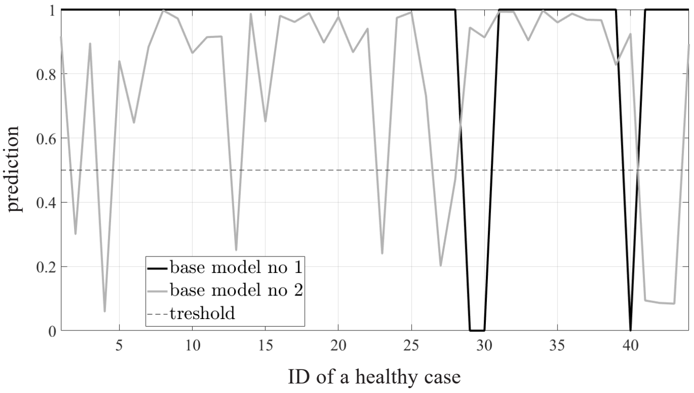


Fig. 10. Results of prediction of both base models for all healthy subjects

8. Ensemble of models

Both of the above machine learning algorithms, i.e. the deep learning approach and feature engineering, are strongly diverse. They use considerably different preprocessing of the same voice recordings and the assumptions behind the methods about how to solve the predictive modeling task are also very different. At the same time, some versions of the algorithms have good skills in the validation data set. All that creates a chance to combine them and use them as base models in a hybrid, stacked ensemble model to improve the accuracy upon the accuracy of the individual ones [33]. As the first base model, the SVM trained with vectors of seven selected *LFCC* coefficients was used as it yielded the best results among other cepstral subsets. The CNN fed with triple resolution spectrograms was used as the second base model for the same reason. The comparison of the prediction results between these two models in the case-relation is depicted in Fig. 9 for all PD patients and Fig. 10 for all healthy subjects. In the case of the CNN model, the output of the softmax layer was used. In the figures, predictions below the threshold 0.5 correspond to the detection of Parkinson's disease and predictions above it denote healthy subjects.

As it can be seen, there are misclassifications corresponding to the confusion matrices presented in Figs. 3D and 8A; however, they are not correlated and the use of the fusion of classifiers gives the hope for better results. The architecture of the stacking model includes a metamodel that learns how to best summarize the predictions of the base models arranged in one layer. The metamodel is usually simple as it is to provide a smooth interpretation of the predictions made by the base models. In this work, the SVM with polynomial kernel of the 2nd order was used for that purpose. The structure of the ensemble is depicted in Fig. 11.

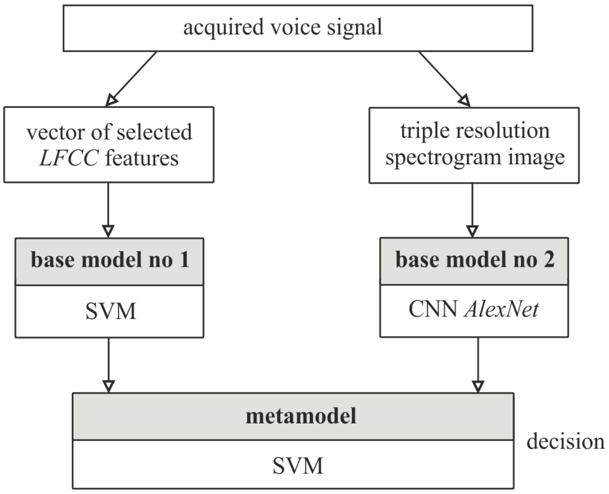
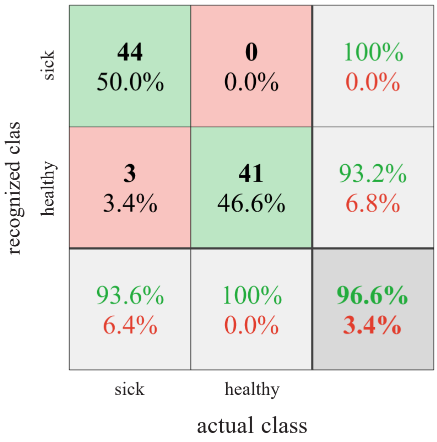


Fig. 11. The structure used as the stacked ensemble model

To avoid overfitting, the training data was divided into two folds. The base models were trained with the use of the data contained in the first fold. Then their predictions made for observations from the second fold were used in training as inputs to the metamodel. The results of recognition in the form of a confusion matrix are shown in Fig. 12. As expected, the hybrid model performed slightly better as compared to single individual models (see Figs. 3D and 8A for comparison). The overall accuracy was almost 97% and all PD cases were recognized properly.



recognized clas				
	sick	44 50.0%	0 0.0%	100% 0.0%
	healthy	3 3.4%	41 46.6%	93.2% 6.8%
		93.6% 6.4%	100% 0.0%	96.6% 3.4%
		sick	healthy	
		actual class		

Fig. 12. Confusion matrix assessing a stacked ensemble model based on deep learning approach and feature engineering

9. Conclusions

The presented material discusses the possibilities of using various methods of contemporary machine learning algorithms as an aid in the medical diagnosis of patients with possible Parkinson's disease on the basis of their voice samples.

An attempt was undertaken to compare the traditional approach used in signal classification and based on feature engineering with the modern one using advanced deep neural classifier finding relevant features automatically. In the latter, the so-called transfer learning was adopted. The results of the comparison including various configurations of decision frameworks with respect to the nomenclature used within the medical community, i.e. the results expressed in terms of sensitivity and specificity of the tests, are summarized in Table 1.

Table 1
Percentage values of sensitivity (*Se*) and specificity (*Sp*) obtained in tests performed with various frameworks examined in this article

Framework		***Se*** **[%]**	***Sp*** **[%]**
single resolution spectrogram image + CNN, short frame 5.8 ms		84.1	75.0
single resolution spectrogram image + CNN, medium frame 23.2 ms		84.1	77.3
single resolution spectrogram image + CNN, long frame 92.9 ms		81.8	72.7
triple resolution spectrogram image + CNN		93.2	79.5
LFCC + SVM	all 12 features	90.9	90.9
	selected 7 features	97.7	93.2
MFCC + SVM	all 12 features	79.5	88.6
	selected 7 features	86.4	88.6
GTCC + SVM	all 12 features	79.5	81.8
	selected 8 features	88.6	90.9
stacked ensemble model		100	93.2

The application of a deep convolutional neural network to recognize the disease on a patient's voice must be preceded by the preparation of image-like data. A very convenient method to do so is to use a spectrogram of the voice signal based on its short-time Fourier transform. However, the results reported in this article show that the quality of the diagnostic test depends on the way the image data is prepared. When the color channels of the image prepared for the neural network contained three spectrograms corresponding to various frame lengths, better sensitivity and specificity were obtained as compared to the cases when the decision was based on spectrogram images corresponding to one fixed resolution.

In the research, only one and relatively simple convolutional network was used. It was the *AlexNet*, which yielded the sensitivity of 93.2% and specificity of 79.5%. It cannot be excluded that using deeper networks, just like those from the example of *VGG* and *ResNet* families, can produce better results. Research on the adaptation of such networks is the aim of the next steps.

Although the deep learning *AlexNet* approach is very attractive, it did not outperform traditional feature engineering in the recognition of the disease based on voice. The application of the cepstral coefficients found thanks to the domain knowledge on voice articulation and the method of sequential feature selection, led to the 97.7% sensitivity and 93.2% specificity.

The presented material extends the findings discussed in the most comparable paper [13] devoted to the application of a convolutional neural network to Parkinson's disease detection by the voice image processing. Although lower overall accuracy was reached: 86.4% with pretrained *AlexNet* vs 91.7% with pretrained *ResNet*, it was shown that better results could be achieved using preprocessing based on the combination of spectrograms of three different resolutions instead of one. It was also shown that the featured engineering approach based on biologically inspired cepstral coefficients yields better results, with an overall accuracy of 95%.

Besides, it was evident that the experiments conducted with the use of a stacked ensemble model based on selected deep learning image processing and cepstral feature engineering algorithms could further improve the results. The improvement was rather slight as compared to the result obtained with the use of *LFCC* features alone but we managed to reach the value of 96.6% of overall accuracy which is over 5% better than the accuracy shown in [13].

The presented method seems to be attractive for a distant medical examination as it potentially excludes personal contacts between the patient and the doctor. In further research, finding the regression models predicting the extent of the disease is planned with the use of one of known clinical scales, e.g. UPDRS.

Acknowledgements. Project co-financed from the funds of the Polish Ministry of National Defense for the implementation of basic research within the research grant No. GBMON/B9/13–996/2018 "Basic research in the field of sensor technology using innovative data processing methods".

References

[1] Y.D. Kumar and A.M. Prasad, "MEMS accelerometer system for tremor analysis", *Int. J Adv. Eng. Global Technol.* 2(5), 685–693 (2014).

[2] P. Pierleoni, "A Smart Inertial System for 24h Monitoring and Classification of Tremor and Freezing of Gait in Parkinson's Disease", *IEEE Sens. J.* 19(23), 11612–11623 (2019).

[3] W. Pawlukowska, K. Honczarenko, and M. Gołąb-Janowska, "Nature of speech disorders in Parkinson disease", *Pol. Neurol. Neurosurg.* 47(3), 263–269 (2013), [in Polish].

[4] S.A. Factor, *Parkinson's Disease: Diagnosis & Clinical Management*, 2nd Edition, 2002.

[5] R. Chiaramonte and M. Bonfiglio, "Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies", *Rev. Neurologia* 70(11), 393–405 (2020).

[6] Jiri Mekyska, *et al.*, "Robust and complex approach of pathological speech signal analysis", *Neurocomputing* 167, 94–111 (2015).

[7] B. Erdogdu Sakar, G. Serbes, C. Sakar, "Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease", *PLoS One* 12, 8 (2017)

[8] L. Berus, S. Klancnik, M. Brezocnik, and M. Ficko, "Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks", *Sensors (Basel)* 19(1), 16 (2019).

[9] L. Jeancolas *et al.*, "Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients", *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2017, pp. 1–6.

[10] D.A. Rahn, M. Chou, J.J. Jiang, and Y.Zhang, "Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis", *J. Voice* 21, 64–71 (2007).

[11] J. Kurek, B. Świderski, S. Osowski, M. Kruk, and W. Barhoumi, "Deep learning versus classical neural approach to mammogram recognition", *Bull. Pol. Acad. Sci. Tech. Sci.* 66(6), 831–840 (2018).

[12] S. Sivaranjini and C.M. Sujatha, "Deep learning based diagnosis of Parkinson's disease using convolutional neural network", *Multimed. Tools Appl.* 79, 15467–15479 (2020).

[13] M. Wodziński, A. Skalski, D. Hemmerling, J.R. Orozco-Arroyave, and E. Noth, "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification" *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 717–720.

[14] J. Chmielińska, K. Białek, A. Potulska-Chromik, J. Jakubowski, E. Majda-Zdancewicz, M. Nojszewska, A. Kostera-Pruszczyk and A. Dobrowolski, "Multimodal data acquisition set for objective assessment of Parkinson's disease", *Proc. SPIE 11442*, *Radioelectronic Systems Conference 2019*, 114420F (2020).

[15] M. Kuhn, K. Johnson, *Applied predictive modeling*, New York: Springer, 2013.

[16] P. Liang, C. Deng, J. Wu, Z. Yang, and J. Zhu, "Intelligent Fault Diagnosis of Rolling Element Bearing Based on Convolutional Neural Network and Frequency Spectrograms" *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, San Francisco, USA, 2019, pp. 1–5.

[17] M.S. Wibawa, I.M.D. Maysanjaya, N.K.D.P. Novianti, and P.N. Crisnapati, "Abnormal Heart Rhythm Detection Based on Spectrogram of Heart Sound using Convolutional Neural Network", *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Parapat, Indonesia, 2019, pp. 1–4.

[18] M. Curilem, J.P. Canário, L. Franco, and R.A. Rios, "Using CNN To Classify Spectrograms of Seismic Events From Llaima Volcano (Chile)", *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brasil, 2018, pp. 1–8.

[19] D. Rethage, J. Pons and X. Serra, "A Wavenet for Speech Denoising", *2018 IEEE Int. Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5069–5073.

[20] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classifi-cation with deep convolutional neural networks", *Neural Infor-mation Processing Systems*, 2012.

[21] J. Jakubowski and J. Chmielińska, "Detection of driver fatigue symptoms using transfer learning", *Bull. Pol. Acad. Sci. Tech. Sci.* 66(6), 869–874 (2018).

[22] A. Benba, A. Jilbab, and A. Hammouch, "Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ", *International conference on circuits*, *systems and signal processing (ICCSSP'14)*, Russia, 2014.

[23] E. Niebudek-Bogusz, J. Grygiel, P. Strumiłło, and M. Śliwińska-Kowalska, "Nonlinear acoustic analysis in the evaluation of occupational voice disorders", *Occupational Medicine*, 64(1), 29–35 (2013), [in Polish].

[24] E. Majda and A.P. Dobrowolski, “Modeling and optimization of the feature generator for speaker recognition systems”, *Electr. Rev.* 88(12A), 131–136 (2012).

[25] Y. Maryn, N. Roy, M. De Bodt, P.B. van Cauwenberge, P. Corthals, “Acoustic measurement of overall voice quality: a meta-analysis”, *J. Acoust. Soc. Am.* 126(5), 2619–2634 (2009), doi: 10.1121/1.3224706.

[26] E. Niebudek-Bogusz, J. Grygiel, P. Strumiłło, and M. Śliwińska-Kowalska, “Mel cepstral analysis of voice in patients with vocal nodules”, *Otorhinolaryngology* 10(4), 176–181 (2011), [in Polish].

[27] A. Krysiak, “Language, speech and communication disorders in Parkinson’s disease”, *Neuropsychiatr. Neuropsychol.* 6(1), 36–42 (2011), [in Polish].

[28] F. Alías, J.C. Socoró, and X. Sevillano, “A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds”, *Appl. Sci.* 6(5), 143 (2016).

[29] X. Valero and F. Alias, “Gammatone Cepstral Coefficients: Biologically InspiredFeatures for Non-Speech Audio Classification”, *IEEE Trans. Multimedia* 14(6), 1684–1689 (2012).

[30] S. Malcolm, “An Efficient Implementation of the Patterson-Holdworth Auditory Filter Bank”, 35. *Apple Computer Technical Report*, 1993.

[31] D.M. Agrawal, H.B. Sailor, M.H. Soni, and H.A. Patil, “Novel TEO-based gammatone features for environmental sound classification”, *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017, pp. 1809–1813.

[32] S. Russel and P. Norvig, *Artificial intelligence – a modern approach*, Upper Saddle River: Pearson Education, 2010.

[33] A. Chatzimparmpas, R.M. Martins, K. Kucher, and A. Kerren, “StackGenVis: Alignment of Data, Algorithms, and Models for Stacking Ensemble Learning Using Performance Metrics”, *IEEE Transactions on Visualization and Computer Graphics* 27(2), 1547–1557 (2021), doi: 10.1109/TVCG.2020.3030352.